Proceedings of Machine Learning Research vol 275:552–605, 2025 4th Conference on Causal Learning and Reasoning

Score matching through the roof: linear, nonlinear, and latent variables causal discovery

Francesco Montagna[*] FRANCESCO.MONTAGNA@EDU.UNIGE.IT
MaLGa, University of Genoa

Philipp M. Faller[*†] PHILIPP.FALLER@PARTNER.KIT.EDU
Karlsruhe Institute of Technology

Patrick Blöbaum BLOEBP@AMAZON.COM
Amazon

Elke Kirschbaum ELKEKI@AMAZON.COM
Amazon

Francesco Locatello FRANCESCO.LOCATELLO@ISTA.AC.AT
Institute of Science and Technology Austria (ISTA)

Editors: Biwei Huang and Mathias Drton

Abstract

Causal discovery from observational data holds great promise, but existing methods rely on strong assumptions about the underlying causal structure, often requiring full observability of all relevant variables. We tackle these challenges by leveraging the score function $\nabla \log p(X)$ of observed variables for causal discovery and propose the following contributions. First, we fine-tune the existing identifiability results with the score on additive noise models, showing that their assumption of nonlinearity of the causal mechanisms is not necessary. Second, we establish conditions for inferring causal relations from the score even in the presence of hidden variables; this result is two-faced: we demonstrate the score's potential to infer the equivalence class of causal graphs with hidden variables (while previous results are restricted to the fully observable setting), and we provide sufficient conditions for identifying direct causes in latent variable models. Building on these insights, we propose a flexible algorithm suited for causal discovery on linear, nonlinear, and latent variable models, which we empirically validate.

Keywords: Causal Discovery, Causality, Score-matching

1. Introduction

The inference of causal effects from observations holds the potential for great impact arguably in any domain of science, where it is crucial to be able to answer interventional and counterfactual queries from observational data (Peters et al., 2017; Pearl, 2009; Spirtes, 2010). Existing causal discovery methods can be categorized based on the information they can extract from the data and the assumptions they rely on (Glymour et al., 2019). Traditional causal discovery methods (e.g. PC, GES (Spirtes et al., 2000; Chickering, 2003)) are general in their applicability but limited to the inference of an equivalence class. Additional assumptions on the structural equations generating effects from the cause are, in fact, imposed to ensure the identifiability of a causal order (Shimizu

[*] Shared first co-author
[†] Part of this work was done while Philipp M. Faller was an intern at Amazon Research Tübingen.

et al., 2006; Hoyer et al., 2008; Peters et al., 2014; Zhang and Hyvärinen, 2009). As a consequence, existing methods for causal discovery require specialized and often untestable assumptions, preventing their application to real-world scenarios.

Further, the majority of existing approaches are hindered by the assumption that all relevant causes of the measured data are observed, which facilitates the interpretation of associations in the data as causal relationships. Despite the convenience of this hypothesis, it is often not met in practice, and the solutions relaxing this requirement face substantial limitations. The FCI algorithm (Spirtes, 2001) can only return an equivalence class from the data. Appealing to additional restrictions ensures the identifiability of some direct causal effects in the presence of latent variables: RCD (Maeda and Shimizu, 2020) relies on the linear non-Gaussian additive noise model, whereas CAM-UV (Maeda and Shimizu, 2021) requires nonlinear additive mechanisms. Nevertheless, the strict conditions on the structural equations hold back their applicability to more general settings.

Our paper tackles these challenges and can be put in the context of a recent line of academic research that derives a connection between the score function $\nabla \log p(X)$ and the causal graph underlying the data-generating process (Ghoshal and Honorio, 2018; Rolland et al., 2022; Montagna et al., 2023b,c,d; Sanchez et al., 2023). The use of the score for causal discovery is practically appealing, as it yields advantages in terms of scalability to high dimensional graphs (Montagna et al., 2023c) and guarantees of finite sample complexity bounds (Zhu et al., 2024). Instead of imposing assumptions that ensure strong, though often impractical, theoretical guarantees, we organically demonstrate different levels of identifiability based on the strength of the modeling hypotheses, always relying on the score function to encode all the causal information in the data. Starting from results of Spantini et al. (2018) and Lin (1997), we show how constraints on the Jacobian of the score $\nabla^2 \log p(X)$ can be used to characterize the Markov equivalence class of causal models with hidden variables. Previous works exploit this connection in case of fully observable causal models (Montagna et al., 2023c; Liu et al., 2024). Further, we prove that the score function identifies the causal direction of additive noise models, with minimal assumptions on the causal mechanisms. This extends the previous findings of Montagna et al. (2023d), limited by the assumption of nonlinearity of the causal effects. On these results, we build the main contributions of our work, enabling the identification of direct causal effects in hidden variable models.

Our main contributions are as follows: *(i)* We present conditions for the identifiability of direct causal effects with the score in the case of latent variables models. *(ii)* We propose AdaScore (Adaptive Score-matching-based causal discovery), a flexible algorithm for causal discovery based on score matching estimation of $\nabla \log p(X)$ (Hyvärinen, 2005). Based on the user's belief about the plausibility of several modeling assumptions on the data, AdaScore can output a Markov equivalence class, a directed acyclic graph, or a mixed graph, accounting for the presence of unobserved variables. To the best of our knowledge, the broad class of causal models handled by our method is unmatched by other approaches in the literature. Our main contributions are presented in Sections 4.2, 4.3 and 5. The preliminary Sections 3 and 4.1 summarise and generalize the existing theory connecting causal discovery and the score function, and Section 2 introduces the formalism of structural causal models, with and without latent variables.

2. Model definition

In this section, we introduce the formalism of structural causal models (SCMs), for the cases with and without hidden variables.

