# Reliable Causal Discovery with Improved Exact Search and Weaker Assumptions

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Many of the causal discovery methods rely on the faithfulness assumption to guarantee asymptotic correctness. However, the assumption can be (approximately) violated in many ways, leading to sub-optimal solutions. Although there is a line of research in Bayesian network learning focusing on weakening the assumption, such as exact search methods with well-defined score functions, they do not scale well to large graphs. To improve the scalability of exact methods, we restrict the search space via learning a sound super-structure, and perform local exact search on the individual variables with their neighbors within two hops in the super-structure. We show theoretically that the proposed framework produces asymptotically correct results under assumptions strictly weaker than faithfulness. Numerical experiments validate the efficacy of our framework, and demonstrate that it scales up to hundreds of nodes with a high accuracy.

## 1 Introduction

Although it is often more reliable to discover causal relationships by making use of interventions or randomized experiments, they are practically challenging or even prohibited owing to ethical considerations. Thus, causal discovery from observational data has received considerable attention in recent decades, and has been widely applied in different fields such as genetics (Peters et al., 2017) and Earth system sciences (Runge et al., 2019).

One major class of causal discovery methods is the constraint-based methods, such as PC (Spirtes and Glymour, 1991) and FCI (Spirtes et al., 1995; Colombo et al., 2011), that first estimate the skeleton of the causal structure and then identify the v-structures based on conditional independence tests. Asymptotically speaking, these methods are guaranteed to return the true Markov equivalence class (MEC) under the Markov and faithfulness assumptions. Several modifications (Ramsey et al., 2006; Spirtes and Zhang, 2014) to these constraint-based methods have been developed to allow certain types of unfaithfulness. However, the resulting methods generally make weaker claims and therefore are not guaranteed to estimate the true MEC.

Another popular approach is the GES (Chickering, 2002) algorithm that searches in the space of MECs greedily by maximizing a well-defined score, such as the Bayesian information criterion (BIC) (Schwarz, 1978) score. It starts with an empty structure and consists of two phases: (1) adding edges until a local maximum is found, and (2) removing edges until a local maximum is reached. In spite of the greedy strategy, GES converges in the large sample limit to the true MEC under the Markov and faithfulness assumptions, similar to the aforementioned constraint-based methods.

Recently, NOTEARS (Zheng et al., 2018) casts the Bayesian network learning task into a continuous constrained optimization problem with the least squares objective, using an algebraic charactezation of directed acyclic graph (DAG). Subsequent work GOLEM (Ng et al., 2020) adopts a continuous unconstrained optimization formulation with a likelihood-based objective. Although these methods focus on learning Bayesian networks, their output can be converted to the MEC for causal inter-

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

pretation, under suitable assumptions (Pearl, 1988; Spirtes et al., 2001). These methods enable the application of numerical solvers and GPU acceleration, which thus are scalable to large graphs. However, they are only guaranteed to find a local optimum of the optimization problem. Therefore, even with the faithfulness assumption, the asymptotic property of these methods remains unclear.

Another line of research focuses on weakening the faithfulness assumption required for asymptotic correctness of the search results, since it has been shown that given finite samples, approximate violations of faithfulness occur surprisingly often, especially when there is a large number of variables (Uhler et al., 2013). For instance, exact search methods find the optimal Bayesian network based on a predefined score function, such as dynamic programming (DP) (Koivisto and Sood, 2004; Ott et al., 2004; Singh and Moore, 2005; Silander and Myllymäki, 2006) and A* (Yuan and Malone, 2013). The DAG estimated by these methods can be converted to the MEC for causal interpretation (Spirtes and Zhang, 2018), and has been shown by Lu et al. (2021) to be asymptotically correct under an assumption that is strictly weaker than faithfulness. It is worth noting that the recently developed approaches based on Boolean satisfiability (SAT) solver (Hyttinen et al., 2013, 2014) and sparsest permutation (Raskutti and Uhler, 2018) can be viewed as instances of exact score-based methods.

Due to the large search space of possible DAGs (Chickering, 1996; He et al., 2015), exact search methods are feasible only for small structures. Therefore, *super-structure* has been adopted to constraint the search space (Perrier et al., 2008; Tsamardinos et al., 2006), which is defined to be an undirected graph that restricts the search to DAGs whose skeleton is its subgraph. However, most of these methods rely on discovering the skeleton of the true DAG for use as super-structure, utilizing estimation methods like MMPC (Tsamardinos et al., 2006), which require the faithfulness assumption to be asymptotically correct. Under (approximate) violations of faithfulness, these skeleton estimation methods may miss some edges owing to unfaithful conditional independencies in the data distribution; thus, further exact search procedures are guaranteed to miss those edges.

**Contributions.** In this work, we develop methods to estimate super-structure and scale up score-based exact search with weaker assumptions than faithfulness, giving rise to a more reliable causal discovery procedure. Our main contributions can be summarized as follows:

- We develop a sound super-structure estimation method based on the support of inverse covariance matrix of the data distribution, and show that it is asymptotically correct under assumptions strictly weaker than faithfulness (or, more specifically, than triangle faithfulness). This can be combined with an exact search method like DP or A* to reduce search space.
- To further scale up exact search, we develop a local search strategy, called Local A*, on the individual variables with their neighbors within two hops in the support of estimated inverse covariance matrix, and prove its consistency under assumptions weaker than faithfulness.
- We demonstrate the effectiveness of our super-structure estimation method and local search strategy by conducting extensive experiments, and show that it scales up to hundreds of nodes with a high accuracy.

