Causal Effect Identification in LiNGAM Models with Latent Confounders

Daniele Tramontano [1] **Yaroslav Kivva** [2] **Saber Salehkaleybar** [3] **Mathias Drton** [1 4] **Negar Kiyavash** [2]

Abstract

We study the generic identifiability of causal effects in linear non-Gaussian acyclic models (LiNGAM) with latent variables. We consider the problem in two main settings: When the causal graph is known a priori, and when it is unknown. In both settings, we provide a complete graphical characterization of the identifiable direct or total causal effects among observed variables. Moreover, we propose efficient algorithms to certify the graphical conditions. Finally, we propose an adaptation of the reconstruction independent component analysis (RICA) algorithm that estimates the causal effects from the observational data given the causal graph. Experimental results show the effectiveness of the proposed method in estimating the causal effects.

1. Introduction

Predicting the impact of an unseen intervention in a system is a crucial challenge in many fields, such as medicine (Sanchez et al., 2022; Michoel & Zhang, 2023), policy evaluation (Athey & Imbens, 2017), fair decision-making (Kilbertus et al., 2017), and finance (de Prado, 2023). Randomized experiments form the gold standard for addressing this challenge, but are often infeasible due to ethical concerns or prohibitively high costs. To tackle situations in which only observational data are available, one needs to make additional assumptions on the underlying causal system. The field of causal inference strives to formalize such assumptions. One notable approach in causal inference is modeling causal relationships through structural equation models (SEM) (Pearl, 2009). In this framework, a random vector is associated with a directed acyclic graph (DAG). Each vector component is associated with a node in the graph and is a function of the random variables corresponding to its parents in the graph and its corresponding exogenous noise.

In this paper, we mainly focus on the identifiability of causal effects in the important subclass of linear SEM and characterize graphically which causal effects can be uniquely determined from the observational data. When the exogenous noises in a linear SEM are Gaussian, the entire information about the model is contained in the covariance matrix among the variables, with all the higher-order moments of the distribution being uninformative. This entails that the causal structure as well as other causal quantities are often not identifiable from mere observational data. The most prominent instance in the context of causal structure learning is that the causal graph is identifiable only up to an equivalence class (e.g., Drton, 2018, §10). This motivated the widespread use of the linear non-Gaussian additive noise model (LiNGAM), where the exogenous noises are non-Gaussian.

The seminal work of Shimizu et al. (2006) showed that in the setting of LiNGAM, the true underlying causal graph is uniquely identifiable when all the variables are observed. Since then, a rich literature on this topic has emerged, focusing mainly on the identification and the estimation of the causal graph; see, e.g., Adams et al. (2021); Shimizu (2022); Yang et al. (2022); Wang et al. (2023); Wang & Drton (2023) for recent results that allow for the presence of hidden variables. Indeed, linear models remain the backbone of problem abstraction in many scientific disciplines, because they offer simple qualitative interpretations and can be learned with moderate sample sizes (Pe'er & Hacohen, 2011, Principle 1). In particular, the LiNGAM model finds application in diverse scientific fields, such as Neuroscience (Chiyohara et al., 2023), Economics (Ciarli et al., 2023), or Epidemiology (Barrera & Miljkovic, 2022).

Within the LiNGAM literature, causal effect identification has received less attention; only few recent work (Kivva et al., 2023; Shuai et al., 2023) have exploited the non-Gaussianity to provide identification formulas that work for specific causal graphs. The only graphical criteria for identification are given in Salehkaleybar et al. (2020); Yang et al. (2022). The main drawback of the aforementioned papers is that they target simultaneous recovery of all the causal effects. However, in many applications, we are interested in causal effects of only some subset of variables on others.

[1]Technical University of Munich, Munich, Germany [2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland [3]Leiden Institute of Advanced Computer Science, Leiden University, Netherlands [4]Munich Center for Machine Learning, Munich, Germany. Correspondence to: Daniele Tramontano <daniele.tramontano@tum.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Indeed, it may be the case that some causal effects are identifiable while others are not; see Figure 2 for an example in the context of proxy variable graphs. In this paper, we provide necessary and sufficient graphical conditions for the generic identifiability (see Section 2.3 for the exact definition) of direct and total causal effects between a given pair of observed variables in a LiNGAM.

1.1. Contribution

Our main contributions are as follows:

- We provide necessary and sufficient graphical criteria for the generic identifiability of the causal effect, both when the causal graph is known a priori (Section 3.1) and when it is unknown (Section 3.2).

- We propose sound and complete algorithms that check our criteria in polynomial time in the size of the graph, in both considered settings (Section 3.4).

- For practical estimation of the effect of interest, we propose an adaptation of the RICA algorithm for Independent Component Analysis (Le et al., 2011). Experimental results show that the proposed method can provide better estimates of causal effects when compared with previous work (Section 3.5).

2. Problem Definition

2.1. Notation

A *directed graph* is a pair $\mathcal{G} = (\mathcal{V}, E)$ where $\mathcal{V} := \{1, \dots, p\}$ is the set of nodes and $E \subseteq \{(i,j) \mid i,j \in \mathcal{V},\ i \neq j\}$ is the set of edges. We denote a pair $(i,j) \in E$ as $i \to j$.

A (directed) path from node i to node j in $\mathcal{G}$ is a sequence of nodes $\pi = (i_1 = i, \dots, i_{k+1} = j)$ such that $i_s \to i_{s+1} \in E$ for $s \in \{1, \dots, k\}$. A cycle in $\mathcal{G}$ is a path from a node i to itself. A *Directed Acyclic Graph* (DAG) is a directed graph without cycles. If $i \to j \in E$, we say that i is a parent of j, and j is a child of i. If there is a path from i to j in $\mathcal{G}$, we say that i is an ancestor of j and j is a descendant of i. The sets of parents, children, ancestors, and descendants of a given node i are denoted by $\mathrm{pa}(i), \mathrm{ch}(i), \mathrm{an}(i)$, and $\mathrm{de}(i)$, respectively. In our work, we distinguish between observed and latent variables by partitioning the nodes in two sets $\mathcal{V} = \mathcal{O} \cup \mathcal{L}$, of respective sizes p_o and p_l. Moreover, we define the set of observed descendants of a node i as $\mathrm{de}_o(i) := (\mathrm{de}(i) \cup \{i\}) \cap \mathcal{O}$.

We write vectors and matrices in boldface. The entry (i,j) of a matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{i,j}$. Let I, J be subsets of the row and column sets of $\mathbf{A}$, respectively. We denote the submatrix containing only the rows in I and the columns in J as $[\mathbf{A}]_{I,J}$. For a permutation σ, $\mathbf{P}_\sigma$ denotes the associated permutation matrix.

2.2. Model

Let $\mathcal{G} = (\mathcal{V}, E)$ be a *fixed* DAG on p nodes. In a fixed probability space, let $\mathbf{V} = (V_0, \dots, V_p)$ be a random vector taking values in $\mathbb{R}^p$ and satisfying the following structural equation model:

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

where $[\mathbf{A}]_{j,i} = 0$ if $i \to j \notin E$, $\mathbf{B} := (\mathbf{I} - \mathbf{A})^{-1}$, and the enteries of the exogenous noise vector $\mathbf{N}$ are assumed to be jointly independent and *non-Gaussian*. $\mathbf{V}$ is partitioned to $[\mathbf{V}_o, \mathbf{V}_l]$, where $\mathbf{V}_o$ is observed of dimension p_o, while $\mathbf{V}_l$ is latent and of dimension p_l. We can rewrite (1) as

$$\begin{bmatrix} \mathbf{V}_o \\ \mathbf{V}_l \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{o,o} & \mathbf{A}_{o,l} \\ \mathbf{A}_{l,o} & \mathbf{A}_{l,l} \end{bmatrix} \begin{bmatrix} \mathbf{V}_o \\ \mathbf{V}_l \end{bmatrix} + \begin{bmatrix} \mathbf{N}_o \\ \mathbf{N}_l \end{bmatrix},$$

which implies that the observed random vector satisfies

$$\mathbf{V}_o = \mathbf{B}'\mathbf{N} = \begin{bmatrix} \mathbf{B}_o & \mathbf{B}_l \end{bmatrix} \begin{bmatrix} \mathbf{N}_o \\ \mathbf{N}_l \end{bmatrix} \tag{2}$$

with $\mathbf{B}' := [(\mathbf{I} - \mathbf{A})^{-1}]_{\mathcal{O},\mathcal{V}}$. We refer to this model as the latent variable LiNGAM (lvLiNGAM).[1]

Salehkaleybar et al. (2020, §3) show that the matrix $\mathbf{B}'$ can be expressed as follows:

$$\mathbf{B}_o = (\mathbf{I} - \mathbf{D})^{-1}, \quad \mathbf{B}_l = (\mathbf{I} - \mathbf{D})^{-1}\mathbf{A}_{o,l}(\mathbf{I} - \mathbf{A}_{l,l})^{-1}, \tag{3}$$

with $\mathbf{D} = \mathbf{A}_{o,o} + \mathbf{A}_{o,l}(\mathbf{I} - \mathbf{A}_{l,l})^{-1}\mathbf{A}_{l,o}$. The matrices $\mathbf{B}'$ and $\mathbf{D}$ contain information on the interventional distributions of $\mathbf{V}_o$. In particular,[2]

$$\begin{aligned} [\mathbf{B}']_{i,j} &= \frac{\partial \mathbb{E}(V_i \mid \mathrm{do}(V_j))}{\partial V_j}, \\ [\mathbf{D}]_{i,j} &= \frac{\partial \mathbb{E}(V_i \mid \mathrm{do}(V_{\mathrm{pa}(i)}))}{\partial V_j}. \end{aligned} \tag{4}$$

In other words, $[\mathbf{B}']_{j,i}$ is the average total causal effect of j on i, while $[\mathbf{D}]_{j,i}$ is the average causal effect of j on i that is not mediated by other observed nodes. With slight abuse of terminology, we refer to the entries of $\mathbf{B}'$ as *total causal effects* and to those of $\mathbf{D}$ as *direct causal effects*.

Hoyer et al. (2008) show that for any lvLiNGAM model, an associated *canonical model* exists, in which, in the corresponding graph, all the latent nodes have at least two

[1] Although our results are presented for the lvLiNGAM model, our analysis only relies on the identifiability of the mixing matrix $\mathbf{B}'$ up to permutation and scaling. This can be also achieved in other settings as explained in Yang et al. (2022, §2.3) or Adams et al. (2021, §3). Hence, our results also hold in these settings.

[2] Please see Pearl (2009, §3) for the definition of *do* intervention.

children and have no parents. We refer to the graph corresponding to a canonical model as a canonical graph. The original and the associated canonical model are *observationally* and *causally* equivalent (Hoyer et al., 2008, §3). Subsequently, without loss of generality, we will assume our model is canonical in this sense. Salehkaleybar et al. (2020, Cor. 11) proved that the number of latent variables is identifiable in canonical models.

Remark 2.1. Throughout the paper, we assume that the number of latent variables p_l is known.

In canonical models, $\mathbf{A}_{l,o} = \mathbf{A}_{l,l} = \mathbf{0}$, and in particular

$$\mathbf{B}_o = (\mathbf{I} - \mathbf{A}_{o,o})^{-1}, \qquad \mathbf{B}_l = (\mathbf{I} - \mathbf{A}_{o,o})^{-1}\mathbf{A}_{o,l}. \quad (5)$$

For every canonical $\mathcal{G}$, let $\mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$ be the set of all $p \times p$ real matrices $\mathbf{A}$ such that $[\mathbf{A}]_{i,j} = 0$ if $j \to i \notin \mathcal{G}$. Let $\mathbb{R}^{\mathcal{G}}$ be the set of all $p_o \times p$ matrices, $\mathbf{B}' = [\mathbf{B}_o, \mathbf{B}_l]$ that can be obtained from a matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$ from (5). Let NG^p be the set of p dimensional, non-degenerate, jointly independent *non-Gaussian* random vectors, and let $\mathcal{M}(\mathcal{G})$ be the set of all p_o dimensional random vectors that can be expressed according to (2), where the matrix $\mathbf{B}' \in \mathbb{R}^{\mathcal{G}}$.

2.3. Identifiability

According to (2), the mechanism generating the observational distribution, i.e., the probability distribution of $\mathbf{V}_o$, only depends on the matrix $\mathbf{B}'$ and the exogenous noise vector $\mathbf{N}$. Therefore, the parameters of interest such as the causal effects in (4), are functions $\phi(\mathbf{B}')$. As generally there are multiple pairs $(\mathbf{B}', \mathbf{N})$, with different $\mathbf{B}'$s, that generate the same observational distribution, it is important to clarify whether a parameter ϕ is *identifiable*. This holds if $\phi(\mathbf{B}')$ takes the same value in all considered pairs $(\mathbf{B}', \mathbf{N})$ generating the same observational distribution. Adding further assumptions about the causal graph may limit the data-generating mechanisms compatible with the observational data, leading to more parameters becoming identifiable.

In this paper, we study the identification of causal effects in two scenarios: when the causal graph $\mathcal{G}$ is known and when it is not. The remainder of the section provides a formal description of the resulting parameter identification problems in the lvLiNGAM setting.

Fix a DAG $\mathcal{G}$. The observational random vector is obtained from the noise and the matrix $\mathbf{B}'$ according to the mapping

$$\begin{aligned} \Phi_{\mathcal{G}} : \mathbb{R}^{\mathcal{G}} \times \mathrm{NG}^p &\to \mathcal{M}(\mathcal{G}) \\ (\mathbf{B}', \mathbf{N}) &\mapsto \mathbf{B}'\mathbf{N} = \begin{bmatrix} \mathbf{B}_o & \mathbf{B}_l \end{bmatrix} \begin{bmatrix} \mathbf{N}_o \\ \mathbf{N}_l \end{bmatrix}. \end{aligned} \quad (6)$$

Let ϕ be a real-valued function on $\mathbb{R}^{\mathcal{G}}$. We say that the parameter ϕ is *globally identifiable without knowledge of the graph* if for every pair $(\mathbf{B}', \mathbf{N}) \in \mathbb{R}^{\mathcal{G}} \times \mathrm{NG}^p$ and the associated observed random vector $\mathbf{V}_o := \Phi_{\mathcal{G}}(\mathbf{B}', \mathbf{N}) \in \mathcal{M}(\mathcal{G})$, there does not exist a DAG $\tilde{\mathcal{G}}$ and a pair $(\tilde{\mathbf{B}}', \tilde{\mathbf{N}}) \in \mathbb{R}^{\tilde{\mathcal{G}}} \times \mathrm{NG}^p$ such that

$$\Phi_{\tilde{\mathcal{G}}}(\tilde{\mathbf{B}}', \tilde{\mathbf{N}}) =_d \mathbf{V}_o \ \text{ but } \ \phi(\mathbf{B}') \neq \phi(\tilde{\mathbf{B}}'). \quad (7)$$

Here, $=_d$ denotes equality in distribution of two random vectors. Note that the above definition allows for consideration of any $\tilde{\mathcal{G}}$, possibly, different from $\mathcal{G}$.

