Statistical Undecidability in Linear, Non-Gaussian Causal Models in the Presence of Latent Confounders

Abstract

Since Spirtes et al. (2000), it is well known that if causal relationships are linear and noise terms are independent and Gaussian, causal orientation is not identified from observational data — even if causal faithfulness is satisfied. Shimizu et al. (2006) showed that linear, **non**-Gaussian (LiNGAM) causal models are identified from observational data, so long as no latent confounders are present. That holds even when faithfulness fails. Genin and Mayo-Wilson (2020) refine that identifiability result: not only are causal relationships identified, but causal orientation is *statistically decidable*. That means that for every $\epsilon > 0$, there is a method that converges in probability to the correct orientation and, at every sample size, outputs an incorrect orientation with probability less than ϵ. These results naturally raise questions about what happens in the presence of latent confounders. Hoyer et al. (2008) and Salehkaleybar et al. (2020) show that, although the causal model is not uniquely identified, causal orientation among observed variables is identified in the presence of latent confounders, so long as faithfulness is satisfied. This paper refines these results. When we allow for the presence of latent confounders, causal orientation is no longer statistically decidable. Although it is possible to converge in probability to the correct orientation, it is not possible to do so with finite-sample bounds on the probability of orientation errors. That is true even if causal faithfulness is satisfied. That limiting result allows us to correctly calibrate our attitudes toward the outputs of causal discovery methods in the LiNGAM framework.

1 Introduction

Spirtes et al. [2000] develop the elements of causal discovery from observational data in the linear Gaussian setting. They show that when functional relationships between variables are linear and noise terms are independent and Gaussian, it is possible to converge to the Markov *equivalence class* of the graph generating the data. Although some causal information can be recovered, causal orientation is often not identified: two linear Gaussian models may differ in causal orientation and nevertheless generate the exact same distribution over the observed variables. Identifiability fails even when causal faithfulness is satisfied and no hidden variables are present. For this reason, it was a significant advance when Shimizu et al. [2006] showed that, when functional relationships between variables are linear and noise terms are independent and *non*-Gaussian, the model could be uniquely identified from observational data, even without assuming faithfulness. However, that early result depends on the absence of unobserved confounders. Since then, the LiNGAM framework, as it came to be called, has been extended to accommodate the presence of hidden variables [Hoyer et al., 2008, Salehkaleybar et al., 2020]. For example, Salehkaleybar et al. [2020] prove that if, in addition to the usual LiNGAM assumptions, we assume causal faithfulness, then causal orientation between observed variables is identified even in the presence of unobserved confounders.[1]

[1] Unfortunately, although all models generating the same distribution over the observed variables must agree on the causal ancestry relations between them, they may disagree on the strength of the causal effects (Figure 2).

34th Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020

These identifiability results are exciting theoretical developments. However, identifiability is a weak criterion and on its own does not entail the existence of a consistent discovery algorithm. Moreover, distinctions ought to be made between consistent algorithms. For example, *uniform* consistency requires that one be able to determine *a priori* the sample size at which the chance of identifying the true model are at least $1 - \alpha$. Unfortunately, it is easy to show that there is no uniformly consistent algorithm for determining the direction of a causal edge, even in unconfounded LiNGAMs.[2] However, Genin and Mayo-Wilson [2020] show that the direction of a causal edge in an unconfounded LiNGAM is *statistically decidable.* Statistical decidability is a reliability concept intermediate between the familiar notions of consistency and uniform consistency. Causal orientation is statistically decidable, if for any $\alpha > 0$, there is a consistent procedure that, *at every sample size,* hypothesizes a false orientation with chance less than α. Such procedures may exist even when uniformly consistent ones do not.

The existence of statistical decision procedures is by no means guaranteed by identifiability. For example, Kelly and Mayo-Wilson [2010] show that in the unconfounded linear Gaussian setting, even in situations in which causal orientation is identifiable, it is possible to force any consistent procedure to "flip" its judgement about whether X causes Y or vice versa, *no matter how strong the effect* of X on Y. Furthermore, even in the absence of confounders, the number of such flips is bounded only by the number of variables in the model. The results of Genin and Mayo-Wilson [2020] show that such flipping behavior can be avoided in the unconfounded LiNGAM setting. The main result of this paper is that flipping returns in the confounded LiNGAM setting, even when we assume causal faithfulness. Although consistent procedures exist for learning causal orientation, consistent *decision* procedures do not.

Table 1: Causal Orientation in LiNGAM models

	Unconfounded	**Potentially Confounded**
Faithful	decidable	decidable in the limit
Unfaithful	decidable	not identified

Let $\mathcal{M}$ be a set of statistical models. We assume there is a function $P : M \mapsto P_M$ that maps each model in $\mathcal{M}$ to a probability measure over a space Ω of observed outcomes, although we often do not distinguish between a random vector and the probability measure induced by its distribution function. Henceforth, we assume $\Omega = \mathbb{R}^p$. We lift $P(\cdot)$ to sets of models in the obvious way: if $\mathcal{A} \subseteq \mathcal{M}$, let $P[\mathcal{A}] = \{P(M) : M \in \mathcal{A}\}$. If $A \subseteq \mathbb{R}^{nd}$, let ∂A be the boundary of A in the usual topology on $\mathbb{R}^{nd}$. Let $\mathcal{P} := P[\mathcal{M}]$ denote the set of all probability measures associated with the models in $\mathcal{M}$. The **weak topology** on $\mathcal{P}$ is defined by letting a sequence of Borel measures P_n converge weakly to P, written $P_n \Rightarrow P$ iff $P_n(A) \to P(A)$, for every A such that $P(\partial A) = 0$. A collection of random vectors $(\mathbf{X}_n)$ converges in distribution to $\mathbf{X}$ iff the probability measures induced by the $\mathbf{X}_n$ converge weakly to the measure induced by $\mathbf{X}$. We write $\mathsf{cl}(\cdot)$ for the closure operator in the weak topology. We write $\mathsf{fr}A$ for $\mathsf{cl}A \setminus A$. We say that a set is **locally closed** iff it is the intersection of an open and a closed set. In metrizable topologies such as the topology of weak convergence, every open set, and therefore every locally closed set, is a countable union of closed sets. For any natural number k, let P_M^k be the k-fold product measure of P_M with itself. This measure describes the probabilities of events in $\mathbb{R}^{kd}$ when we take k iid samples from P_M. If the measures P_n converge weakly to P, the product measures P_n^k also converge weakly to P^k (see Theorem 2.8 in Billingsley [1986]).

We define a **question** $\mathfrak{Q}$ to be a countable set of disjoint subsets of $\mathcal{M}$. The elements of $\mathfrak{Q}$ are called **answers**. For all $M \in \cup\mathfrak{Q}$, let $\mathfrak{Q}(M)$ denote the unique answer in $\mathfrak{Q}$ containing M. The answer to question $\mathfrak{Q}$ is **identified** iff $P(M) \neq P(M')$ whenever $\mathfrak{Q}(M) \neq \mathfrak{Q}(M')$. Given a question $\mathfrak{Q}$, we define a **method** $\lambda = \langle \lambda_n \rangle_{n \in \mathbb{N}}$ to be a sequence of measurable functions $\lambda_n : \Omega^n \to \mathfrak{Q} \cup \{\mathcal{M}\}$, where λ_n maps samples of size n to answers to the question; a method may also take the value $\mathcal{M}$ to indicate that the data do no fit any particular answer sufficiently well, and so we call $\mathcal{M}$ the **uninformative answer**. We require that $\partial\lambda_n^{-1}(\mathcal{A})$ has Lebesgue measure zero for all n and every answer $\mathcal{A}$ in the range of λ_n.

