Ordering-Based Causal Structure Learning in the Presence of Latent Variables

Daniel Irving Bernstein* MIT
Basil Saeed* MIT
Chandler Squires* MIT
Caroline Uhler MIT

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 hybrid 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 of these variables is a central task in many applications (Friedman et al., 2000; Robins et al., 2000; Heckerman et al., 1995). 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 (Richardson and Spirtes, 2002). 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 (Spirtes et al., 2000). This includes the *Fast Causal Inference* (*FCI*) algorithm (Spirtes et al., 2000) and its variants: the *Really Fast Causal Inference* (*RFCI*) algorithm (Colombo et al., 2012), and the *FCI+* algorithm (Claassen et al., 2013). These methods depend on the faithfulness assumption to guarantee soundness and completeness, which has been shown to be restrictive (Uhler et al., 2013). In settings without latent confounders, studies have shown that score-based approaches, including the prominent GES algorithm (Chickering, 2002), achieve superior performance to constraint-based approaches (Nandy et al., 2018). In purely constraint-based approaches such as PC, mistakes made in early stages of the algorithm tend to propagate and lead to later mistakes. Score-based approaches (which are usually greedy) are often more resilient to error propagation, since early mistakes only affect the local structure of the search space but do not affect the scores of later graphs. This motivates the development of an algorithm for causal structure discovery in the presence of latent confounders that shares this resilience with score-based approaches.

In this paper, we propose the *sparsest poset (SPo) algorithm* for causal structure discovery in the presence of latent confounders. Since this algorithm uses both a scoring criterion and conditional independence testing to learn the model, we refer to it as a *hybrid* method. The key idea that we use is that every MAG containing

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

only directed and bidirected edges is consistent with a *partial order* of the observed variables (*poset*) and hence the problem of causal structure discovery can be recast 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 by Zhang and Spirtes (2012) 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 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 $\mathrm{pa}_G(i)$, $\mathrm{sp}_G(i)$, and $\mathrm{an}_G(i)$ respectively to denote the parents, spouses, and ancestors of a node i in G, where we use the typical definitions as in Lauritzen (1996). 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 (Richardson and Spirtes, 2002). 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*.

Richardson and Spirtes (2002) generalized the standard notions of d-separation and d-connectedness for DAGs (see e.g. (Lauritzen, 1996)) to m-separation and m-connectedness for ancestral graphs. 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 without an m-separation relation of the form $i \perp\!\!\!\perp_G j \mid S$ for any $S \subseteq V \setminus \{i, j\}$ (see Richardson and Spirtes (2002)). An ancestral graph is *maximal* if every non-adjacent pair i and j satisfies $i \perp\!\!\!\perp_G j \mid S$ for some $S \subseteq V \setminus \{i, j\}$. Richardson and Spirtes (2002) showed that associated to every graph G is a unique maximal supergraph, denoted $\overline{G}$, with the same set of m-separation statements. They also give an efficient procedure for computing $\overline{G}$ from G. 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 distribution can be connected to the separation relations in G via the *Markov property* (Richardson, 1999); namely, $\mathbb{P}$ is *Markov* with respect to the DMAG G if every $m-$separation relation in G implies the corresponding conditional independence relation in $\mathbb{P}$, i.e.

$$A \perp\!\!\!\perp_G B \mid C \Rightarrow X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$$

for all disjoint $A, B, C \subseteq V$, where $\perp\!\!\!\perp_{\mathbb{P}}$ denotes independence in $\mathbb{P}$. Denoting by $\mathcal{I}(\mathbb{P})$ the set of all CI relations in $\mathbb{P}$, the Markov property is then equivalent to $\mathcal{I}(G) \subseteq \mathcal{I}(\mathbb{P})$. In this case, G is called an *independence map* (*IMAP*) of $\mathbb{P}$; G is called a *minimal IMAP* of $\mathbb{P}$ if there is no edge of G that can be deleted while keeping G both maximal and 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 $\mathcal{M}(G)$. Spirtes and Richardson (1996) provided a combinatorial characterization of graphs in the same Markov equivalence class (MEC). To do this, they used the notion of *discriminating paths for a vertex k*: a path $\gamma = \langle i, \ldots, k, j \rangle$ between non-adjacent i and j is *discriminating for k* if every node between i and k is both a collider and a parent of j, and there is at least one node between i and k. Spirtes and Richardson (1996) show that G and H are Markov equivalent if and only if they have the same skeleta, the same v-structures, and if for any path γ that is discriminating for k in both G and H, k is a collider on γ in G if and only if k is a collider on γ for H.

