Efficient Greedy Equivalence Search for Non-Score-Equivalent Criteria using Sampling

Rafailia Chatzianastasiou
Saarland University

Osman Mian
IKIM – the Institute for Artificial Intelligence in Medicine

Jilles Vreeken
CISPA Helmholtz Center for Information Security

Abstract

Greedy Equivalence Search (GES) is a standard score-based causal discovery algorithm that searches over Markov Equivalence Classes (MECs). Its efficient implementation applies local MEC operators without enumerating all Directed Acyclic Graphs (DAGs), and is sound and complete under score-equivalent criteria such as BIC. We show that this version can fail for non-score-equivalent criteria and propose SGES to address this issue. SGES samples DAGs from the MEC at each step of forward and backward search and scores candidate operations individually. This lets non-score-equivalent criteria exploit directional information, with the sampling rate interpolating between efficient and original GES. We perform initial experiments to show SGES finds more accurate causal structures than GES when score equivalence is violated, and outline future directions for a PAC-style SGES.

1 Introduction

Discovering causal relationships is a central scientific problem. Controlled experiments are the gold standard but are often expensive, ethically problematic, or impossible, so researchers typically rely on observational data, motivating work on causal discovery from such data. Causal discovery methods fall into two main categories: constraint-based approaches which use conditional independence tests to eliminate causal edges and, score-based approaches, which search for the graph maximizing a scoring criterion. A canonical score-based method is Greedy Equivalence Search (GES) [2], which uses criteria such as the Bayesian Information Criterion (BIC) [15] to measure how well a graph explains the data. As the number of possible graphs grows super-exponentially in the number of variables, exhaustive enumeration quickly becomes infeasible and necessitates the need for efficient algorithms. Chickering [2] therefore proposed an efficient version of GES that searches over Markov equivalence classes (MEC) — sets of directed acyclic graphs (DAGs) encoding the same conditional independencies. This efficiency relies on the score equivalence assumption: all DAGs within the same MEC receive the same score, enabling search over equivalence classes instead of DAGs.

Recent work has introduced scores that allow for identification of causal structures beyond the MEC under additional modeling assumptions [6, 9, 11, 7]. These often exploit asymmetries between predicting effects from causes and vice versa. While such scores have been used within efficient GES with empirical success [6, 12], they are not score-equivalent and do not preserve its theoretical guarantees. We address this by extending GES to non–score-equivalent settings in a theoretically consistent way. Our Sampling-based GES (SGES) integrates the polynomial-time DAG sampling algorithm of Wienöbst et al. [18] into the search, sampling DAGs from the current MEC and using the highest-scoring DAG to guide transitions between neighboring MECs. Our preliminary experiments show SGES consistently recovers higher-quality causal structures than standard GES in these settings.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Background and Notation

We consider m continuous-valued random variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ with joint distribution $P(\mathcal{X})$. As required by GES, we assume causal sufficiency (no unobserved confounders) and acyclicity, so causal relationships among $\mathcal{X}$ can be represented by a causal DAG $G_{\mathcal{X}}$, where nodes are variables and directed edges between variables correspond to direct causal influences. We denote the causal parents of X_i with pa_i, and assume the causal Markov and faithfulness conditions [13]. Together these imply that two (conditionally) independent variables in $P(\mathcal{X})$ will not be connected by an edge in $G_{\mathcal{X}}$. The true DAG $G_{\mathcal{X}}$ belongs to a Markov equivalence class (MEC) $\boldsymbol{\mathcal{E}}(G_{\mathcal{X}})$, which contains all DAGs with the same skeleton—the undirected version of the graph—and the same set of v-structures, i.e., triples (X, Y, Z) where Y has incoming edges from both X and Z, and X and Z are non-adjacent. A MEC can be represented by a completed partially directed acyclic graph (CPDAG), in which an edge is directed if its orientation is invariant across all members of the MEC, and undirected otherwise.

Given an i.i.d. sample $\mathbf{X}$ of size n from $P(\mathcal{X})$ with $|\mathbf{X}| = n$ under aforementioned assumptions, GES searches for MEC $\mathcal{E}^*$ with $G_{\mathcal{X}} \in \mathcal{E}^*$ such that,

$$\mathcal{E}^* = \arg\max_{\hat{\mathcal{E}} \in \boldsymbol{\mathcal{E}}^{(m)}} \mathcal{S}(\mathbf{X}; \hat{\mathcal{E}}) \ .$$

