Structure Learning for Cyclic Linear Causal Models

Carlos Améndola   **Philipp Dettling**   **Mathias Drton**   **Federica Onori** *   **Jun Wu**

Department of Mathematics
Technical University of Munich, Germany

Abstract

We consider the problem of structure learning for linear causal models based on observational data. We treat models given by possibly cyclic mixed graphs, which allow for feedback loops and effects of latent confounders. Generalizing related work on bow-free acyclic graphs, we assume that the underlying graph is simple. This entails that any two observed variables can be related through at most one direct causal effect and that (confounding-induced) correlation between error terms in structural equations occurs only in absence of direct causal effects. We show that, despite new subtleties in the cyclic case, the considered simple cyclic models are of expected dimension and that a previously considered criterion for distributional equivalence of bow-free acyclic graphs has an analogue in the cyclic case. Our result on model dimension justifies in particular score-based methods for structure learning of linear Gaussian mixed graph models, which we implement via greedy search.

1 INTRODUCTION

Inferring the structure of a causal model with feedback loops from observational data is a notoriously difficult—if not impossible—problem, particularly if one also seeks to guard against presence of latent confounders [9, 29]. We consider this problem for linear causal models given by mixed graphs (or path diagrams) with directed and bidirected edges. As detailed in Section 2, the vertices of such a graph correspond to the observed variables, and the directed edges encode structural equations

---
*Department of Statistical Sciences, University of Rome La Sapienza, Italy

that relate these variables up to stochastic noise. The bidirected edges indicate possible correlations among the noise terms, as may be induced by latent confounders.

Much work has gone into algorithms that exploit conditional independence relations for learning the structure of causal models, or rather suitable equivalence classes of graphs encoding this structure; see, e.g., [10, 17, 18, 24, 25] or also the review of Spirtes and Zhang in [21, §18]. While methods have been developed that use information about conditional independence relations also in settings with feedback loops or latent variables, there is an inherent limitation to this approach as causal models with feedback loops or latent variables can generally not be characterized using conditional independence constraints alone [8, 27, 31, 32]. Alternatively, structure learning can be approached using score-based search techniques; see, e.g., [3, 28, 30]. For models with feedback loops and/or latent variables, however, the definition of an appropriate statistical score is non-trivial as the model parameters need not be identifiable and, consequently, the model dimension may differ from the number of parameters that are used to specify the model. Although methods exist to detect identifiability or lack thereof [11, 19, 33], it is generally unclear when a linear causal model with feedback loops or latent variables is of the dimension expected from a parameter count [4].

Motivated by these difficulties, prior work has also made attempts to determine more tractable settings. For instance, it has been shown that if an acyclic mixed graph is bow-free, i.e., no pair of nodes is joined by both a directed and a bidirected edge, then the parameters of the induced model are generically identifiable [2], which entails expected dimension. Score-based methods for structure learning in this setting were proposed in [22].

In this paper we generalize results that have been obtained for bow-free acyclic mixed graphs to graphs that may have a cyclic directed part. In other words, we consider mixed graphs that are simple, i.e., have at most

one edge between any pair of nodes, but which need not be acyclic. The presence of cycles brings about many new challenges [1] and, in particular, generic identifiability cannot be guaranteed. However, as our main results show, the considered simple cyclic models are of expected dimension (Theorem 3.1) and a previously considered criterion for distributional equivalence of bow-free acyclic graphs has an analogue in the cyclic case (Theorem 4.1). Using the result on dimension, we propose a model selection score and associated greedy search techniques for structure learning of linear Gaussian mixed graphs. Numerical experiments indicate the need to carefully account for the large size of the set of models/graphs the search considers (Section 5).

2   BACKGROUND

Let $X = (X_i : i \in V)$ be a random vector whose coordinates correspond to observed variables. The considered models assume X to be the solution to a linear equation system of the form

$$X \ = \ \Lambda^T X + \varepsilon, \tag{1}$$

where $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{V \times V}$ is a matrix of unknown parameters, and $\varepsilon = (\varepsilon_i : i \in V)$ is a random vector whose coordinates represent stochastic noise. Suppose ε has (unknown) covariance matrix $\Omega = (\omega_{ij})$. Assuming that $I - \Lambda$ is invertible (I denotes the identity matrix), the system in (1) is solved uniquely by $X = (I - \Lambda)^{-T}\varepsilon$ with covariance matrix

$$\mathrm{Var}[X] \ = \ (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1} \ =: \ \phi(\Lambda, \Omega). \tag{2}$$

Specific models of interest place restrictions on the support of Λ and Ω, and this is naturally represented by a graph. More precisely, we adopt mixed graphs since we have two parameter matrices, Λ and Ω, whose rows and columns are indexed by the same set V.

A mixed graph with vertex set V is a triple $G = (V, D, B)$ where $D, B \subseteq V \times V$ are two sets of edges. The set D comprises ordered pairs (i, j), $i \neq j$, that encode directed edges, which we also denote by $i \to j$. Node j is the head of such an edge. The elements of B are unordered pairs $\{i, j\}$ with $i \neq j$ that encode bidirected edges, also denoted by $i \leftrightarrow j$. These edges have no orientation, and $i \leftrightarrow j \in B$ if and only if $j \leftrightarrow i \in B$. It is convenient to call both endpoints i and j heads of $i \leftrightarrow j$. A collider triple in $G = (V, D, B)$ is a triple of vertices (i, j, k) such that there are edges between i and j and between j and k, with j being a head on both these edges. In other words, the two edges form a path of the form $i \to j \leftarrow k$, $i \leftrightarrow j \leftarrow k$, $i \to j \leftrightarrow k$ or $i \leftrightarrow j \leftrightarrow k$. We emphasize that whether (i, j, k) is a
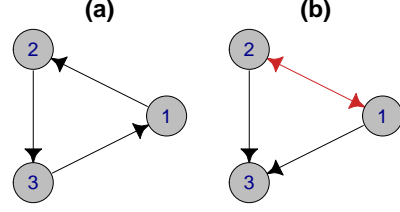

Figure 1: Two simple mixed graphs.

collider triple does not depend on absence or presence of an edge between i and k. Finally, the skeleton of G is the undirected graph obtained by replacing all edges in $(V, D \cup B)$ by undirected edges.

