Maximal Ancestral Graph Structure Learning via Exact Search

Rantanen, Kari

AUAI Press
2021


http://hdl.handle.net/10138/338682

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
**Abstract**

Generalizing Bayesian networks, maximal ancestral graphs (MAGs) are a theoretically appealing model class for dealing with unobserved variables. Despite significant advances in developing practical exact algorithms for learning score-optimal Bayesian networks, practical exact algorithms for learning score-optimal MAGs have not been developed to-date. We develop here methodology for score-based structure learning of directed maximal ancestral graphs. In particular, we develop local score computation employing a linear Gaussian BIC score, as well as score pruning techniques, which are essential for exact structure learning approaches. Furthermore, employing dynamic programming and branch and bound, we present a first exact search algorithm that is guaranteed to find a globally optimal MAG for given local scores. The experiments show that our approach is able to find considerably higher scoring MAGs than previously proposed in-exact approaches.

**INTRODUCTION**

Exact approaches to Bayesian network structure learning [Bartlett and Cussens, 2017; van Beek and Hoffmann, 2015; Yuan and Malone, 2013] have flourished in recent years. Motivated by guarantees on learning globally optimal structures over a justified Bayesian score and the resulting gains in accuracy from limited number of observational data points, advances in scalability of exact structure learning have been achieved through harnessing different search techniques and optimality-preserving score pruning.

However, despite their popularity, Bayesian networks (BNs) offer only a rather limited independence and causal model. In particular, any causal inference based on BNs must assume causal sufficiency, which is rarely satisfied in real-world settings [Spirtes et al., 1993]. Thus, the demand for accurate structure learning in the presence of latent variables is not answered by the various exact approaches developed for BN structure learning (BNSL). Maximal ancestral graphs (MAGs), on the other hand, are a theoretically appealing generalization of BNs to account for unobserved variables and latent confounding [Richardson and Spirtes, 2002, Spirtes et al., 1993]. Since MAGs can represent marginalizations of BNs and retain many of their important properties, they have been extensively used in causal inference [Richardson and Spirtes, 2003; Zhang, 2008a; Jaber et al., 2019; Perkovic et al., 2017; Spirtes et al., 1993].

In contrast to BNs, for which significant advances in exact structure learning algorithms have been made, exact approaches to learning MAGs have not been developed, despite the potential gains in applicability [1]. Several constraint-based learning algorithms for MAGs or their equivalence classes [Ramsey et al., 2012; Colombo and Maathuis, 2014; Zhang, 2008b; Claassen and Heskes, 2012; Bernstein et al., 2020; Claassen et al., 2013] are in-exact and use independence tests, which are not very accurate for large conditioning sets and relatively small sample sizes. As for score-based learning, [Tsirlis et al., 2018] recently presented a greedy score-based approach under the BIC score (a consistent scoring criterion for linear Gaussian parameterization of MAGs [Richardson and Spirtes, 2002]) in which the needed maximum likelihood estimates can be computed with the so-called residual iterative conditional fitting algorithm (RICF) [Drton et al., 2009]. This greedy search approach builds on earlier developments for GSMAG and MMHC [Tsamardinos and Tsamardinos, 2016; Tsamardinos et al., 2006]. Further studies in this direction include Chib and Constantiou [2004].
Figure 1: (a) A (MAG). (b) A mixed graph that is not a MAG: there is an inducing path between \( x \) and \( q \).

We combine structure and causal effect estimation, and GFCI which combines score-based and constraint-based learning. However, all of these approaches employ forms of greedy search without guarantees on the quality of the obtained solutions.

In this work, we develop score-based learning of MAGs via exact search, motivated by the success of exact approaches to BN structure learning. In particular, we formulate the MAGSL problem of learning score-optimal MAGs employing the Gaussian BIC score. We outline sparsity conditions under which exact search is possible, generalizing the parsimony criterion of greedy search with guarantees on the quality of the obtained solutions.

The following lemma further characterizes inducing paths.

**Lemma 1.** Every inducing path of length \( n \geq 3 \) nodes in a MAG between \( x \) and \( y \) is of the form \( x \leftrightarrow z_1 \leftrightarrow \cdots \leftrightarrow z_{n-2} \leftrightarrow y \), where each \( z_i \) is an ancestor of either \( x \) or \( y \). In particular, \( z_1 \) is an ancestor of \( y \) and \( z_{n-2} \) is an ancestor of \( x \).

**Proof.** Let \( G \) be a MAG with an inducing path between \( x \) and \( y \) going through the nodes \( z_1, \ldots, z_{n-2} \). The edges between each \( z_i, z_{i+1} \) must be bi-directed \( z_i \leftrightarrow z_{i+1} \) (otherwise \( z_i \) or \( z_{i+1} \) would not be a collider). Due to maximality, \( G \) has to contain either \( x \leftrightarrow y, x \leftrightarrow y \) or \( x \leftrightarrow y \). Node \( z_1 \) cannot be an ancestor of \( x \) since that would introduce a directed cycle \( x \rightarrow z_1 \rightarrow \cdots \rightarrow x \) in the context of MAGs.

