Learning directed acyclic graphs based on sparsest permutations

Garvesh Raskutti[1]   Caroline Uhler [2]

[1] Statistical and Applied Mathematical Sciences Insitute
[2] Institute of Science and Technology Austria

Abstract

We consider the problem of learning a Bayesian network or directed acyclic graph (DAG) model from observational data. A number of constraint-based, score-based and hybrid algorithms have been developed for this purpose. For constraint-based methods, statistical consistency guarantees typically rely on the faithfulness assumption, which has been show to be restrictive especially for graphs with cycles in the skeleton. However, there is only limited work on consistency guarantees for score-based and hybrid algorithms and it has been unclear whether consistency guarantees can be proven under weaker conditions than the faithfulness assumption.

In this paper, we propose the sparsest permutation (SP) algorithm. This algorithm is based on finding the causal ordering of the variables that yields the sparsest DAG. We prove that this new score-based method is consistent under strictly weaker conditions than the faithfulness assumption. We also demonstrate through simulations on small DAGs that the SP algorithm compares favorably to the constraint-based PC and SGS algorithms as well as the score-based Greedy Equivalence Search and hybrid Max-Min Hill-Climbing method. In the Gaussian setting, we prove that our algorithm boils down to finding the permutation of the variables with sparsest Cholesky decomposition for the inverse covariance matrix. Using this connection, we show that in the oracle setting, where the true covariance matrix is known, the SP algorithm is in fact equivalent to ℓ_0-penalized maximum likelihood estimation.

1 Introduction

A fundamental goal in many scientific problems is to determine causal or directional relationships between variables in a system. While there are a number of challenges involved with inferring directional relations amongst several variables, a useful simplification is to assume that the causal structure is modeled by a directed acyclic graph (DAG) or Bayesian network.

We start with definitions related to graphs and Bayesian networks. A DAG $G = (V, E)$ consists of a set of vertices V and a set of directed edges E with no directed cycle. We usually take $V = \{1, 2, \ldots, p\}$ and associate a random vector $(X_1, X_2, \ldots, X_p)$ with probability distribution $\mathbb{P}$ over the vertices in G. A directed edge from vertex j to k is denoted by (j, k) or $j \to k$. The set $pa(k)$ of *parents* of a vertex k consists of all nodes j such that $(j, k) \in E$. If there is a directed path $j \to \cdots \to k$, then k is called a *descendent* of j and j is an *ancestor* of k. The set $de(k)$ denotes the set of all descendants of a node k. The *non-descendents* of a node k are $nd(k) = V \setminus (\{k\} \cup de(k))$. For a subset $S \subset V$ we define $an(S)$ to be the set of nodes k that are in S or are ancestors of some node in S. Two nodes that are connected by an edge are called *adjacent*. A triple of nodes (j, k, ℓ) is an *unshielded triple* if j and k are adjacent to ℓ but j and k are not adjacent. An unshielded triple (j, k, ℓ) forms a *v-structure* if $j \to \ell$ and $k \to \ell$. In this case ℓ is called a *collider*. Furthermore, an undirected path π from j to k *d-connects* j and k given $S \subset V \setminus \{j, k\}$ if every collider on π is in $an(S)$ and every non-collider on π is not in S. If G has no path that d-connects j and k given a subset S, then j and k are *d-separated* given S. Finally, let $X_j \perp\!\!\!\perp X_k \mid X_S$ with $S \subset V \setminus \{j, k\}$ denote the conditional independence (CI) statement that X_j is conditionally independent (as determined

by $\mathbb{P}$) of X_k given the set of variables $X_S = \{X_\ell \mid \ell \in S\}$, and let $X_j \not\perp\!\!\!\perp X_k \mid X_S$ denote conditional dependence. The Markov condition associates CI relations with a DAG:

Definition 1.1 (Markov condition (Spirtes et al. [18])). A probability distribution $\mathbb{P}$ over a set of vertices V satisfies the *Markov condition* with respect to a DAG $G = (V, E)$ if

$$X_j \perp\!\!\!\perp X_{nd(j)\backslash pa(j)} \mid X_{pa(j)}, \quad \text{for all } j \in V.$$

The Markov condition (also called the *local Markov condition*) is in fact equivalent to the *global Markov condition*: $X_j \perp\!\!\!\perp X_k \mid X_S$ for all triples (j, k, S) for which j is d-separated from k given S (see, for example, [12, ch. 3]). In general, there are many DAGs that are *Markov equivalent*, meaning that they satisfy the (local or global) Markov assumption and hence the same CI relations with respect to a distribution $\mathbb{P}$. Let $\mathcal{M}(G)$ denote the Markov equivalence class consisting of all DAGs with the same CI relations as G. Verma and Pearl [23] showed that the Markov equivalence class of a DAG G is uniquely defined by the *skeleton* (i.e., the undirected edges of G) and the *v-structures*.

The problem we consider in this paper is to determine $\mathcal{M}(G^*)$, or equivalently, to determine the skeleton and the v-structures of the underlying DAG G^*, based on n independent and identically distributed draws $X^{(1)}, \ldots, X^{(n)}$ from a distribution $\mathbb{P}$ on G^*. The underlying graph G^* may also be interpreted as the causal graph. In order to avoid degenerate distributions, we assume throughout that the distribution $\mathbb{P}$ has positive measure everywhere, meaning that $\mathbb{P}(X_S) > 0$ for all subsets $S \subset V$. A number of algorithms have been developed for inferring Bayesian networks and they can broadly be classified as one of three approaches: constraint-based methods, score-based methods and hybrid methods.

1.1 Constraint-based algorithms and faithfulness

Constraint-based methods involve first finding the (undirected) skeleton and then identifying v-structures based on CI testing. The skeleton is inferred by starting with a complete DAG and deleting edge (j, k) if there exists a set $S \subset \{1, 2, ..., p\} \backslash \{j, k\}$ such that $X_j \perp\!\!\!\perp X_k \mid X_S$. The most widely used constraint-based methods are the *SGS algorithm* [9] and the *PC algorithm* [18]. The PC algorithm is a more computationally efficient version of the SGS algorithm, since at each iteration it only searches over subsets S that include neighbors of $\{j, k\}$.

Constraint-based methods are guaranteed to be consistent under the *faithfulness assumption* (in addition to the Markov assumption), which requires that all CI relations are encoded by the DAG:

Definition 1.2 (Faithfulness assumption (Spirtes et al. [18])). A distribution $\mathbb{P}$ is *faithful* with respect to a DAG G if for any $j, k \in V$ and any subset $S \subset V \backslash \{j, k\}$,

$$k \text{ is d-separated from } j \mid S \quad \Longleftrightarrow \quad X_j \perp\!\!\!\perp X_k \mid X_S.$$

The faithfulness assumption is significantly stronger than the Markov assumption since it requires that *all* CI relations of $\mathbb{P}$ are correctly inferred from G^*. Unfortunately, the faithfulness assumption is very sensitive to hypothsis testing errors for inferring CI statements from data and has been shown to be quite restrictive, especially for graphs with undirected cycles in the skeleton (see Uhler et al. [21]). This is the case since multiple paths between pairs of nodes can cause path cancellation and lead to additional unfaithful CI relations. A number of relaxations of the faithfulness assumption have been suggested, such as:

Definition 1.3. A distribution $\mathbb{P}$ satisfies the *restricted-faithfulness assumption* with respect to a DAG G if it satisfies the following two conditions:

(i) *adjacency-faithfulness* (Ramsey et al. [15]): for all $(j, k) \in E$,

$$X_j \not\perp\!\!\!\perp X_k \mid X_S \quad \text{for all } S \subset V \setminus \{j, k\};$$

(ii) *orientation-faithfulness* (Ramsey et al. [15]): for all unshielded triples (j, k, ℓ) and all subsets $S \subset V \setminus \{j, k\}$ such that j is d-connected to k given S,

$$X_j \not\perp\!\!\!\perp X_k \mid X_S.$$

Clearly, the restricted-faithfulness assumption is significantly weaker than the original faithfulness assumption, since it only requires faithfulness for pairs of vertices corresponding to edges and triples of vertices corresponding to v-structures. But for graphs with cycles in the skeleton restricted faithfulness is still very restrictive [21]. For the SGS algorithm adjacency-faithfulness is necessary but not sufficient for inferring the skeleton. However, for the PC algorithm adjacency faithfulness is neither necessary nor sufficient, since the algorithm does not necessarily test all subsets $S \subset V \setminus \{j, k\}$. Assuming the skeleton is inferred correctly, orientation faithfulness is the weakest known sufficient condition for inferring the v-structures of the DAG model for both the PC and SGS algorithms. A number of attempts have been made to modify the constraint-based algorithms to adjust for weaker conditions than faithfulness (e.g., [15, 27]). But these relaxations of the faithfulness assumption have ultimately led to weaker claims which don't guarantee discovery of $\mathcal{M}(G^*)$ (see, e.g., [19]).

1.2 Score-based and hybrid algorithms

Score-based methods posit a scoring criterion $\text{Score}(G; X^{(1)}, \ldots, X^{(n)})$ for each DAG or Bayesian network G and then search for the network with the highest score given the observations $X^{(1)}, \ldots, X^{(n)}$. Typical examples of scoring functions are the Bayesian Information Criterion (BIC) and Minimum Description Length (MDL). One of the challenges for scoring-based methods is that searching over the space of DAGs is NP-hard in general [1].

The *Greedy Equivalence Search* (GES) algorithm developed by Chickering [2] is a computationally efficient way to find locally optimal DAG models. Such local searches require that the scoring function is *decomposable*, meaning that the scoring function can be decomposed into a sum of local scores, which only depend on the parents of a node:

$$\text{Score}\left(G; \ X^{(1)}, \ldots, X^{(n)}\right) = \sum_{j=1}^{p} \text{Score}\left(pa_G(j); \ X^{(1)}, \ldots, X^{(n)}\right).$$

BIC and MDL both have this property. As we will show in Section 2, also the scoring criterion proposed in this paper is decomposable.

Hybrid algorithms attempt to exploit the advantages of both constraint-based and score-based methods. Two of the most widely used hybrid algorithms are the closely related *Sparse Candidate algorithm* [7] and the *Max-Min Hill-Climbing* (MMHC) algorithm [20]. Both methods first infer a skeleton using a local search method based on CI testing and then perform a greedy search over DAG models that respect the skeleton output. MMHC is generally preferred to the Sparse Candidate algorithm since it does not impose a maximum in-degree constraint, which can be quite restrictive.

