Journal of Machine Learning Research 27 (2026) 1-64 Submitted 11/22; Revised 8/26; Published 9/26

Characterization and Greedy Learning of Gaussian Structural Causal Models under Unknown Interventions

Juan L. Gamella
Seminar for Statistics
ETH Zürich
Zürich, Switzerland JUAN@CAUSALCHAMBER.AI

Armeen Taeb
Department of Statistics, University of Washington ATAEB@UW.EDU

Christina Heinze-Deml
Seminar for Statistics, ETH Zürich HEINZEDEML@STAT.MATH.ETHZ.CH

Peter Bühlmann
Seminar for Statistics, ETH Zürich PETER.BUEHLMANN@STAT.MATH.ETHZ.CH

Editor: Vanessa Didelez

Abstract

We consider the problem of recovering the causal structure underlying observations from different experimental conditions when the targets of the interventions in each experiment are unknown. We assume a linear structural causal model with additive Gaussian noise and consider interventions that perturb their targets while maintaining the causal relationships in the system. Different models may entail the same distributions, offering competing causal explanations for the given observations. We fully characterize this equivalence class and offer identifiability results, which we use to derive a greedy algorithm called GnIES to recover the equivalence class of the data-generating model without knowledge of the intervention targets. In addition, we develop a novel procedure to generate semi-synthetic data sets with known causal ground truth but distributions closely resembling those of a real data set of choice. We leverage this procedure and evaluate the performance of GnIES on an array of synthetic and semi-synthetic data sets, and real data from a biological system and a tightly controlled physical system. We provide, in the Python packages `gnies` and `sempler`, implementations of GnIES and our semi-synthetic data generation procedure.

Keywords: causality, causal discovery, graphical models, equivalence classes, greedy search

1. Introduction

Knowledge of the causal relationships underlying a system or phenomenon allows us to predict its behavior as it undergoes manipulations. Examples of such knowledge are ubiquitous in human cognition starting from an early age (Gopnik et al., 2004; Muentener and Bonawitz, 2017). It is also of particular importance in science, where understanding a phenomenon usually means understanding the causal relationships which govern it. Scientists are typically interested in causal relationships because they want to intervene on the system (Didelez, 2005), with ambitious goals ranging from curing diseases to limiting the effects of climate change.

©2026 Juan L. Gamella, Armeen Taeb, Christina Heinze-Deml, Peter Bühlmann.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v27/22-1365.html`.

Causal models (Pearl, 2009a) aim at approximating causal relations. They can be seen as abstractions of more accurate physical or mechanistic models that retain enough power to answer interventional or counterfactual queries (Peters et al., 2017). A causal understanding also benefits traditional prediction and regression problems, as it allows for robust predictions which generalize to new and unseen environments (Subbaswamy et al., 2019; Bühlmann, 2020; Rothenhäusler et al., 2021; Pfister et al., 2021; Heinze-Deml and Meinshausen, 2021).

Inferring causal models from data is commonly referred to in the literature as *causal learning* or *causal discovery*. Akin to statistical learning, this task is challenged by the difficulty of trying to infer properties of a distribution from a finite sample. However, even full knowledge of the underlying distribution does not make causal learning trivial, as different causal models can entail the same distribution, and thus offer competing explanations for the observations gathered from the system. Under additional assumptions about the model class or noise distributions, identifiability of the true causal model is possible (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011; Peters and Bühlmann, 2014; Bühlmann et al., 2014). Without making such assumptions, identifiability can be improved through observations of the system under additional experimental conditions, where some variables have received interventions. This gain in identifiability strongly depends on the nature and targets of the experiments (Eberhardt, 2007; Eberhardt and Scheines, 2007; Hauser and Bühlmann, 2014; Gamella and Heinze-Deml, 2020).

1.1 Related Work

Motivated by the gain in identifiability, several works in the causal discovery literature deal with data collected from different interventional environments, for example, observations gathered under different experimental conditions (Hauser and Bühlmann, 2012; Magliacane et al., 2016; Zhang et al., 2017; Yang et al., 2018). However, they require full knowledge of the intervention targets, an assumption that may not be satisfied in many settings, for example, when interventions may have off-target effects. Methods based on the invariance principle (Peters et al., 2016; Meinshausen et al., 2016; Ghassami et al., 2017; Heinze-Deml et al., 2018; Pfister et al., 2019) do not require such knowledge but still place restrictions on which targets are allowed; as a result, they often estimate only parts of the causal graph, such as the direct causes of a variable of interest. Moreover, while some of these methods allow control against false causal selection, they are often overly conservative.

Other recently proposed methods place weaker assumptions on the intervention targets. Joint causal inference[1] (Mooij et al., 2022) treats environments as additional nodes in the causal graph, with some flexible assumptions on their structure; this allows extending standard methods based on conditional independence testing to this setting, such as the PC (Spirtes et al., 2000), IC (Pearl, 2009a), FCI (Spirtes et al., 1999; Zhang, 2008), and GSP (Solus et al., 2021) algorithms. An extension of the latter, the recent UT-IGSP algorithm additionally employs invariance tests to estimate the causal graph without knowledge of the intervention targets, under flexible assumptions on the nature of the interventions (Squires et al., 2020b, Assumptions 1 and 2). Kocaoglu et al. (2019); Jaber et al. (2020) study the

1. This paper also contains a detailed (but non-exhaustive) summary of methods that deal with data from multiple environments and their assumptions (Mooij et al., 2022, table 4).

same setting in the presence of hidden confounders and provide an algorithm based on the same kind of tests. A drawback of methods based on conditional independence and invariance tests is that these usually require large samples, making them ill-suited for situations where only a few observations of the system per environment are available.

Recent works dealing with a small-sample setting employ gradient-based optimization on a continuous relaxation of the causal discovery problem (Zheng et al., 2018; Yu et al., 2019; Lachapelle et al., 2020; Zheng et al., 2020). Extensions have been made to the setting with interventional data with unknown targets (Ke et al., 2019; Brouillard et al., 2020; Faria et al., 2022; Hägele et al., 2023). However, there are indications that the remarkable performance of some of these methods relies on an artifact of the synthetic evaluation data, and can be matched by a simple algorithm that directly exploits it (Reisach et al., 2021). Furthermore, Seng et al. (2022) found this reliance to be a vulnerability to adversarial attacks.

1.2 Outline and Contributions

We consider linear Gaussian structural causal models for interventional or perturbation data as detailed in Section 2. Such models, often used for different environments or domain sources, have become increasingly popular in the advent of heterogeneous data (Hauser and Bühlmann, 2012, 2015; Peters et al., 2016; Rothenhäusler et al., 2021; Taeb et al., 2024). Our model is a special case of the one used in Taeb et al. (2024), but without considering the setting with hidden variables. The simpler model allows us to derive a precise characterization of identifiable structures (Section 3). Based on this result, we build a greedy algorithm (Section 4) which we call *greedy noise-interventional equivalence search* (GnIES) and which is less heuristic-based than the ad-hoc optimization in Taeb et al. (2024). Yet, our model in Section 2 is still much more flexible than the ones for observational data only (Chickering, 2002; van de Geer and Bühlmann, 2013) or when interventions have known targets (Hauser and Bühlmann, 2012, 2015; Wang et al., 2017).

Our proposed algorithm GnIES can be seen as an extension of the celebrated greedy equivalence search (GES) of the penalized maximum likelihood estimator for observational data (Chickering, 2002) to the setting of interventional data with unknown intervention targets. By aggregating statistical strength across environments, the maximum likelihood estimator becomes particularly attractive when there are many different environments or domain sources with few observations only. In Section 5 we evaluate the performance of the algorithm on synthetic and real data sets. The results show that despite the strong distributional assumptions (linearity and normality), GnIES is robust to an array of model violations, and is competitive in recovering the causal structure in small-sample settings. We study its computational complexity and find it to be on par with competing methods.

The algorithm is made available in the Python package `gnies`. More details about the package and the code to reproduce the experiments and figures from the paper can be found in the repository `github.com/juangamella/gnies-paper`. A detailed summary of the software contributions of the paper, including Python implementations of the baselines and the package `sempler` to generate semi-synthetic data, can be found in Appendix F.

2. Our Model

We consider the setting where we have access to observations of variables under different environments, such as experimental conditions in which some of the variables may have been manipulated, that is, received interventions. To represent this setting, we consider a vector $X^e \in \mathbb{R}^p$ of random variables observed in environment e, and a collection $\mathcal{E}$ of such environments. We denote the resulting collection of distributions as $\{P^e\}_{e\in\mathcal{E}}$. We model these distributions via a parametric model of the form

$$X^e = BX^e + \epsilon^e \quad \forall e \in \mathcal{E}, \tag{1}$$

where:

i) $B \in \mathbb{R}^{p\times p}$ corresponds to the adjacency matrix of a directed acyclic graph (DAG), that is, it has zeros on the diagonal and is lower triangular up to a permutation of rows and columns.

ii) $\epsilon^e = (\epsilon_1^e, ..., \epsilon_p^e) \sim \mathcal{N}(\nu^e, \Omega^e)$ are noise terms with $\epsilon_i^e \perp\!\!\!\perp \epsilon_j^e \ \forall i \neq j$, that is, Ω^e is a diagonal matrix with positive entries on the diagonal.

The independence requirement in (ii) corresponds to the assumption of causal sufficiency. Furthermore, we assume that the observed variables X^e and noise terms ϵ^e are independent across environments. These assumptions mean that for each environment e, the variables X^e follow a structural causal model (SCM) with the same coefficients B across all environments $\mathcal{E}$. The effect of interventions is modeled by allowing the distributions of the noise terms to change across environments. We say that

$$\text{variable } X_i \text{ has received an intervention if } \epsilon_i^e \overset{\text{d}}{\neq} \epsilon_i^f \text{ for some } e, f \in \mathcal{E},$$

that is, if the distribution of its noise term is different for at least two environments. The distribution of the observed variables in each environment is the multivariate Gaussian distribution, that is, $X^e \sim P^e$ with $P^e = \mathcal{N}(\mu^e, \Sigma^e)$ where

$$\mu^e = (I - B)^{-1}\nu^e \ \text{ and } \ \Sigma^e = (I - B)^{-1}\Omega^e(I - B)^{-T}. \tag{2}$$

The matrix B is the adjacency matrix of the graph underlying the SCM, which we denote by

$$\mathcal{G}(B) := (V, E) \text{ where } V = \{1, ..., p\} \text{ and } E = \{(j \to i) : B_{ij} \neq 0\}. \tag{3}$$

Each distribution P^e satisfies the Markov property with respect to the graph $\mathcal{G}(B)$ (Pearl, 2009b; Peters et al., 2017, Proposition 6.31).

The model formulation in (1) defines interventions without alluding to an "observational" environment. Because interventions leave the causal structure intact, the choice of such an environment is arbitrary (Figure 1). This is in contrast to equivalence classes encompassing other interventions, e.g., do-interventions, that are usually defined relative to an observational graph or distribution. We discuss the relationships to other interventional equivalence classes in Section 3.1.

2.1 Equivalent Models

Throughout the paper, we refer to the term "model" as the one from (1), which contains a causal structure reflected in the DAG $\mathcal{G}(B)$ (3). Under our modeling assumptions, we find that different models may entail the same set of distributions $\{P^e\}_{e\in\mathcal{E}}$. Since the distribution over the observed variables is Gaussian, it suffices for two models to entail distributions with the same mean and covariance. Together with (2) we arrive at the following definition:

Definition 1 (distribution equivalent models) *Given a collection of environments $\mathcal{E}$, we call two models $\big(B, \{\nu^e, \Omega^e\}_{e\in\mathcal{E}}\big)$ and $\big(\tilde{B}, \{\tilde{\nu}^e, \tilde{\Omega}^e\}_{e\in\mathcal{E}}\big)$* distribution equivalent *if, for all $e \in \mathcal{E}$,*

i) $(I - B)^{-1}\nu^e = (I - \tilde{B})^{-1}\tilde{\nu}^e$, *and*

ii) $(I - B)^{-1}\Omega^e(I - B)^{-T} = (I - \tilde{B})^{-1}\tilde{\Omega}^e(I - \tilde{B})^{-T}$.

The definition does not include intervention targets because these are assumed to be unknown. They are part of the graphical characterization in Section 3. The true data-generating model, and importantly, its causal structure, may not be identifiable without additional assumptions. As we will show in the next section, if the collection of environments is a singleton ($|\mathcal{E}| = 1$), the class of distribution equivalent models is a superset of the Markov equivalence class. Under faithfulness, and without additional assumptions—such as equal noise variances (Peters and Bühlmann, 2014)—the distribution equivalent models with the sparsest connectivity matrices are the same as the Markov equivalence class (see Proposition 6). Additionally, two immediate results follow from Definition 1. First, the class of equivalent models cannot grow as we add additional environments:

Corollary 1 *Let $\mathcal{E}_1, \mathcal{E}_2$ be collections of environments such that $\mathcal{E}_1 \subseteq \mathcal{E}_2$. If two models are distribution equivalent under $\mathcal{E}_2$, they are also under $\mathcal{E}_1$.*

Second, the class does not shrink under interventions that change only the means of the noise terms. To see this, consider that for any $(B, \{\nu_e\}_{e\in\mathcal{E}})$, condition (i) is satisfied by any $(\tilde{B}, \{\tilde{\nu}^e\}_{e\in\mathcal{E}})$ where $\tilde{\nu}^e = (I - \tilde{B})(I - B)^{-1}\nu^e$ for all $e \in \mathcal{E}$. Without additional assumptions on their targets or size, these interventions do not provide additional identifiability. This argument does not hold for interventions that affect the noise-term variances, as we impose the constraint that $\tilde{\Omega}^e$ be diagonal. Thus, from now on we consider the data to be centered; for a model $\big(B, \{\nu^e, \Omega^e\}_{e\in\mathcal{E}}\big)$, by (2) and the fact that $(I - B)$ is always a full-rank matrix, this assumption results in $\nu^e = 0$ for all $e \in \mathcal{E}$. Therefore, from this point on we will denote a model by a tuple $(B, \{\Omega^e\}_{e\in\mathcal{E}})$, and by $[(B, \{\Omega^e\}_{e\in\mathcal{E}})]$ its class of distribution equivalent models.

2.2 A practical discussion about the intervention model

Do-, hard or perfect interventions (Pearl, 2009b; Eberhardt and Scheines, 2007; Peters et al., 2017) remove the incoming edges of the target, removing the influence of its parents. Thus, they model settings where interventions isolate the target from the effect of other variables in the system. In many settings, such as the real examples in Section 5, interventions can only hope to perturb a variable, and the assumption above may be unrealistic. Instead,

we follow previous work in modeling interventions that *modify* the causal relationships of the target, retaining its parental set (Tian and Pearl, 2001; Squires et al., 2020b; Kocaoglu et al., 2019; Jaber et al., 2020; Taeb et al., 2024). For the class of SCMs we consider, this can be achieved by either changing the parental coefficients of the target in B, or modifying the distribution of the noise term. We leave the former as future work (see Section 6.1 for a discussion) and focus instead on the latter type of intervention. Because the noise term of a variable propagates through the causal path, it models the uncertainty arising from unmeasured causes of the variable, instead of measurement noise—which would not propagate downstream. Therefore, changes in the mean and variance of the noise terms can potentially capture various interventions, as highlighted by the real-data experiments in Section 5 (e.g., Figure 6c).

3. Graphical Characterization of Distributional Equivalence

In this section, we give a graphical characterization of distribution equivalent models, which we then use in Section 4 to construct an algorithm that performs a greedy search over the space of distributional equivalence classes. The proofs for these results can be found in Appendix A. We refer to a *graphical* characterization because we describe the graphs underlying the models in the distributional equivalence class. More formally, for a class $\left[\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)\right]$ of distribution equivalent models, we provide results describing the set

$$\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right) := \left\{G : \exists(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})] \text{ such that } \tilde{B} \sim G\right\}, \tag{4}$$

where $\tilde{B} \sim G$ denotes that the adjacency matrix $\tilde{B}$ satisfies the sparsity pattern of the graph G, that is, that $(i \to j) \notin G \implies \tilde{B}_{ji} = 0$. Our results are based on the notion of *transition sequence equivalence*, introduced by Tian and Pearl (2001). For notational convenience and to account for the different terminology, we refer to it here as $\mathcal{I}$-equivalence, where $\mathcal{I} \subseteq \{1, ..., p\}$ corresponds to a set of intervention targets.

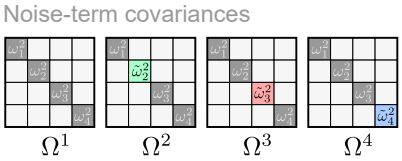
Definition 2 ($\mathcal{I}$-equivalence) *Two DAGs G and G' are said to be* $\mathcal{I}$-equivalent *under interventions on targets $\mathcal{I} \subseteq \{1, ..., p\}$ if*

i) they have the same skeleton,

ii) they have the same v-structures, and

iii) for all $i \in \mathcal{I}$, node i has the same parents in G and G'.

We denote $\mathcal{I}$-equivalence between graphs G and G' by $G \sim_{\mathcal{I}} G'$. Furthermore, we denote by $\mathcal{I}-\text{MEC}(G)$ the class of graphs $\mathcal{I}$-equivalent to G. This nomenclature is motivated by the fact that two graphs that satisfy conditions (i) and (ii) are Markov equivalent (Verma and Pearl, 1990). As such, a class of $\mathcal{I}$-equivalent DAGs is a subset of their Markov equivalence class, and both classes are the same when $\mathcal{I} = \emptyset$. Together, conditions (i) and (iii) imply that two $\mathcal{I}$-equivalent graphs also have the same children for all variables $i \in \mathcal{I}$. In other words, an intervention orients all the edges around its target, and then possibly more following an application of the Meek rules (Meek, 1995, Figure 1).



ref.	Changes wrt. ref. env.	$\mathbb{I}(\{\Omega^1, \ldots, \Omega^4\})$
1	$\{\emptyset, \{X_2\}, \{X_3\}, \{X_4\}\}$	$\{X_2, X_3, X_4\}$
2	$\{\{X_2\}, \emptyset, \{X_2, X_3\}, \{X_2, X_4\}\}$	$\{X_2, X_3, X_4\}$
3	$\{\{X_3\}, \{X_3, X_2\}, \emptyset, \{X_3, X_4\}\}$	$\{X_2, X_3, X_4\}$
4	$\{\{X_4\}, \{X_4, X_2\}, \{X_4, X_3\}, \emptyset\}$	$\{X_2, X_3, X_4\}$

Figure 1: Because the intervention model (Section 2) leaves the causal structure intact, we define targets as variables whose noise-term variance changes across environments (5), as opposed to changes with respect to a reference or observational environment—which in this case is arbitrary. Indeed, given a set of noise-term covariances (left), different reference environments would yield different sets of targets (right).

In the remainder of this section, we give results on the assumptions for which the $\mathcal{I}$-equivalence class of the graph $\mathcal{G}(B)$ underlying a model $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ fully describes the set $\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$ of its distribution equivalent graphs. We use

$$\mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big) := \Big\{j : \exists e, f \in \mathcal{E} \text{ s.t. } \Omega^e_{jj} \neq \Omega^f_{jj}\Big\} \tag{5}$$

to denote the indices of variables that have received an intervention in $\mathcal{E}$. We choose this formulation—as opposed to defining interventions in contraposition to a reference observational environment—because our intervention model entails no such environment. As we illustrate in Figure 1, different choices of reference environment would yield different intervention targets. Under our formulation, the exact changes of noise-term variances across environments are fully identifiable (see Corollary 2 below).

Note that when $|\mathcal{E}| = 1$, then $\mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big) = \emptyset$. Without additional assumptions, we have that the $\mathcal{I}$-equivalence class of the data-generating graph $\mathcal{G}(B)$ is a subset of the class of distribution equivalent graphs:

Lemma 1 *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ with underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}\text{−MEC}\left(\mathcal{G}(B)\right) \subseteq \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right).$$

If we additionally assume faithfulness for the data-generating model, its graph $\mathcal{G}(B)$ has the minimal number of edges[2] and the $\mathcal{I}$-equivalence class is a subset of the sparsest graphs in $\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$:

Proposition 1 *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}\text{−MEC}\left(\mathcal{G}(B)\right) \subseteq \underset{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})}{\arg\min} \|G\|_0,$$

where $\|G\|_0$ is the number of edges in G.

2. See Lemma 14 in Appendix A.

Our goal is to identify the sparsest graphs $\arg\min_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})}\|G\|_0$ that are compatible with the distribution of the data, that is, the most succinct causal explanations which (under faithfulness) comprise the data-generating graph and its equivalents. Without additional assumptions on the parameters of the interventions, Proposition 1 is the best one can do. To see this, consider a data-generating model $(B,\{\Omega^e,\Omega^f\})$ with two interventional environments e and f, and suppose that $\mathbb{I}\big(\{\Omega^e,\Omega^f\}\big)=[p]$, that is, there has been an intervention on all variables. For $\mathcal{I}=\mathbb{I}\big(\{\Omega^e,\Omega^f\}\big)$, it follows from Definition 2 that $\mathcal{I}-\mathrm{MEC}\left(\mathcal{G}(B)\right)$ is a singleton. However, if the interventions simply scale all noise terms by the same factor, that is, $\forall i\ \Omega^e_{ii}/\Omega^f_{ii}=\alpha$, then the set $\arg\min_{G\in\mathbb{G}\left(B,\{\Omega^e,\Omega^f\}\right)}\|G\|_0$ still contains the complete Markov equivalence class of $\mathcal{G}(B)$, which is in general not a singleton. We provide the complete example in Appendix D. To guard against such pathologically selected intervention parameters, we introduce an assumption which we call *intervention-heterogeneity*:

Assumption 1 (Intervention-heterogeneity) *A model $\big(B,\{\Omega^e\}_{e\in\mathcal{E}}\big)$ is said to satisfy* intervention-heterogeneity *if for every pair of environments $e,f\in\mathcal{E}$, it holds that*

$$\frac{\Omega^e_{ii}}{\Omega^f_{ii}}\neq\frac{\Omega^e_{jj}}{\Omega^f_{jj}}\ \textit{for all}\ i\neq j\in\mathbb{I}\big(\{\Omega^e,\Omega^f\}\big).$$

In words, this assumption means that the noise-term variances which change between environments do so by a different factor. This is trivially satisfied when environments contain interventions on a single target. Furthermore, if the intervention variances are independently sampled from an absolutely continuous distribution, violations of Assumption 1 have zero probability.

Under faithfulness and this assumption, we have that the $\mathcal{I}$-equivalence class is the set of sparsest distribution equivalent graphs:

Theorem 1 *Consider a model $\big(B,\{\Omega^e\}_{e\in\mathcal{E}}\big)$ satisfying Assumption 1 and resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I}:=\mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}-\mathrm{MEC}\left(\mathcal{G}(B)\right)=\operatorname*{arg\,min}_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})}\|G\|_0.$$

Whereas Theorem 1 provides an identifiability result for the graphs which underlie distribution equivalent models, the following corollary provides a result on the identifiability of their intervention targets.

Corollary 2 *Consider a model $\big(B,\{\Omega^e\}_{e\in\mathcal{E}}\big)$ satisfying Assumption 1 and resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I}:=\mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then, for any distribution equivalent model $\big(\tilde{B},\{\tilde{\Omega}^e\}_{e\in\mathcal{E}}\big)\in\big[\big(B,\{\Omega^e\}_{e\in\mathcal{E}}\big)\big]$, it holds that*

$$\tilde{\Omega}^e_{ii}=\Omega^e_{ii}\ \forall i\in\mathcal{I}\ ,\forall e\in\mathcal{E}\ \ \textit{and}\ \ \tilde{\Omega}^e_{jj}=\tilde{\Omega}^f_{jj}\ ,\forall j\notin\mathcal{I}\ \forall e,f\in\mathcal{E}.$$

In other words, all equivalent models entail the same interventions, and the noise-term variance of targets for each environment is the same across equivalent models. The proof can be found in Appendix A.

3.1 Relation to other Equivalence Classes

Our graphical characterization is equivalent to the *transition sequence equivalence* introduced by Tian and Pearl (2001, Theorems 2 and 3), which describes how the Markov equivalence class shrinks under *single-target* interventions which do not alter the parental set of the target (Tian and Pearl, 2001, Definition 1). Assumption 1 is a sufficient condition under which this characterization also holds for our parametric setting (see Equation 1) and *multiple-target* interventions. In Appendix D we provide an example of how this equivalence breaks down under violations of Assumption 1. However, further assumptions (e.g., Tian and Pearl, 2001, Definition 2) may be needed for the general non-parametric models treated in Hauser and Bühlmann (2012); Yang et al. (2018); Jaber et al. (2020).

To compare our $\mathcal{I}$-equivalence class with the interventional equivalence classes introduced by the works mentioned above, we first show that $\mathcal{I}$-equivalence can also be expressed in terms of Markov equivalence of augmented graphs. In our case, we construct the augmented graph as follows:

Definition 3 (Augmented graph—this work) *Given a graph G over p variables, and a set of targets $\mathcal{I} \subseteq [p]$, the* augmented graph $G^{\mathcal{I}}$ *is obtained by adding a new node F_i and an edge $F_i \to i$ for all $i \in \mathcal{I}$.*

An example is shown in Figure 2. We can now express $\mathcal{I}$-equivalence as Markov equivalence between these augmented graphs, as stated in Lemma 2 below. The proof can be found in Appendix A.6.