We define MAGSL as follows.

**Problem 1 (MAGSL).** Find a MAG \( G^* = (V, E^*) \) such that

\[
G^* \in \text{argmax}_{G \in \mathcal{G}} s(G).
\]

where \( \mathcal{G} \) denotes the class of MAGs and \( s \) gives a score for each MAG \( G \).

Note that the highest-scoring MAG is a representative of its equivalence class; the common structural features of the members of the equivalence class can then be further inspected by other methods such as FCI.
We use the BIC scoring function \cite{RichardsonSpirtes2002, Tsirlis2018}
\[ s(G) = \ln L_G(\hat{\theta}) - ((|E| + 2|V|)/2) \cdot \ln N, \]
where \( N \) is sample size, \( L_G \) is the multivariate Gaussian likelihood function and \( \hat{\theta} \) are maximum likelihood parameters for the linear Gaussian model over \( G \). The number of parameters accounts here for the mean and variance for each node, and one coefficient per directed or bi-directed edge.

BIC is asymptotically consistent scoring criterion for MAGs \cite{RichardsonSpirtes2002, Tsirlis2018}. Furthermore, since Markov-equivalent MAGs can represent the same multivariate Gaussians distributions \cite{RichardsonSpirtes2002, Corrollary8.19} and Markov-equivalent MAGs share adjacencies \cite{SpirtesRichardson1996}, the BIC score is score-equivalent.

Fortunately, the score function \( s \) factorizes according to so-called c-components \cite{TianPearl2002, Richardson2009}, which we define for the purpose of this paper as follows. We denote the set of parents of variable \( v \) in a MAG \( G \) by \( pa_G(v) \). Furthermore, we denote the parents of a c-component \( C \) over nodes \( v_1, \ldots, v_l \) with a list \( pa_G(C) = (pa_G(v_1), \ldots, pa_G(v_l)) \).

**Definition 2** (local c-component). A local c-component is a pair \( (C_i, pa(C_i)) \), where \( C_i \) is a graph the nodes of which are strongly connected via bidirected edges and \( pa(C_i) \) is a list of the parent sets of the nodes in \( C_i \).

In Figure 1 (a), \((x \leftrightarrow q, \{w\}) \) and \((w \leftrightarrow y, \{q\}) \) are local c-components. The score of a MAG \( G \) containing \( l \) maximal local c-components indexed by \( i \) is
\[ s(G) = \sum_{i=1}^{l} s(C_i, pa_G(C_i)), \]
where \( s(\cdot, \cdot) \) denote the local scores for the c-components.

To make local score computation and exact learning possible for non-trivial instances, we use two restrictions for \( G \), similarly as what is used for BNSL. We use \( c \) to denote the local c-component size limit (in terms of nodes in \( C_i \)). We use \( p \) to denote the limit on parent relations for a c-component, i.e., \( \sum_{x \in C_i} \|pa(x)\| \leq p \). This means that configuration \( c = 1, p \) is BNSL with maximum number of parents \( p \).

## 3 LOCAL SCORE COMPUTATION

The first challenge in learning MAGs is computing local scores. There are vastly more local scores to compute than in BNSL as we need to consider not only c-components consisting of subsets of variables but also all different ways the variables in a c-component can be connected with the bi-directed edges. Further, the intricate properties of the model class imply that an iterative method (with strong convergence guarantees) is needed for local score computation and that special care should be taken when pruning local scores in order to not jeopardize exactness.

### 3.1 COMPUTING A LOCAL SCORE

We compute local scores as \( s(C_i, pa_G(C_i)) = s(G) - s(G'), \) where \( G \) is a MAG with the local c-component \( C_i \) and its parents; \( G' \) is the empty MAG over all parents of \( C_i \) not in \( C_i \). The scores \( s(G') \) are cached to speed up scoring, similarly as in BNSL and in M3HC \cite{GeigerHeckerman1994, Tsirlis2018}. The maximum likelihood estimates for the BIC score are computed using residual iterative conditional fitting (RICF) of \cite{Drton2009}. The strong convergence properties of RICF state that from a given starting point RICF converges to an accumulation point, all of which give the same value for the likelihood function \cite{Drton2009, Theorem13}. There can be several local optima especially for limited sample sizes \cite{Drton2004, Amendola2020}. In practice, for local c-components of limited size, RICF usually converges from random starting points to only a few different likelihood values; see Section 5. For any c-component of size 1, RICF converges in one step and the produced local scores exactly equal the corresponding BNSL local scores computed with Gobnilp \cite{Bartlett2017}.

### 3.2 COMPUTING ALL LOCAL SCORES

Suppose \( s(C, P) \) is a local score where \( C \) is a c-component over nodes \( x_1, \ldots, x_k \) and each \( pa(x_i) \in P \) describes the parents of \( x_i \). We say that \( z \in pa(x_i) \) is an internal parent if \( z \in C \) and otherwise an external parent. If the c-component includes inducing paths or (almost) directed cycles, i.e., violations of the MAG properties, it cannot be a part of an optimal MAG. Thus, we call \( s(C, P) \) a valid local score if and only if the local c-component \( C \) with the parents in \( P \) is a MAG itself. A valid score will always remain valid if we only add external parents to it.

