Statistical Decidability in Linear, Non-Gaussian Causal Models

Konstantin Genin[*]
Cluster of Excellence – Machine Learning for Science
Universität Tübingen

konstantin.genin@uni-tuebingen.de

Conor Mayo-Wilson[†]
Department of Philosophy
University of Washington

conormw@uw.edu

Causal discovery algorithms are becoming increasingly sophisticated, often using fine-grained information about the joint distribution (e.g., that noise is additive) rather than relying on coarse-grained information about which variables are conditionally independent [Hoyer et al., 2009, Peters et al., 2012, Loh and Bühlmann, 2014]. For example, significant advances have been made in the study of linear, non-Gaussian causal models (LiNGAMs) [Shimizu et al., 2011, Hoyer et al., 2008]. If the true model is LiNGAM, then the causal relationships among all measured variables can be identified in the limit [Shimizu et al., 2006]. Under strong additional assumptions, variants of maximum-likelihood estimation provide *uniformly* consistent procedures for identifying the graphical structure of LiNGAM models [Bühlmann et al., 2014].

The main result of this paper is to show that, without any further assumptions, the direction of any causal edge in a LiNGAM is what we call *statistically decidable* [Genin, 2018]. Statistical decidability is a reliability concept that is, in a sense, intermediate between the familiar notions of consistency and uniform consistency. A set of models is statistically decidable if, for any $\alpha > 0$, there is a consistent procedure that, *at every sample size*, hypothesizes a false model with chance less than α. Such procedures may exist even when uniformly consistent ones do not. Uniform consistency requires that one be able to determine the sample size *a priori* at which one's chances of identifying the true model are at least $1 - \alpha$; statistical decidability requires no such pre-experimental guarantees.

It is trivial to show that there is no uniformly consistent algorithm for determining the direction of a causal edge in LiNGAMs; see example in section three. Thus, our main result illuminates how the notions of uniform consistency and statistical decidability come apart.

Our main result also illustrates how discovery of LiNGAMs differs from their Gaussian counterparts. As sample size increases, consistent discovery algorithms for (the Markov equivalence class) of Gaussian models can be forced to "flip" their judgments about whether X causes Y or vice versa, *no matter how strong the effect* of X on Y [Kelly and Mayo-Wilson, 2010]. Further, even in the absence of confounders, the number of such flips is bounded only by the number of variables in the model. Our main result, in contrast, shows that consistent discovery algorithms for LiNGAMs can avoid such flipping; whether existing algorithms do avoid flipping is a matter for further investigation.

Finally, our results suggest a practical, implementable way that existing causal discovery algorithms could be made more sensitive to users' interests. Many existing causal discovery algorithms are functions of the data only, and hence, do not allow the user to designate the kinds of error that she deems acceptable. Our main result shows that, for specified classes of models like LiNGAMs, it is possible to design algorithms that allow users to see not only the models that best fit the data but also, whether the data allow one to draw causal conclusions with the reliability desired by the user. It is an open question whether existing discovery algorithms can be modified in the way we suggest.

[*] https://kgenin.github.io/. Supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645.

[†] https://faculty.washington.edu/conormw/

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

1 Background: Linear Causal Models

A *linear causal model in d variables* M is a triple $\langle \mathbf{X}, \mathbf{e}, B \rangle$, where $\mathbf{X} = \langle X_1, X_2, \ldots, X_d \rangle$ is a vector of d *observable* random variables, $\mathbf{e} = \langle e_1, e_2, \ldots, e_d \rangle$ is a random vector of d *unobserved* noise terms, and B is a $d \times d$ matrix such that

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

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*. If $M = \langle \mathbf{X}, \mathbf{e}, B \rangle$ is a linear causal model then $\mathbf{X} = B\mathbf{X} + \mathbf{e}$. It is clear that, since no X_i causes itself, the matrix B must have all zero diagonal elements. If the X_i are enumerated in agreement with the causal order (i.e., $i < j$ if and only if $B_{ij} = 0$), then B is *lower triangular*, i.e. all elements above the diagonal are zero. The observables also admit the following "dual" representation: $\mathbf{X} = B'\mathbf{e}$, where $B' = (I - B)^{-1}$. Note that the inverse of $I - B$ always exists. By (1), it is clear that the matrix B' must have *unit diagonal*. If the X_i are enumerated in agreement with the causal order, then B' is *also* lower triangular.

Let L_d be the set of all linear causal models on d variables, and let $\mathrm{LG}_d, \mathrm{LNG}_d \subset \mathrm{L}_d$ respectively denote the sets of linear Gaussian and non-Gaussian models. If $M = \langle \mathbf{X}, \mathbf{e}, B \rangle$, let $O(M) = \mathbf{X}$, $E(M) = \mathbf{e}$, and $B(M) = B$. Let $\mathrm{PA}_M(i) = \{j : B_{ij}(M) \neq 0\}$, be the set of parents of i in M.