In Section 4 we will use the concept of treks. A trek is a path without collider triples and thus takes the form:

$$v_l^L \leftarrow \cdots \leftarrow v_1^L \leftrightarrow v_1^R \rightarrow \cdots \rightarrow v_r^R, \text{ or}$$
$$v_l^L \leftarrow \cdots \leftarrow v_0 \rightarrow \cdots \rightarrow v_r^R.$$

Each trek can be decomposed in two directed paths, referred to as the right and the left side of the trek. A trek is simple if its left- and right-hand side do not intersect, with the exception of an intersection at the top node v_0 for the second type of trek.

In the sequel, the key assumption we make about the considered graphs is that they be *simple* mixed graphs. This means that we do not allow more than one edge (of any type) between any two nodes. Note that this allows for presence of directed cycles of length at least 3.

Example 2.1. *Two simple mixed graphs are displayed in Figure 1. The first is a directed 3-cycle, which has no collider triples. The second graph has the collider triple* $(1, 3, 2)$. *Both graphs have the same skeleton.*

Let $\mathbb{R}^D$ be the set of real $V \times V$-matrices $\Lambda = (\lambda_{ij})$ with support in D, that is,

$$\mathbb{R}^D = \{ \Lambda \in \mathbb{R}^{V \times V} : \lambda_{ij} = 0 \text{ if } i \to j \notin D \}. \tag{3}$$

Define $\mathbb{R}^D_{\mathrm{reg}}$ to be the subset of matrices $\Lambda \in \mathbb{R}^D$ for which $I - \Lambda$ is invertible. Similarly, let PD be the cone of positive definite symmetric $V \times V$-matrices, and define $PD(B)$ to be the subcone with support over B, that is,

$$PD(B) = \{ \Omega = (\omega_{ij}) \in PD : \omega_{ij} = 0$$
$$\text{if } i \neq j \text{ and } i \leftrightarrow j \notin B \}.$$

Definition 2.2. *The* linear Gaussian model *given by the mixed graph* $G = (V, D, B)$ *is the family of all multivariate normal distributions on* $\mathbb{R}^V$ *with covariance ma-*

trix in

$$\mathcal{M}_G = \left\{ (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} : \right. $$
$$\left. \Lambda \in \mathbb{R}^D_{\text{reg}}, \ \Omega \in PD(B) \right\}.$$

The covariance parametrization *of the model is the map*

$$\phi_G : \mathbb{R}^D \times PD(B) \mapsto PD, \qquad (4)$$
$$(\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

Note that the model leaves the mean vector of the normal distribution unrestricted. Without loss of generality, we may assume it to be zero.

It is well-known that different mixed graphs may induce the same model $\mathcal{M}_G$, which we describe as follows.

Definition 2.3. *Two mixed graphs G_1 and G_2 are distributionally equivalent if $\mathcal{M}_{G_1} = \mathcal{M}_{G_2}$.*

Distributional equivalence is a stronger requirement than Markov equivalence, which only requires the same conditional independence relations. For acyclic simple mixed graphs, distributional equivalence is implied by having the same skeleton and collider triples [22].

3 EXPECTED DIMENSION

Knowing the dimension of the set of covariance matrices $\mathcal{M}_G$ of a linear causal model is crucial for statistical model selection. In particular, the dimension features in model selection scores/information criteria [7]. Because $\mathcal{M}_G$ is defined by the pair of parameter matrices (Λ, Ω), it is natural to expect that its dimension $\dim(\mathcal{M}_G)$ equals the number of unknown parameters, which is $|V| + |D| + |B|$. For general graphs, however, this may fail to be true in subtle ways [11]. Nevertheless, the following theorem shows that this complication does not arise when considering simple graphs.

Theorem 3.1. *If the graph G is simple, then*

$$\dim(\mathcal{M}_G) = |V| + |D| + |B|.$$

Let J_G be the Jacobian of the covariance parametrization ϕ_G. As discussed, e.g., in [13], the dimension of $\mathcal{M}_G$ equals the maximal rank of J_G. Before proving Theorem 3.1 based on this fact, we extend observations from [11] that simplify studying the rank of J_G.

On the domain $\mathbb{R}^D_{\text{reg}} \times PD$, define the map

$$g : (\Lambda, \Sigma) \mapsto (I - \Lambda)^T \Sigma (I - \Lambda). \qquad (5)$$

A positive definite matrix Σ is in $\mathcal{M}_G$ if and only if there exist $\Lambda \in \mathbb{R}^D_{\text{reg}}$ and $\Omega \in PD(B)$ such that $\Sigma = \phi_G(\Lambda, \Omega)$, which holds if and only if $g(\Lambda, \Sigma) = \Omega$. Let

$$N = \{\{i, j\} : i, j \in V, \ i \neq j, \ \{i, j\} \notin B\}.$$

Then $\Sigma \in \mathcal{M}_G$ if and only if there exists $\Lambda \in \mathbb{R}^D_{\text{reg}}$ such that

$$g_{ij}(\Lambda, \Sigma) = \left[(I - \Lambda)^T \Sigma (I - \Lambda) \right]_{ij} = 0 \qquad (6)$$

for all $\{i, j\} \in N$. Consider then the $N \times D$ Jacobian matrix $\mathbf{J}(\Lambda, \Sigma)$ whose entries are the partial derivatives

$$\mathbf{J}(\Lambda, \Sigma)_{\{i,j\},(k,l)} = \frac{\partial g_{ij}(\Lambda, \Sigma)}{\partial \lambda_{kl}} \qquad (7)$$

with $\{i, j\} \in N$ and $k \to l \in D$. For $i \neq j$, g_{ij} is multilinear in Λ and

$$\frac{\partial g_{ij}(\Lambda, \Sigma)}{\partial \lambda_{kl}} = \begin{cases} -[(I - \Lambda)^T \Sigma]_{jk} & \text{if } l = i, \\ -[(I - \Lambda)^T \Sigma]_{ik} & \text{if } l = j, \\ 0 & \text{if } l \notin \{i, j\}. \end{cases} \qquad (8)$$

