Characterizing Distribution Equivalence and Structure Learning for Cyclic and Acyclic Directed Graphs

AmirEmad Ghassami [1] **Alan Yang** [1] **Negar Kiyavash** [2] **Kun Zhang** [3]

Abstract

The main approach to defining equivalence among acyclic directed causal graphical models is based on the conditional independence relationships in the distributions that the causal models can generate, in terms of the Markov equivalence. However, it is known that when cycles are allowed in the causal structure, conditional independence may not be a suitable notion for equivalence of two structures, as it does not reflect all the information in the distribution that is useful for identification of the underlying structure. In this paper, we present a general, unified notion of equivalence for linear Gaussian causal directed graphical models, whether they are cyclic or acyclic. In our proposed definition of equivalence, two structures are equivalent if they can generate the same set of data distributions. We also propose a weaker notion of equivalence called quasi-equivalence, which we show is the extent of identifiability from observational data. We propose analytic as well as graphical methods for characterizing the equivalence of two structures. Additionally, we propose a score-based method for learning the structure from observational data, which successfully deals with both acyclic and cyclic structures.

1. Introduction

The problem of learning directed graphical models from data has received a significant amount of attention over the past three decades since those models provide a compact and flexible way to represent constraints on the joint distribution of the data (Koller & Friedman, 2009). When interpreted causally, they can model causal relationships among the variables of the system and help make predictions under intervention (Pearl, 2009; Spirtes et al., 2000).

There exists an extensive literature on learning causal graphical models from observational data under the assumption that the model is a directed acyclic graph (DAG) (Zhang et al., 2018). Existing approaches include constraint-based methods (Spirtes et al., 2000; Pearl, 2009), score-based methods (Heckerman et al., 1995; Chickering, 2002), hybrid methods (Tsamardinos et al., 2006), as well as methods which make extra assumptions on the data generating process. For example, the model may be assumed to be linear with non-Gaussian exogenous noise variables (Shimizu et al., 2006) or contain specific types of non-linearity in the causal modules (Hoyer et al., 2009; Zhang & Hyvärinen, 2009).

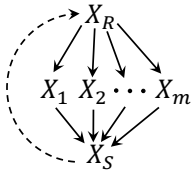

Figure 1.

Most real-life causal systems contain feedback loops, since feedback is generally required to stabilize the system and improve performance in the presence of noise. Hence, the causal directed graph (DG) corresponding to such systems will be cyclic (Spirtes, 1995; Hyttinen et al., 2012). However, there are relatively few works on learning structures that contain cycles. In many state-of-the-art causal models, not only is feedback ignored, it is also explicitly assumed that there are no cycles passing information among the considered quantities. Note that ignoring cycles in structure learning can be very consequential. For instance, in Figure 1, if one uses a conditional independence-based learning method designed for DAGs such as the PC algorithm (Spirtes et al., 2000), in the absence of the dashed feedback loop the skeleton will be estimated correctly on the population dataset and the directions for all edges into X_S can be determined. However, in the presence of the feedback loop, the output is a complete directed graph since no two variables will be independent conditioned on any subset of the rest of the variables.

The lack of attention to cyclic structures in the literature is primarily due to the simplicity of working with acyclic

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA [2]College of Management of Technology, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland [3]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: AmirEmad Ghassami <ghassam2@illinois.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

models (see (Spirtes, 1995)) and the fact that in contrast to DAGs, there exists no generally accepted characterization of statistical equivalence among cyclic structures in the literature. The main method for defining equivalence among DAGs is based on the conditional independence (CI) relationships in the distributions that they imply. That is, two DAGs are equivalent if and only if they imply the same CI relations. CI relationships can be seen from statistical data, and the CI-based equivalence characterization for DAGs is attractive because CI relationships contain all the information in the distribution that can be used for structure learning under the assumption of causal sufficiency. However, when causal sufficiency is violated or cycles are allowed in the structure, conditional independency may not reflect all the information in the distribution that can be used to identify the underlying structure. That is, the joint distribution may contain information that can be used to distinguish among the members of a CI-based equivalence class, which is also known as a Markov equivalence class. This means that it is possible for two graphs to be distinguishable from observational data even though they are in the same Markov equivalence class. For more details, see (Lacerda et al., 2008) for the case of the violation of acyclicity and (Tian & Pearl, 2002; Shpitser et al., 2014) for the case of the violation of causal sufficiency.

With the goal of bridging the gap between cyclic and acyclic DGs, in this paper we present a general characterization of equivalence for linear Gaussian DGs.[1] In the case of DAGs, our approach provides a novel alternative to the customary tests for Markov equivalence. The proposed distribution equivalence characterization (Theorems 1 and 2) not only is capable of characterizing equivalence beyond conditional independencies, but also provides a simpler and more concise evaluation approach compared to (Richardson, 1996b). We summarize our contributions as follows.

- We present a general, unified notion of equivalence based on the set of distributions that the directed graphs are able to generate (Section 2). In our proposed definition of equivalence, two structures are equivalent if they can generate the same *set* of data distributions.
- We propose an algebraic and graphical characterization of the equivalence of two DGs, be they cyclic or acyclic, based on the so-called Givens rotations (Sections 3 and 4).
- We also propose a weaker notion of equivalence called quasi-equivalence, which we show is the extent of identifiability from observational data (Section 5).
- We propose a score-based method for structure learning from observational data with local search. We show that our score asymptotically achieves the extent of identifiability (Section 5). To the best of our knowledge, this is the first local search method capable of learning structures with cycles. The implementation is publicly available at https://github.com/syanga/dglearn.

[1] Note that for non-linear cyclic SEMs, even the Markov property does not necessarily hold (Spirtes, 1995; Pearl & Dechter, 1996; Neal, 2000), and hence, it is not clear if one can make general statements about the equivalence of structures regardless of the involved equations.