Each linear causal model M gives rise to a directed acyclic graph (DAG) in a natural way: the DAG $G(M)$ has a directed edge from j to i if and only if $B_{ij}(M) \neq 0$. Let DAG_d be the set of all DAGs on d variables. For a set of models $\mathcal{M}$, define $G[\mathcal{M}] := \{G(M) : M \in \mathcal{M}\}$. In the reverse direction, given a DAG G, we define $\mathcal{M}_G := \{M \in \mathrm{L}_d : G(M) = G\}$ to be the set of linear models that give rise to G.

2 Identifiability

It is well known that linear, Gaussian causal models are not, in general, identifiable [Richardson and Spirtes, 2002, Theorem 8.14] . In other words, there exist pairs of causal models $M, M' \in \mathrm{LG}_d$ such that $G(M) \neq G(M')$, (and therefore $B(M) \neq B(M')$) but $O(M) = O(M')$. We give a new, simple proof of this fact. The proof invokes the Lukacs-King theorem [1954], which is perhaps not so well known as its consequence, the Darmois-Skitovich theorem [1953, 1953].

Theorem 2.1 (Lukacs-King)**.** *Let $X_1, \ldots, X_m$ be independent random variables, $X' = \sum_i \alpha_i X_i$ and $X'' = \sum_i \beta_i X_i$. Then, X', X'' are independent iff (a) each X_i such that $\alpha_i \beta_i \neq 0$ is Gaussian; and (b) $\sum_{i=1}^{m} \alpha_i \beta_i Var(X_i) = 0$.*

Theorem 2.2. *Suppose that $M = \langle \mathbf{X}, \mathbf{e}, A \rangle$ is an element of LG_d and that there are i, j such that $\mathrm{PA}_M(j) = \mathrm{PA}_M(i) \cup \{i\}$. Then there is $M' = \langle \mathbf{X}', \mathbf{e}', B \rangle \in \mathrm{LG}_d$ such that $O(M) = O(M')$ and $G(M) \neq G(M')$. Typically, M' can be chosen so that $G(M')$ is just like $G(M)$ except $i \leftarrow j \in G(M')$ whereas $i \rightarrow j \in G(M)$.*

Proof of Theorem 2.2. Let $B_{ij} = \frac{A_{ji}\mathrm{Var}(\mathbf{e}_i)}{A_{ji}^2 \mathrm{Var}(\mathbf{e}_i) + \mathrm{Var}(\mathbf{e}_j)}$.
For $k \in \mathrm{PA}_M(i)$, let

$$\begin{aligned} B_{jk} &= A_{ji} A_{ik} + A_{jk}; \\ B_{ik} &= A_{ik} - B_{ij} B_{jk}. \end{aligned}$$

Let all other entries of B be just like A. Let

$$\begin{aligned} \mathbf{e}'_i &= (1 - B_{ij} A_{ji})\mathbf{e}_i - B_{ij}\mathbf{e}_j; \\ \mathbf{e}'_j &= A_{ji}\mathbf{e}_i + \mathbf{e}_j, \end{aligned}$$

and let $\mathbf{e}'_\ell = \mathbf{e}_\ell$ for $\ell \neq i, j$.

We first show that $M' \in \mathrm{LG}_d$. If k is a causal ordering for M, then letting the ordering k' be just like k except $k'(i) = k(j)$ and $k'(j) = k(i)$ yields a causal ordering for M'. To show that the $\mathbf{e}'$ are mutually independent it suffices to show that $\mathbf{e}'_i, \mathbf{e}'_j$ are independent. By the Lukacs-King theorem, this is the case so long as

$$A_{ji}(1 - B_{ij}A_{ji})\mathrm{Var}(\mathbf{e}_i) - B_{ij}\mathrm{Var}(\mathbf{e}_j) = 0,$$

or equivalently,

$$A_{ji}\mathrm{Var}(\mathbf{e}_i) - B_{ij}(A_{ji}^2\mathrm{Var}(\mathbf{e}_i) + \mathrm{Var}(\mathbf{e}_j)) = 0,$$

which is immediate from the definition of B_{ij}. Since $\mathbf{e}'_i, \mathbf{e}'_j$ are mixtures of independent Gaussians, they are Gaussian. Therefore, $M' \in \mathrm{LG}_d$.

It is obvious that $\mathbf{X}_k = \mathbf{X}'_k$ for $k \in \mathrm{PA}_M(i)$. To show that $O(M) = O(M')$, it is sufficient to show that $\mathbf{X}_j = \mathbf{X}'_j$ and $\mathbf{X}_i = \mathbf{X}'_i$. The equality of the other $\mathbf{X}_\ell, \mathbf{X}'_\ell$ follows. For the first equality, note that

$$\begin{aligned}\mathbf{X}'_j &= \sum_{k \in \mathrm{PA}_M(j)\setminus\{i\}} (A_{ji}A_{ik} + A_{jk})\mathbf{X}_k + A_{ji}\mathbf{e}_i + \mathbf{e}_j \\ &= \sum_{k \in \mathrm{PA}_M(j)\setminus\{i\}} A_{jk}\mathbf{X}_k + A_{ji}\mathbf{X}_i + \mathbf{e}_j = \mathbf{X}_j.\end{aligned}$$

