Journal of Machine Learning Research 25 (2024) 1-68 Submitted 8/22; Published 9/24

Learning and scoring Gaussian latent variable causal models with unknown additive interventions

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

Juan L. Gamella JUAN.GAMELLA@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zürich

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

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

Editor: Vanessa Didelez

Abstract

With observational data alone, causal structure learning is a challenging problem. The task becomes easier when having access to data collected from perturbations of the underlying system, even when the nature of these is unknown. Existing methods either do not allow for the presence of latent variables or assume that these remain unperturbed. However, these assumptions are hard to justify if the nature of the perturbations is unknown. We provide results that enable scoring causal structures in the setting with additive, but unknown interventions. Specifically, we propose a maximum-likelihood estimator in a structural equation model that exploits system-wide invariances to output an equivalence class of causal structures from perturbation data. Furthermore, under certain structural assumptions on the population model, we provide a simple graphical characterization of all the DAGs in the interventional equivalence class. We illustrate the utility of our framework on synthetic data as well as real data involving California reservoirs and protein expressions. The software implementation is available as the Python package *utlvce*.

Keywords: directed acyclic graphs, invariance, structural causal models, equivalence classes

1. Introduction

Identifying causal relations from observational data alone is challenging. In the context of (acyclic) structural causal models (Robins et al., 2000; Pearl, 2009), one possibility is to find the *Markov equivalence class* (MEC) of the underlying directed acyclic graph (DAG) under the faithfulness assumption (Verma and Pearl, 1991) or the beta-min condition (van de Geer and Bühlmann, 2013). Some of the well-known algorithms for structure learning of MECs with observational data include the constraint-based PC algorithm (Spirtes et al., 2000), the score-based Greedy Equivalence Search (GES) algorithm (Chickering, 2002), and

hybrid methods that integrate constraint-based and score-based methods such as ARGES (Nandy et al., 2018).

In contrast to the purely observational setting, randomized controlled experiments lie at the opposite pole (Rubin, 2015): they are the gold standard for causal inference but randomizing the treatment is often hindered by cost, feasibility, or ethical concerns. However, under some assumptions, it is possible to exploit unspecific interventions or perturbations in the underlying system of interest which may not have been explicitly designed and controlled by a human experimenter. Such interventions arise in many application domains. For example, in genomics, with the advance of gene editing technologies, high throughput interventional gene expression data is being produced (Kemmeren et al., 2014; Dixit et al., 2016; Meinshausen et al., 2016). While there is typically a particular gene that is targeted by an intervention in a particular experiment, there may be additional off-target effects whose nature is unknown. In this paper, we assume that we have access to interventional data from different so-called "environments" where the location and strength of the respective interventions do not have to be known.

Interventional data can be viewed as *perturbations* to components of the system and can offer substantial gain in identifiability: Hauser and Bühlmann (2012) demonstrated that combining interventional with observational data reduces ambiguity and enhances identifiability to a smaller equivalence class than the MEC, known as the I-MEC (Interventional MEC). A variety of methods have been proposed for causal structure learning from observational and interventional data. This includes the modified GES algorithms by Hauser and Bühlmann (2012); Gamella et al. (2022), permutation-based causal structure learning for observational data (Wang et al., 2017) and for interventional data (Squires et al., 2020), penalized maximum-likelihood procedure in Gaussian models (Hauser and Bühlmann, 2015), the Joint Causal Inference framework based on conditional independence testing (Mooij et al., 2020), and methods based on a causal invariance framework (Meinshausen et al., 2016; Peters et al., 2016; Rothenhäusler et al., 2016, 2019, 2021; Ghassami et al., 2017; Heinze-Deml et al., 2018b; Huang et al., 2020) building on a concept of stability (Dawid and Didelez, 2010; Dawid, 2021). For a more comprehensive list, see also Drton and Maathius (2017); Heinze-Deml et al. (2018a) and the references therein.

One reason why randomized controlled experiments are considered to be the gold standard for causal inference is that the randomization breaks the influence potential hidden confounders have on both the treatment as well as the response variable of interest. In less controlled settings, the presence of latent variables, which may be difficult to measure or are simply unknown, poses a major challenge as the causal graphical model structure is not closed under marginalization. Therefore, the graphical structure corresponding to the marginal distribution of the observed variables consists of potentially many confounding dependencies that are induced due to the marginalization over the latent variables.

In this paper, we propose a modeling framework and estimator that allows for unspecific interventions on some or all of the variables. Figure 1 demonstrates a toy example of our setup among 4 observed variables (X_1, X_2, X_3, X_4), latent variables H, and the environment variables $\mathcal{E}$ representing exogenous effects (to the graphical structure among observed and latent variables) that provide additive interventions to the observed and latent variables.

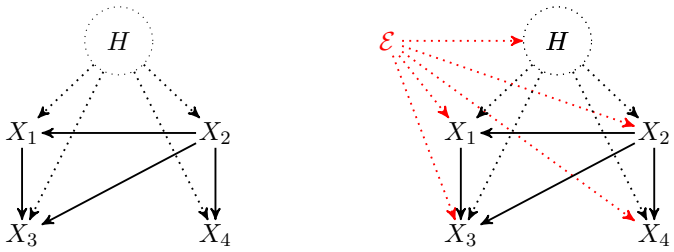


Figure 1: Toy illustration of the setting considered in this paper where X represent the observed variables and H represent the latent variables where solid lines are connections among observed variables and dotted lines are connections between observed and latent variables; **left**: without interventions, **right**: interventions $\mathcal{E}$ on all components indicated with red dotted lines.

Formally, we study a linear structural causal model (SCM) specifying the interventional model and the relationship between p observed variables $X \in \mathbb{R}^p$ and h latent variables $H \in \mathbb{R}^h$. Here, the latent variables are assumed to act exogenously on the observed variables, but unless otherwise specified, no assumption is placed on the dependence structure among the latent variables. The SCM is parameterized by a connectivity matrix encoding the DAG structure among the observed variables, a coefficient matrix encoding the latent variable effects, and parameters involving the noise variances and unknown additive intervention magnitudes and locations among all of the variables. Using data from this model, our objective is to estimate the DAG among the observed variables (e.g. the solid dotted lines in Figure 1) or an equivalence class of DAGs when the underlying structure is not identifiable.

A key property of our modeling framework is that the connectivity matrix and the latent variable coefficient matrix remain *invariant* across all the intervention environments. With this insight, we propose a regularized maximum-likelihood procedure – dubbed (U)nknown (T)arget (L)atent (V)ariable (C)ausal (E)stimator (*UT-LVCE*) – to score any given DAG and estimate the associated parameters. Using the given DAG structure and the learned intervention locations on the observed variables, we then provide a simple graphical approach to identify an equivalence class of DAGs that yield the same fit to the data. We show that to identify the population DAG among the observed variables, it is necessary to impose constraints on the latent effects. Otherwise, the problem is ill-posed. Under certain conditions on the population model, we demonstrate that applying this graphical procedure to the underlying DAG and the intervention locations of the observed variable fully characterizes, in the infinite data limit, the equivalence class of optimally scoring DAGs. Furthermore, under sufficiently many interventions, the optimally scoring DAG is uniquely the population DAG structure among the observed variables. Our characterization of the optimally scoring DAGs is valid under two types of structural assumptions on the latent effects: the first assumption is that the number of latent variables is small (compared to the observed variables) and they affect many observed variables; the second assumption substantially relaxes the first assumption and requires very mild conditions on the latent effects at the expense of approximately knowing the magnitude of the latent interventions.

We envision several use cases for *UT-LVCE*. Firstly, in certain application domains, a DAG structure may be believed to approximate the underlying phenomenon (for example, protein expressions as in Section 5). *UT-LVCE* can be used to learn a latent variable causal

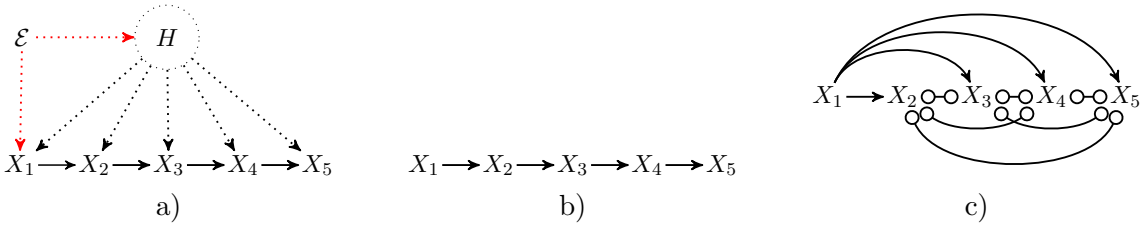


Figure 2: An illustration comparing the output of our approach *UT-LVCE* and approaches that do not impose any constraints on the latent effects such as JCI (Mooij et al., 2020). a) the setup consisting of five observed variables, one latent variable and interventions on the latent variable and observed variable X_1, b) the interventional equivalence class that *UT-LVCE* recovers (under some conditions) consisting uniquely of the subgraph among observed variables, c) the interventional equivalence class that (Mooij et al., 2020) recovers where many causal effects are not identified; here, directed edges denote ancestral relationships and circle marks represent uncertainty about edge marks.

model with respect to this DAG and return an equivalence class of DAGs that fit the data equally well. Secondly, along similar lines, a set of candidate DAGs, instead of only a singleton, may be available based on prior knowledge. In such settings, each DAG in this collection may be scored, and the best scoring ones as well as the DAGs in their respective equivalence class may be returned as output. Thirdly, the input candidate DAGs may be viewed as 'starting points' that may contain spurious edges (obtained by domain expertise or by any structure learning algorithm) where the user aims to improve on these DAGs. Here, we propose to apply *UT-LVCE* on top of a greedy backward deletion approach to remove spurious dependencies due to latent confounding and identify an equivalence class of best scoring DAGs.

1.1 Related work

A large body of causal structure learning methods with latent variables typically characterize a class of graphical independence models called maximal ancestral graphs (MAGs) (Spirtes et al., 2000; Richardson and Spirtes, 2002; Jaber et al., 2020; Bhattacharya et al., 2021). These methods, which allow for arbitrary hidden structure, tend to be overly conservative, recovering only a small subset of the causal effects. For example, suppose a latent variable influences many observed variables. Then, the underlying MAG tends to be dense where many edges cannot be directed. In this work, we take a middle-ground stance and place assumptions on the latent effects; these assumptions then enable us to direct edges and learn the sub-graph among observed variables (see the illustration in Figure 2). A similar perspective was taken in Frot et al. (2019) but without incorporating interventional data.

In the joint observational and interventional setting with unspecified interventions and latent confounders, several methods exist in the literature for either learning the sub-graph among the observed variables or the causal parents (among the observed variables) of a target variable of interest. In particular, with unperturbed latent variables and only so-called shift interventions on the observed covariates, Causal Dantzig (Rothenhäusler et al., 2019) consistently estimates the causal effects on a response variable assuming that the interventions do not directly affect the response variable. Such an assumption is relaxed

Method	Perturbed response	Unperturbed latent	Perturbed latent
IV, ICP, Causal Dantzig	x	✓ single DAG	x
backShift	✓ single DAG	✓ single DAG	x
UT-LVCE	✓ $\mathcal{I}$-MEC	✓ $\mathcal{I}$-MEC	✓ $\mathcal{I}$-MEC

Table 1: Comparison of *UT-LVCE* with competing methods in the following settings: response variable is perturbed, latent variables are unperturbed, and the latent variables are perturbed. The methods are Instrumental Variables IV (Angrist et al., 1996), Invariant Causal Predictions (Peters et al., 2016), Causal Dantzig (Rothenhäusler et al., 2019), backShift (Rothenhäusler et al., 2016) and our proposal *UT-LVCE*. Here, we denote an interventional equivalence class of DAGs by $\mathcal{I}$-MEC.

in the backShift procedure (Rothenhäusler et al., 2016) which still requires that the latent variables remain unperturbed for identifying the causal structure. Both Causal Dantzig and backShift yield a single causal structure, even if the underlying model is not fully identifiable. On the other hand, in addition to allowing for interventions on all the variables, *UT-LVCE* produces an equivalence class of DAGs. For a summary of the assumptions for *UT-LVCE* as compared to competing methods (including Instrumental Variable Regression (IV, Angrist et al. 1996), see Table 1. We will also provide more comparisons throughout the paper.

1.2 Notation

We denote the identity matrix by Id, with the size being clear from context. The collection of $d \times d$ symmetric matrices are denoted by $\mathbb{S}^d$ and positive-semidefinite matrices by $\mathbb{S}^d_+$ and the collection of strictly positive-definite matrices by $\mathbb{S}^d_{++}$. The collection of positive-definite diagonal matrices is denoted by $\mathbb{D}^d_{++}$. For two positive semidefinite matrices A and B, we write $B \succeq A$ if and only if $B - A$ is positive semidefinite. For a positive integer a, we denote the set $\{1, 2, \ldots, a\}$ by $[a]$. We denote the index set of the parents of a random variable X_p by $\mathrm{PA}(p)$. We denote $\mathrm{MEC}(\mathcal{D})$ to be the Markov equivalence class of $\mathcal{D}$, namely DAGs that have the same skeleton and v-structures as $\mathcal{D}$. Here a v-structure is a set of three nodes i, j, k such that there are directed edges from i to k and from j to k, and there is no edge between i and j. For a DAG $\mathcal{D}$ among p variables, and a matrix $B \in \mathbb{R}^{p \times p}$, we use the notation $B \sim \mathcal{D}$ to denote that $B_{ij} \neq 0$ implies there is a directed edge from node i to j in the DAG $\mathcal{D}$. For a set of diagonal and positive-definite matrices $\{\Omega_e\}_{e=1}^m \subseteq \mathbb{D}^p_{++}$ with positive integer $m \geq 2$, we let $\mathbb{I}(\{\Omega_e\}_{e=1}^m) := \{j : \exists\, e, f \text{ such that } [\Omega_e]_{j,j} \neq [\Omega_f]_{j,j}\}$. Finally, we say that a distribution is *faithful* with respect to a DAG $\mathcal{D}$ if the list of conditional independence relationships in the distribution are the same as those encoded in $\mathcal{D}$.

2. Modeling framework and maximum-likelihood estimator

In this section, we describe the data generation process associated with the intervention model sketched in Figure 1. Furthermore, we propose *UT-LVCE*, a regularized maximum-likelihood estimator. Given an input DAG, *UT-LVCE* identifies estimates of the unknown perturbation effects, the latent effects, and the causal relations among the observed variables. Finally, we describe a computationally efficient graphical procedure that uses the estimate obtained from *UT-LVCE* to find a set of equally scoring DAGs.

2.1 Modeling framework

We consider a directed acyclic graph whose $p + h$ nodes correspond to random variables $(X, H) \subseteq \mathbb{R}^p \times \mathbb{R}^h$, where X are observable and H are latent variables. We denote the induced subgraph DAG corresponding to the observed variables by $\mathcal{D}^\star$. We aim to learn $\mathcal{D}^\star$ or an equivalence class of DAGs when there are not enough interventions on the observed variables for full identifiability. Our methodology is also applicable in a setting where one is primarily interested in the causal effects on a particular response variable of interest. As such, we distinguish X_p as the target or response variable.

We assume that the observed and latent variables satisfy the following linear SCM:

$$X = B^\star X + \Gamma^\star H + \epsilon. \tag{1}$$

Here, the connectivity matrix $B^\star \in \mathbb{R}^{p\times p}$ contains zeros on the diagonal and is compatible with $\mathcal{D}^\star$: $B^\star_{ij} = 0$ if X_j is not a parent of X_i in $\mathcal{D}^\star$. Thus, the p-th row vector $B^\star_{p,:}$ encodes the (observable) causal parents of the response variable and the magnitude of their effects. The matrix $\Gamma^\star$ in (1) encodes the effects of the latent variables on the observed variables where $\Gamma^\star_{k,j} = 0$ if the latent variable H_j is not a parent of the node X_k. Further, ϵ is a random vector with independent components. We assume that the latent variables H are exogenous to X, so that ϵ is independent of H. Unless otherwise specified, no assumptions are imposed on the causal structure among the latent variables H.

The compact SCM (1) describes the generating process of X in the observational setting when there are no external perturbations on the system. We next describe how the data generation process alters due to some type of perturbations to the variables (X, H). We consider perturbations that directly shift the distributions of the random variables by some noise acting additively to the system. Specifically, the perturbations $\mathcal{E}$ generate the random pair (X^e, H^e) for each environment $e \in \mathcal{E}$ satisfying the following SCM:

$$X^e = B^\star X^e + \Gamma^\star H^e + \epsilon^e + \delta^e, \tag{2}$$

where for every $e \in \mathcal{E}$, $\epsilon^e \overset{\text{dist}}{=} \epsilon$, $(H^e, \delta^e, \epsilon^e)$ are jointly independent, and the collection $(X^e, H^e, \delta^e, \epsilon^e)$ is independent across e. Further, $\delta^e \in \mathbb{R}^p$ is a vector that represents the additive perturbations on the observed variables. Some of the entries of δ^e could be identically zero indicating that no interventions occurred; the remaining entries are generated from a random distribution. The intervention targets are denoted by $\mathcal{I}^\star := \{j \in [p] : \text{var}(\delta^e_j) \neq 0 \text{ for some } e \in \mathcal{E}\}$[1]. Importantly, the location of the nonzero components (i.e. variables that are intervened) is unknown. Finally, $H^e \in \mathbb{R}^h$ is a random vector that represents the intervened latent variables across the environments. That is, the interventions on the latent variables are absorbed into H^e. Without loss of generality, we assume that all variables are centered.

Given data of observed variables X^e across environments $e \in \mathcal{E}$, our objective is to develop a procedure to estimate the unknown intervention effects, the latent effects, and the causal

1. Here, we consider interventions that vary the variance of the noise terms; see Section 2 of Gamella et al. (2022) for why interventions on the means do not offer any identifiability in linear Gaussian SCMs.

relations among observed variables. To arrive at an estimator, we model the distribution of the random vectors H^e, ϵ^e and the nonzero components of δ^e to be Gaussian. Specifically, we model the random vectors H^e as well as the sum $\epsilon^e + \delta^e$ as follows:

$$H^e \sim \mathcal{N}(0, \Psi_e^\star),\ \Psi_e^\star \in \mathbb{S}_{++}^h,$$
$$\epsilon^e + \delta^e \sim \mathcal{N}(0, \Omega_e^\star),\ \Omega_e^\star \in \mathbb{D}_{++}^p,\ \mathbb{I}(\{\Omega_e^\star\}_{e=1}^m) = \mathcal{I}^\star.$$

The notations of $\mathbb{D}_{++}^{\cdot}$, $\mathbb{S}_{++}^{\cdot}$, $\mathbb{I}(\cdot)$ are defined in Section 1.2. We remark that non-Gaussian linear structural equation models are generally more identifiable than their Gaussian counterparts (Shimizu et al., 2006). While we develop our procedure based on a Gaussian model, we will see that the output of our approach is conservative in the sense that the true set of equivalent DAGs is contained in the estimated set in a non-Gaussian setting.

The compactified SCM (2) characterizes the distribution among all of the observed variables and encodes system-wide invariances. Specifically, (2) insists that for every $k = 1, 2, \ldots, p$, the regression coefficients when regressing X_k^e on the parent sets $\{X_j^e : X_j \text{ parent of } X_k\}$ and $\{H_l^e : H_l \text{ parent of } X_k\}$ remain invariant for all environments $e \in \mathcal{E}$. This is a point of departure from instrumental variable techniques (Angrist et al., 1996) or Invariant Causal Prediction (Peters et al., 2016) in two significant ways: 1) such methods do not allow for interventions on the latent variables or the response variable X_p (i.e. they assume $H^e \overset{\text{dist}}{=} H$ and $\delta_p^e \equiv 0$ for all $e \in \mathcal{E}$) and 2) they only consider "local" invariances arising from the distribution $X_p^e \mid \{(X_j^e, H_l^e) \text{ parents of } X_p\}$. The virtue of considering a joint model over all of the variables and exploiting system-wide invariances is that we can propose a maximum-likelihood estimator *UT-LVCE* which identifies the population DAG structure even under interventions on the response variable and the latent variables.

The SCM (2) is similar in spirit to previous modeling frameworks in the literature. The authors Hauser and Bühlmann (2015) consider jointly observational and interventional Gaussian data where the interventions are limited to do-interventions and there are no latent variables. In the context of (2), this means that $\delta^e \equiv 0$ and $\Gamma^\star \equiv 0$. As such, the framework considered in this paper is a substantial generalization of Hauser and Bühlmann (2015). Further, the backShift (Rothenhäusler et al., 2016) procedure considers the linear SCM (2) with some modifications: i) there are no interventions to the latent variables, i.e. $H^e \overset{\text{dist}}{=} H$ for all $e \in \mathcal{E}$, and ii) $\mathcal{D}^\star$ may be a cyclic directed graph. In addition, the backShift algorithm relies on exploiting invariances of differences of estimated covariance matrices across environments. Our *UT-LVCE* procedure is more in the "culture of likelihood modeling and inference" and has the advantage that it can cope well with having only a few observations per environment. This likelihood perspective also fits much more into the context of inference for mixed models as briefly discussed next.

The framework in (2) bears some similarities to standard random effects mixed models (McLean et al., 1991). In particular, random effects mixed models are widely employed to model grouped data, where some parameter components remain fixed and others are random. In the context of our problem, the fixed parameters are the matrices $B^\star, \Gamma^\star$ and the random parameters are the shift interventions δ^e. However, a difference between our model in (1) and standard mixed models is that the effects of the random parameter δ^e

propagate through the structural equations; and in practice, the order of propagation is usually unknown.

2.2 Scoring DAGs via *UT-LVCE*

In this section, we propose our method *UT-LVCE*, which scores a DAG $\mathcal{D}$ via regularized maximum likelihood estimation. As we will discuss, the scores of a candidate set of DAGs can then be obtained using this procedure to find the best scoring DAG(s). We suppose that there are m environments $|\mathcal{E}| = m$, and for every environment $e = 1, 2, \ldots, m$, we have samples of X^e: $\{X_i^e\}_{i=1}^{n_e}$ for some positive integer n_e which are independent and identically distributed (IID) for each e and independent across e. To obtain a score for a DAG $\mathcal{D}$, *UT-LVCE* identifies a causal model that best fits the data. This model is parameterized by $(B, \Gamma, \mathcal{I}, \Omega_e, \Psi_e)$ for all $e = 1, 2, \ldots, m$, where B is a connectivity matrix, Γ encodes the latent effects, $\mathcal{I}$ is a subset representing the intervention locations, Ω_e represents the noise variances of the observed variables, and Ψ_e encodes the interventions on the latent variables (see (2)). The quantities $(B, \Gamma, \mathcal{I}, \Omega_e, \Psi_e)$ are unknown and estimated by solving the following regularized maximum-likelihood estimator for the DAG structure $\mathcal{D}$ with $\bar{h}$ latent variables:

$$\begin{aligned} \underset{\substack{B\in\mathbb{R}^{p\times p},\Gamma\in\mathbb{R}^{p\times \bar{h}},\mathcal{I}\subseteq\{1,2,\ldots,p\} \\ \{\Omega_e,\Psi_e\}_{e=1}^m\subseteq\mathbb{D}_{++}^p\times\mathbb{S}_{++}^{\bar{h}}}}{\operatorname{argmin}} & \quad \sum_{e=1}^m \hat{\pi}_e \ell(B,\Gamma,\Omega_e,\Psi_e;\hat{\Sigma}_e) + \lambda \mathcal{R}_\gamma(\mathcal{D},\mathcal{I}). \\ \text{subject-to:} & \quad B \sim \mathcal{D} \;\; ; \;\; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I} \end{aligned} \tag{3}$$

Here, $\ell(\cdot)$ is the negative Gaussian log-likelihood

$$\ell(\cdot) := \log\det\left(\Omega_e + \Gamma\Psi_e\Gamma^T\right) + \operatorname{trace}\left(\left[\Omega_e + \Gamma\Psi_e\Gamma^T\right]^{-1}(\mathrm{Id} - B)\hat{\Sigma}_e(\mathrm{Id} - B)^T\right),$$

where the matrix $\hat{\Sigma}_e$ is the sample covariance of the data $\{X_i^e\}_{i=1}^{n_e}$. The quantity $\hat{\pi}_e = n_e / \sum_{e=1}^m n_e$ represents the estimated mixture components. The constraint $B \sim \mathcal{D}$ ensures that the connectivity matrix B satisfies the sparsity pattern of the graph $\mathcal{D}$. Further, the constraint $\mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}$ on the matrices $\{\Omega_e\}_{e=1}^m$ and set $\mathcal{I}$ ensures that the variances corresponding to unintervened coordinates (as specified by $\mathcal{I}$) are the same across all environments. The notations of $\mathbb{D}_{++}^{\cdot}$, $\mathbb{S}_{++}^{\cdot}$, $\cdot \sim \cdot$, $\mathbb{I}(\cdot)$ are defined in Section 1.2. Finally, $\lambda\mathcal{R}_\gamma(\cdot,\cdot)$ represents a regularization term with $\lambda, \gamma \geq 0$ and $\mathcal{R}_\gamma(\cdot,\cdot)$ given by:

$$\mathcal{R}_\gamma(\mathcal{D},\mathcal{I}) := (\|\mathcal{D}\|_{\ell_0} + p \text{ degree}[\text{moral}(\mathcal{D})]) + \gamma|\mathcal{I}|.$$

Here, $\|\mathcal{D}\|_{\ell_0}$ denotes the number of edges in $\mathcal{D}$. Further, moral$(\mathcal{D})$ denotes the moralization of $\mathcal{D}$ which forms an undirected graph of $\mathcal{D}$ by adding edges between nodes that have common children, and degree$[\cdot]$ computes the maximal degree of the undirected graph. The sum $\|\mathcal{D}\|_{\ell_0} + p$ degree$[\text{moral}(\mathcal{D})]$ regularizes DAG $\mathcal{D}$'s complexity[2]; although this term is a constant in the *UT-LVCE* estimator (3), it will play a crucial role for comparing different DAGs that are scored via the *UT-LVCE* estimator (3). The quantity $|\mathcal{I}|$ penalizes

2. One can also add an extra tuning parameter, e.g. $\|\mathcal{D}\|_{\ell_0} + \kappa$ degree$[\text{moral}(\mathcal{D})]$ for some $\kappa \geq 0$; for simplicity, we use a fixed value $\kappa = p$.

the number of interventions on the observed variables. Furthermore, the regularization parameter λ provides overall control of the trade-off between the fidelity of the model to the data and the complexity of the model. Additionally, the regularization parameter γ provides a trade-off between the complexity of the DAG and the number of intervention targets. Overall, $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})$ is akin to the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) score as it prevents overfitting by incorporating the denseness of the DAG $\mathcal{D}$ as well as the number of interventions in the likelihood score.

We note that regularization terms controlling for the complexity of estimated DAGs are commonly employed in causal structural learning (see Drton and Maathius 2017 and the references therein). Previous work on penalized likelihood scores only contain the regularization term $\|\mathcal{D}\|_{\ell_0}$. Thus, our regularization penalty $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})$ contains the novel terms degree[moral($\mathcal{D}$)] and $|\mathcal{I}|$. The quantity degree[moral($\mathcal{D}$)] is an important addition in our context to ensure identifiability of the underlying DAG among observed variables in the presence of latent variables as described in Section 3.2. The quantity $|\mathcal{I}|$ is motivated by the following observation: if the size of the intervention set is not penalized, for any finite sample size, (3) returns $\hat{\mathcal{I}} = [p]$. Intuitively, a model that contains interventions on all of the variables is in its own equivalence class (we formalize this in Section 3). Thus, without the penalty on the size of the intervention target, the optimum of (3) may be unique even if there are multiple DAGs in the equivalence class of the population model.

In summary, the estimator (3) takes as input the DAG $\mathcal{D}$, the tuning parameters $(\lambda, \gamma, \bar{h})$ and the observed empirical covariance matrices $\hat{\Sigma}_e$ to obtain a causal model with the following parameters:

$$\hat{\Theta}(\mathcal{D}, \bar{h}) := \text{ any minimizer } (\hat{B}, \hat{\Gamma}, \hat{\mathcal{I}}, \{\hat{\Omega}_e, \hat{\Psi}_e\}_{e=1}^m) \text{ of (3).}$$

Then the score for the DAG $\mathcal{D}$ given parameters $\hat{\Theta}(\mathcal{D}, \bar{h})$ is computed as:

$$\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h})) := \sum_{e=1}^{m} \hat{\pi}_e \ell(\hat{B}, \hat{\Gamma}, \hat{\Omega}_e, \hat{\Psi}_e; \hat{\Sigma}_e) + \lambda \mathcal{R}_\gamma(\mathcal{D}, \hat{\mathcal{I}}). \tag{4}$$

We select the parameters $(\lambda, \gamma, \bar{h})$ via cross-validation; see Section 4 for more discussion. Note that while the minimizer of (3) is not unique, the associated score (4) is the same for all minimizers of (3).

In comparison to the *UT-LVCE* procedure, backShift (Rothenhäusler et al., 2016) fits the SCM (2) (with some restrictions outlined in Section 2.1) by performing joint diagonalization to the difference of sample covariance matrices. *UT-LVCE* allows for much more modeling flexibility. First, in contrast to backShift where the latent effects are subtracted by computing the difference of covariances, *UT-LVCE* explicitly models these effects. This feature of *UT-LVCE* enables the possibility of interventions to the latent variables and a manner to control the number of estimated latent variables (as opposed to an arbitrary number of latent variables with backShift). We discuss in Section 3 that controlling the number of latent variables may lead to identifiability using *UT-LVCE* with two environments, whereas backShift is guaranteed to fail. Furthermore, *UT-LVCE* allows to pool information over different environments e for the parameter B of interest: this enables *UT-LVCE* to be used

with only a few sample points per environment. Finally, *UT-LVCE* explicitly models the intervention structure among the observed variables (via the set $\mathcal{I}$). We will see in the next section that encoding the intervention structure allows for outputting a set of equally scoring DAGs. The procedure backShift on the other hand returns a single DAG, even if the underlying model is not identifiable.

2.3 Score equivalent DAGs

The estimator (3) can be used in conjunction with (4) to score a collection of DAGs and find the ones that best fit the data. Such an approach raises the following question: are there multiple DAGs that fit the data equally well? In this section, we answer in the affirmative and provide a procedure to obtain a set of score equivalent DAGs. The set of equally scoring DAGs is closely related to an *interventional equivalence class*, which was first introduced for this model without latent variables in Gamella et al. (2022), and we present it below.

