Biometrika (2018), **103**, 1, *pp*. 1–19
Printed in Great Britain Advance Access publication on XX

High-dimensional Causal Discovery Under Non-Gaussianity

BY Y. SAMUEL WANG

Booth School of Business, The University of Chicago
Chicago, Illinois 60615, U.S.A
swang24@uchicago.edu

AND MATHIAS DRTON

Department of Mathematical Sciences, University of Copenhagen
Universitetsparken 5, 2100 Copenhagen Ø, Denmark
Department of Statistics, University of Washington, Box 354322
Seattle, Washington 98195, U.S.A
md5@uw.edu

SUMMARY

We consider graphical models based on a recursive system of linear structural equations. This implies that there is an ordering, σ, of the variables such that each observed variable Y_v is a linear function of a variable specific error term and the other observed variables Y_u with $\sigma(u) < \sigma(v)$. The causal relationships, i.e., which other variables the linear functions depend on, can be described using a directed graph. It has been previously shown that when the variable specific error terms are non-Gaussian, the exact causal graph, as opposed to a Markov equivalence class, can be consistently estimated from observational data. We propose an algorithm that yields consistent estimates of the graph also in high-dimensional settings in which the number of variables may grow at a faster rate than the number of observations, but in which the underlying causal structure features suitable sparsity; specifically, the maximum in-degree of the graph is controlled. Our theoretical analysis is couched in the setting of log-concave error distributions.

Some key words: Causal discovery; Directed graphical model; High-dimensional statistics; Structural equation model; Non-Gaussian data

1. INTRODUCTION

Prior work shows the possibility of causal discovery with observational data in the framework of linear structural equation models with non-Gaussian errors. However, existing methods for estimation of the causal structure are applicable only in low-dimensional settings, in which the number of variables, p, is small compared to the sample size, n. In this paper, we develop a method which, given suitable sparsity, recovers the exact causal structure consistently in high-dimensional regimes where p grows along with n. Careful considerations of computational aspects make our method a practical and statistically sound exploratory tool for the intended high-dimensional settings.

Let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ be multivariate data from an observational study, specifically, the observations form an independent, identically distributed sample. We encode the causal structure generating dependences in the underlying p-variate joint distribution by a graph $G = (V, E)$ with vertex set $V = \{1, \ldots, p\}$. Each node, $v \in V$, corresponds to an observed variable in

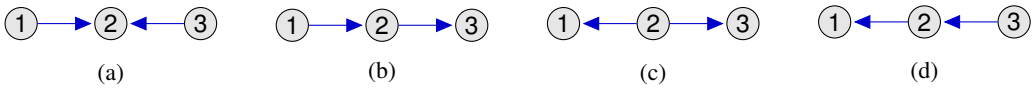


Fig. 1: The Markov equivalence class of graph (a) is a singleton. However, graph (b), (c) and (d) are Markov equivalent and imply the same set of conditional independences.

$Y_i = (Y_{vi})_{v \in V}$, and each directed edge, $(u, v) \in E$, indicates that Y_{ui} has a direct causal effect on Y_{vi}. Thus, positing causal structure is equivalent to selecting a graph. We will only consider directed acyclic graphs (DAGs), directed graphs which do not contain directed cycles. Given the correspondence between a node $v \in V$ and the random variable Y_{vi}, we will at times let v stand in for Y_{vi}; for instance, when stating stochastic independence relations.

Discovery of causal structure from observational data is difficult because of the super-exponential set of possible models, some of which may be indistinguishable from others. Despite this difficulty, many methods for causal discovery have been developed and have seen fruitful applications; see the recent review of Drton and Maathuis (2017). In particular, the celebrated PC algorithm (Spirtes et al., 2000) is a constraint-based method which first infers a set of conditional independence relationships and then identifies the associated Markov equivalence class; this class contains all DAGs compatible with the inferred conditional independences. Kalisch and Bühlmann (2007) show if the maximum total degree of the graph is controlled and the data is Gaussian, then the PC algorithm can consistently recover the true Markov equivalence class even in high-dimensional settings where the number of variables grows with the number of samples. Harris and Drton (2013) extend the result to Gaussian copula models using rank correlations.

However, graphs within the same Markov equivalence class may have drastically different causal and scientific interpretations. For the graphs in Figure 1, conditional independence tests can distinguish model (a) from the rest but cannot distinguish models (b), (c), and (d) from each other. Although Maathuis et al. (2009) provide a procedure for bounding the size of a causal effect over graphs within an equivalence class, interpretation of the set of possibly conflicting graphs can remain difficult. Results on the size and number of Markov equivalence classes, which may be exponentially large, can be found e.g. in Steinsky (2013).

In contrast, it has been shown that under various additional assumptions, the exact graph structure, not just an equivalence class, can be identified from observational data (Loh and Bühlmann, 2014; Peters and Bühlmann, 2014; Rothenhäusler et al., 2018). In particular, Shimizu et al. (2006) show this to be the case under three main assumptions: (1) the data are generated by a linear structural equation model, (2) the error terms in the structural equations are non-Gaussian, and (3) there is no unobserved confounding among the observed variables; i.e., errors are independent. These assumptions yield the linear non-Gaussian acyclic model, abbreviated as LiNGAM, which is described formally in Section 2·1. Under the LiNGAM framework, the four models from Figure 1 are mutually distinguishable. Shimizu et al. (2006) use independent component analysis to estimate the graph structure, and the subsequent DirectLINGAM (Shimizu et al., 2011) and Pairwise LiNGAM (Hyvärinen and Smith, 2013) methods iteratively select a causal ordering by computing pairwise statistics. These methods are motivated by identifiability results that are derived by iteratively forming conditional expectations. In practice, the conditional expectations are estimated using larger and larger regression models. As a result, the methods become inapplicable when the number of variables exceeds the sample size.

We develop a modification of the DirectLiNGAM algorithm that is suitable for high-dimensional data and give guarantees for when our algorithm will consistently recover the true

graph in high-dimensional asymptotic scenarios. Most notably, our analysis considers restricted maximum in-degree of the graph and assumes log-concave distributions. The theory also applies to hub graphs where the maximum out-degree may grow with the size of the graph, which is in contrast to the conditions needed for high-dimensional consistency of the PC algorithm (Kalisch and Bühlmann, 2007). Hub graphs appear in many biological networks (Hao et al., 2012).

2. Causal discovery algorithm

2·1. *Generative model and notation*

We assume that the observations $Y_1, \ldots, Y_n \in \mathbb{R}^p$ are independent, identically distributed replications generated from a linear structural equation model so that the elements of each random vector Y_i satisfy

$$Y_{vi} = \sum_{u \neq v} \beta_{vu} Y_{ui} + \varepsilon_{vi}, \tag{1}$$

where the β_{vu} are unknown real parameters that quantify the direct linear effect of variable u on variable v, and ε_{vi} is an error term of unknown distribution P_v. We assume ε_{vi} has mean 0 and is independent of all other error terms. Our interest is in models that postulate that a particular set of coefficients β_{vu} is zero. In particular, the absence of an edge, $(u, v) \notin E$, indicates that the model constrains the parameter β_{vu} to zero. We assume that the graph, G, representing the model is a DAG, which implies that the structural equation model is recursive; i.e., there exists a permutation of the variables, σ, such that β_{vu} is constrained to be zero unless $\sigma(u) < \sigma(v)$.

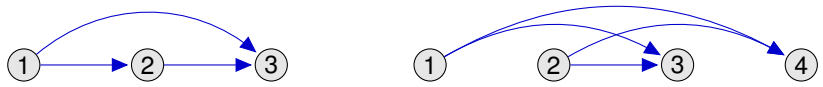
We denote the model given by graph G by $\mathcal{P}(G)$. Each distribution $P \in \mathcal{P}(G)$ is induced through a choice of linear coefficients $(\beta_{vu})_{(u,v)\in E}$ and error distributions $(P_v)_{v\in V}$. Let $B = (\beta_{vu})$ be the $p \times p$ matrix determined by the model constraints and the chosen free coefficients. Then the equations in (1) admit a unique solution with $Y_i = (I - B)^{-1}\varepsilon_i$. The error vectors $\varepsilon_i = (\varepsilon_{vi})_{v\in V}$ are independent and identically distributed and follow the product distribution $\otimes_{v\in V} P_v$. The distribution P is then the joint distribution for Y_i that is induced by the transformation of ε_i.