For the second equality, note that

$$\begin{aligned}\mathbf{X}'_i &= \sum_{k \in \mathrm{PA}_M(i)} (A_{ik} - B_{ij}B_{jk})\mathbf{X}_k + B_{ij}\mathbf{X}'_j + (1 - B_{ij}A_{ji})\mathbf{e}_i - B_{ij}\mathbf{e}_j \\ &= \sum_{k \in \mathrm{PA}_M(i)} A_{ik}\mathbf{X}_k + \mathbf{e}_i = \mathbf{X}_i,\end{aligned}$$

since $B_{ij}\mathbf{X}'_j = \sum_{k \in \mathrm{PA}_M(i)} B_{ij}B_{jk}X_k + B_{ij}A_{ji}\mathbf{e}_i + B_{ij}\mathbf{e}_j$. We have that $G(M) \neq G(M')$, since $B_{ij} > 0 = A_{ij}$. Finally, we have that $G(M')$ is just like $G(M)$ except that the edge between i and j is flipped whenever $B_{jk}, B_{ik} \neq 0$ for all $k \in \mathrm{PA}_M(i)$. It is straightforward to check that this holds whenever $A_{ji}A_{ik} \neq -A_{jk}$ and $A_{jk}A_{ji}\mathrm{Var}(\mathbf{e}_i) \neq A_{ik}\mathrm{Var}(\mathbf{e}_j)$ for all $k \in \mathrm{PA}_M(i)$. □

The situation with respect to identifiability changes dramatically when we restrict attention to the non-Gaussian case.

Theorem 2.3. *Suppose that $M, M' \in \mathrm{LNG}_d$. If $O(M) = O(M')$, then $M = M'$.*

We give a simple proof of this fact. The proof invokes the Lukacs-King theorem, as well as a convenient lemma, which will also be used to prove our main result. Say that a matrix is a **mixing matrix** if and only if some column has two non-zero entries. A **leading principle submatrix** of a matrix A is the result of removing all but the first n rows and columns for some n.

Lemma 2.1. *Suppose that A, B are square matrices of the same dimension having unit diagonals. Suppose that B is lower triangular and the result of the matrix multiplication AB is* **not** *a mixing matrix. Then $A = B^{-1}$.*

Proof of Lemma 2.1. By induction on the dimensions of A, B. For the base case, suppose that A, B are 2×2 matrices. Then:

$$AB = \begin{pmatrix} 1 & a_{12} \\ a_{21} & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ b_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 + a_{12}b_{21} & a_{12} \\ a_{21} + b_{21} & 1 \end{pmatrix}$$

Since by assumption AB is not a mixing matrix, $a_{12} = 0$ and therefore $a_{21} = -b_{21}$. It follows that

$$A = \begin{pmatrix} 1 & 0 \\ -b_{21} & 1 \end{pmatrix} = B^{-1}.$$

For the inductive step, suppose that the lemma holds for matrices of dimension $n \times n$. Suppose that A, B are $(n+1) \times (n+1)$ matrices satisfying the preconditions of the lemma. Since AB is, by assumption, not a mixing matrix, it is sufficient to show that AB has unit diagonal.

Let A', B' be the $n \times n$ leading principle submatrices of A, B respectively. A', B' have unit diagonal since A, B have unit diagonal. B' is lower triangular since B is lower triangular.

Consider the last column of AB :

$$AB = \begin{pmatrix} \cdots & & a_{1,n+1} \\ \cdots & & a_{2,n+1} \\ \cdots & & \vdots \\ \cdots & & 1 \end{pmatrix} \begin{pmatrix} \cdots & & 0 \\ \cdots & & 0 \\ \cdots & & \vdots \\ \cdots & & 1 \end{pmatrix} = \begin{pmatrix} \cdots & & a_{1,n+1} \\ \cdots & & a_{2,n+1} \\ \cdots & & \vdots \\ \cdots & & 1 \end{pmatrix}$$

Since AB is not mixing by assumption, it follows that $a_{i,n+1} = 0$ for $i < n+1$. Since A has unit diagonal $a_{n+1,n+1} = 1$. Therefore,

$$(AB)_{i,j} = \sum_{k=1}^{n+1} a_{i,k} b_{k,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j} = (A'B')_{i,j},$$

for $i, j < n+1$. Since AB is not a mixing matrix, it follows that $A'B'$ is also not a mixing matrix. By the inductive hypothesis $A'B'$ is the identity matrix. Therefore, AB has unit diagonal. Since AB is not a mixing matrix, it must be the identity matrix. Therefore $A = B^{-1}$. □

We can now prove Theorem 2.3.