Lemma 3.2. *For $\Lambda \in \mathbb{R}^D_{\text{reg}}$, $\Omega \in PD(B)$, let $\Sigma = \phi_G(\Lambda, \Omega)$. Then the rank of the Jacobian $J_G(\Lambda, \Omega)$ is equal to*

$$\text{rank} \left(\mathbf{J}(\Lambda, \Sigma) \right) + |B| + |V|.$$

Proof. On $\mathbb{R}^D_{\text{reg}} \times PD(B)$, define the map

$$h : (\Lambda, \Omega) \mapsto (\Lambda, \phi_G(\Lambda, \Omega)).$$

Composing with g from (5), we have

$$(g \circ h)(\Lambda, \Omega) = \Omega. \qquad (9)$$

Differentiating this equation with respect to the free entries (i.e., nonzero) in Λ gives

$$\frac{\partial}{\partial \Lambda} g(\Lambda, \Sigma)|_{\Sigma = \phi_G(\Lambda, \Omega)}$$
$$+ \frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)|_{\Sigma = \phi_G(\Lambda, \Omega)} \frac{\partial}{\partial \Lambda} \phi_G(\Lambda, \Omega) = 0. \quad (10)$$

Similarly, differentiating with respect to the free entries of Ω gives

$$\frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)|_{\Sigma = \phi_G(\Lambda, \Omega)} \frac{\partial}{\partial \Omega} \phi_G(\Lambda, \Omega) = \begin{pmatrix} 0 \\ I_{|B| + |V|} \end{pmatrix}. \quad (11)$$

Here, the rows are indexed by unordered pairs $\{i, j\}$, due to the symmetry of the matrices in (9). In the partitioning of the rows, the pairs in N are listed first.

By (8), again ordering the pairs in N first, we have

$$\frac{\partial}{\partial \Lambda} g(\Lambda, \Sigma) = \begin{pmatrix} \mathbf{J}(\Lambda, \Sigma) \\ 0 \end{pmatrix}.$$

For the Jacobian $J_G = (\frac{\partial \phi_G}{\partial \Lambda}, \frac{\partial \phi_G}{\partial \Omega})$, we obtain that

$$\frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)|_{\Sigma = \phi_G(\Lambda, \Omega)} \cdot J_G(\Lambda, \Omega) =$$
$$\begin{pmatrix} -\mathbf{J}(\Lambda, \Sigma)|_{\Sigma = \phi_G(\Lambda, \Omega)} & 0 \\ 0 & I_{|B|+|V|} \end{pmatrix}, \quad (12)$$

with rows and columns partitioned as $(N, B \cup V)$ and $(D, B \cup V)$, respectively. Restricting g by fixing $\Lambda \in \mathbb{R}^D_{\text{reg}}$ gives the bijection $\Sigma \mapsto (I - \Lambda)^T \Sigma (I - \Lambda)$. Hence, the matrix $\frac{\partial g}{\partial \Sigma}$ is invertible. It follows that the rank of J_G equals the rank of the partitioned matrix on the right-hand side of (12), which in turn equals $\text{rank}\,(\mathbf{J}\,(\Lambda, \Sigma)) + |B| + |V|$, as was our claim. $\square$

We now consider the Jacobian $J_G(0, I)$ obtained by specializing $\Lambda = 0$ and $\Omega = I$.

Lemma 3.3. *The Jacobian $J_G(0, I)$ has full column rank $|V| + |B| + |D|$ if and only if the mixed graph G is simple.*

Proof. The choice $\Lambda = 0$ and $\Omega = I$ yields covariance matrix $\Sigma = \phi_G(0, I) = I$. By Lemma 3.2, it thus suffices to show that the matrix $\mathbf{J}(0, I)$ defined in (8) has full column rank $|D|$ if and only if G is simple.

Suppose G is simple. Let $k \to l \in D$ be any directed edge indexing a column of $\mathbf{J}(0, I)$. With $\Lambda = 0$ and $\Sigma = I$, we have $-(I - \Lambda)^T \Sigma = -I$ and the considered column contains precisely one nonzero entry, namely,

$$\mathbf{J}(0, I)_{\{k.l\}, (k,l)} = -1;$$

note that $k \to l \in D$ implies $\{k, l\} \in N$ if G is simple. Arranging the row indices $\{k, l\}$ for $k \to l \in D$ first, it becomes evident that $\mathbf{J}(0, I)$ has rank $|D|$ as

$$\mathbf{J}(0, I) = \begin{pmatrix} -I_D \\ 0 \end{pmatrix}.$$

Conversely, suppose that G is not simple, with k and l being two nodes joined by at least two edges. Without loss of generality, assume that one of these edges is $k \to l \in D$. We distinguish two cases. First, suppose $k \leftrightarrow l \in B$. Then $\{k, l\} \notin N$. Therefore, the column of $\mathbf{J}(0, I)$ that is indexed by (k, l) is zero, which implies that the rank of $\mathbf{J}(0, I)$ is smaller than $|D|$. Second, suppose $k \leftrightarrow l \notin B$ but $l \to k \in D$. Then the two columns of $\mathbf{J}(0, I)$ indexed by (k, l) and (l, k) are identical. Again the rank of $\mathbf{J}(0, I)$ is smaller than $|D|$. $\square$

Proof of Theorem 3.1. If G is simple, then Lemma 3.3 implies that the maximal rank of the Jacobian J_G is $|V| + |D| + |B|$, which equals the count of parameters. $\square$

For simple acyclic mixed graphs, having the same skeleton is a necessary condition for distributional equivalence [22], but this condition is not necessary for cyclic graphs [14]. However, by Theorem 3.1, two distributionally equivalent simple cyclic mixed graphs must at least have the same number of edges.

4 SUFFICIENT CONDITIONS FOR DISTRIBUTIONAL EQUIVALENCE

In this section, we show that the sufficient condition for distributional equivalence from [22] admits an extension to our setting of possibly cyclic graphs. To this end, we define $\overline{\mathcal{M}_G}$ to be the closure of $\mathcal{M}_G$ (in Euclidean topology). Two mixed graphs G_1 and G_2 are then distributionally equivalent up to closure if $\overline{\mathcal{M}_{G_1}} = \overline{\mathcal{M}_{G_2}}$.

Theorem 4.1. *Let G_1 and G_2 be two simple mixed graphs with same skeleton and collider triples. Then G_1 and G_2 are distributionally equivalent up to closure.*

Note that the likelihood functions of two models that are equal up to closure have the same supremum.

While our proof of Theorem 4.1 (developed in §4.1-4.2) concludes equality up to closure, we do not know any examples where the models are not exactly equal. The condition in Theorem 4.1 is also far from being necessary, e.g., it does not include the well-known characterization of Markov equivalence for directed acyclic graphs (DAGs). However, the theorem is useful to assert equivalence in our simulations (Section 6.1). We are also not aware of better (tractable) conditions in the literature. Indeed, distributional equivalence for cyclic mixed graphs is a subtle problem as the following example shows.

