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

Supplementary Materials

A Proof of Proposition 1

Two DAGs are Markov equivalent if and only if they have the same skeleton and v-structures (Verma & Pearl, 1991). 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 Markov 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 $\text{supp}(Q_1U^{(1)}) \subseteq \text{supp}(Q_{G_2})$, then we can simply choose the entries of $Q_1U^{(1)}$ as the entries of Q_2 (as they are all free variables). Therefore,

$$Q_2Q_2^\top = Q_1U^{(1)}(U^{(1)})^\top Q_1^\top = Q_1Q_1^\top.$$

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 $Q_1Q_1^\top = \Theta$, there exists Q_2 generated by G_2, such that $Q_2Q_2^\top = \Theta$. Since Q_2 is generated by G_2, by definition, $\text{supp}(Q_2) \subseteq \text{supp}(Q_{G_2})$. Also, since $Q_1Q_1^\top = \Theta$ and $Q_2Q_2^\top = \Theta$, we have $Q_2 = Q_1U$, for some orthogonal transformation U, due to the fact that the generating vectors of a Gramian matrix can be determined up to isometry. Therefore, since $Q_2 = Q_1U$ and $\text{supp}(Q_2) \subseteq \text{supp}(Q_{G_2})$, we conclude that $\text{supp}(Q_1U) \subseteq \text{supp}(Q_{G_2})$. It remains to show that there exists a rotation $U^{(1)}$, for which $\text{supp}(Q_1U^{(1)}) \subseteq \text{supp}(Q_{G_2})$. Note that U is an orthogonal transformation and hence, $UU^\top = I$ and $\det(U) = 1$ or -1.

- If $\det(U) = 1$, it means that U is a rotation and we are done by choosing $U^{(1)} = U$.

- If $\det(U) = -1$ (i.e., U is an improper rotation), all we need is to find an orthogonal transformation V, such that (a) $\mathrm{supp}(Q_1 U) = \mathrm{supp}(Q_1 UV)$, i.e., it does not change the support, (b) $\det(V) = -1$, which implies that $\det(UV) = 1$. That is, adding the transformation V to U does not change the support but makes the combination UV into a rotation. Finding such a V is easy, simply choosing a diagonal matrix with an odd number of diagonal entries equal to -1 and the rest equal to 1. This will not change the support and only changes the sign of a subset of the entries. Therefore, we are done by choosing $U^{(1)} = UV$. Note that we are not forced to add a specific reflection at the end, we just add a particular one to do a sign flipping to show that the improper rotation can be changed into a rotation.

C Proof of Proposition 3

- If $\xi_{i,j} = 0$, 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 $\xi_{i,j} = \xi_{i,k} = \times$, then there exists a matrix Q for which zeroing $\xi_{i,j}$ 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 $\xi_{i,j} = \times$ and $\xi_{i,k} = 0$, 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 $\pi/2$ 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 $(\times,0) \to (0,\times)$, $(0,\times) \to (\times,0)$, $(\times,\times) \to (\times,\times)$, and $(0,0) \to (0,0)$, which is equivalent to switching columns j and k.

D Proof of Theorem 1

We first prove the following weaker result:

Theorem 4. *Let ξ_1 and ξ_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 ξ_1 to a subset of ξ_2.*
- *There exists a sequence of support rotations that maps ξ_2 to a subset of ξ_1.*

We need the following lemma for the proof.

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

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

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 ξ are the same on those two columns, the desired result follows from the definition of support rotation. Otherwise,

- If the support of $(Q_{i,j}, Q_{i,k})$ is the same as $(\xi_{i,j}, \xi_{i,k})$, then the effect of the rotation on Q is the same as the effect of the support rotation on ξ, 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 $Q_{i,j} = 0$ and $Q_{i,k} = 0$, and $(\xi_{i,j}, \xi_{i,k}) = (\times, \times)$, then the rotation is a $\pm\pi/2$ while we have an acute rotation for ξ (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, ξ 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 $[Q_{i,j} = 0$ and $Q_{i,k} = 0$, and $(\xi_{i,j}, \xi_{i,k}) = (\times, \times)]$, or $[Q_{i,j} = 0$ and $Q_{i,k} = 0$, and $(\xi_{i,j}, \xi_{i,k}) = (0, \times)]$, or $[Q_{i,j} = 0$ and $Q_{i,k} = 0$, and $(\xi_{i,j}, \xi_{i,k}) = (\times, \times)]$, 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 $[Q_{i,j} = 0$ and $Q_{i,k} = 0$, and $(\xi_{i,j}, \xi_{i,k}) = (\times, 0)]$, 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 ξ, 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 4. By Propositions 2, it suffices to show that there exists a sequence of support rotations $A_1, \cdots A_m$, such that $\xi_1 A_1, \cdots A_m \subseteq \xi_2$ if and only if for all choices of Q_1, there exists a sequence of Givens rotations $G_1, \cdots G_{m'}$ such that $\text{supp}(Q_1 G_1, \cdots G_{m'}) \subseteq \text{supp}(Q_{G_2})$.

