Journal of Machine Learning Research 24 (2023) 1-61 Submitted 11/21; Revised 2/23; Published 8/23

Causal Discovery with Unobserved Confounding and Non-Gaussian Data

Y. Samuel Wang ysw7@cornell.edu
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853, USA

Mathias Drton mathias.drton@tum.de
Department of Mathematics & Munich Data Science Institute
Technical University of Munich
85748 Garching bei München, Germany

Editor: Joris Mooij

Abstract

We consider recovering causal structure from multivariate observational data. We assume the data arise from a linear structural equation model (SEM) in which the idiosyncratic errors are allowed to be dependent in order to capture possible latent confounding. Each SEM can be represented by a graph where vertices represent observed variables, directed edges represent direct causal effects, and bidirected edges represent dependence among error terms. Specifically, we assume that the true model corresponds to a bow-free acyclic path diagram; i.e., a graph that has at most one edge between any pair of nodes and is acyclic in the directed part. We show that when the errors are non-Gaussian, the exact causal structure encoded by such a graph, and not merely an equivalence class, can be recovered from observational data. The method we propose for this purpose uses estimates of suitable moments, but, in contrast to previous results, does not require specifying the number of latent variables a priori. We also characterize the output of our procedure when the assumptions are violated and the true graph is acyclic, but not bow-free. We illustrate the effectiveness of our procedure in simulations and an application to an ecology data set.

Keywords: Causal discovery, Graphical model, Latent variables, Non-Gaussian data, Structural equation model

1. Introduction

We consider the problem of discovering causal structure from multivariate data when only observational data is available, but latent confounding may exist between the observed variables. In our main result, we show that if the data are generated under a recursive linear structural equation model with non-Gaussian idiosyncratic errors, the exact causal structure can be recovered provided the confounding is limited to pairs of variables which do not have a direct effect on each other. These models correspond to *bow-free acyclic path diagrams* (BAPs).

1.1 Linear structural equation models and graphs

Structural equation models (SEMs) are multivariate statistical models that encode causal relationships and are popular in the social and biological sciences (Bollen, 1989; Shipley, 2016). SEMs may be formulated to explicitly include latent unobserved variables, but in this article we consider a setup in which the latent variables have been marginalized out and the models only explicitly refer to effects of the observed variables. This approach has, in particular, been fruitful for causal discovery (Evans, 2019).

A linear SEM assumes that we observe a sample comprised of i.i.d. copies of a random vector $Y = (Y_v : v \in V)$ that solves the equation system

$$Y_v = \sum_{u \neq v} \beta_{v,u} Y_u + \varepsilon_v, \quad v \in V. \tag{1}$$

The direct effect of Y_u on Y_v is encoded by $\beta_{v,u}$, and ε_v is an idiosyncratic error term of mean zero. Note that our setup assumes throughout, and without loss of generality, that Y is *centered.* We collect the effects $\beta_{v,u}$ into a $p \times p$ matrix $B = (\beta_{v,u})_{u,v \in V}$ and the error terms into a vector $\varepsilon = (\varepsilon_v)_{v \in V}$. Because the matrix B encodes the direct causal effect of Y_u onto Y_v for all $u, v \in V$, we will use the term *direct effects* to refer to the matrix B. Each copy of the error vector ε is drawn i.i.d. with expectation 0, but we allow for unobserved confounding between different variables, say Y_v and Y_u, by allowing the corresponding errors, ε_v and ε_u, to be dependant. In vector form (1) reads $Y = BY + \varepsilon$, which is uniquely solved by

$$Y = (I - B)^{-1}\varepsilon \tag{2}$$

when $I - B$ is invertible. Letting $\Omega := \mathbb{E}(\varepsilon\varepsilon^T) = (\omega_{v,u})_{u,v \in V}$ be the covariance matrix of ε, we obtain that the covariance matrix of the observed variables in Y is

$$\Sigma := \mathbb{E}(YY^T) = (I - B)^{-1}\Omega(I - B)^{-T}. \tag{3}$$

Throughout the article, we describe SEMs using the language of graphical models or path diagrams (Maathuis et al., 2019). We represent each SEM by a mixed graph $G = (V, E_\rightarrow, E_\leftrightarrow)$, where each vertex $v \in V$ corresponds to an observed variable, and $E_\rightarrow$ and $E_\leftrightarrow$ are sets of directed edges and bidirected edges, respectively. Let $u, v \in V$ be two distinct vertices. We represent a direct effect of u on v by the directed edge $u \rightarrow v \in E_\rightarrow$ and say that u is a *parent* of *child* v. So, $\beta_{v,u} \neq 0$ only if u is a parent of v. If there exists a sequence of directed edges from u to v, we say that u is an *ancestor* of its *descendant* v. Unobserved confounding between v and u is represented by $v \leftrightarrow u \in E_\leftrightarrow$, and we say that v and u are *siblings.* So, $\omega_{v,u} \neq 0$ only if u and v are siblings. The sibling relation is symmetric; i.e., u being a sibling of v implies that v is a sibling of u. We denote the sets of parents, children, ancestors, descendants, and siblings of v as $\mathrm{pa}(v)$, $\mathrm{ch}(v)$, $\mathrm{an}(v)$, $\mathrm{de}(v)$, and $\mathrm{sib}(v)$, respectively. We let $\mathrm{An}(v) := \mathrm{an}(v) \cup \{v\}$. The problem of interest is then to infer the graph corresponding to a given data-generating SEM.

We assume that the true model is a recursive SEM. In graphical terms, this means that the mixed graph $G = (V, E_\rightarrow, E_\leftrightarrow)$ corresponding to the model does is *acyclic*; i.e., there are no directed cycles such that one can follow a directed path which begins and ends at the same vertex. There then exists a (not necessarily unique) total ordering of V under

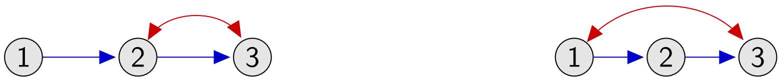


Figure 1: The graph on the left has a bow between 2 and 3 so it is not a BAP. The graph on the right has a bidirected edge between 1 and 3, but since there is no directed edge between 1 and 3, this does not constitute a bow, and thus the graph is a BAP.

which $u \prec v$ implies $v \notin \text{an}(u)$. If, in addition, the graph G contains no bidirected edges, i.e., $E_{\leftrightarrow} = \emptyset$, then G is a *directed acyclic graph* (DAG). A *bow* in G is a subgraph of two vertices u and v that contains both a directed and a bidirected edge; i.e., $u \leftrightarrow v$ and either $u \to v$ or $v \to u$. In this article, we primarily consider mixed graphs that are acyclic and do not contain bows, though in Section 5 we also consider the case where the true graph may contain bows. Following Drton et al. (2009), we refer to these graphs as *bow-free acyclic path* diagrams (BAPs). Said explicitly, a mixed graph is a BAP if (1) it is acylic and (2) for any u, v such that $u \in \text{sib}(v)$, $u \to v \notin E$ and $v \to u \notin E$; Figure 1 gives a simple example of graphs with and without bows. Bow-free structure can arise in particular through conditional randomization of treatments; compare Figure 1 in Drton et al. (2009). The class of BAPs was also considered, e.g., by Brito and Pearl (2002) and Nowzohour et al. (2017).

1.2 Previous work

Most work on causal discovery with latent variables focuses on recovering causal structure in the form of an *ancestral graph*. For settings without selection effects, as considered here, ancestral graphs are special cases of BAPs that satisfy the additional restriction that $\text{an}(v) \cap \text{sib}(v) = \emptyset$ for all nodes v. By adding bidirected edges to $E_{\leftrightarrow}$, every ancestral graph G can be transformed into a *maximal ancestral graph* (MAG) while preserving the conditional independence relations in G. Gaussian MAG models can then be entirely characterized by conditional independence (Richardson and Spirtes, 2002). However, for any MAG there are generally other MAGs that are Markov equivalent, i.e., encode the same set of conditional independence relations. Markov equivalent MAGs have the same adjacencies but the edges may be of different orientations or types. The Markov equivalence class of any MAG can be compactly represented by a *partial ancestral graph* (PAG) (Ali et al., 2009).

Spirtes et al. (2000) propose the Fast Causal Inference algorithm (FCI) to estimate the PAG corresponding to the underlying causal graph. Zhang (2008) added additional orientation rules such that the output of FCI is complete. Colombo et al. (2012), Claassen et al. (2013), and Chen et al. (2021) develop additional variants—RFCI, FCI+, and lFCI, respectively—which only require a polynomial number of conditional independence tests if the degree of the graph is bounded, or in the last case exploit possible local separation properties. Triantafillou and Tsamardinos (2016) select a MAG via a greedy search which maximizes a penalized Gaussian likelihood, and Bernstein et al. (2020) propose a greedy search over partial orderings of the variables. Efforts are also underway to refine the picture provided by conditional independence by considering additional non-parametric constraints imposed by SEM (Verma and Pearl, 1990; Shpitser et al., 2014; Evans, 2016). In a different vein, Nowzohour et al. (2017) propose a greedy search which assumes that the true model is a linear SEM with Gaussian errors which corresponds to a BAP.

The previously mentioned methods have enjoyed great success but operate in a regime in which only an equivalence class of graphs (e.g., via the PAG) can be discovered and different graphs in the equivalence class may have conflicting causal interpretations. In contrast, Shimizu et al. (2006) show that when the true model is a recursive linear SEM with *non-Gaussian* errors, the exact graph—not just an equivalence class—can be identified from observational data using independent component analysis (ICA). Instead of ICA, the subsequent DirectLiNGAM (Shimizu et al., 2011) and Pairwise LiNGAM (Hyvärinen and Smith, 2013) methods use an iterative procedure to estimate a causal ordering. Wang and Drton (2020) give a modified method that is also consistent in high-dimensional settings in which the number of variables p exceeds the sample size n, and Tramontano et al. (2022) consider high-dimensional polytree models. However, all of the above methods for the linear non-Gaussian acyclic model (LiNGAM) do not allow for possible latent confounding.

Hoyer et al. (2008) consider the setting where the data is generated by a LiNGAM model, but some variables are unobserved. Using existing results from overcomplete ICA, they show that the canonical DAG—roughly a DAG in which all unobserved variables have no parents and at least two children—can be identified when all parent-child pairs in the observed set are unconfounded. However, the result critically requires the number of latent variables in the canonical model to be known in advance and requires all unobserved confounding to be linear. For example, suppose $v \in \operatorname{sib}(u)$, and the confounding is caused by a hidden variable Y_h. Then a generative procedure where $\check{\varepsilon}_v \perp\!\!\!\perp \check{\varepsilon}_u$, $\varepsilon_v = \check{\varepsilon}_v + \alpha_v Y_h$, and $\varepsilon_u = \check{\varepsilon}_u + \alpha_u Y_h$ would be allowed; however, $\varepsilon_u = \check{\varepsilon}_u + \alpha_u Y_h^2$ would be precluded. Furthermore, even when the model is correctly specified, Shimizu and Bollen (2014) state "current versions of the overcomplete ICA algorithms are not very computationally reliable since they often suffer from local optima," and indeed Hoyer et al. (2008) use a maximum likelihood procedure with mixtures of Gaussians instead of overcomplete ICA in their simulations.

To avoid using overcomplete ICA and improve practical performance, Entner and Hoyer (2010) and Tashiro et al. (2014) both propose procedures which test subsets of the observed variables and seek to identify as many pairwise ancestral relationships as possible; i.e., either (1) $u \in \operatorname{an}(v)$, (2) $v \in \operatorname{an}(u)$, or (3) $v \notin \operatorname{an}(u)$ and $u \notin \operatorname{an}(v)$. Entner and Hoyer (2010) apply ICA to all subsets of the observed variables which do not have latent confounding. Tashiro et al. (2014) apply an iterative procedure similar to DirectLiNGAM to each subset of variables. They show that the procedure used for certifying ancestral relationships is sound in the presence of confounding, but do not characterize the class of graphs which can be identified. In the appendix, we show a simple ancestral graph that cannot be discovered using the method of Entner and Hoyer (2010). For ParcelLiNGAM, we show in Section 2 that all ancestral relationships can indeed be discovered when the true causal graph itself is ancestral, but that the method will not identify all ancestral relationships for any non-ancestral BAP.

The identifiability results of Section 3 and 4 are found in Chapter 4 of Wang (2018), the Ph.D. dissertation of the first author. Maeda and Shimizu (2020) propose Repetitive Causal Discovery (RCD) for discovering mixed graphs. RCD uses a causal functional model-based algorithm, and similar to our approach—but in constrast to Tashiro et al. (2014)—RCD iteratively uses previously discovered structure to inform later steps. However, in

Section 2 we show that RCD is not able to identify all BAPs. Similar to Hoyer et al. (2008), Salehkaleybar et al. (2020) use overcomplete ICA and thus crucially require all confounding to be linear. They extend the results of Hoyer et al. (2008) by showing that under weak conditions, the total number of variables (unobserved and observed) in the system can be identified and a causal ordering can be identified from population quantities. However, in order to identify causal effects (i.e., determine the graph beyond just ancestral relations), they require a condition which precludes many BAPs (Salehkaleybar et al., 2020, Assumption 2).

1.3 Contribution

As our main contribution, we show that when the data are generated by a linear non-Gaussian SEM that corresponds to a BAP, then the exact BAP—not just an equivalence class—can be consistently recovered. This implies that the causal effects can also be identified. Specifically, we show how to recover the BAP from higher-order moments, avoiding the use of overcomplete ICA in contrast to Hoyer et al. (2008) and Salehkaleybar et al. (2020). Thus, it does not require linearity in how the observed variables depend on the unobserved variables. Our result also does not require knowledge of the number of latent variables or knowledge about the distribution of the errors. It does, however, rely on a genericity assumption for the linear coefficients and error moments that, in particular, rules out Gaussian behavior of the considered moments.

The **B**ow-free **A**cyclic **n**on-**G**aussian (BANG) method we propose for recovery of BAPs uses a series of independence tests between (suitably estimated) regressors and residuals to certify causal structure. When the maximum in-degree (both directed and bidirected edges) is bounded, the total number of tests performed is bounded by a polynomial in the number of variables considered. We also characterize what the BANG procedure will return—given population values—when the model is misspecified and the true generating procedure corresponds to a graph with bows. In simulations, we confirm that the method reliably discovers exact causal structure when given a large enough sample and outperforms existing methods in some settings.

As a secondary contribution, we also show that the previously proposed ParcelLiNGAM (Tashiro et al., 2014) is indeed sound and complete for ancestral graphs. We also show that—in contrast to BANG—ParcelLiNGAM and RCD (Maeda and Shimizu, 2020) are not able to identify every BAP.

1.4 Preliminaries

Throughout, we often let a node $v \in V$ stand in for the variable Y_v. We use superscripts when referring to i.i.d. copies of random variables; i.e., $Y^{(i)}$ refers to the ith copy of random variable Y. For a set $C \subset V$, we let $Y_C = (Y_c : c \in C)$ be the vector of variables indexed by an element of C. Furthermore, for a matrix B and index sets R and C, let $B_{R,C}$ be the submatrix of B corresponding to the Rth rows and Cth columns. For some positive integer z, we also let $[z]$ indicate the set $\{1, \dots, z\}$. When applying a function to a set of arguments, we mean the union of the values obtained by applying the function to each element; e.g., when C is a set of nodes, $\text{pa}(C) = \bigcup_{c \in C} \text{pa}(c)$. In addition, in a slight abuse of notation, we will at times use the notation for a path to refer to the set of nodes on that path; i.e., $\ell = v_1 \to v_2 \to v_3$ may also refer to the set $\{v_1, v_2, v_3\}$.

The notions of *sound* (i.e., correct) and *complete* (i.e., maximally informative) are often used to describe the output of causal discovery methods based on conditional independence tests (Spirtes et al., 2000). We adapt these notions for discovery of ancestral relationships in a mixed graph G. Let $\prec_0$ be a (potentially partial) ordering of V. We say that the ordering $\prec_0$ is *sound* with respect to ancestral relationships in G if $u \prec_0 v$ implies that $v \notin \text{an}(u)$ holds. We say that the ordering $\prec_0$ is *complete* with respect to ancestral relationships in G if $u \in \text{an}(v)$ implies $u \prec_0 v$.

2. Ancestral graphs

Before discussing the main results, we first build intuition for causal discovery with non-Gaussian data by considering the simpler setting of ancestral graphs. We show that given population information, the previously proposed ParcelLiNGAM[1] procedure (Tashiro et al., 2014) is sound and complete for ancestral relationships when the true graph is ancestral. However, it is not complete for certifying ancestral relationships in non-ancestral BAPs. We also give an example of a BAP which the RCD method (Maeda and Shimizu, 2020) can not identify.

2.1 Determining causal relationships in ancestral graphs

Recall from (2) that $Y = (I - B)^{-1}\varepsilon$ so that Y_v is a linear combination of ε_s for all s such that $\left[(I-B)^{-1}\right]_{v,s} \neq 0$. For generic linear coefficients, this set is equal to $\text{an}(v)$. Thus, for $c \notin \text{sib}(v) \cup \text{de}(\text{sib}(v)) \cup \text{de}(v)$, the variable Y_c is a linear combination of error terms which are independent of ε_v, i.e., $Y_c \perp\!\!\!\perp \varepsilon_v$. Thus, for $v \in V$ and a set $C \subseteq V \setminus \{v\}$ such that $\text{pa}(v) \subseteq C \subseteq V \setminus [\text{sib}(v) \cup \text{de}(\text{sib}(v)) \cup \text{de}(v)]$, the population regression coefficients for predicting v from C are

$$\begin{aligned} D_{v,C}^T &= \left[\mathbb{E}\left(Y_C Y_C^T\right)\right]^{-1} \mathbb{E}\left(Y_C Y_v\right) \\ &= \left[\mathbb{E}\left(Y_C Y_C^T\right)\right]^{-1} \mathbb{E}\left(Y_C (Y_C^T B_{v,C}^T + \varepsilon_v)\right) \\ &= \left[\mathbb{E}\left(Y_C Y_C^T\right)\right]^{-1} \left[\mathbb{E}\left(Y_C Y_C^T\right) B_{v,C}^T + \mathbb{E}\left(Y_C \varepsilon_v\right)\right] = B_{v,C}^T, \end{aligned} \tag{4}$$

where $B_{v,C} = (\beta_{v,u})_{u \in C}$ is comprised of the direct effects of C onto v. The last equality in (4) crucially requires that $\mathbb{E}\left(Y_C \varepsilon_v\right) = 0$, as implied by the independences pointed out above. The regression residual $\eta_{v.C}$ obtained from the coefficients in $D_{v,C}$ then satisfies

$$\eta_{v.C} := Y_v - D_{v,C} Y_C = Y_v - B_{v,\text{pa}(v)} Y_{\text{pa}(v)} = \varepsilon_v. \tag{5}$$

As noted, $\eta_{v.C}$ is independent of the regressors Y_C.

In contrast, if C contains a descendant of v, a sibling of v, or a descendant of a sibling of v, then in general $\mathbb{E}\left(Y_C \varepsilon_v\right) \neq 0$, $D_{v,C} \neq B_{v,C}$, and $\eta_{v.C} \neq \varepsilon_v$. It follows, in general, that there exists some $c \in C$ such that $\eta_{v.C} \not\perp\!\!\!\perp Y_c$. Although the first order conditions

[1] Tashiro et al. (2014) give two variants of the ParcelLiNGAM algorithm which they label Algorithm 2 and 3. Algorithm 3 requires less computation and applies Algorithm 2 to a subset of the variables.

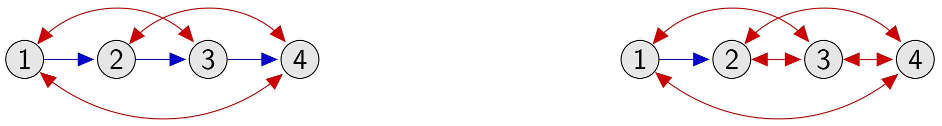
of the least squares criterion ensure that regressors and residuals are uncorrelated, when the errors are non-Gaussian, dependence can still be detected by using a non-parametric independence test (Gretton et al., 2005; Székely and Rizzo, 2009; Bergsma and Dassios, 2014; Pfister et al., 2018) or examining the higher order moments—e.g., $\mathbb{E}(Y_c^k \varepsilon_v)$ for $k > 1$ (Wang and Drton, 2020). Non-Gaussian errors are crucial because for a Gaussian random variable, uncorrelated and independent are equivalent so the residuals are independent of the regressors regardless of C. But when the errors are non-Gaussian, the independence of residuals and regressors can be used to certify that $C \subseteq V \setminus [\mathrm{sib}(v) \cup \mathrm{de}(\mathrm{sib}(v)) \cup \mathrm{de}(v)]$.

This idea can be directly applied to discover a topological ordering of the variables by finding the largest set $C_v^{(\max)}$ such that the residual when regressing v onto $C_v^{(\max)}$ is independent of $Y_{C_v^{(\max)}}$. When G is ancestral then $\mathrm{an}(v) \subseteq C_v^{(\max)} = V \setminus [\mathrm{sib}(v) \cup \mathrm{de}(\mathrm{sib}(v)) \cup \mathrm{de}(v)]$. Thus, to form $\hat{\prec}$, an initial estimate of a topological ordering, we can set $c \hat{\prec} v$ for all $c \in C_v^{(\max)}$. When there is no unique total ordering, there may be pairs u, v such that $u \in C_v^{(\max)} \setminus \mathrm{an}(v)$ and $v \in C_u^{(\max)} \setminus \mathrm{an}(u)$. In this case, we can simply remove either (or both) $u \hat{\prec} v$ or $v \hat{\prec} u$ from the initial ordering to obtain a relation that is a valid ordering.

The basic intuition of certifying an ancestral relationship by testing independence of residuals and regressors motivates the DirectLiNGAM (Shimizu et al., 2011) and Pairwise LiNGAM (Hyvärinen and Smith, 2013) procedures. To begin, one can select a root node, one without any parents or latent confounding, by finding a variable which is independent of all the residuals formed by regressing another variable onto it. Once a root is identified, its effect on the remaining variables can be removed and the root finding procedure recurs on the sub-graph of the remaining variables. The sequence of selected roots forms a topological ordering of the variables. An ordering of the nodes can also be identified in the opposite direction by finding sinks—nodes which have no children or latent confounding—by testing whether the residuals of a variable, when regressed onto all other variables, is independent of all other variables. Once a sink is identified, we simply recur onto the sub-graph of the remaining variables. We use *top-down* to refer to a procedure which successively identifies roots, and we use *bottom-up* to refer to a procedure which successively identifies sinks.

When we allow for latent confounding, a (certifiable) root or sink may not exist in the graph or in one of subsequent sub-graphs considered, so the Pairwise lvLiNGAM (Entner and Hoyer, 2010) and ParcelLiNGAM (Tashiro et al., 2014) procedures aim to estimate ancestral relationships between pairs of variables rather than a total ordering. We show in the appendix that lvLiNGAM may fail to discover even simple ancestral graphs, so we focus our discussion primarily on ParcelLiNGAM. Roughly speaking, ParcelLiNGAM applies both the top-down and bottom-up procedure to all subsets of V and certifies as many ancestral relationships as possible. Tashiro et al. (2014) show that the certification procedure is sound, but do not characterize a class of graphs for which the entire ParcelLiNGAM procedure is complete. In Lemma 1, we show that given population values, ParcelLiNGAM is indeed sound and complete for all ancestral graphs. Details are left for the appendix, but in short, we show that when a graph is ancestral, applying the bottom-up procedure to the subset $\mathrm{An}(v)$ will identify that $\mathrm{an}(v) \prec v$ for all v.

Lemma 1 *Suppose Y is generated by a recursive linear SEM that corresponds to an ancestral graph G. With generic model parameters and population information (i.e., the distri-*



(a)

(b)

Figure 2: The graph in (a) is a non-ancestral BAP which would be correctly identified by BANG but not Pairwise LvLiNGAM, ParcelLiNGAM, or RCD. The graph in (b) shows the graph which would be identified by Pairwise LvLiNGAM, ParcelLiNGAM, and RCD.

bution of Y), the ordering, $\hat{\prec}$, returned by Algorithm 2 of ParcelLiNGAM (Tashiro et al., 2014) is sound and complete for ancestral relationships in G.

2.2 Non-ancestral graphs

When G is not ancestral, the set of ancestral relationships that are certified by the described approach is still sound, but in general it is not complete. Indeed, in a non-ancestral graph, there exists some $v \in V$ such that $\text{sib}(v) \cap \text{an}(v) \neq \emptyset$. Thus, even if $c \in C = \text{pa}(v)$, it is generally not true that $\varepsilon_v \perp\!\!\!\perp Y_c$. This implies that $E\left(Y_C \varepsilon_v\right) \neq 0$ and the population regression coefficients $D_{v,C}$ no longer coincide with direct effects $B_{v,C}$. Indeed, in Lemma 2 we show that ParcelLiNGAM is no longer complete for non-ancestral BAPs; i.e., in every graph G which is bow-free but not ancestral, there are ancestral relations which will not be identified. We also show in the appendix that RCD (Maeda and Shimizu, 2020) cannot identify the BAP shown in Figure 2.a.

Lemma 2 *Suppose Y is generated by a recursive linear SEM that corresponds to a graph G which is bow-free but not ancestral. With generic parameters and population information, both Algorithm 2 and Algorithm 3 of ParcelLiNGAM (Tashiro et al., 2014) will return a partial ordering which is sound, but not complete for ancestral relationships in G.*

As a preview of the work ahead, in Example 1, we exhibit some of the complexities of discovering a non-ancestral graph by testing independence of residuals and regressors.

