Contents lists available at ScienceDirect
International Journal of Approximate Reasoning
www.elsevier.com/locate/ijar

On scoring maximal ancestral graphs with the Max–Min Hill Climbing algorithm

Konstantinos Tsirlis [a,b], Vincenzo Lagani [c,d], Sofia Triantafillou [e], Ioannis Tsamardinos [a,d]

[a] Department of Computer Science, University of Crete, Voutes Campus, 70013 Heraklion, Greece
[b] The Wellcome Center for Human Neuroimaging, University College London, 12 Queen Square, WC1N 3BG, London, United Kingdom
[c] Institute of Chemical Biology, Ilia State University, 3/5, Kakuca Cholokashvili Ave, Tbilisi, Georgia
[d] Gnosis Data Analysis PC, Science and Technology Park of Crete, N. Plastira 100, Vassilika Vouton, 70013 Heraklion, Greece
[e] Department of Bioengineering, University of Pennsylvania, 106 Hayden Hall, 240 S 33rd St, Philadelphia, PA 19104, United States of America

ARTICLE INFO

Article history:
Received 12 December 2017
Received in revised form 30 July 2018
Accepted 3 August 2018
Available online xxxx

Keywords:
Hybrid causal structure learning
Maximal ancestral graphs
Max–Min Hill Climbing
MMPC

ABSTRACT

We consider the problem of causal structure learning in presence of latent confounders. We propose a hybrid method, MAG Max–Min Hill-Climbing (M^3HC) that takes as input a data set of continuous variables, assumed to follow a multivariate Gaussian distribution, and outputs the best fitting maximal ancestral graph. M^3HC builds upon a previously proposed method, namely GSMAG, by introducing a constraint-based first phase that greatly reduces the space of structures to investigate. On a large scale experimentation we show that the proposed algorithm greatly improves on GSMAG in all comparisons, and over a set of known networks from the literature it compares positively against FCI and cFCI as well as competitively against GFCI, three well known constraint-based approaches for causal-network reconstruction in presence of latent confounders.

1. Introduction

Learning causal networks from observational data has been the subject of intense research for at least the last three decades [22]. A notable result obtained so far is that the joint probability distribution of the data can be linked to (a family of) possible causal structures by assuming faithfulness and the causal Markov condition.

Direct Acyclic Graphs (DAGs) are one of the simplest types of causal graphs. In DAGs nodes denote the measured variables while edges indicate direct causal relationship in the form $A \to B$. Several algorithms have been proposed for learning DAGs, most of them falling under two categories, i.e., score-based and constraint-based methods. The first ones attempt to identify the causal structure that optimize a criterion of fitness on the data; the latter use a graphical criterion (d-separation) to constrain the causal structure according to the conditional (in)dependencies that hold in the data. Hybrid methods also exist, as for example the Max–Min Hill-Climbing algorithm (MMHC, [26]). MMHC first utilizes tests of conditional independence to restrict the search space and afterwards uses greedy search to identify the structure that optimizes a scoring function.

Regardless of the approach, usually more than one DAG can represent equally well the same set of relations in a distribution, forming a Markov equivalence class. Complete Partially Directed Acyclic Graphs (CPDAGs) represent Markov equivalence classes by introducing indirect edges, which cannot be oriented on the basis of the information contained in the data.

E-mail addresses: konstantinos.tsirlis@gmail.com (K. Tsirlis), vincenzo.lagani@iliauni.edu.ge (V. Lagani), sof.triantafillou@gmail.com (S. Triantafillou), tsamard.it@gmail.com (I. Tsamardinos).

https://doi.org/10.1016/j.ijar.2018.08.002
0888-613X/© 2018 Published by Elsevier Inc.

DAGs' main drawback is the inability of representing latent variables, i.e., they assume causal sufficiency, an assumption that rarely holds in real world scenarios. Maximal Ancestral Graphs (MAGs) [19] are an extension of DAGs that can represent the presence of latent confounders and selection bias. In this work, we assume no selection bias. MAGs' causal semantics are more complicated than DAGs': directed edges denote causal ancestry, but the relation is not necessarily direct, and bi-directed edges denote the existence of confounding. The most attractive properties of MAGs is that they are closed under marginalization, and every non-adjacency in the graph corresponds to a conditional independence in the probability distribution. Partial Ancestral Graphs (PAGs) represent MAGs Markov equivalence classes by using a special notation (∘) for edge endpoints that cannot be determined.

Until recently, only constraint-based methods existed for learning causal structures in presence of confounding. These algorithms use an extension of d-separation, namely m-separation, that holds also in presence of hidden variables. FCI [22, 27], is the first asymptotically correct constraint-based algorithm for MAGs (PAGs). This algorithm works in two stages. The first one attempts to identify the undirected skeleton of the graph, while the subsequent orientation phase uses a set of rules for orienting the edges. In practice, the orientation phase is prone to error propagation [21]. Conservative FCI [17] is an extension of FCI that performs additional conditional independence tests during the orientation phase and output only robust orientations that are consistent with all tested conditional independences.