Only if side:
For any matrix Q_1, by definition, the support matrix of Q_1 is a subset of ξ_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 θ 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 $Q_1 G_1, \cdots G_m$ is a subset of $\xi_1 A_1, \cdots A_m$. Now, by the assumption, $\xi_1 A_1, \cdots A_m \subseteq \xi_2$, and by definition, $\text{supp}(\xi_2) = \text{supp}(Q_{G_2})$. Therefore, $\text{supp}(Q_1 G_1, \cdots G_m) \subseteq \text{supp}(Q_{G_2})$.

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 ξ_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 ξ_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 $Q_{i,j}$ has become non-zero and later in the sequence there exists a type 1 rotation making $Q_{i,j}$ zero again, we already have zero in position (i, j) and that type 1 rotation should be ignored as well.

□

Similar to the notion of distribution set, for a support matrix ξ we define

$$\Theta(\xi) \coloneqq \{\Theta : \Theta = \tilde{Q}\tilde{Q}^\top, \ \textit{for any } \tilde{Q} \textit{ s.t. } \text{supp}(\tilde{Q}) \subseteq \text{supp}(\xi)\}.$$

Note that unlike Q, the matrix $\tilde{Q}$ is allowed to have zeros on its diagonal.

Definition 1. *A support rotation mapping ξ to ξ' is lossless if $\Theta(\xi) = \Theta(\xi')$.*

Similar to the test for distribution equivalence, losslessness can be evaluated by checking if there exists a sequence of support rotations that maps ξ' back to a subset of ξ. Clearly, reduction, reversible acute rotation, and column swap are lossless, as they are reversible. In most of the cases, irreversible acute rotations are lossy and lead to expansion of $\Theta(\xi)$, as it introduces capacity for having extra free variables. However, this is not necessarily the case.

We have the following observations regarding checking for distribution equivalence.

Lemma 2. *All the support rotations for checking the distribution equivalence of two directed graphs should be lossless.*

We need the following lemma for the proof.

Lemma 3. *If support matrix ξ is mapped to ξ' via a support rotation, then $\Theta(\xi) \subseteq \Theta(\xi')$.*

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

□

Proof of Lemma 2. If support matrix ξ is mapped to ξ' via a lossy support rotation, i.e., $\Theta(\xi) = \Theta(\xi')$ then by Lemma 3, we have $\Theta(\xi) \subsetneq \Theta(\xi')$. Suppose we want to check the equivalence of directed graphs G_1 and G_2 with support matrices ξ_1 and ξ_2, respectively. We note that $\Theta(G_1) = \Theta(\xi_1)$. Suppose ξ_1 is mapped to ξ through a sequence of support rotations, including a lossy rotation, which in turn is mapped to $\xi' \subseteq \xi_2$. Therefore,

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

Therefore,

$$\Theta(G_1) = \Theta(G_2).$$

□

Using Lemma 2, we can prove Theorem 1:

Proof. The if side is clear by Theorem 4. For the only if side, by Theorem 4 and Lemma 2 we show that if ξ_1 can be mapped to ξ_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 ξ_1 to a subset of ξ_2.

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

$$\mathit{diff}_j = \{l : l \in \mathit{diff}, \xi_{l,j} = 0\},$$

$$\mathit{diff}_k = \{l : l \in \mathit{diff}, \xi_{l,k} = 0\},$$

and

$$M = \begin{cases} \max\{m_j, m_k\}, & m_j = m_k, \\ m_j + 1, & \mathit{otherwise}. \end{cases}$$

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

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

Proof of Claim 1. Since $A(i,j,k)$ is lossless, we can map ξ' to a subset of ξ. Therefore, we should be able to introduce zeros in ξ' in indices diff_j of column j and indices diff_k of column k, without removing the existing zeros, except potentially ξ'_{ij}. We first use a reversible acute rotation on columns j and k to move the newly introduce zero in ξ'_{ij} to the first index in diff_j, and we denote the resulting support matrix by $\xi^{(1)}$. 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 ξ to ξ'.