A method is (pointwise) **consistent** for $\mathcal{Q}$ if for all $\epsilon > 0$ and $M \in \cup\mathfrak{Q}$, there is n such that $P_M^k(\lambda_k = \mathfrak{Q}(M)) > 1 - \epsilon$ for all $k \geq n$. We say that $\mathfrak{Q}$ is **decidable in the limit** iff there is a

[2] See Example 1 in Genin and Mayo-Wilson [2020]. For strong additional assumptions ensuring the existence of uniformly consistent procedures, see Bühlmann et al. [2014].

consistent method for $\mathfrak{Q}$. Dembo and Peres [1994] give the following sufficient condition for limiting decidability.

Theorem 1.1 (Dembo and Peres [1994])**.** *$\mathfrak{Q}$ is decidable in the limit if $\{P(A) : A \in \mathfrak{Q}\}$ is disjoint and each $P(A)$ is a countable union of sets closed in the weak topology.*

Proof of Theorem 1.1. See Dembo and Peres [1994], Corollary 2. □

Given some $\alpha > 0$, say that a method λ is an **α-decision procedure** for $\mathfrak{Q}$ if (1) λ is consistent for $\mathfrak{Q}$ and (2) $P^n_M(M \notin \lambda_n) \leq \alpha$ for all $M \in \cup\mathfrak{Q}$ and all sample sizes n. In other words: an α-decision procedure outputs a false hypothesis with probability at most α. A question is **statistically decidable** (or simply decidable) if there is an α-decision procedure for $\alpha > 0$. We give a simple necessary condition for statistical decidability.

Theorem 1.2. *$\mathfrak{Q}$ is statistically decidable only if there are no $\mathcal{A}, \mathcal{B} \in \mathfrak{Q}$ such that $P(\mathcal{A}) \cap \mathsf{cl}(P(\mathcal{B}))$ is not empty and contains a measure absolutely continuous with Lebesgue measure on $\mathbb{R}^p$.*

Proof of Theorem 1.2. Suppose for a contradiction that $P_M \in P(\mathcal{A}) \cap \mathsf{cl}(P(\mathcal{B}))$ is absolutely continuous with Lebesgue measure and λ is an α-decision procedure for $\mathfrak{Q}$. Since λ is consistent for $\mathfrak{Q}$, there must be n such that $P^n_M(\lambda_n^{-1}(\mathcal{A})) > 1 - \alpha$. Since $\partial\lambda_n^{-1}(\mathcal{A})$ has Lebesgue measure zero and P^n_M is absolutely continuous with Lebesgue measure on $\mathbb{R}^{np}$, $P^n_M(\partial\lambda_n^{-1}(\mathcal{A})) = 0$.[3] Therefore, there are (M_i) in $\mathcal{B}$ such that $P^n_{M_i}(\mathcal{A}) \to P^n_M(\mathcal{A})$. But then there is some $M_j \in \mathcal{B}$ such that $P^n_{M_j}(M_j \notin \lambda_n) > \alpha$. Contradiction. □

It is worth introducing some intuitive language for questions $\mathfrak{Q}$ with only one (usually non-exhaustive) answer $\mathcal{A} \subseteq \mathcal{M}$. We say that $\mathcal{A}$ is **statistically verifiable** iff $\mathfrak{Q} = \{\mathcal{A}\}$ is decidable. Say that $\mathcal{A}$ is **statistically refutable** iff $\mathfrak{Q} = \{\mathcal{M} \setminus \mathcal{A}\}$ is decidable. For partial converses of Theorems 1.1 and 1.2, see Genin and Kelly [2017]. Essentially, the converses hold straightforwardly if all distributions are assumed absolutely continuous with Lebesgue measure.

2 Background: Linear Causal Models

A **linear causal model in d variables** M is a triple $\langle \mathbf{X}, \mathbf{e}, A\rangle$, where $\mathbf{X} = \langle X_i\rangle$ is a vector of d random variables, $\mathbf{e} = \langle e_1, e_2, \ldots, e_d\rangle$ is a random vector of d exogenous noise terms, and B is a $d \times d$ matrix such that

1. Each variable X_i is a linear function of variables earlier in the order, plus an unobserved noise term e_i:
$$X_i(\omega) = \sum_{j<i} A_{ij} X_j(\omega) + e_i(\omega);$$
2. the noise terms $e_1, \ldots, e_d$ are mutually independent.

In matrix notation, we have that $\mathbf{X} = A\mathbf{X} + \mathbf{e}$. Because no X_i causes itself, A has only zeroes along its diagonal. By virtue of the causal order, A is lower triangular, i.e. all elements above the diagonal are zero. The random vector $\mathbf{X}$ also admits a "dual" representation: $\mathbf{X} = B\mathbf{e}$, where $B = (I - A)^{-1}$. To see that B always exists, note that the determinant of triangular matrix is equal to the product of its diagonal entries. Since the inverse of a lower triangular matrix is lower triangular, the matrix B is also lower triangular, however its diagonal elements are all equal to one. If $M = \langle \mathbf{X}, \mathbf{e}, A\rangle$, let $|M|$ be equal to the length of the vector $\mathbf{X}$. Moreover, let $\mathbf{X}(M), \mathbf{e}(M), A(M)$ and $B(M)$ be $\mathbf{X}, \mathbf{e}, A$ and $(I - A)^{-1}$, respectively. The relationship between $A(M)$ and $B(M)$ will be made more perspicuous in the following.

Write $j \to_M i$ as a shorthand for $A_{ij}(M) \neq 0$. The relation $\to_M$ defines a directed acyclic graph $G(M)$ over the vertices $\{1, \ldots, |M|\}$. A causal path of length m from i to j in $G(M)$ is a sequence of vertices $\pi = (v_1, \ldots, v_m)$ such that $v_1 = i$, $v_m = j$ and $v_i \to_M v_{i+1}$. Let $\Pi^n_{ij}(M)$ be the set of all causal paths of length n from i to j in $G(M)$. Let $\Pi_{ij}(M)$ be the set of all causal paths from i to j in $G(M)$. Let $\Pi(M)$ be the set of all causal paths in $G(M)$. Write $i \rightsquigarrow_M j$ as a shorthand

[3] It is a basic fact that if $\mu << \nu$ then $\mu^n << \nu^n$.

for $\Pi_{ij}(M) \neq \varnothing$. Write $j \circ_M i$ when $j \not\to_M i$ and $j \not\leftarrow_M i$. If $\pi = (v_1, \ldots, v_n)$ is a sequence of vertices in $\{1, \ldots, |M|\}$, let the **path product** $\bigtimes_M \pi$ be the product of all causal coefficients along the path π in $G(M)$, i.e. $\bigtimes_M \pi = \prod_{i=1}^{n} A_{v_{i+1},v_i}(M)$. Note that if $\pi \in \pi(M)$ iff $\bigtimes_M \pi \neq 0$.

It is easy to verify that

$$A^k_{ij}(M) = \sum_{\pi \in \Pi^k_{ji}(M)} \times_M \pi.$$

In other words, $A^k_{ij}(M)$ is the sum of all path products for paths of length k from i to j. So $A^k_{ij}(M) \neq 0$ implies $j \rightsquigarrow_M i$. The converse is not necessarily true, since non-zero path products may sum to zero. By a result of Carl Neumann's, $B(M) = \sum_{k=0}^{|M|} A^k(M)$.[4] So $B_{ij}(M) = \sum_{\pi \in \Pi_{ji}(M)} \bigtimes_M \pi$. In other words, $B_{ij}(M)$ is the sum of all path products for paths from i to j. So $B_{ij}(M) \neq 0$ implies $j \rightsquigarrow_M i$. The converse does not necessarily hold since non-zero path products may sum to zero. We say that model M is **faithful** if $B_{ij}(M) \neq 0$ whenever $j \rightsquigarrow_M i$.

