Score based vs constraint based causal learning in the presence of confounders

Sofia Triantafillou
Computer Science Dept.
University of Crete
Voutes Campus, 700 13 Heraklion, Greece

Ioannis Tsamardinos
Computer Science Dept.
University of Crete
Voutes Campus, 700 13 Heraklion, Greece

Abstract

We compare score-based and constraint-based learning in the presence of latent confounders. We use a greedy search strategy to identify the best fitting maximal ancestral graph (MAG) from continuous data, under the assumption of multivariate normality. Scoring maximal ancestral graphs is based on (a) residual iterative conditional fitting [Drton et al., 2009] for obtaining maximum likelihood estimates for the parameters of a given MAG and (b) factorization and score decomposition results for mixed causal graphs [Richardson, 2009, Nowzohour et al., 2015]. We compare the score-based approach in simulated settings with two standard constraint-based algorithms: FCI and conservative FCI. Results show a promising performance of the greedy search algorithm.

1 INTRODUCTION

Causal graphs can capture the probabilistic and causal properties of multivariate distributions. Under the assumptions of causal Markov condition and faithfulness, the graph induces a factorization for the joint probability distribution, and a graphical criterion (d-separation) can be used to identify all and only the conditional independencies that hold in the joint probability distribution.

The simplest case of a causal graph is a directed acyclic graph (DAG). A causal DAG $\mathcal{G}$ and faithful probability distribution $\mathcal{P}$ constitute a causal Bayesian network (CBN) [Pearl, 2000]. Edges in the graph of a CBN have a straightforward interpretation: A directed edge $X \to Y$ denotes a causal relationship that is direct in the context of variables included in the DAG. In general, CBNs are considered in the setting where causal sufficiency holds, i.e. the absence of latent confounders. This is restrictive, since in most cases we can/do not observe all variables that participate in the causal mechanism of a multivariate system.

We consider a representation for an equivalence class of models based on maximal ancestral graphs (MAGs) [Richardson and Spirtes, 2002]. MAGs are extensions of CBNs that also consider latent confounders. Latent confounders are represented with bi-directed edges. The set of conditional independencies that hold in a faithful probability distribution can be identified from the graph with the graphical criterion of m-separation. The causal semantics of edges in MAGs are more complicated: Directed edges denote causal ancestry, but the relationship is not necessarily direct. Bi-directed edges denote latent common causes. However, each pair of variables can only share one edge, and causal ancestry has precedence over confounding: If X is a causal ancestor of Y and the two are also confounded, then $X \to Y$ in the MAG. MAGs have several attractive properties: They are closed under marginalization and every non-adjacency corresponds to a conditional independence.

There exist two main approaches for learning causal graphs from data. Constraint-based approaches infer the conditional independencies imprinted in the data and search for a DAG/MAG that entails all (and only) of these independences according to d/m-separation. Score-based approaches try to find the graph $\mathcal{G}$ that maximizes the likelihood of the data given $\mathcal{G}$ (or the posterior), according to the factorization imposed by $\mathcal{G}$. In general, a class of causal graphs, that are called Markov equivalent, fit the data equally well. Constraint-based approaches are more efficient and output a single graph with clear semantics, but give no indication the relative confidence in the model. Moreover, they have been shown to be sensitive to error propagation [Spirtes, 2010]. Score-based methods on the other hand do not have this problem, and they also provide a metric of confidence in the entire output model. Hybrid methods that exploit the best of both worlds have therefore proved successful in learning causal graphs from data [Tsamardinos et al., 2006].

Numerous constraint-based and score-based algorithms exist that learn causal DAGs (classes of Markov equivalent DAGs) from data. Learning MAGs on the other hand is typically done with constraint-based algorithms. A score-based method for mixed causal graphs (not necessarily MAGs) has recently been proposed [Nowzohour et al., 2015] based on relative factorization results [Tian and Pearl, 2003, Richardson, 2009].

Using these decomposition results, we implemented a simple greedy search for learning MAGs from data. We compare the results of this approach with FCI [Spirtes et al., 2000, Zhang, 2008] and conservative FCI [Ramsey et al., 2006] outputs. Greedy search performs slightly worse in most settings in terms of structural hamming distance, and better than FCI in terms of precision and recall.

Based on these results, we believe that score-based approach can be used to improve learning causal graphs in the presence of confounders. Algorithm implementation and code for the detailed results are available in https://github.com/striantafillou.

The rest of the paper is organized as follows: Section 2 briefly reviews causal graphs with and without causal sufficiency. Section 3 gives an overview of constraint-based and score-based methods for DAGs and MAGs. Section 4 describes a greedy search algorithm for learning MAGs. Related work is discussed in Section 5. Section 6 compares the performance of the algorithm against FCI and CFCI. Conclusions and future work are presented in 7.

2 CAUSAL GRAPHS

We begin with some graphical notation: A mixed graph (MG) is a collection of nodes (interchangeably variables) $\mathbf{V}$, along with a collection of edges $\mathbf{E}$. Edges can be directed ($X \rightarrow Y$) or bi-directed ($X \leftrightarrow Y$). A path is a sequence of adjacent edges (without repetition). The first and last node of a path are called endpoints of the path.

A bi-directed path is a path where every edge is bi-directed. A directed path is a path where every edge is directed and oriented in the same direction. We use $X \dashrightarrow Y$ to symbolize a directed path from X to Y. A directed cycle occurs when there exists a directed path $X \dashrightarrow X$. An almost directed cycle is occurs when $X \leftrightarrow Y$ and $X \dashrightarrow Y$. A triplet $\langle X, Y, Z \rangle$ on consequent nodes on a path are form a collider if $X \rightarrow Y \leftarrow Z$. If X and Z are not adjacent, the triplet is an unshielded collider.

