# Invariant Ancestry Search

Phillip B. Mogensen*[1], Nikolaj Thams[1], and Jonas Peters[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark

**Abstract**

Recently, methods have been proposed that exploit the invariance of prediction models with respect to changing environments to infer subsets of the causal parents of a response variable. If the environments influence only few of the underlying mechanisms, the subset identified by invariant causal prediction, for example, may be small, or even empty. We introduce the concept of minimal invariance and propose invariant ancestry search (IAS). In its population version, IAS outputs a set which contains only ancestors of the response and is a superset of the output of ICP. When applied to data, corresponding guarantees hold asymptotically if the underlying test for invariance has asymptotic level and power. We develop scalable algorithms and perform experiments on simulated and real data.

## 1 Introduction

Causal reasoning addresses the challenge of understanding why systems behave the way they do and what happens if we actively intervene. Such mechanistic understanding is inherent to human cognition, and developing statistical methodology that learns and utilizes causal relations is a key step in improving both narrow and broad AI (Jordan, 2019; Pearl, 2018). Several approaches exist for learning causal structures from observational data. Approaches such as the PC-algorithm (Spirtes et al., 2000) or greedy equivalence search (Chickering, 2002) learn (Markov equivalent) graphical representations of the causal structure Lauritzen (1996). Other approaches learn the graphical structure under additional assumptions, such as non-Gaussianity Shimizu et al. (2006) or non-linearity Hoyer et al. (2009); Peters et al. (2014). Zheng et al. (2018) convert the problem into a continuous optimization problem, at the expense of identifiability guarantees.

Invariant causal prediction (ICP) (Peters et al., 2016; Heinze-Deml et al., 2018; Pfister et al., 2019; Gamella & Heinze-Deml, 2020; Martinet et al., 2021) assumes that data are sampled from heterogeneous environments, and identifies direct causes of a target $Y$, also known as causal parents of $Y$. Learning ancestors (or parents) of a response $Y$ yields understanding of anticipated changes when intervening in the system. It is a less ambitious task than learning the complete graph but may allow for methods that come with weaker assumptions and stronger guarantees. More concretely, for predictors $X_1, \ldots, X_d$, ICP

[*] Correspondence to: pbm@math.ku.dk

searches for subsets $S \subseteq \{1, \ldots, d\}$ that are invariant; a set $X_S$ of predictors is called invariant if the error when predicting $Y$ is identically distributed in all environments (or, slightly differently, if the environment is independent of the response $Y$, given $X_S$). ICP then outputs the intersection of all invariant predictor sets $S_{\text{ICP}} \coloneqq \cap_{S \text{ invariant}} S$. Peters et al. (2016) show that if invariance is tested empirically from data at level $\alpha$, the resulting intersection $\hat{S}_{\text{ICP}}$ is a subset of direct causes of $Y$ with probability at least $1 - \alpha$.

In many cases, however, the set learned by ICP forms a strict subset of all direct causes or may even be empty. This is because disjoint sets of predictors can be invariant, yielding an empty intersection, which may happen both for finite samples as well as in the population setting. In this work, we introduce and characterize minimally invariant sets of predictors, that is, invariant sets $S$ for which no proper subset is invariant. We propose to consider the union $S_{\text{IAS}}$ of all minimally invariant sets, where IAS stands for invariant ancestry search. We prove that $S_{\text{IAS}}$ is a subset of causal ancestors of $Y$, invariant, non-empty and contains $S_{\text{ICP}}$. In practice, we estimate minimally invariant sets using a test for invariance. If such a test has asymptotic power against some of the non-invariant sets (specified in Section 5.2), we show that, asymptotically, the probability of $\hat{S}_{\text{IAS}}$ being a subset of the ancestors is at least $1 - \alpha$. This puts stronger assumptions on the invariance test than ICP (which does not require any power) in return for discovering a larger set of causal ancestors. We prove that our approach retains the ancestral guarantee if we test minimal invariance only among subsets up to a certain size. This yields a computational speed-up compared to testing minimal invariance in all subsets, but comes at the cost of potentially finding fewer causal ancestors.

The remainder of this work is organized as follows. In Section 2 we review relevant background material and introduce the concept of minimal invariance in Section 3. Section 4 contains an oracle algorithm for finding minimally invariant sets (and a closed-form expression of $S_{\text{ICP}}$) and Section 5 presents theoretical guarantees when testing minimal invariance from data. In Section 6 we evaluate our method in several simulation studies as well as a real-world data set on gene perturbations. Code is provided at `https://github.com/PhillipMogensen/InvariantAncestrySearch`

# 2 Preliminaries

## 2.1 Structural Causal Models and Graphs

We consider a setting where data are sampled from a structural causal model (SCM) Pearl (2009); Bongers et al. (2021)

$$Z_j \coloneqq f_j(\text{PA}_j, \epsilon_j),$$

for some functions $f_j$, parent sets $\text{PA}_j$ and noise distributions $\epsilon_j$. Following Peters et al. (2016); Heinze-Deml et al. (2018), we consider an SCM over variables $Z \coloneqq (E, X, Y)$ where $E$ is an exogenous environment variable (i.e., $\text{PA}_E = \emptyset$), $Y$ is a response variable and $X = (X_1, \ldots, X_d)$ is a collection of predictors of $Y$. We denote by $\mathcal{P}$ the family of all possible distributions induced by an SCM over $(E, X, Y)$ of the above form.

For a collection of nodes $j \in [d] \coloneqq \{1, \ldots, d\}$ and their parent sets $\text{PA}_j$, we define a directed graph $\mathcal{G}$ with nodes $[d]$ and edges $j' \to j$ for all $j' \in \text{PA}_j$. We denote by $\text{CH}_j$, $\text{AN}_j$ and $\text{DE}_j$ the children, ancestors and descendants of a variable $j$, respectively, neither

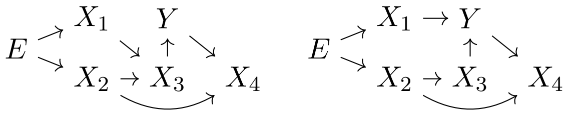

Figure 1: Two structures where $S_{\text{ICP}} \subsetneq \text{PA}_Y$. (*left*) $S_{\text{ICP}} = \emptyset$. (*right*) $S_{\text{ICP}} = \{1\}$. In both, our method outputs $S_{\text{IAS}} = \{1, 2, 3\}$.

containing $j$. A graph $\mathcal{G}$ is called a directed acyclic graph (DAG) if it does not contain any directed cycles. See Pearl (2009) for more details and the definition of $d$-separation.

Throughout the remainder of this work, we make the following assumption.

**Assumption 2.1.** Data are sampled from an SCM over nodes $(E, X, Y)$, such that the corresponding graph is a DAG and the environments are exogenous and do not act directly on $Y$. That is, $\text{PA}_E = \emptyset$ and $E \notin \text{PA}_Y$.

## 2.2 Invariant Causal Prediction

Invariant causal prediction (ICP), introduced by Peters et al. (2016), exploits the existence of heterogeneity in the data, here encoded by an environment variable $E$, to learn a subset of causal parents of a response variable $Y$. A subset of predictors $S \subseteq [d]$ is *invariant* if $Y \perp\!\!\!\perp E \mid S$, and we define $\mathcal{I} := \{S \subseteq [d] \mid S \text{ invariant}\}$ to be the set of all invariant sets. We denote the corresponding hypothesis that $S$ is invariant by

$$H_{0,S}^{\mathcal{I}} : \quad S \in \mathcal{I}.$$

Formally, $H_{0,S}^{\mathcal{I}}$ corresponds to a subset of distributions in $\mathcal{P}$, and we denote by $H_{A,S}^{\mathcal{I}} := \mathcal{P} \setminus H_{0,S}^{\mathcal{I}}$ the alternative hypothesis to $H_{0,S}^{\mathcal{I}}$. Peters et al. (2016) define the oracle output

$$S_{\text{ICP}} := \bigcap_{S : H_{0,S}^{\mathcal{I}} \text{ true}} S \tag{1}$$

(with $S_{\text{ICP}} = \emptyset$ if no sets are invariant) and prove $S_{\text{ICP}} \subseteq \text{PA}_Y$. If provided with a test for the hypotheses $H_{0,S}^{\mathcal{I}}$, we can test all sets $S \subseteq [d]$ for invariance and take the intersection over all accepted sets: $\hat{S}_{\text{ICP}} := \bigcap_{S : H_{0,S}^{\mathcal{I}} \text{ not rejected}} S$; If the invariance test has level $\alpha$, $\hat{S}_{\text{ICP}} \subseteq \text{PA}_Y$ with probability at least $1 - \alpha$.

However, even for the oracle output in Equation (1), there are many graphs for which $S_{\text{ICP}}$ is a strict subset of $\text{PA}_Y$. For example, in Figure 1 (left), since both $\{1, 2\}$ and $\{3\}$ are invariant, $S_{\text{ICP}} \subseteq \{1, 2\} \cap \{3\} = \emptyset$. This does not violate $S_{\text{ICP}} \subseteq \text{PA}_Y$, but is non-informative. Similarly, in Figure 1 (right), $S_{\text{ICP}} = \{1\}$, as all invariant sets contain $\{1\}$. Here, $S_{\text{ICP}}$ contains some information, but is not able to recover the full parental set. In neither of these two cases, $S_{\text{ICP}}$ is an invariant set. If the environments are such that each parent of $Y$ is either affected by the environment directly or is a parent of an affected node, then $S_{\text{ICP}} = \text{PA}_Y$ (Peters et al., 2016, proof of Theorem 3). The shortcomings of ICP thus relate to settings where the environments acts on too few variables or on uninformative ones.

For large $d$, it has been suggested to apply ICP to the variables in the *Markov boundary* Pearl (2014), $\text{MB}_Y = \text{PA}_Y \cup \text{CH}_Y \cup \text{PA}(\text{CH}_Y)$ (we denote the oracle output by $S_{\text{ICP}}^{\text{MB}}$). As

$\mathrm{PA}_Y \subseteq \mathrm{MB}_Y$, it still holds that $S_{\mathrm{ICP}}^{\mathrm{MB}}$ is a subset of the causal parents of the response.[1] However, the procedure must still be applied to $2^{|\mathrm{MB}_Y|}$ sets, which is only feasible if the Markov boundary is sufficiently small. In practice, the Markov boundary can, for example, be estimated using Lasso regression or gradient boosting techniques (Tibshirani, 1996; Meinshausen & Bühlmann, 2006; Friedman, 2001).

## 3 Minimal Invariance and Ancestry

We now introduce the concept of minimally invariant sets, which are invariant sets that do not have any invariant subsets. We propose to consider $S_{\mathrm{IAS}}$, the oracle outcome of invariant ancestry search, defined as the union of all minimally invariant sets. We will see that $S_{\mathrm{IAS}}$ is an invariant set, it consists only of ancestors of $Y$, and it contains $S_{\mathrm{ICP}}$ as a subset.

**Definition 3.1.** Let $S \subseteq [d]$. We say that $S$ is *minimally invariant* if and only if

$$S \in \mathcal{I} \text{ and } \forall S' \subsetneq S: \; S' \notin \mathcal{I};$$

that is, $S$ is invariant and no subset of $S$ is invariant. We define $\mathcal{MI} := \{S \mid S \text{ minimally invariant}\}$.

The concept of minimal invariance is closely related to the concept of minimal $d$-separators (Tian et al., 1998). This connection allows us to state several properties of minimal invariance. For example, an invariant set is minimally invariant if and only if it is non-invariant as soon as one of its elements is removed.

**Proposition 3.2.** *Let $S \subseteq [d]$. Then $S \in \mathcal{MI}$ if and only if $S \in \mathcal{I}$ and for all $j \in S$, it holds that $S \setminus \{j\} \notin \mathcal{I}$.*

The proof follows directly from (Tian et al., 1998, Corollary 2). We can therefore decide whether a given invariant set $S$ is minimally invariant using $\mathcal{O}(|S|)$ checks for invariance, rather than $\mathcal{O}(2^{|S|})$ (as suggested by Definition 3.1). We use this insight in Section 5.1, when we construct a statistical test for whether or not a set is minimally invariant.

To formally define the oracle outcome of IAS, we denote the hypothesis that a set $S$ is minimally invariant by

$$H_{0,S}^{\mathcal{MI}}: \quad S \in \mathcal{MI}$$

(and the alternative hypothesis, $S \notin \mathcal{MI}$, by $H_{A,S}^{\mathcal{MI}}$) and define the quantity of interest

$$S_{\mathrm{IAS}} := \bigcup_{S: H_{0,S}^{\mathcal{MI}} \text{ true}} S \tag{2}$$

with the convention that a union over the empty set is the empty set.