Proof of Theorem 2.3. Let $M = \langle \mathbf{X}, \mathbf{e}, B \rangle$ and $M' = \langle \mathbf{X}', \mathbf{e}', B' \rangle$ be elements of $\textsc{Lng}_d$. Suppose that $\mathbf{X} = \mathbf{X}'$ and, without loss of generality, that the observable variables are enumerated in agreement with the causal order of $G(M)$. Since $\mathbf{X} = \mathbf{X}'$,

$$(I - B')^{-1}\mathbf{e}' = (I - B)^{-1}\mathbf{e},$$

and $\mathbf{e}' = (I - B')(I - B)^{-1}\mathbf{e}$. Suppose for a contradiction that $(I - B')(I - B)^{-1}$ is a mixing matrix. Then there is some k such that the k^{th} column of $(I - B')(I - B)^{-1}$ has non-zero entries a, b in rows i, j, respectively. Then $\mathbf{e}'_i = \cdots + a\mathbf{e}_k + \cdots$ and $\mathbf{e}'_j = \cdots + b\mathbf{e}_k + \cdots$. Since M' is a linear causal model, $\mathbf{e}'_i, \mathbf{e}'_j$ are independent. Therefore, by Lukacs-King, $\mathbf{e}_k$ is Gaussian and $M \in \textsc{L}_d \setminus \textsc{Lng}_d$. Contradiction. Therefore, $(I - B')(I - B)^{-1}$ is not a mixing matrix. Then, by Lemma 2.1, it must be that $B = B'$ and $\mathbf{e} = \mathbf{e}'$. □

3 Progressiveness and Decidability: Between Pointwise and Uniform Consistency

Our main result is that, if one knows the true causal model is LiNGAM, then the orientation of any given causal edge is what we call *statistically decidable*. Moreover, there exists a *progressive* method for estimating the entire DAG. The notions of statistical decidability and progressiveness are not yet widely known, and so in this section, we define the terms precisely in a general statistical setting. Both of these notions are stronger than pointwise consistency, but weaker than uniform consistency. In the LiNGAM context, the results of this section allow us to construct consistent methods for learning the true DAG that do not exhibit the flipping behavior identified by Kelly and Mayo-Wilson [2010].

Let $\mathcal{M}$ be a set of statistical models, which one can often think of as the underlying parameter space. For instance, $\mathcal{M}$ might be $[0, 1]$, representing the value of a Bernoulli parameter. In causal discovery, $\mathcal{M}$ might be $\textsc{L}_d$, $\textsc{Lg}_d$ or $\textsc{Lng}_d$.

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 observable outcomes. If $M \in \textsc{L}_d$, this is the measure on $\mathbb{R}^d$ induced by the $O(M)$. Henceforth, we assume $\Omega = \mathbb{R}^d$. We lift $P(\cdot)$ to sets of models in the obvious way: if $\mathcal{A} \subseteq \textsc{L}_d$, 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 $P(\partial A) = 0$. We write $\mathsf{cl}(\cdot)$ for the closure operator in the weak topology. 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 set of disjoint subsets of $\mathcal{M}$. For example, we might let $\mathfrak{Q} = \{\mathcal{M}_G \cap \text{L}\textsc{ng}_d : G \in \text{D}\textsc{ag}_d\}$ be the question of which DAG represents the causal relationship among the observable variables. We call elements of a question $\mathfrak{Q}$ **answers**. In standard statistical terminology, *estimation* problems typically concern maximally fine questions $\mathfrak{Q} = \{\{M\} : M \in \mathcal{M}\}$ of the model space (e.g., what is the value of Bernoulli parameter in $\mathcal{M} = [0, 1]$?), whereas *model selection* problems concern coarser questions. For all $M \in \mathcal{M}$, let $\mathfrak{Q}(M)$ denote the unique answer/element in $\mathfrak{Q}$ containing M, if one exists, and let $\mathfrak{Q}(M)$ be $\mathcal{M}$, otherwise.

A question $\mathfrak{Q}$ often comes with some sort of metric $d_{\mathfrak{Q}}$ and/or topology $\mathcal{T}_{\mathfrak{Q}}$ that represents how close various answers are. For example, if one is interested in estimating a real-valued parameter $M \in \mathcal{M} = \mathbb{R}$, then the question is $\mathfrak{Q} = \{\{r\} : r \in \mathbb{R}\}$, and it is natural to define $d_{\mathfrak{Q}}(\{r\}, \{q\}) := |r - q|$ to be the standard Euclidean distance. If $\mathfrak{Q}$ is a finite partition of $\mathcal{M}$ (as in some cases of model selection), then the natural metric on $\mathfrak{Q}$ is the discrete one.

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.