While both, the score-based and hybrid methods are believed to be less prone to hypothesis testing errors, there is not much work addressing conditions under which these methods are consistent. In particular, it is unclear whether score-based methods are consistent under stronger, weaker or incomparable conditions than faithfulness.

1.3 Our contributions

In the following, we propose and analyze a score-based method for learning the Markov equivalence class of the underlying DAG based on observational data. Our method involves searching for the DAG with the fewest edges that satisfies the Markov assumption. Hence the scoring function is

$$\text{Score}(G; X^{(1)}, \ldots, X^{(n)}) = \begin{cases} -|G| & \text{if } G \text{ is Markov with respect to } X^{(1)}, \ldots, X^{(n)}, \\ -\infty & \text{otherwise,} \end{cases} \tag{1}$$

where $|G|$ denotes the number of edges in G and a DAG is Markov with respect to the observations $X^{(1)}, \ldots, X^{(n)}$, if it satisfies the Markov assumption with respect to the CI relations inferred from these observations. In Section 2, we describe our sparsest permutation (SP) algorithm in more detail and give a necessary and sufficient condition for consistency of our algorithm in the oracle setting, the *sparsest Markov representation assumption*. Most importantly, we prove that the sparsest Markov representation assumption is strictly weaker than the restricted faithfulness condition. We also discuss connections between the sparsest Markov representation assumption and other minimality assumptions proposed by Pearl [13] and Spirtes et al. [18].

In Section 3 we consider the effects of hypothesis testing on both, our SP and constraint-based algorithms. We propose inferring the CI relations for the SP algorithm using hypothesis testing in the same way as is done for constraint-based algorithms. Using this hypothesis testing framework, we prove uniform consistency of the SP algorithm under a *strong sparsest Markov representation assumption*, which is strictly weaker than the strong faithfulness assumption, the weakest known sufficient condition for uniform consistency of the constraint-based SGS and PC algorithms.

In Section 4, we consider the special case where the variables are Gaussian, i.e. $\mathbb{P} = \mathcal{N}(0, \Sigma)$ where Σ is a positive definite covariance matrix. We use a result due to Pourahmadi [14] to prove that the SP algorithm is equivalent to the problem of finding the sparsest Cholesky factorization of the inverse covariance matrix Σ^{-1}. Using this connection, we show that the SP algorithm corresponds to an oracle or noiseless version of the ℓ_0-penalized maximum likelihood approach described by van de Geer and Bühlmann [22]. Hence, in the oracle setting where Σ is known, the sparsest Markov representation assumption is necessary and sufficient to guarantee success of ℓ_0-penalized maximum likelihood estimation. As a consequence, in the oracle setting ℓ_0-penalized maximum likelihood estimation requires strictly weaker conditions than the faithfulness assumption.

Finally, Section 5 provides an empirical comparison between the SP algorithm, the constraint-based SGS and PC algorithms, the score-based GES algorithm and the hybrid MMHC algorithm. The CI relations are inferred from data using a hypothesis test based on Fisher's z-transform, and we consider random DAG models of up to 8 nodes. Our simulation results demonstrate that the SP algorithm recovers the true skeleton of the DAG more frequently than the state-of-the art competing methods. The bottleneck is that a computationally-intensive search over the set of all permutations is required. Our simulations confirm what our theoretical results suggest, namely that the SP algorithm is consistent under weaker conditions than the constraint-based SGS and PC algorithms.

2 Sparsest permutation algorithm

In this section, we present the SP algorithm and the main results which guarantee point-wise consistency of the SP algorithm under strictly weaker conditions than restricted-faithfulness, a necessary condition for consistency of the constraint-based SGS algorithm. The principle underlying the SP algorithm is that every

DAG gives rise to a partial ordering (or equivalently, a permutation) of the vertices $V = \{1, 2, \ldots, p\}$, and determining the DAG is equivalent to determining the ordering / permutation of the vertices and the skeleton. For any permutation π of the vertices it is in fact possible to construct a DAG $G_\pi = (V, E_\pi)$ that satisfies the Markov assumption in the following way: Let $j < k$, then

$$(\pi(j), \pi(k)) \in E_\pi \iff X_{\pi(j)} \not\perp\!\!\!\perp X_{\pi(k)} \mid X_S \text{ where } S = \{\pi(1), \pi(2), ..., \pi(k-1)\} \setminus \{\pi(j)\}.$$

Then G_π satisfies the Markov and *minimality assumption* meaning that there is no proper sub-DAG (with the same vertex set and a strict subset of the directed edges) that satisfies the Markov assumption. The proof of this result (Lemma 2.1) is given in Section 7.

Lemma 2.1. *Given a positive measure $\mathbb{P}$ on $(X_1, X_2, \ldots, X_p)$, then G_π satisfies the Markov and minimality assumption for all permutations π of the vertices $V = \{1, 2, \ldots, p\}$.*

Lemma 2.1 suggests that there are $p!$ DAGs satisfying the Markov and minimality assumptions (some of which belong to the same Markov equivalence class). The problem is to select a permutation π resulting in a DAG G_π which belongs to $\mathcal{M}(G^*)$. In this paper, we propose to select the permutation π yielding the smallest number of edges, i.e., the *sparsest permutation*. This is the most parsimonious DAG representation of the data satisfying the Markov property. As we will show in this section, the sparsest permutation is in fact in $\mathcal{M}(G^*)$ under some assumptions on G^*.

Let $S(G)$ denote the skeleton of a DAG G and $|G|$ the number of edges in G (or $S(G)$). Then the SP algorithm is defined as follows:

(1) For all permutations π of the vertices {1,2,...,p} construct G_π and let $s_\pi = |G_\pi|$.

(2) Choose the set of permutations $\{\pi^*\}$ for which s_{π^*} is minimal amongst all permutations.

(3) Output G_{π^*} for all π^* such that s_{π^*} is minimal amongst all permutations.

An issue of the SP algorithm is that the output will often contain multiple DAGs, some of which may not belong to the same Markov equivalence class. In such a case, one approach would be to output all DAGs and let the user decide which is most appropriate. However, for the theoretical analysis and simulations in this paper, the presence of multiple possible Markov equivalence classes is interpreted as a failure of the SP algorithm.

Note that for a given permutation only a single CI relation needs to be tested for each edge unlike in the PC algorithm in which (in the worst case) all CI relations need to be tested. On the flip-side, the SP algorithm requires testing over all $p!$ permutations which is an NP-complete problem in general. The focus of this paper is not on computational complexity, but rather on guaranteeing recovery of the correct equivalence class under mild assumptions. We only briefly discuss computational issues in the following paragraph and in Section 4 for the special case when $\mathbb{P}$ is Gaussian.

First, note that since $-|G| = -\sum_{j=1}^{p} |pa_G(j)|$, the score used in the SP algorithm is decomposable. Hence local searches used for the GES algorithm may also be used for the SP algorithm. Note also that the SP algorithm involves a search over the set of orderings rather than the space of networks as is the case for GES. Since the set of orderings is much smaller than the set of Bayesian networks, there are potential computational advantages for heuristic searches when using the SP score compared to GES. However, since the focus of this paper is on consistency guarantees, we assume an exhaustive search over the set of all permutations.

Let $\{G_{SP}\}$ denote the outputted graphs of the SP algorithm. The following lemma (the proof can be found in Section 7) describes properties of the outputted skeletons and justifies choosing the sparsest DAG.

In particular, we show that under restricted-faithfulness the sparsest DAGs G_{SP} are in fact in the Markov equivalence class of G^*. This already suggests that our algorithm requires weaker assumptions than the SGS and PC algorithms.

Lemma 2.2. *Let G^* denote the true underlying DAG whose nodes have distribution $\mathbb{P}$. Then as $n \to \infty$ it holds for all DAGs G_{SP} that*

  (a) $|G_{SP}| \leq |G^*|$,

  (b) if G^* satisfies the adjacency faithfulness assumption with respect to $\mathbb{P}$, then $S(G_{SP}) = S(G^*)$,

  (c) if G^* satisfies the restricted faithfulness assumption with respect to $\mathbb{P}$, then $G_{SP} \in \mathcal{M}(G^*)$.

The SP algorithm tries to overcome one of the problems of constraint-based methods, namely that they tend to miss edges in the skeleton. For an edge between j and k to be included for example using the SGS algorithm, it must hold that $X_j \not\perp\!\!\!\perp X_k \mid X_S$ for *all* subsets $S \subset V \setminus \{j, k\}$. On the other hand, for a given permutation π the SP algorithm includes an edge between j and k as long as $X_j \not\perp\!\!\!\perp X_k \mid X_S$ where $S = \{\pi(1), \pi(2), \ldots, \pi(\max(\pi^{-1}(j), \pi^{-1}(k)))\} \setminus \{j, k\}$. The sparsity criterion in the SP algorithm already determines the sparsest DAG that satisfies the Markov assumption. Deleting more edges as is done in the SGS algorithm is therefore more likely to delete edges that are in G^*. This is reflected in our simulations shown in Section 5.

2.1 Consistency of the SP algorithm in the oracle setting

In the following, we describe a necessary and sufficient condition for point-wise consistency of the SP algorithm. Throughout, we use the term *success* of the SP algorithm to denote recovery of $\mathcal{M}(G^*)$ and *failure* to denote failure to recover $\mathcal{M}(G^*)$.

Definition 2.3 (Sparsest Markov representation (SMR) assumption)**.** A pair $(G^*, \mathbb{P})$ satisfies the *sparsest Markov representation assumption* if $(G^*, \mathbb{P})$ satisfies the Markov assumption and $|G| > |G^*|$ for every DAG G such that $(G, \mathbb{P})$ satisfies the Markov assumption and $G \notin \mathcal{M}(G^*)$.

The SMR assumption asserts that the true DAG G^* is the (unique) sparsest DAG satisfying the Markov assumption. The SMR assumption is in fact equivalent to the scoring-consistency condition defined by Chickering [2] using the scoring function given in (1). In the absence of additional information, the SMR assumption is a necessary condition for any algorithm that uses the CI relations of $\mathbb{P}$ to infer G^*, since if there is a DAG $G \notin \mathcal{M}(G^*)$ which satisfies the Markov assumption and is as sparse or sparser than G^*, there is no reason to select G^* instead of G. Now we present our first main result.