The following proposition states that $S_{\mathrm{IAS}}$ is a subset of the ancestors of the response $Y$. Similarly to $\mathrm{PA}_Y$, variables in $\mathrm{AN}_Y$ are causes of $Y$ in that for each ancestor there is a directed causal path to $Y$. Thus, generically, when intervened, these variables have a causal effect on the response.

[1] In fact, $S_{\mathrm{ICP}}^{\mathrm{MB}}$ is always at least as informative as ICP. E.g., there exist graphs in which $S_{\mathrm{ICP}} = \emptyset$ and $S_{\mathrm{ICP}}^{\mathrm{MB}} \neq \emptyset$, see Figure 1 (left). There are no possible structures for which $S_{\mathrm{ICP}}^{\mathrm{MB}} \subsetneq S_{\mathrm{ICP}}$, as both search for invariant sets over all sets of parents of $Y$.

**Proposition 3.3.** *It holds that* $S_{\text{IAS}} \subseteq \text{AN}_Y$.

The proof follows directly from (Tian et al., 1998, Theorem 2); see also (Acid & De Campos, 2013, Proposition 2). The setup in these papers is more general than what we consider here; we therefore provide direct proofs for Propositions 3.2 and 3.3 in Appendix A, which may provide further intuition for the results.

Finally, we show that the oracle output of IAS contains that of ICP and, contrary to ICP, it is always an invariant set.

**Proposition 3.4.** *The following two statements are true.*
*(i)* $S_{\text{IAS}} \in \mathcal{I}$.
*(ii)* $S_{\text{ICP}} \subseteq S_{\text{IAS}}$*, with equality if and only if* $S_{\text{ICP}} \in \mathcal{I}$.

# 4 Oracle Algorithms

When provided with an oracle that tells us whether a set is invariant or not, how can we efficiently compute $S_{\text{ICP}}$ and $S_{\text{IAS}}$? Here, we assume that the oracle is given by a DAG, see Assumption 2.1. A direct application of Equations (1) and (2) would require checking a number of sets that grows exponentially in the number of nodes. For $S_{\text{ICP}}$, we have the following characterization.[2]

**Proposition 4.1.** $S_{\text{ICP}}$ *is given by* $S_{\text{ICP}} = \text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E))$.

This allows us to efficiently read off $S_{\text{ICP}}$ from the DAG, (e.g., it can naively be done in $\mathcal{O}(\log(d+2)(d+2)^{2.373})$ time) For $S_{\text{IAS}}$, to the best of our knowledge, there is no closed form expression that has a similarly simple structure.

Instead, for IAS, we exploit the recent development of efficient algorithms for computing all minimal $d$-separators (for two given sets of nodes) in a given DAG (see, e.g., Tian et al., 1998; van der Zander et al., 2019). A set $S$ is called a *minimal d-separator* of $E$ and $Y$ if it $d$-separates $E$ and $Y$ given $S$ and no strict subset of $S$ satisfies this property. These algorithms are often motivated by determining minimal adjustment sets (e.g., Pearl, 2009) that can be used to compute the total causal effect between two nodes, for example. If the underlying distribution is Markov and faithful with respect to the DAG, then a set $S$ is minimally invariant if and only if it is a minimal $d$-separator for $E$ and $Y$. We can therefore use the same algorithms to find minimally invariant sets; van der Zander et al. (2019) provide an algorithm (based on work by Takata (2010)) for finding minimal $d$-separators with polynomial delay time. Applied to our case, this means that while there may be exponentially many minimally invariant sets,[3] when listing all such sets it takes at most polynomial time until the next set or the message that there are nor further sets is output. In practice, on random graphs, we found this to work well (see Section 6.1). But since $S_{\text{IAS}}$ is the union of all minimally invariant sets, even faster algorithms may be available; to the best of our knowledge, it is an open question whether finding $S_{\text{IAS}}$ is an NP-hard problem (see Appendix B for details).

[2] To the best of our knowledge, this characterization is novel.

[3] This is the case if there are $d/2$ (disjoint) directed paths between $E$ and $Y$, with each path containing two $X$-nodes, for example (e.g., van der Zander et al., 2019).

We provide a function for listing all minimally invariant sets in our python code; it uses an implementation of the above mentioned algorithm, provided in the R (R Core Team, 2021) package `dagitty` (Textor et al., 2016). In Section 6.1, we study the properties of the oracle set $S_{\mathrm{IAS}}$. When applied to 500 randomly sampled, dense graphs with $d = 15$ predictor nodes and five interventions, the `dagitty` implementation had a median speedup of a factor of roughly 17, compared to a brute-force search (over the ancestors of $Y$). The highest speedup achieved was by a factor of more than 1,900.

The above mentioned literature can be used only for oracle algorithms, where the graph is given. In the following sections, we discuss how to test the hypothesis of minimal invariance from data.

# 5 Invariant Ancestry Search

## 5.1 Testing a Single Set for Minimal Invariance

Usually, we neither observe a full SCM nor its graphical structure. Instead, we observe data from an SCM, which we want to use to decide whether a set is in $\mathcal{MI}$, such that we make the correct decision with high probability. We now show that a set $S$ can be tested for minimal invariance with asymptotic level and power if given a test for invariance that has asymptotic level and power.

Assume that $\mathcal{D}_n = (X_i, E_i, Y_i)_{i=1}^n$ are observations (which may or may not be independent) of $(X, E, Y)$ and let $\phi_n^{\mathcal{MI}}$ : powerset$([d]) \times \mathcal{D}_n \times (0, 1) \to \{0, 1\}$ be a decision rule that transforms $(S, \mathcal{D}_n, \alpha)$ into a decision $\phi_n^{\mathcal{MI}}(S, \mathcal{D}_n, \alpha)$ about whether the hypothesis $H_{0,S}^{\mathcal{MI}}$ should be rejected ($\phi_n^{\mathcal{MI}} = 1$) at significance threshold $\alpha$, or not ($\phi_n^{\mathcal{MI}} = 0$). To ease notation, we suppress the dependence on $\mathcal{D}_n$ and $\alpha$ when the statements are unambiguous.

A test $\psi_n$ for the hypothesis $H_0$ has pointwise asymptotic level if

$$\forall \alpha \in (0, 1) : \quad \sup_{\mathbb{P} \in H_0} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) \leq \alpha \tag{3}$$

and pointwise asymptotic power if

$$\forall \alpha \in (0, 1) : \quad \inf_{\mathbb{P} \in H_A} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) = 1. \tag{4}$$

If the limit and the supremum (resp. infimum) in Equation (3) (resp. Equation (4)) can be interchanged, we say that $\psi_n$ has uniform asymptotic level (resp. power).

Tests for invariance have been examined in the literature. Peters et al. (2016) propose two simple methods for testing for invariance in linear Gaussian SCMs when the environments are discrete, although the methods proposed extend directly to other regression scenarios. Pfister et al. (2019) propose resampling-based tests for sequential data from linear Gaussian SCMs. Furthermore, any valid test for conditional independence between $Y$ and $E$ given a set of predictors $S$ can be used to test for invariance. Although for continuous $X$, there exists no general conditional independence test that has both level and non-trivial power (Shah & Peters, 2020), it is possible to impose restrictions on the data-generating process that ensure the existence of non-trivial tests (e.g., Fukumizu et al., 2008; Zhang et al., 2011; Berrett et al., 2020; Shah & Peters, 2020; Thams et al., 2021). Heinze-Deml et al. (2018) provide an overview and a comparison of several conditional independence tests in the context of invariance.

To test whether a set $S \subseteq [d]$ is minimally invariant, we define the decision rule

$$\phi_n^{\mathcal{MI}}(S) := \begin{cases} 1 & \text{if } \phi_n(S) = 1 \text{ or } \min_{j \in S} \phi_n(S \setminus \{j\}) = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\phi_n^{\mathcal{MI}}(\emptyset) := \phi_n(\emptyset)$. Here, $\phi_n$ is a test for the hypothesis $H_{0,S}^{\mathcal{I}}$, e.g., one of the tests mentioned above. This decision rule rejects $H_{0,S}^{\mathcal{MI}}$ either if $H_{0,S}^{\mathcal{I}}$ is rejected by $\phi_n$ or if there exists $j \in S$ such that $H_{0,S\setminus\{j\}}^{\mathcal{I}}$ is not rejected. If $\phi_n$ has pointwise (resp. uniform) asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has pointwise (resp. uniform) asymptotic level and pointwise (resp. uniform) asymptotic power of at least $1 - \alpha$.

**Theorem 5.1.** *Let $\phi_n^{\mathcal{MI}}$ be defined as in Equation* (5) *and let $S \subseteq [d]$. Assume that the decision rule $\phi_n$ has pointwise asymptotic level and power for $S$ and for all $S \setminus \{j\}, j \in S$. Then, $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level and pointwise asymptotic power of at least $1 - \alpha$, i.e.,*

$$\inf_{\mathbb{P} \in H_{A,S}^{\mathcal{MI}}} \lim_{n \to \infty} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) \geq 1 - \alpha.$$

*If $\phi_n$ has uniform asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has uniform asymptotic level and uniform asymptotic power of at least $1 - \alpha$.*

Due to Proposition 3.3, a test for $H_{0,S}^{\mathcal{MI}}$ is implicitly a test for $S \subseteq \mathrm{AN}_Y$, and can thus be used to infer whether intervening on $S$ will have a potential causal effect on $Y$. However, rejecting $H_{0,S}^{\mathcal{MI}}$ is not evidence for $S \not\subseteq \mathrm{AN}$; it is evidence for $S \notin \mathcal{MI}$.

## 5.2 Learning $S_{\mathrm{IAS}}$ from Data

We now consider the task of estimating the set $S_{\mathrm{IAS}}$ from data. If we are given a test for invariance that has asymptotic level and power and if we correct for multiple testing appropriately, we can estimate $S_{\mathrm{IAS}}$ by $\hat{S}_{\mathrm{IAS}}$, which, asymptotically, is a subset of $\mathrm{AN}_Y$ with large probability.

**Theorem 5.2.** *Assume that the decision rule $\phi_n$ has pointwise asymptotic level for all minimally invariant sets and pointwise asymptotic power for all $S \subseteq [d]$ such that $S$ is not a superset of a minimally invariant set. Define $C := 2^d$ and let $\widehat{\mathcal{I}} := \{S \subseteq [d] \mid \phi_n(S, \alpha C^{-1}) = 0)\}$ be the set of all sets for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}} := \left\{S \in \widehat{\mathcal{I}} \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}\right\}$ and $\hat{S}_{\mathrm{IAS}} := \bigcup_{S \in \widehat{\mathcal{MI}}} S$. It then holds that*

$$\begin{aligned} \lim_{n \to \infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) &\geq \lim_{n \to \infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}) \\ &\geq 1 - \alpha. \end{aligned}$$

A generic algorithm for implementing $\hat{S}_{\mathrm{IAS}}$ is given in Appendix D.

*Remark* 5.3. Consider a decision rule $\phi_n$ that just (correctly) rejects empty set (e.g., because the $p$-value is just below the threshold $\alpha$), indicating that the effect of the environments is weak. It is likely that there are other sets $S' \notin \mathcal{I}$, which the test may not have sufficient power against and are (falsely) accepted as invariant. If one of such sets contains non-ancestors of $Y$, this yields a violation of $\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$. To guard against this, testing $S = \emptyset$ can be done at a lower significance level, $\alpha_0 < \alpha$. This modified IAS approach is conservative and may return $\hat{S}_{\mathrm{IAS}} = \emptyset$ if the environments do not have a strong impact on $Y$, but it retains the guarantee $\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq 1 - \alpha$ of Theorem 5.2.

The multiple testing correction performed in Theorem 5.2 is strictly conservative, because we only need to correct for the number of minimally invariant sets, and there do not exist $2^d$ minimally invariant sets[4]. Thus, in practice, a milder multiple error correction can be used. We propose to use a correction factor of $C = 3^{\lceil d/3 \rceil}$, which is likely to still be conservative in many graphs (see Appendix C for details)

Alternatively, as shown in the following section, we can restrict the search for minimally invariant sets to a predetermined size. This requires milder correction factors and comes with computational benefits.

### 5.3 Invariant Ancestry Search in Large Systems

We now develop a variation of Theorem 5.2, which allows us to search for ancestors of $Y$ in large graphs, at the cost of only identifying minimally invariant sets up to some a priori determined size.