Definition 1 ($\mathcal{I}$-MEC, Gamella et al. (2022)) *Let $\mathcal{D}$ be a DAG and $\mathcal{I} \subseteq [p]$ denotes an intervention set. Furthermore, let* MEC($\mathcal{D}$) *be the standard observational Markov equivalence class of $\mathcal{D}$. Then, we define the following interventional equivalence class:*

$$\mathcal{I}\text{-MEC}(\mathcal{D}) := \left\{ \tilde{\mathcal{D}} \in \text{MEC}(\mathcal{D}) \mid \text{PA}_{\tilde{\mathcal{D}}}(i) = \text{PA}_{\mathcal{D}}(i) \text{ for all } i \in \mathcal{I} \right\}. \tag{5}$$

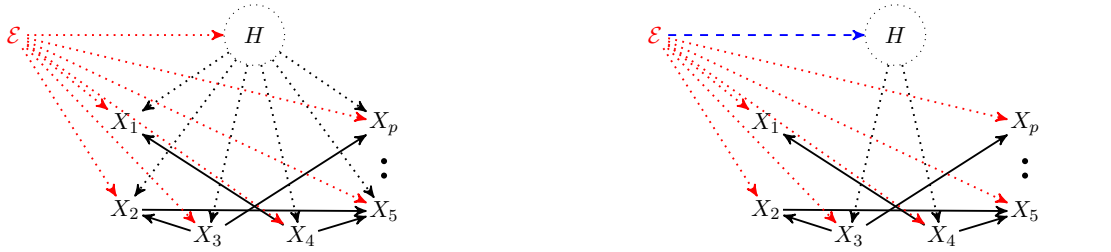
The interventional equivalence class $\mathcal{I}$-MEC($\mathcal{D}$) is closely related to the notion of *transition pair equivalence*, introduced by Tian and Pearl (2001) (see Gamella et al. 2022 for more discussion). This class is a subset of the standard observational Markov equivalence class and consists of DAGs that have the same parents as $\mathcal{D}$ for variables in the intervention target set $\mathcal{I}$. Thus, the intervention target set $\mathcal{I}$ controls the cardinality of $\mathcal{I}$-MEC($\mathcal{D}$), i.e. for any $\mathcal{I}_1 \subseteq \mathcal{I}_2$: $\{\mathcal{D}\} \subseteq \mathcal{I}_2\text{-MEC}(\mathcal{D}) \subseteq \mathcal{I}_1\text{-MEC}(\mathcal{D}) \subseteq \text{MEC}(\mathcal{D})$. As noted in Gamella et al. (2022), the set $\mathcal{I}$-MEC($\mathcal{D}$) can be computed efficiently given $(\mathcal{D}, \mathcal{I})$ using Meek's rules with background knowledge (Meek, 1995). The following theorem statement formally relates the interventional equivalence class $\mathcal{I}$-MEC($\mathcal{D}$) to the set of equally scoring DAGs.

Theorem 2 (score equivalent DAGs) *Consider an SCM* (2) *with structure given by a DAG $\mathcal{D}$ and parameter set $\Theta(\mathcal{D}, \bar{h}) := (B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$. Then, for any DAG $\tilde{D} \in \mathcal{I}$-MEC($\mathcal{D}$), there exists a parameter set $\tilde{\Theta}(\tilde{\mathcal{D}}, \bar{h}) := (\tilde{B}, \tilde{\Gamma}, \tilde{\mathcal{I}}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ with $\tilde{B} \sim \tilde{D}$ such that $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \Theta(\mathcal{D}, \bar{h})) = \texttt{score}_{\lambda,\gamma}(\tilde{\mathcal{D}}, \tilde{\Theta}(\tilde{\mathcal{D}}, \bar{h}))$ for all $\lambda, \gamma \geq 0$.*

The proof of Theorem 2 is shown in Appendix Section B and extends the analysis provided in Gamella et al. (2022) to the setting with latent variables. The results of Theorem 2 enable the characterization of the set of best scoring DAGs. Specifically, let $\hat{\mathcal{D}}_{\text{opt}}$ be an optimal scoring DAG with a corresponding intervention set $\hat{\mathcal{I}}_{\text{opt}}$, i.e. $(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}_{\text{opt}}) \in \text{argmin}_{\mathcal{D}, \bar{h}} \texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$ and $\hat{\mathcal{I}}_{\text{opt}}$ is an intervention set encoded in $\hat{\Theta}(\mathcal{D}_{\text{opt}}, \bar{h}_{\text{opt}})$. Then, all the DAGs inside $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$) are also optimal.

3. Identifiability guarantees with *UT-LVCE*

In this section, we analyze the identifiability guarantees of *UT-LVCE* in the infinite data limit. Specifically, we analyze the DAGs that minimize the score (4), i.e. are solutions to



(a) dense latent effects with $\dim(H) \ll p$ (b) approximately known latent interventions

Figure 3: Structural assumptions needed for equivalence class characterization: a) dense latent effects with a small number of latent variables as compared to the ambient dimension (Section 3.2) and b) known interventions on the latent variables (indicated by blue dashed line) and some confounding dependencies induced by the latent variables (Section 3.3).

$\operatorname{argmin}_{\mathcal{D},\bar{h}} \texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D},\bar{h}))$ when the sample size $n_e \to \infty$ for every $e \in [m]$ and $\lambda \to 0$ (with details on the specific rate described in Appendix C).

In Section 3.1, we show that without imposing any additional structure in the problem, an optimal scoring DAG may be very different than the population DAG $\mathcal{D}^\star$; we will describe how this result is related to the violation of the faithfulness assumption over the graph of observed and latent variables (see Section 1.2 for formal definition of faithfulness) that is typically made in the literature of latent variable causal discovery. Hence, since identifying optimal DAGs is meaningless without any conditions, in Sections 3.2 and 3.3, we impose structural assumptions on the latent effects and constrain the causal parameters appropriately; these conditions ensure that an estimated DAG is inside the interventional equivalence class of the population DAG in the infinite data limit, i.e. for any optimally scoring DAG $\hat{\mathcal{D}}_{\text{opt}}$, we have $\hat{\mathcal{D}}_{\text{opt}} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ with probability tending to one as the sample size in every environment tends to infinity. In Section 3.2, we assume that the number of latent variables is small and they affect many observed variables (visualized in Figure 3a). In Section 3.3, we assume that the latent interventions are approximately known and the latent variables induce some confounding dependencies (visualized in Figure 3b). Throughout, we describe quantitative measures that our software outputs to indicate when deviations from assumptions may have occurred.

We assume that the data is generated according to the intervention model in (2) with population parameters $B^\star, \Gamma^\star, \mathcal{I}^\star, \{(\Omega_e^\star, \Psi_e^\star)\}_{e=1}^m$ (see Section 2.1). We let

$$\hat{\mathcal{D}}_{\text{all.opt}}, \hat{B}_{\text{all.opt}}, \hat{\mathcal{I}}_{\text{all.opt}} \tag{6}$$

be the optimally scoring DAGs and associated connectivity matrice(s) and set(s) of intervention targets, which are all solutions to:

$$\begin{aligned} \operatorname*{argmin}_{\mathcal{D},\bar{h}} \quad \operatorname*{argmin}_{\substack{B,\Gamma,\mathcal{I} \\ \{\Omega_e,\Psi_e\}_{e=1}^m}} \quad & \sum_{e=1}^m \hat{\pi}_e \ell(B,\Gamma,\Omega_e,\Psi_e;\hat{\Sigma}_e) + \lambda \mathcal{R}_\gamma(\mathcal{D},\mathcal{I}) \\ \text{subject-to:} \quad & B \sim \mathcal{D} \;\; ; \;\; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}. \end{aligned} \tag{7}$$

Here, the dimension of the matrices Γ and Ψ_e are $\mathbb{R}^{p\times\bar{h}}$ and $\mathbb{S}^{\bar{h}}$, respectively. Compared to the estimator (3), the estimator (7) searches for the optimal DAG and number of latent

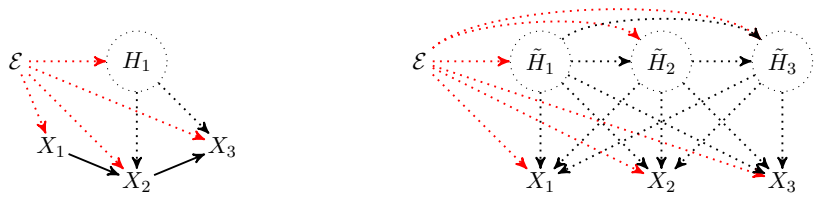


Figure 4: Two equivalent models with respect to the distribution among the observed variables; see text.

variables; thus, $\hat{\mathcal{D}}_{\text{all.opt}} = \text{argmin}_{\mathcal{D},\bar{h}} \texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$ and $\hat{B}_{\text{all.opt}}, \hat{\mathcal{I}}_{\text{all.opt}}$ are the associated model parameters. Throughout, we choose $\lambda \to 0$, with details on the specific rate described in Appendix C. We assume that $\lim_{\text{all } n_e \to \infty} \frac{n^e}{\sum_{e=1}^m n^e} > 0$ for all $e \in [m]$. Finally, for technical reasons, to prove consistency, the parameter space $(B, \Gamma, \{(\Omega_e, \Psi_e)\}_{e=1}^m)$ in (7) is assumed to be compact. Assuming such a compactness constraint enables uniform convergence of M-estimators (van de Geer, 2000); see again Appendix C for more details.

3.1 Impossibility results without imposing structural assumptions

The problem of identifying the underlying DAG is ill-posed if no assumption is placed on the latent effects. Specifically, the distribution among the observed variables in every environment can be expressed as one generated according to an SCM (1) where the graph among the observed variables is arbitrary. We formalize this next.

Proposition 3 (Equivalent SCMs) *Regardless of the intervention set $\mathcal{I}^\star$ and the intervention magnitudes encoded in $\Omega_e^\star$ on the observed variables, for any DAG $\mathcal{D}$ and any connectivity matrix $B \sim \mathcal{D}$, there exists parameters $(\Gamma, \{(\Omega_e, \Psi_e)\}_{e=1}^m)$ such that the associated SCM specified by these parameters is compatible with the data distribution, and consists of $h = p$ latent variables.*

The proof of Proposition 3 is presented in Appendix D.1. Figure 4 provides an illustration of this result among three observed variables. Here, Figure 4(a) and Figure 4(b) represent equivalent models, although the subgraphs among the observed variables are different; see Appendix D.2 for formal description of the parameters of these models. Note that a typical assumption in causal structure learning with latent variables is that the joint distribution of the observed and latent variables is faithful to the graph among these variables. Indeed, many algorithms such as FCI (Spirtes et al., 2000), RFCI (Colombo et al., 2012), and JCI (Mooij et al., 2020) rely on this condition for characterizing and learning an equivalence class of models. The result of Proposition 3 provides a construction of non-faithful models; see Appendix D.2.

Corollary 4 (Optimally scoring DAGs without structural assumptions) *Regardless of the intervention set $\mathcal{I}^\star$ and the intervention magnitudes encoded in $\Omega_e^\star$ on the observed variables, the set $\hat{\mathcal{D}}_{all.opt}$ is a singleton containing the empty graph.*

The proof of Corollary 4 is presented in Appendix D.3. The results of Proposition 4 and Corollary 4 state that searching for the best scoring DAG is meaningless if no structure is imposed on the problem.

3.2 Equivalence class characterization under a small number of latent variables with dense effects

We now analyze the estimates (6) under structural assumptions on the denseness of the latent effects and sparsity of the underlying DAG where the interventions on the latent variables may be arbitrary. We substantially relax this assumption in Section 3.3 at the expense of approximate knowledge of the interventions on the latent variables.

Before proceeding, we present some real-world applications where the assumed structure may be reasonable. For example, Chandrasekaran et al. (2012) showed that a large fraction of the conditional dependencies among stick returns can be explained by a few latent variables. In a similar spirit, Taeb et al. (2017) demonstrated that the California reservoir network is influenced by a few external latent factors (correlated with environmental variables), and these have a system-wide effect; we explore this application further in our experiments. Finally, in an analysis of gene expression data, Zhao et al. (2016) find that a small number of dense latent factors explain much more variability than sparse latent factors, and these dense factors correlated well with some known biological and technical covariates (e.g. batch effects).

How are the aforementioned structural assumptions useful for identifiability? To motivate the utility of these structural assumptions, we first note that the structural equation model (2) yields the covariance model of observed variables $\Sigma_e^\star = (\mathrm{Id} - B^\star)^{-1}(\Omega_e^\star + \Gamma^\star \Psi_e^\star \Gamma^{\star T})(\mathrm{Id} - B^\star)^{-T}$ for every $e \in [m]$. By the Woodbury Inversion lemma, we obtain the following decomposition $\Sigma_e^{\star -1} = S_e^\star - L_e^\star$ of the precision matrix $\Sigma_e^{\star -1}$ for every $e \in [m]$. Here, the matrix $S_e^\star = (\mathrm{Id} - B^\star)^T \Omega_e^{\star -1}(\mathrm{Id} - B^\star)$ is the inverse of the conditional covariance of the observed variables conditioned on the latent variables. The matrix $L_e^\star$ is the rank-h matrix $(\mathrm{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star (\Psi_e^{\star -1} + \Gamma^{\star T} \Omega_e^{\star -1} \Gamma^\star)^{-1} \Gamma^{\star T} \Omega_e^{\star -1}(\mathrm{Id} - B^\star)$ that summarizes the effect of marginalization over latent variables.

Without assuming any additional structure on the population model, the matrices $S_e^\star$ and $L_e^\star$ are not identifiable from $\Sigma_e^{\star -1}$. This lack of identifiability implies $\Sigma_e^{\star -1}$ can be modeled by a different DAG $\mathcal{D} \in \hat{\mathcal{D}}_{\mathrm{all.opt}}$ that may be arbitrarily different from $\mathcal{D}^\star$. Appealing to the previous literature on sparse-plus-low rank decompositions, the matrices $S_e^\star$ and $L_e^\star$ are identifiable from their sum if the matrix $S_e^\star$ is sparse and the matrix $L_e^\star$ is low-rank with its energy spread across the coordinates (Recht et al., 2010; Chandrasekaran et al., 2011; Candès et al., 2011). It is straightforward to check that the entry $[S_e^\star]_{i,j}$ is nonzero if the variables X_i and X_j are connected in the moral graph of $\mathcal{D}^\star$. Thus, a sparse moral graph of $\mathcal{D}^\star$ implies that the matrix $S_e^\star$ is sparse. The assumption on $L_e^\star$ can be interpreted as the number of latent variables being small (as compared to the ambient dimension p) with their effects spread across all the observed variables. We measure the sparsity of the moral graph of a DAG $\mathcal{D}$ by the maximal degree of the moral graph, denoted by degree[moral($\mathcal{D}$)]. Thus, we require degree[moral($\mathcal{D}^\star$)] to be small so that no observed variable is directly connected to "many" other observed variables in the moral graph of $\mathcal{D}^\star$. To measure the "diffuseness"

of the latent effects, we consider the following quantity for any linear subspace $T \subseteq \mathbb{R}^p$ (Candès et al., 2011; Candès and Recht, 2009; Chandrasekaran et al., 2011, 2012):

$$\text{inc}[T] := \max_i \|\mathcal{P}_T(\mathbf{e}_i)\|_2,$$

where $\mathcal{P}_T$ is the projection onto the subspace T and $\mathbf{e}_i$ is a standard coordinate basis. The quantity inc$[T]$ is also known as the "incoherence parameter" (Candès and Recht, 2009; Chandrasekaran et al., 2011). It measures how aligned the subspace T is with respect to standard basis elements and is lower-bounded by $\sqrt{\frac{\dim(T)}{p}}$ and upper-bounded by one. In our setting, the relevant subspace is col-space$((\text{Id} - B^\star)^T\Omega_e^{\star -1}\Gamma^\star))$ which is the column-space of $L_e^\star$. A small value of inc[col-space$((\text{Id} - B^\star)^T\Omega_e^{\star -1}\Gamma^\star))$] ensures the matrix $L_e^\star$ has small rank and cannot have its support concentrated in a few locations.

In summary, to enable identifiability, the population quantities degree[moral$(\mathcal{D}^\star)$] and inc[col-space$((\text{Id} - B^\star)^T\Omega_e^{\star -1}\Gamma^\star)$] are assumed to be sufficiently small. We will first analyze the estimates (6) under these assumptions as well as faithfulness and access to an observational environment. For notational simplicity, we define $d^\star :=$ degree[moral$(\mathcal{D}^\star)$] and $\text{inc}_e^\star :=$ inc[col-space$((\text{Id} - B^\star)^T\Omega_e^{\star -1}\Gamma^\star)$]. Formally, we assume:

Assumption 1 *sparse DAG and incoherent (dense) latent effects across all environments:* $32d^\star \text{inc}_e^{\star 2} < 1$ *for all* $e \in [m]$.

Assumption 2 *the distribution* $X^e|H^e$ *is faithful with respect to* $\mathcal{D}^\star$ *for all* $e \in [m]$.

Assumption 3 *observational environment* $e = 1$ *with no interventions on the observed variables:* $\Omega_e^\star \succeq \Omega_1^\star$ *for all* $e \in [m]$.

Assumption 1 ensures that the population model consists of a sufficiently sparse moral graph and dense latent effects with a small number of latent variables as compared to a relatively large ambient dimension p. This assumption bears resemblance to conditions for identifiability in sparse-plus-low rank decompositions (Chandrasekaran et al., 2011, 2012; Frot et al., 2019), although we demonstrate in Appendix E that our condition is weaker (in terms of high-dimensional scaling) than those imposed in these previous works. We also provide in Appendix E examples of SCMs (2) that satisfy Assumption 1. Further, Assumptions 2-3 are standard conditions for identifiability of an equivalence class of DAGs both in observational and interventional settings (Chickering, 2002; Hauser and Bühlmann, 2015; Wang et al., 2017). Note that regarding Assumption 3, one does not need to know a prioi which environment is observational. We will see later that Assumption 3 can be replaced by conditions on the informativeness of the interventions.

Recall that Proposition 4 and Corollary 4 tells us that solving (7) is meaningless unless the parameters of the estimated causal models are also appropriately constrained. Thus, under Assumptions 1-3, we consider the theoretical properties of (6) when the incoherence of the estimated latent effects is controlled.

Proposition 5 (Equivalence class characterization under incoherent latent effects) *Consider the estimator* (7) *with the additional constraint that* inc[col-space$((\text{Id} - B)^T\Omega_e^{-1}\Gamma)] \leq 2\text{inc}_e^\star$ *for all* $e \in [m]$. *Then, under Assumptions 1-3 and* γ *selected so that*

$d^\star \geq \gamma > 0$, we have that the estimates (6) *satisfy:* $\{\mathcal{D}^\star\} \subseteq \hat{\mathcal{D}}_{\text{all.opt}} \subseteq \text{MEC}(\mathcal{D}^\star)$ *with probability tending to one as the sample size in every environment tends to infinity.*

We present the proof of Proposition 5 in Appendix F. This result states that if the latent effects of the estimated causal models are constrained to be low-dimensional and dense, in infinite data limit and with probability tending to one, the set of equally scoring latent variable causal models $\hat{\mathcal{D}}_{\text{all.opt}}$ are a subset of the Markov equivalence class $\text{MEC}(\mathcal{D}^\star)$ and contain the population DAG $\mathcal{D}^\star$. In Proposition 5 the interventions on the variables indexed by $\mathcal{I}^\star$ do not appear to directly constrain the optimal set of solutions. Indeed, we show in Appendix G.1 that under certain "worst-case configurations" of intervention strengths, the set $\hat{\mathcal{D}}_{\text{all.opt}}$ is precisely equal to the Markov equivalence class $\text{MEC}(\mathcal{D}^\star)$. Thus, to improve identifiability, additional assumptions on the informativeness of the interventions are needed, which we state below:

Assumption 4 *interventions on the observed variables are heterogeneous: for every $i \in \mathcal{I}^\star$, there exists e such that $[\Omega_e^\star]_{i,i} > [\Omega_1^\star]_{i,i}$ and $[\Omega_e^\star \Omega_1^{\star -1}]_{i,i} \neq [\Omega_e^\star \Omega_1^{\star -1}]_{j,j}$ for all $j \neq i$.*

Assumption 5 *interventions on the observed variables are "truthful": for every Markov equivalent connectivity* [3] *B to $B^\star$ w.r.t $X^1|H^1$, and (i,j) where $i \in \mathcal{I}^\star, i \rightsquigarrow j$ in $\mathcal{D}^\star$, $[\text{Id} - B]_{j,:} \not\propto [\text{Id} - B^\star]_{i,:}$.*

Here, the notation $v_1 \not\propto v_2$ for vectors v_1 and v_2 means that the vectors v_1 and v_2 are not proportional. Further, the notation $i \rightsquigarrow j$ in $\mathcal{D}^\star$ means that the variable X_i is an ancestor of the variable X_j in the DAG $\mathcal{D}^\star$. While Assumption 4 ensures that the interventions on the observed variables are sufficiently diverse [4], Assumption 5 excludes a pathological configuration of the intervention strengths and the connectivity matrix $B^\star$ and is similar in spirit to "interventional faithfulness" assumptions imposed in previous work (Gamella and Heinze-Deml, 2020; Squires et al., 2020; Gamella et al., 2022). In summary, Assumptions 4-5 are both rather weak and ensure that interventions on the observed variables improve identifiability:

Theorem 6 (Equivalence class characterization under incoherent latent effects, and truthful and heterogeneous interventions) *Consider the estimator* (7) *with the additional constraint that* $\text{inc}[\text{col-space}((\text{Id} - B)^T \Omega_e^{-1} \Gamma)] \leq 2\text{inc}_e^\star$ *for all $e \in [m]$. Then, under Assumptions 1-2 and 4-5, and if the parameter $d^\star \geq \gamma > 0$: $\{\mathcal{D}^\star\} \subseteq \hat{\mathcal{D}}_{\text{all.opt}} = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$, $B^\star \in \hat{B}_{\text{all.opt}}$ and $\hat{\mathcal{I}}_{\text{all.opt}} = \{\mathcal{I}^\star\}$, all with probability tending to one as the sample size in every environment tends to infinity.*

We present the proof of Theorem 6 in Appendix G.2. Notice that Assumptions 4-5 replace the need for access to an observational environment in Assumption 3, as $\mathcal{I}\text{-MEC}(\mathcal{D}^\star) \subseteq \text{MEC}(\mathcal{D}^\star)$. Theorem 6 states that $\mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ contains the set of optimally scoring DAGs when the latent effects of the estimated causal models are constrained to be low-dimensional and dense and the interventions are informative. Importantly, our proposed procedure *UT-LVCE* cannot directly control the incoherence of the latent effects. Instead, it can only

3. A Markov equivalent connectivity matrix B w.r.t $X^e|H^e$ in Assumption 4 satisfies: compatibility with a DAG $\mathcal{D} \in \text{MEC}(\mathcal{D}^\star)$ and the relation $\Sigma^\star_{X^e|H^e} = (\text{Id} - B)^{-1}\Omega(\text{Id} - B)^{-T}$ for some $\Omega \in \mathbb{D}^p_{++}$.
4. Assumption 4 can be satisfied even if the interventions occur in different environments.

constrain the number of latent variables. In the following corollary, we provide sufficient conditions for when the estimator (7) can obtain a DAG from the set $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$).

Corollary 7 *Suppose that Assumptions 1-2 and 4-5 are satisfied. Let $d^\star \geq \gamma > 0$. Consider the estimator (7) with the constraint that $\bar{h} \leq h_{max}$ for some non-negative integer $h_{max} \geq dim(H)$. Let $\nu^\star$ be a positive integer with $\nu^\star \geq d^\star$. Suppose there exists an estimate $(\hat{\mathcal{D}}, \hat{B}, \hat{\Gamma}, \hat{\mathcal{I}}, \{\Omega_e, \hat{\Psi}_e\}_{e=1}^m)$ that satisfies $48\nu^\star \text{inc}[\text{col-space}((\text{Id}-\hat{B})^T\hat{\Omega}_e^{-1}\hat{\Gamma})] < 1$ for all $e \in [m]$. Then, $\hat{\mathcal{I}} = \mathcal{I}^\star$ and $\hat{\mathcal{I}}$-MEC($\hat{\mathcal{D}}$) = $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) with probability tending to one as the sample size in every environment tends to infinity.*

The proof of Corollary 7 is presented in Appendix G.3. The result in Corollary 7 suggests the following procedure when Assumptions 1-5 are believed to be satisfied and the user has access to $\nu^\star$ that serves as an upper-bound for the maximal degree of the moral graph of the underlying DAG: obtain the best latent variable causal model(s) based on the likelihood score on test data when the regularization parameters $\lambda, \bar{h}$ are varied with $\bar{h}$ smaller than a pre-specified value $h_{\max}$. Then, compute the incoherence of the latent effects of these best scoring models. If for an optimal DAG $\hat{\mathcal{D}}$, the incoherence parameter multiplied by $\nu^\star$ is sufficiently small for all environments, in large data settings and with probability tending to one, $\mathcal{D}^\star$ lies inside $\hat{\mathcal{I}}$-MEC($\hat{\mathcal{D}}$). To highlight when such an assumption is far from being satisfied, our software outputs the following quantitative indicator: $\max_e \text{degree}[\text{moral}(\hat{\mathcal{D}})]\text{inc}[\text{col-space}((\text{Id} - \hat{B})^T\hat{\Omega}_e^{-1}\hat{\Gamma})]$. Here, large values (e.g. far above 1) indicate strong deviations from our assumptions.

3.3 Equivalence class characterization under approximately known latent interventions

In the previous discussion, a central assumption was that the number of latent variables is small and their effects are dense. We next consider a setting where the interventions on the latent variables are approximately known. These assumptions enable an equivalence class characterization of DAGs without needing Assumption 1, i.e. without imposing conditions on the number of latent variables and the denseness (incoherence) of their effects. For technical simplicity, we analyze the setting where the latent variables are independent and identically distributed, i.e. $\Psi_e^\star = \psi_e^\star \text{Id}$ with $\psi_e^\star \in \mathbb{R}_+$.

For illustrative purposes, we first start with an extreme setting where the interventions are exactly known (although the number of latent variables remains unknown) and demonstrate that identifiability is possible with relatively mild assumptions. We then deviate from this extreme setting by assuming that the latent interventions are approximately known and show once again that identifiability is possible under some conditions whose severity depends on the level of the approximation.

Illustrative setting: known latent interventions Without loss of generality, we can take $\psi_1^\star = 1$ and $\psi_e^\star$ to be known positive values that may be different than $\psi_1^\star$. Our theoretical guarantees require Assumptions 2 and 5 as well as modifications to Assumptions 3 and 4 (dubbed 3' - 4'). In particular, we assume that there are two observational environments ($e = 1$ and $e = 2$ without loss of generality) with no interventions on the observed

variables and interventional environments (so that $m \geq 3$) with sufficiently heterogeneous interventions on the observed variables:

Assumption 3' *environments* $e = 1, 2$ *with no interventions on the observed variables:* $\Omega_1^\star = \Omega_2^\star$ *and* $\Omega_e^\star \succeq \Omega_1^\star$ *for all* $e = 3, \ldots, m$.

Assumption 4' *heterogeneous interventions on observed and latent variables: for every* $i \in \mathcal{I}^\star, j \in \mathcal{I}^\star$ *with* $i \neq j$, *there exists* e *such that the collection* $(\psi_1^\star, \psi_2^\star, \psi_e^\star)$ *are distinct* , $[\Omega_e^\star]_{i,i} > [\Omega_1^\star]_{i,i}$ *and* $[(\Omega_e^\star - \psi_e^\star \Omega_1^\star)\, \Omega_1^{\star -1}]_{i,i} \neq [(\Omega_e^\star - \psi_e^\star \Omega_1^\star)\, \Omega_1^{\star -1}]_{j,j}$.

Assumption 3' (analogous to Assumption 3) ensures that there are environments where no interventions act on the observed variables. Assumption 4' (analogous to Assumption 4) ensures that the interventions on the latent variables and observed variables are informative for additional identifiability. One can show that if the parameters $\Omega_e^\star, \Omega_1^\star$ and $\psi_1^\star, \psi_2^\star, \psi_e^\star$ are drawn from continuous distributions, Assumption 4' is satisfied almost surely.

Theorem 8 (Equivalence class characterization under known interventions on the latent variables) *Consider the estimator* (7) *with the additional constraint* $\psi_e \mathrm{Id} = \psi_e^\star \mathrm{Id}$ *for all* $e \in [m]$. *Suppose Assumptions 2, 5 and Assumptions 3'-4' are satisfied. Letting* $\frac{1}{|\mathcal{I}^\star|} > \gamma > 0$, *then* $\{\mathcal{D}^\star\} \subseteq \hat{\mathcal{D}}_{\text{all.opt}} = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ *with probability tending to one as the sample size in every environment tends to infinity.*

The proof of Theorem 8 is presented in Appendix H. This result highlights that at the expense of knowing the latent interventions, no assumptions on the incoherence (denseness) of the latent variables or their number are required for characterizing the equivalence class of optimally scoring DAGs. We note that when the number of latent variables is unconstrained, three environments are necessary for improved identifiability. Indeed, in Appendix I, we show that two environments (regardless of the number of interventions and their strengths) only offer identifiability up to the Markov equivalence class of $\mathcal{D}^\star$.

Approximately known latent interventions Knowing the interventions on the latent variables can be a stringent condition in practice. One can relax this to approximately knowing the interventions at a pre-specified level C_ψ, e.g. $|\tilde{\psi}_e - \psi_e^\star| \leq C_\psi$ where $\tilde{\psi}_e$ is the (approximate) known intervention on the latent variables. A natural choice for the approximate interventions $\tilde{\psi}_e$ would be $\tilde{\psi}_e = 1$ for all e, encoding no interventions on the latent variables: the level C_ψ then describes the deviation from nointerventions on the latent variables. This and versions thereof will be discussed in the remarks below.

*R*emark 2: To account for the latent intervention approximation, the following two assumptions ensure equivalence class characterization. The first assumption is that the latent variables induce some confounding dependencies among the observed variables; this condition becomes more stringent with larger C_ψ (e.g. weaker knowledge of the latent interventions) although we demonstrate in Appendix J that it is generally *far weaker* than the incoherence condition in Assumption 1. The second assumption is that the observed variables in the set $\mathcal{I}^\star$ receive strong enough interventions, Under these two conditions (as well as assumptions 2,3',5 and an assumption similar in spirit to 4'), the estimator (7) obtains (in the infinite data limit) $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) as the set of optimally scoring DAGs. For a formal description of the assumptions and the result, we refer the reader to Appendix J. Finally, as a quantitative

indicator of deviations from assumptions, our software displays the strength of interventions on each variable, i.e. $\xi_j := \frac{1}{m}\sum_{e=1}^{m}([\hat{\Omega}_e]_{j,j} - \frac{1}{m}\sum_{e=1}^{m}[\hat{\Omega}_e]_{j,j})^2$ for each $j \in \hat{\mathcal{I}}_{\text{opt}}$. Here, ξ_j being small for any $j \in \hat{\mathcal{I}}_{\text{opt}}$ indicates that the perturbations on the corresponding variable are weak.

*R*emark 3: Assuming that the latent variables remain unperturbed across all environments is a special case of knowing the latent interventions. In such settings, the equivalence class – when $\mathcal{I}^\star \subset [p]$ – can, in general, be very different than $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) (see Appendix K for a simple illustration), highlighting that interventions on the latent variables may be beneficial for improved identifiability. Nevertheless, when $\mathcal{I}^\star = [p]$, similar to the backShift procedure, *UT-LVCE* attains full identifiability of the population DAG since $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) = $\{\mathcal{D}^\star\}$.

4. Practical use cases of *UT-LVCE*

We next describe how *UT-LVCE* can be used in practice to account for latent effects and obtain a set of DAGs that fit the data well. In Section 4.1, we propose an alternating minimization strategy to solve (3) with the DAG, hence also the support of B, being pre-specified. Building on this, in Section 4.2, we consider the setting where a candidate set of DAGs are available (for example as for the protein expressions dataset in Section 5) and describe how *UT-LVCE* can be used to obtain an optimally scoring equivalence class of DAGs. Finally, in Section 4.3, we extend our algorithmic framework to the setting when a set of DAGs represent starting points, and we deploy *UT-LVCE* to improve on these DAGs by removing any spurious dependencies. A python package containing the implementation of all components of *UT-LVCE* is available at `https://github.com/juangamella/ut-lvce`.