**Organization of the paper.** We review the common assumptions made for causal discovery and the linear structural equation model (SEM) in Section 2. In Section 3, we establish weaker variants of the faithfulness assumption and show how they could be used to learn a sound super-structure. We then formulate an improved exact search strategy in Section 4. The empirical studies in Section 5 validate our theoretical results and the effectiveness of the proposed method. Conclusion and future work are given in Section 6. All proofs are available in Appendix C.

## 2 Background

We first review the concepts of causal Bayesian networks and some commonly used assumptions that are related to our further analysis. We then give a brief overview of the linear SEM.

### 2.1 Causal Bayesian Network and Common Assumptions

Let $G = (\boldsymbol{V}, \boldsymbol{E})$ be a DAG with vertex set $\boldsymbol{V} = \{X_1, \dots, X_d\}$ in which each node $X_i$ corresponds to a random variable. Denote $\boldsymbol{X} = (X_1, \dots, X_d)$ as the random vector concatenating all the variables and its associated probability distribution $\mathbb{P}$. Let $\boldsymbol{X}_{pa(i)}$ be set of parental nodes of $X_i$ in $G$ such that there is a directed edge from $X_j \in \boldsymbol{X}_{pa(i)}$ to $X_i$, or $X_j \to X_i \in \boldsymbol{E}$. We assume *causal sufficiency*

throughout the paper, i.e., no hidden variables. In a *Bayesian network*, the distribution $\mathbb{P}$ is assumed to be Markov w.r.t. to the DAG $G$ via the following definition.

**Assumption 1 (Markov).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, every variable $X$ in $\boldsymbol{V}$ is probabilistically independent of its non-descendants given its parents in $G$.*

A *causal Bayesian network* can be viewed as a Bayesian network where the directed edges are provided a causal meaning, which thereby allows it to answer interventional queries (Koller and Friedman, 2009). In general, there are many DAGs that induce the same conditional independence (CI) relations with the distribution $\mathbb{P}$, and are said to be *Markov equivalent*. The *Markov equivalence class* (MEC) consists of all DAGs that entail the same CI relations as $G$ does, and is uniquely determined by its skeleton and v-structures (Pearl, 2009). *V-structure* is defined to be a collider $X \to Y \leftarrow Z$ where $X$ and $Z$ are not adjacent in $G$, therefore referred to as an *unshielded collider*. If $X$ and $Z$ are adjacent, then it is called a *shielded collider*.

The following faithfulness assumption is commonly used to relate the CI relations in the distribution to the DAG, and can be thought of as the converse to the Markov assumption.

**Assumption 2 (Faithfulness (Spirtes et al., 2001)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, $\mathbb{P}$ implies no CI relations not already entailed by the Markov assumption.*

Under the Markov and faithfulness assumptions, constraint-based methods such as PC have been shown to asymptotically output the correct MEC. However, in the finite sample regime, the faithfulness assumption is sensitive to statistical testing errors when inferring the CI relations, and its (approximate) violations occur surprisingly often, especially when there is a large number of variables (Uhler et al., 2013). Therefore, different relaxations of faithfulness have been proposed, such as the adjacency-faithfulness, orientation-faithfulness, and triangle-faithfulness assumptions, which we review in Appendix A due to space limit.

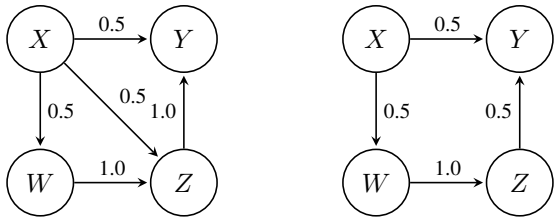
Clearly, the triangle-faithfulness assumption is a consequence of adjacency-faithfulness. Based on these weaker assumptions, different modifications to constraint-based methods have been proposed, such as Conservative PC (Ramsey et al., 2006) and Very Conservative SGS (Spirtes and Zhang, 2014). As their names suggest, these methods make weaker claims about the estimated causal structure and therefore are not guaranteed to estimate the true MEC. On the other hand, Raskutti and Uhler (2018) proposed the following assumption and show that it is strictly weaker than faithfulness.

**Assumption 3 (Sparsest Markov representation (SMR) (Raskutti and Uhler, 2018)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, the MEC of $G$ is the unique sparsest MEC that satisfies the Markov assumption with $\mathbb{P}$.*

Forster et al. (2017) referred to the above assumption as the (unique) frugality assumption, and argued that it has multiple desirable properties compared to the faithfulness assumption. Under the SMR assumption, exact score-based search methods with well-defined score functions have been shown to be asymptotically correct (Lu et al., 2021), of which the sparsest permutation method (Raskutti and Uhler, 2018) and SAT solver are special cases. The study by Lu et al. (2021) demonstrated that causal discovery methods that require the faithfulness assumption produce sub-optimal solutions in various cases, whereas exact methods based on SMR are able to output the right MEC. Raskutti and Uhler (2018) conjectured that the latter line of methods reaches the information-theoretic limit in the sense that the SMR assumption may be the weakest assumption that guarantees asymptotic correctness of any algorithm for learning the true MEC. Therefore, in this work, we aim at improving the scalability of exact score-based methods for reliable causal discovery, by relying on the SMR assumption and relaxing the faithfulness assumption as much as is viable.