Theorem 2.4. *The SP algorithm outputs $\mathcal{M}(G^*)$ if and only if the pair $(G^*, \mathbb{P})$ satisfies the SMR assumption.*

Proof. First, let $(G^*, \mathbb{P})$ satisfy the SMR assumption and assume that $G_{SP} \notin \mathcal{M}(G^*)$. By Lemma 2.2, $|G_{SP}| \leq |G^*|$, which is a contradiction to the SMR assumption, since G_{SP} satisfies the Markov assumption (by Lemma 2.1).

Now assume $(G^*, \mathbb{P})$ does not satisfy the SMR assumption. Then there exists a DAG $G \notin \mathcal{M}(G^*)$ with $|G| \leq |G^*|$ that satisfies the Markov assumption. Without loss of generality, we choose G with minimal number of edges. Let π denote an ordering which is consistent with G. Since G_π satisfies the minimality assumption, $S(G_\pi) = S(G)$. If $S(G) < S(G^*)$, the SP algorithm will output G and hence fail. If $S(G) = S(G^*)$, the SP algorithm will choose both G and G^*. Hence the set $\{G_{SP}\}$ will include DAGs that are not in $\mathcal{M}(G^*)$. $\qquad\square$

6

In the following result, we show that the SMR assumption is strictly weaker than the restricted faithfulness assumption.

Theorem 2.5. *Let $(G^*, \mathbb{P})$ satisfy the Markov condition. Then,*

    (a) the restricted faithfulness assumption implies the SMR assumption,

    (b) there exists a pair $(G^*, \mathbb{P})$ satisfying the SMR assumption that does not satisfy restricted faithfulness.

Proof. (a) We proved in Lemma 2.2 (c) that restricted faithfulness is sufficient for consistency of the SP algorithm. Applying Theorem 2.4 we hence find that restricted faithfulness implies the SMR assumption.

(b) In the following, we construct an example of a 4-node DAG that satisfies the SMR assumption, but for which restricted faithfulness fails. Let $G^* = (\{1, 2, 3, 4\}, E^*)$ be the 4-cycle with $E^* = \{(1, 2), (1, 4), (2, 3), (3, 4)\}$ and let $\mathbb{P}$ satisfy the CI relations $X_1 \perp\!\!\!\perp X_3 \mid X_2$, $X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$ and $X_1 \perp\!\!\!\perp X_2 \mid X_4$. The last CI relation violates adjacency-faithfulness. Such a violation is straightforward to construct: It corresponds to cancellation of the two paths between X_1 and X_2 conditioned on X_4, namely $X_1 \to X_4 \leftarrow X_3 \leftarrow X_2$ and $X_1 \to X_2$. On the other hand, using the SP algorithm it can be shown that any permutation other than the true permutation $(1, 2, 3, 4)$ produces a DAG $G_\pi \notin \mathcal{M}(G^*)$ with more than 4 edges. As an example, we consider the permutation $(1, 4, 2, 3)$, i.e. $\pi(1) = 1$, $\pi(2) = 4$, $\pi(3) = 2$, $\pi(4) = 3$. In this case only the edge $(1, 2)$ would be omitted leading to a DAG with 5 edges. Similarly, one can check all other permutations, proving that $(G^*, \mathbb{P})$ satisfies the SMR assumption. $\square$

The main disadvantage of the constraint-based SGS algorithm compared to the SP algorithm is that any violation of adjacency faithfulness leads to the deletion of a "true" edge, because the SGS algorithm omits an edge if $X_j \perp\!\!\!\perp X_k \mid X_S$ for *any* subset $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$. On the other hand, the SP algorithm exploits the global structure of G^* to ensure that if G^* violates the faithfulness assumption but satisfies the SMR assumption, any permutation π resulting in a G_π that removes an edge from G^* must add at least 2 other edges (cf. the 4-cycle example in the proof of Theorem 2.5). As far as we are aware, standard faithfulness is the weakest known sufficient condition for consistency of the PC and SGS algorithms and this condition is significantly stronger than restricted faithfulness or the SMR assumption. In Section 5, we empirically explore how much weaker the SMR assumption is for random graphs and confirm that, as suggested by the theoretical results in this section, the SP algorithm is consistent more frequently than the PC and SGS algorithms and tends to remove less edges.

2.2 Detecting faithfulness violations

In Section 2.1 we analyzed under what conditions the SP algorithm is consistent. However, the only performance metric we considered was a 0-1-metric where "0" means the algorithm failed to recover $\mathcal{M}(G^*)$ while "1" means the algorithm recovered $\mathcal{M}(G^*)$.

Depending on the problem context, different types of failure may be seen as better or worse. In particular, previous work related to the SGS algorithm has distinguished between two types of errors leading to failure, detectable and undetectable errors [15, 27]. An error is *detectable* when the algorithm fails to return $\mathcal{M}(G^*)$, but there exists no DAG to which the distribution $\mathbb{P}$ is faithful to. An error is *undetectable* when the algorithm returns a DAG $\tilde{G} \notin \mathcal{M}(G^*)$ and $\mathbb{P}$ is faithful with respect to $\tilde{G}$. Zhang and Spirtes [27] show that under *triangle-faithfulness*, meaning that the faithfulness assumption holds on all triangles in the skeleton, any error made by the SGS algorithm is detectable. In particular, for DAGs that contain no triangles in the skeleton any failure of the SGS algorithm is detectable.

7

Figure 1: 4-node examples from Section 2.2.

In Theorem 2.7 we show that, similarly as for the SGS algorithm, under triangle-faithfulness any failure of the SP algorithm is detectable, meaning that the SP algorithm either only outputs DAGs from the true Markov equivalence class $\mathcal{M}(G^*)$, or it outputs a collection of DAGs from different Markov equivalence classes all with the same number of edges, one of which is $\mathcal{M}(G^*)$. We illustrate this result with an example. Let G^* be the 4-cycle shown in Figure 1(a) with $E^* = \{(1,2),(1,4),(2,3),(3,4)\}$ and let $\mathbb{P}$ satisfy the CI relations corresponding to d-separation and one additional CI relation violating adjacency faithfulness, namely $X_1 \perp\!\!\!\perp X_4$. Since G^* contains no triangles, it satisfies triangle-faithfulness and based on the results of Zhang and Spirtes [27], the SGS algorithm makes a detectable error. On the other hand, the SP algorithm would output two DAGs with 4 edges, the true 4-cycle for the original ordering and the DAG G_π shown in Figure 1(b) with $E_\pi = \{(1,2),(1,3),(3,2),(4,3)\}$ for the permutation π with $\pi(1) = 1$, $\pi(2) = 4$, $\pi(3) = 3$ and $\pi(4) = 2$. Since the SP algorithm outputs two different Markov equivalence classes, failure of the SP algorithm is detectable.

The proof of Theorem 2.7 is based on the assumption that violations of faithfulness correspond to path cancellations in the graph. This can be formalized as follows:

Definition 2.6 (Single-path-faithfulness assumption). A distribution $\mathbb{P}$ satisfies the *single-path-faithfulness assumption* with respect to a DAG $G = (V, E)$ if $i \not\perp\!\!\!\perp j \mid S$ for all triples (i, j, S) with $i, j \in V$ and $S \subset V \setminus \{i, j\}$ such that there is a unique path that d-connects i and j given S.

This assumption is satisfied for example in the Gaussian setting, since a partial correlation is a weighted sum of all paths between two vertices [21]. We conjecture that the single-path-faithfulness assumption is satisfied for all linear models. We expect, however, that Theorem 2.7 also holds when removing this assumption. The proof of Theorem 2.7 can be found in Section 7.

Theorem 2.7. *If $(G^*, \mathbb{P})$ satisfies the Markov condition, the single-path-faithfulness assumption and the triangle-faithfulness assumption, then $|G_{SP}| = |G^*|$ and every failure of the SP algorithm is detectable.*

So under triangle-faithfulness, if the SP algorithm outputs a unique Markov equivalence class, one is certain that it is the correct equivalence class. If the SP algorithm outputs multiple equivalence classes, the correct equivalence class is among the outputted ones. If triangle-faithfulness cannot be assumed but detecting faithfulness violations is desirable, an additional step could be included in the SP algorithm making sure that $\{G_{SP}\}$ satisfies the faithfulness assumption. However, the complexity of checking whether the faithfulness assumption is satisfied is equivalent to the complexity of the SGS algorithm.

2.3 Connection to other minimality assumptions

In this section, we provide a comparison between the SMR assumption and two well-known minimality assumptions, one developed by Spirtes, Glymour and Scheines [18] and another by Pearl [13]. We will refer

to these as SGS-minimality and P-minimality, respectively. Both minimality assumptions encourage the selection of DAG models with fewer edges. As we show in this section, they are in fact weaker assumptions than the SMR assumption.

The SGS-minimality condition was encountered already in Lemma 2.1 in Section 2 and asserts that there exists no proper sub-DAG of G^* that satisfies the Markov assumption with respect to $\mathbb{P}$. As shown in Lemma 2.1, there exists an SGS-minimal DAG for every ordering. Hence SGS-minimality is a significantly weaker condition than the SMR assumption.

P-minimality asserts that for DAGs satisfying the Markov assumption, models that satisfy more d-separation statements (or, equivalently, entail more CI statements) are preferred. Stated precisely, let G and G' be two Markovian DAGs and let $D_{sep}(G)$ denote the d-separation statements for G and $D_{sep}(G')$ the d-separation statements for G'. Then G' is *strictly preferred* to G if $D_{sep}(G) \subsetneq D_{sep}(G')$. In other words, the DAG G' entails a strict super-set of CI statements compared to the DAG G. The P-minimality assumption asserts that no DAG is *strictly preferred* to the true DAG G^*.

Zhang [25] proved that P-minimality implies SGS-minimality. In the same paper he also showed that under P-minimality any violation of faithfulness is detectable, since under P-minimality, either G^* satisfies the faithfulness condition or no DAG satisfies the faithfulness condition. We now prove that the SMR assumption is in fact also strictly stronger than P-minimality.