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 $\xi^{(1)}_{\cdot,t_1}$ equal to $\xi^{(1)}_{\cdot,k}$, i.e., $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,k}) = 0$, 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 ξ as well, i.e., $\xi_{\cdot,t_1} = \xi^{(1)}_{\cdot,t_1}$. Now, a reversible acute rotation can be performed on columns t_1 and k to set $d_H(\xi_{\cdot,j}, \xi_{\cdot,j}) = 0$, and then a reduction can be performed to introduce another zero in column j of ξ. 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 diff_j, we need to perform a reduction, for which, we need another column equal to $\xi^{(1)}_{\cdot,j}$. This can be obtained by one of the following cases:
  - There already exists a column t_1, such that $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,j}) = 0$. Similar to Case 1, This implies that column t_1 also exists in ξ. Therefore, ξ has the desired property.
  - There exists a column t_1, such that $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,j}) = 0$, but $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,k}) = 1$. Similar to Case 1, This implies that column t_1 also exists in ξ. Therefore, a reversible acute rotation can transform ξ to a support matrix with the desired property.

  - There exists a column t_1, such that $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,k}) = 0$. Similar to Case 1, This implies that column t_1 also exists in ξ. Therefore, two reductions, one on columns (t_1, k), and then one on columns (t_1, j) can transform ξ 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., $d_H(\xi^{(1)}_{\cdot,t_1}, \xi^{(1)}_{\cdot,k}) = 0$. Similar to Case 1, This implies that column t_1 also exists in ξ. Therefore, ξ 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 ξ.

□

By Claim 1, ξ can be transformed via reduction and reversible acute rotation to a support matrix with the stated property. Therefore, we assume ξ has the property. Therefore, we have columns $\{t_1, \cdots, t_{M-1}, j, k\}$ with any number of zeros $0 \leq i \leq M$ on rows $\mathit{diff} \cup \{i\}$, 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 ξ', can be obtained via columns $\{t_1, \cdots, t_{M-1}, j, k\}$ of ξ, and hence, the irreversible acute rotation could have been replaced by other types of rotations.

□

E Proof of Proposition 4

To show that the property holds for cycle $C = (X_1, \cdots, X_m, X_1)$, we note that our desired support matrix is ξ_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 $\xi_{i,j} = \xi_{i,j+1}$ (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 ξ_2 leads to a subset of ξ_1.

F Proof of Proposition 5

If side:
If columns of ξ_2 are permutation of columns of ξ_1, then ξ_1 can be mapped to ξ_2 and vice versa via a sequence of column swap rotations. Therefore, by Theorem 1, $G_1 \equiv G_2$.

Only if side:
If $G_1 \equiv G_2$, the by Theorem 1, ξ_1 can be mapped to a subset of ξ_2 and ξ_2 can be mapped to a subset of ξ_1, both via only reductions, reversible acute rotations and column swaps. If each pair of column of ξ_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 swaps. Therefore, columns of ξ_2 are permutation of columns of ξ_1.

G Proof of Proposition 6

Only if side:
By definition, directed graph G is reducible if there exists directed graph G' such that $G \equiv G'$ and $\xi' \subset \xi$. By Theorem 1, ξ can be mapped to a subset of ξ' via a sequence of support rotations comprised of reductions, reversible acute rotations and column swaps. 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 swaps 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 ξ is the support matrix corresponding to a directed graph. 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 ξ, if there exists a sequence of reversible support rotations that enables us to apply a reduction to ξ, 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, $\xi_{j,k} = \xi_{j,j} = \times$, and $\xi_{k,j} = \xi_{k,k} = \times$. 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 $\xi_{t_1,k} = 0$, and $\xi_{k,t_1} = \times$, 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, $\xi_{t_1,k}$ will become non-zero, we should have $\xi_{t_1,j} = \times$. This will lead to a 2-cycle unless if $\xi_{j,t_1} = 0$. Now, if $\xi_{j,t_1} = 0$, because all the entries of columns t_1 and k where the same, we also have $\xi_{j,k} = 0$. This gives us two options for $\xi_{k,j}$:

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

- If $\xi_{k,j} = \times$, then in order for columns k and j to have the same number of non-zero entries, there should exist index l such that $\xi_{l,k} = \times$, and $\xi_{l,j} = 0$. Now, we need another column t_2 so that we perform a reversible acute rotation on columns t_2 and k to move $\xi_{j,k} = 0$ to entry $\xi_{l,k}$. This means that columns t_2 and k should be the same on all the entries, except that $\xi_{j,t_2} = \times$, but $\xi_{j,k} = 0$. Therefore, $\xi_{k,t_2} = \xi_{k,k} = \times$ and $\xi_{t_2,k} = \xi_{t_2,t_2} = \times$, 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 ξ_1 and ξ_2 are equivalent if and only if there exist sequences of reversible acute rotations and column swaps that map their support matrices to one another.*

Proof. By Proposition 6, there exists no 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 swaps, and we need to map one support matrix to the other, rather than mapping it to a subset of the other.