where $\boldsymbol{\mathcal{E}}^{(m)}$ denotes the set of all possible MECs over m variables and $\mathcal{S}$ is a scoring criterion that evaluates how well a given MEC $\hat{\mathcal{E}}$ reflects conditional independencies in $\mathbf{X}$. To do so, GES performs a two-phase greedy search. Starting with a DAG containing no edges, it iteratively adds edges in a forward phase to increase the score until no improvement is possible, followed by a backward phase that removes edges to further optimize the score. In particular, let $\mathcal{E}_0$ be the MEC at the current step of GES. Let $\mathcal{G}_0 = \{G \mid \boldsymbol{\mathcal{E}}(G) = \mathcal{E}_0\}$, and let $S_0 = \arg\max_{G \in \mathcal{G}_0} \mathcal{S}(\mathbf{X}; G)$. In forward (resp. backward) phase GES evaluates all possible valid edge additions (resp. deletions) to find the next best MEC $\mathcal{E}_+$. It calculates gain associated with an add (resp. remove) operation $\psi \in \Psi_0$ as,

$$\delta(\psi) = \max_{G \in \mathcal{G}_0} \mathcal{S}(\mathbf{X}; G \oplus \psi) - S_0, \tag{1}$$

where $G \oplus \psi$ is the DAG after applying ψ. Let $\psi^\star = \arg\max_{\psi \in \Psi_0} \delta(\psi)$ with $\delta(\psi^\star) > 0$, GES computes the next G_+ as $G \oplus \psi^\star$, its corresponding score S_+ and the new MEC $\mathcal{E}_+ = \boldsymbol{\mathcal{E}}(G_+)$. Despite its greedy nature, GES asymptotically converges to the correct MEC under a consistent $\mathcal{S}$ such as the BIC [15].

Evaluating Eq. (1) requires exhaustive enumeration of DAGs within $\mathcal{E}_0$ at each step and quickly becomes infeasible due to exponential growth of search space as number of variables increases. To circumvent this Chickering [2] proposes an efficient search scheme which works for score-equivalent criteria like BIC [15]. A criterion is considered score-equivalent if $\mathcal{S}(\mathbf{X}; G) = \mathcal{S}(\mathbf{X}; H) \ \forall G, H \in \mathcal{E}$ for any MEC $\mathcal{E}$. Under score-equivalence Chickering [2] introduces the $INSERT$ resp. $DELETE$ operators such that the score change for each edge addition resp. removal can be directly evaluated on CPDAGs. This bypasses exhaustive enumeration and allows for a faster traversal of search space and still preserves consistency guarantees.

The efficient search however, comes at the cost of restricted identifiability. It renders efficient GES unsuitable for scores that exploit additional assumptions on causal mechanisms, such as additive noise [16, 14] or non-linear functional relationships [5, 9] to give identifiability beyond MEC [6, 10, 11, 7, 8]. While existing work has applied non score-equivalent criteria within GES [6, 12] with empirical success, we show next that this can violate the consistency guarantees of efficient GES.

3 GES for Non-Score-Equivalent Criteria

Consider three variables A, B, C and their corresponding underlying DAG resp. CPDAG shown in Fig. 1. Let $\hat{\mathcal{S}}$ be any consistent but non-score equivalent criterion [6, 11]. Assume that GES using $\hat{\mathcal{S}}$ discovers the correct CPDAG at the end of its forward search phase. Moving to the backward-search, GES evaluates whether or not the edge between B and C should be removed. This can be evaluated in two different ways, namely $DELETE[B \to C | \{A\}]$ (meaning we delete B as a parent of C, while keeping A as a parent) or $DELETE[C \to B | \{A\}]$. For a score-equivalent criterion this makes no difference because the $\max$ operator in Eq. (1) trivially simplifies to a single computation.

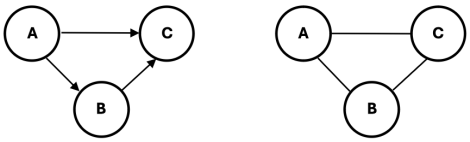

Figure 1: A directed acyclic graph over three variables (left), and its equivalent CPDAG (right).

Things, however, behave differently with $\hat{\mathcal{S}}$: the first operation $DELETE[B \rightarrow C|\{A\}]$ results in deletion of a true parent of C, subsequently worsening the score, whereas the second operation $DELETE[C \rightarrow B|\{A\}]$ implies deleting a redundant parent of B and would improve the score. Whether or not the edge between B and C is removed depends on which $DELETE$ operator is used to test this edge. This demonstrates that efficient implementations of GES may give a different result depending on which $DELETE(\cdot)$ operator is chosen may not necessarily remain consistent[1].