We remark here that searching for optimally scoring DAGs (according to the score (4)) is a very difficult computational task. In particular, an immediate approach that comes to mind is to develop a greedy DAG search over the space of equivalent models akin to GES Chickering (2002). Indeed, Gamella et al. (2022) develops a greedy algorithm to move in the space of interventionally equivalent DAGs for the model (2) *without* latent variables. By employing the *UT-LVCE* score function (4), one may adapt the method of Gamella et al. (2022) to incorporate latent effects. However, a significant conceptual challenge is that the likelihood score (4) is not decomposable according to the DAG structure due to latent confounding; the lack of score decomposability renders greedy-based techniques computationally expensive. Thus, our focus in this paper is to demonstrate the utility of *UT-LVCE* on the use cases described in the previous paragraph.

4.1 Alternating minimization strategy to compute the *UT-LVCE* estimator

We first describe an optimization approach for solving *UT-LVCE* given an input DAG $\mathcal{D}$ and an intervention target set $\mathcal{I}$. While the optimization problem (3) searchers over intervention target set $\mathcal{I}$ when scoring a DAG, in the description in this section, we fix $\mathcal{I}$. We present an optimization strategy for selecting $\mathcal{I}$ in Section 4.2.

Our optimization algorithm is based on the following alternating minimization strategy: starting with an initialization of all of the model parameters, we fix B and perform gradi-

ent updates to find updated estimates for the parameters $(\Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m$ where the noise variances $\{\Omega_e\}_{e=1}^m$ are compatible with the input intervention set $\mathcal{I}$, and then update B by solving a convex program to optimality with the remaining parameters fixed. We find that the alternating method described above is relatively robust to the initialization scheme, but we nonetheless propose the following concrete strategy:

$$\begin{aligned}
&1)\ \text{fit } B^{(0)} \text{ via linear regression with pooled data,}\\
&2)\ \Gamma^{(0)} = UD^{1/2} \text{ where } UDU^T \text{ is SVD of } (\mathrm{Id} - B^{(0)})\hat{\Sigma}_{\text{pooled}}(\mathrm{Id} - B^{(0)})^T,\\
&3)\ \Omega_e^{(0)} = \mathtt{diag}\left\{(\mathrm{Id} - B^{(0)})\hat{\Sigma}_{\text{pooled}}(\mathrm{Id} - B^{(0)})^T\right\} \text{ for every } e \in [m],\\
&4)\ \Psi_e^{(0)} = \mathrm{Id} \text{ for every } e \in [m],
\end{aligned} \tag{8}$$

where $\hat{\Sigma}_{\text{pooled}}$ is the covariance matrix of the pooled data. The first step follows since the DAG structure is known. The entire procedure, involving the initialization step and the parameter updates, is presented in Algorithm 0.

Algorithm 0 Solving *UT-LVCE* to score a given DAG $\mathcal{D}$ for a fixed $(\lambda, \gamma, \bar{h})$ and targets $\mathcal{I}$

1: **Input**: DAG $\mathcal{D}$; intervention targets $\mathcal{I}$; data $\hat{\Sigma}_e$, $\hat{\pi}_e$ for $e \in [m]$; regularization params. $\lambda, \gamma \geq 0$; # of latent vars. $\bar{h}$

2: **Initialize parameters**: via relation (8)

3: **Alternating minimization**: solve for causal parameters

   (a) fixing $(\Gamma^{(t)}, \{(\Omega_e^{(t)}, \Psi_e^{(t)})\}_{e=1}^m)$, update $B^{(t+1)}$ by solving the convex optimization program (3) where $B^{(t+1)} \sim \mathcal{D}$. Fixing $B^{(t+1)}$, perform gradient updates until convergence to find $(\Gamma^{(t+1)}, \{(\Omega_e^{(t+1)}, \Psi_e^{(t+1)})\}_{e=1}^m)$ where $\mathbb{I}(\{\Omega_e^{(t+1)}\}_{e=1}^m) \subseteq \mathcal{I}$

   (b) apply alternating iterates for positive integers t until convergence at iteration T

   (c) obtain estimates $\hat{\Theta}(\mathcal{D}, \bar{h}) := (\hat{B}^{(T)}, \hat{\Gamma}^{(T)}, \mathcal{I}, \{(\hat{\Omega}_e^{(T)}, \hat{\Psi}_e^{(T)})\}_{e=1}^m)$

4: **Output:** causal parameters $\hat{\Theta}(\mathcal{D}, \bar{h})$ and regularized likelihood $\mathtt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$

Step 3(b) involves two convergence criteria: the convergence of the gradient steps for the parameters $(\Gamma^{(t)}, \{(\Omega_e^{(t)}, \Psi_e^{(t)})\}_{e=1}^m)$ as well as the convergence of the alternating procedure. For the first criterion, we terminate the gradient descent when the relative change in the likelihood score is below ϵ_1. For the second criterion, we terminate the alternating minimization at step T when $\|B^{(T)} - B^{(T-1)}\|_\infty \leq \epsilon_2$, where $\|\cdot\|_\infty$ computes the maximum entry in magnitude of an input matrix. In our experiments, we set $\epsilon_1 = 10^{-6}$ and $\epsilon_2 = 10^{-2}$.

4.2 Using *UT-LVCE* to identify the best scoring DAGs from a candidate set

Let $\mathcal{D}_{\text{cand}}$ be a candidate set of DAGs (potentially a singleton). Building on Algorithm 0, we present an algorithm to identify an optimally scoring DAG as well as DAGs in its equivalence class. First, using Algorithm 0, we score each DAG in the candidate set with $\mathcal{I} = [p]$, and obtain an optimally scoring DAG $\hat{\mathcal{D}}_{\text{opt}}$ with noise variances $\{\hat{\Omega}_e\}_{e=1}^m$. To estimate the intervention targets $\hat{\mathcal{I}}_{\text{opt}}$, we measure the variation in each coordinate of $\hat{\Omega}_e$ across the environments as large variations indicate that the corresponding variable has

received an intervention. To quantify the degree of variation, we compute a "variance like" metric $\xi_j := \frac{1}{m}\sum_{e=1}^{m}([\hat{\Omega}_e]_{j,j} - \frac{1}{m}\sum_{e=1}^{m}[\hat{\Omega}_e]_{j,j})^2$ for each $j \in [p]$ (also defined in Section 3.3), where large values of ξ_j provide stronger evidence for the presence of an intervention on variable X_j. We propose a systematic approach to estimate an intervention set $\hat{\mathcal{I}}_{\text{opt}}$ using the values ξ_j: we greedily remove the variable with the smallest variation ξ_j and compute the regularized likelihood with the resulting intervention set. We repeat this process until the likelihood score can no longer be improved. Thus, as output, we return the equivalence class of optimally scoring DAGs $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$). A detailed summary of our procedure is presented in Algorithm 1.

Algorithm 1 Equivalence class of best scoring DAGs from a candidate set via *UT-LVCE*

1: **Input**: candidate DAG(s) $\mathcal{D}_{\text{cand}}$; data $\hat{\Sigma}_e, \hat{\pi}_e$ for $e \in [m]$; regularization params. $\lambda, \gamma \geq 0$; # of latent vars. $\bar{h}$

2: **Obtain likelihood score for each DAG**: for each $\mathcal{D} \in \mathcal{D}_{\text{cand}}$, supply $(\mathcal{D}, \mathcal{I} = [p])$ to Algorithm 0 to obtain the causal parameters $\hat{\Theta}(\mathcal{D}, \bar{h})$ and $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$

3: **Find an optimal scoring DAG**: obtain $\hat{\mathcal{D}}_{\text{opt}} = \text{argmin}_{\mathcal{D}\in\mathcal{D}_{\text{cand}}} \texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$

4: **Greedy backward deletion to estimate intervention set**: initialize $\hat{\mathcal{I}} = \mathcal{I}$ and
   (a) let $\{\hat{\Omega}_e\}_{e=1}^m$ be noise variance encoded in $\hat{\Theta}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h})$
   (b) estimate intervention strengths for each $j \in [p]$: $\xi_j := \frac{1}{m}\sum_e([\hat{\Omega}_e]_{j,j} - \frac{1}{m}\sum_e[\hat{\Omega}_e]_{j,j})^2$
   (c) remove weakest intervention: $\hat{\mathcal{I}}_{\text{opt}} \leftarrow \hat{\mathcal{I}}_{\text{opt}} \setminus \{\text{argmin}_j\, \xi_j : j \in \hat{\mathcal{I}}_{\text{opt}}\}$
   (d) supply $(\hat{\mathcal{D}}_{\text{opt}}, \hat{\mathcal{I}}_{\text{opt}})$ to Algorithm 0 and obtain $\texttt{score}_{\lambda,\gamma}(\hat{\mathcal{D}}_{\text{opt}}, \hat{\Theta}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}))$
   (e) repeat (c,d) until the likelihood score does not improve

5: **Output:** equivalence class $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$) using Definition 1

*R*emark 4: The guarantees of Corollary 7 can be extended to Algorithm 1. Specifically, suppose the conditions of this corollary hold and the candidate set of DAGs $\mathcal{D}_{\text{cand}}$ contains a member of the population interventional equivalence class, i.e. $\mathcal{D}_{\text{cand}} \cap \mathcal{I}^\star$-MEC($\mathcal{D}^\star$) $\neq \emptyset$. Furthermore, suppose that the alternating minimization technique in Algorithm 0 obtains a globally optimal solution; a formal characterization of when global optimality holds is non-trivial, and we leave this open question for future research. Then, we show in Appendix L.1 that in the infinite data limit, the output of Algorithm 1 is consistent, i.e. $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}^\star$ and $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$) $= \mathcal{I}^\star$-MEC($\mathcal{D}^\star$) with probability tending to one.

Selecting $(\lambda, \gamma, \bar{h})$ via cross-validation: In Appendix Section M, we propose an approach to exhaustively search over the equivalence classes indexed by $(\lambda, \gamma, \bar{h})$, and choose an optimal one based on validation with test data. The complexity of our validation approach is $\approx \tilde{t}((\bar{h}_{\max}+1)|\mathcal{D}_{\text{cand}}|+p)$, where $\bar{h}_{\max}$ is the maximum number of latent variables allowed in the model and $\tilde{t}$ represents the time it takes to score a DAG using Algorithm 0. The value for $\tilde{t}$ depends on the DAG and the data generating mechanism; in our numerical experiments, this is typically on the order of 5 seconds for $p = 20$ node graphs.

4.3 Using *UT-LVCE* to remove spurious edges from 'starting point' DAGs

We next consider settings where the DAGs in a candidate set are viewed as 'starting points' and may contain spurious dependencies due to potential latent confounding. Such scenarios naturally arise in practice. For example, a domain expert may be unsure about some of the edges in a DAG and may include them in the analysis to be conservative. In other contexts, the user may have deployed their favorite structure learning algorithm(s) to obtain a set of DAGs. Since many of the computationally efficient structural learning approaches (e.g. GES) do not account for the presence of latent variables, the fitted graph may be more dense than the population DAGs.

Our objective, in contexts where DAGs are viewed as starting points, is to use *UT-LVCE* to remove spurious dependencies and return a refined set of equally scoring DAGs. Our approach is based on the following simple observation: scoring starting point DAGs using Algorithm 0 may yield connectivity matrices that are more dense than the population connectivity matrix, although the magnitude of the spurious edges will be small. To remove these spurious edges, for each DAG in the candidate set, we greedily delete the weakest edge and compute a regularized likelihood score using Algorithm 0. We repeat this process until the score can no longer be improved. After pruning, we obtain a refined collection of candidate DAGs, which are then supplied to Algorithm 1 to identify an optimally scoring set of DAGs. A summary of the entire procedure is presented in Algorithm 2. Similar to Algorithm 1, in all our numerical experiments, we choose the regularization parameters $(\lambda, \gamma, \bar{h})$ via cross-validation; see Appendix Section M. The complexity of our exhaustive validation approach is $\approx \tilde{t}((\bar{h}_{\max}+1)[\sum_{\mathcal{D}\in\tilde{\mathcal{D}}_{\text{cand}}} \|\mathcal{D}\|_{\ell_0}]+p)$.

Algorithm 2 Improving 'starting point' DAGs via *UT-LVCE*

1: **Input**: data $\hat{\Sigma}_e, \hat{\pi}_e$ for $e \in [m]$; candidate DAG(s) $\tilde{\mathcal{D}}_{\text{cand}}$; regularization params. $\lambda, \gamma \geq 0$; # of latent vars. $\bar{h}$

2: **Backward deletion to remove spurious edges**: initialize $\mathcal{D}_{\text{cand}} = \emptyset$; for each $\mathcal{D} \in \tilde{\mathcal{D}}_{\text{cand}}$:
   (a) supply data and $(\mathcal{D}, \mathcal{I} = [p])$ to Algorithm 0 to find score $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$
   (b) let $\mathcal{D}$ be the DAG after deleting the smallest edge in magnitude in $\mathcal{D}$
   (c) repeat (a-b) until the likelihood score does not improve; add $\mathcal{D}$ to $\mathcal{D}_{\text{cand}}$

3: **Output**: supply $\mathcal{D}_{\text{cand}}$ to Algorithm 1 to find best scoring DAGs

*R*emark 5: As with Remark 4, the guarantees of Corollary 7 can be extended to Algorithm 2. In particular, if the candidate set of DAGs $\tilde{\mathcal{D}}_{\text{cand}}$ contains a DAG that is a supergraph of a DAG in the population interventional equivalence class, then, we show in Appendix L.2 that in the infinite data limit, the output of Algorithm 2 is consistent, i.e. $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}^\star$ and $\hat{\mathcal{I}}_{\text{opt}}$-MEC$(\hat{\mathcal{D}}_{\text{opt}}) = \mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$ with probability tending to one.

5. Synthetic and real experiments

Code to reproduce all the experiments can be found here: `https://github.com/juangamella/ut-lvce-paper`.

5.1 Synthetic experiments: recovering the interventional equivalence class

Setup: We consider a collection of $p = 20$ observed variables influenced by $h = 2$ latent variables. The entries of the latent coefficient matrix $\Gamma^\star \in \mathbb{R}^{p\times h}$ are generated IID from the distribution $\mathcal{N}(0, 1/\sqrt{p})$. The noise term ϵ_i for each coordinate i is distributed according to a zero mean Gaussian with variance chosen uniformly and independently from the interval $[0.5, 0.6]$. We suppose there are $m = 5$ environments, an observation environment $e = 1$ and four interventional environments $e \in \{2, 3, 4, 5\}$. For the observational environment, δ^e is all zeros and for the interventional environments and every $i \in \mathcal{I}^\star$, δ_i^e is a zero mean Gaussian distribution whose variance will be specified later. Similarly, the distribution of each latent variable are taken to be $\mathcal{H}_i^e \sim \mathcal{N}(0, \text{Unif}[0.2, 0.3] + \zeta_i^e)$ for every $i \in [h]$, where $\zeta_i^{e=1} = 0$ and is otherwise chosen uniformly and independently from the interval $[0.2, 1]$. The population connectivity matrix $B^\star$, the choice of intervention targets $\mathcal{I}^\star$, and the amount of data in every environment are specified later. In Appendix Section N.1, we provide additional experiments for the following settings: weaker interventions on observed variables and stronger latent effects. Finally, in Appendix Section N.2, we illustrate the performance of our method with a varying number of latent variables ($h = 3, 4, 10$).

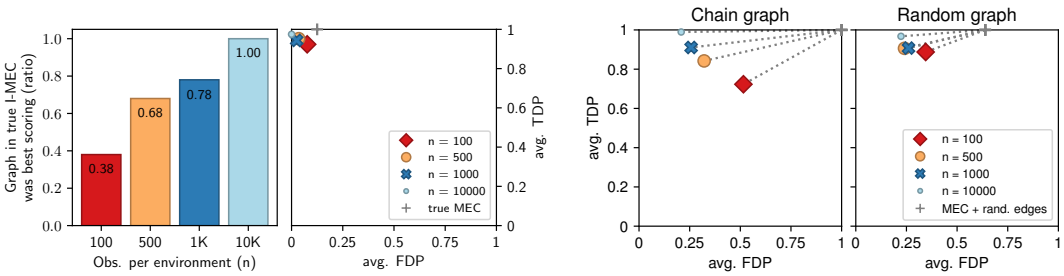
Metrics to assess the quality of an estimated equivalence class: To quantify the 'closeness' of an estimated interventional equivalence class $\hat{\mathcal{I}}$-MEC($\hat{\mathcal{D}}$) to the population interventional equivalence class $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$), we use the following two metrics:

$$\begin{aligned}
\text{FDP}: &\max_{\mathcal{D}_2 \in \hat{\mathcal{I}}\text{-MEC}(\hat{\mathcal{D}})} \min_{\mathcal{D}_1 \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)} [\#\text{ edges in } \mathcal{D}_2 \text{ not in } \mathcal{D}_1]/[\#\text{ edges in } \mathcal{D}_2], \\
\text{TDP}: &\min_{\mathcal{D}_1 \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)} \max_{\mathcal{D}_2 \in \hat{\mathcal{I}}\text{-MEC}(\hat{\mathcal{D}})} [\#\text{ edges in } \mathcal{D}_1 \text{ and in } \mathcal{D}_2]/[\#\text{ edges in } \mathcal{D}_1].
\end{aligned} \tag{9}$$

The metric FDP is akin to false discovery proportion and measures the ratio of spurious edges contained in the DAGs of the estimated interventional equivalence class. The metric TDP is akin to true discovery proportion and measures the proportion of true edges (in the DAGs of the population interventional equivalence class) that are also DAGs in the estimated interventional equivalence class. It is straightforward to check that $\hat{\mathcal{I}}$-MEC($\hat{\mathcal{D}}$) = $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) if and only if FDP = 0 and TDP = 1.

5.1.1 *UT-LVCE* with specified input DAGs

***UT-LVCE* with a candidate set of DAGs:** We generate the population DAG as follows: we first generate an Erdös-Renyi graph with edge probability 0.11 and orient the edges according to a random total ordering of the variables. The edge strengths are drawn uniformly at random from the interval $[0.5, 0.7]$. Let $\mathcal{I}^\star$ be ten indices chosen uniformly at random. The variance of the interventions on the observed variables δ_i^e is taken uniformly and independently from the interval $[6, 12]$. The candidate set of DAGs is taken to be the Markov equivalence class of $\mathcal{D}^\star$, which by definition is a superset of the interventional equivalence class $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$). We generate n observations for each environment, where n is chosen from the set $\{100, 500, 1000, 10000\}$. To illustrate the effectiveness of the scoring function, we supply the data and each candidate DAG, with $\mathcal{I} = [p]$ to Algorithm 0. The left plot in Figure 5a displays the proportion of instances, across 50 independent trials, that the best scoring DAG is inside $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$). Note that 1/3 of the graphs in the candidate



(a) Equivalence class of best scoring DAGs using Algorithm 1

(b) Removing spurious edges using Algorithm 2

Figure 5: a) proportion of instances, among 50 independent trials, that the best scoring DAG in Step 3 of Algorithm 1 is in $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) and average FDP and TDP of the estimated $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$); b) performance of Algorithm 2 with the 'starting point' DAGs MEC($\mathcal{D}^\star$) + 20 random edges.

set of DAGs are also in $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$). We observe that as the sample size increases, the scoring function becomes more accurate and correctly outputs a member of the interventional equivalence class. To obtain an equivalence class of best scoring DAGs, we apply Algorithm 1. As shown in the right plot in Figure 5a, the estimated equivalence class is close to the true equivalence class, even when $n = 500$.

***UT-LVCE* with 'starting point' DAGs:** We consider the setting described above and generate two different population DAGs $\mathcal{D}^\star$: a chain graph and an Erdös-Renyi graph with edge probability 0.11, with the edge strengths of each graph drawn uniformly at random from the interval $[0.5, 0.7]$. In each case, the 'starting point' DAGs are taken to be ones in MEC($\mathcal{D}^\star$) with 20 edges added at random to each graph (picked uniformly, without replacement, from all valid edge additions). We supply the data and the starting point DAGs to Algorithm 2. Ideally, Algorithm 2 removes spurious edges and improves upon the original set of DAGs to identify a class of best-scoring DAGs that is close to the population interventional equivalence class. Figure 5b confirms this to be the case. In particular, we observe that the estimated interventional equivalence class, averaged across 50 independent trials, has a small average FDP and large average TDP. Furthermore, we see that as compared to the 'starting point' DAGs, Algorithm 2 produces an estimate with substantially smaller false discoveries without much loss in power.

5.1.2 *UT-LVCE* as a structure learning procedure and comparisons to other methods

A set of input DAGs may not be available a priori and must be learned from data. Thus, we use GES to obtain a collection of DAGs, although, in principle, any structural learning algorithm may be deployed. Since GES does not account for latent confounding, its output DAGs are typically dense and contain many spurious edges. Thus, as prescribed in Section 4.3, we apply Algorithm 2 to prune GES DAGs and return an interventional equivalence class. We compare the performance of our algorithm to three causal learning methods that account for latent effects: causal Dantzig (Rothenhäusler et al., 2019), backShift (Rothenhäusler et al., 2016), and LRpS-GES (Frot et al., 2019). We note here

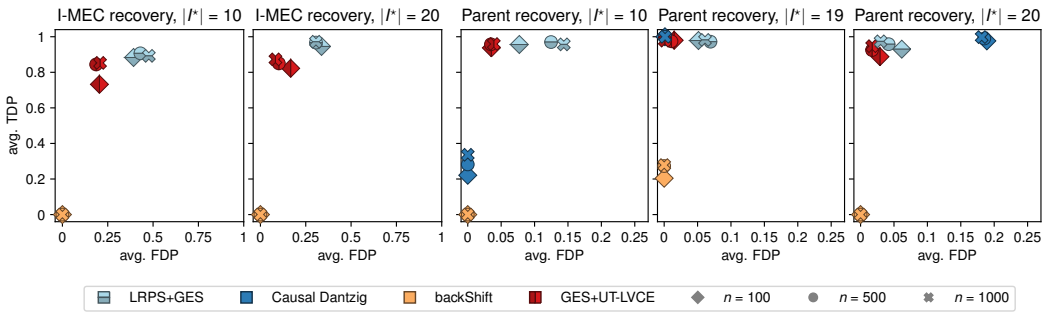


Figure 6: Performance of Algorithm 2 with GES starting DAGs and comparisons with other methods for different size of intervention targets $|\mathcal{I}^\star|$; the left two plots are performances over the entire graph and the right three graphs are performances locally for a target variable. Here, for assessing the quality of local structure recovery, we use the metrics in (9) restricted to the parental set(s) of the target variable.

that the first two methods exploit interventional data while LRpS-GES only operates with observational data. Furthermore, causal Dantzig performs local structural learning around a target variable of interest while the other two methods yield a causal model over the entire graph. Throughout, we consider the synthetic setup at the beginning of this section and generate 50 Erdös-Renyi DAGs with edge probability 0.11 and edge strengths drawn uniformly at random from the interval $[0.5, 0.7]$; we illustrate the robustness of our method to varying graph sparsity and varying number of latent variables in Appendix Section N.2. Furthermore, the magnitude of the interventions ξ_i^e are taken uniformly and independently from the interval $[3, 6]$. Finally, we generate n observations for each environment where n is chosen from the set $\{100, 500, 1000\}$.

Evaluating performance over the entire graph: We consider two settings: $|\mathcal{I}^\star| \in \{10, 20\}$. Figure 6 shows the average FDP and TDP, averaged across all the 50 DAGs and 10 runs for each DAG, for the outputs of *UT-LVCE*, backShift and LRpS-GES. As observed in Figure 6, *UT-LVCE* yields an estimated equivalence class with small average FDP and a large TDP, and performs more favorably compared to the other methods, especially in the setting with partial interventions. We also observe that LRpS-GES produces substantially larger false discoveries as it does not exploit interventional data for improved identifiability, and that backShift yields poor estimates since there are interventions on the latent variables. We note that the performance of *UT-LVCE* is naturally affected by the 'goodness' of the input GES DAGs. In particular, we show in Appendix Section N.3 that if any of the GES DAGs is a supergraph of a DAG in $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$), the performance of *UT-LVCE* substantially improves. Finally, as announced earlier, our procedure *UT-LVCE* can take as input DAGs produced by any structural learning algorithm. As an example, the user may take the DAGs obtained by GES as well as those from backShift and LRpS-GES as input to Algorithm 2, although we do not explore this in our experiments.

Evaluating local structure recovery: We consider a similar setting as above and let X_p be the target variable of interest. When generating our Erdös-Renyi DAGs, we discard DAGs where the target variable has less than two parents and obtain a total of 50 DAGs. We additionally consider the setting with $|\mathcal{I}^\star| = 19$ where all but the target

variable has received an intervention. We observe that for $|\mathcal{I}^\star| = 10$, Causal Dantzig has very low power, which can be attributed to yielding a single estimate even if the parental set is unidentifiable and requiring interventions on all but the target variable for consistency. When $|\mathcal{I}^\star| = 19$, Causal Dantzig obtains accurate recovery of the parental set even though there are interventions on the latent variables. In Appendix Section N.4, we observe that when the magnitude of the latent interventions are made to be stronger, Causal Dantzig performs poorly as compared to *UT-LVCE*. Finally, when $|\mathcal{I}^\star| = 20$, Causal Dantzig yields inaccurate estimates since there are interventions on the target variable of interest.

5.2 Analysis on real data

5.2.1 Protein expressions

We next analyze the protein mass spectroscopy dataset (Sachs et al., 2005). This dataset (downloaded from `https://www.bnlearn.com/research/sachs05/index.html`) contains a large number of measurements of the abundance of 11 phosphoproteins and phospholipids recorded under different experimental conditions in primary human immune system cells. The different experimental conditions are characterized by associated reagents that inhibit or activate signaling nodes, corresponding to interventions at different points of the protein-signaling network. Following the previous works (Mooij and Heskes, 2013; Meinshausen et al., 2016), we take 8 environments consisting of an observational environment and 7 interventional environments.

Multiple papers have applied their structural learning algorithm to identify the causal relationships among the 11 proteins (Sachs et al., 2005; Eaton and Murphy, 2007; Mooij and Heskes, 2013; Meinshausen et al., 2016). Each proposed method returns a DAG (potentially multiple due to non-identifiability) with some commonalities among the output structures, but also many differences. Naturally, the following questions arise: i) how well does each DAG fit the data, and which one is most representative of the data? ii) are there other DAGs in the equivalence class of the best scoring DAG that fit the data equally well? and iii) can any spurious edges in the DAGs be removed to obtain a better fit to data? As outlined next, our procedure *UT-LVCE* is useful for addressing these questions.

Best scoring among reported DAGs in the literature: We use Algorithm 1 to score the DAGs obtained by previous methods. We keep 0.05% of the data for computing test performance. Of the remaining 0.95% of the data, we take 70% for training and the remaining 30% for validation. The number of latent variables $\bar{h}$ and the regularization parameters λ, γ are selected via holdout validation. We obtain a causal model associated with each DAG and evaluate the corresponding negative log-likelihood score on the test set. For reproducibility, we repeat this experiment with 50 different random splits of training/validation datasets. Figure 7a presents the box-plot of the test scores for each DAG. Several remarks are in order. First, the top three best scoring DAGs (displayed alongside with the other 13 DAGs in Appendix Section O) are produced by a method that accounts for latent variables (Meinshausen et al., 2016); the other structural learning procedures assume all relevant variables are observed. Related to the previous point, we find that there are strong latent effects on the protein network. As an example, for the best scoring DAG, our algorithm finds on average 1.16 latent variables. Furthermore, for the top three scoring

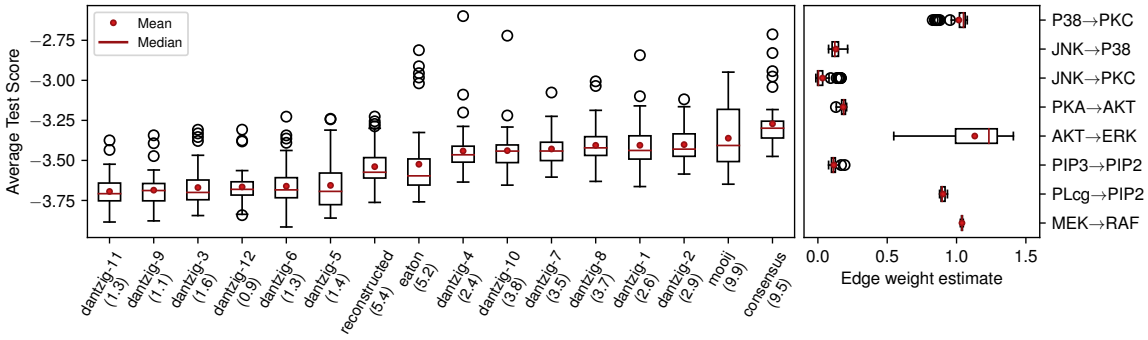


Figure 7: Performance of Algorithms 1 and 2 on the protein expressions dataset with 50 random splits of the data: a) test score of each DAG produced in the literature using Algorithm 1 and the average edges pruned (shown below for each DAG) using Algorithm 2; b) edge strengths of the best scoring DAG "Causal Dantzig 11".

DAGs, we also find that many of the proteins have received a strong intervention; this is likely due to off-target effects that were also reported in Eaton and Murphy (2007). The presence of interventions on many of the variables implies that the equivalence class of all of these top three DAGs are singletons. Finally, in Figure 7b, we present a boxplot of the edge strengths for the top scoring DAG.

Removing spurious edges: We next explore whether any spurious edges can be removed from these DAGs. To that end, we apply Algorithm 2 to each DAG. We observe that the top scoring DAGs produced by Meinshausen et al. (2016) are rather stable as compared to the other DAGs, with ≈ 1 edges removed on average by our procedure. This is consistent with the fact that Meinshausen et al. (2016) accounts for latent confounding and thus is likely to contain fewer spurious edges. We note that for the best scoring DAG, the edge that is removed most often is JNK $\rightarrow$ PKC; indeed, this edge has weak strength (see Figure 7b) and has not been reported in any other DAG in the literature.

5.2.2 California reservoirs

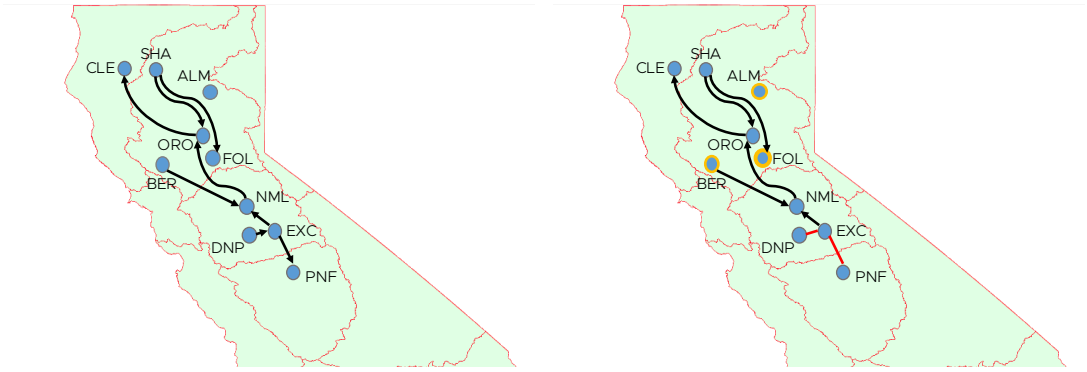
The California reservoir network consists of ≈ 1530 reservoirs that act as buffers against severe drought conditions and are a major source of water for agricultural use, hydropower generation, and industrial use. Water managers of these reservoirs have to assess the likelihood of system-wide failure and the effectiveness of potential policies. Due to similarities in hydrological attributes (e.g. altitude, drainage area, spatial location), the reservoir network is highly interconnected. Thus, effective reservoir management requires an understanding of reservoir interdependencies. Taeb et al. (2017) used historical data of volumes of the largest 55 reservoirs to obtain an undirected graphical model of the California reservoir network. This previous analysis, however, does not provide causal implications: namely, how change in the management of one reservoir (i.e. an intervention) affects the entire system. To that end, we explore the utility of *UT-LVCE* for learning causal relationships among the reservoirs.