Example 1 *Consider discovering ancestral relationships in the BAP displayed in Figure 2.a.*

Nodes 1 and 2*: For this unconfounded pair, the direct approach of regressing Y_2 onto Y_1 yields the regression coefficent $\mathbb{E}(Y_1 Y_2)/\mathbb{E}(Y_1^2) = \beta_{2,1}$ and the residual $Y_2 - \beta_{2,1} Y_1 = \varepsilon_2$ is independent of the regressor $Y_1 = \varepsilon_1$. This independence certifies precedence of 1 before 2 in the graph, a relationship that would be discovered by ParcelLiNGAM and RCD.*

Nodes 2 and 3*: For general distributions in the model, there will not exist $d_{3,2} \in \mathbb{R}$ such that $Y_3 - d_{3,2} Y_2 \perp\!\!\!\perp Y_2$ since Y_2 depends on ε_1 and $1 \in \text{sib}(3)$. However, we can consider replacing Y_2 by an adjusted regressor. Having established that $1 \to 2$, we may take the adjusted regressor to be the residual $Y_2 - \beta_{2,1} Y_1 = \varepsilon_2$ found when regressing Y_2 onto Y_1 in the above first step. The choice $d_{32} = \beta_{3,2}$ then yields that $Y_3 - d_{3,2}\varepsilon_2 = \varepsilon_3 + \beta_{3,2}\beta_{2,1}\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$. This independence establishes that* 2 *precedes* 3.

Nodes 3 and 4*: At this point the latent confounding is such that the strategy considered so far fails. Indeed, for a general distribution in the model, there is no coefficient* d_{43} *such that the residual* $Y_4 - d_{43}Y_3'$ *is independent of the adjusted regressor* $Y_3' = Y_3 - \beta_{32}\varepsilon_2 = \varepsilon_3 + \beta_{32}\beta_{21}\varepsilon_1$ *from the previous step (since* $1 \in \text{sib}(4)$*). Other regressors such as* $Y_3' = Y_3$ *or* $Y_3' = \varepsilon_3$ *also do not yield independence as* $1 \in \text{sib}(4)$ *or* $1 \in \text{sib}(3)$*, respectively.*

Nevertheless, we can progress by giving up on the regression focus. Observe instead that setting $d_{4,3} = \beta_{4,3}$ *yields* $\varepsilon_3 \perp\!\!\!\perp \varepsilon_4 = Y_4 - d_{4,3}Y_3$*. In this case, the independence test focuses on the error term and not the regressor and is able to certify that 3 precedes 4. In the sequel, we will show that this alternative type of independence certificate is in fact sound and complete for parental relationships in BAPs. In the present example, since there are no other parental relationships that can be certified, but dependencies still remain, we may conclude (correctly) that all other pairs are siblings.*

3. Bow-free acyclic path diagrams

We now turn to a larger class of graphs, namely, graphs that are bow-free and acyclic. These are then not necessarily ancestral. We begin by presenting results that will be used to motivate the discovery algorithm presented in subsequent Section 4. Specifically, the results in this section provide a certificate of ancestral relationships. We establish this certificate by showing that an identification formula produces correct direct causal effects when it is applied to suitable sets of ancestors and fails to do so otherwise. The latter statement requires assumptions of genericity of the considered distribution. The certificate is derived from a series of lemmas that investigate specific aspects. Most proofs in this section are deferred to the appendix.

3.1 Comments on genericity assumptions

Throughout, we will consider higher order moments as a proxy for independence; i.e., for random variables X and Z with $\mathbb{E}(X) = \mathbb{E}(Z) = 0$ and $K > 2$, we will use $\mathbb{E}(X^{K-1}Z) = 0$ as a stand-in for $X \perp\!\!\!\perp Z$ and $\mathbb{E}(X^{K-1}Z) \neq 0$ as a stand-in for $X \not\perp\!\!\!\perp Z$. For fixed K, the vanishing of moments and independence are equivalent for random variables derived using generic model parameters, which are the matrix B and the moments of ε. Crucially, the moments are polynomials of the model parameters which will allow us to leverage basic algebraic results to show identifiability. In particular, as done above, we make statements which hold for *generic* parameters. By generic, we mean that the parameters for which the statements do not hold have measure 0 with respect to Lebesgue measure. In the sequel, we will fix $K > 2$, and consider moments of the form $\mathbb{E}(X^{K-1}Z)$ which depend only on the moments of ε up to degree K; i.e., $\mathbb{E}\left(\prod_{v\in V} \varepsilon_v^{r_v}\right)$ for all $r \in \mathbb{Z}^p_{\geq 0}$ with $\|r\|_1 \leq K$. Thus, when we refer to generic error moments, we mean generic values for the error moments up to degree K.

Readers may be familiar with genericity assumptions in the form of faithfulness conditions that in simpler settings may be phrased in terms of conditional independence constraints. In our more involved setting, an exact characterization of the set of parameters for which the statements fail is difficult—this is analogous to the situation in Brito and Pearl (2002) or also Foygel et al. (2012). Although, in principle, one could recursively collect polynomials whose joint non-vanishing is sufficient for our conclusions to hold, this provides little additional insight. This said, existing results on distributional equivalence of BAPS with Gaussian errors (Nowzohour, 2017) do imply that error moments which correspond to a Gaussian distribution must be avoided. Put another way, fixing the covariance for a multivariate Gaussian determines all higher order moments. Thus, it is necessary for identification that the higher-order moments of the data generating distribution be different than the higher-order moments of the Gaussian distribution which shares the same covariance as the data generating distribution.

In addition, our genericity assumptions must avoid parameter values for which the direct effect of u on v—in some marginal model—vanishes for some $u \in pa(v)$. This is like a faithfulness assumption for Gaussian graphical models where certain partial correlations which are generally non-zero may vanish for specific configurations of the model parameters. However, the set of parameters for which the partial correlations vanish is a set of measure zero with respect to Lebesgue measure. In contrast to the Gaussian model which is fully defined by the first and second moments, we explicitly consider higher order moments. Thus our notion of genericity or faithfulness is also with respect to higher order moments.

3.2 Setup

We now define and review some additional notions that will be helpful for the upcoming results. Specifically, we introduce the concept of marginal direct effects, which is used in Section 3.3 to show that the direct effect of a parent onto its child does not disappear even in sub-models where some observed variables have been marginalized away. In addition, in Algorithm 1 defined in Section 4, we keep a running estimate of the direct effects in a matrix D; i.e., D is an estimate of B which we update iteratively. In this section, we describe the involved updates to D, whose entries can be seen as marginal direct effects for some particular marginal model.

Consider the BAP $G = (V, E_{\to}, E_{\leftrightarrow})$ and corresponding parameter $B = (\beta_{u,v})_{u,v \in V}$. For a directed path $l = v_1 \to \ldots \to v_s$, define the *pathweight* of l as $W(l) := \prod_{j=1}^{s-1} \beta_{v_{j+1},v_j}$. Let $\mathcal{L}_{v,u}$ be the set of all directed paths from u to v in $E_{\to}$. Given a set C with $u \in C$, we can partition $\mathcal{L}_{v,u}$ into disjoint sets as $\mathcal{L}_{v,u} = \bigcup_{c \in C} \mathcal{L}^{(c)}_{v,u}(C)$, where $\mathcal{L}^{(c)}_{v,u}(C)$ is the subset of paths in $\mathcal{L}_{v,u}$ such that c is the last node in C to appear on the path. Thus, $\mathcal{L}^{(u)}_{v,u}(C)$ is also the set of paths from u to v which do not pass through $C \setminus u$.

For a set $A \subset V$ and $u, v \in A$, let the *marginal direct effect* be the direct effect between u, v in the sub-model obtained by marginalizing out all variables in $A^c := V \setminus A$. For convenience, let $\Lambda = I - B$. Then the marginal direct effects for all $u, v \in A$ are encoded by the matrix

$$\begin{aligned}\tilde{B}(A) = I - \left[\left(\Lambda^{-1}\right)_{A,A}\right]^{-1} &= I - \left[\left(\Lambda_{A,A} - \Lambda_{A,A^c}(\Lambda_{A^c,A^c})^{-1}\Lambda_{A^c,A}\right)^{-1}\right]^{-1} \\ &= I - \Lambda_{A,A} - \Lambda_{A,A^c}(\Lambda_{A^c,A^c})^{-1}\Lambda_{A^c,A}.\end{aligned} \tag{6}$$

In particular, $\tilde{B}(A)_{v,u} = \beta_{v,u} + \sum_{s \in A^c} \beta_{v,s} \sum_{t \in A^c} \pi'_{s,t} \beta_{t,u}$ where

$$\pi'_{s,t} = ((\Lambda_{A^c,A^c})^{-1})_{s,t}$$

is the total effect of t on s in the sub-graph of G induced by A^c. Graphically, this total effect corresponds to the sum of the pathweights of all paths from t to s which only pass through nodes in A^c. This implies that

$$\tilde{B}(C \cup \{v\})_{v,u} = \sum_{l \in \mathcal{L}^{(u)}_{v,u}(C)} W(l). \tag{7}$$

Moreover, this implies that for $u, v \in A$, $\tilde{B}(A)_{v,u} \neq 0$ only if $u \in \mathrm{an}(v)$.

Let $D \in \mathbb{R}^{p \times p}$ be some estimate of the direct effects B; the support of D and B may differ. Let $E^{(D)}_{\rightarrow}$ be the set of directed edges defined by the support of D. Define the *pseudo-parents* of v given D, $\mathrm{pa}_D(v)$, to be the set of parents of v in $E^{(D)}_{\rightarrow}$ and define the *pseudo-ancestors* of v given D, $\mathrm{an}_D(v)$, to be the ancestors of v in $E^{(D)}_{\rightarrow}$ and $\mathrm{An}_D(v) = \mathrm{an}_D(v) \cup \{v\}$.

Typically we will only consider matrices D such that $\mathrm{pa}_D(v) \subseteq \mathrm{an}(v)$; i.e, $D_{v,u} \neq 0$ only if $u \in \mathrm{an}(v)$. However, it will sometimes be useful to place an additional restriction on D. Consider a family of sets $\mathcal{C} = (C_v)_{v \in V}$ where $C_v \subseteq V \setminus \{v\}$. Such a family defines the matrix-valued function $H_{\mathcal{C}}$ which maps $B \in \mathbb{R}^{p \times p}$ to $D \in \mathbb{R}^{p \times p}$ given by

$$D_{v,u} = \begin{cases} \tilde{B}(C_v \cup \{v\})_{v,u} & \text{if } u \in C_v, \\ 0 & \text{else.} \end{cases} \tag{8}$$

Thus, for any D which is the output of $H_{\mathcal{C}}$, the vth row corresponds to the marginal direct effects of the sub-model induced by $C_v \cup \{v\}$. Each element $D_{v,u}$ is the sum of pathweights for a (not necessarily strict) subset of the paths from v to u, and thus is a polynomial in the elements of B. The specific paths over which the sum is taken—and thus specific form of the polynomial—depends on $\mathcal{C}$. Finally, let $\mathcal{D}$ be the set of functions $H_{\mathcal{C}}$ obtained from all families $\mathcal{C} = (C_v)_{v \in V}$ with $C_v \subseteq V \setminus \{v\}$.

3.3 Certifying ancestral relationships in non-ancestral graphs

In general, we use the symbol γ to denote residuals. Specifically, for $c \in V$, let $\gamma_c(D)$ denote the resulting residual of variable c when positing D to be the matrix of direct effects; i.e.

$$\gamma_c(D) = Y_c - D_{c,V} Y_V. \tag{9}$$

For $v \in V$, we introduce the *debiased direct effect* $\delta_v(C, A, S, D)$ as a function of sets C and A with $C \subseteq A \subseteq V \setminus \{v\}$ and matrices $S, D \in \mathbb{R}^{p \times p}$, where S is the (possibly sample) covariance of Y:

$$\delta_v(C, A, S, D) = \left\{ \left[(I - D)_{C,A} S_{A,C} \right]^{-1} (I - D)_{C,A} S_{A,v} \right\}^T. \tag{10}$$

Overloading the notation slightly, let $\gamma_v(C, S, D)$ denote the residual when using the debiased direct effect in (10) calculated with C, S and D, and setting $A = \mathrm{An}_D(C)$; i.e.,

$$\gamma_v(C, S, D) = Y_v - \delta_v(C, \mathrm{An}_D(C), S, D) Y_C. \tag{11}$$

When the arguments for $\gamma_c(D)$ and $\gamma_v(C, S, D)$ are clear from the context, we will suppress the additional notation.

Lemma 3 *Suppose Y is generated by a linear SEM with parameters B and Ω whose supports respect the sparsity imposed by the BAP G. For node $v \in V$ and sets of nodes $C \subseteq A \subseteq V \setminus \{v\}$, suppose* *(i)* $pa(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$*,*

(ii) $A = \mathrm{An}(C)$*,*

(iii) $D_{A,A} = B_{A,A}$*, and*

(iv) $S_{A\cup\{v\},A\cup\{v\}} = \Sigma_{A\cup\{v\},A\cup\{v\}}$*.*

Then $\delta_v(C, A, S, D) = B_{v,C}$*.*

Proof For any $v \in V$ and set C such that $\mathrm{pa}(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$, it follows from (3) that

$$\left[(I - B)\Sigma(I - B)^T\right]_{u,v} = \omega_{u,v} = 0$$

for all $u \in C$ because $C \cap \mathrm{sib}(v) = \emptyset$. Since $B_{v,V\setminus C} = 0$ and $B_{C,V\setminus \mathrm{an}(v)} = 0$, this yields

$$B_{v,C}\Sigma_{C,\mathrm{an}(v)}(I - B^T)_{\mathrm{an}(v),C} = \Sigma_{v,\mathrm{an}(v)}(I - B^T)_{\mathrm{an}(v),C},$$

so that

$$B_{v,C} = \Sigma_{v,\mathrm{an}(v)}(I - B^T)_{\mathrm{an}(v),C}\left(\Sigma_{C,\mathrm{an}(v)}(I - B^T)_{\mathrm{an}(v),C}\right)^{-1}.$$

■

Lemma 3 states that given the population covariance and direct effects between vertices which are "causally upstream" of v, selecting appropriate sets C and A such that $\mathrm{pa}(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ and $A = \mathrm{An}(C)$ allows recovery of the direct effect of C onto v. Since δ_v only involves matrix inversion and multiplication, it is a rational function of the elements of S and D. The specific form of the function is determined by the sets C and A.

We use the name debiased direct effect because δ_v can be calculated by the following alternative procedure. First form the errors ε_C by regressing each $c \in C$ onto its parents. and regress Y_v onto ε_C. This would yield the total effect of C on v, but since Y_v contains terms involving ε_A, it will be biased by dependence between ε_C and ε_A. Given $B_{A,A}$, however, $\Omega_{A,A}$ can be computed from Σ. Thus, the naive regression coefficients can be debiased to give the true direct effects. The assumption that $C \cap \mathrm{sib}(v) = \emptyset$ ensures that we do not need to also correct for dependence between ε_C and ε_v which we would not be able to calculate with the given information.

Of course, in practice, we do not a priori know the relationships between candidate sets C and v, but the following results show we can certify whether we have selected appropriate sets C and A. Specifically, Algorithm 1, which is proposed in Section 4, will certify that $C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ by testing if

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, S, D)\right) = 0 \qquad \forall c \in C. \tag{12}$$

Corollary 4 *Suppose the conditions in Lemma 3 hold, then for every* $c \in C$*, we have* $\gamma_c(D) \perp\!\!\!\perp \gamma_v(C, S, D)$ *and* $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D)) = 0$*.*

Proof By assumption $D_{C,A} = B_{C,A}$ so that $\gamma_C(D) = \varepsilon_C$. In addition, Lemma 3 implies that $\delta(C, A, S, D) = B_{v,C}$, so $\gamma_v(C, S, D) = \varepsilon_v$. Since we assume $C \cap \mathrm{sib}(v) = \emptyset$, it holds that $\gamma_c \perp\!\!\!\perp \gamma_v$ for all $c \in C$. ■

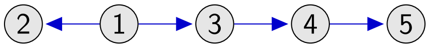


Figure 3: When $D_{2,1} = \beta_{2,1}$ and $D_{5,4} = \beta_{5,4}$, naively testing (12) would mistakenly certify 2 and 5 as ancestors of 3.

In Algorithm 1, we use (12) as a certificate that $C \subseteq \mathrm{an}(v) \backslash \mathrm{sib}(v)$, and indeed Corollary 4 shows that $C \subseteq \mathrm{an}(v) \backslash \mathrm{sib}(v)$ is part of a set of sufficient conditions for (12) to hold. However, it is not necessary and more care is needed to ensure that (12) will not mistakenly certify a set C if $C \not\subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$. We first state Lemma 5 which gives a necessary condition for (12) that will be useful in deriving subsequent results.

Lemma 5 *Let $v \in V$ and $C \subseteq V \setminus \{v\}$. Suppose $D \in \mathbb{R}^{p \times p}$ such that $D_{s,t} \neq 0$ only if $t \in \mathrm{an}(s)$. Then, for generic B and error moments, if $\delta_v(C, \mathrm{an}_D(C), S, D) \neq \tilde{B}(C \cup \{v\})_{v,C}$, then $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, S, D)) \neq 0$ for some $c \in C$.*

Lemma 8 shows that if $C \cap \mathrm{sib}(v) \neq \emptyset$, then there exists some $c \in C$, such that $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D)) \neq 0$. Ensuring that we do not mistakenly certify non-ancestors of v is a bit more delicate because, depending on D, (12) may actually hold for some set $C \not\subseteq \mathrm{an}(v)$ when $C \cap \mathrm{sib}(v) = \emptyset$. In particular there are two cases of potential ways a set can be misscertified. First, C may contain a node $c \notin \mathrm{an}(v) \cup \mathrm{de}(v)$. Alternatively, C may contain a descendant of v which already has the effect of v removed; for instance, for some c such that $v \in \mathrm{an}(c) \setminus \mathrm{pa}(c)$, if $D_{c,V} = B_{c,V}$ then $\gamma_c(D)$ will not contain any term with ε_v. Consider the example in Figure 3, and let $D_{2,1} = \beta_{2,1}$ and $D_{5,4} = \beta_{5,4}$. The set $C = \{2, 5\}$ would satisfy (12) for $v = 3$ because 2 is neither an ancestor or descendant of 3 and 5 is an ancestor of 2, but adjusting 5 by 4 removes the effect of 3.

More generally, suppose that $C \cap \mathrm{sib}(v) = \emptyset$ but $C \not\subseteq \mathrm{an}(v)$, and let $C_1 = C \cap \mathrm{an}(v)$ and $C_2 = C \setminus \mathrm{an}(v)$. Thus, C_1 should rightfully be certified, but C_2 should not. However, if D is such that $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, S, D)) = 0$ for all $c \in C$, then C (including C_2) could be mistakenly certified as a subset of $\mathrm{an}(v) \backslash \mathrm{sib}(v)$. Fortunately, Lemma 6 implies that instead of testing all possible sets $C \subseteq V \setminus \{v\}$, we can use an additional pre-screening procedure to filter out problematic sets, $C \not\subseteq \mathrm{an}(v)$ which would otherwise satisfy (12).

Specifically, Lemma 6 implies that

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1, S, D)) = 0 \qquad \forall c \in C_1. \tag{13}$$

Thus, we still would have certified that $C_1 \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$. Furthermore, after adjusting v for C_1, the resulting residuals of v—$\gamma_v(C_1, S, D)$—would also be independent of γ_c for all $c \in C_2$; i.e.,

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1, S, D)) = 0 \qquad \forall c \in C_2. \tag{14}$$

Thus, we can screen out non-ancestors of v which might otherwise be miscertified, by removing any $c \in C$ such that for some $C' \subseteq C \setminus \{c\}$,

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C', S, D)) = 0. \tag{15}$$

This is implemented in Algorithm 2. A concern is then the question whether the prescreening procedure implemented in Algorithm 2 may mistakenly rule out a parent or sibling of v. To show that this does not happen for generic parameters, we derive Lemma 7 below.

Lemma 7 and 8 both require that $D = H_{\mathcal{C}}(B)$; i.e., for each $v \in V$, $D_{v,C}$ is the marginal direct effect of C on v where C is the set of non-zero entries in $D_{v,V}$. In Algorithm 3, we only update $D_{v,C}$ to $\delta_v(C, A, S, D)$ when $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C,S,D)) = 0$ for all $c \in C$. As shown by Lemma 5, this implies that $\delta_v(C, \mathrm{An}_D(C), S, D)$ is the marginal direct effect of C on v so that the updated $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$.

Lemma 6 *Consider $v \in V$ and $C \subseteq V \setminus \{v\}$. Let $D \in \mathbb{R}^{p\times p}$ such that $D_{s,t} \neq 0$ only if $t \in \mathrm{an}(s)$. Suppose $C \not\subseteq \mathrm{an}(v)$, but that $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C,S,D)) = 0$ for all $c \in C$. Then for generic B and error moments, $C_1 = C \cap [\mathrm{an}(v) \setminus \mathrm{sib}(v)]$,*

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1,S,D)) = 0 \qquad \forall c \in C.$$

Lemma 7 *Suppose $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \mathrm{an}(s) \setminus \mathrm{sib}(s)$. Let $v \in V$ be such that we have $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(\mathrm{pa}_D(v),S,D)) = 0$ for all $c \in \mathrm{pa}_D(v)$. If $q \in (\mathrm{pa}(v) \setminus \mathrm{pa}_D(v)) \cup \mathrm{sib}(v)$, then for generic B and error moments, $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$.*

Lemma 8 *Consider $v \in V$ and C such that $C \subseteq V \setminus \{v\}$. Suppose $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \mathrm{an}(s) \setminus \mathrm{sib}(s)$ for all $v \in V$. If $C \cap \mathrm{sib}(v) \neq \emptyset$, then for generic B and error moments, there exists some $q \in C$ such that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(C,\Sigma,D)\right) \neq 0$.*

Thus far we have been concerned with discovering sets which contain ancestors but not siblings of some node v. Corollary 9 shows that when we have identified such a set which is also a superset of the parents of v, we can prune away ancestors which are not parents. This motivates the pruning procedure described in Algorithm 4.

Corollary 9 *Suppose $D = B$. For $v \in V$ and generic B and error moments, suppose $\mathrm{pa}(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$. If $q \in C \setminus \mathrm{pa}(v)$, then for all $c \in C$*

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\},\Sigma,D)) = 0. \tag{16}$$

If $q \in \mathrm{pa}(v)$, then there exists some $c \in C$ such that

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\},\Sigma,D)) \neq 0. \tag{17}$$

4. Graph estimation algorithm

Using the claims established above, we present the **B**ow-free **A**cyclic **n**on-**G**aussian (BANG) procedure in Algorithm 1 which completely identifies the underlying causal structural of the linear SEM when it corresponds to a BAP.

The algorithm starts with a complete bidirected graph so that the posited siblings for each node, $\widehat{\mathrm{sib}}(v)$, are initialized to $V \setminus v$ and the posited parents $\widehat{\mathrm{pa}}(v)$, are initialized to $\emptyset$. The method then iteratively certifies ancestors which are not siblings by considering whether

(12) holds for progressively larger sets. When we certify that $C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$, C is added to $\widehat{\mathrm{pa}}(v)$, C is removed from $\widehat{\mathrm{sib}}(v)$, and D is updated. This procedure is repeated until no additional ancestral relationships can be certified. Any remaining dependency between the residuals are then assumed to be due to a bidirected edge. In the algorithm, whenever we specify a test for $X \perp\!\!\!\perp W$, we mean testing $\mathbb{E}(X^{K-1}W) = 0$ for some prespecified $K > 2$.

4.1 Graph identification

We first show that when given population values, the BANG procedure will return the correct graph.

Theorem 10 *Suppose Y is generated by a linear SEM which corresponds to a BAP G. Then for generic choices of B and error moments, Algorithm 1 will output $\hat{G} = G$ when given population moments of Y.*

Proof The lemmas in Section 3 make statements about different individual quantities being non-zero for generic B and error moments. Since we will only consider a finite set of these quantities, the union of the null sets to be avoided for each individual quantity is also a null set. Thus, in this proof, we may assume that quantities that are generically non-zero are all actually non-zero.

Our proof proceeds by induction. In particular, let σ be a topological ordering consistent with the directed portion of underlying graph G; i.e., $\sigma(u) < \sigma(v)$ implies that $u \notin \mathrm{de}(v)$. Let the zth induction step be defined as an entire step of testing progressively larger sets C until all parents of $v = \sigma^{-1}(z)$ have been discovered. Thus, there are at most p steps. Note that since we do not know the ordering a priori and simply cycle over all variables and progressively larger sets, it could be that zth induction step is actually completed (i.e., all the parents of $\sigma^{-1}(z)$ are discovered) chronologically before the $(z-1)$th induction step is done.

If, as we show, the induction hypothesis below holds through the final step p, then after Line 10 of Algorithm 1, the procedure obtains $\widehat{\mathrm{pa}}(v)$ such that $\mathrm{pa}(v) \subseteq \widehat{\mathrm{pa}}(v) \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ and $D = B$ where D is an estimate of the direct effects and B are the true direct effects. This implies that $\gamma_v(D) = \varepsilon_v$ for all v so that $\gamma_v(D) \not\perp\!\!\!\perp \gamma_u(D)$ if and only if $u \in \mathrm{sib}(v)$ so that $\widehat{\mathrm{sib}}(v) = \mathrm{sib}(v)$. Then, using Algorithm 4, prunes away from $\widehat{\mathrm{pa}}(v)$ any nodes which are ancestors but not parents so that $\widehat{\mathrm{pa}}(v) = \mathrm{pa}(v)$.