□

Proof of Corollary 1. DAGs do not have 2-cycles. Therefore, by Proposition 7, DAGs are irreducible. Therefore, the result follows from Corollary 3.

□

J Proof of Theorem 2

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 swap 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 swap on columns j and k such that the reversible acute rotation makes the diagonal entry $\xi_{j,j}$ zero and then the column swap 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 $\xi_{j,k}$ 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 swaps 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 swaps, 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 $j \to k$, and the final support matrix corresponds to a graph containing the edge $k \to j$. 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 swaps 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 swap 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 ξ_1 is a support matrix of a directed graphs, it does not have any zeros on the main diagonal. Given the sequence of support rotations, the column swaps 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 swaps, 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 $\xi_{i,i}$ using columns i and j, then $\xi_{i,j}$ should have been non-zero. Therefore, we can instead replace it by zeroing $\xi_{i,j}$, 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 $\xi_{i,i}$ using columns i and j, then $\xi_{i,j}$ should have been non-zero. Therefore, again we can instead replace it by zeroing $\xi_{i,j}$, 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 ξ_1 obtain ξ_1', which does not have any zeros on the main diagonal, and via a sequence of column swaps can be mapped to a subset of ξ_2.

  Now, we perform the reverse of that sequence of column swaps on ξ_2, which gives us a superset of ξ_1' (call it ξ_2''), and hence, does not have any zeros on the main diagonal. Therefore, since ξ_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. ξ_2'' is a superset of ξ_1', and both ξ_2'' and ξ_1' are graphically representable. By Lemma 2, the corresponding directed graph of ξ_2'' is the same (if the directed graph corresponding to ξ_2'' is irreducible) or reducible to the directed graph corresponding to ξ_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 ξ_2'' to ξ_1'. Finally, reversing the reversible acute rotations of the sequence from ξ_1 to ξ_1', we obtain a subset of ξ_1, and the whole sequence from ξ_2 to a subset of ξ_1 is a representable sequence. Similarly, we can construct a representable sequence mapping ξ_1 to a subset of ξ_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 reversions. Therefore, the result follows from Theorem 2.

L Proof of Proposition 8

To violate faithfulness, there are finite number of sets of hard constraints that should be satisfied (since hard constraints are distributional constraints and hence limited). Let θ_i be the set of values satisfying the i-th set of constraints. By the definitions of hard constraints, θ_i is Lebesgue measure zero. Therefore, the set of distributions not g-faithful to G, which is the finite union is also Lebesgue measure zero.

M Proof of Proposition 9

Suppose G^* is the ground truth DG and it generates distribution Θ, and G_1 is a candidate DG which we want to decide whether it is the ground truth or not.

Suppose $G_1 \cong G^*$. Then there exists a set of distribution with non-zero Lebesgue measure that both G_1 and G^* can generate. Suppose Θ is a distribution coming from this intersection which also satisfies Assumption 1. Then clearly, since both DGs can generate Θ, there is no way to realize which one has been the ground truth, and hence, G_1 is non-identifiable from G^*.

For the opposite direction, suppose $G_1 \cong G^*$ then either there is no distribution that they can both generate, or the measure of such distributions is zero. In the first case, Θ is not generatable by G_1 and hence we can identify that G_1 is not the ground truth. In the second case, by Assumption 1, Θ cannot be from the intersection and hence again is not generatable by G_1 and hence we can identify that G_1 is not the ground truth.

N Proof of Theorem 3

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

Let $(\hat{B}_{\ell_0}, \hat{\Omega}_{\ell_0})$ be the output of ℓ_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 Θ 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, $\hat{G}_{\ell_0} \cong G^*$.

O Algorithm for Enumerating Members of a Distribution Equivalence Class and Determining the Equivalence of Two Structures

We first propose an algorithm for enumerating members of the distribution equivalence class of a directed graph with support matrix ξ, based on a depth-first traversal. The algorithm is based on a search tree that is rooted at ξ and branches out via Reduction and AcuteRotation operations. These two operations are defined in Algorithm 1. Since those two rotation operations are independent of column swaps, we perform a similar depth-first traversal of column swaps at the end, leveraging the graphical, cycle reversion representation for efficiency.