Every valid local score \( s(C, P) \) over observed variables \( V \), where the number of nodes in \( C \) is at most \( c \) and number of parent relations is limited by \( \sum_{x \in C} \|pa(x)\| \leq p \), can be computed as follows.

1. For all \( Y \subseteq V \) with \( 1 \leq |Y| \leq c \), iterate over every possible MAG over nodes \( Y \) in which every node belongs to one and the same c-component. This corresponds to iterating all valid local scores \( S_{\text{internal}} \) that only contain internal parents. Ignore scores in which the total number of parent relations is more than \( p \).
2. Iterate over each local score in \( S_{\text{internal}} \) and go over all the ways of adding external parents to them (as long as the parent limit \( p \) is not exceeded), forming \( S_{\text{external}} \).
3. Return the local scores in \( S_{\text{internal}} \) and \( S_{\text{external}} \) which cannot be pruned.
which can be multiple times smaller in size compared to (i.e., an edge). If a MAG is non-maximal, it is not due to a missing edge between any maximality-preserving pair.

Definition 3. Let $x$ and $y$ be nodes in graph $G$. We call

\begin{align*}
{x,y} \text{ a maximality-preserving pair in } G & & \\
(a) & & \text{there is no path of bidirected edges between } x \text{ and } y \text{ containing more than 3 nodes;} & & \\
(b) & & \pa_G(x) \subseteq \pa_G(y) \text{ or } \pa_G(y) \subseteq \pa_G(x).
\end{align*}

The correctness of the following pruning rules will be proven later in this section. The first rule is a non-trivial generalization of the standard pruning in BNSL.

Pruning Rule 1. Let $C$ and $C'$ be c-components over a same set of nodes and let their parent set collections be $P$ and $P'$, respectively. The score $s(C', P')$ can be pruned if:

1. $s(C, P) \geq s(C', P')$;
2. each $\pa(x) \in P$ is a subset of the corresponding $\pa(x) \in P'$;
3. if $x \leftrightarrow y$ is missing from $C'$, then it is missing from $C$ as well; and
4. if $x \leftrightarrow y$ exists in $C'$ but not in $C$, then $\{x, y\}$ must be a maximality-preserving pair in $C$.

The second rule generalizes the first by also considering partitionings of a local c-component $C$ into a set of smaller local c-components $C_1, \ldots, C_n$.

Pruning Rule 2. Let $C_1, \ldots, C_n$ and $C'$ be c-components with parent set collections $P_1, \ldots, P_n$ and $P'$, respectively. Suppose the nodes in $C_1, \ldots, C_n$ partition the nodes in $C'$ to distinct subsets. The score $s(C', P')$ can be pruned if:

1. $\sum_{i=1}^{n} s(C_i, P_i) \geq s(C', P')$;
2. each $\pa(x) \in P$ is a subset of the corresponding $\pa(x) \in P_i$ (if $x$ was partitioned into $C_i$);
3. if $x \leftrightarrow y$ is missing from $C'$, then it is not featured in any $C_i$, either; and
4. if $x \leftrightarrow y$ exists in $C'$ but not in some $C_i$, then either $x$ or $y$ was not partitioned into $C_i$, or $\{x, y\}$ is a maximality-preserving pair in $C_i$.

Figure 3 shows an example of applying Pruning Rule 2. The c-component over $x, w, y, z$ with external parent $q$ for $x$ and $z$ (left) can be pruned if the c-components over $x, y$ and $w, z$ (center, right) have a higher sum of scores. Note that removing edges $x \leftrightarrow w$ and $y \leftrightarrow z$ is permitted since the absence of inducing paths between $x, w$ and $y, z$, respectively, is guaranteed in any resulting graph.

To prove the correctness of the score pruning rules, we need additional theory, which builds on important earlier work on
We are finally ready to establish the correctness of Pruning Rule 2 (and thus also of Pruning Rule 1).

The next three lemmas identify which edges can be removed without making nodes connected by inducing paths non-adjacent, thus preserving maximality (Definition 1).

**Lemma 2.** Let $x$ and $y$ be non-adjacent nodes in a MAG $G$. If $\text{pa}_G(x) \subseteq \text{pa}_G(y)$, then there can be no inducing path between $x$ and $y$.

**Proof.** Assume that $\text{pa}_G(x) \subseteq \text{pa}_G(y)$. This means that any ancestor of $x$ in $G$ is also an ancestor of $y$. By Lemma 1 an inducing path between $x$ and $y$ would imply the existence of the edge $z_{n-2} \leftrightarrow y$ where $z_{n-2}$ is an ancestor of $x$. Clearly, $z_{n-2}$ cannot be an ancestor of $y$ since this would introduce an almost directed cycle. Thus there is no inducing path between $x$ and $y$ in $G$. □

The next three lemmas identify which edges can be removed without making nodes connected by inducing paths non-adjacent, thus preserving maximality (Definition 1).

