ORDERING-BASED CAUSAL STRUCTURE LEARNING
IN THE PRESENCE OF LATENT VARIABLES

DANIEL IRVING BERNSTEIN, BASIL SAEED, CHANDLER SQUIRES, AND CAROLINE UHLER

Abstract. We consider the task of learning a causal graph in the presence of latent confounders given i.i.d. samples from the model. While current algorithms for causal structure discovery in the presence of latent confounders are constraint-based, we here propose a score-based approach. We prove that under assumptions weaker than faithfulness, any sparsest independence map (IMAP) of the distribution belongs to the Markov equivalence class of the true model. This motivates the Sparsest Poset formulation - that posets can be mapped to minimal IMAPs of the true model such that the sparsest of these IMAPs is Markov equivalent to the true model. Motivated by this result, we propose a greedy algorithm over the space of posets for causal structure discovery in the presence of latent confounders and compare its performance to the current state-of-the-art algorithms FCI and FCI+ on synthetic data.

1. Introduction

Determining the causal structure between variables from observational data is a central task in many applications [6, 15, 8]. Causal structure is often modelled by a directed acyclic graph (DAG), where the nodes are associated with the variables of interest and the edges represent the direct causal effects these variables have on one another. In most realistic settings, only some of the variables in an environment are observed at any given time, i.e., only partial observations are available, leading to confounding effects on the observed variables. In such settings, a class of mixed graph models, called maximal ancestral graphs (MAGs) containing directed edges (representing direct causal effects), bidirected edges (representing the effect of a latent confounder on two variables) and undirected edges (representing selection bias), have been proposed to model the structure among the observed variables [14]. In this paper, we concentrate on latent confounders and are concerned with the recovery of mixed graphs containing directed and bidirected edges.

Current methods for estimating MAGs are constraint-based generalizing the prominent PC algorithm for estimating DAGs in the fully observed setting [18]. This includes the Fast Causal Inference (FCI) algorithm [18] and its variants: the Really Fast Causal Inference (RFCI) algorithm [5], and the FCI+ algorithm [4]. These methods depend on the faithfulness assumption to guarantee soundness and completeness, which has been shown to be restrictive [21]. In settings without latent confounders, studies have shown that score-based approaches, including the prominent GES algorithm [3], achieve superior performance to constraint-based approaches [10]. This motivates the development of score-based approaches for causal structure discovery in the presence of latent confounders.

We propose the greedy sparsest poset (GSPo) algorithm, a score-based approach for causal structure discovery in the presence of latent confounders. The key idea is to note that every MAG containing only directed and bidirected edges is consistent with a partial order of the...
observed variables (poset) and to recast the problem of causal structure discovery as the problem of learning a poset. In particular, our main contributions are as follows:

- We define a map that associates to each partial order of the observed variables a MAG, so that the sample-generating distribution is Markov to it.
- We prove that the sparsest such MAG is Markov equivalent to the true graph under conditions that are strictly weaker than faithfulness.
- We propose a greedy search over the space of posets based on the legitimate mark changes, described in [24], to move effectively between MAGs associated with different posets to find the poset yielding the sparsest graph.
- By comparing the performance and speed of our algorithm to FCI and FCI+ on synthetic data, we show that it is competitive to current state-of-the-art methods for causal structure discovery with latent confounders.

2. Preliminaries and related work

In the following, we review relevant concepts and related work; see also Appendix A.

2.1. Directed Maximal Ancestral Graphs. All graphs in this paper can have directed and bidirected edges. Let \( G = (V, D, B) \) be a graph with vertices \( V \), directed (\( \rightarrow \)) edges \( D \), and bidirected (\( \leftrightarrow \)) edges \( B \). We use \( \text{skel}(G) \) to denote the skeleton of \( G \), i.e., the undirected graph obtained by replacing all edges with undirected edges. We denote the number of edges of \( G \) by \( |G| := |D| + |B| \). We use \( \text{pa}_G(i), \text{sp}_G(i) \), and \( \text{an}_G(i) \) respectively to denote the parents, spouses, and ancestors of a node \( i \) in \( G \), where we use the typical definitions as in [9]. 

\( G \) is said to be ancestral if it has no directed cycles, and whenever there is a bidirected edge \( i \leftrightarrow j \) in \( G \), there is no directed path from \( i \) to \( j \) [14]. While ancestral graphs have been defined to also allow for undirected edges, we restrict our treatment to ancestral graphs with only directed and bidirected edges, which we will call directed ancestral graphs.

Standard notions of \( d \)-separation and \( d \)-connectedness for DAGs (see e.g. [9]) were generalized to \( m \)-separation and \( m \)-connectedness for ancestral graphs in [14]. We write \( A \perp \! \! \! \perp_G B \mid C \) to indicate that \( A \) and \( B \) are \( m \)-separated given \( C \) in \( G \). We denote the set of all \( m \)-separation relations of a graph \( G \) by \( \mathcal{I}(G) \). Unlike for DAGs, in the case of ancestral graphs it is possible to have a pair of non-adjacent vertices \( i \) and \( j \) that do not correspond to an \( m \)-separation relation of the form \( i \perp_G j \mid S \) for any \( S \subseteq V \setminus \{i,j\} \) (see [14]). An ancestral graph is maximal if every non-adjacent pair \( i \) and \( j \) satisfies \( i \perp_G j \mid S \) for some \( S \subseteq V \setminus \{i,j\} \). Associated to every graph \( G \) is a unique maximal supergraph, denoted \( G \), with the same set of \( m \)-separation statements [14]. An efficient procedure for computing \( G \) from \( G \) can be found in [14]. We refer to a directed ancestral graph that is maximal as a directed maximal ancestral graph (DMAG).

2.2. Markov Properties of DMAGs. Given a DMAG \( G = (V, D, B) \), we associate to each vertex \( i \in V \) a random variable \( X_i \) such that the random vector \( X_V = (X_i : i \in V) \) has joint distribution \( \mathbb{P} \). This joint distribution can be connected to the separation relations in \( G \) via the Markov property [13]; namely, the joint distribution \( \mathbb{P} \) is Markov with respect to the DMAG \( G \) if every \( m \)-separation relation in \( G \) implies the corresponding conditional independence (CI) relation in \( \mathbb{P} \), i.e.

\[
A \perp_{G} B \mid C \Rightarrow X_A \perp_{\mathbb{P}} X_B \mid X_C
\]
for all disjoint $A, B, C \subseteq V$, where $\cdot \perp \perp \cdot$ denotes conditional independence in $\mathbb{P}$. Denoting by $\mathcal{I}(\mathbb{P})$ the set of all CI relations in $\mathbb{P}$, the Markov property can be phrased as $\mathcal{I}(G) \subseteq \mathcal{I}(\mathbb{P})$. In this case, $G$ is called an independence map (IMAP) of $\mathbb{P}$; $G$ is called a minimal IMAP of $\mathbb{P}$ if deleting any edge produces a graph that is either not maximal, or not an IMAP of $\mathbb{P}$.