Algorithm 1 Reduction and Acute Rotation Operations

```
function Reduction($\xi, i, j$)
    Initialize $\xi' \leftarrow \xi$
    $\xi'_{i,j} \leftarrow 0$
    return $\xi'$
end function

function AcuteRotation($\xi, i, j, k, \ell$)
    Initialize $\xi' \leftarrow \xi$
    $\xi'_{i,j} \leftarrow 0$
    $\xi'_{\ell,j} \leftarrow 1$
    $\xi'_{\ell,k} \leftarrow 1$
    return $\xi'$
end function
```

Each vertex in the search tree corresponds to a support matrix and each of its children corresponds to the outputs of an admissible Reduction and AcuteRotation operation. Algorithm 2 represents the pseudo-code of the function which compiles a set of those operations for a given support matrix.

Algorithm 2 Finding Legal Rotations

```
function FindRotations($\xi$)
    Initialize $Rotations = \emptyset$
    // Find Legal Reductions
    for $j, k$ such that $\|\xi_{\cdot,j} - \xi_{\cdot,k}\|_1 = 0$ do
        for $i$ such that $\xi_{i,j} = 1$ do
            if $i \neq j$ then
                $Rotations \leftarrow Rotations \cup \{$Reduction$(\xi, i, j)\}$
            end if
            if $i \neq k$ then
                $Rotations \leftarrow Rotations \cup \{$Reduction$(\xi, i, k)\}$
            end if
        end for
    end for
    // Find Legal Acute Rotations
    for $j, k$ such that $\|\xi_{\cdot,j} - \xi_{\cdot,k}\|_1 = 1$ do
        $\ell \leftarrow$ index such that $\xi_{\ell,j} \neq \xi_{\ell,k}$
        for $i \neq \ell$ such that $\xi_{i,j} = 1$ do
            if $i \neq j$ then
                $Rotations \leftarrow Rotations \cup \{$AcuteRotation$(\xi, i, j, k, \ell)\}$
            end if
            if $i \neq k$ then
                $Rotations \leftarrow Rotations \cup \{$AcuteRotation$(\xi, i, k, j, \ell)\}$
            end if
        end for
    end for
    return $Rotations$
end function
```

Algorithm 3 enumerates the equivalence class. The algorithm keeps track of the search tree state using a stack S which contain sets of rotated support matrices. The first step of the algorithm enumerates a subset of the equivalence class of ξ^* by finding sequences of Reduction and AcuteRotation operations. The second step enumerates column swaps in a similar depth-first fashion. It is made efficient by using the fact that sequences of legal column swaps correspond to sequences of cycle reversions.

Algorithm 3 Enumerating equivalent structures

```
function ReverseCycles($\xi$)
    $Reversed \leftarrow \emptyset$
    $\mathcal{C} \leftarrow$ list of cycles in $\xi$
    for $C$ in $\mathcal{C}$ do
        $\xi' \leftarrow$ Column-permuted $\xi$ with cycle $C$ reversed
        $Reversed \leftarrow Reversed \cup \{\xi'\}$
    end for
    return $Reversed$
end function

procedure EnumerateEquiv($p \times p$ support matrix $\xi^*$)
    Initialize $Equiv \leftarrow \{\xi^*\}$.
    Initialize empty stack $S$
    $S.push$(FindRotations($\xi^*$))
    while $S$ is not empty do
        $Rotations \leftarrow S.pop()$
        if $|Rotations| = 0$ then
            continue
        else
            $\xi \leftarrow$ a support matrix in the set $Rotations$
            $Rotations \leftarrow Rotations \setminus \{\xi\}$
            $S.push(Rotations)$
            if $\xi$ not in $Equiv$ then
                $Equiv \leftarrow Equiv \cup \{\xi\}$
                $S.push$(FindRotations($\xi$))
            end if
        end if
    end while
    // Enumerate legal column swaps via cycle reversion
    for $\tilde{\xi}$ in $Equiv$ do
        Initialize empty stack $S$
        $S.push$(ReverseCycles($\tilde{\xi}$))
        while $S$ is not empty do
            $Reversals \leftarrow S.pop()$
            if $|Reversals| = 0$ then
                continue
            else
                $\xi \leftarrow$ a support matrix in the set $Reversals$
                $Reversals \leftarrow Reversals \setminus \{\xi\}$
                $S.push(Reversals)$
                if $\xi$ not in $Equiv$ then
                    $Equiv \leftarrow Equiv \cup \{\xi\}$
                    $S.push(ReverseCycles(\xi))$
                end if
            end if
        end while
    end for
end procedure
```

Finally, the procedure EnumerateEquiv in Algorithm 3 may be used to determine whether or not two DGs with respective support matrices ξ_1 and ξ_2 are equivalent by enumerating the equivalence class of ξ_1 and checking whether or not ξ_2 is in that equivalence class.