### 2.2 Linear Structural Equation Model

Given a linear SEM, each random variable follows the relationship $X_i = \boldsymbol{B}_i^\mathsf{T} \boldsymbol{X} + \boldsymbol{N}_i$, where $\boldsymbol{B}_i$ is a coefficient vector and $\boldsymbol{N}_i$ is an exogenous noise variable associated with variable $X_i$. In this work we focus on the linear Gaussian model where the variables $N_i$'s follow the Gaussian distribution. The linear SEM can be written in matrix form as $\boldsymbol{X} = \boldsymbol{B}^\mathsf{T}\boldsymbol{X} + \boldsymbol{N}$, where $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \cdots, \boldsymbol{B}_d]$ corresponds to a weighted adjacency matrix, and $\boldsymbol{N} = (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_d)$ is a noise vector characterized by the covariance matrix $\boldsymbol{\Omega} = \mathrm{cov}[\boldsymbol{N}]$. The nonzero coefficients in $\boldsymbol{B}$ define the structure of $G$, i.e., $X_j \to X_i \in \boldsymbol{E}$ if and only if the coefficient in $\boldsymbol{B}_i$ corresponding to variable $X_j$ is nonzero. Since one can always center the data, we assume $\mathrm{E}[\boldsymbol{X}] = \mathrm{E}[\boldsymbol{N}] = 0$ without loss of generality. The inverse covariance matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ of $\boldsymbol{X}$ is given by $\boldsymbol{\Theta} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{B})^\mathsf{T}$. Note that $\boldsymbol{\Theta}_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j | \boldsymbol{V} \setminus \{X_i, X_j\}$ in the linear Gaussian case.




(a) Violation of SCF. (b) Violation of UCF.

Figure 1: An example illustrating the violations of the SCF and the UCF assumptions, respectively, but the SSCF and SUCF assumptions hold. In both cases, $X \perp\!\!\!\perp Z|Y$, $X \not\perp\!\!\!\perp Z|\{Y, W\}$.

# 3 Weaker Assumptions for Learning Super-Structure

In this section, we establish several weaker variants of the faithfulness assumption, and show how they are sufficient for learning a sound super-structure of the true DAG via inverse covariance estimation.

## 3.1 Weaker Assumptions than Faithfulness

In the following, we describe several relaxed assumptions of faithfulness that are sufficient conditions for our super-structure estimation procedure. We first start with the specific types of faithfulness assumption that are related to different kinds of colliders.

**Assumption 4 (Shielded-collider-faithfulness (SCF)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, let $X \to Y \leftarrow Z$ be any shielded collider in $G$. Then $X$ and $Z$ are dependent conditional on any subset of $\boldsymbol{V} \setminus \{X, Z\}$ that contains $Y$.*

**Assumption 5 (Unshielded-collider-faithfulness (UCF)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, let $X \to Y \leftarrow Z$ be any unshielded collider in $G$. Then $X$ and $Z$ are dependent conditional on any subset of $\boldsymbol{V} \setminus \{X, Z\}$ that contains $Y$.*

The above two assumptions are restrictions of the triangle-faithfulness (Zhang and Spirtes, 2008) and the orientation-faithfulness (Ramsey et al., 2006) assumptions, respectively, to collider structures. Note that they differ only in the type of collider being considered. However, these assumptions require dependence conditioning on any subset of $\boldsymbol{V} \setminus \{X, Z\}$, which may be restrictive in practice. In this work, we further relax them to require only dependence conditioning on $\boldsymbol{V} \setminus \{X, Z\}$, and will show in Section 3.2 that they are sufficient conditions for estimating a sound-super structure of the underlying DAG. An example is illustrated in Figure 1, where the unfaithful CI $X \perp\!\!\!\perp Z|Y$ is constructed via path cancellations. This is formally stated in the following assumptions.

**Assumption 6 (Single shielded-collider-faithfulness (SSCF)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, let $X \to Y \leftarrow Z$ be any shielded collider in $G$. Then $X$ and $Z$ are dependent conditional on $\boldsymbol{V} \setminus \{X, Z\}$.*

**Assumption 7 (Single unshielded-collider-faithfulness (SUCF)).** *Given a DAG $G$ and distribution $\mathbb{P}$ over the variable set $\boldsymbol{V}$, let $X \to Y \leftarrow Z$ be any unshielded collider in $G$. Then $X$ and $Z$ are dependent conditional on $\boldsymbol{V} \setminus \{X, Z\}$.*

Given a distribution $\mathbb{P}$, let $\boldsymbol{G}_{\text{fai}}(\mathbb{P})$, $\boldsymbol{G}_{\text{adj}}(\mathbb{P})$, $\boldsymbol{G}_{\text{ori}}(\mathbb{P})$, $\boldsymbol{G}_{\text{tri}}(\mathbb{P})$, $\boldsymbol{G}_{\text{scf}}(\mathbb{P})$, $\boldsymbol{G}_{\text{ucf}}(\mathbb{P})$, $\boldsymbol{G}_{\text{sscf}}(\mathbb{P})$, and $\boldsymbol{G}_{\text{sucf}}(\mathbb{P})$ be the set of DAGs satisfying the faithfulness, adjacency-faithfulness, orientation-faithfulness, triangle-faithfulness, SCF, UCF, SSCF, and SUCF assumptions, respectively. Denote by $\subset$ the proper subset symbol. Combining with the results by Ramsey et al. (2006); Zhang and Spirtes (2008), we have the following nesting properties, showing that SSCF and SUCF are intuitively much weaker than the faithfulness assumption.