Zhang and Spirtes (2012) provided a transformational characterization for the Markov equivalence class of a DMAG that will play an essential role in this paper. For this, they called the transformation of the edge

$i \to j$ in G into $i \leftrightarrow j$, or of the edge $i \leftrightarrow j$ to $i \to j$ a *legitimate mark change* if there is no other directed path from i to j in G, $\mathrm{pa}_G(i) \subseteq \mathrm{pa}_G(j)$, $\mathrm{sp}_G(i) \setminus \{j\} \subseteq \mathrm{pa}_G(j) \cup \mathrm{sp}_G(j)$, 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^* 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^* from $\mathcal{I}(\mathbb{P})$. The most prominent existing algorithms for learning DMAGs[1] are the Fast Casual Inference (FCI) algorithm (Spirtes et al., 2000) and its variants, most notably FCI+ (Claassen et al., 2013), 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 (Zhang and Spirtes, 2002), which is restrictive in practice, and faithfulness violations lead to the removal of too many edges Uhler et al. (2013).

In the DAG setting (i.e., no latent confounders) it has been shown that score-based approaches may require weaker assumptions for consistency (Van de Geer et al., 2013; Raskutti and Uhler, 2018) and usually achieve superior performance for a given sample size (Nandy et al., 2018). This motivates the development of an algorithm for causal structure discovery that shares these properties with score-based approaches, and works in the presence of latent confounders. Existing score-based approaches that can handle latent confounders require parametric assumptions. For example, Shpitser et al. (2012) requires discreteness and Tsirlis et al. (2018); Nowzohour et al. (2017) requires Gaussianity.

A particular approach that will play an important role in this paper is the *Sparsest Permutation* algorithm, introduced in Raskutti and Uhler (2018), which associates to each permutation π a DAG G_π, which is a minimal IMAP of the data-generating distribution. The Sparsest Permutation algorithm is a hybrid method, combining aspects of the constraint- and score-based paradigms. Like many constraint-based methods, it does not require parametric assumptions, and like many score-based methods, it seems resilient to error-propagation. Since under restricted faithfulness assumptions the sparsest such G_π is Markov equivalent to the true DAG G^*, this motivates a greedy search over the space of permutations to determine the sparsest G_π. In fact, in Solus et al. (2017) 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

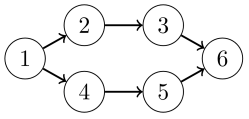
An important assumption for constraint-based methods to recover G^* from $\mathcal{I}(\mathbb{P})$ is the *faithfulness assumption*, which asserts that $\mathcal{I}(\mathbb{P}) = \mathcal{I}(G^*)$. In practice, this assumption is very sensitive to hypothesis testing errors for inferring CI relations from data and almost-violations are frequent (Uhler et al., 2013). This motivates studying restricted versions of the faithfulness assumption Ramsey et al. (2012); Raskutti and Uhler (2018). 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 satisfying*

[1] In fact, all of these methods are able to estimate *MAGs*, which may include undirected edges to model selection bias.

$$X_i = X_{\mathrm{pa}_{G^*}(i)} + \varepsilon_i,\ i \neq 6$$
$$X_6 = X_3 - X_5 + \varepsilon_6$$

(a) Structural equation model



(b) G^*

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^*.

1. Adjacency-faithfulness*: If $(i,j) \in B \cup D$, then $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_S$ 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 $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_S$.*
3. Discriminating-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 $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_S$.*

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 1b. To see that $\mathbb{P}$ is not faithful to G^*, note that $X_1 \perp\!\!\!\perp_{\mathbb{P}} X_6$ 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 $\mathcal{I}(\mathbb{P})$ under the restricted faithfulness assumption by casting this problem into an minimization 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 $i \not\leqslant j$. A set V equipped with a specified partial order $\leq$ is called a *partially ordered set* (*poset*), denoted $(V, \leq)$. Then, V is called the *ground set* of the poset. 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 $\mathcal{P}(V)$. Given a poset $\pi = (V, \leq)$ and $s_1, \ldots, s_k \in V$, define