P Virtual Edge Search Operator

For acyclic DGs, under the Markov and faithfulness assumptions, a variable X_i is adjacent to a variable X_j if and only if X_i and X_j are dependent conditioned on any subset of the rest of the variables. This is not the case for cyclic DGs (Richardson, 1996). Two non-adjacent variables X_i and X_j are dependent conditioned

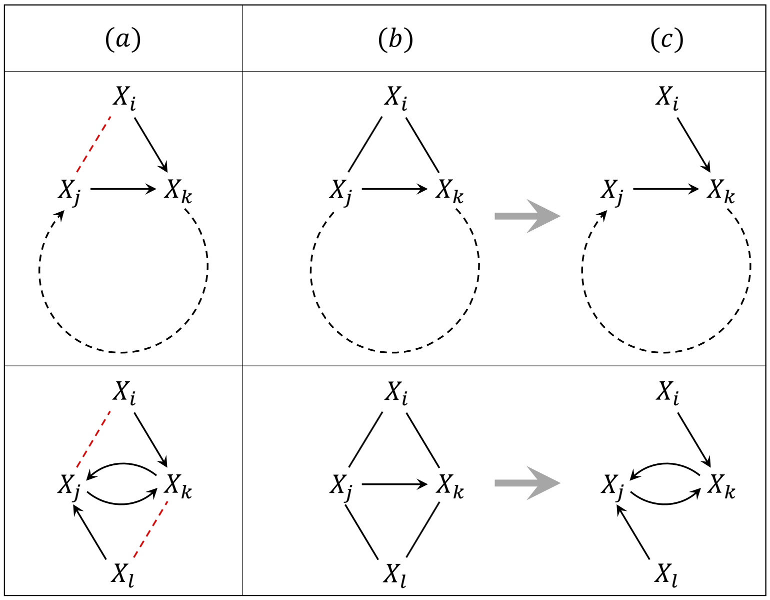


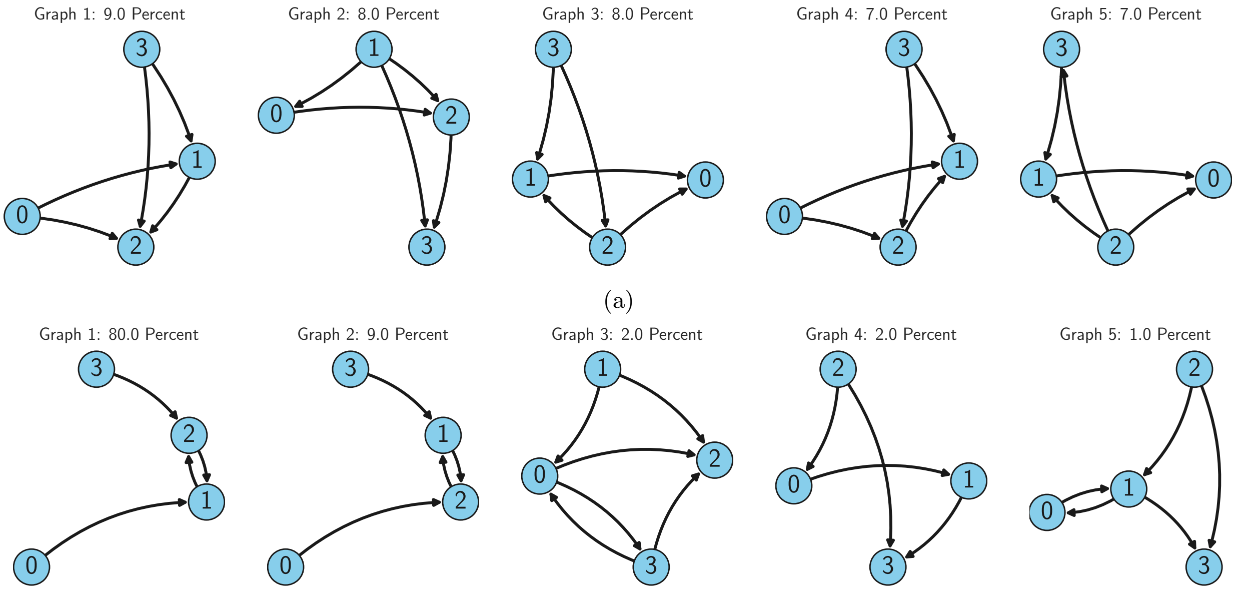
Figure 1: Virtual edge search operator.

