On Different Notions of Redundancy in Conditional-Independence-Based Discovery of Graphical Models

Philipp M. Faller
Karlsruhe Institute of Technology
Karlsruhe, Germany
philipp.faller@partner.kit.edu

Dominik Janzing*
Amazon Research
Tübingen, Germany
janzind@amazon.com

Abstract

Conditional-independence-based discovery uses statistical tests to identify a graphical model that represents the independence structure of variables in a dataset. These tests, however, can be unreliable, and algorithms are sensitive to errors and violated assumptions. Often, there are tests that were not used in the construction of the graph. In this work, we show that these *redundant* tests have the potential to *detect* or sometimes *correct* errors in the learned model. But we further show that not all tests contain this additional information and that such redundant tests have to be applied with care. Precisely, we argue that the conditional (in)dependence statements that hold for every probability distribution are unlikely to detect and correct errors—in contrast to those that follow only from graphical assumptions.

1 INTRODUCTION

Graphical models have become an indispensable tool for understanding complex systems and making informed decisions in various scientific disciplines (Lauritzen, 1996). They provide insights into the structure within a system, and under some additional assumptions, they can be interpreted as *causal models* (Pearl, 2009; Spirtes et al., 2000).

Conditional independence (CI) statements are utilised to infer the graphical structure by algorithms such as PC (Spirtes et al., 2000) or SP (Raskutti and Uhler, 2018). However, a key challenge arises from the statistical hardness of conditional independence tests. As shown by Shah and Peters (2020), CI-tests cannot have a valid false positive control and power against arbitrary alternatives simultaneously. Additionally, constraint-based algorithms often rely on assumptions like *faithfulness*, which means that a graph not only implies *in*dependences but also dependences. Uhler et al. (2013) showed that this assumption can be problematic in the finite-sample regime since even faithful distributions can be close enough to unfaithful ones for a CI-test to fail. It is also violated whenever systems evolve to equilibrium states, as in many biological settings (Andersen, 2013). On the other hand, even a single wrong result of a CI-test can result in arbitrarily large changes in the resulting graphical model (as shown in example 6 in appendix F).

In the worst case, CI-based discovery of graphical models requires exponentially many CI-tests in the number of nodes (Korhonen et al., 2024; Zhang et al., 2024). Despite the large number of required tests, the set of possible graphical models can still be small compared to the set of possible combinations of CI-statements. While e.g. Shiragur et al. (2024) proposed a method to reduce the number of required tests (while sacrificing details of the models), in this paper we will advocate for using additional CI-tests to evaluate the graphs, which has been proposed implicitly or explicitly before (Textor et al., 2016; Eulig et al., 2025; Janzing et al., 2023). In spirit, this follows Raskutti and Uhler (2018), who have hypothesised that there exists a statistical/computational trade-off for causal discovery. We will argue that not all CI-tests carry much additional information (and can therefore be used to evaluate a graphical model), but only those tests for which the result follows from *graphical* restrictions instead of the laws of probability.

Example 1 (non-generic collider) Consider a probability distribution that is Markovian and faithful to the graph $X_1 \to Y \leftarrow X_2$ for random variables

*Work done outside of Amazon. Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

X_1, X_2, Y. Suppose we use the PC algorithm to recover the graph. The algorithm will conduct all pairwise marginal independence tests. These tests already identify the given DAG. But clearly, the graph also entails $X_1 \not\perp\!\!\!\perp X_2 \mid Y$ under the faithfulness assumption. On the other hand, this dependence does not follow for all probability distributions, as we will see by constructing a counterexample. We will use similar constructions multiple times. Assume[1] $Y = (Y_1, Y_2) \in \mathbb{R}^2$. Now assume Y_1 only depends on X_1 and Y_2 only on X_2 as in fig. 1. Then we would get the same marginal (in)dependences as before, but not the conditional dependence from above. On the other hand, if we have $X_1 \not\perp\!\!\!\perp Y$ and $X_1 \perp\!\!\!\perp X_2$ we also have $X_1 \not\perp\!\!\!\perp Y \mid X_2$ for all probability distributions (see appendix B.1 for a derivation). This follows from the Graphoid axioms that we will refer to throughout the paper (see definition 2). In other words, there is a dependence, $X_1 \not\perp\!\!\!\perp X_2 \mid Y$, that follows from the assumption that the underlying distribution can be represented by a *faithful* DAG, but the dependence does not hold for all distributions. At the same time, there is a CI-statement that carries no additional information, namely $X_1 \not\perp\!\!\!\perp Y \mid X_2$.

We can use tests like $X_1 \perp\!\!\!\perp X_2 \mid Y$ in two ways: either to make the result of graph discovery more robust, or for evaluation—similar to held-out samples in statistical cross-validation (Bishop and Nasrabadi, 2006) or additional bits in an error-detecting code (see also appendix D). Hence, we will call them *redundant*. But we will see throughout the paper that these redundant tests can detect and correct (among others) errors from faithfulness violations, which is impossible with methods like statistical cross-validation.

Contributions This paper aims to provide a novel perspective on CI-based discovery of graphical models. Precisely,

- to the best of our knowledge, we are the first to point out that the dependence between CI-statements impacts which tests should be used in graph discovery,
- we show that tests which already follow from previous ones in all distributions can give a misleading impression of evidence for a graphical model, while the ones that only follow through graphical assumptions are more likely to falsify the model,
- we show why the tests that follow for all distributions cannot be used to correct errors, while the ones that follow from graphical assumptions alone can,
- we show how our novel perspective generalises previous results on the robustness of graph discovery.

[1]One could also think about a variable Y taking values in the natural numbers, where disjoint sets of bits in the binary expansion of Y depend on X_1 and X_2, respectively.

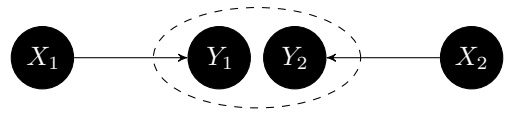


Figure 1: The marginal independence tests identify the faithful DAG (with Y as a single variable). But the collider structure $X_1 \to Y \leftarrow X_2$ implies $X_1 \not\perp\!\!\!\perp X_2 \mid Y$ for all *faithful* distributions. This does not hold for every distribution. On the contrary, given the marginal tests we have, e.g., $X_1 \not\perp\!\!\!\perp Y \mid X_2$ for all distributions.

In other words, we are the first to systematically investigate and categorise the redundancy of CI-statements, and particularly to use these notions to evaluate and improve graphs with CI-tests that are 'held-out' in the sense that they do not follow from the tests used for the graph discovery. This work aims to contribute to the discussion on how graphical models should be evaluated and to question which empirical observations are real evidence and thus capable of corroborating a model.

2 REDUNDANCY OF CIs

Notation We will now introduce some notation and basic concepts. For more detailed definitions, we refer the reader to appendix A. We denote a random variable with an upper case letter X. A set of random variables is denoted with boldface letter $\mathbf{X}$. Let $\mathbf{V}$ be a finite set of variables. An independence model M over $\mathbf{V}$ is a set of triplets $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, where $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$ and $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are disjoint. We say $\mathbf{X}$ is *independent* from $\mathbf{Y}$ given $\mathbf{Z}$ and write $\mathbf{X} \perp\!\!\!\perp_M \mathbf{Y} \mid \mathbf{Z}$, where we often omit the subscript. If $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \notin M$ we say they are dependent and write $\mathbf{X} \not\perp\!\!\!\perp_M \mathbf{Y} \mid \mathbf{Z}$. A *CI-statement* is a quadruple of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and a boolean value, indicating whether the independence holds. For a set of CI-statements L we slightly abuse notation and write $L \subseteq M$ if for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, b) \in L$ we have $b = ((\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M)$. We also sometimes write a CI-statement as function $\mathrm{CI} : \mathbf{X}, \mathbf{Y}, \mathbf{Z} \mapsto (\mathbf{X}, \mathbf{Y}, \mathbf{Z}, b)$. Note that with *in*dependences we refer to statements of the form $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ and with dependences to $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, while with CI-statement we refer to both of them. A probability distribution over $\mathbf{V}$ induces an independence model via probabilistic conditional independence. Graphical models can represent independence models, and we denote a model induced by a graph G as M_G. For an undirected graph G we define $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G$ iff $\mathbf{X}$ is separated from $\mathbf{Y}$ given $\mathbf{Z}$. For DAGs $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G$ iff $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$. In both cases, we also write $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$. By *graphical model* we refer to *either* an undirected graph

or a DAG (and its respective independence model).[2] A graph G is *Markovian* to an independence model M if $(\mathbf{X},\mathbf{Y},\mathbf{Z}) \in M_G \implies (\mathbf{X},\mathbf{Y},\mathbf{Z}) \in M$. It is *faithful* if $(\mathbf{X},\mathbf{Y},\mathbf{Z}) \in M \implies (\mathbf{X},\mathbf{Y},\mathbf{Z}) \in M_G$. Again, we abuse notation and say G is Markovian to a set of CI-statements L when all *in*dependences in G are contained and true in L. As shorthand, we use $\mathrm{CI}(\mathbf{V}) := \{(X,Y,\mathbf{Z}) : X,Y \in \mathbf{V}, \mathbf{Z} \subseteq \mathbf{V} \setminus \{X,Y\}, X \neq Y\}$ for triplets of nodes that often occur in our CI-statements.

2.1 Graphoids, graphs and redundancy

The central observation of this paper is that a graphical model usually entails more CI-statements than the ones necessary to identify its Markov equivalence class, as in example 1. In other words, the space of independence models entailed by a graphical model is typically smaller than the space of possible independence models. The additional tests are our candidates to be used for error detection and correction, which motivates the following definition.

Definition 1 (GM-redundancy) Let L be a set of CI-statements and $s \notin L$ be another CI-statement. Let $\mathcal{G}$ be a set of graphical models. We call s *graphical-model-redundant* (GM-redundant) w.r.t. $\mathcal{G}$ and L if $\{s\} \subseteq M_G$ whenever $L \subseteq M_G$ for any graph $G \in \mathcal{G}$.

Note that this definition is with respect to a set of (previous) CI-tests *and* a class of graphical models. The former depends on the specific algorithm used for discovery, while the latter depends on the assumptions that we make about the data. In example 1, both $X_1 \not\perp\!\!\!\perp X_2 \mid Y$ and $X_1 \not\perp\!\!\!\perp Y \mid X_2$ are GM-redundant with respect to the marginal independence tests, as they hold in every DAG that represents the given marginal CI-statements faithfully.

In coding theory, one can often assume that bit errors occur independently. But it is well-known that the output of CI-tests on the same dataset can be dependent. In particular, there are cases where some (in)dependence statements follow from a set of (in)dependence statements for *all* probability distributions. Pearl and Paz (2022) have provided a sound set of logical rules to derive independences: the (semi) Graphoid axioms.

Definition 2 (Graphoid Axioms) Let M be an independence model over variables $\mathbf{V}$ and $\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{W} \subseteq \mathbf{V}$ be disjoint with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$. We call M a *semi-Graphoid* if the following properties hold

1. Symmetry: $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$

[2] Our insights can be applied to any model with a notion of independence between nodes, such as chain graphs (Lauritzen, 1996), completed partial DAGs, maximal ancestral graphs, partial ancestral graphs (Spirtes et al., 2000), or acyclic directed mixed graphs (Richardson, 2003).

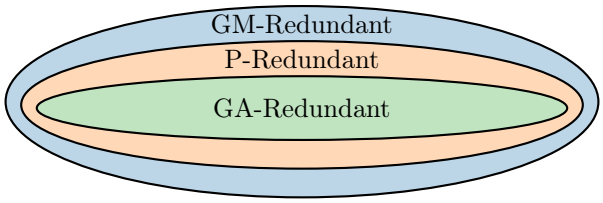


Figure 2: Hierarchy of definitions 1, 3 and 4. We argue to use GM- but not P-redundant CI-statements.

2. Decomposition: $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$
$\implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \ \wedge\ \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$
3. Weak Union: $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$
$\implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \ \wedge\ \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}$
4. Contraction: $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \ \wedge\ \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}$
$\implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$

M is called *Graphoid* if we further have

5. Intersection: $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \ \wedge\ \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}$
$\implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$.

The independence model of a distribution is a semi-Graphoid, and it is a Graphoid if the distribution is positive (Lauritzen, 1996). Although these rules are sound, they are not complete, and there cannot be a finite, sound, and complete set of axioms to describe conditional independence in probability distributions (Studený, 1992). Since the semi-Graphoid rules hold for every distribution, they also hold for every empirical distribution. Moreover, the following proposition shows that for conditional mutual information (which is zero if and only if *in*dependence holds), the Graphoid axioms have a certain continuity property.

Proposition 1 (continuity of a Graphoid) *Let* $\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{W} \subseteq \mathbf{V}$ *be disjoint and* $\epsilon,\delta > 0$. *Then*

1. $I(\mathbf{X}:\mathbf{Y} \mid \mathbf{Z}) \leq \epsilon \iff I(\mathbf{Y}:\mathbf{X} \mid \mathbf{Z}) \leq \epsilon$
2. $I(\mathbf{X}:\mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq \epsilon$
$\implies I(\mathbf{X}:\mathbf{Y} \mid \mathbf{Z}) \leq \epsilon \ \wedge\ I(\mathbf{X}:\mathbf{W} \mid \mathbf{Z}) \leq \epsilon$
3. $I(\mathbf{X}:\mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq \epsilon$
$\implies I(\mathbf{X}:\mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \leq \epsilon \wedge I(\mathbf{X}:\mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$
4. $I(\mathbf{X}:\mathbf{Y} \mid \mathbf{Z}) \leq \epsilon \ \wedge\ I(\mathbf{X}:\mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$
$\implies I(\mathbf{X}:\mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq 2\epsilon$

If we further assume $I(\mathbf{W}:\mathbf{Y} \mid \mathbf{Z}) \leq \delta$ *we get*

5. $I(\mathbf{X}:\mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \leq \epsilon \wedge I(\mathbf{X}:\mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$
$\implies I(\mathbf{X}:\mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq \delta + 2\epsilon$.

All proofs are in appendix B. This means that even if some CI-statements hold only approximately, they are likely to influence other test results according to the Graphoid axioms. In other words, even a very weak dependence that is not detected by a test still influences other tests almost as if there were no dependence. Accordingly, the influenced tests contain less (or no) redundant information in the sense that we are inter-

ested in. Therefore, we want to differentiate between tests that are implied by the graph alone and tests that already follow for all probability distributions.

Definition 3 (P-redundancy) Let L be a set of CI-statements and $s \notin L$ be another CI-statement. We call s *probabilistically redundant* (P-redundant) with respect to L if $\{s\} \subseteq M$ for any independence model induced by a probability distribution with $L \subseteq M$.

The CI-statements we are interested in are the ones that are not P-redundant. Yet, it is hard to operationalise this definition. Although Niepert (2009) shows that the problem of whether a CI-statement follows from a given set of other CI-statements is decidable for variables with finite support, to the best of our knowledge, there are no results on decidability for continuous variables. To render the problem decidable in any case, we will restrict ourselves to CI-statements that follow via the Graphoid axioms.

Definition 4 (GA-redundancy) Let L be a set of CI-statements and $s \notin L$ be another CI-statement. We call s *Graphoid-axiom-redundant* (GA-redundant) with respect to L if $\{s\} \subseteq M$ for any Graphoid independence model with $L \subseteq M$.

Since the Graphoid axioms are sound, GA-redundancy[3] is a sufficient condition for a CI-statement to also be P-redundant. In example 7 in appendix F, we show CI-statements that are P-redundant but not GA-redundant.

The following definition captures the CI-statements that are not already implied by the Graphoid axioms but follow solely from graphical assumptions.