**Lemma 3.** A MAG remains a MAG after removing an edge $x \leftrightarrow y$ if the longest inducing path between $x$ and $y$ has a length less than 4 nodes.

**Proof.** The case where the only inducing path between $x$ and $y$ has length 2 nodes is trivial. Moreover, there cannot be an inducing path of length 3 nodes in a MAG, as it would be of the form $x \leftrightarrow z \leftrightarrow y$ where $z$ is an ancestor of $x$ or $y$ (Lemma 1), which would form an almost directed cycle through $x$ or through $y$. □

**Lemma 4.** If $\text{pa}_{G'}(x) \subsetneq \text{pa}_G(y)$ for nodes $x$ and $y$ in a MAG $G$, then $G$ will remain a MAG after removing edge $x \leftrightarrow y$.

**Proof.** A direct consequence of Lemma 2 □

**Lemma 5.** A MAG remains a MAG after removing an arbitrary edge $x \rightarrow y$.

**Proof.** Let $G$ be a MAG and let $G'$ be $G$ after removing an edge $x \rightarrow y$. Clearly, $G'$ does not introduce new cycles, almost directed cycles or inducing paths. $G'$ could only be non-maximal if there were an inducing path between $x$ and $y$ in $G'$. This would mean that in the original $G$ we had an edge $z_{n-2} \leftrightarrow y$ where $z_{n-2}$ is an ancestor of $x$ (Lemma 1). However, since $G$ has the edge $x \rightarrow y$, this would mean that $G$ contains an almost directed cycle $x \rightarrow y \leftrightarrow z_{n-2} \rightarrow \cdots \rightarrow z_{2} \rightarrow x$, which is not possible. Hence there is no inducing path between $x$ and $y$, and $G'$ is maximal. □

We are finally ready to establish the correctness of Pruning Rule 2 (and thus also of Pruning Rule 1).

Algorithm 1 A dynamic programming algorithm for computing upper bounds, i.e., score-optimal ancestral solutions.

1: function $UB(\text{nodes } U, \text{ reach } R_U, \text{ almost reach } A_U)$
2: if $U = \emptyset$ then return 0
3: Let $u_0 \in U$: lexicographically least node in $U$.
4: Let $C$: the possible c-component and parent set combinations $(C, \text{pa}(C))$ such that $u_0 \in C \subseteq U$.
5: Remove each $(C, \text{pa}(C)) \in C$ that would introduce an (almost) directed cycle given $R_U, A_U$.
6: Let $m \leftarrow -\infty$
7: for each $(C, \text{pa}(C)) \in C$ do
8: Let $R_U \backslash C$ and $A_U \backslash C$ be updated versions of $R_U$ and $A_U$ given $(C, \text{pa}(C))$.
9: $\hat{m} \leftarrow UB(U \setminus C, R_U \setminus C, A_U \setminus C)$
10: $m \leftarrow \max(m, s(C, \text{pa}(C)) + \hat{m})$

return $m$

**Theorem 1.** Let $C'$ and $C_1, \ldots, C_n$ be c-components with parent sets $P'$ and $P_1, \ldots, P_n$ respectively. Suppose the nodes in $C_1, \ldots, C_n$ partition the nodes in $C'$ to distinct subsets. Score $s(C', P')$ can be pruned if:

1. $\sum_i s(C_i, P_i) \geq s(C', P')$, and
2. for all $x \in C_i$, $pa(x) \in P_i$ is a subset of the corresponding path $pa'(x) \in P'$, and
3. if $x \leftrightarrow y \notin C'$ then $x \leftrightarrow y \notin C_i$ for all $x, y \in C_i$, and
4. if $x \leftrightarrow y \in C'$ and $x \leftrightarrow y \notin C_i$ for some $x, y \in C_i$, then $\{x, y\}$ must be maximality-preserving pair in $C_i$.

**Proof.** Suppose a MAG $G'$ includes the local c-component $C'$ with parent sets $P'$. We can replace this component with $C_1, \ldots, C_n$ and $P_1, \ldots, P_n$, to form a valid MAG $G$: (1) $G$ does not contain any edges which are not in $G'$. (2) A MAG remains a MAG after removing edges $x \rightarrow y$ in $G'$ but not in $G$ sequentially (Lemma 2). (3) Suppose that $G$ is missing an edge $x \leftrightarrow y \in G'$, if $x \notin C_i$ or $y \notin C_i$ for all $C_i$, there are no inducing path between $x$ and $y$ in $G$ by Lemma 1 and so $x \leftrightarrow y$ is allowed to be absent. Otherwise, if $x, y \in C_i$ for some $C_i$, we require $\{x, y\}$ to be a maximality-preserving pair in $C_i$. Therefore we have either (a) $pa_{G'}(x) \subseteq pa_G(y)$ or (b) all paths of bidirectional edges between $x$ and $y$ have less than 4 nodes. By Lemmas 4 and 5 resp., the edge $x \leftrightarrow y$ may be absent in either case. As $\sum_i s(C_i, P_i) \geq s(C', P')$, we do not need $P'$ and $C'$ to form an optimal MAG. □