2.1. Causal model with observed variables

Let X be a set of random variables in $\mathbb{R}$ defined according to the set of structural equations

$$X_i \coloneqq f_i(X_{\mathrm{PA}_i^{\mathcal{G}}}, N_i), \quad \forall i = 1, \dots, k. \tag{1}$$

$N_i \in \mathbb{R}$ are mutually independent random variables with strictly positive density, known as *noise* or error terms. The function f_i is the *causal mechanism* mapping the set of *direct causes* $X_{\mathrm{PA}_i^{\mathcal{G}}}$ of X_i and the noise term N_i, to X_i's value. A structural causal model (SCM) is defined as the tuple $(X, N, \mathcal{F}, \mathbb{P}_N)$, where $\mathcal{F} = (f_i)_{i=1}^k$ is the set of causal mechanisms, and $\mathbb{P}_N$ is the joint distribution with the density p_N over the noise terms $N \in \mathbb{R}^k$. We define the *causal graph* $\mathcal{G}$ as a directed acyclic graph (DAG) with nodes $X = \{X_1, \dots, X_k\}$, and the set of edges defined as $\{X_j \to X_i : X_j \in X_{\mathrm{PA}_i^{\mathcal{G}}}\}$, such that $\mathrm{PA}_i^{\mathcal{G}}$ are the indices of the parent nodes of X_i in the graph $\mathcal{G}$. (In the remainder of the paper, we adopt the following notation: given a set of random variables $Y = \{Y_1, \dots, Y_n\}$ and a set of indices $Z \subset \mathbb{N}$, then $Y_Z = \{Y_i | i \in Z, Y_i \in Y\}$.)

Under this model, the probability density of X satisfies the *Markov factorization* (e.g. Peters et al. (2017) Proposition 6.31):

$$p(x) = \prod_{i=1}^{k} p(x_i | x_{\mathrm{PA}_i^{\mathcal{G}}}), \tag{2}$$

where we adopt the convention of lowercase letters referring to realized random variables, and use p to denote the density of different random objects, when the distinction is clear from the argument. In the remainder of the paper, we assume that *faithfulness* is satisfied (Pearl, 2009; Uhler et al., 2012) (Definition 12 in the appendix). Together with the *global Markov condition* (implied by Equation (2), see e.g. Peters et al. (2017) Proposition 6.22), this means that probabilistic and graphical statements of conditional independence are equivalent, such that for $\{X_i, X_j\} \subseteq X$ and $X_Z \subseteq X \setminus \{X_i, X_j\}$

$$X_i \perp\!\!\!\perp X_j | X_Z \iff X_i \perp\!\!\!\perp_{\mathcal{G}}^{d} X_j | X_Z, \tag{3}$$

where $(\cdot \perp\!\!\!\perp \cdot | \cdot)$ denotes probabilistic conditional independence of X_i, X_j given X_Z, and $(\cdot \perp\!\!\!\perp_{\mathcal{G}}^{d} \cdot | \cdot)$ is the notation for *d-separation*, a criterion of conditional independence defined on the graph $\mathcal{G}$ (Definition 8 of the appendix).

The above model assumes that there aren't any unobserved causes of variables in X, other than the noise terms in N. As we are interested in distributions with potential hidden variables, we will now generalize our model to represent data-generating processes that may involve latent causes.

2.2. Causal model with unobserved variables

Under the model (1), we consider the case where the set of variables X is partitioned into the disjoint subsets of *observed* random variables $V = \{V_1, \dots, V_d\}$ and *unobserved* (or *latent*) random variables $U = \{U_{d+1}, \dots, U_p\}$. We assume that the following set of structural equations is satisfied:

$$V_i \coloneqq f_i(V_{\mathrm{PA}_i^{\mathcal{G}}}, U^i, N_i), \qquad \forall i = 1, \dots, d, \tag{4}$$

$$U_i \coloneqq f_i(X_{\mathrm{PA}_i^{\mathcal{G}}}, N_i), \qquad \forall i = d+1, \dots, p, \tag{5}$$

where U^i stands for the set of unobserved parents of V_i, and $V_{\mathrm{PA}_i^{\mathcal{G}}} = \{V_k | k \in \mathrm{PA}_i^{\mathcal{G}}, V_k \in V\}$ are the observed parents of V_i. Some of the causal relations and the conditional independencies implied

by the set of equations (4) can be summarized in a graph obtained as a *marginalization* of the DAG $\mathcal{G}$ onto the observable nodes V. For the next definition, we advise the reader to be comfortable with the notions of *ancestors* (Definition 5) and *inducing paths* (Definition 6) in DAGs.

Definition 1 (Marginal graph, Zhang (2008a)) *Let $X = V \cup U$, V and U disjoint, and $\mathcal{G}$ be a DAG over X. The following construction gives the marginal graph $\mathcal{M}_V^{\mathcal{G}}$, with nodes V and edges found as follows:*

- *pair of nodes V_i, V_j are adjacent in the graph $\mathcal{M}_V^{\mathcal{G}}$ if and only if there is an inducing path between them relative to U in $\mathcal{G}$;*
- *for each pair of adjacent nodes V_i, V_j in $\mathcal{M}_V^{\mathcal{G}}$, orient the edge as $V_i \to V_j$ if V_i is an ancestor of V_j in $\mathcal{G}$, else orient it as $V_i \leftrightarrow V_j$.*

We define the map $\mathcal{G} \mapsto \mathcal{M}_V^{\mathcal{G}}$ as the marginalization of the DAG $\mathcal{G}$ onto V, the observable nodes.

The graph resulting from the above construction is a maximal ancestral graph (MAG, Definition 7), hence we will often refer to it as the *marginal MAG* of $\mathcal{G}$. Intuitively, edges denote dependencies that cannot be removed by conditioning on any of the observed variables; in particular, if the edge is directed it denotes an ancestorship relation.