A linear causal model M is non-Gaussian (a LiNGAM) if in addition to satisfying (1) and (2), each of the noise terms is *non-Gaussian*. Let LIN_d be the class of all linear causal models on d variables, and let LNG_d, FLNG_d respectively denote the classes of non-Gaussian models and faithful non-Gaussian models. Similarly, LIN, LNG and FLNG respectively represent the classes of all linear causal models, all linear non-Gaussian models, and all faithful linear non-Gaussian models over some finite number of variables. It is sometimes reasonable to introduce a priori constraints on the maximum size of a coefficient in a LiNGAM model. For example, if c is the number of particles in the universe, let FLNG^c be the set $\{M \in \text{FLNG} : \max_{i,j} |B_{ij}(M)| < c\}$. Let FLNG^c_d be $\text{FLNG}^c \cap \text{FLNG}_d$.

We introduce some notation for manipulating matrices. Suppose A is a $n \times p$ and U, V are subsets of $\{1, \ldots, n\}, \{1, \ldots, p\}$, respectively. Let $A_{[U;V]}$ be the result of dropping all rows from A that are not in U and all columns that are not in V. Let $A_{(U;V)}$ be the result of dropping all rows from A that are in U and all columns that are in V. For singleton sets $\{i\}$ we drop the braces so that $A_{(i,j)}$ is the result of dropping columns i and j from A. We retain the usual notation $A_{ij} = A_{[i,j]}$. Say that a matrix A has **pairwise linearly independent** columns iff no column of A is proportional to any other.

3 Parsimonious Models

Let $\mathcal{O}$ be the set of all probability distributions on $\mathbb{R}^p$. We are interested in when the same vector of observed random variables could have arisen from distinct causal models. Accordingly, say that a random vector $\mathbf{O} = (O_1, \ldots, O_p) \in \mathcal{O}$ **admits** a LiNGAM model $M \in \text{LNG}_d$ if there is a permutation α of $(1, \ldots, d)$ such that $O_i = X_{\alpha^{-1}(i)}(M)$ for $1 \leq i \leq p$. In other words: $\mathbf{O} = (O_1, \ldots, O_p)$ admits M if there is a way to order the d variables of $X(M)$ such that the first p are identical with $O_1, \ldots, O_p$. We say that the permutation α **embeds** $\mathbf{O}$ into M. If α embeds $\mathbf{O}$ into M, then

$$\mathbf{O} = B_{\mathbf{O}}(M)\mathbf{e}_{\mathbf{O}}(M),$$

where $B_{\mathbf{O}}(M)$ is the first p rows of $P_\alpha B(M) P_\alpha$, $\mathbf{e}_{\mathbf{O}}(M)$ is $P_\alpha \mathbf{e}(M)$ and P_α is the permutation matrix corresponding to α. Extend the causal order over the elements of M to the O_i by setting $O_i \rightsquigarrow_M O_j$ if $\alpha^{-1}(i) \rightsquigarrow_M \alpha^{-1}(j)$ and $O_i \circ_M O_j$ if $\alpha^{-1}(i) \circ_M \alpha^{-1}(j)$.

Say that $\mathbf{O}$ **admits a LiNGAM model** if there is d such that $\mathbf{O}$ admits $M \in \text{LNG}_d$. We say that a model $M \in \text{LNG}_d$ is **parsimonious** for $\mathbf{O}$ if $\mathbf{O}$ admits M and $\mathbf{O}$ admits no M' in LNG_f with $f < d$. It is immediate that if $\mathbf{O}$ admits a LiNGAM model, it admits some parsimonious LiNGAM model. For $A \in \{\text{LNG}_d, \text{LNG}^c_d, \text{FLNG}_d, \text{FLNG}^c_d\}$ Let $\mathcal{O}_A \subset \mathcal{O}$ be the set

$$\{\mathbf{O} \in \mathcal{O} : (\exists M \in A) M \text{ is parsimonious for } \mathbf{O}\}.$$

For $A \in \{\text{LNG}, \text{LNG}^c, \text{FLNG}, \text{FLNG}^c\}$ Let $\mathcal{O}_{A_{\leq d}} = \cup_{j \leq d} \mathcal{O}_{A_j}$ and $\mathcal{O}_{A_{\geq d}} = \cup_{j \geq d} \mathcal{O}_{A_j}$. Let $\mathcal{O}_{A_{<d}}, \mathcal{O}_{A_{>d}}$ be defined similarly. Finally, let $\mathcal{O}_A = \mathcal{O}_{A_{\geq p}}$.

[4] The *spectral radius* $\rho(A)$ *of a square matrix* A *is the largest absolute value of its eigenvalues. Neumann's result states that if* $\rho(A) < 1$ *then* $(I - A)^{-1}$ *exists and is equal to* $\sum_{k=0}^{\infty} A^k$. *Since the eigenvalues of a triangular matrix are exactly its diagonal entries,* $\rho(B(M)) = 0$ *for any linear causal model* M. *By acyclicity, there are no paths longer than* $|M|$, *so* $\sum_{k>d} A^k = 0$.

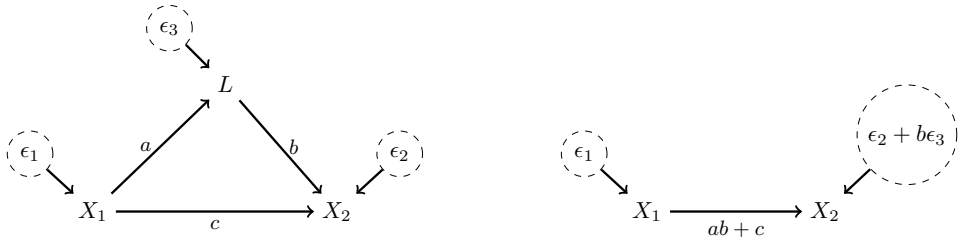

Figure 1: If M is the left-hand model and $\mathbf{O} = (X_1, X_2)$, then $B_{\mathbf{O}}(M) = \begin{pmatrix} 1 & 0 & 0 \\ ab + c & 1 & b \end{pmatrix}$ and the second and third columns are proportional. The right-hand model is in fewer variables and generates the same distribution over $\mathbf{O}$.

The following Lemma says that if M is faithful and parsimonious for $\mathbf{O}$ then no column of $B_{\mathbf{O}}(M)$ is proportional to any other. The proof idea is expressed by Figure 1: by adjusting the edge coefficient from X_1 to X_2, it is possible to "absorb" L into the noise term of X_2 without changing the distribution of (X_1, X_2) or violating the LiNGAM model assumptions. The proof is a relatively straightforward generalization of the example, but since it is rather lengthy we relegate it to the Appendix. A related result is given by Salehkaleybar et al. [2020, Theorem 11].

Lemma 3.1. *Suppose that* $\mathbf{O}$ *admits faithful* $M \in \text{LNG}_d$ *and that some column of* $B_{\mathbf{O}}(M)$ *is proportional to another. Then there is* $M' \in \text{LNG}_{d-1}$ *such that (i)* $\mathbf{O}$ *admits* M' *and (ii)* $O_i \rightsquigarrow_M O_j$ *iff* $O_i \rightsquigarrow_{M'} O_j$ *and (iii)* M' *is faithful.*

Lemma 3.1 raises a question about the converse: is it also the case that if $B_{\mathbf{O}}(M)$ has no two proportional columns, then M is parsimonious for $\mathbf{O}$? The following Theorem from Kagan et al. [1973] allows us to answer the question in the affirmative. We will appeal to this Theorem several times in the following.

