Abstract

The main way for defining equivalence among acyclic directed graphs is based on the conditional independencies of the distributions that they can generate. However, it is known that when cycles are allowed in the structure, conditional independence is not a suitable notion for equivalence of two structures, as it does not reflect all the information in the distribution that can be used for identification of the underlying structure. In this paper, we present a general, unified notion of equivalence for linear Gaussian directed graphs. Our proposed definition for equivalence is based on the set of distributions that the structure is able to generate. We take a first step towards devising methods for characterizing the equivalence of two structures, which may be cyclic or acyclic. Additionally, we propose a score-based method for learning the structure from observational data.

1 Introduction

Directed graphs are one of the main tools for representing causal relations among the variables of a system. The problem of learning a directed graph from data has received a significant amount of attention over the past three decades, as they provide a compact and flexible way to decompose the joint distribution of the data [Koller and Friedman, 2009] and, when interpreted causally, can model causal relations and help to make prediction under interventions.
There exist an extensive literature on learning causal directed acyclic graphs (DAGs). The proposed methods for this task range from purely observational methods to the cases where an experimenter is capable of intervening in the system. However, observational methods have received more interest due to the fact that in many applications, performing an intervention may not be practical. For the case of learning causal DAGs from observational data, existing methods include constraint-based methods [Spirtes et al., 2000, Pearl, 2009], score-based methods [Heckerman et al., 1995], and hybrid methods [Tsamardinos et al., 2006], as well as methods which make extra assumptions on the data generating process, such as linear models with non-Gaussian exogenous noises [Shimizu et al., 2006] and specific types of non-linearity of the causal modules [Hoyer et al., 2009, Zhang and Hyvärinen, 2009]. Score-based methods are a well-established line of work, which started with [Heckerman et al., 1995], in which a greedy search is performed over all DAGs, and has evolved to more sophisticated techniques such as searching over equivalence classes of DAGs [Chickering, 2002], or over permutations of the variables [Teyssier and Koller, 2012, Solus et al., 2017]. Moreover, recently, variations of score-based methods suitable for high-dimensional data were proposed [Van de Geer and Bühlmann, 2013, Fu and Zhou, 2013, Aragam and Zhou, 2015, Zheng et al., 2018].

Cyclic directed graphs are rather ubiquitous in modeling economic processes and natural systems, e.g., in the field of biology. In fact, feedback loops in causal systems are generally helpful to improve system performance in the presence of model uncertainty and achieve stability of the system. However, there are relatively few works on characterizing and learning structures that contain cycles. In many state-of-the-art causal models, not only is feedback ignored, but it is also explicitly assumed that there are no cycles passing information among the considered quantities. This discrimination against cyclic structures in the literature is primarily due to simplicities of working with acyclic models (see [Spirtes, 1995]), and the fact that even a generally accepted definition of equivalence relations does not exist in the literature of cyclic directed graphs. Note that ignoring cycles in structure learning can be very consequential. For instance, in Figure 1 if one performs a conditional independence-based test designed for DAGs, such as the PC algorithm [Spirtes et al. 2000], in the absence of the dashed feedback loop, on the population dataset, the skeleton will be estimated correctly and we can also determine the directions for all edges into $X_S$. However, adding the feedback loop, the output will abruptly change to a complete directed graph, as no two variables will be independent conditioned on any subset of the rest of the variables.
A main way for defining equivalence among DAGs is based on the conditional independencies of the distributions that they can generate. Richardson [1996b] proposed graphical constraints necessary and sufficient for conditional independence-based equivalence, also known as Markov equivalence, for general cyclic directed graphs. However, when cycles are allowed in the structure, conditional independency does not reflect all the information in the distribution that can be used for identification of the underlying structure [Lacerda et al. 2012], and the joint distribution may contain information that can be used to distinguish among the members of a Markov equivalence class. That is, it is possible that two graphs can be distinguishable from observational data even though they are in the same Markov equivalence class. Hence, recovering the ground truth structure requires also other constraints of the generated distribution beyond what is captured by conditional independencies. Lacerda et al. [2012] have proposed a method for learning directed graphs beyond Markov equivalence based on the ICA approach, yet their approach only works for linear systems with non-Gaussian exogenous noises.

With the goal of bridging the gap between cyclic and acyclic directed graphs, in this paper, we present a general definition of equivalence for linear Gaussian directed graphs. Our definition of equivalence is based on the set of distributions that the directed graphs are capable of generating. We take a first step towards devising methods for characterizing the equivalence of two structures, which may be cyclic or acyclic. In the case of acyclic directed graphs, our proposed approach will provide a novel alternative to the customary tests for Markov equivalence. The proposed distribution equivalence test (Theorem 1 and Proposition 8) is not only capable of characterizing equivalence beyond conditional independencies, but also provides a simpler and more concise evaluation approach compared to [Richardson, 1996b]. We also propose a score-based method for learning the structure from observational data, even though the focus of this work is primarily on the description and checking the distribution equivalence.

The rest of the paper is organized as follows: We describe the model and present our formal definition of equivalence in Section 2. The proposed method for checking the equivalence of two directed graphs is presented in Sections 3 and 4. Finally, we discuss the task of learning a directed graph from observational data in Section 5. All the proofs are provided in the Appendix.

## 2 Problem Description

Consider a causally sufficient linear structural causal model over $p$ endogenous variables $\{X_i\}_{i=1}^p$, with Gaussian exogenous variables. Let $X := [X_1 \cdots X_p]^\top$. For $i \in [p]$, variable $X_i$ is generated as $X_i \leftarrow \beta_i^\top X + N_i$, in which $N_i$ is the exogenous variable corresponding to variable $X_i$. Non-zero entries of $\beta_i$ correspond to direct causes of $X_i$. We represent the causal structure among the variables with a directed graph $G = (V(G), E(G))$, in which $X_i \rightarrow X_j \in E(G)$ if $X_i$ is a direct cause of $X_j$. The variables can be represented in matrix form as $X = B^\top X + N$, where, $B = [\beta_1 | \cdots | \beta_p]$ is a $p \times p$ weighted adjacency matrix of $G$, and $N = [N_1 \cdots N_p]^\top$. Elements of $N$ are jointly Gaussian.

---

1 We note that the aforementioned issue is not just specific to the case of the presence of cycles in the causal diagram and can also occur when latent confounders are allowed in a DAG structure [Tian and Pearl, 2002, Shpitser et al. 2014].

2 Note that for non-linear cyclic SEMs, even the Markov property does not necessarily hold [Spirtes, 1995], and hence, it is not clear if one can make general statements regarding the equivalence of structures regardless of the involved equations.
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 $\Sigma$ is the covariance matrix of the joint Gaussian distribution on $X$, and suffices to describe the distribution of $X$. We assume that $\Sigma$ 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 $X$. $\Theta$ can be written as

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

(1)

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

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

**Definition 1 (Independence Equivalence).** Let $\mathcal{I}(G)$ denote the set of all conditional d-separations implied by the directed graph $G$. Directed graphs $G_1$ and $G_2$ are independence equivalent (I-equivalent) denoted by $G_1 \equiv G_2$, if $\mathcal{I}(G_1) = \mathcal{I}(G_2)$.

It is known that when cycles are permitted in the structure, conditional independence is not a suitable notion of equivalence for two structures, as it does not reflect all the information in the distribution that can be used for identification of the underlying structure. That is, there exist directed graphs, which can be distinguished from observational data with probability one, despite having the same conditional d-separations. 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

$$\Theta(G) := \{\Theta : \Theta = (I - B)\Omega^{-1}(I - B)^\top, \text{ for some } (B, \Omega)$$

$$s.t. \Omega \in \text{diag}^+ \text{ and } \text{supp}(B) \subseteq \text{supp}(B_G)\},$$

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

$\Theta(G)$ is the set of all precision matrices (equivalently, distributions) that can be generated by $G$ for different choices of noise variances and edge weights in $G$. Note that in Definition 2 we allow that $B$ has a zero for an entry that $B_G$ is non-zero.

**Definition 3 (Distribution Equivalence).** Directed graphs $G_1$ 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 directed graph $G$ and distribution $\Theta$, having $\Theta \in \Theta(G)$ does not imply that all the constraints of $\Theta$, such as its conditional independences can be read off of $G$. That is, the parameters in $B$ can be designed in a way to represent certain extra constraints in the generated distribution. For instance, a complete DAG does not represent any conditional d-separations, yet all distributions are contained in its distribution set.

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

---

3See [Pearl, 2009] for the definition of d-separation.
Figure 2: Directed graphs related to Examples 1 and 2.

**Proposition 1.** Two DAGs $G_1$ and $G_2$ are equivalent if and only if they are I-equivalent.

Therefore, one does not lose any information by caring only about I-equivalence when dealing with acyclic structures.