Similarly to ICP (see Section 2.2), one could restrict IAS to the variables in $\mathrm{MB}_Y$ but the output may be smaller than $S_{\mathrm{IAS}}$; in particular, there are only non-parental ancestors in $\mathrm{MB}_Y$ if these are parents to both a parent a child of $Y$ (For instance, in the graph $E \to X_1 \to \ldots \to X_d \to Y$, $S_{\mathrm{IAS}} = \{1, \ldots, d\}$ but restricting IAS to $\mathrm{MB}_Y$ would yield the set $\{d\}$.) Thus, we do not expect such an approach to be particularly fruitful in learning ancestors.

Here, we propose an alternative approach and define

$$S^m_{\mathrm{IAS}} \coloneqq \bigcup_{S : S \in \mathcal{MI} \text{ and } |S| \leq m} S \tag{6}$$

as the union of minimally invariant sets that are no larger than $m \leq d$. For computing $S^m_{\mathrm{IAS}}$, one only needs to check invariance of the $\sum_{i=0}^m \binom{d}{i}$ sets that are no larger than $m$. $S^m_{\mathrm{IAS}}$ itself, however, can be larger than $m$: in the graph above Equation (6), $S^1_{\mathrm{IAS}} = \{1, \ldots, d\}$. The following proposition characterizes properties of $S^m_{\mathrm{IAS}}$.

**Proposition 5.4.** *Let $m < d$ and let $m_{\min}$ and $m_{\max}$ be the size of a smallest and a largest minimally invariant set, respectively. The following statements are true:*

*(i) $S^m_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$.*
*(ii) If $m \geq m_{\max}$, then $S^m_{\mathrm{IAS}} = S_{\mathrm{IAS}}$.*
*(iii) If $m \geq m_{\min}$, then $S^m_{\mathrm{IAS}} \in \mathcal{I}$.*
*(iv) If $m \geq m_{\min}$, then $S_{\mathrm{ICP}} \subseteq S^m_{\mathrm{IAS}}$ with equality if and only if $S_{\mathrm{ICP}} \in \mathcal{I}$.*

[4] Under Assumption 2.1, the maximum number of minimally invariant sets in a graph with $d$ predictors can be estimated using simulation, as described in Appendix C.

If $m < m_{\min}$ and $S_{\mathrm{ICP}} \neq \emptyset$, then $S_{\mathrm{ICP}} \subseteq S^m_{\mathrm{IAS}}$ does not hold. However, we show in Section 6.1 using simulations that $S^m_{\mathrm{IAS}}$ is larger than $S_{\mathrm{ICP}}$ in many sparse graphs, even for $m = 1$, when few nodes are intervened on.

In addition to the computational speedup offered by considering $S^m_{\mathrm{IAS}}$ instead of $S_{\mathrm{IAS}}$, the set $S_{\mathrm{IAS}}$ can be estimated from data using a smaller correction factor than the one employed in Theorem 5.2. This has the benefit that in practice, smaller sample sizes may be needed to detect non-invariance.

**Theorem 5.5.** *Let $m \leq d$ and define $C(m) := \sum_{i=0}^{m} \binom{d}{i}$. Assume that the decision rule $\phi_n$ has pointwise asymptotic level for all minimally invariant sets of size at most $m$ and pointwise power for all sets of size at most $m$ that are not supersets of a minimally invariant set. Let $\widehat{\mathcal{I}}^m := \left\{S \subseteq [d] \mid \phi_n(S, \alpha C(m)^{-1}) = 0 \text{ and } |S| \leq m\right\}$, be the set of all sets of size at most $m$ for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}}^m := \left\{S \in \widehat{\mathcal{I}}^m \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}^m\right\}$ and $\hat{S}^m_{\mathrm{IAS}} := \bigcup_{S \in \widehat{\mathcal{MI}}^m} S$. It then holds that*

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq \lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\mathrm{IAS}} = S^m_{\mathrm{IAS}}) \\ \geq 1 - \alpha.\end{aligned}$$

The method proposed in Theorem 5.5 outputs a non-empty set if there exists a non-empty set of size at most $m$, for which the hypothesis of invariance cannot be rejected. In a sparse graph, it is likely that many small sets are minimally invariant, whereas if the graph is dense, it may be that all invariant sets are larger than $m$, such that $S^m_{\mathrm{IAS}} = \emptyset$. In dense graphs however, many other approaches may fail too; for example, it is also likely that the size of the Markov boundary is so large that applying ICP on $\mathrm{MB}_Y$ is not feasible.

# 6 Experiments

We apply the methods developed in this paper in a population-case experiment using oracle knowledge (Section 6.1), a synthetic experiment using finite sample tests (Section 6.2), and a real-world data set from a gene perturbation experiment (Section 6.3).

## 6.1 Oracle IAS in Random Graphs

For the oracle setting, we know that $S_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$ (Proposition 3.3) and $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ (Proposition 3.4). We first verify that the inclusion $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ is often strict in low-dimensional settings when there are few interventions. Second, we show that the set $S^m_{\mathrm{IAS}}$ is often strictly larger than the set $S^{\mathrm{MB}}_{\mathrm{ICP}}$ in large, sparse graphs with few interventions.

In principle, for a given number of covariates, one can enumerate all DAGs and, for each DAG, compare $S_{\mathrm{ICP}}$ and $S_{\mathrm{IAS}}$. However, because the space of DAGs grows super-exponentially in the number of nodes (Chickering, 2002), this is infeasible. Instead, we sample graphs from the space of all DAGs that satisfy Assumption 2.1 and $Y \in \mathrm{DE}_E$ (see Appendix E.1 for details).

In the low-dimensional setting ($d \leq 20$), we compute $S_{\mathrm{ICP}}$ and $S_{\mathrm{IAS}}$, whereas in the larger graphs ($d \geq 100$), we compute $S^{\mathrm{MB}}_{\mathrm{ICP}}$ and the reduced set $S^m_{\mathrm{IAS}}$ for $m \in \{1, 2\}$ when $d = 100$ and for $m = 1$ when $d = 1{,}000$. Because there is no guarantee that IAS outputs a

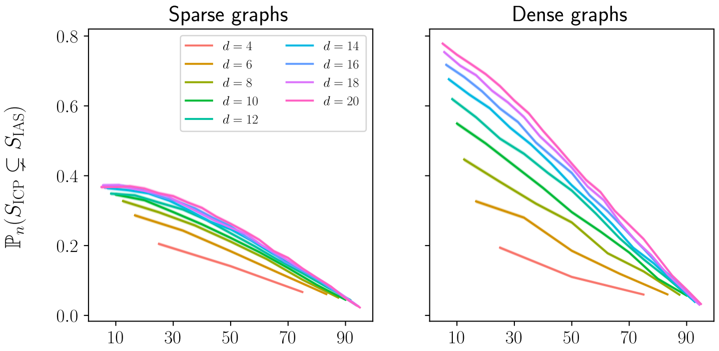

Figure 2: Low-dimensional oracle experiment, see Section 6.1. In all cases, as predicted by theory, $S_{\text{ICP}}$ is contained in $S_{\text{IAS}}$. For many graphs, $S_{\text{IAS}}$ is strictly larger than $S_{\text{ICP}}$. On average, this effect is more expressed when there are fewer intervened nodes. $\mathbb{P}_n$ refers to the distribution used to sample graphs and every point in the figure is based on 50,000 independently sampled graphs; $d$ denotes the number of covariates $X$. Empirical confidence bands are plotted around each line, but are very narrow.

superset of ICP when searching only up to sets of some size lower than $d$, we compare the size of the sets output by either method. For the low-dimensional setting, we consider both sparse and dense graphs, but for larger dimensions, we only consider sparse graphs. In the sparse setting, the DAGs are constructed such that there is an expected number of $d+1$ edges between the $d+1$ nodes $X$ and $Y$; in the dense setting, the expected number of edges equals $0.75 \cdot d(d+1)/2$.

The results of the simulations are displayed in Figures 2 and 3. In the low-dimensional setting, $S_{\text{IAS}}$ is a strict superset of $S_{\text{ICP}}$ for many graphs. This effect is the more pronounced, the larger the $d$ and the fewer nodes are intervened on, see Figure 2. In fact, when there are interventions on all predictors, we know that $S_{\text{IAS}} = S_{\text{ICP}} = \text{PA}_Y$ (Peters et al., 2016, Theorem 2), and thus the probability that $S_{\text{ICP}} \subsetneq S_{\text{IAS}}$ is exactly zero. For the larger graphs, we find that the set $S^m_{\text{IAS}}$ is, on average, larger than $S^{\text{MB}}_{\text{ICP}}$, in particular when $d = 1{,}000$ or when $m = 2$, see Figure 3. In the setting with $d = 100$ and $m = 1$, the two sets are roughly the same size, when 10% of the predictors are intervened on. The set $S^{\text{MB}}_{\text{ICP}}$ becomes larger than $S^1_{\text{IAS}}$ after roughly 15% of the predictors nodes are intervened on (not shown). For both $d = 100$ and $d = 1{,}000$, the average size of the Markov boundary of $Y$ was found to be approximately 3.5.

## 6.2 Simulated Linear Gaussian SCMs

In this experiment, we show through simulation that IAS finds more ancestors than ICP in a finite sample setting when applied to linear Gaussian SCMs. To compare the outputs of IAS and ICP, we use the *Jaccard similarity* between $\hat{S}_{\text{IAS}}$ ($\hat{S}^1_{\text{IAS}}$ when $d$ is large) and $\text{AN}_Y$,

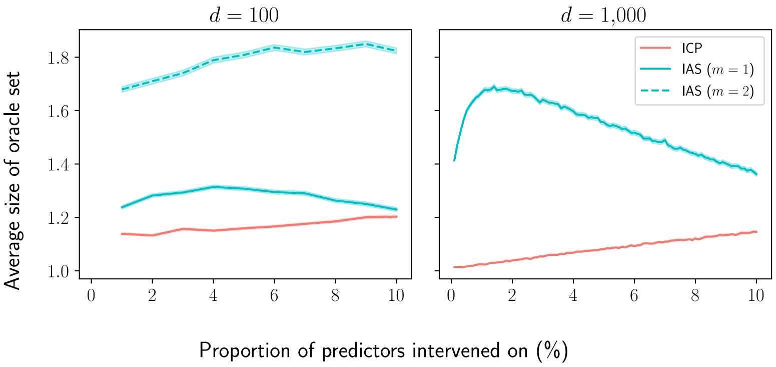

Figure 3: High-dimensional oracle experiment with sparse graphs, see Section 6.1. The average size of the set $S^m_{\mathrm{IAS}}$ is larger than the average size of the set $S^{\mathrm{MB}}_{\mathrm{ICP}}$, both when using IAS to search for sets up to sizes $m = 1$ and $m = 2$. Except for the choice of $d$, the setup is the same as in Figure 2.

and between $\hat{S}_{\mathrm{ICP}}$ ($\hat{S}^{\hat{\mathrm{MB}}}_{\mathrm{ICP}}$ when $d$ is large[5]) and $\mathrm{AN}_Y$.[6]

We sample data from sparse linear Gaussian models with i.i.d. noise terms in two scenarios, $d = 6$ and $d = 100$. In both cases, coefficients for the linear assignments are drawn randomly. We consider two environments; one observational and one interventional; in the interventional environment, we apply do-interventions of strength one to children of $E$, i.e., we fix the value of a child of $E$ to be one. We standardize the data along the causal order, to prevent variance accumulation along the causal order (Reisach et al., 2021). Throughout the section, we consider a significance level of $\alpha = 5\%$. For a detailed description of the simulations, see Appendix E.2.

To test for invariance, we employ the test used in Peters et al. (2016): We calculate a $p$-value for the hypothesis of invariance of $S$ by first linearly regressing $Y$ onto $X_S$ (ignoring $E$), and second testing whether the mean and variance of the prediction residuals is equal across environments. For details, see Peters et al. (2016, Section 3.2.1). Schultheiss et al. (2021) also consider the task of estimating ancestors but since their method is uninformative for Gaussian data and does not consider environments, it is not directly applicable here.

In Theorem 5.2, we assume asymptotic power of our invariance test. When $d = 6$, we test hypotheses with a correction factor $C = 3^{\lceil 9/3 \rceil} = 9$, as suggested in Appendix C, in an attempt to reduce false positive findings. In Appendix E.3, we repeat the experiment of this section with $C = 2^6$ and find almost identical results. We hypothesize, that the effects of a reduced $C$ is more pronounced at larger $d$. When $d = 100$, we test hypotheses with correction the factor $C(1)$ of Theorem 5.5. In both cases, we test the hypothesis of invariance of the empty set at level $\alpha_0 = 10^{-6}$ (cf. Remark 5.3). In Appendix E.4, we investigate the effects on the quantities $\mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$ and $\mathbb{P}(\hat{S}^1_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$ when varying $\alpha_0$, confirming that choosing $\alpha_0$ too high can lead to a reduced probability of $\hat{S}_{\mathrm{IAS}}$ being a subset of ancestors.