We consider the 10 largest reservoirs (with respect to capacity) in California, where daily volume data (downloaded from `https://github.com/armeentaeb/WRR-Reservoir`) are available during the period of study (January 2003–December 2015). Following the preprocessing steps in Taeb et al. (2017), we average the data from daily down to 156 monthly observations. A seasonal adjustment step is performed to remove predictable seasonal patterns. The resulting data was demonstrated in Taeb et al. (2017) to be well-approximated by a multivariate Gaussian distribution.

The reservoir data is not IID as its distribution varies depending on the severity of the drought. In particular, during a drought period, a reservoir manager may decide to reduce the outflow of water, and thus effectively decrease the variability in the reservoir volume; this is in contrast to a wet period where more outflow is allowed, as the reservoir is expected to be replenished. Based on the intuition described above, we organize our reservoir data into four 'environments' or time-blocks based on the severity of the drought conditions: an environment during a normal period (2003-2006, 2010-2012) with no drought conditions, an environment associated to an abnormally dry period (2007, 2013), an environment associated to a moderate drought period (2008-2009), and an environment associated to a severe drought period (2014-2015).

Unlike the protein expression dataset, no candidate DAGs are available a priori for the reservoir dataset. Thus, we employ GES on the first environment (normal period) to obtain a collection of 'starting point' DAGs. These DAGs are then supplied to Algorithm 2, where the number of latent variables $\bar{h}$ as well as the regularization parameters λ, γ are selected via holdout validation with a $(70\%, 30\%)$ training and validation set split for 10 different random splits. For each split, we obtain a causal model and a corresponding equivalence class and then choose the model that obtains the best likelihood score on the overall data. The optimally scoring model consists of two latent variables ($\bar{h} = 2$) and an interventional equivalence class presented in Figure 8a. The connections in the learned DAG are between pairs of reservoirs with at least one of these commonalities: i) similar hydrological attributes (e.g. hydrological zone and elevation) and ii) coordinated management by a district or a state-wide project. For example, the reservoirs New Melones (NML), Don Pedro (NP), New Exchequer (EXC), and Pine Flat (PNF) are all in the San Joaquin district. Further, Shasta (SHA), Trinity (CLE), Oroville (ORO), and Folsom (FOL) are in the network of Central Valley and State Water projects and their reservoir operations are coordinated.

We next analyze the estimated locations and magnitudes of the interventions. Recall that the locations are encoded in the estimate $\hat{\mathcal{I}}_{\text{opt}}$ and the strength of the interventions are computed via the metric ξ_j for every $j \in \hat{\mathcal{I}}_{\text{opt}}$ (see Section 4.2). Our model identifies interventions on all reservoirs except Pine Flat. The strongest estimated intervention is on Lake Almanor, which is consistent with the fact that during the 2014-2015 drought period, there was little to no outflow of water in this reservoir. Finally, aside from the reservoirs {'ALM', 'BER', 'FOL'}, the intervention strengths on the remaining reservoirs are rather small (i.e. below the level $\xi_j \leq 0.01$). However, likely due to the small sample size, these reservoirs were included in the list of intervention targets after validation. Note that overestimating the list of intervention targets may lead to discarding plausible causal mechanisms, as more identifiability is claimed than present in the data. To remain 'conservative', in



(a) output of Algorithm 2 (b) 'conservative' equivalence class (see text)

Figure 8: Graphical structure of the 10 largest reservoirs in California (with respect to capacity): a) equivalence class consisting of a unique DAG obtained by Algorithm 2; b) conservative equivalence class consisting of multiple DAGs (obtained from directing the red edges) after identifying strong interventions (shown in orange) from the output of Algorithm 2.

Figure 8b, we present the interventional Markov equivalence class $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$), where $\hat{\mathcal{I}}_{\text{opt}} = \{$'ALM', 'BER', 'FOL'$\}$ and $\hat{\mathcal{D}}_{\text{opt}}$ is the DAG in Figure 8a. The resulting structure highlights that certain edges (shown in red) may not be identifiable.

6. Discussion and Future Work

In this paper, we proposed a framework to model unspecific interventional data among a collection of observed and latent variables. This framework allows for additive interventions on all components of the system, including a response variable of interest or the latent variables. Further, we presented an algorithm *UT-LVCE* to fit DAGs to this model and obtain an equivalence class of DAGs that best explains the data. There are several interesting directions for further investigation that arise from our work. In Section 4.3, we discussed the setting where no DAGs are available a-priori and proposed using any structural learning algorithm to obtain a set of 'starting point' DAGs; these are then subsequently pruned by Algorithm 2 to arrive at an estimate for the interventional equivalence class. While the empirical results in Section 5 support the utility of our heuristics, there is much room for more rigorous optimization techniques to search over the space of equivalent DAGs with respect to the scoring function (4) (e.g. provably consistent greedy methods). Further, the interventional model (2) assumes a linear relationship between the observed and latent variables. It would be of practical interest to explore extensions of our framework to non-linear settings, or alternatively, characterize the extent to which linear models capture the causal effects.

Acknowledgments

We thank the AE and the referees for their valuable comments that greatly improved the paper. AT received funding in part from the Royalty Research Fund at the University of Washington and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461). JG and PB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461).

Supplementary Material

Appendix A. Notations

For a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, we denote $\|M\|_2$ to be largest singular value. For a symmetric matrix $M \in \mathbb{S}^d$, we denote $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to be the maximum and minimum eigenvalue of M, respectively. For a symmetric matrix $M \in \mathbb{S}^d$, we denote degree(M) to be the maximum number of nonzero elements in any row (or equivalently column) of the matrix. Recall that for an undirected graph $\mathcal{G}$, we denote degree$[\mathcal{G}]$ to be the maximal degree of $\mathcal{G}$.

Appendix B. Proof of Theorem 2

The proof of this theorem relies on some lemmas:

Lemma 9 (A property of $(\mathrm{Id}-\tilde{B})^{-1}(\mathrm{Id}-B)^{-1}$ (Gamella et al., 2022)) *Let $B, \tilde{B} \in \mathbb{R}^{p\times p}$ be two matrices that can be made to be lower-triangular with zeros on the diagonal after row and column permutations (or equivalently, the matrices correspond to two DAGs). If for $\tilde{\Omega}, \Omega \in \mathbb{D}^p_{++}$, $(\mathrm{Id}-B)^{-1}\Omega(\mathrm{Id}-B)^{-T} = (\mathrm{Id}-\tilde{B})^{-1}\tilde{\Omega}(\mathrm{Id}-\tilde{B})^{-T}$, then the following statements are equivalent:*

p1) $\mathrm{support}(\tilde{B}_{i,:}) = \mathrm{support}(B_{i,:})$.

p2) $[(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1}]_{i,:} = \mathbf{e}_i^T$.

p3) $[(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1}]_{:,i} = \mathbf{e}_i$.

Proof For completeness, we include the proof in Gamella et al. (2022). We break down the proof in the following steps:

$\underline{p1 \Leftrightarrow p2}$: The direction $p2 \Rightarrow p1$ follows immediately. For the direction $p1 \Rightarrow p2$, define the variables $\bar{X}$ via the following structural equation model

$$\bar{X} = B\bar{X} + \epsilon \quad , \quad \epsilon \sim \mathcal{N}(0, \Omega). \tag{10}$$

Since $(\mathrm{Id}-B)^{-1}\Omega(\mathrm{Id}-B)^{-T} = (\mathrm{Id}-\tilde{B})^{-1}\tilde{\Omega}(\mathrm{Id}-\tilde{B})^{-T}$, we have that equivalently:

$$\bar{X} = \tilde{B}\bar{X} + \tilde{\epsilon} \quad , \quad \tilde{\epsilon} \sim \mathcal{N}(0, \tilde{\Omega}). \tag{11}$$

Let $S = \mathrm{support}(B_{i,:}) = \mathrm{support}(\tilde{B}_{i,:})$. Then, equations (10) and (11) imply that $B_{i,:}$ and $\tilde{B}_{i,:}$ are both the regression coefficients from regressing $\bar{X}_S$ onto $\bar{X}_i$. Thus, $\tilde{B}_{i,:} = B_{i,:}$.

$\underline{p2 \Rightarrow p3}$ The property $(\mathrm{Id}-B)^{-1}\Omega(\mathrm{Id}-B)^{-T} = (\mathrm{Id}-\tilde{B})^{-1}\tilde{\Omega}(\mathrm{Id}-\tilde{B})^{-T}$ means that $(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1}$ has orthogonal row vectors. Since $\mathbf{e}_i^T(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1} = \mathbf{e}_i^T$, it then follows that $[(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1}]_{:,i} = \mathbf{e}_i$.

$\underline{p3 \Rightarrow p1}$ Notice that:

$$(\mathrm{Id}-\tilde{B})(\mathrm{Id}-B)^{-1}\Omega(\mathrm{Id}-B)^{-T}(\mathrm{Id}-\tilde{B})^T = \tilde{\Omega},$$

which can be rewritten as:

$$(\mathrm{Id}-\tilde{B}) = \tilde{\Omega}(\mathrm{Id}-\tilde{B})^{-T}(\mathrm{Id}-B)^T\Omega^{-1}(\mathrm{Id}-B). \tag{12}$$

Let $\Omega_{i,i}^{-1} = \alpha$ and $\tilde{\Omega}_{i,i} = \beta$, where $\alpha, \beta \neq 0$. Property $p2$ implies that $\mathbf{e}_i^T(\mathrm{Id} - \tilde{B})^{-T}(\mathrm{Id} - B)^T = \mathbf{e}_i^T$. Thus, relation (12) lets us conclude that:

$$\mathbf{e}_i^T(\mathrm{Id} - \tilde{B}) = \alpha\beta\mathbf{e}_i^T(\mathrm{Id} - B),$$

and thus property $p1$. ■

Lemma 10 (equivalent covariance models without latent variables) *Consider an SCM* (2) *with structure given by a DAG $\mathcal{D}$, with no latent confounders (i.e. $h = 0$), connectivity matrix B, noise variances $\{\Omega_e\}_{e=1}^m$ and intervention targets $\mathcal{I}$. For any DAG $\tilde{\mathcal{D}} \in \mathcal{I}$-MEC$(\mathcal{D})$, there exists a connectivity matrix $\tilde{B} \sim \tilde{\mathcal{D}}$ and noise variances $\{\tilde{\Omega}_e\}_{e=1}^m \subseteq \mathbb{D}_{++}^p$ such that $(\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\mathrm{Id} - \tilde{B})^{-T}$ for every $e = [m]$. Furthermore, $\mathbb{I}(\{\tilde{\Omega}_e\}_{e=1}^m) = \mathcal{I}$.*

Proof Decompose Ω_e as: $\Omega_e = \Omega_0 + \Delta\Omega_e$ where $\Omega_0 \in \mathbb{D}_{++}^p$ and $[\Delta\Omega_e]_{i,i} = 0$ for all $e = 1, 2, \ldots, m$ and $i \notin \mathcal{I}$. Note that for any $\tilde{\mathcal{D}} \in \mathcal{I}$-MEC$(\mathcal{D})$, there exists a $\tilde{B} \sim \tilde{\mathcal{D}}$ and $\tilde{\Omega}_0 \in \mathbb{D}_{++}^p$ such that $(\mathrm{Id} - B)^{-1}\Omega_0(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_0(\mathrm{Id} - \tilde{B})^{-T}$. Since $\mathrm{support}(\tilde{B}_{i,:}) = \mathrm{support}(B_{i,:})$ for all $i \in \mathcal{I}$, we have by Lemma 9 that $[(\mathrm{Id} - \tilde{B})(\mathrm{Id} - B)^{-1}]_{:,i} = e_i$ and $[(\mathrm{Id} - \tilde{B})(\mathrm{Id} - B)^{-1}]_{i,:} = e_i^T$ for all $i \in \mathcal{I}$. Then, set $\Delta\tilde{\Omega}_e = \Delta\Omega_e$ and note that $(\mathrm{Id} - \tilde{B})(\mathrm{Id} - B)^{-1}\Delta\Omega_e(\mathrm{Id} - B)^{-T}(\mathrm{Id} - \tilde{B})^T = \Delta\tilde{\Omega}_e$. Defining $\tilde{\Omega}_e = \tilde{\Omega}_0 + \Delta\tilde{\Omega}_e$, we have that $(\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\mathrm{Id} - \tilde{B})^{-T}$ for all $e = 1, 2, \ldots, m$ as desired. Furthermore, since $\tilde{\Omega}_e = (\mathrm{Id} - \tilde{B})(\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T}(\mathrm{Id} - \tilde{B})^T$ and Ω_e is positive-definite matrix, we conclude that $\tilde{\Omega}_e$ is positive-definite. Finally, the property $\mathbb{I}(\{\tilde{\Omega}_e\}_{e=1}^m) = \mathcal{I}$ follows by construction. ■

Lemma 11 (Sufficient condition for equally scoring parameters sets) *Let $\Theta = (B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$ and $\tilde{\Theta} = (\tilde{B}, \tilde{\Gamma}, \tilde{\mathcal{I}}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ be a set of parameters that satisfy for every $e = 1, 2, \ldots, m$:*

$$(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - \tilde{B})^{-1}(\tilde{\Omega}_e + \tilde{\Gamma}\tilde{\Psi}_e\tilde{\Gamma}^T)(\mathrm{Id} - \tilde{B})^{-T},$$

with $\mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m)$ and $\tilde{\mathcal{I}} = \mathbb{I}(\{\tilde{\Omega}_e\}_{e=1}^m)$. Furthermore, suppose that $|\tilde{\mathcal{I}}| = |\mathcal{I}|$, $\|\mathcal{D}\|_{\ell_0} = \|\tilde{\mathcal{D}}\|_{\ell_0}$, moral$(\mathcal{D})$ = moral$(\tilde{\mathcal{D}})$ where $B \sim \mathcal{D}$ and $\tilde{B} \sim \tilde{\mathcal{D}}$. Then, score$_{\lambda,\gamma}(\mathcal{D}, \Theta)$ = score$_{\lambda,\gamma}(\tilde{\mathcal{D}}, \tilde{\Theta})$ for every $\lambda, \gamma \geq 0$.

Proof The parameters Θ and $\tilde{\Theta}$ specify the precision matrices for every $e = 1, 2, \ldots, m$:

$$K_e = (\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B),$$
$$\tilde{K}_e = (\mathrm{Id} - \tilde{B})^T(\tilde{\Omega}_e + \tilde{\Gamma}\tilde{\Psi}_e\tilde{\Gamma}^T)^{-1}(\mathrm{Id} - \tilde{B}).$$

Furthermore, the regularized likelihood score for each model is given by:

$$\mathrm{score}_{\lambda,\gamma}(\mathcal{D}, \Theta) = \sum_{e=1}^m \hat{\pi}_e(-\log\det(K_e) + \mathrm{trace}(K_e\hat{\Sigma}_e)) + \lambda\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}),$$
$$\mathrm{score}_{\lambda,\gamma}(\tilde{\mathcal{D}}, \tilde{\Theta}) = \sum_{e=1}^m \hat{\pi}_e(-\log\det(\tilde{K}_e) + \mathrm{trace}(\tilde{K}_e\hat{\Sigma}_e)) + \lambda\mathcal{R}_\gamma(\tilde{\mathcal{D}}, \tilde{\mathcal{I}}).$$

By the assumptions of the lemma, $K_e = \tilde{K}_e$ for every $e = 1, 2, \ldots, m$, and $\mathcal{R}_\gamma(\tilde{\mathcal{D}}, \tilde{\mathcal{I}}) = \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})$. Thus, $\text{score}_{\lambda,\gamma}(\mathcal{D}, \Theta) = \text{score}_{\lambda,\gamma}(\tilde{\mathcal{D}}, \tilde{\Theta})$. ■

We are now ready to prove Theorem 2.

Proof [Proof of Theorem 2]Take any $\tilde{\mathcal{D}} \in \mathcal{I}\text{-MEC}(\mathcal{D})$. By Lemma 10, we have that that there exists a connectivity matrix $\tilde{B} \sim \tilde{\mathcal{D}}$ and $\{\tilde{\Omega}_e\}_{e=1}^m \subseteq \mathbb{D}_{++}^p$ with $\mathcal{I} = \mathbb{I}(\{\tilde{\Omega}_e\}_{e=1}^m)$ such that $(\text{Id} - B)^{-1}\Omega_e(\text{Id} - B)^{-T} = (\text{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\text{Id} - \tilde{B})^{-T}$ for every $e = 1, 2, \ldots, m$. Furthermore, let $\tilde{\Gamma} = (\text{Id} - \tilde{B})(\text{Id} - B)^{-1}\Gamma$ and $\tilde{\Psi}_e = \Psi_e$ for every $e = 1, 2, \ldots, m$. It is straightforward to show that the parameters $\tilde{\Theta} = (\tilde{B}, \tilde{\Gamma}, \mathcal{I}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ satisfy the condition of Lemma 11 and we can conclude that $\text{score}_{\lambda,\gamma}(\mathcal{D}, \Theta) = \text{score}_{\lambda,\gamma}(\tilde{\mathcal{D}}, \tilde{\Theta})$ for every $\lambda, \gamma \geq 0$. ■

Appendix C. Characterization of optimally scoring DAGs via *UT-LVCE* in the infinite sample regime

Throughout, we consider the asymptotic regime where $p/n_e \to 0$ as $n_e \to \infty$ for all $e \in [m]$. For every $\mathcal{D} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$, let $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ be any set of parameters with $B \sim \mathcal{D}$ that specify the population covariance matrix $\Sigma_e^\star$ in environment e, i.e. $\Sigma_e^\star = (\text{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma)^{-1}(\text{Id} - B)^{-T}$. In our analysis, the parameter space $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ is assumed to be compact. For notational ease, we denote the constraint set for the parameters $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ as $\mathcal{F}$. 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 $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$, $\|(\text{Id} - B)\|_F \leq \tau_1$, $\min_e \sigma_{\min}(\Omega_e + \Gamma\Psi_e\Gamma^T) \geq \tau_2$, and $\max_e \sigma_{\max}(\Omega_e + \Gamma\Psi_e\Gamma^T) \leq \tau_3$. Here, $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ denote maximum and minimum singular values of an input matrix, respectively. Then, the space of connectivity matrices and noise variances has the additional constraints $\|(\text{Id} - B)\|_F \leq \tau_1$, $\min_e \sigma_{\min}(\Omega_e + \Gamma\Psi_e\Gamma^T) \geq \tau_2$, and $\max_e \sigma_{\max}(\Omega_e + \Gamma\Psi_e\Gamma^T) \leq \tau_3$ in the score function (7). As before, we denote the score of a DAG $\mathcal{D}$ with a pre-specified number of latent variables $\bar{h}$ when constrained to the compact parameter space $\mathcal{F}$ as $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \bar{h})$.

With this setup, we have the following characterization of the optimally scoring models, scored according to (7).

Proposition 12 (Characterization of optimally scoring DAGs) *Suppose that the set of parameters is constrained to be in the compact space $\mathcal{F}$. Suppose that $n_e \to \infty$, γ is set to be any bounded scalar, and that $\lambda \to 0$ while satisfying*

$$\lambda \gg \left| \max_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m) \in \mathcal{F}} \sum_{e=1}^m \text{tr}((\text{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\text{Id} - B)(\hat{\pi}_e\hat{\Sigma}_e - \pi_e^\star\Sigma_e^\star)) \right|,$$
$$+ \left| \max_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m) \in \mathcal{F}} \sum_{e=1}^m (\hat{\pi}_e - \pi_e^\star) \log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) \right|,$$

where $\Sigma_e^\star$ represents the true covariance in environment e and $\pi_e^\star = \lim_{all\ n_e \to \infty} \frac{n_e}{\sum_{e=1}^m n_e}$. Then, with probability tending to one, the set of optimally scoring models (according to (7))

are solution to the following optimization problem:

$$\begin{aligned} \underset{\mathcal{D},\bar{h},B,\Gamma,\mathcal{I},\{\Omega_e,\Psi_e\}_{e=1}^m}{\operatorname{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D},\mathcal{I}) \\ \textit{subject-to}: \quad & B \sim \mathcal{D}, \mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m), \textit{ and} \\ & \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \textit{ for every } e \in [m]. \end{aligned} \tag{13}$$

The proof of Proposition 12 relies on the following lemma:

Lemma 13 *The minimizers of the following optimization*

$$\begin{aligned} \underset{\mathcal{D},\bar{h},\mathcal{I}}{\operatorname{argmin}} \; \underset{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}}{\operatorname{argmin}} & \sum_{e=1}^m \pi_e^\star \Big(\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) \\ & \qquad + \operatorname{tr}((\mathrm{Id}-B)^T(\Omega_e+\Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)\Sigma_e^\star)\Big) \\ \textit{subject-to: } & B \sim \mathcal{D} \;\; ; \;\; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I} \end{aligned}$$

are given by parameters $\mathcal{D}, \bar{h}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m$ that satisfy:

$$\begin{aligned} & B \sim \mathcal{D}, \mathcal{I} \subseteq \mathbb{I}(\{\Omega_e\}_{e=1}^m), \textit{ and} \\ & \quad \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \textit{ for every } e \in [m]. \end{aligned}$$

Proof [Proof of Lemma 13] Let $\mathcal{M}(B,\Gamma,\mathcal{I},\Omega_e,\Psi_e)$ denote a model associated with each equation in the SCM (2) (main paper). For notational convenience, we use the short-hand notation $\mathcal{M}_e$ for this model. We let $\Sigma(\mathcal{M}_e) = (\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)(\mathrm{Id}-B)^{-T}$ be the associated covariance model parameterized by the parameters $(B,\Gamma,\mathcal{I},\Omega_e,\Psi_e)$. The optimal solutions of the optimization problem in Lemma 13 can then be equivalently reformulated as:

$$\underset{\{\mathcal{M}_e\}_{e=1}^m}{\operatorname{argmin}} \sum_{e=1}^m \pi_e^\star \mathtt{KL}(\Sigma_e^\star, \Sigma(\mathcal{M}_e)), \tag{14}$$

where $\mathtt{KL}(\cdot,\cdot)$ represents the Gaussian KL-divergence. Notice that for the decision variables $\mathcal{M}_e^\star = (B^\star,\Gamma^\star,\mathcal{I}^\star,\Omega_e^\star,\Psi_e^\star)$ for each $e = 1,2,\ldots,m$, (14) achieves zero loss. Hence, any other optimal solution of (14) must yield zero loss, or equivalently, $\Sigma(\mathcal{M}_e) = \Sigma_e^\star$ for any optimal collection $\{\mathcal{M}_e\}_{e=1}^m$. ■

Proof [Proof of Proposition 12] For a fixed $\mathcal{D}$, number of latent variables $\bar{h}$, and set of intervention targets $\mathcal{I}$, we let $S_{\lambda,\gamma}(\mathcal{D},\bar{h},\mathcal{I})$ be the following score:

$$\begin{aligned} S_{\lambda,\gamma}(\mathcal{D},\bar{h},\mathcal{I}) := \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} & \sum_{e=1}^m \hat{\pi}_e \Big(\log\det(\Omega_e+\Gamma\Psi_e\Gamma^T) \\ & \qquad + \operatorname{tr}((\mathrm{Id}-B)^T(\Omega_e+\Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)\hat{\Sigma}_e)\Big) + \lambda\mathcal{R}_\gamma(\mathcal{D},\mathcal{I}). \\ \text{subject-to: } & B \sim \mathcal{D} \; ; \; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I} \end{aligned}$$

Notice the score $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$ is related to the the score $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$ in (4) via the following simple relation: $\texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h})) = \min_{\mathcal{I}} S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$, and thus:

$$\min_{\mathcal{D},\bar{h}} \texttt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h})) = \min_{\mathcal{D},\bar{h},\mathcal{I}} S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}). \tag{15}$$

We define the population analogue of the score $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$ below:

$$\begin{aligned} S^\star(\mathcal{D}, \bar{h}, \mathcal{I}) := \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} & \sum_{e=1}^m \pi_e^\star \Big(\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) \\ & + \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)\Sigma_e^\star)\Big). \\ \text{subject-to: } & B \sim \mathcal{D} \;\; ; \;\; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}, \end{aligned}$$

where $\Sigma_e^\star$ represents the true covariance in environment e and $\pi_e^\star = \lim_{\text{all } n_e\to\infty} \frac{n_e}{\sum_{e=1}^m n_e}$.

First, we show that for every $\mathcal{D}$, $\bar{h}$, $\mathcal{I}$, the score function $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \xrightarrow{p} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ as $n_e \to \infty$ for every $e \in [m]$. This follows by the compactness of the parameter space leading to uniform convergence of M-estimators, and that $\lambda \to 0$ and bounded γ (van de Geer, 2000). For more details, note that:

$$\begin{aligned} S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) - \lambda\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) = \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} & \sum_{e=1}^m \pi_e^\star \Big(\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) \\ & + \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)\Sigma_e^\star)\Big) \\ & + \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)(\hat{\pi}_e\hat{\Sigma}_e - \pi_e^\star\Sigma_e^\star)) \\ & + (\hat{\pi}_e - \pi_e^\star)\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T). \end{aligned}$$

From there, we have

$$\begin{aligned} S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \geq\; & S^\star(\mathcal{D}, \bar{h}, \mathcal{I}) \\ & - \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} \sum_{e=1}^m \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)(\hat{\pi}_e\hat{\Sigma}_e - \pi_e^\star\Sigma_e^\star)) \\ & - \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} (\hat{\pi}_e - \pi_e^\star)\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) + \lambda\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}), \\ S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \leq\; & S^\star(\mathcal{D}, \bar{h}, \mathcal{I}) \\ & + \max_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} \sum_{e=1}^m \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)(\hat{\pi}_e\hat{\Sigma}_e - \pi_e^\star\Sigma_e^\star)) \\ & + \max_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} (\hat{\pi}_e - \pi_e^\star)\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) + \lambda\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}). \end{aligned}$$

By the compactness constraint,

$$\begin{aligned} \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} \sum_{e=1}^m \operatorname{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)(\hat{\pi}_e\hat{\Sigma}_e - \pi_e^\star\Sigma_e^\star)) \to 0, \\ \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} (\hat{\pi}_e - \pi_e^\star)\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) \to 0, \end{aligned}$$

as $n_e \to \infty$. Furthermore, since $\lambda \to 0$ as $n_e \to \infty$ for every $e \in [m]$, and $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})$ is bounded for a finite γ, we have that $\lambda \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \to 0$. We can thus conclude $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \to S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ in the infinite data limit for every environment.

With the choice of λ in Proposition 12, we have that it is above fluctuations due to sampling error; it follows that for two population score equivalent models $(\mathcal{D}_1, \bar{h}_1, \mathcal{I}_1)$ and $(\mathcal{D}_2, \bar{h}_2, \mathcal{I}_2)$ with $S^\star(\mathcal{D}_1, \bar{h}_1, \mathcal{I}_1) = S^\star(\mathcal{D}_2, \bar{h}_2, \mathcal{I}_2)$, if $\mathcal{R}_\gamma(\mathcal{D}_1, \mathcal{I}_1) < \mathcal{R}_\gamma(\mathcal{D}_2, \mathcal{I}_2)$, then, there exists N such that for $n_e \geq N$ for every e, $S_{\lambda,\gamma}(\mathcal{D}_1, \bar{h}_1, \mathcal{I}_1) < S_{\lambda,\gamma}(\mathcal{D}_2, \bar{h}_2, \mathcal{I}_2)$. This allows us to conclude that:

$$\mathbb{P}\left(\operatorname{argmin} S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) = \operatorname{argmin} \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \text{ subject-to } \mathcal{D}, \bar{h}, \mathcal{I} \in \operatorname{argmin} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})\right) \to 1,$$

as $n_e \to \infty$ for every $e \in [m]$. In other words, we can conclude that in the infinite data regime, minimizers of (7) converge (with probability tending to one) to

$$\operatorname*{argmin}_{\mathcal{D}, \bar{h}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m \in \Theta_{\text{opt}}} \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}), \tag{16}$$

where

$$\begin{aligned}\Theta_{\text{opt}} := \operatorname*{argmin}_{\mathcal{D}, \bar{h}, \mathcal{I}} \operatorname*{argmin}_{(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m) \in \mathcal{F}} & \sum_{e=1}^m \pi_e^\star \Big(\log\det(\Omega_e + \Gamma \Psi_e \Gamma^T) \\ & + \operatorname{tr}((\mathrm{Id} - B)^T (\Omega_e + \Gamma \Psi_e \Gamma^T)^{-1} (\mathrm{Id} - B) \Sigma_e^\star) \Big) \\ \text{subject-to: } & B \sim \mathcal{D} \;\; ; \;\; \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}.\end{aligned}$$

By Lemma 13, we then conclude that with probability tending to one, the minimizers of (7) in the infinite data limit are:

$$\begin{aligned}\operatorname*{argmin}_{\mathcal{D}, \bar{h}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m} & \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \\ \text{subject-to : } & B \sim \mathcal{D}, \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}, \text{ and} \\ & \Sigma_e^\star = (\mathrm{Id} - B)^{-1} (\Omega_e + \Gamma \Psi_e \Gamma^T)^{-1} (\mathrm{Id} - B)^{-T} \text{ for every } e \in [m].\end{aligned}$$

Finally, note that $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})$ is monotonic in the size of $\mathcal{I}$, we can replace the constraint $\mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}$ with $\mathbb{I}(\{\Omega^e\}_{e=1}^m) = \mathcal{I}$ and attain the desired result. ■

Appendix D. Equivalent causal models

D.1 Proof of Proposition 3

Consider a causal model $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ specifying the SCM (2) for the data among observed variables (these parameters can for example be the population parameters), so that:

$$\Sigma_e^\star = (\mathrm{Id} - B)^{-1} (\Omega_e + \Gamma \Psi_e \Gamma^T)^{-1} (\mathrm{Id} - B)^{-T} \text{ for every } e \in [m].$$

We will construct an equivalent SCM with parameters $(\tilde{B}, \tilde{\Gamma}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ where the connectivity matrix $\tilde{B}$ can be arbitrary and compatible with any DAG, and the coefficient matrix $\tilde{\Gamma}$ is any arbitrary $p \times p$ and invertible matrix.

Specifically, Let $\tilde{\mathcal{D}}$ be any DAG and $\tilde{B}$ be any connectivity matrix associated with $\tilde{\mathcal{D}}$. Let $\tilde{\Gamma} \in \mathbb{R}^{p \times p}$ be an arbitrary invertible matrix. For every $e \in [m]$, choose a diagonal positive matrix $\tilde{\Omega}_e \in \mathbb{D}_{++}^p$ such that:

$$(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma \Psi_e \Gamma^T)(\mathrm{Id} - B)^{-T} \succ (\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\mathrm{Id} - \tilde{B})^{-T}.$$

Notice that such a matrix exists since $(\mathrm{Id} - \tilde{B})(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma \Psi_e \Gamma^T)(\mathrm{Id} - B)^{-T}(\mathrm{Id} - \tilde{B})^T$ is a positive definite matrix. Define then for every $e \in [m]$:

$$\tilde{\Psi}_e = \tilde{\Gamma}^{-1}\left[(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma \Psi_e \Gamma^T)(\mathrm{Id} - B)^{-T} - (\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\mathrm{Id} - \tilde{B})^{-T}\right]\tilde{\Gamma}^{-T}.$$

By construction, $\tilde{\Psi}_e \succ 0$ and $(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma \Psi_e \Gamma^T)(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - \tilde{B})^{-1}(\tilde{\Omega}_e + \tilde{\Gamma}\tilde{\Psi}_e\tilde{\Gamma}^T)(\mathrm{Id} - \tilde{B})^{-T}$ for every $e \in [m]$. We have thus shown that:

$$\Sigma_e^\star = (\mathrm{Id} - \tilde{B})^{-1}(\tilde{\Omega}_e + \tilde{\Gamma}\tilde{\Psi}_e\tilde{\Gamma}^T)^{-1}(\mathrm{Id} - \tilde{B})^{-T} \text{ for every } e \in [m].$$

That is, the model specified by the parameters $(\tilde{B}, \tilde{\Gamma}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ specifies an SCM that is compatible with the underlying data distributions.

