Greedy Equivalence Search in the Presence of Latent Confounders - Supplement

Tom Claassen[1] **Ioan Gabriel Bucur**[1]

[1]Institute for Computing and Information Sciences, Radboud University, Nijmegen, (The) Netherlands

Abstract

This article is the supplement to the UAI 2022 paper ‘Greedy Equivalence Search in the Presence of Latent Confounders’. It contains all proofs to the lemmas in the main paper, as well as additional details and background information. Numbering is consistent with the main paper.
Software implementation (Matlab) of all code and experimental settings publicly available via `https://github.com/tomc-ghub/gps_uai2022`.

A REMARK ON SIZE OF MECS

One may wonder whether searching between equivalence classes is actually worth the trouble, given the famous conclusion from Gillispie and Perlman (2002) that the average size of equivalence classes for DAGs is bounded below 4, even as n goes to infinity. This was all the more surprising given that experimental findings from e.g. (Chickering, 2002) reported encountering huge sized equivalence classes.

As demonstrated by He et al. (2015), the main contribution to this bound comes from graphs with a high average density of around $n/2$ that account for the vast majority of graphs over n nodes, and for which nearly every instance is almost fully determined. But for sparse graphs with a density bounded by some constant $d \ll n$ the size of each individual equivalence class can become truly huge as n gets larger. For example (He et al., 2015) report an average equivalence class size of $3.5e19$ for DAGs over 50 nodes with average edge density of 4. Therefore despite some potential overhead, searching over equivalence classes rather than individual MAGs can still bring a sizeable improvement in efficiency.

B PROOFS

Below the proof details for the theoretical results in the main paper.

Lemma 2 *In a MAG $\mathcal{G}$, a triple $\langle a,b,c \rangle$ is in $\mathfrak{C}_i$ (resp. $\mathfrak{D}_i$), if and only if $\langle a,b,c \rangle \in \mathfrak{T}_i$ and $\langle a,b,c \rangle$ is a collider (resp. noncollider) in $\mathcal{G}$.*

Proof Clearly the definitions coincide for triples of order 0. First from old to new: if $\langle a,b,c \rangle \in \mathfrak{T}_1$ then there is a discriminating path $\langle x,a,b,c \rangle$ in $\mathcal{G}$ for which $\langle x,a,b \rangle$ is a collider triple with order 0, hence $\langle x,a,b \rangle \in \mathfrak{C}_0$, and $\langle x,a,c \rangle$ is a noncollider triple with order 0, $\langle x,a,c \rangle \in \mathfrak{D}_0$. Therefore all conditions for order $i = 1$ in the new definition are satisfied, and so $\langle a,b,c \rangle \in \mathfrak{C}_1$ resp. $\mathfrak{D}_1$, depending on whether the triple is a collider or noncollider in $\mathcal{G}$. By induction, suppose the mapping is valid up to order i, and let $\langle a,b,c \rangle \in \mathfrak{T}_{i+1}$. Then there is a discriminating path $\langle x,q_1,..,q_p,a,b,c \rangle$ in $\mathcal{G}$ for which $\langle q_p,a,b \rangle$ is a collider triple with order $k \leq i$, hence $\langle q_p,a,b \rangle \in \mathfrak{C}_k$, and for which $\langle q_p,a,c \rangle$ is a noncollider triple with order $j \leq i$, hence $\langle q_p,a,c \rangle \in \mathfrak{D}_j$. Therefore all conditions for order $i+1$ in the new definition are satisfied, and so $\langle a,b,c \rangle \in \mathfrak{D}_{i+1}$ resp. $\mathfrak{C}_{i+1}$), again depending on whether the triple is a noncollider or collider in $\mathcal{G}$.