For the case of DAGs, there exists a simple graphical test for I-equivalence: DAGs $G_1$ and $G_2$ are I-equivalent if and only if they have the same skeleton and v-structures [Verma and Pearl, 1991]. The case of general directed graphs is drastically more complicated. Graphical conditions for I-equivalence of two general directed graphs was studied in [Richardson, 1996b]. On the other hand, there are no known graphical conditions for distribution equivalence. This is the focus of Section 4.

### 3 Equivalence Test

In order to check for equivalence of structures $G_1$ and $G_2$, a baseline equivalence test would be as follows: We consider a distribution $\Theta \in \Theta(G_1)$ which is resulted from a certain choice of parameters of $G_1$ in expression (1), and then see if there is any choice of parameters for structure $G_2$, which leads to generating $\Theta$. Then we consider $G_2$ as the original generator and repeat the same procedure for $G_1$. More specifically, for directed graph $G_i$, let $Q_i = (I - B)\Omega^{-\frac{1}{2}}$, such that $\text{supp}(B) \subseteq \text{supp}(B_{G_i})$. For any choice of parameters in $G_1$ we form $\Theta = Q_1 Q_1^\top$, and check if $Q_2 Q_2^\top = \Theta$ has real solution, and vice versa.

**Example 1.** In Figure 2 we have $G_1 \equiv G_2$. This shows that unlike DAGs, equivalent directed graphs do not need to have the same skeleton or the same v-structures. Note that any distribution can be generated by $G_1$, and hence, we only need to show whether or not a distribution generated by $G_1$ can be represented by $G_2$.

Consider an arbitrary precision matrix $\Theta$ generated by $G_1$. If $\Theta$ is representable by structure $G_2$, then we should be able to decompose it as $\Theta = QQ^\top$, where $Q$ has the following form.

$$Q = \begin{bmatrix}
\sigma_1^{-1} & -\beta_{12}\sigma_2^{-1} & 0 \\
-\beta_{21}\sigma_1^{-1} & \sigma_2^{-1} & 0 \\
0 & -\beta_{32}\sigma_2^{-1} & \sigma_3^{-1}
\end{bmatrix},$$

where $\beta_{ij}$ denotes the value of the coefficient corresponding to the edge $X_i \rightarrow X_j$, and $\sigma_i$ is the standard
Then we have $\sigma$, consequently, we have $\Theta$. Equivalently, $\Theta$. For this matrix, we should have $\Theta$. Since $\Theta$, $\Theta$, $\Theta$, $\Theta$.

Finally, we should make sure that the obtained values for $a, b, c, d$ satisfy the equality $ac + bd = \Theta$: 

$$(\Theta_{12}e^2 - \Theta_{13}\Theta_{23})^2 = (\Theta_{11}e^2 - \Theta_{13}^2)(\Theta_{22}e^2 - \Theta_{23}^2).$$

Equivalently, 

$$(\Theta_{12}^2 - \Theta_{11}\Theta_{22})e^2 + (\Theta_{11}\Theta_{23}^2 + \Theta_{22}\Theta_{13}^2 - 2\Theta_{12}\Theta_{13}\Theta_{23}) = 0.$$ 

$\Rightarrow \det(\begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix})e^2 = \Theta_{11}\Theta_{23}^2 + \Theta_{22}\Theta_{13}^2 - 2\Theta_{12}\Theta_{13}\Theta_{23}.$

Since $\Theta$ is positive definite, the determinant in the left hand side is positive, and we need the right hand side to be positive.

Since $\Theta$ is generated by $G_1$, based on expression (1), it should have the following form

$$\Theta = \begin{bmatrix} c_1^2 + \beta_2\sigma_2^2 + \beta_3\sigma_3^2 & -\beta_1\sigma_2^2 - \beta_2\sigma_3^2 \\ -\beta_1\sigma_2^2 - \beta_2\sigma_3^2 & c_2^2 + \beta_3\sigma_3^2 \\ -\beta_1\sigma_2^2 - \beta_2\sigma_3^2 & -\beta_3\sigma_3^2 \end{bmatrix}.$$ 

For this matrix, we should have $\Theta_{11}\Theta_{23}^2 + \Theta_{22}\Theta_{13}^2 - 2\Theta_{12}\Theta_{13}\Theta_{23} > 0$, that is

$$(\sigma_1^2 + \beta_2\sigma_2^2 + \beta_3\sigma_3^2)(\beta_3\sigma_2^2 + \sigma_2^2 - (\beta_1\sigma_2^2 + \beta_2\beta_3\sigma_2^2)^2) - 2(\beta_1\beta_2\sigma_2^2 - (\beta_1\sigma_2^2 + \beta_2\beta_3\sigma_2^2)^2)(\beta_3\sigma_2^2)^2 > 0,$$

$\Rightarrow \beta_3^2\sigma_1^2\sigma_2^2 - \beta_3^2\sigma_2^4 + \beta_3^2\sigma_2^4 - \beta_3^2\sigma_3^4 > 0,$

where the last inequality always holds.
Although the baseline equivalence test provides a systematic approach, in many cases it can be tedious to check for the required conditions. In the following, we propose an alternative equivalence test based on the rotations of matrix $Q$.

Let $v_i$ be the $i$-th row of matrix $Q$. Therefore, $\Theta = QQ^\top$ is the Gramian matrix of the set of vectors $\{v_1, \ldots, v_p\}$. The set of generating vectors of a Gramian matrix can be determined up to isometry. That is, given $Q_2Q_2^\top = \Theta$, we have $Q_1Q_1^\top = \Theta$ if and only if $Q_1 = Q_2U$, 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, we should start with all possible choices of parameters of $Q_2$ and check if there exists an orthogonal transformation of $Q_2$, which is possible to be generated for some choice of parameters of $Q_1$, and vice versa. Therefore, only the support of the matrix before and after the orthogonal transformation matters to us. Therefore, we only need to consider rotation transformations. This can be formalized as follows: Let $Q_G := I + B_G$, which is the binary matrix that for all choices of parameters $B$ and $\Omega$, $\text{supp}(Q) \subseteq \text{supp}(Q_G)$.

**Proposition 2.** $G_1$ is distribution equivalent to $G_2$ if and only if both following conditions hold:

- For all choices of $Q_1$, there exists rotation $U^{(1)}$ such that $\text{supp}(Q_1U^{(1)}) \subseteq \text{supp}(Q_G)$.
- For all choices of $Q_2$, there exists rotation $U^{(2)}$ such that $\text{supp}(Q_2U^{(2)}) \subseteq \text{supp}(Q_G)$.

To test the existence of a rotation required in Proposition 2, we propose considering a sequence of special type of planar rotations called Givens rotations [Golub and Van Loan, 2012].

**Definition 4** (Givens rotation). A Givens rotation is a rotation in the plane spanned by two coordinate axis. For a $\theta$ radian rotation in $(j, k)$ plane, the non-zero elements of the Givens rotation matrix $G(j, k, \theta) = [g]_{p \times p}$ in $\mathbb{R}^p$ are $g_{ii} = 1$ for $i \notin \{j, k\}$, $g_{ii} = \cos(\theta)$ for $i \in \{j, k\}$, and $g_{kj} = -g_{jk} = -\sin(\theta)$, 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 $U$ as the product of them. 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.

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

**3.1 Support Rotation**

As mentioned earlier, since we need to consider all choices of parameters in the structure, we are only interested in the existence of a rotation that maps a support to another. Hence, we define support matrix and support rotation as follows.

**Definition 5** (Support matrix). Given a matrix $M$, the support matrix of $M$ is a binary matrix $\xi$ of same size with entries in $\{0, \times\}$, where $\xi_{ij} = \times$, if $M_{ij} \neq 0$, and $\xi_{ij} = 0$, otherwise. For directed graph $G$, we define its support matrix as the 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 $i, j \in [p]$, $Q_{ij}$ can be set to zero using a Givens rotation in the $(j, k)$ plane with angle $\theta = \tan^{-1}(-Q_{ij}/Q_{ik})$. When zeroing $Q_{ij}$, there may exist an index $l$ such that $Q_{lj}$ or $Q_{lk}$ will also become zero. However, since we should consider all possible choices of parameters for $Q$, we cannot take advantage of such accidental zeroings.

**Definition 6 (Support Rotation).** For a given support matrix $\xi$, consider matrix $Q$ with support matrix $\xi$, which leads to maximal $\text{supp}(QG(j, k, \tan^{-1}(-Q_{ij}/Q_{ik})))$. The support rotation $A(i, j, k)$ is a transformation on $\xi$, which zeros $\xi_{ij}$ via a rotation in $(j, k)$ plane. The output of this rotation is the support matrix of $QG(j, k, \tan^{-1}(-Q_{ij}/Q_{ik}))$.