Theorem 3.1. *Suppose that* $\mathbf{X} = A\mathbf{e} = B\mathbf{f}$, *where* A *and* B *are* $p \times r$ *and* $p \times s$ *matrices and* $\mathbf{e} = (e_1, \ldots, e_r)$, $\mathbf{f} = (f_1, \ldots, f_s)$ *are random vectors with independent components. Suppose that no two columns of* A *are proportional to each other. If the* i*-th column of* A *is not proportional to any column of* B*, then* e_i *is normally distributed.*

Theorem 3.2. *Suppose that faithful* $M \in \text{LNG}_d$*. Then,* M *is parsimonious for* $\mathbf{O} = (O_1, \ldots, O_p)$ *iff no column of* $B_{\mathbf{O}}(M)$ *is proportional to any other.*

Proof of Theorem 3.2. The left to right implication is immediate from Lemma 3.1. To prove the converse, suppose that M is not parsimonious for $\mathbf{O}$. Then, $\mathbf{O}$ admits some $M' \in \text{LNG}_f$ with $f < d$. Moreover $\mathbf{O} = B_{\mathbf{O}}(M)\mathbf{e}_{\mathbf{O}}(M) = B_{\mathbf{O}}(M')\mathbf{e}_{\mathbf{O}}(M')$, where $B_{\mathbf{O}}(M)$ is an $p \times d$ matrix and $B_{\mathbf{O}}(M')$ is a $p \times f$ matrix. By Theorem 3.1, every column of $B_{\mathbf{O}}(M)$ is proportional to some column of $B_{\mathbf{O}}(M')$. Since the latter has fewer columns, there must be two distinct columns of $B_{\mathbf{O}}(M)$ that are proportional to the same column of $B_{\mathbf{O}}(M')$ and, therefore, to each other. □

We close this section with an easy corollary of Lemma 3.1, which we will appeal to in the following.

Corollary 3.1. *Suppose that* $\mathbf{O}$ *admits faithful* $M \in \text{LNG}$*. Then there is faithful* $M' \in \text{LNG}$ *such that (i)* $\mathbf{O}$ *admits* M' *(ii)* M' *is parsimonious for* $\mathbf{O}$ *and (iii)* $O_i \rightsquigarrow_M O_j$ *iff* $O_i \rightsquigarrow_{M'} O_j$*.*

Proof of Corollary 3.1. Suppose $\mathbf{O}$ admits faithful $M \in \text{LNG}$ that is not parsimonious for $\mathbf{O}$. By repeated appeal to Lemma 3.1, we must eventually arrive at some M' that is parsimonious for $\mathbf{O}$, either because no column of $B_{\mathbf{O}}(M)$ is proportional to any other (see Theorem 3.2), or because $M' \in \text{LNG}_p$ has no latent variables. □

4 Causal Identifiability

Genin and Mayo-Wilson [2020] prove that if a vector of p observed variables admits a model in p variables, that model is unique.

Theorem 4.1. *Suppose that* $= (O_1, \ldots, O_p)$ *admits* $M, M' \in \text{LNG}_p$*, then* $M = M'$*.*

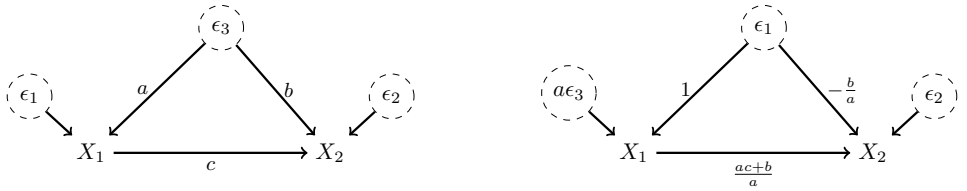

Figure 2: Although the left and right-hand models generate the same distribution over (X_1, X_2) they disagree on the effect of intervening on X whenever $b \neq 0$. When $ac = -b$, the lhs model is unfaithful and the models disagree, not only on the size of the effect, but on the presence of an edge.

Proof of Theorem 4.1. See Theorem 2.3 in Genin and Mayo-Wilson [2020]. □

Unfortunately, that is no longer the case when latent variables are present. Of course, if $\mathbf{O}$ admits a LiNGAM without latents, it also admits one with latents. Figure 1 yields such an example: in the left-hand model the causal relationship is mediated by L and in the right-hand model the causal relationship is unmediated. That situation is not too worrisome, since the effect of an ideal intervention on X_1 would be the same in both circumstances. What is more worrisome is that the vector of observed variables $\mathbf{O}$ may admit two LiNGAM models that *differ* on the effects of interventions. An example, due to Salehkaleybar et al. [2020], is given in Figure 2. The good news is that if a set of observed variables admits two faithful LiNGAM models, the models must agree on the ancestor relationship between them. Although this is shown already by Salehkaleybar et al. [2020, Lemma 5], the following is a simple proof that does not rely on facts about independent component analysis.

Theorem 4.2. *Suppose that* $\mathbf{O} = (O_1, \ldots, O_p)$ *admits faithful* M, M'. *Then* $O_i \rightsquigarrow_M O_j$ *iff* $O_i \rightsquigarrow_{M'} O_j$

Proof of Theorem 4.2. By Corollary 3.1 there are faithful LiNGAMs F, F' such that 1. $\mathbf{O}$ admits F, F'; 2. $O_i \rightsquigarrow_M O_j$ iff $O_i \rightsquigarrow_F O_j$; 3. $O_i \rightsquigarrow_{M'} O_j$ iff $O_i \rightsquigarrow_{F'} O_j$ and 4. $B_{\mathbf{O}}(F)$ and $B_{\mathbf{O}}(F')$ both have pairwise linearly independent columns. By (1) and (2), it suffices to prove that $O_i \rightsquigarrow_F O_j$ iff $O_i \rightsquigarrow_{F'} O_j$. But since the situation is symmetrical, it suffices to prove that $O_i \rightsquigarrow_F O_j$ only if $O_i \rightsquigarrow_{F'} O_j$.

Suppose for a contradiction that $O_i \rightsquigarrow_F O_j$ but $O_i \not\rightsquigarrow_{F'} O_j$. Let α be a permutation embedding $\mathbf{O}$ in F. Let B, C be $B_{\mathbf{O}}(F), B_{\mathbf{O}}(F')$, respectively. Let $\mathbf{e}, \mathbf{f}$ be $\mathbf{e}_{\mathbf{O}}(F), \mathbf{e}_{\mathbf{O}}(F')$, respectively. Then

$$\mathbf{O} = B\mathbf{e} = C\mathbf{f}.$$

Since $O_i \not\rightsquigarrow_{F'} O_j$, $C_{ji} = 0$. Moreover, $C_{ii} = 1$ By faithfulness of F, $O_i \rightsquigarrow_F O_j$ implies that $B_{ji} \neq 0$. By Theorem 3.1, there must be a column $k \neq i$ and real number $a \neq 0$ such that $B_{ik} = aC_{ii} \neq 0$ but $B_{jk} = aC_{ji} = 0$. Since $B_{ik} \neq 0$, it follows that $\alpha^{-1}(k) \rightsquigarrow_F \alpha^{-1}(i)$. Since $O_i \rightsquigarrow_F O_j$ by assumption, it follows that $\alpha^{-1}(i) \rightsquigarrow_F \alpha^{-1}(j)$. By transitivity of $\rightsquigarrow_F$, $\alpha^{-1}(k) \rightsquigarrow_F \alpha^{-1}(j)$. However, $B_{jk} = 0$. So F is unfaithful. Contradiction. □

For $A \in \{\text{LNG}, \text{LNG}^c, \text{LNG}^c_d, \text{FLNG}, \text{FLNG}^c, \text{FLNG}^c_d\}$, let

$$\mathcal{O}_A^{i \to j} = \{\mathbf{O} \in \mathcal{O}_A : (\exists M \in A)\ \mathbf{O} \text{ admits } M \text{ and } O_i \to_M O_j\};$$
$$\mathcal{O}_A^{i \rightsquigarrow j} = \{\mathbf{O} \in \mathcal{O}_A : (\exists M \in A)\ \mathbf{O} \text{ admits } M \text{ and } O_i \rightsquigarrow_M O_j\};$$
$$\mathcal{O}_A^{i \circ j} = \{\mathbf{O} \in \mathcal{O}_A : (\exists M \in A)\ \mathbf{O} \text{ admits } M \text{ and } O_i \circ_M O_j\}.$$

In light of Theorem 4.2, $\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}}$, $\mathcal{O}^{i \nrightsquigarrow j}_{\text{FLNG}}$ and $\mathcal{O}^{i \circ j}_{\text{FLNG}}$ are disjoint.