Graphs $G$ and $H$ are said to be Markov equivalent if $\mathcal{I}(G) = \mathcal{I}(H)$. The set of all graphs that are Markov equivalent to a given $G$ will be denoted $\mathcal{M}(G)$. A combinatorial condition for Markov equivalence of DMAGs was given in \cite{19}. The authors used the notion of discriminating paths: a path $\gamma = \langle i, \ldots, k, j \rangle$ between non-adjacent nodes $i$ and $j$ is discriminating for $k$ if every node between $i$ and $k$ is a collider and a parent of $j$, and there is at least one node between $i$ and $k$. The condition from \cite{19} says that $G$ and $H$ are Markov equivalent if and only if they have the same skeleton, the same v-structures, and if for any path $\gamma$ that is discriminating for $k$ in both $G$ and $H$, $k$ is a collider on $\gamma$ in $G$ if and only if $k$ is a collider on $\gamma$ for $H$.

The following transformational characterization of the Markov equivalence class of a DMAG given in \cite{24} will play an essential role in this paper. The authors say that a transformation of the edge $i \rightarrow j$ in $G$ into $i \leftrightarrow j$, or of the edge $i \leftrightarrow j$ into $i \rightarrow j$ is a legitimate mark change if there is no other directed path from $i$ to $j$ in $G$, $\text{pa}_G(i) \subseteq \text{pa}_G(j)$, $\text{sp}_G(i) \setminus \{j\} \subseteq \text{pa}_G(j) \cup \text{sp}_G(j)$, and there is no discriminating path for $i$ on which $j$ is the endpoint adjacent to $i$. They showed that $G$ and $H$ are Markov equivalent if and only if there is a sequence of legitimate mark changes from $G$ to $H$.

2.3. Causal Structure Discovery Algorithms. The problem of causal structure discovery in the setting of latent confounders is to recover the Markov equivalence class of the underlying DMAG $G^\pi$ from samples on the observed variables. In particular, when the sample size $n \to \infty$, the problem is to recover the Markov equivalence class of the DMAG $G^\pi$ from $\mathcal{I}(\mathbb{P})$. The most prominent existing algorithms for learning DMAGs\footnote{In fact, all of these methods are able to estimate MAGs, which may include undirected edges to model selection bias.} are the Fast Casual Inference (FCI) algorithm \cite{18} and its variants, most notably FCI+ \cite{14}, which has polynomial time complexity for sparse graphs while retaining large-sample consistency. All of these methods are constraint-based; they start by estimating the skeleton of the graph based on the results of CI tests, then use the results of those CI tests to determine some edge orientations. However, constraint-based methods require the faithfulness assumption \cite{23}, which is restrictive in practice, and faithfulness violations lead to the removal of too many edges \cite{21}.

While to the best of our knowledge no score-based approaches have been described so far that can deal with latent confounders and have theoretical guarantees, in the DAG setting (i.e., no latent confounders) it has been shown that score-based approaches may require weaker assumptions for consistency \cite{22, 12} and usually achieve superior performance for a given sample size \cite{10}. This motivates the development of score-based approaches for causal structure discovery in the presence of latent confounders.

A particular score-based approach that will play an important role in this paper is the Sparsest Permutation algorithm, introduced in \cite{12}, which associates to each permutation $\pi$ a DAG $G_{\pi}$, which is a minimal IMAP of the data-generating distribution. Since under restricted faithfulness assumptions the sparsest such $G_{\pi}$ is Markov equivalent to the true DAG $G^\pi$, this motivates a greedy search over the space of permutations to determine the sparsest
\[ X_i = X_{pa_G(i)} + \varepsilon_i, \ i \neq 6 \]
\[ X_6 = X_3 - X_4 + \varepsilon_6 \]

(A) Structural equation model

Figure 1. A structural equation model giving rise to a joint distribution \( \mathbb{P} \) that is restricted-faithful, but not faithful, to the graph \( G^* \).

In fact, in [17] the authors proved that starting in any minimal IMAP there exists a sequence of minimal IMAPs connecting it to the true DAG \( G^* \) by legitimate mark changes such that the number of edges is weakly decreasing. Hence the Greedy Sparsest Permutation (GSP) algorithm is consistent for causal structure discovery in the fully observed setting.

In the following section, we generalize the sparsest permutation algorithm to the setting with latent confounders by using posets instead of permutations. In particular, we show that under restricted faithfulness assumptions the DMAG associated with the Sparsest Poset is Markov equivalent to the true DMAG. This motivates the introduction of a greedy search over posets, which we term Greedy Sparsest Poset (GSPo) algorithm and introduce in Section 4.

Finally, in Section 5 we analyze its performance and compare it to the FCI algorithms on synthetic data.

3. Sparsest poset

This section contains our main results. We first introduce the restricted faithfulness notion required for our results and show that it is strictly weaker than the standard faithfulness assumption. Then we introduce a map from posets to DMAGs, which are minimal IMAPs of the data-generating distribution, and show that the sparsest DMAG in the image of this map is Markov equivalent to the true DMAG \( G^* \).

3.1. Restricted Faithfulness. An important assumption for constraint-based methods to recover \( G^* \) from \( \mathcal{I}(\mathbb{P}) \) is the faithfulness assumption, which asserts that \( \mathcal{I}(\mathbb{P}) = \mathcal{I}(G^*) \). In practice, this assumption is very sensitive to hypothesis testing errors for inferring CI relations from data and almost-violations are frequent [21]. This motivates studying restricted versions of the faithfulness assumption [11, 12]. In the following, we introduce a restricted faithfulness assumption for DMAGs, which we show is sufficient for learning DMAGs.

Definition 1. A distribution \( \mathbb{P} \) is restricted-faithful to a DMAG \( G = (V, D, B) \) if it is Markov to \( G \) and satisfies

1. Adjacency-faithfulness: If \((i, j) \in B \cup D\), then \( X_i \perp \perp X_j \) for any \( S \subseteq V \setminus \{i, j\} \);
2. Orientation-faithfulness: If \( i - k - j \) is contained in the skeleton of \( G \) and \( i \) is \( m \)-connected to \( j \) given some subset \( S \subseteq V \setminus \{i, j\} \), then \( X_i \perp \perp X_j \mid X_S \);
3. Discriminating-path-faithfulness: If \( \langle i, \ldots, k, j \rangle \) is a discriminating path in \( G \) and \( i \) is \( m \)-connected to \( j \) given some subset \( S \subseteq V \setminus \{i, j\} \), then \( X_i \perp \perp X_j \mid X_S \).

It is clear that faithfulness implies restricted-faithfulness. Moreover, restricted-faithfulness is a strictly weaker condition – there exist joint distributions \( \mathbb{P} \) that are restricted-faithful to a DMAG that are not faithful. For example, let \( \mathbb{P} \) be given by the structural equation model in Figure 1a, where each \( \varepsilon_i \sim \mathcal{N}(0, 1) \). Then \( \mathbb{P} \) is restricted-faithful, but not faithful to the
graph $G^*$ displayed in Figure [15]. To see that $\mathbb{P}$ is not faithful to $G^*$, note that $X_1 \perp_{\mathbb{P}} X_6$ even though 1 and 6 are not m-separated in $G^*$.

3.2. Sparsest Poset. In this section, we show that the Markov equivalence class of a DMAG $G^* = (V, D^*, B^*)$ can be determined from $\mathcal{T}(\mathbb{P})$ under the restricted faithfulness assumption by casting this problem as a discrete optimization problem over the space of partial orders of the set $V$. We do this by mapping the space of these partial orders to minimal IMAPs of $G^*$ and minimizing a cost that is a function of such an IMAP.