In contrast, we say that ϕ is *globally identifiable with knowledge of the graph* if for every pair $(\mathbf{B}', \mathbf{N}) \in \mathbb{R}^{\mathcal{G}} \times \mathrm{NG}^p$ and the associated observed random vector $\mathbf{V}_o := \Phi_{\mathcal{G}}(\mathbf{B}', \mathbf{N}) \in \mathcal{M}(\mathcal{G})$, there does not exist another pair $(\tilde{\mathbf{B}}', \tilde{\mathbf{N}}) \in \mathbb{R}^{\mathcal{G}} \times \mathrm{NG}^p$ with

$$\Phi_{\mathcal{G}}(\tilde{\mathbf{B}}', \tilde{\mathbf{N}}) =_d \mathbf{V}_o \ \text{ but } \ \phi(\mathbf{B}') \neq \phi(\tilde{\mathbf{B}}'). \quad (8)$$

In latent variable models such as lvLiNGAM, global identifiability, as defined above, is often too stringent of a condition and fails to apply even in many commonly used models such as the instrumental variable model. Thus, a weaker notion of so-called generic identifiability is often considered for linear models (Maathuis et al., 2019, §16.4). We say that the parameter ϕ is *generically* identifiable with (or without) knowledge of the graph, if the condition in (8) (or (7)) holds for every $(\mathbf{B}', \mathbf{N}) \in (\mathbb{R}^{\mathcal{G}} \setminus \mathcal{B}_{\text{except}}) \times \mathrm{NG}^p$ where $\mathcal{B}_{\text{except}}$ is a Lebesgue measure zero subset of $\mathbb{R}^{\mathcal{G}}$.

Our results give necessary and sufficient conditions for generic identifiability of the parameters $[\mathbf{A}_{o,o}]_{i,j}$ and $[\mathbf{B}_o]_{i,j}$, both with and without knowledge of the underlying graph $\mathcal{G}$.

3. Main Results on Causal Effect Identification

This section presents our main identifiability results for lvLiNGAM. In Sections 3.2 and 3.1, we characterize the identifiable causal effects, respectively, with or without knowledge of the graph. In Section 3.3, we provide several examples in which our identifiability criteria hold; in Section 3.4, we propose an algorithm to certify our criteria efficiently in time. Finally, in Section 3.5, we present an estimation algorithm to estimate the causal effect from the observational data. All proofs appear in the Appendix B.

3.1. Identification with an Unknown Graph

In Salehkaleybar et al. (2020), it is shown that if the total causal effect between any ancestor-descendent pair is non-zero, the mixing matrix $\mathbf{B}'$ can be identified by means of overcomplete independent component analysis (ICA), up to a permutation of the columns corresponding to a pair $(j, l) \in \mathcal{O} \times \mathcal{L}$ such that $\mathrm{de}_o(j) = \mathrm{de}_o(l)$. This leads to a graphical criterion for the identification of the entire mixing matrix, which we rephrase herein using our notation.

Theorem 3.1. *(Salehkaleybar et al., 2020, Theo. 16) For $\mathbf{B}' \in \mathbb{R}^{\mathcal{G}}$, matrix $\mathbf{B}_o$ is* generically *identifiable without knowing the causal graph if and only if there are no pairs $j \neq i \in \mathcal{V}$ such that $\mathrm{de}_o(j) = \mathrm{de}_o(i)$.*

Remark 3.2 (The scaling matrix). Equation (5) implies that as long as we are focused on identifying the causal effect between observed variables alone, the scaling of the latent columns does not make a difference. Hence, without loss of generality, in the sequel, we assume that all the mixing matrices are scaled in such a way that the first non-zero entry in each column is equal to 1. In other words, $\mathbf{A}_{il} = 1$ if i and l are, respectively, observed and latent variables and i is the first child of l in a given causal order. Note that this is always the case if $\mathrm{de}_o(l) = \mathrm{de}_o(i)$.

In this section, we extend the result in Theorem 3.1 by providing *necessary and sufficient* graphical conditions for the *generic* identifiability for the entries of $\mathbf{B}_o$ and $\mathbf{A}_{o,o}$.

Our first result refines the graphical condition of Theorem 3.1 by adding (10). This provides a complete graphical characterization of the identifiable total causal effects.

Theorem 3.3 (Total causal effect)**.** *Consider any two observed variables i and j. The* total causal effect *of j on i is* generically *identifiable without knowledge of the graph* if and only if *there is no $l \in \mathcal{L}$, such that*

$$\mathrm{de}_o(j) = \mathrm{de}_o(l), \text{ and} \tag{9}$$

$$i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l), \tag{10}$$

where $\mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$ is the set of observed descendents of l in the graph obtained from $\mathcal{G}$ by removing all the edges pointing to j.

The next result provides a complete graphical characterization of the identifiable direct causal effects.

Theorem 3.4 (Direct causal effect)**.** *Consider any two observed variables i and j. The direct causal effect of j on i is* generically *identifiable without knowledge of the graph if and only if there are no pairs $(k, l) \in \mathcal{O} \times \mathcal{L}$ such that*

$$\mathrm{de}_o(k) = \mathrm{de}_o(l), \tag{11}$$

$$i \in \mathrm{ch}(l), \tag{12}$$

$$k \in \mathrm{ch}(j) \cup \{j\}. \tag{13}$$

Example 3.5 (Identification with instrumental variables)**.** *The mixing matrix for the instrumental variable (IV) graph (Cunningham, 2021, §7.1) in Figure 1 has the following form*

$$\mathbf{B}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ b_{TI} & 1 & 1 & 0 \\ b_{TI}b_{YT} & b_{YT} & b_{YT} + b_{LY} & 1 \end{bmatrix}.$$

The parameter of interest in the IV graph of Figure 1 is b_{YT}, i.e., the causal effect of the treatment T on the outcome Y. Since $\mathrm{de}_o(T) = \mathrm{de}_o(L) = \{T, Y\}$, we can consider

$$\tilde{\mathbf{B}}' = \mathbf{B}' \cdot \mathbf{P}_\sigma, \quad \tilde{\mathbf{N}} = \mathbf{P}_\sigma^{-1} \cdot \mathbf{N},$$

where σ is the transposition that permutes the columns corresponding to T and L. From (6), one can see that

$$\Phi_{\tilde{\mathcal{G}}_{IV}}(\tilde{\mathbf{B}}', \tilde{\mathbf{N}}) =_d \Phi_{\mathcal{G}_{IV}}(\mathbf{B}', \mathbf{N}), \qquad \forall \mathbf{N} \in \mathrm{NG}^p,$$

where $\tilde{\mathcal{G}}_{IV}$ is the graph obtained from the IV graph after adding an edge from I to Y (see Figure 12 in Appendix B).

Since $[\tilde{\mathbf{B}}']_{YT} \neq [\mathbf{B}']_{YT}$ the total causal effect of T on Y is not identifiable without knowing the true causal graph.

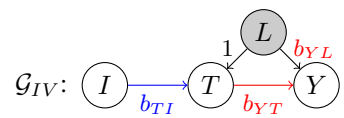


Figure 1. Instrumental variable graph. The parameters in blue are identifiable without knowledge of the graph, while the parameters in red are not identifiable.

Notice that we could derive this result by applying Theorem 3.4 to the pair $(k, l) = (T, Y)$ directly.

3.2. Identification with a Known Graph

A permutation of two columns of the matrix $\mathbf{B}'$ may result in different graphs. This implies that if the graph is known, we can narrow down the set of possible permutations. In this section, we study how this additional assumption allows us to identify a parameter of interest (which is not generically identifiable). We will discuss an instance of this situation in Example 3.7.

The first result of this section provides a characterization of the column permutations that leave the graph unchanged.

Theorem 3.6. *For every $\mathbf{B}'$ outside a Lebesgue zero subset of $\mathbb{R}^{\mathcal{G}}$, let $\tilde{\mathbf{B}}' = \mathbf{B}' \cdot \mathbf{P}_\sigma$, where σ is any permutation. We have $\tilde{\mathbf{B}}' \in \mathbb{R}^{\mathcal{G}}$, if and only if*

$$\mathrm{de}_o(i) = \mathrm{de}_o(\sigma(i)), \qquad \forall i \in \mathcal{V},$$

and there are no $i, j \in \mathcal{O}$ and $l \in \mathcal{L}$ such that

$$i \in \mathrm{pa}(j) \cup \{j\}, \quad \sigma(j) = l, \quad \mathrm{ch}(l) \setminus \mathrm{ch}(i) \neq \emptyset.$$

Example 3.7 (Example 3.5 continued)**.** *We now show that the causal effect of T on Y is identifiable if we assume the graph in Figure 1 is the true underlying graph.*

The adjacency matrix corresponding to $\tilde{\mathbf{B}}'$ has the following form

$$\tilde{\mathbf{A}}_{o,o} = \mathbf{I} - \tilde{\mathbf{B}}_o^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ b_{TI} & 1 & 0 \\ -b_{TI}b_{YL} & b_{YT} + b_{YL} & 1 \end{bmatrix}.$$

This form is not compatible with the graph in Figure 1 since the entry corresponding to the edge from I to Y is nonzero. Hence, if we were told that the graph in Figure 1 is the true underlying causal structure, the only valid permutation of $\mathbf{B}'$ would be the identity. Therefore, we can identify b_{YT}, the parameter of our interest.

Notice that we could derive this result by applying Theorem 3.6 to the triple $(i, j, l) = (I, T, L)$ directly. See Example B.3 for more discussion of the IV graph.

Using Theorem 3.6, we can refine the criteria of Section 3.1 for the setting of known causal graph. In particular, the next two theorems characterize all the total and direct causal effects that are identifiable when the graph is known.

Theorem 3.8 (Total Causal Effect)**.** *Consider any two observed variables i and j. The* total causal effect *of j on i is* generically *identifiable with knowledge of the graph* if and only if *there is no $l \in \mathcal{L}$, such that*

$$\mathrm{de}_o(j) = \mathrm{de}_o(l), \tag{14}$$

$$i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l), \tag{15}$$

$$\mathrm{ch}(l) \setminus \mathrm{ch}(k) = \emptyset, \qquad \forall k \in \mathrm{pa}(j) \cup \{j\}. \tag{16}$$

Theorem 3.9 (Direct Causal Effect)**.** *Consider any two observed variables i and j. The direct causal effect of j on i is* generically *identifiable if and only if there are no pairs $(k, l) \in \mathcal{O} \times \mathcal{L}$ such that*

$$\mathrm{de}_o(k) = \mathrm{de}_o(l), \tag{17}$$

$$i \in \mathrm{ch}(l), \tag{18}$$

$$k \in \mathrm{ch}(j) \cup \{j\}, \tag{19}$$

$$\mathrm{ch}(l) \setminus \mathrm{ch}(k_1) = \emptyset, \qquad \forall k_1 \in \mathrm{pa}(k) \cup \{k\}. \tag{20}$$

We conclude the section by giving the graphical condition for identification of the whole mixing matrix with knowledge of the graph.

Corollary 3.10 (Mixing matrix identification)**.** *For $\mathbf{B}' \in \mathbb{R}^{\mathcal{G}}$, the entire matrix $\mathbf{B}_o$ is* generically *identifiable with knowledge of the graph if and only if there are no $i, j \in \mathcal{O}$, and $l \in \mathcal{L}$ such that* (14)*,* (15)*, and* (16) *are satisfied.*

3.3. Examples

We now highlight different scenarios in which the results in this section allow us to relax standard assumptions in identifying the causal effect. For the sake of simplicity in presentation, in the first two scenarios, we only consider two latent confounders in the system. Although these scenarios can be easily extended to any arbitrary number of latent confounders.

1. **Proxy Variables**

The presence of proxy variables allow the identification of the causal effect in linear models. In particular, it has been shown that the causal effect of treatment T on outcome Y is identifiable if the following two conditions hold (see, e.g., Kuroki & Pearl (2014, §4), Liu et al. (2023, §2)):

(a) There are as many proxies as there are confounders,

(b) $W \perp\!\!\!\perp (T, Y) \mid L$ for every proxy variable W and every latent confounder L.

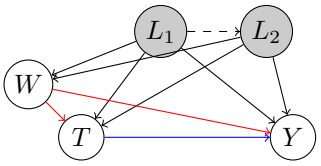


Figure 2. Causal graph with proxy variable. The parameter corresponding to the blue edge is generically identifiable if the graph is known, while those corresponding to the red edges are not. The dashed edge is dropped in the corresponding canonical model.

Consider the causal graph in Figure 2. Both aforementioned conditions are violated as there are two latent confounders but a single proxy and there is an edge from this proxy to the treatment. Yet, in lvLiNGAM, we show that the causal effect from T to Y is identifiable. For both the latent variables we have $W \in \mathrm{de}_o(L) \setminus \mathrm{de}_o(T)$, hence condition (14) cannot be satisfied. Using Theorem 3.8, this implies that the causal effect of our interest is identifiable.