1.1. Related Work

Richardson (1996a;b) proposed graphical constraints necessary and sufficient for Markov equivalence for general cyclic DGs and proposed a constraint-based algorithm for learning cyclic DGs. That algorithm was later extended to handle latent confounders and selection bias (Strobl, 2019). Hyttinen et al. (2013; 2014) also focused on structure learning based on CI relationships for possibly cyclic and causally insufficient data gathered from multiple domains that may contain conflicting CI information. They proposed an approach based on an SAT or ASP solver. Due to generality of their setup, the run time of this approach can be restricting. A similar approach was proposed in (Forré & Mooij, 2018) for the case of nonlinear functional relationships with an extended notion of graphical separation called σ-separation. Also, Hyttinen et al. (2012) provided an algorithm for learning linear models with cycles and confounders that deals with perfect interventions. As mentioned earlier, having the assumption of non-Gaussian exogenous noises and specific types of non-linearity may lead to unique identifiability in DAGs. This idea was also investigated for cyclic DGs. Lacerda et al. (2008) proposed a method for learning DGs based on the ICA approach for linear systems with non-Gaussian exogenous noises, and Mooij et al. (2011) investigated the case of nonlinear causal mechanisms with additive noise.

To the best of our knowledge, there exists no work on learning cyclic linear Gaussian models which utilizes the observational joint distribution itself rather than CI relationships in the distribution.

2. Distribution Equivalence

We consider a linear structural causal model over p observable variables $\{X_i\}_{i=1}^p$, with exogenous Gaussian noise. For $i \in [p]$, variable X_i is generated as $X_i = \sum_{j=1}^p B_{j,i} X_j + N_i$, in which N_i is the exogenous noise corresponding to variable X_i. We assume that $B_{i,i} = 0$, for all $i \in [p]$. Variable X_j is a direct cause of X_i if $B_{j,i} \neq 0$. We represent the causal structure among the variables with a DG $G = (V(G), E(G))$, in which $X_i \to X_j \in G$ if X_i is a direct cause of X_j. Let $X := [X_1 \cdots X_p]^\top$. The model can be represented in matrix form as $X = B^\top X + N$, where B is a $p \times p$ weighted adjacency matrix of G with $B_{j,i}$ as its (j,i)-th entry and $N = [N_1 \cdots N_p]^\top$. Elements of N

are assumed to be jointly Gaussian and independent. Since we can always center the data, without loss of generality, we assume that N, and hence, X is zero-mean. Therefore, $X \sim \mathcal{N}(0, \Sigma)$, where Σ is the covariance matrix of the joint Gaussian distribution on X, and suffices to describe the distribution of X. We assume that Σ is always invertible (the Lebesgue measure of non-invertible matrices is zero). Therefore, equivalently the precision matrix $\Theta = \Sigma^{-1}$ contains all the information regarding the distribution of X. Θ can be written as

$$\Theta = (I - B)\Omega^{-1}(I - B)^\top, \tag{1}$$

where Ω is a $p \times p$ diagonal matrix with $\Omega_{i,i} = \sigma_i^2 = Var(N_i)$. In the sequel, we use the terms precision matrix and distribution interchangeably.

The most common notion of equivalence for DGs in the literature is Markov equivalence (also called independence equivalence) defined as follows:

Definition 1 (Markov Equivalence)**.** *Let $\mathcal{I}(G)$ denote the set of all conditional d-separations[2] implied by the DG G. DGs G_1 and G_2 are Markov equivalent if $\mathcal{I}(G_1) = \mathcal{I}(G_2)$.*

When cycles are permitted, defining equivalence of DGs based on CI relations that they represent is not suitable, as CI relations do not reflect all the information in the distribution that can be used for identification of the underlying structure; e.g., see (Lacerda et al., 2008). That is, there exist DGs which can be distinguished using observational data with probability one despite representing the same CI relations. We define the notion of equivalence based on the set of distributions which can be generated by a structure:

Definition 2 (Distribution Set)**.** *The distribution set of structure G, denoted by $\Theta(G)$, is defined as*

$$\Theta(G) := \{\Theta : \Theta = (I - B)\Omega^{-1}(I - B)^\top, \text{ for any } (B, \Omega) \text{ s.t. } \Omega \in diag^+ \text{ and } supp(B) \subseteq supp(B_G)\},$$

where $diag^+$ is the set of diagonal matrices with positive diagonal entries, B_G is the binary adjacency matrix of G, and $supp(B) = \{(i, j) : B_{ij} = 0\}$.

$\Theta(G)$ is the set of all precision matrices (equivalently, distributions) that can be generated by G for different choices of exogenous noise variances and edge weights in G.

Definition 3 (Distribution Equivalence)**.** *DGs G_1 and G_2 are distribution equivalent, or for short, equivalent, denoted by $G_1 \equiv G_2$, if $\Theta(G_1) = \Theta(G_2)$.*

It is important to note that for DG G and distribution Θ, having $\Theta \in \Theta(G)$ does not imply that all the constraints of Θ, such as its conditional independencies, can be read off of G. For instance, a complete DAG does not represent any conditional d-separations, yet all distributions are contained in its distribution set. This is due to the fact that the parameters in B can be designed to represent certain extra constraints in the generated distribution.

As mentioned earlier, we can have a pair of DGs which are distinguishable using observational data despite having the same conditional d-separations. This is not the case for DAGs. In fact, restricting the space of DGs to DAGs, Definitions 3 and 1 are equivalent.

Proposition 1. *Two DAGs G_1 and G_2 are equivalent if and only if they are Markov equivalent.*

Therefore, one does not lose any information by caring only about Markov equivalence when dealing with acyclic structures. All proofs are provided in the Supplementary Materials.

For general DGs, the graphical test for Markov equivalence is known to be significantly more complex (Richardson, 1996b) than the test for DAGs (Verma & Pearl, 1991). There are currently no known graphical conditions for distribution equivalence. This is the goal of Section 4.

3. Characterizing Equivalence

In order to determine whether DGs G_1 and G_2 are equivalent, a baseline equivalence test is as follows: We consider a distribution $\Theta \in \Theta(G_1)$ which results from a certain choice of parameters of G_1 in expression (1), i.e., a certain choice of exogenous noise variances and edge weights. We then check whether there exists a choice of parameters for which G_2 generates Θ. We then repeat the same procedure for G_1, considering G_2 as the original generator. More specifically, for DG G_i, let $Q_i = (I - B)\Omega^{-\frac{1}{2}}$ for any choice of B such that $\text{supp}(B) \subseteq \text{supp}(B_{G_i})$ for $i \in \{1, 2\}$. For any choice of parameters of G_1 that results in distribution $\Theta = Q_1 Q_1^\top$, we check if $Q_2 Q_2^\top = \Theta$ has real-valued solution, and vice versa. Although this baseline equivalence test provides a systematic approach, it is tedious in many cases to check for the existence of a solution. In the following, we propose an alternative equivalence test based on rotations of Q.