A partial order on a set $V$ is a relation $\leq$ on $V$ that is reflexive, transitive, and antisymmetric. Two elements $i, j \in V$ are said to be incomparable if neither $i \leq j$ nor $j \leq i$ holds. We denote this symbolically by $i \not\leq j$. A set $V$ equipped with a specified partial order $\leq$ is called a partially ordered set (poset), denoted $(V, \leq)$. In this case $V$ is called the ground set of the poset. Given a poset $\pi = (V, \leq)$ and $s_1, \ldots, s_k \in V$, define

$$\text{pre}_\pi(s_1, \ldots, s_k) := \{x \in V : x \leq s_i \text{ for some } 1 \leq i \leq k\}.$$ 

The empty poset is the poset $(V, \leq)$ such that all $i, j \in V$ are incomparable. We denote the set of all posets with a ground set $V$ by $\mathcal{P}(V)$.

Associated to each directed ancestral graph $G = (V, D, B)$ is a partial order $\leq_G$ on $V$, defined by

$$i \leq_G j \iff i \in \text{an}_G(j).$$

Note that the ancestral property implies that if $i \leftrightarrow_G j$, then $i \not\leq_G j$. We denote the poset $(V, \leq_G)$ by $\text{po}(G)$. The map $G \mapsto \text{po}(G)$ gives a bijection from the set of complete DMAGs, i.e., DMAGs whose skeleta are complete graphs, to $\mathcal{P}(V)$, the set of posets with ground set $V$. Since not all DMAGs are complete, the set of DMAGs on $V$ is strictly larger than $\mathcal{P}(V)$. This relationship between ancestral graphs and posets motivates describing the sparsest IMAP of a distribution $\mathbb{P}$ that is restricted-faithful to a DMAG $G^*$ in terms of posets by mapping every poset to an IMAP. To obtain the map, we need the following definition.

Definition 2. Given a joint distribution $\mathbb{P}$ on the random vector $X_V$ and a poset $\pi = (V, \leq_\pi)$. Define $\text{AG}(\pi, \mathbb{P})$ as the ancestral graph with directed edge set

$$\{i \to j : i \leq_\pi j, X_i \perp_{\mathbb{P}} X_j \mid X_{\text{pre}_\pi(i) \setminus \{i,j\}}\}$$

and bidirected edge set

$$\{i \leftrightarrow j : i \not\leq_\pi j, X_i \perp_{\mathbb{P}} X_j \mid X_{\text{pre}_\pi(i) \setminus \{i,j\}}\}.$$

When $\leq_\pi$ is a total order, i.e. a partial order where the relations $i \leq_\pi j$ or $j \leq_\pi i$ hold for all $i, j$, then $\text{AG}(\pi, \mathbb{P})$ defines a map from permutations to DAGs and is the one used in the GSP algorithm [12]. The authors showed in this case that $\text{AG}(\pi, \mathbb{P})$ is a minimal IMAP for $\mathbb{P}$ for all total orders $\leq_\pi$. Unfortunately, as shown in the following example, $\text{AG}(\pi, \mathbb{P})$ may not be an IMAP of $\mathbb{P}$ when $\leq_\pi$ is allowed to be an arbitrary partial order.

Example 1. Let $\mathbb{P}$ be a joint distribution that is restricted-faithful to the DMAG $G^*$ shown in Figure 2a. Let $\pi$ be the poset with ground set $\{1, 2, 3, 4\}$ and relations $2 \leq 3, 1 \leq 4,$ and $i \not\leq j$ otherwise. Then $\text{AG}(\pi, \mathbb{P})$, shown in Figure 2b, is not an IMAP of $\mathbb{P}$. To see this, note that $4 \perp_{\text{AG}(\pi, \mathbb{P})} 3 \mid \{2\}$, but $X_4 \perp_{\mathbb{P}} X_3 \mid \{X_2\}$ since $4 \leftrightarrow 2 \leftarrow 1 \to 3$ is a $\{2\}$-connecting path in $G^*$.
Figure 2. Graphs for Example 1. \( P \) is faithful to \( G^* \) but \( AG(\pi, P) \) is not an IMAP of \( P \).

However, we show in the following proposition, which is proven in Appendix B, that one can construct a minimal IMAP of \( P \) for any poset \( \pi \) by iterating the map \( AG(\cdot, \cdot) \) and taking maximal closure. Define

\[ G^P_\pi := AG(po(AG(\pi, P)), P). \]

where \( P \) and \( \pi \) are as in Definition 2. To simplify notation, we use \( G^P_\pi \) instead of \( G^P_{\pi} \) when \( P \) is clear from context.

**Proposition 1.** Let \( P \) be a joint distribution on \( V \) that is restricted-faithful to a DMAG. Then \( G^P_\pi \) is a minimal IMAP of \( P \) for any poset \( \pi \in P(V) \).

As we show in the following example, it is necessary to take the maximal closure in the definition of \( G^P_\pi \) since the resulting graph may otherwise not be maximal.

**Example 2.** Let \( P \) be a joint distribution faithful to the graph \( G^* \) displayed in Figure 3a. Let \( \pi \) be the poset with ground set \( V = \{1, 2, 3, 4, 5\} \) and ordering relations \( 1 \leq 2, 3 \leq 4, 5 \leq 4 \), and \( i \not\leq j \) otherwise. Then \( AG(po(AG(\pi, P)), P) \), displayed in Figure 3b, is not maximal. To see this, note that \( AG(po(AG(\pi, P)), P) \) lacks an edge between 2 and 4, while there is no set \( S \subseteq V \setminus \{2, 4\} \) that \( m \)-separates 2 and 4 in \( AG(\pi(AG(\pi, P)), P) \).

Having defined a map from posets to minimal IMAPs for DMAGs, we are almost ready to state our main result on the consistency of the sparsest poset. The following theorem establishes that under restricted-faithfulness all sparsest IMAPs of \( G^* \) are Markov equivalent to \( G^* \).

**Theorem 1.** Given a distribution \( P \) and a DMAG \( G^* \) that is an IMAP of \( P \), let

\[ G \in \arg \min \{|H| : H \text{ is an IMAP of } P\}. \]

(a) If \( P \) is adjacency-faithful to \( G^* \), then \( \text{skel}(G) = \text{skel}(G^*) \).

Figure 3. Graphs for Example 2. \( P \) is faithful to \( G^* \) but \( AG(po(AG(\pi, P), P) \) is not maximal.
(b) If $\mathbb{P}$ is restricted-faithful to $G^*$, then $G \in \mathcal{M}(G^*)$.

The proof of this theorem is given in Appendix C; it involves using the adjacency faithfulness condition to obtain $\text{skel}(G) \supseteq \text{skel}(G^*)$ for any IMAP $G$. Then we show that the IMAP condition on $G$, under restricted-faithfulness of $\mathbb{P}$, forces a graph with the same skeleton as $G^*$ to have matching unshielded colliders and matching discriminating paths when these discriminating paths are present in both of these graphs.

The following proposition establishes that $G^*$ is in the image of the map taking each poset $\pi$ to $G_\pi$. Thus, when restricting our search over IMAPs to the the image of this map, the optimum is still in our feasible set. Its proof is in Appendix B.