Lemma 2 *Let G_1 and G_2 be two DAGs over p variables and let $\mathcal{I} \subseteq [p]$ be a set of intervention targets. The following statements are equivalent:*

1. *G_1 and G_2 are $\mathcal{I}$-equivalent, and*
2. *their augmented graphs $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ are Markov equivalent.*

Using this alternative formulation, we compare our equivalence class to those of Hauser and Bühlmann (2012); Yang et al. (2018) and Jaber et al. (2020) in Sections 3.1.1 to 3.1.3 below. Under single-target interventions, our equivalence class results in the same sets of graphs; however, under interventions that simultaneously affect multiple targets, the parametric constraints encoded in our model allow for additional identifiability, as shown with an example in Figure 2. In Appendix D, we show how this additional identifiability crucially depends on Assumption 1, which allows treating an intervention on multiple targets as multiple single-target interventions.

3.1.1 Hauser and Bühlmann (2012)

The authors present an interventional equivalence class for the case of *hard* interventions (Pearl, 2009b, Section 3.2.2). Also called *do-* or *perfect* interventions, these isolate the intervened variable from the effect of its causal parents, effectively removing the incoming edges in the causal graph. One key difference to our formalism is how the intervention targets are represented. While we do so by means of a subset $\mathcal{I} \subseteq [p]$ of the observed variables, Hauser and Bühlmann consider instead a set of subsets, which they refer to as

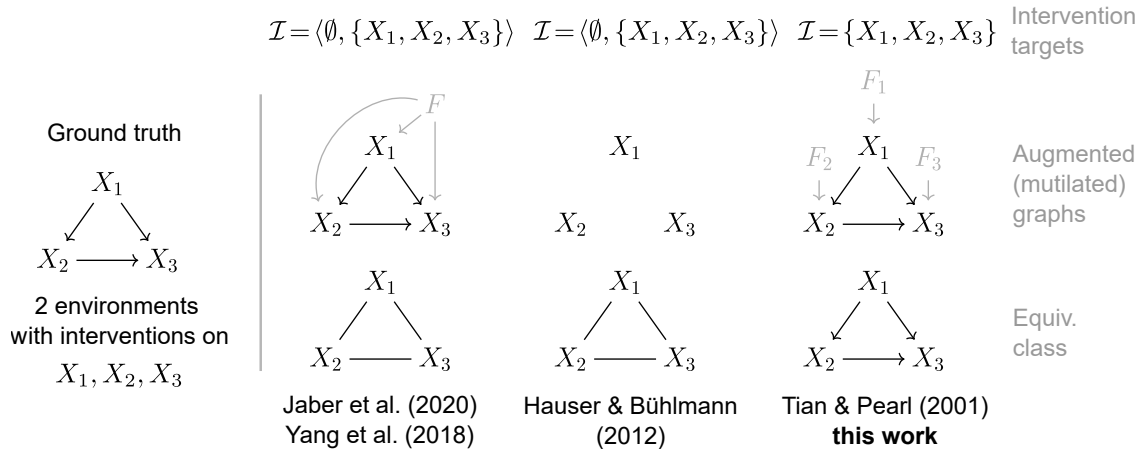


Figure 2: Comparison of the equivalence classes obtained for the ground-truth graph and intervention targets on the left for the different definitions given in Tian and Pearl (2001); Hauser and Bühlmann (2012); Yang et al. (2018); Jaber et al. (2020) and this work. The works differ in their treatment of the intervention targets (see Equation 5 and Figure 1 above) and the resulting equivalence classes. In the example above (left), Jaber et al. (2020) and Yang et al. (2018) result in the same augmented graph and equivalence class, although they are not generally the same (see Jaber et al., 2020, Appendix D.1). The parametric constraints of our model, together with Assumption 1, allow for additional identifiability in the case of interventions that affect multiple targets; see Appendix D for more details.

a *family* of targets. They require that this family $\mathcal{H}$ be *conservative*, meaning that for all $j \in [p]$, there exists $I \in \mathcal{H}$ such that $j \notin I$ (Hauser and Bühlmann, 2012, Definition 6). Note that in the presence of an observational environment, that is, $\emptyset \in \mathcal{H}$, $\mathcal{H}$ is conservative. To avoid confusion with our notation, we denote their equivalence class by $\mathcal{H}-$MEC. It arises from the following equivalence relation:

Proposition 2 (Theorem 10 from Hauser and Bühlmann, 2012) *Let G_1, G_2 be two DAGs and let $\mathcal{H}$ be a conservative family of targets. Then, G_1, G_2 are equivalent if and only if (i) G_1 and G_2 have the same skeleton and the same v-structures, and (ii) $G_1^{(H)}$ and $G_2^{(H)}$ have the same skeleton for all $H \in \mathcal{H}$, where $G_i^{(H)}$ denotes the graph obtained by removing from G_i all edges incoming to the nodes in H.*

While the $\mathcal{I}$-equivalence class and the $\mathcal{H}-$MEC are the same for single-target interventions, the parametric setting we consider (Section 2), together with Assumption 1, yields additional identifiability for multiple-target interventions—see Figure 2 for an example. Thus, the $\mathcal{I}$-equivalence class is a subset of the $\mathcal{H}-$MEC, as we state in the following result. The proof can be found in Appendix A.6.

Proposition 3 *Consider a DAG G and a conservative family of targets $\mathcal{H}$. Let $\mathcal{I} := \cup_{H\in\mathcal{H}} H$. Then $\mathcal{I}-\text{MEC}(G) \subseteq \mathcal{H}-\text{MEC}(G)$.*

3.1.2 Yang et al. (2018)

The authors give an interventional equivalence class for the case of *general* interventions (Yang et al., 2018, Definition 3.3). These can be understood as all interventions for which the resulting interventional distribution is still Markov with respect to the data-generating graph. These exclude interventions that add new parents to a variable or affect several variables in a way that induces a new statistical dependence between them, for example, by adding a hidden confounder. Both the hard interventions from Hauser and Bühlmann (2012) and the noise interventions which we discuss in our paper naturally fall into the category of general interventions. The authors provide a graphical characterization of their equivalence class through the concept of an "interventional DAG". Given a graph G and a family of targets $\mathcal{Y}$, they construct it as follows: for every $Y \in \mathcal{Y} \setminus \{\emptyset\}$ we add a source node to G, with edges from it to the nodes in Y. Then, under the condition that $\emptyset \in \mathcal{Y}$, two DAGs are $\mathcal{Y}$-equivalent if and only if their interventional graphs have the same skeleton and v-structures (Yang et al., 2018, Theorem 3.9). We denote the class of $\mathcal{Y}$-equivalent models by $\mathcal{Y}-$MEC. Since hard interventions are also a special case of general interventions, under causal sufficiency our $\mathcal{I}$-equivalence class is also a subset of the $\mathcal{Y}$-equivalence class, yielding additional identifiability in the case of multiple-target interventions (Figure 2). We state this result in Proposition 4 and provide a proof in Appendix A.6.

Proposition 4 *Consider a DAG G and let $\mathcal{Y}$ be a family of targets such that $\emptyset \in \mathcal{Y}$, and let $\mathcal{I} := \cup_{Y \in \mathcal{Y}} Y$. Then $\mathcal{I}-\text{MEC}(G) \subseteq \mathcal{Y}-\text{MEC}(G)$.*

3.1.3 Jaber et al. (2020)

Building on the results from Kocaoglu et al. (2019), the authors provide a characterization of interventional equivalence (Ψ-Markov equivalence) for the case of soft interventions and violations of causal sufficiency, i.e., hidden confounders. The interventions they consider include those in our model class, and for the case where causal sufficiency is satisfied, the authors provide a graphical characterization which directly applies to our setting (Lemma 3).

Lemma 3 (Corollary 1 from Jaber et al., 2020) *Given causal graphs without latents G_1, G_2 and the corresponding interventional targets $\mathcal{I}_1, \mathcal{I}_2$, the pairs $(G_1, \mathcal{I}_1), (G_2, \mathcal{I}_2)$ are Ψ-Markov equivalent if and only if their augmented graphs have the same skeleton and v-structures.*

The augmented graph is constructed as follows: given a causal graph $G = (V, E)$ and a set of intervention targets $\mathcal{I} = \langle I_1, \dots, I_k \rangle$ with $I_k \subset V$, for each pair of targets $I_i, I_j \in \mathcal{I}$ we add an additional node to the graph, and edges from it to all nodes $v \in \{w \mid (w \in I_i \wedge w \notin I_j) \vee (w \notin I_i \wedge w \in I_j)\}$, that is, nodes which appear in only one of the two targets I_i and I_j. An example is shown in Figure 2.

As for the equivalence class in Yang et al. (2018), $\mathcal{I}$-equivalence and Ψ-Markov equivalence are the same in the case of single-target interventions (Lemma 2 and Lemma 3). However, under interventions that affect several targets in the same environment, the resulting equivalence classes are different: the parametric constraints of our model (Section 2), together with Assumption 1, yield additional identifiability, as shown with an example in Figure 2. In Appendix D, we show how violations of Assumption 1 break this result.

4. Learning the Class of Equivalent Models

Next, we consider the problem of recovering the class of distribution equivalent models from data. In particular, we are interested in recovering the causal structure, that is, the graphs underlying each of the models in the distributional equivalence class of the data-generating model.

As a first approach, one might consider an ℓ_0-penalized maximum likelihood estimator of the form

$$\left(\hat{B}, \{\hat{\Omega}^e\}_{e\in\mathcal{E}}\right) \in \underset{\substack{B\sim\text{DAG} \\ \text{diagonal pos. def. } \{\Omega^e\}_{e\in\mathcal{E}}}}{\arg\max} \sum_{e\in\mathcal{E}} \log P(\boldsymbol{X}^e; B, \Omega^e) - \lambda \left\|\mathcal{G}(B)\right\|_0, \tag{6}$$

where $\boldsymbol{X}^e$ is the sample from environment e, $\lambda > 0$ is a regularization parameter and $\|\mathcal{G}(B)\|_0$ is the number of edges in the graph entailed by B, which we have restricted to be the connectivity matrix of a DAG. The expression for the log-likelihood can be found in Equation 28 in Appendix A.7. This approach suffers from two important drawbacks.

Computing the estimator in (6) is a highly complex task. This stems from the fact that the DAG constraint over B causes the optimization to be highly non-convex. Since the space of DAGs grows super-exponentially with the number of variables, an exhaustive search is infeasible even for a few variables. Even for the observational setting where $|\mathcal{E}| = 1$, a greedy search over this space, that is, starting with an empty graph and greedily adding and removing edges, will often only find local optima and is known to perform poorly (Chickering, 2002; Hauser and Bühlmann, 2012). This constitutes the first drawback.

As a possible way forward, one may note that for $|\mathcal{E}| = 1$, Equation 6 corresponds to the score employed by Chickering (2002) in the GES algorithm. In this case, the class of distribution equivalent models corresponds to the Markov equivalence class[3], and the objective in (6) has some desirable properties which are exploited by the GES algorithm to perform a greedy search over the space of Markov equivalence classes. By iterating over this alternative search space—instead of DAGs—GES is able to escape local minima and return the true equivalence class in the large sample limit. This is possible due to the key property of *score equivalence*: for any finite sample, all members of an equivalence class attain the same score.

In the same manner as GES, we would like to greedily iterate over the space of $\mathcal{I}$-equivalence classes until we arrive at the one containing the data-generating model. Unfortunately, for the setting where $|\mathcal{E}| > 1$, the property of *score equivalence* is not satisfied by (6); $\mathcal{I}$-equivalent graphs may attain different scores on the same finite sample. See Appendix C for a discussion. Motivated by this second drawback, we propose the following score.

4.1 Score Function

For a given DAG G and a set of intervention targets $\mathcal{I}$, we consider the score

$$S(G, \mathcal{I}) = \underset{\substack{B\sim G \\ \text{diagonal pos. def. } \{\Omega^e\}_{e\in\mathcal{E}} \\ \text{s.t. } \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}})\subseteq\mathcal{I}}}{\max} \sum_{e\in\mathcal{E}} \log P(\boldsymbol{X}^e; B, \Omega^e) - \lambda \text{DoF}(G, \mathcal{I}), \tag{7}$$

3. Under faithfulness, see Proposition 6 in Appendix A.

where $\boldsymbol{X}^e$ is the sample from environment e. In words, the constraints placed over the parameters mean that B respects the adjacency of the DAG G, and that the variances of the noise terms remain constant across environments, with the exception of those of the variables in $\mathcal{I}$. The penalization term $\mathrm{DoF}(G, \mathcal{I}) := \|G\|_0 + p + |\mathcal{I}|(|\mathcal{E}| - 1)$ corresponds to the number of free parameters in the model, that is, the number of edges and distinct noise-term variances which need to be estimated.

To develop an efficient and accurate algorithm for identifying the equivalence class of DAGs that best fit the data, we want the score function to satisfy the following properties:

i) *Score equivalence*: for any finite sample, all members of an equivalence class attain the same score, that is, $G \sim_{\mathcal{I}} G' \implies S(G, \mathcal{I}) = S(G', \mathcal{I})$.

ii) *Decomposability*: the score can be written as a sum of terms depending only on a variable and its parents.

iii) *Consistency*: let $G^\star$ be the data-generating DAG and let $\mathcal{I}^\star$ be the true set of intervention targets. Let $\{(G_1, \mathcal{I}_1), ..., (G_l, \mathcal{I}_l)\} = \arg\max_{G,\mathcal{I}} S(G, \mathcal{I})$ be the pairs of DAGs and intervention targets that maximize the score $S(G, \mathcal{I})$. In the large sample limit, where $n^e \to \infty$ for every $e \in \mathcal{E}$ and λ is chosen appropriately, $\mathcal{I}_j = \mathcal{I}^\star$ for all $j \in [l]$, and $\{G_j\}_{j\in[l]} = \mathcal{I}^\star\text{−MEC}(G^\star)$ with probability tending to one.

Indeed, the score (7) satisfies these properties. Score equivalence is attained through the constraint $\mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}) \subseteq \mathcal{I}$ that only the noise-term variances of the targets $\mathcal{I}$ can vary across environments. The proof is presented in Appendix A.7.

Proposition 5 (Score properties) *Without additional assumptions, the score* (7) *satisfies score equivalence* (i) *and decomposability* (ii)*. Under faithfulness and Assumption 1 for the data-generating model, and the assumptions that* $\lim_{\mathit{all}\ n_e\to\infty} \frac{n^e}{\sum_{e\in\mathcal{E}} n^e} > 0$ *for all* $e \in \mathcal{E}$ *and that the parameter space is compact, the score is consistent* (iii)*.*

4.2 Our Greedy Algorithm GnIES

In the remainder of this section, we present the *greedy noise-interventional equivalence search* algorithm (GnIES) to estimate the equivalence class of the data-generating graph under unknown interventions. The algorithm is score-based and composed of two nested, greedy procedures. The inner one proceeds similarly to GES and searches for the optimal equivalence class given a fixed set of intervention targets. The outer procedure greedily searches the space of such intervention targets; its output constitutes the GnIES estimate.

4.2.1 Inner Procedure

For a fixed set of intervention targets $\mathcal{I}$, the inner procedure performs a search over the space of $\mathcal{I}$-equivalence classes, returning the highest-scoring one. The procedure is a modification of the GES algorithm, which we will now describe at a high level with the purpose of illustrating said modification.

As discussed in the previous section, GES performs a search over the space of Markov equivalence classes to find the one that best fits the data, that is, whose constituents maximize the score of choice. This search is carried out in a greedy manner: starting with the equivalence class of the empty graph, the algorithm considers as neighboring classes those whose representatives only have one more edge than those in the current class, and moves to the one which yields the highest score. This "forward-step" is repeated until no transition yields an improvement in the score. With the resulting class as a new starting point, the algorithm then performs the opposite "backward-steps", iteratively transitioning to classes with fewer edges until the score can no longer be improved. The resulting class is returned as the estimate, which is, quite remarkably, consistent (Chickering, 2002, Section 4).

Algorithm 1 Greedy inner procedure of GnIES

1: **Input**: data $\boldsymbol{X}^e$ across environments $e \in \mathcal{E}$, intervention set $\mathcal{I}$.

2: **Set starting point**: set the current CPDAG $C := G_0$ to the empty graph and the current score $s := S(G_0, \mathcal{I})$ to its score.

3: **Forward phase**: Apply the highest scoring GES *insert* operator until the score cannot be improved:

   (a) Compute the change in score for all valid GES *insert* operators which can be applied to the current CPDAG C.

   (b) If the score cannot be improved, end the forward-phase. Otherwise, apply the highest scoring one to C, resulting in the PDAG P and updating the score s.

   (c) $C :=$ `completion_algorithm`$(P, \mathcal{I})$ and repeat steps (a-c).

4: **Backward phase**: Apply the highest scoring GES *delete* operator until the score cannot be improved:

   (a) Compute the change in score for all valid GES *delete* operators which can be applied to the current CPDAG C.

   (b) If the score cannot be improved, end the backward-phase. Otherwise, apply the highest scoring operator to C, resulting in the PDAG P and updating the score s.

   (c) $C :=$ `completion_algorithm`$(P, \mathcal{I})$ and repeat steps (a-c).

5: **Output:** The estimated CPDAG C and its score s.

Internally, GES represents an equivalence class by means of a complete, partially directed acyclic graph (CPDAG) (Chickering, 2002, Section 2.4). Transitions to neighboring classes are implemented as the application of an operator to the CPDAG representing the current equivalence class. Forward steps constitute an application of the so-called *insert* operator, and backward steps of the *delete* operator, respectively adding and removing edges to the CPDAG. Applying an operator results in a PDAG, or partially directed acyclic graph, representing some of the neighboring equivalence class members. This PDAG is then

"completed" using a completion algorithm to transform it into the CPDAG representing the complete class.

The inner procedure of GnIES (Algorithm 1) is exactly GES with the exception of two components: the score (7) and the completion algorithm. All other components, including the operators, remain the same as for GES. The modified completion algorithm takes the PDAG resulting from an operator and the given set of intervention targets, and returns the CPDAG representing the $\mathcal{I}-$MEC, instead of the CPDAG representing the (observational) Markov equivalence class. We describe this completion procedure in Algorithm 3, Appendix E.

4.2.2 Outer Procedure

In Algorithm 2 we describe the outer component of GnIES, which estimates the list of intervention targets from data and returns the estimated equivalence class.

Algorithm 2 GnIES: Equivalence class recovery with unknown targets

1: **Input**: data $\boldsymbol{X}^e$ for $e \in \mathcal{E}$

2: **Compute original score**: initialize the current intervention set $\mathcal{I}_{\text{curr}} := \emptyset$ and compute its score $s_{\text{curr}} := S(\mathcal{I}_{\text{curr}}) := \max_G S(G, \mathcal{I}_{\text{curr}})$ using Algorithm 1.

3: **Greedy addition of target interventions**: add intervention targets until score cannot be improved

   (a) add the intervention target $i = \arg\max_j S(\mathcal{I}_{\text{curr}} \cup \{j\})$ that maximizes the score and let s_{next} be the resulting score.

   (b) if $s_{\text{next}} > s_{\text{curr}}$, set $s_{\text{curr}} := s_{\text{next}}$, $\mathcal{I}_{\text{curr}} := \mathcal{I}_{\text{curr}} \cup \{i\}$ and repeat steps (a-b); otherwise, stop.

4: **Greedy deletion of intervention targets**: delete intervention targets until score cannot be improved

   (a) remove the intervention target $i = \arg\max_{j \in \mathcal{I}_{\text{curr}}} S(\mathcal{I}_{\text{curr}} \setminus \{j\})$ which maximizes the score and let s_{next} be the resulting score.

   (b) if $s_{\text{next}} > s_{\text{curr}}$, set $s_{\text{curr}} := s_{\text{next}}$, $\mathcal{I}_{\text{curr}} := \mathcal{I}_{\text{curr}} \setminus \{i\}$ and repeat steps (a-b); otherwise, stop.

5: **Estimated intervention targets** $\hat{\mathcal{I}} := \mathcal{I}_{\text{curr}}$

6: **Output:** return the intervention targets $\hat{\mathcal{I}}$ and the resulting equivalence class $\hat{\mathcal{I}}-\text{MEC}(\arg\max_G \mathcal{S}(G, \hat{\mathcal{I}}))$, i.e., via Algorithm 1.

Similar to the inner procedure, the outer component of GnIES proceeds greedily: we start from the empty set and greedily add a target until the regularized maximum likelihood score (7) cannot be improved. Then, as in the inner procedure, we move backward by greedily removing an intervention target until the score can no longer be improved. The score is obtained by supplying the data and the set of interventions to the inner procedure

(described in Algorithm 1). Background knowledge, in terms of partially known intervention targets, can be incorporated by keeping them in the estimate $\mathcal{I}_{\text{curr}}$ at every step of Algorithm 2. This functionality is made available through the parameter `known_targets` in the documentation of the `gnies` package.

4.2.3 Consistency of GnIES

Under the assumption that the inner procedure is consistent, the outer procedure is also consistent. That is, as $n^e \to \infty$ for every $e \in \mathcal{E}$ and λ chosen appropriately, the outer procedure returns $\mathcal{I}^\star$ as the intervention targets and $\mathcal{I}^\star\text{−MEC}(G^\star)$ as the equivalence class of graphs with probability tending to one. Formally, for each $\mathcal{I} \subseteq \{1, 2, \dots, p\}$ and DAG G among p nodes, let

$$S^\star(G, \mathcal{I}) = \max_{\substack{B \sim G \\ \text{diagonal pos. def. } \{\Omega^e\}_{e\in\mathcal{E}} \\ \text{s.t. } \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}) \subseteq \mathcal{I}}} \sum_{e\in\mathcal{E}} \mathbb{E}[\log P(\boldsymbol{X}^e; B, \Omega^e)]$$

to be the population analog of $S(G, \mathcal{I})$ in (7).

Theorem 2 *Suppose the assumptions of Proposition 5 are satisfied. Furthermore, suppose that the inner procedure in Algorithm 1 is consistent, i.e., for every $\mathcal{I} \subseteq \{1, 2, \dots, p\}$, it returns $\mathcal{I}$−MEC$(\arg\max_G S^\star(G, \mathcal{I}))$ as $n^e \to \infty$ for every $e \in \mathcal{E}$. Denote by $\hat{\mathcal{I}}_f$ the intervention set resulting from the forward-phase of algorithm 2, and suppose $\mathcal{I}^\star \subseteq \hat{\mathcal{I}}_f$. Then, the output of Algorithm 2 is consistent.*

The proof is presented in Appendix B. Theorem 2 guarantees that the GnIES estimate is consistent, assuming that the inner algorithm 1 is accurate and the greedy addition step (3) of algorithm 2 yields a superset of the true intervention targets. In any case, one could forego this first step altogether by starting from the full set of intervention targets $\mathcal{I} = \{1, ..., p\}$, albeit paying a price in terms of computation time when the true targets are few.

We do not provide a proof for the consistency of the inner procedure, which is GES with a modified score and completion algorithm. For plain GES, the proof consists in first showing that the score is locally consistent and decomposable, and then showing that the insert and delete operators reach all neighboring classes (Chickering, 1995, Theorems 15 and 17). While we show in Proposition 5 that our score satisfies the required properties, we did not extend the results on the operators to our completion algorithm and leave this as future work.

4.2.4 Computational Cost

The computational cost of a single run of the inner procedure is that of GES, plus the nominal cost incurred by the modified completion algorithm and score. The worst-case complexity of GES is polynomial in the number of variables and exponential in some measures of the underlying graph, such as the maximum number of parents (Chickering and Meek, 2015). However, both the experiments carried out in this and the original GES paper (Chickering, 2002) suggest that, in practice, GES is much more efficient than suggested by this worst-case complexity.

The greedy outer procedure of the algorithm, as described in Algorithm 2, incurs a worst-case complexity of $\mathcal{O}(p^2 \times \text{cost of GES})$. As a faster alternative, we propose a "ranking-based" outer procedure whose complexity is instead $\mathcal{O}(p \times \text{cost of GES})$. Under this approach, we first run the inner procedure with interventions on all variables and obtain an ordering of variables based on the variance of their noise-term variance estimates. We then add (remove) targets in increasing (decreasing) order until the score does not improve. At the price of a slightly reduced performance in terms of edge recovery, the ranking procedure yields a substantial speed-up (see Figure 13 in Appendix I).

The computation of the GnIES score for interventional data is slightly more involved than for the observational equivalent used in GES, as it requires an alternating optimization routine to find the maximum likelihood estimate in (7). However, as in GES, we exploit the decomposability of the score for substantial computational savings. Since the contribution of a node to the score remains constant if its parents do not change, we can cache these "local scores" and avoid many unnecessary computations across the steps of the inner procedure of GnIES, resulting in a dramatic speed-up. Furthermore, the cache can be preserved across steps of the outer procedure, as only the local scores of nodes that are added to or removed from the intervention targets need to be recomputed. This results in further savings.