**Remark 1.** $\boldsymbol{G}_{sscf}(\mathbb{P}) \subset \boldsymbol{G}_{scf}(\mathbb{P}) \subset \boldsymbol{G}_{tri}(\mathbb{P}) \subset \boldsymbol{G}_{adj}(\mathbb{P}) \subset \boldsymbol{G}_{fai}(\mathbb{P})$.

**Remark 2.** $\boldsymbol{G}_{sucf}(\mathbb{P}) \subset \boldsymbol{G}_{ucf}(\mathbb{P}) \subset \boldsymbol{G}_{ori}(\mathbb{P}) \subset \boldsymbol{G}_{fai}(\mathbb{P})$.

## 3.2 Inverse Covariance Estimation for Learning Super-Structure

Based on the assumptions described in Section 3.1, we develop a super-structure estimation method via inverse covariance estimation. We first study the specific assumptions required for the support of inverse covariance, denoted as $\text{supp}(\boldsymbol{\Theta})$, to be the same as the moralized graph of the underlying DAG. This is similar to the analysis by Loh and Bühlmann (2014), but we focus on the assumptions from a causal discovery perspective, to shed light on how weak they are as compared to faithfulness.

Then we show that one of the assumptions may be approximately violated in practice, but can be dropped while still guaranteeing the structure defined by $\mathrm{supp}(\mathbf{\Theta})$ to be a sound super-structure of the true DAG. The former case is summarized in the following theorem.

**Theorem 1.** *Given a DAG $G$ and distribution $\mathbb{P}$ that follow a linear Gaussian model with inverse covariance matrix $\mathbf{\Theta}$, under Markov assumption, the SSCF and SUCF assumptions are satisfied if and only if the structure defined by $\mathrm{supp}(\mathbf{\Theta})$ is the same as the moralized graph of the true DAG $G$.*

Although the above theorem guarantees to recover the moralized graph via $\mathrm{supp}(\mathbf{\Theta})$, in practice the SUCF assumption tends to be approximately violated. To illustrate, we simulate the linear Gaussian model and compute the minimum values $\min_{i,j}\{|\mathbf{\Theta}_{ij}| : (X_i, X_j) \text{ is a pair of neighbors in } G\}$ and $\min_{i,j}\{|\mathbf{\Theta}_{ij}| : (X_i, X_j) \text{ is a pair of spouses in } G\}$ over 100 simulations, by considering different expected degrees and number of nodes $d \in \{10, 20, 50, 100, 200, 500, 1000\}$. Note that $(X_i, X_j)$ being a pair of spouses in $G$ implies that it has an unshielded collider, and SUCF requires that $X_i \not\perp\!\!\!\perp X_j | \boldsymbol{V} \setminus \{X_i, X_j\}$, i.e., $\mathbf{\Theta}_{ij} \neq 0$ in the linear Gaussian case. The results for expected degree of 2 is shown in Figure 2, while the visualizations for degrees of 5 and 8 can be be found in Appendix B. Although the values in both cases decrease with larger graphs,[1] it is clear that the minimum values corresponding to the spouses are significantly smaller than those of neighbors, where the difference grows in the number of nodes. For instance, on graphs with the degree of 2 and 1000 nodes, the average minimum value of the former case is 0.025, while that in the latter case is close to 0.1. With finite samples, it is very challenging to discover those undirected edges in the former case, since the weights are very close to zero, leading to approximate violations of SUCF especially when there is a large number of variables. Similar to the strong faithfulness assumption (Zhang and Spirtes, 2003), in the future we plan to examine this observation via the strong SUCF assumption, to rule out the possibility of weak edges in $\mathrm{supp}(\mathbf{\Theta})$ that correspond to pairs of spouses.

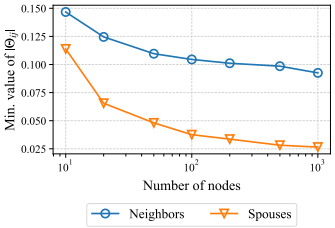



Figure 2: Expected degree of 2.

If we drop the SUCF assumption, the structure defined by $\mathrm{supp}(\mathbf{\Theta})$ may miss some edges as compared to the moralized graph of the true DAG. Fortunately, these edges correspond to the spouses in the DAG, as previously described. That is, $\mathrm{supp}(\mathbf{\Theta})$ is still guaranteed to contain all undirected edges corresponding to neighbors in the DAG, as long as SSCF (in addition to the Markov assumption) holds, indicating that it is a sound super-structure of the true DAG. Based on the simulations in Figure 2, approximate violations of SSCF are much less likely to happen than those of SUCF.

**Theorem 2.** *Given a DAG $G$ and distribution $\mathbb{P}$ that follow a linear Gaussian model with inverse covariance matrix $\mathbf{\Theta}$, under Markov assumption, the SSCF assumption is satisfied if and only if the structure defined by $\mathrm{supp}(\mathbf{\Theta})$ is a super-structure of the true DAG $G$.*

With an asymptotically correct approach to estimate the support of inverse covariance matrix, the above theorem implies that one can asymptotically estimate a sound super-structure of the ground truth DAG under the SSCF assumption that is strictly weaker than triangle-faithfulness (cf. Remark 1), an assumption which is intuitively much weaker than faithfulness. This may be considered as a significant improvement over the existing methods that rely on estimating the skeleton of the underlying DAG for use as super-structure, utilizing estimation methods like MMPC (Tsamardinos et al., 2006), which require the faithfulness assumption to be asymptotically correct.