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.** The effect of the support rotation $A(i, j, k)$ on support matrix $\xi$ will be as follows:

- If $\xi_{ij} = 0$, then applying $A(i, j, k)$ has no effect.
- If $\xi_{ij} = \times$ and $\xi_{ik} = \times$, then applying $A(i, j, k)$ makes $\xi_{ij} = 0$, and for any $l \in [p] \setminus \{i\}$ such that $\xi_{lj} = \times$ or $\xi_{lk} = \times$, it makes $\xi_{lj} = \times$ and $\xi_{lk} = \times$. (see the table below.) This is obtained by an acute rotation.

| $\bar{\xi}_{lj}$ | $\bar{\xi}_{lk}$ | $\tilde{\xi}_{lj}$ | $\tilde{\xi}_{lk}$ |
|-------------------|-------------------|-------------------|-------------------|
| $\times$          | $\times$          | $\times$          | $\times$          |
| $0$               | $\times$          | $\times$          | $\times$          |
| $\times$          | $0$               | $\times$          | $\times$          |
| $0$               | $0$               | $0$               | $0$               |

- If $\xi_{ij} = \times$ and $\xi_{ik} = 0$, then applying $A(i, j, k)$ switches columns $j$ and $k$ of $\xi$. This is obtained by a $\pi/2$ rotation.

See Figure 3 for an example visualization of the effect of support rotations on a support matrix. We classify support rotations based on their effects as follows:

- **Reduction.** $A(i, j, k)$ is a reduction if $\xi_{ij} = \bar{\xi}_{ik} = \times$, and $\xi_{ij} = \bar{\xi}_{lk}$ for all $l \in [p] \setminus \{i\}$.
• **Reversible acute rotation.** $A(i, j, k)$ is a reversible acute rotation if $\xi_{ij} = \xi_{ik} = \times$, and there exists $i' \in [p] \setminus \{i\}$ for which $\xi_{i'j} \neq \xi_{i'k}$, and $\xi_{ij} = \xi_{ik}$ for all $l \in [p] \setminus \{i, i'\}$.

• **Irreversible acute rotation.** $A(i, j, k)$ is an irreversible acute rotation if $\xi_{ij} = \xi_{ik} = \times$, and there exists $\{i', i''\} \in [p] \setminus \{i\}$ for which $\xi_{i'j} \neq \xi_{i'k}$ and $\xi_{i''j} \neq \xi_{i''k}$.

• **Column permutation.** $A(i, j, k)$ is a column permutation if $\xi_{ij} = \times$ and $\xi_{ik} = 0$.

Clearly the four cases above partition all the effects that can be obtained from a support rotation. Reduction rotations only makes one of the non-zero entries zero without changing any other entries. If $\xi$ is transformed to $\xi'$ via a reversible acute rotation $A(i, j, k)$, and $\xi_{i'j} = 0$, then $\xi'$ can be mapped back to $\xi$ via $A(i', j, k)$, hence the name reversible. Column permutations swap two columns of the support matrix and hence, can be used to form any desired permutation of the columns.

We have 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 $\xi$ and $\xi'$, we say $\xi \subseteq \xi'$ if $\text{supp}(\xi) \subseteq \text{supp}(\xi')$.

**Theorem 1.** Let $\xi_1$ and $\xi_2$ be the support matrices of directed graphs $G_1$ and $G_2$, respectively. $G_1$ is distribution equivalent to $G_2$ if and only if both following conditions hold:

• There exists a sequence of reductions, reversible acute rotations, and column permutations that maps $\xi_1$ to a subset of $\xi_2$.

• There exists a sequence of reductions, reversible acute rotations, and column permutations that maps $\xi_2$ to a subset of $\xi_1$.

In the Appendix, we provide an algorithm for enumerating all directed graphs which are distribution equivalent to a given structure. Note that even in the case of DAGs, the number of equivalents of a directed graph can be super-exponential.

**Proposition 4** (Direction of Cycles). A structure $G_1$ containing a directed cycle $C = (X_1, \ldots, X_m, X_1)$ is distribution equivalent to structure $G_2$, in which compared to $G_1$, the direction of cycle $C$ is reversed, i.e., $(X_m, \ldots, X_1, X_m)$, and any variable pointing to $X_i \in C$ in $G_1$, points to $X_j$ in $G_2$ (except $X_j$ itself), where $j = m$ if $i = 1$, and $j = i - 1$ if $1 < i \leq m$. (See Figure 4 for an example.)

The following proposition provides a necessary and sufficient condition for equivalence.
Proposition 5. Consider directed graphs $G_1$ and $G_2$ with support matrices $\xi_1$ and $\xi_2$, respectively. If each pair of columns of $\xi_1$ differ in more than one entry, then $G_1 \equiv G_2$ if and only if columns of $\xi_2$ are permutation of columns of $\xi_1$.

Example 2. In Figure 2, we have (a) $G_1 \equiv G_2$. (b) $G_1 \not\equiv G_3$. (c) $G_1 \equiv G_4$. To see $G_1 \equiv G_2$, we note that

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

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

Part (b) simply follows from Proposition 5. For Part (c), we already have $\xi_1 \subseteq \xi_4$. Also,

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

As seen in Example 2, structures $G_1$ and $G_4$ in Figure 2 are distribution equivalent. Therefore, the extra edge $X_2 \to X_1$ in $G_4$ does not enable this structure to generate any further distributions. In this case, we say structure $G_4$ is reducible. This idea is formalized as follows.

Definition 7 (Reducibility). Directed graph $G$ 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. Directed graph $G$ with support matrix $\bar{\xi}$ is reducible if and only if there exists a sequence of reversible acute rotations that enables us to apply a reduction to $\bar{\xi}$.

Proposition 6 leads to the following necessary condition for reducibility.

Proposition 7. A directed graph with no 2-cycles is irreducible.

Propositions 6 and 7 have the following corollary regarding equivalence for DAGs, which bridges our proposed approach with the classic characterization for equivalence for DAGs.

Corollary 1. DAGs $G_1$ and $G_2$ with support matrices $\xi_1$ and $\xi_2$ are equivalent if and only if there exist sequences of reversible acute rotations and column permutations that map their support matrices to one another.

Example 3. We demonstrate our approach on a familiar equivalence example on DAGs:

$G_1 : X_1 \to X_2 \to X_3 \equiv G_2 : X_1 \leftarrow X_2 \leftarrow X_3,$

$G_1 : X_1 \to X_2 \to X_3 \not\equiv G_3 : X_1 \to X_2 \leftarrow X_3.$
To see that $G_1 : X_1 \rightarrow X_2 \rightarrow X_3 \equiv G_2 : X_1 \leftarrow X_2 \leftarrow X_3$, we note that

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

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

For the second part, we note that $\xi_3$ has two columns with two zeros, while $\xi_1$ has only one column with two zeros. Therefore, reversible acute rotation and column permutations cannot help us to map $\xi_1$ to a subset of $\xi_3$.

4 Graphical Representation of the Rotation-Based Equivalence Test

In Section 3, we observed that one can examine the equivalence of two directed graphs by checking if there exist sequences of reductions, reversible acute rotations, and column permutations, which map the support matrix of each structure to a subset of the support matrix of the other. In this section, we provide the graphical representation for each of those three support rotations and a counterpart to Theorem 1 based on graphical representations. We start with a definition required for the exposition.

**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.

We observed different types of support rotations in Section 3. Those rotations can be represented graphically as follows:

- **Parent reduction.** If $X_j$ and $X_k$ are parent reducible, then any support rotation on columns $\xi_{\cdot,j}$ and $\xi_{\cdot,k}$, which zeros out a non-zero entry on those columns except $\xi_{jj}$ and $\xi_{kk}$, removes one 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, then $\{X_i\} = P_j \triangle P_k$. In this case, any support rotation on columns $\xi_{\cdot,j}$ and $\xi_{\cdot,k}$, which zeros out a non-zero entry on those columns except $\xi_{jj}$ and $\xi_{kk}$, removes one parent from $X_j$ or $X_k$ corresponding to the zeroed entry and adds the missing edge from $X_i$ to $X_j$ or $X_k$. We call this a parent exchange. The support rotation in this case is of column permutation or reversible acute rotation type.

- **Cycle reversion.** We relabel the variables so that variables involved in a cycle $C$ of length $c$ are indexed from $i$ to $i + c$. A series of column permutation rotations which shifts columns $i + 1$ to $i + c$ to left by one and moves column $i$ to location $i + c$, reverses the direction of cycle $C = (X_{i}, \ldots, X_{i+c}, X_{i})$, and any variable pointing to $X_i \in C$ will point to $X_{j'}$ (except $X_{j'}$ itself), where $j' = i + c$ if $j = i$, and $j' = j - 1$ if $i < j \leq i + c$. We call this a cycle reversion.
Figure 5: Elements of a distribution equivalence class.