on any subset of the rest of the variables if they have a common child X_k which is an ancestor of X_i or X_j. In this case, we say there exists a virtual edge between X_i and X_j. Figure 1(a) demonstrates two examples. In this figure, virtual edges are shown with dashed red edges.

There are two cases that detecting a virtual edge as a real edge can trap the greedy search into a local optima which can be improved.

Case 1. This case is shown in the first row of Figure 1. If a greedy search algorithm finds the edges between X_k and X_j but does not find X_k and X_j to be on a cycle, that is, if it does not find the directions correctly, it can significantly increase the likelihood by adding an edge at the location of the virtual edge between X_i and X_j. The algorithm would therefore be trapped in a local optimum shown in Figure 1(b) with one more edge than the ground truth shown in Figure 1(c). To resolve this issue, we propose adding the following 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.

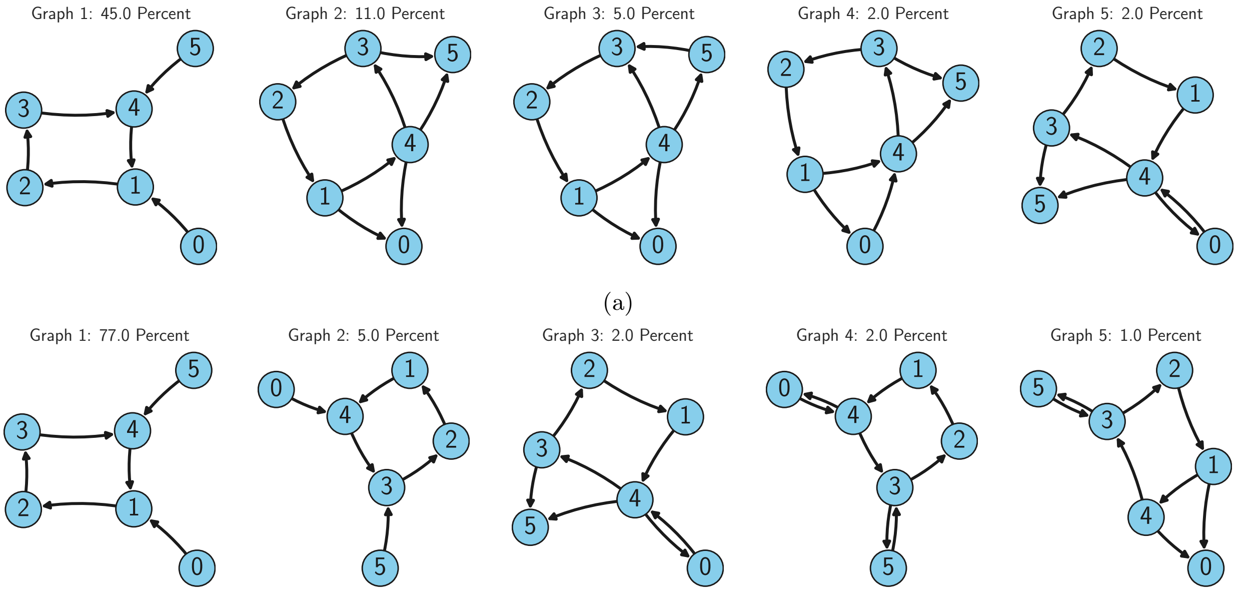
Case 2. This case is shown in the second row of Figure 1. This case involves the case that the cycle over X_j and X_k in the ground truth is a 2-cycle. If a greedy search algorithm finds one edges between X_k and X_j, it can significantly increase the likelihood by adding edges at the location of the virtual edges between X_i and X_j and between X_l and X_k. The algorithm would therefore be trapped in a local optimum shown in Figure 1(b) with one more edge than the ground truth shown in Figure 1(c). To resolve this issue, we propose adding the following search operator: Suppose we have triangles over three variables X_i, X_j and X_k and X_l, X_j and X_k, as shown in the figure. In one atomic move, we delete the edge connecting X_i to X_j and the edge connecting X_l to X_k, and add the edge $X_k \to X_j$. If the likelihood is unchanged, the edge deletion improves the score.