Theorem 2.8. *(a) If a DAG G satisfies the SMR assumption, it also satisfies the P-minimality assumption.*

(b) There exist DAG models $(G^*, \mathbb{P})$ where all DAGs that satisfy the SMR assumption belong to $\mathcal{M}(G^*)$ while there exist DAGs $G \notin \mathcal{M}(G^*)$ that satisfy the P-minimality assumption.

Proof. (a) Let G be a DAG that satisfies the SMR assumption and for contradiction assume G does not satisfy P-minimality. Then there exists a Markovian DAG G' such that $D_{sep}(G) \subsetneq D_{sep}(G')$. Hence $G' \neq G$ and as a consequence of the SMR assumption, $|G'| \geq |G|$. We first consider the case where the skeleton of G and G' are identical. So without loss of generality, there exists a v-structure $i \to k$, $j \to k$ in G' that is not in G. Hence there exists a subset S with $k \in S$ that d-separates i from j in G but does not d-separate i from j in G'. This is in contradiction with $D_{sep}(G) \subsetneq D_{sep}(G')$. As a consequence, the skeletons of G and G' are not identical. Since $|G'| \geq |G|$ this means that there exist vertices i, j that are connected by an edge in G' but not in G. Hence there exists a subset S that d-separates i from j in G but not in G', again contradicting $D_{sep}(G) \subsetneq D_{sep}(G')$.

(b) We consider the same example as in the proof of Theorem 2.5 (b). Let $G^* = (\{1, 2, 3, 4\}, E^*)$ be the 4-cycle with $E^* = \{(1, 2), (1, 4), (2, 3), (3, 4)\}$ and let $\mathbb{P}$ satisfy the CI relations $X_1 \perp\!\!\!\perp X_3 \mid X_2$, $X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$ and $X_1 \perp\!\!\!\perp X_2 \mid X_4$. All DAGs that satisfy the SMR assumption belong to $\mathcal{M}(G^*)$. However, the DAG with edge set $\tilde{E} = \{(1, 4), (1, 3), (4, 2), (4, 3), (2, 3)\}$ is Markov to $\mathbb{P}$, satisfies the P-minimality assumption, but is not in $\mathcal{M}(G^*)$. $\square$

Theorem 2.8 (b) proves that the converse of (a) is false. There are DAGs that satisfy P-minimality but do not satisfy the SMR assumption. In addition, Theorem 2.8 (b) shows that while P-minimality tends to encourage sparser models, it is less powerful than the SMR assumption for determining the true Markov equivalence class of G^*.

3 Effects of inferring CI relations from data

As mentioned earlier, the CI relations used for the SP algorithm and other constraint-based methods need to be estimated from data. In the following, we explain how to infer the CI relations using the standard hy-

pothesis testing framework and discuss what the consequences are for the performance of the SP algorithm.

Applying the standard hypothesis testing framework used in earlier work on the SGS and PC algorithms (see, e.g., [16, 26]), we define for every (j, k) and $S \subset \{1, 2, \ldots, p\} \setminus \{j, k\}$ the hypotheses

$$
\begin{aligned}
\mathcal{H}_0 & : X_j \perp\!\!\!\perp X_k \mid X_S, \\
\mathcal{H}_1 & : X_j \not\perp\!\!\!\perp X_k \mid X_S.
\end{aligned}
$$

Assuming $X_j^{(i)} \in \chi$ for each $1 \leq i \leq n$ and $1 \leq j \leq p$, a test statistic $T_n : \chi^n \to \mathbb{R}$ is used to construct a hypothesis test $\phi_n : \mathbb{R} \to \{0, 1\}$ such that $\phi_n(T_n) = 0$ if $T_n \in R_n \subset \mathbb{R}$ and $\phi_n(T_n) = 1$ if $T_n \in \mathbb{R} \setminus R_n$. If $\phi_n(T_n) = 0$, $\mathcal{H}_0$ is retained implying the CI relation $X_i \perp\!\!\!\perp X_j \mid X_S$, otherwise $\mathcal{H}_0$ is rejected.

For brevity, we only present the formal testing framework when $\mathbb{P}$ has a Gaussian distribution, although our main results apply for any distribution in which suitable CI tests can be constructed. In the Gaussian setting, a CI relation $X_j \perp\!\!\!\perp X_k \mid X_S$ is equivalent to zero partial correlation $\rho_{j,k|S} = \operatorname{corr}(X_j, X_k \mid X_S)$. Hypothesis testing then boils down to

$$
\begin{aligned}
\mathcal{H}_0 : \rho_{j,k|S} & = 0, \\
\mathcal{H}_1 : \rho_{j,k|S} & \neq 0,
\end{aligned}
$$

for which a z-test based on Fisher's z-transform [6] can be built as follows: From the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^n X^{(i)} X^{(i)T}$ compute the sample correlation coefficient $\hat{\rho}_{j,k|S} = [\widehat{\Sigma}]_{jk,jk} - [\widehat{\Sigma}]_{jk,S}([\widehat{\Sigma}]_{S,S})^{-1}[\widehat{\Sigma}]_{S,jk}$ via Schur complement. Next, compute Fisher's z-transform

$$
\hat{z}_{jk|S} = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{j,k|S}}{1 - \hat{\rho}_{j,k|S}} \right).
$$

Then according to Fisher [6], using $T_n = \sqrt{n - |S| - 3} |\hat{z}_{jk|S}|$, and $R_n = (-\Phi^{-1}(1 - \alpha/2), \Phi^{-1}(1 - \alpha/2))$, where Φ denotes the cumulative distribution function of $\mathcal{N}(0, 1)$, leads to a test of size α.

Now we analyze the consequences of inferring the CI relations for the SP algorithm based on the type-I and type-II error rates of these hypothesis tests. Note that a type-I error corresponds to missing a CI relation, whereas a type-II error corresponds to adding a CI relation. In order to contrast the following results for the sample setting from the results in Section 2.1 for the oracle setting, we denote the estimated DAG by $\hat{G}$. The following theorem shows that if the type-II errors outweigh the type-I errors, or more precisely, if there is a permutation π such that $|S(\hat{G}_\pi)| \leq |S(G^*)|$, then the SP algorithm performs at least as well as the SGS algorithm in terms of recovering the skeleton of G^*.

Theorem 3.1. *If the SGS algorithm recovers the skeleton $S(G^*)$ based on the inferred CI relations and there exists a permutation π such that $|S(\hat{G}_\pi)| \leq |S(G^*)|$, then $S(\hat{G}_{SP}) = S(G^*)$.*

Proof. Let $\hat{G}_{SGS}$ denote the output of the SGS algorithm. If the SGS algorithm recovers $S(G^*)$ then $S(\hat{G}_{SGS}) = S(G^*)$. Since by Lemma 2.2, $S(\hat{G}_{SGS}) \subset S(\hat{G}_{SP})$ (regardless of errors made when inferring the CI relations) and by assumption $|S(\hat{G}_\pi)| \leq |S(G^*)|$, it follows that $S(\hat{G}_{SP}) = S(G^*)$. $\qquad \square$

Consequently, if only type-II errors are made in inferring CI relations for the true permutation π, then $|S(\hat{G}_\pi)| \leq |S(G^*)|$ and the SP algorithm performs at least as well as the SGS algorithm in terms of recovering the skeleton. This is in general not the case in the presence of type-I errors, since it might happen that $|S(\hat{G}_\pi)| > |S(G^*)|$ for all π while the SGS algorithm recovers $\mathcal{M}(G^*)$. The following 4-node example illustrates this situation.

Example 3.2. Let $G^* = (\{1, 2, 3, 4\}, E^*)$ be a DAG with $E^* = \{(1, 2), (2, 3), (3, 4)\}$. Now assume a type-I error occurs missing the CI relation $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$, while all other CI relations are inferred correctly. The SGS algorithm recovers the correct skeleton. The correctly inferred CI relation $X_1 \perp\!\!\!\perp X_4 \mid X_3$ leads to the removal of the edge $(1, 4)$, although the CI relation $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$ is missing. On the other hand, the SP algorithm will include the edge $(1, 4)$ for a permutation π^* consistent with G^* and result in $|\hat{G}_\pi| = 4 > |G^*|$. Furthermore, it can be shown that every permutation π results in a DAG $\hat{G}_\pi$ such that $|\hat{G}_\pi| \geq 4$. So in this example, the SP algorithm fails to recover the skeleton while the SGS algorithm succeeds.

This 4-node example illustrates that the SGS algorithm is more robust to type-I errors in inferring the CI relations compared to the SP algorithm. The SGS algorithm infers the presence of an edge (j, k) if $X_j \not\perp\!\!\!\perp X_k \mid X_S$ for all $S \subset \{1, 2, \ldots, p\} \setminus \{j, k\}$. Hence as long as one CI relation $X_j \perp\!\!\!\perp X_k \mid X_S$ is inferred correctly, the SGS algorithm would not add an extra edge although some CI relations are missing. In Section 5, we analyze the frequency of errors made by additional and missing edges with respect to the true DAG for the SP, PC and SGS algorithms. Our results indicate that as long as the size α of the hypothesis test is set sufficiently low, type-I errors arise very rarely and the SP algorithm will generally outperform both the SGS and PC algorithms.

3.1 Uniform consistency of the SP algorithm

Robins et al. [16] proved that the faithfulness assumption is not sufficient to guarantee uniform consistency of the SGS and PC algorithms. Even if the true distribution $\mathbb{P}$ is faithful to G^*, the empirical distribution $\mathbb{P}_n$ might not satisfy the faithfulness assumption and might lead to failure of the PC algorithm. To overcome this problem, Zhang and Spirtes [26] introduced the strong-faithfulness assumption and proved that it ensures uniform consistency of the PC and SGS algorithms. To simplify notation, we define $\Omega_\lambda(\mathbb{P}) = \{(j, k, S) \mid |\text{corr}(j, k \mid S)| \leq \lambda\}$ for $\lambda > 0$.