Definition 5 (PGM-redundancy) Let L be a set of CI-statements and $s \notin L$ be another CI-statement. We call s *purely graphical-model-redundant* (PGM-redundant) with respect to L if s is GM-redundant but not GA-redundant.

In example 1, only $X_1 \not\perp\!\!\!\perp X_2 \mid Y$ is PGM-redundant with respect to the marginal CI-statements.

2.2 Graphical Criteria for Redundancy

These definitions beg the question of whether there is a criterion to find the PGM-redundant CI-statements. We will first present results showing which CI-statements cannot be PGM-redundant, and then see a sufficient graphical criterion that covers a broad class of examples where they are. In particular, corollaries 1 and 2 show cases where all *in*dependences are GA-redundant.

[3]For simplicity, we do not further distinguish between semi-Graphoid-axiom-redundant and Graphoid-axiom-redundant CI-statements.

Corollary 1 (Verma and Pearl (1990) Thm. 2) *Let M be a Graphoid independence model over a set of nodes $\mathbf{V}$, and $\pi : \mathbf{V} \to \mathbb{N}$ be an ordering of $\mathbf{V}$. Let L_π be a list of CI-statements from M such that for $X, Y, \mathbf{Z} \in \mathrm{CI}(\mathbf{V})$ we have $\mathrm{CI}(X, Y \mid \mathbf{Z}) \in L_\pi$ iff $\pi(X) < \pi(Y)$ and*

$$\mathbf{Z} = \{Z \in \mathbf{V} : \pi(X) \neq \pi(Z) < \pi(Y)\}.$$

Let $G(\pi)$ be the DAG over $\mathbf{V}$ with an edge from $X \in \mathbf{V}$ to $Y \in \mathbf{V}$ iff $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \in L_\pi$ for some $\mathbf{Z} \subseteq \mathbf{V}$. Then every independence *in $M_{G(\pi)}$ is GA-redundant with respect to L_π.*

Corollary 2 (Geiger and Pearl (1993) Thm. 12) *Let M be a Graphoid independence model over a set of nodes $\mathbf{V}$. Let L be the set of CI-statements*

$$L = \{\mathrm{CI}(X, Y \mid \mathbf{V} \setminus \{X, Y\}) : X, Y \in \mathbf{V},\, X \neq Y\}$$

from M. Let $G(L)$ be the undirected graph over $\mathbf{V}$ with an edge between $X, Y \in \mathbf{V}$ iff $(X \not\perp\!\!\!\perp Y \mid \mathbf{V} \setminus \{X, Y\}) \in L$. Then every independence *in $M_{G(L)}$ is GA-redundant with respect to L.*

No graph discovery algorithm can rely on the Markov assumption alone. Most of them also use the faithfulness assumption. The latter is especially troublesome, as there are many applications where it may be violated, as we have mentioned before. But in the cases above, none of the *in*dependence statements can be used as additional redundancy. So here, it is only due to faithfulness (and similar assumptions) that we can have PGM-redundancy at all. But also, not all of the dependences are good candidates for error detection and correction. As Bouckaert (1995) shows, there are also dependences that follow from the Graphoid axioms as their contrapositives. In corollary 6 in appendix E, we show how these insights can be applied in the situations of corollaries 1 and 2.

On the other hand, we will now give a criterion under which a dependence statement is guaranteed to be PGM-redundant. So whenever a graphical model implies such a dependence, it is a good candidate for additional redundancy.

Definition 6 (coupling over nodes) Let $(X, Y, \mathbf{Z})$, $(A, B, \mathbf{C}) \in \mathrm{CI}(\mathbf{V})$, G be a graphical model over $\mathbf{V}$, and $s := (X, Y, \mathbf{Z})$. We say a path $(A, \ldots, B)$ is *s-active* given $\mathbf{C}$ if

1. it is active given $\mathbf{C}$ (according to the respective graphical separation) and
2. there is no sub-path $(X, \ldots, Y)$ that is active given $\mathbf{Z}$.

Further, we say A and B are *coupled over s* given $\mathbf{C}$ iff there is at least one active but no s-active path between A and B given $\mathbf{C}$.

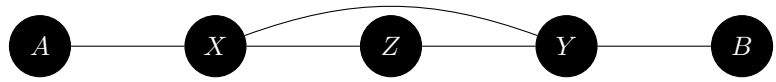


Figure 3: A and B are *coupled over* $(X, Y, \emptyset)$ given $\emptyset$ but are not coupled over (X, Y, Z) given $\emptyset$, since the path $X - Z - Y$ stays active given $\emptyset$ if we 'deactivate' all paths between X and Y that are active given Z.

Intuitively, an s-active path stays active if we 'deactivate' all paths between X and Y, and A, B being coupled over s means that all paths between A and B are 'mediated' by X and Y. See fig. 3 for an example.

Proposition 2 (sufficient criterion) *Let M be the independence model of a positive distribution, L be a set of CI-statements from M and $s := (X, Y, \mathbf{Z}) \in \mathrm{CI}(\mathbf{V})$ with $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \notin L$. Let G be a graphical model such that G is Markovian to L and $X \not\perp_G Y \mid \mathbf{Z}$. If there is no $(A \not\perp\!\!\!\perp B \mid \mathbf{C}) \in L$ such that A and B are coupled over s given $\mathbf{C}$, then $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ is PGM-redundant given L.*

Note that the PC algorithm is not guaranteed to return a graph that is Markovian to the conducted tests if its assumptions are violated (see example 8 in appendix F) and therefore proposition 2 cannot always be applied there.

3 ERROR DETECTION AND CORRECTION

Since GA-redundant tests are unlikely to contradict the given tests, they are unlikely to reveal errors in the given tests (which does not mean that such rare contradictions are uninformative). Moreover, the absence of contradictions might even give the misleading impression of evidence for a model , as we will discuss more formally in example 3. Therefore, we will *detect* and *correct* errors with PGM-redundant tests in the following.

3.1 Error detection

At the beginning of this section, we will focus on spanning trees[4] for two reasons. First, they are a simple class of examples. But also, as we will see, they impose particularly strong graphical assumptions.

Error detection in spanning trees Suppose we use the CI-statements from corollary 2 to find an undirected graph. Although this approach is quite efficient, it is sensitive to errors. Each of the conducted CI-tests directly corresponds to the presence or absence of an edge. In other words, a single error already makes it impossible to find the correct graph. But, in this case, proposition 2 equips us with plenty of CI-statements that could be used to falsify the result under the faithfulness assumption. The following corollaries show general cases where proposition 2 can be applied.

Corollary 3 (transitive dependence) *Consider the independence model that is Markovian and faithful to the undirected graph in fig. 4. Let L be the set of CI-statements from corollary 2. Then every dependence between non-adjacent nodes is PGM-redundant, and all* independences *are GA-redundant.*

This can be seen as follows: by construction, L does not contain a dependence statement involving non-adjacent nodes X_i and X_j where $i, j \in \mathbb{N}$. In particular, L does not contain a dependence between X_k and X_l with $k, l \in \mathbb{N}$ such that X_k and X_l are only connected over a path with more than one edge. By proposition 2, every dependence between non-adjacent nodes is PGM-redundant.[5] The *in*dependences follow via corollary 2. The next corollary shows that the PGM-redundant statements are imposed by *strong* graphical assumptions.

Corollary 4 (implied connectedness) *Suppose again our independence model M is Markovian and faithful to the graph in fig. 4. Now let*

$$L = \{(X_i \perp\!\!\!\perp X_n \,|\, \mathbf{V} \setminus \{X_i, X_n\}) : \, i \in \{1, \ldots, n-2\}\}$$

for $n \in \mathbb{N}_{>2}$. If $\mathcal{G}$ is the set of spanning trees over $\mathbf{V}$, the statement $X_{n-1} \not\perp\!\!\!\perp X_n \mid \mathbf{V} \setminus \{X_{n-1}, X_n\}$ is PGM-redundant.

Would X_n be disconnected from the graph we could still get $X_{n-1} \perp X_n \mid \mathbf{V} \setminus \{X_{n-1}, X_n\}$. But this is not possible in a spanning tree. So here, the PGM-redundancy comes from the spanning tree property. But if we only assume the underlying graph is any undirected graph, this would not follow, as X_n could indeed be disconnected. In terms of error detection, this means that a test with result $X_{n-1} \perp\!\!\!\perp X_n \mid \mathbf{V} \setminus \{X_{n-1}, X_n\}$ would indicate that either this test, one of the tests in L, or our graphical assumptions are wrong.

[4] A spanning tree is an undirected graph, in which every pair of nodes is connected by exactly one path.

[5] In practice, especially the dependences over short paths are of interest. As the data processing inequality (MacKay, 2003) bounds the mutual information along a path, the dependences become weaker the longer a (non-deterministic) path is.

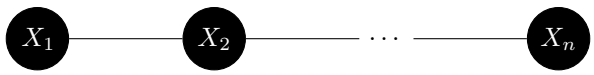


Figure 4: In this graph, every dependence along more than one edge is PGM-redundant given the CI-statements from corollary 2.

Error detection in DAGs First note that in DAGs we can also have PGM-redundancies along (now directed) paths as in corollary 3. But just as in corollary 4, there are also dependences that follow from the particular graphical assumption. To see this, consider the following variant of fig. 1.

Corollary 5 (multiple colliders) *Consider an independence model that is Markovian and faithful to the DAG with edges $X_i \to Y$ for $i = 1, \dots, n \in \mathbb{N}_{>1}$. Assume we learn a DAG using a given topological ordering and the CI-statements from corollary 1. Then all tests $X_i \not\perp\!\!\!\perp X_j \mid Y$ with $i, j \in \{1, \dots, n\}, i \neq j$ are PGM-redundant. Conversely, all* in*dependences and all statements $X_i \not\perp\!\!\!\perp Y \mid \{X_1, \dots, X_n\} \setminus \{X_i\}$ for $i = 1, \dots, n$ are GA-redundant.*

Although Ramsey et al. (2006) do not study the problem through the lens of redundancy, they observe that there can be several conditional (in)dependences that characterise a collider in a DAG. They propose to let their *Conservative PC* algorithm (CPC) indicate when these CI-statements contradict each other. Precisely, suppose PC finds a skeleton H that contains an unshielded triplet $A - B - C$. They then consider all subsets of the neighbours of A and C as potential conditioning sets, i.e. $\mathbf{Z} \subseteq \mathrm{Adj}_H(A)$ or $\mathbf{Z} \subseteq \mathrm{Adj}_H(C)$. If the observed independence model is Markovian and faithful to a DAG, either all or none of the sets $\mathbf{Z}$ with $A \perp\!\!\!\perp C \mid \mathbf{Z}$ contain B. Indeed, we can phrase this in our framework as the CI-statements $A \not\perp\!\!\!\perp C \mid \mathbf{Z} \cup \{B\}$ being GM-redundant for the hypothesis that the underlying graph contains the collider $A \to B \leftarrow C$ or $A \perp\!\!\!\perp C \mid \mathbf{Z} \setminus \{B\}$ otherwise. Further, the following example 2 shows that these tests can indeed be GM-redundant.

Example 2 (CPC and GA-redundancy) Consider again the graph given in fig. 1 and interpret Y_1 and Y_2 as the components of a vector-valued Y. Note that any independence model that is Markovian and faithful to this graph entails $X_1 \not\perp\!\!\!\perp Y$, $Y \not\perp\!\!\!\perp X_2$, and $X_1 \perp\!\!\!\perp X_2$. Further, we have the *in*dependence $X_1 \perp\!\!\!\perp X_2 \mid Y$. Since Y is neither in all sets that separate X_1 and X_2, nor in none of them, CPC would output a contradiction. But this example further shows that such a model exists and thus none of the CI-statements is GA-redundant given the others.

Note that our perspective also includes CI-statements that cannot be detected by CPC (like nodes connected along longer paths in corollary 3), which shows that our work generalises the observations of Ramsey et al. (2006).

3.2 Error correction

One might wonder whether it is also possible to correct certain errors. As noted before, e.g., the procedure in corollary 2, but also other algorithms like PC or SP can be sensitive to the test results in the sense that if a single one of the tests had a different result, the output of the algorithm would change. As we will see, we can again use redundant CI-statements to correct errors. And similarly to the case of error detection, conducting GA-redundant tests might be misleading.

Example 3 (Graphoid prevents correction) Consider the graph in fig. 5 and suppose we use the PC algorithm to learn it. First, the algorithm conducts all marginal CI-tests. Suppose they return the correct result except for[6] $Y \not\perp\!\!\!\perp Z$. In the next step, the algorithm would conduct the CI-tests with a non-empty conditioning set. So, if $Y \not\perp\!\!\!\perp Z$ were the only error, one could hope that $Y \perp\!\!\!\perp Z \mid X$ (implied by the ground truth graph) can still correct this mistake. But note that the marginal tests already imply $X \not\perp\!\!\!\perp Y \mid Z$ and $Y \not\perp\!\!\!\perp Z \mid X$. Intuitively, this is due to the *in*dependence $X \perp\!\!\!\perp Z$, which prevents conditioning on X from changing the relationship between Y and Z. This means, according to proposition 1, we are likely to get the wrong test result for the conditional *in*dependence between Y and Z.

One might wonder whether this is a shortcoming of the PC algorithm. But note how this would also affect our result if we simply select the graph with the fewest contradicting CI-statements to the empirical independence model (which we will discuss in appendix G). If we still get $X \perp\!\!\!\perp Z \mid Y$ right, there are four tests in favour of the actual ground truth model. But there are five that would be explained, e.g., by the graph $X \to Y \to Z$ (see appendix B.1). In other words, by adding GA-redundant tests into our consideration, we might wrongly conclude that we have more evidence for the wrong graph.

Error correction in spanning trees The following result shows that we can correct errors if we consider more tests than necessary. To circumvent the issues raised in example 3, we only consider tests with a conditioning set of size one. This way, we get a set of tests that are not restricted by Graphoid axioms (up to axiom 1) but suffice to identify a spanning tree.

[6] Note that such an error can never be ruled out with standard statistical testing frameworks, as the probability of a type I error cannot be zero for non-trivial procedures.

Figure 5: A false test $Y \not\perp\!\!\!\perp Z$ may lead to the conclusion that the true graph is $X \to Y \to Z$. If this were the only error, the test $Y \perp\!\!\!\perp Z \mid X$ would correct that. But $Y \not\perp\!\!\!\perp Z \mid X$ follows via Graphoid axioms from the marginal tests.

Lemma 1 *Let* $L = \{\mathrm{CI}(X, Y \mid Z) : X, Y, Z \in V, X \neq Y, X \neq Z, Y \neq Z\}$ *such that* $\mathrm{CI}(X, Y \mid Z) = \mathrm{CI}(Y, X \mid Z)$ *for all distinct* $X, Y, Z \in V$*. Then* L *contains no contradictions with respect to the Graphoid axioms and no other CI-statement follows from* L *via Graphoid axioms.*

Then we can indeed simply pick the 'message' whose encoding has the smallest distance to the received code word, i.e., the tree whose independence model differs the least from the observed one.

Proposition 3 (error correction in trees) *Let the set* $S = \{(X, Y, Z) : X, Y, Z \in V, X \neq Y, X \neq Z, Y \neq Z\}$*. Further, let* $\mathcal{T}_n$ *be the set of spanning trees with* $n \in \mathbb{N}_{>3}$ *nodes,* $T^* \in \mathcal{T}_n$ *and* M *be an independence model with* $\mathrm{MD}_S(T^*, M) \leq \lfloor (n-1)/2 \rfloor$*, where*

$$\mathrm{MD}_S(T, M) = \sum_{s \in S} \mathbb{I}\left[(s \in M_T) \neq (s \in M)\right].$$

Then we can correct at least $\lfloor (n-1)/2 \rfloor$ *errors for any spanning tree* T^* *by minimising the distance to* M*, i.e.*

$$T^* = \operatorname*{arg\,min}_{T \in \mathcal{T}_n} \mathrm{MD}_S(T, M).$$

Error correction in DAGs The strong graphical assumptions in the previous section enabled us to derive the guarantee in proposition 3. It would be desirable to have a similar result for DAGs. As example 4 shows, this is not possible without further restrictions.