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

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

$$i \leq_G j \Leftrightarrow i \in \mathrm{an}_G(j).$$

Note that the ancestral property implies that if $i \leftrightarrow_G j$, then $i \not\leqslant_G j$. We denote the poset $(V, \leq_G)$ by $\mathrm{po}(G)$. The map $G \mapsto (V, \leq_G)$ 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. This will lead to the concept of *sparsest posets*; the posets of $\mathcal{P}(V)$ that are mapped to DMAGs in $\mathcal{M}(G^*)$. 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 $AG(\pi, \mathbb{P})$ as the ancestral graph with directed edge set*

$$\{i \to j : i \leq_\pi j, X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{pre}_\pi(i,j)\setminus\{i,j\}}\}$$

and bidirected edge set

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

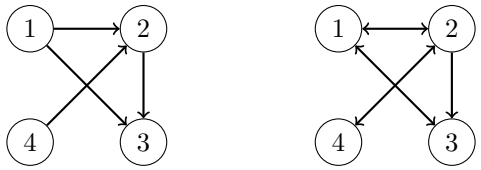
When $\leq_\pi$ is a *total order*, i.e. a partial order where the relations $i \leq_\pi j$ or $j \leq_\pi i$ hold for all i, j, then $AG(\pi, \mathbb{P})$ defines a map from permutations to DAGs and is the one used in the GSP algorithm (Raskutti and Uhler, 2018). The authors showed in this case that $AG(\pi, \mathbb{P})$ is a minimal IMAP for $\mathbb{P}$ for all total orders $\leq_\pi$. Unfortunately, as shown in the following example, $AG(\pi, \mathbb{P})$ 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 π be the poset with ground set $\{1,2,3,4\}$ and relations $2 \leq 3$, $1 \leq 4$, and $i \not\leqslant j$ otherwise. Then $AG(\pi, \mathbb{P})$, shown in Figure 2b, is not an IMAP of $\mathbb{P}$. To see this, note that $4 \perp\!\!\!\perp_{AG(\pi,\mathbb{P})} 3 \mid \{2\}$, but $X_4 \not\perp\!\!\!\perp_{\mathbb{P}} X_3 \mid \{X_2\}$ since $4 \leftrightarrow 2 \leftarrow 1 \to 3$ is a $\{2\}$-connecting path in G^*.*

However, we show in the following proposition, which is proven in Appendix B, that one can construct a minimal IMAP of $\mathbb{P}$ for any poset π using the map $AG(\cdot, \cdot)$ by defining

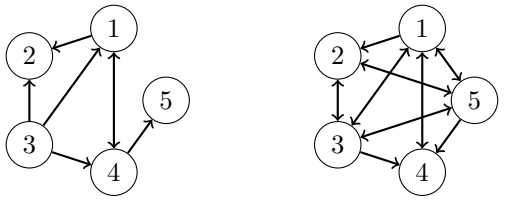
$$G_\pi^{\mathbb{P}} := \overline{AG(\mathrm{po}(AG(\pi, \mathbb{P})), \mathbb{P})}.$$

where $\mathbb{P}$ and π are as in Definition 2. Recall that $\overline{G}$ denotes the maximal closure of G. We want G_π



(a) G^* (b) $AG(\pi, \mathbb{P})$

Figure 2: Graphs for Example 1. $\mathbb{P}$ is faithful to G^* but $AG(\pi, \mathbb{P})$ is not an IMAP of $\mathbb{P}$.



(a) G^* (b) $AG(\text{po}(AG(\pi, \mathbb{P}), \mathbb{P})$

Figure 3: Graphs for Example 2. $\mathbb{P}$ is faithful to G^* but $AG(\text{po}(AG(\pi, \mathbb{P}), \mathbb{P})$ is not maximal.

to be maximal since the results of Zhang and Spirtes (2012) regarding legitimate mark changes apply only to maximal ancestral graphs. To simplify notation, we use G_π instead of $G_\pi^{\mathbb{P}}$ when $\mathbb{P}$ is clear from context.

Proposition 1. *Let $\mathbb{P}$ be a joint distribution on V that is restricted-faithful to a DMAG. Then G_π is a minimal IMAP of $\mathbb{P}$ for any poset $\pi \in \mathcal{P}(V)$.*

As we show in the following example, including the maximal closure in the definition of G_π is required since it may otherwise not be maximal.