A mixed graph is called ancestral if it has no directed and almost directed cycles. An ancestral graph without bi-directed edges is a DAG. X is a parent of Y in a MG $\mathcal{G}$ if $X \rightarrow Y$ in $\mathcal{G}$. We use the notation $Pa_{\mathcal{G}}(X)$, $An_{\mathcal{G}}(X)$ to denote the set of parents and ancestors of X in $\mathcal{G}$.

Under causal sufficiency, DAGs can be used to model causal relationships: For a graph $\mathcal{G}$ over a set of variables $\mathbf{V}$, $X \rightarrow Y$ in $\mathcal{G}$ if X causes Y directly (no variables in $\mathbf{V}$ mediate this relationship). Under the causal Markov condition and faithfulness Pearl [2000], $\mathcal{G}$ is connected to the joint probability distribution $\mathcal{P}$ over $\mathbf{V}$ through the criterion of d-separation (defined below). Equivalently, the causal Markov condition imposes a simple factorization of the joint probability distribution:

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V | Pa_{\mathcal{G}}(V)) \tag{1}$$

Thus, the parameters of the joint probability distribution describe the probability density function of each variable given its parents in the graph. An interesting property of CBNs, that constitutes the basis of constraint-based learning, is the following: Every missing edge in a DAG of a CBN corresponds to a conditional independence. Hence, if X is independent from Y given $\mathbf{Z}$ (symb. $X \perp\!\!\!\perp Y | \mathbf{Z}$) in $\mathcal{P}$, then X and Y are not adjacent in $\mathcal{G}$.

In general, a class of Markov equivalent DAGs fit the data equally well. DAGs in a Markov equivalent class share the same skeleton and unshielded colliders. A Pattern DAG (PDAG) can be used to represent the Markov equivalent class of DAGs: It has the same edges as every DAG in the Markov equivalence class, and the orientations that are shared by all DAGs in the Markov equivalence class.

Confounded relationships cannot be represented in DAGs, and mixed causal graphs were introduced to tackle this problem. The most straightforward approach is with semi-Markov causal models (SMCMs) Tian and Pearl [2003]. The graphs of semi-Markov causal models are acyclic directed mixed graph (ADMGs). Bi-directed edges are used to denote confounded variables, and directed edges denote direct causation. A pair of variables can share up to two edges (one directed, one bi-directed). The conditional independencies that hold in a faithful distribution are represented through the criterion of m-separation:

Definition 2.1 (m-connection, m-separation.) *In a mixed graph $\mathcal{G} = (\boldsymbol{E}, \boldsymbol{V})$, a path p between A and B is **m-connecting** given (conditioned on) a set of nodes $\mathbf{Z}$, $\mathbf{Z} \subseteq \boldsymbol{V} \setminus \{A, B\}$ if*

1. *Every non-collider on p is not a member of $\mathbf{Z}$.*
2. *Every collider on the path is an ancestor of some member of $\mathbf{Z}$.*

*A and B are said to be **m-separated** by $\mathbf{Z}$ if there is no m-connecting path between A and B relative to $\mathbf{Z}$. Otherwise, they are said to be **m-connected** given $\mathbf{Z}$.For graphs without bi-directed edges, m-separation is reduced to the d-separation criterion.*

Markov equivalence classes of semi-Markov causal models do not have a simple characterization, because Markov equivalent SMCMs do not necessarily share the same edges: Absence of an edge in a SMCM does not necessarily correspond to an m-separation. Figure 1 shows an example of two SMCMs that encode the same m-separations but do not have the same edges (figure taken from [Triantafillou and Tsamardinos, 2015]).

Maximal ancestral graphs are also used to model causality and conditional independencies in causally insufficient systems. MAGs are mixed *ancestral* graphs, which means that they can have no directed or almost directed cycles. Every pair of variables X, Y in an ancestral graph is joined by at most one edge. The orientation of this edge represents (non) causal ancestry. A directed edge $X \rightarrow Y$ denotes that X is an ancestor of Y, but the relation is not necessarily direct in the context of modeled variables (see for example edge $A \rightarrow D$ in MAG $\mathcal{M}_1$ of Figure 1). Moreover, X and Y may also be confounded (e.g. edge $B \rightarrow D$ in MAG $\mathcal{M}_1$ of Figure 1). A bi-directed edge $X \leftrightarrow Y$ denotes that X and Y are confounded.

Like SMCMs, ancestral graphs encode the conditional independencies of a faithful distribution according to the criterion of m-separation. *Maximal* ancestral graphs are graphs in which every missing edge (non-adjacency) corresponds to a conditional independence. Every ancestral graph can be extended to a maximal ancestral graph by adding some bi-directed edges [Richardson and Spirtes, 2002]. Thus, Markov equivalence classes of maximal ancestral graphs share the same edges and unshielded colliders, and some additional shielded colliders, discussed in Zhang [2008], Ali et al. [2009]. Partial ancestral graphs (PAGs) are used to represent the Markov equivalence classes of MAGs.

Figure 1 illustrates some differences in SMCMs and MAGs that represent the same marginal of a DAG. For example, A is a causal ancestor of D in DAG $\mathcal{G}_1$, but not a direct cause (in the context of observed variables). Therefore, the two are not adjacent in the corresponding SMCM $\mathcal{S}_1$ over $\{A, B, C, D\}$. However, the two cannot be rendered independent given any subset of $\{B, C\}$, and therefore $A \rightarrow D$ is in the respective MAG $\mathcal{M}_1$.