5 The Topology of Latent Confounding

When $\mathbf{O} = (O_1, \ldots, O_p)$ admits a LiNGAM model without latents, Genin and Mayo-Wilson [2020, Theorem 4.1] prove that orientation hypotheses are topologically well-separated:

Lemma 5.1. $\mathcal{O}^{i \to j}_{\text{LNG}^c_p}, \mathcal{O}^{i \leftarrow j}_{\text{LNG}^c_p}$ *are open and* $\mathcal{O}^{i \circ j}_{\text{LNG}^c_p}$ *is closed in the weak topology on* $\mathcal{O}_{\text{LNG}^c_p}$.

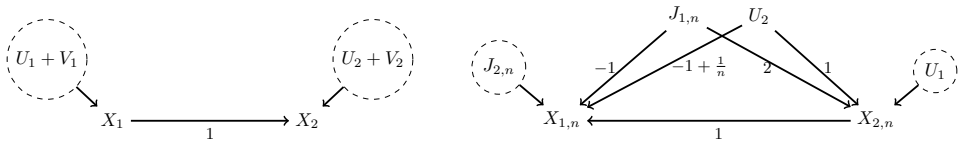

Figure 3: The $(X_{1,n}, X_{2,n})$, which lie in $\mathcal{O}^{1\leftarrow 2}_{\text{FLNG}^c_4}$, converge in probability to (X_1, X_2), lying in $\mathcal{O}^{1\to 2}_{\text{FLNG}^c_2}$. Note that although error terms approach Gaussianity and the model approaches unfaithfulness, no term in the sequence is unfaithful and no noise term is Gaussian. For definitions of error and exogenous terms, see the proof of Lemma 5.2.

The situation changes when we allow for latent variables.

Lemma 5.2. *$\mathcal{O}^{i\to j}_{\text{FLNG}^c_p}$ is not disjoint from $\mathsf{cl}(\mathcal{O}^{i\leftarrow j}_{\text{FLNG}^c_{p+2}})$ in the weak topology on $\mathcal{O}_{\text{FLNG}^c}$. Moreover, there are distributions in the intersection that are absolutely continuous wrt Lebesgue measure.*

Proof of Lemma 5.2. Let $p = 2$. Let $U_1, U_2, W_1, W_2, Z_1, Z_2$ be mutually independent, absolutely continuous random variables. Suppose that all variables except Z_1, Z_2 are non-Gaussian. Let $V_1 = Z_1 + Z_2$ and let $V_2 = Z_1 - Z_2$. By the Lukacs-King theorem, V_1, V_2 are independent. Let $\mathbf{X} = (X_1, X_2) = (U_1 + V_1, U_1 + U_2 + 2Z_1)$. By reference to the lhs model in Figure 3, it is clear that $\mathbf{X} \in \mathcal{O}_{\text{FLNG}^c_2}$. Moreover, since sums of independent absolutely continuous random variables are absolutely continuous, X_1, X_2 are absolutely continuous wrt Lebesgue measure on $\mathbb{R}$. Since products of absolutely continuous measures are absolutely continuous wrt the product, $\mathbf{X}$ is absolutely continuous wrt Lebesgue measure on $\mathbb{R}^2$.

For $n > 2$, let $X_{1,n} = U_1 + V_1 + \frac{1}{n}(W_1 + W_2 + U_2)$ and $X_{2,n} = U_1 + U_2 + 2Z_1 + \frac{2}{n}W_1$. Let $\mathbf{X}_n = (X_{1,n}, X_{2,n})$. It is clear that the $\mathbf{X}_n$ converges in probability, and therefore in distribution, to $\mathbf{X}$. It remains to show that the $\mathbf{X}_n$ lie in $\mathcal{O}_{\text{FLNG}^c_4}$, which we do by reference to the rhs model in Figure 3. Let $J_{1,n} = Z_1 + \frac{1}{n}W_1$ and $J_{2,n} = Z_2 + \frac{1}{n}W_2$. Then $J_{1,n}, J_{2,n}, U_1, U_2$ are independent and non-Gaussian. Let $\mathbf{e}_n^T = (J_{2,n}, U_1, J_{1,n}, U_2)$. Let $A_n = \begin{pmatrix} 0 & 1 & -1 & -1 + \frac{1}{n} \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ and $B_n = \begin{pmatrix} 1 & 1 & 1 & \frac{1}{n} \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. It is easy to check that $B_n = (I - A_n)^{-1}$. Let $M_n = \langle B_n\mathbf{e}_n, A_n, \mathbf{e}_n\rangle$. By inspection of B_n, M_n is faithful. Since the entries of A_n are smaller than c, $M_n \in \text{FLNG}^c_4$. Letting C_n be the first two rows of B_n, it is easy to verify that $(X_{1,n}, X_{2,n})^T = C_n\mathbf{e}_n$. By Theorem 3.2, since, for $n > 2$, no column of C_n is proportional to any other, M_n is parsimonious for $\mathbf{X}_n$. Therefore, $\mathbf{X}_n \in \mathcal{O}_{\text{FLNG}^c_4}$. □

In the following, we will appeal extensively to the following Lemma, given by Kagan et al. [1973].

Lemma 5.3. *Suppose the k-dimensional random vectors $\mathbf{e}_n$ have independent components. Consider the sequence of p-dimensional random vectors $\mathbf{X}_n = B\mathbf{e}_n$, where B is a $p \times k$ matrix. If the $\mathbf{X}_n$ converge in distribution to $\mathbf{X}$, then $X = B\mathbf{e}$, where $\mathbf{e}$ is a k-dimensional random vector with independent components.*

The following is a straightforward Corollary of Lemma 5.3.

Corollary 5.1. *Suppose the k-dimensional random vectors $\mathbf{e}_n$ have independent components. Consider a sequence of p-dimensional random vectors $\mathbf{X}_n = B_n\mathbf{e}_n$, where the B_n are $p \times k$ matrices and $B_n \to B$. If the $\mathbf{X}_n$ converge in distribution to $\mathbf{X}$, then $X = B\mathbf{e}$, where $\mathbf{e}$ is a k-dimensional random vector with independent components.*

Proof of Corollary 5.1. It is a standard fact that if $|X_n - Y_n|$ converge in probability to 0 and the X_n converge in distribution to X, then the Y_n also converge in distribution to X.[5] Clearly, $|B_n\mathbf{e}_n - B\mathbf{e}_n|$ converge in probability to 0. By assumption, the $B_n\mathbf{e}_n$ converge in distribution to $\mathbf{X}$. It follows that

[5] See e.g. `https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_to_a_sequence_converging_in_distribution_implies_convergence_to_the_same_distribution`.

$B\mathbf{e}_n$ converge in distribution to $\mathbf{X}$. By Lemma 5.3, $X = B\mathbf{e}$, where $\mathbf{e}$ is a k-dimensional random vector with independent components. □

Theorem 5.1. *For $d \geq p$, $\mathcal{O}_{\text{FLNG}^c_{>d}}$ is open in the weak topology on $\mathcal{O}_{\text{FLNG}^c}$.*

Proof of Theorem 5.1. Let $e \leq d < f$ Suppose for a contradiction that the $\mathbf{O}_n \in \text{FLNG}^c_e$ converge in distribution to $\mathbf{O} \in \text{FLNG}^c_f$. Let $M \in \text{FLNG}^c_f$ be parsimonious for $\mathbf{O}$ and $M_n \in \text{FLNG}^c_e$ be parsimonious for $\mathbf{O}_n$. Let $B_n = B_{\mathbf{O}_n}(M_n)$ and $A = B_{\mathbf{O}}(M)$. Let $\mathbf{e}_n = \mathbf{e}_{\mathbf{O}_n}(M_n)$ and $\mathbf{e} = \mathbf{e}_{\mathbf{O}}(M)$. While B is a $p \times f$ matrix, each of the B_n are $p \times e$ matrices. By the Bolzano-Weierstrass theorem, since the B_n are uniformly bounded, there is a $p \times e$ matrix B and a convergent subsequence $B_{n_m} \to B$. By assumption, $B_{n_m}\mathbf{e}_{n_m}$ converge in distribution to $\mathbf{O}$. By Corollary 5.1, $\mathbf{O} = B\mathbf{f}$ where $\mathbf{f}$ is a vector of independent components. Therefore $\mathbf{O} = A\mathbf{e} = B\mathbf{f}$. By 3.1 every column of A must be proportional to some column of B. Since A has strictly more columns than B, two columns of A must be proportional to the same column of B and, therefore, to each other. But then, by Theorem 3.2, M is not parsimonious for $\mathbf{O}$. Contradiction.