Standard notation has the set $\mathrm{pa}(v) = \{u : (u, v) \in E\}$ comprise the parents of a given node v. The set of ancestors, $\mathrm{an}(v)$, contains any node $u \neq v$ with a directed path from u to v; we let $\mathrm{An}(v) = \mathrm{an}(v) \cup \{v\}$. The set of descendants, $\mathrm{de}(v)$, contains the nodes u with $v \in \mathrm{an}(u)$.

2·2. *Parental faithfulness*

An important approach to causal discovery begins by inferring relations such as conditional independence and then determines graphs compatible with empirically found relations. For this approach to succeed, the considered relations must correspond to structure in the graph G as opposed to a special choice of parameters. In the context of conditional independence, the assumption that any relation present in an underlying joint distribution $P \in \mathcal{P}(G)$ corresponds to the absence of certain paths in G is known as the faithfulness assumption; see Uhler et al. (2013) for a detailed discussion. For our work, we define a weaker condition, parental faithfulness. In particular, if $u \in \mathrm{pa}(v)$, we require that the total effect of u on v does not vanish when we modify the considered distribution by regressing v onto any set of its non-descendants, as detailed next.

Let $l = (v_1, \ldots, v_z)$ be a directed path in G, so $(v_j, v_{j+1}) \in E$ for $j = 1, \ldots, z - 1$. Given coefficients $(\beta_{vu})_{(u,v)\in E}$, the path has weight $w(l) = \prod_{j=1}^{z-1} \beta_{v_{j+1},v_j}$. Let $\mathcal{L}_{vu}$ be the set of all directed paths from u to v. Then the total effect of u on v is $\pi_{vu} = \sum_{l\in\mathcal{L}_{vu}} w(l)$, with $\pi_{vu} = 0$ if $u \notin \mathrm{An}(v)$ and $\pi_{vu} = 1$ if $u = v$. The effect gives the conditional mean of v under interventions



(a) (b)

Fig. 2: In (a), the choice $\beta_{31} = \beta_{21} = 1$ and $\beta_{32} = -1$ results in parental unfaithfulness because $\pi_{31.\emptyset} = 0$. Also, the choice $\beta_{31} = \beta_{21} = \beta_{32} = 1$ and $E(\varepsilon_1^2) = E(\varepsilon_2^2) = E(\varepsilon_3^2) = 1$ is not faithful because the partial correlation of 2 and 1 given 3 is 0, but is still parentally faithful. In (b), the choice $\beta_{31} = \beta_{32} = \beta_{42} = 1$, $\beta_{41} = 2$, and $E(\varepsilon_1^2) = E(\varepsilon_2^2) = E(\varepsilon_3^2) = 1$ results in parental unfaithfulness because $\pi_{42.3} = 0$.

on u; i.e., $\pi_{vu} = E(Y_{vi} \mid \mathrm{do}(Y_{ui} = y + 1)) - E(Y_{vi} \mid \mathrm{do}(Y_{ui} = y))$ using the do-operator of Pearl (2009). Total effects may be calculated by matrix inversion, $\Pi = (\pi_{vu})_{u,v \in V} = (I - B)^{-1}$.

Let $\Sigma = E(Y_i Y_i^t)$ be the covariance matrix of the, for convenience, centered random vector $Y_i \sim P$. Let Σ_{CC} be the principal sub-matrix for a non-empty set of indices $C \subseteq V$. For $v \in V \setminus C$, let Σ_{Cv} be the sub-vector comprised of the entries in places (c, v) for $c \in C$. Let

$$\beta_{vC} \;=\; (\beta_{vc.C})_{c \in C} = (\Sigma_{CC})^{-1} \Sigma_{Cv} \tag{2}$$

be the population regression coefficients when v is regressed onto C. The quantity $\beta_{vc.C}$ is defined even if $(c, v) \notin E$, and in general $\beta_{vc.C} \neq \beta_{vc}$ even if $(c, v) \in E$. A pair $(u, v) \in E$ is parentally faithful if for any set $C \subseteq V \setminus [\mathrm{de}(v) \cup \{v, u\}]$, the residual total effect defined as

$$\pi_{vu.C} \;=\; \pi_{vu} - \sum_{c \in C} \beta_{vc.C} \pi_{cu} \tag{3}$$

is nonzero. If this holds for every pair $(u, v) \in E$, we say that the joint distribution P is parentally faithful with respect to G. Parental faithfulness only pertains to the linear coefficients and error variances, and the choices for which parental faithfulness fails form a set of Lebesgue measure zero. The concept is exemplified in Figure 2.

2·3. *Test statistic*

Reliable determination of the causal direction between u and v generally requires removal of all confounding. Thus, Shimizu et al. (2011) and Hyvärinen and Smith (2013) adjust v and u for all x such that $\sigma(x) < \sigma(v)$ and $\sigma(x) < \sigma(u)$. However, adjusting by an increasingly larger set of variables propagates error proportional to the number of variables, rendering high-dimensional estimation inconsistent, or impossible when the size of the adjustment set exceeds the sample size. On the other hand, restricting the size of the adjustment sets may not remove confounding completely. The method we present solves this problem via a statistic that is conservative in the sense that it does not mistakenly certify causal direction when confounding is present.

Shimizu et al. (2011) calculate the kernel-based mutual information between v and the residuals of u when it is regressed onto v. The corresponding population information is positive if and only if $v \in \mathrm{de}(u)$ or there is uncontrolled confounding between v and u, that is, u and v have a common ancestor even when certain edges are removed from the graph. Hence, the mutual information can be used to test the hypothesis that $v \notin \mathrm{de}(u)$ versus the hypothesis that $v \in \mathrm{de}(u)$ or there is confounding between u and v. Unfortunately, calculating the mutual information can be computationally burdensome, so Hyvärinen and Smith (2013) propose a different parameter R_{vu}. Without confounding, $R_{vu} > 0$ if $v \in \mathrm{an}(u)$ and $R_{vu} < 0$ if $u \in \mathrm{an}(v)$. With confounding, however, the parameter can take either sign, so it cannot be reliably used if we remain uncertain

about whether or not confounding occurs. We introduce a parameter that shares the favorable properties of the mutual information but admits computationally inexpensive estimators that are rational functions of the sample moments of Y, which facilitates analysis of error propagation.

The parameter we consider is motivated by the following observation. Suppose the true generating mechanism is $Y_1 \to Y_2$ so that $Y_1 = \varepsilon_1$ and $Y_2 = \beta_{21} Y_1 + \varepsilon_2$ for ε_1 independent of ε_2. When the causal direction is correctly specified, the linear coefficient β_{21} is recovered by $\mathbb{E}(Y_1^{K-1} Y_2)/\mathbb{E}(Y_1^K)$ for all integers K greater than 1 for which $\mathbb{E}(Y_1^K) \neq 0$. Of course, letting $K = 2$ gives the typical least squares estimator. This leads to the identity $\mathbb{E}(Y_1^{K-1} Y_2)/\mathbb{E}(Y_1^K) = \mathbb{E}(Y_1 Y_2)/\mathbb{E}(Y_1^2)$ which implies $\mathbb{E}(Y_1^{K-1} Y_2)\mathbb{E}(Y_1^2) - \mathbb{E}(Y_1 Y_2)\mathbb{E}(Y_1^K) = 0$, which holds even when $\mathbb{E}(Y_1^K) = 0$. In general, however, when the errors are non-Gaussian and the roles of Y_1 and Y_2 are swapped, this identity does not hold. When there are more than 2 variables involved, we reduce the problem to a bivariate problem by conditioning on an appropriate set C. To this end, define for any i the residual

$$Y_{vi.C} = Y_{vi} - \sum_{c \in C} \beta_{vc.C} Y_{ci},$$

where $\beta_{vc.C}$ are the population regression coefficients from (2). When $C = \emptyset$, let $Y_{vi.\emptyset} = Y_{vi}$.

THEOREM 1. *Let $P \in \mathcal{P}(G)$ be a distribution in the model given by a DAG G, and let $Y_i \sim P$. For $K > 2$, two distinct nodes u and v, and any set $C \subseteq V \setminus \{u, v\}$, define*

$$\tau^{(K)}_{v.C \to u} \;=\; E_P(Y_{vi.C}^{K-1} Y_{ui}) E_P(Y_{vi.C}^2) - E_P(Y_{vi.C}^K) E_P(Y_{vi.C} Y_{ui}). \tag{4}$$

(i) *If $u \notin \mathrm{pa}(v)$ and $\mathrm{pa}(v) \subseteq C \subseteq V \setminus [\mathrm{de}(v) \cup \{v, u\}]$, then $\tau^{(K)}_{v.C \to u} = 0$.*

(ii) *Suppose $u \in \mathrm{pa}(v)$ with u, v parentally faithful under the covariance matrix of P. If $C \subseteq V \setminus [\mathrm{de}(v) \cup \{v, u\}]$, then $\tau^{(K)}_{v.C \to u} \neq 0$ for generic error moments of order $3, \ldots, K$.*

Estimators $\hat{\tau}^{(K)}_{v.C \to u}$ of the parameter from (4) are naturally obtained from empirical regression coefficients and empirical moments.