On the same DAG, B is another causal ancestor (but not a direct cause) of D. The two variables share the common cause L. Thus, in the corresponding SMCM $\mathcal{S}_1$ over $\{A, B, C, D\}$ $B \leftrightarrow D$ is present. However, a bi-directed edge between B and D is not allowed in MAG $\mathcal{M}_1$, since it would create an almost directed cycle. Thus, $B \rightarrow D$ is in $\mathcal{M}_1$.

Overall, a SMCM has a subset of the adjacencies of its MAG counterpart. These extra adjacencies in MAGs correspond to pairs of variables that cannot be m-separated given any subset of observed variables, but neither directly causes the other, and the two are not confounded. These adjacencies can be checked in a SMCM using a special type of path, called inducing path [Richardson and Spirtes, 2002].

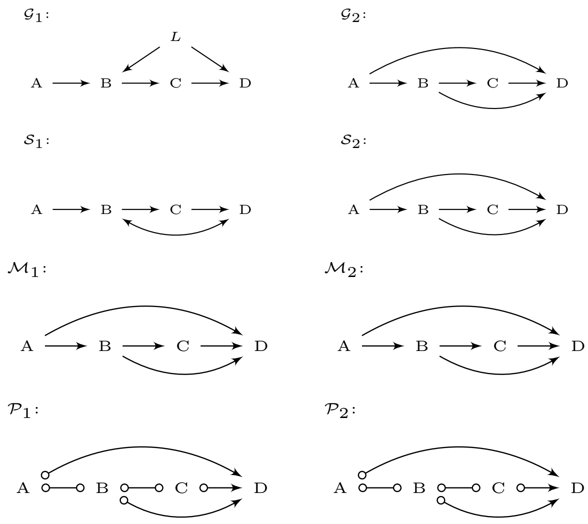

Figure 1: An example of two different DAGs and the corresponding mixed causal graphs over observed variables. From the top: DAGs $\mathcal{G}_1$ over variables $\{A,\ B,\ C,\ D,\ L\}$ (left) and $\mathcal{G}_2$ over variables $\{A,\ B,\ C,\ D\}$ (right). From left to right, on the same row as the underlying causal DAG, the respective SMCMs $\mathcal{S}_1$ and $\mathcal{S}_2$ over $\{A,\ B,\ C,\ D\}$; the respective MAGs $\mathcal{M}_1 = \mathcal{G}_1[_L$ and $\mathcal{M}_2 = \mathcal{G}_2$ over variables $\{A,\ B,\ C,\ D\}$; finally, the respective PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$. Notice that, $\mathcal{M}_1$ and $\mathcal{M}_2$ are identical, despite representing different underlying causal structures.

3 LEARNING THE STRUCTURE OF CAUSAL GRAPHS

As mentioned above, there are two main approaches for learning causal networks from data, constraint-based and score-based. Constraint-based methods estimate from the data which conditional independencies hold, using appropriate tests of conditional independence. Each conditional independence corresponds to an m(d)-separation constraint. Constraint-based algorithms try to eliminate all graphs that are inconsistent with the observed constraints, and ultimately return only the statistically equivalent graphs consistent with all the tests.

Notice that the number of possible conditional independencies are exponential to the number of variables. For graphs that are maximal, i.e. every missing edge corresponds to a conditional independence (DAGs and MAGs but not SMCMs), there exist efficient procedures that can return the skeleton and invariant orientations of the Markov equivalence class of graphs that are consistent with a data set, using only a subset of conditional independencies.

The PC algorithm [Spirtes et al., 2000] is a prototypical, asymptotically correct constraint-based algorithm that identifies the PDAG consistent with a data set. FCI [Spirtes et al., 2000, Zhang, 2008] is the first asymptotically correct constraint-based algorithm that identifies the PAG consistent with a data set. The algorithms work in two stages. The first stage is the skeleton identification stage: Starting from the full graph, the algorithm tries to identify a conditional independence $X \perp\!\!\!\perp Y|\mathbf{Z}$ for each pair of variables X, Y. The corresponding edge is then removed, and the separating set $\mathbf{Z}$ is cached. The second stage is the orientations stage, where the cached conditioning sets are employed to orient the edges.

Given faithfulness, the subset of conditional independencies that have been identified during the skeleton identification stage are sufficient to make all invariant orientation and return the PDAG or PAG that represents the Markov equivalence class of causal graphs that are consistent with all and only the cached conditional independencies. In practice, the orientation stage is sensitive to error propagation [Spirtes, 2010]. Conservative PC (CPC) [Ramsey et al., 2006] proposes a modification of PC that results in more robust orientations: The algorithm performs additional conditional independence tests during the orientation stage, and performs only a subset of robust orientations that are consistent with multiple conditional (in) dependencies. We use the term conservative FCI (CFCI) to describe FCI with the same conservative extension.

Score-based methods on the other hand search over the space of possible graphs trying to maximize a score that reflects how well the graph fits the data. This score is typically related to the likelihood of the graph given the data, $P(D|\mathcal{G})$. For multinomial and Gaussian parametrizations, respectively, BDe[Heckerman et al., 1995] and BGe[Geiger and Heckerman, 1994] are Bayesian scores that integrate over all possible parameters. These scores are employed in most DAG/PDAG-learning algorithms. Other criteria like BIC or MDL can be used to score a graph with specific parameters [Bouckaert, 1995].