[5] $\hat{\mathrm{MB}}$ is a Lasso regression estimate of $\mathrm{MB}_Y$ containing at most 10 variables

[6] The Jaccard similarity between two sets $A$ and $B$ is defined as $J(A, B) := |A \cap B|/|A \cup B|$, with $J(\emptyset, \emptyset) = 0$. The Jaccard similarity equals one if the two sets are equal, zero if they are disjoint and takes a value in $(0, 1)$ otherwise.

In Figure 4 the results of the simulations are displayed. In SCMs where the oracle versions $S_{\text{IAS}}$ and $S_{\text{ICP}}$ are not equal, $\hat{S}_{\text{IAS}}$ achieved, on average, a higher Jaccard similarity to $\text{AN}_Y$ than $\hat{S}_{\text{ICP}}$. This effect is less pronounced when $d = 100$. We believe that the difference in Jaccard similarities is more pronounced when using larger values of $m$. When $S_{\text{IAS}} = S_{\text{ICP}}$, the two procedures achieve roughly the same Jaccard similarities to $\text{AN}_Y$, as expected. When the number of observations is one hundred, IAS generally fails to find any ancestors and outputs the empty set (see Figure 7), indicating that the we do not have power to reject the empty set when there are few observations. This is partly by design; we test the empty set for invariance at reduced level $\alpha_0$ in order to protect against making false positive findings when the environment has a weak effect on $Y$. However, even without testing the empty set at a reduced level, IAS has to correct for making multiple comparisons, contrary to ICP, thus lowering the marginal significance level each set is tested at. When computing the jaccard similarities with either $\alpha_0 = \alpha$ or $\alpha_0 = 10^{-12}$, the results were roughly identical (not shown). We repeated the experiments with $d = 6$ with a weaker influence of the environment (do-interventions of strength 0.5 instead of 1) and find comparable results, with slightly less power in that the empty set is found more often, see Appendix E.5.

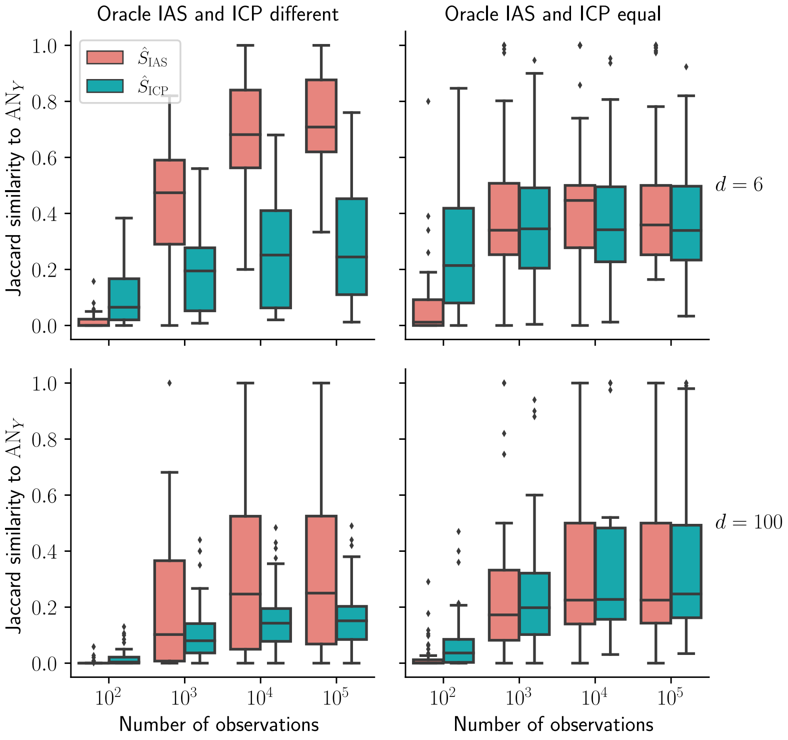

Figure 4: Comparison between the finite sample output of IAS and ICP and $\text{AN}_Y$ on simulated data, see Section 6.2. The plots show the Jaccard similarities between $\text{AN}_Y$ and either $\hat{S}_{\text{IAS}}$ ($\hat{S}^1_{\text{IAS}}$ when $d = 100$) in red or $\hat{S}_{\text{ICP}}$ ($\hat{S}^{\text{MB}}_{\text{ICP}}$ when $d = 100$) in blue and $\text{AN}_Y$. When $S_{\text{ICP}} \neq S_{\text{IAS}}$ (left column), $\hat{S}_{\text{IAS}}$ is more similar to $\text{AN}_Y$ than $\hat{S}_{\text{ICP}}$. The procedures are roughly equally similar to $\text{AN}_Y$ when $S_{\text{ICP}} = S_{\text{IAS}}$ (right column). Graphs represented in each boxplot: 42 (top left), 58 (top right), 40 (bottom left) and 60 (bottom right).

### 6.3 IAS in High Dimensional Genetic Data

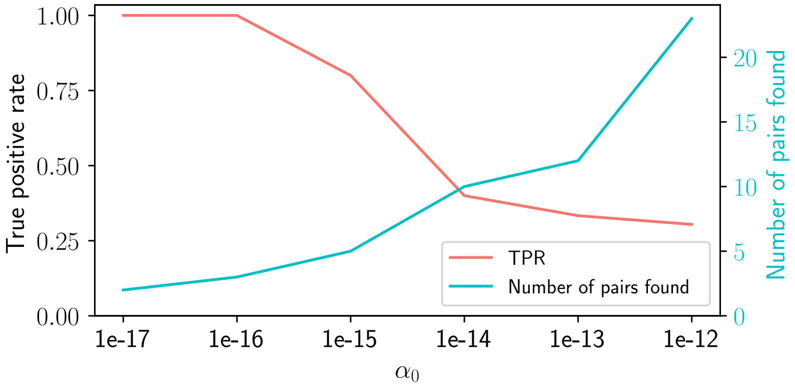

Figure 5: True positive rates and number of gene pairs found in the experiment in Section 6.3. On the $x$-axis, we change $\alpha_0$, the threshold for invariance of the empty set. When $\alpha_0$ is small, we only search for pairs if the environment has a very significant effect on $Y$. For smaller $\alpha_0$, fewer pairs are found to be invariant (blue line), but those found, are more likely to be true positives (red line). This supports the claim that the lower $\alpha_0$ is, the more conservative our approach is.

We evaluate our approach in a data set on gene expression in yeast Kemmeren et al. (2014). The data contain full-genome mRNA expressions of $d = 6{,}170$ genes and consists of $n_{\text{obs}} = 160$ unperturbed observations ($E = 1$) and $n_{\text{int}} = 1{,}479$ intervened-upon observations ($E = 2$); each of the latter observations correspond to the deletion of a single (known) gene. For each response gene $\texttt{gene}_Y \in [d]$, we apply the procedure from Section 5.3 with $m = 1$ to search for ancestors.

We first test for invariance of the empty set, i.e., whether the distribution of $\texttt{gene}_Y$ differs between the observational and interventional environment. We test this at a conservative level $\alpha_0 = 10^{-12}$ in order to protect against a high false positive rate (see Remark 5.3). For 3,631 out of 6,170 response genes, the empty set is invariant, and we disregard them as response genes.

For each response gene, for which the empty set is not invariant, we apply our procedure. More specifically, when testing whether $\texttt{gene}_X$ is an ancestor of $\texttt{gene}_Y$, we exclude any observation in which either $\texttt{gene}_X$ or $\texttt{gene}_Y$ was intervened on. We then test whether the empty set is still rejected, at level $\alpha_0 = 10^{-12}$, and whether $\texttt{gene}_X$ is invariant at level $\alpha = 0.25$. Since a set $\{\texttt{gene}_X\}$ is deemed minimally invariant if the $p$-value exceeds $\alpha$, setting $\alpha$ large is conservative for the task of finding ancestors. Indeed, when estimating $\hat{S}^m_{\text{IAS}}$, one can test the sets of size $m$ at a higher level $\alpha_1 > \alpha$. This is conservative, because falsely rejecting a minimally invariant set of size $m$ does not break the inclusion $\hat{S}^m_{\text{IAS}} \subseteq \text{AN}_Y$. However, if one has little power against the non-invariant sets of size $m$, testing at level $\alpha_1$ can protect against false positives.[7]

[7] Only sets of size exactly $m$ can be tested at level $\alpha_1$; the remaining hypotheses should still be corrected by $C(m)$ (or by the hypothesized number of minimally invariant sets).

We use the held-out data point, where $\texttt{gene}_X$ is intervened on, to determine as ground truth, whether $\texttt{gene}_X$ is indeed an ancestor of $\texttt{gene}_Y$. We define $\texttt{gene}_X$ as a true ancestor of $\texttt{gene}_Y$ if the value of $\texttt{gene}_Y$ when $\texttt{gene}_X$ is intervened on, lies in the $q_{TP} = 1\%$ tails of the observational distribution of $\texttt{gene}_Y$.

We find 23 invariant pairs $(\texttt{gene}_X, \texttt{gene}_Y)$; of these, 7 are true positives. In comparison, Peters et al. (2016) applies ICP to the same data, and with the same definition of true positives. They predict 8 pairs, of which 6 are true positives. This difference is in coherence with the motivation put forward in Section 5.2: Our approach predicts many more ancestral pairs (8 for ICP compared to 23 for IAS). Since ICP does not depend on power of the test, they have a lower false positive rate (25% for ICP compared to 69.6% for IAS).

In Figure 5, we explore how changing $\alpha_0$ and $q_{TP}$ impacts the true positive rate. Reducing $\alpha_0$ increases the true positive rate, but lowers the number of gene pairs found (see Figure 5). This is because a lower $\alpha_0$ makes it more difficult to detect non-invariance of the empty set, making the procedure more conservative (with respect to finding ancestors); see Remark 5.3. For example, when $\alpha_0 \leq 10^{-15}$, the true positive rate is above 0.8; however, 5 or fewer pairs are found. When searching for ancestors, the effect of intervening may be reduced by noise from intermediary variables, so $q_{TB} = 1\%$ might be too strict; in Appendix E.6, we analyze the impact of increasing $q_{TB}$.

# 7 Conclusion and Future Work

Invariant Ancestry Search (IAS) provides a framework for searching for causal ancestors of a response variable $Y$ through finding minimally invariant sets of predictors by exploiting the existence of exogenous heterogeneity. The set $S_{\text{IAS}}$ is a subset of the ancestors of $Y$, a superset of $S_{\text{ICP}}$ and, contrary to $S_{\text{ICP}}$, invariant itself. Furthermore, the hierarchical structure of minimally invariant sets allows IAS to search for causal ancestors only among subsets up to a predetermined size. This avoids exponential runtime and allows us to apply the algorithm to large systems. We have shown that, asymptotically, $S_{\text{IAS}}$ can be identified from data with high probability if we are provided with a test for invariance that has asymptotic level and power. We have validated our procedure both on simulated and real data. Our proposed framework would benefit from further research in the maximal number of minimally invariant sets among graphs of a fixed size, as this would provide larger finite sample power for identifying ancestors. Further it is of interest to establish finite sample guarantees or convergence rates for IAS, possibly by imposing additional assumptions on the class of SCMs. Finally, even though current implementations are fast, it is an open theoretical question whether computing $S_{\text{IAS}}$ in the oracle setting of Section 4 is NP-hard, see Appendix B.

# Acknowledgements

NT and JP were supported by a research grant (18968) from VILLUM FONDEN.

# References

Acid, S. and De Campos, L. M. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, 49(5):2885–2915, 2021.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2008.

Gamella, J. L. and Heinze-Deml, C. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Gaspers, S. and Mackenzie, S. On the number of minimal separators in graphs. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pp. 116–121. Springer, 2015.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, 2009.

Jordan, M. I. Artificial intelligence — the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1), 2019.

Kemmeren, P., Sameith, K., Van De Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

Lauritzen, S. L. *Graphical models.* Clarendon Press, 1996.

Martinet, G., Strzalkowski, A., and Engelhardt, B. E. Variance minimization in the Wasserstein space for invariant causal prediction. *arXiv preprint arXiv:2110.07064*, 2021.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* The Morgan Kaufmann series in representation and learning. Morgan Kaufmann, 2014.

Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Pfister, N., Bühlmann, P., and Peters, J. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Schultheiss, C., Bühlmann, P., and Yuan, M. Higher-order least squares: assessing partial goodness of fit of linear regression. *arXiv preprint arXiv:2109.14544*, 2021.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (10), 2006.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search.* MIT press, 2000.