The familiar notions of statistical consistency and uniform consistency can now be made precise and contrasted with our notions of statistical decidability and progressiveness. Given a question $\mathfrak{Q}$ with topology $\mathcal{T}_{\mathfrak{Q}}$, say a method λ is (pointwise) **consistent** if for all $\epsilon > 0$, all models $M \in \bigcup \mathfrak{Q}$, and all open sets $U \in \mathcal{T}_{\mathfrak{Q}}$ containing $\mathfrak{Q}(M)$, there is some sample size $n \in \mathbb{N}$ such that $P_M^k(\lambda_k \subseteq U) > 1 - \epsilon$ for all $k \geq n$. When $\mathcal{T}_{\mathfrak{Q}}$ is the discrete topology, consistency amounts to the claim that for all $\epsilon > 0$ and models $M \in \bigcup \mathfrak{Q}$, there is some n such that $P_M^k(\lambda_k = \mathfrak{Q}(M)) > 1 - \epsilon$ for all $k \geq n$. If $\mathfrak{Q}$ comes with a metric $d_{\mathfrak{Q}}$, say a method λ is **uniformly consistent** if for any $\delta, \epsilon > 0$, there is some sample size n such that $P_M^k(d_{\mathfrak{Q}}(\lambda_k, \mathfrak{Q}(M)) < \delta) > 1 - \epsilon$ for all models $M \in \bigcup \mathfrak{Q}$. Again, when $d_{\mathfrak{Q}}$ is equivalent to the discrete metric, a method λ is uniformly consistent if for any $\epsilon > 0$, there is some sample size n such that $P_M^k(\lambda_k = \mathfrak{Q}(M)) > 1 - \epsilon$ for all models $M \in \bigcup \mathfrak{Q}$ and all $k \geq n$.

Uniform consistency is an extremely strong demand. For instance, when $\mathfrak{Q}$ is finite, uniform consistency requires that for any positive $\epsilon > 0$ one can name a sample size *a priori* (i.e., before seeing any data) at which the probability of conjecturing the true, *informative* answer exceeds $1 - \epsilon$.

Example 1: We claim there is no uniformly consistent procedure for determining the direction of an edge for $\text{L}\textsc{ng}_d$ on $d = 2$ many variables. The argument is straightforward and easy to generalize to arbitrarily many variables. If one had a uniformly consistent method and could therefore name a sample size *a priori* by which one could tell whether the edge were directed from 1 to 2 or vice versa, one could also name a sample size *a priori* by which one would know whether there was an edge between 1 and 2 at all. But that is impossible because edge coefficients in linear causal models may be arbitrarily small/weak.

In greater detail, let $\mathcal{M} = \text{L}\textsc{ng}_2$. Let G be the DAG $1 \to 2$ and H be the DAG $2 \to 1$. Define $\mathfrak{Q}$ to be the question $\{\mathcal{M}_G, \mathcal{M}_H\}$, equipped with the discrete metric. Let $\mathbf{e} = \langle \mathbf{e}_1, \mathbf{e}_2 \rangle$ be the uniform distribution on the unit square. For each positive $n \in \mathbb{N}$, define $\mathbf{X}^n = \langle X_1^n, X_2^n \rangle$ to be the random vector such that $X_1^n = \mathbf{e}_1$ and $X_2^n = 1/n \cdot \mathbf{e}_1 + \mathbf{e}_2$. Let $M^n = \langle \mathbf{X^n}, \mathbf{e}, B^n \rangle$ be the resulting linear model, where B^n is the 2×2 matrix that has $1/n$ in the upper right corner and zeroes everywhere else. So each M^n generates the DAG G. Similarly, for each n, define $Y_1^n = e_1 + 1/n \cdot e_2$ and $Y_2^n = e_2$; let N^n be the resulting linear model with DAG H. By Slutsky's theorem, $\mathbf{Y}^n, \mathbf{X}^n \Rightarrow \mathbf{e}$ and $P_{M^n}, P_{N^n} \Rightarrow P$, where P is the distribution on $\mathbb{R}^2$ generated by $\mathbf{e}$.

Now suppose for a contradiction that method λ is uniformly consistent for $\mathfrak{Q}$. Then, there must be some sample size k such that for all n, we have that $P^k_{M^n}(\lambda_k = \mathcal{M}_G) > 1/2$ and $P^k_{N^n}(\lambda_k = \mathcal{M}_H) > 1/2$. Since P is the uniform distribution on the unit square, P^k is absolutely continuous with Lebesgue measure on $\mathbb{R}^{2k}$ and $P^k(\partial\lambda_k^{-1}(\mathcal{M}_G)) = 0$. Therefore, $P^k_{M_n}(\lambda_k = \mathcal{M}_G), P^k_{N_n}(\lambda_M = \mathcal{M}_G) \Rightarrow P^k(\lambda_k = \mathcal{M}_G)$ as n grows large. Therefore, for large enough n, $|P^k_{M^n}(\lambda_k = \mathcal{M}_G) - P^k_{N^n}(\lambda_k = \mathcal{M}_G)|$ is small and $P^M_{N_n}(\lambda_k = \mathcal{M}_G) > 1/2$. Contradiction. □

We introduce two intermediate notions of success between pointwise and uniform consistency. The first requires that the method never produce a false answer with probability greater than α, for some fixed $\alpha > 0$. The method can avoid false conclusions by sometimes producing the uninformative answer $\mathcal{M}$. Given some $\alpha > 0$, say that a method λ is an α-**decision procedure** if (1) λ is consistent and (2) $P^n_M(M \notin \lambda_n) \leq \alpha$ for all models $M \in \bigcup \mathfrak{Q}$ and all sample sizes n. Call a question **statistically decidable** (or simply decidable) if there is α-decision procedure for all $\alpha > 0$.