The question remains as how to estimate the support $\mathrm{supp}(\mathbf{\Theta})$. This is straightforward from classical statistic theory (Newey and McFadden, 1994), i.e., maximum likelihood estimation of $\mathbf{\Theta}$ converges to the ground truth parameters in the sample limit, by computing the inverse of sample covariance matrix $\widehat{\mathbf{\Sigma}}^{-1}$. In practice, a sparsity penalty term is commonly imposed owing to finite-sampling errors. Therefore, we adopt the graphical Lasso (GLasso) (Friedman et al., 2008) approach for estimating $\mathrm{supp}(\mathbf{\Theta})$, which has been shown to perform well in practice.

## 4 Improved Exact Search

In this section, we introduce several strategies to improve exact score-based search methods, with a focus on scalability and asymptotic correctness. We first restrict the search space of exact methods using the estimated super-structure described in Section 3, and then develop a local search strategy to further scale up the search procedure.

[1] Note that this is consistent with the analysis of the strong faithfulness assumption by Uhler et al. (2013).

### 4.1 A* Search with Super-Structure

From Section 2.1, traditional methods such as PC and GES may produce sub-optimal solutions when the faithfulness assumption fails. On the other hand, BIC score-based exact search is guaranteed to provide asymptotically correct results when SMR holds.

**Theorem 3 (Yuan and Malone (2013)).** *BIC score-based exhaustive search algorithms find the MEC of the true DAG if and only if the true DAG satisfies the SMR assumption.*

However, it is challenging to scale these algorithms up to large graphs as the task is NP-hard (Chickering, 1996). To remedy this issue, similar to (Perrier et al., 2008), we propose to explicitly constrain the search space of score-based exact search algorithms using super-structure. This is achieved by limiting the size of *parent graph* in the search procedure. For score-based exact search methods such as DP (Singh and Moore, 2005; Silander and Myllymäki, 2006) and A* (Yuan and Malone, 2013), parent graphs act as important roles in constructing the search space. These methods typically construct a parent graph for each variable, and then perform the search on the *order graph* based on that (Singh and Moore, 2005; Yuan and Malone, 2013), by formulating it as the shortest path problem. Each variable has its own parent graph, and all optimal parent sets are selected from the candidate parent sets, by selecting the ones with the best score in the corresponding parent graph.

Therefore, limiting the size of the parent graph is equivalent to limiting the set of candidate optimal parents for each node. By extracting the neighbors of each node in our estimated super-structure, we obtain the set of nodes that are not possible as its candidate parents; thus, the entries involving these set of nodes in the corresponding parent graph can be directly removed to reduce the search space. A by-product of this proposed strategy is that it could also help reduce memory cost, since we do not need to enumerate all candidate parent sets when constructing the parent graphs.

Note that DP and A* differ mainly in the search phase of the order graph. That is, DP has to consider all combinations of nodes (i.e., $2^d$ combinations), while A* uses a heuristic function to only expand the most promising node (Yuan and Malone, 2013), and has been shown empirically to achieve better efficiency. Therefore, we focus on the A* search with the constraint of super-structure as DP is exponentially expensive in computation and quickly becomes infeasible when the graph size increases. It is also straightforward to apply the super-structure-based pruning method of the parent graph to other exact search algorithms, such as DP, to reduce the search space. Under the Markov and SSCF assumptions, our estimated super-structure via inverse covariance estimation is guaranteed to contain the skeleton of the ground-truth DAG (cf. Theorem 2); thus, the proposed constrained search will not suffer from the trade-off between scalability and reliability.

Besides using super-structure, we apply sparse representation (Yuan and Malone, 2013) of the parent graphs to further remove unnecessary entries, in order to reduce the memory cost of the parent graphs. We also adopt the optimal path extension (Karan and Zola, 2016) and dynamic k-cycle conflict heuristics (Yuan and Malone, 2013) to reduce search space during the A* search. We refer the interested reader to the aforementioned references for more details.

### 4.2 Local A*

Even with super-structure and the other techniques mentioned in Section 4, it is still challenging to scale up exact search due to the large search space. Given the popularity of distributed computation, the idea of divide-and-conquer has been explored for scaling up Bayesian network structure learning (Gu and Zhou, 2020; Zhang et al., 2020). However, they either require extensive computation to enumerate potential patterns when merging different clusters (Kojima et al., 2010), or fail to provide optimal results when faithfulness does not hold (Xie and Geng, 2008).

Thus, with a focus on scalability and weaker assumptions, we develop a local search strategy by first studying the specific assumptions required to correctly discover the local structure around each variable, or, more specifically, the undirected edges and v-structures connecting to each variable.