**Proposition 2.** Let $\mathbb{P}$ be Markov and adjacency-faithful to a DMAG $G^*$. If $\pi = \text{po}(G^*)$, then $G_\pi = G^*$.

We are now ready to state our main result.

**Theorem 2** (Sparsest Poset). Let $\mathbb{P}$ be a joint distribution on $X_V$ that is restricted faithful to a DMAG $G^*$. If $\tau \in \arg \min_{\pi \in \mathcal{P}(V)} |G_\pi|$, then $G_\tau$ is Markov equivalent to $G^*$.

**Proof.** Propositions 1 and 2 together imply that there is an IMAP $H = G_\pi$ for some $\pi$ such that $|H| = |G^*|$. Theorem 1 then gives the desired result. □

### 4. Greedy Sparsest Poset

Theorem 2 formulates the problem of finding a graph $G^*$ from $\mathbb{P}$ as a discrete optimization problem over $\mathcal{P}(V)$, the set of all posets on the ground set $V$. In this section, we discuss solving this optimization problem by imposing a graph structure on $\mathcal{P}(V)$ and then performing a greedy search along the edges of the graph. Note that Theorem 2 does not guarantee that a greedy approach returns an optimum. Supported by simulations, we will conjecture that this is indeed the case.

#### 4.1. Greedy Sparsest Poset

Perhaps the most natural graph structure on $\mathcal{P}(V)$ is known as the Hasse diagram of the poset of posets [1], which we denote by $\mathcal{H}_{\mathcal{P}(V)}$. One obtains this by adding an edge to connect posets $(V, \leq_1)$ and $(V, \leq_2)$ whenever there exists a unique pair $i, j \in V$ such that $i \leq_1 j$, but $i \not\leq_2 j$. Figure 4a shows $\mathcal{H}_{\mathcal{P}(V)}$ when $V = \{1, 2, 3\}$. For more details about Hasse diagrams, see [20].

Algorithm 1 is a greedy search along the edges of $\mathcal{H}_{\mathcal{P}(V)}$ to determine a poset $\pi$ yielding the sparsest $G_\pi$. Figure 4b shows an example run of Algorithm 1 when $\mathcal{I}(\mathbb{P}) = \{X_1 \perp \perp X_2 | X_3\}$, where each poset $\pi$ is replaced by its corresponding $G_\pi$, along with a possible path taken when starting at the empty poset.

As the example in Figure 4b shows, $G_\pi = G_\tau$ can happen for $\pi \neq \tau$. To achieve better run-time performance, one might try optimizing directly over the set $\{G_\pi : \pi \in \mathcal{P}(V)\}$ rather than $\mathcal{P}(V)$, so as to avoid moving between posets that give rise to the same graph, similar as in GSP [17]. We propose to do so by obtaining a graph $G'$ from $G_\pi$ via a legitimate mark change, then moving to $G_{\text{po}(G')}$.

**Definition 3** ([21]). Given a DMAG $G$, a legitimate mark change of $G$ is the process of turning an edge $i \rightarrow j$ to $i \leftrightarrow j$, or vice-versa, when
(a) $\mathcal{H}_{\mathcal{P}(V)}$ for $V = \{1, 2, 3\}$. Each of the large squares represents a poset $\pi \in \mathcal{P}(V)$. We represent each $\pi$ by having $i$ lie above $j$ only if $j \leq \pi i$. For example, the square in the upper left corner represents the poset with relations $1 \leq \pi 2 \leq \pi 3$ while the bottom-most square represents the empty poset. Posets $(V, \leq_1)$ and $(V, \leq_2)$ are connected by an edge whenever there exists a unique pair $i, j \in V$ such that $i \leq_1 j$, but $i \not\leq_2 j$.

(b) A relabeling of $\mathcal{H}_{\mathcal{P}(V)}$ be replacing each $\pi$ with its corresponding $G_\pi$ when $\mathcal{I}(\mathcal{P}) = \{X_1 \perp \perp X_2 \mid X_3\}$. The number of edges of each $G_\pi$ is indicated in the bottom right corner of the square containing it. The direction of edges indicates a strict decrease in the number of edges from one graph to the next. A possible path that algorithm 1 could take starting at the bottom square is highlighted in blue, with the graph returned colored green.

(1) there is no directed path from $i$ to $j$ aside from possibly $i \rightarrow j$;
(2) if $k \rightarrow i$, then $k \rightarrow j$. If $k \leftrightarrow i$, then $k \leftrightarrow j$ or $k \rightarrow j$;
(3) there is no discriminating path $\langle k, \ldots, i, j \rangle$.

It was shown in [24] that DMAGs $G$ and $H$ are Markov equivalent if and only if $G$ can be transformed into $H$ via a sequence of legitimate mark changes. An analogous result for DAGs appears in [2] where it is shown that DAGs $G$ and $H$ are Markov equivalent if and

**Algorithm 1**

**Input:** $\mathcal{I}(\mathcal{P})$, with $\mathcal{P}$ restricted-faithful to $G^*$; a starting poset $\pi_0$.

**Output:** A minimal IMAP of $G^*$.

Set $\pi = \pi_0$;

Using a depth-first search on $\mathcal{H}(\mathcal{P}(V))$ with root $\pi$, find a path $\pi_1 := \pi, \ldots, \pi_k := \tau$ such that $\pi_i$ is adjacent to $\pi_{i+1}$ in $\mathcal{H}_{\mathcal{P}(V)}$, $|G_{\pi_i}| \geq |G_{\pi_{i+1}}|$ and $|G_\pi| > |G_\tau|$.

If such $\pi_k$ exists, set $\pi$ to $\pi_k$, and repeat this step.

Otherwise, return $G_\pi$. 
Figure 5. Example of the outgoing edges (in bold) of the node $G_\pi$ in the structure $L_\mathbb{P}$. $\mathbb{P}$ is taken to be faithful to the DMAG $G^*$ from Figure 2b. The graphs $G'_1$ and $G'_2$ are obtained from $G_\pi$ via the legitimate mark changes of the colored dashed edges. The posets $\tau_1$ and $\tau_2$ correspond to $\text{po}(G'_1)$ and $\text{po}(G'_2)$ respectively. Hence the outgoing edges from $G_\pi$ to $G_{\tau_1}$ and $G_{\tau_2}$.

only if $G$ can be transformed into $H$ via a sequence of covered edge flips, which are exactly the moves used by GSP [17].

Given a joint distribution $\mathbb{P}$, we define $L_\mathbb{P}$ to be the directed graph with vertex set \{\(G_\pi : \pi \in \mathcal{P}(V)\)\} with an arc from $G_\pi$ to $G_\tau$ whenever there exists a graph $G'$, obtainable from $G_\pi$ via a single legitimate mark change, such that $\tau = \text{po}(G')$. Figure 5 shows the outgoing edges of a particular minimal IMAP $G_\pi$ in $L_\mathbb{P}$ when $\mathbb{P}$ is faithful to the $G^*$ of Figure 2b. As shown, there are two possible legitimate mark changes that can be performed on the edges of $G_\pi$, namely, the dashed edges. Changing the bidirected dashed edge, for example, would give the graph $G'_1$ with $\tau_1 = \text{po}(G'_1)$. Hence, there is an outgoing edge from $G_\pi$ to $G_{\tau_1}$ in $L_\mathbb{P}$.