Takata, K. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158:1660–1667, 2010.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

Thams, N., Saengkyongam, S., Pfister, N., and Peters, J. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

Tian, J., Paz, A., and Pearl, J. Finding minimal d-separators. Technical report, University of California, Los Angeles, 1998.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

van der Zander, B., Liśkiewicz, M., and Textor, J. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 804–813, 2011.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

# A Proofs

## A.1 A direct Proof of Proposition 3.2

*Proof.* Assume for a contradiction, that an invariant set $S_0 \subsetneq S$ exists. By assumption, $|S \setminus S_0| > 1$, because otherwise $S_0$ would be non-invariant.

We can choose $S_1 \subseteq S$ and $k_0, k_1, \dots, k_l \in S$ with $l \geq 1$ such that for all $i = 1, \dots, l$ : $k_i \notin \mathrm{DE}_{k_0}$ and

$$\begin{array}{rll} & S_0 \cup S_1 \cup \{k_0, \dots, k_l\} = S & \in \mathcal{I} \\ \text{for } 0 \leq i < l: & S_0 \cup S_1 \cup \{k_0, \dots, k_i\} & \notin \mathcal{I} \\ & S_0 \cup S_1 & \in \mathcal{I}. \end{array}$$

This can be done by iteratively removing elements from $S \setminus S_0$, removing first the earliest elements in the causal order. The first invariant set reached in this process is then $S_0 \cup S_1$.

Since $S_0 \cup S_1 \cup \{k_0\}$ is non-invariant, there exist a path $\pi$ between $E$ and $Y$ that is open given $S_0 \cup S_1 \cup \{k_0\}$ but blocked given $S_0 \cup S_1$. Since removing $k_0$ blocks $\pi$, $k_0$ must a collider or a descendant of a collider $c$ on $\pi$:

$$\begin{array}{c} \overbrace{\underbrace{E \rightarrow \cdots \rightarrow c}_{\pi_E} \underbrace{\leftarrow \cdots - Y}_{\pi_Y}}^{\pi} \\ \downarrow \\ \vdots \\ \downarrow \\ k_0 \end{array}$$

Since $\pi$ is open given $S_0 \cup S_1$, the two sub-paths $\pi_E$ and $\pi_Y$ are open given $S_0 \cup S_1$.

Additionally, since $S_0 \cup S_1 \cup \{k_1, \dots, k_l\} = S \setminus \{k_0\}$ is non-invariant, there exists a path $\tau$ between $E$ and $Y$ that is unblocked given $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$ and blocked given $S_0 \cup S_1 \cup \{k_1, \dots, k_l\} \cup \{k_0\}$. It follows that $k_0$ lies on $\tau$ (otherwise $\tau$ cannot be blocked by adding $k_0$) and $k_0$ has at least one outgoing edge. Assume, without loss of generality that there is an outgoing edge towards $Y$. Since $\tau$ is open given $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$, so is $\tau_Y$.

$$\overbrace{E \rightarrow \cdots - \underbrace{k_0 \rightarrow \cdots - Y}_{\tau_Y}}^{\tau}$$

If there are no colliders on $\tau_Y$, then $\tau_Y$ is also open given $S_0 \cup S_1$. But then the path the path $E \overset{\pi_E}{\cdots\!\rightarrow} c \rightarrow \cdots \rightarrow k_0 \overset{\tau_Y}{\rightarrow \cdots}$ is also open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.

If there are colliders on $\tau_Y$, let $m$ be the collider closest to $k_0$, meaning that $m \in \mathrm{DE}_{k_0}$. Since $\tau_Y$ is open given $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$, it means that either $m$ or a descendant of $m$ is in $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$. Since $\{k_1, \dots, k_l\} \cap \mathrm{DE}_{k_0} = \emptyset$, there exist $v \in (S_0 \cup S_1) \cap (\{m\} \cup \mathrm{DE}_m)$. But then $v \in \mathrm{DE}_{k_0} \cap (S_0 \cup S_1)$, meaning that $\pi$ is open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.

$$\begin{tikzcd}
\underbrace{E \to \cdots \to c}_{\pi_E} \arrow[d] \\
\vdots \arrow[d] \\
\overbrace{k_0 \to \cdots - Y}^{\tau_Y}
\end{tikzcd}$$

We could assume that $\tau_Y$ had an outgoing edge from $k_0$ without loss of generality, because if there was instead an outgoing edge from $k_0$ on $\tau_E$, the above argument would work with $\pi_Y$ and $\tau_E$ instead. This concludes the proof. □

## A.2 A direct proof of Proposition 3.3

*Proof.* We show that if $S \in \mathcal{I}$ is not a subset of $\mathrm{AN}_Y$, then $S^* := S \cap \mathrm{AN}_Y \in \mathcal{I}$, meaning that $S \notin \mathcal{MI}$.

Let $p$ be any path between $E$ and $Y$. Since $S \in \mathcal{I}$, $p$ is blocked given $S$. Assume for contradiction that $p$ is open given $S^*$.

Since $p$ is open given the smaller set $S^* \subsetneq S$, all colliders on $p$ is in $S^*$ or has a descendant in $S^*$; therefore all colliders are ancestors of $Y$. Further, there exist at least one non-collider on $p$ that is in $S \setminus \mathrm{AN}_Y$.

However, since all colliders of $p$ are ancestors of $Y$, all nodes on $p$ are ancestors of $Y$. This contradicts the existence of a non-ancestral non-collider and concludes the proof of part 1. □

## A.3 Proof of Proposition 3.4

*Proof.* First, we show that $S_{\mathrm{IAS}} \in \mathcal{I}$. If $S_{\mathrm{IAS}}$ is the union of a single minimally invariant set, it trivially holds that $S_{\mathrm{IAS}} \in \mathcal{I}$. Now assume that $S_{\mathrm{IAS}}$ is the union of at least two minimally invariant sets, $S_{\mathrm{IAS}} = S_1 \cup \ldots \cup S_n$, $n \geq 2$, and assume for a contradiction that there exists a path $\pi$ between $E$ and $Y$ that is unblocked given $S_{\mathrm{IAS}}$.

Since $\pi$ is blocked by a strict subset of $S_{\mathrm{IAS}}$, it follows that $\pi$ has at least one collider; further every collider of $\pi$ either is in $S_{\mathrm{IAS}}$ or has a descendant in $S_{\mathrm{IAS}}$, and hence every collider of $\pi$ is an ancestor of $Y$.

Therefore, $\pi$ has the shape displayed below,

$$\overbrace{E \to \cdots \to c_1}^{\pi_1} \overbrace{\leftarrow \cdots \to}^{\pi_2} c_2 \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_3, \ldots, \pi_k} c_k \overbrace{\leftarrow \cdots \to Y}^{\pi_{k+1}}$$

where the paths $\pi_1, \ldots, \pi_{k+1}$, $k \geq 1$, do not have any colliders and are unblocked given $S_{\mathrm{IAS}}$. In particular, $\pi_1, \ldots, \pi_{k+1}$ are unblocked given $S_1$.

The path $\pi_{k+1}$ must have a final edge pointing to $Y$, because otherwise it would be a directed path from $Y$ to $c_k$, which contradicts acyclicity since $c_k$ is an ancestor of $Y$.

As $c_1$ is an ancestor of $Y$, there exists a directed path, say $\rho_1$, from $c_1$ to $Y$. Since $\pi_1$ is open given $S_1$ and since $S_1$ is invariant, it follows that $\rho_1$ must be blocked by $S_1$ (otherwise the path $E \overset{\pi_1}{\to} c_1 \overset{\rho_1}{\to} Y$ would be open). For this reason, $S_1$ contains a descendant of the collider $c_1$.

Similarly, if $\rho_2$ is a directed path from $c_2$ to $Y$, then $S_1$ blocks $\rho_2$, because otherwise the path $E \overset{\pi_1}{\to} c_1 \leftarrow \overset{\pi_2}{\cdots} \to c_2 \overset{\rho_2}{\to} Y$ would be open. Again, for this reason, $S_1$ contains a descendant of $c_2$.

Iterating this argument, it follows that $S_1$ contains a descendant of every collider on $\pi$, and since $\pi_1, \dots, \pi_{k+1}$ are unblocked by $S_1$, $\pi$ is open given $S_1$. This contradicts invariance of $S_1$ and proves that $S_{\mathrm{IAS}} \in \mathcal{I}$.

We now show that $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ with equality if and only if $S_{\mathrm{ICP}} \in \mathcal{I}$. First, $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ because $S_{\mathrm{IAS}}$ is a union of the minimally invariant sets, and $S_{\mathrm{ICP}}$ is the intersection over all invariant sets. We now show the equivalence statement.

Assume first that $S_{\mathrm{ICP}} \in \mathcal{I}$. As $S_{\mathrm{ICP}}$ is the intersection of all invariant sets, $S_{\mathrm{ICP}} \in \mathcal{I}$ implies that there exists exactly one invariant set, that is contained in all other invariant sets. By definition, this means that there is only one minimally invariant set, and that this set is exactly $S_{\mathrm{ICP}}$. Thus, $S_{\mathrm{IAS}} = S_{\mathrm{ICP}}$.

Conversely assume that $S_{\mathrm{ICP}} \notin \mathcal{I}$. By construction, $S_{\mathrm{ICP}}$ is contained in any invariant set, in particular in the minimally invariant sets. However, since $S_{\mathrm{ICP}}$ is not invariant itself, this containment is strict, and it follows that $S_{\mathrm{ICP}} \subsetneq S_{\mathrm{IAS}}$.

□

## A.4 Proof of Proposition 4.1

*Proof.* First we show $\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq S_{\mathrm{ICP}}$. If $j \in \mathrm{PA}_Y \cap \mathrm{CH}_E$, any invariant set contains $j$, because otherwise the path $E \to j \to Y$ is open. Similarly, if $j \in \mathrm{PA}_Y \cap \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)$, any invariant set contains $j$ (there exists a node $j'$ such that $E \to j' \to \cdots \to Y$ and $E \to j' \leftarrow j \to Y$, and any invariant set $S$ must contain $j'$ or one of its descendants; thus, it must also contain $j$ to ensure that the path $E \to j' \leftarrow j \to Y$ is blocked by $S$.) It follows that for all invariant $S$,

$$\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq S,$$

such that

$$\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq \bigcap_{S \text{ invariant}} S.$$

To show $S_{\mathrm{ICP}} \subseteq \mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E))$, take any $j \notin \mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E))$. We argue, that an invariant set $\bar{S}$ not containing $j$ exists, such that $j \notin S_{\mathrm{ICP}} = \bigcap_{S \text{ invariant}} S$. If $j \notin \mathrm{PA}_Y$, let $\bar{S} = \mathrm{PA}_Y$, which is invariant. If $j \in \mathrm{PA}_Y$, define

$$\bar{S} = (\mathrm{PA}_Y \setminus \{j\}) \cup \mathrm{PA}_j \cup (\mathrm{CH}_j \cap \mathrm{AN}_Y) \cup \mathrm{PA}(\mathrm{CH}_j \cap \mathrm{AN}_Y).$$

Because $j \notin \mathrm{CH}_E$ and $j \notin \cap \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)$, we have $E \notin \bar{S}$. Also observe that $\bar{S} \subseteq \mathrm{AN}_Y$. We show that any path between $E$ and $Y$ is blocked by $\bar{S}$, by considering all possible paths:

$\cdots \mathbf{j'} \to \mathbf{Y}$ **for** $\mathbf{j'} \neq \mathbf{j}$**:** Blocked because $j' \in \mathrm{PA}_Y \setminus \{j\}$.

$\cdots \mathbf{v} \to \mathbf{j} \to \mathbf{Y}$**:** Blocked because $v \in \mathrm{PA}_j \subseteq \bar{S}$ and $E \notin \mathrm{PA}_j$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \mathrm{PA}_j(\mathrm{CH}_j \cap \mathrm{AN}_Y)$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $\bar{S} \subseteq \mathrm{AN}_Y$, and since $c \notin \mathrm{AN}_Y$, $\bar{S} \cap \mathrm{DE}_c = \emptyset$ and the path is blocked given $\bar{S}$ because of the collider $c$.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \mathrm{AN}_c$ and $c \in \mathrm{AN}_Y$, so $v \in \mathrm{CH}_j \cap \mathrm{AN}_Y \subseteq \bar{S}$.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$**:** Same reason as for the case '$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ and $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$'.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{Y}$ Since $\bar{S} \subseteq \mathrm{AN}_Y$, we must have $\bar{S} \cap \mathrm{DE}_c = \emptyset$ (otherwise this would create a directed cycle from $Y \to \cdots \to Y$). Hence the path is blocked given $\bar{S}$ because of the collider $c$.