We have shown that for every $\mathbf{O} \in \text{FLNG}^c_f$, there is an open set separating $\mathbf{O}$ from FLNG^c_e. Since $e < f$ was taken to be arbitrary, there is such an open set E_g separating $\mathbf{O}$ from each FLNG^c_g with $p \leq g < f$. Since there are only finitely many of the E_g, the intersection of the E_g is open and separates $\mathbf{O}$ from $\mathcal{O}_{\text{FLNG}^c_{\leq d}}$. That shows that for every $\mathbf{O} \in \mathcal{O}_{\text{FLNG}^c_{>d}}$, there is an open set $E_{\mathbf{O}}$ separating $\mathbf{O}$ from $\mathcal{O}_{\text{FLNG}^c_{\leq d}}$. Therefore, $\mathcal{O}_{\text{FLNG}^c_{>d}} = \cup_{\mathbf{O} \in \mathcal{O}_{\text{FLNG}^c_{>d}}} E_{\mathbf{O}}$ is a union of open sets and, therefore, open in $\mathcal{O}_{\text{FLNG}^c}$. □

It is worth reflecting briefly on the consequences of Lemma 5.1. As a special case, it entails that $\mathcal{O}_{\text{FLNG}^c_{>p}}$ is open. By Genin and Kelly [2017, Theorem 4.1], this means that it is statistically verifiable whether an unobserved confounder must be introduced in order to accommodate the distribution of $\mathbf{O}$, at least when all distribution are assumed to be absolutely continuous wrt Lebesgue measure. As expected, the hypothesis of un-confoundedness is statistically testable. On the other hand, the precise hypothesis FLNG^c_d is neither statistically verifiable nor refutable, even under the background assumption that the distribution $\mathbf{O}$ was generated by some model in FLNG^c. To see this, note that for $d > p$, FLNG^c_d is neither open nor closed, since more parsimonious models can approximate simpler models. Therefore it is properly locally closed. Although it is neither verifiable nor decidable, it is decidable in the limit by Theorem 1.1.[6] We shall see that the same is true for the hypothesis of orientation $\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c}$. The following theorem shows that if we knew exactly how many latent variables were necessary to accommodate the observed distribution, the hypothesis of orientation would be topologically well-separated.

Theorem 5.2. *For $d \geq p$, $\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c_d}, \mathcal{O}^{i \leftsquigarrow j}_{\text{FLNG}^c_d}$ are open and $\mathcal{O}^{i \circ j}_{\text{FLNG}^c_d}$ is closed in the weak topology on $\mathcal{O}_{\text{FLNG}^c_d}$.*

Proof of Theorem 5.2. Suppose for a contradiction that the $\mathbf{O}_n \in \mathcal{O}^{i \not\rightsquigarrow j}_{\text{FLNG}^c_d}$ converge in distribution to $\mathbf{O} \in \mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c_d}$. Let $M \in \text{FLNG}^c_d$ be parsimonious for $\mathbf{O}$ and $M_n \in \text{FLNG}^c_d$ be parsimonious for $\mathbf{O}_n$. Let $B_n = B_{\mathbf{O}_n}(M_n)$ and $A = B_{\mathbf{O}}(M)$. Let $\mathbf{e}_n = \mathbf{e}_{\mathbf{O}_n}(M_n)$ and $\mathbf{e} = \mathbf{e}_{\mathbf{O}}(M)$. By the Bolzano-Weierstrass theorem, since the B_n are uniformly bounded, there is a $p \times e$ matrix B and a convergent subsequence $B_{n_m} \to B$. By assumption, $B_{n_m}\mathbf{e}_{n_m}$ converge in distribution to $\mathbf{O}$. By Corollary 5.1, $\mathbf{O} = B\mathbf{f}$ where $\mathbf{f}$ is a vector of independent components. Therefore $\mathbf{O} = A\mathbf{e} = B\mathbf{f}$. Since $(B_n)_{ji} = 0$ for all n, $B_{ji} = 0$. Moreover, $B_{ii} = 1$. Since A and B have equal dimensions, by Theorem 3.1 there must be a column k such that $A_{jk} = 0$ and $A_{ik} \neq 0$. But then $O_k \rightsquigarrow_M O_i$ and $O_i \rightsquigarrow_M O_j$ and, therefore, $O_k \rightsquigarrow_M O_j$. But since $A_{jk} = 0$, M must be unfaithful. Contradiction. We have shown that $\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c_d}$ is open in the weak topology on $\mathcal{O}_{\text{FLNG}^c_d}$. Since the situation is symmetrical, $\mathcal{O}^{i \leftsquigarrow j}_{\text{FLNG}^c_d}$ is also open. Since $\mathcal{O}^{i \circ j}_{\text{FLNG}^c_d}$ is the complement of $\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c_d} \cup \mathcal{O}^{i \leftsquigarrow j}_{\text{FLNG}^c_d}$, it is closed in $\mathcal{O}_{\text{FLNG}^c_d}$. □

Corollary 5.2. *$\mathcal{O}^{i \rightsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i \leftsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i \circ j}_{\text{FLNG}^c}$ are disjoint countable unions of sets closed in $\mathcal{O}_{\text{FLNG}^c}$.*

[6] Recall that, in metrizable spaces such as the weak topology, every locally closed set is a countable union of closed sets.

Proof of Corollary 5.2. In general if A is open/closed in a subspace, it is the intersection of an open/closed set with the subspace. By Theorem 5.1, $\mathcal{O}_{\text{FLNG}_d^c}$ is locally closed in $\mathcal{O}_{\text{FLNG}^c}$. By Theorem 5.2, $\mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}_d^c}, \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}_d^c}, \mathcal{O}^{i\circ j}_{\text{FLNG}_d^c}$ are either open or closed in $\mathcal{O}_{\text{FLNG}_d^c}$. Therefore, $\mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}_d^c}, \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}_d^c}, \mathcal{O}^{i\circ j}_{\text{FLNG}_d^c}$ are locally closed in $\mathcal{O}_{\text{FLNG}^c}$. It follows that each of $\mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\circ j}_{\text{FLNG}^c}$ is a countable union of locally closed sets. In a metrizable space such as the weak topology, each open set, and therefore each locally closed set, is a countable union of closed sets. Therefore, each of $\mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\circ j}_{\text{FLNG}^c}$ is a countable union of closed sets. They are disjoint by Theorem 4.2. □

6 Main Result and Discussion

We are almost in position to state and prove the main results. Let $\mathcal{M}$ be the set of all pairs $\langle M, \alpha\rangle$, where $M \in \text{FLNG}^c$ and α is a permutation of $\{1, \ldots, |M|\}$. Let $P(\langle M, \alpha\rangle) = (\mathbf{X}_{\alpha^{-1}(1)}(M), \ldots, \mathbf{X}_{\alpha^{-1}(p)}(M))$. Let $\mathcal{M}^{i\rightsquigarrow j} = \{\langle M, \alpha\rangle \in \mathcal{M} : \alpha^{-1}(i) \rightsquigarrow_M \alpha^{-1}(j)\}$ and $\mathcal{M}^{i\circ j} = \{\langle M, \alpha\rangle \in \mathcal{M} : \alpha^{-1}(i) \circ_M \alpha^{-1}(j)\}$.