**Theorem 4.** *Suppose $(G, \mathbb{P})$ follows a linear Gaussian model that satisfies the Markov and SMR assumptions. Then in the results produced by exact score-based search on the variable set containing the variable $X$ and its parents, children, spouses, and grandparents, the discovered undirected edges between $X$ and its direct neighbors and the v-structures involving $X$ are asymptotically correct.*

The above theorem implies that, for each node, only certain variables (i.e., its parents, children, spouses, and grandparents) will affect the search results on its local structure, and adding other variables will not influence the results. The question remains as how to correctly identify a set that

**Algorithm 1** The proposed Local A* approach for causal discovery

**Require:** Structure defined by the support of estimated inverse covariance matrix $G(\mathrm{supp}(\widehat{\mathbf{\Theta}}))$

Initialize $\widehat{\mathcal{M}}$ as a $n \times n$ matrix, all elements 0
Let $N(i)$ denote the set of neighbors of variable $X_i$ indicated by $G(\mathrm{supp}(\widehat{\mathbf{\Theta}}))$
Obtain a variable ordering $\pi$ by sorting the variables $X_i, i = 1 \ldots, d$ based on the cardinality of their corresponding local cluster $C_i = X_i \cup N(i) \cup N(N(i))$
**for** $i = \pi(1), \pi(2), \ldots, \pi(d)$ **do**
    Obtain the local cluster $C_i$
    From $\widehat{\mathcal{M}}$, obtain the previously discovered undirected edges and v-structures involving $X_i$
    With these edges fixed, run improved A* search on $C_i$
    From the A* search result, obtain the trusted edges and save them to $\widehat{\mathcal{M}}$
        (i.e., v-structures and undirected edges involving $X_i$)
**end for**
Output the final MEC $\widehat{\mathcal{M}}$

contains those variables. From Section 3, if the Markov and SSCF assumptions are satisfied, the super-structure defined by the support of estimated inverse covariance matrix $G(\mathrm{supp}(\widehat{\mathbf{\Theta}}))$ asymptotically contains the underlying DAG, indicating that each node in the super-structure has no missing parent or child (i.e., direct neighbor). It is therefore straightforward that all neighbors within two hops of any target node indicated by the super-structure $G(\mathrm{supp}(\widehat{\mathbf{\Theta}}))$ must include its parents, children, spouses, and grandparents. With this *local cluster* formed, including the target variable itself, we can apply exact search method such as A* on it. Based on Theorem 4, the correctness of the undirected edges and v-structures involving the target node can be guaranteed under Markov, SSCF, and SMR assumptions. Therefore, one is able to perform local search independently on the local cluster of each variable, then combine the results that can be trusted based on Theorem 4. In the asymptotic case, the combined output is guaranteed to be the right MEC under the aforementioned assumptions.

The above procedure searches on the local clusters formed by each variable and its neighbors within two hops in the super-structure $G(\mathrm{supp}(\widehat{\mathbf{\Theta}}))$. Given unlimited computational resources, one is certainly able to run local search for all variables in *parallel*, in which the complexity depends on the maximum local cluster size, which however may not be feasible for large graphs. Furthermore, there is a huge overlap between different local clusters, leading to the redundancy of computation during the search procedure. To alleviate this, we consider an *iterative* approach starting from the smallest to the largest local cluster that keeps some edges fixed to take full advantage of the information obtained from the previous searches. More specifically, in each iteration, we save the discovered undirected edges and v-structures from the local search that can be trusted (i.e., shown to be asymptotically correct) based on Theorem 4. Then in the next iteration for, e.g., variable $X$, we look up the undirected edges and v-structures involving $X$ that have been discovered and saved from previous iterations, and keep these edges fixed in the local search for variable $X$. By trusting the information conveyed by these edges and keeping them fixed, we can drastically constrain the search space for the following clusters. This improved procedure, which we call *Local A**, is illustrated in Algorithm 1, and its effectiveness is verified in Section 5.3. Note that parallel computing can be used to further accelerate this proposed iterative search procedure.

# 5 Experiments

We first conduct experiments to compare the effectiveness of GLasso and MMPC for estimating super-structures of the true DAG. We then compare the proposed modifications of A* to different baselines. Lastly, we compare different approaches to scale up A* search, including the proposed ones. The baselines include Triplet A* (Lu et al., 2021), PC (Spirtes and Glymour, 1991; Ramsey et al., 2006), FGES (Ramsey et al., 2017), MMHC (Tsamardinos et al., 2006), and SP (Raskutti and Uhler, 2018). The implementation details of our method and the baselines are given in Appendix D.

The experiment setup is as follows: the ground truth DAGs are simulated using the Erdös–Rényi model (Erdös and Rényi, 1959) with different expected degrees and number of nodes. We then construct the weighted adjacency matrix of each sampled DAG with edge weights sampled uniformly from $[-0.8, -0.2] \cup [0.2, 0.8]$ to ensure that it is bounded away from zero. Based on the weighted

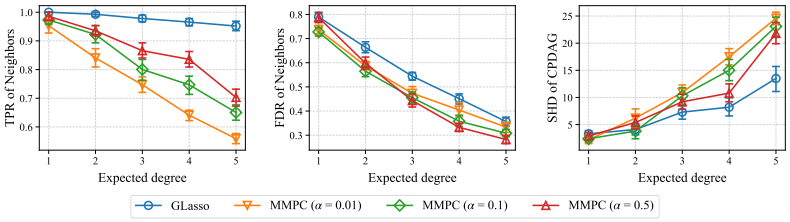



Figure 3: Results of different super-structure estimation methods on 10-node graphs with different degrees. The sample size is $n = 300$. Lower is better, except for TPR.

matrix constructed, we simulate $n \in \{300, 10000\}$ samples using the linear Gaussian model with exogenous noise variances sampled uniformly from $[1, 2]$. The estimated MECs are evaluated using the structural Hamming distance (SHD) computed on the complete partial DAGs (CPDAGs). We also report the F1 score of the estimated skeleton and directed edges in the CPDAGs.