**4 EXACT STRUCTURE SEARCH**

We turn to developing a first exact search procedure for MAGSL. Several approaches for BNSL [Koivisto and Sood 2004] Silander and Myllymäki, 2006, Yuan and Malone 2013] can use the fact that if a sink node and edges to it are removed from an optimal BN structure, then the remaining BN structure is optimal with respect to the remaining vari-
Algorithm 2 Exact branch-and-bound algorithm for finding score-optimal maximal and ancestral solutions $S^*$ (MAGs).

\begin{algorithm}
\begin{algorithmic}[1]
\Function{\textsc{search}}{(nodes $U$, partial solution $S$)}
\If{$S$ is non-maximal} \Return \EndIf
\If{$U = \emptyset$} \Return \EndIf
\If{$s(S') < s(S)$} \State $S^* \leftarrow S$ \EndIf
\State $R_U, A_U \leftarrow$ reachability in $S$
\If{$s(S) + \UB(U, R_U, A_U) \leq s^*$} \State \Return \EndIf
\State Let $\hat{S}$: $S$ extended with the UB solution.
\If{$\hat{S}$ is maximal} \State \Return \EndIf
\State \If{$s(S^*) < s(\hat{S})$} \State $S^* \leftarrow \hat{S}$ \EndIf
\State Let $u_0 \in U$: lexicographically least node in $U$.
\State Let $C$: all possible c-component and parent set combinations $(C, \pa(C))$ such that $u_0 \in C \subseteq U$.
\State Remove each $(C, \pa(C)) \in C$ that would introduce an (almost) directed cycle given $R_U, A_U$.
\For{each $(C, \pa(C)) \in C$} \State Let $S'$ be $S$ with $(C, \pa(C))$ added to it.
\State \EndFor
\State \EndFunction
\end{algorithmic}
\end{algorithm}

Although all MAGs do have sink nodes, unfortunately removing a sink node does not always preserve optimality: to see this, consider the covariance matrix

$$C_x = \begin{bmatrix}
1.00 & 0.00 & 0.20 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.20 \\
0.20 & 0.00 & 1.04 & 0.50 \\
0.00 & 0.20 & 0.50 & 1.04
\end{bmatrix}$$

with $N = 100$ samples. The optimal solution over all four nodes is $v_1 \rightarrow v_3 \leftarrow v_4 \leftarrow v_2$, while without $v_4$ the optimal solution is the empty graph. Intuitively, the correlation between $v_1$ and $v_2$ is only worth the additional parameter under the BIC score if the correlation of $v_4$ on $v_3$ is also taken into account. Symmetrically, without $v_3$ the optimal solution is also the empty graph. The decomposability of the score in Eq. 3 though directly implies that removing a sink c-component (from which there are no edges out) from an optimal MAG, retains optimality. However, not all MAGs have so called sink c-components. Figure 1(u) shows a MAG for which there is no order for the c-components and thus no sink c-components [Richardson and Spirtes, 2002].

4.1 A BRANCH-AND-BOUND APPROACH

Nevertheless, the conditions of a mixed graph being a MAG (Definition 1) are rather restrictive in terms of search. Placing a single edge forbids the presence of a number of other edges through the MAG conditions. Our approach takes advantage of this: after placing a local c-component (with its parents) we disregard a large number of other c-components whose edges would now result in non-MAG structures.

To use this, we keep track of the reach $R_U$ and almost reach $A_U$ for any subset of nodes $U \subseteq V$. A node $x$ reaches a node $y$ in a graph $G$ if there exists a directed path from $x$ to $y$ in $G$. Moreover, $x$ almost reaches $y$ in $G$ if there exists nodes $v_1, \ldots, v_n$ with $x = v_1, y = v_n$ such that the edge $v_j \leftrightarrow v_{j+1}$ exists in $G$ for some $j \in \{1, \ldots, n-1\}$ and for all other $i \in \{1, \ldots, n-1\}$ the edge $v_i \rightarrow v_j$ is in $G$. $R_U$ ($A_U$) implies an (almost) directed cycle if a node can (almost) reach itself.

Algorithm 1 uses dynamic programming [Koivisto and Sood, 2004; Silander and Myllymäki, 2006; Tian and He, 2009] to compute an upper bound for the score of an optimal solution for nodes $U$ such that, together with $R_U, A_U$, (almost) directed cycles are not formed (Line 5). The algorithm iterates over the local c-components $C$ containing a node $u_0$ (Line 8-10). This solves a relaxation of the MAGSL optimization problem as solutions need only be ancestral: inducing paths between non-adjacent nodes are allowed. We cache the solutions of each $\UB(U, R_U, A_U)$ to avoid solving subproblems repeatedly. For efficiency, we do not store the almost reach from node $x$ to node $y$ if $x$ can reach $y$.