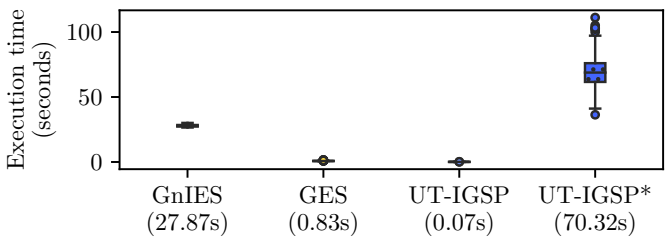


Figure 3: Execution times of GnIES and additional baselines on the light-tunnel dataset from Section 5.3 ($\mathcal{I}^\star = \{\tilde{\theta}_1, \tilde{\theta}_2\}$). We report the execution times over 100 bootstrap samples of the data, and show the average time in parenthesis under each method. The regularization parameters of each method correspond to the ones shown in Figure 6b.

To put the computational cost of GnIES in context with other methods, we compare it in Figure 3 to the baselines used in the real-data setting of Section 5.3: GES (Chickering, 2002), UT-IGSP (Squires et al., 2020b) with partial correlation tests, and UT-IGSP with kernel-based tests (Zhang et al., 2011), denoted by UT-IGSP*. Due to the additional speedups described above, GnIES is about 30 times slower than GES, which is below the time expected by the worst-case complexity given for $p = 10$ in this setting. While the method is faster than UT-IGSP with kernel-based tests, it is much slower than the efficient implementation of UT-IGSP for partial correlation tests (Squires et al., 2020a).

5. Experimental Results

Throughout this section, we evaluate the performance of GnIES across a variety of settings[4]. These include synthetic data (Section 5.1) and two real-data settings consisting of biological data (Section 5.2) and data from a controlled physical system (Section 5.3).

The estimates produced by GnIES and some of the baseline methods will be a set of graphs, whose closeness with the ground-truth graph or equivalence class needs to be measured. We use the metrics introduced in Taeb et al. (2024) for this purpose, which we restate here:

Definition 4 (Metrics) *Let $\mathbb{G}^\star$ be the true $\mathcal{I}-$*MEC *and $\hat{\mathbb{G}}$ its estimate. We define the* true (false) discovery proportion *respectively as*

$$\text{TDP} : \max_{G^\star \in \mathbb{G}^\star} \min_{\hat{G} \in \hat{\mathbb{G}}} \big| \textit{edges in } G^\star \textit{ and } \hat{G} \big| / \big| \textit{edges in } G^\star \big|,$$

$$\text{FDP} : \max_{\hat{G} \in \hat{\mathbb{G}}} \min_{G^\star \in \mathbb{G}^\star} \big| \textit{edges in } \hat{G} \textit{ not in } G^\star \big| / \big| \textit{edges in } \hat{G} \big|.$$

In the case where both the estimate and the truth are singletons, the TDP and FDP correspond to the true and false positive rates in terms of edges. The estimate and truth are equal if and only if FDP $= 0$ and TDP $= 1$.

5.1 Synthetic Data

We evaluate the performance of GnIES on simulated data and compare it to other algorithms that can be applied to the same problem setting. Additional results for variations of the generation settings and a misspecified intervention model are given in Appendix H.

Data generation. For the results shown in Figure 4, the data is sampled from 100 randomly generated linear Gaussian SCMs and intervention targets, following the model in Equation 1. The underlying DAGs are Erdös-Renyi graphs with $p = 10$ nodes and an average degree of 2.7. We provide results for denser graphs in Appendix H. The edge weights are sampled uniformly at random from $[0.5, 1]$ to bound them away from zero; the variances of the noise terms are sampled from $[1, 2]$ and their means are set to zero. We generate 10 different data sets from each SCM consisting of samples of the same observational and interventional distributions, with a sample size of $n = 10, 100$, or 1000 observations. Each interventional distribution stems from an environment where a single variable is intervened on at random, with targets being different across environments. In Appendix H, we show the results for multiple-target interventions. For Figure 4, the interventions satisfy the modeling assumptions, that is, they change the noise term distribution of the target by setting its variance to a value sampled uniformly at random from $[3, 4]$. In Appendix H, we provide results for a misspecified intervention model where the interventions are instead *do-* or hard interventions (Pearl, 2009b, Section 3.2.2).

While having access to an observational sample is not a requirement for GnIES, it is needed for a fair comparison with the other baselines. In particular, an observational environment is required by UT-IGSP; furthermore, it implies that the family of targets is *conservative*, an assumption required by GIES (Hauser and Bühlmann, 2012, Definition 6).

4. The code to reproduce the experiments can be found at `github.com/juangamella/gnies-paper`.

Varsortability. As shown by Reisach et al. (2021), synthetic data as generated above incurs a high degree of *varsortability*, that is, that marginal variance tends to increase along the causal ordering of the data-generating graph. This fact can be exploited for causal discovery, and the authors find that some recent continuous-optimization methods (Zheng et al., 2018; Ng et al., 2020) inadvertently rely on this property of synthetic data. However, this reliance means that their synthetic performance may not carry over to real-world data, where varsortability may be moderate and depends on the measurement scale; indeed, their performance broke down after standardization of the data, and when applied to the real data set from Sachs et al. (2005). To check to what extent the performance of GnIES depends on varsortability, we perform our experiments on both the standardized (Figures 4, 11) and raw (Figure 14, Appendix I) synthetic data, as advised by Reisach et al. (2021). As a sanity check, we additionally run the *sortnregress* algorithm proposed by Reisach et al. (2021) on the data pooled across environments. The method shows remarkable performance in recovering the data-generating graph when there is a high varsortability in the data, and breaks down when there is not. We find that neither GnIES nor the other algorithms employed in the experiments show a significant change in performance after standardization of the data (see Figure 14 in Appendix I), an indication that they do not exploit varsortability in their inference.

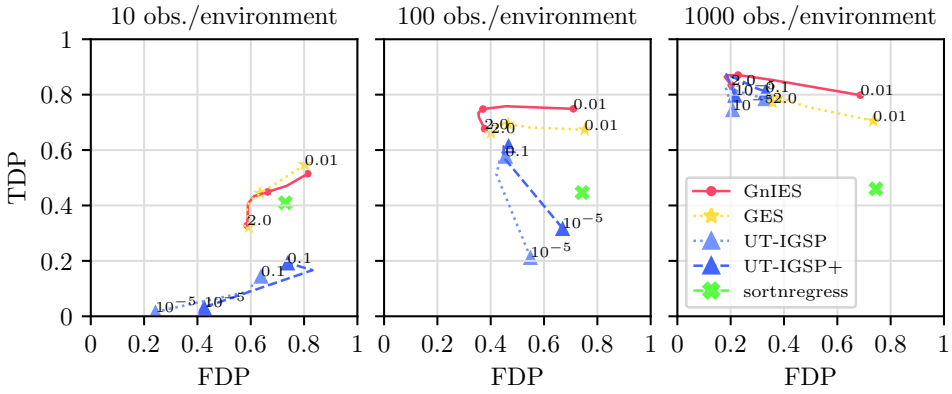


Figure 4: (correctly specified model) Recovery of the true equivalence class for different sample sizes, as measured by the metrics described in Definition 4; the point $(0, 1)$ corresponds to perfect recovery. The results are averaged over 1000 data sets originating from 100 linear Gaussian SCMs following the model in Equation 1. Because UT-IGSP returns a single DAG as estimate, we use its estimated targets and the completion algorithm of GnIES to produce an estimate of the equivalence class (displayed as UT-IGSP+). For each method, the regularization path is shown, together with the smallest and largest values of the regularizers; the additional point shown on the regularization path for GnIES and GES corresponds to the BIC score.

Due to its flexibility in the assumptions placed on the data-generating process, we employ UT-IGSP (Squires et al., 2020b) as a baseline in all our experiments. An extension of the

greedy sparse permutation algorithm (Solus et al., 2021), UT-IGSP works by measuring the conditional independencies and invariances in the data through statistical hypothesis tests, and finds the sparsest DAG and intervention targets that satisfy these constraints. The choice of tests can vary according to the setting; for the synthetic experiments of this section, we use standard partial correlation independence and invariance tests provided by the authors in their implementation of the algorithm. A drawback of the method is that it returns a single graph as estimate, when in practice there may be competing equivalent graphs, as happens in the setting that we study here. To ensure a fair comparison, we take the graph and targets estimated by UT-IGSP and compute the corresponding $\mathcal{I}$-equivalence class with the GnIES completion algorithm (see Appendix E).

As a baseline for the setting where the model assumptions are satisfied (Figure 4), we additionally run GES (Chickering, 2002) on the data pooled across environments. The motivation is that this could be an initial approach employed by a practitioner to recover the class of equivalent models. Its performance suggests that this would not be entirely misguided, particularly at the smallest sample size. At this size, a possible explanation for the similar performance of GnIES and GES is that the effect of the interventions is less pronounced; due to its penalization, GnIES includes fewer intervention targets in its estimate and thus searches over classes that are often observational, essentially solving the same problem as GES. The poor performance of UT-IGSP in this sample size is not due to the diminished effect of interventions, but rather its reliance on conditional independence and invariance tests, which perform better with larger samples. Indeed, for $n = 10$, in about 9% of the runs the correlation matrix computed internally by UT-IGSP becomes singular, and the method fails to produce an estimate; we exclude these cases when computing the average results in the figures. As expected of the setting where all its assumptions are satisfied, GnIES performs competitively in recovering the true equivalence class and does so better than GES due to the additional identifiability.

For the results shown in Figure 4 and the additional figures in Appendix H, we evaluate UT-IGSP at different levels of the conditional independence and invariance tests, taken from the range $[10^{-5}, 0.1]$. We evaluate GnIES and GES with different values of the regularization parameter $\lambda := \lambda' \log(N)$, where N is the total number of observations and λ' is taken from the range $[0.01, 2]$.

5.2 Biological and Semi-synthetic Data

We now consider the protein expression data set[5] from Sachs et al. (2005). It consists of measurements of the abundance of 11 phosphoproteins and phospholipids in human immune cells, recorded via a multiparameter flow cytometer under different experimental conditions. These conditions consist of adding reagents to the cell medium (Sachs et al., 2005, table 1), which inhibit or activate different nodes of the protein signaling network. As has been done in previous works (Squires et al., 2020b; Meinshausen et al., 2016; Mooij and Heskes, 2013), we take the observations under different conditions as originating from different interventional distributions, resulting in a data set with a total of 7466 observations across nine interventional samples, each ranging from 707 to 913 observations. Furthermore, because

5. The data set was downloaded from `science.org/doi/suppl/10.1126/science.1105809/suppl_file/sachs.som.datasets.zip`.

the observations correspond to independent measurements of individual cells, we assume them to be independent and identically distributed. Since GnIES and the other baselines aim to estimate the causal structure underlying the observations, we require a ground truth to evaluate their performance. To this end, we consider the "consensus network" presented in (Sachs et al., 2005, Figure 2).

The data set is an interesting challenge because of an array of violations of the GnIES modeling assumptions. The functional relationships between variables appear to be non-linear and the distributional assumption of normality is strongly violated (see Figures 7 and 8 in Appendix G). Furthermore, the conventionally accepted structure of the protein signaling network contains cycles (see nodes PIP2, PIP3 and PLCγ in Sachs et al., 2005, Figure 2) and, as for any measurements of real-world systems, the presence of unmeasured confounders cannot be discarded.

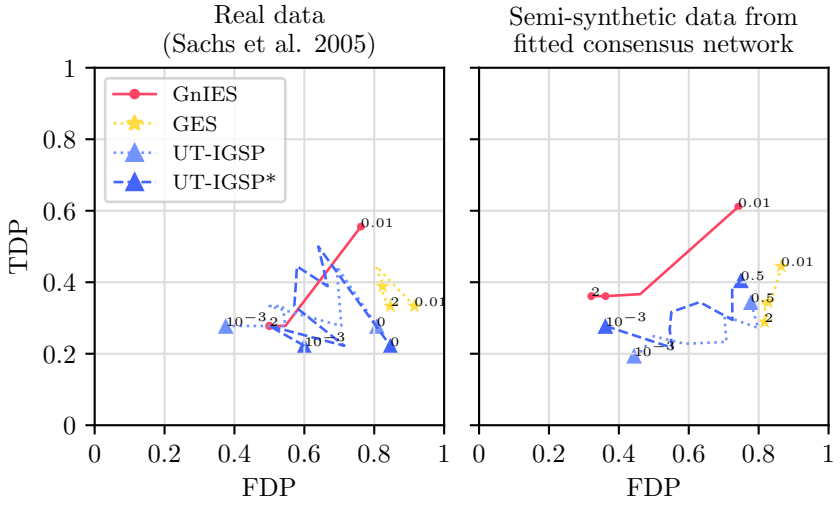


Figure 5: (Sachs data set) Recovery of the "consensus network" from the original Sachs data (left) and data sampled from a non-parametric SCM fitted to the original data with the consensus network structure (right). For the semi-synthetic data, the results are the average over 10 different samples from the same fitted interventional distributions. UT-IGSP corresponds to the method run with partial correlation conditional independence and invariance tests, and UT-IGSP* to the method with kernel-based tests (Zhang et al., 2011). Because GnIES and UT-IGSP estimate interventions on all variables, their class estimates are a singleton; thus, the TDP and FDP metrics for these methods correspond to the true and false positive rates in terms of directed edge recovery. GES is run on the pooled data across environments and returns the Markov equivalence class estimated from the data. For each method, the regularization path is shown. For GnIES and GES, the additional point shown on the regularization path corresponds to the BIC score.

Evaluation on semi-synthetic data. Another issue is that the consensus network, which is constructed from effects that appear at least once in the literature (Sachs et al., 2005, table S1), may not be a complete or accurate description of the causal relationships underlying the observations of the protein signaling network. As a safeguard, and to better understand the effect that the different model violations have on their performance, we evaluate the methods on both the original data set and semi-synthetic data sampled from a non-parametric Bayesian network, fitted to each interventional sample according to the structure of the consensus network[6]. The sample sizes match those in the real data set. The motivation is the following: the marginal and conditional distributions of the semi-synthetic data closely resemble those of the real data and preserve the linearity and Gaussianity violations (Figures 7 and 8, Appendix G). However, it respects acyclicity, causal sufficiency (the absence of hidden confounders), and the conditional independence relations implied by the given "ground-truth" graph. Thus, we can evaluate GnIES and the other baselines on two data sets of increasing difficulty, and still receive a certain degree of validation in the case of a faulty ground truth. The non-parametric SCM is fitted and sampled from using distributional random forests (Cevid et al., 2022); the procedure is described in detail in Appendix G and can be accessed through the Python package `sempler`.

As for the synthetic experiments, we run GnIES and the other methods over a range of their regularization parameters. Again, we run GES on the observations pooled across all samples. For UT-IGSP, we run the method with the same Gaussian tests as in the synthetic experiments and kernel-based tests (Zhang et al., 2011) provided in the implementation of the algorithm; we respectively label each approach as UT-IGSP and UT-IGSP*. Due to the nature of the experiments, it is not straightforward to select one as the observational environment for UT-IGSP. For a fair comparison, we compare the performance over every option and display the best result in Figure 5; the others can be found in Figure 12 of Appendix I.

Taking the consensus network as ground truth, there is no uniformly best performer for the real data. The performance of UT-IGSP seems to depend heavily on the level of its tests. GES performs uniformly worse than the other methods. For the semi-synthetic data and the selected regularization parameters, GnIES performs uniformly better than the other methods. The performance of UT-IGSP with Gaussian tests remains relatively unchanged and decreases with kernel tests. Several competing explanations exist for this change in performance between real and semi-synthetic data. In principle, the potential differences in the data stem from (1) the acyclicity of the data-generating model, (2) the absence of hidden confounders, or (3) a different ground truth. The interplay between these differences and the inner workings of the algorithms is poorly understood at the moment of writing. We acknowledge that without precise knowledge of the causal structure and data collection process which generated the Sachs data set, our statements about this issue are at best speculative.

5.3 Real Data from a Controlled Physical System with Causal Ground Truth

As a real-data setting with a well-justified ground truth, we consider one of the light-tunnel datasets from the causal chambers (Gamella et al., 2025). The light tunnel (Figure 6a)

6. After removing the edge PIP2 $\to$ PIP3 to satisfy the acyclicity constraint.

is an elongated chamber with a controllable light source at one end, two linear polarizers mounted on rotating frames, and sensors to measure different physical variables such as the light intensity at different locations of the tunnel. A detailed diagram of the tunnel can be found in Gamella et al. (2025, Figure 2). As a ground truth for causal inference tasks, the authors provide a directed graph summarizing the causal effects in the system (Gamella et al., 2025, Figure 3a), built from background knowledge and further validated using randomized experiments.

The dataset[7] contains data from experiments in which the light-source colors (R, G, B) and polarizer positions (θ_1, θ_2) are sampled independently from different distributions, and measurements are collected under different configurations of the sensor parameters (e.g., oversampling rate, measurement duration, etc.). For our validation task, we also consider the measurements of the polarizer angles $(\tilde{\theta}_1, \tilde{\theta}_2)$, the current drawn by the light source $(\tilde{C})$, and the infrared-light intensity before $(\tilde{I}_1)$, between $(\tilde{I}_2)$ and after the polarizers $(\tilde{I}_3)$. As observational data, we take observations from the dataset's reference experiment. As interventional data, we consider two types of interventions: shifts in the distributions of the polarizer positions θ_1 and θ_2, and changes in the reference voltages of the angle sensors $\tilde{\theta}_1$ and $\tilde{\theta}_2$. When compared to the reference environment, the former results in the relative angle between the polarizers being close to 90°, dimming the light reaching the third sensor and effectively scaling the effect that the light-source colors (R, G, B) have on its reading $\tilde{I}_3$ (Figure 6d). Reducing the reference voltage of the angle sensors increases their sensitivity, affecting the variance of the resulting measurements $\tilde{\theta}_1$ and $\tilde{\theta}_2$ (Figure 6c). For more details about these effects, we refer the reader to Gamella et al. (2025, Appendix IV.2.1 and Figure 14, respectively).

We run GES, GnIES, and UT-IGSP on the light tunnel data and evaluate their recovery of the ground-truth graph (Figure 6b) for different values of their hyperparameters (Figure 6e). The effects of the light-source colors (R, G, B) on the drawn current $(\tilde{C})$ and light intensities $(\tilde{I}_1, \tilde{I}_2, \tilde{I}_3)$ are approximately linear and are reliably recovered by all three methods from the observational data alone ($\mathcal{I}^\star = \emptyset$ in Figure 6). However, both GES and GnIES fail to orient the edges between the polarizer positions (θ_1, θ_2) and the corresponding angle measurements $(\tilde{\theta}_1, \tilde{\theta}_2)$. As predicted by the identifiability results in Section 3, including data from the interventions on the reference voltages allows GnIES to correctly orient these edges and reliably recover the ground-truth graph ($\mathcal{I}^\star = \{\tilde{\theta}_1, \tilde{\theta}_2\}$ in Figure 6b). On the other hand, including data from the interventions on the polarizer positions ($\mathcal{I}^\star = \{\theta_1, \theta_2\}$) has a catastrophic effect on the performance of GnIES. While the interventions on θ_1 and θ_2 should also allow the method to orient the edges around these variables, the effect of these interventions—which scale the effect (coefficients) of R, G, B on $\tilde{I}_3$ (Figure 6d)—cannot be expressed by the GnIES model (1). As a result, the method reverses many of the edges in its initial estimate to find an optimal scoring structure, yielding poor results. UT-IGSP is not affected by the non-linear effect of these interventions, even when partial correlation tests are used (UT-IGSP). However, its performance is virtually unchanged when interventional data is added. The results are similar when the method is run with kernel-based tests (Zhang et al., 2011), shown in Figure 6 as UT-IGSP*.

7. The data and its documentation can be found under the `lt_interventions_standard_v1` dataset at `github.com/juangamella/causal-chamber`.

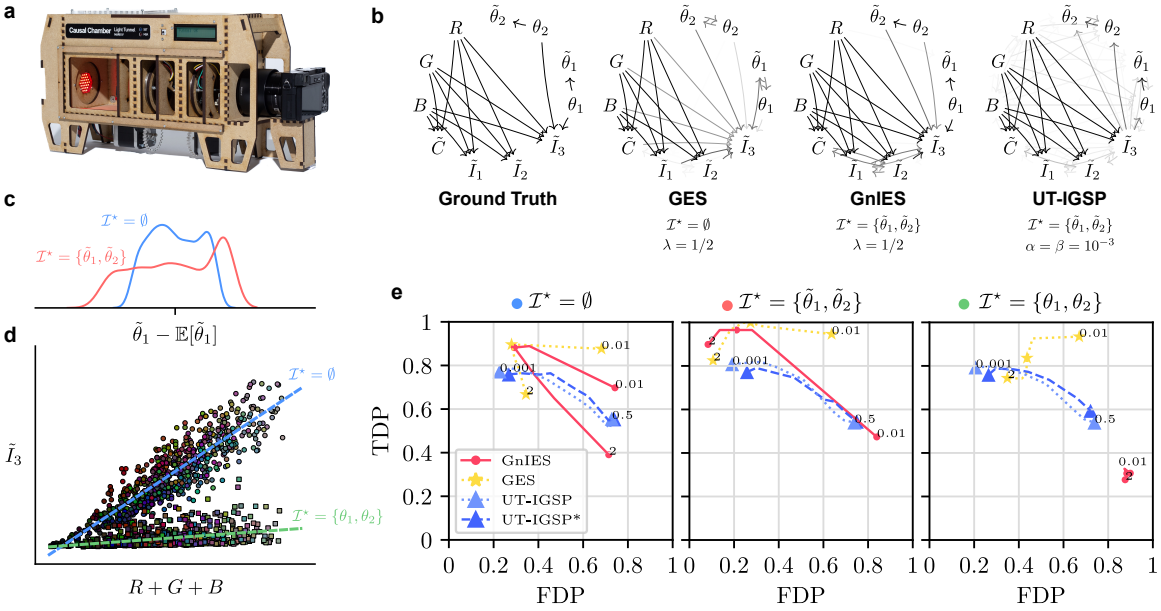


Figure 6: (Causal chamber dataset) (a) The light tunnel. (b) The ground-truth graph and a visualization of the estimates produced by each method for 100 bootstrap samples of the data. The color of each edge denotes the frequency with which it appears in the method's estimates, with darker edges appearing at a higher frequency. For GES and GnIES, which return a set of graphs, we pool all graphs to compute the edge frequencies. (c) Effect of the interventions on $\tilde{\theta}_1, \tilde{\theta}_2$: plot of the (centered) marginal density of $\tilde{\theta}_1$ in the reference experiment (blue) and the experiment where we lower the reference voltage of the corresponding sensor (red). (d) Effect of the interventions on the polarizer positions θ_1, θ_2: the change in the relative angle between the polarizers reduces the amount of light reaching the light sensor $\tilde{I}_3$, scaling the effect of the light-source colors R, G, B. (e) Recovery of the ground-truth graph as measured by the metrics in Definition 4. For each method, we show the regularization path averaged over 100 bootstrap samples of the data.

6. Discussion

This paper considers a parametric approach to the problem of causal structure discovery from interventional data when the targets of interventions are unknown. For linear SCMs with additive Gaussian noise, we provide identifiability results under noise interventions and use these to develop a greedy algorithm to recover the equivalence class of the data-generating model. We provide an easy-to-use Python implementation in the package `gnies`. A summary of our software contributions can be found in Appendix F.

When evaluated on synthetic data, the algorithm performs competitively with other algorithms built for the same setting, particularly when the available samples are small. We tested our algorithm on real data from one of the controlled physical systems described in Gamella et al. (2025); the results show that recovery improves with interventional data,

following the theoretical results in Section 3. We also evaluated the algorithm on the biological data set from Sachs et al. (2005) and found its performance in recovering the "consensus network" relatively poor, but in line with other methods despite the stronger distributional assumptions. To tease apart the issues underlying this result, we developed a semi-synthetic data generation procedure, which we believe to be of general interest to the causal discovery community. The procedure allows generating semi-synthetic data sets with known causal ground truth but distributions resembling those in real data sets, providing a more realistic challenge than purely synthetic data. We detail the procedure in Appendix G and make it publicly available through the Python package `sempler`.

6.1 Future Work

While we have employed $\mathcal{I}$-equivalence to characterize sets of distribution equivalent Gaussian models under noise interventions, the notion is an instance of *transition sequence equivalence*, which was introduced by Tian and Pearl (2001, Theorems 2 and 3). As such, it applies beyond the Gaussian models and noise interventions we consider here. This suggests that an extension of our algorithm to more general models is justified and may be possible by appropriately adapting its score. Within the realm of Gaussian models, one could instead consider other intervention types still covered by the framework of Tian and Pearl (2001). In particular, these include interventions that modify the parental coefficients of their target, as we pointed out in Section 2.2. Given that the graphical characterization would remain the same, GnIES could be extended to this setting by allowing the coefficients of a target to vary between environments. However, we leave this task and deriving the necessary assumptions for the equivalent of Theorem 1 as future work.