In Theorem 1(ii), the term generic indicates that the set of error moments for which this statement does not hold has Lebesgue measure zero. Given that there is a finite number of sets $C \subset V$, the union of all exceptional sets is also a null set. A detailed proof of Theorem 1 is included in the supplement. Claim (i) can be shown via direct calculation, and we give a brief sketch of (ii) here. For fixed coefficients $(\beta_{vu})_{(u,v) \in E}$ and set $C \subset V$, $\tau^{(K)}_{v.C \to u}$ is a rational function of the error moments. Thus existence of a single choice of error moments for which $\tau^{(K)}_{v.C \to u} \neq 0$ is sufficient to show that the statement holds for generic error moments. As the argument boils down to showing that a certain polynomial is not the zero polynomial (Okamoto, 1973), the choice considered need not necessarily be realizable by a particular distribution. In particular, we choose all moments of order less than K equal to those of the centered Gaussian distribution with variance $\sigma_v^2 = E(\varepsilon_v^2)$, but for the Kth moment we add an offset $\eta_v > 0$, so

$$E(\varepsilon_v^K) = \begin{cases} \eta_v & \text{if } K \text{ is odd,} \\ (K-1)!! \sigma_v^K + \eta_v & \text{if } K \text{ is even,} \end{cases} \tag{5}$$

where $q!! = \prod_{z=0}^{\lceil q/2 \rceil - 1} (q - 2z)$ is the double factorial of q. If there is no confounding between $Y_{v.C}$ and Y_u, that is, no ancestor of u is the source of a directed path to v that avoids $C \cup \{u\}$, then

$$\tau^{(K)}_{v.C \to u} \;=\; \pi_{vu.C} \left(\pi_{vu.C}^{K-2} \, \eta_u \sigma_v^2 - \eta_v \sigma_u^2 \right) \tag{6}$$

with $\pi_{vu.C} \neq 0$, by the assumed parental faithfulness. Thus, a choice of offsets with $\pi_{vu.C}^{K-2}\eta_u\sigma_v^2 \neq \eta_v\sigma_u^2$ implies $\tau^{(K)}_{v.C\to u} \neq 0$. A more involved but similar argument can be made in the case of confounding. Under a slightly stronger form of faithfulness, $\tau^{(k)}_{v.C\to u} \neq 0$ if there is confounding regardless of whether $u \in \mathrm{pa}(v)$; see supplement Remark 1.

COROLLARY 1. *Let P_v and P_u be two distributions that each have all moments up to order K equal to those of some Gaussian distribution. Then there exists a graph G, for which $u \in \mathrm{pa}(v)$, and distributions P which are parentally faithful with respect to G, but $\tau^{(K)}_{v.C\to u} = 0$ for some set $C \subseteq V \setminus [\mathrm{de}(v) \cup \{v, u\}]$.*

Proof. The moments of P_v and P_u satisfy (5) with $\eta_v = \eta_u = 0$. Consequently, if there exists a set C such that there is no confounding between $Y_{v.C}$ and Y_u, then $\tau_{v.C\to u}$ satisfies (6), the right-hand side of which is zero when $\eta_v = \eta_u = 0$. For example, if ε_v and ε_u are both Gaussian and the graph is $u \to v$, $\tau^{(K)}_{v\to u} = 0$ for all choices of β_{vu} and all K. □

Corollary 1 confirms that the null set to be avoided in Theorem 1(ii) contains points for which all error moments are consistent with some Gaussian distribution. Thus, our identification of causal direction requires that the error moments of order at most K be inconsistent with all Gaussian distributions. In practice, we consider the case $K = 3, 4$ and recommend $K = 4$ unless one is certain the errors are not symmetric. We refer readers to Hoyer et al. (2008) for a full characterization of when graphs with both Gaussian and non-Gaussian errors are identifiable.

At each step, the high-dimensional LiNGAM algorithm presented in Section 3·1 considers a sub-graph and searches for a root node, i.e., a node without any parents. Suppose $\mathrm{de}(V_2) = V_2 \subseteq V$; if $v \in V_2$ is not a root in the sub-graph induced by V_2, then there must exist some $u \in V_2$ with $\tau_{v.C\to u} \neq 0$ for all sets C which are upstream of v and V_2. However, if v is a root, then $\tau_{v.C\to u} = 0$ for all $u \in V_2$ when $C = \mathrm{pa}(v)$. Thus, to test whether v is a root in V_2, we aggregate the various τ parameters corresponding to $u \in V_2$ and conditioning sets C. Corollary 2 describes two ways to do this aggregation. If v is a root, this quantity will be 0, and if v is not a root, it will be positive.

COROLLARY 2. *Let $P \in \mathcal{P}(G)$, let $v \in V$, and consider two disjoint sets $V_1, V_2 \subseteq V \setminus \{v\}$. For a chosen non-negative integer J, define*

$$T_1^{(K)}(v, V_1, V_2) = \min_{C\in V_1(J)} \max_{u\in V_2} |\tau^{(K)}_{v.C\to u}|, \qquad T_2^{(K)}(v, V_1, V_2) = \max_{u\in V_2} \min_{C\in V_1(J)} |\tau^{(K)}_{v.C\to u}|,$$

where $V_1(J) = \{C \subseteq V_1 : |C| = J\}$ if $J \leq |V_1|$ and $V_1(J) = V_1$ if $J \geq |V_1|$.

(i) *If $|\mathrm{pa}(v)| \leq J$ and $\mathrm{pa}(v) \subseteq V_1 \subseteq V \setminus \mathrm{de}(v)$, then*

$$T_1^{(K)}(v, V_1, V_2) = T_2^{(K)}(v, V_1, V_2) = 0.$$

(ii) *Suppose $\beta_{vu} \neq 0$ for all $u \in \mathrm{pa}(v)$. If $\mathrm{de}(V_2 \cup \{v\}) \subseteq V_2 \cup \{v\}$ and $\mathrm{pa}(v) \cap V_2 \neq \emptyset$, then for generic error moments of order up to K, we have $T_1^{(K)}(v, V_1, V_2) > 0$ and $T_2^{(K)}(v, V_1, V_2) > 0$.*

Proof. (i) The statement follows immediately from Theorem 1. (ii) Since, $\mathrm{pa}(v) \cap V_2 \neq \emptyset$, but $\mathrm{de}(V_2 \cup \{v\}) \subseteq V_2 \cup \{v\}$, there exists some $u \in \mathrm{pa}(v) \cap V_2$ such that $\mathrm{de}(u) \cap \mathrm{pa}(v) = \emptyset$. For that u and any $C \subseteq V_1$, the residual total effect is $\pi_{vu.C} = \beta_{vu} - \sum_{c\in C} \beta_{vc.C}\pi_{cu} = \beta_{vu}$ because the assumed facts $\mathrm{de}(V_2 \cup \{v\}) \cap V_1 = \emptyset$ and $\mathrm{de}(v) \cap \mathrm{pa}(v) = \emptyset$ imply that $\pi_{cu} = 0$ for all $c \in C$ and $\pi_{vu} = \beta_{vu}$. We have assumed $\beta_{vu} \neq 0$, so, by Theorem 1, generic error moments ensure that $|\tau^{(K)}_{v.C\to u}| > 0$ for all C, which in turn implies $T_j^{(K)}(v, V_1, V_2) > 0$ for $j = 1, 2$. □

When (i) is satisfied, there may be more than one set C which makes all pairwise statistics 0. T_1 is calculated by finding a single conditioning set C which minimizes the maximum pairwise statistic τ across all $u \in V_2$; in contrast, T_2 allows for a different conditioning set for each u. For fixed v, V_1, and V_2, the signs, either positive or zero, of T_1 (min-max) and T_2 (max-min) will always agree, but when $\mathrm{pa}(v) \cap V_2 \neq \emptyset$ and both quantities are positive, $T_1 \geq T_2$. Thus, the min-max statistic may be more robust to sampling error when testing if the parameters are non-zero. However, as discussed in Section 3·1, T_2 can be computed more efficiently than T_1.