Let v_i be the i-th row of matrix Q. Therefore, $\Theta = QQ^\top$ is the Gramian matrix of the set of vectors $\{v_1, \cdots v_p\}$. The set of generating vectors of a Gramian matrix can be determined up to isometry. That is, given $Q_1 Q_1^\top = \Theta$, we have $Q_2 Q_2^\top = \Theta$ if and only if $Q_2 = Q_1 U$ for some orthogonal transformation U. Therefore, Q_1 should be transformable to Q_2 by a rotation or an improper rotation (a rotation followed by a reflection).

In our problem of interest, for *any* parameterization of Q_1 (resp. Q_2) it is necessary to check if there exists an orthogonal transformation of Q_1 (resp. Q_2) which can be generated

[2] See (Pearl, 2009) for the definition of d-separation.

for *some* parameterization of Q_2 (resp. Q_1). Therefore, only the support of the matrix before and after the orthogonal transformation matters. Hence, we only need to consider rotation transformations. This can be formalized as follows: Let Q_G be B_G with 1s on its diagonal, i.e. $Q_G := I + B_G$. This is the binary matrix that for all choices of parameters B and Ω, $supp(Q) \subseteq supp(Q_G)$.

Proposition 2. *$G_1 \equiv G_2$ if and only if for any choice of Q_1, there exists rotation $U^{(1)}$ such that $supp(Q_1 U^{(1)}) \subseteq supp(Q_{G_2})$, and for any choice of Q_2, there exists rotation $U^{(2)}$ such that $supp(Q_2 U^{(2)}) \subseteq supp(Q_{G_1})$.*

To test the existence of a rotation required in Proposition 2, we propose utilizing a sequence of a special type of planar rotations called *Givens rotations* (Golub & Van Loan, 2012).

Definition 4 (Givens rotation)**.** *A Givens rotation is a rotation in the plane spanned by two coordinate axes. For a θ-radian rotation in the (j,k) plane, the entries of the Givens rotation matrix $G(j,k,\theta) = [g]_{p\times p}$ in $\mathbb{R}^p$ are $g_{i,i} = 1$ for $i \notin \{j,k\}$, $g_{i,i} = \cos(\theta)$ for $i \in \{j,k\}$, and $g_{k,j} = -g_{j,k} = -\sin(\theta)$, and the rest of the entries are zero.*

Any rotation in $\mathbb{R}^p$ can be decomposed into a sequence of Givens rotations. Hence, in Proposition 2, we need to find a sequence of Givens matrices and define U to be their product. The advantage of this approach is that the effect of a Givens rotation is easy to track: The effect of $G(j,k,\theta)$ on a row vector v is as follows.

$$[v_1 \cdots v_j \cdots v_k \cdots v_p]G(j,k,\theta) = [v_1 \cdots \cos(\theta)v_j + \sin(\theta)v_k \cdots -\sin(\theta)v_j + \cos(\theta)v_k \cdots v_p]. \quad (2)$$

3.1. Support Rotation

As previously mentioned, since all choices of parameters in the structure need to be considered, it is necessary to determine the existence of a rotation that maps one support to another. We define support matrix and support rotation as follows.

Definition 5 (Support matrix)**.** *For any matrix Q, its support matrix is a binary matrix ξ of the same size with entries in $\{0,\times\}$, where $\xi_{i,j} = \times$ if $Q_{i,j} = 0$ and $\xi_{i,j} = 0$ otherwise. For directed graph G, we define its support matrix as support matrix of Q_G.*

Givens rotations can be used to introduce zeros in a matrix, and hence, change its support. Consider input matrix Q. Using expression (2), for any $i,j \in [p]$, $Q_{i,j}$ can be set to zero using a Givens rotation in the (j,k) plane with angle $\theta = \tan^{-1}(-Q_{i,j}/Q_{i,k})$. When zeroing $Q_{i,j}$, there may exist an index l such that $Q_{l,j}$ or $Q_{l,k}$ will also become zero. However, since we consider all parameterizations of Q, we cannot take advantage of such accidental zeroings.

Definition 6 (Support Rotation)**.** *The support rotation $A(i,j,k)$ is a transformation that takes a support matrix ξ as the input and sets $\xi_{i,j}$ to zero using a Givens rotation in the (j,k) plane. The output is the support matrix of $QG(j,k,\tan^{-1}(-Q_{i,j}/Q_{i,k}))$, where $Q \in \arg\max_{Q'} |supp(Q'G(j,k,\tan^{-1}(-Q'_{i,j}/Q'_{i,k})))|$ such that the support matrix of Q' is ξ. Note that $G(j,k,\tan^{-1}(-Q'_{ij}/Q'_{i,k}))$ is the Givens rotation in the (j,k) plane which zeros $Q'_{i,j}$.*

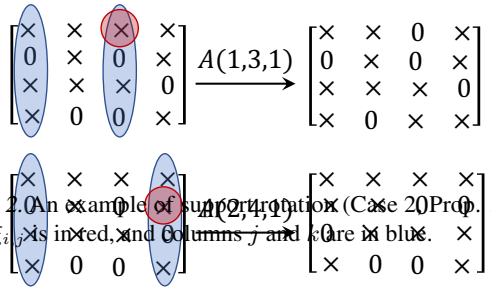

Figure 2. An example of support rotation (Case 2 Prop. 3). Element $\xi_{i,j}$ is in red, and columns j and k are in blue.

Note that due to (2), $A(i,j,k)$ only affects the j-th and k-th columns of the input. The general effect of support rotation $A(i,j,k)$ is described in the following proposition.

Proposition 3. *Support rotation $A(i,j,k)$ can have three possible effects on support matrix ξ:*

1. *If $\xi_{i,j} = 0$, $A(i,j,k)$ has no effect.*
2. *If $\xi_{i,j} = \times$ and $\xi_{i,k} = \times$, $A(i,j,k)$ makes $\xi_{i,j} = 0$, and for any $l \in [p] \setminus \{i\}$ such that at least one of $\xi_{l,j}$ and $\xi_{l,k}$ is $\times$, $A(i,j,k)$ makes $\xi_{l,j} = \times$ and $\xi_{l,k} = \times$. This is obtained by an acute rotation.*
3. *If $\xi_{i,j} = \times$ and $\xi_{i,k} = 0$, $A(i,j,k)$ switches columns j and k of ξ. This is obtained by a $\pi/2$ rotation.*

Figure 2 visualizes an example of a support rotation. Observe that the following four cases partition all the effects that can be obtained from a support rotation $A(i,j,k)$.