The number of possible graphs is super-exponential to the number of variables. For DAGs, efficient search-and-score learning is based on the factorization in Equation 1. Thus, the likelihood of the graph can be decomposed into a product of individual likelihoods of each node given its parents. This can make greedy search very efficient: In each step of the search, all the graphs that occur with single changes of the current graph are considered. Using the score decomposition, one only needs to recompute the scores of nodes that are affected by the change (i.e. the set of parents changes). Unfortunately, the factorization presented in Equation 1 does not apply to probability distributions that are faithful to mixed graphs. This happens because variables connected with bi-directed edges (confounded) are no longer independent given their parents in the graph. Thus, SMCMs and MAGs do not admit a simple factorization, where each node has a single contribution to the likelihood. However, the joint probability of a set of variables $\mathbf{V}$ according to an ADMG $\mathcal{G}$ can be factorized based on sets of variables, called the c-components [Tian and Pearl, 2003], or districts [Richardson, 2009] of the graph. The c-components correspond to the connected components of the bi-directed part of $\mathcal{G}$ (denoted $\mathcal{G}_{\leftrightarrow}$), the graph stemming from $\mathcal{G}$ after the removal of all directed edges.

Parametrizations of the set of distributions obeying the conditional independence relations given by an ADMG are available for multivariate discrete [Richardson, 2009] and multivariate Gaussian distributions [Richardson and Spirtes, 2002, Drton et al., 2009]. Gaussian parametrizations for SMCMs are not always identifiable, but they have been shown to be almost everywhere identifiable for ADMGs without $\overset{\longleftrightarrow}{\longrightarrow}$ edges (called bows in the structural equation model literature) [Brito and Pearl, 2002]. If, in addition, an ADMG does not have almost directed cycles (i.e. is ancestral), the parametrization is everywhere identifiable [Richardson and Spirtes, 2002].

4 GSMAG: GREEDY SEARCH FOR MAXIMAL ANCESTRAL GRAPHS

Let $\mathbf{V} = \{V_i : i = 1, \ldots, V\}$ be a random vector of V variables following a multivariate normal distribution $\mathcal{N}(O, \Sigma)$ with positive definite covariance matrix Σ. Let $\mathcal{G}$ be a MAG. Then graph $\mathcal{G}$ defines a system of linear equations:

$$V_i = \sum_{j \in Pa_{\mathcal{G}}(V_i)} \beta_{ij} V_j + \epsilon_i, \quad i \in \{1, \ldots, V\} \tag{2}$$

Let $\mathbf{B}(\mathcal{G})$ be the collection of all real $V \times V$ matrices $B = (\beta_{ij})$ such that (i) $\beta_{ij} = 0$ when $j \to i$ is not in $\mathcal{G}$, and (ii) $(I - B)$ is invertible. Let $\mathbf{\Omega}(\mathcal{G})$ be all the $V \times V$ matrices $\Omega = (\omega_{ij})$ such that (i) Ω is positive definite and (ii) $\omega_{ij} = 0$ if $j \leftrightarrow i$ is not in $\mathcal{G}$.

Then the system of linear equations (2) can be written as $\mathbf{V} = B\mathbf{V} + \epsilon$, and for $B \in \mathbf{B}(\mathcal{G}), Cov(\epsilon) = \Omega \in \mathbf{\Omega}(\mathcal{G})$ it has a unique solution that is a multivariate normal vector with covariance matrix $\Sigma = (I - B)^{-1}\Omega(I - B)^{-T}$, where the superscript $-T$ denotes the transpose inverse. The family of distributions with covariance matrix in the set $\{\Sigma = (I - B)^{-1}\Omega(I - B)^{-T}\}$ is called the normal linear model associated with $\mathcal{G}$ (symb $\mathbf{N}(\mathcal{G})$). For MAGs, the normal linear model is everywhere identifiable.

Let D be a $V \times N$ matrix of observations for the variables $\mathbf{V}$. Then the empirical covariance matrix is defined as

$$S = \frac{1}{n} DD^T.$$

For $N \geq V + 1$, S is almost surely positive definite. For a MAG $\mathcal{G}$, the log likelihood of the model is

$$\begin{aligned} l_{\mathcal{G}}(B, \Omega|S) = -\frac{N}{2} ln(|2\pi\Omega| - \\ \frac{N-1}{N} tr[(I - B)^T \Omega^{-1} (I - B)S]) \end{aligned} \tag{3}$$

```
input  : Data set $D$ over $\mathbf{V}$ with N samples, tolerance $tol$
output: MAG $\mathcal{G}$, score $sc$
$S \leftarrow corr(D)$;
$\mathcal{G} \leftarrow$ empty graph;
$\mathbf{C} \leftarrow \{V \in \mathbf{V}\}$;
foreach $C_k \in \mathbf{C}$ do
    $s_k \leftarrow$ scoreContrib$(V, 1, N)$;
end
$curScore \leftarrow -2\sum_k s_k + ln(N)(2V + E)$
$minScore \leftarrow curScore$;
repeat
    foreach pair $(X, Y) \in \mathbf{V}$ do
        foreach action in {addLeft, addRight,
        addBidirected, orientLeft, orientRight,
        orientBidirected, remove, reverse} do
            if action is applicable and does not create
            directed or almost directed cycles then
                $(\mathbf{s}', \mathbf{C}', \mathcal{G}') \leftarrow$
                updateScores$(X, Y, action, \mathbf{s}, \mathbf{C}, \mathcal{G}$,tol,N)
                $curScore \leftarrow -2\sum_k s'_k + ln(N)(2V+E)$;
                if $curScore < minScore$ then
                    $(\mathbf{s}, \mathbf{C}, \mathcal{G}) \leftarrow (\mathbf{s}', \mathbf{C}', \mathcal{G}')$;
                end
            end
        end
    end
until no action reduces curScore;
$sc \leftarrow curScore$ ;
```

Algorithm 1: GSMAG

Maximum likelihood estimates $\hat{B}, \hat{\Omega}$ that maximize (3) can be found using the residual iterative conditional fitting (RICF) algorithm presented in Drton et al. [2009], and the corresponding implied covariance matrix is $\hat{\Sigma} = (I - \hat{B})^{-1}\hat{\Omega}(I - \hat{B})^{-T}$.