To make GES compatible for $\hat{S}$, ideally we must compute Eq. (1) exactly by enumerating all DAGs within a MEC. But we already know that this quickly becomes computationally prohibitive. As a middle ground, we propose a sampling-based alternate — compute Eq. 1 approximately by sampling a user-specified fraction of DAGs for each ψ. Formally,

$$\tilde{\delta}(\psi) = \max_{G \in \tilde{\mathcal{G}}_0} \mathcal{S}(\mathbf{X}; G \oplus \psi) - S_0 \tag{2}$$

with $\tilde{\mathcal{G}}_0 \subseteq \mathcal{G}_0$. Defining Eq. (2) in this way allows us to naturally interpolate between the efficient and original GES. We refer to this new variant as Sampling GES (SGES).

To achieve this goal we can use any existing algorithm for sampling DAGs within a given MEC. We propose to use the algorithm of Wienöbst et al. [18] due to its polynomial time-complexity in contrast to similar approaches having an exponential worst-case [4, 17, 1, 3]. The modularity of our proposed approach makes it straightforward to implement and allows us to directly incorporate the sampling procedure in any available implementation of GES. In the following, we implement our proposed modification and analyze initial results.

4 Results

We implement SGES in Python, using the implementation of GES in causal-learn library [19]. We directly incorporated the polynomial-time sampling algorithm from Wienöbst et al. [18] into the forward and backward phases of GES. At each forward resp. backward phase step, we approximate Eq. (1) using Eq. (2). We generate a variety of synthetic data involving variables following linear, polynomial, Gaussian process, randomly initialized Neural networks, and Sigmoid relationships with independent additive Gaussian noise. We consider graphs of sizes $m \in \{5, 7, 10\}$, evaluate sample sizes $n \in \{200, 500, 700, 1000\}$ and measure performance on varying DAG sampling rate $f \in \{0, 0.05, 0.1, 0.2, 0.3\}$, where $f = 0$ means that we run the efficient GES.

We consider one score-equivalent criterion BIC [15] and one non-score equivalent criterion, MDL score of Mian et al. [11] in our evaluation. We use BIC as a test-case to show that it may not behave as a score-equivalent criteria for non-linear causal mechanisms and can still benefit from a sampling-based procedure — a conclusion that we confirmed holds in our experiments. We evaluate the goodness of discovered causal structures using the Structural Hamming Distance (SHD) which counts the number of edges where the predicted causal graph differs from the true graph. For comparability across different network sizes, we normalize SHD to be between 0 and 1 by dividing it by $m(m-1)$.

We report the results for different causal mechanisms in Fig. 2 where, with an exception of linear mechanisms, we see a performance improvement going from GES to SGES for both BIC and MDL scores. While the improvement for MDL is expected based on our theoretical understanding of Eq. (1), we observe that BIC also benefits from a sampling-based procedure for non-linear case. This is because for the latter case, the model itself does not stay score-equivalent and results in a model misspecification for BIC. This in turn, is alleviated by the use of sampling. For the linear-case, since score-equivalence holds within the model, we do not see any improvement between GES and SGES.

[1]We could construct a similar example using $INSERT$ but we find the $DELETE$ example more intuitive

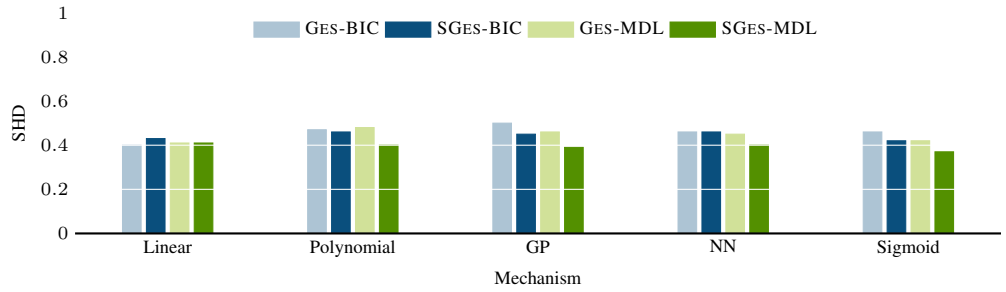

Figure 2: **[SHD Lower is better]** for random graphs of sizes $m \in \{5, 10, 15\}$. SGES improves over efficient GES for all cases involving non-linear causal mechanisms.