Algorithm 1 BANG procedure

```
 1: Input: Data $Y \in \mathbb{R}^{p \times n}$ and $S \in \mathbb{R}^{p \times p}$ which is the (potentially sample) covariance of $Y$
 2: For all $v \in V$, set $\widehat{\mathrm{pa}}(v) = \emptyset$ and $\widehat{\mathrm{sib}}(v) = V \setminus \{v\}$
 3: Set all elements of $D \in \mathbb{R}^{p \times p}$ to be 0 and $l = 1$
 4: while $\max_v |\widehat{\mathrm{sib}}(v)| \geq l$ do
 5:     for $v \in V$ do
 6:         Prune $\widehat{\mathrm{sib}}(v)$ using Algorithm 2
 7:         Certify pseudo-parents of $v$ and update $\widehat{\mathrm{pa}}(v)$, $\widehat{\mathrm{sib}}(v)$, and $D$ using Algorithm 3
 8:     end for
 9:     if $D$ was updated, reset $l = 1$;  else  set $l = l + 1$
10: end while
11: Remove ancestors which are not parents from $\widehat{\mathrm{pa}}(v)$ for all $v \in V$ using Algorithm 4
12: Return: $\hat{E}_{\rightarrow} = \{(u, v) : u \in \widehat{\mathrm{pa}}(v)\}$, $\hat{E}_{\leftrightarrow} = \{\{u, v\} : u \in \widehat{\mathrm{sib}}(v)\}$
```

Algorithm 2 Prune $\widehat{\text{sib}}(v)$

```
Input: $v$, $\widehat{\text{sib}}(v)$, $Y$, $D$
Set $\gamma(D) = Y - DY$
for $u \in \widehat{\text{sib}}(v)$ do
    if $\gamma_u(D) \perp\!\!\!\perp \gamma_v(D)$ then
        Remove $u$ from $\widehat{\text{sib}}(v)$ and remove $v$ from $\widehat{\text{sib}}(u)$
    end if
end for
Return: $\widehat{\text{sib}}(v)$ for all $v \in V$,
```

Algorithm 3 Certify pseudo-parents

```
Input: $v$, $\widehat{\text{pa}}(v)$, $\widehat{\text{sib}}(v)$, $D$, $S$, $Y$, $l$
Set $C^\star = \emptyset$
for $C \subseteq \widehat{\text{sib}}(v)$ such that $|C| = l$ do
    if $\gamma_C(D) \perp\!\!\!\perp \gamma_v(C \cup \widehat{\text{pa}}(v), S, D)$ then
        $C^\star = C^\star \cup C$
    end if
end for
$\widehat{\text{pa}}(v) = \widehat{\text{pa}}(v) \cup C^\star$
$D_{v,\widehat{\text{pa}}(v)} = \delta_v(\widehat{\text{pa}}(v), S, D)$
$\widehat{\text{sib}}(v) = \widehat{\text{sib}}(v) \setminus \widehat{\text{pa}}(v)$
$\widehat{\text{sib}}(s) = \widehat{\text{sib}}(s) \setminus \{v\} \qquad \forall s \in \widehat{\text{pa}}(v)$
Return: $D$, $\widehat{\text{sib}}(v)$ and $\widehat{\text{pa}}(v)$ for all $v \in V$,
```

As the induction hypothesis for step z, let $v = \sigma^{-1}(z)$ and suppose that:

1. For $A = \sigma^{-1}([z-1])$, $D_{A,A} = B_{A,A}$ and $\widehat{\text{pa}}(a) \supseteq \text{pa}(a)\ \forall a \in A$;
2. $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ where $\forall s \in V$, $\widehat{\text{pa}}(s) \subseteq \text{an}(s) \setminus \text{sib}(s)$ (i.e., $C_s \subseteq \text{an}(s) \setminus \text{sib}(s)$);
3. For all $u \in V$, $\widehat{\text{sib}}(u) \supseteq \text{sib}(u)$ and $\text{pa}(u) \subseteq \{\widehat{\text{sib}}(u) \cup \widehat{\text{pa}}(u)\}$.

The first condition assumes all directed edges upstream of v have been identified. The second condition assumes that each row in the current value of D corresponds to a marginal direct effect and that nothing has been misscertified into $\widehat{\text{pa}}(v)$. The third condition assumes no siblings or parents have been incorrectly pruned.

For the base case, when $z = 1$, the first condition is trivially satisfied because by definition there are no edges upstream of $v = \sigma^{-1}(1)$. The second condition is satisfied because $\widehat{\text{pa}}(v)$ is initialized to be empty and D is initialized to be all 0's which is the marginal direct effect for the empty set. Finally, since $\widehat{\text{sib}}(v)$ is initialized to $V \setminus \{v\}$, the third condition also holds. We now show that when the conditions hold for the zth step—after completion—the induction conditions will hold for step $z + 1$.

Condition 3: By assumption, $D_{A,A} = B_{A,A}$ and $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$, so $\text{An}_D(\text{pa}(v)) = \text{An}(\text{pa}(v))$. Lemma 7 implies that for all $u \in V$, Algorithm 2 does not mistakenly remove any siblings or parents of u from $\widehat{\text{sib}}(v)$. Furthermore, Lemma 8 implies that Algorithm 3 will not remove any siblings from $\widehat{\text{sib}}(v)$. Thus $\text{pa}(u) \subseteq \big(\widehat{\text{sib}}(u) \cup \widehat{\text{pa}}(u)\big)$ and $\widehat{\text{sib}}(u) \supseteq \text{sib}(u)$ continue to hold, and Condition 3 is satisfied for the next step.

Condition 2: Algorithm 3 only adds C to $C^\star$ if $C \cup \widehat{\text{pa}}(v)$ satisfies (12). Lemma 8 implies that any set C such that $C \cap \text{sib}(v) \neq \emptyset$ will not be added. We now show that any

Algorithm 4 Prune ancestors which are not parents

```
Input: $\widehat{\text{pa}}(v)$ and $\widehat{\text{sib}}(v)$ for all $v \in V$, $D$, $S$, $Y$, $l$
Form topological ordering $\sigma$ such that $\sigma(u) < \sigma(v)$ implies $v \notin \text{an}_D(u)$
for $v \in \sigma^{-1}([p])$ do
    for $s \in \widehat{\text{pa}}(v)$ do
        if $\gamma_c \perp\!\!\!\perp \gamma_v(\widehat{\text{pa}(v)} \setminus \{s\}, S, D)$ for all $c \in \hat{\text{p}}\text{a}(v)$ then
            $\widehat{\text{pa}}(v) = \widehat{\text{pa}}(v) \setminus \{s\}$
            $D_{v,s} = 0$
        end if
    end for
end for
```

set $C \not\subseteq \text{an}(v)$ will not be added to $C^\star$ (and eventually $\widehat{\text{pa}}(v)$) because it either will not be considered by Algorithm 3 or will not satisfy (12).

For some $v \in V$, let $C \subseteq V \setminus \widehat{\text{pa}}(v)$ and $C_1 = C \cap \text{an}(v)$. If $C \not\subseteq \text{an}(v)$ and $C_1 \neq C$, the set $C \cup \widehat{\text{pa}}(v)$ will only be considered by Algorithm 3 after the set $C_1 \cup \widehat{\text{pa}}(v)$. If $C \cup \widehat{\text{pa}}(v)$ satisfies (12), then Lemma 6 implies that $C_1 \cup \widehat{\text{pa}}(v)$ also satisfies (12). Thus, C_1 would first be certified into $\widehat{\text{pa}}(v)$. Lemma 6 further implies that for any $c \in C$, $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1 \cup \widehat{\text{pa}}(v), S, D) = 0$, so that after C_1 is placed in $\widehat{\text{pa}}(v)$ and D is updated, we have $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(D)) = 0$ for all $c \in C \setminus C_1$. Hence, Algorithm 2 will subsequently remove $C \setminus C_1$ from $\widehat{\text{sib}}(v)$. Thus, C will only be considered if $C \subseteq \text{an}(v)$ or if $C \cup \widehat{\text{pa}}(v)$ does not satisfy (12). Thus, any updates preserve $C^\star \subseteq \text{an}(v) \setminus \text{sib}(v)$. Because $C^\star$ is the union of certified sets, Lemma 28 implies that $C^\star$ will also be certified. Finally, Lemma 5 implies that after $C^\star$ is added to $\widehat{\text{pa}}(v)$, the resulting update to the vth row of D is a marginal direct effect so that $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$. Thus, Condition 2 continues to hold.

Condition 1: By the acyclicity assumption, $|\text{pa}(v)| \leq z - 1$. So by successively testing larger sets, and resetting the counter after each update, if we do not first certify all parents of v as part of smaller sets, we will eventually consider $C = \text{pa}(v)$. The induction hypothesis and Lemma 4 ensure that $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D)) = 0$ for all $c \in C$ so that C will be certified into $\widehat{\text{pa}}(v)$. Lemma 5 implies the resulting update which sets $D_{v,\widehat{\text{pa}}(v)} = \delta_v(\widehat{\text{pa}}(v), \text{An}_D(\widehat{\text{pa}}(v)), S, D)$ will result in $D_{v,V} = B_{v,V}$. Thus, induction step z will be completed, and Condition 1 continues to hold.

After p steps, $D = B$, so $\gamma_v(D) = \varepsilon_v$ for all v and $\mathbb{E}(\gamma_u(D)^{K-1}\gamma_v(D)) \neq 0$ if and only if $u \in \text{sib}(v)$. If $\widehat{\text{sib}}(v) \neq \emptyset$ for any v then after the last update to D, there will be at least one more pass through Algorithm 2 so any non-siblings will be removed and $\widehat{\text{sib}}(v) = \text{sib}(v)$ for all $v \in V$. If $\widehat{\text{sib}}(v) = \emptyset$ for all v, then by the induction conditions, $\text{sib}(v) \subseteq \widehat{\text{sib}}(v) = \emptyset$, so again, $\text{sib}(v) = \widehat{\text{sib}}(v)$.

By Condition 2, $\text{pa}(v) \subseteq \widehat{\text{pa}}(v) \subseteq \text{an}(v) \setminus \text{sib}(v)$, and Corollary 9 implies that Algorithm 4 removes any ancestors from $\widehat{\text{pa}}(v)$ which are not parents but does not remove any parents. Thus, $\widehat{\text{pa}}(v) = \text{pa}(v)$. ■

Theorem 10 shows that the graph is correctly identified given population values by successively testing whether a quantity is zero or non-zero. However, the quantities considered are non-linear functions of the data so in finite samples, in addition to sample variability, the sample quantities will typically be biased. Nonetheless, the following corollary shows that there exists a cut-off $\eta_1 > 0$ such that checking whether each sample statistic is greater than or less than $\eta_1/2$ as a proxy for independence will yield consistent estimates of G as long as the sample moments of Y are consistent for the population moments. The value of η_1 depends on the model parameters, but some $\eta_1 > 0$ must exist for generic B and error moments. This implies pointwise consistency of BANG when the tests are "appropriately" tuned. Of course η_1 depends on quantities that are unknown in practice, so in applications we find ourselves in a similar position as for other existing constraint-based algorithms in causal discovery (e.g., PC or FCI algorithm) where algorithm output delicately depends on suitable specification of levels for statistical tests which act as tuning parameters.

Corollary 11 *Suppose Y is a sample comprised of i.i.d. vectors $Y^{(1)}, \dots, Y^{(n)}$ generated by a linear SEM that corresponds to the BAP G. Then, for generic choices of B and error moments, there exist $\eta_1, \eta_2 > 0$ such that when the sample moments are within an η_2-ball of the population moments of Y, Algorithm 1 will output $\hat{G} = G$ when comparing the absolute value of the sample statistics to $\eta_1/2$ as a proxy for the independence tests.*

Proof In Theorem 10, we showed that BANG will correctly identify the true BAP as long as certain expectations encoding absence of edges and paths are all 0 and further expectations encoding presence of edges/paths are all non-zero (which holds for generic B and error moments). For a fixed BAP G, let S_0 be the set of expectations which should be 0, and let S_1 be the set of expectations which should be non-zero. Let $\eta_1 = \min_{S_1} |\mathbb{E}(\gamma_c^{k-1}\gamma_v)|$. For generic parameters, $\eta_1 > 0$.

We note that when $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$, the maps which take moments of Y to $E(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D))$ are rational functions and are thus Lipschitz within a sufficiently small ball around the population moments of Y. Thus, there must exist some $\eta_2 > 0$ such that when the sample moments are within η_2 of the population moments, the sample quantities in S_0 and S_1 are within $\eta_1/2$ of the population quantities.

This implies that all estimates corresponding to quantities which are 0 are less than $\eta_1/2$ in absolute value, and all estimates that correspond to quantities which are generically non-zero are greater than $\eta_1/2$ in absolute value. Thus, comparing the absolute value of the sample quantities to $\eta_1/2$ accurately determines whether the parameters belong to S_0 or S_1 and thus yields a correct estimate $\hat{G}$. ■

4.2 Practical concerns

For any BAP, identification with population values holds for all but a null set of B and error moments. As discussed in Section 3, this is similar to the faithfulness assumption required in Gaussian graphical models. However, the typical Gaussian faithfulness condition only regards the linear coefficients and error covariances, whereas our notion involves linear coefficients as well as the higher order moments of the errors. Given a finite number of samples, consistent recovery of the graph would also require an analogue to "strong faithfulness;" i.e., the quantities which we require to be non-zero must be bounded away from 0 by a sufficiently large amount. As shown in Uhler et al. (2013), a careful characterization of strong faithfulness is already difficult in the Gaussian setting, and even more so for our setting. Nonetheless, we can make some general conclusions. The quantities we consider for identification are continuous in the error moments. As previously stated, we require that the higher-order moments of the data generating distribution must be different than the higher-order moments of the Gaussian distribution which shares the same covariance as our data generating distribution. Thus, if the higher-order moments of the errors are close to the higher-order moments of the Gaussian distribution which shares the same covariance, then the quantities we require to be non-zero would likely still be close to 0. Thus, we would require a very large sample size to consistently determine that the population quantity is non-zero. As we see in the simulations, when the errors come from a multivariate T distribution with moderately large degrees of freedom, which is not too different from a Gaussian, the finite sample performance suffers considerably.

Throughout the proof, we examine high order moments as a proxy for independence. Since these quantities are polynomials of the parameters, it allows us to make algebraic arguments that facilitate the analysis. However, in practice, one could use any non-parametric independence test instead (Gretton et al., 2005; Székely and Rizzo, 2009; Bergsma and Dassios, 2014; Pfister et al., 2018). In Section 6, we use dHSIC (Pfister et al., 2018) which performs well when the sample size is small. However, simply calculating the statistic requires $O(n^2)$ time rendering the permutation or bootstrap procedures required for calibrating a null distribution prohibitively expensive; we thus use the "gamma approximation" to the null distribution. However, even this becomes infeasible for $n > 2000$ and $p = 6$.

When the sample size is large, we choose an implementation which is tied to the theoretical analysis and test whether moments are zero or non-zero. Specifically, we use empirical likelihood to test the joint hypothesis that $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$. Empirical likelihood is useful as it does not require explicit estimation of the variances of $\gamma_c^{K-1}\gamma_v$ in order to form a well-calibrated test statistic, and the empirical likelihood ratio statistic converges to a known reference distribution under mild conditions. In addition, pooling together all the tests into one omnibus test helps mitigate multiple testing. The empirical likelihood approach is typically less powerful than dHSIC at detecting dependence; however, the computation time required for the test is an order of magnitude smaller. When testing whether moments are zero or non-zero, a specific value of K must be selected. This should correspond to a moment of the errors which is not consistent with the Gaussian distribution. If the data is skewed, $K = 3$ could suffice since the third moment of the Gaussian is zero, but the third moment of the data is non-zero. If the data is either heavy or light tailed relative to the Gaussian (regardless of whether the data is skewed or symmetric), $K = 4$ should suffice. One can also combine the results of multiple values and test $\mathbb{E}(\gamma_c^{k-1}\gamma_v) = 0$ for all $k = \{3, \dots, K\}$ for some arbitrarily large K, but in practice, using larger values of K requires more samples for accurate estimation and testing.

Given an oracle independence test, if the in-degree of each node (counting both directed and bidirected edges) is bounded by some constant s, then the total number of tests required is bounded by a polynomial of the number of variables. Again, let σ be a topological ordering of the nodes consistent with G. As shown in the proof of Theorem 10, at step z, once all the ancestors of $\sigma^{-1}([z-1])$ have been identified, then we need only test sets C up to size $|\text{pa}(\sigma^{-1}(z))| \leq s$ to certify the parents of $\sigma^{-1}(z)$ and subsequently update D. Thus, l, the size of sets considered, will never exceed s. In between any update to D, for each node there will no more than $\sum_{k=1}^{s} \binom{p}{k} \leq p^s$ sets considered. In addition, each time l is incremented, for each node we screen no more than $p - 1$ potential siblings using Algorithm 2. By the acyclicity assumption, there are at most $p(p-1)/2 < p^2$ ancestral relationships to discover, which would cause an update to D. Thus, to fully discover B, there will be no more than $p^2 \times p(p^s + sp)$ independence tests. Once D is fully updated so that $D = B$, then $\widehat{\mathrm{sib}}(v) = \mathrm{sib}(v)$ for all $v \in V$. So that for each $v \in V$ there will be at most an additional cycle through all sets with size less than s which will again result in $p(p^s + sp)$ additional tests. We conclude that Algorithm 4 will check at most $p(p-1)/2$ discovered ancestral relationships. Thus, there must be less than $O(p^{s+3})$ total independence tests. Practically speaking, the computational effort for calculating the residuals is small, and indeed the computational expense for each step is primarily due to the independence test.

Finally, we note that after a graph has been estimated, the resulting D and empirical covariance of γ could be used as estimates for B and Ω. Alternatively, the empirical likelihood procedure of Wang and Drton (2017) could be used for both point estimates and simultaneous confidence intervals.

5. Discovery in the presence of bows

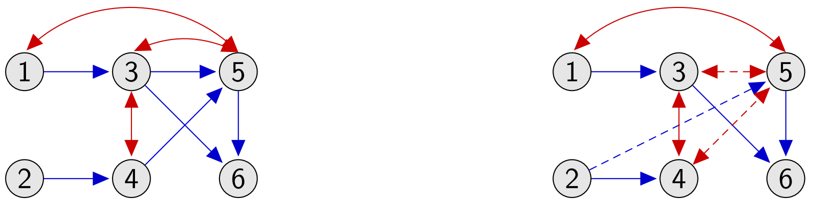
In Section 4, we show that BANG recovers the true graph when the data is generated by a linear SEM corresponding to a BAP. It is unclear, however, how to test this assumption directly in practice. Hence, it is interesting to study what BANG will output if the true data generating process is a linear SEM which corresponds to an acyclic mixed graph $G = (V, E_\to, E_\leftrightarrow)$ which is not bow-free.

In this section, we show that, given population values, BANG will return a BAP $\bar{G} = \left(V, \bar{E}_\to, \bar{E}_\leftrightarrow\right)$ which we subsequently define. We use $\bar{\mathrm{pa}}(v)$ and $\bar{\mathrm{sib}}(v)$ to denote the parents and siblings of node v in $\bar{G}$. Although $\bar{G}$ can be quite different from G, certain key properties important for interpretation are preserved. In particular, $\bar{\mathrm{pa}}(v) \subseteq \mathrm{an}(v)$, so that any directed edge in the output is not in the opposite direction of the true effect. Thus, roughly speaking, in the presence of bows, the procedure is sound—though potentially not complete—for identifying ancestral relations. In addition, any member of $\bar{\mathrm{sib}}(v)$ must be connected to v by a bi-directed path in G, so in the presence of bows, the procedure is complete—though potentially not sound—for identifying confounded relationships. However, roughly speaking, a bidirected edge indicates less certainty about a causal relationship than a directed or absent edge, so the BANG procedure can be considered "conservative" in the sense of not being overconfident when positing causal relationships. In this sense, it is similar to some existing procedures; e.g., RCD (Maeda and Shimizu, 2020). In Example 2, we show a graph with bows and its corresponding "projection" $\bar{G}$.

For $v \in V$, we recursively define a set of nodes which we will call the *irremovable* nodes. We denote this set by $\mathrm{irr}(v)$. Roughly speaking, $\mathrm{irr}(v)$ contains all nodes, whose total effect will never be fully removed from v by the BANG procedure.

Definition 12 *Let $v \in V$, and define $\mathrm{Irr}(v)_0 = \{v\}$. For $k = 1, \ldots, p$, we define recursively $\mathrm{Irr}(v)_k = [\mathrm{pa}(\mathrm{Irr}(v)_{k-1}) \cap \mathrm{sib}(\mathrm{Irr}(v)_{k-1})] \cup \mathrm{Irr}(v)_{k-1}$. Then the set of irremovable nodes is defined as $\mathrm{irr}(v) = \mathrm{Irr}(v)_p$.*

Every $w \in \mathrm{irr}(v)$ is connected to v by both a directed path which only passes through $\mathrm{irr}(v)$ as well as a path of bidirected edges which also only passes through nodes which are in $\mathrm{irr}(v)$. In addition, $\mathrm{irr}(v) \subseteq \mathrm{sib}(\mathrm{irr}(v))$ and $\mathrm{Irr}(v)_1$ contains all nodes which form a bow which ends at v. Given $\mathrm{irr}(v)$ we now define the bidirected edges in $\bar{G}$.



(a) True model, G.

(b) Bow-free graph, $\bar{G}$.

Figure 4: Graphs used in Example 2. The "projection" of (a) into a BAP is shown in (b) where dotted lines indicate edges that differ from the "true model."

Definition 13 *The siblings of a node $v \in V$ in $\bar{G}$ are*

$$\bar{\operatorname{sib}}(v) = \{u : \operatorname{sib}(\operatorname{irr}(v)) \cap \operatorname{irr}(u) \neq \emptyset\} \setminus \{v\}. \tag{18}$$

In other words, the elements of $\bar{\operatorname{sib}}(v)$ are all nodes which are irremovable from v, the siblings of those nodes, and any other nodes whose irremovable nodes have a sibling in $\operatorname{irr}(v)$.

Definition 14 *The parents of a node $v \in V$ in $\bar{G}$ are*

$$\bar{\operatorname{pa}}(v) = \operatorname{pa}(\operatorname{irr}(v)) \setminus \bar{\operatorname{sib}}(v). \tag{19}$$

Note that by definition $\bar{\operatorname{pa}}(v) \cap \bar{\operatorname{sib}}(v) = \emptyset$, which prevents bows.

Since every $u \in \operatorname{irr}(v)$ has a directed path to v which only passes through $\operatorname{irr}(v)$, Definition 14 implies that for every $u \in \bar{\operatorname{pa}}(v)$, there exists a path $l = u \to s_1 \to \ldots \to s_{|l|-2} \to v$ such that $\{s_1, \ldots, s_{|l|-2}\} \subseteq \operatorname{irr}(v)$. By construction, every path from $w \in V$ to v either is entirely contained in $\operatorname{irr}(v)$ or passes through $\bar{\operatorname{pa}}(v)$.

Definition 15 *Let $G = \{V, E_{\to}, E_{\leftrightarrow}\}$ be an acyclic directed mixed graph. Let $\bar{E}_{\to}$ and $\bar{E}_{\leftrightarrow}$ be given by Definitions 13 and 14. We term $\bar{G} = \{V, \bar{E}_{\to}, \bar{E}_{\leftrightarrow}\}$ the BAP projection of G.*

Example 2 *Consider the graph G from panel (a) of Figure 4. There is a bow between 3 and 5. Thus, $\operatorname{Irr}(5)_1 = \{3\} \cup \{5\}$.*

Furthermore, $1 \in \operatorname{pa}(3) \cap \operatorname{sib}(5) \subseteq \operatorname{pa}(\operatorname{Irr}(5)_1) \cap \operatorname{sib}(\operatorname{Irr}(5)_1)$; thus, $1 \in \operatorname{Irr}(5)_2$. Similarly, $4 \in \operatorname{pa}(5) \cap \operatorname{sib}(3) \subseteq \operatorname{pa}(\operatorname{Irr}(5)_1) \cap \operatorname{sib}(\operatorname{Irr}(5)_1)$; thus $4 \in \operatorname{Irr}(5)_2$. No other nodes are in $\operatorname{pa}(\{3, 5\})$ so $\operatorname{Irr}(5)_2 = \{1, 4\} \cup \{3, 5\}$.

Next, note that $2 \in \operatorname{pa}(4)$ but $2 \notin \operatorname{sib}(\operatorname{Irr}(5)_2)$ so it is not a member of $\operatorname{Irr}(5)_3$. Also, $6 \notin \operatorname{pa}(\operatorname{Irr}(5)_2)$, so it is also not a member of $\operatorname{Irr}(5)_3$. Thus, for $s = 3, \ldots, 6$, $\operatorname{Irr}(5)_s = \operatorname{Irr}(5)_2 = \{1, 3, 4, 5\}$ and $\operatorname{irr}(5) = \{1, 3, 4, 5\}$.

There are no other bows in the graph, so $\operatorname{irr}(v) = \{v\}$ for all other nodes and $\bar{\operatorname{sib}}(5) = \{1, 3, 4\}$. As $2 \in \operatorname{pa}(\operatorname{irr}(5)) \setminus \bar{\operatorname{sib}}(5)$, we have $2 \in \bar{\operatorname{pa}}(5)$.

As we will show, when applied to population values corresponding to the graph G in (a), BANG will recover the BAP projection $\bar{G}$ displayed in panel (b). The dotted lines in (b) indicate edges in $\bar{G}$ which are different from the edges in G.

To develop our result on estimation of $\bar{G}$, we first show that the distribution of Y is also in the model implied by BAP $\bar{G}$.