Algorithm 2 uses branch and bound [Suzuki, 1996; Tian, 2000; Rantanen et al., 2017] to find an optimal, maximal and ancestral solution $S^*$, by extending the partial solution $S$ over $V \setminus U$ to also cover the remaining nodes $U$. The algorithm is invoked by $\textsc{search}(V, S_0)$ where $S_0$ contains no nodes and $S^*$ is initialized to a valid MAG (e.g. the empty MAG). At each step, we ensure that $S$ is maximal by checking that there are no inducing paths between any unadjacent $x, y \in V \setminus U$ (Line 2). If there are no remaining nodes and $S$ has a better score than $S^*$, we update $S^*$ (Line 4). Otherwise, if improving $S^*$ by assigning the remaining nodes $U$ to $S$ is impossible, we backtrack in the search (Line 7). On Line 8 we create $\hat{S}$ by extending $S$ with the upper bound solution obtained from Algorithm 1. If $\hat{S}$ is maximal, we backtrack and also update $S^*$ if $\hat{S}$ has better score.

4.2 AN ALTERNATIVE ASP-BASED APPROACH

Motivated by the successes of declarative programming techniques, in particular answer set programming (ASP) [Gebser et al., 2012], for learning globally optimal structures from various graphical model classes and settings [Hyttinen et al., 2014; Sonntag et al., 2015; Magliacane et al., 2016; Fores and Moot, 2018; Zhalama et al., 2019], we developed a first ASP-based approach to MAGSL as a comparative baseline exact approach. As the approach turns out to show weaker performance than the branch-and-bound approach, we provide the detailed encoding as part of our code package. In short, the ASP encoding is based on representing in a natural way the search space of MAGs. This requires encoding the ruling out of directed and almost directed cycles as well as confirming that inducing paths only appear between adjacent nodes. Existence of specific edges is implied by the
chosen c-components, and each node is required to be contained in exactly one c-component. The objective of finding a best-scoring MAG structure is represented as weighted soft constraints on the input c-component candidates.

## 5 EMPIRICAL EVALUATION

We empirically evaluate the efficiency of the score computation approach, and our branch-and-bound algorithm ("BB") against competing learning algorithms in terms of runtime efficiency and quality of the MAGs found. We implemented the local score computation and Algorithms 1 and 2 in C++. The score computation software includes an efficient C++ implementation of the RICF algorithm as well as the score pruning. The experiments were run on on 2.83-GHz Intel Xeon E5440 nodes under 100-GB memory limit.

**Data** We generated synthetic data by sampling MAGs over $n = 7..12$ nodes with expected degree is 2 or 3 and the presence of directed and bi-directed edges equally probable. The coefficients and covariances for the directed and bidirected edges were drawn uniformly from $[-0.2, 0.8]$. Variances of disturbances were drawn as $[0.5 \cdot s, 1]$, with $s \sim N(0, 1)$. Rejection sampling was used to ensure the models were valid linear Gaussian MAGs. For each model we generated $N = 200$ samples. In addition to synthetic MAGs, we use four Gaussian Bayesian network benchmarks from the bnlearn network repository: ecoli70, magic-niab, magic-irri and arth150 [Scutari, 2010]. We sampled from $N = 100$ to $N = 1600$ samples from each benchmark. To inspect structure learning in the presence of latent variables, we marginalized the data to datasets of $n = 7..11$ highly correlated variables, reducing the set of variables by repeatedly removing the variable for which the sum of the absolute values of correlations to the remaining variables was minimized.

**Score Computation and Pruning** We first look at the performance of the local score computation part on the synthetic data. Figure 4 shows the number of local scores produced for different numbers of nodes and c-component size limits without (blue x) and with pruning (box plot). Note that for $c = 2$ we use parent limit $p = 8$ and for $c = 3$ we use $p = 4$, as these are still feasible choices. We observe that increasing the c-component size limit from 2 to 3 results in substantial increase in the number of local scores. The increase in the number of local scores for each additional node is perhaps more modest. There is considerable instance-specific variation on the number local scores after pruning. Figure 5 shows the runtime for local score computation for different configurations. The runtime grows exponentially with the number of nodes but we can compute local scores on a single thread for up to 12 nodes in 10 hours. This is inline with BNSL, also there the computation of local scores may take considerable amount of time.

**Runtime Performance** Figure 6 left and middle show runtimes (sorted by increasing runtime for each line) of our branch and bound under the c-component size limits $c = 2$ and $c = 3$ on the synthetic data. Within approximately 1 hour, the approach scales up to 8-variable instances for $c = 2, 3$. For $c = 2$ we can solve many instances with 9-12 nodes within the time limit of 10 hours. Runtime comparison of BB with the alternative ASP-based approach (Section 4.2) using the state-of-the-art Clingo ASP solver [Gebser et al., 2012] is shown in Figure 6 right for the benchmark datasets marginalized to 7-node instances. BB is considerably faster than the ASP approach. Table 1 ($c = 2$) and Table 2 ($c = 3, 4$) show further per-instance runtimes of our BB approach on the benchmark instances. There is some variation to the runtimes depending on the dataset. BB scales best on magic-irri and magic-niab, solving many of the 10-node instances for $c = 2$ within 10 hours.