Definition 3.3 (λ-strong faithfulness assumption (Zhang and Spirtes [26]))**.** Given $\lambda \in (0, 1)$, a Gaussian distribution $\mathbb{P}$ satisfies the λ-*strong-faithfulness assumption* with respect to a DAG G if for any $j, k \in V$ and $S \subset V \setminus \{j, k\}$,

$$j \text{ is d-separated from } k \text{ given } S \iff (j, k, S) \in \Omega_\lambda(\mathbb{P}).$$

Along the lines of the strong-faithfulness assumption for constraint-based methods, we need to introduce a strong-SMR assumption in order to guarantee uniform consistency of the SP algorithm. We will define this assumption for Gaussian distributions, although it can be extended to other distributions by replacing partial correlation tests with mutual information tests. Before stating the strong-SMR assumption, we point out that the faithfulness assumption, the Markov assumption, and the SMR assumption, which are defined for distributions $\mathbb{P}$, naturally extend to sets of CI relations. In particular, we will be discussing these assumptions with respect to $\Omega_\lambda(\mathbb{P})$.

Definition 3.4 (λ-strong sparsest Markov representation assumption)**.** Given $\lambda > 0$, a Gaussian distribution $\mathbb{P}$ satisfies the λ-*strong sparsest Markov representation assumption* with respect to a DAG G^* if G^* satisfies the Markov assumption with respect to the CI relations in $\Omega_\lambda(\mathbb{P})$ and $|G| > |G^*|$ for every DAG G such that G satisfies the Markov assumption with respect to $\Omega_\lambda(\mathbb{P})$ and $G \notin \mathcal{M}(G^*)$.

The λ-strong SMR assumption asserts that G^* satisfies the SMR assumption with respect to the CI relations in $\Omega_\lambda(\mathbb{P})$. Since strong-faithfulness asserts that G^* is faithful to the CI relations in $\Omega_\lambda(\mathbb{P})$, it

follows from Theorem 2.5 that the λ-strong SMR assumption is strictly weaker than the λ-strong faithfulness assumption.

Using the framework developed in [16] and used in [26], we state and prove uniform consistency of the SP algorithm.

Theorem 3.5. *Under the λ-strong SMR assumption, there exists a sequence of hypothesis tests $\phi_n : \chi^n \to \{0, 1\}$ such that the SP algorithm is uniformly consistent.*

A precise definition of uniform consistency in this setting is provided in Theorem 2 in [26].

Proof. Let $\hat{\rho}_{jk|S}$ for $S \subset V\{j, k\}$ denote a sample partial correlation and $\rho_{jk|S}$ the corresponding partial correlation. By Chebyshev's Inequality we get

$$\mathbb{P}\left(|\hat{\rho}_{jk|S} - \rho_{jk|S}| \geq \epsilon\right) \leq \frac{\text{var}(\hat{\rho}_{jk|S})}{\epsilon^2}.$$

The asymptotic distribution of $\hat{\rho}_{jk|S}$ can be computed from the asymptotic distribution of the z-transform [6]:

$$\sqrt{n - |S| - 3}\left(\hat{z}_{jk|S} - z_{jk|S}\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Since $\hat{\rho}_{jk|S} = \tanh(\hat{z}_{jk|S})$, using the delta method we get

$$\sqrt{n - |S| - 3}\left(|\hat{\rho}_{jk|S} - \rho_{jk|S}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(1 - \rho_{jk|S}^2\right)^2\right),$$

and as a consequence, asymptotically,

$$\mathbb{P}\left(|\hat{\rho}_{jk|S} - \rho_{jk|S}| \geq \epsilon\right) \leq \frac{1}{(n - |S| - 3)\epsilon^2}.$$

More general distributional results on $\hat{\rho}$ can be found in [4]. So, for n sufficiently large it holds with high probability that $|\hat{\rho}_{jk|S} - \rho_{jk|S}| \leq \epsilon$. Using the sample partial correlations $\hat{\rho}_{jk|S}$ for hypothesis testing, we can build a sequence of correlation hypothesis tests $\phi_n^{\lambda} : \chi^n \to \{0, 1\}$ such that the inferred CI relations equal the CI relations in $\Omega_{\lambda}(\mathbb{P})$ with probability approaching 1. By applying the λ-strong SMR assumption, the SP algorithm recovers $\mathcal{M}(G^*)$ with probability approaching 1. Hence the SP algorithm is uniformly consistent. $\square$

Theorem 3.5 ensures that the SP algorithm is uniformly consistent under the λ-strong SMR assumption which is strictly weaker than the λ-strong faithfulness assumption.

4 SP algorithm for Gaussian DAGs

In this section, we analyze the SP algorithm in the case when $\mathbb{P}$ is a Gaussian distribution. Assuming that the vertices of the DAG G^* are *topologically ordered*, meaning that $j < k$ for all $(j, k) \in E$, the distribution $\mathbb{P}$ is defined by the following linear structural equations:

$$X_k = \sum_{j<k} a_{jk} X_j + \epsilon_k, \qquad j = 1, 2, ..., p, \tag{2}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p) \sim \mathcal{N}(0, D)$, $D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ and $a_{jk} \neq 0$ if and only if $(j, k) \in E$. We also assume that $\sigma_j^2 > 0$ for all j to ensure that $\mathbb{P}$ has positive measure everywhere. In matrix form, the structural equations can be expressed as follows:

$$(I - A)^T X = \epsilon,$$

where $X = (X_1, X_2, \ldots, X_p)$ and $A \in \mathbb{R}^{p \times p}$ is an upper triangular matrix with $A_{jk} = a_{jk}$ for $j < k$. Since $\epsilon \sim \mathcal{N}(0, D)$,

$$X \sim \mathcal{N}(0, [(I - A)D(I - A)^T]^{-1}).$$

To simplify notation we define $U = I - A$ and the inverse covariance matrix $K = UDU^T$. Then UDU^T is the upper Cholesky decomposition of K. Note that the Cholesky decomposition for any symmetric positive definite matrix exists and is unique. The goal is then to determine the Markov equivalence class of a DAG G based on the inverse covariance matrix K which encodes all CI relations.

Using a result by Pourahmadi [14] we show that in the Gaussian setting finding the sparsest permutation π according to the SP algorithm is equivalent to finding the sparsest Cholesky decomposition of the inverse covariance matrix. Let K^π denote the inverse covariance matrix K with the rows and columns permuted according to the permutation π. Let $K^\pi = U^\pi D^\pi U^{\pi T}$ denote the upper Cholesky decomposition of K^π. The following result links the entries of U^π to conditional covariances:

Theorem 4.1 (Pourahmadi [14]). *For every permutation π and $j < k$,*

$$U_{jk}^\pi = 0 \qquad \Longleftrightarrow \qquad cov(X_{\pi(j)}, X_{\pi(k)} | X_S) = 0, \tag{3}$$

where $S = \{\pi(1), \pi(2), \ldots, \pi(k-1)\} \setminus \{\pi(j)\}$.

Applying this result, we can adapt the SP algorithm to the Gaussian setting and replace step (1) of the algorithm by:

> For each permutation π of the nodes $(1, 2, \ldots, p)$ form K^π and compute the Cholesky decomposition $K^\pi = U^\pi D^\pi U^{\pi T}$. Let $E^\pi \subset \{(1, 2), (1, 3), \ldots, (1, p), (2, 3), \ldots, (p - 1, p)\}$, where $(i, j) \in E^\pi$ if and only if $U_{ij}^\pi \neq 0$.

Let $\|U^\pi\|_0 = \sum_{i<j} \mathbf{1}(U_{ij}^\pi \neq 0)$ denote the number of non-zero off-diagonal elements of U^π. Then in the Gaussian setting the SP algorithm is equivalent to:

$$\begin{aligned} \arg\min_\pi \quad & \|U^\pi\|_0 \\ \text{subject to} \quad & K^\pi = U^\pi D^\pi U^{\pi T}. \end{aligned}$$

Hence for Gaussian random variables the SP algorithm boils down to finding the permutation of the rows and columns of the inverse covariance matrix that provides the sparsest Cholesky factorization. Although finding the sparsest Cholesky decomposition is NP-complete [24], a number of heuristic methods exist (see [3] for an overview and [8] for a review of the minimum degree algorithm).

An important consequence of the equivalence we established between the SP algorithm and the problem of finding the sparsest Cholesky factorization allows us to show that in the oracle setting, when the true covariance matrix Σ is given, the SP algorithm is equivalent to ℓ_0-penalized maximum likelihood estimation as described in [22]. We now discuss this connection in more detail.

Given a sample covariance matrix $\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X^{(i)} X^{(i)T}$ based on n i.i.d. samples $X^{(i)} \sim \mathcal{N}(0, \Sigma^*)$, van de Geer and Bühlmann [22] suggest using ℓ_0-penalized maximum likelihood estimation to learn $\mathcal{M}(G)$ as follows:

$$\arg\min_{K,P,U,D} \quad \text{trace}(K\widehat{\Sigma}_n) - \log\det(K) + C^2\|U\|_0$$
$$\text{subject to} \quad K = PUDU^TP^T,$$

where P is a permutation matrix, U an upper triangular matrix with all diagonal entries equal to 1, D a positive diagonal matrix, and C^2 a regularization parameter. They prove that if $C^2 = \mathcal{O}(\frac{\log p}{n})$ and under a permutation beta-min condition (see page 9 in [22]), the penalized maximum likelihood estimator of U corresponds to a DAG that belongs to the data-generating Markov equivalence class.