Theorem 1(ii) requires parental faithfulness since we consider arbitrary $u \in \mathrm{pa}(v)$, whereas Corollary 2(ii) only requires that $\beta_{vu} \neq 0$ since we maximize over V_2. Use of sample moments yields estimates $\hat{\tau}_{v.C\to u}$, which in turn yields estimates $\hat{T}_j^{(K)}(v, V_1, V_2)$ of $T_j^{(K)}(v, V_1, V_2)$ for $j = 1, 2$. In the remainder of the paper, we drop the subscript j in statements that apply to both parameters/estimators. Moreover, as we always fix K, we lighten notation by omitting the superscript, writing $T(v, V_1, V_2)$, $\tau_{v.C\to u}$ and $\hat{\tau}_{v.C\to u}$.

3. Graph estimation procedure

3·1. *Algorithm*

We now present a modified DirectLiNGAM algorithm which estimates the underlying causal structure (Algorithm 1). As in the original algorithm, we identify a root and recur on the subgraph that has the identified root removed. After step z, we have a z-tuple, $\Theta^{(z)}$, which gives an ordering of the roots identified so far, and the remaining nodes $\Psi^{(z)} = V \setminus \Theta^{(z)}$. In contrast to DirectLiNGAM, the proposed algorithm does not adjust for all non-descendants, but only for subsets of limited size. This gives meaningful regression residuals also when the number of variables exceeds the sample size and limits error propagation from the estimated linear coefficients.

At each step z, we consider subsets of $\mathcal{C}_v^{(z)} \subseteq \Theta^{(z-1)}$, which we use to denote the set of possible parents for v. Naively allowing $\mathcal{C}_v^{(z)} = \Theta^{(z-1)}$ is not precluded by theory, but the number of subsets $C \subset \Theta^{(z-1)}$ such that $|C| = J$ grows at $\mathcal{O}(z^J)$. Thus, for computational reasons, we prune nodes which are not parents of v by letting

$$\mathcal{C}_v^{(z)} = \left\{ p \in \mathcal{C}_v^{(z-1)} : \min_{C \in D_v^{(z)}} |\hat{\tau}_{v.C\to p}| > g^{(z)} \right\} \cup \Theta_{z-1}^{(z-1)} \tag{7}$$

where $D_v^{(z)} = \bigcup_{d<z} \{C : C \subseteq \mathcal{C}_v^{(d)} \setminus \{p\}; |C| \leq J\}$, $\Theta_{z-1}^{(z-1)}$ is the node selected at the previous step, and $g^{(z)}$ is some cut-off value. Selecting a good value for $g^{(z)}$ is difficult because it should depend on the unknown signal strength. However, under the assumptions of Theorem 2, if r is the root selected at step $z-1$ and α is some tuning parameter in $[0, 1]$, then letting $g^{(z)} = \max(g^{(z-1)}, \alpha \hat{T}(r, \mathcal{C}_r^{(z)}, \Psi^{(z-1)}))$ will not mistakenly prune parents from $\mathcal{C}_v^{(z)}$. In Algorithm 1, we do not update $\mathcal{C}_v^{(z)}$ after v is selected as a root. Since the final cut-off, $g^{(p)}$, may be larger than the cut-off used to select $\mathcal{C}_v^{(z)}$, a final pruning step uses the criteria from (7) with $g^{(p)}$ to prune away nodes in $\mathcal{C}_v^{(p)}$ that may be ancestors but not parents.

A larger value of α prunes more aggressively, decreasing the computational effort. However, setting α too large could result in incorrect estimates if some parent of v is incorrectly pruned from $\mathcal{C}_v^{(z)}$. Section 3·2 discusses selecting an appropriate α and a more detailed discussion of computational savings from the pruning procedure is given in the supplement.

As discussed in Section 2·3, T_1 may be more robust to sampling error than T_2 but comes at greater computational cost. At each step, $\Psi^{(z)}$ decreases by a single node and $\mathcal{C}_v^{(z)}$ may grow by

Algorithm 1. Estimate Causal DAG.

Set $\Theta^{(0)} = \emptyset$ and $\Psi^{(0)} = [p]$.
For $z = 1, \ldots, p$:
  For $v \in \Psi^{(z-1)}$:
    Select the set of possible parents $\mathcal{C}_v^{(z)} \subseteq \Theta^{(z-1)}$ and compute $\hat{T}(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)} \setminus \{v\})$.
  Let $r = \arg\min_{v \in \Psi^{(z-1)}} \hat{T}(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)} \setminus \{v\})$.
  Append r to $\Theta^{(z-1)}$ to form $\Theta^{(z)}$ and set $\Psi^{(z)} = \Psi^{(z-1)} \setminus \{r\}$.
Prune ancestors to form parents $\mathcal{C}_v^\star$ for all $v \in V$.
Return $\Theta^{(p)}$ as the topological ordering; $\{\mathcal{C}_v^\star\}_{v \in V}$ as the set of parents.

one node. If the $|\Psi^{(z)}|^2$ values of $\min_{C \in \mathcal{C}_v^{(z-1)}} \hat{\tau}_{v.C \to u}$ have been stored, updating $\hat{T}_2$, the max-min, only requires testing the $\binom{|\mathcal{C}^{(z-1)}|}{J-1}$ subsets of $\mathcal{C}_v^{(z)}$ which include the variable selected at the previous step. Updating the min-max statistic $\hat{T}_1$ without redundant computation would require storing the $\mathcal{O}\left((p-z)^2 z^J\right)$ values of $|\tau_{v.C \to u}|$. In practice, we completely recompute it at each step. Section 4 demonstrates this trade-off between computational burden and robustness.

3·2. *Deterministic statement*

Theorem 2 below makes a deterministic statement about sufficient conditions under which Algorithm 1 will output a specific graph G when given data $Y = (Y_1, \ldots, Y_n)$. We assume each $Y_i \sim P_Y$ but allow model misspecification so that P_Y may not be in $\mathcal{P}(G)$ for any DAG G. However, we require that the sample moments of Y are close enough to the population moments for some distribution $P \in \mathcal{P}(G)$. For notational convenience, for $H \subseteq V$ and $\alpha \in \mathbb{R}^{|H|}$, let $\hat{m}_{H,\alpha} = \frac{1}{n} \sum_i^n \left(\prod_{v \in H} Y_{vi}^{\alpha_v} \right)$ denote a sample moment estimated from data Y, and let $m_{H,\alpha} = E_P \left(\prod_{v \in H} Z_v^{\alpha_v} \right)$ denote a population moment for $Z \sim P$.

Condition 1. For some p-variate distribution P, there exists a DAG G with $|\mathrm{pa}(v)| \leq J$ for all $v \in V$ such that:

(a) For all $v, u \in V$ and $C \subseteq V \setminus \{u, v\}$ with $|C| \leq J$ and $C \cap \mathrm{de}(v) = \emptyset$; if $u \in \mathrm{pa}(v)$ then the population quantities for P satisfy $\left| \tau^{(K)}_{v.C \to u} \right| > \gamma > 0$.
(b) For all $v, u \in V$ and $C \subseteq V \setminus \{v, u\}$ with $|C| \leq J$ and $\mathrm{pa}(v) \subseteq C \subseteq V \setminus \mathrm{de}(v)$, if $u \notin \mathrm{pa}(v)$, then the population quantities for P satisfy $\tau^{(K)}_{v.C \to u} = 0$.

Condition 2. All $J \times J$ principal submatrices of the population covariance of P have minimum eigenvalue greater or equal to $\lambda_{\min} > 0$.

Condition 3. All population moments of P up to degree K, $m_{V,\alpha}$ for $\sum_v \alpha_v \leq K$, are bounded by a constant $\infty > M > \max(1, \lambda_{\min}/J)$ for positive integer J.

Condition 4. All sample moments of Y up to degree K, $\hat{m}_{V,\alpha}$ for $\sum_v \alpha_v \leq K$, are within $\delta_1 < \lambda_{\min}/(2J)$ of the corresponding population values of P.

The constraint in Condition 3 that $M > \max(1, \lambda_{\min}/J)$ is only used to facilitate simplification of the error bounds and is not otherwise necessary. Condition 1 is a faithfulness type assumption on P, and in Theorem 2 we make a further assumption on γ which ensures strong faithfulness. However, it is not strictly stronger or weaker than the Gaussian strong faithfulness type assumption. In particular we require the linear coefficients and error moments considered

to be jointly "sufficiently parentally faithful and non-Gaussian." So for a fixed sample size, there may be cases where the linear coefficients and error covariances do not satisfy Gaussian strong faithfulness, but do satisfy the non-Gaussian condition because the higher order moments are sufficiently non-Gaussian. However, the opposite may also occur where a set of linear coefficients and error moments satisfy Gaussian strong faithfulness but not the non-Gaussian condition.