**Quality of Solutions** We evaluate the performance of our exact branch-and-bound search for MAGSL in terms of solution quality against competing in-exact approaches: the hill-climbing-based in-exact MAGSL algorithm M3HC [Tsirlis et al., 2018], the constraint-based FCI [Zhang, 2008b], Spirtes et al. [1993] and its variant GFCI [Ogarrio et al., 2018].

---

**Figure 4:** Score counts for the synthetic datasets.

**Figure 5:** Scoring and pruning time on synthetic data.
Figure 6: Runtimes: BB for $c = 2$ (left) and $c = 3$ (middle) on synthetic data; BB vs ASP on benchmark data (right).

Table 1: Results on the benchmark datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>N</td>
<td>n</td>
<td>N</td>
<td>n</td>
<td>N</td>
<td>n</td>
<td>N</td>
</tr>
<tr>
<td>arthropod</td>
<td>100 8</td>
<td>9633</td>
<td>-233.27</td>
<td>-283.38</td>
<td>-228.19</td>
<td>-257.75</td>
<td>-235.54</td>
</tr>
<tr>
<td>arthropod</td>
<td>200 8</td>
<td>162569</td>
<td>-412.39</td>
<td>-465.41</td>
<td>-457.46</td>
<td>-586.98</td>
<td>-457.29</td>
</tr>
<tr>
<td>arthropod</td>
<td>400 8</td>
<td>30276</td>
<td>-777.73</td>
<td>-819.2</td>
<td>-769.37</td>
<td>-905.43</td>
<td>-744.04</td>
</tr>
<tr>
<td>arthropod</td>
<td>800 8+</td>
<td>timeout</td>
<td>-1423.85</td>
<td>-1492.67</td>
<td>-1367.92</td>
<td>-1567.07</td>
<td>-1450.01</td>
</tr>
<tr>
<td>arthropod</td>
<td>1600 8+</td>
<td>timeout</td>
<td>-2780.39</td>
<td>-2928.61</td>
<td>-2584.09</td>
<td>-3057.74</td>
<td>-2643.85</td>
</tr>
<tr>
<td>ecoli70</td>
<td>100 8</td>
<td>5439</td>
<td>-701.22</td>
<td>-739.24</td>
<td>-844.42</td>
<td>-745.36</td>
<td>-701.22</td>
</tr>
<tr>
<td>ecoli70</td>
<td>200 8</td>
<td>8285</td>
<td>-1378.43</td>
<td>-1386.35</td>
<td>-1378.43</td>
<td>-1386.35</td>
<td>-1378.43</td>
</tr>
<tr>
<td>ecoli70</td>
<td>400 8</td>
<td>12256</td>
<td>-2684.62</td>
<td>-2696.84</td>
<td>-2413.51</td>
<td>-2696.84</td>
<td>-2684.62</td>
</tr>
<tr>
<td>ecoli70</td>
<td>800 8+</td>
<td>timeout</td>
<td>-18996</td>
<td>-5330.34</td>
<td>-5332.49</td>
<td>-5332.61</td>
<td>-5332.61</td>
</tr>
<tr>
<td>ecoli70</td>
<td>1600 8</td>
<td>19717</td>
<td>-10728.61</td>
<td>-10728.61</td>
<td>-10728.61</td>
<td>-10728.61</td>
<td>-10728.61</td>
</tr>
</tbody>
</table>

Figure 7 shows the Bayes factors of the found MAGs with respect to the null model, which combines score-based and constraint-based learning. We employ the FCI and GFCI implementation from the causal-cmd v1.2.1 package with Fisher-z-test $\alpha = 0.05$ and sem-bic-score. We directly obtained a single representative MAG of the equivalence class returned by these algorithms. As the other methods are considered faster per a single run (finishing often within minutes), we also bootstrapped the data 50 times and then report the best result based on score w.r.t. the original data. We also report (as “BN”) the scores of the optimal Bayesian networks found by using Gobnilp [Bartlett and Cussens, 2017].
Table 2: Results on the benchmark datasets.