2. **Longitudinal Data**

The causal graph associated with a longitudinal data model (Imai & Kim, 2019) is given in the left plot in Figure 3, where L is the latent confounder, T is the time-varying treatment, and Y is the time-varying outcome. The common identifiability assumptions in a linear setting are that the causal effect is constant through time, i.e., $[\mathbf{A}]_{Y_1,T_1} = [\mathbf{A}]_{Y_2,T_2}$ and there is no time-varying confounding, see, e.g., Cunningham (2021, §8). In lvLiNGAM, these assumptions can be relaxed by having access to some covariates. In particular, suppose that there is a covariate C_i such that $\mathrm{de}_o(L_i) = \mathrm{de}_o(C_i)$ for every time period i (see the right plot in Figure 3). We have $C_i \in \mathrm{de}_o(L_i) \setminus \mathrm{de}_o(T_i)$ and the condition (14) is not satisfied. This implies that the causal effect of T_i on Y_i is identifiable for every time period i.

3. **Underspecified Instruments**

A standard assumption for identifying the causal effect using instrumental variables is that there are at least as many instruments as treatments (Brito & Pearl, 2002, Thm. 1). Recently, Ailer et al. (2023) showed that in the linear underspecified setting, i.e., when the number of instruments is less than the number of treatments, one can identify the projection of the treatment on the instrument space, but this can be different from the causal effect. We now show that in the lvLiNGAM model, the treatment effects are identifiable in the underspecified case.

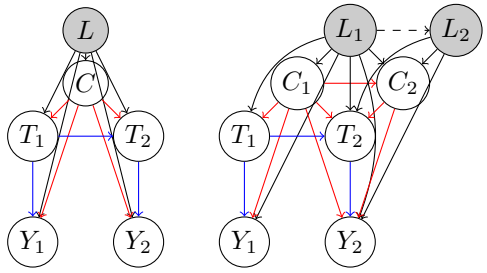


Figure 3. The causal graphs for longitudinal data. The graph on the right allows for a time-varying confounder. The parameters corresponding to the blue edges are generically identifiable with the knowledge of the graph, while the ones corresponding to the red edges are not. The dashed line is dropped in the corresponding canonical model.

We say that I is a valid instrument for the treatments $T_1, \dots, T_n$ on Y if the following conditions hold

$$I \in \mathrm{pa}(T_i) \quad \forall i \in \{1, \dots, n\},$$
$$I \perp_{\mathcal{G}_{\setminus T}} Y,$$

where $\perp$ denotes d-separation (Pearl, 2009, §1.2), and $\mathcal{G}_{\setminus T}$ is the graph obtained from $\mathcal{G}$ by removing all the edges from T_i to Y. See Figure 4 for an instance with two treatments and one instrument.

For every latent variable L_i such that $\mathrm{de}_o(T_i) = \mathrm{de}_o(L_i)$ and $Y \in \mathrm{ch}(L_i)$, we have $Y \in \mathrm{ch}(L_i) \setminus \mathrm{ch}(I)$. Hence condition (16) does not hold. By Theorem 3.8, in lvLiNGAM, the causal effect of T_i on Y is identifiable for every i, even when only one instrumental variable is available.

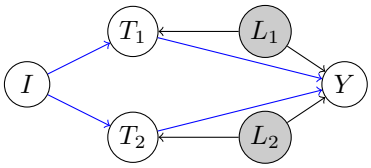


Figure 4. An example of causal graph for underspecified instrumental variable. The parameters corresponding to the blue edges are generically identifiable with knowledge of the graph.

3.4. Certifying Identifiability

In this section, we prove that the graphical condition for identification given in Theorem 3.8 can be certified in polynomial time in the size of the graph, and we provide a sound and complete algorithm (Algorithm 1) for this task.

The algorithms for the other identification criteria can be found in Appendix B.3.

Theorem 3.11. *Algorithm 1 is sound and complete for certifying the generic identifiability of the total causal effect of j on i with knowledge of the graph $\mathcal{G}$. The computational complexity of the algorithm is $\mathcal{O}(p_l(p_o^2 + |E|)) = \mathcal{O}(p^3)$.*

Algorithm 1 Total Causal Effect Identification with Known Graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \mathcal{G}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}, (j, i)$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: Compute $\mathrm{de}_o(j)$
4: **while** ID $==$ TRUE **and** $|\mathcal{L}| > 0$ **do**
5: $\quad l \leftarrow \mathcal{L}[1]$ {The first element in the list}
6: $\quad$ Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
7: $\quad$ **if** $\mathrm{ch}(l)[1] = j$ **then**
8: $\quad\quad$ Compute $\mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$
9: $\quad\quad$ **if** $i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$ **then**
10: $\quad\quad\quad$ Compute $\mathrm{de}_o(l)$
11: $\quad\quad\quad$ **if** $\mathrm{de}_o(l) = \mathrm{de}_o(j)$ **then** {(15)}
12: $\quad\quad\quad\quad$ ID $\leftarrow$ FALSE {(14)}
13: $\quad\quad\quad\quad$ Compute $\mathrm{pa}(j)$
14: $\quad\quad\quad\quad$ **for all** $k \in \mathrm{pa}(j) \cup \{j\}$ **do**
15: $\quad\quad\quad\quad\quad$ **if** $\mathrm{ch}(l) \setminus \mathrm{ch}(k) \neq \emptyset$ **then**
16: $\quad\quad\quad\quad\quad\quad$ ID $\leftarrow$ TRUE {(16)}
17: $\quad \mathcal{L} \leftarrow \mathcal{L} \setminus \{l\}$
18: **RETURN:** ID

Remark 3.12. Algorithm 1 is simple in the sense that it directly checks the graphical conditions in Theorem 3.8. This is not the case for checking most identifiability results in linear models, which often requires building an auxiliary graph and solving a maximum-flow problem on it (which becomes prohibitive for large graphs), (Brito & Pearl, 2006; Kumor et al., 2020; Barber et al., 2022).

3.5. Estimation Algorithms

When the graph structure is unknown, Salehkaleybar et al. (2020, Alg. 1) proposed an algorithm that first solves an overcomplete ICA problem and then post-process the estimated mixing matrix to enumerate all the possible causal effects. This usually entails solving a high-dimensional non-convex optimization problem. If the DAG structure is known, one can enforce this knowledge to reduce the problem's dimensionality from p^2 to $|E|$ and improve the performance. We follow this approach and propose an adaptation of the RICA algorithm for recovering the mixing matrix (Le et al., 2011).

The objective function optimized by RICA is a weighted sum of two terms; the first term is a contrast function that measures the non-Gaussianity of the exogenous noise, e.g., the l_1-loss, and the second term is a reconstruction loss that enforces the orthonormality of the rows of the mixing matrix. The only instance in which the rows of a matrix in $\mathbb{R}^{\mathcal{G}}$ might be orthonormal is when all the causal effects are zero. Hence, we drop the reconstruction loss and only optimize the contrast function.

For a given DAG $\mathcal{G}$, given observed data $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}^{p_o}$, and a contrast function g, our algorithm solves the following optimization problem

$$\textbf{Graphical RICA:} \arg\min_{\mathbf{B}' \in \mathbb{R}^{\mathcal{G}}} \frac{1}{N} \sum_{i=1}^{N} g((\mathbf{B}')^T \cdot \mathbf{X}_i).$$

We evaluate the performance of our algorithm in comparison with existing methods in Section 5.2.

4. Related Work

There is a rich literature on graphical criteria for the identifiability of causal effects. In the nonparametric setting, the ID algorithm is a sound and complete algorithm that solves the global identification problem given the causal graph (Shpitser & Pearl, 2006; Shpitser, 2023).

In the parametric case, most results are for the semi-Markovian and linear Gaussian models; for these models, a necessary and sufficient criterion for global identifiability is known, (Drton et al., 2011), while a complete characterization for generic identifiability remains unknown, (Kumor et al., 2020). For Gaussian models with explicit linear confounders, a sufficient graphical criterion for generic identifiability was proposed in Barber et al. (2022).

For linear non-Gaussian models, Salehkaleybar et al. (2020) proposed necessary and sufficient graphical criterion for the identifiability of the whole mixing matrix, Yang et al. (2022) defines a notion of equivalence class for lvLiNGAM models, and Cai et al. (2023) provide sufficient graphical conditions for the identification of the mixing matrix using explicit moment equations. Kivva et al. (2023) proposed an identification formula that works for the causal graph with one proxy variable; in contrast, Shuai et al. (2023) proved that if one assumes that only the treatment is non-Gaussian and pre-treatment covariates are available, then the causal effect can be identified in the presence of latent confounders. Tramontano et al. (2024) proposed a necessary and sufficient graphical criterion for the identifiability of the direct causal effect in linear models for acyclic-directed mixed graphs.

5. Experimental Results[3]

5.1. Identification

We used Algorithms 1 and 2 to check the identifiability of a causal effect for randomly selected edges in random graphs. The graphs are generated according to an Erdős-Rényi model in which we ensure that the sampled graphs are canonical. The probability of the causal effect of a randomly selected edge being identifiable versus the probability of accepting an edge in the graph generation model is plotted in Figure 5. For each setup, we randomly sample 500 graphs.

Interestingly, we found that for all the graphs and all the edges we sampled, the corresponding causal effects are identifiable with the knowledge of the graph. As expected, when we do not assume the graph is known, the probability of identifying the causal effect of randomly selected edge drops and it depends both on the density of the graph and the proportion of observed nodes (see Figure 5).

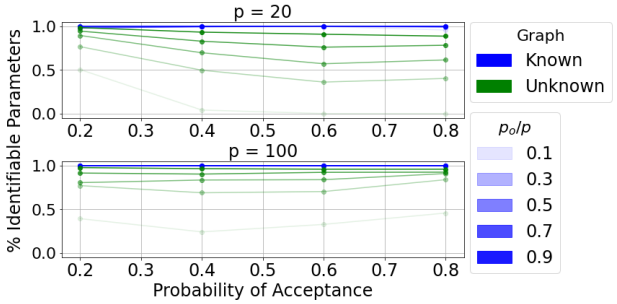


Figure 5. On the x-axis, the probability of acceptance of an edge. On the y-axis, the percentage of identifiable parameters.

The average run time of Algorithm 1 for different graph sizes is shown in Figure 6. It is noteworthy that our algorithm can handle graphs with a thousand nodes in about a second; this is due to the simplicity of our identification criteria, as explained in Remark 3.12.

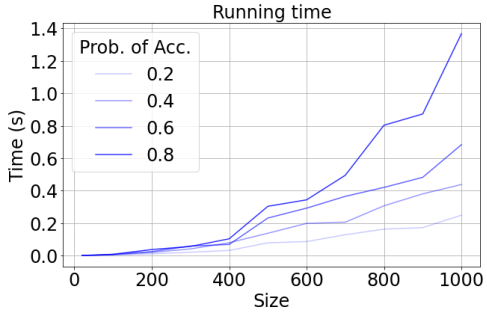


Figure 6. On the x-axis, the size of the graph. On the y-axis, the average running time in seconds. p_o/p is fixed to 0.5.

5.2. Causal Effect Estimation

In this section, we provide experimental results for Graphical RICA (GRICA), which we introduced in Section 3.5. We present experiments for data generated according to the causal structures in Figure 7, and compare the performance of GRICA with the state-of-the-art. As a measure of performance, we used the relative error metric given by: $error = |\text{Estimated Value} - \text{True Value}|/|\text{True Value}|$. Further experimental results are provided in Appendix C.

[3]The code to replicate the experiments can be found at : https://github.com/danieletramontano/Causal-Effect-Identification-in-LiNGAM-Models-with-Latent-Confounders.

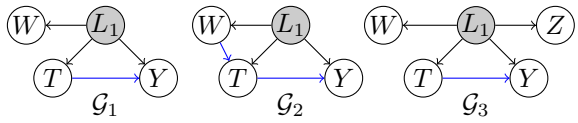
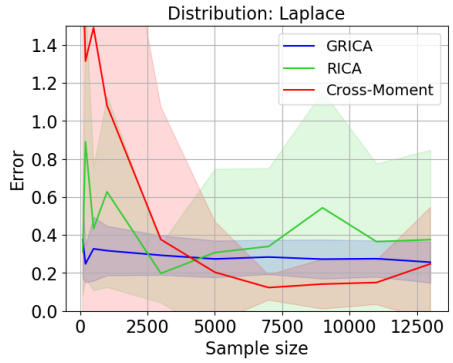
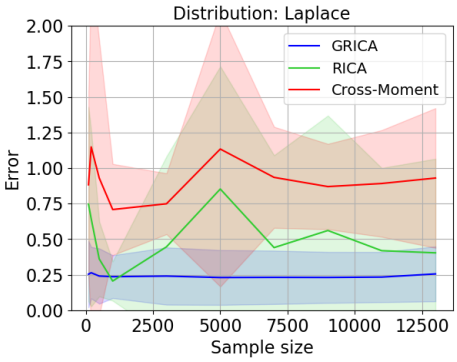


Figure 7. The causal graphs considered in the experiments.



(a) $\mathcal{G}_1$.



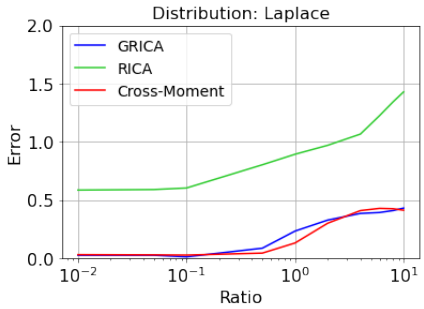
(b) $\mathcal{G}_2$.

Figure 8. Relative error vs sample size

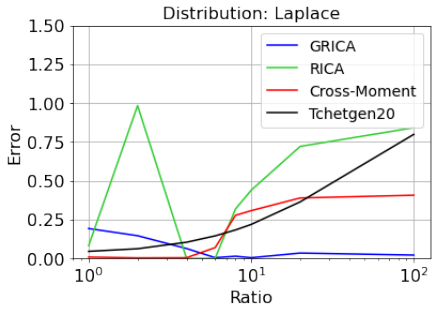
1. Relative error vs sample size.