- **Reduction.** If $\xi_{i,j} = \xi_{i,k} = \times$ and $\xi_{l,j} = \xi_{l,k}$ for all $l \in [p] \setminus \{i\}$, then only $\xi_{i,j}$ becomes zero.
- **Reversible acute rotation.** If $\xi_{i,j} = \xi_{i,k} = \times$ and there exists a row i' such that the j-th and k-th columns differ only in that row, then $\xi_{i,j}$ becomes zero and both $\xi_{i',j}$ and $\xi_{i',k}$ become $\times$.
- **Irreversible acute rotation.** If $\xi_{i,j} = \xi_{i,k} = \times$ and the j-th and k-th columns differ in at least two rows, then $\xi_{i,j}$ becomes zero and all entries on the j-th and k-th columns become $\times$ on the rows on which they differed.
- **Column swap.** If $\xi_{i,j} = \times$ and $\xi_{i,k} = 0$, then columns j and k are swapped.

Note that if ξ is transformed to ξ' via a reversible acute rotation $A(i,j,k)$, and $\xi_{i',j} = 0$, then ξ' can be mapped back to ξ via $A(i',j,k)$, hence the name reversible.

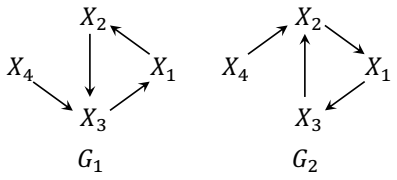

Figure 3. Example related to Proposition 4.

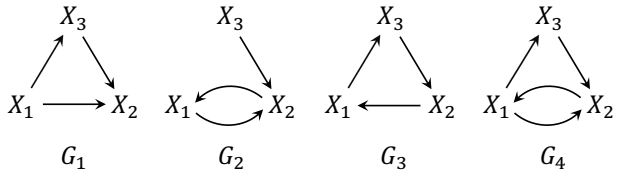

Figure 4. DGs related to Example 1.

3.2. Characterizing Equivalence via Support Rotations

We give the following necessary and sufficient condition for distribution equivalence of two structures using the introduced support operations. We show that irreversible acute rotations are not needed for checking equivalence. Here, for two support matrices ξ and ξ', we say $\xi \subseteq \xi'$ if $\text{supp}(\xi) \subseteq \text{supp}(\xi')$.

Theorem 1. *Let ξ_1 and ξ_2 be the support matrices of DGs G_1 and G_2, respectively. G_1 is distribution equivalent to G_2 if and only if there exists a sequence of reductions, reversible acute rotations, and column swaps that maps ξ_1 to a subset of ξ_2, and a sequence that maps ξ_2 to a subset of ξ_1.*

Theorem 1 converts the problem of determining the equivalence of two structures into a search problem for two sequences of support rotations. We propose to use a depth-first search algorithm that performs all column swaps at the end of the sequences. Due to space constraints, the pseudo-code is presented in the Supplementary Materials.

The following result is a nontrivial application of Theorem 1 regarding reversing cycles in DGs.

Proposition 4 (Direction of Cycles)**.** *Suppose structure G_1 contains a directed cycle C. Let G_2 be a structure that differs from G_1 in two ways. (1) The direction of cycle C is reversed and (2) any variable pointing to $X_i \in C$ in G_1 via an edge which is not part of C is, in G_2, pointing to the preceder of X_i in C in G_1. In this case, G_1 is distribution equivalent to G_2. (See Figure 3 for an example.)*

Richardson (1996b) presented a result similar to Proposition 4 for the case of using CI relationships in the data and concluded that "it is impossible to orient a cycle merely using CI information." Proposition 4 extends that result by concluding that it is impossible to orient a cycle merely using observational data.

The following proposition provides a necessary and sufficient condition for equivalence for a specific class of DGs.

Proposition 5. *Consider DGs G_1 and G_2 with support matrices ξ_1 and ξ_2, respectively. If every pair of columns of ξ_1 differ in more than one entry, then $G_1 \equiv G_2$ if and only if the columns of ξ_2 are a permutation of columns of ξ_1.*

Example 1. *In Figure 4, (a) $G_1 \equiv G_2$, (b) $G_1 \not\equiv G_3$, and (c) $G_1 \equiv G_4$.*

(a) shows that unlike DAGs, equivalent DGs do not need to have the same skeleton or the same v-structures. To see $G_1 \equiv G_2$, we note that

$$\xi_1 = \begin{bmatrix} \times & \times & \times \\ 0 & \times & 0 \\ 0 & \times & \times \end{bmatrix} \xrightarrow{A(1,3,1)} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & 0 \\ \times & \times & \times \end{bmatrix} \xrightarrow{A(3,1,2)} \begin{bmatrix} \times & \times & 0 \\ \times & \times & 0 \\ 0 & \times & \times \end{bmatrix} \subseteq \xi_2.$$

$$\xi_2 = \begin{bmatrix} \times & \times & 0 \\ \times & \times & 0 \\ 0 & \times & \times \end{bmatrix} \xrightarrow{A(2,1,2)} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & 0 \\ \times & \times & \times \end{bmatrix} \xrightarrow{A(3,1,3)} \begin{bmatrix} \times & \times & \times \\ 0 & \times & 0 \\ 0 & \times & \times \end{bmatrix} \subseteq \xi_1.$$

(b) follows from Proposition 5 since each pair of columns of ξ_3 differ in more than one entry. For (c), we already have $\xi_1 \subseteq \xi_4$. For the other direction,

$$\xi_4 = \begin{bmatrix} \times & \times & \times \\ \times & \times & 0 \\ 0 & \times & \times \end{bmatrix} \xrightarrow{A(2,1,2)} \begin{bmatrix} \times & \times & \times \\ 0 & \times & 0 \\ \times & \times & \times \end{bmatrix} \xrightarrow{A(3,1,3)} \begin{bmatrix} \times & \times & \times \\ 0 & \times & 0 \\ 0 & \times & \times \end{bmatrix} \subseteq \xi_1.$$

As seen in Example 1, structures G_1 and G_4 in Figure 4 are distribution equivalent. Therefore, the extra edge $X_2 \to X_1$ in G_4 does not enable this structure to generate any additional distributions. In this case, we say structure G_4 is reducible. This idea is formalized as follows.