<table>
<thead>
<tr>
<th>Benchmark dataset</th>
<th>Time</th>
<th>BIC score</th>
</tr>
</thead>
<tbody>
<tr>
<td>arith150 100</td>
<td>528s</td>
<td>-224.74</td>
</tr>
<tr>
<td>arith150 200</td>
<td>1482s</td>
<td>-377.96</td>
</tr>
<tr>
<td>arith150 400</td>
<td>1426s</td>
<td>-722.47</td>
</tr>
<tr>
<td>arith150 800</td>
<td>1447s</td>
<td>-1380.36</td>
</tr>
<tr>
<td>arith150 1600</td>
<td>1771s</td>
<td>-2069.91</td>
</tr>
<tr>
<td>ecoli70 100</td>
<td>519s</td>
<td>-602.8</td>
</tr>
<tr>
<td>ecoli70 200</td>
<td>631s</td>
<td>-1417s</td>
</tr>
<tr>
<td>ecoli70 400</td>
<td>748s</td>
<td>-2321.64</td>
</tr>
<tr>
<td>ecoli70 800</td>
<td>914s</td>
<td>-4614.11</td>
</tr>
<tr>
<td>ecoli70 1600</td>
<td>935s</td>
<td>-15366.12</td>
</tr>
<tr>
<td>magic-irri 100</td>
<td>17s</td>
<td>-977.01</td>
</tr>
<tr>
<td>magic-irri 200</td>
<td>51s</td>
<td>-1951.54</td>
</tr>
<tr>
<td>magic-irri 400</td>
<td>52s</td>
<td>-3802.84</td>
</tr>
<tr>
<td>magic-irri 800</td>
<td>198s</td>
<td>-7520.48</td>
</tr>
<tr>
<td>magic-irri 1600</td>
<td>259s</td>
<td>-15051.99</td>
</tr>
<tr>
<td>magic-niab 100</td>
<td>1144s</td>
<td>-1091.97</td>
</tr>
<tr>
<td>magic-niab 200</td>
<td>312s</td>
<td>-2169.11</td>
</tr>
<tr>
<td>magic-niab 400</td>
<td>493s</td>
<td>-4228.14</td>
</tr>
<tr>
<td>magic-niab 800</td>
<td>2028s</td>
<td>-8358.48</td>
</tr>
<tr>
<td>magic-niab 1600</td>
<td>278s</td>
<td>-16718.88</td>
</tr>
<tr>
<td>magic-naib 100</td>
<td>&lt;1s</td>
<td>-845.77</td>
</tr>
<tr>
<td>magic-naib 200</td>
<td>3s</td>
<td>-1616.51</td>
</tr>
<tr>
<td>magic-naib 400</td>
<td>2s</td>
<td>-3593.07</td>
</tr>
<tr>
<td>magic-naib 600</td>
<td>6s</td>
<td>-6618.23</td>
</tr>
<tr>
<td>magic-naib 1000</td>
<td>8s</td>
<td>-132206.57</td>
</tr>
<tr>
<td>magic-naib 1600</td>
<td>3s</td>
<td>-974.23</td>
</tr>
<tr>
<td>magic-naib 1600</td>
<td>74s</td>
<td>-1850.11</td>
</tr>
<tr>
<td>magic-naib 800</td>
<td>61s</td>
<td>-4109.82</td>
</tr>
<tr>
<td>magic-naib 1600</td>
<td>131s</td>
<td>-8776.99</td>
</tr>
<tr>
<td>magic-irri 1600</td>
<td>470s</td>
<td>-15306.12</td>
</tr>
</tbody>
</table>

There is noticeable variation (see Figure 7) with respect to the dataset: for many the MAGs found are 10× more likely than BNs, for some BNs are equally likely. FCI and GFCI perform somewhat poorly, suggesting that they cannot find very high scoring MAGs over the 200 data samples given. The MAGs found by bootstrapped M3HC are better, often more likely than BNs. The MAGs found by our approach are still better, 2-10 times more likely than the optimal BNs. Note that our method is feasible with different parent relation limits p, i.e., p = 8 for c = 2 and p = 4 for c = 3, 4. There is an interesting trade-off as to whether to allow for larger c-components or more parent relations.

Tables 1 and 2 give the BIC scores obtained on the benchmark data, with highest score for each instance in bold. Optimal BNs achieve the highest score for a minority of the instances. Relative performance of the in-exact algorithms depends noticeably on the dataset and the number of samples. FCI, GFCI and M3HC can in some cases (especially for arith10 and ecoli70) find high-scoring solutions. For magic-niab and magic-irri our BB obtains clearly better solutions (using c = 2, 3, 4) on this metric than its in-exact competitors. Moreover, the solutions of BB are guaranteed not find very high scoring MAGs over the 200 data samples given. The MAGs found by bootstrapped M3HC are better, often more likely than BNs. The MAGs found by our approach are still better, 2-10 times more likely than the optimal BNs. Note that our method is feasible with different parent relation limits p, i.e., p = 8 for c = 2 and p = 4 for c = 3, 4. There is an interesting trade-off as to whether to allow for larger c-components or more parent relations.

6 CONCLUSION

MAGs are a generalization of BNs, allowing for unobserved variables. We developed algorithmic solutions—incorporating
score computation, score pruning, and a practical exact search algorithm—to learning score-optimal MAG, motivated by the impact of such solutions in the realm of BNs. Potential avenues for developing further approaches to MAG structure learning include approaches with improved scalability [Lee and van Beek 2017, Tsamardinos et al. 2006, Chickering 2002], liftings of techniques currently mostly limited to DAG structures, e.g., Bayesian model averaging through exact computation [Liao et al. 2019, Koivisto and Sood 2004, Tian and He 2009] or MCMC sampling [Kuipers and Moffa 2017, Koller and Friedman 2009, Silva and Ghahramani 2006], and techniques for focusing search on promising local c-components towards scaling to large-scale structures [Scanagatta et al. 2015].

Acknowledgements

This work was funded by Academy of Finland under grants 316771, 322869 and 328718.

References