Figure 4 displays the two equivalent models. Figure 4(a) can be taken for example to be the population model with a single latent variable. Figure 4(b) is a model with three latent variables that is an equally good representation of the data among the observed variables.

D.2 Connection to violation of faithfulness via an illustration

For simplicity, we first consider the scenario without any interventions, e.g. there are no nodes $\mathcal{E}$ in Figure 4. Suppose that the graphical models in Figure 4(a) and Figure 4(b) specify the distribution among the observed variables. Notice that Figure 4(a) implies that $X_1 \perp X_3$, while the same conclusion cannot be made in Figure 4(b). In other words, if the model in Figure 4(b) is the population DAG, the conditional independence relationships among the observed variables in Figure 4(b) are not encoded in the data distribution. Thus, the faithfulness assumption is not satisfied.

Now we consider the scenario where there are interventions $\mathcal{E}$. In our modeling assumption, we assume that the interventions are independent among the observed variables, that is the matrix Ω_e encoding noise variances among the observed variables is diagonal. Then, again, Figure 4(a) concludes that $X_1 \perp X_3$, while Figure 4(b) does not.

D.3 Proof of Corollary 4

Let $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ be the population parameters. Consider the construction in the proof of Proposition 3. Let $\tilde{\mathcal{D}}$ be the empty graph and $\tilde{B} = 0$. Let $\tilde{\Omega}_e = \alpha \mathrm{Id}$, where α is chosen such that:

$$(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma \Psi_e \Gamma^T)(\mathrm{Id} - B)^{-T} \succ \alpha(\mathrm{Id} - \tilde{B})^{-1}(\mathrm{Id} - \tilde{B})^{-T}.$$

Setting $\tilde{\Gamma}$ and $\tilde{\Psi}_e$ as in the proof of Proposition 3, we have that $(\tilde{B}, \tilde{\Gamma}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ is an equivalent model. We have found a model satisfying the constraint in the optimization (16). By the construction of $\tilde{\Omega}_e$, we have that the set of intervention targets encoded by this equivalent model is empty, i.e. $\tilde{\mathcal{I}} = \mathbb{I}(\{\tilde{\Omega}_e\}) = \emptyset$. Furthermore, $\mathcal{R}_\gamma(\tilde{\mathcal{D}}, \tilde{\mathcal{I}}) = 0$, which is the minimal value the regularization function $\mathcal{R}_\gamma(\cdot, \cdot)$ can attain for any value of γ. We thus conclude from Proposition 12 that the constructed model with an empty graph is optimal.

Appendix E. Discussions of Assumption 1

E.1 When is the incoherence parameter $\text{inc}_e^\star$ small?

Recall that Assumption 1 states that the product of an incoherence parameter $\text{inc}_e^\star$ (capturing the denseness of the latent effects) and the degree of the moral graph is $\mathcal{O}(1)$. We provide a simple illustration of a model (2) for which the incoherence parameter $\text{inc}_e^\star$ is small in that it is close to its lower-bound $\sqrt{\frac{h}{p}}$.

Illustration: We consider a model (2) where there is one latent variable (i.e. $h = 1$) and the latent coefficient matrix $\Gamma \in \mathbb{R}^{p \times 1}$ has identical entries so that the effect of the latent variable on the observed variables is *equally spread out*. Let k be the maximum number of parents for any node in the DAG $\mathcal{D}^\star$ among the observed variables. Let $\text{cond}(\Omega_e^\star) = \frac{\max_i [\Omega_e^\star]_{i,i}}{\min_i [\Omega_e^\star]_{i,i}}$ represent the condition number of the noise variance matrix $\Omega_e^\star$. Suppose that the edge weights of the DAG $\mathcal{D}^\star$, i.e. the nonzero entries of $B^\star$ are sufficiently small, i.e. are smaller in magnitude than $1/(k\ \text{cond}(\Omega_e^\star))$. Then, some manipulations yield the following bound for $\text{inc}_e^\star$:

$$\begin{aligned}\text{inc}_e^\star = \text{inc}[\text{col-space}((\text{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star] &= \frac{\max_i |\left[(\text{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star\right]_{i,1}|}{\|(\text{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star\|_2} \\ \leq \frac{\max_i |\left[(\text{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star\right]_{i,1}|}{\sqrt{p} \min_i |\left[(\text{Id} - B^\star)^T \Omega_e^{\star -1} \Gamma^\star\right]_{i,1}|} &\leq \sqrt{\frac{h}{p}} \frac{\text{cond}(\Omega_e^\star)(1 + k\|B^\star\|_\infty)}{1 - k\|B^\star\|_\infty \text{cond}(\Omega_e^\star)}.\end{aligned}$$

Then, supposing that the noise variances on each observed variable are not too different, i.e. $\text{cond}(\Omega_e^\star) = \mathcal{O}(1)$, we have that $\text{inc}_e^\star = \mathcal{O}\left(\sqrt{\frac{h}{p}}\right)$.

We have shown via the above illustration that when the effect of the latent variables is spread out among all the observed variables, the incoherence parameter is small with $\text{inc}_e^\star = \mathcal{O}\left(\sqrt{\frac{h}{p}}\right)$.

E.2 Examples of models (2) that satisfy Assumption 1

Throughout, we consider models in which the low-rank matrix $L_e^\star$ is almost maximally incoherent (dense), that is $\text{inc}[\text{col-space}(L_e^\star)] = \mathcal{O}\left(\sqrt{\frac{h}{p}}\right)$ so the effect of marginalization over the latent variables is diffuse across all the observed variables (see Section E.1). We will suppress the constants involved in Assumption 1 and focus on the trade-off between $\text{inc}[\text{col-space}(L_e^\star)]$ and maximal degree of the moral graph of $\mathcal{D}^\star$ represented by the quantity

degree[moral($\mathcal{D}^\star$)]. So we study models (2) in which:

$$\text{inc[col-space}(L_e^\star)]^2\text{degree[moral}(\mathcal{D}^\star)] = \mathcal{O}\left(\frac{h}{p}\text{degree[moral}(\mathcal{D}^\star)]\right) = \mathcal{O}(1). \tag{17}$$

As we describe next, there are nontrivial classes of models in which the above condition holds.

Bounded degree: the first class of models that we consider are the moral graph of the DAG $\mathcal{D}^\star$ among the observed variables has constant degree:

$$\text{degree[moral}(\mathcal{D}^\star)] = \mathcal{O}(1) \;\; ; \;\; h = \mathcal{O}(p).$$

Here, consistent estimation of the underlying equivalence class of DAGs is possible even when the number of latent variables is on the same order as the number of observed variables.

Polynomial degree: The next class of models we consider are those in which the degree of the moral graph of $\mathcal{D}^\star$ grows polynomially with p:

$$\text{degree[moral}(\mathcal{D}^\star)] = \mathcal{O}(p^q) \;\; ; \;\; h = \mathcal{O}\left(\frac{p}{p^q}\right),$$

where $q \in (0, 1)$. Here, according to the theorems in Section 3.2, consistent estimation of the underlying equivalence class of DAGs is possible with the number of latent variables growing with p.

E.3 Comparison to assumptions in Chandrasekaran et al. (2011, 2012); Frot et al. (2019)

Building on the methodology and results of Chandrasekaran et al. (2011, 2012), Frot et al. (2019) impose a similar but more stringent condition (than Assumption 1) on the denseness of the latent effects for equivalence class recovery in observational settings. In particular, Frot et al. (2019) provide guarantees for models in which:

$$\text{inc[col-space}(L^\star)]\text{degree[moral}(\mathcal{D}^\star)] = \mathcal{O}\left(\sqrt{\frac{h}{p}}\text{degree[moral}(\mathcal{D}^\star)]\right) = \mathcal{O}(1), \tag{18}$$

where $L^\star$ is the matrix $L_e^\star$ for an observational environment e. Comparing (17) with (18), we see that our guarantees are applicable to a broader class of models [5]. Furthermore, as described in the main text, the method in Frot et al. (2019) is only appropriate for observational settings, whereas our method also exploits interventional data for additional identifiability.

5. The method in Frot et al. (2019) uses a nuclear norm penalty to induce low-rank structure; due to the facial structure of the nuclear norm ball, the incoherence condition that they impose is more stringent than the one in our paper.

Appendix F. Proof of Proposition 5

The proof relies on a few lemmas. We first state a known result (see Chandrasekaran et al. (2011)) on the spectral norm of a sparse matrix. For completeness, we include a proof.

Lemma 14 (spectral norm of a low-degree matrix) *Let $N \in \mathbb{S}^p$ and* degree(N) *be the maximum number of non-zeros in any column of N. Then, $\|N\|_2 \leq \text{degree}(N)\|N\|_\infty$.*

Proof Let $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$. Notice that for any standard basis element e_i:

$$\begin{aligned} |e_i^T N v|_\infty &\leq \|e_i^T N\|_2 \sqrt{\sum_{j:N_{i,j}\neq 0} v_j^2} \\ &\leq \sqrt{\text{degree}(N)}\|N\|_\infty \sqrt{\sum_{j:N_{i,j}\neq 0} v_j^2}. \end{aligned}$$

Combining this with the inequality $\sum_{i=1}^p \sum_{j:N_{i,j}\neq 0} v_j^2 \leq \text{degree}(N)$, $\|Nv\|_2^2$ is bounded as follows:

$$\begin{aligned} \|Nv\|_2^2 &\leq \sum_{i=1}^p \text{degree}(N)\|N\|_\infty^2 \sum_{j:N_{i,j}\neq 0} v_j^2 \\ &= \text{degree}(N)\|N\|_\infty^2 \left[\sum_{i=1}^p \sum_{j:N_{i,j}\neq 0} v_j^2\right] \\ &\leq \text{degree}(N)^2\|N\|_\infty^2. \end{aligned}$$

Since v was arbitrary, we arrive at the desired result. ■

Lemma 15 (sparse/low rank incoherence) *Let $K \in \mathbb{S}^p$ and $L \in \mathbb{S}^p$. If* degree(K)inc[*col-space(L)]$^2 < 1$, *then, $K = L$ if and only if $K = L = 0$.*

Proof The proof follows very similarly to the proof of Lemma 2 in Chandrasekaran et al. (2011). Let Ω be the following subspace induced by K:

$$\Omega = \{M \in \mathbb{S}^p : M_{ij} = 0 \text{ if } K_{ij} = 0\}.$$

Let T be the following subspace induced by L:

$$T = \{\mathcal{P}_{\text{col-space}(L)} M \mathcal{P}_{\text{col-space}(L)} \text{ for } M \text{ symmetric and non-singular}\}.$$

It suffices to show that under the condition stated above $\Omega \cap T = \{0\}$. Note that:

$$\max_{\|N\|_2=1, N\in T} \|\mathcal{P}_\Omega(N)\|_2 < 1 \Rightarrow \Omega \cap T = \{0\},$$

since if $N \in \Omega \cap T$ with $\|N\|_2 = 1$, $\|\mathcal{P}_\Omega(N)\|_2 = \|N\|_2 = 1$, leading to a contradiction. Furthermore, by Lemma 14, we have that:

$$\begin{aligned}
\max_{\|N\|_2=1, N\in T} \|\mathcal{P}_\Omega(N)\|_2 &\leq \text{degree}(K) \max_{\|N\|_2=1, N\in T} \|\mathcal{P}_\Omega(N)\|_\infty \\
&\leq \text{degree}(K) \max_{\|N\|_2=1, N\in T} \|N\|_\infty \\
&= \text{degree}(K) \max_{\|N\|_2=1, N\in T} \max_{i,j} |e_i^T \mathcal{P}_{\text{col-space}(L)} N \mathcal{P}_{\text{col-space}(L)} e_j| \\
&\leq \text{degree}(K) \max_i \|\mathcal{P}_{\text{col-space}(L)}(e_i)\|_2^2 \\
&= \text{degree}(K) \text{inc}[\text{col-space}(L)]^2.
\end{aligned}$$

■

Lemma 16 (Sum of incoherent matrices) *Let $L_1, L_2 \in \mathbb{S}$ with column spaces $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. Then:*

$$\text{inc}[\textit{col-space}(L_1 + L_2)] \leq 2\min\{\text{inc}[\mathcal{C}_1], \text{inc}[\mathcal{C}_2]\} + \max\{\text{inc}[\mathcal{C}_1], \text{inc}[\mathcal{C}_2]\}.$$

Proof Without loss of generality, let $\text{inc}[\mathcal{C}_1] \leq \text{inc}[\mathcal{C}_2]$. First, notice that for any $v \in \text{col-space}(L_1 + L_2)$, we have that $v = \text{span}(u_1, u_2)$ where $u_1 \in \mathcal{C}_1$ and $u_2 \in \mathcal{C}_2$. Notice that:

$$\begin{aligned}
\text{inc}[\text{col-space}(L_1 + L_2)] &= \max_i \max_{v \in \text{col-space}(L_1+L_2)} |v^T e_i|/\|v\|_2 \\
&= \max_i \max_{\substack{u_1\in\mathcal{C}_1, u_2\in\mathcal{C}_2, \|u_1\|_2=\|u_2\|_2=1 \\ v=c_1u_1+c_2u_2}} |v^T e_i|/\|v\|_2 \\
&= \max_i \max_{\substack{u_1\in\mathcal{C}_1, u_2\in\mathcal{C}_2, \|u_1\|_2=\|u_2\|_2=1 \\ u_3=u_2-(u_2^T u_1)u_1 \\ v=c_1u_1+c_2u_3}} |v^T e_i|/\|v\|_2 \\
&\leq \max_i \max_{\substack{u_1\in\mathcal{C}_1, u_2\in\mathcal{C}_2, \|u_1\|_2=\|u_2\|_2=1 \\ u_3=u_2-(u_2^T u_1)u_1 \\ v=c_1u_1+c_2u_3}} \frac{|c_1|}{\sqrt{c_1^2+c_2^2}} |u_1^T e_i| + \frac{|c_2|}{\sqrt{c_1^2+c_2^2}} |u_3^T e_i| \\
&\leq \max_i \left[\max_{u_1\in\mathcal{C}_1, \|u_1\|_2=1} 2|u_1^T e_i| + \max_{u_2\in\mathcal{C}_2, \|u_2\|_2=1} |u_2^T e_i| \right] \\
&\leq 2\min\{\text{inc}[\mathcal{C}_1], \text{inc}[\mathcal{C}_2]\} + \max\{\text{inc}[\mathcal{C}_1], \text{inc}[\mathcal{C}_2]\}.
\end{aligned}$$

■

Proof [Proof of Proposition 5] Consider the optimization problem

$$\begin{aligned} \underset{\mathcal{D},\bar{h},B,\Gamma,\mathcal{I},\{\Omega_e,\Psi_e\}_{e=1}^m}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D},\mathcal{I}) \\ \text{subject-to}: \quad & B \sim \mathcal{D}, \mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m), \text{ and} \\ & \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \text{ for every } e \in [m] \\ & \mathrm{inc}(\text{col-space}((\mathrm{Id}-B)^T\Omega_e^{-1}\Gamma)) \leq 2\mathrm{inc}_e^\star \text{ for every } e \in [m], \end{aligned} \tag{19}$$

Where compared to (16), we have added the incoherence constraint inc(col-space($(\mathrm{Id} - B)^T\Omega_e^{-1}\Gamma)$) $\leq 2\mathrm{inc}_e^\star$. Following exactly similar logic as proof of Proposition 12, one can show that with probability tending to one, the optimal solutions of (7) with the incoherence constraint equal to the optimal solution of (19). Thus, we will analyze the estimates produced by (19).

Let

$$(\mathcal{D}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$$

be any optimal set of parameters in (19). Since the parameters $(\mathcal{D}^\star, B^\star, \Gamma^\star, \mathcal{I}^\star, \{\Omega_e^\star, \Psi_e^\star\}_{e=1}^m)$ are feasible in (19), we have that:

$$p \text{ degree}[\text{moral}(\mathcal{D})] + \|\mathcal{D}\|_{\ell_0} + \gamma|\mathcal{I}| \leq pd^\star + \|\mathcal{D}^\star\|_{\ell_0} + \gamma|\mathcal{I}^\star|.$$

Since $\|\mathcal{D}^\star\|_{\ell_0} \leq pd^\star$, $|\mathcal{I}^\star| \leq p$, and $\gamma \in (0, d^\star]$, we have the inequality degree[moral($\mathcal{D}$)] $\leq 3d^\star$. Again, by the feasibility of the parameters $(\mathcal{D}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$ in (19), we have that

$$(\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)(\mathrm{Id}-B)^{-T} = (\mathrm{Id}-B^\star)^{-1}(\Omega_e^\star+\Gamma^\star\Psi_e^\star\Gamma^{\star T})(\mathrm{Id}-B^\star)^{-T} \text{ for all } e \in [m].$$

Equivalently, taking the inverse of both sides in the previous equation, and using the Woodbury Inversion Lemma, we have for $e = 1, 2, \ldots, m$

$$\begin{aligned} &(\mathrm{Id}-B)^T\Omega_e^{-1}(\mathrm{Id}-B) - (\mathrm{Id}-B)^T\Omega_e^{-1}\Gamma(\Psi_e^{-1} + \Gamma^T\Omega_e^{-1}\Gamma)^{-1}\Gamma^T\Omega_e^{-1}(\mathrm{Id}-B) \\ &= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}(\mathrm{Id}-B^\star) - (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}\Gamma^\star(\Psi_{e\,e}^{\star-1} + \Gamma^{\star T}\Omega_e^{\star-1}\Gamma^\star)^{-1}\Gamma^{\star T}\Omega_e^{\star-1}(\mathrm{Id}-B^\star). \end{aligned} \tag{20}$$

Define for every $e = 1, 2, \ldots, m$ the following quantities:

$$\begin{aligned} K_e &:= (\mathrm{Id}-B)^T\Omega_e^{-1}(\mathrm{Id}-B), \\ K_e^\star &= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}(\mathrm{Id}-B^\star), \\ L_e &:= (\mathrm{Id}-B)^T\Omega_e^{-1}\Gamma(\Psi_e^{-1} + \Gamma^T\hat{\Omega}_e^{-1}\Gamma)^{-1}\Gamma^T\Omega_e^{-1}(\mathrm{Id}-B), \\ L_e^\star &:= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}\Gamma^\star(\Psi_e^{\star-1} + \Gamma^{\star T}\Omega_e^{\star-1}\Gamma^\star)^{-1}\Gamma^{\star T}\Omega_e^{\star-1}(\mathrm{Id}-B^\star). \end{aligned} \tag{21}$$

Notice that degree($K_e^\star$) = $d^\star$. Since degree[moral($\mathcal{D}$)] $\leq 3d^\star$, we have that degree(K_e) $\leq 3d^\star$. Furthermore, by the constraint on the optimization (19), inc[col-space(L_e] $\leq 2\mathrm{inc}_e^\star$.

Notice that (20) can be equivalently written for every $e = 1, 2, \ldots, m$:

$$K_e - K_e^\star = L_e - L_e^\star. \tag{22}$$

Since degree$(A+B) \leq$ degree$(A)+$ degree(B) for matrices $A, B \in \mathbb{S}^p$, degree$(K_e - K_e^\star) \leq 4d^\star$ for every e. Furthermore, by Lemma 16, we have that inc[col-space$(L_e - L_e^\star)$] $\leq$ 4inc$_e^\star$ for every $e = 1, 2, \ldots, m$. Thus, by Assumption 1, we have that: degree$(K_e - K_e^\star)$inc[col-space$(L_e - L_e^\star)$] < 1 for all $e = 1, 2, \ldots, m$. Hence, by Lemma 15, $K_e - K_e^\star = 0$ or equivalently $(\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T} = (\mathrm{Id} - B^\star)^{-1}\Omega_e^\star(\mathrm{Id} - B^\star)^{-T}$ for every $e = 1, 2, \ldots, m$.

Let $\mathcal{D}_{\text{all.opt}}, \mathcal{I}_{\text{all.opt}}$ be the collection of optimal DAGs and intervention targets in the optimization (19). Letting $\Sigma^\star_{X^e|H^e}$ be the covariance of $X^e|H^e$, the analysis above allows us to conclude that:

$$\begin{aligned}(\mathcal{D}_{\text{all.opt}}, \mathcal{I}_{\text{all.opt}}) = \underset{\mathcal{D},\mathcal{I}}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \\ \text{subject-to} \quad & \text{there exists } B \sim \mathcal{D} \ : \ \mathbb{I}(\{\Omega_e\}_{e=1}^m) = \mathcal{I} \\ & \Sigma^\star_{X^e|H^e} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T} \text{ for all } e = 1, 2, \ldots, m.\end{aligned} \tag{23}$$

We first show that for any feasible $\mathcal{I}$ in (23), $|\mathcal{I}| = |\mathcal{I}^\star|$. For simplicity, let $e = 1$ be the observational environment. Then, the relations $\Sigma^\star_{X^e|H^e} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)$ for all $e = 1, 2, \ldots, m$ imply for every $e = 2, 3, \ldots, m$

$$(\mathrm{Id} - B^\star)^{-1}\left[\Omega_e^\star - \Omega_1^\star\right](\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}\left[\Omega_e - \Omega_1\right](\mathrm{Id} - B)^{-T}.$$

Since $\Omega_e^\star - \Omega_1^\star \succeq 0$ by Assumption 3, the relation above lets us conclude that $\Omega_e - \Omega_1 \succeq 0$. Hence, $|\mathcal{I}^\star| = \text{rank}\left(\sum_{e=2}^m \Omega_e^\star - \Omega_1^\star\right)$ and $|\mathcal{I}| = \text{rank}\left(\sum_{e=2}^m \Omega_e - \Omega_1\right)$. Finally, again by the relation $\Sigma^\star_{X^e|H^e} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T}$ for all $e = 1, 2, \ldots, m$, we have that:

$$(\mathrm{Id} - B^\star)^{-1}\left[\sum_{e=2}^m \Omega_e^\star - \Omega_1^\star\right](\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}\left[\sum_{e=2}^m \Omega_e - \Omega_1\right](\mathrm{Id} - B)^{-T},$$

which lets us conclude that $\text{rank}\left(\sum_{e=2}^m \Omega_e^\star - \Omega_1^\star\right) = \text{rank}\left(\sum_{e=2}^m \Omega_e - \Omega_1\right)$ and consequently $|\mathcal{I}| = |\mathcal{I}^\star|$.

Next, we show that any optimal DAG in (23) must be in the Markov equivalence class of $\mathcal{D}^\star$. Consider the distribution $X^e|H^e$ which is specified by the covariance $\Sigma^\star_{X^e|H^e}$. For any DAG $\mathcal{D}$ compatible[6] with $X^e|H^e$, the following are satisfied:

$$\begin{aligned}\{(i,j,S) : X_i^e \perp\!\!\!\perp X_j^e | X_S^e, H^e \text{ for some set } S\} &\subseteq \{(i,j,S) : X_S \text{ d-separates } X_i \text{ and } X_j \text{ in } \mathcal{D}\}, \\ \{(i,j) : X_i^e \perp\!\!\!\perp X_j^e | X_{\setminus\{i,j\}}^e, H^e\} &\subseteq \{(i,j) : X_i^e \text{ and } X_j^e \text{ are not connected in moral}(\mathcal{D})\},\end{aligned} \tag{24}$$

where set equality in the relations above hold if $X^e|H^e$ is faithful with respect to the DAG $\mathcal{D}$. By Assumption 2 (i.e. faithfulness of $X^e|H^e$ for every environment with respect to the DAG $\mathcal{D}^\star$), and relation (24), we have for any DAG $\mathcal{D}$ consistent with $X^e|H^e$:

$$\|\mathcal{D}^\star\|_{\ell_0} \leq \|\mathcal{D}\|_{\ell_0} \quad ; \quad \text{moral}(\mathcal{D}^\star) \subseteq \text{moral}(\mathcal{D}), \tag{25}$$

6. By compatible, we mean that there exists B compatible with $\mathcal{D}$ and $\Omega \in \mathbb{D}_{++}^p$ such that $\Sigma^\star_{X^e|H^e} = (\mathrm{Id} - B)^{-1}\Omega(\mathrm{Id} - B)^{-T}$.

where equality holds if and only if $\mathcal{D} \in \text{MEC}(\mathcal{D}^\star)$. Combining this fact with $|\mathcal{I}| = |\mathcal{I}^\star|$ for any feasible $\mathcal{I}$ in (23), we conclude that $\mathcal{D}_{\text{all.opt}} \subseteq \text{MEC}(\mathcal{D}^\star)$.

■

Appendix G. Uninformative interventions and proof of Theorem 6

G.1 Worst-case interventions

In Section F, we proved that in the setting of Proposition 5, the optimal DAGs $\mathcal{D}_{\text{all.opt}}$ in the optimization (19) satisfy $\mathcal{D}_{\text{all.opt}} \subseteq \text{MEC}(\mathcal{D}^\star)$. We next show that there are worst-case intervention configurations such that $\mathcal{D}_{\text{all.opt}} = \text{MEC}(\mathcal{D}^\star)$. As an example, suppose for every $e, e' \in 1, 2, \ldots, m$, there exists $\alpha_{e,e'} > 0$ such that

$$\Omega_e^\star = \alpha_{e,e'} \Omega_{e'}^\star. \tag{26}$$

By our construction (26), $\Sigma_{X^e|H^e} = \alpha_{e,e'} \Sigma_{X^{e'}|H^{e'}}$. Recall that the optimal DAGs satisfy the relation (23). However, since $\Sigma_{X^e|H^e} = \alpha_{e,e'} \Sigma_{X^{e'}|H^{e'}}$, it is straightforward to see that when (26) is satisfied, there is no additional information gained over just data in a single environment, i.e. (23) is simplified to:

$$\begin{aligned} \mathcal{D}_{\text{all.opt}} = \underset{D}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D}, \{1, 2, \ldots, p\}) \\ \text{subject-to} \quad & \text{there exists } B \sim \mathcal{D}, \Omega \in \mathbb{D}_{++}^p \text{ such that} \\ & \Sigma_{X^1|H^1}^\star = (\text{Id} - B)^{-1} \Omega (\text{Id} - B)^{-T}. \end{aligned} \tag{27}$$

Since $X^e|H^e$ is faithful with respect to $\mathcal{D}^\star$, following relation (25), we conclude that $\mathcal{D}_{\text{all.opt}} = \text{MEC}(\mathcal{D}^\star)$.

G.2 Proof of Theorem 6

The proof of this theorem relies on two lemmas.

Lemma 17 *Let $B, \tilde{B} \in \mathbb{R}^{p \times p}$ be two matrices that can be made to be lower-triangular with zeros on the diagonal after row and column permutations (or equivalently, the matrices correspond to two DAGs). Suppose that there exists $\tilde{\Omega}, \Omega \in \mathbb{D}_{++}^p$ such that $(\text{Id} - B)^{-1}\Omega(\text{Id} - B)^{-T} = (\text{Id} - \tilde{B})^{-1}\tilde{\Omega}(\text{Id} - \tilde{B})^{-T}$. Then, if for some $\alpha \neq 0$, $[(\text{Id} - \tilde{B})^{-1}(\text{Id} - B)]_{:,i} = \alpha \mathbf{e}_j$, then $[\text{Id} - B]_{i,:} \propto [\text{Id} - \tilde{B}]_{j,:}$.*

Proof [Proof of Lemma 17] By the condition of the lemma, we have that: $\alpha \mathbf{e}_j^T (\text{Id} - \tilde{B})^{-T} (\text{Id} - B)^T = \mathbf{e}_i^T$ and that $(\text{Id} - \tilde{B}) = \Omega(\text{Id} - \tilde{B})^{-T}(\text{Id} - B)^T \Omega^{-1}(\text{Id} - B)$. Combining these two, it follows that for some constant $\beta \neq 0$, $\mathbf{e}_j^T(\text{Id} - \tilde{B}) = \alpha\beta \mathbf{e}_i^T (\text{Id} - B)$. ■

Lemma 18 (Equivalent characterization of $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$)) *Under Assumptions 4 and 5, the following statements are equivalent for $\mathcal{D} \in \text{MEC}(\mathcal{D}^\star)$:*

p1) *There exists a connectivity matrix B compatible with $\mathcal{D}$ and $\{\Omega_e\}_{e=1}^m \subseteq \mathbb{D}_p^{++}$ such that $(\mathrm{Id} - B^\star)^{-1}\Omega_e^\star(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T}$ for all $e = 1, 2, \ldots, m$*

p2) $\mathcal{D} \in \mathcal{I}^\star$-MEC($\mathcal{D}^\star$)

Proof [Proof of Lemma 18] The direction $p2 \Rightarrow p1$ follows from Lemma 10. We next prove $p1 \Rightarrow p2$. In particular, we must show that if for a $\mathcal{D} \in \mathrm{MEC}(\mathcal{D}^\star)$ with a compatible connectivity matrix B satisfying $(\mathrm{Id} - B^\star)^{-1}\Omega_e^\star(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T}$ for all $e = 1, 2, \ldots, m$, then the parents of nodes indexed by $\mathcal{I}^\star$ in $\mathcal{D}$ are fixed to be the same as parents of $\mathcal{D}^\star$. More concretely, we will show that:

$$B_{i,:} = B^\star_{i,:} \qquad \text{for all } i \in \mathcal{I}^\star.$$

We define the following parameters:

$$M := (\mathrm{Id} - B)(\mathrm{Id} - B^\star)^{-1} \quad ; \quad N_e := (\mathrm{Id} - B)(\mathrm{Id} - B^\star)^{-1}\Omega_e^\star \ , \ e = 1, 2, \ldots, m.$$

Notice that the condition of $p1$ implies

$$M_{k,:} \perp [N_e]_{l,:} \text{ for } k \neq l. \tag{28}$$

Since the rows of the matrix M are linearly independent, the relation (28) implies that for every $k = 1, 2, \ldots, p$, the vectors $\{[N_e]_{k,:}\}_{e=1}^m$ live in a one-dimensional null space of the matrix formed by concatenating the vectors $\{M_{l,:}\}_{l \neq k}$, i.e.

$$\dim\left(\mathrm{span}\left(\{[N_e]_{k,:}\}_{e=1}^m\right)\right) = 1. \tag{29}$$

We focus on a particular $i \in \mathcal{I}^\star$. Take an environment e satisfying Assumption 4 for $i \in \mathcal{I}^\star$. Then, (29) implies that for every k, there exists a constant $c_k \neq 0$ (nonzero since the matrix $N_{e'}$ is non-singular for every e') such that $[N_1]_{k,:} = c_k[N_e]_{k,:}$ or equivalently $M_{k,:}\Omega_1^\star = c_k M_{k,:}\Omega_e^\star$. Thus, combining this fact with Assumption 3, we conclude that:

$$\begin{aligned} &\text{for every } k = 1, 2, \ldots, p, \\ &[M]_{k,i} = 0 \quad \text{OR} \quad [M]_{k,i} \text{ has one nonzero and } [M]_{k,\{1,2,\ldots,p\}\setminus i} = 0. \end{aligned} \tag{30}$$

Combining (30) with the fact that the rows of M are linearly independent, we conclude that:

$$_{:,i} = \alpha \mathbf{e}_j \text{ for some standard basis element } \mathbf{e}_j \text{ and } \alpha \neq 0. \tag{31}$$

Appealing to Lemma 17, we conclude that (31) can be equivalently written as:

$$_{i,:} = \alpha \mathbf{e}_j^T \text{ for some standard basis element } \mathbf{e}_j \text{ and } \alpha \neq 0. \tag{32}$$

We consider two scenarios. Scenario 1 is when $j \neq i$ in (31) and Scenario 2 is when $j = i$. Our proof strategy is to show that under Assumption 5, Scenario 1 *cannot* occur, implying that only Scenario 2 is possible. For Scenario 2, we conclude that $B_{i,:} = B^\star_{i,:}$.