The similarity between GES and the inner procedure of GnIES suggests another line of work, based on successful extensions of GES. Under certain assumptions, constraining the maximum degree of nodes in the GES search space yielded significant improvements in sample efficiency (Chickering, 2020); a hybrid approach, using background knowledge or conditional independence tests to restrict the available operators in each step adaptively, has led to a successful scaling of GES to high-dimensional settings with thousands of variables (Nandy et al., 2018). Adapting these constraints to the $\mathcal{I}$-equivalence class may produce similar improvements for GnIES.

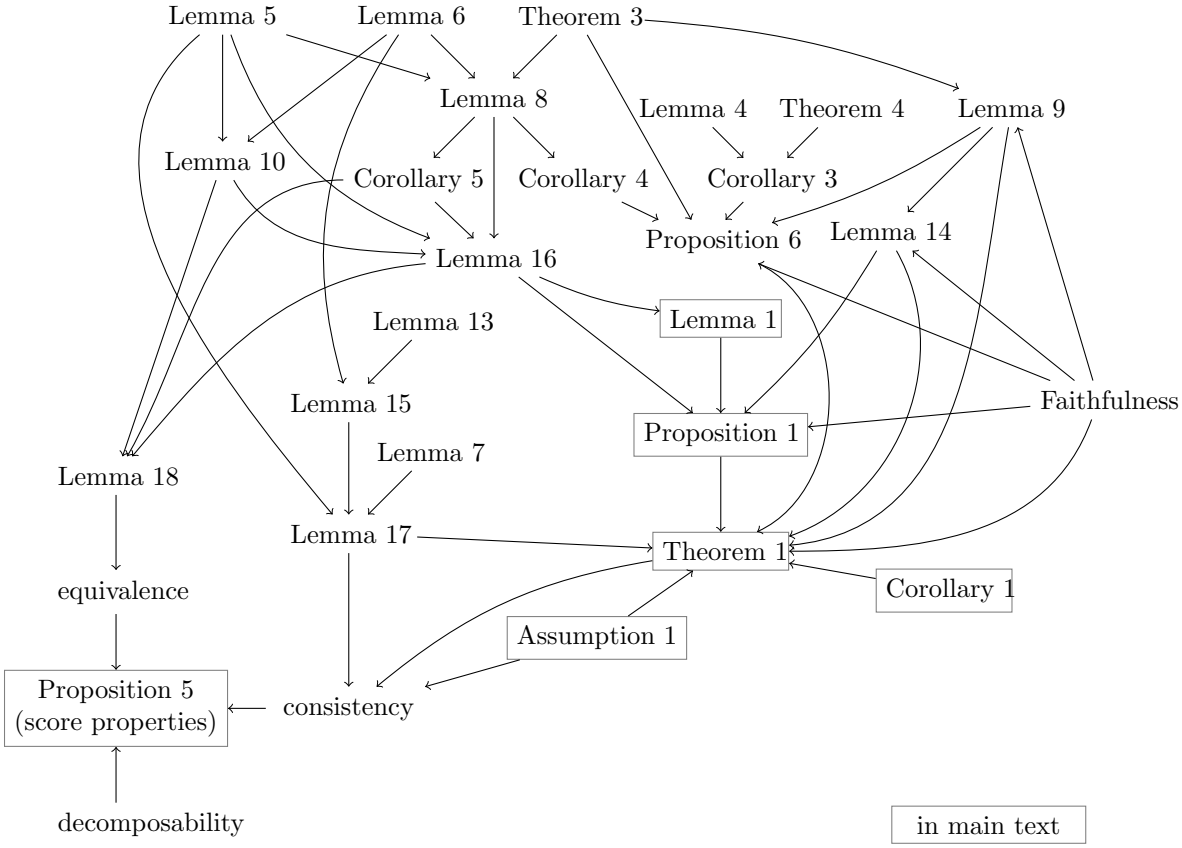
Acknowledgments

We would like to thank Jonas Peters, Jinzhou Li, and Malte Londschien for their valuable discussions and comments on the manuscript. J.L. Gamella, A. Taeb, and P. Bühlmann have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 786461).

Appendix A. Proofs

This section contains the proofs for all results presented in the main text of the paper. An emphasis has been placed on readability and completeness, arguably at the expense of brevity. To aid in the reading of the proofs, we summarize here the notation used and provide a graph of how the results relate to each other.

Proof map. How the presented results relate to each other. An edge $i \to j$ means that result i is used to construct the proof of result j.



A.1 Restated Results

We restate results from Chickering (1995, 2002) which are used in the proofs of this section. The notation is slightly adapted to fit ours.

Definition 5 (Covered edge, from Chickering, 2002) *For any DAG G, we say an edge $j \to i$ in G is covered if*

$$\mathrm{PA}_G(i) = \mathrm{PA}_G(j) \cup \{j\}.$$

Definition 6 (Independence Map (I-MAP) from Chickering, 2002) *A DAG H is an independence map of a DAG G if every d-separation statement in H also holds in G, that is, $i \perp\!\!\!\perp_H j \mid S \implies i \perp\!\!\!\perp_G j \mid S$.*

Theorem 3 (Restated Theorem 4 from Chickering, 2002) *Let G and H be any pair of DAGs such that H is an independence map of G. Let r be the number of edges in H that have opposite orientation in G, and let m be the number of edges in H that do not exist in either orientation in G. There exists a sequence of at most $r + 2m$ edge reversals and additions in G with the following properties:*

1. *Each edge reversed is a covered edge.*
2. *After each reversal and addition, G is a DAG and H is an independence map of G.*
3. *After all reversals and additions, $G = H$.*

Theorem 4 (Restated Theorem 2 from Chickering, 1995) *Let G and G' be any pair of DAGs that are Markov equivalent and for which there are δ edges in G that have opposite orientation in G'. Then there exists a sequence of δ distinct edge reversals in G with the following properties:*

1. *Each edge reversed in G is covered.*
2. *After each edge reversal, G is a DAG and is Markov equivalent to G'.*
3. *After all reversals, $G = G'$.*

Lemma 4 (Restated Lemma 1 from Chickering, 1995) *Let G be any DAG model, and let G' be the result of reversing the edge $X \to Y$ in G. Then G' is a DAG that is Markov equivalent to G if and only if $X \to Y$ is covered in G.*

Corollary 3 (Markov equivalence and covered edge reversals) *Let G and G' be two DAGs. From Theorem 4 and Lemma 4, it follows that they are Markov equivalent if and only if there exists a sequence of covered edge reversals that yield G' when applied to G.*

A.2 Supporting Results from Linear Algebra

Lemma 5 *Let $A \in \mathbb{R}^{p \times p}$. The following two statements are equivalent,*

1. *AA^T is diagonal with positive entries, and*
2. *A has full rank and orthogonal row vectors.*

Proof To show that 1 $\implies$ 2, assume AA^T is diagonal with positive entries but

(i) $\text{rank}(A) < p$, or

(ii) $\exists i, j$ such that $i \neq j$ and $\langle A_{i:}, A_{j:} \rangle \neq 0$.

If (i) is true, then $\text{rank}(A) = \text{rank}(AA^T) < p$, which means that $\det(AA^T) = 0$ and thus, since AA^T is diagonal, $\exists i$ such that $A_{ii} = 0$, which is a contradiction. If (ii) is true, then $(AA^T)_{ij} = \langle A_{i:}, A_{j:} \rangle \neq 0$, which is a contradiction.

That 2 $\implies$ 1 is trivial, as for $i \neq j$, $(AA^T)_{ij} = \langle A_{i:}, A_{j:} \rangle = 0$ by definition, and $(AA^T)_{ii} = \langle A_{i:}, A_{i:} \rangle > 0$ as A has full rank and $A_{i:} \neq \vec{0}$ for all i. ■

Lemma 6 *Let $A, B \in \mathbb{R}^{p\times p}$ with B invertible, $\alpha \in \mathbb{R}$ and let $e_i \in \mathbb{R}^p$ be the i^{th} canonical unit vector.*

$$A_{i:} = \alpha B_{i:} \iff (AB^{-1})_{i:} = \alpha e_i^T.$$

Proof " $\implies$ " $(AB^{-1})_{i:} = A_{i:}B^{-1} = \alpha B_{i:}B^{-1} = \alpha e_i^T$.
" $\impliedby$ " $(AB^{-1})_{i:} = \alpha e_i^T \implies A_{i:}B^{-1} = \alpha e_i^T \implies A_{i:} = \alpha e_i^T B = \alpha B_{i:}$. ■

Lemma 7 (Scaling of a matrix with orthogonal rows) *Let $M \in \mathbb{R}^{p\times p}$ be a full-rank matrix with orthogonal row vectors, such that every row has more than one non-zero entry. Let $D \in \mathbb{R}^{p\times p}$ be a diagonal matrix with positive diagonal entries for which $\exists i \in [p]$ such that $D_{ii} \neq D_{jj}$ for all $j \neq i$. Then MD does not have orthogonal row vectors.*

Proof Assume that MD does in fact have orthogonal row vectors, i.e., that $MDDM^T$ is a diagonal matrix. Denote $\tilde{M} := MDD$, and pick[8] k such that $M_{ki} \neq 0$. We have that

$$\langle M_{k:}, M_{j:}\rangle = 0 \text{ and } \langle \tilde{M}_{k:}, M_{j:}\rangle = 0 \text{ for } j \neq k. \tag{8}$$

Since $\{M_{j:}\}_{j\neq k}$ are $p-1$ linearly independent vectors in $\mathbb{R}^p$, (8) implies that $\tilde{M}_{k:}$ and $M_{k:}$ belong to the same one-dimensional subspace. Thus, for some $\alpha \in \mathbb{R}$ we have that

$$\begin{aligned}\tilde{M}_{k:} =& M_{k:}DD = \alpha M_{k:}\\ &\implies M_{kl}D_{ll}^2 = \alpha M_{kl} \text{ for all } l \in [p]\\ &\implies D_{ll}^2 = \alpha \text{ for all } l : M_{kl} \neq 0.\end{aligned} \tag{9}$$

Note that we chose k s.t. $M_{ki} \neq 0$, and because the rows of M contain more than one non-zero entry, $\exists j \neq i$ such that $M_{kj} \neq 0$. Thus, (9) implies that $D_{ii}^2 = \alpha = D_{jj}^2$. Since the diagonal entries of D are positive, the above means that $D_{ii} = D_{jj}$, arriving at a contradiction with our constraints on D. Thus, MD cannot have orthogonal row vectors, completing the proof. ■

A.3 Supporting Results for the Observational Case ($|\mathcal{E}| = 1$)

While the results of this subsection are known in the literature, for completeness we prove them again here for the models we consider.

Lemma 8 (I-MAPs are distribution equivalent) *Let (B, Ω) be a model with a single environment, which entails a distribution $P \sim \mathcal{N}(0, \Sigma)$ with $\Sigma := (I-B)^{-1}\Omega(I-B)^{-T}$. Let G be an independence map of $\mathcal{G}(B)$, that is, all the d-separation statements in G hold also in $\mathcal{G}(B)$. Then,*

$$\exists \text{ a model } (B', \Omega') \text{ s.t. } B' \sim G \text{ and } \Sigma = (I-B')^{-1}\Omega'(I-B')^{-T}.$$

8. Such k exists as M has full rank.

Proof

Case I: The edges in $\mathcal{G}(B)$ are a subset of the edges in G. Then, trivially $B' := B$ and $\Omega' := \Omega$, completing the proof.

Case II: Let the edges in $\mathcal{G}(B)$ be a subset of the edges in G, except for a single edge $j \to i$ which has been reversed to $j \leftarrow i$ in G. Furthermore, assume that $j \leftarrow i$ is covered in G. This means that

$$\mathrm{PA}_{\mathcal{G}(B)}(k) \subseteq \mathrm{PA}_G(k) \text{ for all } k \notin \{i, j\}, \tag{10}$$

$$\mathrm{PA}_{\mathcal{G}(B)}(i) \setminus \{j\} \subseteq \mathrm{PA}_G(i), \tag{11}$$

$$\mathrm{PA}_{\mathcal{G}(B)}(j) \cup \{i\} \subseteq \mathrm{PA}_G(j), \text{ and} \tag{12}$$

$$\mathrm{PA}_G(i) \cup \{i\} = \mathrm{PA}_G(j). \tag{13}$$

We proceed to construct a connectivity matrix B' and noise term covariance Ω' such that the result holds. First, set all but the i^{th} and j^{th} rows of B' to be the same as in B, i.e., $B'_{k:} = B_{k:}\ \forall k \notin \{i, j\}$. We construct the remaining i^{th} and j^{th} rows of B' by the following rule:

$$(I - B')_{i:} := \beta^i \Omega^{-1/2}(I - B) \text{ and } (I - B')_{j:} := \beta^j \Omega^{-1/2}(I - B) \tag{14}$$

for some $\beta^i, \beta^j \in \mathbb{R}^p$ such that

i) $\beta^i_k = \beta^j_k = 0$ for all $k \notin \{i, j\}$,

ii) $[\beta^i \Omega^{-1/2}(I - B)]_i = 1$ and $[\beta^j \Omega^{-1/2}(I - B)]_j = 1$.

Additionally, we impose the constraints that (iii) $\beta^i \beta^{j^T} = 0$ and (iv) $(I - B')_{ij} = [\beta^i \Omega^{-1/2}(I - B)]_j = 0$; their purpose will become clear later. We can express all the constraints in matrix form. For the vector β^i we have

$$\beta^i C^i = (0, ..., 0, 1) \text{ where } C^i := \left[\{e_k\}_{k \notin \{i,j\}}, \Omega^{-1/2}(I - B)_{:j}, \Omega^{-1/2}(I - B)_{:i}\right], \tag{15}$$

and for β^j

$$\beta^j C^j = (0, ..., 0, 1) \text{ where } C^j := \left[\{e_k\}_{k \notin \{i,j\}}, \beta^{i^T}, \Omega^{-1/2}(I - B)_{:j}\right], \tag{16}$$

where e^k is the k^{th} canonical unit vector[9]. Up to a permutation of columns, C^i is lower triangular with non-zero diagonal entries, and thus it has full rank; therefore β^i, as a solution to (15), exists and is unique. Plugging it into (16) yields a full rank matrix; to see this, consider that the last two columns are linearly independent from the first $p - 2$ columns[10]. The last two columns are orthogonal to each other by construction (iv). Thus, β^j also exists and is unique. Now we show that the resulting B' satisfies the requirements.

9. That is, $e^k_k = 1$ and $e^k_i = 0\ \forall i \neq k$.
10. They are unit vectors with zero i^{th} and jth entries, whereas β^i and $(I - B)_{:j}$ have non-zero elements in these positions.

B' is compatible with the graph G, i.e., $B' \sim G$. We proceed by showing that for each node l, $\text{supp}(B'_{l:}) \subseteq \text{PA}_G(l)$, which implies $B' \sim G$. Since $B'_{k:} = B_{k:}\ \forall k \notin \{i,j\}$, by (10) it holds that $\text{supp}(B'_{k:}) = \text{supp}(B_{k:}) = \text{PA}_{\mathcal{G}(B)}(k) \subseteq \text{PA}_G(k)$ for all $k \notin \{i,j\}$. Now we show that it also holds for i and j. By (14), we have that $B'_{i:} = \beta^i \Omega^{-1/2} B - \beta^i \Omega^{-1/2} - I_{i:}$, subject to the constraints that $B'_{ii} = 0$ (ii) and $B'_{ij} = 0$ (iv). Thus,

$$
\begin{aligned}
\text{supp}(B'_{i:}) &= \Big(\text{supp}(\beta^i \Omega^{-1/2} B) \cup \text{supp}(\beta^i \Omega^{-1/2}) \cup \text{supp}(I_{i:})\Big) \setminus \{i,j\} \\
\textit{by constraint (i)} \quad &\subseteq (\text{supp}(B_{j:}) \cup \text{supp}(B_{i:}) \cup \{i,j\} \cup \{i\}) \setminus \{i,j\} \\
&= \text{supp}(B_{j:}) \setminus \{i,j\} \cup \text{supp}(B_{i:}) \setminus \{i,j\} \\
&= \text{PA}_{\mathcal{G}(B)}(j) \setminus \{i\} \cup \text{PA}_{\mathcal{G}(B)}(i) \setminus \{j\} \\
\textit{by (11, 12)} \quad &\subseteq \text{PA}_G(j) \setminus \{i\} \cup \text{PA}_G(i) \\
\textit{by (13)} \quad &= \text{PA}_G(i) \cup \text{PA}_G(i) \\
&= \text{PA}_G(i).
\end{aligned}
$$

Similarly, we have that $B'_{j:} = \beta^j \Omega^{-1/2} B - \beta^j \Omega^{-1/2} - I_{j:}$, subject to $B'_{jj} = 0$ (ii), and thus

$$
\begin{aligned}
\text{supp}(B'_{j:}) &= \Big(\text{supp}(\beta^j \Omega^{-1/2} B) \cup \text{supp}(\beta^j \Omega^{-1/2}) \cup \text{supp}(I_{j:})\Big) \setminus \{j\} \\
\textit{by constraint (i)} \quad &\subseteq (\text{supp}(B_{j:}) \cup \text{supp}(B_{i:}) \cup \{i,j\} \cup \{i\}) \setminus \{j\} \\
&= (\text{supp}(B_{j:}) \cup \{i\} \cup \text{supp}(B_{i:})) \setminus \{j\} \\
&= \big(\text{PA}_{\mathcal{G}(B)}(j) \cup \{i\} \cup \text{PA}_{\mathcal{G}(B)}(i)\big) \setminus \{j\} \\
&= \text{PA}_{\mathcal{G}(B)}(j) \cup \{i\} \cup \text{PA}_{\mathcal{G}(B)}(i) \setminus \{j\} \\
\textit{by (11, 12)} \quad &\subseteq \text{PA}_G(j) \cup \text{PA}_G(i) \\
\textit{by (13)} \quad &= \text{PA}_G(j).
\end{aligned}
$$

(B', Ω') is a model such that $\Sigma = (I - B')^{-1} \Omega' (I - B')^{-T}$. Since $\mathcal{G}(B')$ corresponds to a DAG[11], it holds that B' is lower triangular up to a permutation of rows and columns and has zero diagonal entries. Now we look at the noise term variances; let $\Omega' := (I - B')(I - B)^{-1} \Omega (I - B)^{-T} (I - B')^T$. By Lemma 5, Ω' is a diagonal matrix with positive entries if and only if $(I - B')(I - B)^{-1} \Omega^{1/2}$ has full rank and orthogonal row vectors. The first condition holds as it is the product of three invertible matrices. For the orthogonality of the row vectors, note that because $B'_{k:} = B_{k:}\ \forall k \notin \{i,j\}$, by Lemma 6 we have that $[(I - B')(I - B)^{-1}]_{k:} = e_k^T$, and thus these rows are orthogonal to each other. Furthermore, since by construction $[(I - B')(I - B)^{-1} \Omega^{1/2}]_{ik} = \beta^i_k = 0$ and $[(I - B')(I - B)^{-1} \Omega^{1/2}]_{jk} = \beta^j_k = 0\ \forall k \notin \{i,j\}$, they are also orthogonal to the i^{th} and j^{th} rows. These two rows are also orthogonal to one another, as by constraint (iv) we have $\beta^i \beta^{j^T} = 0$. Thus (B', Ω') is a valid model for which $\Sigma = (I - B')^{-1} \Omega' (I - B')^{-T}$. This completes the proof for case II.

Case III: Suppose now G is any independence map of $\mathcal{G}(B)$. By Theorem 3, there exists a sequence of graphs $(\mathcal{G}(B) = G_1, G_2, \ldots, G_k = G)$, where consecutive graphs differ in

11. $\mathcal{G}(B')$ has its edges contained in the DAG G, and is thus a DAG.

the reversal of a single covered edge or the addition of a single edge to G_i. We let $B^1 := B$, and proceeding iteratively, for each matrix B^i and graph G_{i+1} we apply the result of case I or II to obtain a new model $(B^{i+1}, \Omega^{i+1}) \in [(B, \Omega)]$ with $B^{i+1} \sim G_{i+1}$. At the end of this process, we can set $B' := B^k$ and $\Omega' := \Omega^k$, completing the proof. ■

Corollary 4 (Markov equivalence implies distributional equivalence) *Let (B, Ω) be a model with a single environment. It follows from Lemma 8 that*

$$G \in \mathit{MEC}(\mathcal{G}(B)) \implies G \in \mathbb{G}(B, \Omega).$$

Corollary 5 (Same coefficients) *From the construction of B' in the proof of Lemma 8, it follows that $B_{i:} = B'_{i:}$ for all i such that $\mathrm{PA}_{\mathcal{G}(B)}(i) \subseteq \mathrm{PA}_G(i)$.*

Lemma 9 (Under faithfulness, the true model has the minimal number of edges) *Let (B, Ω) be a model with a single environment, which entails a distribution $P \sim \mathcal{N}(0, \Sigma)$ with $\Sigma := (I - B)^{-1}\Omega(I - B)^{-T}$. Let $\mathcal{G}(B)$ be the corresponding graph and assume that P is faithful with respect to it. Then it holds that*

$$\min_{G \in \mathbb{G}(B,\Omega)} \|G\|_0 = \|\mathcal{G}(B)\|_0.$$

Proof Let $\mathcal{C}(P)$ denote the set of conditional independence relationships in the distribution P, and let d-sep(G) denote the d-separation statements entailed by some graph G.

Take any $G \in \mathbb{G}(B, \Omega)$. There exist (B', Ω') such that $B' \sim G$ and $\Sigma = (I - B')^{-1}\Omega'(I - B')^{-T}$; in other words, there exists a linear Gaussian SCM with connectivity B', noise term variances Ω' and underlying graph $\mathcal{G}(B')$ which entails the distribution P. As such, this distribution P is Markov with respect to $\mathcal{G}(B')$; because G contains all edges in $\mathcal{G}(B')$, the distribution is also Markov wrt. G, i.e., all d-separation relations in the graph are matched by a conditional independence relationship in P. Because P is faithful with respect to $\mathcal{G}(B)$, it follows that d-sep$(G) \subseteq$ d-sep$(\mathcal{G}(B)) = \mathcal{C}(P)$. As such, G is an independence map of $\mathcal{G}(B)$, and by Theorem 3 there exists a sequence of edge reversals and additions that yield G when applied to $\mathcal{G}(B)$. It follows that G cannot have fewer edges than $\mathcal{G}(B)$. ■

Proposition 6 (Markov and minimal-edge distributional equivalence) *Let (B, Ω) be a model as in (1) with a single environment, which entails a distribution $P \sim \mathcal{N}(0, \Sigma)$ with $\Sigma := (I - B)^{-1}\Omega(I - B)^{-T}$. Let $\mathcal{G}(B)$ be the corresponding graph and assume that P is faithful with respect to it. Then, it holds that*

$$\mathit{MEC}(\mathcal{G}(B)) = \arg \min_{G \in \mathbb{G}(B,\Omega)} \|G\|_0,$$

where $\|G\|_0$ denotes the number of edges in G.

Proof We show that both sets contain each other.

$\subseteq$ Let G' be a Markov equivalent graph to $\mathcal{G}(B)$. By Corollary 4 there exists (B', Ω') in $[(B, \Omega)]$ such that $B' \sim G'$, and thus $G' \in \mathbb{G}(B, \Omega)$. By Lemma 9 we have that $\min_{G \in \mathbb{G}(B,\Omega)} \|G\|_0 = \|\mathcal{G}(B)\|_0$; since $\mathcal{G}(B)$ and G' are Markov equivalent, they have the same number of edges, and thus $G' \in \arg\min_{G \in \mathbb{G}(B,\Omega)} \|G\|_0$.

$\supseteq$ Let $\mathcal{C}(P)$ denote the set of conditional independence relationships in the distribution P, and let d-sep(G) denote the d-separation statements entailed by some graph G. Take $G' \in \arg\min_{G \in \mathbb{G}(B,\Omega)} \|G\|_0$; there exist (B', Ω') such that $B' \sim G'$ and $\Sigma = (I - B')^{-1}\Omega'(I - B')^{-T}$; in other words, there exists a linear Gaussian SCM with connectivity B', noise term variances Ω' and underlying graph $\mathcal{G}(B')$ which entails the distribution P. As such, this distribution P is Markov with respect to $\mathcal{G}(B')$ and its independence map G'. Because P is faithful with respect to $\mathcal{G}(B)$, it follows that d-sep$(G') \subseteq$ d-sep$(\mathcal{G}(B)) = \mathcal{C}(P)$, and thus G' is an independence map of $\mathcal{G}(B)$. By Theorem 3 there exists a sequence S of edge additions and covered-edge reversals that yield G' when applied to $\mathcal{G}(B)$. By Lemma 9, G' and $\mathcal{G}(B)$ have the same number of edges; therefore, the sequence S contains only covered-edge reversals and no edge additions. By Corollary 3, this means that G' and $\mathcal{G}(B)$ are Markov equivalent, completing the proof. ■

Lemma 10 (structure of $(I - B')(I - B)^{-1}$ for same parents) *Consider a model (B, Ω) with a single environment and let $(B', \Omega') \in [(B, \Omega)]$ be an equivalent model. Let e_i be the i^{th} canonical unit vector. The following statements are equivalent:*

i) $B_{i:} = B'_{i:}$,

ii) $\left[(I - B')(I - B)^{-1}\right]_{i:} = e_i^T$, *and*

iii) $\left[(I - B')(I - B)^{-1}\right]_{:i} = e_i$.