We evaluated performance of the GRICA with respect to the observational sample size. We considered the settings that are compatible with the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 7 to recover the direct causal effect of variable T on Y. We compared the performance of GRICA with the adaptation of the RICA algorithm implemented in Salehkaleybar et al. (2020) and the recently developed Cross-Moment method in Kivva et al. (2023). We assumed all exogenous noises have the same distribution. Moreover, all direct causal coefficients in matrix $\mathbf{A}$ are generated uniformly at random from $[-1, -0.5] \cup [0.5, 1]$.[4] In Figure 8, we observe that GRICA consistently recovers the correct causal effect for both graphs, even with a few number of samples. Note that the Cross-Moment is a consistent estimator for the causal graph $\mathcal{G}_1$. It performs better than the rest when there are enough samples to compute high-order moments accurately. Furthermore, the RICA algorithm often gets stuck in bad local minima, and as a result is unstable.

[4]We consider this interval to ensure a fair comparison with RICA's implementation for causal effect estimation in Salehkaleybar et al. (2020) as it requires the absolute values of all causal coefficients to be smaller than one.



(a) $\mathcal{G}_1$.



(b) $\mathcal{G}_3$.

Figure 9. Relative error vs noise scaling ratio.

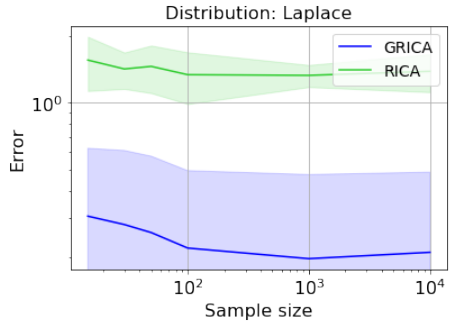
2. Relative error vs observations noise. In Figure 9, we illustrate how the variances of certain exogenous noise impact the accuracy of the estimation. All causal coefficients in both settings are set to one.

For the experiment over $\mathcal{G}_1$, we scaled the standard deviation of exogenous noise N_W corresponding to variable W by a factor displayed on the x-axis of Figure 9a. We observed that GRICA performs similarly to the Cross-Moment method, which is specifically designed for the graph $\mathcal{G}_1$.

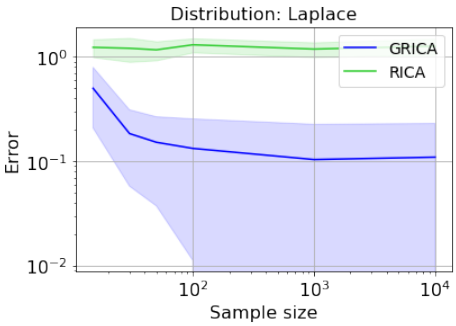
The experiment for $\mathcal{G}_3$ is similar to the one performed in Kivva et al. (2023). Here, we scaled N_W and N_Z by $Ratio$ and $1/Ratio$, respectively, where $Ratio$ is plotted on the x-axis of Figure 9b. We compared the performance of GRICA with RICA and methods developed specifically for this causal graph in (Kivva et al., 2023; Tchetgen et al., 2020). As depicted in Figure 9b, GRICA can benefit from the noiseless observation of L_1 through proxy Z, while all other algorithms are affected by the presence of noise in the observations from W.

3. Causal effect estimation on random graphs. In this section, we will present experiments for larger causal graphs. In

particular, we compare the GRICA algorithm with the RICA algorithm on graphs randomly sampled from an Erdős-Rényi model. The probability of accepting an edge is set to $1/2$. Similarly to the previous experiments, we generated all causal coefficients in the matrix $\mathbf{A}$ uniformly at random from $[-1, -0.5] \cup [0.5, 1]$. As a measure of performance, we use the normalized Frobenius loss between the estimated mixing matrix and the true one, i.e., $||\tilde{\mathbf{B}} - \mathbf{B}||_F / ||\mathbf{B}||_F$. Note that according to the experiments in Section 5.1, we expect all the causal effects of interest to be identifiable. Figures 10a and 10b show the experimental results for the Erdős-Rényi model in graphs with one latent and five observed variables and graphs with two latent and ten observed variables, respectively. The results are averaged over ten trials. As can be seen, GRICA significantly outperforms the RICA algorithm in larger causal graphs.



(a) One latent and five observed variables.



(b) Two latent and ten observed variables.

Figure 10. Results for the Erdős-Rényi model.

5.3. Consistency Guarantees

The OICA problem is known to be identifiable but not separable; see (Eriksson & Koivunen, 2004). This implies that, as opposed to the complete ICA, minimizing a measure of non-Gaussianity does not necessarily lead to the identification of the mixing matrix.

In the past, various Expectation-Maximization (EM) algorithms have been proposed to work under specific parametric models for the exogenous noise model, e.g., Mixture of Gaussians (Olshausen & Millman, 1999) or Laplace distribution (Lewicki & Sejnowski, 2000). These algorithms have been proven to be consistent if the assumptions are satisfied. At the same time, they are very computationally demanding and have been shown to perform poorly in practice. For this reason, the core of the research on the topic has shifted to heuristic methods, in which the mixing matrix is found as a solution to a suitable smooth optimization problem. The RICA algorithm (Le et al., 2011) is arguably the most prominent in this class of algorithms and has already been used in causality, see, e.g., Yang et al. (2022). However, due to the complications stated above, the RICA algorithm has not yet been equipped with consistency guarantees.

Being an adaptation of the RICA algorithm, also the large sample size performance of our algorithm is not well understood in rigorous mathematical terms. However, we point out that the same parametrization of the mixing matrix can be utilized in any OICA algorithm to reduce the dimensionality of the problem. What our experiments suggest is that this simple step can improve remarkably the performances of OICA algorithms, when applied to the estimation of causal effects.

6. Conclusions

We considered the problem of generic identifiability of causal effects in LiNGAM models when only observational data are available. We solved the problem by providing *efficiently implementable*, graphical criteria for identification of the causal effect with and without knowledge of the graph. To estimate the effect, we proposed a flexible adaptation of the RICA algorithm, (Le et al., 2011), that incorporates the knowledge of the graph to reduce the dimension in the optimization problem.

To conclude, we highlight possible future directions.

Cyclic Models. When the acyclicity assumption is dropped, Lacerda et al. (2008) showed that, in the fully observed case, the structure of the graph can be recovered up to an equivalence class. There are no explicit graphical criteria for the identification of the causal effects in cyclic LiNGAM models. It is worth exploring whether our proof techniques could be extended to this setting.

Relaxing non-Gaussianity. Ng et al. (2023) proved that in the complete ICA case, non-Gaussianity of the noises can be relaxed if the mixing matrix has a specific sparse structure. Extensions of these results to the overcomplete case might offer results on identification of the causal effects when some of the exogenous noises are allowed to be Gaussian.

Non-linear ICA. There is an active line of work that aims to exploit non-linear ICA, (Hyvärinen et al., 2024), to construct causal discovery algorithms, (Reizinger et al., 2023). There is no literature that leverages these results to understand the identifiability of the causal effects.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818) and supported in part by the SNF project 200021_204355/1, Causal Reasoning Beyond Markov Equivalencies. DT's PhD scholarship is funded by the IGSSE/TUM-GS via a Technical University of Munich–Imperial College London Joint Academy of Doctoral Studies.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adams, J., Hansen, N., and Zhang, K. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Ailer, E., Hartford, J., and Kilbertus, N. Sequential underspecified instrument selection for cause-effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Athey, S. and Imbens, G. W. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 2017.

Axler, S. *Linear algebra done right*. Springer, Cham, 3rd ed. edition, 2015.

Barber, R., Drton, M., Sturma, N., and Weihs, L. Half-trek criterion for identifiability of latent variable models. *The Annals of Statistics*, 50, 2022.

Barrera, E. L. and Miljkovic, D. The link between the two epidemics provides an opportunity to remedy obesity while dealing with covid-19. *Journal of Policy Modeling*, 44(2), 2022.

Brito, C. and Pearl, J. Generalized instrumental variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02. Morgan Kaufmann Publishers Inc., 2002.

Brito, C. and Pearl, J. Graphical condition for identification in recursive sem. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06. AUAI Press, 2006.

Cai, R., Huang, Z., Chen, W., Hao, Z., and Zhang, K. Causal discovery with latent confounders based on higher-order cumulants. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Chiyohara, S., Furukawa, J.-i., Noda, T., Morimoto, J., and Imamizu, H. Proprioceptive short-term memory in passive motor learning. *Scientific Reports*, 13(1), 2023.

Ciarli, T., Coad, A., and Moneta, A. Does exporting cause productivity growth? evidence from chilean firms. *Structural Change and Economic Dynamics*, 66, 2023.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. MIT Press, Cambridge, MA, third edition, 2009.

Cox, D. A., Little, J., and O'Shea, D. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015. An introduction to computational algebraic geometry and commutative algebra.

Cunningham, S. *Causal inference: The mixtape*. Yale university press, 2021.

de Prado, M. M. L. *Causal Factor Investing: Can Factor Investing Become Scientific?* Cambridge University Press, 2023.

Drton, M. Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77 of *Adv. Stud. Pure Math.* Math. Soc. Japan, Tokyo, 2018.

Drton, M., Foygel, R., and Sullivant, S. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2), 2011.

Eriksson, J. and Koivunen, V. Identifiability, separability, and uniqueness of linear ica models. *Signal Processing Letters, IEEE*, 11, 2004.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2), 2008. Special Section on Probabilistic Rough Sets and Special Section on PGM'06.

Hyvärinen, A., Khemakhem, I., and Monti, R. Identifiability of latent-variable and structural-equation models: From linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76, 2024.

Imai, K. and Kim, I. S. When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science*, 63(2), 2019.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

Kivva, Y., Salehkaleybar, S., and Kiyavash, N. A cross-moment approach for causal effect estimation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kumor, D., Cinelli, C., and Bareinboim, E. Efficient identification in linear structural causal models with auxiliary cutsets. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2), 2014.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08. AUAI Press, 2008.

Le, Q., Karpenko, A., Ngiam, J., and Ng, A. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Lewicki, M. and Sejnowski, T. Learning overcomplete representations. *Neural Computation*, 12:337–365, 02 2000.

Liu, J., Tchetgen, E. J. T., and Varjão, C. Proximal causal inference for synthetic control with surrogates. *arXiv:2308.09527*, 2023.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (eds.). *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Michałek, M. and Sturmfels, B. *Invitation to nonlinear algebra*, volume 211 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2021.

Michoel, T. and Zhang, J. D. Causal inference in drug discovery and development. *Drug Discovery Today*, 28 (10), 2023.

Ng, I., Zheng, Y., Dong, X., and Zhang, K. On the identifiability of sparse ica without assuming non-gaussianity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Okamoto, M. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1 (4), 1973.

Olshausen, B. A. and Millman, K. J. Learning sparse codes with a mixture-of-gaussians prior. In Solla, S. A., Leen, T. K., and Müller, K. (eds.), *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pp. 841–847. The MIT Press, 1999.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. Models, reasoning, and inference.

Pe'er, D. and Hacohen, N. Principles and strategies for developing network models in cancer. *Cell*, 144(6), 2011.

Reizinger, P., Sharma, Y., Bethge, M., Schölkopf, B., Huszár, F., and Brendel, W. Jacobian-based causal discovery with nonlinear ICA. *Transactions on Machine Learning Research*, 2023.

Salehkaleybar, S., Ghassami, A., Kiyavash, N., and Zhang, K. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39), 2020.

Sanchez, P., Voisey, J., Xia, T., Watson, H., O'Neil, A., and Tsaftaris, S. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9, 2022.

Shimizu, S. *Statistical Causal Discovery: LiNGAM Approach*. Springer, 2022.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2006.

Shpitser, I. When does the id algorithm fail? *arXiv:2307.03750*, 2023.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06. AAAI Press, 2006.

Shuai, K., Luo, S., Zhang, Y., Xie, F., and He, Y. Identification and estimation of causal effects using non-gaussianity and auxiliary covariates. *arXiv:2304.14895*, 2023.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tramontano, D., Drton, M., and Etesami, J. Parameter identification in linear non-gaussian causal models under general confounding. *arXiv:2405.20856*, 2024.

Wang, Y. S. and Drton, M. Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271), 2023.

Wang, Y. S., Kolar, M., and Drton, M. Confidence sets for causal orderings. *arXiv:2305.14506*, 2023.

Yang, Y., Ghassami, A., Nafea, M., Kiyavash, N., Zhang, K., and Shpitser, I. Causal discovery in linear latent variable models subject to measurement error. *Advances in Neural Information Processing Systems*, 35, 2022.

A. Notions of Non-Linear Algebra

In this section, we give the basic definitions of *non-linear* algebra we will need for the proofs; we refer the interested reader to Cox et al. (2015); Michałek & Sturmfels (2021) for more details.

Definition A.1. For every natural number n, we denote the ring of polynomials in n variables $x_1, \dots, x_n$ by $\mathbb{R}[x_1, \dots, x_n]$. Let S be a, possibly infinite, subset of $\mathbb{R}[x_1, \dots, x_n]$. The affine variety associated to it is defined as $\mathcal{V}(S) = \{x \in \mathbb{R}^n \mid f(x) = 0, \forall f \in S\}$. The vanishing ideal associated to a variety $\mathcal{V}$ is $\mathcal{I}(\mathcal{V}) = \{f \in \mathbb{R}[x_1, \dots, x_n] \mid f(x) = 0 \, \forall x \in \mathcal{V}\}$. The coordinate ring of $\mathcal{V}$ is defined as $\mathbb{R}[\mathcal{V}] = \mathbb{R}[x_1, \dots, x_n]/\mathcal{I}(\mathcal{V})$.

Lemma A.2 (Lemma (Okamoto, 1973))**.** *Let $f(x_1, \dots, x_n)$ be a polynomial in real variables $x_1, \dots, x_n$, which is not identically zero. The set of zeros of the polynomial is a Lebesgue measure zero subset of $\mathbb{R}^n$.*

Remark A.3 (Notation). For a given matrix $\mathbf{A}$, we denote the submatrix in which the i-th row and the j-th column are excluded by $[\mathbf{A}]_{\setminus i, \setminus j}$.