In the case of DAGs, d-separation encodes the probabilistic conditional independence relations between the variables of X in the graph $\mathcal{G}$, as explicit by Equation (3). Such notion of graphical separation has a natural generalization to maximal ancestral graphs, known as *m-separation* (Definition 8 of the appendix). Zhang (2008a) shows that *m-separation* and *d-separation* are in fact equivalent (see Lemma 11 of the appendix), such that given $\{V_i, V_j\} \subset V$ and $V_Z \subset V \setminus \{V_i, V_j\}$, the following holds:

$$V_i \perp\!\!\!\perp_{\mathcal{G}}^{d} V_j | V_Z \iff V_i \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} V_j | V_Z, \tag{6}$$

where $(\cdot \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} \cdot \mid \cdot)$ denotes *m-separation* relative to the graph $\mathcal{M}_V^{\mathcal{G}}$. Just like with DAGs, MAGs that imply the same set of conditional independencies define an equivalence class. Usually, the common structure of these graphs is represented by partial ancestral graphs (PAGs, Definition 10 of the appendix). We use $\mathcal{P}_{\mathcal{M}_V^{\mathcal{G}}}$ to denote the PAG relative to $\mathcal{M}_V^{\mathcal{G}}$.

Problem definition. In this work, our goal is to provide theoretical guarantees for the identifiability of the Markov equivalence class of the marginal graph $\mathcal{M}_V^{\mathcal{G}}$ and its direct causes with the score, where variables V_i are defined according to Equation (4).

Without further assumptions on the data-generating process, we can identify the graph $\mathcal{M}_V^{\mathcal{G}}$ only up to its partial ancestral graph. This information is encoded in the Jacobian of the score, as discussed in the next section.

3. A score-matching-based criterion for m-separation

In this section, we show that for $V \subseteq X$ generated according to Equation (4) the Hessian matrix of $\log p(V)$ identifies the equivalence class of the marginal MAG $\mathcal{M}_V^{\mathcal{G}}$. It has already been proven that

cross-partial derivatives of the log-likelihood are informative about a set of conditional independence relationships between random variables: Spantini et al. (2018) (Lemma 4.1) and previously Lin (1997) show that, given $V_Z \subseteq X$ such that $\{V_i, V_j\} \subseteq V_Z$, then

$$\frac{\partial^2}{\partial V_i \partial V_j} \log p(V_Z) = 0 \iff V_i \perp\!\!\!\perp V_j | V_Z \setminus \{V_i, V_j\}. \tag{7}$$

Equation (3) resulting from faithfulness and the directed global Markov property implies that this expression can be used as a test of graphical separation to identify the Markov equivalence class of the graph $\mathcal{M}_V^{\mathcal{G}}$, as commonly done in constraint-based causal discovery (for reference, see e.g. Section 3 in Glymour et al. (2019)). This result generalizes Lemma 1 of Montagna et al. (2023c), where it is used to define constraints to infer edges in the causal structure without latent variables, under the assumption of nonlinear models with additive Gaussian noise.

Proposition 2 (Corollary of Spantini et al. (2018)) [1] *Let V be a set of random variables with strictly positive density generated according to the structural equations (4). For each set $V_Z \subseteq V$ of nodes in $\mathcal{M}_V^{\mathcal{G}}$ such that $\{V_i, V_j\} \subseteq V_Z$, then the following holds:*

$$\frac{\partial^2}{\partial V_i \partial V_j} \log p(V_Z) = 0 \iff V_i \perp\!\!\!\perp^m_{\mathcal{M}_V^{\mathcal{G}}} V_j | V_Z \setminus \{V_i, V_j\}.$$

The result of Proposition 2 presents an alternative way to assess graphical separation in *constraint-based* approaches to causal discovery: the equivalence class of the graph $\mathcal{M}_V^{\mathcal{G}}$ can be identified using the cross-partial derivatives of the log-likelihood to characterize conditional independencies between variables, much in the spirit of the Fast Causal Inference (FCI) algorithm (Spirtes, 2001). Identifying the Markov equivalence class is the most we can hope to achieve without further restrictions on the hypothesis class. As we will see in the next section, the score function can also help leverage additional restrictive assumptions on the causal mechanisms of Equation (4) to identify direct causal effects.

4. A theory of identifiability from the score

In this section, we show that, under additional assumptions on the data-generating process, we can identify some direct causal relations with the score even in the case of latent variable models.

As a preliminary step before diving into causal discovery with hidden nodes, we show how the properties of the score function identify edges in directed acyclic graphs, that is in the absence of unmeasured variables (when $U = \emptyset$ and $\mathcal{G} = \mathcal{M}_V^{\mathcal{G}}$). The goal of the next section is two-sided: first, it introduces the fundamental ideas connecting the score function to causal discovery that also apply to hidden variable models, second, it extends the existing theory of causal discovery with score matching to additive noise models with both linear and nonlinear mechanisms.

1. In their Lemma 4.1 Spantini et al. (2018) provides the connection between vanishing cross-partial derivatives of the log-likelihood and conditional independence of random variables. Note that this result does not depend on the assumption of a generative model, thus holding beyond the set of structural equations (4) and (5). Our result exploits their finding to the case when observations are generated according to a causal model with potentially latent variables.