Proof We proceed in order.

($i \implies ii$) Follows immediately from Lemma 6.

($ii \implies iii$) Since $(B', \Omega') \in [(B, \Omega)]$, it holds that $(I-B')(I-B)^{-1}\Omega(I-B)^{-T}(I-B')^T$ is a diagonal matrix with positive entries. By Lemma 5, it follows that $(I - B')(I - B)^{-1}\Omega^{1/2}$ has full rank and orthogonal row vectors. Now, since (ii) is true, the i^{th} row vector is $\left[(I - B')(I - B)^{-1}\Omega^{1/2}\right]_{i:} = \Omega_{ii}^{1/2} e_i^T$, and so the remaining vectors are orthogonal to it iff they all have zeros on the i^{th} coordinate, i.e., $\left[(I - B')(I - B)^{-1}\right]_{:i} = e_i$.

($iii \implies i$) Note that (iii) can be rewritten as $(I - B')(I - B)^{-1}e_i = e_i$. It follows that

$$e_i^T = e_i^T(I - B')^{-T}(I - B)^T. \tag{17}$$

Now, since $(B', \Omega') \in [(B, \Omega)]$, we have that

$$(I - B')(I - B)^{-1}\Omega(I - B)^{-T}(I - B')^T = \Omega',$$

which we can rewrite as

$$(I - B') = \Omega'(I - B')^{-T}(I - B)^T\Omega^{-1}(I - B).$$

Multiplying on the left by e_i^T, and letting $\Omega'_{ii} = \alpha$ and $\Omega_{ii}^{-1} = \beta$ we arrive at

$$\begin{aligned} e_i^T(I-B) &= e_i^T\Omega'(I-B')^{-T}(I-B)^T\Omega^{-1}(I-B) \\ &= \alpha e_i^T(I-B')^{-T}(I-B)^T\Omega^{-1}(I-B) \\ &= \alpha e_i^T\Omega^{-1}(I-B) \quad \text{(c.f. Equation 17)} \\ &= \alpha\beta e_i^T(I-B). \end{aligned}$$

We have arrived at $(I-B)_{i:} = \alpha\beta(I-B)_{i:}$. Since both B' and B have zeros on the diagonal, $(I-B)_{ii} = (I-B')_{ii} = 1 = \alpha\beta(I-B)_{ii}$, which implies that $\alpha\beta = 1$. Therefore, $(I-B')_{i:} = (I-B)_{i:}$ and $B'_{i:} = B_{i:}$. ■

Lemma 11 *(independence of noise terms and parents) Let (B, Ω) be a model and define the random vector $X := (I-B)^{-1}\epsilon$ with $\epsilon \sim \mathcal{N}(0, \Omega)$. For any $i \in \{1, ..., p\}$, it holds that $\epsilon_i \perp\!\!\!\perp X_{\text{supp}}(B_{i:})$.*

Proof We have that

$$\begin{pmatrix} \epsilon \\ X \end{pmatrix} = \begin{pmatrix} I \\ (I-B)^{-1} \end{pmatrix} \Omega^{1/2} Z \text{ with } Z = (Z_1, ..., Z_p) \sim \mathcal{N}(0, I),$$

that is, ϵ and X are jointly normal with covariance

$$\Gamma := \begin{pmatrix} \Omega & P \\ P^T & \Sigma \end{pmatrix}$$

where $P := \Omega(I-B)^{-T}$ and $\Sigma := (I-B)^{-1}\Omega(I-B)^{-T}$. Denote $S := \text{supp}(B_{i:})$; the joint normality implies that $\epsilon \mid X_S$ follows a normal distribution with mean

$$\mathbb{E}[\epsilon_i] + P_{i,S}\Sigma_{S,S}^{-1}(x - \mathbb{E}[X_S]) \tag{18}$$

and variance

$$\Omega_{ii} - P_{i,S}\Sigma_{S,S}^{-1}P_{S,i}. \tag{19}$$

Now, without loss of generality, assume B is lower triangular with zeros on the diagonal. It follows that

$$j < i \quad \forall i \in \{1, ..., p\}, \forall j \in \text{supp}(B_{i:}). \tag{20}$$

Since B is lower triangular, so are $(I-B)$ and $(I-B)^{-1}$, and thus $P = \Omega(I-B)^{-T}$ is upper triangular. Therefore $P_{ij} = 0$ for $j < i$, and together with (20), it follows that $P_{i,S} = 0$. Thus, from (18, 19) we see that $\epsilon \mid X_S$ has mean zero and variance Ω_{ii}, that is, $\epsilon_i \perp\!\!\!\perp X_S$. ■

Lemma 12 *(Same structure implies same coefficients) Consider a model (B, Ω) with a single environment and let $(B', \Omega') \in [(B, \Omega)]$ be an equivalent model. Then,*

$$\text{supp}(B_{i:}) = \text{supp}(B'_{i:}) \iff B_{i:} = B'_{i:}.$$

Proof The " $\Longleftarrow$ " direction is trivial. For the " $\Longrightarrow$ " direction, define the random vectors

$$X := (I - B)^{-1}\epsilon \text{ and } X' := (I - B')^{-1}\epsilon' \text{ with } \epsilon \sim \mathcal{N}(0, \Omega),\ \epsilon' \sim \mathcal{N}(0, \Omega').$$

Let $S := \text{supp}(B_{i:}) = \text{supp}(B'_{i:})$; we have that

$$X_i = B_{iS}X_S + \epsilon_i \text{ and } X'_i = B'_{iS}X'_S + \epsilon'_i.$$

Because $(B', \Omega') \in [(B, \Omega)]$, X and X' follow the same distribution, and it holds that for all $x_S \in \mathbb{R}^{|S|}$ that

$$\mathbb{E}[X_i \mid X_S = x_S] = \mathbb{E}[X'_i \mid X'_S = x_S].$$

Plugging in the expression for X_i and X'_i into the above, we arrive at

$$B_{iS}x_S + \mathbb{E}[\epsilon_i \mid X_S = x_S] = B'_{iS}x_S + \mathbb{E}[\epsilon'_i \mid X'_S = x_S],$$

and by Lemma 11, at

$$B_{iS}x_S + \mathbb{E}[\epsilon_i] = B'_{iS}x_S + \mathbb{E}[\epsilon'_i].$$

Since $\mathbb{E}[\epsilon_i] = \mathbb{E}[\epsilon'_i] = 0$ the above results in the condition that $(B_{iS} - B'_{iS})x_S = 0$ for all $x_S \in \mathbb{R}^{|S|}$, which is true if and only if $B_{iS} = B'_{iS}$, completing the proof. ■

Lemma 13 *Consider a linear structural equation model* (1) *with connectivity matrix* B *that specifies a DAG* $\mathcal{G}(B)$ *(see Equation 3). Then, we have that both results hold:*

(i) $[(I - B)^{-1}]_{ii} = 1$ *for all* $i \in [p]$

(ii) $[(I - B)^{-1}]_{ji} \neq 0$ *implies that* i *is an ancestor of* j *in the DAG specified by* B*.*

Proof B specifies a DAG $\mathcal{G}(B)$ if and only if B is nilpotent. Thus, we have that $B^k = 0$ for $k \geq p$, and we can write the decomposition

$$(I - B)^{-1} = I + B + B^2 + \ldots, B^{p-1}.$$

For any h, l and $k \leq p - 1$,

$$B^k_{hl} = \sum_{(v_1,\ldots,v_{k-1}) \in [p]^{k-1}} B_{hv_1}B_{v_1v_2}\cdots B_{v_{k-1}l},$$

where $(v_1, \ldots, v_{k-1}) \in [p]^{k-1}$ is a collection of nodes (without repetition) of size $k - 1$. The product $B_{hv_1}B_{v_1v_2}\ldots B_{v_{k-1}l}$ is nonzero if each term is nonzero, i.e., in the DAG $\mathcal{G}(B)$ there is a path $l \to v_{k-1} \to \cdots \to v_1 \to h$. In other words, $B^k_{hl} \neq 0$ implies that there exists a directed path of length k from l to h.

The first part (i) follows from using the decomposition above and noting that $B^k_{ii} \neq 0$ would imply that there is a directed path from i to i, violating acylicity. For the second part (ii), note from the decomposition that since $[(I - B)^{-1}]_{ji} \neq 0$, there must exist a $k \leq p - 1$ such that $B^k_{ji} \neq 0$. This implies there exists a directed path of length k from i to j. This completes the proof. ■

A.4 Supporting Results for the Case $|\mathcal{E}| > 1$

Lemma 14 (Under faithfulness, the true model has the minimum number of edges) *Consider a model $\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$ resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Then it holds that*

$$\min_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0 = \|\mathcal{G}(B)\|_0.$$

Proof Let $\mathcal{E}^0 \subset \mathcal{E}$ be a singleton. By Lemma 9 we have that $\mathcal{G}(B)$ attains the minimum number of edges in $\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}^0}\right)$. It is trivial to see[12] that $\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right) \subseteq \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}^0}\right)$; since $\mathcal{G}(B) \in \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$, it follows that $\mathcal{G}(B)$ also attains the minimum number of edges in $\mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$. ■

Lemma 15 (structure of $(I-\tilde{B})(I-B)^{-1}$ for different parents) *Let $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ be a model with a single environment, and let $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$ be an equivalent model such that the diagonal entries of the matrix $M := (I-\tilde{B})(I-B)^{-1}$ are all non-zero. Then, it holds for any $i = 1, 2, \ldots, p$ that*

$$\mathrm{supp}(B_{i:}) \neq \mathrm{supp}(\tilde{B}_{i:}) \implies \left|\mathrm{supp}(\left[(I-\tilde{B})(I-B)^{-1}\right]_{i:})\right| > 1.$$

Proof Let $M := (I-\tilde{B})(I-B)^{-1}$ and assume that

$$\mathrm{supp}(B_{i:}) \neq \mathrm{supp}(\tilde{B}_{i:}) \text{ but } |\mathrm{supp}(M_{i:})| \leq 1.$$

Since M has full rank it cannot be that $|\mathrm{supp}(M_{i:})| = 0$. If $|\mathrm{supp}(M_{i:})| = 1$ and $M_{i:} = [(I-\tilde{B})(I-B)^{-1}]_{i:} = \alpha e_i^T$ for $\alpha \neq 0$, by Lemma 6 we have that $(I-\tilde{B})_{i:} = \alpha(I-B)_{i:}$ and thus $\mathrm{supp}(B_{i:}) = \mathrm{supp}(\tilde{B}_{i:})$, resulting in a contradiction. We will now show that the remaining case, $M_{i:} = \alpha e_j^T$ for $j \neq i$ and $\alpha \neq 0$, also results in a contradiction. We first present two implications of $M_{i:} = \alpha e_j^T$, and then show that these implications lead to a contradiction.

Implication 1 Suppose $M_{i:} = \alpha e_j^T$ for $\alpha \neq 0$ and $j \neq i$. Then, $M_{:j} = b e_i$ with $b \neq 0$, i.e., the j^{th} column of M is proportional to the i^{th} unit vector.

Proof of Implication 1 Note the following chain of equalities for any $k = 1, 2, \ldots, p$, following from $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$:

$$\begin{aligned}
\tilde{\Omega} &= M\Omega M^T, \\
e_i^T\tilde{\Omega}e_k &= e_i^T M\Omega M^T e_k, \\
\tilde{\Omega}_{ik} &= M_{i:}\Omega M_{:k}^T, \\
&= \alpha e_j^T \Omega M_{:k}^T, \\
&= \alpha\Omega_{jj} e_j^T M_{:k}^T, \\
&= \alpha\Omega_{jj} M_{kj},
\end{aligned}$$

that is, $M_{kj} = \frac{\tilde{\Omega}_{ik}}{\alpha\Omega_{jj}}$. Because $\Omega, \tilde{\Omega}$ are positive diagonal matrices, it follows that $M_{:j} = b e_i$, where $b = \frac{\tilde{\Omega}_{ii}}{\alpha\Omega_{jj}} \neq 0$. This completes the proof of Implication 1.

12. See also Corollary 1.

Implication 2 Suppose $M_{i:} = \alpha e_j^T$ for $\alpha \neq 0$ and $j \neq i$. Then, $M_{:i}^{-1} = \frac{1}{\alpha} e_j$, i.e., the i^{th} column of M^{-1} is proportional to the j^{th} unit vector.

Proof of Implication 2 Since M is a full rank and thus an invertible matrix, we have for any $r = 1, 2, \ldots, p$

$$\begin{aligned} MM^{-1} &= I, \\ e_r^T M M^{-1} e_i &= e_r^T e_i. \end{aligned}$$

If $r = i$, the above becomes

$$\begin{aligned} e_i^T M M^{-1} e_i &= e_i^T e_i, \\ M_{i:} M^{-1} e_i &= 1, \\ \alpha e_j^T M^{-1} e_i &= 1, \\ M_{ji}^{-1} &= \frac{1}{\alpha}. \end{aligned}$$

Now, for $r \neq i$, the above becomes

$$\begin{aligned} M_{r:} M_{:i}^{-1} &= 0, \\ \sum_k M_{rk} M_{ki}^{-1} &= 0, \\ M_{rj} M_{ji}^{-1} + \sum_{k \neq j} M_{rk} M_{ki}^{-1} &= 0, \\ \sum_{k \neq j} M_{rk} M_{ki}^{-1} &= 0, \end{aligned}$$

where the last step follows from Implication 1 ($M_{:j} = b e_i$). Because the above holds for all $r \neq i$, we can rewrite it as the matrix product

$$A\, M_{k \neq j, i}^{-1} = \mathbf{0}, \quad \text{where } A := \left(M_{rl}\right)_{\substack{r \neq i \\ l \neq j}} \in \mathbb{R}^{p-1 \times p-1}. \tag{21}$$

Note here that $M_{k \neq j, i}^{-1}$ is a $p-1$ dimensional vector. By cofactor expansion along the j^{th} column, we have the following decomposition of the determinant of the matrix M:

$$\det(M) = \sum_k M_{kj} (-1)^{k+j} \det\left(\left(M_{rl}\right)_{\substack{r \neq k \\ l \neq j}}\right).$$

By Implication 1 ($M_{:j} = b e_i$) the decomposition reduces to

$$\det(M) = b(-1)^{i+j} \det\left(\left(M_{rl}\right)_{\substack{r \neq i \\ l \neq j}}\right) = b(-1)^{i+j} \det(A).$$

Since M is full-rank and thus has nonzero determinant, and $b(-1)^{i+j} \neq 0$, we have that A has nonzero determinant, and is thus full rank. As a result, (21) implies that $M_{ki}^{-1} = 0$ for all $k \neq j$, completing the proof for Implication 2.

Completing proof of Lemma 15 We can now proceed with the main argument of the proof. For any $l = 1, 2, \ldots, p$, we have the following equalities:

$$\begin{aligned} M(I - B) &= (I - \tilde{B}), \\ e_i^T M(I - B)e_l &= e_i^T(I - \tilde{B})e_l, \\ \alpha e_j^T(I - B)e_l &= e_i^T(I - \tilde{B})e_l, \\ \alpha e_j^T e_l - \alpha B_{jl} &= e_i^T e_l - \tilde{B}_{il}. \end{aligned}$$

The above results in two constraints:

$$B_{ji} = -\frac{1}{\alpha} \text{ (from } l = i) \quad \text{and} \quad \tilde{B}_{ij} = -\alpha \text{ (from } l = j).$$

Note how the latter implies that the edge $j \to i$ exists in the graph $\mathcal{G}(\tilde{B})$. Now, starting from the same point as before, see that

$$\begin{aligned} (I - \tilde{B}) &= M(I - B), \\ (I - \tilde{B})^{-1} &= (I - B)^{-1}M^{-1}, \\ e_j^T(I - \tilde{B})^{-1}e_i &= e_j^T(I - B)^{-1}M^{-1}e_i \\ &= e_j^T(I - B)^{-1}\frac{1}{\alpha}e_j, \qquad \text{(by Implication 2)} \end{aligned}$$

that is, $[(I - \tilde{B})^{-1}]_{ji} = \frac{1}{\alpha}[(I - B)^{-1}]_{jj}$. By the first part of Lemma 13, we have that $[(I - B)^{-1}]_{jj} = 1$. Thus, $[(I - \tilde{B})^{-1}]_{ji} = \frac{1}{\alpha} \neq 0$. By the second part of Lemma 13, this implies that there is a directed path from i to j in $\mathcal{G}(\tilde{B})$. Together with the edge $j \to i$ above, this results in a cycle in $\mathcal{G}(\tilde{B})$, which is a contradiction.

We have shown that if $\text{supp}(B_{i:}) \neq \text{supp}(\tilde{B}_{i:})$, the i^{th} row of M cannot have no or a single non-zero entry, and thus $\left|\text{supp}(\left[(I - \tilde{B})(I - B)^{-1}\right]_{i:})\right| > 1$.

■

Lemma 16 (I-MAPs are distribution equivalent) *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ with underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} = \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received an intervention in at least one of the environments in $\mathcal{E}$. Let G be an independence map of $\mathcal{G}(B)$ such that $\text{PA}_{\mathcal{G}(B)}(i) \subseteq \text{PA}_G(i)$ for all $i \in \mathcal{I}$. Then, $G \in \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$.*

Proof We want to show that there exists a model $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$ such that $\tilde{B} \sim G$ and

$$(I - B)^{-1}\Omega^e(I - B)^{-T} = (I - \tilde{B})^{-1}\tilde{\Omega}^e(I - \tilde{B})^{-T} \text{ for all } e \in \mathcal{E}. \tag{22}$$

That is equivalent to the condition that for all $e \in \mathcal{E}$

$$\tilde{\Omega}^e := (I - \tilde{B})(I - B)^{-1}\Omega^e(I - B)^{-T}(I - \tilde{B})^T$$

is a diagonal matrix with positive entries. By Lemma 5, in turn this is equivalent to $(I - \tilde{B})(I - B)^{-1}(\Omega^e)^{1/2}$ being a full-rank matrix with orthogonal row vectors, for all $e \in \mathcal{E}$.

Now, let $\mathcal{E}^0 \subset \mathcal{E}$ be a singleton and let $\Omega^0 := \operatorname{diag}\left(\omega_1^2, ..., \omega_p^2\right)$ be the noise-term variances associated to it. By Lemma 8, there exists $\tilde{B}$ such that $\tilde{B} \sim G$, and $(I-\tilde{B})(I-B)^{-1}(\Omega^0)^{1/2}$ is a full-rank matrix with orthogonal row vectors. Now, for every $e \in \mathcal{E}$, we can rewrite

$$(I - \tilde{B})(I - B)^{-1}(\Omega^e)^{1/2} = (I - \tilde{B})(I - B)^{-1}(\Omega^0)^{1/2}(\Omega^0)^{-1/2}(\Omega^e)^{1/2}. \tag{23}$$

If we let $\gamma_1^2, ..., \gamma_p^2$ be the diagonal entries of Ω^e, we can write

$$(\Omega^0)^{-1/2}(\Omega^e)^{1/2} = \operatorname{diag}\left(\left|\frac{\gamma_1}{\omega_1}\right|, ..., \left|\frac{\gamma_p}{\omega_p}\right|\right),$$

i.e., a diagonal matrix that constitutes a positive scaling of some of the columns of $(I - \tilde{B})(I - B)^{-1}(\Omega^0)^{1/2}$. Note[13] that this scaling can only occur in those columns with index in $\mathcal{I}$. Since we have assumed that $\mathrm{PA}_{\mathcal{G}(B)}(i) \subseteq \mathrm{PA}_G(i)$ for all $i \in \mathcal{I}$, by Corollary 5 we have that $B_{i:} = \tilde{B}_{i:}$ for $i \in \mathcal{I}$, and by Lemma 10, these columns have exactly one non-zero element. Scaling them leaves the matrix (23) with full rank and orthogonal row vectors, that is, for all $e \in \mathcal{E}$

$$\tilde{\Omega}^e := (I - \tilde{B})(I - B)^{-1}\Omega^e(I - B)^{-T}(I - \tilde{B})^T$$

is a diagonal matrix with positive elements. Thus, (22) holds and since $\tilde{B} \sim G$ we have completed the proof. ■

Lemma 17 (Targets have the same parents) *Consider a model $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ satisfying Assumption 1 and let $\mathcal{I} := \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}})$ be the indices of variables that have received interventions in $\mathcal{E}$. Let $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$ be a distribution equivalent model. Then, $B_{i:} = \tilde{B}_{i:}$ for all $i \in \mathcal{I}$.*

Proof If $|\mathcal{E}| = 1$, then $\mathcal{I} = \emptyset$ and we are done. For $|\mathcal{E}| > 1$, see that because $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$ it holds that, for all $e \in \mathcal{E}$,

$$\tilde{\Omega}^e = (I - \tilde{B})(I - B)^{-1}\Omega^e(I - B)^{-T}(I - \tilde{B})^T,$$

which implies that $M^e := (I-\tilde{B})(I-B)^{-1}(\Omega^e)^{1/2}$ is a full-rank matrix with orthogonal row vectors (Lemma 5). Let $S := \{i : B_{i:} = \tilde{B}_{i:}\}$; then, $M^e_{i:} = (\Omega^e)^{1/2}e_i^T$ and $M^e_{:i} = (\Omega^e)^{1/2}e_i$ for all $i \in S$. In other words, for any $e \in \mathcal{E}$ we have that M^e is a block-diagonal matrix with two blocks

i) M^e_{SS}, a diagonal matrix; and

ii) $M^e_{S^\perp S^\perp}$, where $S^\perp := [p] \setminus S$.

Because of this block-diagonal form, the rows of one block are orthogonal to those of the other. Thus, M^e has orthogonal row vectors if and only if so do both blocks. This is

13. Because $\mathcal{I} := \mathbb{I}\left(\{\Omega^e\}_{e\in\mathcal{E}}\right)$ and (5).

immediately true for the diagonal block (i). Take some $f \in \mathcal{E} : f \neq e$ and let $D := (\Omega^f)^{-1/2}(\Omega^e)^{1/2}$. We can rewrite

$$M^e = (I - \tilde{B})(I - B)^{-1}(\Omega^f)^{1/2}(\Omega^f)^{-1/2}(\Omega^e)^{1/2} = M^f D$$

and $M^e_{S^\perp S^\perp} = M^f_{S^\perp S^\perp} D_{S^\perp S^\perp}$. Because $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$, by Lemma 5 we have that M^e and M^f have full rank and orthogonal row vectors, and because of their block structure, so do $M^e_{S^\perp S^\perp}$ and $M^f_{S^\perp S^\perp}$. From Lemma 15, we additionally have that $M^f_{S^\perp S^\perp}$ has more than one non-zero element per row. Now, assume $\mathcal{I} \cap S^\perp \neq \emptyset$, that is, $\exists i \in \mathcal{I}$ such that $B_{i:} \neq \tilde{B}_{i:}$. Then, by intervention-heterogeneity (Assumption 1), $D_{S^\perp S^\perp}$ has at least one diagonal entry different from all others. These facts allow us to apply Lemma 7, by which $M^e_{S^\perp S^\perp} = M^f_{S^\perp S^\perp} D_{S^\perp S^\perp}$ does not have orthogonal rows, arriving at a contradiction. This completes the proof. ■

A.5 Main Results of Section 3

Now we provide the proofs for Lemma 1, Proposition 1, and Theorem 1.

Lemma 1 *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ with underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}\text{−MEC}\left(\mathcal{G}(B)\right) \subseteq \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right).$$

Proof Let $G \in \mathcal{I}\text{−MEC}(\mathcal{G}(B))$. By Definition 2, this implies that G is an independence map of $\mathcal{G}(B)$, and that $\mathrm{PA}_{\mathcal{G}(B)}(i) = \mathrm{PA}_G(i)$ for all $i \in \mathcal{I}$. Thus, by Lemma 16, $G \in \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$, completing the proof. ■

Proposition 1 *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}\text{−MEC}\left(\mathcal{G}(B)\right) \subseteq \operatorname*{arg\,min}_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0,$$

where $\|G\|_0$ is the number of edges in G.

Proof Let $G' \in \mathcal{I}\text{−MEC}(\mathcal{G}(B))$. By Lemma 1, $G' \in \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$. By Lemma 14 we have that $\min_{G\in\mathbb{G}(B,\Omega)} \|G\|_0 = \|\mathcal{G}(B)\|_0$; since $\mathcal{G}(B)$ and G' are Markov equivalent, they have the same number of edges, and thus $G' \in \arg\min_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0$, completing the proof. ■

Theorem 1 *Consider a model $\big(B, \{\Omega^e\}_{e\in\mathcal{E}}\big)$ satisfying Assumption 1 and resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\big(\{\Omega^e\}_{e\in\mathcal{E}}\big)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then,*

$$\mathcal{I}\text{−MEC}\left(\mathcal{G}(B)\right) = \operatorname*{arg\,min}_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0.$$

Proof We proceed by showing both sets are contained in each other.