Example 4 (almost complete DAG) Let G be a complete DAG over $\mathbf{V}$, i.e. for $n \in \mathbb{N}_{>1}$ the graph with nodes $\{X_1, \ldots, X_n\} = \mathbf{V}$ and edges $X_i \to X_j$ whenever $i < j$. Suppose our tests unfaithfully show $X_{n-1} \perp\!\!\!\perp X_n \mid \mathbf{V} \setminus \{X_{n-1}, X_n\}$. This observed independence model would be explained by the graph

$$G - (X_{n-1} \to X_n).$$

Even though this is only a single error, there is another graph that perfectly explains the independence model, so we would prefer this graph over G.

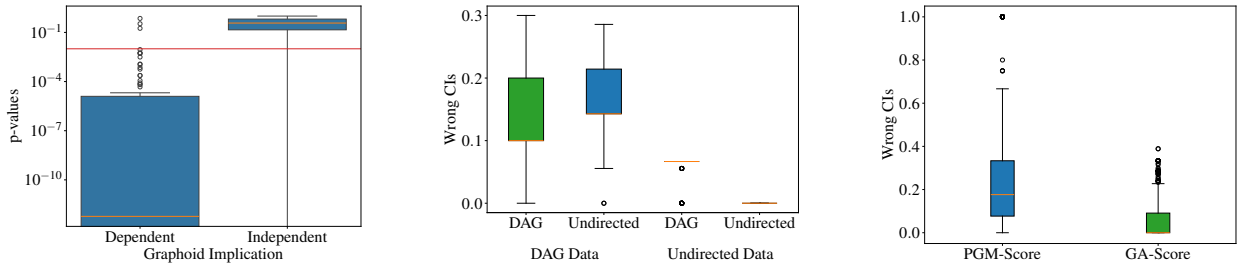
One might consider a subset of tests that cannot contain implications about each other, like we already did in proposition 3. As we have seen in example 2, a potential candidate could be the tests that the CPC algorithm by Ramsey et al. (2006) utilises to orient colliders. Indeed, Colombo et al. (2014) have proposed to do a majority vote over these tests (although they did not investigate the GA-redundancy of these tests). Clearly, this method is capable of correcting errors. It would be interesting to characterise further such 'local' criteria, where a subgraph of the learned DAG can be corrected. But note how Colombo et al. (2014) rely on the assumption that the learned skeleton is correct. Without such an assumption, it is not obvious how a local error correction could be established.

A different approach would be to study under which conditions the optimization over *all* tests works. Indeed, recently there has been a lot of interest in methods like the *sparsest permutation* (SP) algorithm (Raskutti and Uhler, 2018; Lam et al., 2022; Andrews et al., 2023), which outputs the sparsest graph among the graphs that can be constructed like in corollary 1 for all permutations of the nodes. Raskutti and Uhler (2018) show that the required assumption is strictly weaker than faithfulness. Moreover, they postulate a 'statistical/computational trade-off', i.e., that additional computations can help to reduce statistical uncertainty. Although they do not formally analyse this statement, the idea is in line with our work. To emphasise this, we will briefly study SP from the perspective of redundancy. The following lemma shows that SP relies on two key aspects.

Lemma 2 (characterisation of SP output) *The SP algorithm outputs a DAG* G^* *iff*

(a) there is a topological ordering π^* *with respect to* G^* *such that all tests in* L_{π^*} *are as in* M_{G^*}*.*
(b) for all other permutations π' *there are no less than* $|G^*|$ *dependences in* $L_{\pi'}$*.*

Evidently, if (a) fails, the algorithm cannot recover from this error. But in principle, SP can correct arbitrary errors in $L_{\pi'}$, as long as property (b) holds. If we assume that the underlying independence model is Graphoid, the errors that can occur are already restricted. Corollary 1 implies that given (a) all *in*dependences are GA-redundant and thus we are unlikely to find a contradiction here. Further, corollary 6 from appendix E also characterises dependences that follow from (a). This means the errors that the SP algorithm can correct are, for example, of the kind of proposition 2. In principle, it would be desirable to have an algorithm that can also handle violations of (a), while still requiring weaker assumptions than faithfulness. In analogy to proposition 3, we construct an algorithm that fulfils this requirement in appendix G.



(a) The p-values of different CI-tests that are implied by known tests to be either dependent or independent.

(b) Incorrect predictions of PGM-redundant CI-statements for different models and data.

(c) Incorrect predictions of PGM- versus GA-redundant CI-statements on real data.

Figure 6: Experimental results

4 EXPERIMENTS

See appendix I for more details on all experiments. The source code can be found under `https://github.com/PhilippFaller/RedundancyInGraphDiscovery`. In the first experiment, we checked the hypothesis that empirical tests rarely contradict the Graphoid axioms. To this end, we generated synthetic data from a multivariate Gaussian distribution with four variables and conducted several CI-tests. Before each test, we check with the `Z3` solver (De Moura and Bjørner, 2008) whether the result is already implied by the previous tests via Graphoid axioms. If so, we track what result the axioms imply and the resulting p-value of a CI-test. As we can see in fig. 6a, the p-values mostly follow the predictions (where the red line indicates the significance level 0.01). This corroborates our hypothesis that GA-redundant tests provide little additional information and could thus give a false impression of evidence.

In the next experiments, we thus exclude GA-redundant tests from the evaluation of our model. In the former, we synthetically generated two datasets with four binary variables. One follows the DAG with edges $X \leftarrow W \rightarrow Y$ and $X \rightarrow Z \leftarrow Y$, while the other one follows the undirected model with the same skeleton. We then learn a DAG and an undirected model with the procedures described in corollaries 1 and 2 and identify GA-redundant CI-statements using corollaries 1, 2 and 6. We then check whether they hold empirically in the data. Figure 6b shows the fraction of these tests where the graphical implication and the empirical tests disagree. We can see that the models tend to make fewer wrong predictions when they match the respective data-generating process (green) than the other model (blue).

Finally, we learned a DAG on the protein signaling data from Sachs et al. (2005). We use the procedure from corollary 1 and identified PGM-redundant CI-statements via proposition 2. Figure 6c shows the fraction of tests where the graph contradicts the empirical result. We further calculate this fraction with additional GA-redundant CI-statements that we detect via corollaries 1 and 6. In fig. 6c we can see that the PGM-redundant tests indicate more errors than the GA-redundant tests. This is in line with the fact that none of the recovered graphs were consistent with the ground truth provided by Sachs et al. (2005). In appendix H we repeated the experiment on different synthetic datasets.

5 RELATED WORK

We are the first to study how graphical and probabilistic constraints on CI-statements interplay in the *detection* and *correction* of errors in graph discovery. The fact that causal graphs can entail implications about parts of the distribution that were not seen before has been noted by Tsamardinos et al. (2012); Janzing et al. (2023) and in terms of structural causal models by (Gresele et al., 2022). Building on these insights, Faller et al. (2024); Schkoda et al. (2025) have proposed to falsify causal models by exploiting these constraints. Although this is the basic observation of GM-redundancy, we are the first to contrast this with P-redundancy. Mazaheri et al. (2025) show that there are different kinds of dependences between CI-statements, but do not discuss how this influences their confirmatory power with respect to graphical models. On the other hand, Faltenbacher et al. (2025) propose to use the CI-statements that PC would conduct anyway to detect contradictions without discussing the dependences between tests. Bromberg and Margaritis (2009) have also noted that contradictions between CI-tests can be corrected to improve discovery results, and Kim et al. (2024) correct errors by using Graphoid axioms to conduct a set of tests that are statistically better conditioned. In contrast to us, they both focus on violations of the Graphoid axioms. We have argued before that such violations are comparatively

rare, which does not mean that they are not informative if they occur. But additionally, we pointed out that the absence of these violations does not constitute evidence for a model. Hyttinen et al. (2014); Russo et al. (2024) propose to use symbolic reasoning to resolve conflicts in the provided CI-statements. But neither makes the distinction between GM- and P-redundant tests. In appendix C, we discuss more work on the general robustness of PC.

Score-based graph learning methods have also been shown to recover the independence model under certain conditions (Aragam et al., 2015). Arguably, these methods have the advantage that they 'weigh' conflicting CI-statements by their influence on the score (which is often related to the likelihood). On the other hand, it is not clear how, for such algorithms, one could determine which information has not been used in the optimization, or in other words, what test could be used to independently evaluate a learned graphical model. Studying this constitutes a research project in its own right.

Zhang and Spirtes (2016) argue that the faithfulness assumption serves three functions in graph discovery. Our work can be interpreted as adding a fourth 'face' to their three faces of faithfulness by showing that, in many cases, we get error detection and correction properties only through graphical assumptions (such as faithfulness) that are stronger than the laws of probability.

6 DISCUSSION AND LIMITATIONS

We have defined different notions of redundancy to distinguish between CI-statements that are already implied by the laws of probability and the ones that follow only from graphical assumptions. As the former, GA-redundant ones, can wrongly give the impression of additional evidence, we have characterised conditions where we have the latter, PGM-redundancy, and conditions under which we only have GA-redundancy. We are the first to propose to use PGM-redundant tests similarly to held-out data in cross-validation, or redundant bits in coding theory.

Our work shows that numerous correctly predicted CI-statements only provide evidence if they represent 'independent degrees of freedom' of the underlying distribution. This means that the mere number of correct CI-statements provides a superficial image and cannot corroborate a model. This exacerbates the more tests are used (for robustness or evaluation), rendering it all the more important to avoid P-redundancy. On the other hand, our theory is fundamentally limited by the fact that graphical models can never be *verified* without ground truth data or at least assumptions about the error model of the CI-tests used. Specifically, this means that we also cannot know for additional tests whether they are correct or how they might depend on each other in different ways than through Graphoid axioms. Nonetheless, we think that our insights can be a first step towards methods that not only detect or correct errors but also allow us to define confidence regions outside of which a model should be rejected, as the model differs too much from the observed evidence. Another limitation is that additional CI-tests add to the long runtime of state-of-the-art methods, which cannot be avoided unless $\mathcal{P} = \mathcal{NP}$ (Chickering, 1996). Regardless, we think that more robust discovery, and discovery with some form of quality estimate, is more useful for downstream tasks than a highly scalable method without that, as results are often sensitive even to small errors.

Acknowledgements

We thank Elke Kirschbaum and William Roy Orchard for helpful discussions and proofreading. Philipp M. Faller was supported by a doctoral scholarship of the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation). This work does not relate to Dominik Janzing's position at Amazon.

References

Holly Andersen. When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683, 2013.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of dags using the best order score search and grow shrink trees. *Advances in Neural Information Processing Systems*, 36:63945–63956, 2023.

Ankur Ankan and Johannes Textor. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024.

Bryon Aragam, Arash A Amini, and Qing Zhou. Learning directed acyclic graphs with penalized neighbourhood regression. *arXiv preprint arXiv:1511.08963*, 2015.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, pages 147–160, 2021.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Remco Ronaldus Bouckaert. *Bayesian belief networks:*

from construction to inference. PhD thesis, Utrecht University, 1995.

Facundo Bromberg and Dimitris Margaritis. Improving the reliability of causal discovery from small data sets using argumentation. *Journal of Machine Learning Research*, 10(2), 2009.

David Maxwell Chickering. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pages 121–130, 1996.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

Elias Eulig, Atalanti A Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26778–26786, 2025.

Philipp M Faller, Leena C Vankadara, Atalanti A Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pages 4132–4140. PMLR, 2024.

Sofia Faltenbacher, Jonas Wahl, Rebecca Herman, and Jakob Runge. Internal incoherency scores for constraint-based causal discovery algorithms. *arXiv preprint arXiv:2502.14719*, 2025.

Dan Geiger and Judea Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The annals of statistics*, 21(4):2001–2021, 1993.

Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In *International Conference on Machine Learning*, pages 7793–7824. PMLR, 2022.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Dominik Janzing, Philipp M Faller, and Leena Chennuru Vankadara. Reinterpreting causal discovery as the task of predicting unobserved joint statistics. *arXiv preprint arXiv:2305.06894*, 2023.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 (3), 2007.

Markus Kalisch and Peter Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789, 2008.

Jonghwan Kim, Inwoo Hwang, and Sanghack Lee. Causal discovery with deductive reasoning: One less problem. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, pages 1999–2017. PMLR, 2024.

Murat Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Advances in Neural Information Processing Systems*, 36:74140–74179, 2023.

Tuukka Korhonen, Fedor Fomin, and Pekka Parviainen. Structural perspective on constraint-based learning of markov networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1855–1863. PMLR, 2024.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Honghao Li, Vincent Cabeli, Nadir Sella, and Hervé Isambert. Constraint-based causal structure learning with consistent separating sets. *Advances in neural information processing systems*, 32, 2019.

Junning Li and Z Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(2), 2009.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Bijan Mazaheri, Jiaqi Zhang, and Caroline Uhler. Meta-dependence in conditional independence testing. *arXiv preprint arXiv:2504.12594*, 2025.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

Mathias Niepert. Logical inference algorithms and matrix representations for probabilistic conditional independence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 428–435, 2009.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Azaria Paz. Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 189–200. Association for Computing Machinery, 2022.

Joseph Ramsey. Improving accuracy and scalability of the pc algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*, 2016.

Joseph Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 401–408, 2006.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Raanan Y Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias. *Advances in Neural Information Processing Systems*, 34:2454–2465, 2021.

Fabrizio Russo, Anna Rapberger, and Francesca Toni. Argumentative causal discovery. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 938–949, 2024.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Daniela Schkoda, Philipp Michael Faller, Dominik Janzing, and Patrick Blöbaum. Cross-validating causal discovery via leave-one-variable-out. In *Causal Learning and Reasoning*, pages 659–692. PMLR, 2025.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48: 1514 – 1538, 2020.

Kirankumar Shiragur, Jiaqi Zhang, and Caroline Uhler. Causal discovery with fewer conditional independence tests. In *International Conference on Machine Learning*, pages 45060–45078. PMLR, 2024.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Eric V Strobl, Peter L Spirtes, and Shyam Visweswaran. Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–37, 2019.

Milan Studený. Conditional independence relations have no finite complete characterization. In *Information Theory, Statistical Decision Functions and Random Processes. Transactions of the 11th Prague Conference vol. B*, pages 377–396, 1992.

Johannes Textor, Benito Van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Ioannis Tsamardinos, Sofia Triantafillou, and Vincenzo Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *The Journal of Machine Learning Research*, 13(1):1097–1157, 2012.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 3412–3420. PMLR, 2025.

Marcel Wienöbst and Maciej Liskiewicz. Recovering causal structures from low-order conditional independencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10302–10309, 2020.

Jiaqi Zhang, Kirankumar Shiragur, and Caroline Uhler. Membership testing in markov equivalence classes via independence queries. In *International Conference on Artificial Intelligence and Statistics*, pages 3925–3933. PMLR, 2024.

Jiji Zhang and Peter Spirtes. The three faces of faithfulness. *Synthese*, 193:1011–1027, 2016.

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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator. If your work uses existing assets. Yes.
   (b) The license information of the assets, if applicable. Yes.
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

On Different Notions of Redundancy in Conditional-Independence-Based Discovery of Graphical Models Supplementary Materials

A DETAILED DEFINITIONS

We denote a random variable with upper case letter X. A set of random variables is denoted with boldface letters $\mathbf{X}$. For single variables and sets of variables, we write the attained values with a lower case letter x and the domain with calligraphic letter $\mathcal{X}$.