Since there are no open paths from $E$ to $Y$ given $\bar{S}$, $\bar{S}$ is invariant, and $S_{\mathrm{ICP}} \subseteq \bar{S}$. Since $j \notin \bar{S}$, it follows that $j \notin S_{\mathrm{ICP}}$. This concludes the proof. □

### A.5 Proof of Theorem 5.1

*Proof.* Consider first the case where all marginal tests have pointwise asymptotic power and pointwise asymptotic level.

**Pointwise asymptotic level:** Let $\mathbb{P}_0 \in H_{0,S}^{\mathcal{MI}}$. By the assumption of pointwise asymptotic level, there exists a non-negative sequence $(\epsilon_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \epsilon_n = 0$ and $\mathbb{P}_0(\phi_n(S) = 1) \leq \alpha + \epsilon_n$. Then

$$\begin{aligned}
\mathbb{P}_0(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j\in S} (\phi_n(S \setminus \{j\}) = 0)\right) \\
&\leq \mathbb{P}_0(\phi_n(S) = 1) + \sum_{j\in S} \mathbb{P}_0(\phi_n(S \setminus \{j\}) = 0) \\
&\leq \alpha + \epsilon_n + \sum_{j\in S} \mathbb{P}_0(\phi_n(S \setminus \{j\}) = 0) \\
&\to \alpha + 0 \qquad \text{as } n \to \infty \\
&= \alpha.
\end{aligned}$$

The convergence step follows from

$$H_{0,S}^{\mathcal{MI}} = H_{0,S}^{\mathcal{I}} \cap \bigcap_{j\in S} H_{A,S\setminus\{j\}}^{\mathcal{I}}$$

and from the assumption of pointwise asymptotic level and power. As $\mathbb{P}_0 \in H_{0,S}^{\mathcal{MI}}$ was arbitrary, this shows that $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level.

**Pointwise asymptotic power:** To show that the decision rule has pointwise asymptotic power, consider any $\mathbb{P}_A \in H_{A,S}^{\mathcal{MI}}$. We have that

$$H_{A,S}^{\mathcal{MI}} = H_{A,S}^{\mathcal{I}} \cup \left(H_{0,S}^{\mathcal{I}} \cap \bigcup_{j\in S} H_{0,S\setminus\{j\}}^{\mathcal{I}}\right). \tag{7}$$

As the two sets $H^{\mathcal{I}}_{A,S}$ and

$$H^{\mathcal{I}}_{0,S} \cap \bigcup_{j \in S} H^{\mathcal{I}}_{0,S\setminus\{j\}}$$

are disjoint, we can consider them one at a time. Consider first the case $\mathbb{P}_A \in H^{\mathcal{I}}_{A,S}$. This means that $S$ is not invariant and thus

$$\begin{aligned}
\mathbb{P}_A(\phi^{\mathcal{MI}}_n(S) = 1) &= \mathbb{P}_A\left((\phi_n(S) = 1) \cup \bigcup_{j \in S}(\phi_n(S \setminus \{j\}, \alpha) = 0)\right) \\
&\geq \mathbb{P}_A(\phi_n(S) = 1) \\
&\to 1 \qquad \text{as } n \to \infty
\end{aligned}$$

by the assumption of pointwise asymptotic power.

Next, assume that there exists $j' \in S$ such that $\mathbb{P}_A \in (H^{\mathcal{I}}_{0,S} \cap H^{\mathcal{I}}_{0,S\setminus\{j'\}})$. Then,

$$\begin{aligned}
\mathbb{P}_A(\phi^{\mathcal{MI}}_n(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j \in S}(\phi_n(S \setminus \{j\}) = 0)\right) \\
&\geq \mathbb{P}_A(\phi_n(S \setminus \{j'\}) = 0) \\
&\geq 1 - \alpha - \epsilon_n \\
&\to 1 - \alpha \qquad \text{as } n \to \infty.
\end{aligned}$$

Thus, for arbitrary $\mathbb{P}_A \in H^{\mathcal{MI}}_{A,S}$ we have shown that $\mathbb{P}_A(\phi^{\mathcal{MI}}_n(S) = 1) \geq 1 - \alpha$ in the limit. This shows that $\phi^{\mathcal{MI}}_n$ has pointwise asymptotic power of at least $1 - \alpha$. This concludes the argument for pointwise asymptotic power.

Next, consider the case that the marginal tests have uniform asymptotic power and uniform asymptotic level. The calculations for showing that $\phi^{\mathcal{MI}}_n$ has uniform asymptotic level and uniform asymptotic power of at least $1 - \alpha$ are almost identical to the pointwise calculations.

**Uniform asymptotic level:** By the assumption of uniform asymptotic level, there exists a non-negative sequence $\epsilon_n$ such that $\lim_{n\to\infty} \epsilon_n = 0$ and $\sup_{\mathbb{P} \in H^{\mathcal{I}}_{0,S}} \mathbb{P}(\phi_n(S) = 1) \leq \alpha + \epsilon_n$.

Then,

$$\begin{aligned}
\sup_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) &= \sup_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\\
&\leq \sup_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}} \left(\mathbb{P}(\phi_n(S)=1)+\sum_{j\in S}\mathbb{P}(\phi_n(S\setminus\{j\})=0)\right)\\
&\leq \sup_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}} \mathbb{P}(\phi_n(S)=1)+\sum_{j\in S}\sup_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}}\mathbb{P}(\phi_n(S\setminus\{j\})=0)\\
&\leq \alpha+\epsilon_n+\sum_{j\in S}\left(1-\inf_{\mathbb{P}\in H_{0,S}^{\mathcal{MI}}}\mathbb{P}(\phi_n(S\setminus\{j\})=1)\right)\\
&\to \alpha+0+\sum_{j\in S}(1-1) \qquad \text{as } n\to\infty\\
&= \alpha.
\end{aligned}$$

**Uniform asymptotic power:** From (7), it follows that

$$\inf_{\mathbb{P}\in H_{A,S}^{\mathcal{MI}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) = \min\left\{\inf_{\mathbb{P}\in H_{A,S}^{\mathcal{I}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1), \inf_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap\bigcup_{j\in S}H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1)\right\}.$$

We consider the two inner terms in the above separately. First,

$$\begin{aligned}
\inf_{\mathbb{P}\in H_{A,S}^{\mathcal{I}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) &= \inf_{\mathbb{P}\in H_{A,S}^{\mathcal{I}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\\
&\geq \inf_{\mathbb{P}\in H_{A,S}^{\mathcal{I}}} \mathbb{P}(\phi_n(S)=1)\\
&\to 1 \qquad \text{as } n\to\infty.
\end{aligned}$$

Next,

$$\begin{aligned}
\inf_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap\bigcup_{j\in S}H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) &= \inf_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap\bigcup_{j\in S}H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\\
&= \min_{j\in S}\left\{\inf_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\right\}\\
&\geq \min_{j\in S}\left\{\inf_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}(\phi_n(S\setminus\{j\})=0)\right\}\\
&= \min_{j\in S}\left\{1-\sup_{\mathbb{P}\in H_{0,S}^{\mathcal{I}}\cap H_{0,S\setminus\{j\}}^{\mathcal{I}}} \mathbb{P}(\phi_n(S\setminus\{j\})=1)\right\}\\
&\geq 1-\alpha-\epsilon_n\\
&\to 1-\alpha \qquad \text{as } n\to\infty.
\end{aligned}$$

This shows that $\phi_n^{\mathcal{MI}}$ has uniform asymptotic power of at least $1-\alpha$, which completes the proof. □

### A.6 Proof of Theorem 5.2

*Proof.* We have that

$$\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y) \geq \lim_{n\to\infty} \mathbb{P}(\hat{S}_{\text{IAS}} = S_{\text{IAS}})$$

as $S_{\text{IAS}} \subseteq \text{AN}_Y$ by Proposition 3.4. Furthermore, we have

$$\mathbb{P}(\hat{S}_{\text{IAS}} = S_{\text{IAS}}) \geq \mathbb{P}(\widehat{\mathcal{MI}} = \mathcal{MI}).$$

Let $A := \{S \mid S \notin \mathcal{I}\} \setminus \{S \mid \exists S' \subsetneq S \text{ s.t. } S' \in \mathcal{MI}\}$ be those non-invariant sets that do not contain a minimally invariant set and observe that

$$(\widehat{\mathcal{MI}} = \mathcal{MI}) \supseteq \bigcap_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S\in A} (\phi_n(S, \alpha C^{-1}) = 1). \tag{8}$$

To see why this is true, note that to correctly recover $\mathcal{MI}$, we need to 1) accept the hypothesis of minimal invariance for all minimally invariant sets and 2) reject the hypothesis of invariance for all non-invariant sets that are not supersets of a minimally invariant set (any superset of a set for which the hypothesis of minimal invariance is not rejected is removed in the computation of $\widehat{\mathcal{MI}}$). Then,

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{MI}} = \mathcal{MI}) &\geq \mathbb{P}\left(\bigcap_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S\in A} (\phi_n(S, \alpha C^{-1}) = 1)\right) \\
&\geq 1 - \mathbb{P}\left(\bigcup_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 1)\right) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S\in\mathcal{MI}} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 1) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S\in\mathcal{MI}} (\alpha C^{-1} + \epsilon_{n,S}) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - |\mathcal{MI}|\alpha C^{-1} + \sum_{S\in\mathcal{MI}} \epsilon_{n,S} - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \alpha + \sum_{S\in\mathcal{MI}} \epsilon_{n,S} - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\to 1 - \alpha \quad \text{as } n \to \infty,
\end{aligned}$$

where $(\epsilon_{n,S})_{n\in\mathbb{N}, S\in\mathcal{MI}}$ are non-negative sequences that converge to zero and the last step follows from the assumption of asymptotic power. The sequences $(\epsilon_{n,S})_{n\in\mathbb{N}, S\in\mathcal{MI}}$ exist by the assumption of asymptotic level. □

### A.7 Proof of Proposition 5.4

*Proof.* We prove the statements one by one.

**(i)** Since $S^m_{\text{IAS}}$ is the union over some of the minimally invariant sets, $S^m_{\text{IAS}} \subseteq S_{\text{IAS}}$. Then the statement follows from Proposition 3.3.

**(ii)** If $m \geq m_{\max}$, all $S \in \mathcal{MI}$ satisfy the requirement $|S| \leq m$.

**(iii)** If $m \geq m_{\min}$, then $S^m_{\text{IAS}}$ contains at least one minimally invariant set. The statement then follows from the first part of the proof of Proposition 3.4 given in Appendix A.3.

**(iv)** $S^m_{\text{IAS}}$ contains at least one minimally invariant set and, by (iii), it is itself invariant. Thus, if $S_{\text{ICP}} \notin \mathcal{I}$, then $S_{\text{ICP}} \subsetneq S^m_{\text{IAS}}$. If $S_{\text{ICP}} \in \mathcal{I}$, then there exists only one minimally invariant set, which is $S_{\text{ICP}}$ (see proof of Proposition 3.4), and we have $S_{\text{ICP}} = S^m_{\text{IAS}}$. This concludes the proof. □

### A.8 Proof of Theorem 5.5

*Proof.* The proof is identical to the proof of Theorem 5.2, when changing the correction factor $2^{-d}$ to $C(m)^{-1}$, adding superscript $m$'s to the quantities $\widehat{\mathcal{MI}}$, $\hat{S}_{\text{IAS}}$ and $S_{\text{IAS}}$, and adding the condition $|S| \leq m$ to all unions, intersections and sums. □

## B Oracle Algorithms for Learning $S_{\text{IAS}}$

In this section, we review some of the existing literature on minimal $d$-separators, which can be exploited to give an algorithmic approach for finding $S_{\text{IAS}}$ from a DAG. We first introduce the concept of $M$-minimal separation with respect to a constraining set $I$.

**Definition B.1** (van der Zander et al. (2019), Section 2.2)**.** Let $I \subseteq [d]$, $K \subseteq [d]$, and $S \subseteq [d]$. We say that $S$ is a *$K$-minimal separator* of $E$ and $Y$ with respect to a constraining set $I$ if all of the following are true:
(i) $I \subseteq S$.
(ii) $S \in \mathcal{I}$.
(iii) There does not exists $S' \in \mathcal{I}$ such that $K \subseteq S' \subsetneq S$.
We denote by $M_{K,I}$ the set of all $K$-minimal separating sets with respect to constraining set $I$.

(In this work, $S \in \mathcal{I}$ means $E \perp\!\!\!\perp Y \mid S$, but it can stand for other separation statements, too.) The definition of a $K$-minimal separator coincides with the definition of a minimally invariant set if both $K$ and the constraining set $I$ are equal to the empty set. An $\emptyset$-minimal separator with respect to constraining set $I$ is called a *strongly-minimal separator with respect to constraining set I*.