Example 4.2. *Let G_1 and G_2 be the two simple mixed graphs displayed in Figure 1(a) and (b), respectively. By Theorem 3.1, both $\mathcal{M}_{G_1}$ and $\mathcal{M}_{G_2}$ are full-dimensional (i.e., 6-dimensional) subsets of the cone of positive definite 3×3 matrices. Graph G_2 is acyclic, and $\mathcal{M}_{G_2}$ is easily seen to be equal to PD. However, as observed in [5], the set $\mathcal{M}_{G_1}$ is a strict subset of $\mathcal{M}_{G_2} = PD$.*

4.1 Useful Lemmas

Let $G_1 = (V, D_1, B_1)$ and $G_2 = (V, D_2, B_2)$ be two mixed graphs. Let $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$ be parameters for G_1. The essence of the proof of Theorem 4.1 is a strategy to find parameters $(\Lambda_2, \Omega_2) \in \mathbb{R}^{D_2}_{\text{reg}} \times PD(B_2)$ such that $\Phi_{G_2}(\Lambda_2, \Omega_2) = \Phi_{G_1}(\Lambda_1, \Omega_1)$. The key steps of the construction are a reduction to correlation matrices and an edge-relabeling considered in the acyclic case by [22]. However, the cyclic case brings about new subtleties in this approach.

Let $\mathcal{R} : PD \rightarrow PD$ be the standardization map that takes covariance matrices to correlation matrices via $\mathcal{R}(\Sigma)_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$.

Lemma 4.3. *Let $G = (V, D, B)$ be simple and $\Sigma \in PD$. Then $\Sigma \in \mathcal{M}_G$ if and only if $\mathcal{R}(\Sigma) \in \mathcal{M}_G$.*

Proof. We show one direction as the converse can be verified similarly. If $\Sigma \in \mathcal{M}_G$ then

$$\Sigma = \Phi_G(\Lambda, \Omega) = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}.$$

for some matrices $\Lambda \in \mathbb{R}^D_{\text{reg}}, \Omega \in PD(B)$. Setting Δ diagonal with entries $\Delta_{ii} = \Sigma_{ii}^{-\frac{1}{2}}$ it holds that

$$
\begin{aligned}
\mathcal{R}(\Sigma) &= \Delta\Sigma\Delta \\
&= (\Delta^{-1} - \Delta^{-1}\Lambda)^{-T}\Omega(\Delta^{-1} - \Delta^{-1}\Lambda)^{-1} \\
&= \Phi_G(\tilde{\Lambda}, \tilde{\Omega})
\end{aligned}
$$

with $\tilde{\Lambda} = \Delta^{-1}\Lambda\Delta \in \mathbb{R}^D_{\text{reg}}, \tilde{\Omega} = \Delta\Omega\Delta \in PD(B)$. $\square$

Throughout the rest of this section, let $G_1 = (V, D_1, B_1)$ and $G_2 = (V, D_2, B_2)$ be two mixed graphs. If the graphs have the same skeleton, then there is a natural way to copy the edge labels from one graph to the other. To describe the procedure, we decompose an error covariance matrix, Ω, into its diagonal and off-diagonal parts, denoted Ω^d and Ω^{od}, respectively. So, $\Omega = \Omega^d + \Omega^{od}$.

Definition 4.4. *Let G_1 and G_2 be simple mixed graphs with the same skeleton. Given a choice $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$, the induced edge labeling on G_2 is the pair of matrices $(\Lambda_2, \Omega_2^{od})$ obtained as*

$$(\Lambda_2)_{ij} = \begin{cases} (\Lambda_1)_{ij} & \text{if } i \rightarrow j \in G_1, \, i \rightarrow j \in G_2, \\ (\Lambda_1)_{ji} & \text{if } i \leftarrow j \in G_1, \, i \rightarrow j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1, \, i \rightarrow j \in G_2, \\ 0 & \text{if } i \rightarrow j \notin G_2, \end{cases}$$

$$(\Omega_2^{od})_{ij} = \begin{cases} (\Lambda_1)_{ij} & \text{if } i \rightarrow j \in G_1, i \leftrightarrow j \in G_2, \\ (\Lambda_1)_{ji} & \text{if } i \leftarrow j \in G_1, i \leftrightarrow j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1, i \leftrightarrow j \in G_2, \\ 0 & \text{if } i \leftrightarrow j \notin G_2 \text{ or } i = j. \end{cases}$$

For the construction from Definition 4.4, it holds that $\Lambda_2 \in \mathbb{R}^{D_2}_{\text{reg}}$. Moreover, Ω_2^{od} can be turned into a matrix in $PD(B_2)$ by addition of a diagonal matrix.

Lemma 4.5. *Let G_1 and G_2 be simple mixed graphs with same skeleton and collider triples, and let $(\Lambda_i, \Omega_i) \in \mathbb{R}^{D_i} \times PD(B_i)$ for $i = 1, 2$. If $(\Lambda_2, \Omega_2^{od})$ equals the edge labeling induced by (Λ_1, Ω_1) then*

$$\det(I - \Lambda_1) = \det(I - \Lambda_2).$$

In particular, if $\Lambda_1 \in \mathbb{R}^{D_1}_{\text{reg}}$ then $\Lambda_2 \in \mathbb{R}^{D_2}_{\text{reg}}$.

Proof. The determinants depend on the values of cycle products [6, Lemma 1]. Let S_V be the group of permutations of the nodes in V. For $\sigma \in S_V$, let $V(\sigma)$ be the set of nodes contained in a non-trivial cycle of σ. Then

$$\det(I - \Lambda) = \sum_{\sigma \in S_V(G)} (-1)^{\text{sgn}(\sigma)} \prod_{i \in V(\sigma)} \Lambda_{\sigma(i),i} \quad (13)$$

where $S_V(G)$ is the subset of permutations such that $i = \sigma(i)$ or $i \rightarrow \sigma(i) \in D$ for all $i \in V$. We remark that even though the lemma in [6] is stated for $\Lambda \in \mathbb{R}^D_{\text{reg}}$, the proof relies on Laplace expansion of the determinant which holds even if $I - \Lambda$ is not invertible.