Example 2. *Let $\mathbb{P}$ be a joint distribution faithful to the graph G^* displayed in Figure 3a. Let π 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 $AG(\text{po}(AG(\pi, \mathbb{P})), \mathbb{P})$, displayed in Figure 3b, is not maximal. To see this, note that $AG(\text{po}(AG(\pi, \mathbb{P})), \mathbb{P})$ 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 $AG(\pi(AG(\pi, \mathbb{P}), \mathbb{P})$.*

Having defined a map from posets to minimal IMAPs for DMAGs, we are almost ready to state our 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 $\mathbb{P}$ and a DMAG G^* that is an IMAP of $\mathbb{P}$, let*

$$G \in \arg \min_{\{H:\, H \text{ is an IMAP of } \mathbb{P}\}} |H|. \tag{1}$$

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

(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 $\pi \mapsto G_\pi$; its proof is given in Appendix D. Thus, when restricting our search over IMAPs to the the image of this map, the optimum is still in our feasible set.

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

We are now ready to state our main result.

Theorem 2 (Sparsest Poset)**.** *Let $\mathbb{P}$ be a distribution on V that is restricted faithful to a DMAG G^*. If*

$$\tau \in \arg \min_{\pi \in \mathcal{P}(V)} |G_\pi|,$$

then G_τ is Markov equivalent to G^*.

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

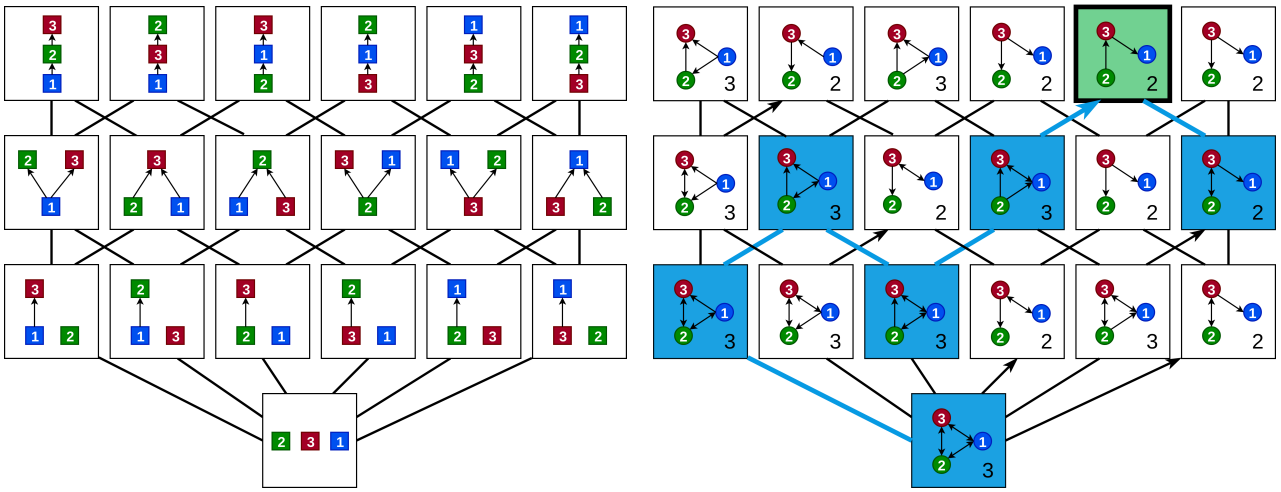
4 GREEDY SPARSEST POSET

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* (Bouc, 2013), 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 gives an example of $\mathcal{H}_{\mathcal{P}(V)}$ when $V = \{1, 2, 3\}$. For more details about Hasse diagrams, see Stanley (2011).

Algorithm 1 is a greedy search along the edges of $\mathcal{H}_{\mathcal{P}(V)}$ to determine a poset π yielding the sparsest



(a) $\mathcal{H}_{\mathcal{P}(V)}$ for $V = \{1, 2, 3\}$.

(b) A relabeling of $\mathcal{H}_{\mathcal{P}(V)}$ be replacing each π with its corresponding G_π when $\mathcal{I}(\mathbb{P}) = \{X_1 \perp\!\!\!\perp_{\mathbb{P}} X_2 \mid X_3\}$.