Definition 7 (Reducibility)**.** *DG G is reducible if there exists G' such that $G \equiv G'$ and $E(G') \subset E(G)$. In this case, we say edges in $E(G) \setminus E(G')$ are reducible, and G is reducible to G'.*

Proposition 6. *DG G with support matrix ξ is reducible if and only if there exists a sequence of reversible acute rotations that enables us to apply a reduction to ξ.*

Proposition 6 implies the following necessary condition for reducibility.

Proposition 7. *A DG with no 2-cycles is irreducible.*

A 2-cycle is a cycle over only two variables, such as the cycle over X_1 and X_2 in G_2 in Figure 4. Propositions 6 and 7 lead to the following corollary regarding equivalence for DAGs, which bridges our proposed approach with the classic characterization for equivalence of DAGs.

Corollary 1. *DAGs G_1 and G_2 with support matrices ξ_1 and ξ_2 are equivalent if and only if there exists a sequence of reversible acute rotations and column swaps that maps ξ_1 to a subset of ξ_2, and one that maps ξ_2 to a subset of ξ_1.*

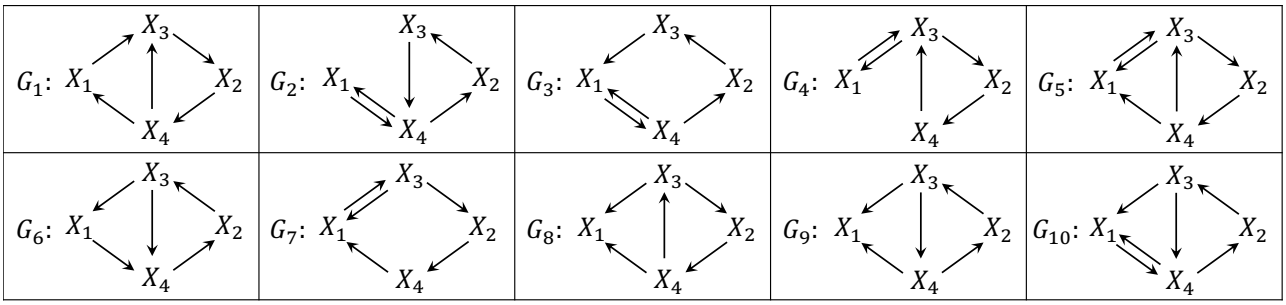

Figure 5. Elements of a distribution equivalence class.

Example 2. *We demonstrate our approach on a familiar equivalence example on DAGs: Let $G_1 : X_1 \to X_2 \to X_3$, $G_2 : X_1 \leftarrow X_2 \leftarrow X_3$, and $G_3 : X_1 \to X_2 \leftarrow X_3$. (a) $G_1 \equiv G_2$. (b) $G_1 \not\equiv G_3$.*

To see $G_1 \equiv G_2$, we note that

$$\xi_1 = \begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{A(1,2,1)} \begin{bmatrix} \times & 0 & 0 \\ \times & \times & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{A(2,3,2)} \begin{bmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ 0 & \times & \times \end{bmatrix} \subseteq \xi_2.$$

$$\xi_2 = \begin{bmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ 0 & \times & \times \end{bmatrix} \xrightarrow{A(3,2,3)} \begin{bmatrix} \times & 0 & 0 \\ \times & \times & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{A(2,1,2)} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix} \subseteq \xi_1.$$

For the second part, we note that ξ_3 has two columns with two zeros, while ξ_1 has only one column with two zeros. Therefore, reversible acute rotations and column swaps cannot map ξ_1 to a subset of ξ_3. Therefore $G_1 \not\equiv G_3$.

4. Graphical Characterization of Equivalence

In this section, we present a graphical counterpart to Theorem 1 by providing graphical counterparts to the rotations required by that Theorem.

Definition 8. *For vertices X_1 and X_2, let $P_1 := Pa(X_1) \cup \{X_1\}$ and $P_2 := Pa(X_2) \cup \{X_2\}$, where $Pa(X)$ denotes the set of parents of vertex X. X_1 and X_2 are parent reducible if $P_1 = P_2$ and parent exchangeable if $|P_1 \triangle P_2| = 1$, where $\triangle$ is the symmetric difference operator, which identifies elements which are only in one of the sets.*

The three rotations in Theorem 1 lead to the following graphical operations:

- **Parent reduction.** If X_j and X_k are parent reducible, any support rotation on columns $\xi_{\cdot,j}$ and $\xi_{\cdot,k}$ which zeros a non-zero entry on those columns except $\xi_{j,j}$ and $\xi_{k,k}$ removes the parent from X_j or X_k corresponding to the zeroed entry. We call this edge removal a parent reduction. The support rotation in this case is of reduction rotation type.
- **Parent exchange.** If X_j and X_k are parent exchangeable, by definition there exists X_i such that $P_j \triangle P_k = \{X_i\}$. In this case, any support rotation on columns $\xi_{\cdot,j}$ and $\xi_{\cdot,k}$ which zeros a non-zero entry on those columns except $\xi_{j,j}$ and $\xi_{k,k}$ removes the parent from X_j or X_k corresponding to the zeroed entry. Additionally, the missing edge from X_i to X_j or X_k is added. We call this a parent exchange. The support rotation in this case is of column swap or reversible acute rotation type.
- **Cycle reversion.** A cycle reversion swaps the column of each member of a cycle C with the column corresponding to its preceder in the cycle. This reverses the direction of the cycle C and changes any edge outside of C connecting to an $X_i \in C$ in the original DG to point instead to the preceder of X_i in C.

Note that in the graphical operations above, we exclude support rotations that lead to zeroing a diagonal entry, since they do not have a graphical representation (by Def. 5).

Equipped with the graphical operations, we present a graphical counterpart to Theorem 1.

Theorem 2. *G_1 is distribution equivalent to G_2 if and only if there exists a sequence of parent reductions, parent exchanges, and cycle reversions that maps G_1 to a subgraph of G_2, and a sequence that maps G_2 to a subgraph of G_1.*

Example 3. *Figure 5 shows the elements of a distribution equivalence class. Suppose G_1 is the original structure. Cycle reversion on the cycle (X_2, X_4, X_3, X_2) results in G_2, cycle reversion on the cycle $(X_1, X_3, X_2, X_4, X_1)$ results in G_3, parent exchange $A(4, 1, 3)$ results in G_4, and parent exchange $A(1, 3, 1)$ results in G_8.*

Remark 1. *Given observational data from any of the structures in Figure 5, CI-based structure learning methods such as CCD (Richardson, 1996a) may output a structure (for example G_1 without edges $X_4 \to X_1$) which is not distribution equivalent to the ground truth. This can be prevented by leveraging other statistical information in the distribution beyond CI relationships.*

We have the following corollary regarding equivalence for DAGs. The reasoning is the same as in Corollary 1.