Now, since collider triples are preserved, an edge that is part of a directed cycle of G_1 cannot be bidirected in G_2. Furthermore, if G_1 contains a cycle which has a directed edge that is reversed in G_2, then the cycle must be chordless in G_1 (that is, every node in the cycle can have only one child in the cycle) and must be fully reversed in G_2. Since the labels agree, the cycle products in (13) remain unchanged and therefore $\det(I-\Lambda_1) = \det(I-\Lambda_2)$. $\square$

4.2 Constructing Covariance Matrices

The key to completing the proof of Theorem 4.1 is to show that a correlation matrix obtained from a generic choice of parameters $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$ also belongs to $\mathcal{M}_{G_2}$. Let $\circ$ denote the Hadamard (entrywise) product of matrices, and define $\mathcal{H} : \mathbb{R}^D_{\text{reg}} \rightarrow \mathbb{R}^{V \times V}$ by

$$\mathcal{H}(\Lambda) := (I - \Lambda)^{-1} \circ (I - \Lambda)^{-1}.$$

We denote the spectral radius of a matrix Λ by $\rho(\Lambda)$.

Lemma 4.6. *Let G_1, G_2 be simple mixed graphs with same skeleton and collider triples. Let $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{reg} \times PD(B_1)$ such that $\Sigma = \Phi_{G_1}(\Lambda_1, \Omega_1) \in \mathcal{M}_{G_1}$ is a correlation matrix and consider the induced edge labeling $(\Lambda_2, \Omega_2^{od})$. If*

(i) $\rho(\Lambda_j) < 1$ for $j = 1, 2$, and

(ii) $\det(\mathcal{H}(\Lambda_2)) \neq 0$,

then there exists a unique diagonal matrix Ω_2^d such that with $\Omega_2 = \Omega_2^d + \Omega_2^{od}$ it holds that $(\Lambda_2, \Omega_2) \in \mathbb{R}^{D_2}_{reg} \times PD(B_2)$ and $\Sigma = \Phi_{G_2}(\Lambda_2, \Omega_2) \in \mathcal{M}_{G_2}$.

Proof. By Lemma 4.5, we have indeed that $\Lambda_2 \in \mathbb{R}^{D_2}_{\text{reg}}$. We need to construct Ω_2^d such that

$$
\begin{aligned}
\Phi_{G_2}(\Lambda_2, \Omega_2) &= (I - \Lambda_2)^{-T}\Omega_2^d(I - \Lambda_2)^{-1} \\
&\quad + (I - \Lambda_2)^{-T}\Omega_2^{od}(I - \Lambda_2)^{-1} = \Sigma.
\end{aligned}
$$

Since $\Sigma_{ii} = 1$, this requires for all $i \in V$ that

$$((I - \Lambda_2)^{-T} \Omega_2^d (I - \Lambda_2)^{-1})_{ii}$$
$$= 1 - ((I - \Lambda_2)^{-T} \Omega_2^{od} (I - \Lambda_2)^{-1})_{ii}.$$

Solving for the diagonal of Ω_2^d is equivalent (see [16, Lemma 5.1.3]) to the linear system $Ax = b$ where

$$A = \mathcal{H}(\Lambda_2) = (I - \Lambda_2)^{-1} \circ (I - \Lambda_2)^{-1}$$

and the coordinates of the vector b are

$$b_i = 1 - ((I - \Lambda_2)^{-T} \Omega_2^{od} (I - \Lambda_2)^{-1})_{ii}.$$

By hypothesis, $\det(\mathcal{H}(\Lambda_2)) \neq 0$ and the system has a unique solution. It thus remains to show that $\Phi_{G_2}(\Lambda_2, \Omega_2)$ also matches Σ in all off-diagonal entries.

In general, if $\Phi(\Lambda, \Omega)$ is a correlation matrix over a mixed graph G and $\rho(\Lambda) < 1$, by [22, Theorem 4], the entries for $i \neq j$ are given by

$$\Phi_G(\Lambda, \Omega)_{ij} = \sum_{\tau \in S^{ij}} \prod_{s \to t \in \tau} \Lambda_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st}, \qquad (14)$$

where S_G^{ij} is the set of simple treks from i to j. By assumption, $\rho(\Lambda_1), \rho(\Lambda_2) < 1$, and we may apply the representation in (14) to G_1 and G_2. In general, $S_{G_1}^{ij} \neq S_{G_2}^{ij}$. However, the fact that the graphs have the same skeleton and share collider triples implies that when replacing (Λ, Ω) by (Λ_j, Ω_j), $j = 1, 2$, in (14), the induced edge labeling guarantees that the right hand sides of the expression are equal. Hence,

$$\Phi_{G_1}(\Lambda_1, \Omega_1) = \Sigma = \Phi_{G_2}(\Lambda_2, \Omega_2)$$

as was the claim. □

With these preparations in place, we may complete the proof of the main result of this section.

Proof of Theorem 4.1. First, observe that the covariance parametrization

$$\Phi_{G_1} : \mathbb{R}^{D_1} \times PD(B_1) \to \mathcal{M}_{G_1} \subseteq PD$$

is a rational map. Next, consider the algebraic map

$$\Psi : \mathcal{U} \subset \mathbb{R}^{D_1} \times PD(B_1) \to \mathcal{M}_{G_2} \subseteq PD$$

defined as follows. First, apply the standardization map on the parameter (Λ_1, Ω_1) to obtain $(\tilde{\Lambda}_1, \tilde{\Omega}_1, \Delta)$, as in the proof of Lemma 4.3. We denote this map by $\tilde{\mathcal{R}}$. As $(\tilde{\Lambda}_1, \tilde{\Omega}_1)$ define a correlation matrix we may obtain $(\tilde{\Lambda}_2, \tilde{\Omega}_2)$ from the procedure in Lemma 4.6, for representation of the same correlation matrix. Finally, de-standardize $(\tilde{\Lambda}_2, \tilde{\Omega}_2)$ with the matrix Δ from the standardization map, and apply Φ_{G_2}.

$$(\Lambda_1, \Omega_1) \xrightarrow{\tilde{\mathcal{R}}} (\tilde{\Lambda}_1, \tilde{\Omega}_1, \Delta) \xrightarrow{H} (\tilde{\Lambda}_2, \tilde{\Omega}_2, \Delta)$$

$$\downarrow \tilde{\mathcal{R}}^{-1}$$

$$(\Lambda_2, \Omega_2)$$

$$\downarrow \Phi_{G_2}$$

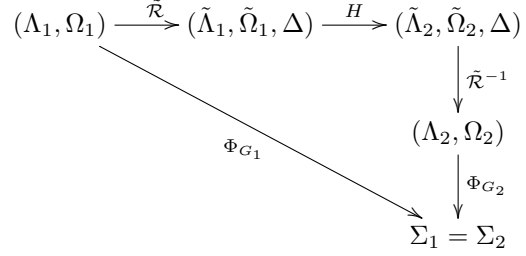$$\Phi_{G_1} \qquad \Sigma_1 = \Sigma_2$$

Figure 2: Commutative diagram illustrating two ways of obtaining a matrix $\Sigma \in \mathcal{M}_{G_1} \cap \mathcal{M}_{G_2}$. One is the parametrization Φ_{G_1}, while the other is a composition of maps (including the map H defined via Lemma 4.6), that we denote Ψ.