4.1. Warm up: identifiability without latent confounders

In this section, we summarise and extend the theoretical findings presented in Montagna et al. (2023d), where the authors show how to derive constraints on the score function that identify the causal order of the DAG $\mathcal{G}$ where all the variables in the set X are observed. Define the structural relations of (1) as:

$$X_i \coloneqq h_i(X_{\mathrm{PA}_i^{\mathcal{G}}}) + N_i, i = 1, \dots, k, \tag{8}$$

with three times continuously differentiable mechanisms h_i, noise terms centered at zero, and strictly positive density p_X. Further, we assume that the SCM on X is a *restricted* additive noise model (Definition 16), which is necessary to ensure identifiability. Given the Markov factorization of Equation (2), and the change of variable formula for densities, the components of the score function $\nabla \log p(x)$ are:

$$\partial_{X_i} \log p(x) = \partial_{N_i} \log p(n_i) - \sum_{j \in \mathrm{CH}_i^{\mathcal{G}}} \partial_{X_i} h_j(x_{\mathrm{PA}_j^{\mathcal{G}}}) \partial_{N_j} \log p(n_j), \tag{9}$$

where $\mathrm{CH}_i^{\mathcal{G}}$ denotes the set of children of node X_i. We observe that if a node X_s is a *sink*, i.e. a node satisfying $\mathrm{CH}_s^{\mathcal{G}} = \emptyset$, then the summation over the children vanishes, implying $\partial_{X_s} \log p(x) = \partial_{N_s} \log p(n_s)$. The key point is that the score component of a sink node is a function of its structural equation noise term, such that one could learn a consistent estimator of $\partial_{X_s} \log p_X$ from a set of observations of the noise term N_s. Given that, in general, one has access to X samples rather than observations of the noise random variables, authors in Montagna et al. (2023d) show that N_s of a sink node can be consistently estimated from i.i.d. realizations of X. For each node $X_1, \dots, X_k$, we define the quantity:

$$R_i \coloneqq X_i - \mathbb{E}[X_i | X_{\setminus X_i}], \tag{10}$$

where $X_{\setminus X_i}$ are the random variables in the set $X \setminus \{X_i\}$. $\mathbb{E}[X_i | X_{\setminus X_i}]$ is the optimal least squares predictor of X_i from all the remaining nodes in the graph, and R_i is the regression residual. For a sink node X_s, the residual satisfies:

$$R_s = N_s, \tag{11}$$

which can be seen by rewriting $\mathbb{E}[X_s | X_{\setminus X_s}] = h_s(X_{\mathrm{PA}_s^{\mathcal{G}}}) + \mathbb{E}[N_s | X_{\mathrm{DE}_s^{\mathcal{G}}}, X_{\mathrm{ND}_s^{\mathcal{G}}}] = h_s(X_{\mathrm{PA}_s^{\mathcal{G}}}) + \mathbb{E}[N_s]$, where $X_{\mathrm{DE}_s^{\mathcal{G}}}$ and $X_{\mathrm{ND}_s^{\mathcal{G}}}$ denotes the descendants and non-descendants of X_s, respectively. Equations (9) and (11) together imply that the score $\partial_{N_s} \log p(N_s)$ is a function of R_s, such that it is possible to find a consistent approximator of the score of a sink from observations of R_s.

Proposition 3 (Generalization of Lemma 1 in Montagna et al. (2023d)) *Let X be a set of random variables, generated by a restricted additive noise model (Definition 16) with structural equations (8), and let $X_j \in X$. Then:*

$$X_j \text{ is a sink} \iff \mathbb{E}\left[\left(\mathbb{E}\left[\partial_{X_j} \log p(X) \mid R_j\right] - \partial_{X_j} \log p(X)\right)^2\right] = 0. \tag{12}$$

Our result generalizes Lemma 1 in Montagna et al. (2023d), as they assume X generated by an identifiable additive noise model with nonlinear mechanisms, which we can replace by the weaker hypothesis of *restricted* additive noise model. These findings are not surprising, in light of previous literature connecting the precision matrix and the causal structure of linear SCMs (Ghoshal and Honorio, 2018), as we discuss in Appendix D.1. Inspired by these results, in the next section we demonstrate the identifiability via the score of direct causal effects between a pair of variables in the marginal MAG $\mathcal{M}_V^{\mathcal{G}}$ when $U \neq \emptyset$.

4.2. Identifiability in the presence of latent confounders

We now introduce our main theoretical result, that is: given a pair of nodes V_i, V_j that are adjacent in the graph $\mathcal{M}_V^{\mathcal{G}}$ with $U \neq \emptyset$, we can use the score function to identify the presence of a direct causal effect between V_i and V_j which can not be detected when inference is limited to the PAG $\mathcal{P}_{\mathcal{M}_V^{\mathcal{G}}}$ (representing the Markov equivalence class). Given that the causal model of Equation (4) ensures identifiability only up to the equivalence class, we need additional restrictive assumptions. In particular, we enforce an additive noise model with respect to both the observed and unobserved noise variables. This corresponds to an additive noise model on the observed variables with the noise terms recentered by the latent causal effects.

Assumption 1 (SCM assumptions) *The set of structural equations of the observable variables specified in (4) is now defined as:*

$$V_i \coloneqq f_i(V_{\mathrm{PA}_i^{\mathcal{G}}}) + g_i(U^i) + N_i, \forall i = 1, \dots, d, \tag{13}$$

assuming the mechanisms f_i to be of class $\mathcal{C}^3(\mathbb{R}^{|V_{\mathrm{PA}_i^{\mathcal{G}}}|})$, and mutually independent noise terms with strictly positive density function. The N_i's are assumed to be non-Gaussian when f_i is linear in some of its arguments.

Crucially, our hypothesis is weaker than those required by two state-of-the-art approaches, CAM-UV (Maeda and Shimizu, 2021) and RCD (Maeda and Shimizu, 2020): CAM-UV assumes a Causal Additive Model (CAM) with structural equations with nonlinear mechanisms in the form $V_i \coloneqq \sum_{k \in \mathrm{PA}_i^{\mathcal{G}}} f_{ik}(V_k) + \sum_{U_k^i} g_{ik}(U_k^i) + N_i$, and RCD requires an additive noise model with linear effects of both the latent and observed causes. Thus, our model encompasses and extends the nonlinear and linear settings of CAM-UV and RCD, such that the theory developed in the remainder of the section is valid for a broader class of causal models.