Based on the factorization of MAGs presented in Richardson [2009], the likelihood can be decomposed according to the c-components of $\mathcal{G}$ as follows [Nowzohour et al., 2015]:

$$l_{\mathcal{G}}(\hat{\Sigma}|S) = -\frac{N}{2}\sum_k \Big(|C_k| ln(2\pi) + ln(\frac{|\Sigma_{\mathcal{G}_k}|}{\prod_{j \in Pa_{\mathcal{G}_k}} \sigma^2_{kj}}) + \frac{N-1}{N} tr[\Sigma^{-1}_{\mathcal{G}_k} S_{\mathcal{G}_k} - |Pa_{\mathcal{G}}(C_k) \setminus \{C_k\}|]\Big), \quad (4)$$

where the $\mathcal{G}_k$ is the graph consisting only of nodes in $C_k \cup Pa_{\mathcal{G}}(C_k)$ without any edges among variables in $Pa_{\mathcal{G}}(C_k) \setminus C_k$, and the subscript $\mathcal{G}_k$ denotes the restriction of a matrix to the rows and columns participating in $\mathcal{G}_k$. σ^2_{kj} denotes the diagonal entry of $\Sigma_{\mathcal{G}_k}$ corresponding to parent node k. The log likelihood is now a sum of c-component scores.

The scoring function is typically the negative log likelihood regularized by a penalty for the number of parameters to avoid over-fitting. The BIC score for MAGs is:

```
input  : Pair $X, Y$, Action $action$, c-components $\mathbf{C}$,
         scores $\mathbf{s}$, MAG $\mathcal{G}$, covariance matrix $S$, tolerance
         tol, sample size $N$
output: c-components $\mathbf{C}'$, scores $\mathbf{s}'$, MAG $\mathcal{G}'$
$\mathcal{G}' \leftarrow$ action( X, Y, $\mathcal{G}$);
if action==(orientBidirected $\vee$ addBidirected) then
    $m \leftarrow m : X \in C_m$;
    $l \leftarrow l : Y \in C_l$;
    $C_m \leftarrow C_m \cup C_l$;
    $\mathbf{C}' \leftarrow \mathbf{C} \setminus C_l$;
    $\hat{\Sigma}_m \leftarrow$ RICF$(\mathcal{G}_m, S_m, tol)$;
    $s'_m \leftarrow$ scoreContrib$(C_m, \Sigma_m, N)$;
end
else if $X \leftrightarrow Y$ in $\mathcal{G}$ then
    $m \leftarrow m : X, Y \in C_m$;
    $\mathbf{C}_{new} \leftarrow$ connectedComponents$(\mathcal{G}'_m)$;
    $\mathbf{C} \leftarrow (\mathbf{C} \setminus \{C_m\}) \cup \mathbf{C}_{new}$;
    foreach $C \in \mathbf{C}_{new}$ do
        $m \leftarrow$ index of $C$ in $\mathbf{C}'$;
        $\hat{\Sigma}_m \leftarrow$ RICF$(\mathcal{G}_m, S_m, tol)$;
        $s'_m \leftarrow$ scoreContrib$(C_m, \Sigma_m, N)$;
    end
end
else
    $m \leftarrow C_m : \mathcal{G}_m \neq \mathcal{G}'_m$;
    $\hat{\Sigma}_m \leftarrow$ RICF$(\mathcal{G}_m, S_m, tol)$;
    $s'_m \leftarrow$ scoreContrib$(C_m, \Sigma_m, N)$;
end
```

Algorithm 2: updateScores

$$BIC(\hat{\Sigma}, \mathcal{G}) = -2ln(l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{S})) + ln(N)(2V + E), \quad (5)$$

where $l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{G})$ is the likelihood of the graph $\mathcal{G}$ with the MLE parameters $\hat{B}, \hat{\Omega}$. BIC is an asymptotically correct criterion for selecting among Gaussian ancestral graph models [Richardson and Spirtes, 2002].

A simple greedy strategy starts from a MAG $\mathcal{G}$ with a score sc and then checks the local neighborhood (i.e. the graphs that stem from the current graph after making a single edge change) for the lowest-scoring network. The algorithm continues this “hill-climbing” until no single edge change reduces the score.

Algorithm 1 begins with the empty graph, where each node is a component. At every subsequent step, every possible edge change is considered: For absent edges, the possible actions are addLeft, addRight, addBidirected. For directed edges, the possible actions are reverse, orientBidirected, remove. For bi-directed edges the possible actions are orientLeft, orientRight, remove.

Score decomposition described in Equation 4 is used to avoid re-fitting the entire MAG. Instead, only the likelihood of the c-components affected by the change need to be re-estimated. Algorithm 1 describes a simple greedy search strategy for learning MAG structure.