Figure 4: (a) shows $\mathcal{H}_{\mathcal{P}(V)}$ for $V = \{1, 2, 3\}$. Each of the large squares represents a poset $\pi \in P(V)$. We represent each π 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) shows a relabeling of $\mathcal{H}_{\mathcal{P}(V)}$ by replacing each π with G_π when $\mathcal{I}(\mathbb{P}) = \{X_1 \perp\!\!\!\perp_{\mathbb{P}} X_2 \mid X_3\}$. The number of edges of each G_π 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.

G_π. Figure 4b shows an example run of Algorithm 1 when $\mathcal{I}(\mathbb{P}) = \{X_1 \perp\!\!\!\perp_{\mathbb{P}} X_2 \mid X_3\}$, where each poset π is replaced by its corresponding G_π, 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 optimize directly over the set $\{G_\pi : \pi \in \mathcal{P}(V)\}$ rather than $\mathcal{P}(V)$, thus avoiding moving between posets that give rise to the same graph, similar as in GSP (Solus et al., 2017; Mohammadi et al., 2018). We propose to do this by moving from G_π to $G_{\text{po}\, G'}$ where G' is obtained from G_π via a legitimate mark change, the definition of which we now restate.

Algorithm 1

Input: $\mathcal{I}(\mathbb{P})$, with $\mathbb{P}$ restricted-faithful to G^*; a starting poset π_0.
Output: A minimal IMAP of G^*.
Set $\pi = \pi_0$;
Via depth-first search on $\mathcal{H}_{\mathcal{P}(V)}$ with root π, find a path $\pi_1 := \pi, \ldots, \pi_k := \tau$ such that π_i is adjacent to π_{i+1} in $\mathcal{H}_{\mathcal{P}}$, $|G_{\pi_i}| \geq |G_{\pi_{i+1}}|$ and $|G_\pi| > |G_\tau|$.
If such π_k exists, set π to π_k, and repeat this step.
Otherwise, return G_π.

Definition 3 (Zhang and Spirtes (2012))**.** *Given a DMAG G, a* legitimate mark change of G *is the process of turning an edge $i \to j$ to $i \leftrightarrow j$, or vice-versa, when*

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

Zhang and Spirtes (2012) showed 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. This is analogous to the result by Chickering (1995) that DAGs G and H are Markov equivalent if and only if G can be transformed into H via a sequence of covered edge flips, which are exactly the moves used by GSP (Solus et al., 2017; Mohammadi et al., 2018). Using this notion of edge change gives a different search space, defined in terms of the IMAPs G_π. Namely, given a distribution $\mathbb{P}$, define $\mathcal{L}_{\mathbb{P}}$ to be the directed graph with vertex set $\{G_\pi : \pi \in \mathcal{P}(V)\}$ with an arc from G_π to G_τ when there exists a graph G', obtainable from G_π via a single legitimate mark change, such that $\tau = \text{po}(G')$.

Figure 5 shows the outgoing edges of a particular minimal IMAP G_π in $\mathcal{L}_{\mathbb{P}}$ when $\mathbb{P}$ is faithful to G^* of Figure 2a. As shown, there are two possible legitimate

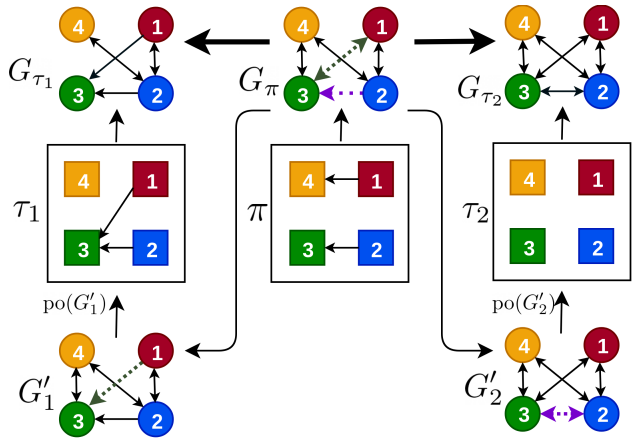


Figure 5: Example of the outgoing edges (in bold) of the node G_π in $\mathcal{L}_\mathbb{P}$ where $\mathbb{P}$ is faithful to G^* from Figure 2b. The graphs G'_1 and G'_2 are obtained from G_π via legitimate mark changes of the colored dashed edges. The posets τ_1, τ_2 are $\text{po}(G'_1), \text{po}(G'_2)$ respectively so $\mathcal{L}_\mathbb{P}$ has edges from G_π to G_{τ_1} and G_{τ_2}.

mark changes that can be performed on G_π shown as dashed. Changing the bidirected dashed edge, for example, would result in G'_1 with $\tau_1 = \text{po}(G'_1)$. Hence, there is an outgoing edge from G_π to G_{τ_1} in $\mathcal{L}_\mathbb{P}$.