$\subseteq$ Follows from Proposition 1.

$\supseteq$ Let $\tilde{G} \in \arg\min_{G \in \mathbb{G}(B, \{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0$. We begin by showing that $\tilde{G}$ and $\mathcal{G}(B)$ have the same skeleton and v-structures. Let $\mathcal{E}^0 \subset \mathcal{E}$ be a singleton and Ω^0 be the noise-term variances associated with it. Clearly (c.f. Corollary 1), $\tilde{G} \in \mathbb{G}\left(B, \Omega^0\right)$. Because we are under faithfulness, by Lemmas 9 and 14

$$\min_{G\in\mathbb{G}(B,\Omega^0)} \|G\|_0 = \|\mathcal{G}(B)\|_0 = \min_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0,$$

and thus $\tilde{G} \in \arg\min_{G\in\mathbb{G}(B,\Omega^0)} \|G\|_0$. By Proposition 6, $\arg\min_{G\in\mathbb{G}(B,\Omega^0)} \|G\|_0 = \mathrm{MEC}(\mathcal{G}(B))$ and thus $\tilde{G}$ and $\mathcal{G}(B)$ have the same skeleton and v-structures.

If $\mathcal{E}$ is a singleton, then $\mathcal{I} = \emptyset$ and we have completed the proof. For the remaining cases, we will now show that the variables in the intervention set $\mathcal{I}$ have the same parents in $\tilde{G}$ and $\mathcal{G}(B)$. Since $\tilde{G} \in \mathbb{G}\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$, there exists a model $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) \in [(B, \{\Omega^e\}_{e\in\mathcal{E}})]$ such that $\tilde{B} \sim \tilde{G}$. First, by Lemma 17, it holds that $\tilde{B}_{i:} = B_{i:}$ for all $i \in \mathcal{I}$. Second, since $\tilde{G} \in \arg\min_{G\in\mathbb{G}(B,\{\Omega^e\}_{e\in\mathcal{E}})} \|G\|_0$, it must hold that $\mathcal{G}(\tilde{B}) = \tilde{G}$ or we would have a contradiction[14]. From these two facts, it follows that $\mathrm{PA}_{\tilde{G}}(i) = \mathrm{PA}_{\mathcal{G}(\tilde{B})}(i) = \mathrm{PA}_{\mathcal{G}(B)}(i)$ for all $i \in \mathcal{I}$, completing the proof. ■

Corollary 2 *Consider a model $\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)$ satisfying Assumption 1 and resulting in a set of distributions which are faithful with respect to its underlying graph $\mathcal{G}(B)$. Let $\mathcal{I} := \mathbb{I}\left(\{\Omega^e\}_{e\in\mathcal{E}}\right)$ be the indices of variables that have received interventions in $\mathcal{E}$. Then, for any distribution equivalent model $\left(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}}\right) \in \left[\left(B, \{\Omega^e\}_{e\in\mathcal{E}}\right)\right]$, it holds that*

$$\tilde{\Omega}^e_{ii} = \Omega^e_{ii}\ \forall i \in \mathcal{I}\ , \forall e \in \mathcal{E} \quad \textit{and} \quad \tilde{\Omega}^e_{jj} = \tilde{\Omega}^f_{jj}\ , \forall j \notin \mathcal{I}\ \forall e, f \in \mathcal{E}.$$

Proof We prove each statement separately.

($\tilde{\Omega}^e_{ii} = \Omega^e_{ii}$) From Definition 1, see that

$$\begin{aligned}\tilde{\Omega}^e_{ii} &= e_i^T(I-\tilde{B})(I-B)^{-1}\Omega^e(I-B)^{-T}(I-\tilde{B})^T e_i \\ &= \left[(I-\tilde{B})(I-B)^{-1}\right]_{i:} \Omega^e \left[(I-B)^{-T}(I-\tilde{B})\right]_{:i}.\end{aligned}$$

By Lemma 17 and Lemma 10, the above becomes $\tilde{\Omega}^e_{ii} = e_i^T\Omega^e e_i = \Omega_{ii}$.

($\tilde{\Omega}^e_{jj} = \tilde{\Omega}^f_{jj}$) We proceed by contradiction: let $e, f \in \mathcal{E}$ and $j \in [p]$ such that $\tilde{\Omega}^e_{jj} \neq \tilde{\Omega}^f_{jj}$. From Definition 1, we have that

$$(I-B)^{-1}(\Omega^e - \Omega^f)(I-B)^{-T} = (I-\tilde{B})^{-1}(\tilde{\Omega}^e - \tilde{\Omega}^f)(I-\tilde{B})^{-T}.$$

Because $(I-B)$ and $(I-\tilde{B})$ have full rank, $(\Omega^e - \Omega^f)$ and $(\tilde{\Omega}^e - \tilde{\Omega}^f)$ must have equal rank. From (5), we know that $\Omega^e_{kk} = \Omega^f_{kk}\ \forall k \notin \mathcal{I}$, and thus,

$$\mathrm{rank}(\Omega^e - \Omega^f) = |\{k : \Omega^e_{kk} \neq \Omega^f_{kk} \wedge k \in \mathcal{I}\}|.$$

14. Otherwise $\mathcal{G}(\tilde{B})$ would have less edges than $\tilde{G}$ and thus $\tilde{G} \notin \arg\min_{G\in\mathbb{G}\left(B,\{\Omega^e\}_{e\in\mathcal{E}}\right)} \|G\|_0$.

Following the same reasoning, by the first statement of this corollary, and because $\tilde{\Omega}^e_{jj} \neq \tilde{\Omega}^f_{jj}$, it follows that

$$\text{rank}(\tilde{\Omega}^e - \tilde{\Omega}^f) > \text{rank}(\Omega^e - \Omega^f),$$

which is a contradiction. ■

A.6 Proofs for Section 3.1

Lemma 2 *Let G_1 and G_2 be two DAGs over p variables and let $\mathcal{I} \subseteq [p]$ be a set of intervention targets. The following statements are equivalent:*

1. *G_1 and G_2 are $\mathcal{I}$-equivalent, and*
2. *their augmented graphs $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ are Markov equivalent.*

Proof ($1 \implies 2$) G_1 and G_2 have the same skeleton (c.f. Definition 2); thus, so do $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$. Furthermore, G_1 and G_2 have the same v-structures (c.f. Definition 2), and the new v-structures in their augmented graphs are $\{j \to i \leftarrow F_i \mid i \in \mathcal{I} \wedge j \in \text{PA}_{G_1}(i)\}$ and $\{j \to i \leftarrow F_i \mid i \in \mathcal{I} \wedge j \in \text{PA}_{G_2}(i)\}$, respectively. Because $\text{PA}_{G_1}(i) = \text{PA}_{G_2}(i)\ \forall i \in \mathcal{I}$, both sets of v-structures are the same, and $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ are Markov equivalent.

($2 \implies 1$) Because $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ have the same skeleton, so do G_1 and G_2, satisfying condition (i) of Definition 2. Assume now that condition (ii) is false, that is, that G_1 and G_2 have different v-structures. Since the new nodes in the augmented graphs $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ are source nodes, adding them can only create new v-structures. Thus, $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ would also have different v-structures, which contradicts that they are Markov equivalent. Finally, assume now that there exists $i \in \mathcal{I}$ such that $\text{PA}_{G_1}(i) \neq \text{PA}_{G_2}(i)$; then, the addition of a new node and edge $F_i \to i$ would create different v-structures in the augmented graphs $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$, which is again a contradiction; this proves condition (iii) of Definition 2, completing the proof. ■

Proposition 3 *Consider a DAG G and a conservative family of targets $\mathcal{H}$. Let $\mathcal{I} := \cup_{H \in \mathcal{H}} H$. Then $\mathcal{I}-\text{MEC}(G) \subseteq \mathcal{H}-\text{MEC}(G)$.*

Proof Let $\tilde{G} \in \mathcal{I}-\text{MEC}(G)$. By definition G and $\tilde{G}$ are Markov equivalent, satisfying (i) in Proposition 2. Since $\forall i \in \mathcal{I} : \text{PA}_G(i) = \text{PA}_{\tilde{G}}(i)$, the graphs $\tilde{G}^{(H)}$ and $G^{(H)}$ have the same skeleton for all $H \in \mathcal{H}$, satisfying (ii) and completing the proof. ■

Proposition 4 *Consider a DAG G and let $\mathcal{Y}$ be a family of targets such that $\emptyset \in \mathcal{Y}$, and let $\mathcal{I} := \cup_{Y \in \mathcal{Y}} Y$. Then $\mathcal{I}-\text{MEC}(G) \subseteq \mathcal{Y}-\text{MEC}(G)$.*

Proof Let $\tilde{G} \in \mathcal{I}-\text{MEC}(G)$ and let $\tilde{G}^{\mathcal{Y}}$ and $G^{\mathcal{Y}}$ be the corresponding interventional DAGs. By Definition 2, $\tilde{G}$ and G have the same skeleton; because the edges introduced in the interventional DAGs depend only on $\mathcal{Y}$, $\tilde{G}^{\mathcal{Y}}$ and $G^{\mathcal{Y}}$ have the same additional edges

and also have the same skeleton. We now show that they also have the same v-structures. Without loss of generality, suppose there exists a v-structure in $G^{\mathcal{Y}}$ which is not in $\tilde{G}^{\mathcal{Y}}$. Because the nodes $Y \in \mathcal{Y}$ do not have any parents, they can only appear in the tails of the v-structure, which can therefore be of three types:

a) $i \to k \leftarrow j$ where $i, k, j \in [p]$;

b) $Y_1 \to k \leftarrow Y_2$ where $k \in [p]$ and $Y_1, Y_2 \in \mathcal{Y}$; or

c) $Y \to k \leftarrow j$ where $Y \in \mathcal{Y}$ and $j, k \in [p]$.

By Definition 2, $\tilde{G}$ and G have the same v-structures, which rules out (a). Furthermore, because the edges from nodes in $\mathcal{Y}$ to nodes in $[p]$ are the same for both $\tilde{G}^{\mathcal{Y}}$ and $G^{\mathcal{Y}}$, and they have the same skeleton, a v-structure like (b) would appear in both graphs. We now consider (c): since both graphs have the same skeleton and nodes in $\mathcal{Y}$ are always source nodes, the only option is that (c) appears instead as $Y \to k \to j$ in $\tilde{G}^{\mathcal{Y}}$. However, the edge from $Y \to k$ means that $k \in Y$, and by extension $k \in \mathcal{I}$; thus, by Definition 2 k would have the same parents in $\tilde{G}$ and G, arriving at a contradiction. Thus, we have shown that the interventional DAGs $\tilde{G}^{\mathcal{Y}}$ and $G^{\mathcal{Y}}$ have the same skeleton and v-structures, which means that $\tilde{G} \in \mathcal{Y}-\mathrm{MEC}(G)$, completing the proof. ■

A.7 Proofs of the Score Properties

We begin by stating some supporting results.

Lemma 18 ($\mathcal{I}$-equivalent graphs entail the same distributions) *Consider a set of intervention targets $\mathcal{I} \subseteq [p]$ and two graphs G and G' such that $G \sim_{\mathcal{I}} G'$. Then, for every model $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ such that $B \sim G$ and $\mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}) = \mathcal{I}$, there exists a model $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$ such that $\tilde{B} \sim G'$, $\mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) = \mathcal{I}$, and*

$$(I - B)^{-1}\Omega^e(I - B)^{-T} = (I - \tilde{B})^{-1}\tilde{\Omega}^e(I - \tilde{B})^{-T}.$$

Proof G is an independence map of $\mathcal{G}(B)$ as it contains all it edges. Because G and G' are $\mathcal{I}$-equivalent, G' is also an independence map of $\mathcal{G}(B)$, and since

$$\mathrm{PA}_{\mathcal{G}(B)}(i) \subseteq \mathrm{PA}_G(i) = \mathrm{PA}_{G'}(i), \tag{24}$$

by Lemma 16 there exists $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$ such that $\tilde{B} \sim G'$ and

$$(I - B)^{-1}\Omega^e(I - B)^{-T} = (I - \tilde{B})^{-1}\tilde{\Omega}^e(I - \tilde{B})^{-T}. \tag{25}$$

Now we will show that $\mathcal{I} = \mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) = \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}})$. As for notation, let $M := (I - \tilde{B})(I - B)^{-1}$. Together with Corollary 5, (24) implies that $B_{i:} = \tilde{B}_{i:}$ and therefore[15] $M_{i:} = e_i^T$ and $M_{:i} = e_i$ for $i \in \mathcal{I}$. First we will show that $i \in \mathbb{I}(\Omega^e\}_{e\in\mathcal{E}} \implies i \in \mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$. Note that (c.f. Equation 5)

$$i \in \mathcal{I} = \mathbb{I}(\Omega^e\}_{e\in\mathcal{E}}) \iff \exists e, f \in \mathcal{E} : [\Omega^e(\Omega^f)^{-1}]_{ii} \neq 1. \tag{26}$$

15. Pick any $g \in \mathcal{E}$: see that by Corollary 1 $(\tilde{B}, \tilde{\Omega}^g) \in [(B, \Omega^g)]$ and apply Lemma 10.

We can rewrite (25) as $\tilde{\Omega}^e = M\Omega^e M^T$, and thus $\tilde{\Omega}^e(\tilde{\Omega}^f)^{-1} = M\Omega^e(\Omega^f)^{-1}M^{-1}$. For $i \in \mathcal{I}$ we have that

$$\begin{aligned}
[\tilde{\Omega}^e(\tilde{\Omega}^f)^{-1}]_{ii} &= M_{i:}\Omega^e(\Omega^f)^{-1}M_{:i}^{-1} \\
&= e_i^T\Omega^e(\Omega^f)^{-1}M_{:i}^{-1} \quad (M_{i:} = e_i^T, \text{ see above}) \\
&= [\Omega^e(\Omega^f)^{-1}]_{ii}e_i^T M_{:i}^{-1} \quad (\Omega^e, \Omega^f \text{ are diagonal}) \\
&= [\Omega^e(\Omega^f)^{-1}]_{ii}M_{i:}M_{:i}^{-1} \\
&= [\Omega^e(\Omega^f)^{-1}]_{ii} \neq 1,
\end{aligned}$$

and thus by (26) $i \in \mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$. Now we will show that $j \notin \mathbb{I}(\Omega^e\}_{e\in\mathcal{E}} \implies j \notin \mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$. Note that

$$j \notin \mathcal{I} = \mathbb{I}(\Omega^e\}_{e\in\mathcal{E}}) \iff [\Omega^e - \Omega^f]_{jj} = 0 \; \forall e, f \in \mathcal{E}. \tag{27}$$

Now, pick any $e, f \in \mathcal{E}$ and see from (25) that $\tilde{\Omega}^e - \tilde{\Omega}^f = M(\Omega^e - \Omega^f)M^T$. Because $M_{:i} = e_i$ for $i \in \mathcal{I}$, it follows that $M_{ji} = 0$ for $j \notin \mathcal{I}$. Thus, we have that

$$[\tilde{\Omega}^e - \tilde{\Omega}^f]_{jj} = M_{j:}(\Omega^e - \Omega^f)M_{:j}^T = \sum_{i\in\mathcal{I}} M_{ij}[\Omega^e - \Omega^f]_{ii} + \sum_{k\notin\mathcal{I}} M_{kj}[\Omega^e - \Omega^f]_{kk} = 0,$$

as by (27) $[\Omega^e - \Omega^f]_{kk} = 0$ in the second sum. Because we picked e, f arbitrarily, this means that $[\tilde{\Omega}^e - \tilde{\Omega}^f]_{jj} = 0$ for all $e, f \in \mathcal{E}$ and thus, by the same argument as in (27), $j \notin \mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$. Together with the previous results, this shows that $\mathbb{I}(\{\tilde{\Omega}^e\}_{e\in\mathcal{E}}) = \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}) = \mathcal{I}$, completing the proof. ■

Proposition 5 (Score properties) *Without additional assumptions, the score* (7) *satisfies score equivalence* (i) *and decomposability* (ii)*. Under faithfulness and Assumption 1 for the data-generating model, and the assumptions that* $\lim_{\text{all } n_e\to\infty} \frac{n^e}{\sum_{e\in\mathcal{E}} n^e} > 0$ *for all* $e \in \mathcal{E}$ *and that the parameter space is compact, the score is consistent* (iii)*.*

Proof of score equivalence Consider a set $\mathcal{I} \subseteq [p]$ of intervention targets and two graphs G and G' such that $G \sim_{\mathcal{I}} G'$. We want to show that they attain the same score over any sample, that is $S(G, \mathcal{I}) = S(G', \mathcal{I})$. Note that we can rewrite

$$\log P\left(\boldsymbol{X}^e; B, \Omega^e\right) = -n_e \ln\det(K^e) - n_e\text{tr}(K^e\hat{\Sigma}^e), \tag{28}$$

where $\hat{\Sigma}^e$ is the sample covariance for environment e, and $K^e := [(I-B)^{-1}\Omega^e(I-B)^{-T}]^{-1}$ is the inverse of the covariance matrix entailed by the model for environment e. Thus, we can rewrite the score (7) as

$$S(G, \mathcal{I}) = \max_{\substack{B\sim G \\ \text{diagonal } \{\Omega^e\}_{e\in\mathcal{E}} \\ \text{s.t. } \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}})=\mathcal{I}}} \sum_{e\in\mathcal{E}} n_e \ln\det(K^e) - n_e\text{tr}(K^e\hat{\Sigma}^e) - \lambda\text{DoF}(G, \mathcal{I}). \tag{29}$$

Now, let $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ be the model parameters which maximize the score for G. By Lemma 18, there exists $(\tilde{B}, \{\tilde{\Omega}^e\}_{e\in\mathcal{E}})$ for which $\tilde{B} \sim G'$ and

$$(I-B)^{-1}\Omega^e(I-B)^{-T} = (K^e)^{-1} = (I-\tilde{B})^{-1}\Omega^e(I-\tilde{B})^{-T}.$$

Furthermore, because G and G' have the same number of edges, $\mathrm{DoF}(G,\mathcal{I}) = \mathrm{DoF}(G',\mathcal{I})$ and $S(G,\mathcal{I}) \leq S(G',\mathcal{I})$. If we apply the same reasoning to G' and its maximizing parameters, we arrive at $S(G,\mathcal{I}) \geq S(G',\mathcal{I})$. It must hold that $S(G,\mathcal{I}) = S(G',\mathcal{I})$, completing the proof.

Proof of decomposability Note that if $B \in \mathbb{R}^{p\times p}$ has zeros on the diagonal and is lower triangular up to a permutation of rows and columns, then $(I-B)$ can be expressed as PLP for some permutation matrix P and some lower-triangular matrix L with ones on the diagonal. Thus, $\det((I-B)^T) = \det((I-B)) = 1$, and we can rewrite the logarithm term in (28) as

$$\ln\det(K^e) = \ln[\det((I-B)^T)\det((\Omega^e)^{-1})\det(I-B)] = \ln(\prod_{i\in[p]} 1/\Omega^e_{ii}) = -\sum_{i\in[p]} \ln\Omega^e_{ii}$$

For the trace term, we have that

$$\mathrm{tr}(K^e\hat{\Sigma}^e) = \mathrm{tr}((I-B)^T(\Omega^e)^{-1}(I-B)\hat{\Sigma}^e) = \mathrm{tr}((\Omega^e)^{-1}(I-B)\hat{\Sigma}^e(I-B)^T),$$

where the last equality stems from the invariance of the trace to cyclic permutations. By rewriting the trace of products as the sum of entry-wise products, we have that

$$\mathrm{tr}(K^e\hat{\Sigma}^e) = \sum_{i\in[p]}\sum_{j\in[p]} 1/\Omega^e_{ij}[(I-B)\hat{\Sigma}^e(I-B)^T)]_{ij} = \sum_{i\in[p]} 1/\Omega^e_{ii}[(I-B)\hat{\Sigma}^e(I-B)^T)]_{ii},$$

and thus $\mathrm{tr}(K^e\hat{\Sigma}^e) = \sum_{i\in[p]} 1/\Omega^e_{ii}(I-B)_{i:}\hat{\Sigma}^e(I-B)^T)_{:i}$. Together with the new expression for the logarithm term, we can rewrite the score (29) as

$$S(G,\mathcal{I}) = \sum_{i\in[p]} \max_{\substack{B_{i:},\{\Omega^e_{ii}\}_{e\in\mathcal{E}} \\ \text{s.t.} \\ \mathrm{supp}(B_{i:})\subseteq \mathrm{PA}_G(i) \\ \Omega_{ii}>0}} \sum_{e\in\mathcal{E}} -n_e\left[\ln\Omega^e_{ii} + 1/\Omega^e_{ii}(I-B)_{i:}\hat{\Sigma}^e(I-B)^T)_{:i}\right] - \lambda\mathrm{DoF}(i,G,\mathcal{I}),$$

where $\mathrm{DoF}(i,G,\mathcal{I}) = |\mathrm{PA}_G(i)| + \max(|\mathcal{E}|\,\mathbf{1}_{\{i\in\mathcal{I}\}}, 1)$ are the number of parameters estimated for variable i. In other words, the score decouples into a sum of terms depending only on each variable and its parents, which completes the proof.

Proof of score consistency Throughout, we consider p fixed and $n_e \to \infty$ for every $e \in \mathcal{E}$.

For every $G \in \mathcal{I}^\star\text{−MEC}(G^\star)$, let $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ be the unique connectivity matrix and noise variances that specify the population covariance matrices $\{\Sigma^e\}_{e\in\mathcal{E}}$. In our analysis, the parameter space $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ is assumed to be compact. For notational ease, we denote the constraint set for the parameters $(B, \{\Omega^e\}_{e\in\mathcal{E}})$ as Θ. Assuming such a compactness constraint enables uniform convergence of M-estimators (van de Geer, 2000). As an example of a compactness constraint, let $\tau_1, \tau_2, \tau_3 > 0$ be scalars where for every such $(B, \{\Omega^e\}_{e\in\mathcal{E}})$, $\|(I-B)\|_F \leq \tau_1$, $\min_{i,e}|\Omega^e_{ii}| \geq \tau_2$, and $\max_{i,e}|\Omega^e_{ii}| \leq \tau_3$. Then, the space of connectivity matrices and noise variances has the additional constraints $\|(I-B)\|_F \leq \tau_1$, $\min_{i,e}|\Omega^e_{ii}| \geq \tau_2$ and $\max_{i,e}|\Omega^e_{ii}| \leq \tau_3$ in the score function. We denote the score of a DAG G and intervention set $\mathcal{I}$ when constrained to the compact parameter space Θ as $S(G,\mathcal{I})$.

For any G and intervention set $\mathcal{I}$, we let $S^\star(G,\mathcal{I})$ be the score in population, namely:

$$S^\star(G,\mathcal{I}) = \max_{\substack{B\sim G \\ \text{diagonal } \{\Omega^e\}_{e\in\mathcal{E}} \\ (B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta \\ \text{s.t. } \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}})=\mathcal{I}}} \sum_{e\in\mathcal{E}} \pi^{e,\star}\left[-\ln\det(\Omega^e) - \operatorname{tr}((I-B)^T(\Omega^e)^{-1}(I-B)\Sigma^{e,\star})\right],$$

where $\Sigma^{e,\star}$ represents the true covariance in environment e and $\pi^{e,\star} = \lim_{\text{all } n_e\to\infty} \frac{n^e}{\sum_{e\in\mathcal{E}} n^e}$. Notice that the maximizers $\arg\max_{G,\mathcal{I}} S^\star(G,\mathcal{I})$ are the distributionally equivalent models to the population model $(B^\star, \{\Omega^{e,\star}\}_{e\in\mathcal{E}})$.

First, we show that for every G and intervention set $\mathcal{I}$, the score function $S(G,\mathcal{I}) \xrightarrow{p} S^\star(G,\mathcal{I})$ as $n_e \to \infty$ for every $e \in \mathcal{E}$. This follows by the compactness of the parameter space leading to uniform convergence of M-estimators, and that $\lambda \to 0$ (van de Geer, 2000). For more details, note that:

$$\begin{aligned} S(G,\mathcal{I}) + \lambda\mathrm{DoF}(G,\mathcal{I}) = \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} &\sum_{e\in\mathcal{E}} \pi^{e,\star}\left[-\ln\det(\Omega^e) - \operatorname{tr}((I-B)^T(\Omega^e)^{-1}(I-B)\Sigma^{e,\star})\right] \\ &- \operatorname{tr}((I-B)^T(\Omega^e)^{-1}(I-B)(\hat\pi^e\hat\Sigma^e - \pi^{e,\star}\Sigma^{e,\star})) \\ &- (\hat\pi^e - \pi^{e,\star})\ln\det(\Omega^e) \end{aligned} \tag{30}$$

where $\hat\pi^e = \frac{n^e}{\sum_{e\in\mathcal{E}} n^e}$. From there, we have

$$\begin{aligned} S(G,\mathcal{I}) \geq S^\star(G,\mathcal{I}) &- \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat\pi^e\hat\Sigma^e - \pi^{e,\star}\Sigma^{e,\star}))\right) \\ &- \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat\pi^e - \pi^{e,\star})\ln\det(\Omega^e) - \lambda\ \mathrm{DoF}(G,\mathcal{I}), \\ S(G,\mathcal{I}) \leq S^\star(G,\mathcal{I}) &+ \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat\pi^e\hat\Sigma^e - \pi^{e,\star}\Sigma^{e,\star})\right) \\ &+ \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat\pi^e - \pi^{e,\star})\ln\det(\Omega^e). \end{aligned} \tag{31}$$

By the compactness constraint,

$$\begin{aligned} \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat\pi^e\hat\Sigma^e - \pi^{e,\star}\Sigma^{e,\star})\right) &\to 0 \\ \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat\pi^e - \pi^{e,\star})\ln\det(\Omega^e) &\to 0 \end{aligned}$$

as $n^e \to \infty$. Furthermore, since $\lambda \to 0$ as $n^e \to \infty$ for every $e \in \mathcal{E}$, and $\mathrm{DoF}(G,\mathcal{I})$ is bounded, we have that $\lambda\mathrm{DoF}(G,\mathcal{I}) \to 0$. We can thus conclude $S(G,\mathcal{I}) \to S^\star(G,\mathcal{I})$ in the infinite data limit for every environment.