Our first step is rewriting the structural relations in (13) as:

$$\begin{aligned} V_i &\coloneqq f_i(V_{\mathrm{PA}_i^{\mathcal{G}}}) + \tilde{N}_i, \\ \tilde{N}_i &\coloneqq g_i(U^i) + N_i, \forall i = 1, \dots, d, \end{aligned} \tag{14}$$

which provides an additive noise model in the form of (8). Next, we define the following regression residuals for any node V_k in the graph $\mathcal{M}_V^{\mathcal{G}}$:

$$R_k(V_Z) \coloneqq V_k - \mathbb{E}[V_k \mid V_{Z \setminus \{k\}}], \tag{15}$$

where $V_{Z \setminus \{k\}}$ denotes the set of random variables $V_Z \setminus \{V_k\}$.

Given these definitions, we are ready to show that the score can identify the presence of direct causal effects between pairs of observed variables in V which can not be detected by conditional independence testing (e.g., as in FCI). In particular, if the mechanisms are nonlinear, direct parents in the DAG can be identified, while in the linear case identifiability is limited to ancestorship relations.

4.2.1. IDENTIFIABILITY OF DIRECTED EDGES

Let V_i, V_j be adjacent nodes in the PAG $\mathcal{P}_{\mathcal{M}_V^{\mathcal{G}}}$: we want to investigate when a direct causal effect $V_i \in V_{\mathrm{PA}_j^{\mathcal{G}}}$ can be identified from the score. Consider the set of variables $V_Z = V_{\mathrm{PA}_j^{\mathcal{G}}} \cup \{V_j\}$ and

the graph $\mathcal{M}^{\mathcal{G}}_{V_Z}$ resulting from the marginalization of $\mathcal{G}$ on V_Z. By Equation (14), we have that:

$$V_j \coloneqq f_j(V_{\mathrm{PA}^{\mathcal{G}}_j}) + \tilde{N}_j, \quad \tilde{N}_j \coloneqq g_j(U^j) + N_j.$$

The key observation is that for $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp^{\mathcal{G}}_d U^j$, then $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp \tilde{N}_j$ (e.g. K. Blitzstein and Hwang (2019), Theorem 3.8.5): this allows considering $\tilde{N}_j$ as an *exogenous* noise term independent of other variables in the structural equation of V_j. By Equation (15) this implies that

$$R_j(V_Z) = \tilde{N}_j - \mathbb{E}[\tilde{N}_j], \tag{16}$$

where we use $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp \tilde{N}_j$ to write $\mathbb{E}[\tilde{N}_j|V_{\mathrm{PA}^{\mathcal{G}}_j}] = \mathbb{E}[\tilde{N}_j]$. Moreover, $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp \tilde{N}_j$ implies that $p(V_j|V_{\mathrm{PA}^{\mathcal{G}}_j}) = p(\tilde{N}_j)$, and from simple manipulations we can show that $\partial_{V_j} \log p(V_Z) = \partial_{N_j} \log p(\tilde{N}_j)$. We conclude that in analogy to the case without latent variables the score $\partial_{V_j} \log p(V_Z)$ is a function of $\tilde{N}_j$, the error term in the additive noise model of Equation (14). Then the score of V_j can be consistently predicted from observations of the residual $R_j(V_Z)$, which is the sufficient and necessary condition to discover the direct causes (parents and ancestors) of V_j with the score.

Proposition 4 *Let $X = V \cup U$, V and U disjoint, be generated by a restricted additive noise model with causal graph $\mathcal{G}$. Let V satisfying the set of structural equations (13), and f_i nonlinear for each $i = 1, \dots, d$. Consider V_i, V_j adjacent in $\mathcal{P}_{\mathcal{M}^{\mathcal{G}}_V}$, the PAG relative to marginalization $\mathcal{M}^{\mathcal{G}}_V$. Further, assume that for each subset $V_Z \subseteq V$ the score component $\partial_{V_j} \log p(V_Z)$ is non-constant for uncountable values of V_Z. Then,*

$$\begin{aligned} &\exists V_Z \subseteq V, \{V_i, V_j\} \in V_Z, \text{s.t. } \mathbb{E}\left[\left(\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]\right)^2\right] = 0 \\ &\iff V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp^{d}_{\mathcal{G}} U^j \wedge V_i \in V_{\mathrm{PA}^{\mathcal{G}}_j}. \end{aligned} \tag{17}$$

If f_i linear for each $i = 1, \dots, d$ we have

$$\begin{aligned} &\exists V_Z \subseteq V, \{V_i, V_j\} \in V_Z, \text{s.t. } \mathbb{E}\left[\left(\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]\right)^2\right] = 0 \\ &\iff V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp^{d}_{\mathcal{G}} U^j \wedge V_i \in V_{\mathrm{AN}^{\mathcal{G}}_j}. \end{aligned} \tag{18}$$

The proof is found in Appendix C.3. Here, we present the intuition about the content of the proposition. Given two adjacent nodes V_i, V_j in the PAG, they must be graphically connected by one edge between $\circ\!\!\rightarrow, \circ\!\!-\!\!\circ, \leftrightarrow$ or $\rightarrow$, the latter denoting ancestral relation in the DAG (which may not be a direct parent). Equation (17) provides the condition under which a parent-child relation can be identified in place of the less informative PAG edges: in particular, for nonlinear additive noise models, that is when the independence $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp \tilde{N}_j$ implied by $V_{\mathrm{PA}^{\mathcal{G}}_j} \perp\!\!\!\perp^{d}_{\mathcal{G}} U^j$ allows to interpret $\tilde{N}_j$ as an *exogenous* noise term independent of other variables in the structural Equation (14) of V_j. This condition is necessary: given an active path such that $V_{\mathrm{PA}^{\mathcal{G}}_j} \not\perp\!\!\!\perp^{d}_{\mathcal{G}} U^j$, the score could not identify a direct causal effect $V_i \in V_{\mathrm{PA}^{\mathcal{G}}_j}$. We remark on the novelty of our theory compared to Maeda and Shimizu (2021), the nearest neighbor to our work in the literature: they demonstrate identifiability of direct parents under the assumptions of additive and nonlinear mechanisms (i.e.