Corollary 2. *DAGs G_1 and G_2 are equivalent if and only if there exists a sequence of parent exchanges that maps G_1 to G_2, and one that maps G_2 to G_1.*

5. Learning Directed Graphs from Data

Structure G imposes constraints on the entries of precision matrix Θ. We will refer to such constraints as the *distributional constraints* of G. Every distribution in $\Theta(G)$ should satisfy the distributional constraints of G. Clearly, two DGs are distribution equivalent if and only if they have the same distributional constraints. We call a distributional constraint a *hard constraint* if the set of the values satisfying that constraint is Lebesgue measure zero over the space of the parameters involved in the constraint. For instance in DAGs, if X_i and X_j are non-adjacent and have no common children, we have the hard constraint $\Theta_{i,j} = 0$. We denote the set of hard constraints of a DG G by $H(G)$.

Recall that distribution equivalence of two structures G_1 and G_2 implies that any distribution that can be generated by G_1 can also be generated by G_2, and vice versa. Therefore, no distribution can help us distinguish between G_1 and G_2. However, in practice we usually have access to only one distribution which is generated from a ground truth structure, and it may be the case that this distribution can be generated by another structure which is not equivalent to the ground truth. Therefore, finding the distribution equivalence class of the ground truth structure from one distribution is in general not possible, and extra considerations are required for the problem to be well defined. Below we will accordingly provide a weaker notion of equivalence and show that the ground truth can be recovered up to this equivalence.

The aforementioned issue also arises when learning DAGs and considering Markov equivalence. The most common approach to dealing with this issue in the literature is to assume that the distribution is *faithful* to the ground truth structure. This requires a one-to-one correspondence between the conditional d-separations of the ground truth structure and the CI relationships in the distribution (Spirtes et al., 2000). This is a sensible assumption from the perspective that the Lebesgue measure of the parameters which lead to extra CIs in the generated distribution is zero (Meek, 2013).

The case of general DGs is more complex since they can require other distributional constraints besides CIs. In particular, we may have distributional constraints other than hard constraints due to cycles. Hence, in this case the Lebesgue measure of the parameters which lead to extra distributional constraints in the generated distribution is not necessarily zero. This motivates the following weaker notion of equivalence for structure learning from observational data.

Definition 9 (Quasi Equivalence)**.** *Let θ_G be the set of linearly independent parameters needed to parameterize any distribution $\Theta \in \Theta(G)$. For two DGs G_1 and G_2, let μ be the Lebesgue measure defined over $\theta_{G_1} \cup \theta_{G_2}$. G_1 and G_2 are quasi equivalent, denoted by $G_1 \cong G_2$, if $\mu(\theta_{G_1} \cap \theta_{G_2}) = 0$.*

Roughly speaking, two DGs are quasi equivalent if the set of distributions that they can both generate has a non-zero Lebesgue measure. Note that Definition 9 implies that if DGs G_1 and G_2 are quasi equivalent they share the same hard constraints. We have the following assumption for structure learning, which is a generalization of faithfulness:

Definition 10 (Generalized faithfulness)**.** *A distribution Θ is generalized faithful (g-faithful) to structure G if Θ satisfies a hard constraint κ if and only if $\kappa \in H(G)$.*

Assumption 1. *The generated distribution is g-faithful to the ground truth structure G^*, and for irreducible DG G^*, if there exists a DG G such that $H(G) \subseteq H(G^*)$ and $|E(G)| \leq |E(G^*)|$, then $H(G) = H(G^*)$.*

The following justifies the first part of Assumption 1:

Proposition 8. *With respect to Lebesgue measure over θ_G, the set of distributions not g-faithful to G is measure zero.*

The second part of Assumption 1 requires that if the ground truth structure G^* has no reducible edges and there exists another DG G that has only relaxed some of the hard constraints of G^*, then G must have more edges than G^*. This is clearly the case for DAGs.

Proposition 9. *Under Assumption 1, quasi equivalence is the extent of identifiability from observational data.*

5.1. Score-Based Structure Learning

We propose a score-based method for structure learning based on local search. Score-based methods are well-established in the literature for learning DAGs. The predominant approach is to maximize the regularized likelihood of the data by performing a greedy search over all DAGs (Heckerman et al., 1995), equivalence classes of DAGs (Chickering, 2002), or permutations of the variables (Teyssier & Koller, 2012; Solus et al., 2017). Also, works such as (Van de Geer & Bühlmann, 2013; Fu & Zhou, 2013; Aragam & Zhou, 2015; Raskutti & Uhler, 2018; Zheng et al., 2018) specifically consider the problem of learning a linear Gaussian acyclic model via penalized parameter estimation.

To the best of our knowledge, there are no existing score-based structure learning approaches for the cyclic linear Gaussian model. In light of our theory, we propose to use the ℓ_0-regularized negative log likelihood function as the score, which is a standard choice of the score in the literature of learning DAGs, and show that it is able to recover the quasi equivalence class of the underlying DG. Let $\mathbf{X}$ be the $n \times p$ data matrix. The ℓ_0-regularized ML estimator solves

the following unconstrained optimization problem:

$$\min_{G} \min_{(B,\Omega):\text{supp}(B)\subseteq\text{supp}(B_G)} \mathcal{L}(\mathbf{X}:B,\Omega) + \lambda\|B\|_0, \quad (3)$$

where $\mathcal{L}(\mathbf{X}:B,\Omega) = -n\log(\det(I-B)) + \sum_{i=1}^{p}\frac{n}{2}\log(\sigma_i^2) + \frac{1}{2\sigma_i^2}\|\mathbf{X}_{\cdot,i} - \mathbf{X}B_{\cdot,i}\|_2^2$ is the negative log-likelihood of the data, $\|B\|_0 := \sum_{i,j}\mathbb{1}_{x\neq 0}(B_{i,j})$, and similar to the BIC score, we set $\lambda = 0.5\log n$.