Lemma A.4. *Let $\mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$ and $\mathbb{R}^{\mathcal{G}}$ defines as in Section 2.2. Then we have $\mathbb{R}^{\mathcal{G}} \sim \mathbb{R}^{\mathcal{G}}_{\mathbf{A}} \sim \mathbb{R}^{|e|}$, where with the symbol $\sim$, we denote an isomorphism of affine varieties, see, e.g., Cox et al. (2015, Def. 6, §5) for a definition. Moreover $\mathbb{R}[\mathcal{G}]$, $\mathbb{R}[\mathcal{G}_{\mathbf{A}}]$, and $\mathbb{R}[a_{i,j} \mid j \to i \in \mathcal{G}]$ are isomorphic as rings.*

Proof. The isomorphism $\mathbb{R}^{\mathcal{G}}_{\mathbf{A}} \sim \mathbb{R}^{|e|}$ comes directly from its definition. Indeed it is easy to see that $\mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$ is an $|e|$-dimensional linear subspace of $\mathbb{R}^{p \times p} = (a_{i,j})_{i,j \in p \times p}$, defined by the linear equations $a_{i,i} = 1$, and $a_{i,j} = 0$, $\forall i, j \in \mathcal{V}$ such that $j \to i \notin \mathcal{G}$.

To prove the isomorphism $\mathbb{R}^{\mathcal{G}} \sim \mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$, we need to prove that there is a polynomial bijective map between the two spaces. From (5), and using $[\mathbf{B}_o]_{i,j} = [(\mathbf{A}_{o,o})^{-1}]_{i,j} = (-1)^{i+j} \det([\mathbf{A}_{o,o}]_{\setminus j, \setminus i})$, where we used that $\det(\mathbf{A}_{o,o}) = 1$. It is clear that $\mathbb{R}^{\mathcal{G}}$ is the image of polynomial map of $\mathbb{R}^{\mathcal{G}}_{\mathbf{A}}$. Let us call this polynomial map ψ and assume $\psi(\mathbf{A}) = \psi(\tilde{\mathbf{A}})$. Then from the definition of ψ we have $(I - \mathbf{A}_{o,o})^{-1} = (I - \tilde{\mathbf{A}}_{o,o})^{-1}$ that implies $\mathbf{A}_{o,o} = \tilde{\mathbf{A}}_{o,o}$. Moreover, $(I - \mathbf{A}_{o,o})^{-1}\mathbf{A}_{o,l} = (I - \tilde{\mathbf{A}}_{o,o})^{-1}\tilde{\mathbf{A}}_{o,l}$ that implies $\mathbf{A}_{o,l} = \tilde{\mathbf{A}}_{o,l}$ and so $\mathbf{A} = \tilde{\mathbf{A}}$.

The isomorphisms between the rings come from Cox et al. (2015, §5, Thm. 9). □

Corollary A.5. *Let $f \in \mathbb{R}[\mathcal{G}]$ be a non-zero polynomial. Then the subset of $\mathbb{R}^{\mathcal{G}}$ on which f vanishes is a Lebesgue measure 0 subset of $\mathbb{R}^{\mathcal{G}}$.*

Proof. Thanks to the isomorphism in Lemma A.4, we can apply Lemma A.2 to $\mathbb{R}^{\mathcal{G}}$. □

Definition A.6 (Leibniz Expansion of the Determinant)**.** For any $M \in \mathbb{R}^{n \times n}$, the determinant of M can be computed using the following formula:

$$\det(M) = \sum_{\sigma_P \in \mathcal{S}(n)} (-1)^{\sigma_P} \prod_{i=1}^{n} M_{\sigma_P(i), i}, \tag{21}$$

where $\mathcal{S}(n)$ is the set of all the permutations of n elements, and $(-1)^{\sigma_P}$ is the sign of the permutation. See, e.g., Axler (2015, Def. 10.33) for more details.

Definition A.7. Let $\pi \in \mathcal{P}(j, i)$. The path monomial associated to it is defined as

$$a^{\pi} = a_{i_1, i_2} \cdot \dots \cdot a_{i_k, i_{k+1}} \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}].$$

Lemma A.8. *Let $\mathbf{A}$ defined as in (1). We have*

$$\mathbf{B} = (I - \mathbf{A})^{-1} = \sum_{i=0}^{\infty} \mathbf{A}^i = I + \mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^p,$$

$$[\mathbf{B}]_{i,j} = \sum_{P \in \mathcal{P}(i,j)} a^P.$$

In particular $[\mathbf{B}]_{i,j} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]$ if and only if $P \in \mathcal{P}(i, j) = \emptyset$.

Definition A.9. Let $I = \{i_1, \dots, i_n\}, J = \{j_1, \dots, j_n\} \subset \mathcal{V}$. $P = P_1, \dots, P_n$ is system of paths between I, J, if there exists a permutation $\sigma_P \in S_n$, such that $P_k \in \mathcal{P}(i_k, j_{\sigma_P(k)})$. We denote the set of all such systems by $\mathcal{P}(I, J)$. Moreover, a system of paths is non-intersecting if $P_k \cap P_l = \emptyset$ for $k \neq l$. The set of all such systems is denoted by $\tilde{\mathcal{P}}(I, J)$. The path monomial associated to P, is defined as:

$$a^P = a^{P_1} \cdot \dots \cdot a^{P_n}.$$

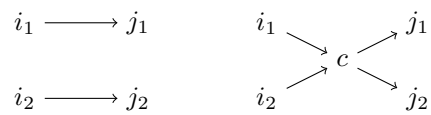


Figure 11. The system on the left has no sided intersection while the system on the right has.

Theorem A.10 (Gessel-Viennot-Lindström Lemma)**.** *Let $I, J \subset \mathcal{V}$ having the same size; then it holds that*

$$\det(I - \mathbf{A})^{-1}_{I,J} = \sum_{P \in \mathcal{P}(J,I)} (-1)^{\sigma_P} a^P = \sum_{P \in \tilde{\mathcal{P}}(J,I)} (-1)^{\sigma_P} a^P.$$

In particular, $\det(I - \mathbf{A})^{-1}_{I,J} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]$ *if and only if* $\tilde{\mathcal{P}}(J, I) = \emptyset$.

B. Proofs and Lemmas

B.1. Proofs of Section 3.1

Proof of Theorem 3.3. From Salehkaleybar et al. (2020, Theo. 15), we know that the entire j-th column $\mathbf{B}'$ is identifiable if and only if there is no latent variable l such that $\mathrm{de}_o(j) = \mathrm{de}_o(l)$. Thus, we know that if this condition is satisfied, the entry $[\mathbf{B}']_{i,j}$ is identifiable. Hence, we can assume that such an l exists to conclude the proof.

We will prove that $[\mathbf{P}_\sigma \mathbf{B}']_{i,j} = [\mathbf{B}']_{i,l} \neq [\mathbf{B}']_{i,j}$, *in general*, if and only if $i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$, where σ is the transposition that swaps j and l. In particular, we will show that $[\mathbf{B}']_{i,j} - [\mathbf{B}']_{i,l} \in \mathbb{R}[\mathcal{G}]$ where $\mathbb{R}[\mathcal{G}]$ (the coordinate ring associate to $\mathcal{G}$ as defined in Definition A.1), is a *non-zero* polynomial if and only if the graphical condition is satisfied. Notice that from Corollary A.5, we know this is enough to show that the parameters are generically identifiable.

Using Lemma A.8, we can write the entries of $\mathbf{B}'$ as

$$[\mathbf{B}']_{i,l} = (I - \mathbf{A})^{-1}_{i,l} = \sum_{P \in \mathcal{P}(l,i)} a^P,$$

where $\mathcal{P}(l, i)$ denotes the set of all the paths from l to i. Now let $\mathcal{P}_j(l, i)$ be the set of directed path from l to i that passes through j, and $\mathcal{P}_{\setminus j}(l, i) = \mathcal{P}(l, i) \setminus \mathcal{P}_j(l, i)$. We can rewrite the formula above as follows

$$[\mathbf{B}']_{i,l} = \sum_{P \in \mathcal{P}(l,i)} a^P = \sum_{P \in \mathcal{P}_j(l,i)} a^P + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q.$$

Note that $P \in \mathcal{P}_j(l, i)$ if and only if there are $P_j \in \mathcal{P}(l, j)$ and $P_i \in \mathcal{P}(j, i)$ such that $a^P = a^{P_j} a^{P_i}$. Finally, we can write $[\mathbf{B}']_{i,l}$ as:

$$\begin{aligned}
&\sum_{P \in \mathcal{P}_j(l,i)} a^P + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q = \sum_{\substack{P_i \in \mathcal{P}(j,i) \\ P_j \in \mathcal{P}(l,j)}} a^{P_j} a^{P_i} + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q \\
&= \underbrace{\sum_{P_i \in \mathcal{P}(j,i)} a^{P_i}}_{[\mathbf{B}']_{i,j}} \underbrace{\sum_{P_j \in \mathcal{P}(l,j)} a^{P_j}}_{[\mathbf{B}']_{i,l}} + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q = [\mathbf{B}']_{i,j} [\mathbf{B}']_{j,l} + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q \\
&= [\mathbf{B}']_{i,j} + \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q,
\end{aligned} \tag{22}$$

where we used $[\mathbf{B}']_{j,l} = 1$ from Remark 3.2 that we can assume without loss of generality. Finally, we have $[\mathbf{B}']_{i_0,j} - [\mathbf{B}']_{i_0,l} = \sum_{Q \in \mathcal{P}_{\setminus j}(l,i_0)} a^Q \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]$. From Lemma A.8, we know it is the zero polynomial if and only if $\mathcal{P}_{\setminus j}(l, i_0) = \emptyset$, i.e., $i_0 \notin \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$. Notice that we proved that $[\mathbf{B}']_{i,j} - [\mathbf{B}']_{i_0,l}$ is non zero in $\mathbb{R}[\mathcal{G}_{\mathbf{A}}]$, that is enough to show thanks to Lemma A.4. □

Lemma B.1. *Let j and l be observed and latent variables, respectively, such that $\mathrm{de}_o(j) = \mathrm{de}_o(l)$. Let $\tilde{\mathbf{B}}' = \mathbf{B}'\mathbf{P}_\sigma$, where σ is the transposition swapping the columns corresponding to l and j. For every $i, k \in \mathcal{O}$, and $h \in \mathcal{L}$ we have:*

$$[\tilde{\mathbf{A}}_{o,o}]_{i,k} = [\mathbf{A}_{o,o}]_{i,k} + \mathbf{1}_{\mathrm{ch}(l)}(i) c_{l,i}^j [I - \mathbf{A}_{o,o}]_{j,k}, \tag{23}$$

$$[\tilde{\mathbf{A}}_{o,l}]_{i,h} = [\mathbf{A}_{o,l}]_{i,h} - \mathbf{1}_{\mathrm{ch}(l)}(i) c_{l,i}^j [\mathbf{A}_{o,l}]_{j,h}, \tag{24}$$

where $c_{l,i}^j = (-1)^{i+j} \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q$, and $\mathbf{1}_X$ represent the indicator function of a set X.

In particular, $[\tilde{\mathbf{A}}_{o,o}]_{i,k} = [\mathbf{A}_{o,o}]_{i,k}$, if $i \notin \mathrm{ch}(l)$ or $k \notin \mathrm{pa}(j) \cup \{j\}$.

Proof of Lemma B.1. We first show the equality in (23). According to (22), $\tilde{\mathbf{B}}_o$ can be written as $\mathbf{B}_o + C$:

$$C_{i,k} = \begin{cases} \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q & \text{if } k = j \text{ and } i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l), \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

From the formula for the inverse of a matrix, we know that

$$[\mathbf{B}_o + C]^{-1}_{i,k} = (-1)^{i+k} \det([\mathbf{B}_o + C]_{\setminus k, \setminus i}).$$

Note that $\det(\mathbf{B}_o + C) = 1$ since it is a lower triangular matrix with one on the diagonal. Thus, we can imply that $[\tilde{\mathbf{A}}_{o,o}]_{j,k} = [(\mathbf{B}_o + C)^{-1}]_{j,k} = [\mathbf{B}_o]^{-1}_{j,k} = [\mathbf{A}_{o,o}]_{j,k}$. Hence, we only need to consider the entries of $[\tilde{\mathbf{A}}_{o,o}]_{i,k}$ where $i \neq j$.

To compute $\det([\mathbf{B}_o + C]_{\setminus k, \setminus i})$ with respect to the j-th column, we have:

$$\begin{aligned} &\det([\mathbf{B}_o + C]_{\setminus k, \setminus i}) = \sum (-1)^{s+j} [\mathbf{B}_o]_{s,j} \det([\mathbf{B}_o]_{\setminus (j,s), \setminus (i,j)}) + \sum (-1)^{s+j} C_{s,j} \det([\mathbf{B}_o]_{\setminus (j,s), \setminus (i,j)}) \\ &= \det([\mathbf{B}_o]_{\setminus k, \setminus i}) + \sum (-1)^{s+j} C_{s,j} \det([\mathbf{B}_o]_{\setminus (j,s), \setminus (i,j)}) \\ &= (-1)^{i+j} [\mathbf{A}_{o,o}]_{i,k} + \sum_{s \in \mathrm{de}_o^{\mathcal{G}_{\setminus k}}(l)} (-1)^{s+j} \sum_{Q \in \mathcal{P}_{\setminus j}(l,s)} a^Q \det([\mathbf{B}_o]_{\setminus (k,s), \setminus (i,j)}), \end{aligned}$$

where in the last equality, we plug in $C_{s,j}$ according to (25).