Note that in the representations above, we are only excluding support rotations that lead to zeroing a diagonal entry, which do not have a graphical representation.

Equipped with the graphical representations, we present a graphical counterpart to Theorem 1:

**Proposition 8.** $G_1$ is distribution equivalent to $G_2$ if and only if both following conditions hold:

- There exists a sequence of parent reductions, parent exchanges, and cycle reversions that maps $G_1$ to a subgraph of $G_2$.
- There exists a sequence of parent reductions, parent exchanges, and cycle reversions that maps $G_2$ to a subgraph of $G_1$.

**Example 4.** Figure 5 shows an example of the elements of a distribution equivalence class. Suppose $G_1$ is the original structure. Then, 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 $G_4$. Parent exchange $A(1, 3, 1)$ results in $G_8$.

Note that given observational data from any of the structures in Figure 5, conditional independence-based tests such as CCD [Richardson, 1996a] may output a structure (for example $G_1$ without edges $X_4 \rightarrow X_1$), which is not distribution equivalent to the ground truth, which can be prevented using other statistical information in the distribution beyond the conditional independencies.

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

**Corollary 2.** DAGs $G_1$ and $G_2$ are equivalent if and only if there exist sequences of parent exchanges that map them to one another.

## 5 Learning Directed Graphs from Data

A structure $G$ requires constraints on the entries of a precision matrix $\Theta$. We will refer to such constraints as distributional constraints of $G$. Every distribution in $\Theta(G)$ should satisfy the distributional constraints of $G$. Clearly, two directed graphs are distribution equivalent if and only if they have the same distributional constraints. Distributional constraints are of the forms of equalities and inequalities. We refer to the former as hard constraints, denoted by $H(G)$, and the latter as soft constraints, denoted by $S(G)$.
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 distinguishing between $G_1$ and $G_2$. However, in practice, we usually have access only to one distribution, which is generated from a ground truth structure, and it may be also possible to 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.

The aforementioned issue arises in the case of learning DAGs and considering I-equivalence as well. The most common approach to deal with this issue in the literature is to assume faithfulness assumption, which requires that there is a one-to-one correspondence between the conditional d-separations of the ground truth structure and the conditional independences of the distribution. This is a reasonable assumption due to the fact that the Lebesgue measure of the parameters which lead to extra conditional independences in the generated distribution is zero. The case of general directed graphs is more complex, as a directed graph can require other distributional constraints besides conditional independences. Specially, we may have constraints of the form of inequalities due to cycles. Therefore, the Lebesgue measure of the parameters which lead to extra distributional constraints in the generated distribution is not necessarily zero anymore. This motivates the following weaker notion of equivalence for the task of structure learning from data.

**Definition 9 (Quasi Equivalence).** Directed graphs $G_1$ and $G_2$ are quasi equivalent, denoted by $G_1 \cong G_2$, if they share the same hard distributional constraints.

We have the following assumption for structure learning, which is a generalization of faithfulness:

**Definition 10 (Generalized faithfulness).** A distribution $\Theta$ is generalized faithful (Gen-faithful) to structure $G$ if for a hard distributional constraint $\kappa$ we have $\Theta$ satisfies $\kappa$ if and only if $\kappa \in H(G)$.

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

The second part of Assumption 1 requires that if the ground truth structure $G^*$ has no reducible edges and there exists another directed graph $G$ 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.

We propose to use a regularized maximum likelihood (ML) estimator for the task of structure learning. This is a common approach in the literature of learning Bayesian networks and was popularized by works such as [Heckerman et al., 1995] and [Chickering, 2002]. Also, works such as [Van de Geer and Bühlmann, 2013], [Fu and Zhou, 2013], [Aragam and Zhou, 2015], and [Raskutti and Uhler, 2018] considered the problem of learning a linear Gaussian Bayesian network via penalized parameter estimation. Let $X$ be the $n \times p$ data matrix. We consider an $\ell_0$-regularized ML estimator, obtained via the following unconstraint continuous optimization problem:

$$\min_{B, \Omega} \mathcal{L}(X : B, \Omega) + \lambda \|B\|_0,$$

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

13
The $\ell_0$-regularized ML estimator in (3) can be solved efficiently via a greedy search. This is the preferred choice in many score-based works on learning Bayesian networks [Heckerman et al., 1995, Chickering, 2002, Van de Geer and Bühlmann, 2013].

Remark 1. The estimator in (3) will never output a reducible directed graph, as it can increase the score by removing redundant edges. In fact, this is in line with the Occam’s Razor principle. Considering the number of edges as the causal complexity of the structure, Occam’s Razor principle suggests that among equivalent hypotheses the one with minimal complexity should be chosen.

Theorem 2. Under Assumption 1, the $\ell_0$-regularized maximum likelihood estimator outputs $\hat{G}_{\ell_0} \cong G^*$ on the population dataset.

6 Conclusion

For cyclic directed graphs, it has been known that d-separation relations cannot capture all the information in the distribution that can be used for structure learning and a suitable characterization of equivalence has been unknown so far. This is particularly unfortunate since cyclic structures are ubiquitous in real applications. We presented a general, unified notion of equivalence for linear Gaussian directed graphs and took a first step towards devising methods for characterizing the equivalence of two structures, which may be cyclic or acyclic. We also proposed a score-based method for learning the structure from observational data. We hope that our proposal in this work spurs more research in this direction in the future.
Appendices

A Proof of Proposition 1

Two DAGs are I-equivalent if and only if they have the same skeleton and v-structures [Spirtes et al., 2000]. Therefore, it suffices to show that two DAGs $G_1$ and $G_2$ are distribution equivalent if and only if they have the same skeleton and v-structures.

By Corollary 2, DAGs $G_1$ and $G_2$ are equivalent if and only if there exist sequences of parent exchanges that map them to one another. Suppose $G_1$ and $G_2$ are distribution equivalent. Therefore there exists a sequence of parent exchanges mapping one to another. Since DAGs do not have 2-cycles, parent exchange for them will only result in flipping an edge, and since the other parents of the vertices at the two ends of that edge should be the same, it does not generate or remove a v-structure. Therefore, the sequence of parent exchanges does not change the skeleton or change the set of v-structures. Therefore, $G_1$ and $G_2$ are I-equivalent.

If two DAGs $G_1$ and $G_2$ have the same skeleton and v-structures, then their difference can be demonstrated as a sequence of edge flips such that in each flip, all the parent of the two ends have been the same, which means this flip is a parent exchange. Therefore, by Corollary 2, DAGs $G_1$ and $G_2$ are distribution equivalent.

B Proof of Proposition 2

If side:
If $\text{supp}(Q_1 U^{(1)}) \subseteq \text{supp}(Q_{G_2})$, then we can simply choose the entries of $Q_1 U^{(1)}$ as the entries of $Q_2$ (as they are all free variables). Therefore,
\[
Q_2 Q_2^T = Q_1 U^{(1)} (U^{(1)})^T Q_1^T = Q_1 Q_1^T.
\]
That is, $Q_2$ can generate the distribution which was generated by $Q_1$. Since this is true for all choices of $Q_1$, and since the reverse (i.e., starting with $Q_2$) is also true, by definition, $G_1$ is distribution equivalent to $G_2$.

Only if side:
If $G_1$ is distribution equivalent to $G_2$, then for all choices of $Q_1$, generating $Q_1 Q_1^T = \Theta$, there exists $Q_2$ generated by $G_2$, such that $Q_2 Q_2^T = \Theta$. Since $Q_2$ is generated by $G_2$, by definition, $\text{supp}(Q_2) \subseteq \text{supp}(Q_{G_2})$. Also, since $Q_1 Q_1^T = \Theta$ and $Q_2 Q_2^T = \Theta$, we have $Q_2 = Q_1 U$, for some orthogonal transformation $U$, due to the fact that the generating vectors of a Grammian matrix can be determined up to isometry. Therefore, since $Q_2 = Q_1 U$ and $\text{supp}(Q_2) \subseteq \text{supp}(Q_{G_2})$, we conclude that $\text{supp}(Q_1 U) \subseteq \text{supp}(Q_{G_2})$. It remains to show that there exists a rotation $U^{(1)}$, for which $\text{supp}(Q_1 U^{(1)}) \subseteq \text{supp}(Q_{G_2})$. If $U$ is a rotation, we are done by choosing $U^{(1)} = U$. If $U$ is an improper rotation, there exists another reflection $U'$, the one which fixes the zeros and changes the sign of a subset of the rest of the entries, which does not change the support of $Q_1 U$, i.e., $\text{supp}(Q_1 U) = \text{supp}(Q_1 U U')$, and $U U'$ is a rotation. Therefore, we are done by choosing $U^{(1)} = U U'$.
Proof of Proposition 3

- If $\xi_{ij} = 0$, then by definition, the Givens rotation corresponding to $A(i, j, k)$ is a zero degree rotation. Therefore, applying $A(i, j, k)$ has no effect.
- If $\xi_{ij} = \xi_{ik} = \times$, then there exists a matrix $Q$ for which zeroing $\xi_{ij}$ is an acute rotation and the other rows of $Q$ either have no element in the $(j, k)$ plane, or if they do, they will not become aligned with either $j$ or $k$ axis in the $(j, k)$ plane after the rotation. Therefore, support $(0, 0)$ will stay at $(0, 0)$, and any other support will become $(\times, \times)$.
- If $\xi_{ij} = \times$ and $\xi_{ik} = 0$, then the $i$-th row has been aligned with the $j$ axis in the $(j, k)$ plane before the rotation and since the rotation is planar, will become aligned with the $k$ axis after the rotation, and hence we have a a $\pi/2$ rotation. Therefore, all other rows aligned with one axis will become aligned with the other axis, and any vector not aligned with either axes will remain the same. Therefore, we have support transformations $(\times, 0) \rightarrow (0, \times)$, $(0, \times) \rightarrow (\times, 0)$, $(\times, \times) \rightarrow (\times, \times)$, and $(0, 0) \rightarrow (0, 0)$, which is equivalent to switching columns $j$ and $k$.