For the reverse, from new to old: at order $i = 1$, if $\langle a,b,c \rangle \in \mathfrak{D}_1$ then by definition there is a $\exists x : \langle x,a,c \rangle \in \mathfrak{D}_0$ as noncolllider triple, and also as collider triple $\langle x,a,b \rangle \in \mathfrak{C}_0$. But that implies $\langle x,a,b,c \rangle$ is a discriminating path in $\mathcal{G}$, and so $\langle a,b,c \rangle \in \mathfrak{T}_1$ as we already saw $\langle x,a,b \rangle \in \mathfrak{T}_0$. Similarly when $\langle a,b,c \rangle \in \mathfrak{C}_1$. Again by induction assuming the mapping is valid up to order i, and let $\langle a,b,c \rangle \in \mathfrak{D}_{i+1}$. Then $\exists q_p : \langle q_p,a,c \rangle \in \mathfrak{D}_{j \leq i}$ and $\langle q_p,a,b \rangle \in \mathfrak{C}_{k \leq i}$. If $j > 0$, then again there is a $q_{p-1} : \langle q_{p-1},q_p,c \rangle \in \mathfrak{D}_{m<j}$ and $\langle q_{p-1},q_p,a \rangle \in \mathfrak{C}_{n<k}$. The same holds for all subsequent triples until we arrive at some triple with order 0 for which $\langle x,q_1,c \rangle \in \mathfrak{D}_0$ and $\langle x,q_1,q_2 \rangle \in \mathfrak{C}_r$. Then $\langle x,q_1,..,q_p,a,b,c \rangle$ is a

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

discriminating path, where all required collider triples are of lower order than i and so also in $\bigcup \mathfrak{C}_{j<i}$. This implies $\langle a, b, c\rangle \in \mathfrak{T}_i$, which proves the lemma. ■

We can store triples $\langle a, b, c\rangle$ as value c stored in a list at entry (a, b) in an $N \times N$ array. For sparse graphs with node degree bounded by d, each entry has at most d such entries, meaning that when searching for a matching triple for, say, collider $\langle a, b, c\rangle$, we do not need to scan the full noncollider list $\mathcal{D}$, but only at most d such entries at the corresponding index (a, b) for list $\mathcal{D}$.

Corollary 3 *Two MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.*

Proof Lemma 2 implies a MEC $\mathcal{M}(\mathcal{G})$ is unique and in a one-to-one correspondence with Lemma 1 which guarantees 'if and only if' Markov equivalence. ■

In order to prove Lemma 4, we first prove the soundness of rule $\mathcal{R}4'$ when applied to the core PAG from definition 4:

$\mathcal{R}4'$: Let Z be a district among the parents of a node y. If $x \ast\!\!\rightarrow z \longrightarrow y$, with $z \in Z$ and x and y not adjacent, then orient all $u \circ\!\!\rightarrow y$ with $u \ast\!\!\rightarrow z'$ for some $z' \in Z$ (possibly $z' = z$) as $u \longrightarrow y$.

Lemma 1 *When applied to a (not necessarily completed) PAG $\mathcal{P}$ that contains all invariant marks of the core PAG, rule $\mathcal{R}4'$ is sound.*

Proof We prove the triggering condition implies the existence of a discriminating path for u which means the mark at u on the edge to y must be invariant. Then we note that an invariant arrowhead at u would already have been oriented in the core PAG, which implies any remaining circle mark must become $u \longrightarrow y$.

Firstly, if $Z = \{z_1, .., z_n\}$ is a district among parents of y in $\mathcal{P}$, then all $z_i \in Z$ have $z_i \longrightarrow y$ in $\mathcal{P}$, and all z_i are connected among each other by a sequence of one or more bidirected edges. Suppose $\mathcal{R}4'$ applies with $z = z_1$ and $z' = z_k$ (possibly $z_1 = z_k$). Then there is a path $x \ast\!\!\rightarrow z_1 (\longleftrightarrow ..z_i... \longleftrightarrow z_k) \leftarrow\!\!\ast\, u \ast\!\!\rightarrow y$ in $\mathcal{P}$. This path is also a discriminating path for u, as it contains at least three edges, x is not adjacent to y, and every vertex $\langle z_1, .., z_k\rangle$ is both collider on this path and also a parent of y. That means standard FCI orientation rule $\mathcal{R}4$ applies, and so the triple $\langle z_k, u, y\rangle$ is either an invariant collider of the form $z_k \longleftrightarrow u \longleftrightarrow y$, or an invariant noncollider$z_k \leftarrow\!\!\ast\, u \longrightarrow y$. But if it was an invariant collider, then by Lemma 1 the arrowhead $u \leftarrow\!\!\ast\, y$ must have been part of *some* collider with order (otherwise there would be two MAGs that are not Markov equivalent with the same skeleton and colliders with order). But this does not mean that the triple $z_k \ast\!\!\rightarrow u \leftarrow\!\!\ast\, y$ itself is necessarily a (collider) triple with order, as definition 1 only implies that every higher order triple with order corresponds to a discriminating path, but not the other way around.