Algorithm 2 is the resulting greedy search for the sparsest G_π over $\mathcal{L}_\mathbb{P}$. We call this algorithm the *greedy sparsest poset algorithm* (*GSPo*). We conjecture, supported by simulations on the order of 100,000s of examples (see Appendix E), that GSPo is consistent under the restricted-faithfulness assumption (using a sufficiently large depth d in the search), i.e., it yields a DMAG that is Markov equivalent to G^* no matter the starting point. This conjecture generalizes the consistency result proven for GSP in the fully observed setting (Solus et al., 2017).

Conjecture 1. *Let $\mathbb{P}$ be a probability distribution that is restricted-faithful to a DMAG G^*. If π_0 is any poset, then there exists a directed path $\pi_0 \to \pi_1 \to \cdots \to \pi_k$ in $\mathcal{L}_\mathbb{P}$ such that G_{π_k} is sparsest, and such that π_i always has weakly fewer edges than π_{i-1}.*

4.2 Implementation

A crucial practical consideration for GSPo is the choice of the starting poset π_0, since a sparser initial IMAP would be favorable. The empty poset $\emptyset$ provides a simple starting place, with $G_\emptyset = \{i \leftrightarrow j \mid X_i \not\perp\!\!\!\perp_\mathbb{P} X_j\}$, but in general will not be sparse. An effective alternative is to start at a sparse DAG that is a minimal IMAP (e.g., 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 (Solus et al., 2017). We compare these initialization schemes in Section 5.

Algorithm 2 GREEDY SPARSEST POSET (GSPO)

Input: $\mathcal{I}(\mathbb{P})$, with $\mathbb{P}$ restricted-faithful to G^*; starting poset π_0; maximum depth d.
Output: A minimal IMAP of $\mathbb{P}$.
Set $\pi = \pi_0$;
Via depth-first search with root π and depth at most d, find path $\pi_0, \ldots, \pi_k$ such that π_i and π_{i+1} are adjacent in $\mathcal{L}_\mathbb{P}$, $|G_{\pi_i}| \geq |G_{\pi_{i+1}}|$ and $|G_{\pi_0}| > |G_{\pi_k}|$.
If such π_k exists, set π to π_k, and repeat this step.
Otherwise, return G_π.

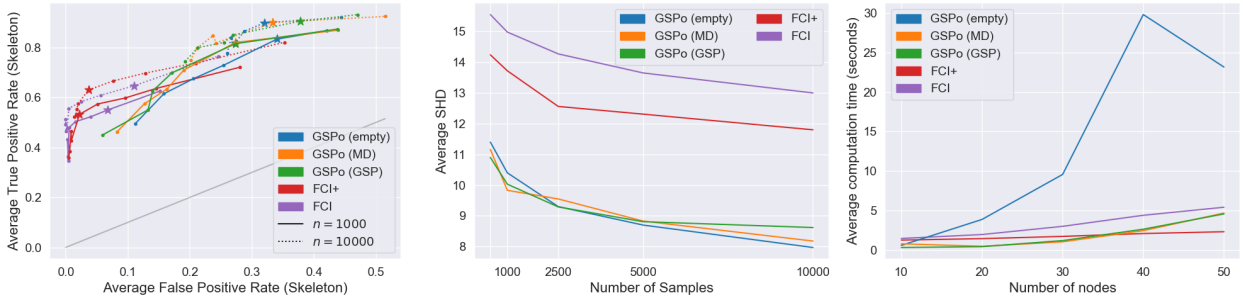
5 EXPERIMENTAL RESULTS

In this section, we compare the performance of GSPo to FCI and FCI+ in recovering DMAGs from samples of the observed nodes. In each simulation, we sample 100 Erdös-Rényi DAGs on $p+K$ nodes with s expected neighbors per node, then form DMAGs by marginalizing over the first K nodes, to obtain DMAGs on p nodes. If i, j is an edge $i \to j$ in the DAG, we assign an edge weight w_{ij} drawn uniformly at random from $[-1, -.25] \cup [.25, 1]$; we set $w_{ij} = 0$ otherwise. 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 remove the first K columns of the data matrix.