Proof of Theorem 1

We first prove the following weaker result:

Theorem 3. Let $\xi_1$ and $\xi_2$ be the support matrices of directed graphs $G_1$ and $G_2$, respectively. $G_1$ is distribution equivalent to $G_2$ if and only if both following conditions hold:

- There exists a sequence of support rotations that maps $\xi_1$ to a subset of $\xi_2$.
- There exists a sequence of support rotations that maps $\xi_2$ to a subset of $\xi_1$.

We need the following lemma for the proof.

Lemma 1. Consider a matrix $Q$ and a support matrix $\xi$. If the support matrix of $Q$ is a subset of $\xi$, then for all $i, j, k$, the support matrix of $QG(j, k, \theta)$ is subset of $\xi A(i, j, k)$, where,

$$
\theta = \begin{cases} 
0, & \text{if } Q_{ij} = Q_{ik} = 0 \text{ and } \xi_{ij} = \xi_{ik} \neq 0, \\
0, & \text{if } Q_{ij} = Q_{ik} = 0 \text{ and } \xi_{ik} \neq \xi_{ij} = 0, \\
\pi/2, & \text{if } Q_{ij} = Q_{ik} = 0 \text{ and } \xi_{ij} \neq \xi_{ik} = 0, \\
\tan^{-1}(-Q_{ij}/Q_{ik}), & \text{otherwise}.
\end{cases}
$$

Proof. The rotation and the support rotation do not alter any columns except the $j$-th and $k$-th columns. Hence we only need to see if the desired property is satisfied by those two columns. If the support of $Q$ and $\xi$ are the same on those two columns, the desired result follows from the definition of support rotation. Otherwise,

- If the support of $(Q_{ij}, Q_{ik})$ is the same as $(\xi_{ij}, \xi_{ik})$, then the effect of the rotation on $Q$ is the same as the effect of the support rotation on $\xi$, except that if we are in the second case of Proposition 3, the support rotation cannot introduce any extra zeros in rows $[p] \setminus \{i\}$, while this is possible for the rotation on $Q$. Therefore, the support matrix of $QG(j, k, \theta)$ is subset of $\xi A(i, j, k)$.
• If $Q_{ij} \neq 0$ and $Q_{ik} = 0$, and $(\xi_{ij}, \xi_{ik}) = (\ast, \ast)$, then the rotation is a $\pm \pi/2$ while we have an acute rotation for $\xi$ (second case of Proposition 3). Hence, if a zero entry of $Q$ in a row in $[p] \setminus \{i\}$ has become non-zero after the rotation, $\xi$ has non-zero entries in both entries of that row. Therefore, the support matrix of $QG(j, k, \theta)$ is subset of $\xi A(i, j, k)$.

• If $[Q_{ij} = 0$ and $Q_{ik} \neq 0$, and $(\xi_{ij}, \xi_{ik}) = (\ast, \ast)$, or $[Q_{ij} = 0$ and $Q_{ik} = 0$, and $(\xi_{ij}, \xi_{ik}) = (0, \ast)$, or $[Q_{ij} = 0$ and $Q_{ik} = 0$, and $(\xi_{ij}, \xi_{ik}) = (\ast, \ast)$, then the rotation has no effect on $Q$, while the support rotation can only turn some of the zero entries in rows $[p] \setminus \{i\}$ to non-zero. Therefore, the support matrix of $QG(j, k, \theta)$ is subset of $\xi A(i, j, k)$.

• Finally, if $[Q_{ij} = 0$ and $Q_{ik} = 0$, and $(\xi_{ij}, \xi_{ik}) = (\ast, 0)$, then by the statement of the lemma, the rotation on $Q$ will be $\pi/2$. Due to this fact and part three of Proposition 3, for both $Q$ and $\xi$, columns $j$ and $k$ will be flipped. Therefore, the support matrix of $QG(j, k, \theta)$ is subset of $\xi A(i, j, k)$.

Proof of Theorem 3. By Propositions 2, it suffices to show that there exists a sequence of support rotations $A_{11}, \ldots, A_{m}$, such that $\xi_{1} A_{1}, \ldots, A_{m} \subseteq \xi_{2}$ if and only if for all choices of $Q$, there exists a sequence of Givens rotations $G_{1}, \ldots, G_{m}$ such that $\text{supp}(Q_{1} G_{1}, \ldots, G_{m}) \subseteq \text{supp}(Q_{G_{2}})$.

Only if side:

For any matrix $Q_{1}$, by definition, the support matrix of $Q_{1}$ is a subset of $\xi_{1}$. In the sequence of support rotations, use the first support rotation $A_{1}(i, j, k)$ to generate Givens rotation $G_{1}(j, k, \theta)$, where $\theta$ is defined in the statement of Lemma 1. Therefore, by Lemma 1, the support matrix of $Q_{1} G_{1}(j, k, \theta)$ is a subset of $\xi_{1} A_{1}(i, j, k)$. Repeating this procedure, we see that the support matrix of $Q_{1} G_{1}, \ldots, G_{m}$ is a subset of $\xi_{1} A_{1}, \ldots, A_{m}$. Now, by the assumption, $\xi_{1} A_{1}, \ldots, A_{m} \subseteq \xi_{2}$, and by definition, $\text{supp}(\xi_{2}) = \text{supp}(Q_{G_{2}})$. Therefore, $\text{supp}(Q_{1} G_{1}, \ldots, G_{m}) \subseteq \text{supp}(Q_{G_{2}})$.

If side:

Consider Givens rotation $G(j, k, \theta)$ applied to matrix $Q$. The effect of this rotation is one of the following:

1. For an acute rotation, zeroing a subset of entries in columns $j$ and $k$.
2. For a $\pm \pi/2$ rotation, swapping the support of columns $j$ and $k$.
3. For an acute rotation, making no entries zero, while making a subset of the entries in columns $j$ and $k$ non-zero.
4. For an acute rotation, no change to $\text{supp}(Q)$.

Since the assumption is true for all $Q$, we focus on matrices with support matrix $\xi_{1}$ (i.e., none of the free parameters are set at zero). If in case 1 above the subset has more than one element, more than one rows of $Q$ have been aligned on the $(j, k)$ plane, not on the $j$ and $k$ axes. Therefore, there exists another $Q$ (i.e., another choice of free parameters), in which those rows are not aligned. Consider $Q^*$ for which no such alignment happens, and hence, each of the Givens rotations in its sequence of rotations that causes case 1 above, only makes one entry zero. Therefore, its corresponding sequence of rotations acts exactly the same as support rotations for effects 1 and 2 above, in terms of their effect on the support.