Theorem 3.1. *Suppose that $P : M \mapsto P_M$ is injective and every element of $P[\mathcal{M}]$ is absolutely continuous with Lebesgue measure. Suppose $\mathfrak{Q}$ is countable. Then, $\mathfrak{Q}$ is decidable iff $P[\mathcal{A}]$ is open in the weak topology on $P[\mathcal{M}]$ for each $\mathcal{A} \in \mathfrak{Q}$.*

Proof. The theorem is a minor modification of Theorem 3.2.4 in [Genin, 2018] □

Many interesting questions are not decidable. We introduce another intermediate success notion that is achievable when no decision procedures are available. Say that a consistent method is $\alpha-$**progressive** if the chance of outputting the correct answer increases "almost monotonically" as samples increase, i.e. $P^{n_1}_M(\lambda_{n_1} = \mathfrak{Q}(M)) - P^{n_2}_M(\lambda_{n_2} = \mathfrak{Q}(M)) < \alpha$ for all $n_1 < n_2$ and all $M \in \bigcup \mathfrak{Q}$. Say that a question $\mathfrak{Q}$ is **progressively solvable** iff there exists an α-progressive method for $\mathfrak{Q}$ for every $\alpha > 0$. Progressiveness ensures that collecting a larger sample is never a disastrous idea. Failing to satisfy progressiveness amounts to building in a disposition to fail to replicate true results. Surprisingly, many standard frequentist methods do not satisfy this criterion [Chernick and Liu, 2002]. However, it is often possible to satisfy progressiveness even though you cannot bound the chance of producing false conclusions. Genin [2018] proves the following (see Theorem 3.6.3).

Theorem 3.2. *Suppose that $P : M \mapsto P_M$ is injective and every element of $P[\mathcal{M}]$ is absolutely continuous with Lebesgue measure. Suppose that (1) $\mathfrak{Q}$ is a countable partition of $\mathcal{M}$ and (2) there exists an enumeration $\mathcal{A}_1, \mathcal{A}_2, \ldots,$ of the elements of $\mathfrak{Q}$ such that $\mathcal{A}_i \cap \mathsf{cl}(\mathcal{A}_j) = \varnothing$ whenever $i > j$. Then, for every $\alpha > 0$ there exists a consistent, α-progressive method for $\mathfrak{Q}$.*

We are now ready to state and prove our main result.

4 Main Result

Say that the model $M \in \mathrm{L}_d$ has **causal coefficients bounded above by** c if $|B_{ij}(M)| \leq c$ for all i, j. Let $\textsc{Lng}^c_d \subseteq \textsc{Lng}_d$ bet the set of all $M \in \textsc{Lng}_d$ such that (1) $P(M)$ is absolutely continuous with Lebesgue measure and (2) M has causal coefficients bounded above by c. In most applications, if we know a priori that the true model belongs to $\textsc{Lng}_d$ (and that noise is continuous), then there is c such that we also know that the true model belongs to $\textsc{Lng}^c_d$. Typically, it suffices to let c equal the number of particles in the universe.

In this section, we prove the following theorem.

Theorem 4.1. *Let $\mathcal{M}_{i\to j} = \{M \in \textsc{Lng}^c_d : i \to j \in G(M)\}$. Then, $P[\mathcal{M}_{i\to j}]$ is open in the weak topology on $P[\textsc{Lng}^c_d]$.*

By the example in the previous section, it follows that there are statistical decision procedures for determining the orientation of a causal edge although there is no uniformly consistent procedure for the same question.

The following are easy corollaries of Theorem 4.1.

Corollary 4.1. *The question $\mathfrak{Q} = \{\mathcal{M}_{i\to j}, \mathcal{M}_{i\leftarrow j}\}$ is statistically decidable.*

Corollary 4.2. *Let $\mathcal{M}_{i\circ j} = \textsc{Lng}^c_d \setminus \{\mathcal{M}_{i\to j}, \mathcal{M}_{i\leftarrow j}\}$. The question $\mathfrak{Q} = \{\mathcal{M}_{i\circ j}, \mathcal{M}_{i\to j}, \mathcal{M}_{i\leftarrow j}\}$ is progressively solvable.*

Corollary 4.3. *Let $\mathcal{M}_G = \{M \in \text{LNG}_d^c : G(M) = G\}$. Then, $\mathfrak{Q} = \{\mathcal{M}_G : G \in \text{DAG}_d\}$ is progressively solvable.*

Proof of Corollary 4.1. Immediate from Theorems 2.3, 3.1 and 4.1. □

Proof of Corollary 4.2. By Theorem 2.3, the elements of $\mathfrak{Q}$ are disjoint. It is sufficient to show that the ordering $\mathcal{M}_{i \circ j}, \mathcal{M}_{i \to j}, \mathcal{M}_{i \leftarrow j}$ satisfies the conditions of Theorem 3.2. By Theorem 4.1, $\mathcal{M}_{i \to j}, \mathcal{M}_{i \leftarrow j}$ are both open in the weak topology and so is their union. Therefore $\mathcal{M}_{i \circ j}$ is closed and $\varnothing = \mathcal{M}_{i \to j} \cap \mathsf{cl}(\mathcal{M}_{i \circ j}) = \mathcal{M}_{i \leftarrow j} \cap \mathsf{cl}(\mathcal{M}_{i \circ j})$. By Theorem 4.1, $\mathcal{M}_{i \leftarrow j}$ and $\mathcal{M}_{i \to j}$ are separated by open sets in the weak topology, so $\mathcal{M}_{i \leftarrow j} \cap \mathsf{cl}(\mathcal{M}_{i \to j}) = \varnothing$. □