Remark 2. *The estimator in (3) will never output a reducible DG, since removing redundant edges improves the score. This is in line with the minimality assumption in the literature for DAGs (Pearl, 1988; Raskutti & Uhler, 2018).*

Theorem 3. *Under Assumption 1, the global minimizer of (3) with $\lambda = 0.5\log n$ outputs $\hat{G} \cong G^*$ asymptotically.*

Hence, by Prop. 9 and Theorem 3, the score (3) is consistent, i.e., it asymptotically achieves the extent of identifiability.

5.1.1. Structure Search

We solve the outer optimization problem in (3) via local search over the structures. We choose the search space to contain all DGs and use the standard operators (i.e., local changes) of edge addition, deletion, and reversal. See (Koller & Friedman, 2009) for a discussion regarding the necessity of these operators. Two main issues arise when cycles are allowed in the structure:

Virtual edges. There exists a virtual edge between non-adjacent vertices X_i and X_j if they have a common child X_k which is an ancestor of X_i or X_j (Richardson, 1996b). If a greedy search algorithm does not find X_k and X_i (or X_j) to be on a cycle, it can significantly increase the likelihood by adding an edge at the location of the virtual edge. The algorithm would therefore be trapped in a local optimum with one more edge than the ground truth. To resolve this issue, we propose adding the following fourth search operator: Suppose we have a triangle over three variables X_i, X_j and X_k, and there exists an additional sequence of edges connecting X_j and X_k. In one atomic move, we perform a series of edge reversals to form a cycle containing $X_j \to X_k$ along the sequence, delete the edge connecting X_i to X_j, and orient the edge $X_i \to X_k$. If the likelihood is unchanged, the edge deletion improves the score. In the case that the oriented cycle is of length two, additional considerations are needed; see the Supplementary Materials for details as well as simulations justifying this fourth operator.

Score decomposability. When the DG is acyclic, the distribution generated by a linear Gaussian structural equation model satisfies the local Markov property. This implies that the joint distribution can be factorized into the product of the distributions of the variables conditioned on their parents. The benefit of this factorization is that the computational complexity of evaluating the effect of operators can be dramatically reduced since a local change in the structure does not change the score of other parts of the DAG. In contrast, for the case of cyclic DGs the distribution does not necessarily satisfy the local Markov property. However, the distribution still satisfies the global Markov property (Spirtes, 1995). Therefore, our search procedure factorizes the joint distribution into the product of conditional distributions. Each of these distributions is over the variables in a maximal strongly connected subgraph (MSCS), conditioned on their parents outside of the MSCS. After applying an operation, the likelihoods of all involved MSCSs are updated; see the Supplementary Materials for additional details.

The implementation of the approach is publicly available at https://github.com/syanga/dglearn.

6. Experiments

We generated 100 random ground truth DGs of orders $p \in \{5, 20, 50\}$, all with maximum degree 4. The DGs are constrained to have maximum cycle lengths 5, 5, and 10, respectively. For each structure, we sampled the edge weights uniformly from $B_{i,j} \in [-0.8, -0.2]\cup[0.2, 0.8]$ and the exogenous noise variances uniformly from $\sigma_i^2 \in [1, 3]$ to generate the data matrix $\mathbf{X}$ of size $10^4 \times p$. We constrained the ground truth B matrices to be stable via an accept-reject approach; the modulus of all eigenvalues of B should be strictly less than one. The stability of a model guarantees that the effects of one-time noise dissipate. Our search algorithms were also constrained to only output stable structures. We used the following standard local search methods: 1. Hill climbing 2. Tabu search (Koller & Friedman, 2009).

Evaluating the performance of a learning approach is not trivial for the case of general DGs. As seen before, equivalent cyclic DGs may have very different skeletons. Hence, conventional evaluation metrics such as structural Hamming distance (SHD) with the ground truth DG or comparison of the learned and ground truth adjacency matrices cannot be used. We propose the following evaluation methods:

1. SHD Evaluation. We enumerate the set of all DGs equivalent to the ground truth DG using Algorithm 1 in the Supplementary Materials to form the distribution equivalence class of the ground truth. We then compute the smallest SHD between the algorithm's output DG and the members of the equivalence class as a measure of the performance.

2. Multi-Domain Evaluation. Suppose the input data is sampled from a distribution Θ generated by ground truth DG G^*, and let $\hat{G}$ denote an algorithm's output structure. Due to finite sample size and the possible violation of Assumption 1, $\hat{G}$ may be able to maximize the likelihood yet not be (quasi) equivalent to G^*. In general, we expect such an output to be compatible with only the given data and not with data sampled from other distributions generated by G^*. We therefore propose the following evaluation approach.

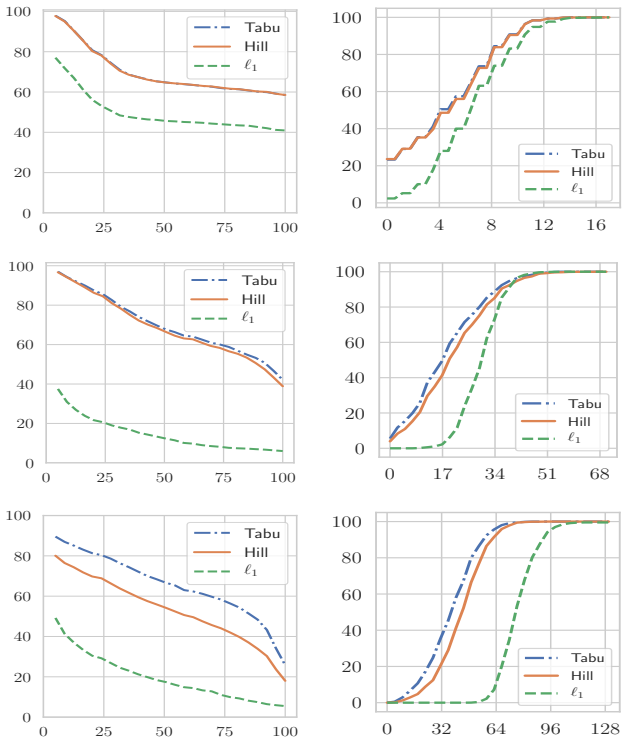

Figure 6. Results for $p = 5, 20, 50$, top to bottom. **Left column:** multi-domain evaluation. The percentage of outputs with success rate larger than a certain value is plotted vs. success percentages; e.g., for $p = 20$, 80% of the outputs could generate more than 25% of the distributions generated by their corresponding ground truth. **Right column:** SHD evaluation. The percentage of outputs with SHD less than or equal to a certain value is plotted vs. SHD.