In each run of GSPo, we set the depth parameter d to 4, and run the algorithm 5 times for each graph (using different initializations). For DAGs, a depth of 4 has been used to reflect the empirically-observed average size of the MECs (Gillispie and Perlman, 2001; Solus et al., 2017). Although we are not aware of results on the average size of the MECs of DMAGs, we found little benefit in using values larger than 4.

In Figure 6, we chose $p = 10$, $K = 3$, and $s = 3$. The resulting graphs have on average about 4 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 performance of GSPo with three initialization schemes as compared to FCI and FCI+ on recovering the skeleton of the true MAG. 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 mis-



(a) Skeleton edge recovery (ROC) (b) Skeleton Edge Recovery (SHD) (c) Median Runtime

Figure 6: In (a) and (b), $p = 10$, $K = 3$, and $s = 3$. In (a), each variant of GSPo was run on 8 α values from 10^{-10} to .7, and each variant of FCI was run on 7 α values from 10^{-20} to .5. The best α for each algorithm was selected for (b); 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 best α was selected for (c), except for FCI, which was run with $\alpha = 10^{-3}$ since higher α values were extremely slow.

takenly removing edges. ROC curves for $p = 50$ nodes are reported in Appendix F, with similar findings.

Figure 6b shows the structural Hamming distance (SHD)[2] of the skeleton of the true DMAG to the skeleton of the DMAG estimated by each algorithm. For each algorithm, the value of α was picked from among the values used in Figure 6a in order to minimize the average SHD over all sample sizes; the corresponding values are marked by stars on the ROC curves. All variants of GSPo outperform both variants of FCI for all sample sizes in terms of SHD.

Figure 6c shows the median computation time required for each algorithm for graphs of varying number of vertices. Average computation time is in Appendix F. For each algorithm, we chose the parameter α based on the best-performing value in Figure 6b; 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 slower than FCI+ as the number of nodes increases. Given that CI tests in the construction of G_π involve all ancestors of pairs of nodes, this poor scaling is expected. Fortunately, this suggests that improvements along the lines of those in FCI+ may bring the scaling of GSPo in line with that of FCI+.

6 DISCUSSION

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. 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 (Solus et al., 2017) for learning DAGs to the setting with latent confounders, thereby providing a general hybrid 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.

Consistency of our greedy algorithm remains an open question, and is an interesting issue for future work. Furthermore, it may be possible to improve the statistical and computational performance of GSPo through modifications such as: more efficiently 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.

[2]the SHD between two undirected graphs is equal to the minimum number of edge additions/deletions required to transform from one graph to another

Acknowledgements

Daniel Irving Bernstein was funded by an NSF Mathematical Sciences Postdoctoral Research Fellowship (DMS-1802902). Basil Saeed was partially supported by the Abdul Latif Jameel Clinic for Machine Learning in Health at MIT. 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

Serge Bouc. The poset of posets. *arXiv preprint arXiv:1311.2219*, 2013.

David Maxwell 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.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. *arXiv preprint arXiv:1309.6824*, 2013.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

Steven B Gillispie and Michael 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.

David Heckerman, Abe Mamdani, and Michael P Wellman. Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–26, 1995.

Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

F. Mohammadi, C. Uhler, C. Wang, and J. Yu. Generalized permutohedra from probabilistic graphical models. *SIAM Journal on Discrete Mathematics*, 32:64–93, 2018.

Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, Peter Bühlmann, et al. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11 (2):5342–5374, 2017.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

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

Thomas Richardson. Markov properties for acyclic directed mixed graphs. Technical report, Technical Report, 1999.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Ilya Shpitser, Thomas S Richardson, James M Robins, and Robin Evans. Parameter and structure learning in nested markov models. *arXiv preprint arXiv:1207.5058*, 2012.

Liam Solus, Yuhao Wang, Lenka Matejovicova, and Caroline Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

Peter Spirtes and Thomas 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.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in Advanced Mathematics*, 2011.

Konstantinos Tsirlis, Vincenzo Lagani, Sofia Triantafillou, and Ioannis Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85, 2018.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, Bin Yu, et al. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41 (2):436–463, 2013.

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

Jiji Zhang and Peter 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.

Jiji Zhang and Peter Spirtes. A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1419*, 2012.