Finally, let $\mathcal{P}_{F_K}(G)$ be the subset of distributions $P \in \mathcal{P}(G)$ with $\tau^{(K)}_{v.C\to u} \neq 0$ whenever $u \in \mathrm{pa}(v)$ and $C \subseteq V \setminus (\{u, v\} \cup \mathrm{de}(v))$. Then the set of linear coefficients and error moments that induce an element of $\mathcal{P}(G) \setminus \mathcal{P}_{F_K}(G)$ has measure zero. This set difference includes distributions which are not parentally faithful with respect to G and distributions for which there exist a parent/child pair for which both error distributions have Gaussian moments up to order K.

THEOREM 2. *For some p-variate distribution P and data $Y = (Y_1, \ldots, Y_n)$:*

(i) Suppose Condition 1 *holds. Then among all DAGs with maximum in-degree at most J, there exists a unique DAG G such that $P \in P_{F_K}(G)$*

(ii) Suppose Conditions 1-4 *hold for constants which satisfy*

$$\begin{aligned}\gamma/2 > \delta_3 := 4M\delta_1 &\left\{16(3^K)(J+K)^K K \frac{J^{(K+4)/2}M^{K+1}}{\lambda_{\min}^{K+1}}\right\} \\ &+ 2\left[\delta_1\left\{16(3^K)(J+K)^K K \frac{J^{(K+4)/2}M^{K+1}}{\lambda_{\min}^{K+1}}\right\}\right]^2.\end{aligned} \tag{8}$$

Then with pruning parameter $g = \gamma/2$, Algorithm 1 *will output $\hat{G} = G$.*

The main result of Theorem 2 is part (ii). The identifiability of a DAG was previously shown by Shimizu et al. (2006) by appealing to results for independent component analysis; however, our direct analysis of rational functions of Y allows for an explicit tolerance for how sample moments of Y may deviate from corresponding population moments of P. This implicitly allows for model misspecification; see Corollary 3. The proof of Theorem 2 requires Lemmas 1-3, which we develop first. The lemmas are proven in the supplement. Recall that β_{vC} are the population regression coefficients from (2), and let $\hat{\beta}_{vC}$ denote the coefficients estimated from Y.

LEMMA 1. *Suppose Conditions* 2, 3, *and* 4 *hold. Then for any $v \in V$, $C \subseteq V$, and $|C| \leq J$,*

$$\|\hat{\beta}_{vC} - \beta_{vC}\|_\infty < \delta_2 = 4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}.$$

Recall, that $Y_{vi.C} = Y_{vi} - \sum_{c\in C}\beta_{vc.C}Y_{ci}$. Let $Z_{v.C}$ denote the analogous quantity for $Z \sim P$, and let $\hat{Y}_{vi.C} = Y_{vi} - \sum_{c\in C}\hat{\beta}_{vc.C}Y_{ci}$.

LEMMA 2. *Suppose that Conditions* 2, 3, *and* 4 *hold. Let s, r be non-negative integers such that $s + r \leq K$, and let $Z \sim P$. For any $v, u \in V$ and $C \subseteq V \setminus \{u, v\}$ such that $|C| \leq J$,*

$$\left|\frac{1}{n}\sum_i \hat{Y}^s_{vi.C}Y^r_{ui} - E\left(Z^s_{v.C}Z^r_u\right)\right| < \delta_1\Phi(J, K, M, \lambda_{\min})$$

where

$$\Phi(J, K, M, \lambda_{\min}) = \left\{16(3^K)(J+K)^K K \frac{J^{(K+4)/2}M^{K+1}}{\lambda_{\min}^{K+1}}\right\}. \tag{9}$$

The proof of Lemma 2 relies on the fact that the map from moments of Z to the quantities of interest are Lipschitz continuous within a bounded domain.

LEMMA 3. *Suppose that Conditions* 2*,* 3*, and* 4 *hold. Then*

$$|\hat{\tau}_{v.C\to u} - \tau_{v.C\to u}| < 4M\delta_1\Phi(J,K,M,\lambda_{\min}) + 2\left\{\delta_1\Phi(J,K,M,\lambda_{\min})\right\}^2 = \delta_3$$

for the function $\Phi(J,K,M,\lambda_{\min})$ *given in Lemma* 2*.*

The proof of Lemma 3 is an application of the triangle inequality.

Proof of Theorem 2. (ii) We proceed by induction. By Lemma 3 and assuming (8), each statistic $\hat{\tau}_{v.C\to u}$ is within $\delta_3 < \gamma/2$ of the corresponding population quantity. Thus, any statistic corresponding to a parameter with value 0 is less than $\gamma/2$ and, by Condition 1 and the condition on γ in (8), all statistics corresponding to a non-zero parameter are greater than $\gamma/2$.

Recall that $\Theta^{(z)}$ is a topological ordering of nodes. Assume for some step z, that $\Theta^{(z-1)}$ is consistent with a valid ordering of G. Let $R^{(z)} = \{v \in \Psi^{(z-1)} : \mathrm{an}(r) \subseteq \Theta^{(z-1)}\}$ so that any $r \in R^{(z)}$ is a root in the subgraph induced by $\Psi^{(z-1)}$ and $\Theta^{(z)} = (\Theta^{(z-1)} \cup \{r\})$ is consistent with G. The base case for $z = 1$ is trivially satisfied since $\Theta^{(0)} = \emptyset$.

Setting $g = \gamma/2$ does not incorrectly prune any parents, so $\mathrm{pa}(r) = \mathcal{C}_r^{(z)}$, which implies $\hat{T}(r, \mathcal{C}_r^{(z)}, \Phi^{(z-1)}) < \gamma/2$ for all $r \in R^{(z)}$. Similarly, for any $v \in \Psi^{(z-1)} \setminus R^{(z)}$, there exists $u \in \Psi^{(z-1)}$ with $|\hat{\tau}_{v.C\to u}| > \gamma/2$ for all $C \subseteq \Theta^{(z-1)}$. Thus, $\hat{T}\big(r, \mathcal{C}_r^{(z)}, \Psi^{(z-1)}\big) < \hat{T}\big(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)}\big)$ for every $r \in R^{(z)}$ and $v \in \Psi^{(z-1)} \setminus R^{(z)}$. This implies the next root selected, $\arg\min_{v\in\Psi^{(z-1)}} \hat{T}\big(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)}\big)$ must be in $R^{(z)}$, and thus $\Theta^{(z)}$ remains consistent with G.

(i) The fact that $P \in \mathcal{P}_{F_K}(G)$ follows directly from the definition. To show uniqueness, we use population quantities so that $\delta_1 = 0$ which in turn implies $\delta_3 = 0$. Then for any $\gamma > 0$, Algorithm 1 will return G. Thus, by 2(ii), G must be unique. □

Remark 1. As stated Theorem 2 concerns an explicit cut-off g, whereas in practice we specify a tuning parameter α that is easier to interpret and tune. If $\alpha \leq 1$, it holds under the conditions of Theorem 2 that Algorithm 1 returns a topological ordering consistent with G, but $\hat{E}$ may be a superset of E. However, there exists $\alpha \geq 1$ which will recover the exact graph.

To see this note that $\alpha \leq 1$ ensures that $g^{(z)} < \gamma/2$ under the specified conditions, so no parents are pruned incorrectly and the estimated topological ordering is correct. This, however, may not remove all ancestors that are not parents, so the estimated edge set may be a superset of the true edge set. Letting instead

$$\alpha = \frac{\min_v \min_{a\in\mathrm{pa}(v)} \min_{C\cap\mathrm{de}(v)=\emptyset} |\hat{\tau}_{v.C\to a}|}{\max_v \max_{a\in\mathrm{an}(v)\setminus\mathrm{pa}(v)} \min_{C\cap\mathrm{de}(v)=\emptyset} |\hat{\tau}_{v.C\to a}|}, \tag{10}$$

will correctly prune ancestors and not parents. Because all sample moments are close to their population values, the denominator must be less than $\gamma/2$ and strong parental faithfulness further implies that the numerator is greater than $\gamma/2$ so (10) is greater than 1. However, setting α too large may result in an incorrect estimate of the ordering since a true parent may be errantly pruned. Thus, we advocate a more conservative approach of setting $\alpha \leq 1$ which is more robust to violations of strong faithfulness.