As a result, it is possible that there is a discriminating path for triple $z_k \ast\!\!-\!\!\ast\, u -\!\!\ast\, y$, where u is an invariant noncollider along the path, but where $\langle z_k, u, y\rangle$ is *not* a triple with order. There is no guarantee that in that case the edge $u -\!\!\ast\, y$ would be part of some other noncollider triple $\langle \ast, u, y\rangle$ with order ≥ 1 (as Lemma 1 only relates to colliders with order), and hence the invariant tail mark $u -\!\!\ast\, y$ is not necessarily present in the core PAG. But that also means that if we encounter a discriminated node that has not obtained an explicit edge mark in the core PAG, then it must be noncollider along that discriminating path, and hence get oriented as $u \longrightarrow y$ in the completed PAG. ■

The reader will notice that the rule $\mathcal{R}4'$ definition via 'district among parents' applies to discriminated nodes in general, and indeed the standard FCI orientation rule $\mathcal{R}4$ can be implemented in the same way, without having to look for specific discriminating paths, at a significant increase in processing speed.

Lemma 4 *For a valid MEC $\mathcal{M}$, algorithm 2 will output the corresponding completed PAG $\mathcal{P}$.*

Proof (Rules following the notation in (Zhang, 2008).) Given the core PAG, all *v*-structures from rule $\mathcal{R}0$ are already included. In the eliminated discriminating path rule $\mathcal{R}4$, for the final 3 nodes $\langle .., \alpha, \beta, \gamma\rangle$ along a discriminating path all invariant edge marks at β on the edge to γ are also already covered in the core PAG via triples with order $k \geq 1$.

All other elements oriented by rule $\mathcal{R}4$ will get oriented by $\mathcal{R}2$. In particular: both branches of $\mathcal{R}4$ will also orient an arrowhead at γ on the edge to β, but this also follows directly from the second case triggering $\mathcal{R}2$, as $\langle \alpha, \beta, \gamma\rangle$ together with already established arc $\alpha \rightarrow \gamma$ satisfy the precondition for $\mathcal{R}2$ with the roles of α and β reversed, leading to the invariant arrowhead $\beta \ast\!\!\rightarrow \gamma$. For the remaining arrowhead orientation at $\alpha \ast\!\!\rightarrow \beta$ from the second branch of rule $\mathcal{R}4$, the final three nodes also satisfy the first precondition for $\mathcal{R}2$, except now with the roles of β and γ reversed.

All other individual orientation rules remain sound, so that all other rules triggered in creating the PAG by FCI can/will also be triggered when starting from the MEC, which means the output PAG is also sound and complete. ■

C SCORING MECS

This section describes the details behind the BIC score for MAGs (Richardson and Spirtes, 2002), used to score MECs as indicated in section 5.1.

To connect a MAG to a linear Gaussian model, we can associate a MAG $\mathcal{G}$ over $n = |\mathbf{V}|$ variables with a collection of $n \times n$ matrices of structural parameters $\mathbf{B}(\mathcal{G})$, with $B_{ij} = 0$ iff $i = j$ or $j \rightarrow i \notin \mathcal{G}$, and a collection of positive definite covariance matrices of error/noise terms $\mathbf{\Omega}(\mathcal{G})$, where $\Omega_{ij} = 0$ iff $i \neq j$ and $i \longleftrightarrow j \notin \mathcal{G}$. Then the system of (normal) linear equations $\mathbf{V} = B\mathbf{V} + \boldsymbol{\epsilon}$ with $B \in \mathbf{B}(\mathcal{G})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Omega \in \mathbf{\Omega}(\mathcal{G}))$ implies a multivariate Gaussian distribution over $\mathbf{V}$ with covariance matrix $\Sigma = (I - B)^{-1}\Omega(I - B)^{-\mathrm{T}}$.