In the oracle setting in which $\widehat{\Sigma}_n = \Sigma^*$ and $C \to \infty$, the ℓ_0-penalized maximum likelihood approach reduces to

$$\arg\min_{P,U,D} \quad \|U\|_0$$
$$\text{subject to} \quad (\Sigma^*)^{-1} = PUDU^TP^T,$$

which is in fact equivalent to the SP algorithm in the Gaussian setting. This proves that in the oracle setting ℓ_0-penalized maximum likelihood estimation requires strictly weaker conditions for consistency than the faithfulness assumption.
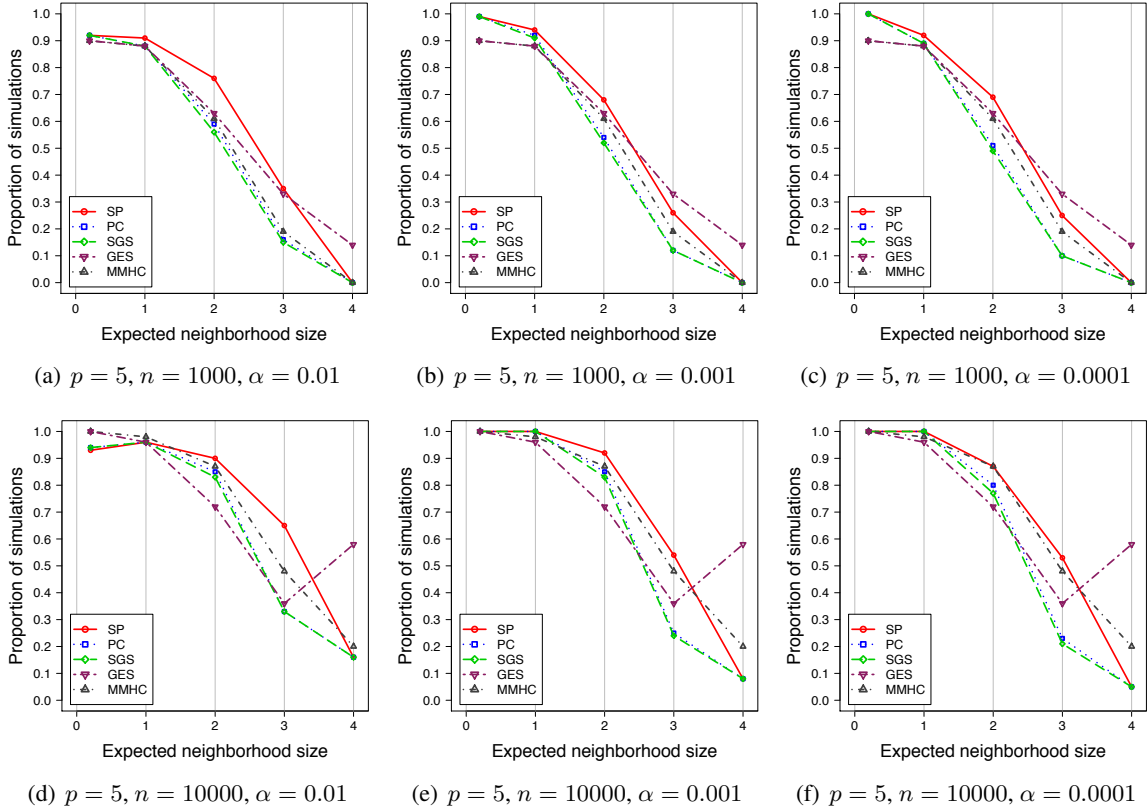

Figure 2: Proportion of simulations out of 100 simulations in which the algorithms recovered the skeleton $S(G^*)$ for 5-node random DAG models.
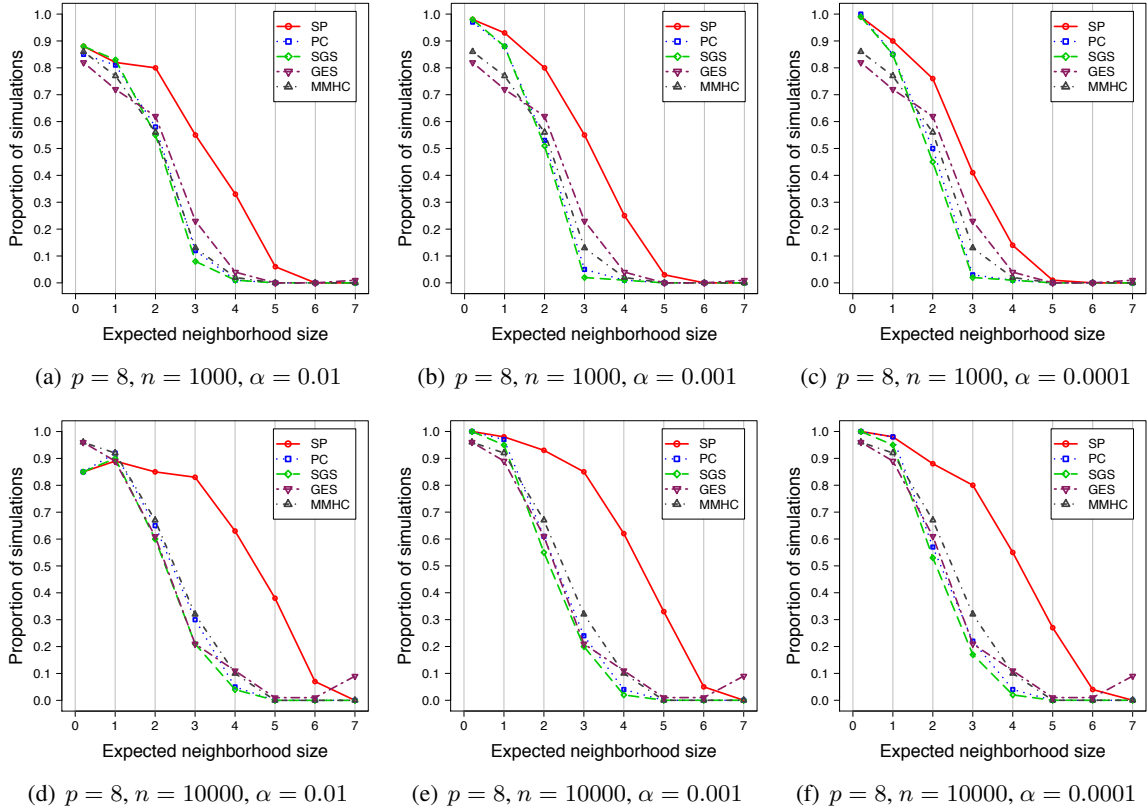
Figure 3: Proportion of simulations out of 100 simulations in which the algorithms recovered the skeleton $S(G^*)$ for 8-node random DAG models.

5   Simulation comparison between SP and other algorithms

In Sections 2.1 and 3.1, we showed that the (strong) SMR assumption is strictly weaker than the (strong) faithfulness assumption. This suggests a higher recovery rate of $\mathcal{M}(G^*)$ for the SP algorithm compared to the SGS and PC algorithms, where the weakest known sufficient condition is the strong faithfulness assumption. In this section, we support our theoretical results with simulations on small Gaussian DAGs of up to 8 nodes in the sample setting where the CI relations are inferred from data. In addition to the constraint-based PC and SGS algorithms, we compare our SP algorithm to the popular score-based GES and hybrid MMHC algorithms. For the PC and SGS algorithms we used the R package 'pcalg' [11], while for the GES and MMHC algorithms, we used the R package 'bnlearn' [17].

The simulation study was conducted as follows: 100 realizations of a p-node random Gaussian DAG model were constructed with $p \in \{3, 4, 5, 6, 7, 8\}$ for different expected neighborhood sizes (i.e. edge probabilities) with edge weights a_{jk} chosen uniformly in $[-1, -0.25] \cup [0.25, 1]$, ensuring the edge weights are bounded away from 0. The considered expected neighborhood sizes range from 0.2 (very sparse DAG) to $p - 1$ (complete DAG). Subsequently, n samples were drawn from the distribution induced by the Gaussian DAG model. We report the results for $n \in \{1000, 10000\}$ and $p \in \{5, 8\}$. The CI relations were estimated for both algorithms using the hypothesis test outlined in Section 3 based on Fisher's z-transform with size $\alpha \in \{0.01, 0.001, 0.0001\}$. These values of α were selected since empirically they were optimal for the

15

(a) $p = 8$, $n = 10000$, $\alpha = 0.01$ (b) $p = 8$, $n = 10000$, $\alpha = 0.001$ (c) $p = 8$, $n = 10000$, $\alpha = 0.0001$
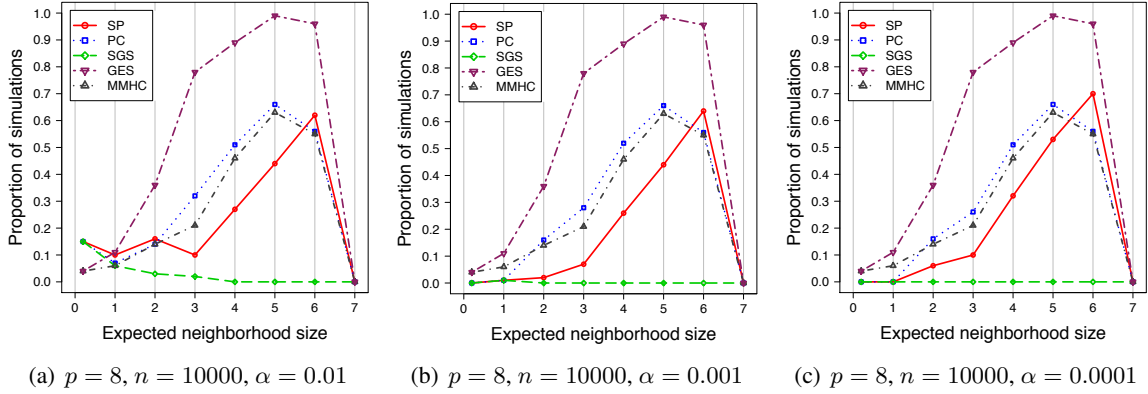
Figure 4: Proportion of simulations out of 100 simulations in which the algorithms outputted a skeleton with additional edges for 8-node random DAG models.

PC and SGS algorithms on the considered examples. This is consistent with the findings in the simulation results in [10].

Figure 2 and Figure 3 display the proportion of simulations out of 100 simulations in which the algorithms recovered the skeleton $S(G^*)$ for different expected neighborhood sizes for 5-node and 8-node random DAG models, respectively. When the skeleton recovered by the SP algorithm was not unique, this was recorded as a failure of the SP algorithm, making the comparison less favorable for the SP algorithm. Figure 4 shows the proportion of simulations out of 100 simulations in which the algorithms outputted a skeleton with additional edges for 8-node random DAG models, whereas Figure 5 shows the proportion of simulations in which the algorithms outputted a skeleton with missing edges. Figures 4 and 5 allow us to analyze the strengths and weaknesses of each algorithm.