Scenario 1: $j \neq i$ in (32) Since B is a connectivity matrix associated with a DAG $\mathcal{D} \in \mathrm{MEC}(\mathcal{D}^\star)$, by Assumption 4, this case is not allowed.

Scenario 2: $j = i$ in (31) Here, we have that $[M]_{:,i} = \alpha \mathbf{e}_i$ for nonzero α. Hence, $M_{i,i} \neq 0$. By (30), we then conclude that $M_{i,:} = \alpha \mathbf{e}_i^T$, or equivalently $[\mathrm{Id} - B]_{i,:} = \alpha[\mathrm{Id} - B^\star]_{i,:}$. Since $\mathrm{Id} - B$ and $\mathrm{Id} - B^\star$ have ones on the diagonal, $\alpha = 1$ and thus $B_{i,:} = B^\star_{i,:}$.
We repeat the above arguments for every $i \in \mathcal{I}^\star$ to conclude that $B_{i,:} = B^\star_{i,:}$. ■

Proof [Proof of Theorem 6] Let $\mathcal{D}_{\text{all.opt}}, B_{\text{all.opt}}, \mathcal{I}_{\text{all.opt}}$ be the collection of optimal DAGs and intervention targets in the optimization (19). From (23) and Proposition 5, we have that any $\mathcal{D} \in \mathcal{D}_{\text{all.opt}} \subseteq \text{MEC}(\mathcal{D}^\star)$ with an associated connectivity matrix B and noise variances Ω_e satisfies:

$$(\mathrm{Id} - B)(\mathrm{Id} - B^\star)^{-1}\Omega_e^\star(\mathrm{Id} - B^\star)^1(\mathrm{Id} - B)^T = \Omega_e \quad e = 1, 2, \ldots, m,$$

where $\{\Omega_e\}_{e=1}^m$ are diagonal and $\mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m)$. By Lemma 18, we have that $\mathcal{D} \in \mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. Thus, the associated connectivity matrix B satisfies for every $i \in \mathcal{I}^\star$: support$(B_{i,:})$ = support$(B^\star_{i,:})$. Then, appealing to Lemma 9, we conclude that $[\Omega_e^\star]_{i,i} = [\Omega_e]_{i,i}$ for every $i \in \mathcal{I}^\star$ and $e \in [m]$. Further, since the matrix $(\mathrm{Id} - B)(\mathrm{Id} - B^\star)^{-1}$ is invertible, we have that $\mathrm{rank}(\Omega_e^\star - \Omega_f^\star) = \mathrm{rank}(\Omega_e - \Omega_f)$. Combining the previous two facts, we conclude that $\mathcal{I} = \mathcal{I}^\star$ so that $\mathcal{I}_{\text{all.opt}} = \{\mathcal{I}^\star\}$. Appealing to Lemma 18, we conclude that (23) can be equivalently expressed as:

$$\mathcal{D}_{\text{all.opt}} = \arg \min_{\text{DAG } \mathcal{D}} \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}^\star) \;\text{ subject-to }\; \mathcal{D} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star).$$

Since $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star) \subseteq$ MEC$(\mathcal{D}^\star)$, the regularizer $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}^\star) = \mathcal{R}_\gamma(\mathcal{D}^\star, \mathcal{I}^\star)$ for all $\mathcal{D} \in \mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. We thus conclude that $\mathcal{D}_{\text{all.opt}} = \mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. Finally, since the population parameters are feasible in the optimization (19) and that $\mathcal{D}^\star$ and $\mathcal{I}^\star$ achieve the optimum loss $\mathcal{R}_\gamma(\cdot, \cdot)$, we conclude that $B^\star \in B_{\text{all.opt}}$. ■

G.3 Proof of Corollary 7

Consider the optimization problem

$$\begin{aligned} \underset{\mathcal{D}, \bar{h}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \\ \text{subject-to}: \quad & B \sim \mathcal{D}, \mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m), \text{ and} \\ & \Sigma_e^\star = (\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)^{-T} \text{ for every } e \in [m] \\ & \bar{h} \leq h_{\max}, \end{aligned} \tag{33}$$

Where compared to (16), we have added a constraint $\bar{h} \leq h_{\max}$ on the number of latent variables included in the model. Following exactly similar logic as proof of Proposition 12, one can show that with probability tending to one, in the infinite data regime, the optimal solutions of (7) with the constraint on the number of latent variables equal to the optimal solution of (19). Thus, we will analyze the estimates produced by (33).

We follow a very similar proof technique as proof of Theorem 6. Specifically, let $(\mathcal{D}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$ be any optimal set of parameters in (33). Then, we can arrive at the equality

(22) where the matrices $K_e, K_e^\star, L_e, L_e^\star$ are defined in (21). Note that from Lemma 15 that for all $e = 1, 2, \ldots, m$:

$$\text{degree}(K_e - K_e^\star)\text{inc}[\text{col-space}(L_e - L_e^\star)]^2 < 1 \Rightarrow K_e = K_e^\star. \tag{34}$$

In this analysis, we show that under the conditions described in Corollary 7, $\text{degree}(K_e - K_e^\star)\text{inc}[\text{col-space}(L_e - L_e^\star)]^2 < 1$, allowing us to conclude that $K_e = K_e^\star$. Following then an exact line of reasoning as the last paragraph of proof of Theorem 6, we conclude that $\mathcal{D}_{\text{all.opt}} = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$, $B^\star \in B_{\text{all.opt}}$ and $\mathcal{I}_{\text{all.opt}} = \{\mathcal{I}^\star\}$, where $\mathcal{D}_{\text{all.opt}}, B_{\text{all.opt}}, \mathcal{I}_{\text{all.opt}}$ is the collection of optimal DAGs, connectivity matrices and intervention sets, respectively, in the optimization (33).

Notice that $\text{degree}(K_e - K_e^\star) \leq \text{degree}[\text{moral}(\mathcal{D})] + d^\star$. Further, by Lemma 16, $\text{inc}[\text{col-space}(L_e - L_e^\star)] \leq 2(\text{inc}[\text{col-space}(L_e)] + \text{inc}_e^\star)$. Thus, it suffices to show for all $e \in [m]$ that:

$$4(\text{degree}[\text{moral}(\mathcal{D})] + d^\star)(\text{inc}[\text{col-space}(L_e)]^2 + (\text{inc}_e^\star)^2) < 1. \tag{35}$$

Since the population parameters satisfy the constraint of the optimization problem (33), we have that $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \leq \mathcal{R}_\gamma(\mathcal{D}^\star, \mathcal{I}^\star)$. Thus, we can conclude with the choice of γ that $\text{degree}[\text{moral}(\mathcal{D})] \leq 3d^\star$. Therefore, the following conditions are satisfied for every $e = 1, 2, \ldots, m$ due to Assumption 1, the conditions of Corollary 7, and the bound $d^\star \leq \nu^\star$:

$$\begin{aligned}
&\text{degree}[\text{moral}(\mathcal{D})]\text{inc}[\text{col-space}(L_e)]^2 \leq 3\nu^\star\text{inc}[\text{col-space}(L_e)]^2 < 1/16,\\
&\text{degree}[\text{moral}(\mathcal{D})]\text{inc}[\text{col-space}(L_e^\star)]^2 \leq 3\nu^\star\text{inc}[\text{col-space}(L_e^\star)] < 1/16,\\
&d^\star\text{inc}[\text{col-space}(L_e)]^2 \leq \nu^\star\text{inc}[\text{col-space}(L_e)]^2 < 1/16,\\
&d^\star(\text{inc}_e^\star)^2 < 1/16.
\end{aligned}$$

Combining these relations, we arrive at the inequality in (35).

Appendix H. Proof of Theorem 8 with known latent interventions

Consider the optimization problem

$$\begin{aligned}
\underset{\mathcal{D}, \bar{h}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I})\\
\text{subject-to}: \quad & B \sim \mathcal{D}, \mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m), \text{ and}\\
& \Sigma_e^\star = (\text{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\text{Id} - B)^{-T} \text{ for every } e \in [m]\\
& \Psi_e = \psi_e^\star \text{Id} \text{ for every } e \in [m],
\end{aligned} \tag{36}$$

Where compared to (16), we have added the known latent interventions constraint $\Psi_e = \psi_e^\star\text{Id}$. Following exactly similar logic as proof of Proposition 12, one can show that with probability tending to one, the optimal solutions of (7) with the known latent interventions constraint equal to the optimal solution of (36). Thus, we will analyze the estimates produced by (36). We will let $\mathcal{D}_{\text{all.opt}}, B_{\text{all.opt}}, \mathcal{I}_{\text{all.opt}}$ be optimal DAGs, connectivity matrices, and intervention sets according to the optimization (36).

Consider any optimal set of parameters $(\mathcal{D}, B, \Gamma, \{\Omega_e, \Psi_e)$ of (36). Then for every $e = 1, 2, \ldots, m$

$$(\mathrm{Id} - B^\star)^{-1}(\Omega_e^\star + \psi_e^\star \Gamma^\star \Gamma^{\star T})(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}(\Omega_e + \psi_e^\star \Gamma \Gamma^T)(\mathrm{Id} - B)^{-T}. \quad (37)$$

The relations (37) imply:

$$\begin{aligned}(\mathrm{Id} - B^\star)^{-1}\left(\Omega_2^\star - \psi_2^\star/\psi_1^\star \Omega_1^\star\right)(\mathrm{Id} - B^\star)^{-T} &= (\mathrm{Id} - B)^{-1}\left(\Omega_2 - \psi_2^\star/\psi_1^\star \Omega_1\right)(\mathrm{Id} - B)^{-T}, \\ (\mathrm{Id} - B^\star)^{-1}\left(\Omega_e^\star - \psi_e^\star/\psi_1^\star \Omega_1^\star\right)(\mathrm{Id} - B^\star)^{-T} &= (\mathrm{Id} - B)^{-1}\left(\Omega_e - \psi_e^\star/\psi_1^\star \Omega_1\right)(\mathrm{Id} - B)^{-T}.\end{aligned} \quad (38)$$

Appealing to Assumption 3', the first relation in (38) yields:

$$(\mathrm{Id} - B^\star)^{-1}\Omega_1^\star(\mathrm{Id} - B^\star)^{-T} = \frac{1}{(1 - \psi_2^\star/\psi_1^\star)}(\mathrm{Id} - B)^{-1}\left(\Omega_2 - \psi_2^\star/\psi_1^\star \Omega_1\right)(\mathrm{Id} - B)^{-T}. \quad (39)$$

Relation (39) states that B is an equivalent connectivity with respect to distribution $X^1|H^1$ in the sense $\Sigma^\star_{X^1|H^1} = (\mathrm{Id} - B)^{-1}\Omega(\mathrm{Id} - B)^{-T}$ for some $\Omega \in \mathbb{D}^p_{++}$. Due to Assumption 2 (faithfulness condition), we appeal to relation (25), and conclude that: $\|\mathcal{D}^\star\|_{\ell_0} \leq \|\mathcal{D}\|_{\ell_0}$ and $\mathrm{moral}(\mathcal{D}^\star) \subseteq \mathrm{moral}(\mathcal{D})$, where equality holds if and only if $\mathcal{D} \in \mathrm{MEC}(\mathcal{D}^\star)$. Thus, since $\gamma \leq \frac{1}{|\mathcal{I}^\star|}$, the following holds:

for any DAG $\mathcal{D}$ with a connectivity matrix B satisfying (39) where $\mathcal{D} \notin \mathrm{MEC}(\mathcal{D}^\star)$,
$\Rightarrow$,
$\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) > \mathcal{R}_\gamma(\mathcal{D}^\star, \mathcal{I}^\star)$ for all $\mathcal{I} \subseteq [p]$.

The relation above allows us to conclude that $\mathcal{D} \in \mathrm{MEC}(\mathcal{D}^\star)$.

We will next show that $B_{i,:} = B^\star_{i,:}$ for all $i \in \mathcal{I}^\star$. Consider a particular $i \in \mathcal{I}^\star$. Let e be the environment satisfying Assumption 4'. Let $M := (\mathrm{Id} - B)^{-1}(\mathrm{Id} - B^\star)$ and $N_1 := M(\Omega_2^\star - \psi_2^\star/\psi_1^\star \Omega_1^\star) = (1 - \psi_2^\star/\psi_1^\star)\Omega_1^\star$, $N_2 := M(\Omega_e^\star - \psi_e^\star/\psi_1^\star \Omega_1^\star)$. From Assumption 4', we have that N_1, N_2 are non-singular. Further, from (38):

$$M_{k,:} \perp [N_1]_{l,:} \text{ and } M_{k,:} \perp [N_2]_{l,:} \quad \text{for } k \neq l.$$

Since the rows of the matrix M are linearly independent, we have for every $k = 1, 2, \ldots, p$: $[N_1]_{k,:}$ and $[N_2]_{k,:}$ are linearly independent. Thus, there exists constant $c \neq 0$ such that $c(1 - \psi_2^\star/\psi_1^\star)M_{k,:}\Omega_1^\star = M_{k,:}(\Omega_e^\star - \psi_e^\star/\psi_1^\star \Omega_1^\star)$. Suppose $M_{k,i} \neq 0$. We will argue that $M_{k,j} = 0$ for all $j \neq i$. Specifically, suppose $M_{k,j} \neq 0$ for $j \neq i$. We would have that $[\Omega_e^\star - \psi_e^\star/\psi_1^\star \Omega_1^\star]_{i,i}[\Omega_1^\star]^{-1}_{j,j} = [\Omega_e^\star - \psi_e^\star/\psi_1^\star \Omega_1^\star]_{j,j}[\Omega_1^\star]^{-1}_{j,j}$. However, we arrive at a contradiction by Assumption 4'. We have thus argued that the matrix M has the structure described in (31). Furthermore, going through the scenarios described in the proof of Lemma 18 and appealing to the intervention truthfulness condition in Assumption 5 (using relation (39)), we conclude that $B_{i,:} = B^\star_{i,:}$ for all $i \in \mathcal{I}^\star$. In other words, we have now shown that $\mathcal{D}_{\mathrm{opt}} \subseteq \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$.

We must now show that for all $\mathcal{D} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$, $\mathcal{D} \in \mathcal{D}_{\mathrm{opt}}$. Let $(\tilde{\mathcal{D}}, \tilde{B}, \tilde{\Gamma}, \tilde{\mathcal{I}}, \{\tilde{\Omega}_e, \tilde{\Psi}_e\}_{e=1}^m)$ be any optimal set of parameters of (36) where we have shown that $\tilde{\mathcal{D}} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$. By

Lemma 10, there exists a connectivity matrix $B \sim \mathcal{D}$, and noise variances Ω_e such that for all $e = 1, 2, \ldots, m$

$$\mathrm{Id} - \tilde{B})^{-1}\tilde{\Omega}_e(\mathrm{Id} - \tilde{B})^{-T} = (\mathrm{Id} - B)^{-1}\Omega_e(\mathrm{Id} - B)^{-T}. \tag{40}$$

Furthermore, let $\Gamma = \tilde{\Gamma}$ and $\Psi_e = \psi_e^\star \mathrm{Id}$. We then have that the model $(\mathcal{D}, B, \Gamma, \mathcal{I}, \{\Omega_e, \Psi_e\}_{e=1}^m)$ is feasible in (36). It remains to check that $\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) = \mathcal{R}_\gamma(\tilde{\mathcal{D}}, \tilde{\mathcal{I}})$. Since $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are in the same Markov equivalence class, it suffices to show that $\mathcal{I} = \tilde{\mathcal{I}}$. Since $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are both in the set $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$, we have that $\tilde{B}_{i,:} = B_{i,:}$ for every $i \in \tilde{\mathcal{I}}$. From Lemma 9, we have that: $[\Omega_e]_{i,i} = [\tilde{\Omega}_e]_{i,i}$ for every $i \in \tilde{\mathcal{I}}$. Let $\mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m)$. We have shown that $\tilde{\mathcal{I}} \subseteq \mathcal{I}$. Suppose that there exists $j \in \mathcal{I} \setminus \tilde{\mathcal{I}}$. Then, there must exist e, f such that $[\Omega_e]_{j,j} = [\Omega_f]_{j,j}$. From (40), we have that:

$$\mathrm{Id} - \tilde{B})^{-1}(\tilde{\Omega}_e - \tilde{\Omega}_f)(\mathrm{Id} - \tilde{B})^{-T} = (\mathrm{Id} - B)^{-1}(\Omega_e - \Omega_f)(\mathrm{Id} - B)^{-T}. \tag{41}$$

We have that $[\tilde{\Omega}_e]_{i,i} - [\tilde{\Omega}_f]_{i,i} = 0$ for every $i \notin \tilde{\mathcal{I}}$. We also have that $[\tilde{\Omega}_e]_{i,i} - [\tilde{\Omega}_f]_{i,i} = [\Omega_e]_{i,i} - [\Omega_f]_{i,i}$ for every $i \in \mathcal{I}$. Since $[\Omega_e]_{j,j} - [\Omega_f]_{j,j} \neq 0$, we have that $\mathrm{rank}(\Omega_e - \Omega_f) > \mathrm{rank}(\tilde{\Omega}_e - \tilde{\Omega}_f)$ which is a contradiction given (41). Therefore, we conclude that $\mathcal{I} = \tilde{\mathcal{I}}$.

Appendix I. Three environments are required if the number of latent variables is not constrained

Without imposing a condition on the number of latent variables, two environments can only offer identifiability up the Markov equivalence class, even if all the variables are perturbed. To offer some intuition, we sketch a quick argument below. Consider the setting in Section 3.3 where the number of latent variables is allowed to be arbitrary. For simplicity, we assume there are two environments with no interventions on the latent variables and parameters $(B^\star, \Gamma^\star, \{\Omega_e^\star\}_{e=1}^2)$ where $B^\star$ represents the connectivity matrix, $\Gamma^\star$ encodes the effect of latent variables, and $\Omega_e^\star$ is a diagonal matrix encoding the noise terms on the observed variables for environments $e = 1, 2$. Here, for example, $\Omega_1^\star$ is the noise variance associated with an observational environment and $\Omega_2^\star$ is the noise term associated with an interventional environment with $\Omega_2^\star \succ \Omega_1^\star$ (since there are interventions on all variables). Any compatible causal model, parameterized by $(B, \Gamma, \{\Omega_e\}_{e=1}^2)$ must entail the same covariance model, i.e. for $e = 1, 2$:

$$(\mathrm{Id} - B^\star)^{-1}(\Gamma^\star \Gamma^{\star T} + \Omega_e^\star)^{-1}(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}(\Gamma\Gamma^T + \Omega_e)^{-1}(\mathrm{Id} - B)^{-T}. \tag{42}$$

An equivalent reformulation of (42) is:

$$\begin{aligned}(\mathrm{Id} - B^\star)^{-1}(\Omega_2^\star - \Omega_1^\star)(\mathrm{Id} - B^\star)^{-T} &= (\mathrm{Id} - B)^{-1}(\Omega_2 - \Omega_1)(\mathrm{Id} - B)^{-T}, \\ (\mathrm{Id} - B^\star)^{-1}(\Omega_1^\star + \Gamma^\star \Gamma^{\star T})(\mathrm{Id} - B^\star)^{-T} &= (\mathrm{Id} - B)^{-1}(\Gamma\Gamma^T + \Omega_1)^{-1}(\mathrm{Id} - B)^{-T}.\end{aligned} \tag{43}$$

It is straightforward to check that for any DAG in the Markov equivalence class of the population DAG, there exists a connectivity matrix B and diagonal matrix D such that $(\mathrm{Id} - B^\star)^{-1}(\Omega_2^\star - \Omega_1^\star)(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}D(\mathrm{Id} - B)^{-T}$. Let Ω_1 be any positive definite diagonal matrix such that $(\mathrm{Id} - B)(\mathrm{Id} - B^\star)^{-1}(\Omega_2^\star - \Omega_1^\star)(\mathrm{Id} - B^\star)^{-T}(\mathrm{Id} - B)^T \succ$

Ω_1. Furthermore, let Γ be any matrix square root of $(\mathrm{Id}-B)(\mathrm{Id}-B^\star)^{-1}(\Omega_2^\star-\Omega_1^\star)(\mathrm{Id}-B^\star)^{-T}(\mathrm{Id}-B)^T-\Omega_1$. Finally, let $\Omega_2=\Omega_1+D$. Notice that by construction, the parameters $(B,\Gamma,\{\Omega_e\}_{e=1}^2)$ satisfy the relations in (43). In other words, we have shown that even though all observed variables have received an intervention, we are not able to attain any additional identifiability than the Markov equivalence class. A similar analysis can also be done in the case where there are interventions on the latent variables.

Appendix J. Approximately known latent interventions

In this section, we relax the assumption of knowing the latent interventions to having access to approximate values, where the level of approximation is given by C_ψ. In particular, we are given approximations $\tilde{\psi}_1,\tilde{\psi}_2,\ldots,\tilde{\psi}_m$ where $\max\{\tilde{\psi}_e-\psi_e^\star,|1/\tilde{\psi}_e-1/\psi_e^\star|\}\leq C_\psi$ for every $e=1,2,\ldots,m$. We then apply *UT-LVCE* with the additional constraint that: $\max\{|1/\psi_e-1/\tilde{\psi}_e|,|\psi_e-\tilde{\psi}_e|\}\leq C_\psi$ for every $e=1,2,\ldots,m$. Having only an approximation to the latent interventions comes at the expense of additional assumptions for partial identifiability. Specifically, we assume that the interventions on the observed variables in $\mathcal{I}^\star$ are sufficiently strong as compared to the approximation error in the latent interventions. Furthermore, we assume that the latent effects induce some confounding dependencies among the observed variables, although this assumption is generally far weaker than the incoherence condition in Assumption 1 (see Section 3.2).

We impose the following assumptions where as introduced in the main paper, $d^\star=\text{degree}\,[\text{moral}(\mathcal{D}^\star)]$:

Assumption 1' *latent effects induce some confounding dependencies: for every $i\in\mathcal{I}^\star$ and $|\kappa|\leq\frac{4C_\psi}{\max_{e,e'}\psi_{e'}^\star/\psi_e^\star-1},\kappa\neq 0,|\delta|\leq C_\psi$, there exists $e\in[m]$ such that the following conditions hold: i) $[\Omega_e^\star]_{i,i}>[\Omega_1^\star]_{i,i}$, ii)* $\text{degree}\left((\mathrm{Id}-B^\star)^T(\Omega_e^\star-\Omega_1^\star(\psi_e^\star+\delta)+\kappa\Gamma^\star\Gamma^{\star T})^{-1}(\mathrm{Id}-B^\star)\right)\geq 3d^\star$ *and iii)* $\text{degree}\left((\mathrm{Id}-B^\star)^T(\Omega_1^\star+\kappa\Gamma^\star\Gamma^{\star T})^{-1}(\mathrm{Id}-B^\star)\right)\geq 3d^\star$.

Assumption 4" *heterogeneous interventions on observed and latent variables: i) for every $e\neq e'\in[m],\max\left\{\frac{\psi_e^\star}{\psi_{e'}^\star},\frac{\psi_{e'}^\star}{\psi_e^\star}\right\}>1+\frac{C_\psi(\lambda_{max}(\Omega_1^\star)+\|\Gamma^\star\|_2^2)}{\lambda_{min}(\Omega_1^\star)}$, ii) for every $i,j\in\mathcal{I}^\star,i\neq j$, there exists $e\in[m]$ such that $[\Omega_e^\star]_{i,i}>[\Omega_1^\star]_{i,i}$ and* $\left[(\Omega_e^\star-\psi_e^\star\Omega_1^\star)(\Omega_1^\star)^{-1}\right]_{i,i}\neq\left[(\Omega_e^\star-\psi_e^\star\Omega_1^\star)(\Omega_1^\star)^{-1}\right]_{j,j}$.

Assumption 6 *sufficiently strong interventions on the observed variables in $\mathcal{I}^\star$: for every $i\in\mathcal{I}^\star$, there exists $e\in[m]$ such that $[\Omega_e^\star]_{i,i}>(2\psi_e^\star+1)[\Omega_1^\star]_{i,i}$.*

Assumption 1' ensures that the latent effects induce some confounding dependencies. This condition is far weaker than an incoherence-type assumption such as Assumption 1; for a comprehensive discussion, see Section J.1. Assumption 4" (analogous to Assumption 4) ensures that the interventions on the latent variables and observed variables are informative for additional identifiability. One can show that if the parameters $\Omega_e^\star,\Omega_1^\star$ and $\psi_1^\star,\psi_2^\star,\psi_e^\star$ are drawn from continuous distributions, Assumption 4" is satisfied almost surely. Finally, Assumption 6 requires that the interventions on the observed variables in $\mathcal{I}^\star$ are sufficiently strong.

Theorem 19 (Equivalence class characterization under approximately known latent interventions) *Consider the estimator* (6) *with* $\Psi_e = \psi_e \mathrm{Id}$ *and the constraints* $\max\{|\tilde{\psi}_e - \psi_e|, |1/\tilde{\psi}_e - 1/\psi_e|\} \leq C_\psi$ *where* C_ψ *is chosen conservatively so that* $\max\{|\tilde{\psi}_e - \psi_e^\star|, |1/\tilde{\psi}_e - 1/\psi_e^\star|\} \leq C_\psi$. *Suppose Assumptions 1',2,3',4", and 6 are satisfied. Letting* $\frac{1}{|\mathcal{I}^\star|} > \gamma > 0$, *then* $\hat{\mathcal{D}}_{\text{all.opt}} = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ *with probability tending to one.*

To prove Theorem 19, we consider the optimization problem

$$\begin{aligned} \underset{\mathcal{D},\bar{h},B,\Gamma,\mathcal{I},\{\Omega_e,\psi_e\mathrm{Id}\}_{e=1}^m}{\text{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D},\mathcal{I}) \\ \text{subject-to}: \quad & B \sim \mathcal{D}, \mathcal{I} = \mathbb{I}(\{\Omega_e\}_{e=1}^m), \text{ and} \\ & \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e + \psi_e \Gamma\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \text{ for every } e \in [m] \\ & \max\{|\tilde{\psi}_e - \psi_e|, |1/\tilde{\psi}_e - 1/\psi_e|\} \leq C_\psi \end{aligned} \tag{44}$$

Where compared to (16), we have added the constraint $\Psi_e = \psi_e \mathrm{Id}$ (the latent variables are iid), and the constraint $\max\{|\tilde{\psi}_e - \psi_e|, |1/\tilde{\psi}_e - 1/\psi_e|\} \leq C_\psi$ that controls the deviation to which we know the interventions on the latent variables. Following exactly similar logic as proof of Proposition 12, one can show that with probability tending to one, the optimal solutions of (7) with the constraint on the number of latent variables equal to the optimal solution of (44). Thus, we will analyze the estimates produced by (44).

The proof of Theorem 19 relies on the following lemmas:

Lemma 20 *Let* $\mathcal{D}$ *and* B *be any DAG and associated connectivity matrix that is optimal with respect to* (44). *Suppose there exists a non-singular diagonal matrix* $\Omega \in \mathbb{D}^p$ *satisfying the following relation for any* $e \in [m]$, $|\kappa| \leq 4C_\psi / \max_{e,e'}(\psi_{e'}^\star/\psi_e^\star - 1)$, $|\delta| \leq C_\psi$:

$$(\mathrm{Id}-B^\star)^{-1}(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta) - \kappa\Gamma^\star\Gamma^{\star T})(\mathrm{Id}-B^\star)^{-T} = (\mathrm{Id}-B)^{-1}\Omega(\mathrm{Id}-B)^{-T}. \tag{45}$$

Then, under Assumption 1', $\kappa = 0$.

Lemma 21 (Sufficient conditions for estimated and true latent interventions to be equal) *Under Assumptions 1',2,3',4" and 6, we have that for any optimal solution* $(B,\Gamma,\mathcal{I},\{\Omega_e,\Psi_e\}_{e=1}^m)$ *of* (44), $\psi_1 = \psi_1^\star$, $\psi_2 = \psi_2^\star$. *Furthermore, for any* $i \in \mathcal{I}^\star$, *letting* $e \in [m]$ *be the environment satisfying Assumption 4", we have that* $\psi_e = \psi_e^\star$.

Proof [Proof of Theorem 19] We appeal to Lemma 21 to conclude that for any optimal solution $(B,\Gamma,\mathcal{I},\{\Omega_e,\Psi_e\}_{e=1}^m)$, $\psi_1 = \psi_1^\star$, $\psi = \psi_2^\star$ and for every e satisfying Assumption 6 on the intervention on the observed variables, $\psi_e^\star = \psi_e$. We then follow a similar strategy to the proof of Theorem 8 to conclude the desired result. ■

Proof [Proof of Lemma 20] Taking matrix inverses of both sides of the equation in the lemma, we have that:

$$(\mathrm{Id}-B^\star)^T(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta) - \kappa\Gamma^\star\Gamma^{\star T})^{-1}(\mathrm{Id}-B^\star) = (\mathrm{Id}-B)^T\Omega^{-1}(\mathrm{Id}-B). \tag{46}$$

By Assumption 1' and the relation (46), we have that $\text{degree}[\text{moral}(\mathcal{D})] \geq 3d^\star$. We arrive at a contradiction with $\mathcal{D}$ being optimal however since for any $\mathcal{I}$,

$$\mathcal{R}_\gamma(\mathcal{D},I) > \text{degree}[\text{moral}(\mathcal{D})] \geq 3d^\star \geq \mathcal{R}_\gamma(\mathcal{D}^\star,\mathcal{I}^\star).$$

■

Proof [Proof of Lemma 21] Let $(B, \Gamma, \mathcal{I}, \{\Omega_e, \psi_e \mathrm{Id}\}_{e=1}^m)$ be any optimal parameters of (44). Then, by feasibility, we have for every $e \in [m]$

$$(\mathrm{Id} - B^\star)^{-1}(\Omega_e^\star + \psi_e^\star \Gamma^\star \Gamma^{\star T})(\mathrm{Id} - B^\star)^{-T} = (\mathrm{Id} - B)^{-1}(\Omega_e + \psi_e \Gamma \Gamma^T)(\mathrm{Id} - B)^{-T}. \tag{47}$$

Notice that without loss of generality, $\psi_1 = 1$ as Γ can be appropriately scaled. We will first show that the assumptions imply that $\psi_2^\star = \psi_2$. We prove $\psi_2 = \psi_2^\star$ both in the settings where $\psi_2^\star > \psi_1^\star$ and $\psi_2^\star < \psi_1^\star$.