A *graph* G is a tuple $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a finite set of nodes and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is the set of edges. We say G is *undirected* if $\mathbf{E}$ is a symmetric relation. To emphasise the direction of an edge we also write $X \to Y$ for $(X, Y) \in \mathbf{E}$, $X \leftarrow Y$ when $(Y, X) \in \mathbf{E}$ and $X - Y$ when the graph is undirected and $(X, Y) \in \mathbf{E}$. We slightly abuse notation and write $(X \to Y) \in G$ if $(X, Y) \in \mathbf{E}$ and analogously for $X \leftarrow Y$ and $X - Y$. We denote the graph $G - (X \to Y) := (\mathbf{V}, \mathbf{E} \setminus \{(X, Y)\})$. We say two nodes $X, Y \in \mathbf{V}$ are *adjacent* if we have $(X, Y) \in \mathbf{E}$ or $(Y, X) \in \mathbf{E}$ and we denote with $\mathrm{Adj}(X)$ the set of nodes that are adjacent to X. $\mathrm{PA}(Y)$ denotes the set of nodes $X \in \mathbf{V}$ such that $(X, Y) \in \mathbf{E}$. A *path* $p = (X_1, \ldots, X_n)$ is a sequence of $n \in \mathbb{N}_{>1}$ nodes such that $X_i \in \mathbf{V}$ for $i = 1, \ldots, n$ and $X_i \in \mathrm{Adj}(X_{i+1})$ for $i = 1, \ldots, n-1$. Further, p is called a *cycle* if $X_1 = X_n$. If a graph contains no cycles, it is called a *directed acyclic graph* (DAG). If in an undirected graph G any distinct nodes $X, Y \in \mathbf{V}$ are connected by at most one path, G is called a *tree*. If it is exactly one path, G is called a *spanning tree*. For a DAG $G = (\mathbf{V}, \mathbf{E})$ we define the *skeleton* of G as the undirected graph $H = (\mathbf{V}, \mathbf{E}')$, with $(X, Y) \in \mathbf{E}'$ if $(X, Y) \in \mathbf{E}$ or $(Y, X) \in \mathbf{E}$, for $X, Y \in \mathbf{V}$. For all graphs $G = (\mathbf{V}, \mathbf{E})$ we denote $|G| := |\mathbf{E}|$. A DAG $G = (\mathbf{V}, \mathbf{E})$ is called a *complete DAG* if we have $|\mathbf{E}| = |\mathbf{V}|(|\mathbf{V}| - 1)/2$.

Let $\mathbf{V}$ be a set of random variables. We call a set $M \subseteq \mathcal{P}(\mathbf{V})^3$ an *independence model*[7] if all $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M$ are disjoint and $\mathbf{X}$ and $\mathbf{Y}$ are not empty. Then we say $\mathbf{X}$ is *independent* from $\mathbf{Y}$ given $\mathbf{Z}$ and write $\mathbf{X} \perp\!\!\!\perp_M \mathbf{Y} \mid \mathbf{Z}$, where we often omit the subscript. If $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is not in M we say $\mathbf{X}$ is *dependent* on $\mathbf{Y}$ given $\mathbf{Z}$ and write $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$. A *CI-statement* is a quadruple of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and a boolean value, indicating whether the independence holds. For a set of CI-statements L we slightly abuse notation and write $L \subseteq M$ if for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, b) \in L$ we have $b = ((\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M)$. We also sometimes write a CI-statement as function $\mathrm{CI} : \mathbf{X}, \mathbf{Y}, \mathbf{Z} \mapsto (\mathbf{X}, \mathbf{Y}, \mathbf{Z}, b)$. As we noted before, with *in*dependences we refer to statements of the form $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ and with dependences to $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, while with CI-statement we refer to both of them.

A probability distribution over $\mathbf{V}$ entails an independence model by the standard definition of probabilistic conditional independence, e.g. $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ with respect to the discrete distribution P iff

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = P(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}) \cdot P(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}),$$

for all $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \mathbf{z} \in \mathcal{Z}$. We often refer to the distribution and its independence model interchangeably.

We can also use graphs to represent independence models. First, we define a graphical notion and then see how that corresponds to conditional independence. Let G be an undirected graph with nodes $\mathbf{V}$ and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$. We say $\mathbf{X}$ and $\mathbf{Y}$ are *separated* given $\mathbf{Z}$ if every path from $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ contains a node in $\mathbf{Z}$. If a path contains no node in $\mathbf{Z}$, we say it is *active*. For a DAG $G = (\mathbf{V}, \mathbf{E})$, and a path $p = (X_1, \ldots, X_n)$ in G we say X_i is a *collider* on p if $(X_{i-1} \to X_i), (X_i \leftarrow X_{i+1}) \in \mathbf{E}$, for $n \in \mathbb{N}_{>2}, X_1, \ldots, X_n \in \mathbf{V}, i \in \{2, \ldots, n-1\}$. We say a path p is *active* given $\mathbf{Z}$ if for all colliders C on p, a descendant of C or C itself is in $\mathbf{Z}$. The sets $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* if there are no $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ such that there is an active path given $\mathbf{Z}$ between X and Y. Now we denote a model induced by a graph G as M_G. For an undirected graph G we define $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G$ iff $\mathbf{X}$ is separated from $\mathbf{Y}$ given $\mathbf{Z}$. For DAGs $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G$ iff $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$. By *graphical model*

[7] Note that Bouckaert (1995) formally distinguishes between an *independence* and *dependence* model. For our discussion it should suffice to consider the latter only implicitly.

we refer to *either* an undirected graph or a DAG (and its respective independence model). As we have noted before, we restrict our attention to undirected graphs and DAGs. But our insights can be applied to any model that is equipped with a notion of independence between nodes, such as chain graphs (Lauritzen, 1996), completed partial DAGs, maximal ancestral graphs, partial ancestral graphs (Spirtes et al., 2000), or acyclic directed mixed graphs (Richardson, 2003). If for two graphs G, G' we have $M_G = M_{G'}$, we say G and G' are *Markov equivalent* and for any graph G we call the set of DAGs that are Markov equivalent to G the *Markov equivalence class* of G.

A graph is *Markovian* to an independence model M if $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G \implies (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M$ and it is *faithful* if $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M \implies (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in M_G$. A graph is called a *minimal I-map* if it is Markovian but removing any edge would make it non-Markovian. Again, we slightly abuse notation and say G is Markovian to a set of CI-statements L when all *in*dependences in G are contained in L. For independence models M, M' over $\mathbf{V}$ we define the *Markov distance* with respect to $S \subseteq \mathrm{CI}(\mathbf{V})$ via

$$\mathrm{MD}_S(M, M') = \sum_{s \in S} \mathbb{I}\left[(s \in M) \neq (s \in M')\right].$$

We omit the subscript if $S = \mathrm{CI}(\mathbf{V})$. In this case, Wahl and Runge (2025) call this *s/c-metric* and show that it is a proper metric for the space of Markov equivalence classes. We extend the definition to a graph G by considering the induced independence model M_G, i.e. we define $\mathrm{MD}_S(G, M) = \mathrm{MD}_S(M_G, M)$ and $\mathrm{MD}_S(M, G) = \mathrm{MD}_S(M, M_G)$.

Bouckaert (1995) uses the following definition to graphically characterise CI-statements that follow via the Graphoid axioms (definition 2), as used in corollary 6.

Definition 7 (coupling) Let G be an undirected graph over $\mathbf{V}$ and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$. Then $\mathbf{X}$ and $\mathbf{Y}$ are *coupled* given $\mathbf{Z}$ if there are $X \in \mathbf{X}, Y \in \mathbf{Y}$ or $Y \in \mathbf{X}, X \in \mathbf{Y}$ such that

$$(X - Y) \in G \quad \text{and} \quad \mathrm{Adj}(X) \subset \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}.$$

Now let G be a DAG. Then $\mathbf{X}$ and $\mathbf{Y}$ are *coupled* given $\mathbf{Z}$ if there are $X \in \mathbf{X}, Y \in \mathbf{Y}$ or $Y \in \mathbf{X}, X \in \mathbf{Y}$ such that

$$(X \to Y) \in G, \qquad \mathrm{PA}(Y) \subset \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},$$

and there is a $\mathbf{Q} \subseteq \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} \setminus \{X, Y\}$ such that

$$\mathbf{Z} \subseteq \mathbf{Q} \text{ and } X \perp_{G-(X \to Y)} Y \mid \mathbf{Q}.$$

B OMITTED PROOFS

Proof (Proof for proposition 1) Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint and $\epsilon, \delta > 0$. Property 1) follows immediately from the definition of the conditional mutual information. For 2), suppose we have $I(\mathbf{X} : \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq \epsilon$. This means

$$\epsilon \geq I(\mathbf{X} : \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) = I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) + I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z}) \geq I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z}),$$

where the equality follows from the chain rule of conditional mutual information and the last inequality from the non-negativity of conditional mutual information. We get $\epsilon \geq I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z})$ symmetrically. For 3), we can also apply the chain rule. Let $I(\mathbf{X} : \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \leq \epsilon$. Then

$$\epsilon \geq I(\mathbf{X} : \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) = I(\mathbf{X} : \mathbf{Y} \mid \mathbf{W} \cup \mathbf{Z}) + I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z}) \geq I(\mathbf{X} : \mathbf{Y} \mid \mathbf{W} \cup \mathbf{Z})$$

and analogously $I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$.

For 4), let $I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z}) \leq \epsilon$ and $I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$. Then we have

$$I(\mathbf{X} : \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) = I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) + I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z}) \leq 2\epsilon.$$

Finally, for 5), let $I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \leq \epsilon$, $I(\mathbf{X} : \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \leq \epsilon$ and $I(\mathbf{W} : \mathbf{Y} \mid \mathbf{Z}) \leq \delta$. We apply the chain rule to the quantity $I(\mathbf{X} \cup \mathbf{W} : \mathbf{Y} \mid \mathbf{Z})$ in two ways:

$$I(\mathbf{X} \cup \mathbf{W} : \mathbf{Y} \mid \mathbf{Z}) = I(\mathbf{W} : \mathbf{Y} \mid \mathbf{Z}) + \underbrace{I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W})}_{\leq \epsilon} = I(\mathbf{X} : \mathbf{Y} \mid \mathbf{Z}) + \underbrace{I(\mathbf{W} : \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{X})}_{\geq 0}.$$

Rearranging the equality and dropping the non-negative term:

$$I(\mathbf{X}:\mathbf{Y}\mid\mathbf{Z}) \leq I(\mathbf{W}:\mathbf{Y}\mid\mathbf{Z}) + \epsilon \leq \delta + \epsilon.$$

Now apply the chain rule to $I(\mathbf{X}:\mathbf{Y}\cup\mathbf{W}\mid\mathbf{Z})$:

$$I(\mathbf{X}:\mathbf{Y}\cup\mathbf{W}\mid\mathbf{Z}) = I(\mathbf{X}:\mathbf{Y}\mid\mathbf{Z}) + \underbrace{I(\mathbf{X}:\mathbf{W}\mid\mathbf{Z}\cup\mathbf{Y})}_{\leq\epsilon} \leq \delta + 2\epsilon.$$

Lemma 3 *Let M be a Graphoid independence model over a set of nodes $\mathbf{V}$, and $\pi:\mathbf{V}\to\mathbb{N}$ be an ordering of $\mathbf{V}$. Let L_π be given as follows: for $X,Y,\mathbf{Z}\in\mathrm{CI}(\mathbf{V})$ we have $\mathrm{CI}(X,Y\mid\mathbf{Z})\in L_\pi$ iff*

$$\pi(X)<\pi(Y) \text{ and } \mathbf{Z}=\{Z\in\mathbf{V}:\ \pi(X)\neq\pi(Z)<\pi(Y)\}.$$

For $X\in\mathbf{V}$ let $\mathbf{V}_X=\{V\in\mathbf{V}\ :\pi(V)<\pi(X)\}$, and $\mathbf{P}_X$ be the smallest subset of $\mathbf{V}_X$ such that $X\perp\!\!\!\perp\mathbf{V}_X\backslash\mathbf{P}_X\mid\mathbf{P}_X$ if such a set exists and else $\mathbf{P}_X=\mathbf{V}_X$. Then the statements

$$X\perp\!\!\!\perp\mathbf{V}_X\setminus\mathbf{P}_X\mid\mathbf{P}_X$$

follow via Graphoid axioms for all $X\in\mathbf{V}$ from L_π if such a set exists. Further, $\mathbf{P}_X$ contains exactly the nodes $Y\in\mathbf{V}_X$ with $(X\not\perp\!\!\!\perp Y\mid\mathbf{V}_X\setminus\{Y\})\in L_\pi$. I.e. the statements from the causal input list *(Bouckaert, 1995) follow from L_π.*

Proof (Proof for lemma 3) We prove this statement by induction. For $|\mathbf{V}|=2$ and w.l.o.g. $\mathbf{V}=\{X_1,X_2\}$ the statement holds trivially, as L contains $X_1\perp\!\!\!\perp X_2\mid\emptyset$ iff they are independent. Let $|\mathbf{V}|=k\in\mathbb{N}_{>2}$ and the statement be true for $k-1$. Then the statement holds for any $X\in\mathbf{V}$ with $\pi(X)<k-1$, since $\mathbf{V}_X$ already follows from the tests that do not include the last node.

Let X_k be the last node in the ordering, and

$$\mathbf{Q}_k=\{V\in\mathbf{V}_{X_k}:V\not\perp\!\!\!\perp X_k\mid\mathbf{V}_{X_k}\setminus\{V\}\},$$

i.e., our candidate set for $\mathbf{P}_{X_k}$. We first want to see that

$$X_k\perp\!\!\!\perp\mathbf{V}_{X_k}\setminus\mathbf{Q}_k\mid\mathbf{Q}_k.$$

If $\mathbf{Q}_k=\mathbf{V}_{X_k}$ this holds trivially and with $|\mathbf{V}_{X_k}\setminus\mathbf{Q}_k|=1$ the required statement is contained in L_π. So let $X_i,X_j\in\mathbf{V}_{X_k}$ be two distinct nodes such that $(X_k\perp\!\!\!\perp X_i\mid\mathbf{V}_{X_k}\setminus\{X_i\})\in L_\pi$ and $(X_k\perp\!\!\!\perp X_j\mid\mathbf{V}_{X_k}\setminus\{X_j\})\in L_\pi$. By application of Graphoid axiom number 5, we get

$$X_k\perp\!\!\!\perp X_i\mid\mathbf{V}_{X_k}\setminus\{X_i\})\ \wedge\ X_k\perp\!\!\!\perp X_j\mid\mathbf{V}_{X_k}\setminus\{X_j\}\implies X_k\perp\!\!\!\perp\{X_i,X_j\}\mid\mathbf{V}_{X_k}\setminus\{X_i,X_j\}.$$

Suppose there is a third $X_l\in\mathbf{V}_{X_k}\setminus\mathbf{Q}_k$. Analogously we get

$$\begin{aligned}&X_k\perp\!\!\!\perp X_l\mid\mathbf{V}_{X_k}\setminus\{X_l\})\ \wedge\ X_k\perp\!\!\!\perp\{X_i,X_j\}\mid\mathbf{V}_{X_k}\setminus\{X_i,X_j\}\\ &\implies X_k\perp\!\!\!\perp\{X_i,X_j,X_l\}\mid\mathbf{V}_{X_k}\setminus\{X_i,X_j,X_l\}.\end{aligned}$$

This can be repeated until we get the required statement.