Triantafillou and Tsamardinos [25] proposed a score-based algorithm able to deal with confounding, namely the Greedy Search for MAGs (GSMAG). This algorithm builds upon a score function for mixed graphs [12], coupled with a simple greedy search upon the space of all possible MAGs. The Greedy Fast Causal Inference (GFCI, [13]) uses a different strategy, where a first approximation of the causal graph is obtained using FGES [18], a score-based method that ignores latent variables and then FCI orientation rules are used for identifying possible confounding, as well removing some of the edges added by FGES.

In this paper we move along the lines of Triantafillou and Tsamardinos [25], by adapting the MMHC algorithm mentioned above to be able to incorporate latent confounders. The resulting algorithm, M^3HC (MMHC for MAGs) is contrasted against GSMAG and against FCI, cFCI and GFCI on simulated data. The results of the experiments show that M^3HC consistently outperforms or is on par with GSMAG, while it often provides improved performances with respect to all FCI-based algorithms over known networks. Code for reproducing our results will be made available at https://github.com/mensxmachina.

2. The M^3HC algorithm

In this section we present the M^3HC algorithm. Let $\mathcal{D}$ be a dataset containing N i.i.d. samples measured over $\mathbf{V}$ continuous variables, jointly following a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ with positive definite covariance matrix Σ. Our objective is to identify a MAG $\mathcal{G}$ that belongs to the Markov equivalence class of MAGs representing the data at hand. According to Richardson [19], the Bayesian Information Score (BIC) is an asymptotically correct criterion for scoring MAGs:

$$BIC(\hat{\Sigma}, \mathcal{G}) = -2ln(l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{G})) + ln(N)(2|\mathbf{V}| + |\mathbf{E}|), \quad (1)$$

where $l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{G}))$ is $\mathcal{G}$ likelihood, while $|\mathbf{V}|$ and $|\mathbf{E}|$ are the number of vertices and edges $\mathbf{E}$ in $\mathcal{G}$, respectively. The matrix $\hat{\Sigma}$ is the maximum likelihood estimate of the covariance matrix Σ according to the normal linear model associated with $\mathcal{G}$. Briefly, the MAG $\mathcal{G}$ defines a set of linear equations $\mathbf{V} = B \cdot \mathbf{V} + \epsilon$, where (a) $B = \{\beta_{ij}\}$ is a matrix such that $(I - B)$ is invertible and $\beta_{ij} = 0$ if $j \to i$ is not in $\mathcal{G}$ and (b) ϵ is a vector of random errors such that $\Omega = Cov(\epsilon) = \{\omega_{ij}\}$ is positive definite and $\omega_{ij} \neq 0$ if $j \leftrightarrow i$. Given the empirical covariance matrix S, maximum likelihood estimates $\hat{B}$ and $\hat{\Omega}$ can be identified by maximizing the log-likelihood $l_{\mathcal{G}}(B, \Omega|\Sigma)$ with the residual iterative conditional fitting (RICF) algorithm provided by Drton et al. [7]. The corresponding covariance matrix is computed as $\hat{\Sigma} = (I - \hat{B})\hat{\Omega}((I - \hat{B})^{-1})^T$.

A general score decomposition property is known from Tian et al. (2005) [23] and Richardson et al. (2009) [20]. They show that $l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{G})$ can be decomposed according to the c-components of $\mathcal{G}$. The c-components of a graph $\mathcal{G}$ correspond to the connected components of its bi-directed part, i.e., the graph stemming from $\mathcal{G}$ after the removal of all directed edges.

For the special case of the Gaussian parametrization Nowzohour et al. (2015) demonstrated that the likelihood of $\mathcal{G}$ becomes $l_{\mathcal{G}}(\hat{\Sigma}|\mathcal{G})) = -\frac{N}{2}\sum_k s_k$, where each s_k correspond to a c-component and can be computed with the closed formula [12]:

$$s_k = |C_k| \cdot ln(2\pi) + ln(\frac{|\hat{\Sigma}_{\mathcal{G}_k}|}{\prod_{j \in Pa_{\mathcal{G}_k}} \sigma^2_{kj}}) + \frac{N-1}{N} \cdot tr\left[\hat{\Sigma}^{-1}_{\mathcal{G}_k} S_{\mathcal{G}_k} - |Pa_{\mathcal{G}}(C_k) \setminus \{C_k\}|\right] \quad (2)$$

where C_k are the nodes in the c-component k, while $Pa_{\mathcal{G}}(C_k)$ are the parents of nodes in C_k, $\mathcal{G}_k$ represents the marginalization of $\mathcal{G}$ to nodes in $C_k \cup Pa_{\mathcal{G}}(C_k)$, and σ^2_{kj} is the diagonal entry of $\hat{\Sigma}_{\mathcal{G}_k}$ that corresponds to the parent node k.

Algorithm 1 shows M^3HC operation. The first phase of the algorithm ("skeleton identification") identifies the set $\mathbf{P}$ of variable pairs that are allowed to be directly connected by an edge in the final graph. In other words, the final MAG $\mathcal{G}$ can have an edge between variables $\{X, Y\}$ only if $\{X, Y\} \in \mathbf{P}$. In the second phase the empty graph is first set as initial solution and its score as current score. The algorithm then proceed in identifying for each pair of variables $\{X, Y\} \in \mathbf{P}$ the action that improves the score the most. Possible actions include adding an edge pointing at X, Y or both (*addLeft*, *addRight* and