$V_i := \sum_{k \in \mathrm{PA}_i^{\mathcal{G}}} f_{ik}(V_k) + \sum_{U_k^i} g_{ik}(U_k^i) + N_i)$, which is more restrictive than our modeling hypothesis. In the case of linear mechanisms, we can only identify $V_i \in V_{\mathrm{AN}_j^{\mathcal{G}}}$, which subsumes the findings of Maeda and Shimizu (2020), as they do not consider the case of unobserved mediators. See Appendix D.3 for more details on the relation of these results.

Connections to FCI. In Appendix D.2 we present a thorough analysis of the relation between the identifiability guarantees of Proposition 4 and those provided by the FCI algorithm. Intuitively, under standard assumptions, FCI discovers all aspects of the causal structure that are uniquely determined by facts of probabilistic dependence and independence (Zhang, 2008c). Adding the assumption of an additive noise model, as in our case, allows further identifying causal directions that are not distinguishable purely by probabilistic (in)dependence facts.

Connections to local causal discovery. We note that the result of Proposition 4 has an important application in the domain of *local causal discovery*, which is based on the idea that the identification of causal effects of the covariates on the response under interventions only requires knowledge of the local causal structure around the treatment (Maathuis et al., 2008). Then, the discovery of parental relations between a targeted subset of nodes, rather than on the entire set of observed variables, may be sufficient for some downstream tasks in causal inference. The benefits of this approach are clear in terms of computational efficiency. Proposition 4 shows that for any pair of nodes, the score function can discover a parent-child (or ancestral) relationship while being agnostic of the structure of the other nodes in the graph. This is in contrast with existing approaches to causal discovery with score matching that are concerned with the inference of the global topological order, which might be unnecessarily computationally expensive. While we do not elaborate on this connection in the following algorithmic section, we believe this to be an important potential application building on our results.

We have established theoretical guarantees of identifiability for additive noise models, even in the presence of hidden variables: we find that, for nonlinear models, the score function is a means for the identifiability of all direct parental relations that are not influenced by unobserved variables; all the remaining arrowheads of the edges in the graph $\mathcal{M}_V^{\mathcal{G}}$ are identified no better than in the equivalence class. For linear models, identifiability is guaranteed for ancestral relations. Based on these insights, we propose AdaScore, a score matching-based algorithm for the inference of Markov equivalence classes and direct causal effects, in the presence of latent variables.

4.3. A score-matching-based algorithm for causal discovery

Building on our theory, we propose AdaScore, a generalization of NoGAM to linear and nonlinear additive noise models with latent variables. The main strength of our approach is the adaptivity of its theoretical guarantees for this broad class of structural assumptions, as illustrated in Table 1. In practice, we design our method to be flexible in its output: based on the user's belief about the plausibility of several modeling assumptions on the data, AdaScore can output an equivalence class (using the condition of Proposition 2 to infer conditional independence in an FCI-like algorithm), a directed acyclic graph (as in NoGAM), or a mixed graph, accounting for the presence of unobserved variables. We now describe the version of our algorithm whose output is a mixed graph, where we rely on score matching estimation of the score and its Jacobian (Appendix E.3).

At an intuitive level, we find unoriented edges using Proposition 2, i.e. checking for dependencies in the form of non-zero entries in the Jacobian of the score via hypothesis testing on the mean, and find the edges' directions via the condition of Proposition 4, i.e. by estimating residuals of each node X_i (via kernel ridge-regression, as we motivate in Appendix D.4) and checking whether they can correctly predict the i-th score entry (the vanishing mean squared errors are verified by hypothesis test of independent residuals, see Appendix E.3). It would be tempting to simply find the skeleton (i.e. the graphical representation of the constraints of an equivalence class) first via the well-known adjacency search of the FCI algorithm and then iterate through all neighborhoods of all nodes to orient edges using Proposition 4. This would be prohibitively expensive.

Instead, we propose an alternative solution: exploiting the fact that some nodes may not be influenced by latent variables, we first use Proposition 3 to find sink nodes that are not affected by latents (using hypothesis testing to find vanishing mean squared error in the score predictions from the residuals), in the spirit of the NoGAM algorithm. If there is such a sink, we search all its adjacent nodes via Proposition 2 (plus an optional pruning step for better accuracy, Appendix E.3), and orient the inferred edges towards the sink. Else, if no sink can be found, we pick a node in the graph and find its neighbors by Proposition 2, orienting its edges using the condition in Proposition 4 (score estimation by residuals under latent effects). This way, we get an algorithm that is polynomial in the best case (Appendix E.4). Details on AdaScore are provided in Appendix E, while a pseudo-code summary is provided in the Algorithm 1 box.

Algorithm 1: Simplified pseudo-code of AdaScore

```
while nodes remain do
    Find sink candidate using Proposition 4
    if Proposition 4 finds a sink and output is mixed graph
    or output is DAG then
        Add edges from adjacent nodes to sink
    else
        Pick some remaining node V_i ∈ V
        Prune neighbourhood of V_i using Proposition 2
        if output is not PAG then
            Orient edges adjacent to V_i using Proposition 4
        end
        if V_i has outgoing directed edge to some V_j ∈ V then
            continue with V_j
        else
            Remove V_i from remaining nodes
        end
    end
end
Prune remaining potential edges using Proposition 2
if output is PAG then
    Do PAG orientations using Proposition 2
end
```