It remains to show that there is no smaller set with the same property. Suppose for a contradiction there is a set $\mathbf{Q}'_k$ with

$$X_k\perp\!\!\!\perp\mathbf{V}_{X_k}\setminus\mathbf{Q}'_k\mid\mathbf{Q}'_k.$$

Then there is at least one node $X_i\in\mathbf{Q}_k\setminus\mathbf{Q}'_k$. Then we can rewrite the statement above and apply Graphoid axiom 3 to get

$$X_k\perp\!\!\!\perp X_i\cup(\mathbf{V}_{X_k}\setminus\mathbf{Q}'_k)\setminus\{X_i\}\mid\mathbf{Q}'_k\implies X_k\perp\!\!\!\perp X_i\mid\mathbf{V}_{X_k}\setminus\{X_i\}.$$

But by construction, $\mathbf{Q}_k$ contains all nodes V that are dependent on X_k given $\mathbf{V}_{X_k}\setminus\{V\}$. So this is a contradiction.

Proof (Proof for corollary 1) Let M be a Graphoid independence model over a set of nodes $\mathbf{V}$, and $\pi : \mathbf{V} \to \mathbb{N}$ be an ordering of $\mathbf{V}$. Let L_π be the list of tests from the statement, i.e. for $X, Y, \mathbf{Z} \in \mathrm{CI}(\mathbf{V})$ we have $\mathrm{CI}(X, Y \mid \mathbf{Z}) \in L_\pi$ iff $\pi(X) < \pi(Y)$ and

$$\mathbf{Z} = \{Z \in \mathbf{V} : \pi(X) \neq \pi(Z) < \pi(Y)\}.$$

Let $G(\pi)$ be the DAG over $\mathbf{V}$ with an edge from $X \in \mathbf{V}$ to $Y \in \mathbf{V}$ iff $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \in L_\pi$ for some $\mathbf{Z} \subseteq \mathbf{V}$.

By lemma 3 the set L_π from corollary 1 gives

$$X \perp\!\!\!\perp \mathbf{V}_X \setminus \mathbf{P}_X \mid \mathbf{P}_X$$

for the smallest subset of $\mathbf{V}_X$ such that this holds via Graphoid axioms and the graph $G(\pi)$ constructed contains an edge from every node in $\mathbf{P}_X$ to X. Then we can apply Theorem 2 and Corollary 2 from Verma and Pearl (1990) to see that this graph is a minimal I-map of the underlying independence model M. By definition, this means that for all disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$ we get

$$\mathbf{X} \perp_{G(\pi)} \mathbf{Y} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp_M \mathbf{Y} \mid \mathbf{Z},$$

and therefore all *in*dependence statements are GA-redundant with respect to L_π.

Proof (Proof for corollary 2) Let M be a Graphoid independence model over a set of nodes $\mathbf{V}$. Let L be the set of CI-statements

$$L = \{\mathrm{CI}(X, Y \mid \mathbf{V} \setminus \{X, Y\}) : X, Y \in \mathbf{V},\ X \neq Y\}.$$

Let $G(L)$ be the undirected graph over $\mathbf{V}$ with an edge between $X, Y \in \mathbf{V}$ iff $(X \not\perp\!\!\!\perp Y \mid \mathbf{V} \setminus \{X, Y\}) \in L$.

By direct application of Theorem 12 by Geiger and Pearl (1993), we get

$$\mathbf{X} \perp_{G(L)} \mathbf{Y} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp_M \mathbf{Y} \mid \mathbf{Z},$$

i.e. that all *in*dependences are GA-redundant with respect to L.

Proof (Proof for proposition 2) Let M be the independence model of a positive distribution, L be a set of CI-statements and $s := (X, Y, \mathbf{Z}) \in \mathrm{CI}(\mathbf{V})$ with $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \notin L$. Let G be a graphical model such that G is Markovian to L and $X \not\perp_G Y \mid \mathbf{Z}$. We will prove the statement by constructing a Graphoid independence model that contains all CI-statements in L but not $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ by building a distribution that is Markovian but not faithful to a DAG.

If X and Y are only d-connected given $\mathbf{Z}$ by a direct edge, we set $G' = (V, E \setminus \{(X, Y), (Y, X)\})$.

Otherwise, let $N(X \sim Y \mid \mathbf{Z})$ be the set of nodes on the active paths between X and Y given $\mathbf{Z}$ (without X and Y). We construct a new graph $G' = (V', E')$ with more nodes and more edges. See fig. 7 for an example. First, add all nodes in $V \setminus N(X \sim Y \mid \mathbf{Z})$ to V'. For each $W \in N(X \sim Y \mid \mathbf{Z})$, we add nodes W_1, W_2 to V'. For all edges $W \to W'$ with $W, W' \in N(X \sim Y \mid \mathbf{Z})$, add the edges $W_1 \to W_1'$ and $W_2 \to W_2'$. If $W' \in V \setminus N(X \sim Y \mid \mathbf{Z})$ add $W_1 \to W'$ and $W_2 \to W'$. Analogously add $W \to W_1'$ and $W \to W_2'$ if $W \in V \setminus N(X \sim Y \mid \mathbf{Z})$. Finally add $X \to W_1$ and $X \leftarrow W_1$ whenever $X \to W$ or $X \leftarrow W$ for $W \in N(X \sim Y \mid \mathbf{Z})$ and $Y \to W_2$ and $Y \leftarrow W_2$, respectively.

Clearly, in this graph we have $X \perp_{G'} Y \mid \mathbf{Z}'$, where $\mathbf{Z}'$ contains all nodes in $\mathbf{Z} \setminus N(X \sim Y \mid \mathbf{Z})$ and W_1, W_2 for $W \in \mathbf{Z} \cap N(X \sim Y \mid \mathbf{Z})$. Let P be a distribution with M as independence model (so G is Markovian to P). We can now construct a distribution P' as an intermediate step towards a distribution P'' that is Markovian and faithful to L. If $D \in \mathbf{V}$ has parents in the set of copies of $N(X \sim Y \mid \mathbf{Z})$, we copy the dependence to the original node. For other parents, we also keep the dependence the same. More formally, let $E^1, \ldots, E^n$ be the $n \in \mathbb{N}$ parents of D that are in $\mathbf{V} \setminus N(X \sim Y \mid \mathbf{Z})$, $F^1, \ldots, F^m$ be the $m \in \mathbb{N}$ parents of D that are copies of nodes in $N(X \sim Y \mid \mathbf{Z})$ and $F_0^1, \ldots, F_0^m$ be the corresponding original nodes in $N(X \sim Y \mid \mathbf{Z})$. In other words, for discrete distributions we set

$$P'(D = d \mid \mathrm{PA}(D) = \mathrm{pa}(D)) := P(D = d \mid E^1 = e^1, \ldots, E^n = e^n, F_0^1 = f^1, \ldots, F_0^m = f^m)$$

and analogously for continuous distributions. If we now construct a distribution P'' by considering all copies $W_1, W_2, \ldots$ for $W \in N(X \sim Y \mid \mathbf{Z})$ as a single vector-valued variable (and keep the canonical mapping between

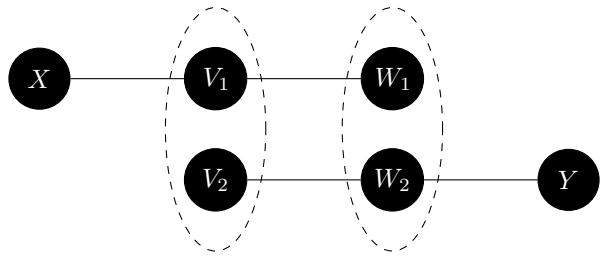


Figure 7: Example for a modified graph G' from the proof of proposition 2. All intermediate nodes on the path $X - V - W - Y$ are replaced by two copies and each new path only connects to either X or Y.

node names $W = (W_1, W_2, \dots)$), this distribution, again, contains all CI-statements we had in L but $X \perp\!\!\!\perp_{P''} Y \mid \mathbf{Z}$, as we will argue next. Further, it is also positive if P was and therefore its independence model is Graphoid.

To see the former, let $(A \perp\!\!\!\perp B \mid \mathbf{C}) \in L$. In our construction we did not add any dependences. If A and B were not connected given $\mathbf{C}$ in G, they still are not, as we did not add an edge between nodes that were disconnected and we did not introduce any colliders between these nodes. If they were connected, but still independent in P, they also still are, since each Markov kernel of P' is defined using P and a subset of the parents from G. So overall we get $A \perp\!\!\!\perp_{P''} B \mid \mathbf{C}$.

Now suppose $(A \not\perp\!\!\!\perp B \mid \mathbf{C}) \in L$, i.e., there is an active path between A and B given $\mathbf{C}$ (as otherwise G would not be Markovian to L). By assumption, A and B are not coupled over s given $\mathbf{C}$. This means there is an s-active path between them, which entails that this path does not contain any subpath from X to Y that is active given $\mathbf{Z}$. As we only modified the paths between X and Y that are active given $\mathbf{Z}$, the path between A and B is still active. So we also still have $A \not\perp\!\!\!\perp_{P''} B \mid \mathbf{C}$.

It now suffices to note that every positive distribution (and therefore particularly the one we just constructed) implies a Graphoid independence model.

Proof (Proof for lemma 1) Let $L = \{\mathrm{CI}(X, Y \mid Z) : X, Y, Z \in V, X \neq Y, X \neq Z, Y \neq Z\}$ such that $\mathrm{CI}(X, Y \mid Z) = \mathrm{CI}(Y, X \mid Z)$ for all distinct $X, Y, Z \in V$. Then for any CI-statement $\mathrm{CI}(X, Y \mid Z) \in L$, only axiom 1 from definition 2 is applicable, since all the other axioms require at least one of the operands to be a set of size larger than one, which we don't have in L. By assumption, we have no contradictions between the statements in L and the statements that follow from axiom 1. And also L contains all statements that can be derived with axiom 1.

Proof (Proof for proposition 3) Let $S = \{(X, Y, Z) : X, Y, Z \in V, X \neq Y, X \neq Z, Y \neq Z\}$. Further let $T^* \in \mathcal{T}_n$ and let M be an independence model with $\mathrm{MD}_S(T^*, M) \leq \lfloor (n-1)/2 \rfloor$ for $n \in \mathbb{N}_{>3}$. To arrive at a contradiction we assume

$$\mathrm{MD}_S(T', M) < \mathrm{MD}_S(T^*, M), \tag{1}$$

for some tree $T' \in \mathcal{T}_n$ with $T' \neq T^*$. Since the graphs differ and both are trees, there are nodes X, Y that were connected by a direct edge in T^* but are not in T'. But since they are spanning trees there still exists a path of length $k \in \mathbb{N}_{>1}$ between X and Y in T'. So there is at least one node Z such that we have $X \perp_{T'} Y \mid Z$ but $X \not\perp_{T^*} Y \mid Z$, because of the edge between X and Y in T^*. So we have at least one CI-statement difference between T' and T^*. For the second difference, there are two cases how Z is connected to X in T^*.

1. If Y lies between X and Z, i.e. $X - Y \overset{*}{-} Z$ we get $X \not\perp_{T'} Z \mid Y$ but $X \perp_{T^*} Z \mid Y$.
2. If X lies between Y and Z, i.e. $Z \overset{*}{-} X - Y$ we get $Z \not\perp_{T'} Y \mid X$ but $Z \perp_{T^*} Y \mid X$.

So in any case we have another difference between T^* and T'.

W.l.o.g. we can assume that Z is the node closest to X on the path between X and Y, i.e. the node that has a direct edge to X. Now let W be another node (which exists, because $n > 3$). There are three cases how this node can be connected to X along the unique path in T'.

1. If Y is between Z and W, i.e. if we have
$$X - Z \overset{*}{-} Y \overset{*}{-} W.$$
Then we have $X \perp_{T'} W \mid Z$ but $X \not\perp_{T^*} W \mid Z$, since in T^* we had the direct edge from X to Y.

2. If Y does not lie on the paths between X and W and X not on the one between Y and W, i.e. if we have the structure in fig. 8, we will further subdivide this into two subcases depending on T^*.
   (a) In T^* we have Y between X and W, i.e. $X - Y \overset{*}{-} W$. Then $X \not\perp_{T'} W \mid Y$ but $X \perp_{T^*} W \mid Y$.
   (b) In T^* we have X between Y and W, i.e. $W \overset{*}{-} X - Y$. Then we have $Y \not\perp_{T'} W \mid X$ but $Y \perp_{T^*} W \mid X$.
3. If X is between W and Y, i.e.
$$W \overset{*}{-} X - Z \overset{*}{-} Y.$$
   Then this case works symmetrically to the first one.

Since we chose W arbitrarily, we can repeat this for every node other than X, Y, Z. This means, we get $n-3$ more contradictions between the independence model of T^* and T'. So overall we have $n-1$ CI-statements where T' and T^* disagree. So even if T' fits all the $\lfloor (n-1)/2 \rfloor$ tests where T^* and M differ, it will also differ by another $\lceil (n-1)/2 \rceil$ statements from T^* and therefore also from M. This is a contradiction to eq. (1).

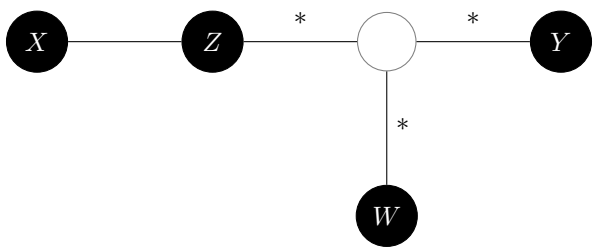


Figure 8: Visualisation how the node W from the proof of proposition 3 can be connected to X and Y in case 2 of the proof.

Proof (Proof for lemma 2) Let G^* be a DAG and consider some independence model M (that is not necessarily Markovian and faithful to G^*). Concerning (1): Suppose there is no permutation π^* such that all CI-statements in L_{π^*} are as in M_{G^*}. So for every π there is a CI-statement $\mathrm{CI}(X, Y \mid \mathbf{Z})$ that is not as in M_{G^*}. If $X \perp_{G^*} Y \mid \mathbf{Z}$, then G^* does not have an edge between X and Y. But due to $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ the graph $G(\pi)$ would have this edge. Analogously, if $X \not\perp_{G^*} Y \mid \mathbf{Z}$ then $G(\pi)$ would not have this edge. Since this holds for all permutations, SP won't output G^*.

Concerning (2): Suppose there is a permutation π with less than $|G^*|$ dependences. But since SP only adds an edge to $G(\pi)$ for every dependence, $|G(\pi)| < |G^*|$ and SP does not output G^* as it only outputs the sparsest graphs.

B.1 Omitted derivations of smaller claims

Example 1 '[...] *if we have $X_1 \not\perp\!\!\!\perp Y$ and $X_1 \perp\!\!\!\perp X_2$ we also have $X_1 \not\perp\!\!\!\perp Y \mid X_2$*'. Suppose we have $X_1 \not\perp\!\!\!\perp Y$, and $X_1 \perp\!\!\!\perp X_2$. For a contradiction assume $X_1 \perp\!\!\!\perp Y \mid X_2$ holds. Then we can apply Graphoid axiom 4 (contraction) and get
$$X_1 \perp\!\!\!\perp X_2 \wedge X_1 \perp\!\!\!\perp Y \mid X_2 \implies X_1 \perp\!\!\!\perp \{X_2, Y\}.$$
By applying axiom 2 (decomposition) we get $X_1 \perp\!\!\!\perp Y$. This is a contradiction to our assumptions.

Example 3 '[...] *note that the marginal tests already imply $X \not\perp\!\!\!\perp Y \mid Z$ and $Y \not\perp\!\!\!\perp Z \mid X$*'. Suppose we have $X \not\perp\!\!\!\perp Y, X \perp\!\!\!\perp Z$, and $Y \not\perp\!\!\!\perp Z$. Assume for a contradiction that we have $X \perp\!\!\!\perp Y \mid Z$. Like before, we can apply Graphoid axiom 4 and get
$$X \perp\!\!\!\perp Z \wedge X \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp \{Y, Z\}.$$
By applying axiom 2 (decomposition), we get $X \perp\!\!\!\perp Y$, which is a contradiction. If we assume $Y \perp\!\!\!\perp Z \mid X$, axioms 4 and 1 (symmetry) also gives us
$$Z \perp\!\!\!\perp X \wedge Z \perp\!\!\!\perp Y \mid X \implies Z \perp\!\!\!\perp \{X, Y\},$$
which results in $Z \perp\!\!\!\perp Y$ via axiom 2. This contradicts our assumptions.