Note that the map Ψ is well-defined for input that satisfies the two conditions in Lemma 4.6. This domain includes an open subset $\mathcal{U} \subset \mathbb{R}^{D_1} \times PD(B_1)$. This subset is nonempty because $(0, I) \in \mathcal{U}$. The final application of Φ_{G_2} to (Λ_2, Ω_2) gives a matrix in $\mathcal{M}_{G_2}$, which by construction and Lemma 4.6 coincides with $\Phi_{G_1}(\Lambda_1, \Omega_1)$. The diagram in Figure 2 illustrates the situation.

The map Ψ is a composition of a rational map with algebraic maps that involve radicals (i.e., square roots in the standardization $\mathcal{R}$). Since Ψ coincides with the rational map Φ_{G_1} on the open set $\mathcal{U}$, they must be equal outside of an algebraic hypersurface (i.e., the zero set of a multivariate polynomial). This exceptional set has Lebesgue measure zero (see, e.g., the lemma in [23]). Covariance matrices in $\mathcal{M}_{G_1}$ that are given by parameters (Λ_1, Ω_1) outside the exceptional set are also in $\mathcal{M}_{G_2}$. We may conclude that $\mathcal{M}_{G_1} \subseteq \overline{\mathcal{M}_{G_2}}$ because the elements of the exceptional set are limits of sequences off the exceptional set. By symmetry, $\overline{\mathcal{M}_{G_1}} = \overline{\mathcal{M}_{G_2}}$ as claimed. □

5 GREEDY SEARCH

Based on our result on the dimension of models given by simple mixed graphs, we may form scores that trade off model dimension and model fit. For model selection, we may then maximize such a score over the considered set of graphs. Given the large number of possible graphs, we follow prior work and consider a greedy search that starts from some initial graph and iteratively selects the highest-scoring graph from a local neighborhood of graphs. The procedure stops when no higher score can be found in the local neighborhood or a fixed maximum number of iterations is reached. To mitigate getting trapped in local optima, the search is running from different (random) starting points. In the present context, we take the local neighborhood of a graph G to be the union of all simple mixed graphs that can be obtained from G by adding one edge, by removing one edge, or by reversing one directed edge; compare [22].

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix, assumed to hold in its rows the realizations of n i.i.d. and centered Gaussian random vectors. Let $S = \mathbf{X}^T \mathbf{X}/n$ be the sample covariance matrix. The Gaussian log-likelihood function is

$$\ell(\Sigma; S) = -\frac{n}{2}\left[\log \det(2\pi\Sigma) + \operatorname{tr}(\Sigma^{-1}S)\right].$$

Our proposed score for a mixed graph $G = (V, D, B)$ then takes the form

$$s(G) = \frac{1}{n}\left(\max_{\Sigma \in \mathcal{M}_G} \ell(\Sigma; S) - \operatorname{penalty}(p, k, n)\right), \quad (15)$$

where $p = |V|$ and $k = |D| + |B|$ is the number of edges. To compute the maximum log-likelihood in (15) we apply the block coordinate-descent algorithm from [6]. The standard Bayesian information criterion (BIC) uses $\operatorname{penalty}(p, k, n) = \frac{1}{2}(p + k)\log(n)$; here $p + k$ is the model dimension. The authors of [22] double this penalty when searching over acyclic simple mixed graphs as it improved performance in experiments. An increased penalty is supported by related work on selecting sparse graphical models [12].

There the increased penalty is induced from a prior distribution over graphs under which the number of edges is uniformly distributed. Adopting these ideas in our case we propose to define the score $s(G)$ from (15) with

$$\operatorname{penalty}(p, k, n) = \frac{1}{2}(p + k)\log(n) + \log(p^{2k}3^k). \quad (16)$$

The last term reflects that there are $\binom{p(p+1)/2}{k}3^k \sim p^{2k}3^k$ simple mixed graphs with k edges. This penalty is formulated with a view towards sparser graphs as encountered in the application we consider in Section 6.2. We will also explore its use in simulated non-sparse problems of small scale (Section 6.1). Note also that the proposed penalty ignores the issue of distributional equivalence, i.e., different graphs inducing the same model $\mathcal{M}_G$. Unfortunately this equivalence issue is still poorly understood.

6 NUMERICAL EXPERIMENTS

In this section we present numerical experiments, in which we apply the proposed greedy search to simulated data and well-known protein expression data [26].

6.1 Simulation Studies

We consider graphs with $p \in \{5, 6\}$ nodes. In each case generate 100 simple mixed graphs uniformly at random using an MCMC algorithm; in analogy to [22]. The parameters for the graphs' edges are sampled uniformly from $[-0.9, -0.5] \cup [0.5, 0.9]$. The diagonal entries in

Table 1: Proportion of estimated graphs that share the dimension (Dim), skeleton (Skel) and both skeleton and set of collider triples (Skel & Coll) with the true graph, and minimal structural hamming distance (SHD*) averaged over simulations. Estimates use BIC with standard (1) and increased penalty (2), and search initialized at random (R) or at the true graph (TG).

BIC	n	Start	Dim	Skel	Skel & Coll	SHD*
1	10^2	R	**0.39**	**0.13**	**0.07**	**3.79**
		TG	0.8	0.8	0.25	1.15
	10^3	R	**0.63**	**0.43**	**0.26**	**2.44**
		TG	0.88	0.88	0.53	0.63
	10^4	R	**0.76**	**0.59**	**0.45**	**2.29**
		TG	0.92	0.92	0.74	0.34
2	10^2	R	**0.24**	**0.14**	**0.1**	**3.55**
		TG	0.92	0.92	0.36	1.03
	10^3	R	**0.48**	**0.34**	**0.21**	**2.78**
		TG	0.9	0.9	0.52	0.65
	10^4	R	**0.71**	**0.61**	**0.38**	**2.02**
		TG	0.93	0.93	0.71	0.42

Ω are set by adding independent χ_1^2 draws to the absolute row sums to ensure positive definiteness by diagonal dominance. For each graph, we generate three Gaussian data sets of size $n \in \{10^2, 10^3, 10^4\}$.