We can now represent (2) using this notation. $M_{\emptyset,\emptyset}$ contains the minimally invariant sets and thus

$$S_{\text{IAS}} := \bigcup_{S \in M_{\emptyset,\emptyset}} S.$$

Listing the set $M_{I,I}$ of all $I$-minimal separators with respect to the constraining set $I$ (for any $I$) can be done in polynomial delay time $\mathcal{O}(d^3)$ (van der Zander et al., 2019; Takata, 2010), where delay here means that finding the next element of $M_{I,I}$ (or announcing that there is no further element) has cubic complexity. This is the algorithm we exploit, as described in the main part of the paper.

Furthermore, we have

$$i \in S_{\text{IAS}} \quad \Leftrightarrow \quad M_{\emptyset,\{i\}} \neq \emptyset.$$

This is because $i \in S_{\text{IAS}}$ if and only if there is a minimally invariant set that contains $i$, which is the case if and only if there exist a strongly minimal separating set with respect to constraining set $\{i\}$. Thus, we can construct $S_{\text{IAS}}$ by checking, for each $i$, whether there is an element in $M_{\emptyset,\{i\}}$. Finding a strongly-minimal separator with respect to constraining set $I$, i.e., finding an element in $M_{\emptyset,I}$, is NP-hard if the set $I$ is allowed to grow (van der Zander et al., 2019). To the best of our knowledge, however, it is unknown whether finding an element in $M_{\emptyset,\{i\}}$, for a singleton $\{i\}$ is NP-hard.

## C The Maximum Number of Minimally Invariant Sets

If one does not have a priori knowledge about the graph of the system being analyzed, one can still apply Theorem 5.2 with a correction factor $2^d$, as this ensures (with high probability) that no minimally invariant sets are falsely rejected. However, we know that the correction factor is strictly conservative, as there cannot exists $2^d$ minimally invariant sets in a graph. Thus, correcting for $2^d$ tests, controls the familywise error rate (FWER) among minimally invariant sets, but increases the risk of falsely accepting a non-invariant set relatively more than what is necessary to control the FWER. Here, we discuss the maximum number of minimally invariant sets that can exist in a graph with $d$ predictor nodes and how a priori knowledge about the sparsity of the graph and the number of interventions can be leveraged to estimate a less strict correction that still controls the FWER.

As minimally invariant sets only contain ancestors of $Y$ (see Proposition 3.3), we only need to consider graphs where $Y$ comes last in a causal ordering. Since $d$-separation is equivalent to undirected separation in the moralized ancestral graph (Lauritzen, 1996), finding the largest number of minimally invariant sets is equivalent to finding the maximum number of minimal separators in an undirected graph with $d+2$ nodes. It is an open question how many minimal separators exists in a graph with $d+2$ nodes, but it is known that a lower bound for the maximum number of minimal separators is in $\Omega(3^{(d+2)/3})$ (Gaspers & Mackenzie, 2015). We therefore propose using a correction factor of $C = 3^{\lceil d/3 \rceil}$ when estimating the set $\hat{S}_{\text{IAS}}$ from Theorem 5.2 if one does not have a priori knowledge of the number of minimally invariant sets in the DAG of the SCM being analyzed. This is a heuristic choice and is not conservative for all graphs.

Theorem 5.2 assumes asymptotic power of the invariance test, but as we can only have a finite amount of data, we will usually not have full power against all non-invariant sets that are not supersets of a minimally invariant set. Therefore, choosing a correction factor that is potentially too low represents a trade-off between error types: if we correct too little, we stand the risk of falsely rejecting a minimally invariant set but not rejecting a superset of it, whereas when correcting too harshly, there is a risk of failing to reject non-invariant sets due to a lack of power.

If one has a priori knowledge of the sparsity or the number of interventions, these can be leveraged to estimate the maximum number of minimally invariant sets using simulation, by the following procedure:

1. For $b = 1, \ldots, B$:
   (a) Sample a DAG with $d$ predictor nodes, $N_{\text{interventions}} \sim \mathbb{P}_N$ interventions and $p \sim \mathbb{P}_p$ probability of an edge being present in the graph over $(X, Y)$, such that $Y$ is last in a causal ordering. The measures $\mathbb{P}_N$ and $\mathbb{P}_p$ are distributions representing a priori knowledge. For instance, in a controlled experiment, the researcher may have chosen the number $N_0$ of interventions. Then, $\mathbb{P}_N$ is a degenerate distribution with $\mathbb{P}_N(N_0) = 1$.
   (b) Compute the set of all minimally invariant sets, e.g., using the `adjustmentSets` algorithm from `dagitty` (Textor et al., 2016).
   (c) Return the number of minimally invariant sets.
2. Return the largest number of minimally sets found in the $B$ repetitions above.

Instead of performing $B$ steps, one can continually update the largest number of minimally invariant sets found so far and end the procedure if the maximum has not updated in a predetermined number of steps, for example.

# D A Finite Sample Algorithm for Computing $\hat{S}_{\text{IAS}}$

In this section, we provide an algorithm for computing the sets $\hat{S}_{\text{IAS}}$ and $\hat{S}^m_{\text{IAS}}$ presented in Theorems 5.2 and 5.5. The algorithm finds minimally invariant sets by searching for invariant sets among sets of increasing size, starting from the empty set. This is done, because the first (correctly) accepted invariant is a minimally invariant set. Furthermore, any set that is a superset of an accepted invariant set, does not need to be tested (as this set cannot be minimal). Tests for invariance can be computationally expensive if one has large amounts of data. Therefore, skipping unnecessary tests offers a significant speedup. In the extreme case, where all singletons are found to be invariant, the algorithm completes in $d + 1$ steps, compared to $\sum_{i=0}^{m} \binom{d}{i}$ steps ($2^d$ if $m = d$). This is implemented in lines 8-10 of Algorithm 1.

# E Additional Experiment Details

## E.1 Simulation Details for Section 6.1

We sample graphs that satisfy Assumption 2.1 with the additional requirement that $Y \in \text{DE}_Y$ by the following procedure:

1. Sample a DAG $\mathcal{G}$ for the graph of $(X, Y)$ with $d + 1$ nodes, for $d \in \{4, 6, \ldots, 20\} \cup \{100, 1{,}000\}$, and choose $Y$ to be a node (chosen uniformly at random) that is not a root node.

2. Add a root node $E$ to $\mathcal{G}$ with $N_{\text{interventions}}$ children that are not $Y$. When $d \leq 20$, $N_{\text{interventions}} \in \{1, \ldots, d\}$ and when $d \geq 100$, $N_{\text{interventions}} \in \{1, \ldots, 0.1 \times d\}$ (i.e., we consider interventions on up to ten percent of the predictor nodes).

**Algorithm 1** An algorithm for computing $\hat{S}_{\mathrm{IAS}}$ from data

```
input  A decision rule $\phi_n$ for invariance, significance thresholds $\alpha_0, \alpha$, max size of sets to
    test $m$ (potentially $m = d$) and data
output  The set $\hat{S}_{\mathrm{IAS}}$
Initialize $\widehat{\mathcal{MI}}$ as an empty list.
$PS \leftarrow \{S \subseteq [d] \mid |S| \leq m\}$
if $\phi_n(\emptyset, \alpha_0) = 0$ then
    End the procedure and return $\hat{S}_{\mathrm{IAS}} = \emptyset$
end if
Sort $PS$ in increasing order according the set sizes
for $S \in PS$ do
    if $S \supsetneq S'$ for any $S' \in \widehat{\mathcal{MI}}$ then
        Skip the test of $S$ and go to next iteration of the loop
    else
        Add $S$ to $\widehat{\mathcal{MI}}$ if $\phi_n(S, \alpha) = 0$, else continue
    end if
    if The union of $\widehat{\mathcal{MI}}$ contains all nodes then
        Break the loop
    end if
end for
Return $\hat{S}_{\mathrm{IAS}}$ as the union of all sets in $\widehat{\mathcal{MI}}$
```

3. Repeat the first two steps if $Y \notin \mathrm{DE}_E$.

## E.2 Simulation Details for Section 6.2

We simulate data for the experiment in Section 6.2 (and the additional plots in Appendix E.4) by the following procedure:

1. Sample data from a single graph by the following procedure:

   (a) Sample a random graph $\mathcal{G}$ of size $d + 1$ and sample $Y$ (chosen uniformly at random) as any node that is not a root node in this graph.

   (b) Sample coefficients, $\beta_{i \to j}$, for all edges $(i \to j)$ in $\mathcal{G}$ from $U((-2, 0.5) \cup (0.5, 2))$ independently.

   (c) Add a node $E$ with no incoming edges and $N_{\text{interventions}}$ children, none of which are $Y$. When $d = 6$, we set $N_{\text{interventions}} = 1$ and when $d = 100$, we sample $N_{\text{interventions}}$ uniformly from $\{1, \ldots, 10\}$.

   (d) If $Y$ is not a descendant of $E$, repeat steps (a), (b) and (c) until a graph where $Y \in \mathrm{DE}_E$ is obtained.

   (e) For $n \in \{10^2, 10^3, 10^4, 10^5\}$:

      i. Draw 50 datasets of size $n$ from an SCM with graph $\mathcal{G}$ and coefficients $\beta_{i \to j}$ and with i.i.d. $N(0, 1)$ noise innovations. The environment variable, $E$, is sampled independently from a Bernoulli distribution with probability parameter $p =$

0.5, corresponding to (roughly) half the data being observational and half the data interventional. The data are generated by looping through a causal ordering of $(X, Y)$, starting at the bottom, and standardizing a node by its own empirical standard deviation before generating children of that node; that is, a node $X_j$ is first generated from $\mathrm{PA}_j$ and then standardized before generating any node in $\mathrm{CH}_j$. If $X_j$ is intervened on, we standardize it prior to the intervention.

ii. For each sampled dataset, apply IAS and ICP. Record the Jaccard similarities between IAS and $\mathrm{AN}_Y$ and between ICP and $\mathrm{AN}_Y$, and record whether or not is was a subset of $\mathrm{AN}_Y$ and whether it was empty.

iii. Estimate the quantity plotted (average Jaccard similarity in Figure 4 or probability of $\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$ or $\hat{S}_{\mathrm{IAS}} = \emptyset$ in Figure 7) from the 50 simulated datasets.

(f) Return the estimated quantities from the previous step.

2. Repeat the above 100 times and save the results in a data-frame.

### E.3 Analysis of the Choice of $C$ in Section 6.2

We have repeated the simulation with $d = 6$ from Section 6.2 but with a correction factor of $C = 2^6$, as suggested by Theorem 5.2 instead of the heuristic correction factor of $C = 9$ suggested in Appendix C. Figure 6 shows the results. We see that the results are almost identical to those presented in Figure 4. Thus, in the scenario considered here, there is no change in the performance of $\hat{S}_{\mathrm{IAS}}$ (as measured by Jaccard similarity) between using a correction factor of $C = 2^6$ and a correction factor of $C = 3^{\lceil 9/3 \rceil} = 9$. In larger graphs, it is likely that there is a more pronounced difference. E.g., at $d = 10$, the strictly conservative correction factor suggested by Theorem 5.2 is $2^{10} = 1024$, whereas the correction factor suggested in Appendix C is only $3^{\lceil 10/3 \rceil} = 3^4 = 81$, and at $d = 20$ the two are $2^{20} = 1{,}048{,}576$ and $3^{\lceil 20/3 \rceil} = 3^7 = 2187$.

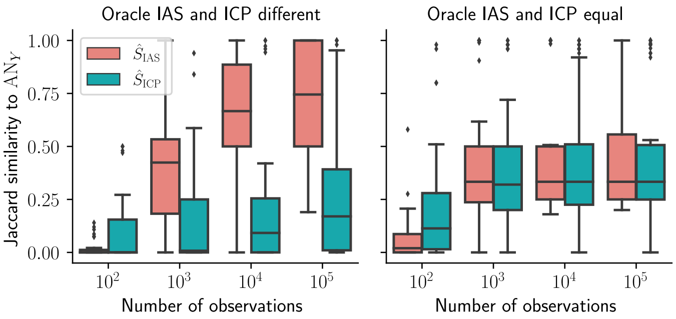

Figure 6: The same figure as in Figure 4, but with a correction factor of $C = 2^6 = 64$ instead of $C = 3^{\lceil 6/3 \rceil} = 9$. Only $d = 6$ shown here, as the correction factor for $d = 100$ is unchanged. Here, the guarantees of Theorem 5.2 are not violated by a potentially too small correction factor, and the results are near identical to those given in Figure 4 using a milder correction factor.