'[...] *there are four tests in favour of the actual ground truth model. But there are five that would be explained, e.g., by the graph* $X \to Y \to Z$*'.* The tests in agreement with the ground truth model are

$$X \not\perp\!\!\!\perp Y, \quad X \not\perp\!\!\!\perp Y \mid Z, \quad X \perp\!\!\!\perp Z \mid Y, \quad \text{and} \quad X \perp\!\!\!\perp Z.$$

The tests that match the alternative are

$$X \not\perp\!\!\!\perp Y, \quad X \not\perp\!\!\!\perp Y \mid Z, \quad X \perp\!\!\!\perp Z \mid Y, \quad Y \not\perp\!\!\!\perp Z, \quad \text{and} \quad Y \not\perp\!\!\!\perp Z \mid X.$$

C ADDITIONAL RELATED WORK

There is also a vast literature on making the discovery of graphical models robust against statistical uncertainty. Notable examples include Kalisch and Bühlman (2007), who propose a stronger version of faithfulness under which the PC algorithm is uniformly consistent, and Bhattacharyya et al. (2021), who show finite-sample bounds for tree learning. Other approaches focus on controlling the statistical error: Strobl et al. (2019); Li and Wang (2009) use techniques from multi-hypothesis testing to control the error rate of edges. Robustness towards violations of parametric assumptions of the PC algorithm is studied by Kalisch and Bühlmann (2008); Harris and Drton (2013). Additional strategies include Wienöbst and Liskiewicz (2020); Kocaoglu (2023), who propose to use a subset of tests that is assumed to be statistically more robust and investigate which graphical structures can be identified by them. Ramsey (2016) proposes to always pick the separating sets with the highest p-value, and Li et al. (2019) choose them such that the separating sets are consistent with the final graph. Finally, Rohekar et al. (2021) proposes an algorithm that is anytime valid.

D GRAPHS AS ERROR-CORRECTING CODES

To motivate why we call some conditional independence tests 'redundant', we will phrase graph discovery as a coding problem. Suppose a sender picks a graph G from a set of graphs. Since the Markov equivalence class of this graph is identified by a sequence of CI-statements, she can encode this equivalence class in a binary string $s \in \{0,1\}^k$ for some $k \in \mathbb{N}$, where each bit represents whether a certain CI-statement holds or not. If a receiver knows the sequence of CI-statements, she can perfectly recover the equivalence class of G from s. In this scenario, the mapping $G \mapsto s$ is a coding scheme. Then a (deterministic) CI-based discovery algorithm implicitly defines a decoding scheme $s \mapsto G$. Unfortunately, the discovery algorithm rarely receives s, but a noisy version of it, since the CI-tests can have erroneous outputs. If we assume, for now, that the errors of each bit are independent, Shannon's noisy-channel coding theorem (MacKay, 2003) asserts that there is a coding scheme such that messages can be transmitted with arbitrarily small error probability as the number of sent bits approaches infinity.[8] This begs the question of what it would mean to have these *redundant* bits added to s. In a (literal) noisy channel, one could resend bits, but clearly redoing a CI-test does not give us additional information. Also, techniques like bootstrapping or cross-validation cannot help if the errors come from faithfulness violations. The main goal of this work is to address the question of which tests are suitable to add redundancy to our encoding.

E GRAPHICAL CRITERION FOR GA-REDUNDANT DEPENDENCES

Building on insights from Bouckaert (1995), we will now show a graphical criterion for dependence statements that follow in the situation of corollaries 1 and 2.

Corollary 6 (Bouckaert (1995) Thm. 3.6, 3.10) *Let M be a Graphoid independence model over $\mathbf{V}$, and G be a graph constructed like in corollaries 1 and 2. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$. If $\mathbf{X}, \mathbf{Y}$ are coupled in G given $\mathbf{Z}$, then $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is GA-redundant.*

Proof (Proof for corollary 6) Let us first consider the case of DAGs. By application of lemma 3, Theorem 2, and Corollary 2 from Verma and Pearl (1990), we get that the graph $G(\pi)$ from corollary 1 is a minimal I-map of the independence model M. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint such that $\mathbf{X}$ and $\mathbf{Y}$ are coupled given $\mathbf{Z}$ in $G(\pi)$. By Theorem 3.10 by Bouckaert (1995) we get

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z},$$

[8] It is worth noting that the relationship to the noisy-channel coding theorem is just an analogy and the theorem cannot be applied to our setting. The reason is that the theorem holds when the number of sent bits approaches infinity, while we can only conduct finitely many CI-tests.

i.e. that this dependence holds in every independence model that contains L_π. This means the dependence is GA-redundant with respect to the class of DAGs and L_π.

Similarly, for undirected graphs we also get that $G(L)$ is a minimal I-map from Theorem 12 by Geiger and Pearl (1993). Then we can apply Theorem 3.6 from Bouckaert (1995) to get

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$$

again, i.e. whenever L is contained in the independence model we get the dependence.

Further, this criterion is not only necessary but also sufficient in the scenario of corollaries 1 and 2.

Corollary 7 (coupling is necessary and sufficient) *Let M be a Graphoid independence model over $\mathbf{V}$, and $G(\pi)$ be a graph constructed like in corollary 1 and $(X, Y, \mathbf{Z}) \in \mathrm{CI}(\mathbf{V})$. Then X, Y are coupled in $G(\pi)$ given $\mathbf{Z}$ if and only if $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ is GA-redundant with respect to L_π.*

Now let $G(L)$ be a graph constructed like in corollary 2. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint with $\mathbf{X} \neq \emptyset \neq \mathbf{Y}$. Then $\mathbf{X}, \mathbf{Y}$ are coupled in $G(L)$ given $\mathbf{Z}$ if and only if $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is GA-redundant with respect to L.

Proof (Proof for corollary 7) Similarly to before we can use lemma 3 to get the statements

$$X \perp\!\!\!\perp \mathbf{V}_X \setminus \mathbf{P}_X \mid \mathbf{P}_X,$$

for every $X \in \mathbf{V}$, where $\mathbf{V}_X = \{V \in \mathbf{V} \ : \pi(V) < \pi(X)\}$, and $\mathbf{P}_X$ is the smallest subset of $\mathbf{V}_X$ such that $X \perp\!\!\!\perp \mathbf{V}_X \setminus \mathbf{P}_X \mid \mathbf{P}_X$ if such a set exists and else $\mathbf{P}_X = \mathbf{V}_X$. This also entails $X \not\perp\!\!\!\perp Y \mid \mathbf{P}_X \setminus \{Y\}$ for all $Y \in \mathbf{P}_X$, which can be seen by constructing the following contradiction: suppose we have $X \perp\!\!\!\perp Y \mid \mathbf{P}_X \setminus \{Y\}$. Together with the *in*dependence above, Graphoid axiom 4 implies

$$X \perp\!\!\!\perp \{Y\} \cup \mathbf{V}_X \setminus \mathbf{P}_X \mid \mathbf{P}_X \setminus \{Y\} = X \perp\!\!\!\perp \mathbf{V}_X \setminus (\mathbf{P}_X \setminus \{Y\}) \mid \mathbf{P}_X \setminus \{Y\},$$

which is a contradiction to the minimality of $\mathbf{P}_X$. In other words, we have the *dependency base* (Bouckaert, 1995). The first statement then follows from Theorem 3.11 from Bouckaert (1995).

Also like before, we get that $G(L)$ is a minimal I-map from Theorem 12 by Geiger and Pearl (1993). Then the second statement follows by applying Theorem 3.6 from Bouckaert (1995).

Yet, this does not hold anymore if we consider multiple additional CI-statements, as the following example shows.

Example 5 (iterated coupling) Consider an independence model that is Markovian and faithful to the graph in fig. 9. Suppose we want to learn the graph with the method from corollary 1 with a valid causal ordering π. Now suppose we conduct the test $\mathrm{CI}(X, W \mid Z)$ afterwards. X and W are not coupled given Z in $G(\pi)$, since there is no edge between X and W. But now the tests in $L_\pi \cup \{X \not\perp\!\!\!\perp W \mid Z\}$ imply $X \not\perp\!\!\!\perp Y \mid Z$. This can be seen as follows: from corollary 1 we get $X \perp\!\!\!\perp W \mid \{X, Y\}$, since an *in*dependence follows from L_π if the nodes are separated in the graph. We will now derive a contradiction. Suppose we would have $X \perp\!\!\!\perp Y \mid Z$. Then we could apply Graphoid axiom 4 to $X \perp\!\!\!\perp W \mid \{Z, Y\}$ and $X \perp\!\!\!\perp Y \mid Z$ to get $X \perp\!\!\!\perp \{Y, W\} \mid Z$. Using axiom 3 we get $X \perp\!\!\!\perp W \mid Z$. But we have conducted the respective test and found $X \not\perp\!\!\!\perp W \mid Z$. So $X \not\perp\!\!\!\perp Y \mid Z$ follows. But X and Y are not coupled given Z in $G(\pi)$. This means the dependence only follows from $L_\pi \cup \{W \not\perp\!\!\!\perp X \mid Z\}$, not from L_π alone. Particularly, coupling is not a necessary criterion for GA-redundancy anymore if we conduct more tests than the ones in L_π.

Figure 9: In the setting of corollary 1, $X \not\perp\!\!\!\perp Y \mid Z$ only follows from $L_\pi \cup \{X \not\perp\!\!\!\perp W \mid Z\}$, not from L_π alone.

This highlights the utility of proposition 2 and especially corollary 8, which we will see later. If we want to conduct multiple redundant tests, a sufficient criterion for GA-redundancy is not enough to exclude all GA-redundant tests. On the other hand, the criterion in proposition 2 may miss some PGM-redundancies, yet all tests from this criterion are guaranteed to be PGM-redundant.

F ADDITIONAL REMARKS AND EXAMPLES

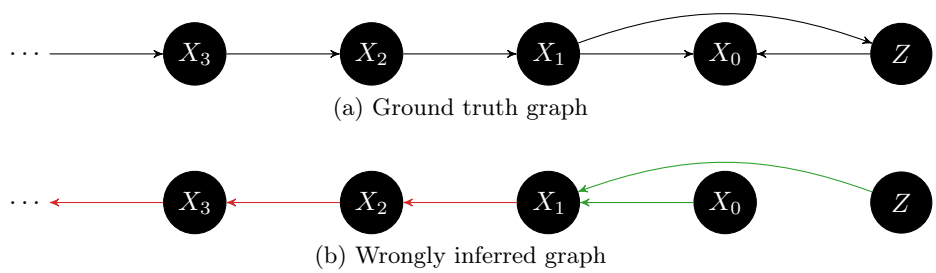
F.1 Examples

Example 6 (Wrong Collider-Structure) In this example, we assume we want to find a graph using the PC algorithm Pearl (2009). Consider the family of graphs depicted in fig. 10a. Assume there are nodes Z and X_i for $i = 0, \dots, k \in \mathbb{N}_{>1}$. Further assume all CI-test results are Markovian and faithful to the graph, except for

$$X_0 \perp\!\!\!\perp Z.$$

Then the algorithm would wrongly detect a collider structure between X_0, X_1 and Z, as X_1 is not in the separating set of X_0 and Z. Then, by application of Meek rules (Meek, 1995), this orientation propagates along the path. As in the true graph, there is no collider between X_2, X_1 and X_0, the triplet will not be oriented as collider, and by Meek rule R1 the edge $X_1 - X_2$ will become $X_1 \to X_2$. The same argument holds for all further X_{i+1}, X_i with $i > 2$, which will cause the algorithm to orient the whole chain of X_i the wrong way around, as visualised in fig. 10b. In other words, a single false CI-statement might cause $k + 1$ wrongly directed edges.

If the purpose of the graphical model is simply a concise representation of the CI-statements in the data, this might be acceptable. But if the model is interpreted, e.g., as a causal model and it is used in some downstream task, this might be worrisome. Although graph discovery is already algorithmically expensive, in such cases, a practitioner might be willing to incur an additional computational overhead to make the results more robust.



(a) Ground truth graph

(b) Wrongly inferred graph

Figure 10: When the collider-structure $X_0 \to X_1 \leftarrow Z$ is wrongly identified, all edges $X_{i+1} \to X_i$ will be oriented the wrong way around.

Example 7 (P- but not GA-redundant) In this example, we will see a case where some CI-statements are P-redundant but not GA-redundant. Suppose we have random variables $\mathbf{V} = \{X, Y, Z, W\}$ and let

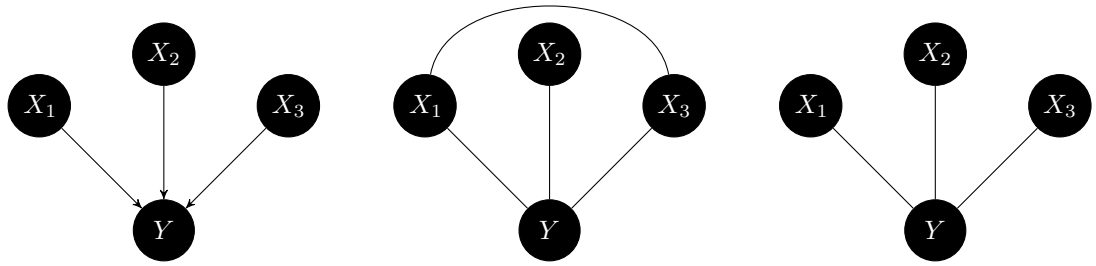
$$L = \{X \perp\!\!\!\perp Y \mid \{Z, W\}, \quad X \perp\!\!\!\perp Y \mid \emptyset, \quad Z \perp\!\!\!\perp W \mid X, \quad Z \perp\!\!\!\perp W \mid Y\}.$$

Studený (1992) (Proposition 5) showed that for all probability distributions, this entails also

$$X \perp\!\!\!\perp Y \mid Z, \quad X \perp\!\!\!\perp Y \mid W, \quad Z \perp\!\!\!\perp W \mid \{X, Y\}, \quad Z \perp\!\!\!\perp W \mid \emptyset.$$

But none of these statements follows from the Graphoid axioms, so they are P-redundant but not GA-redundant.

Example 8 (PC results can be non-Markovian) Suppose we have an independence model that is Markovian and faithful to the graph in fig. 11a, except for the dependence $X_1 \not\perp\!\!\!\perp X_3$ and the *in*dependence $X_1 \perp\!\!\!\perp X_3 \mid Y$. Further, suppose we want to recover this graph with the PC algorithm. In the first round of the algorithm, it will conduct all marginal independence tests. This will give us the intermediate skeleton in fig. 11b. In the following rounds, the algorithm will conduct all tests with conditioning set of size one and two, which will result in the graph in fig. 11c. But now in the orientation phase, we have conflicting evidence for the existence of colliders, and depending on how exactly the algorithm resolves them, we will get different results. For example, in the default implementation in `causal-learn` (Zheng et al., 2024), the algorithm simply picks the first orientation according to the (non-predetermined) ordering in which it conducts the CI-tests. Suppose the algorithm first checks whether the unshielded triplet $X_1 - Y - X_2$ is a collider. This is the case, as Y is not in the separation set of X_1 and X_2, which is the empty set. The same holds for X_2 and X_3. If the algorithm sticks with these orientations, it ignores that Y is indeed a member of all separating sets of X_1 and X_3. Therefore, it would output the graph in fig. 11a. But note that this graph does imply $X_1 \perp\!\!\!\perp X_3$, which does not hold in our independence model. It can be checked (e.g., with `Z3`) that the CI-tests that we used are indeed a valid Graphoid.