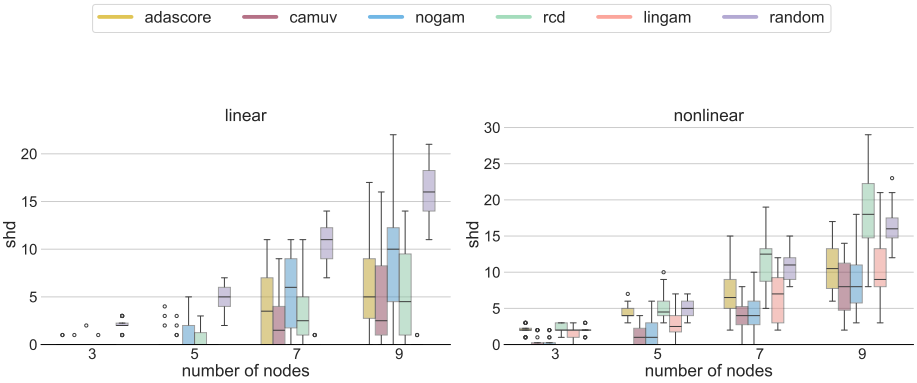
5. Experiments

The code for all experiments is available under https://github.com/amazon-science/causal-score-matching. We use the `causally`[2] Python library Montagna et al. (2023a) to generate synthetic data with known ground truths, created as Erdös-Rényi sparse and dense graphs, respectively with probability of edge between pair of nodes equals 0.3 and 0.5. We sample the data according to linear and nonlinear mechanisms with additive noise, where the nonlinear functions are parametrized by a neural network with random weights, a common approach in the literature (Montagna et al., 2023b,a; Lippe et al., 2022; Ke et al., 2023; Brouillard et al., 2020). Noise terms are sampled from a uniform distribution in the $[-2, 2]$ range. We introduce hidden variables by randomly picking two nodes and dropping the corresponding column from the data matrix (datasets with no confounding effect are discarded and re-sampled to ensure that experiments are meaningful). See Appendix F.1 for further details on the data generation. As metric, we consider the structural Hamming distance (SHD) (Tsamardinos et al., 2006; Triantafillou and Tsamardinos, 2016), a simple count of the number of incorrect edges, where missing and wrongly directed edges count as one error. We fix the level of the hypothesis tests of AdaScore to 0.05, which is a common choice in the absence of prior knowledge. We compare AdaScore to NoGAM, CAM-UV, RCD, and DirectLiNGAM, whose assumptions are detailed in Table 1. We also adopt a random baseline, described in Appendix F.3. In the main manuscript, we consider inference of *sparse* graphs, where each dataset contains 1000 observations (boxplots are obtained sampling datasets with 20 different random seeds). Additional experiments including those on dense networks are presented in Appendix G. Our synthetic data are standardized by their empirical variance to remove shortcuts in the data (Montagna et al., 2023b; Reisach et al., 2021). We limit our synthetic experiments to graphs with 9 nodes, as we empirically observed that AdaScore and CAM-UV struggle to scale with the number of variables. A thorough analysis of the elapsed computational time is provided in Appendix G.6.

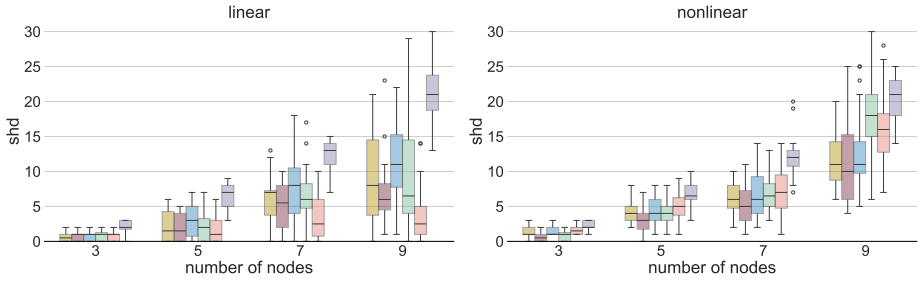
Further, we show results for three real and pseudo-real benchmark datasets. Namely, a biological dataset on cell signaling (Sachs et al., 2005), the `AutoMPG` dataset[3] concerning fuel consumption in cars (Bache and Lichman, 2013), where we use the causal ground truth given by (Wang and Mueller, 2017), and the synthetic FMRI dataset `Sim2` (Smith et al., 2011). For each dataset, we randomly pick two variables and drop them to introduce hidden variables. For `Sachs` and `Sim2` we also randomly[4] select 1000 samples. The experiments are repeated 20 times.

Discussion. Our experimental results on models without latent variables of Figure 1(*a*) show that when causal relations are linear, AdaScore can recover the causal graph with accuracy that is comparable with all the other benchmarks, with the exception of DirectLiNGAM. On nonlinear data, AdaScore outperforms RCD accuracy, while being slightly worse than CAM-UV and NoGAM. When inferring under latent causal effects, Figure 1(*b*), our method is comparable to CAM-UV and NoGAM, and appears to be preferrable to RCD, which degrade its performance with scale and is no better than random with 9 nodes. Finally, on the real benchmarks of Figure 2, AdaScore presents promising results compared to the other methods. On cell signaling data, our algorithm emerges as the best option, while it retains competitive performance on fuel consumption and FMRI data. Overall, we observe that our method is robust to a variety of structural assumptions, with accuracy

2. https://causally.readthedocs.io/en/latest/
3. Some of the features in this dataset are not continuous. For simplicity, we still treat them as if they were.
4. `Sim2` is subdivided into samples from different (synthetic) subjects. We picked our subsample across all subjects.