The results in Figures 2 and 3 show that the SP algorithm recovers the true skeleton $S(G^*)$ more frequently than the constraint-based SGS and PC algorithms. In addition, Figures 4 and 5 show that this is mainly due to the fact that the SGS and PC algorithms miss edges in the skeleton. This supports our theoretical findings. As expected, the SGS and PC algorithms are very similar in terms of performance, but the PC algorithm tends to include additional edges more often than the SGS algorithm (see Figure 4), while the SGS algorithm tends to miss edges more often than the PC algorithm (see Figure 5). The SP algorithm also outperforms the GES and MMHC algorithms as long as the DAG is not fully connected (i.e. neighborhood size of $p - 1$). The increased performance of GES for fully connected DAGs can be explained by the tendency of GES to include many additional edges as seen in Figure 4. We only show the results for $p = 5$ and $p = 8$, but it is clear from our simulation results on $p \in \{3, 4, 5, 6, 7, 8\}$ nodes that the performance of the SP algorithm increases compared to the performance of the PC, SGS, GES and MMHC algorithms for increasing number of nodes.

Since the focus of this work is on consistency guarantees and not on computation or scalability, a thorough computational comparison has not been conducted. It must be pointed out, however, that using our current implementation of the SP algorithm which exhaustively searches over all permutations, running the SP algorithm requires significantly greater computational resources compared to the other four algorithms. Speeding up the SP algorithm by developing efficient searches through the space of all permutations remains an open and important research direction.

16

(a) $p = 8, n = 10000, \alpha = 0.01$ (b) $p = 8, n = 10000, \alpha = 0.001$ (c) $p = 8, n = 10000, \alpha = 0.0001$
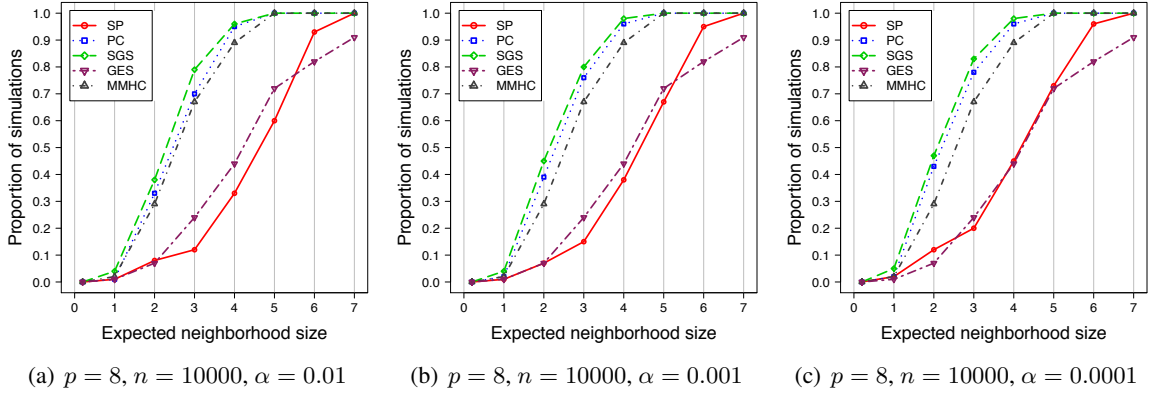
Figure 5: Proportion of simulations out of 100 simulations in which the algorithms outputted a skeleton with missing edges for 8-node random DAG models.

6 Discussion and future work

In this paper, we developed a new method for learning DAGs. The proposed SP algorithm is a score-based method that chooses the permutation which yields the DAG with the fewest number of edges. We proved that the SP algorithm requires strictly weaker conditions than the restricted faithfulness assumption which is weaker than the weakest known sufficient conditions for the PC and SP algorithms. When the variables follow a Gaussian distribution, we showed that in the oracle setting when $n \to \infty$ the SP algorithm is equivalent to finding the sparsest Cholesky factorization of the inverse covariance matrix and the penalized maximum likelihood estimator in [22].

Although the SP algorithm requires checking all $p!$ permutations which is an NP-complete problem, we believe that the connection to sparse Cholesky factorization, for which there are good heuristics, might make our approach computationally feasible also for large graphs. It would then be interesting to define a computationally-efficient version of the SP algorithm also for the high-dimensional setting with imposed sparsity. One would like to understand under which conditions the λ-strong SMR assumption is also necessary for uniform consistency and describe how the parameter λ depends on n and the graph structure.

We proved uniform consistency of the SP algorithm under the λ-strong SMR assumption, which is strictly weaker than the λ-strong faithfulness assumption introduced in [26]. It would be interesting to study how the strong SMR assumption compares to the β-min condition in [22], which is sufficient for the penalized maximum likelihood approach, and also how it compares to consistency conditions for the GES and MMHC algorithms. One approach is to find a geometric description of the strong-SMR assumption to bound the proportion of distributions that satisfy this assumption, similarly as for the strong-faithfulness assumption in [21]. However, this would be more difficult than in [21], since it would have to involve also a combinatorial argument.

7 Proofs

Proof of Lemma 2.1 We first prove that G_π satisfies the Markov assumption: Let E_π denote the edge set of G_π and let $\pi(i), \pi(j) \in V$ such that $\pi(j) \in nd(\pi(i)) \setminus pa(\pi(i))$. We need to prove that $X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_{pa(\pi(i))}$. First note that since $\pi(j)$ is a non-descendant of $\pi(i)$, the two vertices are not adjacent in G_π and

hence by the definition of G_π,

$$X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_S \text{ where } S = \{\pi(1), \ldots, \pi(\max(i,j))\} \setminus \{\pi(i), \pi(j)\}. \tag{4}$$

We define $D = \{\pi(1), \ldots, \pi(i-1)\} \setminus (pa(\pi(i)) \cup \{\pi(j)\})$ and show that $X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_{D \cup pa(\pi(i))}$. This is immediate if $j < i$. If $i < j$ this requires a short argument: Since $\pi(j)$ is a non-descendant of $\pi(i)$ at least one of the two following CI relations must hold:

(i) $X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j-1)} \mid X_S$, where $S = \{\pi(1), \ldots, \pi(i-1)\} \cup \{\pi(i+1), \ldots, \pi(j-2)\}$,

(ii) $X_{\pi(j)} \perp\!\!\!\perp X_{\pi(j-1)} \mid X_S$, where $S = \{\pi(1), \ldots, \pi(i-1)\} \cup \{\pi(i)\} \cup \{\pi(i+1), \ldots, \pi(j-2)\}$.

In case (i) it follows from the contraction axiom for conditional independencies (see, e.g., page 62 in [5]) and the CI relation (4) that

$$X_{\pi(i)} \perp\!\!\!\perp (X_{\pi(j-1)}, X_{\pi(j)}) \mid X_S, \text{ where } S = \{\pi(1), \ldots, \pi(i-1)\} \cup \{\pi(i+1), \ldots, \pi(j-2)\}.$$

In case (ii) it follows from the intersection axiom for conditional independencies (see, e.g., page 63 in [5]) and the CI relation (4) that

$$X_{\pi(j)} \perp\!\!\!\perp (X_{\pi(i)}, X_{\pi(j-1)}) \mid X_S, \text{ where } S = \{\pi(1), \ldots, \pi(i-1)\} \cup \{\pi(i+1), \ldots, \pi(j-2)\}.$$

In both cases it follows that

$$X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_S, \text{ where } S = \{\pi(1), \ldots, \pi(i-1)\} \cup \{\pi(i+1), \ldots, \pi(j-2)\},$$

and by induction we get that

$$X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_S, \text{ where } S = \{\pi(1), \ldots, \pi(i-1)\},$$

which corresponds to

$$X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_{D \cup pa(\pi(i))}. \tag{5}$$

It is important to note that all nodes in D are non-adjacent to $\pi(i)$, since otherwise they would be parents of $\pi(i)$. Hence by the definition of G_π we get that

$$X_{\pi(i)} \perp\!\!\!\perp X_d \mid X_{D \setminus \{d\} \cup pa(\pi(i))} \text{ for all } d \in D.$$

Let $d_1, d_2 \in D$, then

$$X_{\pi(i)} \perp\!\!\!\perp X_{d_1} \mid X_{D \setminus \{d_1, d_2\} \cup pa(\pi(i)) \cup \{d_2\}} \text{ and } X_{\pi(i)} \perp\!\!\!\perp X_{d_2} \mid X_{D \setminus \{d_1, d_2\} \cup pa(\pi(i)) \cup \{d_1\}}.$$

By the intersection axiom it follows that

$$X_{\pi(i)} \perp\!\!\!\perp (X_{d_1}, X_{d_2}) \mid X_{D \setminus \{d_1, d_2\} \cup pa(\pi(i))}$$

and by induction

$$X_{\pi(i)} \perp\!\!\!\perp (X_D) \mid X_{pa(\pi(i))}. \tag{6}$$

By applying one last time the contraction axiom to the CI relations (5) and (6) we get that $X_{\pi(i)} \perp\!\!\!\perp X_{\pi(j)} \mid X_{pa(\pi(i))}$.

In order to prove minimality, assume for contradiction that there exists a sparser DAG $G'_\pi \subset G_\pi$ with an edge $(\pi(i), \pi(j)) \in E_\pi$ but $(\pi(i), \pi(j)) \notin E'_\pi$ which satisfies the Markov assumption. Since $(\pi(i), \pi(j)) \in E_\pi$ it follows from the definition of G_π that

$$X_{\pi(i)} \not\perp\!\!\!\perp X_{\pi(j)} \mid X_S \text{ where } S = \{\pi(1), \ldots, \pi(j-1)\} \setminus \{\pi(i)\}.$$

But this contradicts the Markov assumption since $\pi(i)$ is d-separated from $\pi(j)$ given S in G'_π. ∎

Proof of Lemma 2.2 (a) Let π^* denote an ordering consistent with G^*, let E^* denote the edges in G^* and let G_{π^*} denote the DAG resulting from applying the SP algorithm to the ordering π^*. Note that two nodes $\pi^*(i)$ and $\pi^*(j)$ with $i < j$ and $(\pi^*(i), \pi^*(j)) \notin E^*$ are d-separated in G^* by the set $S = \{\pi^*(1), \ldots, \pi^*(j-1)\} \setminus \{\pi^*(i)\}$. Then, as a consequence of the Markov assumption, the distribution $\mathbb{P}$ satisfies $X_{\pi^*(i)} \perp\!\!\!\perp X_{\pi^*(j)} \mid X_S$ and hence, $(\pi^*(i), \pi^*(j)) \notin E_{\pi^*}$. This shows that $|G_{\pi^*}| \leq |G^*|$ and together with $|G_{SP}| \leq |G_{\pi^*}|$ completes the proof.