(a)

(b)

Figure 2: Example 1. Comparison of 5 most commonly learned structures.



(a)

(b)

Figure 3: Example 2. Comparison of 5 most commonly learned structures.

In order to evaluate the proposed search operator, we performed two experiments. The first involves the ground truth structure shown in Figure 2b, Graph 1. This graph has one equivalent structure, which is Graph 2 in the same figure. We run the tabu search algorithm with and without the proposed search operator for

100 instantiations of the edge weights and variances. The 5 most commonly found structures found by tabu search without and with the proposed operator are shown in Figures 2a and 2b, respectively. While the proposed algorithm finds an equivalent structure 89% of the time, the nominal tabu search never finds an equivalent structure.

Next, we consider the ground truth structure shown in Figure 3b, Graph 1. This structure has one equivalent, which is Graph 2 in the same figure. While the nominal tabu search algorithm finds an equivalent structure 45% of the time, the proposed algorithm is much more reliable, finding an equivalent structure 83% of the time.

Q 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 as follows.

$$P(V) = \prod_{X_i \in V} P(X_i | Pa(X_i)).$$

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. This can be shown as follows, where an MSCS is denoted by S.

$$P(V) = \prod_{S_i \subseteq V} P(S_i | Pa(S_i)).$$

After applying an operation, the likelihoods of all involved MSCSs are updated. Note that an operation can merge several MSCSs or break one into several smaller MSCSs. We perform the updates as follows:

- If the change adds an edge from MSCS S_1 to S_2, These two MSCSs and any MSCS on any path from S_2 to S_1 will fused into a new large MSCS.
- If the change is performed inside an MSCS, the score of the rest of MSCSs do not change.
- If the change removes or reverses an edge inside an MSCS, we find the MSCSs in that subset again, as it may be divided into smaller MSCSs.

R Effect of Sample Size on the Performance

In this section, we compare the performance of the discussed structure learning algorithms in the case of $p = 5$ variables and three different sample sizes: $n = 10^3, 10^4$, and 10^5. The results of the comparison are shown in Figure 4. As can be seen in the figure, the performance of the ℓ_0-regularized local search methods show marked improvement as sample size is increased.

For all experiments, including those in the main text, we use the following hyperparameters for the search algorithms. For the ℓ_1-regularized MLE, we use a regularization coefficient of 0.1, and threshold the learned

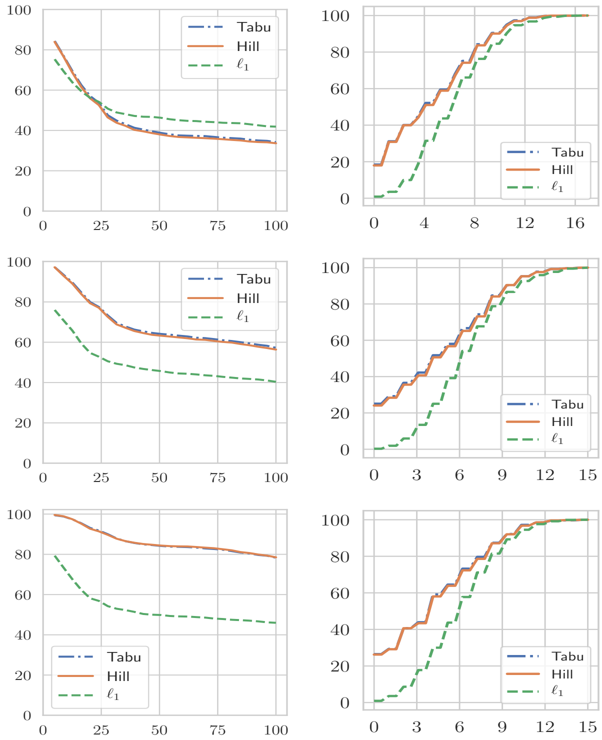


Figure 4: Results for $n = 10^3, 10^4, 10^5$, top to bottom. **Left column:** multi-domain evaluation. The percentage of outputs with success rate larger than a certain value is plotted vs. success percentages. **Right column:** SHD evaluation. The percentage of outputs with SHD less than or equal to a certain value is plotted vs. SHD.

B matrix at 0.05. See (Koller & Friedman, 2009) for details on greedy hill search and tabu search and its parameters. For tabu search, we use a tabu length of 5 for the $p = 5$ case and 10 for the $p = 20$ and $p = 50$ cases. In all cases, we used a tabu search patience of 5.

References

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

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., 1996.

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.

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