Remark 2. Suppose $P_Y \in \mathcal{P}(G)$ but is not necessarily parentally faithful with respect to G. If $\alpha = 0$ and $\beta_{vu} \neq 0$ for all $(u, v) \in E$, then for generic error moments a correct ordering will still be recovered consistently as $\delta_1 \to 0$.

Indeed, Corollary 2(ii) holds without parental faithfulness. So for generic error moments, there exists $\gamma > 0$ such that $T(v, \mathcal{C}_v^{(z-1)}, \Phi^{(z-1)}) > \gamma$ for all $v \in \Phi^{(z-1)} \setminus R^{(z)}$ for all steps z. However, without parental faithfulness, a parent node may be errantly pruned if $\alpha > 0$. To ensure Corollary 2(i) holds, we need $\mathrm{pa}(r) \subseteq \mathcal{C}_v^{(z)}$ for all $r \in R^{(z)}$, which is satisfied by letting $\mathcal{C}_r^{(z)} = \Theta^{(z-1)}$. For fixed γ, since $\delta_3 \to 0$ as $\delta_1 \to 0$, there exists a δ_1 so that $\gamma > 2\delta_3$.

3·3. *High-dimensional consistency*

We now consider a sequence of graphs, observations, and distributions indexed by the number of variables p. For notational brevity, we do not explicitly include the index p in the notation, and keep simply writing G, Y, P_Y and P for these sequences. The following corollary states conditions sufficient for the conditions of Theorem 2 to hold with probability tending to 1. We first make explicit assumptions on P_Y, with $m^\star_{V,\alpha}$ denoting the population moments of P_Y. Again, we allow for misspecification, but require control of the L_∞ distance between population moments of P_Y and some $P \in \mathcal{P}_{F_K}(G)$.

Condition 5. P_Y is a log-concave distribution.

Condition 6. All population moments of P_Y up to degree $2K$, $m^\star_{V,\alpha}$ for $\sum_v \alpha_v \leq 2K$, are bounded by $M - \xi > \max(1, \lambda_{\min}/J)$.

Condition 7. Each population moment of Y up to degree K, $m^\star_{V,\alpha}$ for $\sum_v \alpha_v \leq K$, is within ξ of the corresponding population moment of P.

When Y is actually generated from a recursive linear structural equation model, Condition 7 trivially holds with $\xi = 0$ and log-concave errors imply that Y is log-concave.

COROLLARY 3. *For a sequence of distributions P and data Y assume Conditions* 1*,* 2*,* 5*,* 6*, and* 7 *hold. For pruning parameter $g = \gamma/2$, Algorithm* 1 *will return the graph $\hat{G} = G$ with probability tending to* 1 *if*

$$\frac{\log(p)}{n^{1/(2K)}} \frac{J^{5/2} K^{5/2} M^2}{\gamma^{1/2} \lambda_{\min}^{3/2}} \to 0, \qquad \xi \frac{3^K K^{K+1} J^{(3K)/2+2} M^{K+2}}{\gamma \lambda_{\min}^{K+1}} \to 0 \tag{11}$$

when $p \to \infty$ and $\gamma, \lambda_{\min} < 1 < M$.

Proof. Conditions 6 and 7 imply Condition 3. It remains to be shown that Condition 4 and (8) hold for the γ specified in Condition 1. Solving the inequality in Lemma 3 for δ_1 shows (8) will be satisfied if the sample moments of Y are within δ of the population moments such that $\delta + \xi \leq \delta_1$ with δ_1 less than

$$\min \left[\frac{-8M\Phi + \left\{ (8M\Phi)^2 + 16\Phi^2\gamma \right\}^{1/2}}{8\Phi^2}, \frac{\lambda_{\min}}{2J}, M \right] = \min \left\{ \frac{\left(M^2 + \gamma/4\right)^{1/2} - M}{\Phi}, \frac{\lambda_{\min}}{2J} \right\}$$

for Φ defined in (9). Since $J, K, M > 1$, $\gamma, \lambda_{\min} < 1$ ensure that first term is the relevant term. We further simplify the expression since

$$\left(M^2 + \gamma/4\right)^{1/2} \geq M + \gamma \min_{t \in (0,\gamma)} \left. \frac{\partial \left(M^2 + \gamma/4\right)^{1/2}}{\partial \gamma} \right|_{\gamma = t} = M + \frac{\gamma}{8\left(M^2 + \gamma/4\right)^{1/2}}.$$

Thus, the conditions of Theorem 2 will be satisfied if

$$\delta + \xi \leq \frac{\gamma}{8\left(M^2 + \gamma/4\right)^{1/2} \Phi} =: \delta_4.$$

Specifically, we analyze the case when $\xi < \delta_4/2$ and $|\hat{m}_{V,a} - m_{V,a}| < \delta < \delta_4/2$ for all $|a| \leq K$. If Y_v follows a log-concave distribution, we can apply Lemma B.3 of Lin et al. (2016) which states for f, some K degree polynomial of log-concave random variables $Y = (Y_1, \ldots, Y_n)$, and some absolute constant, L, if

$$\frac{2}{L}\left(\frac{\delta}{(e)\left[\operatorname{var}\left\{f(Y)\right\}\right]^{1/2}}\right)^{1/K} \geq 2$$

then

$$\Pr\left[|f(Y) - E\left\{f(Y)\right\}| > \delta\right] \leq \exp\left\{\frac{-2}{L}\left(\frac{\delta}{\left[\operatorname{var}\left\{f(Y)\right\}\right]^{1/2}}\right)^{1/K}\right\}.$$

Letting $f(Y)$ be the sample moments of Y up to degree K, Condition 6 implies the variance is bounded by M/n. When $p > 2$, there are $\binom{p+K}{p} < p^K$ moments with degree at most K, then by a union bound, when $0 < \xi < \delta_4/2$,

$$\begin{aligned}\Pr\left(\hat{G} = G\right) &\geq 1 - \Pr\left(|\hat{m}_{V,a} - m_{V,a}| > \delta_4/2 \text{ for any } |a| \leq K\right) \\ &\geq 1 - p^K \exp\left[\frac{-2}{L}\left\{\frac{\delta_4/2}{(M/n)^{1/2}}\right\}^{1/K}\right]\end{aligned}$$

when

$$\frac{2n^{1/(2K)}}{L}\left(\frac{\delta_4/2}{eM^{1/2}}\right)^{1/K} \geq 2. \tag{12}$$

In the asymptotic regime, where p is increasing,

$$\frac{LM^{1/(2K)}K\log(p)}{(\delta_4/2)^{1/K}\, n^{1/(2K)}} \to 0$$

implies that the inequality in (12) will be satisfied and

$$p^K \exp\left[\frac{-2}{L}\left\{\frac{\delta_4/2}{(M/n)^{1/2}}\right\}^{1/K}\right] \to 0.$$

Plugging in the expression for δ_4, we find

$$\begin{aligned}\frac{LM^{1/(2K)}K\log(p)}{(\delta_4/2)^{1/K}\, 2n^{1/(2K)}} = \frac{LM^{1/(2K)}K\log(p)}{2n^{1/(2K)}} \times \\ \left\{\frac{16\left(M^2+\gamma/4\right)^{1/2} 16(3^K)(J+K)^K K J^{(K+4)/2} M^{K+1}}{\gamma\lambda_{\min}^{K+1}}\right\}^{1/K}.\end{aligned}$$

This quantity is of order $\mathcal{O}\big(\big(\log(p)J^{5/2}K^{5/2}M^2\big)/\big(n^{1/(2K)}\gamma^{1/2}\lambda_{\min}^{3/2}\big)\big)$ when assuming that $\gamma < M$. In addition, $\xi < \delta_4/2$ will be satisfied if $\frac{2\xi}{\delta_4} \to 0$. This ratio is

$$\frac{2\xi}{\delta_4} = 2\xi\left\{\frac{16\left(M^2+\gamma/4\right)^{1/2} 16(3^K)(J+K)^K K J^{(K+4)/2} M^{K+1}}{\gamma\lambda_{\min}^{K+1}}\right\}$$