For each data set, we run the greedy search starting from (i) 300 randomly selected graphs but also from (ii) the true graph. For every single restart, we set the maximum number of iterations of greedy search to be 10^4. In Table 1 we report on how often the greedy search gives a model of correct dimension, correct graph skeleton, and both correct collider triples and skeleton when $p = 5$. For each initialization scheme, we consider the BIC with both standard penalty and penalty as in (16). Furthermore, in Table 2, the frequency distribution of the difference between the dimension for the true graph and for the estimated graph is reported for the case $p = 5$ and BIC with standard penalty.

In Table 1, we also report a structural Hamming distance (SHD); counting edge additions, deletions and reversals needed to move from one mixed graph to another. The distance we give uses our theorems on dimension and collider triples to bound the true minimal SHD between a graph representing the selected model and one representing the true model. In other words, we minimize the SHD over pairs of graphs $(\bar{G}_1, \bar{G}_2)$, where $\bar{G}_1$ has the same skeleton and collider triples as the true graph and $\bar{G}_2$ has the same skeleton and collider triples as the estimated graph. While this upper bound SHD* needs not be tight, it is on average only half as large as a naive SHD computed for estimated and true graph directly.

By Theorem 3.1, the frequency of having the same dimension gives an upper bound for the frequency of getting equivalent models. Additionally, Theorem 4.1 indicates that the frequency of having both the same set of

Table 2: Absolute frequency distribution of the difference between the dimension for the true graph (TG) and the dimension of the estimated graph (EST) in 100 simulations for p=5 and BIC with standard penalty.

n	Start	Dim(EST) - Dim(TG)					
		-3	-2	-1	0	1	2
10^2	R	5	14	35	39	7	0
	TG	0	0	0	80	20	0
10^3	R	1	3	25	63	8	0
	TG	0	0	0	88	12	0
10^4	R	0	2	14	76	8	0
	TG	0	0	0	97	7	1

Table 3: Summary statistics on the number of edges in the estimated graphs for the 14 protein expression datasets.

Type of edges	Min	Median	Max
All	10	13	13
Directed	2	8	11
Bidirected	2	4.5	8

collider triples and the same skeleton is a lower bound for the frequency of getting equivalent models. Hence, according to our experiment with 100 simulations, when starting from the true graph, the estimated graph is distributionally equivalent to the true graph between 74% and 92% of times when $p = 5$ and $n = 10^4$ (BIC with standard penalty, see Table 1) and between 61% and 80% when $p = 6$ (results not given in Table). On the other hand, if the greedy search algorithm is started from a random graph, the estimated graph belongs to the equivalence class of the true graph between 38% and 57% of times when $p = 5$ and between 21% and 60% when $p = 6$. The standard penalty seems to slightly outperform the increased penalty when $p \in \{5, 6\}$, although this is not so evident from the upper bound on the true minimal structural Hamming distance. Additional partial experiments we carried out for $p = 10$ suggest that for a larger number of nodes, smaller Hamming distances result from the increased penalty.

6.2 Protein Expression Data

For further illustration, we consider a frequently studied collection of data sets on expression of $p = 11$ proteins in human T-cells [26]. The collection comprises 14 data sets, each obtained under different experimental conditions. The sample size of these data sets ranges from 707 to 927. The focus on 11 proteins is due to limitations in the experimental technology. Figure 2 in [26] suggests presence of further relevant but unobserved proteins and leaves open the possibility of a feedback cycle.

To accommodate departures from our linear Gaussian models, at least at the level of marginal distributions, we consider a Gaussian copula version of the models [15, 20]. In other words, each observed variable is assumed to be a deterministic and isotonic function of a Gaussian latent variable. As shown in [20], consistent estimation in the Gaussian copula models is achieved by replacing the sample covariance matrix in the Gaussian

likelihood function (and the model selection score) by a bias-corrected Kendall's tau correlation matrix. The entries of this matrix are $\sin(\frac{\pi}{2}\hat{\tau}_{ij})$, where $\hat{\tau}_{ij}$ is Kendall's τ for the pair of variables (X_i, X_j). With this substitution, we actually project a $\frac{p(p+1)}{2}$-dimensional covariance matrix to the $\frac{p(p-1)}{2}$-dimensional space of correlation matrices. However, our sufficient condition for equivalence remains unchanged, and the result on expected dimension still holds as long as the graph has no more than $\frac{p(p-3)}{2}$ edges.

For each dataset, the greedy search based on the bias-corrected Kendall's tau matrix was repeated 100 times, each time starting from a random graph and using BIC with increased penalty. The highest scoring graph for each data set was then determined. As reported in Table 3, the total number of edges in the 14 estimated graph ranges from 10 to 13, with the median being 13. The table also gives these statistics for the count of directed and bidirected edges. The proportion of directed edges in each graph ranges from a minimum of 0.2 to a maximum of 0.85. A total of 4 of the 14 graphs contain a directed cycle: datasets 3, 6 and 7 each yield a graph with one 3-cycles, and dataset 4 leads to one 4-cycle.

We display two of the selected graphs in Figure 3, one with minimum (dataset 1) and one with maximum (dataset 6) number of edges, the latter also displaying a 3-cycle. Although further work is needed to fully determine possible equivalences, there is no obvious reason (e.g., by Theorem 4.1) for a distributionally equivalent graph without a cycle to exist. We conjecture that this is indeed not the case. Considering all 14 graph estimates together it is reassuring to observe that some structure is shared. Figure 4 shows the (undirected) edges that appear in all/at least 11 of the skeletons of the estimated graphs.