Only actions that do not create directed or almost directed cycles are attempted. To efficiently check for cycle creation, a matrix of ancestral relationships[1] of the current MAG is cached. Edge removals can never create directed cycles. Using the cached ancestral matrix, it is straightforward to check whether the addition of a directed edge will create a directed cycle, or if the addition of a bi-directed edge will create an almost directed cycle. Almost directed cycles can also be created when adding directed edges: For each edge $X \leftrightarrow Y$, adding edge $J \rightarrow I$ will create a semi-directed cycle if I is an ancestor if $X(Y)$ and $Y(X)$ is an ancestor of J.

At the end of each iteration, the matrix of ancestral relationships is updated. If a previously missing edge is added, the update takes $O(V^2)$ time. If an edge is removed, the matrix is recomputed using Warshall' s algorithm for transitive closure [Warshall, 1962].

The c-components are only updated in case a bi-directed edge is added or altered in any way. When adding a bi-directed edge, the corresponding c-components of the end-points are merged if separate. When an existing bi-directed edge is removed (completely or becomes directed), the corresponding c-component C_k is divided in the new connected components. The scores of the affected components are recomputed using new RICF estimates. In any other case, the c-components remain the same, and the score of the c-component whose corresponding graph $\mathcal{G}_k$ is affected by the change is recomputed. This procedure is described in Algorithm 2.

When no single-edge change improves the current score, the algorithm terminates and the current network is returned. Greedy hill-climbing procedures can be stuck in local optima (minima). To tackle this problem, they are often augmented with meta-heuristics such as random restarts, TABU lists or simulated annealing. For the scope of this work we use no such heuristic. In preliminary experiments, however, we found that augmenting Algorithm 1 with a TABU heuristic did not significantly improve performance.

5 RELATED WORK

Several constraint-based algorithms exist for learning a Markov equivalence class of MAGs from an observational data set: FCI [Spirtes et al., 2000, Zhang, 2008] is a sound and complete algorithm that returns the complete PAG. RFCI Colombo et al. [2012] and FCI+[Claassen et al., 2013] are modifications of FCI that try to avoid the computationally expensive possible d-separating stage in the skeleton search of FCI. Conservative FCI [Ramsey et al., 2006] is a modification of FCI that makes fewer, but more robust orientations, to avoid error propagation.

Nowzohour et al. [2015] propose a greedy search with random restarts for learning "bow-free" ADMGs from data and introduce the score decomposition showed in Equation 4. Since Markov equivalence for ADMGs that are not MAGs has not yet been characterized, they use a greedy strategy for obtaining the *empirical* Markov equivalence class, based on score similarity. The authors use the estimated ADMGs to compute causal effects and show promising results. However, since they do not necessarily find maximal ancestral graphs, they do not compare against constraint-based methods or evaluate the accuracy of the learnt graphs.

Marginalizing out variables from causal DAGs results in some additional equality constraints that are not conditional independencies. Nested Markov models Shpitser et al. [2013] extend SMCMs and are used to also model these additional constraints. Shpitser et al. [2012] use a penalized likelihood score and a greedy search with TABU list to identify a nested Markov model from discrete data.

6 COMPARISON OF GSMAG WITH FCI, CFCI

We compared the performance of Algorithm 1 against FCI and CFCI in simulated data. We simulated 100 random DAGs over 10, 20 and 50 variables. To control the sparseness of the DAGs, we set the maximum parents of each node. We present results for sparse networks, where each variable was allowed to have up to 3 parents, and denser networks where each variables was allowed to have up to 5 parents. For each DAG, 10% of the variables were marginalized (1, 2 and 5 variables respectively). The ground truth PAG $\mathcal{P}_{GT}$ was then created for each marginal DAG.

Data sets with 100, 1000 and 5000 samples were simulated for each DAG and random parameters with absolute values in $\{0.1, 0.9\}$. The corresponding marginal data sets were input in Algorithm 1, FCI and CFCI. FCI and CFCI were run with a significance threshold of 0.05 and a maximum conditioning size 5. Algorithm 1 outputs a MAG. To compare the outputs of the algorithms, the corresponding PAG

[1] Changing edge orientation is equivalent to a removing the edge and then adding it re-oriented. To test for possible cycles efficiently, a matrix of all the non-trivial ancestral relationships (more than one variable in the path) is also cached. Reversing an edge $X \rightarrow Y$ creates a directed cycle only if X is a non-trivial ancestor of Y.

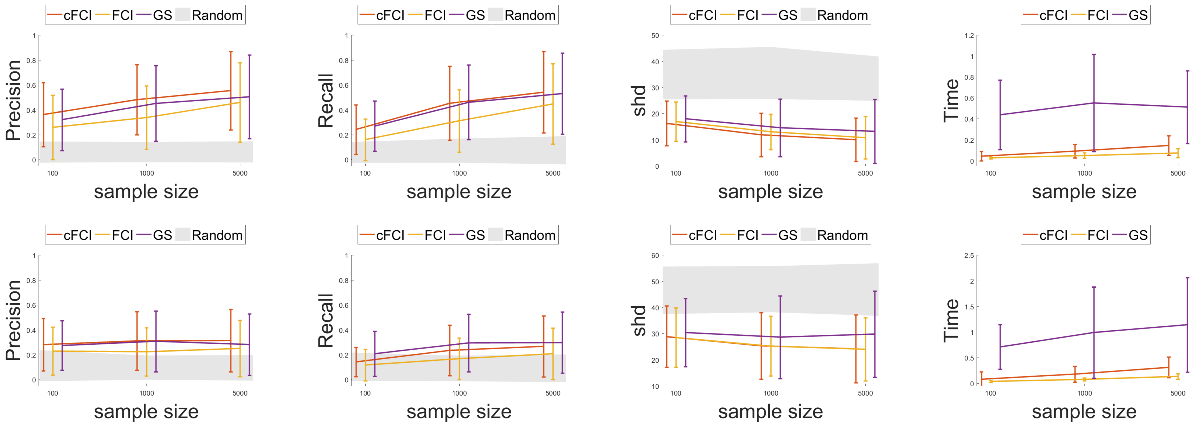

Figure 2: Performance of FCI, CFCI and GSMAG for networks with 9 observed variables (top) 3 maximum parents per variables (bottom) 5 maximum parents per variable.