Lemma 16 *Let G be an acyclic directed mixed graph, and suppose the random vector Y follows a distribution in the linear SEM given by G. Let $\bar{G} = \big(V, \bar{E}_{\to}, \bar{E}_{\leftrightarrow}\big)$ be the projection of G given by Definition 15. Then $\bar{G}$ is a BAP. Furthermore, the distribution of Y is equal to the distribution in the linear SEM given by $\bar{G}$ when defining edge weights and errors as follows:*

$$\begin{aligned}
\bar{\beta}_{v,u} &= \begin{cases} \sum_{l \in \mathcal{L}^{(u)}_{v,u}(\bar{\mathrm{pa}}(v))} W(l) & \text{if } u \in \bar{\mathrm{pa}}(v), \\ 0 & \text{else,} \end{cases} \\
\bar{\varepsilon}_v &= \varepsilon_v + \sum_{w \in \mathrm{irr}(v)} \varepsilon_w \sum_{\substack{l \in \mathcal{L}_{v,w} \\ l \subseteq \mathrm{irr}(v)}} W(l).
\end{aligned} \tag{20}$$

Proof Acyclicity of G implies acyclicity of the projection $\bar{G}$ because $\bar{\mathrm{pa}}(v) \subseteq \mathrm{an}(v)$. Furthermore, by definition, $\bar{\mathrm{pa}}(v)$ does not include $\bar{\mathrm{sib}}(v)$, so $\bar{G}$ does not contain any bows. Thus, $\bar{G}$ is a BAP.

We now show that the distribution of Y belongs to the SEM given by $\bar{G}$. Recall the definition of $\mathcal{L}^{(w)}_{v,w}(\mathrm{an}(v))$ from Section 3.2. Note that $\mathcal{L}^{(w)}_{v,w}(\mathrm{an}(v)) = \emptyset$ for any $w \notin \bar{\mathrm{pa}}(v) \cup \bar{\mathrm{sib}}(v)$ because if $w \notin \bar{\mathrm{sib}}(v)$ and there was a path from w to v for which w was the last node, then w would be in $\bar{\mathrm{pa}}(v)$. Hence,

$$\begin{aligned}
Y_v &= \varepsilon_v + \sum_{w \in \mathrm{an}(v)} \pi_{v,w} \varepsilon_w = \varepsilon_v + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \sum_{l \in \mathcal{L}_{v,w}} W(l) \\
&= \varepsilon_v + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \left(\sum_{\substack{l \in \mathcal{L}_{v,w} \\ \bar{\mathrm{pa}}(v) \cap l \neq \emptyset}} W(l) + \sum_{\substack{l \in \mathcal{L}_{v,w} \\ \bar{\mathrm{pa}}(v) \cap l = \emptyset}} W(l) \right) \\
&= \varepsilon_v + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \left(\sum_{s \in \bar{\mathrm{pa}}(v)} \sum_{l \in \mathcal{L}^{(s)}_{v,w}(\bar{\mathrm{pa}}(v))} W(l) + \sum_{\substack{l \in \mathcal{L}_{v,w} \\ l \subseteq \mathrm{irr}(v)}} W(l) \right) \\
&= \varepsilon_v + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \left(\sum_{s \in \bar{\mathrm{pa}}(v)} \pi_{s,w} \sum_{l \in \mathcal{L}^{(s)}_{v,s}(\bar{\mathrm{pa}}(v))} W(l) + \sum_{\substack{l \in \mathcal{L}_{v,w} \\ l \subseteq \mathrm{irr}(v)}} W(l) \right) \\
&= \varepsilon_v + \sum_{s \in \bar{\mathrm{pa}}(v)} \sum_{w \in \mathrm{an}(v)} \pi_{s,w} \varepsilon_w \sum_{l \in \mathcal{L}^{(s)}_{v,s}(\bar{\mathrm{pa}}(v))} W(l) + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \sum_{\substack{l \in \mathcal{L}_{v,w} \\ l \subseteq \mathrm{irr}(v)}} W(l).
\end{aligned} \tag{21}$$

Because $\mathrm{an}(s) \subseteq \mathrm{an}(v)$ if $s \in \bar{\mathrm{pa}}(v) \subseteq \mathrm{an}(v)$, we have that

$$
\begin{aligned}
(21) = \varepsilon_v + \sum_{s \in \bar{\mathrm{pa}}(v)} Y_s \sum_{l \in \mathcal{L}_{v,s}^{(s)}(\bar{\mathrm{pa}}(v))} W(l) + \sum_{w \in \mathrm{an}(v)} \varepsilon_w \sum_{\substack{l \in \mathcal{L}_{v,w} \\ l \subseteq \mathrm{irr}(v)}} W(l) \\
= \sum_{s \in \bar{\mathrm{pa}}(v)} Y_s \underbrace{\sum_{l \in \mathcal{L}_{v,s}^{(s)}(\bar{\mathrm{pa}}(v))} W(l)}_{\bar{\beta}_{v,s}} + \underbrace{\varepsilon_v + \sum_{w \in \mathrm{irr}(v)} \varepsilon_w \sum_{\substack{l \in \mathcal{L}_{v,w} \\ \bar{l} \subseteq \mathrm{irr}(v)}} W(l)}_{\bar{\varepsilon}_v} .
\end{aligned} \tag{22}
$$

Our claim now follows because the coefficients $\bar{\beta}_{v,u}$ respect the constraints given by $\bar{G}$. Indeed, $\bar{\beta}_{v,u} = 0$ if $u \notin \bar{\mathrm{pa}}(v)$. Furthermore, if $\mathrm{irr}(u) \cap \mathrm{sib}(\mathrm{irr}(v)) = \emptyset$, then $\bar{\varepsilon}_v$ only contains terms which are independent of the terms in $\bar{\varepsilon}_u$. Thus, $\varepsilon_v \perp\!\!\!\perp \varepsilon_u$ if $u \notin \bar{\mathrm{sib}}(v)$. ■

Though, as we subsequently show, it is true that given population values BANG returns $\bar{G}$, Lemma 16 does not immediately imply that BANG will discover $\bar{G}$; it simply implies that the distributions implied by G are a subset of distributions implied by $\bar{G}$. It must further be shown that for generic parameters (of the full model), the distribution of Y does not lie in any sub-model of $\bar{G}$.

Replacing B with $\bar{B}$, Lemma 3, Corollary 4, Lemma 5, and Lemma 6 still directly hold in this setting. We restate these results below for completeness, but note that the proofs follow in analogy to the previously proved lemmas. Thus, to prove an analogue to Theorem 10, it remains to show analogues to Lemma 7, Lemma 8, and Corollary 9.

Corollary 17 (Analogue of Lemma 3 and Corollary 4) *Suppose Y is generated by a linear SEM which corresponds to a acyclic mixed graph G. Let $\bar{G}$ be the projection of G and $\bar{B}$ be the corresponding direct effects defined in* (20). *For a node $v \in V$ and a set $C \subseteq A \subseteq V \setminus v$, suppose:*

1. $\bar{\mathrm{pa}}(v) \subseteq C \subseteq \bar{\mathrm{an}}(v) \setminus \bar{\mathrm{sib}}(v)$
2. $A = \bar{\mathrm{An}}(C)$
3. $D_{A,A} = \bar{B}_{A,A}$
4. $S_{\{A,v\},\{A,v\}} = \Sigma_{\{A,v\},\{A,v\}}$

Then $\delta_v(C, A, S, D) = \bar{B}_{v,C}$. Furthermore, for every $c \in C$, $\gamma_c(D) \perp\!\!\!\perp \gamma_v(C, S, D)$ and $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D))$.

Proof By Lemma 16, the distribution of Y is equivalent to the BAP defined by $\bar{G}$, $\bar{B}$ and $\bar{\varepsilon}$. Thus, the result directly follows by applying Lemma 3 and Corollary 4 to $\bar{G}$, $\bar{B}$ and $\bar{\varepsilon}$. ■

In the following statements, we will at times make statements about sets of nodes in $\bar{G}$; however, the requirement of generic parameters will always refer to parameters in the model which may have bows defined by G. Let $\breve{B}(C \cup \{v\})_{v,C}$ be the marginal direct effect of $C \cup \{v\}$ in $\bar{G}$; i.e., the analogue of $\tilde{B}$, as defined in (6), but using $\bar{B}$ instead of B.

Corollary 18 (Analogue of Lemma 5) *Let $v \in V$, and consider any set $C \subseteq A \subseteq V \setminus \{v\}$. Suppose $D \in \mathbb{R}^{p \times p}$ with $D_{s,t} \neq 0$ only if $t \in \bar{\mathrm{an}}(s)$. Then, for generic B and error moments, if $\delta_v(C, A, S, D) \neq \check{B}(C \cup v)_{v,C}$, then $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D)) \neq 0$ for some $c \in C$.*

Corollary 19 (Analogue of Lemma 6) *Consider $v \in V$ and set $C \subseteq V \setminus \{v\}$. Let $D \in \mathbb{R}^{p \times p}$ such that $D_{s,t} \neq 0$ only if $t \in \bar{\mathrm{an}}(s)$. Suppose $C \not\subseteq \bar{\mathrm{an}}(v)$, but $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, S, D)) = 0$ for all $c \in C$. Then for generic B and error moments, $C_1 = C \cap \big[\bar{\mathrm{an}}(v) \setminus \bar{\mathrm{sib}}(v)\big]$,*

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1, S, D)) = 0$$

for all $c \in C$.

Lemma 20 (Analogue of Lemma 7) *Suppose $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \bar{\mathrm{an}}(s) \setminus \bar{\mathrm{sib}}(s)$ for all $s \in V$. Let $v \in V$ be such that we have $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(D)) = 0$ for all $c \in \mathrm{pa}_D(v)$. If $q \in \big(\bar{\mathrm{pa}}(v) \setminus \mathrm{pa}_D(v)\big) \cup \bar{\mathrm{sib}}(v)$, then for generic B and error moments, $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$.*

Lemma 21 (Analogue of Lemma 8) *Consider $v \in V$ and sets A, C such that $C \subseteq A \subseteq V \setminus \{v\}$. Suppose $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \bar{\mathrm{an}}(s)\bar{\mathrm{sib}}(s)$ for all $s \in V$. Suppose $u \in C$ and $u \in \bar{\mathrm{sib}}(v)$, then for generic B and error moments, there exists some $q \in C$ such that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0$.*

Corollary 22 (Analogue of Corollary 9) *Suppose $D = \bar{B}$. Then, for $v \in V$ and generic B and error moments, suppose $\bar{\mathrm{pa}}(v) \subseteq C \subseteq \bar{\mathrm{an}}(v) \setminus \bar{\mathrm{sib}}(v)$ and $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)) = 0$ for all $c \in C$. If $q \in C \setminus \bar{\mathrm{pa}}(v)$, the for all $c \in C$*

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = 0. \tag{23}$$

If $q \in \mathrm{pa}(v)$, then there exists some $c \in C$ such that

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) \neq 0. \tag{24}$$

Lemma 20 and 21 require two intermediate results which we state here for completeness.

Lemma 23 *Let $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \bar{\mathrm{an}}(s) \setminus \bar{\mathrm{sib}}(s)$ for all $s \in V$. Suppose there exists some path $l = s_1 \to s_2 \to \dots s_{|l|-1} \to v$ such that $l \cap C_v = \emptyset$. Further suppose that $u \in \mathrm{sib}(s_1) \setminus l$ and $u \notin C_v$. Then for generic parameters*

$$\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_{s_1}(D)\right) \neq 0 \qquad \text{and} \qquad \mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0, \tag{25}$$

so that s_1 and u will not be pruned from $\widehat{\mathrm{sib}}(v)$ by Alg 2. Furthermore, for any C such that $u \in C$, for generic parameters there exists some $c \in C$ such that

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0, \tag{26}$$

so that u will not be certified into $\widehat{\mathrm{pa}}(v)$.

Lemma 24 *Suppose* $\mathrm{irr}(u) \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$ *and* $D = H_{\mathcal{C}}(\bar{B})$ *for some* $\mathcal{C} = (C_s)_{s \in V}$ *such that* $C_s \subseteq \bar{\mathrm{an}}(s) \setminus \bar{\mathrm{sib}}(s)$ *for all* $s \in V$*. Then, for generic parameters*

$$\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0 \tag{27}$$

so that u *will not be pruned from* $\widehat{\mathrm{sib}}(v)$ *by Alg 2. Furthermore, for any* $C \subseteq V \setminus v$ *such that* $u \in C$*, for generic parameters, there exists some* $c \in C$ *such that*

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0, \tag{28}$$

so that u *will not be certified into* $\widehat{\mathrm{pa}}(v)$*.*

Theorem 25 *Suppose* Y *is generated under the linear SEM given by an acyclic directed mixed graph* G *with BAP projection* $\bar{G}$ *as defined by* (18) *and* (19)*. Then for generic choices of* B *and error moments, BANG will output* $\hat{G} = \bar{G}$ *when given population values of* Y*.*

Proof The proof exactly mirrors the proof of Theorem 10, but using the lemmas developed for the misspecified case. ■

6. Numerical results

We consider two implementations of BANG[2]: one which uses empirical likelihood to test whether moments are zero or non-zero and another which uses dHSIC (Pfister et al., 2018) to test whether certain variables are independent or not. We compare these implementations against ParcelLiNGAM (Tashiro et al., 2014), RCD (Maeda and Shimizu, 2020), and two methods for Gaussian data—FCI+ (Claassen et al., 2013) with Gaussian conditional independence tests (i.e. vanishing partial correlations) and Greedy BAP Search (GBS) (Nowzohour et al., 2017). For ParcelLiNGAM we use the Matlab implementation available from the author's website[3]; for RCD we use the `lingam` python package[4]; for FCI+, we use the `R` package `pcalg` (Kalisch et al., 2012); and for GBS we use the `R` package `greedyBaps` (Nowzohour, 2017).

Our experiments consider structure learning in two settings: ancestral graphs and BAPs. In each case, we simulate errors from gamma, lognormal, and uniform distributions. In addition, we include a setting with errors drawn from T_{13} as a counter example to show how performance can suffer when the errors are close to Gaussian. Finally, we show that when applied to ecology data the BANG method recovers a model close to the ground truth.

We also note that in the implementation, we include a symbolic version that allows interested readers to track the population steps of the algorithm.

6.1 Comparison with ParcelLiNGAM

ParcelLiNGAM is designed to discover ancestral relationships, not graph structure. Also, as shown in Section 2 it is sound and complete for ancestral graphs, but not non-ancestral

[2] Available at `https://github.com/ysamwang/ngBap`

[3] `https://sites.google.com/site/sshimizu06/Plingamcode`

[4] `https://github.com/cdt15/lingam`

graphs. Thus, to give ParcelLiNGAM the most favorable comparison, we only consider ancestral graphs, and to compare performance, we measure the accuracy in identifying ancestral relationships; i.e., for each (u, v), is $u \in \text{an}(v)$ or instead is $u \not\in \text{an}(v)$? The accuracy is defined as (True Positives + True Negatives)/Total Cases.

We let $p = 6$ and consider three settings with varying levels of sparsity with d directed edges and b bidirected edges: *sparse* ($d = p/2$, $b = p/2$), *medium* ($d = p$, $b = p$), and *dense* ($d = 3p/2$, $b = p$). We let $n = 500, 1000, 1500$.

To generate a random ancestral graphs, we first select d directed edges uniformly from the set $\{(i, j) : i < j\}$, and then select b bidirected edges uniformly from the set $\{(i, j) : i \not\in \text{an}(j) \text{ and } j \not\in \text{an}(i)\}$ when generating ancestral graphs if the set of possible bidirected edges is less than b, we select as many as possible. We then draw the directed edgeweights uniformly from $(.6, 1)$. Note that the graphs are all ancestral, but may not be maximal.

For the idiosyncratic errors, we first form the covariance Ω by drawing $\omega_{ij} = \omega_{ji}$ uniformly from $(.3, .5)$ for all $(i, j) \in E_{\leftrightarrow}$, and setting all other elements to 0. To ensure that Ω is positive definite, we set $\Omega_{ii} = 1 + \sum_{j \neq i} |\omega_{ij}|$. We consider five settings where the errors marginally follow uniform, gamma, lognormal, and T with $d.f. = 13$. We draw the gamma errors using `lcmix` (Dvorkin, 2012), uniform using `MultiRNG` (Demirtas et al., 2019), T_{13} by using univariate copulas to transform a multivariate normal with covariance Ω, and the lognormal errors are formed by exponentiating multivariate normal draws with covariance Ω.

We then set $Y^{(i)} = (I - B)^{-1}\epsilon^{(i)}$. Finally, because the output generally depends on the ordering of the variables in the data matrix, we also randomly permute the labeling of the variables so that $1, \ldots, p$ is generally not a valid ordering. This entire process is repeated 200 times for each setting.

The results are given in Table 1. For BANG with dHSIC or empirical likelihood, the value is bolded if the accuracy is significantly larger than the ParcelLiNGAM accuracy (measured using a two-sample paired T-test with with $\alpha = .05$). For ParcelLiNGAM, the value is bolded if the accuracy is significantly larger than both of the BANG implementations. In general, we see that ParcelLiNGAM tends to do better when the graph is dense; this is particularly drastic when the errors are uniform, but less so for the other error types. Under the "medium" and "sparse" graph regimes, BANG tends to outperform ParcelLiNGAM, particularly the dHSIC implementation. In the sparse regime, the difference is quite drastic for all error settings. In general, when the errors are T and not too far from Gaussian, all three methods suffer.

6.2 Comparison with FCI+, GBS, and RCD

We now compare BANG against FCI+ and GBS, which both identify an equivalence class of graphs as well as RCD which selects a unique graph. We compare FCI+, GBS, and BANG on ancestral graphs generated as described in the previous section. Furthermore, we compare GBS, RCD and BANG on possibly non-ancestral BAPs. The graphs are generated by the same procedure, except when generating BAPs we do not enforce the ancestral condition and instead draw bidirected edges from the set $\{(i, j) : i \not\in \text{pa}(j) \text{ and } j \not\in \text{pa}(i)\}$. We let $n = 2500, 5000, 7500, 10000, 25000, 50000$, but for computational reasons we only use RCD and the dHSIC implementation of BANG for $n \leq 7500$.

Regime	dist	n	DH	EL	PL
Dense	Gamma	500	0.887	0.892	0.893
		1000	0.921	0.907	**0.928**
		1500	0.925	0.908	**0.937**
	Lognormal	500	0.886	0.865	0.889
		1000	0.880	0.868	0.887
		1500	0.892	0.873	**0.902**
	T	500	0.510	0.524	0.550
		1000	0.479	0.504	**0.597**
		1500	0.508	0.523	**0.643**
	Unif	500	0.732	0.679	**0.848**
		1000	0.786	0.713	**0.859**
		1500	0.809	0.718	**0.865**
Medium	Gamma	500	**0.867**	**0.899**	0.808
		1000	**0.911**	0.861	0.893
		1500	**0.951**	0.884	0.922
	Lognormal	500	**0.914**	**0.885**	0.860
		1000	**0.915**	0.872	0.888
		1500	**0.922**	0.880	0.909
	T	500	**0.587**	0.563	0.540
		1000	0.548	0.558	0.582
		1500	0.532	0.552	**0.606**
	Unif	500	0.761	0.753	**0.784**
		1000	0.817	0.819	0.808
		1500	0.839	0.833	0.829
Sparse	Gamma	500	**0.930**	**0.913**	0.664
		1000	**0.960**	**0.930**	0.697
		1500	**0.979**	**0.959**	0.710
	Lognormal	500	**0.965**	**0.902**	0.686
		1000	**0.965**	**0.893**	0.708
		1500	**0.970**	**0.895**	0.715
	T	500	**0.747**	**0.695**	0.520
		1000	**0.707**	**0.706**	0.530
		1500	**0.703**	**0.696**	0.550
	Unif	500	**0.869**	**0.897**	0.646
		1000	**0.911**	**0.943**	0.675
		1500	**0.918**	**0.951**	0.685

Table 1: The average accuracy of each procedure in identifying ancestral relationships across 200 replications. DH: BANG using dHSIC independence tests, EL: BANG using empirical likelihood moment tests, PL: ParcelLiNGAM. All procedures use independence tests with $\alpha = .01$.

To compare performance, we record the proportion of the times each method recovers the equivalence class corresponding to the true graph (or a graph in the equivalence class of the true graph). For BANG and RCD, we also record the proportion of times it recovers the exact graph. We also show the structural Hamming distance of the estimated graph to the true graph (i.e., the number of edges which would would need to be added/deleted/modified to transform the estimated graph into the true graph). For FCI+, RCD, and BANG we set the nominal level of each hypothesis test performed to $\alpha = .05, .01, .001$. In the plots, we show the results for the best performing α in each setting. For GBS, we allow 100 random restarts, the same number used in the simulations by Nowzohour (2017). For BANG with EL, we set $K = 3$ for the gamma and lognormal errors (since they are skewed) and let $K = 4$ for the uniform and T_{13} (since they are symmetric). In the simulations with ancestral graphs, to check whether the equivalence class is recovered, we take the graph estimated by BANG, RCD, and GBS and project it into a PAG. In the simulations with BAPs, since there is no known graphical characterization for the equivalence class of BAPs, we follow Nowzohour (2017) and say that the estimated and true graph are in the same equivalence class if the score of the estimated graph is within 10^{-10} of score of the true graph. We repeat the experiment 200 times for each simulation setting. The results for ancestral graphs are shown in Figures 5 and 6, and the results for BAPs are shown in Figures 7 and 8.

When only considering ancestral graphs in Figures 5 and 6, there are a number of settings—particularly in the medium and dense regime—in which BANG and RCD are able to recover the exact graph more often than GBS and FCI recover the equivalence class. In most cases, the dHSIC implementation of BANG outperforms the EL implementation. Performance of RCD and BANG is generally comparable, although BANG seems to slightly outperform RCD when the errors are Gamma and Uniform. As expected, BANG and RCD both perform poorly when the errors are drawn from a T distributions, though RCD slightly outperforms BANG.

When considering BAPs, we allow the graph to be non-ancestral, but do not explicitly enforce that the graph is non-ancestral. Thus, under the generating procedure, the random graph is still ancestral with probability $.87, .27, .06$ under the sparse, medium, and dense regimes respectively. As shown in Figures 7 and 8, in the sparse setting, although BANG-dHSIC generally outperforms RCD, RCD still shows decent performance because the vast majority of graphs are still ancestral.

In the medium and dense settings, when the graph is more likely to be non-ancestral, we see that BANG clearly outperforms RCD in recovering the true graph when the errors are Gamma, Lognormal and Uniform. As expected, both RCD and BANG perform poorly when the errors are drawn from a T distribution, though in contrast to the other settings, it seems that RCD slightly outperforms BANG.s GBS tends to perform well in the sparse setting; in particular it does well with the T errors when all other procedures do poorly. However, it performs poorly in the medium or dense setting across all error types.

6.3 Computational feasibility of larger problems

To demonstrate the feasibility on larger problems, we apply BANG to problems where $p = 15$ and $n = 5000, 10000, 25000, 50000$. We draw random BAPS and SEM parameters as before and consider graphs with $d = 25$ directed edges and $b = 20$ bidirected edges. We draw the errors from a gamma distribution and use empirical likelihood tests with a nominal level of $\alpha = .05, .01, .001$. Table 2 records the results of 50 replications at each setting. We see that BANG performs well statistically when n is large, and at the largest sample size is able to recover the exact graph over half the time. Moreover, the procedure is computationally feasible even with the largest sample size $n = 50000$. We show timing

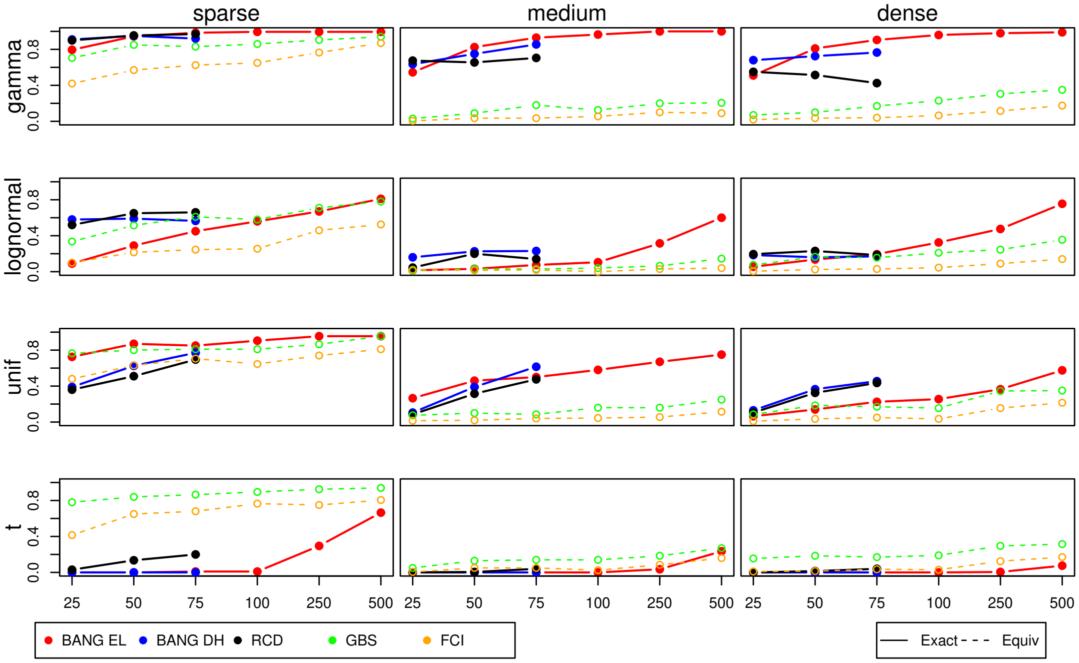


Figure 5: The performance of each method on random ancestral graphs with $p = 6$ across 200 replications. The dotted lines indicate the proportion of times the estimated graph corresponds to the PAG of the true graph; the solid lines indicate the proportion of times BANG or RCD identifies the exact graph. The horizontal axis shows sample size in hundreds.

results for a single cpu, however, this could be further improved since many of the required tests can be performed in parallel.