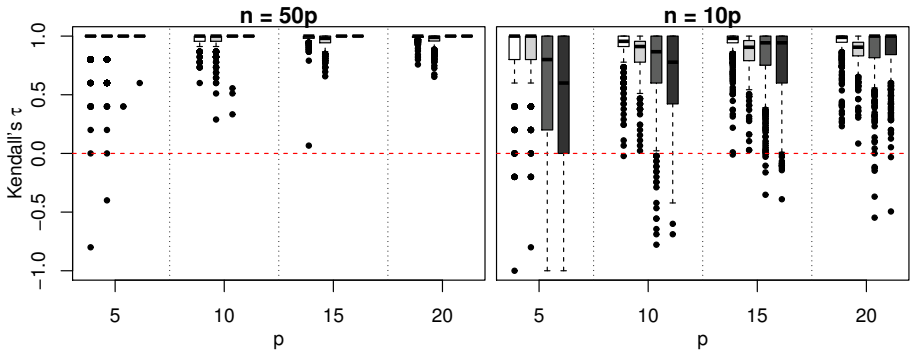


Fig. 3: Each bar represents the results from 500 randomly drawn graphs and data. In each group, from left to right, the bars represent (1) min-max $\hat{T}_1$, (2) max-min $\hat{T}_2$, (3) Shimizu et al. (2011), and (4) Hyvärinen and Smith (2013). In the left panel $n = 50p$ and the right panel $n = 10p$.

which is $\mathcal{O}\big(\left(\xi 3^K K^{K+1} J^{(3K)/2+2} M^{K+2}\right)/\left(\gamma \lambda_{\min}^{K+1}\right)\big)$ when $\gamma < M$. □

When fixing the other terms, Corollary 3 requires $\log(p) = o(n^{1/(2K)})$. Corollary 3 does not preclude J from growing with n and p; however, the computational complexity of Algorithm 1 is exponential in J, so in practice J must remain relatively small.

4. Numerical results

4·1. *Simulations: low dimensional performance*

We first compare the proposed method using: (1) min-max $\hat{T}_1$ and (2) max-min $\hat{T}_2$ against (3) DirectLiNGAM (Shimizu et al., 2011) and (4) Pairwise LiNGAM (Hyvärinen and Smith, 2013, Section 3.2). We randomly generate graphs and corresponding data with the following procedure. For each node v, select the number of parents d_v uniformly from $1, \ldots, \min(v, J)$. We include edge $(v-1, v)$ to ensure that the ordering is unique and draw $\beta_{v,v-1}$ uniformly from $(-1, -.5) \cup (.5, 1)$. The remaining parents are selected uniformly from $[v-2]$ and the corresponding edge weights are set to $\pm 1/5$. The n error terms for variable v are generated by selecting $\sigma_v \sim \text{unif}(.8, 1)$ and then drawing $\varepsilon_{vi} \sim \sigma_v \text{unif}(-\sqrt{3}, \sqrt{3})$.

We use $K = 4$, fix the max in-degree $J = 3$, let $p = 5, 10, 15, 20$, and let $n = 50p$ and $n = 10p$. We set $\alpha = .8$ and compare performance by measuring Kendall's τ between the returned ordering and the true ordering; i.e., the number of concordant pairs in the ordering minus the number of discordant pairs, normalized by the number of total pairs. The procedure is repeated 500 times for each setting of p and n.

Figure 3 shows that in the low-dimensional case with $n = 50p$, the Pairwise LiNGAM and DirectLiNGAM methods outperform the proposed method, with either statistic. However, already with $n = 10p$, our method begins to give improvements. The min-max statistic T_1 does slightly better than T_2, the max-min. However, Figure 4 shows a large difference in computational effort; $p = 40, 80$ are included for further contrast. In the sequel, we use the max-min statistic, T_2.

The proposed method compares favorably to the DirectLiNGAM method in computational effort because of the expensive kernel mutual information calculation and is comparable to the Pairwise LiNGAM. However, we refrain from a direct timing comparison because DirectLiNGAM and Pairwise LiNGAM are both implemented in Matlab while our proposed method is implemented in R and C++ (R Core Team, 2017; Eddelbuettel and François, 2011). In the supplement,

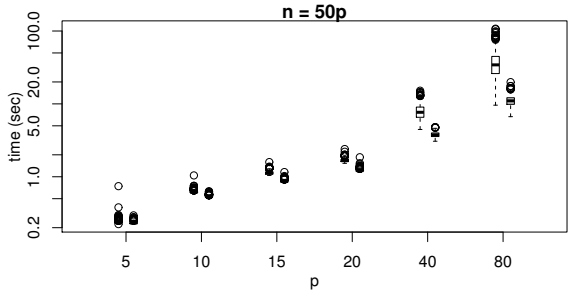


Fig. 4: Timing results from 500 randomly drawn graphs and data with $n = 50p$. In each pair, the left represents min-max, $\hat{T}_1$ and the right max-min, $\hat{T}_2$. The y-axis is on a log scale.

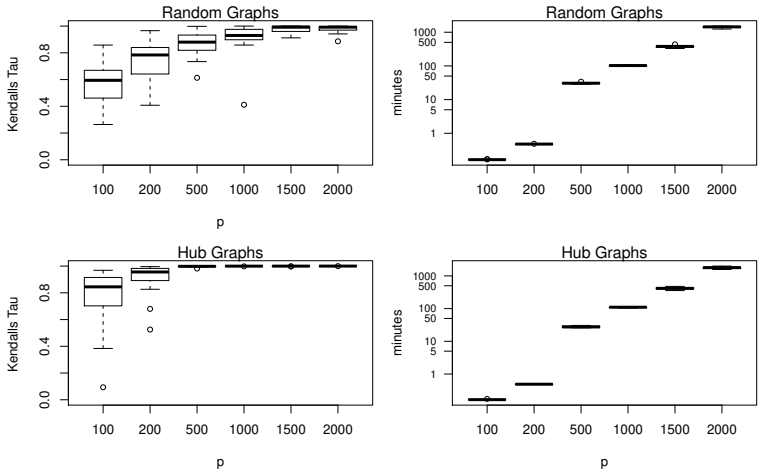


Fig. 5: Each boxplot represents the results of 20 simulations. In all cases, we let $n = 3/4p$. The top panels show results from randomly drawn DAGs while the bottom panel shows results from DAGs constructed to have hub structure. The left plots show performance as measured by Kendall's τ and the right plots show computational time when using 16 CPUs in parallel.

we also provide a direct comparison between the proposed statistic and those used by Shimizu et al. (2011) and Hyvärinen and Smith (2013).

4·2. *Simulations: high-dimensional consistency*

To illustrate high-dimensional consistency, we generate the graph and coefficients as in Section 4·1 but with $n = 3/4p$ for $p = 100, 200, 500, 1000, 1500, 2000$. We first consider random DAGs and data generated as before, but with $J = 2$. We also consider graphs with hubs, that is, nodes with large out-degree. These are generated by including a directed edge from $v - 1$ to v for all nodes $v = 2, \ldots, p$ and drawing the edge weight uniformly from $(-1, -.65) \cup (.65, 1)$. We then set nodes $\{1, 2, 3\}$ as hubs and include an edge with weight $\pm 1/5$ to each non-hub node from a randomly selected hub. Thus, the out-degree for each of the hub nodes grows linearly with p, but the maximum in-degree remains bounded by 2. For both cases, the results for 20 runs at each value of p are shown in Figure 5. In the supplement, we show simulations with gamma errors and also consider a setting with Gaussian errors, where our method should not be consistent.

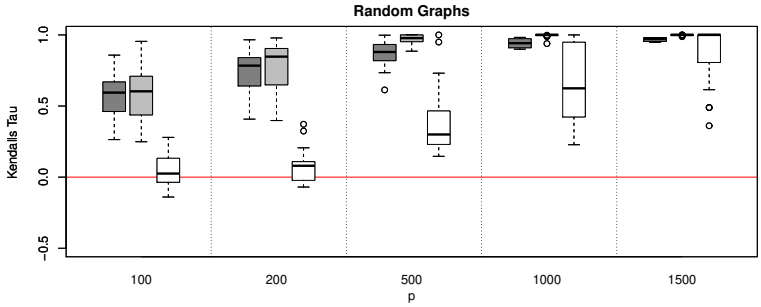


Fig. 6: Each boxplot represents 20 simulations with random DAGs when using a pre-selection step; in each case $n = 3/4p$. From left to right the methods are: the proposed high-dimensional LiNGAM procedure, same as Figure 5; the proposed high-dimensional LiNGAM procedure with pre-selection; the two stage pairwise procedure from Hyvärinen and Smith (2013).

4·3. *Pre-selection of neighborhoods*

As with the original DirectLiNGAM procedure, any edges or non-edges known in advance can be accounted for. Such information could, for instance, be obtained by applying neighborhood selection (Meinshausen and Bühlmann, 2006) to estimate the Markov blanket of each node. This blanket consists of parents, children, and parents of children. For sparse graphs, Hyvärinen and Smith (2013, Section 3.3) propose first using such a pre-selection step, then directly estimating the direction of each edge using pairwise measures without any additional adjustment. To create a total ordering, Alg B and Alg C of Shimizu et al. (2006) can be used. This does not require specifying a maximum in-degree, but in general, the neighborhood selection procedure will only be consistent if the total degree is controlled.