Proof of Corollary 4.3. Let $\preceq$ be the partial order on DAGs over d variables induced by setting $G \preceq G'$ iff G' has all the edges that G has. By the order extension principle, it is possible to extend this partial order to a total order $\preceq^*$ over all DAGs on d variables. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ be an enumeration of the elements of $\mathfrak{Q}$ that agrees with $\preceq^*$, i.e. if $i < j$ then all $M \in \mathcal{A}_i, N \in \mathcal{A}_j$ have $G(M) \preceq^* G(N)$. Then, if $i < j$, there is some edge $a \to b$ such that all $M \in \mathcal{A}_j$ have $a \to b \in G(M)$ but no $N \in \mathcal{A}_i$ have $a \to b \in G(N)$. By Theorem 4.1, $\mathcal{A}_j$ is separated from $\mathcal{A}_i$ by the open set $\mathcal{M}_{a \to b}$ and $\mathcal{A}_j \cap \mathsf{cl}(\mathcal{A}_i) = \varnothing$.The corollary follows by Theorems 2.3 and 3.2. □

By Corollary 4.2, there exist progressive methods for learning causal orientation, even when we do not know *a priori* whether any given edge exists. By conjoining the conclusions of a collection of such methods (one for each possible edge), it is possible to construct consistent methods that converge to the true DAG as sample sizes increase, without exhibiting the flipping behavior identified by Kelly and Mayo-Wilson [2010].

It remains to prove Theorem 4.1. The proof of Theorem 4.1 relies mainly on Lemma 2.1 and the Lukacs-King theorem. However, it also depends on the following lemma.

Lemma 4.1. *Suppose that the random vector (X, Y) is absolutely continuous with Lebesgue measure and that X, Y are dependent. Then, if $(X_i, Y_i) \Rightarrow (X, Y)$ all but finitely many of the X_i, Y_i are dependent.*

Proof. Let P be the measure on $\mathbb{R}^2$ induced by (X, Y) and P_n be the measure induced by (X_n, Y_n). If (X, Y) are dependent, there must be $b, c \in \mathbb{R}$ such that $A = \{(x, y) : x \leq c, y \leq d\}, B = \{(x, y) : x \leq b\}, C = \{(x, y) : y \leq c\})$ and $P(A) \neq P(B)P(C)$. Since P is a.c. with Lebesgue measure we have that $P(\partial A) = P(\partial B) = P(\partial C) = 0$. Therefore $P_n(A) \Rightarrow P(A)$, $P_n(B) \Rightarrow P(B)$ and $P_n(C) \Rightarrow P(C)$, from which it follows that for all but finitely many n, $P_n(A) \neq P_n(B)P_n(C)$. □

Proof of Theorem 4.1. Suppose for a contradiction that $P[\mathcal{M}_{i \to j}]$ is not open in the weak topology. Then there is $M = \langle \mathbf{X}, \mathbf{e}, B \rangle$ in $\mathcal{M}_{i \to j}$ and $M_n = \langle \mathbf{X_n}, \mathbf{e_n}, A_n \rangle$ all in $\text{LNG}_d^c \setminus \mathcal{M}_{i \to j}$ such that $\mathbf{X}_n \Rightarrow \mathbf{X}$. Suppose, without loss of generality, that the observable variables are enumerated in agreement with the causal order of $G(M)$. Since $\mathbf{X}_n \Rightarrow \mathbf{X}$:

$$(I - A_n)^{-1}\mathbf{e_n} \Rightarrow (I - B)^{-1}\mathbf{e}.$$

Since we have assumed that the A_n are bounded, it follows by the Bolzano-Weierstrass theorem that there must be some subsequence (A_{n_i}) converging in the Euclidean metric to a matrix A. It follows by Slutsky's theorem that

$$\mathbf{e_{n_i}} \Rightarrow (I - A)(I - B)^{-1}\mathbf{e}.$$

The matrix $(I - A)$ has unit diagonal since each of $(I - A_n)$ does. $(I - B)^{-1}$ also has unit diagonal. Because the observable variables are, by assumption, enumerated in agreement with the causal order of $G(M)$, we have that $(I - B)^{-1}$ is lower triangular.

Suppose that $(I - A)(I - B)^{-1}$ is a mixing matrix. Let

$$\mathbf{e}' = (I - A)(I - B)^{-1}\mathbf{e}$$

Then, by Lukacs-King, there must be two elements of $\mathbf{e}'$ that are dependent. Therefore, by Lemma 4.1, all but finitely many of the same elements of $\mathbf{e_{n_i}}$ must also be dependent. Contradiction.