### 5.1 Different Super-Structure Estimation Methods

To demonstrate the effectiveness of GLasso for estimating super-structures with weaker assumptions, we compare the quality of super-structures estimated by GLasso to those by MMPC. We evaluate the ability of these methods for discovering the direct neighbors in the ground truth, by computing the true positive rates (TPRs) and false discovery rates (FDRs) of estimating the true neighbors. In particular, we consider 10-node graphs with varying sample sizes and expected degrees. We also report the performance of MMPC with different significance levels $\alpha$.

Due to space limit, the results for sample size $n = 300$ are shown in Figure 3, while results for $n = 10000$ are in Figure 7 in Appendix E.1. In the first two panels of Figure 3, one observes that GLasso achieves TPRs close to one across all cases, indicating that it rarely misses any direct neighbor, as compared to MMPC with lower TPRs. Notice also that the difference grows as the degree increases. Although GLasso has slightly higher FDRs than MMPC, we believe that the higher TPRs justify the cost of having a larger search space for exact search. It is interesting to observe that even with a high significance level $\alpha$ such as $0.5$ (i.e., the hypothesis tests used by MMPC tend to produce more edges), MMPC has low TPRs on denser graphs. For instance, its TPR is close to $0.7$ on graphs with degree $5$. Similar observations can be made for the case of $n = 10000$. These may be ascribed to the almost violations of faithfulness in the simulated data, and therefore MMPC may miss some neighbors because of unfaithful conditional independencies. On the other hand, GLasso requires only the SSCF assumption, which is intuitively much weaker than faithfulness.

In the third panels of Figures 3 and 7, given the super-structures estimated by GLasso and MMPC, we evaluate the performance of A* method for estimating the CPDAGs. Using super-structures estimated by GLasso shows improvement over MMPC in most settings, especially on graphs with high expected degrees. This is not surprising because the performance of recovering the true CPDAGs is upper-bounded by the proportion of direct neighbors discovered by the super-structure.

### 5.2 Numerical Results

To compare the proposed method, i.e., A* with super-structure and Local A*, with the baselines, we consider graphs with varying sample sizes $n \in \{300, 10000\}$ and expected degrees from $\{1, 2, 3, 4, 5\}$. The baselines include Triplet A*, MMHC, PC, FGES, and SP. We start with 7-node graphs since the computation of SP may be too slow on graphs larger than that.

We report the results for $n = 300$ in Figure 4, and those for other sample sizes, such as $n = 10000$, in Figure 8 in Appendix E.2. In the case of 300 samples, A* and Local A* outperform other baselines across all metrics in most settings. Not surprisingly, the results of them are exactly the same in all settings, which indicates that, similar to A*, Local A* manages to find the globally optimal solutions (given an accurate super-structure), and also verifies the consistency of the proposed local search strategy (cf. Theorem 4). With a larger sample size $n = 10000$, the advantage of A* and Local A* slightly diminishes as their performance is similar to the other baselines. A possible reason is that approximate violations of faithfulness are less likely to occur with more samples; thus, methods like PC and GES are able to find the right solutions. This also demonstrates that exact methods like A* and Local A* that rely on the SMR assumption are relatively reliable in practice since they are less susceptible to approximate violations of faithfulness. Notice also that SP performs the best for 10000 samples; this is not surprising as it is guaranteed to find the globally optimal solutions. For A* and

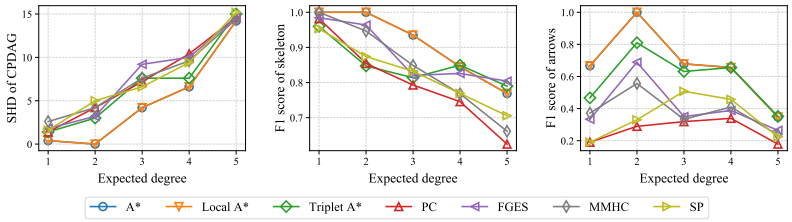



Figure 4: Results of different structure learning methods on 7-node graphs with different degrees. The sample size is $n = 300$. Higher is better, except for SHD. Since each panel has a number of lines, for better readability we do not report the error bars.

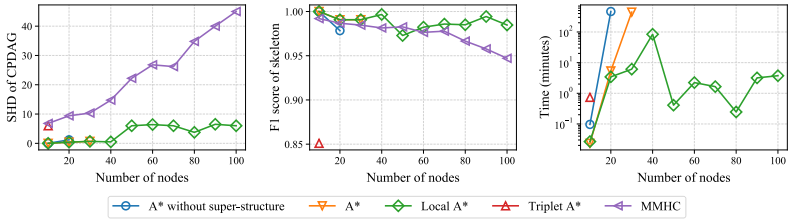



Figure 5: Results and time for different variants of hybrid search on graphs with different sizes. The sample size is $n = 10000$. Lower is better, except for F1 score.

Local A* that rely on super-structure, their performance may be upper-bounded by the proportion of direct neighbors discovered by the super-structure estimation method (see Section 5.1). However, SP does not scale well to large graphs and can handle at most 8 nodes due to the huge search space.