Based on Theorem A.10,

$$\det([\mathbf{B}_o]_{\setminus (k,s), \setminus (i,j)}) = \det([(I - \mathbf{A}_{o,o})^{-1}]_{\setminus (k,s), \setminus (i,j)}) = \sum_{P \in \tilde{\mathcal{P}}(\mathcal{O} \setminus \{i,j\}, \mathcal{O} \setminus \{k,s\})} (-1)^{\sigma_P} a^P.$$

In the following, we show that that $\tilde{\mathcal{P}}(\mathcal{O} \setminus \{i,j\}, \mathcal{O} \setminus \{k,s\})$ can be non-empty only if:

$$\begin{cases} s = i \text{ or } s \to i \in \mathcal{G}, \\ k = j \text{ or } k \to j \in \mathcal{G}. \end{cases} \tag{26}$$

To show this, let

$$P = \{\pi_1, \dots, \pi_{j-1}, \pi_{j+1}, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_p\}$$

be a system of paths in $\tilde{\mathcal{P}}(\mathcal{O} \setminus \{i,j\}, \mathcal{O} \setminus \{k,s\})$, where with π_a we denote the path starting at the node a. Consider the path $\pi_s = sa_0 \dots a_t$. If there is a $x \in \{0, \dots, t\}$ such that $a_x \notin \{i,j\}$, then π_s would intersect π_{a_x} and thus P would have an intersection. The same argument applies for π_k. Therefore,

$$\pi_s = sa_0, \qquad \pi_k = kb_0, \qquad \{a_0, b_0\} = \{i,j\}.$$

To conclude the argument, please note that, since from (25) we see that $C_{s,j} = 0$ if $s \notin \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$, we can restrict ourselves to the case in which $s \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l) \subseteq \mathrm{de}(j)$. Since the graph is acyclic, this implies $a_0 = i$ and $b_0 = j$.

We now prove that if for all the paths from l to i, there is an observed variable $s_0 \neq i$ on the path, the quantity

$$\sum_{s \in \mathrm{de}_o^{\mathcal{G}_{\setminus k}}(l)} (-1)^{s+j} \sum_{Q \in \mathcal{P}_{\setminus j}(l,s)} a^Q \det([\mathbf{B}_o]_{\setminus(k,s),\setminus(i,j)}),$$

is equal to 0. We can rewrite the sum above as follows:

$$\sum_{s \in \mathrm{de}_o^{\mathcal{G}_{\setminus k}}(l)} \sum_{\substack{Q \in \mathcal{P}_{\setminus j}(l,s) \\ P \in \tilde{\mathcal{P}}(\mathcal{O}\setminus\{i,j\},\mathcal{O}\setminus\{k,s\})}} (-1)^{s+j+\sigma_P} a^Q a^P. \tag{27}$$

Let π_0 be a path from l to i that can be decomposed in the following way:

$$\overbrace{\underbrace{l \to \cdots \to s_0}_{Q_0} \to i}^{\pi_0},$$

for some $s_0 \in \mathcal{V}$. If s_0 is observed, the monomial associated with this path will appear in the sum twice in (27); the first time when considering $Q = Q_0$ and $s = s_0$ while the second time when considering $Q = \pi_0$ and $s = i$. The last thing to prove is that the sign will differ in the two cases. This comes from the Leibniz expansion for the determinant, (21). Indeed, the sign associated with each monomial is the sign associated with the permutation in the Leibniz formula. In the first case, the associated permutation is the following:

$$\sigma_1(a) = \begin{cases} a & \textit{if } a \notin \{s_0, i, j\}, \\ j & \textit{if } a = s_0, \\ s_0 & \textit{if } a = i, \\ k & \textit{if } a = j, \end{cases}$$

while for the second case, the associated permutation is:

$$\sigma_2(a) = \begin{cases} a & \textit{if } a \notin \{j, i\}, \\ j & \textit{if } a = i, \\ k & \textit{if } a = j. \end{cases}$$

As the two permutations can be obtained one from the other via the transposition corresponding to swapping the columns corresponding to j and s_0, they have different signs.

The same argument also implies that all the elements in (27), involving $s \in \mathrm{pa}(i)$ cancel out allowing us to rewrite the sum as:

$$\sum_{\substack{Q \in \mathcal{P}_{\setminus j}(l,i) \\ P \in \tilde{\mathcal{P}}(\mathcal{O}\setminus\{i,j\},\mathcal{O}\setminus\{i,k\})}} (-1)^{i+j+\sigma_P} a^Q a^P.$$

With the same argument used to prove (26), one can see that the only element in $\tilde{\mathcal{P}}(\mathcal{O} \setminus \{i,j\}, \mathcal{O} \setminus \{i,k\})$, is the system of paths that sends k to j directly and all the other elements remain fixed. This system of paths has a negative sign if $k \neq j$ and a positive sign otherwise, implying that the sum is equal to:

$$(-1)^{i+j}[I - \mathbf{A}_{o,o}]_{j,k} \sum_{Q \in \mathcal{P}_{\setminus j}(l,i)} a^Q,$$

which concludes the first part of the proof. We now prove (24). Using (5) we can write $[\tilde{\mathbf{A}}_{o,l}]_{i,h} = [(\tilde{\mathbf{B}}_{o,o})^{-1}(\tilde{\mathbf{B}}_{o,l})]_{i,h} = \sum_{s \in \mathcal{O}} [(\tilde{\mathbf{B}}_{o,o})^{-1}]_{i,s} [(\tilde{\mathbf{B}}_{o,l})]_{s,l}$.

We first prove the case $h = l$. In this case, by definition of $\tilde{\mathbf{B}}'$, we have $[\tilde{\mathbf{B}}_{o,l}]_{s,l} = [\mathbf{B}_o]_{s,j}$. Thus, plugging in (23) we can write $[\tilde{\mathbf{A}}_{o,l}]_{i,l}$ as

$$\begin{aligned}&\sum_{s\in\mathcal{O}}[(\mathbf{B}_o)^{-1}]_{i,s}[\mathbf{B}_o]_{s,j} - c^j_{l,i}\sum_{s\in\mathcal{O}}[(\mathbf{B}_o)^{-1}]_{j,s}[\mathbf{B}_o]_{s,j}\\ &= \delta_{i,j} - c^j_{l,i} = \delta_{i,j} - c^j_{l,i}[\mathbf{A}]_{j,l},\end{aligned}$$

where, again, we used $[\mathbf{A}]_{j,l} = 1$, that comes from Remark 3.2. For the case $h \neq l$ we use again (23), and write $[\tilde{\mathbf{A}}_{o,l}]_{i,h}$ as

$$\begin{aligned}&\sum_{s\in\mathcal{O}}[(\mathbf{B}_o)^{-1}]_{i,s}[\mathbf{B}_l]_{s,h} - c^j_{l,i}\sum_{s\in\mathcal{O}}[(\mathbf{B}_o)^{-1}]_{j,s}[\mathbf{B}_l]_{s,h}\\ &= [\mathbf{A}_{o,l}]_{i,h} - c^j_{l,i}[\mathbf{A}_{o,l}]_{j,h},\end{aligned}$$

where for the last equality we only used $\mathbf{A}_{o,l} = (\mathbf{B}_o)^{-1}\mathbf{B}_l$ that comes from (5). □

Proof of Theorem 3.4. From Lemma B.1, we can conclude that if there are no k and l satisfying the conditions in Equations (11)-(13), then the entry $[\mathbf{A}_{\mathbf{o},\mathbf{o}}]_{i,j}$ remains unchanged when swapping the columns of $\mathbf{B}'$, proving that the condition is sufficient. To prove the necessity, it is enough to show that $\sum_{Q\in\mathcal{P}_{\setminus j}(l,i)} a^Q \neq 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]$, which is equivalent to proving that $\mathcal{P}_{\setminus j}(l,i) \neq \emptyset$. This is true since $i \in \mathrm{ch}(l)$ from Equation (12). □

B.2. Proofs of Section 3.2

Lemma B.2. *For every $\mathbf{B}'$ outside of a Lebesgue zero subset of $\mathbb{R}^{\mathcal{G}}$, let $\tilde{\mathbf{B}}' = \mathbf{B}' \cdot \mathbf{P}_\sigma$, where σ is the transposition that swaps j and l, and $\mathbf{P}_\sigma$ is the associated permutation matrix. We have $\tilde{\mathbf{B}}' \in \mathbb{R}^{\mathcal{G}}$ if and only if*

$$\mathrm{de}_o(I) = \mathrm{de}_o(\sigma(I)), \qquad \forall i \in \mathcal{V},$$

and there is no $k \in \mathrm{pa}(j) \cup \{j\}$, such that $\mathrm{ch}(l) \setminus \mathrm{ch}(k) \neq \emptyset$.

Proof of Lemma B.2. From Salehkaleybar et al. (2020, Thm. 15), we know that the only permutations that result in matrices in $\mathbb{R}^{\tilde{\mathcal{G}}}$ for some DAG $\tilde{\mathcal{G}}$, are the ones for which

$$\mathrm{de}_o(I) = \mathrm{de}_o(\sigma(I)), \qquad \forall i \in \mathcal{V}.$$

Hence this condition is necessary.

The edges in $\tilde{\mathcal{G}}$ are given by the support of the matrix $\tilde{\mathbf{A}}$. In particular $\mathcal{G} = \tilde{\mathcal{G}}$ if and only if

$$\begin{aligned}[\mathbf{A}_{o,o}]_{i,k} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}] &\iff [\tilde{\mathbf{A}}_{o,o}]_{i,k} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]\\ [\mathbf{A}_{o,l}]_{i,l} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}] &\iff [\tilde{\mathbf{A}}_{o,l}]_{i,l} = 0 \in \mathbb{R}[\mathcal{G}_{\mathbf{A}}]\end{aligned} \tag{28}$$

We provide the proof only for $\mathbf{A}_{o,o}$ since the one for $\mathbf{A}_{o,l}$ follows the same argument only using (24) instead of (23).

From Lemma B.1 we know that $[\tilde{\mathbf{A}}_{o,o}]_{i,k} = [\mathbf{A}_{o,o}]_{i,k}$, if $i \notin \mathrm{ch}(l)$ or $k \notin \mathrm{pa}(j) \cup \{j\}$. Thus, the only the entries to consider are $i \in \mathrm{ch}(l)$ and $k \in \mathrm{pa}(j) \cup \{j\}$.
We know that $[\mathbf{A}_{o,o}]_{i,k} \neq 0$ if and only if $i \in \mathrm{ch}(k)$, while $[\tilde{\mathbf{A}}_{o,o}]_{i,k}$ is always different from 0. Therefore, the condition in (28) fails if and only if there is an $i \in \mathrm{ch}(l) \setminus \mathrm{ch}(j)$. □

Proof of Theorem 3.6. In Lemma B.2, we have shown that the statement is true if σ is a transposition. We now assume that $\sigma = \sigma_n \cdots \sigma_1$ where σ_s is a transposition for every $s \in \{1, \dots, n\}$, such that if $\sigma_s(j) \neq j$ for some j then also $\sigma(j) \neq j$. This can be done without loss of generality since every permutation can be written this way.

We can see that the condition is sufficient from Lemma B.2; if it is satisfied, the support of $\mathbf{A}^\sigma$ cannot change. In order to prove that the condition is also necessary, we need to verify that if there is a σ_s such that $[\mathbf{A}^{\sigma_s}]_{i,k} \neq 0$ then $[\mathbf{A}^\sigma]_{i,k} \neq 0$ as well. Again, we will prove the result for a pair of observed variables i and k since the result follows the same way when considering latent variables.

In particular, we are going to prove the following

$$[\mathbf{A}^{\sigma}_{o,o}]_{i,k} = [\mathbf{A}_{o,o}]_{i,k} + \sum_{s} \mathbf{1}_{\mathrm{ch}(l_s)}(i) c^{j_s}_{l_s,i} [I - \mathbf{A}_{o,o}]_{j_s,k} + r_{\sigma}, \tag{29}$$

where σ_s is the transposition that swaps j_s with l_s and r_σ is either 0 or a polynomial of degree at least two in $\mathbb{R}[\mathcal{G}_{\mathbf{A}}]$. We proceed by induction on n. If n is equal to 1, then we apply Lemma B.1 with $r_\sigma = 0$.

In order to proceed with the induction step, let us define $\sigma_{[s]} = \sigma_s \cdots \sigma_1$ so that we can write σ as $\sigma_{k+1} \cdot \sigma_{[k]}$. From the induction, we know that

$$[\mathbf{A}^{\sigma_{[n]}}_{o,o}]_{i,k} = [\mathbf{A}_{o,o}]_{i,k} + \sum_{s=1}^{n} \mathbf{1}_{\mathrm{ch}(l_s)}(i) c^{j_s}_{l_s,i} [I - \mathbf{A}_{o,o}]_{j_s,k} + r_{\sigma_{[n]}},$$

and by construction we can write $\mathbf{B}^{\sigma} = \mathbf{B}^{\sigma_{[n]}} + C^{(n+1)}$ where

$$C^{(n+1)}_{i,k} = \begin{cases} \sum\limits_{Q \in \mathcal{P}_{\setminus j_{n+1}}(l_{n+1},i)} a^{Q} & \textit{if } k = j_{n+1} \textit{ and } i \in \mathrm{de}^{\mathcal{G}_{\setminus j}}_{o}(l_{n+1}), \\ 0 & \textit{otherwise}. \end{cases} \tag{30}$$

Following the same steps of the proof of Lemma B.1, with the only difference of using $\mathbf{B}^{\sigma}$ in place of $\tilde{\mathbf{B}}'$, and $\mathbf{B}^{\sigma_{[n]}}$ in place of $\mathbf{B}'$ we obtain

$$[\mathbf{A}^{\sigma}]_{i,k} = [\mathbf{A}^{\sigma_{[n]}}]_{i,k} + \sum_{s} (-1)^{s+j_{n+1}} C^{(n+1)}_{s,j_{n+1}} \det([\mathbf{B}^{\sigma_{[n]}}_{o}]_{/(k,s),/(i,j_{n+1})}). \tag{31}$$

Thus, the only thing that is left to prove is that the last term on the right-hand side of the equation is equal to $\mathbf{1}_{\mathrm{ch}(l_{n+1})}(i) c^{j_{n+1}}_{l_{n+1},i} [I - \mathbf{A}_{o,o}]_{j_{n+1},j} + r_{\sigma_{n+1}}$. In order to do so note that using the same formula as above we can write $\mathbf{B}^{\sigma_{[n]}}_{o} = \mathbf{B}^{\sigma_{[n-1]}}_{o} + C^{n}$, and thus

$$\det([\mathbf{B}^{\sigma_{[n]}}_{o}]_{/(k,s),/(i,j_{n+1})}) = \det([\mathbf{B}^{\sigma_{[n-1]}}_{o}]_{/(k,s),/(i,j_{n+1})}) + r'_{\sigma_{[n-1]}}.$$