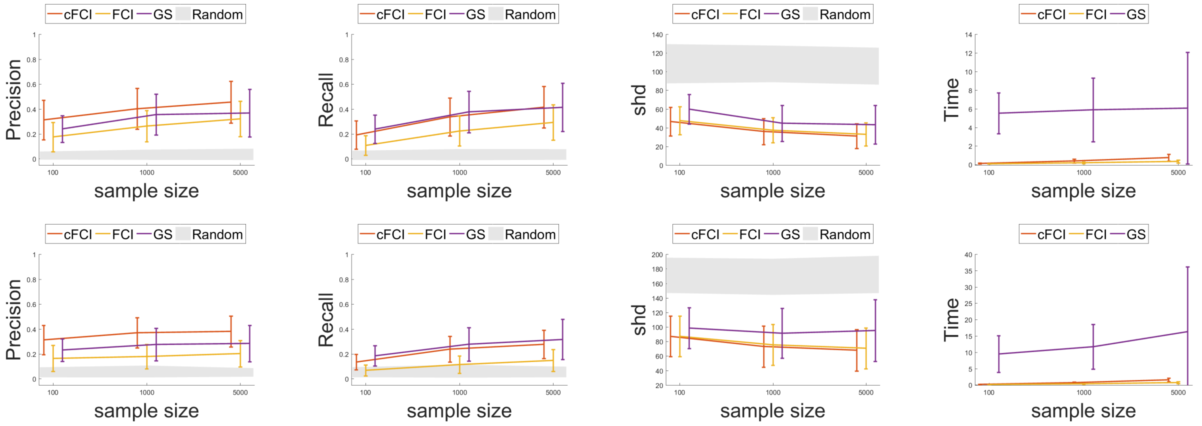

Figure 3: Performance of FCI, CFCI and GSMAG for networks with 18 observed variables (top) 3 maximum parents per variable (bottom) 5 maximum parents per variable.

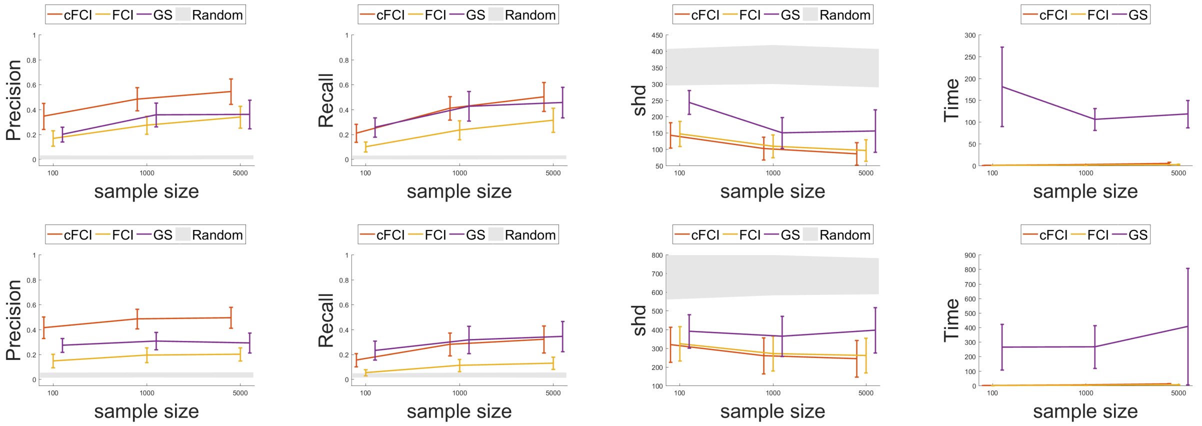

Figure 4: Performance of FCI, CFCI and GSMAG for networks with 45 observed variables (top) 3 maximum parents per variable (bottom) 5 maximum parents per variable.

was created for each MAG output. We use $\mathcal{P}_{FCI}, \mathcal{P}_{CFCI}$ and $\mathcal{P}_{GS}$ to denote the outputs of FCI, CFCI and Algorithm 1, respectively.

Summarizing PAG differences is not trivial, and many dif-

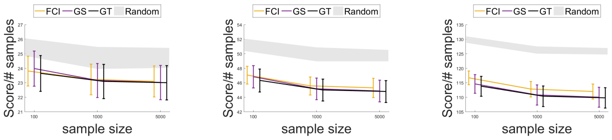

Figure 5: Score divided by number of samples for FCI, GS and the ground truth network for sparse networks (3 maximum parents) for 9(left), 18(middle) and 45(right) observed variables.

ferent approaches are used in the literature. As a general metric of how different two PAGs are, we use the structural hamming distance (shd) for PAGs, defined as follows: Let $\hat{\mathcal{P}}$ be the output PAG and $\mathcal{P}$ be the ground truth PAG. For each change (edge addition, edge deletion, change arrowhead, change tail) required to transform $\hat{\mathcal{P}}$ into $\mathcal{P}$, shd is increased by 1.