Suppose that $(I - A)(I - B)^{-1}$ is not a mixing matrix. By Lemma 2.1, $A = B$. But since $M \in \mathcal{M}_{i \to j}$ and the M_n are in $\textsc{Lng}_d^c \setminus \mathcal{M}_{i \to j}$, we have that $B_{ij} \neq 0$ but $(A_{n_i})_{ij} = 0$ for all n_i. So the A_{n_i} cannot converge to B. Contradiction. □

Thusfar, we have proven the *existence* of statistical decision procedures and the *existence* of progressive methods for several important questions about LiNGAMs. It is therefore natural to ask, "Which causal discovery algorithms that are *currently in use*, if any, are α-statistical decision procedures or α-progressive, and for which questions and which values of α?" We do not know the answers to those questions, but it is easy to show that some existing algorithms have worst-case error rates that empirical scientists might find alarming.

Example 2: We used the DirectLiNGAM algorithm [Shimizu et al., 2011, Hyvärinen and Smith, 2013] to analyze 1000 simulated samples of size 50 and 500, each drawn the following LiNGAM. The underlying DAG is $X_1 \to X_2 \to X_3$. The exogenous variable $X_1 = \mathbf{e}_1$ is uniform on $\{1, 2 \ldots, 20\}$, and the variables X_2 and X_3 have independent Bernoulli error terms $\mathbf{e}_2$ and $\mathbf{e}_3$ respectively, each with parameter $1/2$. We let $X_2 = X_1 + \mathbf{e}_2$ and $X_3 = X_2 + \mathbf{e}_3$, so that all edge coefficients are either zero or one. The results are displayed in the table below.

Sample Size	Models with $X_2 \to X_3$	Models with $X_3 \to X_2$
50	55%	45%
500	92%	8%

The good news is that DirectLiNGAM, as one would suspect, identifies the correct direction of the $X_2 \to X_3$ edge over 90% of the time at sample size 500. The bad news is the algorithm is only slightly better than chance at sample size 50. Thus, DirectLiNGAM appears not to be an α-decision procedure for the edge orientation question, even for values of α near $1/2$.

□

In Example 2, DirectLiNGAM's poor performance at sample size 50 is to be expected: the number of possible values for X_1 is large in comparison to the sample size, and so the simulated samples often do not contain every combination of values in the support of the random variables. There are likely no algorithms that are reliable in such settings. So we are *not* suggesting that DirectLiNGAM or related algorithms should be abandoned. But Example 2 indicates two ways in which some discovery algorithms could be modified to better address users' interests, which is important as algorithms are adopted by working scientists and policy-makers.

First, algorithms should, we think, be designed to indicate "not enough evidence" in response to weak data (which is represented by returning the uninformative answer in our framework). In our simulations, DirectLiNGAM *always* orients the edge between X_2 and X_3 in some way, even at low sample sizes. Taking a stand is not always necessary. Second, to determine when to take a stand, users could be prompted for their desired error probability α, as long as the question is decidable or progressively solvable.

5 Conclusions and Future Research

We have proven that, when the data is generated by a LiNGAM, there *exist* statistical decision procedures and progressive methods, depending upon one's question. As noted above, our work immediately raises the question, "For which questions about LiNGAMs and for which values of α, if any, are existing algorithms α-statistical decision procedures? α-progressive?"

Our research raises at leas three other important questions for future research. First, our results assume that the observable variables are *causally sufficient*, i.e., that there are no unobserved common causes of two observed variables. Which questions about LiNGAMs are decidable or progressively solvable in the presence of confounders? Second, for what other classes of causal models are questions about edge orientation (or questions about the entire DAG) decidable and/or progressively solvable? Finally, in many applications, not all variables in a causal model can be observed simultaneously [Mayo-Wilson, 2013, 2018]. Which causal questions about LiNGAMs and other non-parametric models, if any, are decidable and/or progressively solvable when only a few variables can be observed at any given time?

References

P. Billingsley. *Probability and measure*. John Wiley & Sons, New York, second edition, 1986.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. Publisher: Institute of Mathematical Statistics.

M.R. Chernick and C.Y. Liu. The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. *The American Statistician*, 56(2): 149–155, 2002. Publisher: Taylor & Francis.

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

K. Genin. *The Topology of Statistical Inquiry*. Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA, 2018.

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

P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.

A. Hyvärinen and S.M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

K.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.

P. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014. Publisher: JMLR. org.

E. Lukacs and E.P. King. A property of the normal distribution. *The Annals of Mathematical Statistics*, 25(2):389–394, 1954. Publisher: Institute of Mathematical Statistics.

C. Mayo-Wilson. The Limits of Piecemeal Causal Inference. *The British Journal for the Philosophy of Science*, 2013. doi: 10.1093/bjps/axs030.

C. Mayo-Wilson. Causal identifiability and piecemeal experimentation. *Synthese*, May 2018. ISSN 1573-0964. doi: 10.1007/s11229-018-1826-4. URL `https://doi.org/10.1007/s11229-018-1826-4`.

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962–1030, 2002.

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

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

V.P. Skitivic. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, (89):217–219, 1953.