Hence, the proof is complete by showing that cases 3 and 4 can be ignored, because we assumed that the support matrix of $Q^*$ is $\xi_{1}$, and each not ignored Givens rotation corresponds to a support rotation, and by definition, $\text{supp}(Q_{G_{2}}) = \text{supp}(\xi_{2})$. Clearly, case 4 can be ignored as it has no effect.
on the support. For case 3, we note that this effect only adds elements to the support, and hence we want the support after rotations to be a subset of \( \text{supp}(Q_{G_2}) \), the rotations of this type do not serve for that purpose. Therefore, if we ignore such rotations, the resulting support would be smaller compared to the case of considering these rotations. Note that if due to such rotation entry \( Q_{ij} \) has become non-zero and later in the sequence there exists a type 1 rotation making \( Q_{ij} \) zero again, we already have zero in position \((i,j)\) and that type 1 rotation should be ignored as well.

\[
\text{Lemma 2. All the support rotations for checking the distribution equivalence of two digraphs should be lossless.}
\]

We need the following lemma for the proof.

\[
\text{Lemma 3. If support matrix } \xi \text{ is mapped to } \xi' \text{ via a support rotation, then } \Theta(\xi) \subseteq \Theta(\xi').
\]

\[
\text{Proof. For reduction, reversible acute rotation, and column permutation, we have } \Theta(\xi) = \Theta(\xi'), \text{ and irreversible acute rotation only introduces extra free variables, and hence, leads to } \Theta(\xi) \subseteq \Theta(\xi'). \text{ To make the argument regarding irreversible acute rotation rigorous, consider irreversible acute rotation } A(i, j, k), \text{ which zeros } \xi_{ij}. \text{ For all } l \in [p] \setminus \{i\}, \text{ if } \xi_{lj} \neq 0, \text{ this rotation results in } (\xi_{ij}, \xi_{lk}) = (\times, \times). \text{ Suppose } (\xi_{i'j'}, \xi_{i'k'}) = (0, \times). A(i', j, k) \text{ will be a reversible acute rotation for } \xi' \text{ and leads to } \xi'' \text{ such that } \xi \subseteq \xi''. \text{ Therefore, } \Theta(\xi) \subseteq \Theta(\xi'') = \Theta(\xi'). \]

\[
\text{Proof of Lemma 2. If support matrix } \xi \text{ is mapped to } \xi' \text{ via a lossy support rotation, i.e., } \Theta(\xi) \neq \Theta(\xi') \text{ then by Lemma 3 we have } \Theta(\xi) \subsetneq \Theta(\xi'). \text{ Suppose we want to check the equivalence of digraphs } G_1 \text{ and } G_2 \text{ with support matrices } \xi_1 \text{ and } \xi_2, \text{ respectively. We note that } \Theta(G_1) = \Theta(\xi_1). \text{ Suppose } \xi_1 \text{ is mapped to } \xi \text{ through a sequence of support rotations, including a lossy rotation, which in turn is mapped to } \xi' \subseteq \xi_2. \text{ Therefore, }
\]

\[
\Theta(G_1) = \Theta(\xi_1) \subsetneq \Theta(\xi) \subseteq \Theta(\xi') \subseteq \Theta(\xi_2) = \Theta(G_2).
\]

Therefore,

\[
\Theta(G_1) \neq \Theta(G_2).
\]
Proof. The if side is clear by Theorem 3. For the only if side, by Theorem 3 and Lemma 2, we show that if \( \xi_1 \) can be mapped to \( \xi_2 \) via a sequence of lossless support rotations (i.e., \( \Theta(\xi_1) = \Theta(\xi_2) \)) including an irreversible acute rotation, then there exists a sequence of support rotations which does not include any irreversible acute rotations that maps \( \xi_1 \) to a subset of \( \xi_2 \).

We show that every irreversible acute rotation can be replaced by other types of support rotation. Consider the first irreversible acute rotation \( \Lambda(i, j, k) \) in the sequence, which maps \( \xi \) to \( \xi' \). Applying this rotation, we have \( (\xi_{ij}', \xi_{ik}') = (0, \times) \), and columns \( \xi_{-, j}' \) and \( \xi_{-, k}' \) agree on the rest of the entries. Suppose, prior to applying this rotation, columns \( \xi_{-, j} \) and \( \xi_{-, k} \) disagree on \( m \) entries in rows with indices \( \text{diff} = \{s_1, \ldots, s_m\} \). Let

\[
diff_j = \{l : l \in \text{diff}, \xi_{lj} = 0\},
\]

\[
diff_k = \{l : l \in \text{diff}, \xi_{lk} = 0\},
\]

and

\[
M = \begin{cases} 
\max\{m_j, m_k\}, & m_j \neq m_k, \\
m_j + 1, & \text{otherwise.}
\end{cases}
\]

where \( m_j = |\text{diff}_j| \) and \( m_k = |\text{diff}_k| \). We can always swap two columns, hence, without loss of generality, assume \( M = m_j + 1_{\{m_j=m_k\}} \).

Claim 1. \( \xi \) can be transformed via reduction and reversible acute rotation to a support matrix, in which there exist columns with indices \( \{t_1, \ldots, t_{M-1}\} \) such that the sub-matrix of \( \xi \) on columns \( \{t_1, \ldots, t_{M-1}, j, k\} \) and rows \( \text{diff} \cup \{i\} \) has a column with \( i \) zeros, for all \( i \in \{0, 1, \ldots, M\} \), and the sub-matrix of \( \xi \) on columns \( \{t_1, \ldots, t_{M-1}, j, k\} \) and the rest of the rows has equal columns.

Proof of Claim 1. Since \( \Lambda(i, j, k) \) is lossless, we can map \( \xi' \) to a subset of \( \xi \). Therefore, we should be able to introduce zeros in \( \xi' \) in indices \( \text{diff}_j \) of column \( j \) and indices \( \text{diff}_k \) of column \( k \), without removing the existing zeros, except potentially \( \xi'_{ij} \). We first use a reversible acute rotation on columns \( j \) and \( k \) to move the newly introduced zero in \( \xi'_{ij} \) to the first index in \( \text{diff}_j \) and we denote the resulting support matrix by \( \xi_{n(1)} \). We note that reduction is the only support rotation, which increases the number of zeros in the support matrix. Therefore, we need one reduction for reviving each of the \( m - 1 \) other removed zeros in the transformation of \( \xi \) to \( \xi' \).

The claim can be proven by induction. The base of the induction, i.e., for \( M = 2 \) can be proven as follows:

- **Case 1**: \( m_j = m_k = 1 \). In order to have the zero in column \( k \), we need to perform a reduction, for which, we need another column \( \xi_{-, j}^{(1)} \) equal to \( \xi_{-, j}^{(1)} \), i.e., \( d_H(\xi_{-, j}^{(1)}, \xi_{-, j}^{(1)}) = 0 \), where \( d_H(\cdot, \cdot) \) denotes the Hamming distance between its two arguments. Since the original irreversible acute rotation was on the \( (j, k) \) plane and did not affect other columns, the column \( t_1 \) with the aforementioned property exists in the original support matrix \( \xi \) as well, i.e., \( \xi_{-, j_1} = \xi_{-, j_1}^{(1)} \). Now, a reversible acute rotation can be performed on columns \( t_1 \) and \( k \) to set \( d_H(\xi_{-, j_1}, \xi_{-, j_1}) = 0 \), and then a reduction can be performed to introduce another zero in column \( j \) of \( \xi \). The resulting support matrix has the desired property stated in the claim.
• Case 2: $m_j = 2, m_k = 0$. In order to have the zero in the second index of $\text{diff}_j$, we need to perform a reduction, for which, we need another column equal to $\xi_{i,j}^{(1)}$. This can be obtained by one of the following cases:

  - There already exists a column $t_1$, such that $d_H(\xi_{i,j_1}^{(1)}, \xi_{i,j}^{(1)}) = 0$. Similar to Case 1, this implies that column $t_1$ also exists in $\xi$. Therefore, $\xi$ has the desired property.
  
  - There exists a column $t_1$, such that $d_H(\xi_{i,j_1}^{(1)}, \xi_{i,j}^{(1)}) \neq 0$, but $d_H(\xi_{i,j_1}^{(1)}, \xi_{i,k}^{(1)}) = 1$. Similar to Case 1, this implies that column $t_1$ also exists in $\xi$. Therefore, a reversible acute rotation can transform $\xi$ to a support matrix with the desired property.

  - There exists a column $t_1$, such that $d_H(\xi_{i,j_1}^{(1)}, \xi_{i,j}^{(1)}) = 0$. Similar to Case 1, this implies that column $t_1$ also exists in $\xi$. Therefore, two reductions, one on columns $(t_1, k)$, and then one on columns $(t_1, j)$ can transform $\xi$ to a support matrix with the desired property.

• Case 3: $m_j = 2, m_k = 1$. In order to have the zero in column $k$, we need to perform a reduction, for which, we need another column $t_1$ equal to column $k$, i.e., $d_H(\xi_{i,j_1}^{(1)}, \xi_{i,k}^{(1)}) = 0$. Similar to Case 1, this implies that column $t_1$ also exists in $\xi$. Therefore, $\xi$ has the property desired in the claim.

Now, suppose the property holds for $M = n$. To show that it also holds for $M = n + 1$, a reasoning same as the one provided for the base case of the induction can be used, and it can be shown that for the required extra reduction, an extra column $t_n$ should exist in $\xi$.

By Claim 1, $\xi$ can be transformed via reduction and reversible acute rotation to a support matrix with the stated property. Therefore, we assume $\xi$ has the property. Therefore, we have columns $\{t_1, \ldots, t_{M-1}, j, k\}$ with any number of zeros $0 \leq i \leq M$ on rows $\text{diff} \cup \{i\}$, and it is easy to see the $i$ zeros in these columns can be relocated to any other indices via only reversible acute rotations amongst these columns. Therefore, any effect sought to be achieved via columns $j$ and $k$ of $\xi'$, can be obtained via columns $\{t_1, \ldots, t_{M-1}, j, k\}$ of $\xi$, and hence, the irreversible acute rotation could have been replaced by other types of rotations.