Theorem 6.1. *The question $\mathfrak{Q} = \{\mathcal{M}^{i\rightsquigarrow j}, \mathcal{M}^{i\circ j}, \mathcal{M}^{i\leftsquigarrow j}\}$ is identified and decidable in the limit, but not statistically decidable.*

Proof of Theorem 6.1. It is immediate from defintions that $P(\mathcal{M}^{i\rightsquigarrow j}) = \mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}^c}$, $P(\mathcal{M}^{i\leftsquigarrow j}) = \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}^c}$ and $P(\mathcal{M}^{i\circ j}) = \mathcal{O}^{i\circ j}_{\text{FLNG}^c}$. Since $\mathcal{O}^{i\rightsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\leftsquigarrow j}_{\text{FLNG}^c}, \mathcal{O}^{i\circ j}_{\text{FLNG}^c}$ are disjoint by Theorem 4.2, the question is identified. Since they are each countable unions of closed sets (by Corollary 5.2), Theorem 1.1 implies that the question is decidable in the limit. The question is not decidable by Theorem 1.2 and Lemma 5.2. □

Theorem 6.1 shows that learning causal orientation in faithful, but potentially confounded, LiNGAM models is a difficult problem. Not so difficult that it is impossible to construct consistent methods, but difficult enough that no consistent method can guarantee a finite-sample bound on the probability of orientation errors. In view of the positive results given by Genin and Mayo-Wilson [2020] in the unconfounded setting, this negative results is something of a disappointment. However, limiting results of this kind serve to forestall wishful thinking and calibrate our attitude toward the output of our best causal discovery methods. If it is not possible to give finite-sample bounds on the probability of orientation errors, then *a fortiori* it is not possible to construct confidence intervals for causal effects enjoying finite-sample coverage. However, under similar assumptions and using similar techniques, it is possible to show that such intervals can be constructed in randomized experiments, or whenever an instrumental variable is available. Furthermore, this work may stimulate a search for similar, and perhaps equally plausible, assumptions that yield stronger senses of identifiability. In recent years, the field of causal discovery has produced many exciting new identifiability results under a variety of modeling assumptions. But demonstrating identifiability proves only that a problem is not completely hopeless — it is only the first step in understanding how difficult a problem is. Perhaps the topological approach taken in this work can serve as a model for future investigations into the difficulty of causal discovery under varied modeling assumptions.

7 NeurIPS Paper Checklist

1. (a) Do the **main claims** made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Yes. Except for a very small concession to readability, the claims made in the abstract and intro correspond very closely to the results given in Section 6.

   (b) Have you read the **ethics review guidelines** and ensured that your paper conforms to them?

   Yes. The submission is intended to be a contribution to the methodology of causal discovery from observational data. Moreover, it is a limiting result encouraging responsible use of these methods more so than inventing a new method. Surely, causal discovery can be used for good or bad ends, but I think this work raises only very remote ethical concerns. Since the main contribution is a theorem and its proof, it raises no issues of data use or privacy.

(c) Did you discuss any potential **negative societal impacts** of your work?

No. Of course I would like the work to have a positive societal impact. Causal discovery is at the heart of social science. A sober and reliable causal discovery methodology is essential for responsible and progressive social science. I don't think this work has any significant potential for negative social impact — at least no more than, say, a discussion of the limitations of regression for causal inference would. I think it's tendency is to encourage understanding of the difficulty of reliable causal discovery. I think a discussion of negative societal impact in the body of the text would take me rather far into the realm of remote speculation.

(d) Did you describe the **limitations** of your work.

Not explicitly but in a sense this work is all about limitations. If I interpret the question narrowly, then: yes, I attempted to state the preconditions of the theorems clearly and explicitly.

2. If you are including theoretical results ...

   (a) Did you state the full set of **assumptions** of all theoretical results?

   Yes.

   (b) Did you include complete **proofs** of all theoretical results?

   Yes, to the best of my knowledge all proofs are complete. Of course, it is possible that I have made a mistake someplace.

3. If you ran experiments...

   This work did not involve experiments.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   Not applicable.

5. If you used crowdsourcing or conducted research with human subjects...

   Not applicable.

References

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, second edition, 1986.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Amir Dembo and Yuval Peres. A topological criterion for hypothesis testing. *The Annals of Statistics*, 22(1):106–117, 1994.

Konstantin Genin and Kevin T Kelly. The Topology of Statistical Verifiability. *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK)*, 2017. URL `https://arxiv.org/abs/1707.09378v1`.

Konstantin Genin and Conor Mayo-Wilson. Statistical Decidability in Linear, Non-Gaussian Models. In *NeurIPS 2020:Workshop on Causal Discovery and Causality-Inspired Machine Learning*, 2020.

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

A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. John Wiley & Sons, 1973.

Kevin T. Kelly and C. Mayo-Wilson. Causal Conclusions that Flip Repeatedly and Their Justification. *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence*, pages 277–286, 2010.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

Shohei Shimizu, Patrick O. Hoyer, Aapo Hyvarinen, and Antti J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

8 Appendix

Proof of Lemma 3.1. Suppose that α embeds $\mathbf{O} = (O_1, \ldots, O_p)$ in $M \in \text{LNG}_d$ and that $D = B_{\mathbf{O}}(M)$. Suppose that $D_{[,v]} = aD_{[,u]}$. First, we show that at least one of u, v must be strictly greater than p. Suppose for a contradiction that $u, v \leq p$. Since D has ones everywhere on the diagonal of the principal $p \times p$ submatrix, $D_{uu} = D_{vv} = 1$. By assumption, we have that $D_{uv} = a > 0$ and $D_{vu} = 1/a > 0$. But then $O_u \rightsquigarrow_M O_v$ and $O_v \rightsquigarrow_M O_u$, which contradicts acyclicity. Therefore, without loss of generality, we suppose that $v > p$.

Let $y, z = \alpha^{-1}(u), \alpha^{-1}(v)$. Let $A = A(M)$ and $B = B(M)$. Let β be the permutation of $\{1, ..., d\}$ sending $i \mapsto i$ for $i < z$ and $i \mapsto i - 1$ for $i > z$. Define the $d - 1 \times d - 1$ matrix A' in the following way:

$$A'_{ij} = A_{\beta^{-1}(i,j)} + A_{z,\beta^{-1}(j)} A_{\beta^{-1}(i),z}.$$

Since A has zeros on the diagonal and, by acyclicity, one of $A_{z,\beta^{-1}(i)}, A_{\beta^{-1}(i),z}$ must be zero, A' has zeros on the diagonal. We show that A' is lower triangular, i.e. that $A'_{ij} = 0$ whenever $j > i$. There are three cases to consider: (1) $i < z, j < z$; (2) $i < z, j \geq z$ and (3) $i \geq z, j \geq z$. In the first case $A'_{ij} = A_{ij} + A_{zj} A_{iz}$. Since A is lower triangular, $A_{ij} = A_{iz} = 0$. In the second case, $A'_{ij} = A_{i,j+1} + A_{z,j+1} A_{iz}$. Since A is lower triangular, $A_{i,j+1} = A_{iz} = 0$. In the final case, $A'_{ij} = A_{i+1,j+1} + A_{z,j+1} A_{i+1,z}$. Since A is lower triangular, $A_{i+1,j+1} = A_{z,j+1} = 0$.