Algorithm 2 is the resulting greedy search for the sparsest $G_\pi$ over the space $L_\mathbb{P}$. We call this algorithm the \textit{greedy sparsest poset algorithm} ($\text{GSP}_\mathbb{P}$). We conjecture, supported by simulations on the order of 100,000s of different examples (see Appendix D), that $\text{GSP}_\mathbb{P}$ is consistent under the restricted-faithfulness assumption (using a sufficiently large depth $d$ in the depth-first search), i.e., it yields a DMAG that is Markov equivalent to $G^*$ no matter the starting point. This conjecture generalizes the consistency result for GSP in the fully observed setting proven in [17].

**Conjecture 1.** Let $\mathbb{P}$ be a probability distribution that is restricted-faithful to a DMAG $G^*$. Let $\pi_0$ be any poset. Then there exists a directed path $\pi_0 \rightarrow \pi_1 \rightarrow \cdots \rightarrow \pi_K$ in $L_\mathbb{P}$ such that $G_{\pi_K}$ is sparsest (i.e., $G_{2\pi_K} = G^*$), and such that $\pi_k$ has weakly fewer edges than $\pi_{k-1}$ for all $k = 1, \ldots, K$.

4.2. Implementation. A crucial practical consideration for $\text{GSP}_\mathbb{P}$ is the choice of the starting poset $\pi_0$, since a sparser initial IMAP requires fewer CI tests to reach the true Markov equivalence class. Letting $\pi_0$ be the empty poset, i.e. the poset where all pairs are incomparable, is a natural choice. In this case $G_{\pi_0} = \{i \leftrightarrow j \mid X_i \not\perp\!\!\!\!\perp _\mathbb{P} X_j\}$. However, this can be
Algorithm 2 Greedy Sparsest Poset (GSPo)

**Input:** $\mathcal{I}(\mathbb{P})$, with $\mathbb{P}$ restricted-faithful to $G^*$; starting poset $\pi_0$; maximum depth $d$.

**Output:** A minimal IMAP of $\mathbb{P}$.

Set $\pi = \pi_0$;

Using a depth-first search with root $\pi$ and depth at most $d$, find a path $\pi_0, \ldots, \pi_k$ such that $G_{\pi_i}$ and $G_{\pi_{i+1}}$ are adjacent in $L_{\mathbb{P}}$, $|G_{\pi_i}| \geq |G_{\pi_{i+1}}|$ and $|G_{\pi_0}| > |G_{\pi_k}|$.

If such $\pi_k$ exists, set $\pi$ to $\pi_k$, and repeat this step.

Otherwise, return $G_{\pi}$.

dense even for sparse $G^*$. An effective alternative is to start at a sparse DAG that is a minimal IMAP (i.e., given by a permutation), either by running a DAG-learning algorithm such as GSP or by simply using the same starting heuristic as GSP based on the minimum-degree (MD) algorithm [17]. We compare these initialization schemes in Section 5.

5. Experimental results

In this section, we compare GSPo to FCI and FCI+ on the task of recovering a DMAG from samples of the observed nodes. In each simulation, we sample 100 DMAGs from a marginalized Erdős-Rényi model with $p$ nodes, $K$ latent variables and $s$ expected neighbors per node. More precisely, we sample Erdős-Rényi DAGs on $p + K$ nodes with $s$ expected neighbors per node, then form DMAGs by marginalizing over the first $K$ nodes. To each edge $i \to j$ in the DAG, we assign an edge weight $w_{ij}$ drawn uniformly at random from $[−1, −.25] \cup [.25, 1]$; for each non-edge, $w_{ij} = 0$. Finally, we generate $n$ samples from the structural equation model $X = W^\top X + \epsilon$ where $\epsilon \sim \mathcal{N}(0, I_{K+p})$ and we remove the first $K$ columns of the data matrix.

In each run of GSPo, we set the depth parameter to $d = 4$, and run the algorithm 5 times for each graph using different starting points. In the case of DAGs, a depth of 4 has been used to reflect the empirically-observed average size of the Markov equivalence class of a DAG [7, 17]. Although we are not aware of any studies on the average size of Markov equivalence classes for DMAGs, we find that $d = 4$ works well in practice, and there is little benefit using larger values of $d$.

In Figure 6, we chose $p = 10$, $K = 3$, and $s = 3$. The resulting graphs have approximately 4 expected neighbors per node, and have varying proportions of bidirected edges, from 0% bidirected to 75% bidirected, with roughly 30% bidirected on average.

Figure 6a shows the performance of GSPo with three different initialization schemes as compared to FCI and FCI+ on recovering the skeleton of the true MAG. We observe that regardless of the initialization scheme, GSPo generally estimates denser graphs than FCI and FCI+, with the densest graphs estimated when starting at the empty poset. The performance of initializing GSPo by the MD algorithm and GSP are comparable, so for simplicity we recommend initializing by the MD algorithm. While FCI and FCI+ achieve better performance in the low false positive rate regime, GSPo begins to surpass FCI and FCI+ in the middle regime. This indicates that even with a large number of samples, FCI+ suffers from near-faithfulness violations, which leads to mistakenly removing edges. ROC curves for $p = 50$ nodes are reported in Appendix E with similar findings.
In (a) and (b), $p = 10$, $K = 3$, and $s = 3$. In (a), each variant of GSPo was run on $8$ $\alpha$ values from $10^{-10}$ to $.7$, and each variant of FCI was run on $7$ $\alpha$ values from $10^{-20}$ to $.5$ The top-performing value of $\alpha$ for each algorithm was selected for (b), and the corresponding point is marked by $\star$ in (a). These values were $\alpha = .1$ for each variant of GSPo and for FCI+, and $\alpha = .7$ for FCI. In (c), $p = 10, 20, 30, 40, 50, K = 3$, and $s = 3$. Again the top-performing $\alpha$ values were selected for (b), with the exception of FCI, which was run with $\alpha = 10^{-3}$ since higher $\alpha$ values were extremely slow.

Figure 6c shows the median computation time required for each algorithm for graphs of varying number of vertices. Average computation time is presented in Appendix E. For each algorithm, we chose the parameter $\alpha$ based on the best-performing value in Figure 6b. However, for FCI we were limited to $\alpha = 10^{-3}$ due to its poor scaling for dense graphs. Thus, the median runtime for FCI is a conservative lower bound. We observe that GSPo with GSP initialization is faster than FCI or FCI+ for small graphs, but takes longer than FCI+ as the number of nodes increases. Given that CI tests in the construction of $G_\pi$ involve all ancestors of two nodes, which in general may be close to the total number of nodes in the graph, this poor scaling is expected. Fortunately, this suggests that improvements along the lines of those in the FCI+ algorithm may bring the scaling of GSPo in line with that of FCI+.

6. DISCUSSION

In this work, we provided a new characterization of the Markov equivalence class of a DMAG in terms of the set of sparsest minimal IMAPs, which allows structure learning in the presence of latent confounders to be expressed as a discrete optimization problem and hence the development of score-based approaches for this problem. To restrict the search space for this problem, we introduced a map from posets to minimal IMAPs whose image
contains the true DMAG. Then, we proposed a greedy algorithm in the space of minimal IMAPs to determine the sparsest minimal IMAP and hence a graph that is Markov equivalent to the true DMAG. This algorithm extends the Greedy Sparsest Permutation algorithm \cite{Heckerman1995} for learning DAGs to the setting with latent confounders, thereby providing the first score-based approach for causal structure discovery in this setting. We also demonstrated that it outperforms the current constraint-based methods FCI and FCI+ in some relevant settings.