6.4 Data example

Grace et al. (2016) use a structural equation model to examine the relationships between land productivity and the richness of plant diversity. They consider measurements taken at 1126 plots which are locations across 39 different sites. A graphical model from the original

	SHD			Exact Recov.			Time		
n	$\alpha = .05$	.01	.001	.05	.01	.001	.05	.01	.001
5000	**25.0**	**25.0**	28.0	0.00	0.00	0.00	174.3	98.1	**58.6**
10000	**14.5**	19.0	24.5	0.04	**0.06**	0.02	380.3	312.8	**241.7**
25000	**10.5**	12.0	11.0	**0.14**	**0.14**	0.12	2004.8	1881.5	**1490.3**
50000	3.0	0.5	**0.0**	0.26	0.50	**0.52**	6783.4	**3691.5**	4237.1

Table 2: The results of random BAPs with $p = 15$. The SHD columns show the median structural Hamming distance; the Exact Recov. columns show the proportion of the time the exact graph is recovered, and the Time columns show the computational time in seconds.

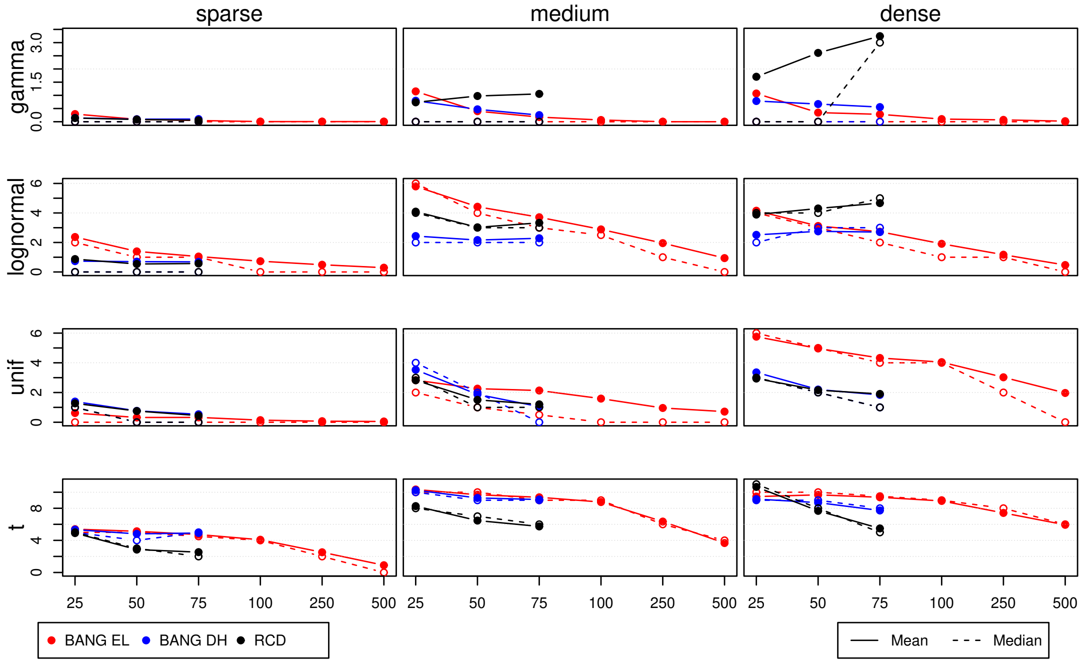


Figure 6: The performance of each method on random ancestral graphs with $p = 6$ across 200 replications measured by structural Hamming distance. The solid line indicates the average, the dotted line indicates the median. The horizontal axis shows sample size in hundreds.

paper is in Figure 9. We consider their plot level model which includes: plot productivity, plot biomass, plot shade, plot richness, plot soil suitability, site richness, site biomass, site productivity.

Beginning with the graphical model shown in Figure 9, we first remove any edges which they found were not significant (denoted by NS in Figure 9). Note that this removes the cycle in the plot specific measurements, but there is still a cycle between site productivity, biomass and richness. The nodes for climate, disturbance and suitability, actually represent multiple variables which are used in the SEMs. For climate and disturbance, the separate measures are both highly correlated, so it seems reasonable to use bidirected edges between site productivity, biomass and richness when marginalizing out those variables, despite the fact that they are actually separate measures. To keep the bow-free assumption, we do not include the directed edges between site productivity, site biomass and site richness. This results in ancestral relationships in the full model which are not otherwise captured in the marginalized model. Thus, we add directed edges from site productivity to plot biomass and plot richness; from site biomass to plot productivity and plot richness; from site richness to plot productivity and plot biomass. For suitability, there is both a site suitability, which is a parent of site richness, and a plot suitability which is a parent of plot richness. Although there is no explicit specification in their SEM of how site suitability relates into plot suitability, it seems reasonable to assume that site suitability has a direct effect on plot suitability, as is the case for all other site vs plot measures. Thus, we include

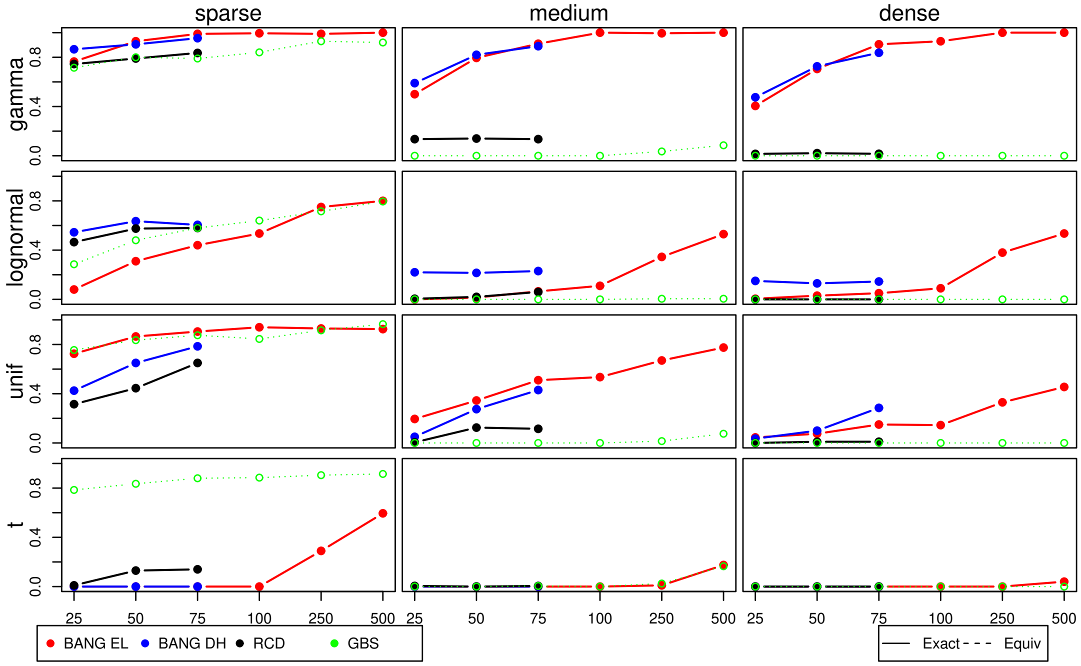


Figure 7: The performance of each method on random bow-free acyclic graphs with $p = 6$ across 200 replications. The dotted lines indicate the proportion of times the estimated graph is in the equivalence class of the truth; the solid lines indicate the proportion of times BANG or RCD identifies the exact graph. The horizontal axis shows sample size in hundreds.

a bidirected edge between plot suitability and site richness. This results in the BAP shown in Figure 10a. We consider this model the ground truth.

For BANG, we use empirical likelihood and selected the nominal test level, .01, so that there are roughly the same number of directed edges in the estimated and ground truth graphs, 11 and 13 respectively. The discovered graph is shown in Figure 10b. Of the 28 pairs of nodes, BANG correctly identifies the correct relation ($\rightarrow, \leftarrow, \leftrightarrow$ or no edge) for 16 of the pairs. Naively, letting the probability of guessing each relationship to be $1/4$, this results in a binomial probability of $P(X \geq 16) = .00029$. This probability does not account for the dependency between edges since there is an acyclic restriction, but it suggests that BANG is discovering reasonable structure. There are 7 bidirected edges in the estimated graph compared to 4 in the ground truth model. This behavior is somewhat expected since there is still likely to be uncontrolled confounding which is either not actually fully accounted for in the ground truth model or direct causes which cannot be fully explained by a linear relationship. For comparison, we also use the GBS procedure with 500 random restarts. In Figure 11, we plot the resulting score against the number of correct edges for each of the 500 runs. There seems to be a positive association between the score and the correct number of edges. Although one initialization resulted in a graph with 16 correct edges it did not have the highest score, and each of the resulting estimated graphs with maximum score (up to optimization error) only has 12 correct edges.

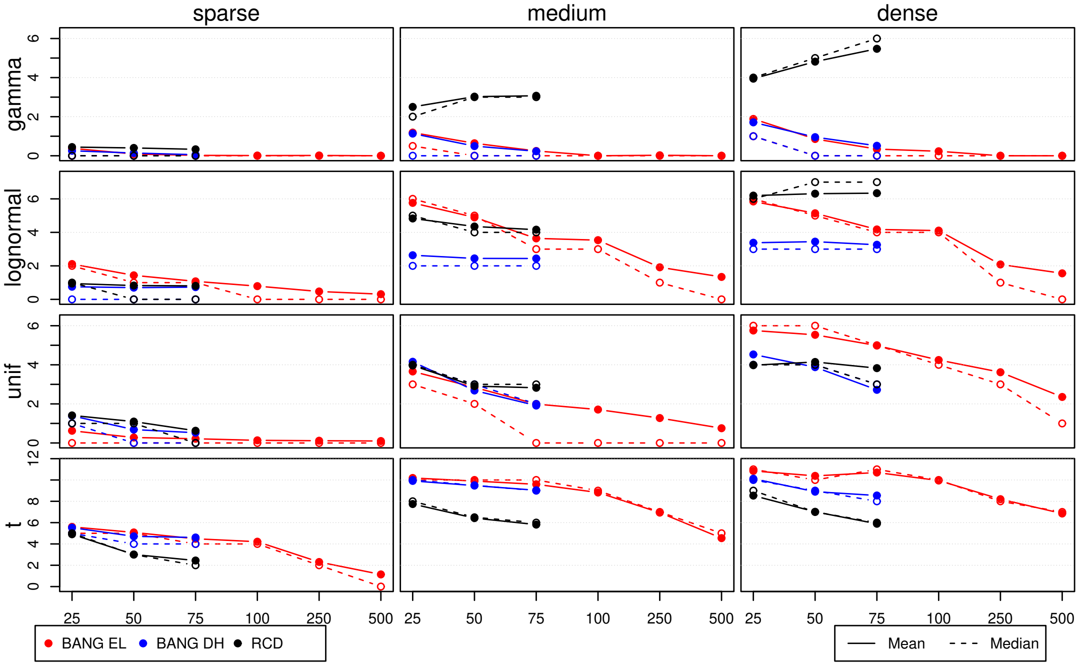


Figure 8: The performance of each method on random bow-free acyclic graphs with $p = 6$ across 200 replications as measured by structural Hamming distance. The solid line indicates the average, the dotted line indicates the median. The horizontal axis shows sample size in hundreds.

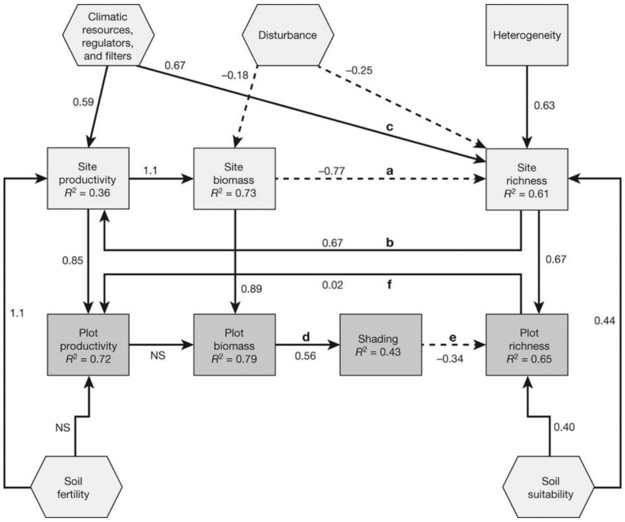
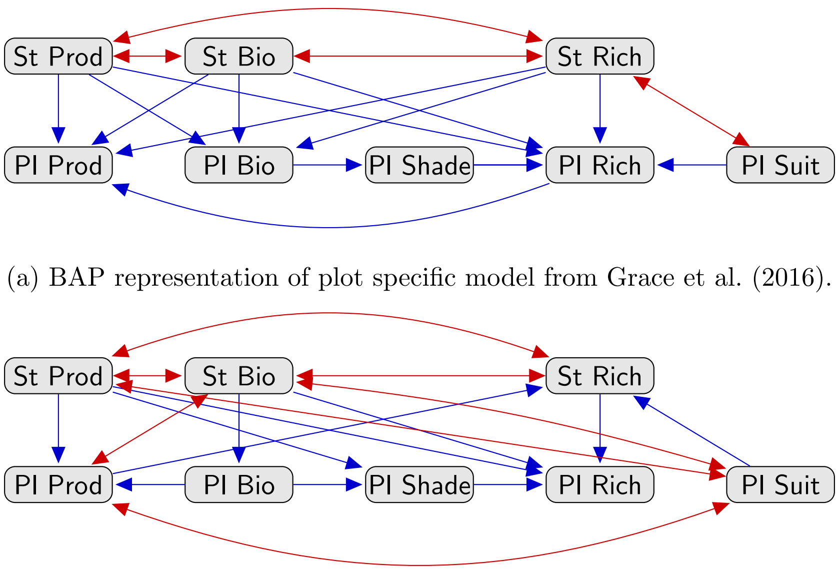

Climatic resources, regulators, and filters
Disturbance
Heterogeneity
0.67
−0.18
−0.25
0.59
0.63
c
Site productivity $R^2 = 0.36$
1.1
Site biomass $R^2 = 0.73$
−0.77
a
Site richness $R^2 = 0.61$
0.67
b
0.85
0.89
0.02
f
0.67
1.1
0.44
Plot productivity $R^2 = 0.72$
NS
Plot biomass $R^2 = 0.79$
d
0.56
Shading $R^2 = 0.43$
e
−0.34
Plot richness $R^2 = 0.65$
NS
0.40
Soil fertility
Soil suitability

Figure 9: Full model from Grace et al. (2016).

7. Discussion

Borrowing intuition from the LiNGAM line of work (Shimizu et al., 2006), we show that when a SEM corresponds to a BAP and the errors are non-Gaussian, one can identify the



(a) BAP representation of plot specific model from Grace et al. (2016).

(b) Discovered model (BANG).

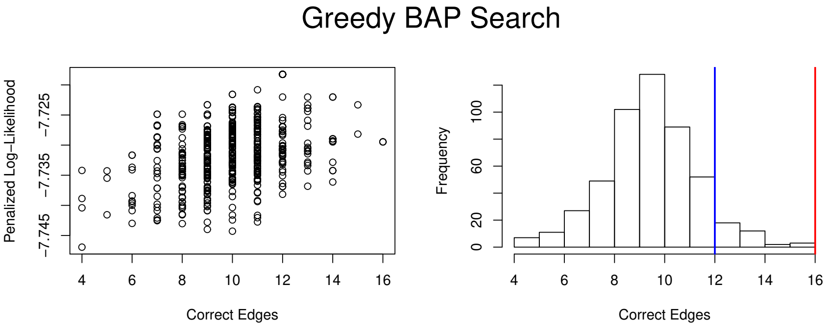


Figure 11: The left panel shows the score and number of correct edges for each of the 500 random initializations of GBS on the Grace et al. (2016) ecology data. The estimated graphs with the highest score has 12 correct edges. The right panel shows a histogram of the number of correct edges from the 500 random restarts. The blue line represents the graph with the highest score and the red line represents the number of correct edges for the BANG procedure.

exact causal structure from observational data. We propose the BANG algorithm and show that it consistently identifies the graph. This extends previous work on BAPs by Nowzohour et al. (2017) by identifying an exact graph rather than a larger equivalence class. In addition, this extends the work on non-Gaussian SEMs with confounding by not requiring advance knowledge of the number of latent variables, not requiring the effect of confounders to be linear, or provably recovering a larger class of graphs. Finally, we also show that in the presence of bows, our proposed procedure is "conservative" in certifying causal relationships and explicitly characterize the returned output in the population setting.

Since the number of independence tests considered is a polynomial of the number of variables, under additional assumptions, future work might investigate conditions under which the graph might also be consistently recovered in a sparse high dimensional setting where the number of variables is larger than the number of samples. Theoretical results may be straightforward; however, considering the results in Section 6 where very large sample sizes are needed for recovery with high probability, this may require significant methodological improvements. One such improvement is a pre-screening procedure. Loh and Bühlmann (2014) show for DAGs, even with non-Gaussian errors, the precision matrix encodes causal structure. A similar statement can be shown for BAPs, where a non-zero entry in the precision implies that two nodes are in the same *mixed component*—roughly a set of nodes which are connected by bidirected edges plus the parents of those nodes; see Tian (2005); Foygel et al. (2012) for a formal definition. Thus, starting with a sparse estimate of the precision could reduce the search space and improve empirical performance.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818) as well as from the U.S. National Science Foundation under Grant No. DMS 1712535. The authors also thank Thomas S. Richardson for helpful comments on an early copy of the manuscript. This research was supported in part through the computational resources and staff contributions provided for the Mercury high performance computing cluster at The University of Chicago Booth School of Business which is supported by the Office of the Dean.

References

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *Ann. Statist.*, 37(5B):2808–2837, 2009.

Wicher Bergsma and Angelos Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028, 2014.

Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4098–4108, Online, 26–28 Aug 2020. PMLR.

Kenneth A. Bollen. *Structural equations with latent variables.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1989. A Wiley-Interscience Publication.

Carlos Brito and Judea Pearl. A new identification condition for recursive models with correlated errors. *Struct. Equ. Model.*, 9(4):459–474, 2002.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Wenyu Chen, Mathias Drton, and Ali Shojaie. Causal structural learning via local graphs, 2021. arXiv:2107.03597.

Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, 40(1):294–321, 2012.

Hakan Demirtas, Rawan Allozi, and Ran Gao. *MultiRNG: Multivariate Pseudo-Random Number Generation*, 2019. R package version 1.2.2.

Mathias Drton. Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77 of *Adv. Stud. Pure Math.*, pages 35–86. Math. Soc. Japan, Tokyo, 2018.

Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *J. Mach. Learn. Res.*, 10:2329–2348, 2009.

Daniel Dvorkin. *lcmix: Layered and chained mixture models*, 2012. R package version 0.3/r5.

Doris Entner and Patrik O. Hoyer. Discovering unconfounded causal relationships using linear non-gaussian models. In Takashi Onada, Daisuke Bekki, and Eric McCready, editors, *New Frontiers in Artificial Intelligence - JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers*, volume 6797 of *Lecture Notes in Computer Science*, pages 181–195. Springer, 2010.

Robin Evans. Markov properties for mixed graphical models. In *Handbook of graphical models*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 39–60. CRC Press, Boca Raton, FL, 2019.

Robin J. Evans. Graphs for margins of Bayesian networks. *Scand. J. Stat.*, 43(3):625–648, 2016.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.*, 40(3):1682–1713, 2012.

James B Grace, T Michael Anderson, Eric W Seabloom, Elizabeth T Borer, Peter B Adler, W Stanley Harpole, Yann Hautier, Helmut Hillebrand, Eric M Lind, Meelis Pärtel, et al. Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, 529(7586):390–393, 2016.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, volume 3734 of *Lecture Notes in Comput. Sci.*, pages 63–77. Springer, Berlin, 2005.

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

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

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

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

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models.* Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Takashi Nicholas Maeda and Shohei Shimizu. Causal discovery of linear non-Gaussian acyclic models in the presence of latent confounders. *CoRR*, abs/2001.04197, 2020.

Christopher Nowzohour. *greedyBAPs: Greedy BAP Learning Using Penalised Maximum Likelihood Score*, 2017. R package version 0.0.0.9000.

Christopher Nowzohour, Marloes H. Maathuis, Robin J. Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electron. J. Stat.*, 11(2):5342–5374, 2017.

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

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(1):5–31, 2018.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30 (4):962–1030, 2002.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-Gaussian causal models in the presence of latent variables. *J. Mach. Learn. Res.*, 21(39):1–24, 2020.

Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *J. Mach. Learn. Res.*, 15:2629–2653, 2014.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248, 2011.

Bill Shipley. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R.* Cambridge University Press, 2016.

Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *Ann. Appl. Stat.*, 3 (4):1236–1265, 2009.

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. ParceLiNGAM: a causal ordering method robust against latent confounders. *Neural Comput.*, 26(1):57–83, 2014.

Jin Tian. Identifying direct causal effects in linear models. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 346–353. AAAI Press / The MIT Press, 2005.

Daniele Tramontano, Anthea Monod, and Mathias Drton. Learning linear non-Gaussian polytree models. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1960–1969. PMLR, 01–05 Aug 2022.

Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In Frederick Eberhardt, Elias Bareinboim, Marloes H. Maathuis, Joris M. Mooij, and Ricardo Silva, editors, *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), Jersey City, USA, June 29, 2016.*, volume 1792 of *CEUR Workshop Proceedings*, pages 59–67. CEUR-WS.org, 2016.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463, 2013.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, MIT, Cambridge, MA, USA, July 27-29, 1990*, pages 255–270, 1990.

Y. Samuel Wang. *Linear Structural Equation Models with Non-Gaussian Errors: Estimation and Discovery.* PhD thesis, University of Washington, 2018.

Y. Samuel Wang and Mathias Drton. Empirical likelihood for linear structural equation models with dependent errors. *Stat*, 6:434–447, 2017.

Y. Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107(1):41–59, 2020.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Appendix A. Proofs from Section 2

We first restate two lemmas from Tashiro et al. (2014) which imply that the sink and source certification procedures used by ParcelLiNGAM are sound. Strictly speaking, the lemmas require linear confounders, but the results trivially generalize to our setting in which the effects of confounders are represented via correlated errors.

Also, as stated, the lemmas do not require faithfulness because they consider the full latent variable LiNGAM model where $\beta_{v,p} \neq 0$ for all $p \in \mathrm{pa}(v)$. Because the graph is acyclic, this implies that every non-source has at least one parent with a non-zero total effect. However, this does not hold when considering sub-models induced by marginalizing out subsets of the variables; i.e., $\beta_{v,p} \neq 0$ for all $p \in \mathrm{pa}(v)$ in the full model does not imply that all parents (or ancestors) in a sub-model induced by marginalization have non-zero total effects on their children (or descendants). A simple example is given in Figure 12. Thus, to show that ParcelLiNGAM is sound and complete, we require that the marginal direct effect of an ancestor on its descendants does not disappear for any model induced by marginalization. This is similar to the notion of parental faithfulness required in Wang and Drton (2020) and is true for generic linear coefficients. Hence, in the proofs of Lemma 1 and 2 we assume generic model parameters, and then apply Lemmas 26 and 27 assuming that they hold for all subsets of the variables as well. Finally, we use the notation K_{head} and K_{tail} which was used in ParcelLiNGAM. In particular, ParcelLiNGAM is an iterative procedure which keep two orderings of nodes: K_{head}, which is an estimate of the causal ordering from the top downwards (i.e., from root to the leaves), and K_{tail}, which is an estimate of the causal ordering from the bottom upwards (i.e., from the leaves to the roots).

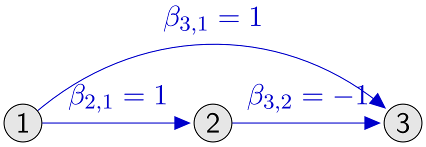


Figure 12: When considering the entire graph with nodes $\{1, 2, 3\}$, the lemmas holds since every non-source has at least one parent with a non-zero total effect. However, when only considering the sub-graph induced by $\{1, 3\}$, the marginal direct effect of 1 on 3 is 0 so 3 is a source in that sub-model despite the fact that it is a sink in the full model.

Lemma 26 *(Lemma 1 in Tashiro et al. (2014)) Assume all model assumptions of the latent variable LiNGAM are met. Denote by $r_i^{(j)}$ the population residuals when Y_i are regressed onto Y_j. Then a variable Y_j is exogenous in the sense that is has no parent observed variable or latent confounder if and only if Y_j is independent of its residuals $r_i^{(j)}$ for all $i \neq j$.*

Lemma 27 *(Lemma 2 in Tashiro et al. (2014)) Assume all model assumptions of the latent variable LiNGAM are met. Denote by $Y_{(-j)}$ a vector that contains all the variables other than Y_j. Denote by $r_j^{(-j)}$ the population residuals when Y_j is regressed onto $Y_{(-j)}$. Then a variable Y_j is a sink in the sense that is has no parent observed variable or latent confounder if and only if $Y_{(-j)}$ is independent of its residual $r_j^{(-j)}$.*

A.1 Lemma 1

Suppose Y is generated by a recursive linear SEM that corresponds to an ancestral graph G. With generic model parameters and population information (i.e., the distribution of Y), the ordering, $\hat{\prec}$, returned by Algorithm 2 of ParcelLiNGAM is sound and complete for ancestral relationships in G.