1. For ground truth structure G^*, generate d distributions $\{\Theta_1, ..., \Theta_d\}$ by sampling edge weights and variances.
2. For each Θ_i, run the algorithm to obtain $\hat{G}_i$.
3. For each $\hat{G}_i$, optimize its edge weights and variances to generate distributions $\{\hat{\Theta}_{i,1}, ..., \hat{\Theta}_{i,d}\}$ such that $\hat{\Theta}_{i,j}$ minimizes the KL-divergence to $\Theta_j \in \{\Theta_1, ..., \Theta_d\}$.
4. The success rate of $\hat{G}_i$ is the percentage of domains for which the minimizing KL-divergence computed in step 3 is below a threshold η.

Since domain distributions are generated randomly, if the success rate of output $\hat{G}_i$ is large, there is a non-negligible subset of the distribution set of G^* that $\hat{G}_i$ can generate as well. Hence, $\hat{G}_i$ is quasi equivalent to G^*. In our evaluations, we used $d = 50$ and $\eta = p \times 10^{-3}$. We emphasize that multi-domain data is *only* used for evaluation. In the learning stage, only one distribution is used.

We cannot compare the performance of our approach with the performance of methods based on CI relationships (such as CCD), since those approaches return a PAG representing all Markov equivalent DGs, which usually represents a much larger set of DGs than the distribution equivalence class. We therefore only compared our approach with an ℓ_1-regularized maximum likelihood estimator which directly solves the optimization problem $\min_{B,\Omega} \mathcal{L}(\mathbf{X} : B, \Omega) + \lambda \|B\|_1$, which does not need a separate structure search. The results are given in Figure 6. The figure shows that our proposed approach successfully finds DGs capable of generating distributions generated by the ground truth structure. While the SHD evaluation shows that the outputs are not always distribution equivalent, the multi-domain evaluation provides evidence that many are quasi equivalent to the ground truth. We also evaluated the effect of sample size on the performance in the Supplementary Materials.

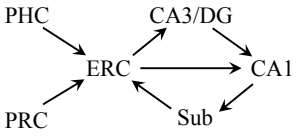

Figure 7. Ground truth structure for the fMRI hippocampus dataset.

6.1. fMRI hippocampus data

We considered the fMRI hippocampus dataset (Poldrack et al., 2015), which contains signals from six separate brain regions: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate Gyrus (CA3) in the resting state. We used the anatomical connections (Bird & Burgess, 2008; Zhang et al., 2017) as the ground truth, depicted in Figure 7. We applied our proposed method on one of the domains in the dataset and found that two out of eight structures equivalent to the ground truth were (local) optima for the score even though there is no evidence that the data are linear Gaussian.

7. Conclusion

We presented a general, unified notion of equivalence for linear Gaussian DGs and proposed methods for characterizing the equivalence of two structures. We also proposed a score-based structure learning approach that asymptotically achieves the extent of identifiability. Our results are instrumental to the fields of causality and graphical models. From the causality perspective, consider for example Figure 5. Our results guarantee a direct causal effect between X_2 and X_4 and show that a direct causal effect does not necessarily exist between X_3 and X_4. From the graphical models perspective, our results provide the tools to handle distributions that lack a DAG representation but can be modeled by a cyclic DG. We hope that this work spurs further research in the study of directed graphs.

Acknowledgements

This work was supported in part by ONR grants W911NF15-1-0479 and N00014-19-1-2333, NSF CCF 1704970, and NSF CNS 16-24811. KZ would like to acknowledge the support by National Institutes of Health under Contract No. NIH-1R01EB022858-01, FAIN-R01EB022858, NIH-1R01LM012087, NIH5U54HG008540-02, and FAIN-U54HG008540, and by the United States Air Force under Contract No. FA8650-17-C7715. We thank Dr. Jia-En Liang from Unisound for helpful discussions.

References

Aragam, B. and Zhou, Q. Concave penalized estimation of sparse gaussian bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328, 2015.

Bird, C. M. and Burgess, N. The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, 9(3):nrn2335, 2008.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Forré, P. and Mooij, J. M. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*, 2018.

Fu, F. and Zhou, Q. Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501):288–300, 2013.

Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU press, 2012.

Heckerman, D., Geiger, D., and Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439, 2012.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Jarvisalo, M. Discovering cyclic causal models with latent variables: A general sat-based procedure. *arXiv preprint arXiv:1309.6836*, 2013.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pp. 340–349, 2014.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Lacerda, G., Spirtes, P. L., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. pp. 366–374, 2008.

Meek, C. Strong completeness and faithfulness in bayesian networks. *arXiv preprint arXiv:1302.4973*, 2013.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, pp. 639–647, 2011.

Neal, R. M. On deducing conditional independence from d-separation in causal graphs with feedback (research note). *Journal of Artificial Intelligence Research*, 12: 87–91, 2000.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 1988.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. and Dechter, R. Identifying independencies in causal graphs with feedback. pp. 420–426, 1996.

Poldrack, R., Laumann, T., Koyejo, O., Gregory, B., Hover, A., Chen, M., Luci, J., Joo, S., Handwerker, D., Liang, J., Boyd, R., Hunicke-Smith, S., Simpson, Z., Caven, T., Sochat, V., Shine, J., Gordon, E., Snyder, A., Adeyemo, B., and ... Mumford, J. https://openfmri.org/dataset/ds000031/, 2015.

Raskutti, G. and Uhler, C. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

Richardson, T. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pp. 454–461. Morgan Kaufmann Publishers Inc., 1996a.

Richardson, T. A polynomial-time algorithm for deciding markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pp. 462–469. Morgan Kaufmann Publishers Inc., 1996b.

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

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Solus, L., Wang, Y., Matejovicova, L., and Uhler, C. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

Spirtes, P. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 491–498. Morgan Kaufmann Publishers Inc., 1995.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Strobl, E. V. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8 (1):33–56, 2019.

Teyssier, M. and Koller, D. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

Tian, J. and Pearl, J. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 519–527. Morgan Kaufmann Publishers Inc., 2002.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Van de Geer, S. and Bühlmann, P. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Verma, T. and Pearl, J. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), Montreal, Canada*, 2009.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. Causal discovery in the presence of distribution shift: Skeleton estimation and orientation determination. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 2017.

Zhang, K., Schölkopf, B., Spirtes, P., and Glymour, C. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2018.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.