Our selected graphs show good agreement with regulatory relationships described in [26], e.g., the interplay PLCG-PIP2-PIP3 (in at least 12 of the inferred graphs); the connection PKC-P38-PJNK (all 14 graphs); the connection P44/42 (named ERK in [26]) and PKA-PAKTS473 (named AKT in [26], all 14 graphs). Moreover, three expected relationships that are well-reported from the field-related literature emerge in our work that were undetected in [26]. This is the case for the connections: PIP2 to PKC (dataset 14), PLCG to PKC (dataset
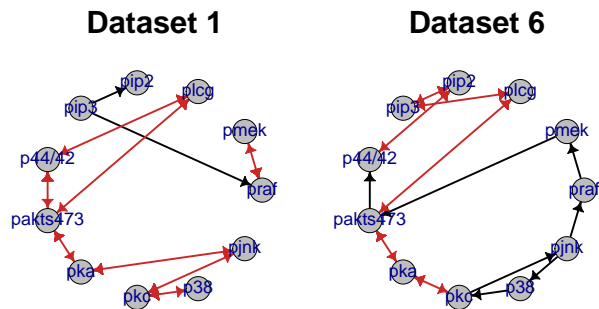
Figure 3: Estimated graphs corresponding to the case of minimum number of edges (10 edges, dataset 1) and maximum number of edges (13 edges, dataset 6).
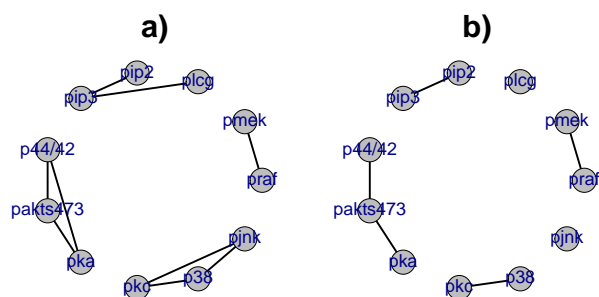

Figure 4: Edges appearing in at least 9 (a) and 13 (b) skeletons from the estimated graphs.

14) and PIP3 to PAKTS473 (dataset 6, 7 and 12).

Finally, in order to illustrate the behavior of the greedy search itself we focus again on datasets 1 and 6. Figure 5 shows the respective search paths in terms of the score achieved at each iteration. While local optima are possible, we observe that most search paths end with a score near the overall maximum.

7 CONCLUSION

We considered structure learning for linear causal models with Gaussian errors that may exhibit feedback loops and correlation induced by latent variables. In order to gain tractability in this difficult problem, we restricted our attention to simple mixed graphs. Such graphs have the favorable property of always inducing a model whose dimension is as one expects from counting parameters. This property allows one to form meaningful model selection scores. While a search over simple mixed graphs remains challenging, computationally and statistically, our experiments suggest that useful information can be learned from greedy search methods. This generalizes similar conclusions for acyclic simple graphs [22].
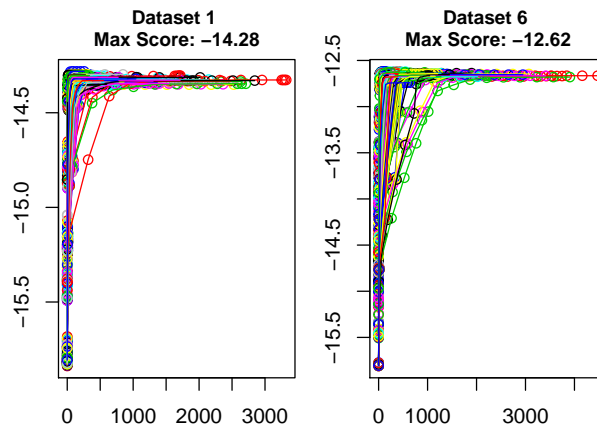

Figure 5: Curves of scores versus the time (seconds) in 300 random restarts greedy search for dataset 1 and dataset 6.

We also showed that an existing sufficient condition for distributional equivalence admits a natural generalization from acyclic to cyclic simple mixed graphs. However, the condition is very restrictive. It would be important to find more broadly applicable conditions for distributional equivalence.

References

[1] S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Theoretical aspects of cyclic structural causal models, 2016.

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

[3] D. M. Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002.

[4] M. 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.

[5] M. Drton, C. Fox, A. Käufl, and G. Pouliot. The maximum likelihood threshold of a path diagram. *Ann. Statist.*, 47(3):1536–1553, 2019.

[6] M. Drton, C. Fox, and Y. S. Wang. Computation of maximum likelihood estimates in cyclic structural equation models. *Ann. Statist.*, 47(2):663–690, 2019.

[7] M. Drton and M. Plummer. A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79:323–380, 2017.

[8] M. Drton, E. Robeva, and L. Weihs. Nested covariance determinants and restricted trek separation in Gaussian graphical models. *arXiv:1807.07561*, 2018.

[9] R. J. Evans. Model selection and local geometry, 2018.

[10] P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *UAI*, 2018.

[11] R. Foygel, J. Draisma, and M. Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.*, 40(3):1682–1713, 2012.

[12] R. Foygel and M. Drton. Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.*, 23:2020–2028, 2010.

[13] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *Ann. Statist.*, 29(2):505–529, 2001.

[14] A. Ghassami, K. Zhang, and N. Kiyavash. Characterizing distribution equivalence for cyclic and acyclic directed graphs, 2019.

[15] N. Harris and M. Drton. PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14(1):3365–3383, 2013.

[16] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[17] A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *J. Mach. Learn. Res.*, 13:3387–3439, 2012.

[18] A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, 2014.

[19] D. Kumor, B. Chen, and E. Bareinboim. Efficient identification in linear structural causal models with instrumental cutsets. In *Adv. Neural Inf. Process. Syst. 32*, pages 12477–12486. 2019.

[20] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

[21] M. Maathuis, M. Drton, and M. Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

[22] C. Nowzohour, M. H. Maathuis, R. J. Evans, and P. Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electron. J. Stat.*, 11(2):5342–5374, 2017.

[23] M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, pages 763–765, 1973.

[24] K. Rantanen, A. Hyttinen, and M. Järvisalo. Discovering causal graphs with cycles and latent confounders: an exact branch-and-bound approach. *Internat. J. Approx. Reason.*, 117:29–49, 2020.

[25] T. Richardson. A discovery algorithm for directed cyclic graphs. In *UAI*, 1996.

[26] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[27] I. Shpitser, R. Evans, T. Richardson, and J. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[28] L. Solus, Y. Wang, and C. Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms, 2017.

[29] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, Cambridge, MA, second edition, 2000.

[30] S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.*, 41(2):536–567, 2013.

[31] T. van Ommen and J. M. Mooij. Algebraic equivalence of linear structural equation models. In *UAI*, 2017.

[32] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *UAI*. 1991. UCLA Cognitive Systems Laboratory, Tech. Report (R-150).

[33] L. Weihs, B. Robinson, E. Dufresne, J. Kenkel, K. Kubjas, R. McGee, II, N. Nguyen, E. Robeva, and M. Drton. Determinantal generalizations of instrumental variables. *Journal of Causal Inference*, 6:1, 2017.