(a) PC will output this graph, even when the independence model contains $X_1 \not\perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_3 \mid Y$. But then the graph is not Markovian to the given input.

(b) Intermediate skeleton that PC finds after conducting all marginal independence tests.

(c) Final skeleton that PC finds before orienting the edges.

Figure 11: This example illustrates how the PC algorithm arrives at a graph that contradicts a dependence statement used as input to the algorithm.

F.2 Iterated sufficient criterion for PGM-redundancy

Note that proposition 2 requires a graph G to be Markovian to a set of CI-statements L. If we want to find several PGM-redundant CI-statements like in the experiment shown in fig. 6c, we can only apply proposition 2 as long as the additional tests imply dependences. If one test returns an *in*dependence and we continue to wrongly apply proposition 2, we could end up with CI-statements that follow from the previously conducted tests, as the following example shows.

Example 9 Let G be the graph given in fig. 12. Suppose we have L_π as in corollary 1 with respect to the ordering X, Y, Z, W. From proposition 2 we know that $X \not\perp\!\!\!\perp Z \mid W$ is PGM-redundant. Suppose now, we conduct this additional test and actually find $X \perp\!\!\!\perp Z \mid W$. If now we would try to read further PGM-redundant CI-statements from proposition 2, we might find statements that follow from L_π in conjunction with $X \perp\!\!\!\perp Z \mid W$. To see this, note that also $X \not\perp\!\!\!\perp Z \mid \emptyset$ would be PGM-redundant according to the criterion proposition 2. But if we already have $X \perp\!\!\!\perp Z \mid W$, also $X \perp\!\!\!\perp Z \mid \emptyset$ follows. This can be seen as follows: According to corollary 1 we can read off $X \perp\!\!\!\perp W \mid Z$ from G. Then we can apply Graphoid axiom 4 and get

$$X \perp\!\!\!\perp Z \mid \{W\} \cup \emptyset \;\wedge\; X \perp\!\!\!\perp W \mid \{Z\} \cup \emptyset \implies X \perp\!\!\!\perp \{W, Z\} \mid \emptyset,$$

and by application of Graphoid axiom 2, we get the required statement.

Figure 12: Both $X \not\perp\!\!\!\perp Z \mid W$ and $X \not\perp\!\!\!\perp Z$ are PGM-redundant. But if we conduct a test and find $X \perp\!\!\!\perp Z \mid W$ also $X \perp\!\!\!\perp Z$ follows.

But the following corollary states that we can continue in a similar fashion if we check separations with the graph G' from the proof of proposition 2.

Corollary 8 (Iterated PGM-redundancy) *Let M be the independence model of a positive distribution, L be a set of CI-statements and $s := (X, Y, \mathbf{Z}) \in \mathrm{CI}(\mathbf{V})$ with $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \notin L$. Let G be a graphical model such that G is Markovian to L and $X \not\perp_G Y \mid \mathbf{Z}$. Let G' be a graphical model like in the proof of proposition 2, i.e. a graphical model such that* separation *of (or given) a node W that has been copied is defined as the usual separation of (or given) the set of its $k \in \mathbb{N}$ copies $\{W_1, \ldots, W_k\}$. Further let G' be Markovian to L and $X \not\perp_{G'} Y \mid \mathbf{Z}$. If there is no $(A \not\perp\!\!\!\perp B \mid \mathbf{C}) \in L$ such that A is coupled over s with B given $\mathbf{C}$, then s is PGM-redundant given L.*

Further, the graph G' constructed in the proof of proposition 2 is Markovian (in terms of the above definition of separation) to $L \cup \{\neg s\}$.

Proof (Proof for corollary 8) The proof works analogously to the proof of proposition 2.

F.3 Relationship between coupling over nodes and coupling

The similarity in names between definitions 6 and 7 implies that these concepts are related. Obviously, they slightly differ in their scope, as coupling is defined for three sets of variables, while decoupling is concerned with two triplets consisting of two variables and a set of variables, respectively. In the following, we will see how these notions can coincide.

Lemma 4 (coupling implies coupling over themselves) *Let G be a DAG or undirected graph over $\mathbf{V}$. If $(X, Y, \mathbf{Z}) \in \mathrm{CI}(\mathbf{V})$ are coupled then $(X, Y, \mathbf{Z})$ are coupled over $(X, Y, \mathbf{Z})$.*

Proof Suppose X and Y are coupled given $\mathbf{Z}$. Then by definition, there is an edge between X and Y. If G is a DAG, we have $X \perp_{G-(X\to Y)} Y \mid \mathbf{Z}$ (since $\mathbf{X}$ and $\mathbf{Y}$ are singleton sets and thus $\mathbf{Q} = \mathbf{Z}$). But this means, the edge $X \to Y$ is the only active path between X and Y given $\mathbf{Z}$. Analogously, if G is an undirected graph, we have $\mathrm{Adj}(X) \subseteq \{Y\} \cup \mathbf{Z}$ (or symmetrically for $\mathrm{Adj}(Y)$). Again, this means the edge $X - Y$ is the only active path given $\mathbf{Z}$. But this path is not an $(X, Y, \mathbf{Z})$-active path, since it contains X and Y and is active given $\mathbf{Z}$. So, $(X, Y, \mathbf{Z})$ are coupled over $(X, Y, \mathbf{Z})$.

The question that remains is whether coupling and coupling over s encode the same concept. The next example shows a case where they differ.

Example 10 (coupling and coupling over nodes differ) Let G be the graph in fig. 13. Then A and B are coupled over $(X \not\perp\!\!\!\perp Y \mid \emptyset)$ given $\emptyset$, since there is one active path (A, X, Z, Y, B) but also the sub-path (X, Z, Y) is active given $\emptyset$. So in particular, if $(A \not\perp\!\!\!\perp B \mid \emptyset) \in L$, we cannot conclude that $X \not\perp\!\!\!\perp Y$ is PGM-redundant. But neither A and B nor X and Y are coupled given $\emptyset$, as they are not adjacent.

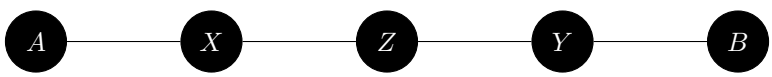
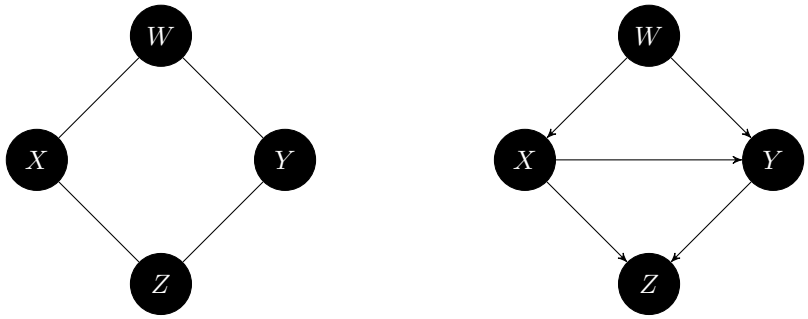


Figure 13: A and B are *coupled over* $(X, Y, \emptyset)$ given $\emptyset$ (as defined in definition 6) but neither A and B nor X and Y are *coupled* given $\emptyset$ (as defined in definition 7) as they are not adjacent.

Since coupling can be a sufficient and necessary condition in certain cases, as corollary 7 shows, it is natural to ask whether proposition 2 is also necessary in this setting. The following example shows that this is not the case.

Example 11 (coupling over nodes is not necessary) Suppose we have an independence model that is Markovian and faithful to the graph G in fig. 14. If we estimate a graph using the procedure in corollary 1 with the ordering $\pi = (W, X, Y, Z)$ and tests L_π, we will arrive at the graph $\hat{G}$. Now G entails $X \perp\!\!\!\perp Y \mid \{W, Z\}$ while $\hat{G}$ entails the opposite. So clearly, this statement is PGM-redundant with respect to L_π. Yet, X and Y are coupled over $(X, Y, \{W, Z\})$ given W in $\hat{G}$. So proposition 2 does not detect this PGM-redundancy.



(a) Graph G (b) Graph $\hat{G}$

Figure 14: The *in*dependence $X \perp\!\!\!\perp Y \mid \{W, Z\}$ in G is PGM-redundant given the test from corollary 1, since G and $\hat{G}$ differ in this statement but not in the previous tests. But X and Y are coupled over $(X, Y, \{W, Z\})$ given W, so proposition 2 cannot be applied.

G THE MMD ALGORITHM

In the following, we will propose an assumption and an algorithm that can correct violations of (a) in lemma 2 in some cases, but at the price of being even more expensive than the SP algorithm. The main purpose of this is not to propose a practical alternative to SP, but rather to demonstrate that the utility of redundant CI-statements is not tied to a specific algorithm or assumption.

Assumption 1 (Minimum Markov Distance) *Let $\mathcal{G}$ be a set of graphical models, $G^* \in \mathcal{G}$, and M an independence model. We say G^* and M fulfil the* minimum Markov distance assumption *(MMD) iff*

$$G^* \in \arg\min_{G \in \mathcal{G}} \mathrm{MD}(G, M).$$

Not surprising, assumption 1 is weaker than faithfulness as proposition 4 shows.

Proposition 4 *Let $G^* \in \mathcal{G}$ be a graphical model, and M be an independence model that is Markovian and faithful to G^*. Then G^* and M also fulfil MMD. Further, there are independence models M that fulfil MMD relative to G but are not Markovian and faithful to any graph $G \in \mathcal{G}$.*

Proof (Proof of proposition 4) Let $G^* \in \mathcal{G}$ be a graphical model and M be an independence model that is Markovian and faithful to G^*. Then by definition we have $M = M_{G^*}$. This means $\mathrm{MD}(M, G^*) = 0$ and since the Markov distance is non-negative, surely $\mathrm{MD}(M, G^*) \leq \mathrm{MD}(M, G)$ for any DAG G. To see that there are cases where MMD holds but M is not faithful to any DAG, consider an independence model M that is Markovian and faithful to the graph G in fig. 15 except for the *in*dependences $X_1 \perp\!\!\!\perp X_2 \mid X_4$ and $X_2 \perp\!\!\!\perp X_3 \mid X_1$. For faithfulness, a graph G would have to contain no edge between X_1 and X_2. But then the dependence $X_1 \not\perp\!\!\!\perp X_2$ would violate the Markov property. Therefore, there is no DAG that is Markovian and faithful to the given independence model, yet MMD holds.

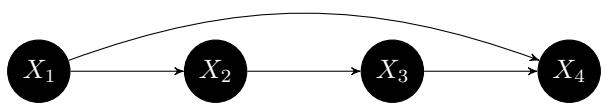


Figure 15: If this graph is the ground truth but we have the unfaithful *in*dependences $X_1 \perp\!\!\!\perp X_2 \mid X_4$ and $X_2 \perp\!\!\!\perp X_3 \mid X_1$, there is no Graph that is Markovian and faithful to the observed independence model. Further, MMD is fulfilled but SP outputs a different graph.

Perhaps more interesting, example 12 shows a case where SP fails but MMD still recovers the correct graph.

Example 12 (MMD holds but not SP) Consider again an independence model M that is Markovian and faithful to the graph G in fig. 15 except for the *in*dependences $X_1 \perp\!\!\!\perp X_2 \mid X_4$ and $X_2 \perp\!\!\!\perp X_3 \mid X_1$. It can be verified easily that G and M fulfil MMD with $\mathrm{MD}(G, M) = 2$. But SP cannot recover G, as the algorithm would output $G - (X_2 \to X_3)$.

Yet, MMD is not strictly weaker than the SP algorithm, as example 13 demonstrates.

Example 13 (SP works but not MMD) Consider the graph G with nodes X, Y, Z but no edges. Assume further that due to false positive CI-tests, we get the *in*dependence model M with $Y \not\perp\!\!\!\perp Z$, $X \not\perp\!\!\!\perp Y \mid Z$, $Y \not\perp\!\!\!\perp Z \mid X$ and *in*dependence otherwise (note that this model is not Graphoid). Clearly $\mathrm{MD}(G, M) = 3$. But for the graph G' with the additional edge $Y \to Z$, we have $\mathrm{MD}(G', M) = 2$, so G and M do not fulfil MMD. Yet, SP would return G.

G.1 MMD Experiments

We generated a multivariate Gaussian distribution that is Markovian to a spanning tree over five nodes as described below. We then used the algorithm from proposition 3 (MMD algorithm) and a simplified version of PC to recover the tree, which we will call *TreePC*. As we can see in fig. 16, MMD can indeed profit from considering additional PGM-redundant tests.

For the data generation, we first pick a random spanning tree with five nodes. To this end, we initialise a matrix with uniformly distributed numbers from $[0, 1)$, interpret it as an adjacency matrix of a weighted graph, and find

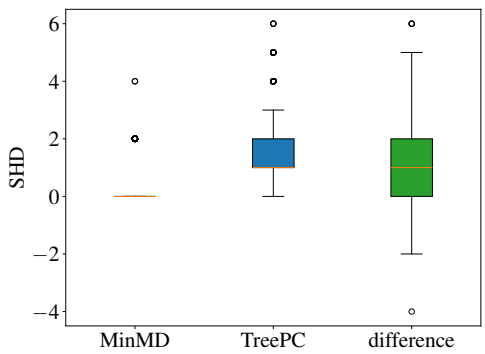


Figure 16: Structural Hamming distance for simplified PC algorithm and MMD algorithm from proposition 3. MMD outperforms PC.

a maximum weight spanning tree using Kruskal's algorithm. We then pick a node as root uniformly at random and orient all edges away from the root in a depth-first search to get a Markov equivalent DAG. Then we can recursively draw samples from this graph: we uniformly pick coefficients for a linear structural causal model from $(-1, -0.1] \cup [0.1, 1)$ and draw noise from a standard normal distribution. For each dataset, we generate 500 samples.

To find the underlying tree under the MMD assumption, we conduct all CI-tests in the set S from proposition 3 using the Fisher Z test from `causal-learn` with α-threshold 0.01. We calculate the Markov distance with respect to S for all possible spanning trees over the nodes.

As a baseline, we consider a simplified version of the PC algorithm. This way, we only use tests of the same conditioning set size and can rule out that the difference is due to the statistical condition of the problem. First, we skip the initial phase, where PC would conduct all marginal independence tests, as in a spanning tree, all nodes are dependent anyway. We then proceed with the tests with a single conditioning variable as usual. Since, in the limit of infinite data, these tests are already sufficient to identify the graph, we do not consider larger conditioning sets. Further, we stop once the current graph is a tree, as any further CI-tests could only violate the spanning property.

For each of the resulting graphs, we calculate the structural Hamming distance (Tsamardinos et al., 2006), i.e., the number of differing edges, to the ground truth graph. We repeat the experiment for 1000 datasets.

Finally, we conducted a Mann-Whitney U test for the null-hypothesis that the distributions of the SHD of MMD and TreePC are not stochastically ordered. The test yields a p-value of $p = 6.54 \cdot 10^{-103}$.

As another check to demonstrate the utility of the PGM-redundant tests, we have also conducted an additional experiment, similar to fig. 15. We compared three different sets of tests that we could use in the optimization. One set is the baseline, one contains additional GA-redundant tests, and the third one contains additional non-GA-redundant tests. We observed that the additional GA-redundant tests did not change the results, while the additional non-GA-redundant tests significantly improved the results.