Scenario $\psi_2^\star > \psi_1^\star$: The relation (47) implies that:

$$\begin{aligned} &(\mathrm{Id} - B^\star)^{-1}\left((1-\psi_2^\star)\Omega_1^\star + (\psi_2^\star - \psi_2)(\Gamma^\star \Gamma^{\star T} + \Omega_1^\star)\right)(\mathrm{Id} - B^\star)^{-T} \\ &= (\mathrm{Id} - B)^{-1}(\Omega_2 - \psi_2 \Omega_1)(\mathrm{Id} - B)^{-T}. \end{aligned} \tag{48}$$

By Assumption 4", we have that the matrix $(1-\psi_2^\star)\Omega_1^\star + (\psi_2^\star - \psi_2)(\Gamma^\star \Gamma^{\star T} + \Omega_1^\star)$ is non-singular and $\psi_2 \neq 1$. Rearranging the left-hand side of (48), we have that:

$$\begin{aligned} &(\mathrm{Id} - B^\star)^T\left(\Omega_1^\star - (\psi_2^\star - \psi_2)/(\psi_2 - 1)\Gamma^\star \Gamma^{\star T}\right)^{-1}(\mathrm{Id} - B^\star) \\ &= (\psi_2 - 1)(\mathrm{Id} - B)^T(\Omega_2 - \psi_2 \Omega_1)^{-1}(\mathrm{Id} - B). \end{aligned} \tag{49}$$

Let $\kappa := (\psi_2^\star - \psi_2)/(\psi_2 - 1)$. By the constraint on how close ψ_e is to $\psi_e^\star$ in (44), we have that $|\kappa| \leq 2C_\psi/\psi_2^\star - 1$. Appealing to Lemma 20, we have that $\kappa = 0$ or equivalently $\psi_2^\star = \psi_2$.

Scenario $\psi_2^\star < \psi_1^\star$: The relation (47) implies that:

$$\begin{aligned} &(\mathrm{Id} - B^\star)^{-1}\left((1/\psi_2^\star - 1)\Omega_2^\star + (1/\psi_2^\star - 1/\psi_2)(\Gamma^\star \Gamma^{\star T} - \Omega_2^\star)\right)(\mathrm{Id} - B^\star)^{-T} \\ &= (\mathrm{Id} - B)^{-1}(\Omega_1 - 1/\psi_2 \Omega_2)(\mathrm{Id} - B)^{-T}. \end{aligned} \tag{50}$$

By Assumption 4", we have that that the matrix $(1/\psi_2^\star - 1)\Omega_2^\star + (1/\psi_2^\star - 1/\psi_2)(\Gamma^\star \Gamma^{\star T} - \Omega_2^\star)$ is non-singular and $\psi_2 \neq 1$. Rearranging the left-hand side of (50), we have that:

$$\begin{aligned} &(\mathrm{Id} - B^\star)^{-1}\left(\Omega_1^\star - (1/\psi_2^\star - 1/\psi_2)/(1/\psi_2 - 1)\Gamma^\star \Gamma^{\star T}\right)(\mathrm{Id} - B^\star)^{-T} \\ &= \frac{1}{1/\psi_2 - 1}(\mathrm{Id} - B)^{-1}(\Omega_{e_1} - 1/\psi_2 \Omega_2)(\mathrm{Id} - B)^{-T}. \end{aligned} \tag{51}$$

Let $\kappa := (1/\psi_2^\star - 1/\psi_2)/(1/\psi_2 - 1)$. By the constraint on how close ψ_e is to $\psi_e^\star$ in (44), we have that $|\kappa| \leq 2C_\psi/(1/\psi_2^\star - 1)$. Appealing to Lemma 20, we have that $\kappa = 0$ or equivalently $\psi_2^\star = \psi_2$.

Consider any $hi \in \mathcal{I}^\star$ and let e be an environment satisfying Assumption 4", i.e. the observed variable X_i receives strong heterogeneous interventions at environment e. We will show that $\psi_e^\star = \psi_e$. Again, we consider two settings: $\psi_e^\star > \psi_1^\star$ and $\psi_e^\star < \psi_1^\star$ (notice that $\psi_1^\star \neq \psi_e^\star$ by Assumption 4"):

$\underline{\psi_e^\star > \psi_1^\star}$: The relation (47) implies that:

$$\begin{aligned}&(\mathrm{Id} - B^\star)^{-1}\left(\Omega_e^\star - \psi_e^\star\Omega_1^\star - (\psi_e^\star - \psi_e)(\Gamma^\star\Gamma^{\star T} + \Omega_1^\star)\right)(\mathrm{Id} - B^\star)^{-T} \\ &= (\mathrm{Id} - B)^{-1}(\Omega_e - \psi_e\Omega_1)(\mathrm{Id} - B)^{-T}.\end{aligned} \tag{52}$$

Letting $\sigma_{\min}(\cdot)$ be the minimum singular value of an input matrix, we have by Assumption 6 that:

$$\sigma_{\min}(\Omega_e^\star - \psi_e^\star\Omega_1^\star) \geq \min\{|1 - \psi_e^\star|\sigma_{\min}(\Omega_1^\star), (1 + \psi_e^\star)\sigma_{\min}(\Omega_1^\star)\} \geq |1 - \psi_e^\star|\sigma_{\min}(\Omega_1^\star). \tag{53}$$

By Assumption 4", we have that: $(\Omega_e^\star - \psi_e^\star\Omega_1^\star - (\psi_e^\star - \psi_e)(\Gamma^\star\Gamma^{\star T} + \Omega_1^\star)$ is non-singular. Therefore, re-arranging (52), we have that:

$$(\mathrm{Id} - B^\star)^T\left(\Omega_e^\star - \Omega_1^\star\psi_e - (\psi_e^\star - \psi_e)\Gamma^\star\Gamma^{\star T}\right)^{-1}(\mathrm{Id} - B^\star) = (\mathrm{Id} - B)^T(\Omega_e - \psi_e\Omega_1)^{-1}(\mathrm{Id} - B). \tag{54}$$

Let $\kappa := (\psi_e^\star - \psi_e)$. By the constraint on how close ψ_e is to $\psi_e^\star$ in (44), we have that $|\kappa| \leq 2C_\psi/\psi_e^\star - 1$. Appealing to Lemma 20, we have that $\kappa = 0$ or equivalently $\psi_e^\star = \psi_e$.

$\underline{\psi_e^\star < \psi_1^\star}$: The relation (47) implies that:

$$\begin{aligned}&(\mathrm{Id} - B^\star)^{-1}\left(\Omega_e^\star/\psi_e^\star - \Omega_1^\star + (1/\psi_e^\star - 1/\psi_e)(\Gamma^\star\Gamma^{\star T} - \Omega_e^\star)\right)(\mathrm{Id} - B^\star)^{-T} \\ &= (\mathrm{Id} - B)^{-1}(\Omega_1 - 1/\psi_e\Omega_e)(\mathrm{Id} - B)^{-T}.\end{aligned} \tag{55}$$

We have by Assumption 6 that:

$$\sigma_{\min}(1/\psi_e^\star\Omega_e^\star - \Omega_1^\star) \geq \min\{|1/\psi_e^\star - 1|\sigma_{\min}(\Omega_1^\star), (1/\psi_e^\star + 1)\sigma_{\min}(\Omega_1^\star)\} \geq |1/\psi_e^\star - 1|\sigma_{\min}(\Omega_1^\star) \tag{56}$$

By Assumption 4", we have that that the matrix $\left(\Omega_e^\star/\psi_e^\star - \Omega_1^\star + (1/\psi_e^\star - 1/\psi_e)(\Gamma^\star\Gamma^{\star T} - \Omega_e^\star)\right)$ is non-singular and $\psi_e \neq 0$. Rearranging the left-hand side of (50), we have that:

$$\begin{aligned}&(\mathrm{Id} - B^\star)^{-1}\left(\Omega_e^\star - \Omega_1^\star\psi_e - \psi_e(1/\psi_e^\star - 1/\psi_e)\Gamma^\star\Gamma^{\star T}\right)(\mathrm{Id} - B^\star)^{-T} \\ &= \psi_e(\mathrm{Id} - B)^{-1}(\Omega_e/\psi_e - \Omega_1)(\mathrm{Id} - B)^{-T}.\end{aligned} \tag{57}$$

Let $\kappa := (1/\psi_e^\star - 1/\psi_e)/(1/\psi_e)$. By the constraint on how close ψ_e is to $\psi_e^\star$ in (44), we have that $|\kappa| \leq 2C_\psi/(1/\psi_e^\star - 1)$. Appealing to Lemma 20, we have that $\kappa = 0$ or equivalently $\psi_e^\star = \psi_e$. ■

J.1 Assumptions 1' is weaker than an incoherence-type condition on the latent effects

We first show that Assumption 1' is generally satisfied even when the number of latent variables is large, whereas the incoherence-type assumption requires that the number of latent variables is far smaller than the ambient dimension. Specifically, using the Woodbury

inversion lemma, consider the following decomposition of $(\mathrm{Id} - B^\star)^T(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta) + \kappa\Gamma^\star\Gamma^{\star T})^{-1}(\mathrm{Id} - B^\star)$ (which appears in Assumption 1') when $\kappa \neq 0$

$$(\mathrm{Id} - B^\star)^T(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta) + \kappa\Gamma^\star\Gamma^{\star T})^{-1}(\mathrm{Id} - B^\star) = S_e - L_e,$$

where

$$\begin{aligned} S_e &= (\mathrm{Id} - B^\star)^T(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta))^{-1}(\mathrm{Id} - B^\star), \\ L_e &= (\mathrm{Id} - B^\star)^T(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta))^{-1}\Gamma^\star(\kappa\mathrm{Id} + \Gamma^{\star T} \\ &\quad (\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta))^{-1}\Gamma^\star)^{-1}\Gamma^{\star T}(\Omega_e^\star - \Omega_1^\star(\psi_e^\star + \delta))^{-1}(\mathrm{Id} - B^\star). \end{aligned}$$

Assumption 1' imposes a lower-bound on the degree of the matrix sum $S_e - L_e$, which we show is not very stringent and holds even when the number of latent variables is large. Specifically, notice that $\mathrm{degree}(S_e) \leq \mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)]$. Furthermore, generally, even when the number of latent variables is large, $\mathrm{degree}(L_e)$ is large relative to the ambient dimension. Due to the basic inequality $\mathrm{degree}(S_e + L_e) \geq \mathrm{degree}(L_e) - \mathrm{degree}(S_e)$, it is then straightforward to see that the condition in Assumption 1' is generally satisfied.

Furthermore, even in settings where the number of latent variables is far smaller than the ambient dimension, Assumption 1' can be far weaker than the incoherence condition in Assumption 1'. For illustration, we consider a simple setting when the number of latent variables is equal to one and show that an incoherence-type condition implies the condition $\mathrm{degree}(S_e + L_e) \geq 3\mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)]$ in Assumption 1'. Specifically, some linear algebraic manipulations yield that for any rank-1 symmetric matrix M, $\mathrm{degree}[M] \geq 1/\mathrm{inc}[\text{col-space}(M)]^2$. Employing this inequality in conjunction with the bound $\mathrm{degree}(S_e + L_e) \geq \mathrm{degree}(L_e) - \mathrm{degree}(S_e)$, we find that

$$\begin{aligned} \mathrm{degree}(S_e + L_e) &\geq 1/\mathrm{inc}[\text{col-space}(L_e)]^2 - \mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)] \\ &= \mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)](1/(\mathrm{inc}[\text{col-space}(L_e)]^2\mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)]) - 1). \end{aligned}$$

The above inequality suggests than the incoherence-type condition $\mathrm{inc}[\text{col-space}(L_e)]^2\mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)] < 4$ would imply the condition in Assumption 1' .

Appendix K. Illustration with unperturbed latent variables

We consider the following illustration to show that if no constraints are imposed on the number of latent variables, the equivalence class of optimally scoring DAGs when $|\mathcal{I}^\star| < p$ could be very different than $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. In this section, we will construct a simple example where the equivalence class of optimally scoring DAGs is the empty graph, even when the population graph has multiple edges. Specifically, consider the following structural equation model (specialization of (2)) among DAG of three nodes and a single normally distributed latent variable for all environments $e \in \mathcal{E}$:

$$\begin{aligned} X_1^e &= c_1 H + \epsilon_1, \\ X_2^e &= c_2 H + \epsilon_2, \\ X_3^e &= c_3 H + \alpha X_1^e + \epsilon_3 + \delta^e. \end{aligned} \tag{58}$$

Here, $\epsilon_1, \epsilon_2, \epsilon_3 \sim \mathcal{N}(0,1)$ and δ^e is identically zero for environment $e = 1$ (observational) and $\delta^e \sim \mathcal{N}(0, w^e)$ for $e > 1$. Note that by construction, $\mathcal{I}^\star = \{3\}$. The SCM (58) is then parameterized by the following quantities:

$$\begin{aligned} B^\star &= \begin{pmatrix} 1 & 0 & 0 \\ \beta & 1 & 0 \\ \alpha & 0 & 1 \end{pmatrix} \quad ; \quad \Omega_1^\star = \mathrm{Id} \quad ; \quad \Omega_e^\star = \Omega_1^\star + \mathrm{diag}\begin{pmatrix} 0 & 0 & w_e \end{pmatrix} \text{ for } e > 1 \\ \Gamma^\star &= \begin{pmatrix} c_1 & c_2 & c_3 \end{pmatrix}^T . \end{aligned} \tag{59}$$

We then have the following theorem statement:

Proposition 22 (Equivalence class with unperturbed latent variables) *Consider the SCM* (58)*. In population, for any $\gamma > 0$, $\mathcal{D}^{\gamma}_{regul.opt}$ is the empty graph.*

Proof We will construct Γ, Ω_e such that together with the connectivity matrix associated to an empty DAG, they entail the same covariance model as the population. Specifically, consider the following set of parameters:

$$\begin{aligned} &B = 0 \quad ; \quad \Omega_1 = \lambda_{\min}(\Sigma_e^\star)/2\mathrm{Id} \; ; \; \Omega_e = \Omega_1 + \mathrm{diag}\begin{pmatrix} 0 & 0 & w_e \end{pmatrix} \text{ for } e > 1 \; ; \\ &\Gamma = \text{matrix square root of } \Sigma_e^\star - \lambda_{\min}(\Sigma_e^\star)/2\mathrm{Id}, \end{aligned} \tag{60}$$

where $\Sigma_e^\star$ is the population covariance. Thus, the intervention set encoded by the model (60) is $\mathcal{I} = \{3\}$. It is straightforward to see that the parameters (B, Γ, Ω_e) entail the same covariance as the population model, that is:

$$(\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma\Gamma^T)(\mathrm{Id} - B)^{-T} = \Sigma_e^\star \quad \text{for all } e \in \mathcal{E}.$$

Note that the graph encoded by B is the empty graph. Furthermore, note that any model that yields the same covariance as the population must contain at least a single intervention target since $\Sigma_e^\star - \Sigma_1^\star$ has rank-1. We have concluded the result. ■

Appendix L. Consistency guarantees of Algorithm 1 and Algorithm 2

Throughout, we assume that for a given DAG $\mathcal{D}$, number of latent variables $\bar{h}$ and intervention targets $\mathcal{I}$, Algorithm 0 obtains the global optimum solution. As required in Corollary 7, we will assume that the input number of latent variables $\bar{h}$ is greater than the true number of latent variables, i.e. $\bar{h} \geq \dim(H)$.

L.1 Consistency guarantees of Algorithm 1

We will denote the output $\hat{\Theta}(\hat{\mathcal{D}}_{\mathrm{opt}}, \bar{h})$ of Algorithm 1 by $(\hat{\mathcal{D}}_{\mathrm{opt}}, \hat{B}_{\mathrm{opt}}, \hat{\Gamma}_{\mathrm{opt}}, \{\hat{\Omega}_{\mathrm{opt},e}, \hat{\Psi}_{\mathrm{opt},e}\}_{e=1}^m)$. We also take $\lambda \to 0$ with the rate given in Proposition 12, and assume that the parameters are in a compact space $\mathcal{F}$ for technical reasons; see Section C. We will prove the following formal statement, where recall that $d^\star := \mathrm{degree}[\mathrm{moral}(\mathcal{D}^\star)]$.

Theorem 23 *Consider a set of candidate DAGs $\mathcal{D}_{cand}$ and suppose $\exists \mathcal{D} \in \mathcal{D}_{cand}$ with the following two properties:*

1. *there are parameters* $(B, \Gamma, \{\Psi_e, \Omega_e\}_{e=1}^m)$ *with* $B \sim \mathcal{D}$ *and* $\Sigma_e^\star = (\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)^{-T}$ *for every* $e \in [m]$.
2. $\mathrm{degree}[\mathit{moral}(\mathcal{D})] + \|\mathcal{D}\|_{\ell_0} \leq \mathrm{degree}[\mathit{moral}(\mathcal{D}^\star)] + \|\mathcal{D}^\star\|_{\ell_0}$

Suppose that Assumptions 1-2 and 4-5 are satisfied. Let $d^\star \geq \gamma > 0$. *Let* $\nu^\star$ *be a positive integer with* $\nu^\star \geq d^\star$. *Suppose that* $48\nu^\star \mathrm{inc}[\text{col-space}((\mathrm{Id} - \hat{B}_{opt})^T\hat{\Omega}_{opt,e}^{-1}\hat{\Gamma}_{opt})] < 1$ *for all* $e \in [m]$. *Then,* $\hat{\mathcal{I}}_{opt} = \mathcal{I}^\star$ *and* $\hat{\mathcal{I}}_{\mathrm{opt}}\text{-MEC}(\hat{\mathcal{D}}_{opt}) = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ *with probability tending to one.*

We immediately have the following corollary noting that DAGs in the same Markov equivalence class have the same moral graph and the same number of edges.

Corollary 24 *Suppose that* $\mathcal{D}_{cand} \cap \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star) \neq \emptyset$. *Then,* $\hat{\mathcal{I}}_{opt} = \mathcal{I}^\star$ *and* $\hat{\mathcal{I}}_{\mathrm{opt}}\text{-MEC}(\hat{\mathcal{D}}_{opt}) = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ *with probability tending to one.*

Proof [Proof of Theorem 23] The proof will rely on the following fact about the candidate set of DAGs $\mathcal{D}_{\mathrm{cand}}$: there exists a DAG $\mathcal{D} \in \mathcal{D}_{\mathrm{cand}}$ and associated parameters $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ with $B \sim \mathcal{D}$ that is compatible with the data distribution, i.e. $\Sigma_e^\star = (\mathrm{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)^{-T}$ for every $e \in [m]$. Indeed, By the assumption of the theorem, we have that $\mathcal{D}_{\mathrm{cand}}$ contains at least one of the DAGs in $\mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$. We will denote this DAG by $\tilde{\mathcal{D}}^\star$. By Theorem 2, we have that there exists a model associated with the DAG $\tilde{\mathcal{D}}^\star$ that is compatible with the data distributions.

We will analyze different components of Algorithm 1.

Steps 2-3 of Algorithm 1: As described in these steps in the main text, we take $\mathcal{I} = [p]$. Step 2 of Algorithm 1 scores different DAGs in the candidate set, with the score:

$$S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) := \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} \sum_{e=1}^m \hat{\pi}_e \Big(\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) + \mathrm{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)\hat{\Sigma}_e)\Big) + \lambda\mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}).$$
$$\text{subject-to: } B \sim \mathcal{D}, \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}$$

Notice the score $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$ is the same as the score $\mathtt{score}_{\lambda,\gamma}(\mathcal{D}, \hat{\Theta}(\mathcal{D}, \bar{h}))$ in (4). We define the population analogue of the score $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$ below:

$$S^\star(\mathcal{D}, \bar{h}, \mathcal{I}) := \min_{(B,\Gamma,\{\Omega_e,\Psi_e\}_{e=1}^m)\in\mathcal{F}} \sum_{e=1}^m \pi_e^\star \Big(\log\det(\Omega_e + \Gamma\Psi_e\Gamma^T) + \mathrm{tr}((\mathrm{Id} - B)^T(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id} - B)\Sigma_e^\star)\Big).$$
$$\text{subject-to: } B \sim \mathcal{D}, \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I}$$

Under the rate of regularization λ described in Proposition 12, following very similar proof strategy to that of Proposition 12, we conclude that $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \to S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ in the infinite data limit. As described earlier, we have that there exists a model associated with the DAG $\tilde{\mathcal{D}}^\star$ that is compatible with the data distributions. By Lemma 13, we have that

that the minimizers of $\operatorname{argmin}_{\mathcal{D}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ are precisely models that are compatible with the data distributions; that is:

$$(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m) \text{ optimal } \Leftrightarrow \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \text{ for every } e \in [m],$$

where optimal here means that they result in the smallest score according to $S^\star(\cdot,\cdot,\cdot)$. Since the DAG $\tilde{\mathcal{D}}^\star$ attains optimal score, any candidate DAG that is selected in this step must also attain this optimal score.

With the choice of λ in Proposition 12, we have that it is above fluctuations due to sampling error; it follows that for two population score equivalent models $(\mathcal{D}_1, \bar{h}, \mathcal{I})$ and $(\mathcal{D}_2, \bar{h}, \mathcal{I})$ with $S^\star(\mathcal{D}_1, \bar{h}, \mathcal{I}) = S^\star(\mathcal{D}_2, \bar{h}, \mathcal{I})$, if $\mathcal{R}_\gamma(\mathcal{D}_1, \mathcal{I}) < \mathcal{R}_\gamma(\mathcal{D}_2, \mathcal{I})$, then, there exists N such that for $n_e \geq N$, $S_{\lambda,\gamma}(\mathcal{D}_1, \bar{h}, \mathcal{I}) < S_{\lambda,\gamma}(\mathcal{D}_2, \bar{h}, \mathcal{I})$. Thus, in the infinite data limit, with probability tending to one, the output of steps 3 of Algorithm 1 is with probability tending to one the minimizer of the following optimization problem with $\mathcal{I} = [p]$

$$\begin{aligned} \underset{\mathcal{D}\in\mathcal{D}_{\text{cand}}, B, \Gamma, \{\Omega_e,\Psi_e\}_{e=1}^m}{\operatorname{argmin}} \quad & \mathcal{R}_\gamma(\mathcal{D}, \mathcal{I}) \\ \text{subject-to :} \quad & B \sim \mathcal{D}, \mathbb{I}(\{\Omega_e\}_{e=1}^m) \subseteq \mathcal{I} \text{ and} \\ & \Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T} \text{ for every } e \in [m]. \end{aligned} \tag{61}$$

Let $\hat{\mathcal{D}}_{\text{opt}}$ be the output of Step 3 of Algorithm 2. From the analysis above, we have that in the infinite data limit and with probability tending to one, $\hat{\mathcal{D}}_{\text{opt}}$ and its associated parameters of minimizers of (61). Furthermore, since $\tilde{\mathcal{D}}^\star$ and its associated parameters are feasible in (61), we can make two conclusions with probability tending to one:

$$\begin{aligned} &1.\ p \text{ degree}[\text{moral}(\hat{\mathcal{D}}_{\text{opt}})] + \|\hat{\mathcal{D}}_{\text{opt}}\|_{\ell_0} \leq pd^\star + \|\mathcal{D}^\star\|_{\ell_0}, \\ &2.\ S^\star(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, [p]) = \min_{\mathcal{D},\mathcal{I}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I}). \end{aligned} \tag{62}$$

Here, the first fact is based on every member of the Markov equivalence class having the same moral graph and the same number of edges. The second fact is based on noting that the population score $S^\star(\cdot,\cdot,\cdot)$ is minimized by models that are compatible with the data distribution, and for any DAG $\mathcal{D}$, $S^\star(\mathcal{D}, \bar{h}, \mathcal{I}_1) \leq S^\star(\mathcal{D}, \bar{h}, \mathcal{I}_2)$ with $\mathcal{I}_1 \supseteq \mathcal{I}_2$.

Step 4 of Algorithm 1: Recall that Step 4 of Algorithm 1 takes the best scoring model (among the candidate sets). With this best model, it greedily removes intervention targets until the score of the resulting model does not improve any further. From the analysis of Steps 2-3, we have with probability tending to one that the best scoring DAG $\hat{\mathcal{D}}_{\text{opt}}$ is a minimizer of the optimization (61). As we described in property 2 of (62), this model minimizes the population score $\operatorname{argmin}_{\mathcal{D},\mathcal{I}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$. Furthermore, recall that for any $\mathcal{I}$, $S_{\lambda,\gamma}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}) \to S^\star(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I})$. Therefore, at every step of the greedy algorithm, an intervention target can only be removed if the resulting intervention targets $\mathcal{I}$ satisfies: $S^\star(\mathcal{D}_{\text{opt}}, \bar{h}, \mathcal{I}) = S^\star(\mathcal{D}_{\text{opt}}, \bar{h}, [p])$. Since the model obtained after step 3 is already optimally scoring (i.e. minimizing $S^\star(\cdot,\cdot,\cdot)$ with probability tending to one), we have that every iteration of step 4, the associated model $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ satisfies the condition $\Sigma_e^\star = (\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)^{-1}(\mathrm{Id}-B)^{-T}$ for every $e \in [m]$, with $B \sim \hat{\mathcal{D}}_{\text{opt}}$.

Putting things together: For notational ease, let $(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m)$ be the output of Algorithm 1. From the analysis above we have that:

$$(\mathrm{Id}-B)^{-1}(\Omega_e+\Gamma\Psi_e\Gamma^T)(\mathrm{Id}-B)^{-T} = (\mathrm{Id}-B^\star)^{-1}(\Omega_e^\star+\Gamma^\star\Psi_e^\star\Gamma^{\star T})(\mathrm{Id}-B^\star)^{-T} \text{ for all } e \in [m].$$

Equivalently, taking the inverse of both sides in the previous equation, and using the Woodbury Inversion Lemma, we have for $e = 1, 2, \ldots, m$

$$\begin{aligned}&(\mathrm{Id}-B)^T\Omega_e^{-1}(\mathrm{Id}-B) - (\mathrm{Id}-B)^T\Omega_e^{-1}\Gamma(\Psi_e^{-1}+\Gamma^T\Omega_e^{-1}\Gamma)^{-1}\Gamma^T\Omega_e^{-1}(\mathrm{Id}-B)\\ &= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}(\mathrm{Id}-B^\star) - (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}\Gamma^\star(\Psi_e^{\star-1}+\Gamma^{\star T}\Omega_e^{\star-1}\Gamma^\star)^{-1}\Gamma^{\star T}\Omega_e^{\star-1}(\mathrm{Id}-B^\star).\end{aligned}$$

Define for every $e = 1, 2, \ldots, m$ the following quantities:

$$\begin{aligned}K_e &:= (\mathrm{Id}-B)^T\Omega_e^{-1}(\mathrm{Id}-B),\\ K_e^\star &= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}(\mathrm{Id}-B^\star),\\ L_e &:= (\mathrm{Id}-B)^T\Omega_e^{-1}\Gamma(\Psi_e^{-1}+\Gamma^T\hat{\Omega}_e^{-1}\Gamma)^{-1}\Gamma^T\Omega_e^{-1}(\mathrm{Id}-B),\\ L_e^\star &:= (\mathrm{Id}-B^\star)^T\Omega_e^{\star-1}\Gamma^\star(\Psi_e^{\star-1}+\Gamma^{\star T}\Omega_e^{\star-1}\Gamma^\star)^{-1}\Gamma^{\star T}\Omega_e^{\star-1}(\mathrm{Id}-B^\star).\end{aligned}$$

Note that from Lemma 15, we have the following implication for any $e \in [m]$

$$\mathrm{degree}(K_e - K_e^\star)\mathrm{inc}[\text{col-space}(L_e - L_e^\star)]^2 < 1 \Rightarrow K_e = K_e^\star. \tag{63}$$

In this analysis, we show that under the conditions described in Theorem 23, $\mathrm{degree}(K_e - K_e^\star)\mathrm{inc}[\text{col-space}(L_e - L_e^\star)]^2 < 1$, allowing us to conclude that $K_e = K_e^\star$. Notice that $\mathrm{degree}(K_e - K_e^\star) \leq \mathrm{degree}[\mathrm{moral}(\mathcal{D})] + d^\star$. Further, by Lemma 16, $\mathrm{inc}[\text{col-space}(L_e - L_e^\star)] \leq 2(\mathrm{inc}[\text{col-space}(L_e)] + \mathrm{inc}[\text{col-space}(L_e^\star)])$. Thus, it suffices to show for all $e \in [m]$ that:

$$4(\mathrm{degree}[\mathrm{moral}(\mathcal{D})] + d^\star)(\mathrm{inc}[\text{col-space}(L_e)]^2 + \mathrm{inc}[\text{col-space}(L_e^\star)]^2) < 1. \tag{64}$$

From the property 1 in (62), we have that $\mathrm{degree}[\mathrm{moral}(\mathcal{D})] \leq 2d^\star$. Therefore, the following conditions are satisfied for every $e \in [m]$ due to Assumption 1, the conditions of Corollary 7, and the bound $d^\star \leq \nu^\star$:

$$\begin{aligned}&\mathrm{degree}[\mathrm{moral}(\mathcal{D})]\mathrm{inc}[\text{col-space}(L_e)]^2 \leq 3\nu^\star\mathrm{inc}[\text{col-space}(L_e)]^2 < 1/16,\\ &\mathrm{degree}[\mathrm{moral}(\mathcal{D})]\mathrm{inc}[\text{col-space}(L_e^\star)]^2 \leq 3\nu^\star\mathrm{inc}[\text{col-space}(L_e^\star)] < 1/16,\\ &d^\star\mathrm{inc}[\text{col-space}(L_e)]^2 \leq \nu^\star\mathrm{inc}[\text{col-space}(L_e)]^2 < 1/16,\\ &d^\star\mathrm{inc}[\text{col-space}(L_e^\star)]^2 < 1/16.\end{aligned}$$

Combining these relations, we arrive at the inequality in (64). We have thus concluded

$$(\mathrm{Id}-B)(\mathrm{Id}-B^\star)^{-1}\Omega_e^\star(\mathrm{Id}-B^\star)^1(\mathrm{Id}-B)^T = \Omega_e \quad e = 1, 2, \ldots, m,$$

Following the same steps as proof of Theorem 6, we conclude that $\mathbb{I}(\{\Omega_e\}_{e=1}^m) = \mathcal{I}^\star$ and $\hat{\mathcal{D}}_{\mathrm{opt}} \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$. Finally, it remains to check that the output of the intervention targets $\hat{\mathcal{I}}_{\mathrm{opt}}$ is equal to $\mathcal{I}^\star$. Since, $\hat{\mathcal{I}}_{\mathrm{opt}} \supseteq \mathbb{I}(\{\Omega_e\}_{e=1}^m)$, we clearly have that $\hat{\mathcal{I}}_{\mathrm{opt}} \supseteq \mathcal{I}^\star$. Suppose that there exists a $j \in \hat{\mathcal{I}}_{\mathrm{opt}} \setminus \mathcal{I}^\star$. Notice that $S^\star(\hat{\mathcal{D}}_{\mathrm{opt}}, \bar{h}, \hat{\mathcal{I}}_{\mathrm{opt}} \setminus \{j\}) = \mathrm{argmin}_{\mathcal{D},\mathcal{I}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$. With the choice of λ in Proposition 12, we have that it is above fluctuations due to sampling error;

this, if follows that for two population score equivalent models $(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_1)$ and $(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_2)$ with $S^\star(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_1) = S^\star(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_2)$, if $\mathcal{R}_\gamma(\hat{\mathcal{D}}_{\text{opt}}, \mathcal{I}_1) < \mathcal{R}_\gamma(\hat{\mathcal{D}}_{\text{opt}}, \mathcal{I}_2)$, then, there exists N such that for $n_e \geq N$, $S_{\lambda,\gamma}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_1) < S_{\lambda,\gamma}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}, \mathcal{I}_2)$. Letting $\mathcal{I}_1 = \hat{\mathcal{I}}_{\text{opt}} \backslash \{j\}, \mathcal{I}_2 = \hat{\mathcal{I}}_{\text{opt}}$, we can conclude that the target j could have been removed to improve the score. Repeating this argument, we can conclude that $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}^\star$. ■

L.2 Consistency guarantees of Algorithm 2

We will denote the output $\hat{\Theta}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h})$ of Algorithm 2 by $(\hat{\mathcal{D}}_{\text{opt}}, \hat{B}_{\text{opt}}, \hat{\Gamma}_{\text{opt}}, \{\hat{\Omega}_{\text{opt},e}, \hat{\Psi}_{\text{opt},e}\}_{e=1}^m)$. We also take $\lambda \to 0$ with the rate given in Proposition 12, and assume that the parameters are in a compact space $\mathcal{F}$ for technical reasons; see Section C. We will prove the following formal statement, where recall that $d^\star := \text{degree}[\text{moral}(\mathcal{D}^\star)]$.