Proof We first consider Algorithm 2 which applies Algorithm 1 to all sets in the powerset of V. Since Lemma 26 and 27 explicitly concern the certificate used to place nodes into K_{head} or K_{tail}, they trivially imply that any output K_{head} and K_{tail} method is sound. It remains to be shown that the procedure is complete for ancestral relationships. By the ancestral assumption, every $v \in V$ is a sink in the set $\text{An}(v)$ since it does not share a confounder with any ancestor, so when applying Algorithm 1 to $\text{An}(v)$ either all of $\text{An}(v)$ will be put into K_{head} or v will be put into K_{tail}. Regardless, it will be identified that $u \prec v$ for all $u \in \text{An}(v) \setminus v = \text{an}(v)$. Thus, Algorithm 2 is thus complete.

Now consider Algorithm 3 which first applies Algorithm 1 to V and identifies K_{head} and K_{tail}. It then applies Algorithm 2 to $U_{\text{res}} := V \setminus \{K_{\text{head}} \cup K_{\text{tail}}\}$. K_{head} and K_{tail} are both total orderings so the orientation rules in Step 4 of Algorithm 1 will completely identify all ancestral relationships between any u, v such that (1) $u, v \in K_{\text{head}} \cup K_{\text{tail}}$, (2) $u \in K_{\text{head}}$ and $v \in U_{\text{res}} \cup K_{\text{tail}}$, or (3) $u \in U_{\text{res}}$ and $v \in K_{\text{tail}}$. Thus, it remains to show that the remaining steps of Algorithm 3 completely discover all ancestral relationships between any pair u, v such that $u, v \in U_{\text{res}}$.

By the soundness of the certification procedure, $K_{\text{head}} \cup \text{an}(K_{\text{head}}) = K_{\text{head}}$ and $K_{\text{head}} \cap S = \emptyset$ where $S = \{v \in V : \text{sib}(v) \neq \emptyset\}$. Similarly $K_{\text{tail}} \cup \text{de}(K_{\text{tail}}) = K_{\text{tail}}$. It is well known that when $A \subset V$ is an ancestral set, the residuals when regressing $V \setminus A$ onto A correspond to a model which can be represented by the sub-graph induced by $V \setminus A$ (e.g., Chen et al. (2019, Lemma 2)). In addition, removing a set which contains all of its descendants does not change the induced sub-graph; for instance see Drton (2018, Section 5). Thus, the residuals formed in Step 4, R_{res}, correspond to the sub-graph induced by U_{res}, which is also ancestral. Thus, applying the proof for Algorithm 2 implies that Step 5 of Algorithm 1 discovers all ancestral relations for $u, v \in U_{\text{res}}$. Thus Algorithm 3 is also complete. ■

A.2 Lemma 2

Suppose Y is generated by a recursive linear SEM that corresponds to a graph G which is bow-free but not ancestral. With generic parameters and population information, both Algorithm 2 and Algorithm 3 of ParcelLiNGAM will return a partial ordering which is sound, but not complete for ancestral relationships in G.

Proof Algorithm 2 applies Algorithm 1 to the powerset of V, and we first consider the output of Algorithm 1 on a set $M \subseteq V$.

Let $S = \{v \in V : \text{sib}(v) \neq \emptyset\}$. In a graph which is not ancestral, there must exist some $u, v \in V$ such that $u \in \text{sib}(v) \cap \text{an}(v)$. Let

$$Z(u, v) = \{z \, : \, \{u \cup \text{de}(u)\} \setminus \text{de}(v)\}. \tag{29}$$

Now consider testing any set $M \subseteq V$ such that $v \in M$ and $M \cap Z(u,v) \neq \emptyset$. Let

$$Z_{\text{top}} = \{z \in Z(u,v) \,:\, \text{an}(z) \cap \{M \cap Z(u,v)\} = \emptyset\},$$

so that Z_{top} are nodes in $M \cap Z(u,v)$ which are not downstream of any other nodes in $M \cap Z(u,v)$. Thus any $z \in Z_{\text{top}}$ will not be exogenous since it shares a latent confounder (acting through u) with v and similarly v will not be a sink. Thus, Lemma 26 implies that no $z \in Z(u,v)$ will be placed into K_{head} which further implies no $\text{de}(Z(u,v))$ will be placed into K_{head}. Similarly, Lemma 27 implies that v will not be placed into K_{tail} which further implies no ancestor of v will be put into K_{tail}. Together, this implies that $M \cap Z(u,v) \subseteq U_{\text{res}}$ so that running Algorithm 1 on M will return inconclusive ancestral relationships between all $z \in Z(u,v)$. Since this holds for any $M \subseteq V$ such that $v \in M$ and $M \cap Z(u,v) \neq \emptyset$, Algorithm 2 will not discover that $z \prec v$ for any $z \in Z(u,v)$. Since $Z(u,v) \cap \text{an}(v) \neq \emptyset$ Algorithm 2 is not complete.

Algorithm 3 uses additional steps (Steps 2-4) before applying Algorithm 2. We show that these additional steps do not rectify the problem. First note that when applying Algorithm 1 to V (Step 2), $K_{\text{head}} \subseteq V \setminus \{S \cup \text{de}(S)\}$ and $K_{\text{tail}} \subseteq V \setminus \text{An}(S)$. This is true because, by definition, any $s \in S$ is not exogenous since it shares a common confounder with some other $s' \in S$. Thus, no $s \in S$ will be put into K_{head} and subsequently no $\text{de}(S)$ will be put into K_{head}. For the same reason, no $s \in S$ will be put into K_{tail} since it is not a sink and subsequently no $\text{an}(S)$ will be put into K_{tail}.

Since $Z(u,v) \subseteq \{u \cup \text{de}(u)\}$ and $u \in S$, then $Z(u,v) \cap K_{\text{head}} = \emptyset$. Thus, Step 4 will not remove from any $z \in Z(u,v)$ the effect of u or the effect of the latent confounder shared by u and v. Thus, as shown above, Step 5 of Algorithm 3 (applying Algorithm 2 to R_{res}) will still fail to identify that $z \in \text{an}(v)$ for any $z \in Z(u,v)$. ■

A.3 Counterexamples: Pairwise LvLiNGAM

Pairwise LvLiNGAM will fail to discover any relationships in the simple ancestral graph shown in Figure 13. This is because all subsets of $V = \{1, 2, 3\}$ are confounded.

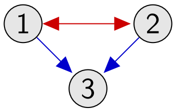


Figure 13: An ancestral graph which Pairwise LvLiNGAM will fail to identify.

A.4 Counterexamples: RCD

We step through the RCD procedure when applied to the graph in Figure 2. We follow the notation from Algorithm 1 in Maeda and Shimizu (2020): x_j is the observed data for variable j; at each step we consider a set $U \subseteq V$ where $|U| = l + 1$ for some counter l; M_i is the set of verified ancestors of i; $H_U = \bigcap_{j \in U} M_j$, and y_j is the resulting residual when x_j is regressed onto H_U. When regressing x_j onto some set H, we let $d_{j,u.H}$ denote the population regression coefficient corresponding to $u \in H$. When performing a HSIC minimizing regression of y_i onto y_j for $j \in U \setminus i$ we let λ be a potential solution and S_i^U is the resulting residual. Finally S denotes a possible sink which might be certified at each step.

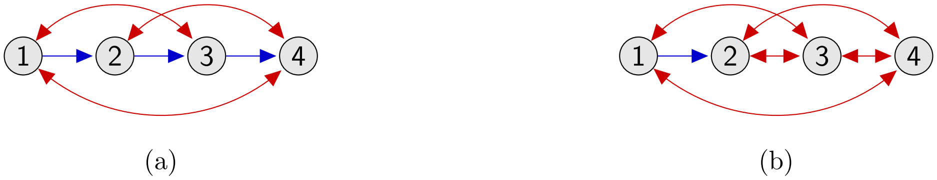


Figure 14: The graph in (a) is a non-ancestral BAP which would be correctly identified by BANG but not Pairwise LvLiNGAM, ParcelLiNGAM, or RCD. The graph in (b) shows the graph which would be identified by Pairwise LvLiNGAM, ParcelLiNGAM, and RCD.

We walk through the RCD procedure with the correction that line 9 of Algorithm 1 of Maeda and Shimizu (2020) is actually the intersection of sets—i.e., $H_U = \bigcap_{j \in U} M_j$—instead of the union as stated in the original text.

- $\boldsymbol{U = \{1,2\} : \qquad M = \{\emptyset, \emptyset, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $y_1 = \varepsilon_1$; $y_2 = \varepsilon_2 + \beta_{2,1} x_1$
  - Let $i = 1$: $1 \in \mathrm{pa}(2)$ so there is no value of λ such that $y_2 \perp\!\!\!\perp y_1 - \lambda y_2$
  - Let $i = 2$: Setting $\lambda = \beta_{2,1}$ yields
  $$y_2 - \lambda y_1 = \varepsilon_2 \perp\!\!\!\perp y_1.$$
  - $S = 2$. Update $M_2 = \{1\}$.
- $\boldsymbol{U = \{1,3\} : \qquad M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $1 \in \mathrm{sib}(3)$ so there is no updates to M.
- $\boldsymbol{U = \{1,4\} : \qquad M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $1 \in \mathrm{sib}(4)$ so there is no update to M.
- $\boldsymbol{U = \{2,3\} : \qquad M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $y_2 = \varepsilon_2 + \beta_{2,1}\varepsilon_1$; $y_3 = \varepsilon_3 + \beta_{3,2}(\varepsilon_2 + \beta_{2,1}\varepsilon_1)$
  - Let $i = 2$: $2 \in \mathrm{pa}(3)$ so there is no value of λ such that $y_3 \perp\!\!\!\perp y_2 - \lambda y_3$.
  - Let $i = 3$: In order for $S_3^U \perp\!\!\!\perp y_2$, it is necessary that S_3^U not contain a ε_2 term. This implies that λ must be $\beta_{3,2}$ so that $S_3^U = y_3 - \lambda y_2 = \varepsilon_3$. However, since $1 \in \mathrm{sib}(3)$, then
  $$y_3 - \lambda y_2 = \varepsilon_3 \not\!\perp\!\!\!\perp y_2 = \varepsilon_2 + \beta_{2,1}\varepsilon_1.$$
  so there is no update to M.
- $\boldsymbol{U = \{2,4\} : \qquad M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $2 \in \mathrm{sib}(4)$ so there is no update.
- $\boldsymbol{U = \{3,4\} : \qquad M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; \, l = 1 \qquad H_U = \emptyset}$
  - $y_3 = \varepsilon_3 + \beta_{3,2} x_2$; $y_4 = \varepsilon_4 + \beta_{4,3} x_3$
  - Let $i = 3$: $3 \in \mathrm{pa}(4)$ so there is no value of λ such that $y_4 \perp\!\!\!\perp y_3 - \lambda y_4$
  - Let $i = 4$: For $S_4^U \perp\!\!\!\perp y_3$, it is necessary that S_4^U not contain a ε_3 term. This implies that λ must be $\beta_{4,3}$, so that $S_4^U = y_4 - \lambda y_3 = \varepsilon_4$. However, since $1 \in \mathrm{sib}(4)$, then
  $$S_4^U = y_4 - \lambda y_3 = \varepsilon_4 \not\!\perp\!\!\!\perp \varepsilon_3 + \beta_{3,2}(\varepsilon_2 + \beta_{2,1}\varepsilon_1)$$
  so there is no update to M.

This is all subsets of size 2, but since an update has occurred, l remains 1, and the procedure will cycle through all subsets of size 2 again. However, since M_2 is the only non-empty set, H_U is the same for all pairs, so the outcomes are the same as before the second time through. M is not updated so $l = 2$ and we test all sets of size 3.

- $\boldsymbol{U = \{1,2,3\}:} \qquad \boldsymbol{M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; l = 2} \qquad \boldsymbol{H_U = \emptyset}$
  - Let $i = 1$: $1 \in \text{pa}(2)$ so no update can be made.
  - Let $i = 2$: $M_2 \cap U \neq \emptyset$ so it is not tested.
  - Let $i = 3$: $y_1 = \varepsilon_1$ and $y_2 = x_2$ both do not contain a ε_3 term, so S_3^U must contain a ε_3 term. Since $1 \in \text{sib}(3)$, then S_3^U cannot be independent of y_1, so no update is made.
  - No update is made.
- $\boldsymbol{U = \{1,2,4\}:} \qquad \boldsymbol{M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; l = 2} \qquad \boldsymbol{H_U = \emptyset}$
  - Let $i = 1$: $1 \in \text{pa}(2)$ so no update can be made.
  - Let $i = 2$: $M_2 \cap U \neq \emptyset$ so it is not tested.
  - Let $i = 4$: $y_1 = \varepsilon_1$ and $y_2 = x_2$ both do not contain a ε_4 term, so S_4^U must contain a ε_4 term. Since $1 \in \text{sib}(4)$, then S_4^U cannot be independent of y_1, so no update is made.
  - No update is made.
- $\boldsymbol{U = \{1,3,4\}:} \qquad \boldsymbol{M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; l = 2} \qquad \boldsymbol{H_U = \emptyset}$
  - Let $i = 1$: $1 \in \text{pa}(2)$ so no update can be made.
  - Let $i = 3$: $3 \in \text{pa}(4)$ so no update can be made.
  - Let $i = 4$: $y_1 = \varepsilon_1$ and $y_3 = x_3$ both do not contain a ε_4 term, so S_4^U must contain a ε_4 term. Since $1 \in \text{sib}(4)$, then S_4^U cannot be independent of y_1, so no update is made.
  - No update is made.
- $\boldsymbol{U = \{2,3,4\}:} \qquad \boldsymbol{M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; l = 2} \qquad \boldsymbol{H_U = \emptyset}$
  - Let $i = 2$: $2 \in \text{pa}(3)$ so no update can be made.
  - Let $i = 3$: $3 \in \text{pa}(4)$ so no update can be made.
  - Let $i = 4$: $y_2 = x_2$ and $y_3 = x_3$ both do not contain a ε_4 term, so S_4^U must contain a ε_4 term. Since $2 \in \text{sib}(4)$, then S_4^U cannot be independent of y_2, so no update is made.
  - No update is made.

Since no updates have been made, $l = 3$.

- $\boldsymbol{U = \{1,2,3,4\}:} \qquad \boldsymbol{M = \{\emptyset, \{1\}, \emptyset, \emptyset\}; l = 3} \qquad \boldsymbol{H_U = \emptyset}$

- Let $i = 1$: $1 \in \mathrm{pa}(2)$ so no update can be made.
- Let $i = 2$: $2 \in \mathrm{pa}(3)$ so no update can be made.
- Let $i = 3$: $3 \in \mathrm{pa}(4)$ so no update can be made.
- Let $i = 4$: $y_1 = \varepsilon_1$, $y_2 = x_2$, and $y_3 = x_3$ both do not contain a ε_4 term, so S_4^U must contain a ε_4 term. Since $\mathrm{sib}(4) = \{1, 2\}$, then S_4^U cannot be independent of y_1 or y_2, so no update is made.
- No update is made.

The algorithm will terminate and has only discovered $1 \to 2$.

Appendix B. Proofs from Section 3

B.1 Lemma 5

Let $v \in V$ and $C \subseteq V \setminus \{v\}$. Suppose $D \in \mathbb{R}^{p \times p}$ such that $D_{s,t} \neq 0$ only if $t \in \mathrm{an}(s)$. Then, for generic B and error moments, if $\delta_v(C, \mathrm{an}_D(C), S, D) \neq \tilde{B}(C \cup \{v\})_{v,C}$, then $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, S, D)) \neq 0$ for some $c \in C$.

Proof Since $\mathbb{E}\left(\gamma_c^{K-1}\gamma_v\right)$ is a rational function of the model parameters, by Okamoto (1973, Lemma 1), showing that the quantity is non-zero for some parameters is sufficient for showing that it vanishes only over a null set. Without loss of generality, let C be ordered such that $C = \{c_1, \dots, c_{|C|}\}$ where c_i is not a descendant of c_j for any $j < i$. Note that

$$\begin{aligned}\gamma_v &= \varepsilon_v + \sum_{a \in \mathrm{an}(v)} \pi_{v,a}\varepsilon_a - \sum_{c \in C} \delta_{v,c} Y_c \\ &= \varepsilon_v + \sum_{a \in \mathrm{an}(v)} \pi_{v,a}\varepsilon_a - \sum_{c \in C} \delta_{v,c}(\varepsilon_c + \sum_{a \in \mathrm{an}(c)} \pi_{c,a}\varepsilon_a).\end{aligned} \tag{30}$$

Suppose i is the minimum index for which $\delta_{c_i} \neq \tilde{B}_{v,c_i}$ so that $\delta_{c_j} = \tilde{B}_{v,c_j}$ for all $j < i$. Then, the coefficient of ε_{c_i} in $Y_v - \sum_{j<i} \delta_{v,c_j} Y_{c_j}$ is

$$\begin{aligned}\pi_{v,c_i} - \sum_{j<i} \delta_{v,c_j}\pi_{c_j,c_i} &= \pi_{v,c_i} - \sum_{j<i} \tilde{\beta}_{v,c_j}\pi_{c_j,c_i} \\ &= \sum_{l \in \mathcal{L}_{v,c_i}} W(l) - \sum_{j<i} \left[\left(\sum_{l \in \mathcal{L}^{(c_j)}_{v,c_j}(C)} W(l)\right)\left(\sum_{l \in \mathcal{L}_{c_j,c_i}} W(l)\right)\right] \\ &= \sum_{l \in \mathcal{L}_{v,c_i}} W(l) - \sum_{j<i} \left[\sum_{l \in \mathcal{L}^{(c_j)}_{v,c_i}(C)} W(l)\right] \\ &= \sum_{l \in \mathcal{L}^{(c_i)}_{v,c_i}} W(l) = \tilde{B}(C)_{v,c_i}.\end{aligned} \tag{31}$$

For all $j > i$, c_j is not a descendant of c_i so Y_{c_j} does not include any terms of ε_{c_i}. By assumption, $\delta_{c_i} \neq \tilde{B}_{v,c_i}$, so let $\delta_{c_i} = \tilde{B}_{v,c_i} - \alpha$ for $\alpha \neq 0$ so that

$$\gamma_v = \alpha\varepsilon_{c_i} + \eta \qquad \text{and} \qquad \gamma_{c_i} = \varepsilon_{c_i} + \zeta, \tag{32}$$

where η and ζ do not contain ε_{c_i}. Then,

$$\begin{aligned}\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right) &= \mathbb{E}\left([\varepsilon_{c_i} + \zeta]^{K-1}[\alpha\varepsilon_{c_i} + \eta]\right) \\ &= \mathbb{E}\left(\left[\varepsilon_{c_i}^{K-1} + \sum_{k=0}^{K-2} \varepsilon_{c_i}^{k}\zeta^{K-1-k}\right][\alpha\varepsilon_{c_i} + \eta]\right) \\ &= \alpha\mathbb{E}\left(\varepsilon_{c_i}^{K}\right) + \mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2} \varepsilon_{c_i}^{k}\zeta^{K-1-k}\right][\alpha\varepsilon_{c_i} + \eta]\right).\end{aligned}$$

Since the last two terms do not involve $\mathbb{E}(\varepsilon_{c_i}^K)$, we can always select some $\mathbb{E}(\varepsilon_{c_i}^K)$ such that

$$\mathbb{E}\left(\varepsilon_{c_i}^K\right) \neq -\frac{\left(\mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_{c_i}^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_{c_i}+\eta\right]\right)\right)}{\alpha}$$

which ensures that $\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right) \neq 0$. ■

B.2 Proof of Lemma 6

Consider $v \in V$ and $C \subseteq V \setminus \{v\}$. Let $D \in \mathbb{R}^{p\times p}$ such that $D_{s,t} \neq 0$ only if $t \in \mathrm{an}(s)$. Suppose $C \not\subseteq \mathrm{an}(v)$, but that $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C,S,D)) = 0$ for all $c \in C$. Then for generic B and error moments, $C_1 = C \cap [\mathrm{an}(v) \setminus \mathrm{sib}(v)]$,

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1,S,D)) = 0 \qquad \forall c \in C.$$

Proof For convenience, let $A = \mathrm{An}_D(C)$, $A_1 = \mathrm{An}_D(C_1)$, $A_2 = A \setminus A_1$, and $\Lambda = I - D$ and $\Pi = (I-D)^{-1}$. Note that $A_2 \cap \mathrm{de}(A_1) = \emptyset$; this implies $D_{A_1,A_2} = 0$ and $\left[(I-D_{A,A})^{-1}\right]_{A_1,A_2} = 0$. So that

$$\begin{aligned}(I-D)_{C_1,A}S_{A,C} &= \begin{bmatrix}\Lambda_{C_1,A_1} & \Lambda_{C_1,A_2}\end{bmatrix}\begin{bmatrix}S_{A_1,C_1}\\ S_{A_2,C_1}\end{bmatrix} = \begin{bmatrix}\Lambda_{C_1,A_1} & 0\end{bmatrix}\begin{bmatrix}S_{A_1,C_1}\\ S_{A_2,C_1}\end{bmatrix}\\ &= \Lambda_{C_1,A_1}S_{A_1,C_1},\end{aligned}$$

and

$$\begin{aligned}(I-D)_{C_1,A}\Sigma_{A,v} &= \begin{bmatrix}\Lambda_{C_1,A_1}\Lambda_{C_1,A_2}\end{bmatrix}\begin{bmatrix}\Sigma_{A_1,v}\\ \Sigma_{A_2,v}\end{bmatrix} = \begin{bmatrix}\Lambda_{C_1,A_1}0\end{bmatrix}\begin{bmatrix}\Sigma_{A_1,v}\\ \Sigma_{A_2,v}\end{bmatrix}\\ &= (I-D)_{C_1,A_1}\Sigma_{A_1,v}.\end{aligned}$$

Thus,

$$\begin{aligned}\delta_v(C_1,A,S,D) &= \left[(I-D)_{C_1,A}S_{A,C_1}\right]^{-1}(I-D)_{C_1,A}\Sigma_{A,v}\\ &= \left[(I-D)_{C_1,A_1}S_{A_1,C_1}\right]^{-1}(I-D)_{C_1,A_1}\Sigma_{A_1,v}\\ &= \delta_v(C_1,A_1,S,D).\end{aligned}$$

By Lemma 5, for generic B and error moments, if

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C,S,D)) = 0,$$

then for every $q \notin C_1$, $\delta_{vq}(C,A,S,D) = 0$.

$$\begin{aligned}\gamma_v(C,S,D) &= Y_v - Y_C\delta_v(C,A,S,D)\\ &= Y_v - Y_{C_1}(C_1,A_1,S,D)\\ &= \gamma_v(C_1,A_1,S,D).\end{aligned}$$

So if for all $c \in C$,

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)\right) = 0, \tag{33}$$

then for all $c \in C$

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C_1, S, D)\right) = 0. \tag{34}$$

■

B.3 Proof of Lemma 7

Suppose $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \text{an}(s) \setminus \text{sib}(s)$. Let $v \in V$ be such that we have $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(\text{pa}_D(v), S, D)) = 0$ for all $c \in \text{pa}_D(v)$. If $q \in (\text{pa}(v) \backslash \text{pa}_D(v)) \cup \text{sib}(v)$, then for generic B and error moments, $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$.
Proof For notational convenience, let $C = \text{pa}_D(v)$. First consider $q \in \text{pa}(v) \setminus \text{pa}_D(v)$. $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \text{An}_D(C), S, D)) = 0$ for all $c \in \text{pa}_D(v)$ implies that

$$\begin{aligned}
\gamma_v(D) &= Y_v - Y_{\text{pa}_D(v)}(D_{v,\text{pa}_D(v)})^T \\
&= \left(\pi_{v,q} - \sum_{c \in C} \tilde{B}(C \cup \{v\})_{v,c}\pi_{c,q}\right)\epsilon_q + \eta \\
&= \left(\pi_{v,q} - \sum_{c \in C \cap \text{de}(q)} \tilde{B}(C \cup \{v\})_{v,c}\pi_{c,q}\right)\epsilon_q + \eta \\
&= \alpha\epsilon_q + \eta
\end{aligned} \tag{35}$$

where η does not involve ϵ_q. For any $c \in \text{de}(q)$, $\tilde{B}(C \cup \{v, q\})_{v,C} = \tilde{B}(C \cup \{v\})_{v,C}$ because there are no paths from c to v which pass through q, so marginalizing q does not change the marginal direct effect. Thus, as shown in Lemma 5,

$$\begin{aligned}
\alpha &= \pi_{q,v} - \sum_{c \in C \cap \text{de}(q)} \tilde{B}(C \cup \{q, v\})_{v,C}\pi_{c,q} \\
&= \tilde{B}(C \cup \{q, v\})_{v,q}.
\end{aligned} \tag{36}$$

The set of points, B such that $q \in \text{pa}(v)$, but the marginal direct effect $\tilde{B}(C \cup \{q, v\})_{vq} = 0$ have Lebesgue measure 0, so by the same argument as Lemma 5 when $\alpha \neq 0$, for generic error moments, $\mathbb{E}(\gamma_q^{K-1}\gamma_v) \neq 0$.

Now consider $q \in \text{sib}(v)$. Since $\text{pa}_D(v) \subseteq \text{an}(v)$ for all $v \in V$, then $\gamma_v = \epsilon_v + \eta$ where η does not involve ε_v and $\gamma_q = \epsilon_q + \zeta$ where ζ does not involve ϵ_q. Then, using the same argument as the previous lemmas, selecting

$$\mathbb{E}(\epsilon_q^{K-1}\epsilon_v) \neq -\mathbb{E}\left(\sum_{t=0}^{K-2}\binom{K-1}{t}\varepsilon_q^t\zeta^{K-1-t}(\epsilon_v + \eta) + \epsilon_q^{K-1}\eta\right) \tag{37}$$

ensures that $\mathbb{E}(\gamma_q^{K-1}\gamma_v) \neq 0$ ■

B.4 Proof of Lemma 8

Consider $v \in V$ and C such that $C \subseteq V \setminus \{v\}$. Suppose $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \text{an}(s) \setminus \text{sib}(s)$ for all $v \in V$. If $C \cap \text{sib}(v) \neq \emptyset$, then for generic B and error moments, there exists some $q \in C$ such that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0$.