Plugging the above equation in (31), concludes the proof, following the same steps as in the proof of Lemma B.1. □

Proof of Theorem 3.8. The only difference with respect to Theorem 3.3 is the condition

$$\mathrm{ch}(l) \setminus \mathrm{ch}(k_1) = \emptyset, \quad \forall k_1 \in \mathrm{pa}(k) \cup \{k\},$$

that comes from Theorem 3.6. Indeed, if this condition is not satisfied, then the permutation that swaps the columns corresponding to k and l, cannot result in a model in $\mathbb{R}^{\mathcal{G}}$. □

Proof of Theorem 3.9. As in the proof of Theorem 3.8, the difference from the conditions of Theorem 3.4 and Theorem 3.9 is condition (20). This is, again, a direct consequence of Theorem 3.6. □

Proof of Corollary 3.10. The mixing matrix $\mathbf{B}_o$ is identifiable if and only if all of its parameters are. Hence, the result is a direct consequence of Theorem 3.8. □

Example B.3 (Examples 3.5 and 3.7 continued)**.** *Here, we report the mixing matrices corresponding to the two models that are compatible with the observed distribution*

$$\mathbf{B}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ b_{TI} & 1 & 1 & 0 \\ b_{TI}b_{YT} & b_{YT} & b_{YT} + b_{LY} & 1 \end{bmatrix}, \quad \tilde{\mathbf{B}}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ b_{TI} & 1 & 1 & 0 \\ b_{TI}b_{YT} & b_{YT} + b_{LY} & b_{YT} & 1 \end{bmatrix},$$

together with the corresponding adjacency matrices

$$\mathbf{A}_{o,o} = \begin{bmatrix} 1 & 0 & 0 \\ b_{TI} & 1 & 0 \\ 0 & b_{YT} & 1 \end{bmatrix}, \quad \tilde{\mathbf{A}}_{o,o} = \begin{bmatrix} 1 & 0 & 0 \\ b_{TI} & 1 & 0 \\ -b_{TI}b_{YL} & b_{YT} + b_{YL} & 1 \end{bmatrix}.$$

The two models are depicted in Figure 12. A parameter is generically identifiable without knowledge of the graph if it is the same for both models. Since there is only one model that is compatible with $\mathcal{G}_{IV}$, all the parameters are generically identifiable with knowledge of the graph. These results are summarized in Table 1.

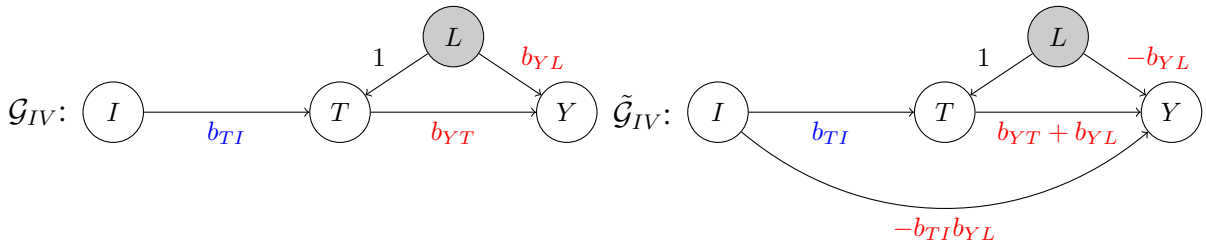


Figure 12. The graphs corresponding to the two models.

Table 1. Summary of the identifiable parameters in the IV graph.

	Known DAG		Unknown DAG	
	TCE	DCE	TCE	DCE
$I \to T$	✓	✓	✓	✓
$I \to Y$	✓	✓	✓	✗
$T \to Y$	✓	✓	✗	✗

B.3. Proofs of Section 3.4

Theorem B.4. *Algorithms 1, 2, 3, 4, 5, and 6 are sound and complete for solving the corresponding identifiability queries, as summarized in Table 2.*

Moreover, the computational complexity the algorithms is $\mathcal{O}(p_l(p_o^2 + |E|)) = \mathcal{O}(p^3)$.

Table 2. Summary of the identification algorithms, with the corresponding identifiability queries.

	Known DAG		Unknown DAG	
	TCE	DCE	TCE	DCE
Given Pair	Algorithm 1	Algorithm 3	Algorithm 2	Algorithm 4
Complete Matrix	Algorithm 5	Algorithm 5	Algorithm 6	Algorithm 6

Proof of Theorem 3.11 and Theorem B.4. In order to prove that correctness of Algorithm 1, it is enough to show that if $\mathrm{de}_o(l) = \mathrm{de}_o(j)$ then $j < k$ for every $k \in \mathrm{ch}(l) \setminus \{j\}$. Assume by contradiction that there is $k < j$ in $\mathrm{ch}(l)$. Then k cannot be a descendent of j and so $\mathrm{de}_o(l) \neq \mathrm{de}_o(j)$. This implies that the total causal effect is identifiable if the condition at line 7 : in Algorithm 1 is not satisfied for any latent variable. The rest of the algorithm is just a translation of the conditions in Theorem 3.8.

The nodes can be arranged in a topological order in $\mathcal{O}(p+|E|)$ time with one run of depth-first search, see e.g., Cormen et al. (2009, §22.4). Computing the descendants of a node can be done again with depth-first search in the graph in which only the observed nodes are considered since we assume no edges from observed to latent variables. Therefore, it has a cost of $\mathcal{O}(p_o + |E|)$, sorting $\mathrm{ch}(l_k)$ in line 5 of the algorithm is $\mathcal{O}(p_o log(p_o))$, while computing the parents of a node can be done in $\mathcal{O}(p_o^2)$. Therefore, the internal loop that starts at line 4 has the complexity of $\mathcal{O}(p_o^2 + |E|)$ and this is repeated at most p_l times. Hence, the loop costs $\mathcal{O}(p_l(p_o^2 + |E|))$. The final cost is $\mathcal{O}(p_l(p_o^2 + |E|) + p + |E|) = \mathcal{O}(p_l(p_o^2 + |E|)) = \mathcal{O}(p^3)$.

The result for Theorem B.4 follows in the same way. Note that a consequence of Lemma A.4 is that when one is interested in identifying the whole matrix, it is not necessary to distinguish between direct and total causal effects. □

Algorithm 2 Total Causal Effect Identification without knowledge of the graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \mathcal{G}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}, (j, i)$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: Compute $\mathrm{de}_o(j)$
4: **while** ID $==$ TRUE **and** $|\mathcal{L}| > 0$ **do**
5: $\quad l \leftarrow \mathcal{L}[1]\{$The first element in the list$\}$
6: $\quad$ Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
7: $\quad$ **if** $\mathrm{ch}(l)[1] = j = j$ **then**
8: $\quad\quad$ Compute $\mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$
9: $\quad\quad$ **if** $i \in \mathrm{de}_o^{\mathcal{G}_{\setminus j}}(l)$ **then** $\{(10)\}$
10: $\quad\quad\quad$ Compute $\mathrm{de}_o(l)$
11: $\quad\quad\quad$ **if** $\mathrm{de}_o(l) = \mathrm{de}_o(j)$ **then**
12: $\quad\quad\quad\quad$ ID $\leftarrow$ FALSE $\{(9)\}$
13: $\quad \mathcal{L} \leftarrow \mathcal{L} \setminus \{l\}$
14: **RETURN:** ID

Algorithm 3 Direct Causal Effect Identification with knowledge of the graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \mathcal{G}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}, (j, i)$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: **for all** $l \in \mathcal{L}$ **do**
4: $\quad$ Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
5: $\quad$ Compute $\mathrm{de}_o(l)$
6: $\quad$ **if** $i \in \mathrm{ch}(l)$ **then** $\{(18)\}$
7: $\quad\quad$ **for all** $k \in \mathrm{ch}(j) \cup \{j\}$ **do**
8: $\quad\quad\quad$ Compute $\mathrm{de}_o(k)$
9: $\quad\quad\quad$ **if** $\mathrm{de}_o(k) = \mathrm{de}_o(l)$ **then** $\{(17)\}$
10: $\quad\quad\quad\quad$ ID $\leftarrow$ FALSE
11: $\quad\quad\quad\quad$ Compute $\mathrm{pa}(j)$
12: $\quad\quad\quad\quad$ **for all** $k_1 \in \mathrm{pa}(j) \cup \{j\}$ **do**
13: $\quad\quad\quad\quad\quad$ **if** $\mathrm{ch}(l) \setminus \mathrm{ch}(k_1) \neq \emptyset$ **then** $\{(20)\}$
14: $\quad\quad\quad\quad\quad\quad$ ID $\leftarrow$ TRUE
15: **RETURN:** ID

Algorithm 4 Direct Causal Effect Identification without knowledge of the graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}, (j, i)$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: **for all** $l \in \mathcal{L}$ **do**
4: $\quad$ Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
5: $\quad$ Compute $\mathrm{de}_o(l)$
6: $\quad$ **if** $i \in \mathrm{ch}(l)$ **then** $\{(12)\}$
7: $\quad\quad$ **for all** $k \in \mathrm{ch}(j) \cup \{j\}$ **do**
8: $\quad\quad\quad$ Compute $\mathrm{de}_o(k)$
9: $\quad\quad\quad$ **if** $\mathrm{de}_o(k) = \mathrm{de}_o(l)$ **then** $\{(11)\}$
10: $\quad\quad\quad\quad$ ID $\leftarrow$ FALSE
11: **RETURN:** ID

Algorithm 5 Matrix Identification with knowledge of the graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \mathcal{G}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: **for all** $l \in \mathcal{L}$ **do**
4: Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
5: $j \leftarrow \mathrm{ch}(l)[1]$
6: Compute $\mathrm{de}_o(l)$ and $\mathrm{de}_o(j)$
7: **if** $\mathrm{de}_o(j) = \mathrm{de}_o(l)$ **then**
8: ID $\leftarrow$ FALSE
9: **for all** $k \in \mathrm{ch}(j)$ **do**
10: **for all** k_1 $\mathrm{pa}(k) \cup \{k\}$ **do**
11: **if** $\mathrm{ch}(l) \setminus \mathrm{ch}(k_1) \neq \emptyset$ **then**
12: ID $\leftarrow$ TRUE
13: **RETURN:** ID

Algorithm 6 Matrix Identification without knowledge of the graph

INPUT: $\mathcal{V} = \mathcal{O} \cup \mathcal{L}, \{\mathrm{ch}(i) \mid i \in \mathcal{V}\}, \mathcal{G}$

1: ID $\leftarrow$ TRUE
2: Sort $\mathcal{V}$ according to an ascending topological order
3: **for all** $l \in \mathcal{L}$ **do**
4: Sort $\mathrm{ch}(l)$ according to the topological order defined in step 2
5: $j \leftarrow \mathrm{ch}(l)[1]$
6: Compute $\mathrm{de}_o(l)$ and $\mathrm{de}_o(j)$
7: **if** $\mathrm{de}_o(j) = \mathrm{de}_o(l)$ **then**
8: ID $\leftarrow$ FALSE
9: **RETURN:** ID

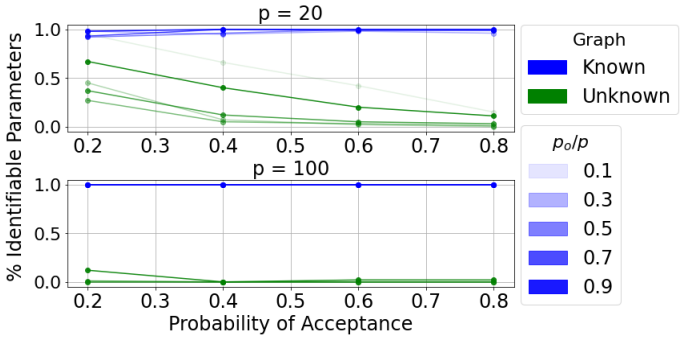


Figure 13. On the x-axis, the probability of acceptance of an edge. The y-axis shows the percentage of graphs in which all the parameters are identifiable

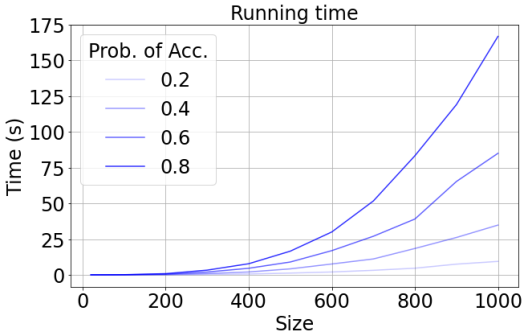


Figure 14. On the x-axis, the size of the graph. On the y-axis, the average running time in seconds, p_o/p is fixed to 0.5.

C. Details on the Experimental Setting and Additional Experiments

C.1. Identification

C.1.1. Mixing Matrix Identification Experiments

We used Algorithm 5 and Algorithm 6 on randomly generated graphs. In Figure 5, we report the percentage of graphs in which all the causal effects are identifiable for graphs of size 10 and 100. The graphs are generated according to an Erdős-Rényi model in which we ensure that the graph we sample is canonical. The probability of acceptance of an edge is plotted on the x-axis. For each setup, we randomly sample 500 graphs.

For the case in which the graph is known, we find the same qualitative behavior observed in Section 5.1. In contrast, when we do not assume the graph to be known, the probability that all the parameters in the graph are identifiable drops drastically. We found the same qualitative behavior with larger graphs.

The average running time for different graph sizes is shown in Figure 14.