The question of consistency of our greedy algorithm remains open, and is an interesting issue for future work. Furthermore, it may be possible to improve both the statistical and computational performance of GSPo through a number of modifications. These could include a new way of obtaining minimal IMAPs after legitimate mark changes, using dynamic connectivity algorithms to keep track of ancestral relations, and better heuristics for initialization.

By introducing a method for structure learning for DMAGs that is not a variant of FCI, we open the door to comparisons between the behavior of different types of methods on issues besides just statistical and computational performance, such as behavior of the algorithms under misspecification of parametric or modeling assumptions (e.g., non-i.i.d. data or non-Gaussianity when using partial correlation tests). It would also be interesting to use the idea of an ordering-based search as provided in this paper for the problem of learning general MAGs (i.e., including selection bias). To the best of our knowledge, there is no known transformational characterization for Markov equivalence classes of general MAGs yet, which is a key ingredient in the development of such a greedy algorithm.

Acknowledgements. We thank Thomas Richardson and Peter Spirtes for helpful discussions. Daniel Irving Bernstein was funded by an NSF Mathematical Sciences Postdoctoral Research Fellowship (DMS-1802902). Chandler Squires was supported by an NSF Graduate Research Fellowship and an MIT Presidential Fellowship. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, a Sloan Fellowship and a Simons Investigator Award.

References

[1] S. Bouc. The poset of posets. arXiv preprint arXiv:1311.2219, 2013.
[2] D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pages 87–98. Morgan Kaufmann Publishers Inc., 1995.
[3] D. M. Chickering. Optimal structure identification with greedy search. Journal of Machine Learning Research, 3(Nov):507–554, 2002.
[4] T. Claassen, J. Mooij, and T. Heskes. Learning sparse causal models is not NP-hard. arXiv preprint arXiv:1309.6824, 2013.
[5] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, pages 294–321, 2012.
[6] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7(3-4):601–620, 2000.
[7] S. B. Gillispie and M. D. Perlman. Enumerating Markov equivalence classes of acyclic digraph models. In Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, pages 171–177. Morgan Kaufmann Publishers Inc., 2001.
[8] D. Heckerman, A. Mamdani, and M. P. Wellman. Real-world applications of Bayesian networks. Communications of the ACM, 38(3):24–26, 1995.
[9] S. L. Lauritzen. Graphical Models, volume 17. Clarendon Press, 1996.
[10] P. Nandy, A. Hauser, M. H. Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

[11] J. Ramsey, J. Zhang, and P. L. Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

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

[13] T. Richardson. Markov properties for acyclic directed mixed graphs. Technical report, Technical Report, 1999.

[14] T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[15] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[16] K. Sadeghi, S. Lauritzen, et al. Markov properties for mixed graphs. *Bernoulli*, 20(2):676–696, 2014.

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

[18] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

[19] P. Spirtes and T. Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

[20] R. P. Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in Advanced Mathematics*, 2011.

[21] C. Uhler, G. Raskutti, P. Buhlmann, B. Yu, et al. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

[22] S. Van de Geer, P. Buhlmann, et al. ℓ0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

[23] J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc., 2002.

[24] J. Zhang and P. Spirtes. A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1419*, 2012.
Appendix A. Graph Theory

This section provides additional graph-theoretic notations that are standard in the literature and are provided for ease of access. Let \( G = (V, D, B) \) be a mixed graph. If there is any edge between nodes \( i \) and \( j \), they are called adjacent and denoted by \( i \sim j \). Otherwise they are called non-adjacent and denoted by \( i \nsim j \). We use \( \circ \) as a “wildcard” for edge marks, i.e. \( i \circ j \) denotes that either \( i \to j \) or \( i \leftrightarrow j \). We use subscripts on these vertex relations as a shorthand way to indicate the presence or absence of an edge, or the presence of a particular kind of edge. For example, \( i \leftrightarrow_G j \) and \( k \nsim_G l \) respectively indicate that \( G \) has a bidirected edge between \( i \) and \( j \), and no edge between \( k \) and \( l \). A graph with only directed edges is called a directed graph.

A path \( \gamma = \langle v_1, v_2, \ldots, v_k \rangle \) is a sequence of distinct nodes such that \( v_i \) and \( v_{i+1} \) are adjacent. A cycle is a path together with any type of edge between \( v_k \) and \( v_{k+1} = v_1 \). A path or a cycle is called directed if all edges are directed toward later nodes, i.e. \( v_i \to v_{i+1} \).

We extend the notation \( \text{pa}_G(i), \text{sp}_G(i), \) and \( \text{an}_G(i) \) to allow arguments that are subsets of vertices by taking unions. For example, when \( S \subseteq V \), we have
\[
\text{pa}_G(S) := \cup_{i \in S} \text{pa}_G(i).
\]
We add an asterisk to denote that the arguments are not included in the set, e.g.
\[
\text{pa}^*_G(S) := \text{pa}_G(S) \setminus S.
\]
The colliders on a path \( \gamma \) are the nodes where two arrowheads meet, i.e., \( v_i \) is a collider if \( v_{i-1} \circ v_i \leftarrow v_{i+1} \). A triple of nodes \( (i, j, k) \) is called a \( v \)-structure if \( j \) is a collider on the path \( (i, j, k) \) and \( i \nsim k \).

Given graphs \( G \) and \( H \), then \( G \) is Markov with respect to \( H \), or equivalently, \( H \) is an IMAP of \( G \), if \( \mathcal{I}(H) \subseteq \mathcal{I}(G) \). This is symbolically denoted by \( G \leq H \). The reason that the inequality symbol goes in the given direction is because when \( H \) is an IMAP of \( G \) and \( H \) is maximal, the skeleton of \( G \) is a subgraph of the skeleton of \( H \).

Appendix B. Proof of Propositions 1 and 2

We will prove Propositions 1 and 2 via several intermediate Lemmas. The following lemma, which reduces the number of \( m \)-separation statements we must consider, will be used in both.

Lemma 1. Let \( G^* \) and \( H \) be DMAGs. Then \( G^* \leq H \) if and only if whenever \( i \nsim_H j \), \( i \) is \( \text{an}_H(\{i, j\}) \)-separated from \( j \) in \( G^* \), i.e. \( i \perp_{G^*} j | \text{an}_H(\{i, j\}) \).

Proof. This is an immediate consequence of Theorem 3 in [16]. \(\square\)

Throughout the remainder of this section, let it be understood that \( G^* \) is a DMAG that is restricted-faithful to some fixed joint distribution \( \mathbb{P} \). We will not repeat this assumption. Moreover, we will suppress \( \mathbb{P} \) in our notation and write \( \text{AG}(\pi) \) instead of \( \text{AG}(\pi, \mathbb{P}) \). Also, note that when \( H \) is a DMAG, \( \text{po}(H) = \text{po}(\overline{H}) \) since \( \overline{H} \) is obtained from \( H \) by adding only bidirected edges [14]. We will make repeated use of this fact.