### E.4 Analysis of the Choice of $\alpha_0$ in Section 6.2

Here, we investigate the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}^1_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}^1_{\text{IAS}} = \emptyset)$ using the same simulation setup as described in Section 6.2. Furthermore, we also ran the simulations for values $\alpha_0 = \alpha$ (testing all hypotheses at the same level), $\alpha_0 = 10^{-6}$ (conservative, see Remark 5.3) as in Section 6.2 and $\alpha_0 = 10^{-12}$ (very conservative). The results for $\alpha = 10^{-6}$ (shown in Figure 7) were recorded in the same simulations that produced the output for Figure 4. For $\alpha_0 \in \{\alpha, 10^{-12}\}$ (shown in Figure 8 and Figure 9, respectively) we only simulated up to 10,000 observations, to keep computation time low.

Generally, we find that the probability of IAS being a subset of the ancestors seems to generally hold well and even more so with large sample sizes. (see Figures 7 to 9), in line with Theorem 5.2. When given 100,000 observations, the probability of IAS being a subset of ancestors is roughly equal to one for almost all SCMs, although there are a few SCMs, where IAS is never a subset of the ancestors (see Figure 7). For $\alpha_0 = 10^{-6}$, the median probability of IAS containing only ancestors is one in all cases, except for $d = 100$ with 1,000 observations – here, the median probability is 87%.

In general, varying $\alpha_0$ has the effect hypothesized in Remark 5.3: lowering $\alpha_0$ increases the probability that IAS contains only ancestors, but at the cost of increasing the probability that it is empty (see Figures 7 to 9). For instance, the median probability of IAS being a subset of ancestors when $\alpha_0 = 10^{-12}$ is one for all sample sizes, but the output is always empty when there are 100 observations and empty roughly half the time even at 1,000 observations when $d = 100$ (see Figure 9). In contrast, not testing the empty set at a reduced level, means that the output of IAS is rarely empty, but the probability of IAS containing only ancestors decreases. Still, even with $\alpha_0 = \alpha$, the median probability of IAS containing only ancestors was never lower than 80% (see Figure 8). Thus, choosing $\alpha_0$ means choosing a trade-off between finding more ancestor-candidates, versus more of them being false positives.

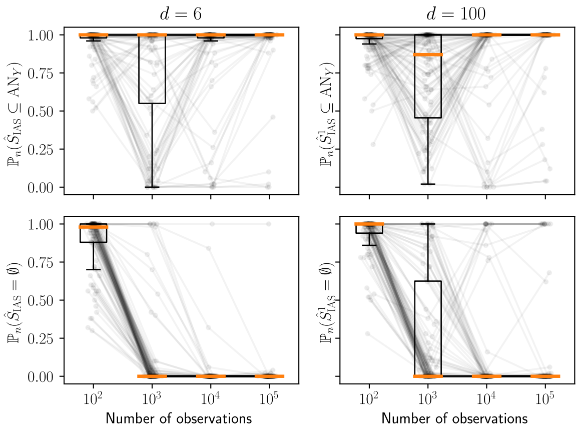

Figure 7: The empirical probabilities of recovering a subset of $\text{AN}_Y$ (top row) and recovering an empty set (bottom row), when testing the empty set for invariance at level $\alpha_0 = 10^{-6}$. Generally, our methods seem to hold level well, especially when sample sizes are large. When the sample size is small, the output is often the empty set. When $d = 6$, we estimate $\hat{S}_{\text{IAS}}$ (left column) and when $d = 100$, we estimate $\hat{S}^1_{\text{IAS}}$ (right column). The results here are from the simulations that also produced Figure 4. Medians are displayed as orange lines through each boxplot. Each point represents the probability that the output set is ancestral (resp. empty) for a randomly selected SCM, as estimated by repeatedly sampling data from the same SCM for every $n \in \{10^2, 10^3, 10^4, 10^5\}$. Observations from the same SCM are connected by a line. Each figure contains data from 250 randomly drawn SCMs. Points have been perturbed slightly along the $x$-axis to improve readability.

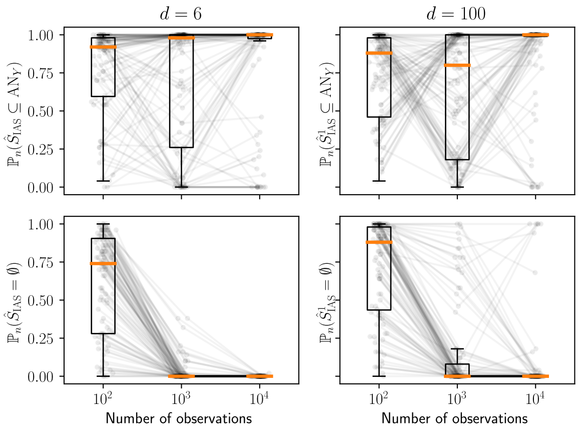

Figure 8: The same figure as Figure 7, but with $\alpha_0 = \alpha = 0.05$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at the non-conservative level $\alpha_0 = \alpha$ means that the empty set is output less often for small sample sizes, but decreases the probability that the output is a subset of ancestors. Thus, we find more ancestor-candidates, but make more mistakes when $\alpha_0 = \alpha$. However, the median probability of the output being a subset of ancestors is at least 80% in all configurations.

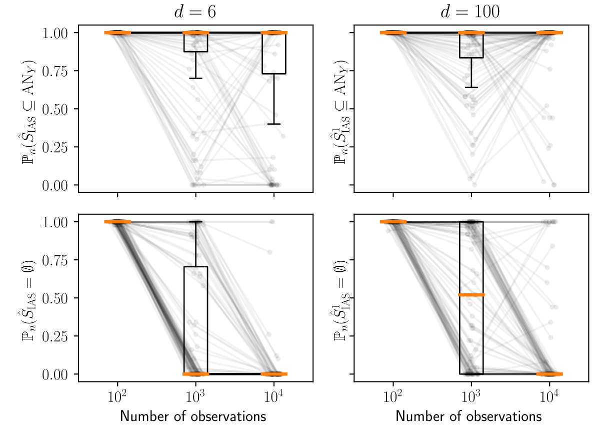

Figure 9: The same figure as Figure 7, but with $\alpha_0 = 10^{-12}$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at at very conservative level $\alpha_0 = 10^{-12}$ means that the empty set is output more often (for one hundred observations, we only find the empty set), but increases the probability that the output is a subset of ancestors. Thus, testing at a very conservative level $\alpha_0 = 10^{-12}$ means that we do not make many mistakes, but the output is often non-informative.

### E.5 Analysis of the strength of inverventions in Section 6.2

Here, we repeat the $d = 6$ simulations from Section 6.2 with a reduced strength of the environment to investigate the performance of IAS under weaker interventions. We sample from the same SCMs as sampled in Section 6.2, but reduce the strength of the interventions to be 0.5 instead of 1. That is, the observational distributions are the same as in Section 6.2, but interventions to a node $X_j$ are here half as strong as in Section 6.2.

The Jaccard similarity between $\hat{S}_{\text{IAS}}$ and $\text{AN}_Y$ is generally lower than what we found in Figure 4 (see Figure 10). This is likely due to having lower power to detect non-invariance, which has two implications. First, lower power means that we may fail to reject the empty set, meaning that we output nothing. Then, the Jaccard similarity between $\hat{S}_{\text{IAS}}$ and $\text{AN}_Y$ is zero. Second, it may be that we correctly reject the empty set, but fail to reject another non-invariant set which is not an ancestor of $Y$ which is then potentially included in the output. Then, the $\hat{S}_{\text{IAS}}$ and $\text{AN}_Y$ is lower, because we increase the number of false findings.

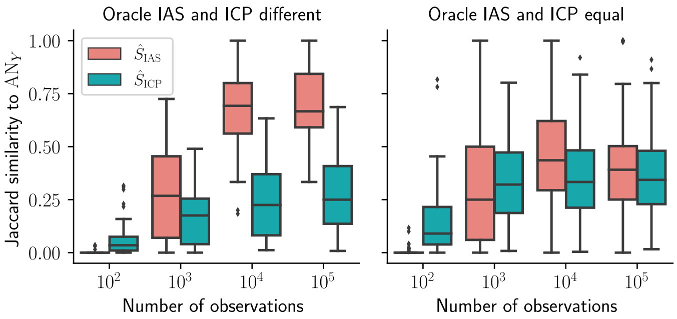

Figure 10: The same figure as the one presented in Figure 4, but with weaker environments (do-interventions of strength 0.5 compared to 1 in Figure 4). Generally, IAS performs the same for weaker interventions as for strong interventions, when there are more than 10,000 observations Graphs represented in each boxplot: 42 (left), 58 (right).

We find that the probability that $\hat{S}_{\text{IAS}}$ is a subset of ancestors is generally unchanged for the lower intervention strength, but the probability of $\hat{S}_{\text{IAS}}$ generally increases for small sample sizes (see Table 1). This indicates that IAS does not make more mistakes under the weaker interventions, but it is more often uninformative. We see also that in both settings, $\hat{S}_{\text{IAS}}$ is empty more often than $\hat{S}_{\text{ICP}}$ for low sample sizes, but less often for larger samples (see Table 1). This is likely because IAS tests the empty set at a much lower level than ICP does ($10^{-6}$ compared to 0.05). Thus, IAS requires more power to find anything, but once it has sufficient power, it finds more than ICP (see also Figure 10). The median probability of ICP returning a subset of the ancestors was always at least 95% (not shown).

Table 1: Summary of the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}_{\text{ICP}} = \emptyset)$ for weak and strong do-interventions (strength 0.5 and 1, respectively) when $d = 6$. Numbers not in parentheses are means, numbers in parentheses are medians. The level is generally unchanged when the environments have a weaker effect, but the power is lower, in the sense that the empty set is output more often.

| | | $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$ | $\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$ | $\mathbb{P}(\hat{S}_{\text{ICP}} = \emptyset)$ |
|---|---|---|---|---|
| Strong interventions | $n = 100$ | 96.6% (100%) | 89.6% (98%) | 52.3% (52%) |
| | $n = 1{,}000$ | 75.7% (100%) | 10.0% (0%) | 30.4% (14%) |
| | $n = 10{,}000$ | 83.7% (100%) | 1.0% (0%) | 24.9% (10%) |
| | $n = 100{,}000$ | 93.8% (100%) | 0.2% (0%) | 22.9% (10%) |
| Weak interventions | $n = 100$ | 99.3% (100%) | 98.7% (100%) | 72.0% (84%) |
| | $n = 1{,}000$ | 81.1% (100%) | 40.2% (26%) | 36.9% (24%) |
| | $n = 10{,}000$ | 80.8% (100%) | 1.7% (0%) | 27.5% (15%) |
| | $n = 100{,}000$ | 92.6% (100%) | 1.1% (0%) | 24.8% (14%) |

### E.6 Analysis of the Choice of $q_{TB}$ in Section 6.3

In this section, we analyze the effect of changing the cut-off $q_{TB}$ that determines when a gene pair is considered a true positive in Section 6.3. For the results in the main paper, we use $q_{TB} = 1\%$, meaning that the pair $(\texttt{gene}_X, \texttt{gene}_Y)$ is considered a true positive if the value of $\texttt{gene}_Y$ when intervening on $\texttt{gene}_X$ is outside of the 0.01- and 0.99-quantiles of $\texttt{gene}_Y$ in the observational distribution. In Figure 11, we plot the true positive rates for several other choices of $q_{TB}$. We compare to the true positive rate of random guessing, which also increases if the criterion becomes easier to satisfy. We observe that the choice of $q_{TB}$ does not substantially change the excess true positive rate of our method compared to random guessing. This indicates that while the true positives in this experiments are inferred from data, the conclusions drawn in Figure 5 are robust with respect to some modelling choices of $q_{TB}$.

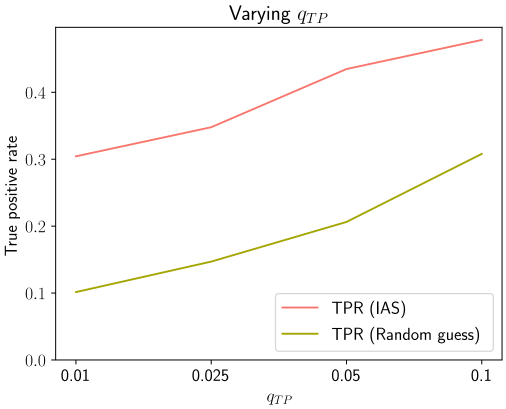

Figure 11: True positive rates (TPRs) for the gene experiment in Section 6.3. $q_{TB}$ specifies the quantile in the observed distribution that an intervention effect has to exceed to be considered a true positive. While the TPR increases for our method when $q_{TB}$ is increased, the TPR of random guessing increases comparably. This validates that changing the definition of true positives in this experiment by choosing a different $q_{TB}$ does not change the conclusion of the experiment substantially.