In our proposed procedure, we may incorporate estimated Markov blankets by limiting, at each step z, for each remaining node v, the set of potential parents, $C_v^{(z)}$, to the intersection of the estimated Markov blanket of v and the previously ordered nodes, $\Theta^{(z-1)}$. We do not otherwise prune the set of potential parents. Figure 6 shows results from using the pre-selection step under the setting from Section 4·2 for general random graphs. The pre-selection procedure improves the performance of our proposed high-dimensional LiNGAM procedure, but the proposed procedure without pre-selection still outperforms the two-stage procedure of Hyvärinen and Smith (2013, Section 3.3). Similar results for the hub graph setting are shown in the supplement.

4·4. *Data example: high-dimensional performance*

We estimate causal structure among the stocks in the Standard and Poor's 500. Specifically, we consider the percentage increase/decrease for each share price for each trading day between Jan 2007 to Sep 2017. We consider the $p = 442$ companies for which data is available for the entire period, and we scale and center the data so that each variable has mean 0 and variance 1. As structure may vary over time, we estimate the causal structure for each of the following periods separately with $J = 3$ and $K = 4$: 2007-2009, 2010-2011, 2012-2013, 2014-2015, 2016-2017 (ending in September). Across these periods, the sample size, n, ranges from 425 to 755.

The underlying structure is unlikely to be causally sufficient or acyclic. In addition, although it is common to assume that daily returns are independent, this assumption may not hold in practice. Nonetheless, the method still recovers reasonable structure. We first consider the most recent Jan 2016 - Sep 2017 period. Figure 7 shows a boxplot for the estimated ordering of the companies within each sector. The sectors are sorted top to bottom by median ordering. Near

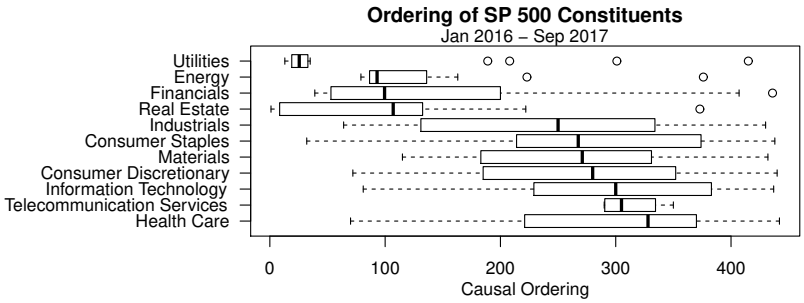


Fig. 7: Estimated causal ordering of the stocks in the Standard and Poor's 500 for Jan 2016 - Sep 2017. The stocks are grouped by sector, and the sectors are arranged by median causal ordering.

the top, we see utilities, energy, real estate, and finance. Since energy is an input for almost every other sector, intuitively price movements in energy should be causally upstream of other sectors. The estimated ordering of utilities might seem surprising; however, utility stocks are typically thought of as a proxy for bond prices. Thus, the estimated ordering may reflect the fact that changes in utility stocks capture much of the causal effect of interest rates, which had stayed constant for much of 2011-2015 but began moving again in 2016. Real estate and finance, sectors that are highly impacted by interest rates, are also estimated to be early in the causal ordering.

Figure 8 ranks each sector by the median topological ordering for each period. The orderings are relatively stable over time, but there are a few notable changes. In 2007, real estate was estimated to be the "root sector" while finance is in the middle. This aligns with the idea that the root of the 2008 financial crisis was actually failing mortgage backed securities in real estate, which had a causal effect on finance. However, over time, real estate has moved more downstream.

5. DISCUSSION

We proposed a causal discovery method that was proven consistent for specific test statistics and log concave errors. Similar analyses could be given for other statistics that are Lipschitz continuous in the sample moments over a bounded domain, can distinguish causal direction, and indicate the presence of confounding. This would include a normalized version of the proposed test statistics which accounts for the scaling of the data. Log-concavity was assumed for exponential concentration of sample moments and other distributional assumptions could be considered instead if analogous concentration results can be obtained and traced throughout the analysis.

The proposed algorithm requires selecting a bound on the in-degree J and a pruning parameter α. The in-degree is typically unknown, but a reasonable upper bound may be used as a "bet on sparsity". If the maximum in-degree of the true graph is larger than the specified J but the "extra edges" have small enough edge-weights, the "closest" DAG with maximum in-degree J is still recovered with high probability. The pruning parameter α plays a similar role to the nominal level for each conditional independence test in the PC algorithm. Both parameters have an effect on the sparsity of the estimated graph and regulate the maximum size of conditioning sets.

At each step, instead of taking the minimum $|\tau|$ over all subsets of potential parents, one could also pick parents for every unordered node using a variable selection procedure and then only calculate $|\tau|$ using the selected parents. Such a procedure would also consistently estimate the causal ordering as long as the variable selection procedure is consistent. Slightly different conditions, such as a beta-min condition, would be needed when adopting standard methods based on

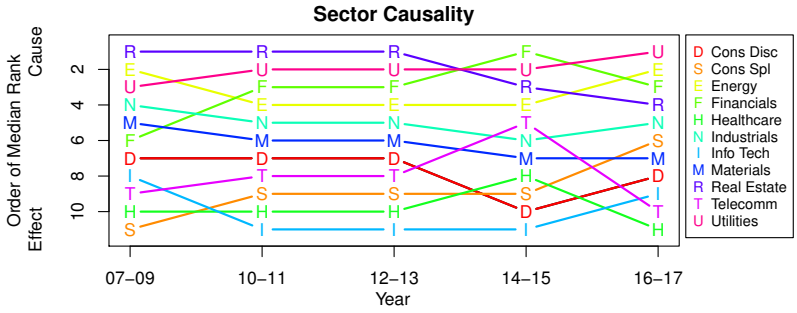


Fig. 8: Sectors ranked by median estimated topological ordering across each time period.

least squares, but in practice the resulting method performs quite well as shown in simulations in the supplement. This could be explained as being due to use of only second moments for the variable selection.

In (8), we have made a key restriction that the error moments must be adequately different from the moments of any Gaussian and the edge weights must be strongly parentally faithful. In practice, this is a difficult condition to satisfy, and Uhler et al. (2013) show that strong faithfulness type restrictions can be problematic in practice. However, even if the distribution is not strongly parentally faithful, we can still consistently recover the correct ordering as long as each individual linear coefficient is non-zero and the errors are sufficiently non-Gaussian. Sokol et al. (2014) consider identifiability of independent component analysis for fixed p when the error terms are Gaussians contaminated with non-Gaussian noise. In particular, when the effect of the non-Gaussian contamination decreases at an adequately slow rate, the entire mixing matrix is identifiable asymptotically. In our analysis, the measure of non-Gaussianity is treated by our assumptions on γ. Our results suggest that the results of Sokol et al. (2014) can also be extended, given suitable sparsity, to the asymptotic regime where the number of variables is increasing.

The modified procedure we propose retains the existing benefits of the original DirectLiNGAM procedure. In particular, the output of algorithm is independent of the ordering of the variables in the input data. Although this is typically not an issue in the low-dimensional case, in the high-dimensional setting, the output of causal discovery methods may be highly dependent on ordering (Colombo and Maathuis, 2014).

ACKNOWLEDGMENT

This work was supported by the U.S. National Science Foundation under Grant No. DMS 1712535. Thomas S. Richardson gave helpful feedback on an advance copy of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Theorem 1 and Lemmas 1,2, and 3. Additional simulations are also included.

REFERENCES

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741–3782.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Hao, D., Ren, C., and Li, C. (2012). Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC Systems Biology*, 6(1):34.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14:3365–3383.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, pages 282–289.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.*, 14:111–152.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res.*, 15:3065–3105.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, 1:763–765.

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

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

R Core Team (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rothenhäusler, D., Ernest, J., and Bühlmann, P. (2018). Causal inference in partially linear structural equation models. *Ann. Statist.*, 46(6A):2904–2938.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

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

Sokol, A., Maathuis, M. H., and Falkeborg, B. (2014). Quantifying identifiability in independent component analysis. *Electron. J. Stat.*, 8(1):1438–1459.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

Steinsky, B. (2013). Enumeration of labelled essential graphs. *Ars Combin.*, 111:485–494.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463.

[*Received* 30 Mar 2018. *Editorial decision on* 30 Sep 2018]