More formally, we start with the TreePC algorithm that we used as a baseline before. Denote the set of tests conducted by TreePC with S', the set of tests from MinMD as in proposition 3 with S, and the empirical independence model as M. Note that $S' \subseteq S$. To find GA-redundant CI-statements, we applied axiom 5 of the Graphoid axioms (Intersection) to all combinations of tests in S. This way, we get GA-redundant marginal independence statements that we denote with U (this is not a comprehensive list of GA-redundant statements). Note that the tests in U should statistically be better conditioned than the tests in S. For the derivation of the statements in U, we did not use all tests in S. Denote with W the subset of S of CI-statements that we used to derive U. As a baseline, we took the spanning tree with minimal Markov distance on $S' \cup W$, i.e., $\min_{T \in \mathcal{T}} \mathrm{MD}_{S' \cup W}(T, M)$. Call this algorithm A_1. The rationale for including S' is that we have at least a set of tests that is sufficient to identify the graph in the limit of infinite data. Then we include the GA-redundant tests and optimize $\min_{T \in \mathcal{T}} \mathrm{MD}_{S' \cup U \cup W}(T, M)$. Let this be A_2. We finally compare this with the MinMD version from

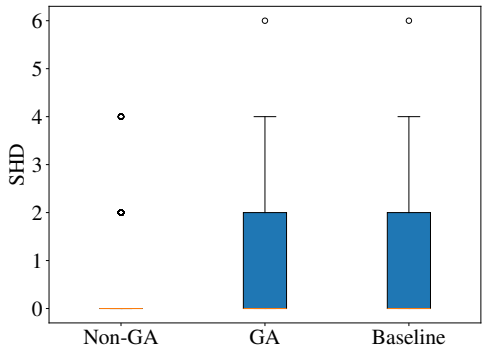


Figure 17: Structural Hamming distance when we modify MMD to optimize over different sets of tests. Additional GA-redundant tests do not change the result. Additional non-GA-redundant tests improve the results significantly.

the original experiment (which optimizes over S) and call it A_3. This way, we have one algorithm that optimizes over $S' \cup W$ (A_1), one that uses the same tests plus GA-redundant tests (A_2), and one that uses additional non-GA-redundant tests (A_3).

In our experiment, visualised in fig. 17, we see that A_1 outperforms TreePC already, as it is restricted to the correct model class and also has some PGM-redundant tests. We further see that A_1 and A_2 do not differ at all (in the considered experiments). This is what we would expect, as the GA-redundant tests contain no additional information. Finally, A_3 performs significantly (with respect to the Mann-Whitney U test) better than A_1 and A_2 with a p-value of $p = 2.49 \cdot 10^{-12}$.

The fraction of GA-redundant tests ranges between 0 and 41.9%, with 15.5% on average. Under faithfulness and in the limit of infinite data, we should not be able to apply axiom 5, i.e., we are only able to derive GA-redundant statements with axiom 5 when there are at least two tests with insufficient power. Thus, we also tried the same experiment with a sample size of 100 to increase the fraction of GA-redundant tests. Indeed, we can find more GA-redundant tests like this, namely a fraction of 29.1% on average. But we still get qualitatively the same result with p-value of $p = 4.18 \cdot 10^{-7}$.

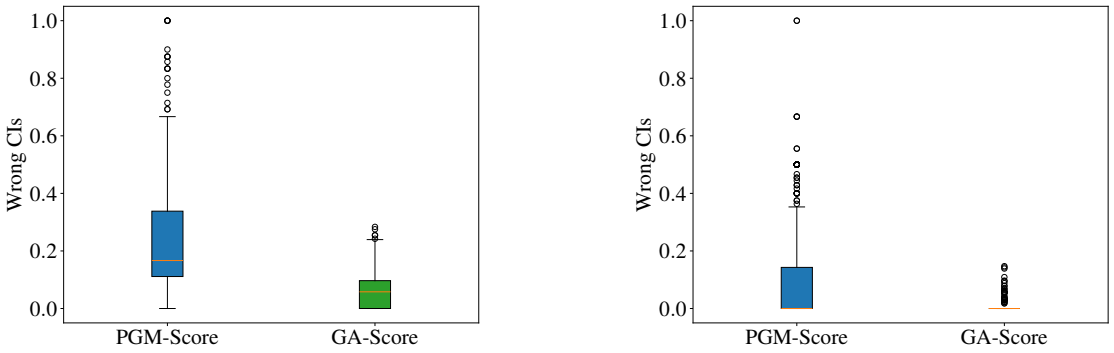
H ADDITIONAL EXPERIMENTS

In the experiment in fig. 6c, we have seen that the PGM-redundant CI-tests seem to indicate more wrong predictions than the GA-redundant tests. In the following experiments, we want to investigate how the redundant tests behave in other scenarios where we would expect them to indicate errors and scenarios where the learned graphs are mostly correct. To this end, we generate a DAG with five nodes according to an Erdös-Rényi model with edge probability 0.3. We then uniformly draw coefficients for a linear structural equation model from $(-1, -0.1] \cup [0.1, 1)$. For each variable, we add Gaussian noise with zero mean and unit variance. For fig. 18a we generate 50 samples, while for fig. 18b we generate 2000 samples.

We learn DAGs using the procedure from corollary 1 with the Fisher Z test from `causal-learn` and a topological ordering from the ground truth graph. We then find PGM-redundant CI-statements via proposition 2 and corollary 8 and GA-redundant CI-statements via corollary 1 and corollary 6. After conducting a test, we add it to the set of previously conducted tests. We report the fraction of wrong predictions by the graphical model, where we ignore examples where we did not find any PGM-redundant or GA-redundant CI-statements, respectively. This reduces the effective sample size to 364 samples in fig. 18a and 750 in fig. 18b.

The small dataset is supposed to yield worse graph discovery results and therefore is expected to show plenty of wrongly predicted CI-statements, and indeed we observe that the correct ground truth graph is only recovered for 21.8% of the learned graphs, while with the larger sample we recover the correct graph in 98% of the cases. And indeed we see that in the latter case the difference between the tests is considerably smaller, as we were hoping. Finally, we conducted two Mann-Whitney U tests for the null-hypotheses that the distributions of the test errors are not stochastically ordered. For fig. 18a we get a p-value of $p = 8.94 \cdot 10^{-36}$. For fig. 18b we get $p = 2.55 \cdot 10^{-72}$,

which is also clearly significant. We hypothesise that mainly the extreme values of the distribution are responsible for the latter p-value, as the empirical median is identical, unlike in fig. 18a. In conclusion, we think that these experiments demonstrate that the additional errors indicated by the PGM-redundant CI-tests in fig. 6c cannot be explained by, e.g., the PGM-redundant tests being more error-prone by nature. This corroborates our hypothesis that they are a promising tool to evaluate graph discovery.



(a) Dataset with 50 samples.

(b) Dataset with 2000 samples.

Figure 18: Incorrect predictions of PGM- versus GA-redundant CI-statements on synthetic datasets with five nodes and different sample sizes.

I EXPERIMENTAL DETAILS

Figure 6a For the experiment in fig. 6a, we first generate a DAG with four nodes according to an Erdös-Rényi model with edge probability 1/2. We then uniformly draw coefficients for a linear structural equation model (Pearl, 2009) from $(-1, -0.1] \cup [0.1, 1)$. For each variable, we add Gaussian noise with zero mean and unit variance. We generate 2000 samples for each dataset. Then, we randomly permute a list containing all triplets in $\mathrm{CI}(\mathbf{V})$, where $\mathbf{V}$ is the set of nodes. For every triplet $(X, Y, \mathbf{Z})$ in this list, we conduct a Fisher Z test for independence, implemented in `causal-learn` (Zheng et al., 2024). We also check with the `Z3` SMT solver (De Moura and Bjørner, 2008) whether the result of $\mathrm{CI}(X, Y \mid \mathbf{Z})$ follows from the previously conducted tests (with α-threshold 0.01) via Graphoid axioms. If so, we add the p-value of the test to either the list of implied dependences or *in*dependences, respectively. We repeat the same experiment 16 times and keep adding the p-values to the same lists. The plot in fig. 6a shows these two lists of p-values and α. Figure 19 shows how many of the tests were found to be GA-redundant.

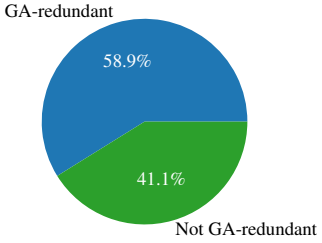


Figure 19: Fraction of all tests that were identified as GA-redundant in the experiment in fig. 6a.

Figure 6b For the experiment in fig. 6b, we generate two kinds of datasets, one generated by a DAG and the other by an undirected graphical model. The ground truth graphs can be seen in fig. 20. All variables are binary. For the DAG, we draw coefficients for the conditional distributions of each variable given all possible values of its

parents. We then recursively sample the values along the topological ordering of the graph, starting from W. We drew $P(X = 1 \mid W = 0)$, $P(Y = 1 \mid W = 0)$, $P(Z = 1 \mid X = 0, Y = 0)$ and $P(Z = 1 \mid X = 1, Y = 1)$ uniformly from $[0.3, 0.7)$ and then set

$$\begin{aligned} P(X = 1 \mid W = 1) &= 1 - P(X = 1 \mid W = 0) \\ P(Y = 1 \mid W = 1) &= 1 - P(Y = 1 \mid W = 0) \\ P(Z = 1 \mid X = 0, Y = 1) &= 1 - P(Z = 1 \mid X = 0, Y = 0) \\ P(Z = 1 \mid X = 1, Y = 0) &= 1 - P(Z = 1 \mid X = 1, Y = 1). \end{aligned}$$

For the undirected model, we used the `pgmpy` package (Ankan and Textor, 2024). For each edge $X - Y$ in the graph, we add a factor ϕ to a factor graph model, where we pick the value $\phi(X = 0, Y = 0)$ and $\phi(X = 1, Y = 1)$ uniformly from $[0.1, 0.3)$ and set

$$\begin{aligned} \phi(X = 0, Y = 1) &= 1 - \phi(X = 0, Y = 0) \\ \phi(X = 1, Y = 0) &= 1 - \phi(X = 1, Y = 1). \end{aligned}$$

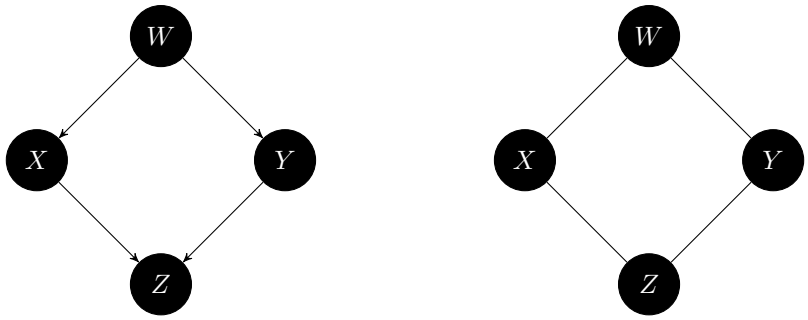
We then draw the distribution using Gibbs sampling with the default parameters of `pgmpy`. Note that neither of the graphs' independence models is Markovian and faithful to the other. For each dataset we generated 2000 samples.

We then use the methods described in corollaries 1 and 2 to construct an undirected graphical model and a DAG on each dataset, where we used the χ^2-test implemented in `causal-learn` with α-threshold 0.01.

Finally, we identify GA-redundant tests using corollaries 1, 2 and 6. We exclude these tests and report the fraction of remaining unused CI-tests, where the graphical implication contradicts the empirical test result. There were no cases where all statements were detected as GA-redundant CI-statements.

In most practical settings it would be more relevant to use the PGM-tests detected by proposition 2. Yet, in this toy scenario, we have already seen in example 11 that the only test that would reveal the graph misspecification between the undirected ground truth and a learned DAG would not be detected by proposition 2 (although it is PGM-redundant). The experiment nonetheless shows that we pick up the relevant signal if we exclude all detected GA-redundant tests.

We repeated the experiment 1000 times. Additionally, we conduct Mann-Whitney U tests for the null-hypotheses that the distributions of errors of the DAGs and the undirected graphs are not stochastically ordered. For the datasets generated by DAGs, we get $p = 3.03 \cdot 10^{-23}$ and for the datasets from the undirected model we get a p-value below machine precision.



(a) Ground truth for the DAG datasets in fig. 6b.

(b) Ground truth for the undirected graph datasets in fig. 6b.

Figure 20: Ground truth graphs for the experiment in fig. 6b.

Figure 21 shows the fraction of the tests that were identified as GA-redundant by the criterion from corollaries 1, 2 and 6 among $\mathrm{CI}(\mathbf{V}) \setminus L$ and $\mathrm{CI}(\mathbf{V}) \setminus L_\pi$, respectively. In other words, the fraction of tests that follow from the tests used in corollaries 1 and 2 among all others.

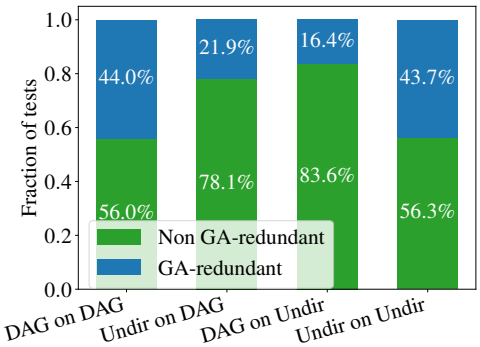
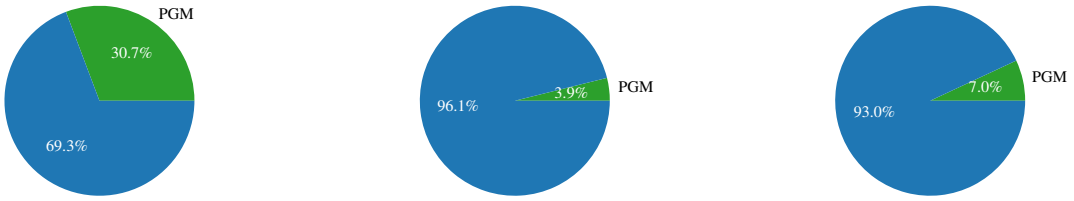


Figure 21: Fraction of all unused tests that were identified as GA-redundant in the experiments in fig. 6b.

Figure 6c For the experiment in fig. 6c, we use the dataset from Sachs et al. (2005) as provided in the `causal-learn` package and the ground truth graph given in the original paper. We repeat the experiment 1000 times. For each run, we draw a causally sufficient subset (i.e., a subset such that according to the ground truth graph, there is no hidden confounder) of five variables using rejection sampling, and we draw a bootstrap sample of the same size as the original dataset. We then apply the algorithm from corollary 2 using a topological ordering of the ground truth graph and the Fisher Z test to get a DAG. As the significance threshold of the tests, we used $\alpha = 0.001$. In no instance the expected ground truth was recovered. From this DAG, we derive PGM-redundant CI-statements via proposition 2 and corollary 8. We conduct these tests and report the fraction of CI-tests, where the graphical implication contradicts the empirical test result. After conducting a test, we add it to the set of previously conducted tests. Finally, we conducted a Mann-Whitney U test for the null-hypothesis that the distribution of errors is not stochastically ordered. We found a p-value of $p = 2.92 \cdot 10^{-144}$.

Figure 22 shows the fraction of the tests that were identified as PGM-redundant with respect to L_π and following tests by the criterion from proposition 2 and corollary 8 among $\mathrm{CI}(\mathbf{V}) \setminus L_\pi$, with L_π as in corollary 1.



(a) Sachs dataset. (b) Dataset with 50 samples. (c) Dataset with 2000 samples.

Figure 22: Fraction of all unused tests that were identified as PGM-redundant in the experiments in figs. 6c and 18.

All experiments, including the ones in appendices G and H, were run on an Apple M3 Pro processor with 18 GB RAM. The experiment for fig. 6a took roughly a day, while all other experiments finished in a couple of minutes.