Suppose the regularization parameter $\lambda \to 0$ and is chosen such that for every $e \in \mathcal{E}$, it is above fluctuations due to sampling error:

$$\lambda \gg \left| \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\pi^{e,\star}\Sigma^{e,\star} - \hat{\pi}^e\hat{\Sigma}^e)\right)\right| + \left| \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e)\right|.$$

With this choice of the regularization, it follows that among any distributionally equivalent models $(G,\mathcal{I})$ and $(G',\mathcal{I}')$, if $\mathrm{DoF}(G,\mathcal{I}) < \mathrm{DoF}(G',\mathcal{I}')$, then, there exists N such that for any $n^e \geq N$, $S(G,\mathcal{I}) > S(G',\mathcal{I}')$. This allows us to conclude that:

$$\mathbb{P}\left(\arg\max S(G,\mathcal{I}) = \arg\min \mathrm{DoF}(G,\mathcal{I}) \text{ subject-to } G,\mathcal{I} \in \arg\max S^\star(G,\mathcal{I})\right) \to 1,$$

as $n^e \to \infty$ for every $e \in \mathcal{E}$. In other words, in the infinite data limit, the maximizers of $\arg\max_{G,\mathcal{I}} S(G,\mathcal{I})$ are given by:

$$\begin{aligned} \arg\min_{G,\mathcal{I}} \quad & \mathrm{DoF}(G,\mathcal{I}) \\ \text{subject-to} \quad & B \sim G, \mathcal{I} = \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}), \text{ and} \\ & \Sigma_e^\star = (I-B)^{-1}\Omega^e(I-B)^{-T} \quad \text{for every } e \in \mathcal{E}. \end{aligned} \tag{32}$$

Our remaining goal is to show that $\mathcal{I}^\star\text{−MEC}(\mathcal{G}(B^\star))$ is the set of minimizers in (32). By definition, $\Sigma_e^\star = (I-B^\star)^{-1}\Omega^{e,\star}(I-B^\star)^{-T}$ for every $e \in \mathcal{E}$. Since the model $(B^\star, \{\Omega^{e,\star}\}_{e\in\mathcal{E}})$ satisfies Assumption 1, we appeal to Lemma 17 to conclude that a feasible B in (32) satisfies $B_{i,:} = B^\star_{i,:}$ for all $i \in \mathcal{I}^\star$. Since $(I-B^\star)^{-1}\Omega^{e,\star}(I-B^\star)^{-T} = (I-B)^{-1}\Omega^e(I-B)^{-T}$, the previous conclusion implies that $\Omega^{e,\star}_{ii} = \Omega^e_{ii}$ for all $i \in \mathcal{I}^\star$ and feasible $\{\Omega^e\}_{e\in\mathcal{E}}$. Thus, any feasible $\mathcal{I}$ satisfies $\mathcal{I} \supseteq \mathcal{I}^\star$. Furthermore, from Theorem 1 we know that:

$$\begin{aligned} \mathcal{I}^\star\text{−MEC}(G^\star) = \arg\min_{G} \quad & \mathrm{DoF}(G) \\ \text{subject-to} \quad & B \sim G \\ & \Sigma_e^\star = (I-B)^{-1}\Omega^e(I-B)^{-T} \quad \text{for every } e \in \mathcal{E}. \end{aligned} \tag{33}$$

Notice that $\mathrm{DoF}(G,\mathcal{I}) = \mathrm{DoF}(G) + \mathrm{DoF}(\mathcal{I})$. Combining this with the relation in (33) and the fact that any feasible $\mathcal{I}$ in (32) satisfies $\mathcal{I} \supseteq \mathcal{I}^\star$, we arrive at the desired result.

Appendix B. Proof of Theorem 2

Proof Throughout, we consider p fixed and $n_e \to \infty$ for every $e \in \mathcal{E}$. We consider a similar compactness constraint as the one described in Section A.7. Since $S^\star(G,\mathcal{I})$ is a maximum-likelihood objective, the maximizers $\arg\max_{G,\mathcal{I}} S^\star(G,\mathcal{I})$ are the distributionally equivalent models to the population model $(B^\star, \{\Omega^{e,\star}\}_{e\in\mathcal{E}})$, that is the set:

$$\{(G,\mathcal{I}) : \exists (B, \{\Omega^e\}_{e\in\mathcal{E}}) \text{ such that } \mathcal{I} = \mathbb{I}(\{\Omega^e\}_{e\in\mathcal{E}}) \text{ and } \Sigma_e^\star = (I-B)^{-1}\Omega^e(I-B)^{-T}\}.$$

Note here that the tuple $(B^\star, \mathcal{I}^\star)$ belongs to this set so that $\max_{G,\mathcal{I}} S^\star(G,\mathcal{I}) = S^\star(G^\star,\mathcal{I}^\star)$. Since the model $(B^\star, \{\Omega^{e,\star}\}_{e\in\mathcal{E}})$ satisfies Assumption 1, we appeal to Lemma 17 to conclude

that a feasible B in (32) satisfies $B_{i,:} = B^\star_{i,:}$ for all $i \in \mathcal{I}^\star$. Since $(I-B^\star)^{-1}\Omega^{e,\star}(I-B^\star)^{-T} = (I-B)^{-1}\Omega^e(I-B)^{-T}$, the previous conclusion implies that $\Omega^{e,\star}_{ii} = \Omega^e_{ii}$ for all $i \in \mathcal{I}^\star$ and feasible $\{\Omega^e\}_{e\in\mathcal{E}}$. Thus, all optimal $\mathcal{I}$ that are maximizers of $\arg\max_{G,\mathcal{I}} S^\star(G,\mathcal{I})$ satisfy the property $\mathcal{I} \supseteq \mathcal{I}^\star$. Furthermore, it is straightforward to see that all $\mathcal{I} \supseteq \mathcal{I}^\star$ are also all optimal with respect to the optimization problem $\arg\max_{G,\mathcal{I}} S^\star(G,\mathcal{I})$.

We are now ready to prove the theorem statement. Let $\mathcal{I}$ be the set of intervention targets at the initial step of the backward phase of Algorithm 2. From the assumption of Theorem 2, we have that $\mathcal{I}^\star \subseteq \mathcal{I}$. Given the consistency of the inner procedure, the backward phase of Algorithm 2 finds $j^\star \in \mathcal{I}$ given by $j^\star = \arg\max_{G,j} S(G, \mathcal{I}\setminus j)$. If $\max_G S(G,\mathcal{I}\setminus j^\star) > \max_G S(G,\mathcal{I})$, then the algorithm removes intervention target j from $\mathcal{I}$. First, we show that the algorithm never removes $j \in \mathcal{I}^\star$. Specifically, decomposing $S(G,\mathcal{I})$ as $S(G,\mathcal{I}) - S^\star(G,\mathcal{I}) + S^\star(G,\mathcal{I})$ and from (30), we obtain:

$$
\begin{aligned}
\max_G S(G,\mathcal{I}) \geq{}& \max_G S^\star(G,\mathcal{I}) - \lambda \max_G \mathrm{DoF}(G,\mathcal{I}) \\
&- \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta, B\sim G} \sum_{e\in\mathcal{E}} \mathrm{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat{\pi}^e\hat{\Sigma}^e - \pi^{e,\star}\Sigma^{e,\star})\right) \\
&- \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta, B\sim G} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e), \\
\max_G S(G,\mathcal{I}\setminus\{j\}) \leq{}& \max_G S^\star(G,\mathcal{I}\setminus\{j\}) \\
&+ \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta, B\sim G} \sum_{e\in\mathcal{E}} \mathrm{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat{\pi}^e\hat{\Sigma}^e - \pi^{e,\star}\Sigma^{e,\star})\right) \\
&+ \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta, B\sim G} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e).
\end{aligned}
$$

By the compactness constraint,

$$
\begin{aligned}
\max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \mathrm{tr}\left((I-B)^T(\Omega^e)^{-1}(I-B)(\hat{\pi}^e\hat{\Sigma}^e - \pi^{e,\star}\Sigma^{e,\star})\right) &\to 0, \\
\max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e) &\to 0,
\end{aligned}
$$

as $n^e \to \infty$. Furthermore, since $\lambda \to 0$ as $n^e \to \infty$ for every $e \in \mathcal{E}$, and $\mathrm{DoF}(G,\mathcal{I})$ is bounded, we have that $\lambda\mathrm{DoF}(G,\mathcal{I}) \to 0$. From the earlier analysis, we have that $\max_G S^\star(G,\mathcal{I}) = S^\star(\mathcal{G}^\star,\mathcal{I}^\star)$ and that $S^\star(\mathcal{G}^\star,\mathcal{I}^\star) > \max_G S^\star(G,\mathcal{I}\setminus\{j\})$. Putting everything together, we conclude that there exists an integer N such that for $n^e \geq N$ for every environment $e \in \mathcal{E}$, $\max_G S(G,\mathcal{I}) > \max_G S(G,\mathcal{I}\setminus\{j\})$. In other words, in the infinite data regime, any node $j \in \mathcal{I}^\star$ will not be removed by the update on $\mathcal{I}$.

We now consider any potential $j \in \mathcal{I}\setminus\mathcal{I}^\star$. We will argue that in the infinite data regime, $\max_G S(G,\mathcal{I}) < \max_G S(G,\mathcal{I}\setminus\{j\})$ so that our algorithm will remove j from $\mathcal{I}$. From (30), we can conclude that:

$$\max_G S(G,\mathcal{I}) \geq \max_G S^\star(G,\mathcal{I}) - \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\Big((I-B)^T(\Omega^e)^{-1}(I-B)(\hat{\pi}^e\hat{\Sigma}^e - \pi^{e,\star}\Sigma^{e,\star})\Big)$$
$$- \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e) - \max_G \lambda \operatorname{DoF}(G,\mathcal{I}),$$
$$\max_G S(G,\mathcal{I}) \leq \max_G S^\star(G,\mathcal{I}) + \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\Big((I-B)^T(\Omega^e)^{-1}(I-B)(\hat{\pi}^e\hat{\Sigma}^e - \pi^{e,\star}\Sigma^{e,\star})\Big)$$
$$+ \max_G \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e).$$

Again, appealing to the compactness constraint, and that $\lambda \to 0$ as $n^e \to \infty$ for every $e \in \mathcal{E}$, we conclude that $\max_G S(G,\mathcal{I}) \xrightarrow{p} \max_G S^\star(G,\mathcal{I})$. Thus, as $n^e \to \infty$, the maximizers $\arg\max_G S(G,\mathcal{I}) \subseteq \arg\max_G S^\star(G,\mathcal{I})$. Since $\mathcal{I} \supseteq \mathcal{I}^\star$, recall that the maximizers $\arg\max_G S^\star(G,\mathcal{I})$ are distributional equivalent models. Further, among distributional equivalent models $(G,\mathcal{I})$ and $(G',\mathcal{I})$, if $\operatorname{DoF}(G,\mathcal{I}) < \operatorname{DoF}(G,\mathcal{I})$, and the regularization λ is above fluctuations due to sampling error

$$\lambda \gg \left| \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} \operatorname{tr}\Big((I-B)^T(\Omega^e)^{-1}(I-B)(\pi^{e,\star}\Sigma^{e,\star} - \hat{\pi}^e\hat{\Sigma}^e)\Big) \right|$$
$$+ \left| \max_{(B,\{\Omega^e\}_{e\in\mathcal{E}})\in\Theta} \sum_{e\in\mathcal{E}} (\hat{\pi}^e - \pi^{e,\star}) \ln\det(\Omega^e) \right|,$$

then there exists N such that for any $n^e \geq N$, $S(G,\mathcal{I}) > S(G',\mathcal{I})$. This allows us to conclude that the maximizers of $\arg\max_G S(G,\mathcal{I})$ are given by the solutions to the following optimization problem

$$\arg\min_G \ \operatorname{DoF}(G) \quad \text{subject-to} \quad B \sim G \ \ ; \ \ \Sigma_e^\star = (I-B)^{-1}\Omega^e(I-B)^{-T} \quad \text{for every } e \in \mathcal{E}.$$

From Theorem 1, we know in the infinite data regime, as $n^e \to \infty$ for every $e \in \mathcal{E}$, $\lambda \to 0$ and λ being above the fluctuations due to sampling error (described above), the maximizes $\arg\max_G S(G,\mathcal{I})$ converge to $\mathcal{I}^\star-\text{MEC}(G^\star)$.

Since $j \notin \mathcal{I}^\star$, a similar argument as above allows us to conclude that $\max_G S(G,\mathcal{I}\setminus\{j\}) \xrightarrow{p} S^\star(G^\star,\mathcal{I}^\star)$ as $n^e \to \infty$ and the maximizers of $\arg\max_G S(G,\mathcal{I}\setminus\{j\})$ converge to $\mathcal{I}^\star-\text{MEC}(G^\star)$. Let N be the integer such that for $n^e \geq N$, maximizers of $\arg\max_G S(G,\mathcal{I}\setminus\{j\})$ and $\arg\max_G S(G,\mathcal{I})$ are both $\mathcal{I}^\star-\text{MEC}(G^\star)$. Since $\operatorname{DoF}(G)$ is the same for every $G \in \mathcal{I}^\star-\text{MEC}(G^\star)$, and that $\operatorname{DoF}(G,\mathcal{I}\setminus\{j\}) + 1 = \operatorname{DoF}(G,\mathcal{I})$, we conclude that for $n^e \geq N$, $\max_G S(G,\mathcal{I}) < \max_G S(G,\mathcal{I}\setminus\{j\})$, as desired.

We have thus concluded that in the first step of the backward phase of the algorithm, we remove some $j \in \mathcal{I}\setminus\mathcal{I}^\star$. We can repeat the same argument to conclude that estimated intervention targets estimated by our algorithm converges to $\mathcal{I}^\star$ and the estimated equivalence class of graphs converges to $\mathcal{I}^\star-\text{MEC}(G^\star)$.

■

Appendix C. Violation of score equivalence by the ℓ_0-penalized score

We offer some discussion and an example of how the score corresponding to the objective in (6), i.e., the ℓ_0-penalized likelihood score from Chickering (2002), does not satisfy score equivalence. Namely, graphs in the same $\mathcal{I}$-equivalence class may obtain different scores on the same finite sample. For compactness, we state the score corresponding to (6) here

$$\mathcal{S}(G) := \max_{\substack{B \sim G \\ \text{diagonal pos. def. } \{\Omega^e\}_{e\in\mathcal{E}}}} \sum_{e\in\mathcal{E}} \log P(\boldsymbol{X}^e; B, \Omega^e) - \lambda \left\|\mathcal{G}(B)\right\|_0, \tag{34}$$

The likelihood term can be formulated in terms of the sample covariances and covariances implied by the model for each environment $e \in \mathcal{E}$, that is

$$\log P\left(\boldsymbol{X}^e; B, \Omega^e\right) = -n_e \ln \det(K^e) - n_e \mathrm{tr}(K^e \hat{\Sigma}^e), \tag{35}$$

where $\hat{\Sigma}^e$ is the sample covariance for environment e, and $K^e := [(I-B)^{-1}\Omega^e(I-B)^{-T}]^{-1}$ is the inverse of the covariance matrix implied by B and Ω^e for the graph $\mathcal{G}$. Thus, a necessary condition for two graphs to attain the same score on any finite sample is that they produce the same sets of covariance matrices under the constraints in (34). For a single environment ($|\mathcal{E}| = 1$, i.e., $\mathcal{I} = \emptyset$), this is true for any two Markov-equivalent graphs (see proof of Proposition 5 in Appendix A.7). However this does not hold for $|\mathcal{E}| > 1$, because—without the additional (sufficient) constraints placed on the noise-term matrices $\{\Omega^e\}_{e\in\mathcal{E}}$ by the GnIES score in (7)—two $\mathcal{I}$-equivalent graphs may yield disjoint sets of covariance matrices for each of the environments in $\mathcal{E}$.

As an example, consider the following two graphs: $X_1 \to X_2$, which we denote by $\mathcal{G}$, and $X_1 \leftarrow X_2$, which we denote by $\tilde{\mathcal{G}}$. The two are $\mathcal{I}$-equivalent for $\mathcal{I} = \emptyset$. Let $(B, \{\Omega, \Omega'\})$ and $(\tilde{B}, \{\tilde{\Omega}, \tilde{\Omega}'\})$ be any two models that satisfy the constraints in (34) for $\mathcal{G}$ and $\tilde{\mathcal{G}}$, respectively. Additionally, let $\Omega_{11} = \Omega'_{11}$, $\Omega'_{22} \neq \Omega_{22}$, and $B_{21} = 1$. For the first environment, the covariance implied by the first model (graph $\mathcal{G}$) is given by

$$\Sigma = (I-B)^{-1}\Omega(I-B)^{-T} = \begin{pmatrix} \Omega_{11} & B_{21}\Omega_{11} \\ B_{21}\Omega_{11} & B_{21}^2\Omega_{11} + \Omega_{22} \end{pmatrix},$$

and the covariance implied by $\tilde{\mathcal{G}}$ is

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{B}_{12}^2\tilde{\Omega}_{22} + \tilde{\Omega}_{11} & \tilde{B}_{12}\tilde{\Omega}_{22} \\ \tilde{B}_{12}\tilde{\Omega}_{22} & \tilde{\Omega}_{22} \end{pmatrix}.$$

The two covariances are equal if and only if (note: $B_{21} = 1$)

$$\tilde{\Omega}_{11} = \Omega_{11} - \frac{\Omega_{11}}{\Omega_{11}+\Omega_{22}},\ \tilde{\Omega}_{22} = \Omega_{11} + \Omega_{22},\ \text{and}\ \ \tilde{B}_{12} = \frac{\Omega_{11}}{\Omega_{11}+\Omega_{22}}. \tag{36}$$

For the second environment, the covariances matrices are given by

$$\Sigma = \begin{pmatrix} \Omega'_{11} & B_{21}\Omega'_{11} \\ B_{21}\Omega'_{11} & B_{21}^2\Omega'_{11} + \Omega'_{22} \end{pmatrix} \text{ and } \tilde{\Sigma}' = \begin{pmatrix} \tilde{B}_{12}^2\tilde{\Omega}'_{22} + \tilde{\Omega}'_{11} & \tilde{B}_{12}\tilde{\Omega}'_{22} \\ \tilde{B}_{12}\tilde{\Omega}'_{22} & \tilde{\Omega}'_{22} \end{pmatrix}.$$

As before, the two covariances are equal if and only if

$$\tilde{\Omega}'_{11} = \Omega'_{11} - \frac{\Omega'_{11}}{\Omega'_{11} + \Omega'_{22}},\ \tilde{\Omega}'_{22} = \Omega'_{11} + \Omega'_{22},\ \text{and}\ \ \tilde{B}_{12} = \frac{\Omega'_{11}}{\Omega'_{11} + \Omega'_{22}} = \frac{\Omega_{11}}{\Omega_{11} + \Omega'_{22}}.$$

However, this is a contradiction since the value obtained for $\tilde{B}_{12}$ above differs from that in (36), which is not possible as $\tilde{B}_{12}$ is defined as being constant across environments.

Appendix D. Example: Violation of Assumption 1

We provide an example of how violations of Assumption 1 affect our main result in Theorem 1, and thus break the correspondence between our equivalence class and transition sequence equivalence introduced by Tian and Pearl (2001, Theorems 2 and 3).

Consider a model $(B, \{\Omega\})$ defined over a single environment. Assuming faithfulness, Proposition 6 states that the class of distribution equivalent models with minimal edges corresponds to the Markov equivalence class of the graph $\mathcal{G}(B)$. Consider the case where this class is not a singleton, and let $(\tilde{B}, \{\tilde{\Omega}\}) \in [(B, \{\Omega\})]$ be a distinct model that is distribution equivalent and has the same number of edges. By Definition 1, it holds that

$$\tilde{\Omega} = (I - \tilde{B})(I - B)^{-1}\Omega(I - B)^{-T}(I - \tilde{B})^T$$

is a diagonal matrix with positive elements. Assume now that—in violation of Assumption 1—we add an additional environment, where the noise-term variances of all variables have been increased by a factor $\alpha > 0$, that is, $\Omega' := \alpha\Omega$. Because this amounts to an intervention on all variables, by Theorem 1 the equivalence class of $(B, \{\Omega, \Omega'\})$ should be a singleton, i.e., there should exist no $\tilde{\Omega}'$ such that

$$\tilde{\Omega}' = (I - \tilde{B})(I - B)^{-1}\Omega'(I - B)^{-T}(I - \tilde{B})^T$$

is a diagonal matrix with positive entries. However, substituting $\Omega' = \alpha\Omega$ yields

$$\tilde{\Omega}' = (I - \tilde{B})(I - B)^{-1}\alpha\Omega(I - B)^{-T}(I - \tilde{B})^T = \alpha\tilde{\Omega}$$

which is indeed diagonal with positive entries. Thus, even after this intervention on all variables, $(\tilde{B}, \{\tilde{\Omega}, \tilde{\Omega}'\})$ remains inside the equivalence class of $(B, \{\Omega, \Omega'\})$.

Appendix E. GnIES Completion Algorithm

Below we detail the completion algorithm used at the end of each step of the inner procedure from GnIES. The algorithm is also employed in the experiments to compute an estimate of the equivalence class from the graph and intervention targets returned by UT-IGSP.

Algorithm 3 Completion algorithm for the inner procedure of GnIES

1: **Input**: PDAG P, intervention set $\mathcal{I}$.
Precondition: The PDAG P has no undirected in- or out-going edges from the nodes in $\mathcal{I}$; this is guaranteed during the inner procedure of GnIES.
2: **Obtain the CPDAG** C by completing P using the GES completion algorithm (Chickering, 2002, Appendix C)
3: **Orient edges**: For every node in $\mathcal{I}$, orient all in- and out-going edges to match the orientation in the original PDAG P
4: **Apply the Meek rules** (Meek, 1995) iteratively to C until no more edges can be oriented.
5: **Output:** The I-CPDAG C representing the $\mathcal{I}-$MEC.

Appendix F. Software Contributions

The code, data sets and instructions to reproduce the experiments and figures in the manuscript can be found in the repository `github.com/juangamella/gnies-paper`. The Python implementation of the GnIES algorithm is available through the `gnies` package (see details below). During our work on this paper, we developed additional software which we make available through the Python packages listed below. The packages are well documented, actively maintained, and available through the python package index[16], that is, they can easily be installed with Python's package manager by typing `pip install <package_name>` into a suitable terminal. More details and installation instructions can be found in their respective repositories, listed below.

Package	Description	Details and documentation
`gnies`	The GnIES algorithm.	`github.com/juangamella/gnies`
`ges`	A pure Python implementation of the GES algorithm (Chickering, 2002).	`github.com/juangamella/ges`
`gies`	A pure Python implementation of the GIES algorithm from Hauser and Bühlmann (2012).	`github.com/juangamella/gies`
`sempler`	Used to generate the synthetic and semi-synthetic data used in the experiments.	`github.com/juangamella/sempler`

Appendix G. Semi-synthetic Data Generation Procedure

In this section we summarize the procedure used to generate the semi-synthetic data set from Section 5, generated by combining the biological data set and consensus graph given

16. See `https://pypi.org/`.

in Sachs et al. (2005). The procedure is general in that it can be used with any other data set and graph; an implementation is provided in the python package[17] `sempler`.

As input, we are given a directed acyclic graph $\mathcal{G}$ over p variables and a data set consisting of their observations under different environments. The procedure then fits a nonparametric Bayesian network which approximates the conditional distributions entailed by the factorization of the graph $\mathcal{G}$ via distributional random forests (Cevid et al., 2022). Once fitted, we can sample from this collection of forests to produce a new, semi-synthetic data set that respects the conditional independence relationships entailed by $\mathcal{G}$, while its marginal and conditional distributions closely match those of the original data set (Figures 7 and 8).