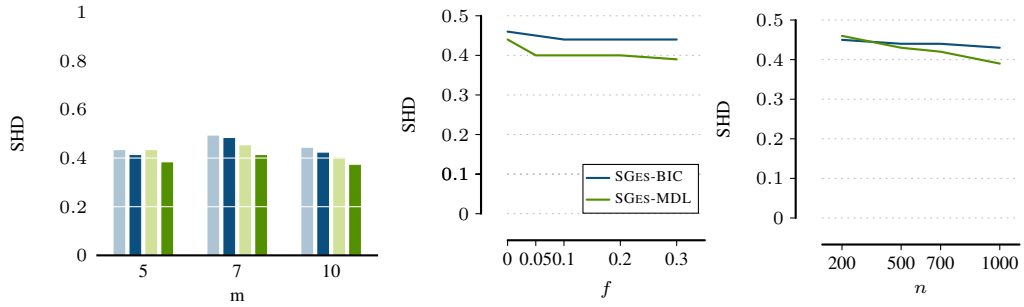

Figure 3: **[SHD Lower is better]** for random graphs of sizes $m \in \{5, 7, 10\}$, sampling rate $f \in \{0, 0.05, 0.1, 0.2, 0.3\}$, and sample sizes $n \in \{200, 500, 700, 1000\}$. (Left) Average SHD's for MDL resp. BIC scores with (dark color) sampling and without (light) sampling inside GES. (Middle) Change in SHD as we increase the sampling frequency. We see that introducing sampling improves performance. (Right) Changes in SHD with increasing n, performance improves as n increases.

For performance across varying graph sizes, we see in Fig. 3 that SGES improves over GES in all cases. Our initial experiments further show that increasing the number of sample for $\tilde{\mathcal{G}}_0$ (ref. Eq. (2)) improves both BIC and MDL scores. The improvement however is more profound for the MDL score as sampling increases the likelihood of evaluating the neighbouring states using the correct candidate. We observe a similar trend as the number of samples used for causal discovery increases; specifically, a higher sampling rate and a larger sample size lead to better performance.

5 Outlook and Conclusion

In this preliminary work, we identified a key limitation in current GES implementations: their handling of non-score-equivalent criteria and the theoretical mismatch that arises when such criteria are used. As an initial step toward addressing this gap, we proposed a sampling-based variant, SGES, which samples a subset of DAGs from the MEC at each search step and evaluates the score using the best DAG in the sample, in spirit similar to the original GES formulation. Our experiments over diverse settings, show that SGES consistently outperforms GES, producing causal graphs that more closely match the ground truth. While these results are promising, they mark only the beginning of a broader research direction.

A clear drawback of SGES is its increased runtime due to the additional DAG sampling step. Although partial mitigation is possible through score caching, future work will focus on improving sampling efficiency. A natural extension is to replace uniform DAG sampling with an *adaptive sampling strategy* that leverages prior information from the score function itself. That is, sampling graphs with higher-scoring parent sets more frequently. Such priors could enable SGES to achieve the same performance as uniform sampling with a lower sampling rate, reducing computational overhead. Moreover, they could pave the way for a *PAC-style version* of SGES for certain additive-noise-based scores [11, 12, 7], further narrowing the gap between theoretical guarantees and practical performance, and is our current line of ongoing work.

References

[1] Ali AhmadiTeshnizi, Saber Salehkaleybar, and Negar Kiyavash. Lazyiter: a fast algorithm for counting markov equivalent dags and designing experiments. In *International conference on machine learning*, pages 125–133. PMLR, 2020.

[2] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[3] Robert Ganian, Thekla Hamm, and Topi Talvitie. An efficient algorithm for counting markov equivalent dags. *Artificial Intelligence*, 304:103648, 2022.

[4] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Counting and sampling from markov equivalent dags using clique trees. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3664–3671, 2019.

[5] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, volume 21. Curran, 2009.

[6] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1551–1560, 2018.

[7] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Discovering invariant and changing mechanisms from data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1242–1252, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850.

[8] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. In *Advances in Neural Information Processing Systems*, volume 36, pages 75595–75622. Curran Associates, Inc., 2023.

[9] Alexander Marx and Jilles Vreeken. Telling cause from effect by local and global regression. *Knowledge and Information Systems*, 60(3):1277–1305, 2019.

[10] Alexander Marx and Jilles Vreeken. Identifiability of cause and effect using regularized regression. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2019.

[11] Osman Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[12] Osman Mian, David Kaltenpoth, Michael Kamp, and Jilles Vreeken. Nothing but regrets—privacy-preserving federated causal discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 8263–8278. PMLR, 2023.

[13] Judea Pearl. *Causality*. Cambridge university press, 2009.

[14] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

[15] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[16] Shohei Shimizu, Patrik O. Hoyer, Aapo Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72): 2003–2030, 2006.

[17] Topi Talvitie and Mikko Koivisto. Counting and sampling markov equivalent directed acyclic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7984–7991, 2019.

[18] Marcel Wienöbst, Max Bannach, and Maciej Liśkiewicz. Polynomial-time algorithms for counting and sampling markov equivalent dags with applications. *Journal of Machine Learning Research*, 24(213):1–45, 2023.

[19] Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.