Proof We again appeal to Okamoto (1973, Lemma 1), and show that the quantity is non-zero for generic B and the error moments by constructing a single point (of B and the error moments) at which the quantity of interest is non-zero. In particular, select $q \in C \cap \text{sib}(v)$. We then represent γ_v as

$$\begin{aligned}\gamma_v &= \varepsilon_v + \sum_{a \in \text{an}(v)} \pi_{v,a}\varepsilon_a - \sum_{c \in C} \delta_{v,c} \sum_{z \in \text{An}(c)} \pi_{c,z}\varepsilon_z \\ &= \alpha\varepsilon_q + \eta,\end{aligned} \tag{38}$$

where

$$\begin{aligned}\alpha &= \pi_{v,q} + \sum_{c \in C} \delta_{v,c}\pi_{c,q} \\ \eta &= (1 - \sum_{c \in C} \delta_{v,c}\pi_{c,v})\epsilon_v + \sum_{a \in \text{an}(v) \setminus q} \pi_{v,a}\varepsilon_a - \sum_{c \in C} \delta_{v,c} \sum_{z \in \text{An}(c) \setminus q} \pi_{c,z}\varepsilon_z\end{aligned}$$

and $\delta_{v,c}$ is the c-th element of δ_v from (10). Similarly, we represent γ_q

$$\begin{aligned}\gamma_q &= \varepsilon_q + \sum_{a \in \text{an}(q)} \pi_{v,a}\varepsilon_a - \sum_{s \in \text{pa}_D(q)} d_{q,s} \sum_{t \in \text{An}(s)} \pi_{s,t}\varepsilon_t \\ &= \varepsilon_q + \zeta\end{aligned} \tag{39}$$

where ζ does not involve ε_q. The coefficient on ε_q is 1 since $D = H_{\mathcal{C}}(B)$ implies that $d_{q,s} \neq 0$ only if $s \in \text{an}(q)$. For $S = \Sigma$ and any $H_{\mathcal{C}} \in \mathbf{D}$, α is a rational function of B and Ω because both Π and δ only involve matrix inversions and multiplications of D and S which in turn are rational functions of B and Ω. We now show that for some point B and Ω, $\alpha \neq 0$. In particular, let $B = 0$ and $\omega_{qv} \neq 0$, but $\omega_{ij} = 0$ for all other $i \neq j$. At this point, $\pi_{v,q} = \pi_{c,q} = 0$ for all $c \in C \setminus q$ so that

$$\alpha = \delta_{v,q}. \tag{40}$$

$B = 0$ implies that $D = 0$ for all $D \in \mathcal{D}$ and $S_{C,C} = \Omega_{C,C}$. In addition, $S_{C \setminus q, v} = 0$ since all treks between nodes in C or between treks $C \setminus \{q\}$ and v have path weights of 0. However, there is a single trek between q and v, namely the bidirected edge, so $S_{qv} = \omega_{qv}$. Then,

$$\alpha = \delta_{v,C} = \left[S_{C,C}\right]^{-1} S_{C,v} = \frac{\omega_{qv}}{\omega_{qq}} \neq 0. \tag{41}$$

Thus, for generic choice of B and Ω, $\alpha \neq 0$. Now, we finally examine the quantity of interest, which is a rational function of the error moments and B, and play the same game as before. In particular,

$$\begin{aligned}\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) &= \mathbb{E}\left([\varepsilon_q + \zeta]^{K-1}\left[\alpha\varepsilon_q + \eta\right]\right) \\ &= \mathbb{E}\left(\left[\varepsilon_q^{K-1} + \sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q + \eta\right]\right) \\ &= \alpha\mathbb{E}\left(\varepsilon_q^K\right) + \mathbb{E}\left(\varepsilon_q^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q + \eta\right]\right).\end{aligned}$$

The last two terms do not involve $\mathbb{E}(\varepsilon_q^K)$ so we select $\mathbb{E}\left(\varepsilon_q^K\right)$ such that

$$\mathbb{E}\left(\varepsilon_q^K\right) \neq -\frac{\left(\mathbb{E}\left(\varepsilon_q^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q + \eta\right]\right)\right)}{\alpha} \tag{42}$$

to ensure that $\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) \neq 0$. Thus, there exists some point such that $\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) \neq 0$. This implies there is a null set of B and error moments which we must avoid for each $H_{\mathcal{C}} \in \mathcal{D}$, but since $|\mathcal{D}|$ is finite, then the union of these null sets is again a null set. ■

B.5 Proof of Corollary 9

Suppose $D = B$. For $v \in V$ and generic B and error moments, suppose $\text{pa}(v) \subseteq C \subseteq \text{an}(v) \setminus \text{sib}(v)$. If $q \in C \setminus \text{pa}(v)$, then for all $c \in C$

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = 0. \tag{43}$$

If $q \in \text{pa}(v)$, then there exists some $c \in C$ such that

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) \neq 0. \tag{44}$$

Proof Suppose $q \in C \setminus \text{pa}(v)$ and without loss of generality, assume that q is the last element in C. Then, Lemma 3 implies that

$$\begin{aligned}\delta_v(C, \text{An}_D(C), \Sigma, D) = B_{v,(C\setminus\{q\},q)} &= \begin{bmatrix}B_{v,(C\setminus\{q\})} & 0\end{bmatrix} \\ &= \begin{bmatrix}\delta_v(C \setminus \{q\}, \Sigma, D) & 0\end{bmatrix}\end{aligned} \tag{45}$$

so that for all $c \in C$

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = \mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C, \Sigma, D)) = 0. \tag{46}$$

Now consider the second statement when $q \in \text{pa}(v)$. If $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) \neq 0$ for some $c \in C \setminus \{q\}$ then the statement trivially holds. Thus, it remains to be shown that $\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) \neq 0$ when $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = 0$ for all $c \in C \setminus \{q\}$. This is directly implied by Lemma 7 ■

Lemma 28 *Fix $v \in V$ and sets $C_1, C_2 \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$, and suppose $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \mathrm{an}(s) \setminus \mathrm{sib}(s)$ for all $v \in V$. For generic B and error moments, if $\mathbb{E}(\gamma_c(D)\gamma_v(C_1, \Sigma, D)) = 0$ for all $c \in C_1$ and $\mathbb{E}(\gamma_c(D)\gamma_v(C_2, \Sigma, D)) = 0$ for all $c \in C_2$, then $\mathbb{E}(\gamma_c(D)\gamma_v(C_3, \Sigma, D)) = 0$ for all $c \in C_3 = C_1 \cup C_2$. That is, if C_1 and C_2 are certified by Eq.* (12), *then $C_1 \cup C_2$ will also be certified.*

Proof We begin by showing that for generic parameters, certification of $C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ via (12) is equivalent to a graphical condition for v and C given D. Furthermore, we show that when this graphical condition holds for C_1 and C_2 given D, then it also holds for $C_1 \cup C_2$ given D. Thus, if C_1 and C_2 are certified, then $C_1 \cup C_2$ will also be certified.

Let $an(v \mid C) = \{a \in \mathrm{an}(v) \,:\, \exists$ a path from a to v which does not pass through $C\}$. Then, we say that v is "uncounfounded" with C given D if for each $a \in \mathrm{an}(v \mid C)$, we have $\mathrm{an}(C \mid \mathrm{pa}_D(C)) \cap \{a\} \cup \mathrm{sib}(a) = \emptyset$.

We first show that certification of C for generic parameters implies that v is "uncounfounded" with C given D.

By Lemma 5, for generic parameters, C is certified only if $\delta_{v,C} = \check{B}(C \cup \{v\})$. Thus,

$$
\begin{aligned}
\gamma_v(C, \Sigma, D) &= Y_v - \check{B}(C \cup \{v\})_{v,C} Y_C \\
&= \varepsilon_v + \sum_{a\in\mathrm{an}(v)} \pi_{v,a}\varepsilon_a - \sum_{c\in C}\sum_{z\in\mathrm{An}(c)} \left[\left(\sum_{l\in\mathcal{L}^{(c)}_{v,c}(C)} W(l) \right) \left(\sum_{l\in\mathcal{L}_{c,z}} W(l) \right) \right] \varepsilon_z \qquad (47) \\
&= \varepsilon_v + \sum_{a\in\mathrm{an}(v)} \left[\sum_{l\in\mathcal{L}_{v,a}} W(l) - \sum_{l\in\bigcup_{c\in C}\mathcal{L}^{(c)}_{v,a}(C)} W(l) \right] \varepsilon_a.
\end{aligned}
$$

The set of paths $\mathcal{L}_{v,a}$ and $\bigcup_{c\in C} \mathcal{L}^{(c)}_{v,a}(C)$ are equal—which for generic parameters is equivalent to the coefficient for ε_a in γ_v being 0—if and only if all directed paths from a to v pass through C. Thus, ε_a appears in $\gamma_v(C, \Sigma, D)$ for all $a \in \mathrm{an}(v \mid C)$. Similarly, ε_z for $z \in \mathrm{an}(c)$ appears in $\gamma_c(D)$ if and only if $z \in \mathrm{an}(c \mid \mathrm{pa}_D(c))$ because $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ implies that $D_{c,\mathrm{pa}_D(c)} = \check{B}(\mathrm{pa}_D(c) \cup \{c\})$. As previously shown, when C is certified by (12), if ε_a appears in $\gamma_v(C, \Sigma, D)$, then ε_z for $z \in \{a\} \cup \mathrm{sib}(a)$ cannot appear in any γ_c for any $c \in C$. Thus, if C is certified, then for every $a \in \mathrm{an}(v \mid C)$, we have $\mathrm{an}(C \mid \mathrm{pa}_D(C)) \cap \{a\} \cup \mathrm{sib}(a) = \emptyset$, and v is "unconfounded" with C given D.

We now show the reverse direction that when v is "uncounfounded" with C given D, C will be certified. In particular, we show that distribution of Y lies within another model (which we denote using $\tilde{Y}$) for which C satisfies the conditions in Lemma 1; thus, $\mathbb{E}(\tilde{\gamma}_c\tilde{\gamma}_v(C, S, \tilde{D}) = 0$ for all $c \in C$.

For all $a \in \mathrm{an}(v)$, we construct $\tilde{Y}$ as follows. Let $\tilde{\varepsilon}_a = \gamma_a(D)$ and let

$$
\tilde{Y}_a := \sum_{z\in\mathrm{pa}_D(a)} d_{a,z}\tilde{Y}_z + \tilde{\varepsilon}_a.
$$

Since $Y_a = \sum_{z\in\mathrm{pa}_D(a)} d_{a,z}Y_z + \gamma_a(D)$, then $\tilde{Y}_a = Y_a$ for all $a \in \mathrm{an}(v)$. Furthermore, let $\pi_{v,a.C} := \sum_{l\in\mathcal{L}_{v,a}\setminus\bigcup_{c\in C}\mathcal{L}^{(c)}_{v,a}(C)} W(l)$ be the sum of all pathweights of paths from a to v which do not pass through C. Finally, define

$$\tilde{\varepsilon}_v = Y_v - \sum_{c \in C} \check{B}(C \cup v)_{v,c} Y_c = \varepsilon_v + \sum_{a \in \mathrm{an}(v|C)} \pi_{v,a.C} \varepsilon_a$$

and set

$$\tilde{Y}_v := \sum_{c \in C} \check{B}(C \cup v)_{v,c} \tilde{Y}_c + \tilde{\varepsilon}_v,$$

so that $\tilde{Y}_v = Y_v$. By construction $D_{\mathrm{an}(v),\mathrm{an}(v)}$ are the true direct effects between $\tilde{Y}_{\mathrm{an}}(v)$. In addition, because $D = H_{\mathcal{C}}(B)$ for some $H_{\mathcal{C}} \in \mathcal{D}$ such that $C_s \subseteq \mathrm{an}(s) \setminus \mathrm{sib}(s)$ for all $v \in V$, for each $c \in C$ and v is "unconfounded" with C given D, then ε_z does not appear in $\gamma_{\mathcal{C}}(D)$ for any $z \in \{v\} \cup \mathrm{an}(v \mid C) \cup \mathrm{sib}(\mathrm{an}(v \mid C))$. Thus, $\tilde{\varepsilon}_c \perp\!\!\!\perp \tilde{\varepsilon}_v$, so that $\tilde{\mathrm{pa}}(v) = C \subseteq \tilde{\mathrm{an}}(v) \backslash \tilde{\mathrm{sib}}(v)$ and the conditions for Lemma 3 are satisfied.

Because $\mathrm{an}(v \mid C_1 \cup C_2) \subseteq \mathrm{an}(v \mid C_1) \cap \mathrm{an}(v \mid C_2)$, if C_1 and C_2 are "unconfounded" with v given D, then $C_1 \cup C_2$ is also unconfounded with v given D. This implies that $C_1 \cup C_2$ will also be certified by (12). ■

Appendix C. Proofs from Section 5

C.1 Proof of Corollary 18

Let $v \in V$, and consider any set $C \subseteq A \subseteq V \setminus \{v\}$. Suppose $D \in \mathbb{R}^{p \times p}$ with $D_{s,t} \neq 0$ only if $t \in \bar{\text{an}}(s)$. Then, for generic B and error moments, if $\delta_v(C, A, S, D) \neq \check{B}(C \cup v)_{v,C}$, then $\mathbb{E}(\gamma_c^{K-1}(D)\gamma_v(C, S, D)) \neq 0$ for some $c \in C$.

Proof The proof exactly follows that of Lemma 5; however, some of the quantities in G are replaced with the corresponding quantities in $\bar{G}$.

Without loss of generality, let C be ordered such that $C = \{c_1, \dots, c_{|C|}\}$ where $c_i \notin \text{de}(c_j)$ (note that this is in the original graph G) for any $j < i$. Note that

$$\begin{aligned}
\gamma_v(C, S, D) &= \bar{\varepsilon}_v + \sum_{a \in \bar{\text{an}}(v)} \bar{\pi}_{v,a}\bar{\varepsilon}_a - \sum_{c \in C} \delta_{v,c} Y_c \\
&= \bar{\varepsilon}_v + \sum_{a \in \bar{\text{an}}(v)} \bar{\pi}_{v,a}\bar{\varepsilon}_a - \sum_{c \in C} \delta_{v,c}(\bar{\varepsilon}_c + \sum_{a \in \bar{\text{an}}(c)} \bar{\pi}_{c,a}\bar{\varepsilon}_a).
\end{aligned} \tag{48}$$

Suppose i is the minimum index for which $\delta_{c_i} \neq \check{B}_{v,c_i}$ so that $\delta_{c_j} = \tilde{B}_{v,c_j}$ for all $j < i$. Then, the coefficient of ε_{c_i} in $Y_v - \sum_{j<i} \delta_{v,c_j} Y_{c_j}$ is

$$\begin{aligned}
\bar{\pi}_{v,c_i} - \sum_{j<i} \delta_{v,c_j}\bar{\pi}_{c_j,c_i} &= \bar{\pi}_{v,c_i} - \sum_{j<i} \check{\beta}_{v,c_j}\bar{\pi}_{c_j,c_i} \\
&= \sum_{l \in \bar{\mathcal{L}}_{v,c_i}} W(l) - \sum_{j<i} \left[\left(\sum_{l \in \bar{\mathcal{L}}^{(c_j)}_{v,c_j}(C)} W(l) \right) \left(\sum_{l \in \bar{\mathcal{L}}_{c_j,c_i}} W(l) \right) \right] \\
&= \sum_{l \in \bar{\mathcal{L}}_{v,c_i}} W(l) - \sum_{j<i} \left[\sum_{l \in \bar{\mathcal{L}}^{(c_j)}_{v,c_i}(C)} W(l) \right] \\
&= \sum_{l \in \bar{\mathcal{L}}^{(c_i)}_{v,c_i}} W(l) = \check{B}(C \cup v)_{v,c_i}.
\end{aligned}$$

For all $j > i$, $c_j \notin \text{de}(c_i)$ so Y_{c_j} does not include any terms of ε_{c_i} (note this is not $\bar{\varepsilon}_{c_i}$). By assumption, $\delta_{v,c_i} \neq \check{B}_{v,c_i}$, so let $\delta_{v,c_i} = \check{B}_{v,c_i} - \alpha$ for $\alpha \neq 0$ so that

$$\gamma_v = \alpha\varepsilon_{c_i} + \eta \qquad \text{and} \qquad \gamma_{c_i} = \varepsilon_{c_i} + \zeta, \tag{49}$$

where η and ζ do not contain ε_{c_i}. Then,

$$\begin{aligned}
\mathbb{E}\left(\gamma_{c_i}^{K-1}(D)\gamma_v(C, S, D)\right) &= \mathbb{E}\left([\varepsilon_{c_i} + \zeta]^{K-1} [\alpha\varepsilon_{c_i} + \eta] \right) \\
&= \mathbb{E}\left(\left[\varepsilon_{c_i}^{K-1} + \sum_{k=0}^{K-2} \varepsilon_{c_i}^k \zeta^{K-1-k} \right] [\alpha\varepsilon_{c_i} + \eta] \right) \\
&= \alpha\mathbb{E}\left(\varepsilon_{c_i}^K\right) + \mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2} \varepsilon_{c_i}^k \zeta^{K-1-k} \right] [\alpha\varepsilon_{c_i} + \eta] \right).
\end{aligned}$$

Since the last two terms do not involve $\mathbb{E}(\varepsilon_{c_i}^K)$, we can always select some $\mathbb{E}(\varepsilon_{c_i}^K)$ such that

$$\mathbb{E}\left(\varepsilon_{c_i}^K\right) \neq -\frac{\left(\mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2} \varepsilon_{c_i}^k \zeta^{K-1-k}\right]\left[\alpha\varepsilon_{c_i} + \eta\right]\right)\right)}{\alpha}$$

which ensures that $\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right) \neq 0$. ■

C.2 Proof of Corollary 19

Consider $v \in V$ and set $C \subseteq V \setminus \{v\}$. Let $D \in \mathbb{R}^{p\times p}$ such that $D_{s,t} \neq 0$ only if $t \in \overline{\mathrm{an}}(s)$. Suppose $C \not\subseteq \overline{\mathrm{an}}(v)$, but that $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C,S,D)) = 0$ for all $c \in C$. Then for generic B and error moments, $C_1 = C \cap \left[\overline{\mathrm{an}}(v) \setminus \overline{\mathrm{sib}}(v)\right]$,

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1,S,D)) = 0$$

for all $c \in C$.

Proof The proof exactly follows the proof of Lemma 6, except replaces all quantities in G with quantities in $\bar{G}$. ■

C.3 Proof of Lemma 20

Suppose $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \overline{\mathrm{an}}(s) \setminus \overline{\mathrm{sib}}(s)$ for all $s \in V$. Let $v \in V$ be such that we have $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(D)) = 0$ for all $c \in \mathrm{pa}_D(v)$. If $q \in \left(\overline{\mathrm{pa}}(v)\setminus\mathrm{pa}_D(v)\right)\cup\overline{\mathrm{sib}}(v)$, then for generic B and error moments, $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$.

Proof Suppose $q \in \overline{\mathrm{pa}}(v) \setminus \mathrm{pa}_D(v)$. Then there exists a directed path l from q to v such that $l \setminus \{q\} \subseteq \mathrm{irr}(v)$ from q to v. Since $C_v \subseteq \overline{\mathrm{an}}(v) \setminus \overline{\mathrm{sib}}(v)$, then $\mathrm{irr}(v) \cap C_v = \emptyset$ because $\mathrm{irr}(v) \subseteq \overline{\mathrm{sib}}(v)$. Thus, $l \cap C_v = \emptyset$ so Lemma 23 implies that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$ for generic parameters.

Suppose $q \in \overline{\mathrm{sib}}(v)$. Then either $q \in \mathrm{sib}(\mathrm{irr}(v))$ or $\mathrm{irr}(q) \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$. If $q \in \mathrm{sib}(\mathrm{irr}(v))$, then there exists some path l such that $l \setminus \{q\} \subseteq \mathrm{irr}(v)$ from s_1 to v where either $q = s_1$ (if $q \in \mathrm{irr}(v)$) or $q \in \mathrm{sib}(s_1)$ (if $q \in \mathrm{sib}(\mathrm{irr}(v)) \setminus \mathrm{irr}(v)$). In this case, $l \subseteq \mathrm{irr}(v) \cap C_v = \emptyset$, so Lemma 23 implies that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$ for generic parameters. If $\mathrm{irr}(q) \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$, then Lemma 24 implies the desired result. ■

C.4 Proof of Lemma 21

Consider $v \in V$ and sets A, C such that $C \subseteq A \subseteq V \setminus \{v\}$. Suppose $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \overline{\mathrm{an}}(s)\overline{\mathrm{sib}}(s)$ for all $s \in V$. Suppose $u \in C$ and $u \in \overline{\mathrm{sib}}(v)$, then for generic B and error moments, there exists some $q \in C$ such that $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(C,\Sigma,D)\right) \neq 0$.

Proof If $q \in \overline{\mathrm{sib}}(v)$, then either $q \in \mathrm{sib}(\mathrm{irr}(v))$ or $\mathrm{irr}(q) \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$. If $\mathrm{irr}(q) \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$, then Lemma 24 implies the desired result. Now, consider the first case where $q \in \mathrm{sib}(\mathrm{irr}(v))$. Then either $q \in \mathrm{sib}(v)$ or $q \in \mathrm{sib}(\mathrm{irr}(v) \setminus \mathrm{sib}(v)$. If $q \in \mathrm{sib}(v)$, then the desired result is implied by Lemma 8. If $q \in \mathrm{sib}(\mathrm{irr}(v) \setminus \mathrm{sib}(v)$ then there exists some path $l \subseteq \mathrm{irr}(v)$ from s_1 to v where $q \in \mathrm{sib}(s_1)$. Furthermore, $l \subseteq \mathrm{irr}(v) \cap C_v = \emptyset$, so Lemma 23 implies that there exists some $c \in C$ such that $\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(D)\right) \neq 0$ for generic parameters. ■

C.5 Proof of Corollary 22

Suppose $D = \bar{B}$. Then, for $v \in V$ and generic B and error moments, suppose $\overline{\mathrm{pa}}(v) \subseteq C \subseteq \overline{\mathrm{an}}(v) \setminus \overline{\mathrm{sib}}(v)$ and $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)) = 0$ for all $c \in C$. If $q \in C \setminus \overline{\mathrm{pa}}(v)$, the for all $c \in C$

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = 0. \tag{50}$$

If $q \in \mathrm{pa}(v)$, then there exists some $c \in C$ such that

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) \neq 0. \tag{51}$$

Proof Corollary 17 implies for any $q \in C \setminus \overline{\mathrm{pa}}(v)$,

$$\delta_v(C, \mathrm{An}_D(C), \Sigma, D) = \bar{B}_{v,(C\setminus\{q\},q)} = \begin{bmatrix}\bar{B}_{v,(C\setminus\{q\})} & 0\end{bmatrix} = \begin{bmatrix}\delta_v(C \setminus \{q\}, \mathrm{An}_D(C \setminus \{q\}), \Sigma, D) & 0\end{bmatrix} \tag{52}$$

so that

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \Sigma, D)) = \mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C, \Sigma, D)) = 0. \tag{53}$$

Now consider the second statement when $q \in \overline{\mathrm{pa}}(v)$. If $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C\setminus\{q\}, \Sigma, D)) \neq 0$ for some $c \in C \setminus \{q\}$ then the statement trivially holds. Thus, it remains to be shown that $\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C\setminus\{q\}, \Sigma, D)) \neq 0$ when $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C\setminus\{q\}, \Sigma, D)) = 0$ for all $c \in C\setminus\{q\}$. This is directly implied by Lemma 20. ■

C.6 Proof of Lemma 23

Let $D = H_{\mathcal{C}}(\bar{B})$ for some $H_{\mathcal{C}} \in \mathcal{D}$ with $\mathcal{C} = (C_s)_{s \in V}$ such that $C_s \subseteq \overline{\mathrm{an}}(s) \setminus \overline{\mathrm{sib}}(s)$ for all $s \in V$. Suppose there exists some path $l = s_1 \to s_2 \to \ldots s_{|l|-1} \to v$ such that $l \cap C_v = \emptyset$. Further suppose that $u \in \mathrm{sib}(s_1) \setminus l$ and $u \notin C_v$. Then for generic parameters

$$\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_{s_1}(D)\right) \neq 0 \qquad \text{and} \qquad \mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0, \tag{54}$$

so that s_1 and u will not be pruned from $\widehat{\mathrm{sib}}(v)$ by Alg 2. Furthermore, for any C such that $u \in C$, for generic parameters there exists some $c \in C$ such that

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0, \tag{55}$$

so that u will not be certified into $\widehat{\mathrm{pa}}(v)$.

Proof Because $C_v \subseteq \mathrm{an}(v)$, we can write:

$$\gamma_v(D) = \varepsilon_v + \sum_{a \in \mathrm{an}(v)} \zeta_{v,a}\varepsilon_a. \tag{56}$$

We first show that $\zeta_{v,s} \neq 0$ for all $s \in l$. Note that $\zeta_{v,s} = \pi_{v,s} - \sum_{a\in\mathrm{an}(v)} d_{v,a}\pi_{a,s}$ is a rational function of the parameters because $d_{v,a}$, $\pi_{a,s}$, and $\pi_{v,s}$ are rational functions of the parameters. Thus, showing that $\zeta_{v,s} \neq 0$ for specific parameter values implies that it is non-zero for generic parameter values.