### 5.3 Scaling up A* Search to Large Graphs

We have proposed several strategies to scale up exact method like A*, such as local search and constraining the parent graphs based on super-structure. To study the contribution of each strategy, we compare different variants of A*, including A* without super-structure, A* (with super-structure), and Local A*. Here we adopt Triplet A* and MMHC as our baselines since they have been shown to have similar performance to the other baselines in Section 5.2. Furthermore, they are the most relevant to our proposed Local A* method as they are also two-state hybrid methods, i.e., they first estimate a super-structure, and then perform score-based search on it. We conduct experiments with different graph sizes up to 300 nodes, and terminate the experiments that run for more than one day.

For better readability, the results and time for graphs with 100 nodes or less are depicted in Figure 5, and those for larger graphs are provided in Appendix E.3. We do not report the time for MMHC in Figure 5 since it generally finishes within two minutes. One observes that Local A* and MMHC significantly outperform the others in terms of scalability. Both of them can easily scale up to more than 300 nodes, while none of the rest successfully finish within one day with graph size larger than 30.[2] It is worth noting that A* (with super-structure) scales up to 30 nodes, while A* without super-structure can only be scaled up to 20 nodes in our experiments. This verifies that incorporating super-structure with A* indeed helps reduces search space. Moreover, the search time for Local A* increases gently in the number of nodes, indicating the potential of its scalability even for much larger graphs. On large graphs, Local A* has much better structure recovery results than MMHC.

## 6 Conclusion

We studied the problem of reliable causal discovery with assumptions strictly weaker than faithfulness. Specifically, in our proposed procedure, we adopted the SMR assumption and relaxed the faithfulness assumption as much as is viable. We proposed (1) a sound super-structure estimation method based on assumption that is intuitively much weaker than faithfulness, (2) an improved exact score-based search method with the constraint of the estimated super-structure, and (3) a local search strategy that improves the scalability of exact search without introducing sub-optimal results. The effectiveness and correctness of the proposed method have been validated in our experiments conducted across various settings. Future works include extending the proposed procedure for non-Gaussian cases, and examining the discovery in Section 3.2 for general structure estimation methods.

[2]For fair comparison, we run all experiments with 12 CPUs in parallel, including Triplet A*.

## References


D. M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V*. Springer, 1996.

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

D. Colombo, M. Maathuis, M. Kalisch, and T. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40:294–321, 2011.

P. Erdös and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 69, 04 2017.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9:432–41, 2008.

J. Gu and Q. Zhou. Learning big gaussian bayesian networks: Partition, estimation and fusion. *Journal of Machine Learning Research*, 21(Aug):1–31, 2020.

Y. He, J. Jia, and B. Yu. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 16:2589–2609, 2015.

A. Hyttinen, P. Hoyer, F. Eberhardt, and M. Jarvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Conference on Uncertainty in Artificial Intelligence*, 2014.

S. Karan and J. Zola. Exact structure learning of Bayesian networks by optimal path extension. In *IEEE International Conference on Big Data*, pages 48–55. IEEE Computer Society, 2016.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5(Dec):549–573, 2004.

K. Kojima, E. Perrier, S. Imoto, and S. Miyano. Optimal search on clustered structural constraint for learning bayesian network structure. *Journal of Machine Learning Research*, 11(Jan):285–310, 2010.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(88):3065–3105, 2014.

N. Y. Lu, K. Zhang, and C. Yuan. Improving causal discovery by optimal Bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994.

I. Ng, A. Ghassami, and K. Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, 2020.

S. Ott, S. Imoto, and S. Miyano. Finding optimal models for small gene networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 9:557–67, 2004.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

E. Perrier, S. Imoto, and S. Miyano. Finding optimal bayesian network given a super-structure. *Journal of Machine Learning Research*, 9(74):2251–2286, 2008.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 2006.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

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

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. V. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2019.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Conference on Uncertainty in Artificial Intelligence*, 2006.

A. P. Singh and A. W. Moore. Finding optimal Bayesian networks by dynamic programming. Technical report, Carnegie Mellon University, 2005.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

P. Spirtes and J. Zhang. A uniformly consistent estimator of causal effects under the $k$-triangle-faithfulness assumption. *Statistical Science*, 29(4):662–678, 2014.

P. Spirtes and K. Zhang. Search for causal models. In M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors, *Handbook of Graphical Models*, chapter 18. CRC Press, Inc., 2018.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Conference on Uncertainty in Artificial Intelligence*, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2001.

I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

X. Xie and Z. Geng. A recursive method for structural learning of directed acyclic graphs. *The Journal of Machine Learning Research*, 9(March):459–483, 2008.

C. Yuan and B. Malone. Learning optimal bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48(1):23–65, 2013.

H. Zhang, S. Zhou, C. Yan, J. Guan, X. Wang, J. Zhang, and J. Huan. Learning causal structures based on divide and conquer. *IEEE Transactions on Cybernetics*, 2020.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 2003.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18, 06 2008.

X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.1
   (b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs of all theoretical results are shown in Appendix C
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Error bars are included only in the figures with a few lines. For those with many lines, to make them more readable, we run the experiments with the same set of random seeds and report the average instead of error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix D.
   (b) Did you mention the license of the assets? [Yes] See Appendix D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code and data are available as the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]