Since $I - A'$ is lower triangular and its diagonal entries are all equal to one, the inverse matrix $B' = (I - A')^{-1}$ exists. We argue that $B'_{ij} = B_{\beta^{-1}(i,j)}$. Let Π_{ij} be the set of all paths from i to j over the vertices $\{1, \ldots, d\}$. Let Π_{ikj} be the set of all paths from i to j over the vertices $\{1, \ldots, d\}$ passing through k and let $\Pi_{i \not k j} = \Pi_{ij} \setminus \Pi_{ikj}$. Let Π'_{ij} bet the set of all paths from i to j over the vertices $\{1, \ldots, d-1\}$. From our previous observation, we have that

$$B'_{ji} = \sum_{\pi \in \Pi'_{ij}} \prod_{i=1}^{|\pi|} A'_{\pi_{i+1}, \pi_i} \tag{1}$$

$$= \sum_{\pi \in \Pi'_{ij}} \prod_{i=1}^{|\pi|} \left(A_{\beta^{-1}(\pi_{i+1}, \pi_i)} + A_{z,\beta^{-1}(\pi_i)} A_{\beta^{-1}(\pi_{i+1}),z} \right) \tag{2}$$

$$= \sum_{\pi \in \Pi_{\beta^{-1}(i) \not \beta^{-1}(j)}} \prod_{i=1}^{|\pi|} \left(A_{\pi_{i+1}, \pi_i} + A_{z,\pi_i} A_{\pi_{i+1},z} \right). \tag{3}$$

Note that for any $\pi \in \Pi_{ij}$ there can be at most one π_i such that

$$A_{z,\pi_i} A_{\pi_{i+1},z} \neq 0.$$

If this were not the case, there would be causal paths in $G(M)$ passing through z twice, contradicting acyclicity. Let π_{i*} be the unique such π_i, if it exists, and let $\pi_{i*} = \pi_1$, otherwise. Then:

$$B'_{ji} = \sum_{\pi \in \Pi_{\beta^{-1}(i) \not\sim \beta^{-1}(j)}} \prod_{i=1}^{|\pi|} (A_{\pi_{i+1},\pi_i}) + A_{z,\pi_{i^*}} A_{\pi_{i^*+1},z} \prod_{i \neq i*} (A_{\pi_{i+1},\pi_i}) \tag{4}$$

$$= \sum_{\pi \in \Pi_{\beta^{-1}(i) \not\sim \beta^{-1}(j)}} \prod_{i=1}^{|\pi|} (A_{\pi_{i+1},\pi_i}) + \sum_{\pi \in \Pi_{\beta^{-1}(i) z \beta^{-1}(j)}} \prod_{i=1}^{|\pi|} (A_{\pi_{i+1},\pi_i}) \tag{5}$$

$$= \sum_{\pi \in \Pi_{\beta^{-1}(i) \not\sim \beta^{-1}(j)}} \times_M \pi + \sum_{\pi \in \Pi_{\beta^{-1}(i) z \beta^{-1}(j)}} \times_M \pi \tag{6}$$

$$= \sum_{\pi \in \Pi_{\beta^{-1}(i,j)}} \times_M \pi \tag{7}$$

$$= B_{\beta^{-1}(j,i)}. \tag{8}$$

Let the $d-1$ element column vector $\mathbf{e}'$ be just like the first $d-1$ rows of $P_\beta \mathbf{e}(M)$ except $\mathbf{e}'_{\beta(y)} = \mathbf{e}_y + a\mathbf{e}_z$. Since the sum of independent non-Gaussian variables is non-Gaussian, each element of $\mathbf{e}'$ is not Gaussian. Moreover, since functions of independent random variables are independent, the $\mathbf{e}'_i$ are mutually independent.

Let $\mathbf{X}' = B'\mathbf{e}'$ Since B' is lower triangular with ones on the diagonal and $\mathbf{e}'$ is a vector of mutually independent, non-Gaussian random variables, we have that $M' = \langle \mathbf{X}', \mathbf{e}', A' \rangle$ is in LNG_{d-1}.

We are now in a position to prove part (i) of the theorem. We claim that $\beta^{-1} \circ \alpha$ embeds $\mathbf{O}$ into M'. In other words, we claim that $O_i = \mathbf{X}_{\beta(\alpha^{-1}(i))}(M')$:

$$\begin{aligned}
\mathbf{X}_{\beta(\alpha^{-1}(i))}(M') &= \sum_{j=1}^{d-1} B'_{\beta(\alpha^{-1}(i)),j} \mathbf{e}'_j \\
&= B'_{\beta(\alpha^{-1}(i)),\beta(y)} \mathbf{e}'_{\beta(y)} + \sum_{j<z, j \neq \beta(y)} B'_{\beta(\alpha^{-1}(i)),j} \mathbf{e}'_j + \sum_{j \geq z, j \neq \beta(y)} B'_{\beta(\alpha^{-1}(i)),j} \mathbf{e}'_j \\
&= B_{\alpha^{-1}(i),y}(\mathbf{e}_y + a\mathbf{e}_z) + \sum_{j<z, j \neq \beta(y)} B_{\alpha^{-1}(i),\beta^{-1}(j)} \mathbf{e}'_j + \sum_{j \geq z, j \neq \beta(y)} B_{\alpha^{-1}(i),\beta^{-1}(j)} \mathbf{e}'_j \\
&= B_{\alpha^{-1}(i),y}(\mathbf{e}_y + a\mathbf{e}_z) + \sum_{j<z, j \neq y} B_{\alpha^{-1}(i),j} \mathbf{e}_j + \sum_{j \geq z, j \neq y} B_{\alpha^{-1}(i),j+1} \mathbf{e}_{j+1} \\
&= B_{\alpha^{-1}(i),y}(\mathbf{e}_y + a\mathbf{e}_z) + \sum_{j<z, j \neq y} B_{\alpha^{-1}(i),j} \mathbf{e}_j + \sum_{j>z, j \neq y} B_{\alpha^{-1}(i),j} \mathbf{e}_j \\
&= aB_{\alpha^{-1}(i),y} \mathbf{e}_z + \sum_{j \neq z} B_{\alpha^{-1}(i),j} \mathbf{e}_j \\
&= \mathbf{X}_{\alpha^{-1}(i)}(M) \\
&= O_i.
\end{aligned}$$

The third and fourth equalities follows from the fact that $B'_{ij} = B_{\beta^{-1}(i,j)}$ and that $\beta^{-1}(i) = i$ when $i < z$ and $\beta^{-1}(i) = i+1$ when $i \geq z$. The penultimate equality follows from the fact that $aB_{\alpha^{-1}(i),y}\mathbf{e}_z = aB_{\alpha^{-1}(i),\alpha^{-i}(u)}\mathbf{e}_z = aD_{i,u}\mathbf{e}_z = D_{i,v}\mathbf{e}_z = B_{\alpha^{-1}(i),\alpha^{-i}(v)}\mathbf{e}_z = B_{\alpha^{-1}(i),z}\mathbf{e}_z$. The final equality follows from the fact that α embeds $\mathbf{O}$ in M.

We now prove (ii). Suppose that $O_i \rightsquigarrow_M O_j$. Since M is faithful, $B_{\alpha^{-1}(j,i)} \neq 0$ and therefore $B'_{\beta(\alpha^{-1}(j,i))} \neq 0$. That entails that $\beta(\alpha^{-1}(i)) \rightsquigarrow_{M'} \beta(\alpha^{-1}(j))$ and, since $\beta^{-1} \circ \alpha$ embeds $\mathbf{O}$ in M', $O_i \rightsquigarrow_{M'} O_j$. For the converse it suffices to show that $i \rightarrow_{M'} j$ entails $\beta^{-1}(i) \rightsquigarrow_M \beta^{-1}(j)$. Suppose the antecedent holds. Then $A'_{ji} \neq 0$. Therefore, either $A_{\beta^{-1}(j,i)} \neq 0$ or $A_{z,\beta^{-1}(i)} A_{\beta^{-1}(j),z} \neq 0$. In either case, $\beta^{-1}(i) \rightsquigarrow_M \beta^{-1}(j)$.

It remains to prove (iii). Suppose that $i \rightsquigarrow_{M'} j$. Then, by (ii), $\beta^{-1}(i) \rightsquigarrow \beta^{-1}(j)$. Since M is faithful, $B_{\beta^{-1}(j,i)} \neq 0$. But since $B_{\beta^{-1}(j,i)} = B'_{j,i}$, $B'_{j,i} \neq 0$, as required. □