The procedure is motivated by a fundamental concept from graphical models. We say a distribution P satisfies the *global Markov* property with respect to a DAG $\mathcal{G}$ if and only if it factorizes according to its parental relations (Pearl, 1988), that is, if

$$P(X_1, ..., X_p) = \prod_{j \in [p]} P(X_j \mid X_{\mathrm{PA}_{\mathcal{G}}(j)}).$$

Thus, we can model the joint data-generating distribution by independently modeling the conditional distribution of each variable given its parents in $\mathcal{G}$. We take this approach and model the collection of data-generating distributions $\{P^e\}_{e \in \mathcal{E}}$ as follows: for each environment $e \in \mathcal{E}$ and each node $j \in [p]$, we estimate $\hat{P}^e_{X_j | X_{\mathrm{PA}_{\mathcal{G}}(j)} = x}$ by fitting a distributional random forest (Cevid et al., 2022). Note that while the distributions $\{P^e\}_{e \in \mathcal{E}}$ of the data may not satisfy the Markov condition with respect to $\mathcal{G}$, the estimated distributions $\{\hat{P}^e\}_{e \in \mathcal{E}}$ do.

We can now generate new data by sampling from the estimated conditional distributions, following the causal ordering of the DAG $\mathcal{G}$. We proceed as follows:

1. First, we obtain the causal ordering $O_{\mathcal{G}}$ of $\mathcal{G}$.

2. Then, for every environment $e \in \mathcal{E}$ and the next (or first) node in $O_{\mathcal{G}}$:

   (a) if j is a source node, we generate a sample $\left\{\hat{x}^e_j(i)\right\}_{i=1,...,n_e}$ via bootstrapping, where n_e is the number of observations from environment e in the original data set.

   (b) If j is not a source node, for every new observation $i = 1, ..., n_e$ we sample

   $$\hat{X}^e_j(i) \sim \hat{P}^e_{X_j | X_{\mathrm{PA}_{\mathcal{G}}(j)} = \hat{x}^e_{\mathrm{PA}_{\mathcal{G}}(j)}(i)}$$

   using the technique provided by Cevid et al. (2022), where $\hat{x}^e_{\mathrm{PA}_{\mathcal{G}}(j)}(i)$ is the previously sampled observation for the parents of j.

3. We collect all the samples, resulting in a new semi-synthetic data set.

17. See `github.com/juangamella/sempler` for more details.

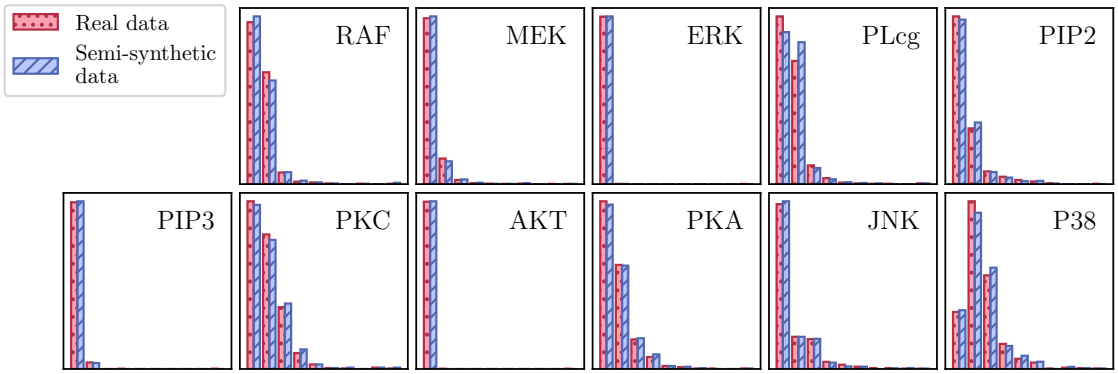


Figure 7: Histograms of the observations of each variable measured in the Sachs data set, for the real data (red, dots) and the semi-synthetic data (blue, stripes) sampled from the network fitted to the data set using distributional random forests (Cevid et al., 2022) following the consensus network. The marginals of the semi-synthetic data closely resemble those of the real data, and both show strong indications of non-normality.

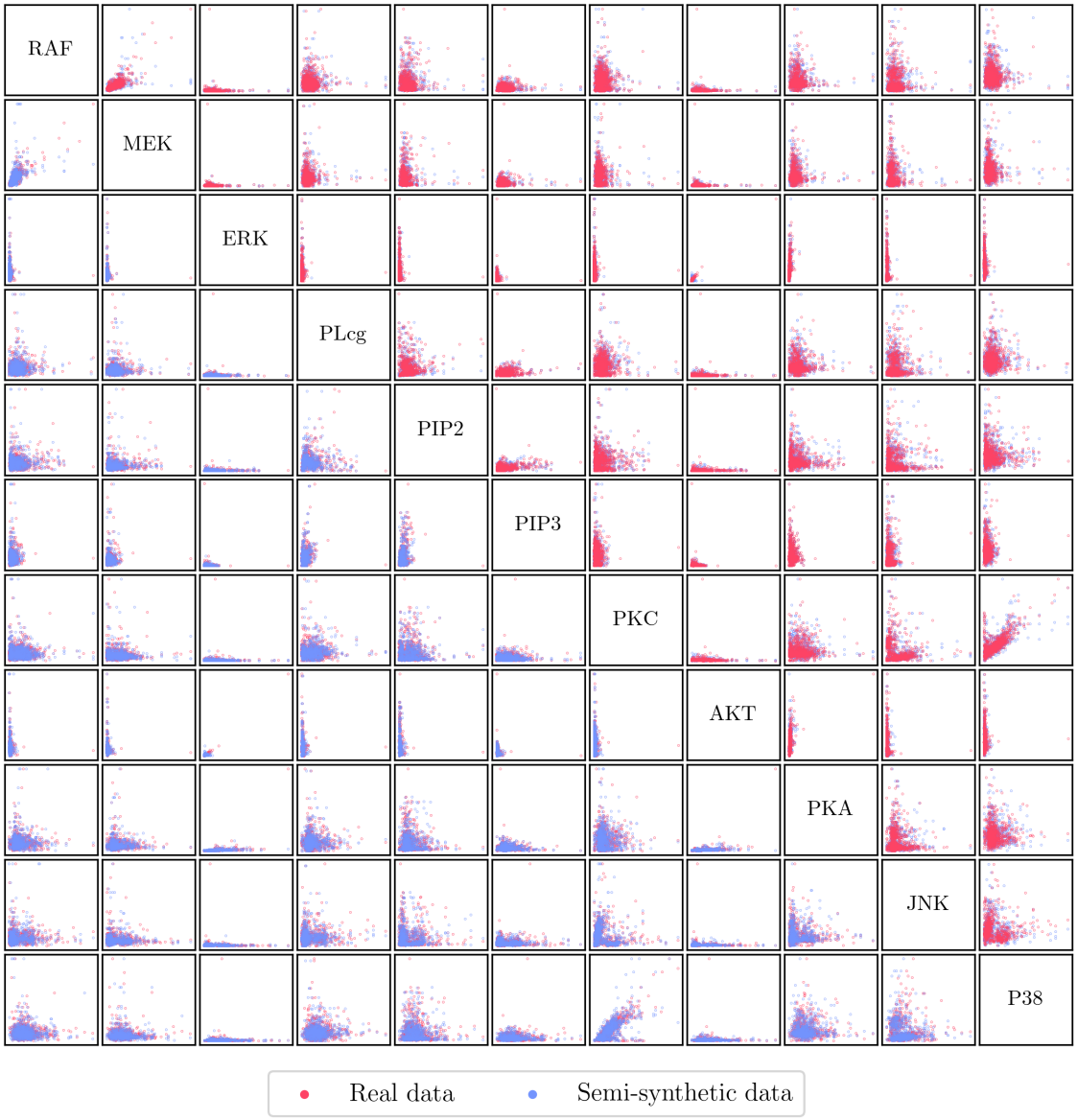


Figure 8: Pairwise observations of all variables in the Sachs data set, for the real (red) and semi-synthetic (blue) data. The conditional distributions of the semi-synthetic data set closely follow those of the real data set, and both exhibit a high degree of non-linearity.

Appendix H. Additional Synthetic Experiments

In this section we provide the results of additional synthetic experiments to evaluate GnIES and the compared methods on data from denser ground-truth graphs (Figure 9), multiple target interventions (Figure 10), and a misspecified intervention model (Figure 11). In all

figures, we evaluate the recovery of the ground-truth equivalence class using the metrics described in Definition 4; the point $(0, 1)$ corresponds to perfect recovery. The results are an average across 1000 data sets originating from 100 linear Gaussian SCMs, which are otherwise generated following the procedure in Section 5.1. We show the regularization path for all methods. For GnIES, GIES and GES, the regularizer is set to $\lambda := \lambda' \log(N)$ where N is the total number of observations and $\lambda' \in [0.01, 2]$; the additional point shown on the regularization path of these methods corresponds to the BIC score ($\lambda' = 1/2$). For UT-IGSP, the regularizers are the levels of the conditional independence and invariance tests, taken between 10^{-5} and 0.1. The poor performance of sortnregress in all experiments indicates the low varsortability in the data after it is standardized.

H.1 Misspecified intervention model

We evaluate the performance of GnIES on data from a misspecified intervention model, where the interventions are *do*- or hard interventions (Pearl, 2009b, Section 3.2.2) instead of the noise interventions described in Section 2. In particular, the interventions remove the causal effect of the parents on the target and set it to a normal distribution with mean zero and variance sampled uniformly at random from $[3, 4]$. The data is otherwise generated as in Section 5.1.

We show the results in Figure 11. We employ GIES (Hauser and Bühlmann, 2012) as an additional baseline, as this setting precisely satisfies the modeling assumptions of the method save for one aspect: that knowledge of the intervention targets is required. To overcome this we provide GIES with the true intervention targets, thus acting as an indicator of the "optimum performance" we would hope to obtain if we had such knowledge. Furthermore, the equivalence class of the data-generating model is no longer the $\mathcal{I}$-equivalence class described in this paper, but rather the interventional Markov equivalence class introduced by Hauser and Bühlmann (2012). In turn, we now use the completion algorithm from GIES to compute the equivalence class estimated by UT-IGSP, from its estimated DAG and intervention targets; we note that, in practice, this would require knowledge of the model violation. Because UT-IGSP allows for a wider range of intervention types, its performance is essentially undisturbed by the change to hard interventions on the same targets. The results are shown in Figure 11 below. As expected, GnIES is sensitive to the misspecification and performs worse than GIES. The relative advantage of GnIES over UT-IGSP is reduced and then reversed as the sample size increases.

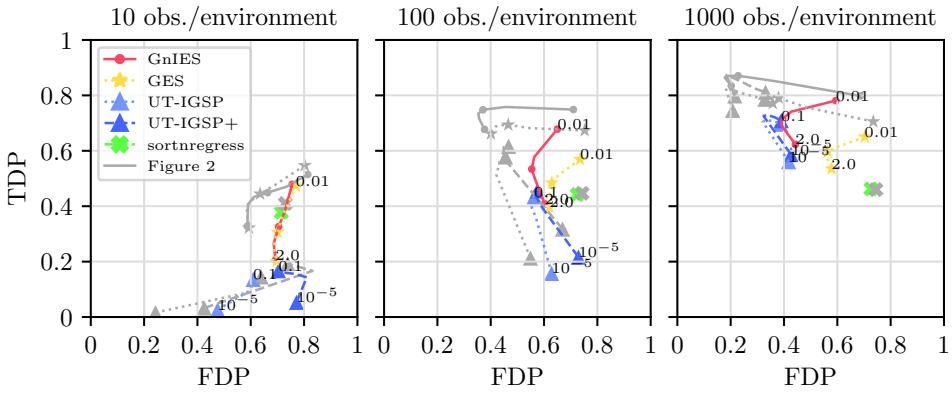


Figure 9: (denser graphs) Comparing the effect of denser ground-truth graphs with the results from Figure 4 (shown in gray). The ground-truth models and datasets are generated as in Section 5.1, but we increase the average degree of the underlying Erdös-Renyi graphs from 2.7 to 4. As expected, the performance of all methods deteriorates, but their relative performance remains similar.

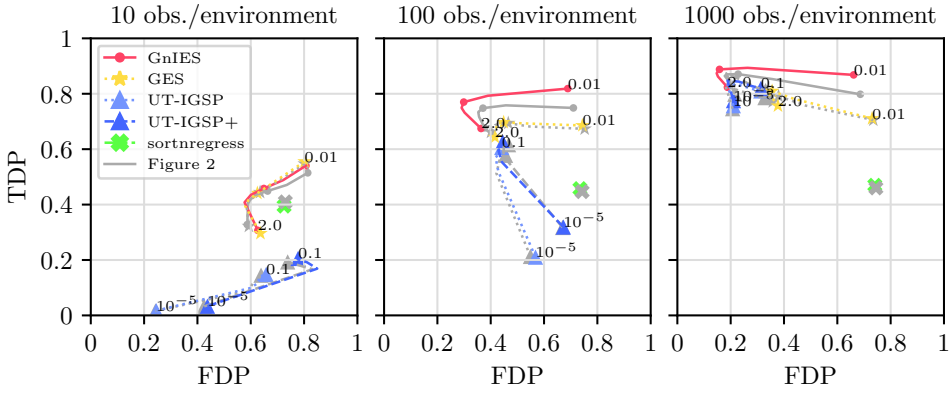


Figure 10: (multiple target interventions) Comparing the effect of multiple-target interventions with the results from Figure 4 (shown in gray). The datasets are generated as in Section 5.1, but the number of intervention targets in each environment is increased to 3. The performance of the compared methods remains virtually unchanged or even improves slightly for GnIES in the larger sample sizes.

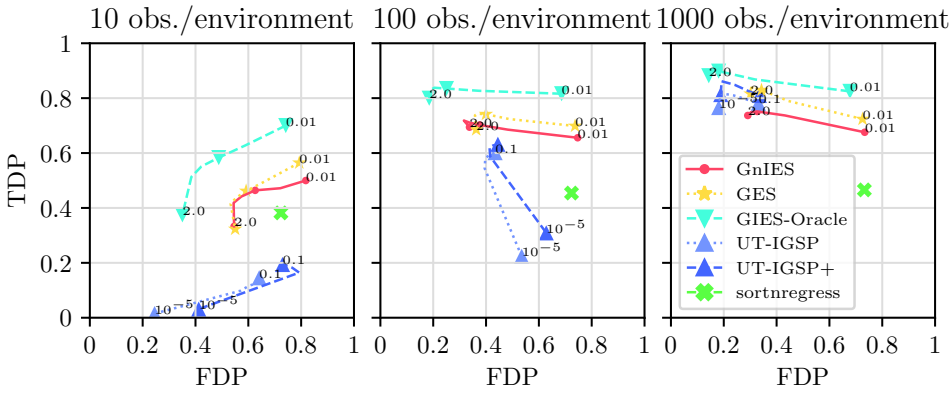


Figure 11: (misspecified intervention model) In this case, the true equivalence class is the interventional Markov equivalence class given in (Hauser and Bühlmann, 2012) for hard interventions (c.f. $\mathcal{H}-$MEC in Section 3.1.1). GIES is run with full knowledge of the true intervention targets. Because UT-IGSP returns a single DAG as an estimate, we use its estimated targets and the completion algorithm of GIES to produce an estimate of the equivalence class (displayed as UT-IGSP+). For each method, the regularization path is shown.

Appendix I. Additional Figures

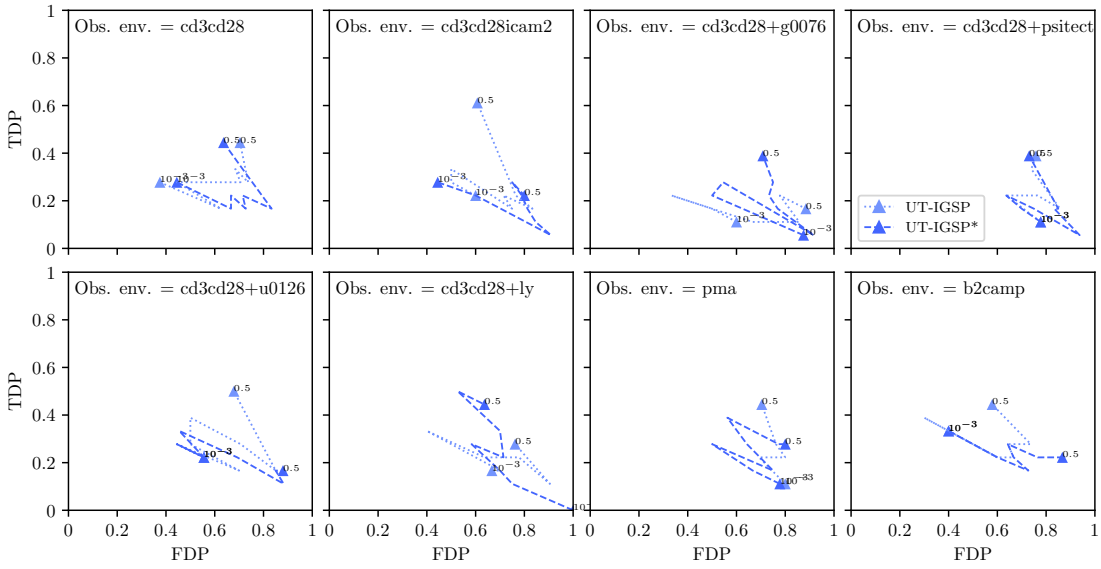


Figure 12: Performance of UT-IGSP on the Sachs data set for other choices of the observational environment. The remaining environment `cd3cd28+aktinhib` is displayed in Figure 5 in the main text.

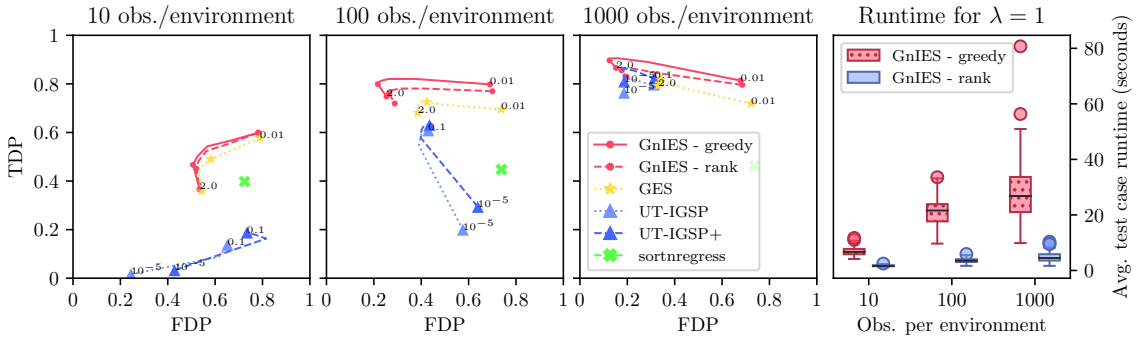


Figure 13: Recovery of the interventional Markov equivalence class for GnIES with the ranking outer procedure (red, circles, dashed) compared to the greedy approach and other algorithms. Bottom right: its runtime compared to GnIES with a greedy outer procedure. The data sets, metrics, and other algorithms are the same as in Figure 4. At the cost of a slightly reduced performance in class recovery, the ranking procedure results in a significant speed-up. The relative weight of the penalization term in the score decreases as the sample size increases, resulting in more targets being added and a higher runtime for both approaches.

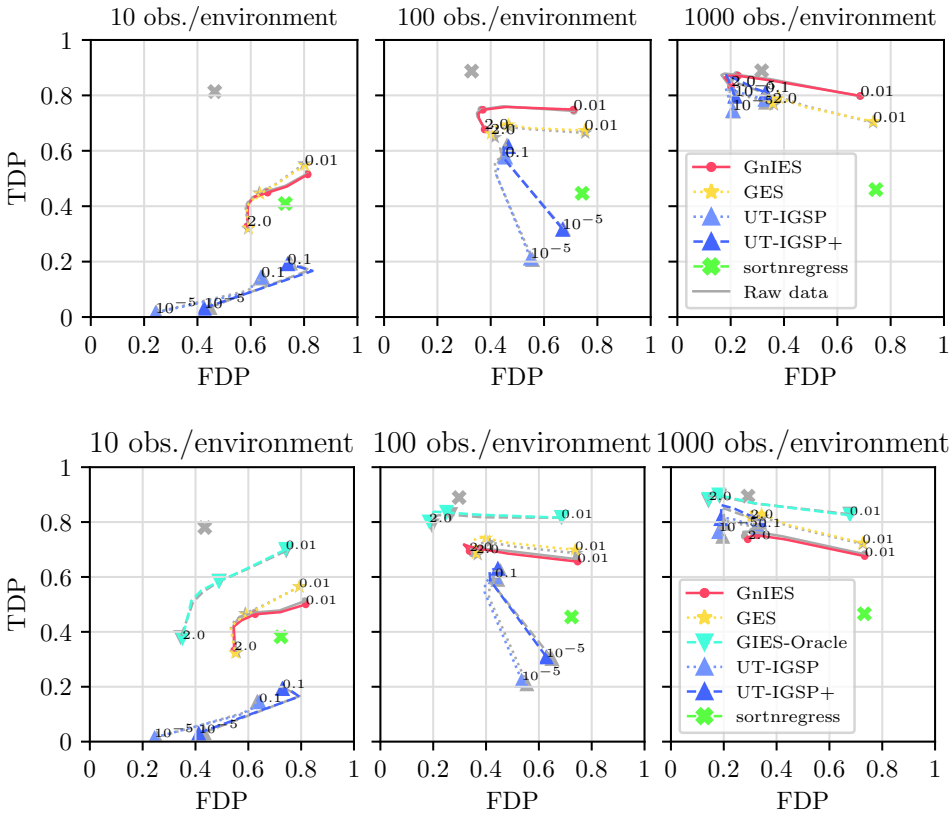


Figure 14: Performance of GnIES and the other algorithms when fed standardized (color) and raw (gray) data, for the settings with a model match (top, as in Figure 4) and a model mismatch in the form of *hard* interventions (bottom, as in Figure 11). The sortnregress algorithm explicitly exploits varsortability, and its performance is heavily affected when it is removed from the data via standardization. The other algorithms, including GnIES, show no significant change in performance; this can be taken as an indication that they do not rely on varsortability for inference. In both settings, the average varsortability score (Reisach et al., 2021, Section 3) is around 0.9 for the raw data and 0.5 for the synthetic data.

References

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526

– 2556, 2014.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 87–98, 1995.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

David Maxwell Chickering. Statistically efficient greedy equivalence search. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 241–249, 2020.

David Maxwell Chickering and Christopher Meek. Selective greedy equivalence search: finding optimal Bayesian networks using a polynomial number of score evaluations. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 2015.

Vanessa Didelez. Statistical causality. *Consilience Interdisciplinary Communications*, 2996: 114–120, 2005.

Frederick Eberhardt. Causation and intervention. *PhD Thesis, Carnegie Mellon University*, 2007.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

Gonçalo Rui Alves Faria, Andre Martins, and Mario A. T. Figueiredo. Differentiable causal discovery under latent interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 253–274, 2022.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475, 2020.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-024-00964-x.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. *Advances in Neural Information Processing Systems*, 30:3015–3025, 2017.

Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, 111(1):3, 2004.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. BaCaDI: Bayesian causal discovery with unknown interventions. In *Proceedings of the Twenty-Sixth International Conference on Artificial Intelligence and Statistics*, pages 1411–1436, 2023.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55 (4):926–939, 2014.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):291–318, 2015.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

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

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 22:689–696, 2009.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: characterization and learning. *Advances in Neural Information Processing Systems*, 33:9551–9561, 2020.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C. Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. *Advances in Neural Information Processing Systems*, 32:14369–14379, 2019.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. *Advances in Neural Information Processing Systems*, 29:4473–4481, 2016.

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

Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Joris M. Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 431–439, 2013.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(Jun):1–108, 2022.

Paul Muentener and Elizabeth Bonawitz. The development of causal reasoning. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press, 2017.

Preetam Nandy, Alain Hauser, and Marloes Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009a.

Judea Pearl. *Causality*. Cambridge University Press, 2009b.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 589–598, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

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

Niklas Pfister, Evan G. Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3): 1220–1246, 2021.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 83(2):215–246, 2021.

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

Jonas Seng, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Tearing apart NOTEARS: Controlling the graph prediction via variance manipulation. *arXiv preprint arXiv:2206.07195*, 2022.

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

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, Causation, and Discovery*, 21:211–252, 1999.

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

Chandler Squires, Joshua Amaniampong, and Caroline Uhler. Efficient permutation discovery in causal DAGs. *arXiv preprint arXiv:2011.03610*, 2020a.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048, 2020b.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

Armeen Taeb, Juan L. Gamella, Christina Heinze-Deml, and Peter Bühlmann. Learning and scoring Gaussian latent variable causal models with unknown additive interventions. *Journal of Machine Learning Research*, 25(293):1–68, 2024.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 512–521, 2001.

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Sara van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30:5824–5833, 2017.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the Thirty-Fifth International Conference on Machine Learning*, pages 5541–5550, 2018.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the Thirty-Sixth International Conference on Machine Learning*, pages 7154–7163, 2019.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, volume 2017, pages 1347–1353, 2017.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P Xing. DAGs with no tears: continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31:9492–9503, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics*, pages 3414–3425, 2020.