Lemma 2. Let \( \pi \) be a partial order on the random variables of \( \mathbb{P} \) such that \( G_\pi = \overline{\text{AG}(\pi)} \). Then \( G_\pi \) is an IMAP of \( \mathbb{P} \).

Proof. Lemma 1 implies that it suffices to show that whenever \( i \nsim_{G_\pi} j \), then \( X_i \perp_{\mathbb{P}} X_j | X_{\text{an}_{G^*_\pi}(i, j)} \). So assume \( i \nsim_{G_\pi} j \). Since \( G_\pi = \overline{\text{AG}(\text{po}(\text{AG}(\pi)))} \), then \( i \nsim_{G_\pi} j \) implies \( X_i \perp_{\mathbb{P}} X_j \). \(\square\)
$X_j \mid X_{\text{pre}^*_{\text{po}(AG(\pi))}(i,j)}$. This completes the proof, since $\text{pre}^*_{\text{po}(AG(\pi))}(i,j) = \text{an}^*_{AG(\pi)}(i,j) = \text{an}^*_{AG(\pi)}(i,j)$ and by assumption $G_{\pi} = \overline{AG(\pi)}$. \qed

**Lemma 3.** Let $\pi$ be a partial order on the random variables of $\mathbb{P}$. Then $\text{po}(G_{\pi}) = \text{po}(AG(\pi))$.

**Proof.** We must show that

$$\text{po}(AG(\text{po}(AG(\pi)))) = \text{po}(AG(\pi)).$$

If $i \leq j$ in $\text{po}(AG(\text{po}(AG(\pi))))$, then there exists a directed path $i = i_0 \rightarrow \cdots \rightarrow i_k = j$ in $AG(\text{po}(AG(\pi)))$ and so $i = i_0 \leq \cdots \leq i_k = j$ in $AG(\pi)$.

We now proceed to show that if $i \leq j$ in $\text{po}(AG(\pi))$, then the same is true in $\text{po}(AG(\text{po}(AG(\pi))))$. We do this by showing that if $i \rightarrow_{AG(\pi)} j$, then $i \rightarrow_{AG(\text{po}(AG(\pi)))} j$. So for the sake of contradiction, assume that $i \rightarrow_{AG(\pi)} j$ but not $i \rightarrow_{AG(\text{po}(AG(\pi)))} j$. By the definition of $AG$, this implies that $i \not\rightarrow_{AG(\text{po}(AG(\pi)))} j$ and so $i$ is m-separated from $j$ given an*$_{AG(\pi)}(i,j)$ in $G^*$. But $i \rightarrow_{AG(\pi)} j$ implies that $i$ is m-connected to $j$ given pre*$_{\pi}(i,j)$ in $G^*$. Let $P$ be an m-connecting path from $i$ to $j$ given pre*$_{\pi}(i,j)$ in $G^*$. Since an*$_{AG(\pi)}(i,j) \subseteq$ pre*$_{\pi}(i,j)$, we can write

$$\text{pre}^*_{\pi}(i,j) = \text{an}^*_{AG(\pi)}(i,j) \cup S$$

for some nonempty set $S$, disjoint from an*$_{AG(\pi)}(i,j)$. Since $i$ is m-separated from $j$ given an*$_{AG(\pi)}(i,j)$ in $G^*$, $P$ must contain a collider with a descendent in $S$, but no descendant in an*$_{AG(\pi)}(i,j)$. Let $d$ be such a collider that is closest to $j$ along $P$ and let $s$ be a po($G^*$)-minimal descendant of $d$ from $S$.

We now construct a path $Q$ in $G^*$ that m-connects $j$ and $s$ given pre*$_{\pi}(j,s)$. Since $S \subseteq$ pre*$_{\pi}(j,s)$, this would imply existence of the edge $s \rightarrow_{AG(\pi)} j$, contradicting $Q$ is a subpath of $P$ from $j$ to $s$. Otherwise, we let $Q$ be obtained by concatenating the subpath of $P$ from $j$ to $d$, followed by a directed path from $d$ to $s$. Since $P$ is m-connecting given pre*$_{\pi}(i,j)$ and $i, s \leq j$ in $\pi$, it follows that when $Q$ is a subpath of $P$, $Q$ is m-connecting given pre*$_{\pi}(j,s)$. When $Q$ additionally has a directed path from $d$ to $s$, $Q$ is m-connecting given pre*$_{\pi}(j,s)$ since the non-$P$ segment has no colliders, and assumptions on $d$ and po($G^*$)-minimality of $s$ imply that no element of this segment is in an*$_{AG(\pi)}(s,j)$. \qed

**Proof of Proposition 4.** Define $\tau := \text{po}(AG(\pi))$. Since $\text{po}(H) = \overline{\text{po}(\overline{H})}$ for any DMAG $H$, we have

$$G_{\tau} = \overline{AG(\text{po}(G_{\pi}))}.$$

Lemma 3 implies that this is equal to $\overline{AG(\text{po}(AG(\pi))))}$, which is equal to both $G_{\pi}$ and $\overline{AG(\tau)}$. Thus we have shown that $G_{\pi} = G_{\tau} = \overline{AG(\tau)}$ and so Lemma 2 implies that $G_{\pi}$ is an IMAP of $\mathbb{P}$.

It remains to show that $G_{\pi}$ is a minimal IMAP of $\mathbb{P}$, i.e. that removing any edge results in a directed ancestral graph that is either not maximal, or not an IMAP of $\mathbb{P}$. Let $i, j$ be such that $i \sim_{G_{\pi}} j$ and let $G'$ be the graph obtained from $G_{\pi}$ by removing the edge between $i$ and $j$. If $G'$ is still maximal, then Lemma 1 implies that $i$ is m-separated from $j$ given an*$_{G'}(i,j)$ in $G'$. If $G' \leq G^*$, then $i$ is m-separated from $j$ given an*$_{G'}(i,j)$ in $G^*$. Note that an*$_{G'}(i,j) = \text{an}^*_{G_{\pi}}(i,j)$, and that Lemma 2 implies that an*$_{G_{\pi}}(i,j) = \text{pre}^*_{\text{po}(AG(\pi))}(i,j)$. But if $i$ were m-separated from $j$ in $G^*$, then $X_i \perp_{\mathbb{P}} X_j \mid X_{\text{pre}^*_{\text{po}(AG(\pi))}(i,j)}$ and so $i \not\rightarrow_{AG(\pi)} j$. This would imply that $AG(\pi)$ is a subgraph of $G'$. Since $G'$ is maximal, $G_{\pi}$ would be a subgraph as well, contradicting $i \sim_{G_{\pi}} j$. \qed
Proof of Proposition \[2\] We show that $AG(\pi) = G^*$. Since $\pi = po(G^*)$ and $G^*$ is maximal, this would give

$$G_\pi = AG(po(AG(\pi))) = AG(\pi) = G^* = G^*.$$ 

Since $\pi = po(G^*)$, it suffices to show that $AG(\pi)$ and $G^*$ have the same skeleton. If $i \not\sim_{AG(\pi)} j$ then $i \perp \perp j | an^*_G(i, j)$. Adjacency faithfulness then implies that $i \not\sim_{G^*} j$. If $i \sim_{AG(\pi)} j$ then $i \not\perp \perp j | an^*_G(i, j)$. In this case we must have $i \sim_{G^*} j$. Otherwise, Lemma \[1\] would imply $i \not\perp \perp j | an^*_G(i, j)$ contradicting that $G^*$ is an IMAP of $\mathbb{P}$.

\[\square\]

Appendix C. Proof of Theorem \[1\]

We begin by proving the following lemma, which extends classic results for the case of DAGs and deals with discriminating paths.