Theorem 25 *Consider a set of candidate DAGs $\tilde{\mathcal{D}}_{cand}$ and suppose $\exists \bar{\mathcal{D}} \in \tilde{\mathcal{D}}_{cand}, \bar{\mathcal{D}}^\star \in \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ such that $\bar{\mathcal{D}} \supseteq \bar{\mathcal{D}}^\star$. Suppose that Assumptions 1-2 and 4-5 are satisfied. Let $d^\star \geq \gamma > 0$. Let $\nu^\star$ be a positive integer with $\nu^\star \geq d^\star$. Suppose that $48\nu^\star \text{inc}[\text{col-space}((\text{Id} - \hat{B}_{opt})^T \hat{\Omega}_{opt,e}^{-1} \hat{\Gamma}_{opt})] < 1$ for all $e \in [m]$. Then, $\hat{\mathcal{I}}_{opt} = \mathcal{I}^\star$ and $\hat{\mathcal{I}}_{\text{opt}}\text{-MEC}(\hat{\mathcal{D}}_{opt}) = \mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$ with probability tending to one.*

Proof Recall that Algorithm 2 starts with a candidate set of 'starting point' DAGs $\tilde{\mathcal{D}}_{\text{cand}}$ and greedily deletes spurious edges in each DAG to obtain a modified set of candidate DAGs $\mathcal{D}_{\text{cand}}$. Our proof proceeds by showing that with probability tending to 1, the candidate set of DAGs $\mathcal{D}_{\text{cand}}$ contains a DAG $\mathcal{D}$ with associated parameters $(B, \Gamma, \{\Psi_e, \Omega_e\}_{e=1}^m)$ such that i) $\Sigma_e^\star = (\text{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\text{Id} - B)^{-T}$ for every $e \in [m]$, and ii) $\text{degree}[\text{moral}(\mathcal{D})] + \|\mathcal{D}\|_{\ell_0} \leq \text{degree}[\text{moral}(\mathcal{D}^\star)] + \|\mathcal{D}^\star\|_{\ell_0}$. Then, we appeal to Theorem 23 to obtain the desired result.

Step 2a of Algorithm 2: Recall the scores $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$ and its population analogue $S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ defined in the proof of Theorem 23. Step 2a computes the score $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I})$. Under the rate of regularization λ described in Proposition 12, following very similar proof strategy to that of Proposition 12, we can conclude that $S_{\lambda,\gamma}(\mathcal{D}, \bar{h}, \mathcal{I}) \to S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ in the infinite data limit for every $\mathcal{D}, \mathcal{I}, \bar{h}$. Let $\bar{\mathcal{D}} \in \tilde{\mathcal{D}}_{\text{cand}}$ be a DAG in the original candidate set that is a supergraph of a DAG in $\mathcal{I}^\star\text{-MEC}(\mathcal{D}^\star)$. By Theorem 2, we have that there exists a model associated with the DAG $\bar{\mathcal{D}}$ that is compatible with the data distributions. By Lemma 13, we have that that the minimizers of $\text{argmin}_{\mathcal{D}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$ are precisely models that are compatible with the data distributions; that is:

$$(B, \Gamma, \{\Omega_e, \Psi_e\}_{e=1}^m) \text{ optimal } \Leftrightarrow \Sigma_e^\star = (\text{Id} - B)^{-1}(\Omega_e + \Gamma\Psi_e\Gamma^T)^{-1}(\text{Id} - B)^{-T} \text{ for every } e \in [m],$$

where optimal here means that they result in the smallest score according to $S^\star(\cdot, \cdot, \cdot)$.

Step 2b of Algorithm 2: Since the DAG $\bar{\mathcal{D}}$ attains optimal score, any removal of the edges that is selected in this step must also attain this optimal score. Let $\hat{\mathcal{D}}$ be the DAG after removing spurious edges. Then, the associated model is also compatible with the data distribution.

Suppose that $\hat{\mathcal{D}}$ is a strict supergraph of $\bar{\mathcal{D}}^\star$ (see theorem statement for definition). We will show that the score of $\hat{\mathcal{D}}$ will be larger than that of $\bar{\mathcal{D}}^\star$, so we conclude that the output of Step 2 must be a DAG $\hat{\mathcal{D}}$ that is a subgraph of $\bar{\mathcal{D}}^\star$ and thus p degree[moral($\hat{\mathcal{D}}$)] $+ \|\hat{\mathcal{D}}\|_{\ell_0} <$ p degree[moral($\bar{\mathcal{D}}$)] $+ \|\bar{\mathcal{D}}\|_{\ell_0}$.

To show the above statement, let (i, j) be a pair of edges that are connected in $\hat{\mathcal{D}}$ but are not connected in $\bar{\mathcal{D}}^\star$. Since the DAG $\hat{\mathcal{D}} \setminus (i \to j)$ that is obtained by removing the edge between the pair (i, j) from $\hat{\mathcal{D}}$ is still a supergraph of $\bar{\mathcal{D}}^\star$, we have that $S^\star(\hat{\mathcal{D}} \setminus (i \to j), \bar{h}, [p]) = \text{argmin}_{\mathcal{D},\mathcal{I}} S^\star(\mathcal{D}, \bar{h}, \mathcal{I})$. With the choice of λ in Proposition 12, we have that it is above fluctuations due to sampling error; it follows that for two population score equivalent models $(\mathcal{D}_1 \bar{h}, [p])$ and $(\mathcal{D}_2, \bar{h}, [p])$ with $S^\star(\mathcal{D}_1, \bar{h}, [p]) = S^\star(\mathcal{D}_2, \bar{h}, [p])$, if $\mathcal{R}_\gamma(\mathcal{D}_1, [p]) < \mathcal{R}_\gamma(\mathcal{D}_2, [p])$, then, there exists N such that for $n_e \geq N$, $S_{\lambda,\gamma}(\mathcal{D}_1, \bar{h}, [p]) <$ $S_{\lambda,\gamma}(\mathcal{D}_2, \bar{h}, [p])$. Furthermore for DAGs $\mathcal{D}_1, \mathcal{D}_2$ with $\mathcal{D}_1$ strict subgraph of $\mathcal{D}_2$, we have that p degree[moral($\mathcal{D}_1$)] $+ \|\mathcal{D}_1\|_{\ell_0} < p$ degree[moral($\mathcal{D}_2$)] $+ \|\mathcal{D}_2\|_{\ell_0}$. Letting $\mathcal{D}_2 = \hat{\mathcal{D}}$ and $\mathcal{D}_1 = \hat{\mathcal{D}} \setminus (i \to j)$, we can conclude that the edge between the pair (i, j) could have been removed to improve the score. Repeating this argument, we can conclude the desired result.

■

Appendix M. Selecting a cross-validated causal model

The regularization parameters (λ, γ) and the number of latent variables $\bar{h}$ in Algorithm 1 and Algorithm 2 are generally unknown. Here, we will propose an efficient cross-validation approach to select a causal model (and the associated interventional equivalence class). Our approach is based on the following reparameterization of the regularization parameters: $\lambda_B := \lambda, \lambda_I := \lambda\gamma$, so that we must select $\bar{h}, \lambda_B, \lambda_I$. This particular reparameterization and the nature of the procedures in Algorithm 1 and Algorithm 2 lead to the following important simplifications: for a given $\bar{h}$, we can first do a 1-dimensional grid search for λ_B to choose a DAG and another 1-dimensional grid search for λ_I to choose an intervention set. Furthermore, building on the previous simplification, the regularization parameter λ_B selects a DAG $\mathcal{D}$ from a finite set of available DAGs $\mathcal{D}_{\text{set}}$; these are all the candidate DAGs and all the pruned DAGs (according to the greedy backward deletion). Similarly, the regularization parameter λ_I selects intervention targets from a finite collection of available sets $\mathcal{I}_{\text{set}}$; these are the collection of intervention targets according to the greedy backward deletion. Thus, choosing an optimal λ_B, λ_I based on validation is equivalent to choosing an optimal DAG and optimal intervention targets from the sets $\mathcal{D}_{\text{set}}$ and $\mathcal{I}_{\text{set}}$.

With the observations above, we can propose an efficient cross-validation approach to select a causal model for both Algorithms 1 and 2. For simplicity, we assume that the data is split into a training set and two test sets. The first test data will be used to determine the number of latent variables and the DAG (among the candidate set), and the second test data will be used to select the intervention set $\mathcal{I}$. The training data is parameterized by the covariance matrices $\hat{\Sigma}_e^{\text{train}}$ and mixture values $\hat{\pi}_e^{\text{train}}$ for every $e = 1, 2, \ldots, m$; similarly, the test data is parameterized by $\hat{\Sigma}_e^{\text{test1}}, \hat{\Sigma}_e^{\text{test2}}$ and mixture values $\hat{\pi}_e^{\text{test1}}$,$\hat{\pi}_e^{\text{test2}}$. We measure the likelihood of a given causal model $\hat{\Theta} = (\hat{B}, \hat{\Gamma}, \{\Omega_e, \Psi_e\}_{e=1}^m)$ on the first split of the test

is:

$$\texttt{score-test}(\hat{\Theta}) := \sum_{e=1}^{m} \hat{\pi}_e^{\text{test1}} \ell(\hat{B}, \hat{\Gamma}, \hat{\Omega}_e, \hat{\Psi}_e; \hat{\Sigma}_e^{\text{test1}}).$$

A similar measure can be defined to quantify the likelihood on the second split of the test data.

We are now ready to state a cross-validated version of Algorithm 1, which is presented below.

Algorithm 3 Equivalence class of best scoring DAGs from a candidate set via *cross-validated UT-LVCE*

1: **Input**: candidate DAG(s) $\mathcal{D}_{\text{cand}}$; training dataset and and two test datasets; max # of latent vars. $h_{\max}$; initial intervention set $\mathcal{I} = [p]$

2: **Causal parameters for each DAG and # latent vars.**: for each $\mathcal{D} \in \mathcal{D}_{\text{cand}}$ and $\bar{h} \leq \bar{h}_{\max}$, supply training data as well as $\mathcal{D}$, $\bar{h}$, and $\mathcal{I}$ to Algorithm 0 to obtain the causal parameters $\hat{\Theta}_{\mathcal{I}}(\mathcal{D}, \bar{h})$

3: **Find an optimal DAG and # latent variables using test data**: use the first test dataset to obtain the best DAG and number of latent variables $(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}_{\text{opt}}) = \arg\min_{\mathcal{D} \in \mathcal{D}_{\text{cand}} \bar{h} \leq \bar{h}_{\max}} \texttt{score-test}(\hat{\Theta}_{\mathcal{I}}(\mathcal{D}, \bar{h}))$

4: **Updating intervention targets $\mathcal{I}$ using test data**: initialize $\mathcal{I}_{\text{set}} = \{\mathcal{I}\}$ and
   (a) let $\{\hat{\Omega}_e\}_{e=1}^m$ be noise variance encoded in $\hat{\Theta}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}_{\text{opt}})$
   (b) estimate intervention strengths via $\xi_j := \frac{1}{m}\sum_{e=1}^m ([\hat{\Omega}_e]_{j,j} - \frac{1}{m}\sum_{e=1}^m [\hat{\Omega}_e]_{j,j})^2$ for each $j \in [p]$
   (c) remove the variable with weakest estimated intervention: $\mathcal{I} \leftarrow \mathcal{I} \setminus \{\arg\min_{j \in \hat{\mathcal{I}}} \xi_j\}$; add $\{\mathcal{I}\}$ to $\mathcal{I}_{\text{set}}$ and use Algorithm 0 to obtain the causal parameters $\hat{\Theta}_{\mathcal{I}}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}_{\text{opt}})$
   (d) obtain optimal intervention set: $\mathcal{I}_{\text{opt}} = \arg\min_{\mathcal{I} \in \mathcal{I}_{\text{set}}} \texttt{score-test}(\hat{\Theta}_{\mathcal{I}}(\hat{\mathcal{D}}_{\text{opt}}, \bar{h}_{\text{opt}}))$

5: **Output:** equivalence class $\hat{\mathcal{I}}_{\text{opt}}$-MEC$(\hat{\mathcal{D}}_{\text{opt}})$

Here, we start with a candidate set of DAGs and a starting intervention set that is taken to be the full set of targets $[p]$. Step 2 of the algorithm scores all DAGs in the candidate set with the number of latent variables varied between $0, 1, \ldots, h_{\max}$. Step 3 of the algorithm then computes the likelihood of each of these causal models on test data and chooses the best DAG and number of latent variables. For this selected DAG and the number of latent variables, Step 4 of the algorithm uses the second test set to choose the intervention targets.

Algorithm 4 Improving "starting point" DAGs via cross-validated *UT-LVCE*

1: **Input**: candidate DAG(s) $\tilde{\mathcal{D}}_{\text{cand}}$; training dataset; max # of latent vars. $h_{\max}$; initial intervention set $\mathcal{I} = [p]$

2: **Backward deletion to remove spurious edges**: initialize $\mathcal{D}_{\text{cand}} = \tilde{\mathcal{D}}_{\text{cand}}$; for each $\mathcal{D} \in \tilde{\mathcal{D}}_{\text{cand}}$ and $\bar{h} \leq \bar{h}_{\max}$:
   (a) supply data and $\hat{\mathcal{I}} = [p]$ to Algorithm 0 to obtain $\hat{\Theta}_{\mathcal{I}}(\mathcal{D}, \bar{h})$
   (b) let $\mathcal{D}$ be the DAG after deleting smallest edge in magnitude in $\mathcal{D}$; add $\mathcal{D}$ to $\mathcal{D}_{\text{cand}}$
   (c) repeat (a-b) until DAG $\mathcal{D}$ is empty

3: **Output:** supply $\mathcal{D}_{\text{cand}}$ to Algorithm 3 to obtain an equivalence class $\hat{\mathcal{I}}_{\text{opt}}$-MEC($\hat{\mathcal{D}}_{\text{opt}}$)

Here, as with Algorithm 3, we start with a candidate set of DAGs and a starting intervention set that is taken to be the full set of targets $[p]$. For every DAG in the candidate set and the number of latent variables varied between $0, 1, \ldots, h_{\max}$, step 2 scores the DAG and removes the weakest edge greedily until the DAG is empty. Each DAG along the path is included in the candidate set $\mathcal{D}_{\text{cand}}$. We then feed the candidate set $\mathcal{D}_{\text{cand}}$ to Algorithm 3 to obtain an output equivalence class.

Appendix N. Additional synthetic experiments

N.1 Robustness to the strength of interventions on observed and latent variables

We explore the robustness of *UT-LVCE* as a structural learning algorithm to different strengths of interventions on the observed variables and latent variables. In particular, we consider the setting described at the beginning of Section 5.1 and different configurations for the strengths of interventions on observed and latent variables:

$$\begin{aligned}
&\text{Soft interventions on the observed variables: variance of } \delta_i^e \text{ in the interval } [3, 6] \text{ for all } i \in \mathcal{I}^\star,\\
&\text{Strong interventions on the observed variables: variance of } \delta_i^e \text{ in the interval } [6, 12] \text{ for all } i \in \mathcal{I}^\star,\\
&\text{Soft interventions on the latent variables: } \xi_i^e \text{ chosen uniformly and independently from } [0.1, 1],\\
&\text{Strong interventions on the latent variables: } \xi_i^e \text{ chosen uniformly and independently from } [1, 5].
\end{aligned} \tag{65}$$

Figure 9 shows the performance of *UT-LVCE* using GES starting points as well as LRpS (Frot et al., 2019) for four settings: i) soft interventions on the observed and latent variables, i) soft interventions on the observed variables and hard interventions on the latent variables, iii) hard interventions on the observed variables and soft interventions on the latent variables, and iv) hard interventions on both the observed and the latent variables. In all settings, $|\mathcal{I}^\star| = 10$. We notice that the performance of *UT-LVCE* is robust across all the settings.

N.2 Effect of graph density and number of latent variables on performance

We explore the effect that different generating-graph densities and the number of latent variables have on the performance of *UT-LVCE* and the other methods. In particular, we consider the setting from the leftmost panel in Figure 6, and vary the edge probability of

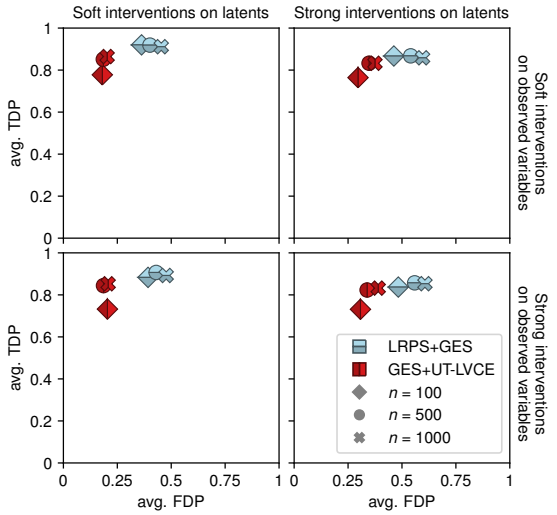


Figure 9: Performance of *UT-LVCE* as a structural learning procedure with GES initialization for different strengths of interventions on the observed and latent variables as outlined in (65).

the data generating graph $(0.10, 0.13, 0.16)$ and the number of hidden variables $(2, 3, 4, 10)$. Figure 10 shows the performance of *UT-LVCE* using GES starting points as well as LRpS (Frot et al., 2019) and Backshift (Rothenhäusler et al., 2016), for each combination of edge probability and number of latents. The performance of both methods slowly deteriorates as the number of hidden variables increases.

N.3 Goodness of GES input DAGs and performance of *UT-LVCE*

We explore how the performance of *UT-LVCE* is affected by the quality of the input GES DAGs. To that end, we consider the synthetic setting in Section 5.1.2 for both $|\mathcal{I}^\star| \in \{10, 20\}$. Among the 50 randomly generated DAGs and 5 runs, we bin the GES input solution into two categories: a category where there exists a DAG in the GES equivalence class that is a supergraph of a member of $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$, and another category where no DAG in the GES equivalence class is a supergraph of a member of $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. Figure 11 shows the performance of *UT-LVCE* for these two categories. We observe that there is a substantial improvement in the performance of *UT-LVCE* if the GES initialization contains a DAG that is a supergraph of a member of $\mathcal{I}^\star$-MEC$(\mathcal{D}^\star)$. Interestingly, we also observe that as the number of interventions increases, GES input DAGs are more likely to meet the aforementioned criteria.

N.4 Violations of Causal Dantzig and comparisons to *UT-LVCE*

We next explore the performance of Causal Dantzig (Rothenhäusler et al., 2019) under latent interventions of different strengths. Specifically, we consider soft interventions on the observed variables and both soft and harder interventions on the latent variables (see (65)). Further, we consider the size of interventions to be $|\mathcal{I}^\star| \in \{19, 20\}$. Figure 12 shows the performance of Causal Dantzig and *UT-LVCE* with GES initialization. Here, the

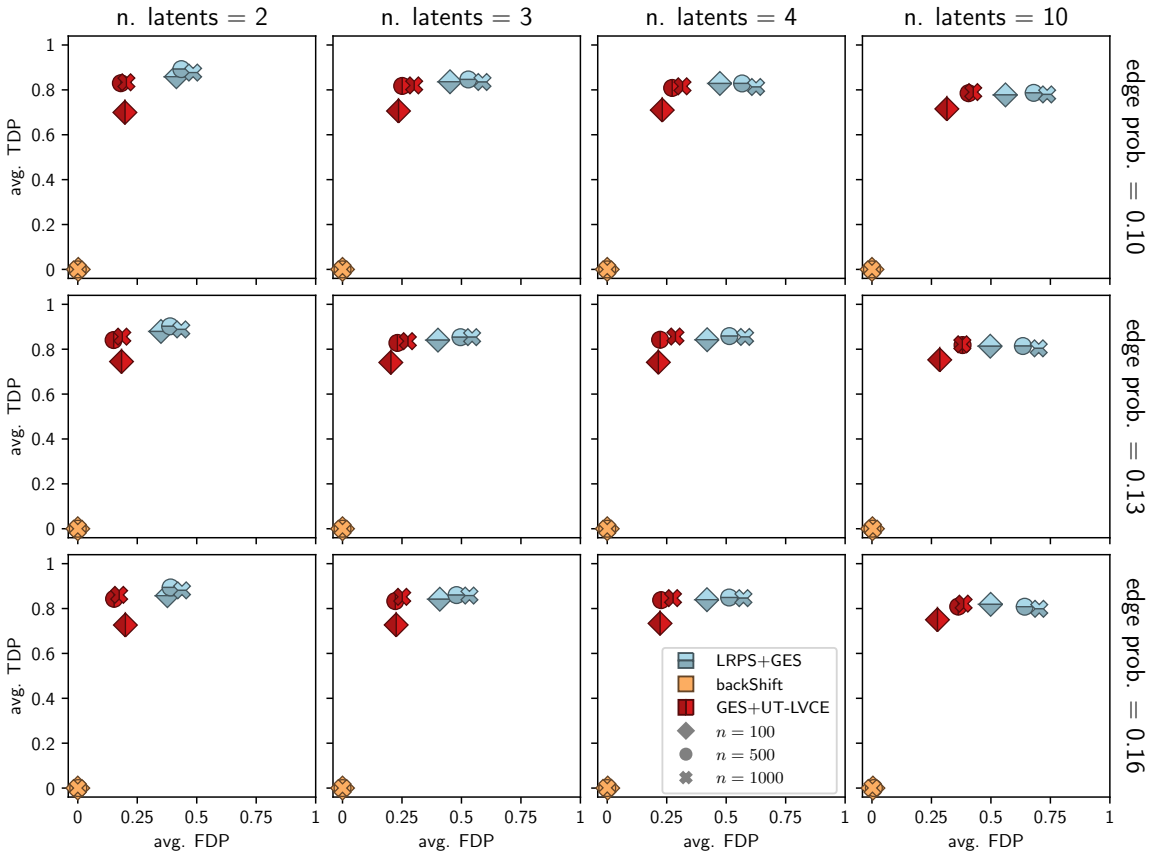


Figure 10: Performance of *UT-LVCE* as a structural learning procedure with GES initialization for different numbers of latent variables and edge probabilities in the data-generating graph.

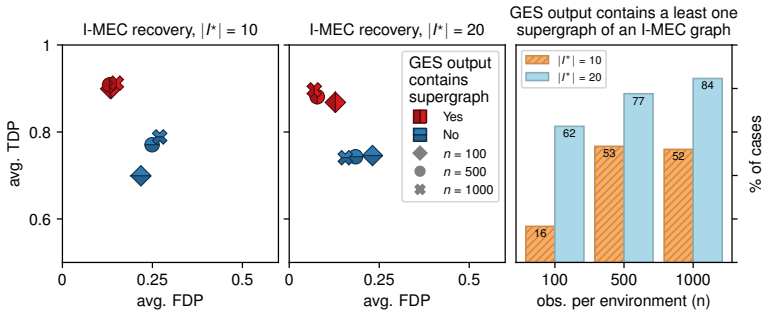


Figure 11: Left and middle: categorizing the performance of *UT-LVCE* into a setting where the GES input has a DAG that is a supergraph of a member of $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$) and the complement setting; right: proportion of instances where the GES input has a DAG that is a supergraph of a member of $\mathcal{I}^\star$-MEC($\mathcal{D}^\star$).

performance is with respect to accurate recovery of the parental sets, and the average FDP and TDP are taken over 50 DAGs and 5 runs per DAG. We observe that when there is an intervention on the target variable or when there are strong interventions on the latent variables, Causal Dantzig performs poorly as compared to *UT-LVCE*. This is consistent with the fact that the assumptions for consistency with Causal Dantzig require interventions on all observed variables except the target variable as well as no interventions on the latent variables.

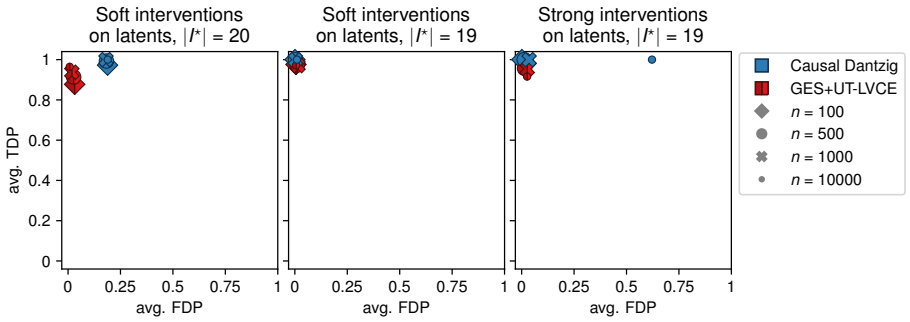
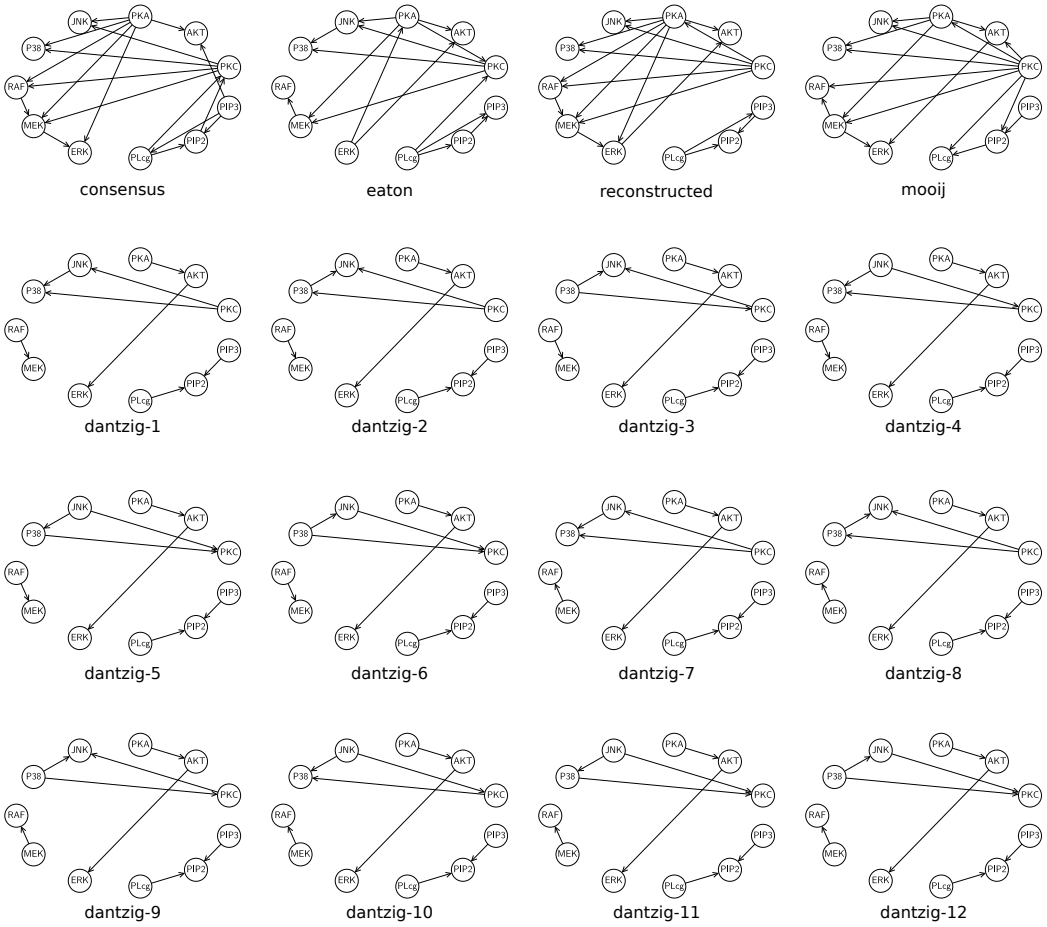


Figure 12: Performance of Causal Dantzig and *UT-LVCE* for different intervention strengths on the latent variables and for $|\mathcal{I}^\star| \in \{19, 20\}$.

Appendix O. DAGs for the protein expressions dataset



References

J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2021.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 55(6):111–119, 2009.

E Candès, X Li, Y Ma, and J Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

V. Chandrasekaran, V. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal of Optimization*, 21:572–596, 2011.

V. Chandrasekaran, P. Parillo, and A. Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40:1935–1967, 2012.

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

D. Colombo, M. Maathuis, M. Kalisch, and T. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40:294–321, 2012.

A. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9:39–77, 2021.

A. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statistical Surveys*, 4:184–231, 2010.

A. Dixit, O. Parnas, and B. Li. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167:1853–1866, 2016.

M. Drton and M. Maathius. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, 2007.

B. Frot, P. Nandy, and M. Maathius. Robust causal structure learning with hidden variables. *Journal of the Royal Statistical Society, Series B*, 81:459–487, 2019.

J. Gamella and C. Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Neural Information Processing Systems*, 2020.

J. Gamella, A. Taeb, C. Heinze-Deml, and P. Bühlmann. Characterization and greedy learning of gaussian structural causal models under unknown interventions. *arXiv 2211.14897*, 2022.

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. Learning causal structures using regression invariance. In *In: Advances in Neural Information Processing Systems*, 2017.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13: 2409–2464, 2012.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318, 2015.

C. Heinze-Deml, M. Maathuis, and N. Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018a.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6:1–35, 2018b.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:1–51, 2020.

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.

P. Kemmeren et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157:740–752, 2014.

A. McLean, L. Sanders, and W. Walter. A unified approach to mixed linear models. *Journal of American Statistical Association*, 45:54–64, 1991.

C. Meek. Causal inference and causal explanation with background knowledge. In *Uncertainty in Artificial Intelligence*, 1995.

N. Meinshausen, A. Hauser, J. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceeding of National Academy of Sciences*, 113:7361–7368, 2016.

J. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. In *Uncertainty in Artificial Intelligence*, 2013.

J. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:1–108, 2020.

P. Nandy, A. Hauser, and M. Maathius. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46:3151–3183, 2018.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78:947–1012, 2016.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

Thomas S. Richardson and Peter L. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30:962–1030, 2002.

J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

D. Rothenhäusler, C. Heinze-Deml, J. Peters, and N. Meinshausen. backshift: Learning causal cyclic graphs from unknown shift interventions. In *Neural Information Processing Systems*, 2016.

D. Rothenhäusler, P. Bühlmann, and N. Meinshausen. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions.

Annals of Statistics, 47:1688–1722, 2019.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *Journal of the Royal Statistical Society, Series B*, 83: 215–246, 2021.

D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331, 2015.

K. Sachs, O. Perez, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Cambridge: MITPress, 2000.

C. Squires, Y. Wang, and C. Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Uncertainty in Artificial Intelligence*, 2020.

A. Taeb, J. Reager, M. Turmon, and V. Chandrasekaran. A statistical graphical model of the California reservoir system. *Water Resources Research*, 53:9721–9739, 2017.

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

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41:536–567, 2013.

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge university press, 2000.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, 1991.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation based causal inference algorithms with interventions. In *Neural Information Processing Systems*, 2017.

S. Zhao, C. Gao, S. Mukherjee, and B. Engelhardt. Bayesian group factored analysis with structured sparsity. *Journal of Machine Learning Research*, 17:1–47, 2016.