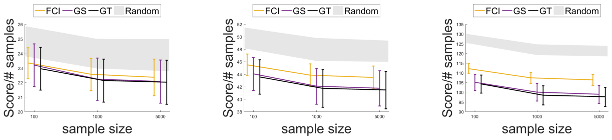

Figure 6: Score divided by number of samples for FCI, GS and the ground truth network (5 maximum parents) for 9(left), 18(middle) and 45(right) observed variables.

We also use precision and recall, as described in Tillman and Spirtes [2011]: Precision is defined as the number of edges in the output PAG with the correct orientations, divided by the number of edges in the output PAG. Recall is defined as the number of edges in the output PAG with correct orientations, divided by the number of edges in the ground truth PAG. These metrics are very conservative, since they penalize even small differences. For example, an edge that is $\rightarrow$ in the ground truth but $\circ\!\!\!\rightarrow$ in the output PAG will be classified as a false positive.

Figures 2, 3, and 4 show the performance results for networks of 10, 20 and 50 variables, respectively. Mean values over 100 iterations are presented for all experiments. All algorithms perform better in sparser networks.

Greedy search has larger structural hamming distances than FCI and CFCI. More specifically, out of 900 cases (over all variable and sample sizes), FCI outperforms GSMAG in 657 cases for sparse networks and in 715 cases for dense networks, while CFCI outperforms GSMAG in 682 cases for sparse networks and in 710 cases for dense networks. In terms of precision, CFCI is again the best of the three (outperforms GSMAG in 601 and 659 out of 900 cases for sparse and dense networks, respectively). FCI has the poorest precision: it outperforms GSMAG in 318 and 221 cases for sparse and dense networks, respectively). Finally, GSMAG has the best recall out of all algorithms, with CFCI being second. Specifically, terms of recall, FCI outperforms GSMAG in 139 and 83 cases, while CFCI outperforms GSMAG in 357 and 249 cases for sparse networks and dense networks, respectively. Naturally, greedy search is much slower than both conservative and plain FCI.

Intriguingly, GSMAG's performance declines for the largest attempted sample size (5000 samples), particularly for larger networks. This happens because greedy search tends to include many false positive edges. It is possible that this is related to the naive greedy search, and could be improved by augmenting some kind of heuristic for escaping local minima, or by adjusting the scoring criterion.

Figure 6 shows the score of the output MAG for Algorithm 1 and the ground truth MAG. To compare also with FCI, we used the method presented in Zhang [2008] to obtain a MAG from $\mathcal{P}_{FCI}$. Notice that this cannot be applied to the output of CFCI, since it is not a complete PAG (due to unfaithful triplets). Greedy search typically scores closer to the ground truth, particularly for denser networks.

7 FUTURE WORK

We present an implementation of a greedy search algorithm for learning MAGs from observations, and compare it to FCI and CFCI. To the best of our knowledge, this is the first comparison of score-based and constraint-based search in the presence of confounders.

The algorithm uses the decomposition presented in Nowzohour et al. [2015] for bow-free SMCMs. Compared to SMCMs, MAGs are less expressive in terms of causal statements. However, since they have no almost directed cycles, fitting procedures for obtaining maximum likelihood estimates always converge. Semi-Markov causal models that are Markov equivalent to the output MAG could be identified as a post-processing step.

Heuristic procedures for escaping local minima could also be explored, to improve the performance of GSMAG. Algorithm efficiency could also possibly be improved by updating without recomputing inverse matrices and where applicable.

Other interesting directions include taking weighted averages for specific PAG features, or using both constraint-based and score-based techniques for hybrid learning. Greedy search in the space of PAGs instead of MAGs could also be explored, since a transformational characterization for Markov equivalent MAGs exists [Zhang and Spirtes, 2005].

Acknowledgments

This work was funded by European Research Council (ERC) and is part of the CAUSALPATH - Next Generation Causal Analysis project, No 617393.

References

RA Ali, TS Richardson, and P Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B): 2808–2837, October 2009.

RR Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, University of Utrecht, 1995.

C Brito and J Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002.

T Claassen, JM Mooij, and T Heskes. Learning sparse causal models is not NP-hard. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence*, 2013.

D Colombo, MH Maathuis, M Kalisch, and TS Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 02 2012.

M Drton, M Eichler, and TS Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *The Journal of Machine Learning Research*, 10:2329–2348, 2009.

D Geiger and D Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, 1994.

D Heckerman, D Geiger, and DM Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

C Nowzohour, M Maathuis, and P Bühlmann. Structure learning with bow-free acyclic path diagrams. *arXiv preprint arXiv:1508.01717*, 2015.

J Pearl. *Causality: Models, Reasoning and Inference*, volume 113 of *Hardcover*. Cambridge University Press, 2000.

J Ramsey, P Spirtes, and J Zhang. Adjacency faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

TS Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.

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

I Shpitser, TS Richardson, JM Robins, and R Evans. Parameter and structure learning in nested markov models. In *In UAI (Workshop on Causal Structure Learning)*, 2012.

I Shpitser, R Evans, TS Richardson, and JM Robins. Sparse nested Markov models with log-linear parameters. In *Proceedings of the 29h Conference on Uncertainty in Artificial Intelligence*, 2013.

P Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11:1643–1662, 2010.

P Spirtes, C Glymour, and R Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, January 2000.

J Tian and J Pearl. On the identification of causal effects. Technical Report R-290-L, UCLA Cognitive Systems Laboratory, 2003.

RE Tillman and P Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

I Tsamardinos, LE Brown, and CF Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

S Warshall. A theorem on boolean matrices. *J. ACM*, 9(1): 11–12, 1962.

J Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.

J Zhang and P Spirtes. A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, pages 667–674, 2005. cited By 8.