(b) Under adjacency faithfulness $S(G_{SP}) = S(G^*)$. So to show equality of the skeletons it suffices to prove that $S(G^*) \subset S(G_\pi)$ for all permutations π. Under adjacency faithfulness, $X_j \not\perp\!\!\!\perp X_k \mid X_S$ for all edges $(X_j, X_k) \in E^*$ and for all subsets $S \subset \{1, 2, \ldots, p\} \setminus \{j, k\}$. Then by the definition of G_π any edge in G^* must be an edge in G_π and hence $S(G^*) \subset S(G_\pi)$ for all permutations π.

(c) Since the restricted faithfulness assumption entails the adjacency faithfulness assumption, $S(G_{SP}) = S(G^*)$. In the following we prove that $G_{SP} \in \mathcal{M}(G^*)$ as a consequence of the orientation faithfulness assumption. Let (i, j, k) be an unshielded triple in G^*. We consider two cases, first, when (i, j, k) is a v-structure. Then $X_i \not\perp\!\!\!\perp X_j \mid X_S$ for all $S \subset V \setminus \{i, j\}$ such that $k \in S$. Let π^* be a permutation giving rise to G_{SP} and let $a, b, c \in V$ such that $\pi^*(a) = i$, $\pi^*(b) = j$, $\pi^*(c) = k$. Note that $(\pi^*(a), \pi^*(b)) \notin E^*$ and hence it is also not an edge in G_{SP}. As a consequence $X_{\pi^*(a)} \perp\!\!\!\perp X_{\pi^*(b)} \mid X_S$ where $S = \{\pi^*(1), \ldots, \pi^*(\max(a, b))\} \setminus \{\min(a, b)\}$ and therefore $c > \max(a, b)$. Hence (i, j, k) is also a v-structure in G_{SP}. The second case when (i, j, k) is not a v-structure is analogous. ∎

Proof of Theorem 2.7 Let π^* denote an ordering consistent with G^*, let E^* denote the edges in G^* and let G_{π^*} denote the DAG resulting from applying the SP algorithm to the ordering π^*. We first show that $S(G_{\pi^*}) = S(G^*)$. As a consequence of the single-path-faithfulness assumption, we can partition the CI relations satisfied by $\mathbb{P}$ into two classes, namely, CI relations corresponding to d-separation in G^* and unfaithful CI relations resulting from path cancellations. As seen in the proof of Lemma 2.2 (a), all CI relations in $\mathbb{P}$ emerging from d-separation (corresponding to non-edges in G^*), also get omitted by the SP algorithm. We now show that all additional unfaithful CI relations in $\mathbb{P}$ emerging through path cancellations do not lead to further edge deletions and hence $S(G_{\pi^*}) = S(G^*)$: Every path between two nodes $\pi^*(i)$ and $\pi^*(j)$ goes either through a non-collider $b <_{\pi^*} \pi^*(j)$ or a collider $c >_{\pi^*} \pi^*(j)$. Conditioning on b blocks the path through b, while not conditioning on c blocks the path through c. As a consequence, none of the CI relations emerging from path cancellations are of the form $X_{\pi^*(i)} \perp\!\!\!\perp X_{\pi^*(j)} \mid X_S$, where $S = \{\pi^*(1), \ldots, \pi^*(j-1)\} \setminus \{\pi^*(i)\}$ and hence $S(G_{\pi^*}) = S(G^*)$.

We now prove that applying the SP algorithm to any other permutation π results in a DAG G_π with $|G_\pi| \geq |G^*|$. Similarly as in the previous paragraph, we partition the CI relations in $\mathbb{P}$ into two classes: CI relations corresponding to d-separation in the graph G^* (with ordering π^*) and CI relations emerging from violations of adjacency faithfulness corresponding to path cancellations. In the previous paragraph, we have shown that for the ordering π^* only the CI relations corresponding to d-separation are active (i.e. lead to the omission of an edge in the SP algorithm). We will now show that for every activation of a CI relation

corresponding to a violation of adjacency faithfulness at least one CI relation corresponding to d-separation gets de-activated. As a consequence, for every edge $e \in E^*$ which gets deleted from E_π, at least one edge $\tilde{e} \notin E^*$ gets added to E_π, resulting in $|G_\pi| \geq |G^*|$.

Let $X_{\pi^*(i)} \perp\!\!\!\perp X_{\pi^*(j)} \mid X_S$ with $i < j$ denote a CI relation corresponding to a violation of adjacency faithfulness. Hence $(\pi^*(i), \pi^*(j)) \in E^*$ and, as a consequence of triangle-faithfulness, $(\pi^*(i), \pi^*(j))$ is not part of any triangle in G^*. There are two cases, either the edge $(\pi^*(i), \pi^*(j))$ cancels out with (case 1) a path going through $k <_{\pi^*} \pi^*(j)$ for all nodes k on the path, or (case 2) a path going through a collider $k >_{\pi^*} \pi^*(j)$.

Case 1: Let K denote the set of nodes on the considered path between $\pi^*(i)$ and $\pi^*(j)$. Since $(\pi^*(i), \pi^*(j))$ is not part of any triangle, $|K| \geq 2$ and for all $k \in K$ either $(k, \pi^*(i)) \notin S(G^*)$ or $(k, \pi^*(j)) \notin S(G^*)$. In order for the SP algorithm to omit the edge $(\pi^*(i), \pi^*(j))$, the ordering $<_\pi$ must satisfy $\max(\pi^*(i), \pi^*(j)) <_\pi k$ for all $k \in K$. Let $\tilde{k}$ denote the minimal element in K with respect to $<_\pi$. Without loss of generality we can assume that $(\tilde{k}, \pi^*(i)) \notin S(G^*)$ (otherwise $(\tilde{k}, \pi^*(j)) \notin S(G^*)$). Hence there exists a subset $S \subset V \setminus \{\tilde{k}, \pi^*(i)\}$ that d-separates $\tilde{k}$ from $\pi^*(i)$ given S in the DAG G^*. Note that S necessarily contains a node $\tilde{l} \in K \setminus \{\tilde{k}\}$. This d-separation leads to the CI relation $X_{\tilde{k}} \perp\!\!\!\perp X_{\pi^*(i)} \mid X_S$. Since $\tilde{k} <_\pi \tilde{l}$ and $\pi^*(i) <_\pi \tilde{l}$, but $\tilde{l} \in S$, this CI relation does not hold for the permutation π and hence $(\tilde{k}, \pi^*(i)) \in S(G_\pi)$.

Case 2: Let C denote the set of colliders on the considered path and $\bar{C}$ the set of non-colliders. Since $(\pi^*(i), \pi^*(j))$ is not part of any triangle, $\bar{C} \neq \emptyset$ and for all $c \in C$ either $(\pi^*(i), c) \notin S(G^*)$ or $(\pi^*(j), c) \notin S(G^*)$. In order for the SP algorithm to omit the edge $(\pi^*(i), \pi^*(j))$, the ordering $<_\pi$ must satisfy $c <_\pi \max(\pi^*(i), \pi^*(j)) <_\pi k$ for all $c \in C$ and all $k \in \bar{C}$. Let $\tilde{c}$ denote the maximal element in C with respect to $<_\pi$. Without loss of generality we assume that $(\tilde{c}, \pi^*(i)) \notin S(G^*)$. Hence there exists a subset $S \subset V \setminus \{\tilde{c}, \pi^*(i)\}$ that d-separates $\tilde{c}$ from $\pi^*(i)$ given S in the DAG G^*. Note that S necessarily contains a node $\tilde{l} \in \bar{C}$. This d-separation leads to the CI relation $X_{\tilde{c}} \perp\!\!\!\perp X_{\pi^*(i)} \mid X_S$. Since $\tilde{c} <_\pi \tilde{l}$ and $\pi^*(i) <_\pi \tilde{l}$, but $\tilde{l} \in S$, this CI relation does not hold for the permutation π and hence $(\tilde{c}, \pi^*(i)) \in S(G_\pi)$. ∎

Acknowledgements

References

[1] D. M. Chickering. Learning Bayesian networks is NP-complete. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, 1995.

[2] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[3] T. A. Davis. *Direct methods for sparse linear systems*. SIAM, Philadelphia, 2006.

[4] M. Drton and M. D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22:430–449, 2007.

[5] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*, volume 40 of *Oberwolfach Seminars*. Birkhäuser, Basel, 2009.

[6] R. A. Fisher. Frequency distribution of the values of the correlation coefficient samples of an indefinitely large population. *Biometrika*, 10:507–521, 1915.

[7] N. Friedman, I. Nachman, and D. Pe'er. Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI 1999)*, pages 196–205, 1999.

[8] A. George and J. Liu. The evolution of the minimum degree ordering algorithm. *SIAM Review*, 31:1–19, 1989.

[9] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Strucure*. Academic Press, 1987.

[10] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[11] M. Kalisch, M. Mächler, D. Colombo, M.H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47:1–26, 2012.

[12] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

[13] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[14] M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterization. *Biometrika*, 86:677–690, 1999.

[15] J. Ramsey, J. Zhang, and P. Spirtes. Adjancency-faithfulness and conservative causal inference. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 401–408, 2006.

[16] J. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

[17] M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35:1–22, 2010.

[18] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2000.

[19] P. Spirtes and J. Zhang. A uniformly consistent estimator of causal effects under the k-triangle faithfulness assumption. *Unpublished manuscript*, 2012.

[20] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. In *Machine Learning*, 2006.

[21] C. Uhler, G. Raskutti, P. Buhlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *Annals of Statistics*, 41:436–463, 2013.

[22] S. van de Geer and P. Bühlmann. Penalized maximum likelihood estimation for sparse directed acyclic graphs. *Annals of Statistics*, 41:536–567, 2013.

[23] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, pages 220–227, 1990.

[24] M. Yannakakis. Computing the minimum fill-in is NP-complete. *SIAM Journal of Algebraic Discrete Methods*, 2:77–79, 1981.

[25] J. Zhang. A comparison of three Occam's razors for Markovian causal models. *British Journal for the Philosophy of Science*, 64:423–448, 2013.

[26] J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, pages 632–639, 2005.

[27] J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.