For any given choice of B and Ω we can compute the likelihood of the observed sample covariance matrix S. But for a given MAG $\mathcal{G}$ we only have the structure, not the parameters. As a reasonable approximation, for a given graph $\mathcal{G}$ we therefore compute the parameters that maximize this likelihood. For DAGs this boils down to straightforward regression, but for MAGs in general no such expression exists, even though they are uniquely identifiable. Instead we can employ the *residual iterative conditional fitting* (RICF) method developed by Drton and Richardson (2008) which iteratively finds the maximum likelihood solution for the parameters in the model given the graph $\mathcal{G}$ and observed sample covariance matrix S, and outputs the implied covariance matrix $\hat{\Sigma}$, from which we can compute the (log) likelihood of the sample covariance matrix S under the model covariance $\hat{\Sigma}$ for $\mathcal{G}$.

An attractive property, as shown by Nowzohour et al. (2017), is that this log-likelhood can be decomposed into a sum of distinct contributions over the separate districts (connected bidirected components) in the graph $\mathcal{G}$. With each district D_k a so-called *c-component* C_k is associated, consisting of the subgraph $\mathcal{G}_k$ of $\mathcal{G}$ over the nodes in $D_k \cup \mathrm{pa}_{\mathcal{G}}(D_k)$, but with all edges between $\mathrm{pa}_{\mathcal{G}}(C_k) \equiv \mathrm{pa}_{\mathcal{G}}(D_k) \setminus D_k$ removed. With this the log-likelihood given N samples becomes:

$$l(S|\hat{\Sigma}_{\mathcal{G}}) = -\frac{N}{2}\sum_k \Big(|C_k| \log 2\pi + \log \frac{|\Sigma_{\mathcal{G}_k}|}{\prod_{j \in \mathrm{pa}(C_k)} \sigma^2_{kj}} + \frac{N-1}{N} \operatorname{tr}(\Sigma^{-1}_{\mathcal{G}_k} S_{\mathcal{G}_k} - |\mathrm{pa}(C_k)|) \Big) \quad (1)$$

As a result, when computing the score for a modified MEC we only need to recompute the score for the c-components that changed relative to the source MEC, providing a significant speed improvement for the overall computational cost. Note that here the use of the arc-augmented MAG extension for a PAG minimizes the size of the districts, which also benefits the speed and convergence of the RICF step for each district in the computation of the score.

To avoid overfitting, the negative log-likehood is typically regularized by adding a complexity penalty for the number of free parameters. For that we will use the BIC score for MAGs from (Richardson and Spirtes, 2002), with n and e resp. the number of variables and edges in $\mathcal{G}$; see also (Triantafillou and Tsamardinos, 2016).

$$BIC(\hat{\Sigma}, \mathcal{G}) = 2l(S|\hat{\Sigma}_{\mathcal{G}}) - \log(N)(2n + e) \quad (2)$$

Two final remarks: in practice, the score (2) is not guaranteed to be a fully equivalent score, as different MAG instances in the same equivalence class can have different sized districts, making it harder for the RICF step in 1 to converge to the same value. However, in theory in the large sample limit any MAG instance from the true equivalence class should obtain a higher score than any MAG that is not. Secondly, the current likelihood score (1) is only defined for directed graphs, meaning that MAGs with invariant undirected edges (identifiable selection bias) cannot be scored and are therefore skipped in the evaluation. It is possible to extend the score to include selection bias as well, but that is left to another article.

References

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Drton, M. and Richardson, T. S. (2008). Graphical methods for efficient likelihood inference in Gaussian covariance models. *Journal of Machine Learning Research*, 9:893–914.

Gillispie, S. B. and Perlman, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, 141(1-2):137–155.

He, Y., Jia, J., and Yu, B. (2015). Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 16(1):2589–2609.

Nowzohour, C., Maathuis, M. H., Evans, R. J., and Bühlmann, P. (2017). Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374.

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

Triantafillou, S. and Tsamardinos, I. (2016). Score-based vs constraint-based causal learning in the presence of confounders. In *Cfa@ uai*, pages 59–67.

Zhang, J. (2008). 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.