\section*{E Proof of Proposition 4}

To show that the property holds for cycle $C = (X_1, \ldots, X_m, X_1)$, we note that our desired support matrix is $\xi_1$, when columns 2 to $m$ are all shifted to left by one, and column 1 is moved to location $m$. Therefore, it suffices to first flip columns 1 and 2, then 2 and 3, all the way to $m - 1$ and $m$. For each flip, we use the third part of Proposition 3. For instance, for flipping columns $j$ and $j + 1$, we find row $i$ such that $\xi_{i,j} \neq \xi_{i,j+1}$ (if there is no such row, then no flip for those columns is needed as they are already the same). If, say $\xi_{i,j} = \times$, we use support rotation $A(i, j, j + 1)$ for flipping columns $j$ and $j + 1$. Following the same reasoning, we see that support rotation of $\xi_2$ leads to a subset of $\xi_1$. 

20
If side:
If columns of $\xi_2$ are permutation of columns of $\xi_1$, then $\xi_1$ can be mapped to $\xi_2$ and vice versa via a sequence of column permutation rotations. Therefore, by Theorem 1, $G_1 \equiv G_2$.

Only if side:
If $G_1 \equiv G_2$, the by Theorem 1, $\xi_1$ can be mapped to a subset of $\xi_2$ and $\xi_2$ can be mapped to a subset of $\xi_1$, both via only reductions, reversible acute rotations and column permutations. If each pair of column of $\xi_1$ are different in more than one entry, then we are not able to perform any reversible acute rotations and reductions. Therefore, we have been able to perform the mapping merely via column permutations. Therefore, columns of $\xi_2$ are permutation of columns of $\xi_1$.

G  Proof of Proposition 6

Only if side:
By definition, digraph $G$ is reducible if there exists digraph $G'$ such that $G \equiv G'$ and $\xi' \subset \xi$. By Theorem 1, $\xi$ can be mapped to a subset of $\xi'$ via a sequence of support rotations comprised of reductions, reversible acute rotations and column permutations. We note that reduction is the only support rotation, which increases the number of zeros in the support matrix. Therefore, there should be a reduction in the sequence. We can always swap any two columns and the location of two columns does not influence the feasibility of reduction or reversible acute rotations. Therefore, column permutations can be ignored in reducibility.

If side:
Suppose the performed reduction turns a non-zero entry in column $j$ to zero, using a reduction on columns $j$ and $k$. Note that prior to the reduction, these columns have the same number of zeros and in order to be able to perform the reduction a sequence of reversible acute rotations have been performed to prepare column $k$ such that the hamming distance of columns $j$ and $k$ be equal to zero. That is, its zeros have been moved to match the zero pattern of column $j$. We can always assume that we only moved the zeros of column $k$, as if there are columns to move the zeros of column $j$, they can be used to move the zeros of column $k$ as well. The only concern is that the zeroed entry may be on the diagonal. In this case, a reversible acute rotation can be performed on columns $j$ and $k$ to move the new zero to another index of column $j$. Also, entry $(j,j)$ cannot be the only non-zero entry of column $j$; otherwise, column $k$ should also have only one non-zero entry, which should initially be located at $(k,k)$. Therefore, to perform a reversible acute rotation on any other column $l$ and $k$, column $l$ should have only two non-zero entries, on $(k,l)$ and $(j,l)$, while one of them should initially be located at $(l,l)$. This reasoning can be repeated $p$ times and leads to the contradiction that the final column is not allowed to have a non-zero entry on the diagonal, which contradicts the fact that $\xi$ is the support matrix corresponding to a digraph. Finally, all the performed reversible acute rotations can be done in the reverse direction to obtain the initial zero pattern for columns $[p] \setminus \{j\}$. 
H Proof of Proposition 7

Using Proposition 6, we show that for directed graph $G$ with support matrix $\xi$, if there exists a sequence of reversible support rotations that enables us to apply a reduction to $\xi$, then $G$ has a 2-cycle. Suppose the reduction is performed on columns $j$ and $k$, to turn a non-zero entry of column $j$ to zero. If no reversible support rotations prior to the reduction is needed, it implies that already columns $j$ and $k$ are identical. Therefore, $\xi_{jk} = \xi_{jj} = \times$, and $\xi_{kj} = \xi_{kk} = \times$. Therefore, there exists a 2-cycle between $j$ and $k$ and the proof is complete. Therefore, we assume some reversible support rotations are needed.

Consider the first rotation in the sequence of reversible support rotations applied to column $k$. Assume it is performed on columns $t_1$ and $k$. Therefore, the support of column $t_1$ has one element more than the support of column $k$, and the Hamming distance between these two columns is one. The only way that this does not cause a 2-cycle between $t_1$ and $k$ is that $\xi_{t_1k} = 0$, and $\xi_{kt_1} = \times$, and all the entries show be the same. This rotation is supposed to move the extra zero in column $k$ to an index, which is zero in column $j$ (to reduce the Hamming distance between columns $j$ and $k$). Therefore, since after this rotation, $\xi_{t_1k}$ will become non-zero, we should have $\xi_{t_1j} = \times$. This will lead to a 2-cycle unless $\xi_{jt_1} = 0$. Now, if $\xi_{jt_1} = 0$, because all the entries of columns $t_1$ and $k$ where the same, we also have $\xi_{jk} = 0$. This gives us two options for $\xi_{kj}$:

- If $\xi_{kj} = 0$, then we need another column $t_2$ so that we perform a reversible acute rotation on columns $t_2$ and $k$ to move $\xi_{jk} = 0$ to entry $\xi_{kk}$, which is currently non-zero. This means that columns $t_2$ and $k$ should be the same on all the entries, except that $\xi_{jt_2} = \times$, but $\xi_{jk} = 0$. Therefore, $\xi_{kt_2} = \xi_{kk} = \times$ and $\xi_{t_2k} = \xi_{t_2t_2} = \times$, which implies that there is a 2-cycle between $t_2$ and $k$.

- If $\xi_{kj} = \times$, then in order for columns $k$ and $j$ to have the same number of non-zero entries, there should exist index $l$ such that $\xi_{lk} = \times$, and $\xi_{lj} = 0$. Now, we need another column $t_2$ so that we perform a reversible acute rotation on columns $t_2$ and $k$ to move $\xi_{jk} = 0$ to entry $\xi_{lk}$. This means that columns $t_2$ and $k$ should be the same on all the entries, except that $\xi_{jt_2} = \times$, but $\xi_{jk} = 0$. Therefore, $\xi_{kt_2} = \xi_{kk} = \times$ and $\xi_{t_2k} = \xi_{t_2t_2} = \times$, which implies that there is a 2-cycle between $t_2$ and $k$.

I Proof of Corollary 1

We first prove the following corollary:

Corollary 3. Irreducible directed graphs $G_1$ and $G_2$ with support matrices $\xi_1$ and $\xi_2$ are equivalent if and only if there exist sequences of reversible acute rotations and column permutations that map their support matrices to one another.

Proof. By Proposition 6, there does not exist a sequence of reversible acute rotations that enables us to apply a reduction to the support matrix. Therefore, we only need to consider reversible acute rotations and column permutations, and we need to map one support matrix to the other, rather than mapping it to a subset of the other. 

22
Proof. DAGs do not have 2-cycles. Therefore, by Proposition 7, DAGs are irreducible. Therefore, the result follows from Corollary 3.

J Proof of Proposition 8

If side:
If there exist sequences of parent reduction, parent exchange, and cycle reversion, mapping one graph to a subgraph of the other, then there exist sequences of reduction, reversible acute rotation, and column permutation mapping the support matrix of one graph to a subset of the support matrix of the other. Therefore, by Theorem 1, \( G_1 \) is distribution equivalent to \( G_2 \).

Only if side:
The proof of the only if side consists of two steps:

- **Step 1.** We note that
  1. All support rotations of reduction type, that do not make a diagonal entry zero are representable by a parent reduction. This is clear from the definitions of reduction and parent reduction.
  2. All reversible acute rotations, that do not make a diagonal entry zero are representable by a parent exchange. This is clear from the definitions of reversible acute rotation and parent exchange.
  3. If we have a reversible acute rotation and a column permutation on columns \( j \) and \( k \) such that the reversible acute rotation makes the diagonal entry \( \xi_{jj} \) zero and then the column permutation swaps columns \( j \) and \( k \) (we call such a pair a flip pair), then this pair can be replaced by a reversible acute rotation that makes the non-diagonal entry \( \xi_{jk} \) zero, and hence, is representable by a parent exchange.
  4. If we start with a support matrix with no diagonal entries equal to zero and by performing a sequence of column permutations reach another support matrix with no diagonal entries equal to zero, then this sequence is representable by a cycle reversion. To see this, we note that if after the sequence of column permutations, column \( j \) has moved to location \( k \), it implies that its \( j \)-th and \( k \)-th elements are non-zero. Therefore, the original support matrix corresponds to a graph containing the edge \( j \rightarrow k \), and the final support matrix corresponds to a graph containing the edge \( k \rightarrow j \). This reasoning identifies the cycle before, and the reversed cycle after the transformation.