Lemma 4. Let $G^*$ and $H$ be DMAGs and let $\mathbb{P}$ be a distribution that is Markov to both $G^*$ and $H$. If $\mathbb{P}$ is adjacency-faithful to $G^*$, then

(a) $skel(G^*) \subseteq skel(H)$.

If $\mathbb{P}$ is furthermore orientation-faithful to $G^*$, then

(b) If $i \rightarrow{k} \leftarrow{j}$ is a v-structure in $G^*$, then either $i \rightarrow{k} \leftarrow{j}$ is a v-structure in $H$ or $i \not\sim_H j$.

(c) If $i \rightarrow{k} \leftarrow{j}$ is a v-structure in $H$, then either $i \rightarrow{k} \leftarrow{j}$ is a v-structure in $G^*$, or $i \not\sim_{G^*} k$ or $j \not\sim_{G^*} k$.

Finally, if $\mathbb{P}$ is also discriminating-path-faithful to $G^*$, then

(d) If $\gamma := \langle i, \ldots, k, j \rangle$ is a discriminating path in both $H$ and $G^*$, then $k$ is a collider in $\gamma$ in $H$ if and only if $k$ is a collider in $\gamma$ in $G^*$.

Proof. (a) If $i \not\sim_H j$, then by the pairwise Markov property \[14\], $X_i \perp \perp X_j | X_{an^*_H(i,j)}$, and by adjacency-faithfulness, $i \not\sim_{G^*} j$ in $G^*$.

(b) Let $i \not\sim_H j$, so $X_i \perp \perp X_j | X_{an^*_H(i,j)}$. Suppose $k$ is a parent of either $i$ or $j$. Since $k \in an^*_H(i,j)$, $i$ is m-connected to $j$ in $G^*$ given $an^*_H(i,j)$ by the path $i \rightarrow{k} \leftarrow{j}$, and thus $X_i \perp \perp X_j | X_{an_H(i,j)}$ by orientation faithfulness. Hence, $H$ is not an I-MAP of $\mathbb{P}$.

(c) Suppose $i \sim_H k$ and $j \sim_H k$. We have $X_i \perp \perp X_j | X_{an^*_H(i,j)}$, and thus by orientation faithfulness $i$ and $j$ are m-separated given $an^*_H(i,j)$ in $G^*$. Since $H$ is ancestral, $k \not\in an^*_H(i,j)$. Thus, to ensure the required m-separation in $G$, $k$ must be a collider in $G$ on the path $i - k - j$.

(d) Assume $\gamma = \langle i, C_1, \ldots, C_l, k, j \rangle$. If $k$ is a non-collider in $\gamma$ in $G^*$, then $i$ is m-connected to $j$ given $S$ for every $S$ containing $C_1, \ldots, C_l$ but not $k$. Discriminating-path-faithfulness implies $X_i \not\perp \perp X_j | X_S$ for every such $S$. Then $k$ must also be a non-collider in $\gamma$ in $H$, since otherwise there would exist some $S$ containing $C_1, \ldots, C_l$ but not $k$ such that $i$ is m-separated from $j$ given $S$ in $H^*$, contradicting $\mathcal{I}(H) \subseteq \mathcal{I}(\mathbb{P})$. If $K$ is a collider in $\gamma$ in $G^*$, then $i$ is m-connected to $j$ given $S$ for every $S$ containing $C_1, \ldots, C_l, k$. Again, discriminating-path-faithfulness implies $X_i \not\perp \perp X_j | X_S$ for every such $S$. Then $K$ must also be a collider in $\gamma$ in $H$, since otherwise there would exist some $S$ containing $C_1, \ldots, C_l, k$ such that $i$ is m-separated from $j$ given $S$ in $H^*$.

\[\square\]

We proceed to proving the theorem.
Proof of Theorem 1. (a) is implied by Lemma 4(a).

Since restricted faithfulness implies adjacency faithfulness, \( \text{ske}(G) = \text{ske}(G^*) \). It remains to show that \( G \) and \( G^* \) have the same v-structures, and that if \( \gamma \) is a discriminating path for \( k \) in both \( G \) and \( G^* \), then \( k \) is a collider on \( \gamma \) in \( G \) if and only if it is a collider on \( \gamma \) in \( G^* \).

Equality of the skeleta together with Lemma 4(b) and (c) imply that \( G \) and \( H \) have the same v-structures. If \( \gamma := (i, C_1, \ldots, C_l, k, j) \) is a discriminating path in both \( G^* \) and \( G \), then Lemma 4(d) implies that \( k \) is a collider in \( \gamma \) in \( G^* \) if and only if \( k \) is a collider in \( \gamma \) in \( G \), which completes the proof.

\[ \square \]

Appendix D. Conjecture Simulations

In Figure 7, we display a scatter plot of the number of edges of the graphs that we tested the oracle version of our algorithm on. The plot includes over 200,000 points, representing graphs with varying number of bidirected edges and total number of edges.

\textbf{Figure 7.} A scatter plot of the number of edges of the graphs that we tested the oracle version of our algorithm on. The plot includes over 200,000 points, representing graphs with varying number of bidirected edges and total number of edges.

Appendix E. Additional Simulations

In this section, we followed the same procedure for DMAG sampling as described in Section 5. In Figure 8, we use \( p = 50 \) nodes, \( K = 12 \) latent variables, and \( s = 3 \) expected neighbors per node in the DAG before marginalization. For 100 graphs, we find that this results in
(A) Average performance over 100 MAGs for each algorithm, when \( p = 50, K = 12, \) and \( 10, 20, 30, 40, 50, K = 3, \) and \( s = 3. \) Each variant of GSPo was run on 8 \( \alpha \) values from \( 10^{-10} \) to .7, and each variant of FCI+ were run with \( \alpha = \) values from \( 10^{-20} \) to .5 extremely long runtime of higher \( \alpha \) values.

(B) Average runtime over 100 MAGS for \( p = 10, 20, 30, 40, 50, K = 3, \) and \( s = 3. \) Each variant of GSPo and FCI+ were run with \( \alpha = .1, \) while FCI was run with \( \alpha = 10^{-3} \) due to the extremely long runtime of higher \( \alpha \) values.

Figure 8. Additional Simulation Results

MAGs with an average of 43% bidirected edges, ranging from 14% to 71% bidirected edges, and an average of 5 neighbors per node in the MAGs. Due to the slow runtime of FCI, GSPo with empty initialization, and FCI+ with high \( \alpha \) values, our comparison between the algorithms for larger graphs is limited, and mainly serves to demonstrate that GSPo has similar performance on larger graphs for the same range of \( \alpha \) values.

In Figure 8b, we use the same set of DMAGs as used in 6c, in particular, \( p = 10, 20, 30, 40, 50, K = 3, \) and \( s = 3, \) but report the average computation time instead of the median computation time. We can observe that GSPo with the empty initialization and FCI both have much higher average computation times than median computation times, indicating that they are more susceptible to outlier instances from our sampled MAGs.