Specifically, consider the set of edgeweights where all edges in the path l are set to 1, and all other edges are set to 0. Then, $\pi_{v,s'} = \pi_{s',s''} = 1$, for any $s', s'' \in l$ and $\pi_{v,w'} = \pi_{w',w''} = 0$ if either w' or w'' is not in l. Then, for any $D = H_{\mathcal{C}}(\bar{B})$ where $\mathrm{pa}_D(v) \cap l = \emptyset$, $d_{v,w} = 0$ for all $w \notin l$ since the only non-zero path to v is l. Furthermore, let $\omega_{s,s} = 1$ for all $s \in V$ and let $\omega_{s_1,u} = 1$.

Since $l \cap \mathrm{pa}_D(v) = \emptyset$, we have

$$\begin{aligned}\gamma_v(D) &= Y_v - \sum_{w\in\mathrm{pa}_D(v)} d_{v,w}Y_w \\ &= \sum_{s\in\mathrm{an}(v)} \pi_{v,s}\varepsilon_s - \sum_{w\in\mathrm{pa}_D(v)} d_{v,w}Y_w \\ &= \sum_{s\in l} \pi_{v,s}\varepsilon_s.\end{aligned} \tag{57}$$

Thus, $\zeta_{v,s} = \pi_{v,s} - \sum_{w\in\mathrm{pa}_D(v)} d_{v,w}\pi_{w,s} = \pi_{v,s} = 1$ for all $s \in l$. This implies that $\zeta_{v,s} \neq 0$ for generic B.

Furthermore, since all directed edges pointing into s_1 are 0, then $\gamma_{s_1}(D) = \varepsilon_{s_1}$. For notational convenience, let $\phi = \sum_{s\in l\setminus s_1} \zeta_{v,s}\varepsilon_s$ so that $\gamma_v(D) = \zeta_{v,s_1}\varepsilon_{s_1} + \phi$. Then

$$\begin{aligned}\mathbb{E}(\gamma_v(D)^{K-1}\gamma_{s_1}(D)) &= \mathbb{E}\left((\zeta_{v,s_1}\varepsilon_{s_1} + \phi)^{K-1}\varepsilon_{s_1}\right) \\ &= \mathbb{E}\left(\zeta_{v,s_1}^{K-1}\varepsilon_{s_1}^{K} + \sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k+1}\phi^{K-1-k}\right).\end{aligned} \tag{58}$$

Since $\sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k+1}\phi^{K-1-k}$ does not include any terms with $\varepsilon_{s_1}^K$, we can pick

$$\mathbb{E}(\varepsilon_{s_1}^K) \neq -\frac{\mathbb{E}\left(\sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k+1}\phi^{K-1-k}\right)}{\zeta_{v,s_1}^{K-1}} \tag{59}$$

so that $\mathbb{E}(\gamma_v(D)^{K-1}\gamma_{s_1}(D)) \neq 0$. Because $\mathbb{E}(\gamma_v(D)^{K-1}\gamma_{s_1}(D))$ is a rational function of the model parameters, this implies that $\mathbb{E}(\gamma_v(D)^{K-1}\gamma_u(D))$ is non-zero for generic model parameters.

Similarly, $\gamma_u(D) = \varepsilon_u$ so that

$$\begin{aligned}\mathbb{E}(\gamma_v(D)^{K-1}\gamma_u(D)) &= \mathbb{E}\left((\zeta_{v,s_1}\varepsilon_{s_1} + \phi)^{K-1}\varepsilon_u\right) \\ &= \mathbb{E}\left(\zeta_{v,s_1}^{K-1}\varepsilon_{s_1}^{K-1}\varepsilon_u + \varepsilon_u\sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k}\phi^{K-1-k}\right).\end{aligned} \tag{60}$$

Since $\varepsilon_u\sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k}\phi^{K-1-k}$ does not include any terms with $\varepsilon_{s_1}^{K-1}\varepsilon_u$ we can pick

$$\mathbb{E}(\varepsilon_{s_1}^{K-1}\varepsilon_u) \neq -\frac{\mathbb{E}\left(\varepsilon_u\sum_{k=0}^{K-2} \zeta_{v,s_1}^{k}\varepsilon_{s_1}^{k}\phi^{K-1-k}\right)}{\zeta_{v,s_1}^{K-1}} \tag{61}$$

so that $\mathbb{E}(\gamma_v(D)^{K-1}\gamma_u(D)) \neq 0$. This implies that $\mathbb{E}(\gamma_v(D)^{K-1}\gamma_u(D)) \neq 0$ for generic parameters.

We now show that u will not be certified into $\widehat{\text{pa}}(v)$. Recall that

$$\delta_v(C, A, \Sigma, D) = \left\{\left[(I-D)_{C,A}\Sigma_{A,C}\right]^{-1}(I-D)_{C,A}\Sigma_{A,v}\right\}^T = (\Sigma_{C,C})^{-1}\Sigma_{C,v}. \tag{62}$$

There are no non-zero treks between w and v for any $w \notin l \cup u$, so $\Sigma_{w,v} = 0$ for all $w \in C \setminus u$. Furthermore, $\Sigma_{u,v} = 1$ since there is a single trek between u and v and all edge weights on that trek are 1. Furthermore, $\Sigma_{C,C}$ is a diagonal matrix with $\omega_{c,c} = 1$ on the diagonals (i.e., $\Sigma_{C,C}$ is the identity) since there are no non-zero treks between any nodes in C. Thus, $\delta_v(C, A, \Sigma, D)$ is 0 except for the element corresponding to u which is $1/\omega_{u,u} = 1$.

Thus,

$$\gamma_v(C, \Sigma, D) = \sum_{s \in l} \varepsilon_s - \frac{1}{\omega_{u,u}}\varepsilon_u. \tag{63}$$

For notational convenience, let $\phi = \sum_{s\in l}\varepsilon_s$ and let $\zeta_{v,u} = -1/\omega_{u,u}$. Using a similar argument as before, picking

$$\mathbb{E}(\varepsilon_u^K) \neq -\frac{\mathbb{E}\left(\sum_{k=0}^{K-2}\zeta_{v,u}^k \varepsilon_u^{k+1}\phi^{K-1-k}\right)}{\zeta_{v,u}^{K-1}} \tag{64}$$

implies that $\mathbb{E}(\gamma_v(C, \Sigma, D)^{K-1}\gamma_u(D)) \neq 0$. ■

C.7 Proof of Lemma 24

Suppose $\text{irr}(u) \cap \text{sib}(\text{irr}(v)) \neq \emptyset$ and $D = H_{\mathcal{C}}(\bar{B})$ for some $\mathcal{C} = (C_s)_{s\in V}$ such that $C_s \subseteq \bar{\text{an}}(s) \setminus \bar{\text{sib}}(s)$ for all $s \in V$. Then, for generic parameters

$$\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0 \tag{65}$$

so that u will not be pruned from $\widehat{\text{sib}}(v)$ by Alg 2. Furthermore, for any $C \subseteq V \setminus v$ such that $u \in C$, for generic parameters, there exists some $c \in C$ such that

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, \Sigma, D)\right) \neq 0, \tag{66}$$

so that u will not be certified into $\widehat{\text{pa}}(v)$.

Proof We first show that u will not be pruned out of $\widehat{\text{sib}}(v)$ by Alg 2. We subsequently show that u will not be certified into $\widehat{\text{pa}}(v)$.

Not pruned from $\widehat{\text{sib}}(v)$: Let $q \in \text{irr}(u) \cap \text{sib}(\text{irr}(v))$ and $w \in \text{sib}(q) \cap \text{irr}(v)$. Then there exists a directed path l_1 from w to v such that $l_1 \subseteq \text{irr}(v)$ so that $C_v \cap l_1 = \emptyset$. As shown in Lemma 23, this implies that $\gamma_v(D) = \zeta_{v,w}\varepsilon_w + \phi_v$ where $\zeta_{v,w} \neq 0$ for generic parameters and some term ϕ_v which does not include ε_w. A similar statement can be made for q and u so that $\gamma_u(D) = \zeta_{u,q}\varepsilon_q + \phi_u$. Thus,

$$\begin{aligned}
\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) &= \mathbb{E}\left(\left[\zeta_{v,w}\varepsilon_w + \phi_v\right]^{K-1}\left[\zeta_{u,q}\varepsilon_q + \phi_u\right]\right) \\
&= \mathbb{E}\left(\zeta_{v,w}^{K-1}\varepsilon_w^{K-1}\left[\zeta_{u,q}\varepsilon_q + \phi_u\right] + \left[\zeta_{u,q}\varepsilon_q + \phi_u\right]\sum_{k=0}^{K-2}\zeta_{v,w}^{k}\varepsilon_w^{k}\phi_v^{K-k-1}\right) \\
&= \mathbb{E}\left(\zeta_{v,w}^{K-1}\varepsilon_w^{K-1}\zeta_{u,q}\varepsilon_u + \zeta_{v,w}^{K-1}\varepsilon_v^{K-1}\phi_q\right. \\
&\qquad \left. + \left[\zeta_{u,q}\varepsilon_q + \phi_u\right]\sum_{k=0}^{K-2}\zeta_{v,w}^{k}\varepsilon_w^{k}\phi_v^{K-k-1}\right).
\end{aligned} \tag{67}$$

Since $\zeta_{v,w}^{K-1}\varepsilon_w^{K-1}\phi_u + [\zeta_{u,q}\varepsilon_q + \phi_u]\sum_{k=0}^{K-2}\zeta_{v,w}^{k}\varepsilon_w^{k}\phi_v^{K-k-1}$ does not involve any terms with $\varepsilon_w^{K-1}\varepsilon_q$, we can select

$$\mathbb{E}\left(\varepsilon_w^{K-1}\varepsilon_q\right) \neq -\frac{\mathbb{E}\left(\zeta_{v,w}^{K-1}\varepsilon_w^{K-1}\phi_u + [\zeta_{u,q}\varepsilon_q + \phi_u]\sum_{k=0}^{K-2}\zeta_{v,w}^{k}\varepsilon_w^{k}\phi_v^{K-k-1}\right)}{\zeta_{v,w}^{K-1}\zeta_{u,q}} \tag{68}$$

so that $\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0$. This implies that $\mathbb{E}\left(\gamma_v(D)^{K-1}\gamma_u(D)\right) \neq 0$ for generic parameters.

Not certified into $\widehat{\mathrm{pa}}(v)$: We now show that u will not be certified into $\widehat{\mathrm{pa}}(v)$ so that u will remain in $\widehat{\mathrm{sib}}(v)$. Specifically, we show that any set C will not be certified if $u \in C$. If $u \in \mathrm{sib}(\mathrm{irr}(v))$ or if $C \cap \mathrm{sib}(\mathrm{irr}(v)) \neq \emptyset$, then Lemma 23 directly completes the proof. Thus, it remains to be shown that C will not be certified even if $u \notin \mathrm{sib}(\mathrm{irr}(v))$ and $C \cap \mathrm{sib}(\mathrm{irr}(v)) = \emptyset$. Without loss of generality, assume that $1, \ldots, p$ is a valid causal ordering of V. We consider two cases: (1) $\mathrm{irr}(u) \cap \mathrm{irr}(v) \neq \emptyset$ and (2) $\mathrm{irr}(u) \cap \mathrm{irr}(v) = \emptyset$.

For the first case, let $w = \max(\mathrm{irr}(u) \cap \mathrm{irr}(v))$. Then there exist two directed paths l_1 and l_2 such that l_1 is a directed path from w to v with $l_1 \subseteq \mathrm{irr}(v)$, l_2 is a directed path from w to u with $l_2 \subseteq \mathrm{irr}(u)$, and $\{w\} = l_1 \cap l_2$. Let $s_0 = \min\{s \in C : C_s \cap l_2 = \emptyset\}$ so that s_0 is the most upstream node in l_2 whose currently estimated parent set, C_s, does not contain any other nodes in l_2. The set over which the min is taken is non-empty because $l_2 \subset \mathrm{irr}(u)$ so that $C_u \cap l_2 = \emptyset$. For each $s \in C$, let $m_s = \max(l_2 \cap C_s)$ where $m_s = 0$ if $l_2 \cap C_s$ is empty. Now set the directed edges on l_1 and all directed edges on l_2 before s_0 to $1/p^3$. Set $\omega_{v,v} = 1$ for all $v \in V$, and set all other directed and bidirected edgeweights to 0. Finally, for any node s in l_2, let $L(s)$ denote the position of s in l_2; i.e., if $l = s_1 \to s_2 \to s_3 \ldots$ then $L(s_i) = i$.

Suppose $s \in C \cap l_2$ and $s < s_0$. Since the only non-zero directed edges are within l_2 and $m_s = \max(C_s \cap l_2)$, then all non-zero directed paths from C_s to s must pass through m_s. Thus, for C_s the marginal direct effect is 0 for any $t \neq m_s$ and the marginal direct effect of

m_s is $\pi_{s,m_s} = p^{-3(L(s)-L(m_s))}$. Thus, for $s < s_0$

$$\begin{aligned}\gamma_s(D) &= Y_s - D_{s,C_s}Y_{C_s} = Y_s - \pi_{s,m_s}Y_{m_s} \\ &= \varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'}\varepsilon_{s'} + \pi_{s,m_s}Y_{m_s} - \pi_{s,m_s}Y_{m_s} \\ &= \varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'}\varepsilon_{s'}\end{aligned} \tag{69}$$

and

$$Y_s = \varepsilon_s + \sum_{\substack{s' \in l_2 \\ s' < s}} \pi_{s,s'}\varepsilon_{s'}. \tag{70}$$

Because $C_{s_0} \cap l_2 = \emptyset$, no node in C_{s_0} has a non-zero directed path to s. Thus, $D_{s_0,V}$ is the zero vector and

$$\gamma_{s_0}(D) = Y_{s_0} = \varepsilon_s + \pi_{s_0,w}\varepsilon_w + \sum_{s \in l_2 \setminus w} \pi_{s_0,s}\varepsilon_s. \tag{71}$$

For all other $s \in C$, then either $s \notin l_2$ or $s \in l_2$ but $s > s_0$. Then all directed paths into s have weight 0, so $D_{s,V} = 0$. Thus,

$$\gamma_s(D) = Y_s = \varepsilon_s. \tag{72}$$

Notably, $\gamma_s(D)$ only contains a term with ε_w if $s = s_0$. We now show that under these parameters, when checking the certificate for C, $\delta_{v,s_0}(C, D, \Sigma) \neq 0$.

Note that $(I-D)_{C,A}\Sigma_{A,C} = \mathbb{E}\left((Y_C - D_{C,A}Y_A)Y_C^T\right) = \mathbb{E}\left(\gamma_C Y_C^T\right)$ and $(I-D)_{C,A}\Sigma_{A,v} = \mathbb{E}\left((Y_C - D_{C,A}Y_A)Y_v\right) = \mathbb{E}\left(\gamma_C Y_v\right)$. Then, letting $M = (I-D)_{C,A}\Sigma_{A,C}$ for convenience, we first show that M is diagonally dominant so that it is non-singular and $M_{V\setminus s_0, V \setminus s_0}$ is also non-singular.

For any $s \in C$, since all edgeweights are positive and $\mathbb{E}(\varepsilon_s\varepsilon_r) = 0$ for $s, r \in C$, we have

$$M_{s,s} = \mathbb{E}\left([\varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'}\varepsilon_{s'}][\varepsilon_s + \sum_{\substack{s' \in l_2 \\ s' < s}} \pi_{s,s'}\varepsilon_{s'}]\right) > \mathbb{E}(\varepsilon_s^2) = 1. \tag{73}$$

If $s > s_0$ then $\gamma_s = \varepsilon_s$, but since all directed edges downstream of s are set to 0, then ε_s does not appear in any other $Y_{C\setminus s}$ so $M_{s,C\setminus s} = 0$. Similarly, since $Y_s = \varepsilon_s$, then then ε_s does not appear in any other $\gamma_{C \setminus s}$ so $M_{C\setminus s,s} = 0$. It then remains to characterize $M_{r,s}$ when r, sMs_0. In this case,

$$\begin{aligned}M_{s,r} &= \mathbb{E}\left([\varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'}\varepsilon_{s'}][\varepsilon_r + \sum_{\substack{s' \in l_2 \\ s' < r}} \pi_{r,s'}\varepsilon_{s'}]\right) \\ &= \sum_{\substack{s' \in l_2 \\ s' < r \\ s > s' > m_s}} \pi_{s,s'}\pi_{r,s'}\mathbb{E}\left(\varepsilon_{s'}^2\right) \\ &< p \max_{j,k \in V}(\pi_{j,k}^2) < \frac{1}{p^2}.\end{aligned} \tag{74}$$

Thus, $|M_{s,r}| < \frac{1}{p^2}$ for any $s \neq r$ and $M_{s,s} \geq 1$. This implies that M and $M_{V \setminus s_0, V \setminus s_0}$ are both diagonally dominant so that

$$|(M^{-1})_{s_0,s_0}| = \left| \frac{\det(M_{V \setminus s_0, V \setminus s_0})}{\det(M)} \right| \neq 0. \tag{75}$$

Letting $F = (I - D)_{C,A} \Sigma_{A,v}$, we have:

$$(F)_s = \mathbb{E}(\gamma_s Y_v) = \mathbb{E}\left(\left[\varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'} \varepsilon_{s'} \right] \left[\varepsilon_v + \sum_{s'' \in l_1} \pi_{v,s''} \varepsilon_{s''} \right] \right). \tag{76}$$

Since $l_1 \cap l_2 = w$ and all covariances are set to 0, then $F_s \neq 0$ only if γ_s contains ε_w. Thus, $F_s = 0$ for all $s \neq s_0$ and $F_{s_0} = \mathbb{E}(\pi_{s_0,w} \pi_{v,w} \varepsilon_w^2) \neq 0$. Combining all the results, we then have that the s_0 element of $\delta_{v,C}(C, D, \Sigma)$ is $M_{s_0,s_0} F_{s_0} \neq 0$. Since Y_{s_0} is the only Y which has an ε_{s_0} term, it holds that

$$\gamma_v(C, \Sigma, D) = Y_v - \delta_{v,C} Y_C = \delta_{v,s_0} \varepsilon_{s_0} + \phi_v, \tag{77}$$

where ϕ_v does not involve ε_{s_0}. Similarly, we can write

$$\gamma_{s_0}(D) = Y_{s_0} - D_{s_0,V} Y = \varepsilon_{s_0} + \phi_{s_0}, \tag{78}$$

where ϕ_{s_0} does not involve ε_{s_0}. Then,

$$\begin{aligned} \mathbb{E}\left(\gamma_{s_0}(D)^{K-1} \gamma_v(C, \Sigma, D)\right) &= \mathbb{E}\left([\varepsilon_{s_0} + \phi_{s_0}]^{K-1} [\delta_{v,s_0} \varepsilon_{s_0} + \phi_v] \right) \\ &= \mathbb{E}\left(\varepsilon_{s_0}^{K-1} [\delta_{v,s_0} \varepsilon_{s_0} + \phi_v] + [\delta_{v,s_0} \varepsilon_{s_0} + \phi_v] \sum_{k=0}^{K-2} \varepsilon_{s_0}^k \phi_{s_0}^{K-k-1} \right) \\ &= \mathbb{E}\left(\varepsilon_{s_0}^K \delta_{v,s_0} + \varepsilon_{s_0}^{K-1} \phi_v + [\delta_{v,s_0} \varepsilon_{s_0} + \phi_v] \sum_{k=0}^{K-2} \varepsilon_{s_0}^k \phi_{s_0}^{K-k-1} \right). \end{aligned} \tag{79}$$

Therefore, selecting

$$\mathbb{E}\left(\varepsilon_{s_0}^K\right) \neq -\frac{\mathbb{E}\left(\varepsilon_{s_0}^K \delta_{v,s_0} + \varepsilon_{s_0}^{K-1} \phi_v + [\delta_{v,s_0} \varepsilon_{s_0} + \phi_v] \sum_{k=0}^{K-2} \varepsilon_{s_0}^k \phi_{s_0}^{K-k-1} \right)}{\delta_{v,s_0}} \tag{80}$$

implies that $\mathbb{E}\left(\gamma_{s_0}(D)^{K-1} \gamma_v(C, \Sigma, D)\right) \neq 0$ and thus $\mathbb{E}\left(\gamma_{s_0}(D)^{K-1} \gamma_v(C, \Sigma, D)\right) \neq 0$ for generic parameters.

Now we slightly modify the argument above to the case where $\text{irr}(u) \cap \text{irr}(v) = \emptyset$. Let $w \in \text{sib}(\text{irr}(v)) \cap \text{irr}(u)$, and let $q \in \text{sib}(w) \cap \text{irr}(v)$. Select two paths, l_1 and l_2, such that $l_1 \cap l_2 = \emptyset$ where l_1 is be a path from q to v which only passes through $\text{irr}(v)$ and l_2 to be a path from w to u which only passes through $\text{irr}(u)$. Similar to before, let $s_0 = \min(\{s \in C : C_s \cap l_2 = \emptyset\})$ so that s_0 is the most upstream node in l_2 whose currently estimated parent set, C_s, does not contain any other nodes in l_2. The set over which the min is taken is non-empty because $l_2 \subset \text{irr}(u)$ so that $C_u \cap l_2 = \emptyset$. For all $s \in C \cap l_2$ such that $s < s_0$, let $m_s = \max(l_2 \cap C_2)$ where C_s is the set of sets in $\mathcal{C}$. Now consider the set of parameters where all directed edges on l_1 and all directed edges before s_0 on l_2 are set to $1/p^3$. Let $\mathbb{E}(\varepsilon_w \varepsilon_q) = 1$, and set all other directed and bidirected edgeweights to 0. In addition, set $\omega_{v,v} = 1$ for all $v \in V$.

Using the same argument as before, $M = (I - D)_{C,A}\Sigma_{A,C}$ is diagonally dominant so $(M^{-1})_{s_0,s_0} \neq 0$. Furthermore, letting $F = (I - D)_{C,A}\Sigma_{A,v}$, we have:

$$(F)_s = \mathbb{E}(\gamma_s Y_v) = \mathbb{E}\left(\left[\varepsilon_s + \sum_{\substack{s' \in l_2 \\ s > s' > m_s}} \pi_{s,s'}\varepsilon_{s'}\right]\left[\varepsilon_v + \sum_{s \in l_1} \pi_{v,s}\varepsilon_s\right]\right). \tag{81}$$

Since $l_1 \cap l_2 = \emptyset$ and the only errors with non-zero covariance are w and q, then $(F)_s \neq 0$ only if γ_s contains ε_w. By construction, only γ_{s_0} contains ε_w so F is zero except for the element corresponding to s_0 and $F_{s_0} = \mathbb{E}(\pi_{s_0,w}\varepsilon_w \pi_{v,q}\varepsilon_q) = \pi_{s_0,w}\pi_{v,q}\mathbb{E}(\varepsilon_w\varepsilon_q) \neq 0$. This implies that $\delta(C, \Sigma, D)_{s_0} = (M^{-1})_{s_0,s_0} F_{s_0} \neq 0$.

Since Y_{s_0} is the only Y which has an ε_{s_0} term, we obtain that

$$\gamma_v(C, \Sigma, D) = Y_v - \delta_{v,C} Y_c = \delta_{v,s_0}\varepsilon_{s_0} + \phi_v, \tag{82}$$

where ϕ_v does not include any terms with ε_{s_0}. Similarly, $\gamma_{s_0}(D) = \varepsilon_{s_0} + \phi_{s_0}$ where ϕ_{s_0} does not involve ε_{s_0}. Then,

$$\begin{aligned}
\mathbb{E}\left(\gamma_{s_0}(D)^{K-1}\gamma_v(C,\Sigma,D)\right) &= \mathbb{E}\left([\varepsilon_{s_0} + \phi_{s_0}]^{K-1}[\delta_{v,s_0}\varepsilon_{s_0} + \phi_v]\right) \\
&= \mathbb{E}\left(\varepsilon_{s_0}^{K-1}[\delta_{v,s_0}\varepsilon_{s_0} + \phi_v] + [\delta_{v,s_0}\varepsilon_{s_0} + \phi_v]\sum_{k=0}^{K-2}\varepsilon_{s_0}^{k}\phi_{s_0}^{K-k-1}\right) \\
&= \mathbb{E}\left(\varepsilon_{s_0}^{K}\delta_{v,s_0} + \varepsilon_{s_0}^{K-1}\phi_v + [\delta_{v,s_0}\varepsilon_{s_0} + \phi_v]\sum_{k=0}^{K-2}\varepsilon_{s_0}^{k}\phi_{s_0}^{K-k-1}\right)
\end{aligned} \tag{83}$$

Thus, selecting

$$\mathbb{E}\left(\varepsilon_{s_0}^{K}\right) \neq -\frac{\mathbb{E}\left(\varepsilon_{s_0}^{K}\delta_{v,s_0} + \varepsilon_{s_0}^{K-1}\phi_v + [\delta_{v,s_0}\varepsilon_{s_0} + \phi_v]\sum_{k=0}^{K-2}\varepsilon_{s_0}^{k}\phi_{s_0}^{K-k-1}\right)}{\delta_{v,s_0}}. \tag{84}$$

implies that $\mathbb{E}\left(\gamma_{s_0}(D)^{K-1}\gamma_v(C,\Sigma,D)\right) \neq 0$ and thus $\mathbb{E}\left(\gamma_{s_0}(D)^{K-1}\gamma_v(C,\Sigma,D)\right) \neq 0$ for generic parameters. ■