C.2. Estimation

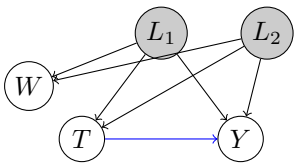
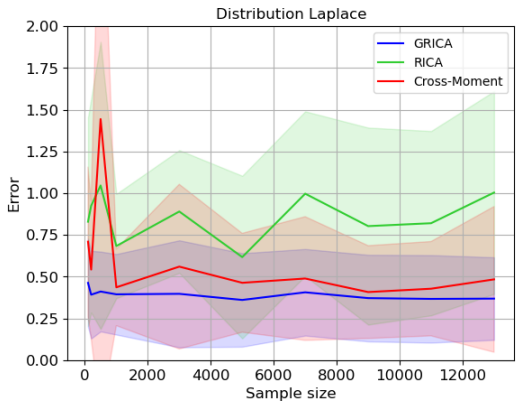
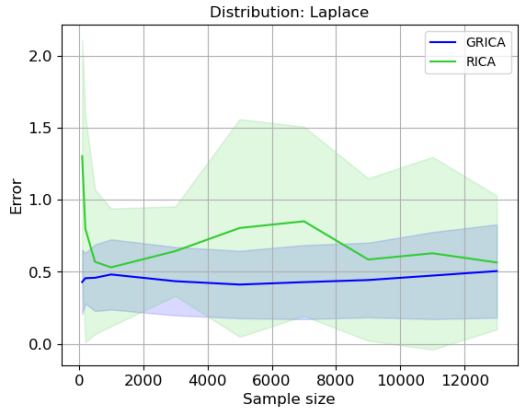


Figure 15. $\mathcal{G}_4$



(a) $\mathcal{G}_4$



(b) Underspecified instrument

Figure 16. Laplace distribution

C.2.1. Sample size vs relative error

In this subsection we present additional experiments on synthetic data to illustrate the performance of GRICA algorithm in comparison with the RICA algorithm implemented by Salehkaleybar et al. (2020) and Cross-Moment method by Kivva et al. (2023).

Experimental setup. All the experiments in this subsection are done on the synthetic data generated according to the specific causal structure established for it. To generate synthetic data we specify all exogenous noises to be i.i.d., and select all non-zero entries within the matrix $\mathbf{A}$ through uniform sampling from $[-1, -0.5] \cup [0.5, 1]$. In the following, we display the results for the setups described in Table 3.

Table 3. Summary of the experimental setups.

Figure	Causal Graph	Distribution			Parameter of Interest
		Family	μ	$scale$	
16a	$\mathcal{G}_4$	Laplace	0	1	$T \to Y$
16b	IV, Figure 4	Laplace	0	1	$T_2 \to Y$
17a	$\mathcal{G}_1$	Exponential	-	1	$T \to Y$
17b	$\mathcal{G}_2$	Exponential	-	1	$T \to Y$
17c	$\mathcal{G}_4$	Exponential	-	1	$T \to Y$
17d	IV, Figure 4	Exponential	-	1	$T_2 \to Y$

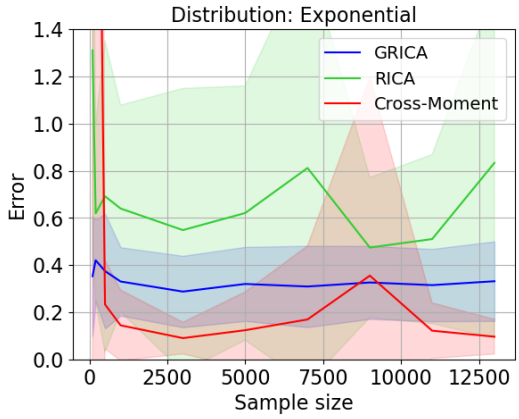
On the figures, we plot the average relative error over 10 independent experiments, with its standard deviation visualized with a transparent area filled with the respective color. From the experiments, we see that GRICA outperforms other methods in all setups, except the one for the graph $\mathcal{G}_1$. Note that for this specific graph, the Cross-Moment method gives us an exact statistical solution that requires computing high-order moments. This results in a less statistically robust estimation when the sample size is small, but it is more accurate if the sample size is large enough.

C.2.2. Model misspecification

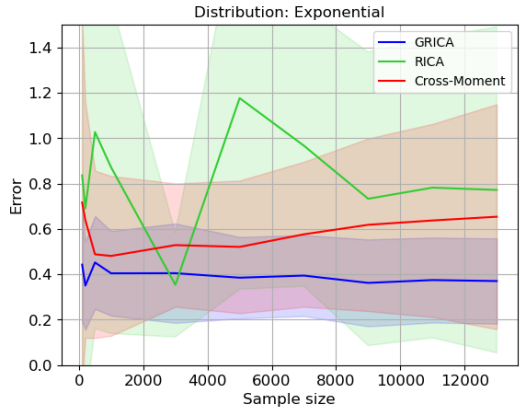
Here we consider the robustness of GRICA algorithm when data generation process is not linear SEM, but close to it. For these experiments the data is generated with the following machanism:

$$\begin{cases} V_j = \tanh\left(\sum_{i=1}^{p} \mathbf{A}_{j,i} V_i\right) + N_j, \text{ if } V_j \neq Y, \\ V_j = \tanh\left(\sum_{i=1, i\neq k}^{p} \mathbf{A}_{j,i} V_i\right) + A_{j,k} V_k + N_j, \text{ if } V_j = Y \text{ and } V_k = T, \end{cases}$$

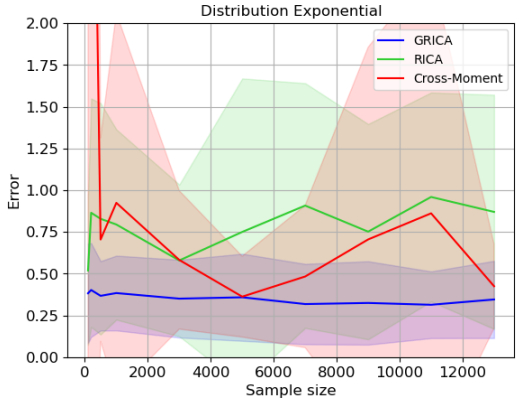
where $T \to Y$ is a direct causal effect that we want to estimate.



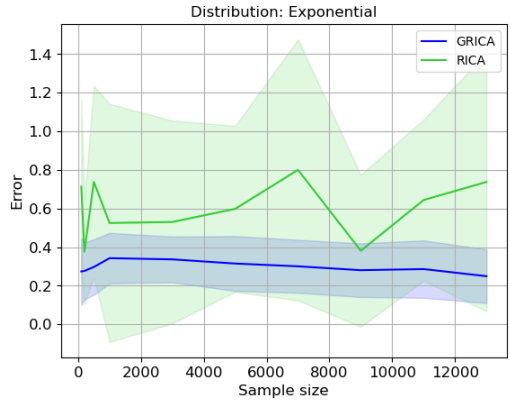
(a) $\mathcal{G}_1$



(b) $\mathcal{G}_2$

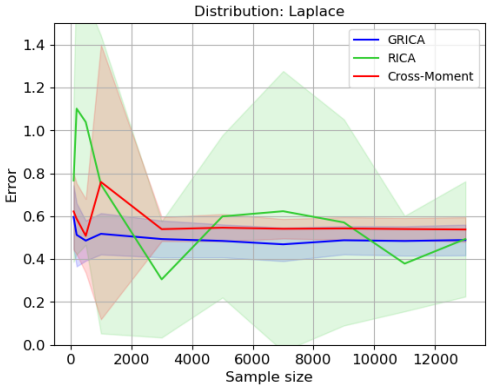


(c) $\mathcal{G}_4$

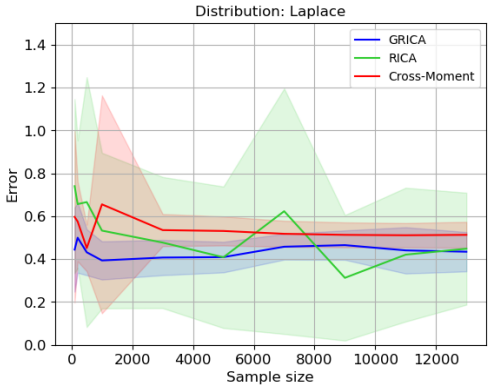


(d) Underspecified instrument

Figure 17. Exponential distribution



(a) $\mathcal{G}_1$



(b) $\mathcal{G}_2$

Figure 18. Laplace distribution

Experimental setup. For each experiment we generate the synthetic data accordingly to the causal structure chosen for it. All exogenous noises are i.i.d., and all non-zero non-diagonal entries within the matrix $\mathbf{A}$ sampled uniformly from $[-1, -0.5] \cup [0.5, 1]$. In the following, we display the results for the setups described in Table 4.

Table 4. Summary of the experimental setups.

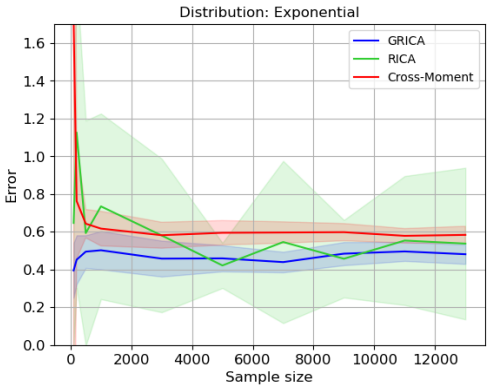
Figure	Causal Graph	Distribution			Parameter of Interest
		Familiy	μ	$scale$	
18a	$\mathcal{G}_1$	Laplace	0	1	$T \to Y$
18b	$\mathcal{G}_2$	Laplace	0	1	$T \to Y$
19a	$\mathcal{G}_1$	Exponential	-	1	$T \to Y$
19a	$\mathcal{G}_2$	Exponential	-	1	$T \to Y$

On the figures, we plot the average relative error over 10 independent experiments, with its standard deviation visualized with transparent area filled with respective color. We observe that GRICA outperforms all other methods for these experiments.

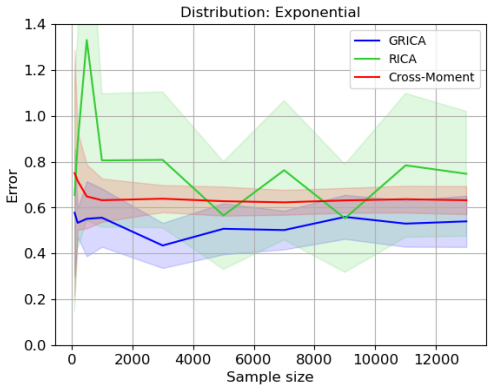
C.2.3. Relative error vs observations noise

These experiments illustrate how the variances of certain exogenous noise impact the performance of GRICA algorithm in comparison to the state-of-the-art algorithms for the considered settings. Herein we considered three main setups. In each of the setup we initialize all the non-zero entries of matrix $\mathbf{A}$ to be equal to 1 and all exogenous noises are modeled as i.i.d. distributions.

1. Causal structure: $\mathcal{G}_1$. The experiment for this causal structure is illustrated in Figure 20a. In this experiment, we initialize all exogenous noises as Exponential distributions. Then we scaled the standard deviation of exogenous noise N_W corresponding to the variable W by a factor displayed on x-axis of the Figure 20a and plotted the performance of GRICA algorithm against the performance of RICA algorithm.
2. Causal structure: $\mathcal{G}_3$. The experiment for this causal structure is illustrated in Figure 20b. In this experiment, we initialize all exogenous noises as Exponential distributions. Then we scaled N_W and N_Z by $Ratio$ and $1/Ratio$, respectively, where $Ratio$ is plotted on the x-axis of Figure 20a and plotted the performance of GRICA algorithm against the performance of RICA, Cross-Moment method and method proposed by Tchetgen et al. (2020).
3. Causal structure: underspecified instrument. The experiments for this causal structure are illustrated in Figures 21a and 21b. In Figure 21a all exogenous noises are modeled as a Laplace distribution and in Figure 21b all exogenous noises are modeled as an Exponential distribution. Then we scaled N_{L_1} and N_{L_2} by a factor displayed on the $x - axis$ of each Figure and estimated the direct causal effect from T_2 to Y with GRICA and RICA algorithms. The relative error of the estimations is plotted in the Figures.

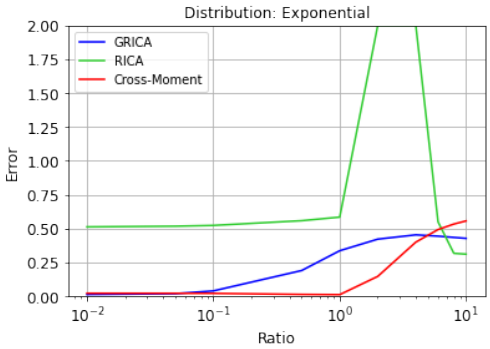


(a) $\mathcal{G}_1$

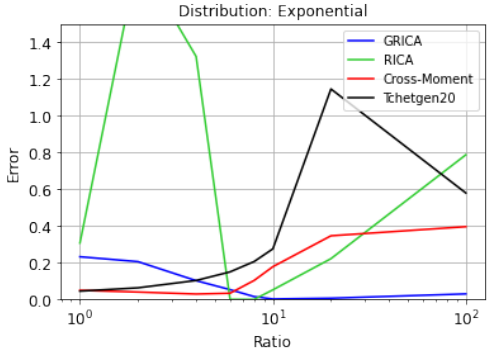


(b) $\mathcal{G}_2$

Figure 19. Exponential distribution

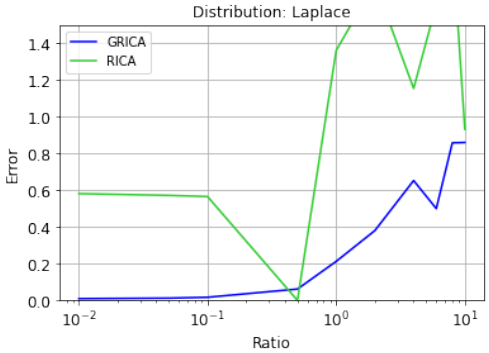


(a) $\mathcal{G}_1$

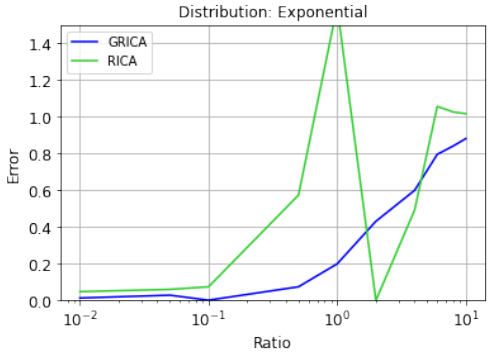


(b) $\mathcal{G}_3$

Figure 20. Exponential distribution



(a) Laplace distribution



(b) Exponential distribution

Figure 21. Underspecified insturment