Step 1 implies that if we have a sequence of support rotations which includes 1. reduction rotations, that do not make a diagonal entry zero, 2. reversible acute rotations, that do not make a diagonal entry zero, 3. flip pairs, and 4. sequence of column permutations starting and ending on a support matrix with non-zero diagonal entries, (we call such a sequence, a representable sequence) then we can represent this sequence with a sequence of parent reductions, parent exchanges, and cycle reversions.

- **Step 2.** If \( G_1 \) is distribution equivalent to \( G_2 \), then by Theorem 1, there exists a sequence of reduction, reversible acute rotations, and column permutation mapping the support matrix
of one to the other. We show that in this case, there exists a representable sequence as well that maps the support matrix of one to the other. Therefore, by Step 1 the only if side will be concluded.

We note that since $\xi_1$ is a support matrix of a directed graph, it does not have any zeros on the main diagonal. Given the sequence of support rotations, the column permutations do not enable us or prevent us from performing reversible acute rotations and reductions, and merely change the indices of the columns. Therefore, we can have an equivalent sequence of support rotations, in which we have moved all the column permutations, except those involved in flip pairs, to the end of the sequence. Consider the first rotation in the sequence of the rotations which zeros out a diagonal entry. If this rotation is of reduction type and has zeroed out $\xi_{ii}$ using columns $i$ and $j$, then $\xi_{ij}$ should have been non-zero. Therefore, we can instead replace it by zeroing $\xi_{ij}$, and use column $j$ instead of column $i$ in the next steps. If this rotation is of reversible acute rotation type and has zeroed out $\xi_{ii}$ using columns $i$ and $j$, then $\xi_{ij}$ should have been non-zero. Therefore, again we can instead replace it by zeroing $\xi_{ij}$, and use column $j$ instead of column $i$ in the next steps. Therefore, we can perform all the reductions and reversible acute rotations and from $\xi_1$ obtain $\xi'_1$, which does not have any zeros on the main diagonal, and via a sequence of column permutations can be mapped to a subset of $\xi_2$.

Now, we perform the reverse of that sequence of column permutations on $\xi_2$, which gives us a superset of $\xi'_2$ (call it $\xi''_2$), and hence, does not have any zeros on the main diagonal. Therefore, since $\xi_2$ is a support matrix of a directed graph and hence, it also does not have any zeros on the main diagonal, by part 4 of Step 1, this is equivalent to a cycle reversion. $\xi''_2$ is a superset of $\xi'_2$, and both $\xi''_2$ and $\xi'_1$ are graphically representable. By Lemma 2, the corresponding directed graph of $\xi''_2$ is the same (if the directed graph corresponding to $\xi''_2$ is irreducible) or reducible to the directed graph corresponding to $\xi'_1$. Therefore, by Proposition 6 we can perform the reduction via a sequence of reversible acute rotations. Similar to the reasoning in the previous paragraph, since we start with a support matrix with no zeros on the main diagonal, this can be done without zeroing any element of the main diagonal, and hence, we can map $\xi''_2$ to $\xi'_1$. Finally, reversing the reversible acute rotations of the sequence from $\xi_1$ to $\xi'_1$, we obtain a subset of $\xi_1$, and the whole sequence from $\xi_2$ to a subset of $\xi_1$ is a representable sequence. Similarly, we can construct a representable sequence mapping $\xi_1$ to a subset of $\xi_2$, which completes the proof.

K Proof of Corollary 2

DAGs do not have 2-cycles. Therefore, by Proposition 7 DAGs are irreducible. Hence, a parent reduction cannot be performed. Also, DAGs do not have cycles. Hence, there will not be any cycle reversion. Therefore, the result follows from Proposition 8.
L  Proof of Theorem 2

Let $G^*$ and $\Theta$ be the ground truth structure and the generated distribution, and for an ML estimator, assume we are capable of finding a correct pair $(\hat{B}_{ML}, \hat{\Theta}_{ML})$, such that $(I - \hat{B}_{ML})\hat{\Theta}_{ML}^{-1}(I - \hat{B}_{ML})^\top = \Theta$ and denote the directed graph corresponding to $\hat{B}_{ML}$ by $\hat{G}_{ML}$. We have $\Theta \in \Theta(\hat{G}_{ML})$, which implies that $\Theta$ contains all the distributional constraints of $\hat{G}_{ML}$. Therefore, under Assumption 1, we have $H(\hat{G}_{ML}) \subseteq H(G^*)$.

Let $(\hat{B}_{\ell_0}, \hat{\Theta}_{\ell_0})$ be the output of $\ell_0$-regularized ML estimator, and denote the directed graph corresponding to $\hat{B}_{\ell_0}$ by $\hat{G}_{\ell_0}$. Since the likelihood term increases much faster with the sample size compared to the penalty term, asymptotically, we still have the desired properties that $\Theta$ contains all the distributional constraints of $\hat{G}_{\ell_0}$, and hence, under Assumption 1, we again have $H(\hat{G}_{\ell_0}) \subseteq H(G^*)$.

Now, consider an irreducible equivalent of $G^*$, denoted by $G^\dagger$. Since $H(G^*) = H(G^\dagger)$, we have $H(\hat{G}_{\ell_0}) \subseteq H(G^\dagger)$. Also, because of the penalty term we have $|E(\hat{G}_{\ell_0})| \leq |E(G^\dagger)|$, otherwise the algorithm would have outputted $G^\dagger$. Therefore, by Assumption 1, we have $H(\hat{G}_{\ell_0}) = H(G^\dagger)$, and hence $H(\hat{G}_{\ell_0}) = H(G^*)$. Therefore, by definition, $\hat{G}_{\ell_0} \cong G^*$.

M  An Algorithm for Enumerating Members of an Equivalence Class

**Algorithm 1** Enumerating equivalent structures

1. **Input:** $p \times p$ support matrix $\zeta$.
2. **Initiation:** $S = \emptyset$.
3. Add all column permutations of $\zeta$ to the set $S$.
4. $S = f(\zeta, S)$.
5. Remove all support matrices in $S$ that have zero on their main diagonal.
6. **Output:** $S$.

7. function $S = f(\zeta, S)$
8. for $i = 1$ to $p - 1$ do
9.   for $j = i + 1$ to $p$ do
10.  if $d_H(\zeta_{i,j}, \zeta_{j,i}) = 0$ then
11.    for $k = 1$ to $p$ do
12.       $\zeta' = \zeta$.
13.       if $\zeta'_{k,i} \neq 0$ then
14.          $\zeta'_{k,j} = 0$.
15.        if $\zeta' \notin S$ then
16.           Add all column permutations of $\zeta'$ to the set $S$.
17.           $S = f(\zeta', S)$.
18.        end if
19.       end if
20.     end for
21.   end if
22. end if
23. if $d_H(\zeta_{i,j}, \zeta_{j,i}) = 1$ then
24.   $l = \arg\min_{i,j} \{d_H(\zeta_{i,j}, 0), d_H(\zeta_{j,i}, 0)\}$
25. for $k = 1$ to $p$ do
26.   $\zeta' = \zeta$.
27.   if $\zeta'_{k,l} 
eq 0$ then
28.     $\zeta'_{k,l} = 0$.
29.   end if
30. if $\zeta' \notin S$ then
31.   Add all column permutations of $\zeta'$ to the set $S$.
32.   $S = f(\zeta', S)$.
33. end if
34. end if
35. end for
36. end for
37. return $S$.

In the algorithm, $d_H(\cdot, \cdot)$ denotes Hamming distance.
References

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

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

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

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

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

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

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

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

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

Thomas Richardson. 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*, pages 462–469. Morgan Kaufmann Publishers Inc., 1996b.

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

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

Liam Solus, Yuhao Wang, Lenka Matejovicova, and Caroline Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.
Peter Spirtes. Directed cyclic graphical representations of feedback models. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pages 491–498. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. Causation, prediction, and search. MIT press, 2000.

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

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

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

Sara Van de Geer and Peter Bühlmann. ℓ₀-penalized maximum likelihood for sparse directed acyclic graphs. The Annals of Statistics, 41(2):536–567, 2013.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. UCLA, Computer Science Department, 1991.

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

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In Advances in Neural Information Processing Systems, pages 9472–9483, 2018.