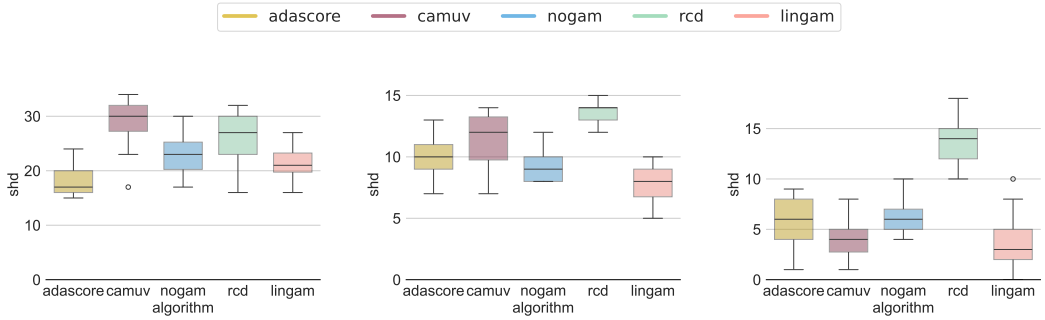


(*a*) Fully observable model



(*b*) Latent variables model

Figure 1: Empirical results on sparse graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the SHD accuracy (lower is better). We note that Adascore is comparable to the other methods in all settings (except for DirectLiNGAM on linear data), and always significantly better than random.



(*a*) Cell signaling data (*b*) Fuel consumption data (*c*) FMRI data

Figure 2: Empirical results on real and pseudo-real datasets from Sachs et al. (2005), Bache and Lichman (2013) , and Smith et al. (2011). We report the SHD accuracy (lower is better). AdaScore has the lowest SHD among all tested methods on the gene dataset and appears to be competitive compared to other methods on the fuel consumption and FMRI data.

Table 1: Algorithms of our experiments. Each cell denotes denotes whether the method has (✓) or has not (✗) guarantees of identifiability under the condition specified in the related row.

	CAM-UV	RCD	NoGAM	DirectLiNGAM	AdaScore
Linear additive noise model	✗	✓	✗	✓	✓
Nonlinear additive noise model	✗	✗	✓	✗	✓
Nonlinear CAM	✓	✗	✓	✗	✓
Latent variables effects	✓	✓	✗	✗	✓
Output	Mixed	Mixed	DAG	DAG	Mixed

that is comparable and sometimes better than competitors. We remark that although AdaScore does not clearly outperform the other baselines, its broad theoretical guarantees of identifiability are not matched by any available method in the literature; this makes it an appealing option for inference in realistic scenarios that are hard to investigate with synthetic data, where the structural assumptions of the causal model underlying the observations are unknown.

6. Conclusion

The existing literature on causal discovery shows a connection between score matching and structure learning in the context of nonlinear ANMs: in this paper, (i) we formalize and extend these results to linear SCMs, and (ii) we show that the score retains information on the causal structure even in the presence of unobserved variables. While previous works posit the accent on finding the causal order through the score, we study its potential to identify the Markov equivalence class with a *constraint-based* strategy, as well as to identify direct causal effects. Our theoretical insights result in AdaScore: unlike existing approaches for the estimation of causal directions, our algorithm provides theoretical guarantees for a broad class of identifiable models, namely linear and nonlinear, with additive noise, in the presence of latent variables. Even though AdaScore does not clearly outperform the existing baselines on our synthetic benchmark, it appears promising on realistic datasets, and its adaptivity to different structural hypotheses is a step towards causal discovery that is less reliant on prior assumptions, which are often untestable and thus hindering reliable inference in real-world problems. A promising research direction in relation to our work involves extending and applying our theory to algorithms for local causal discovery.

Acknowledgments

Philipp M. Faller was supported by a doctoral scholarship of the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation). This work has been supported by AFOSR, grant n. FA8655-20-1-7035. FM is supported by Programma Operativo Nazionale ricerca e innovazione 2014-2020. We thank Atalanti A. Mastakouri, Kun Zhang and Haoyue Dai for the insightful discussions.

References

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34:22822–22833, 2021.

Kevin Bache and Moshe Lichman. Uci machine learning repository. 2013.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6), dec 2014. URL https://doi.org/10.1214%2F14-aos1260.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *The Twelfth International Conference on Learning Representations*, 2024.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/ghoshal18a.html.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. URL https://api.semanticscholar.org/CorpusID:1152227.

Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 249–257, 2009.

Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability Second Edition*. 2019.

Nan Rosemary Ke, Silvia Chiappa, Jane X Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Theophane Weber, Matthew Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hp_RwhKDJ5.

Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Juan Lin. Factorizing multivariate function classes. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997. URL https://proceedings.neurips.cc/paper_files/paper/1997/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=eYciPrLuUhG.

Wenqin Liu, Biwei Huang, Erdun Gao, Qiuhong Ke, Howard Bondell, and Mingming Gong. Causal discovery with mixed linear and nonlinear additive noise models: A scalable approach. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1237–1263. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/liu24b.html.

Marloes Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observation data. *The Ann. Stat.*, 37, 10 2008. doi: 10.1214/09-AOS685.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 735–745. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/maeda20a.html.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47339–47378. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/93ed74938a54a73b5e4c52bbaf42ca8e-Paper-Conference.pdf.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, and Francesco Locatello. Shortcuts for causal discovery of nonlinear models by score matching, 2023b.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *2nd Conference on Causal Learning and Reasoning*, 2023c. URL https://openreview.net/forum?id=6VvoDjLBPQV.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *2nd Conference on Causal Learning and Reasoning*, 2023d. URL https://openreview.net/forum?id=rVO0Bx90deu.

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

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15(1):2009–2053, jan 2014. ISSN 1532-4435.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:239998404.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753. PMLR, 17–23 Jul 2022.

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

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Idusfje4-Wq.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, dec 2006. ISSN 1532-4435.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7 (2), 2006.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings, 2018.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Peter Spirtes. An anytime algorithm for causal inference. In Thomas S. Richardson and Tommi S. Jaakkola, editors, *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pages 278–285. PMLR, 04–07 Jan 2001. URL https://proceedings.mlr.press/r3/spirtes01a.html. Reissued by PMLR on 31 March 2021.

Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(54): 1643–1662, 2010. URL http://jmlr.org/papers/v11/spirtes10a.html.

Peter Spirtes and Thomas Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500. Citeseer, 1996.

Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Cfa@ uai*, pages 59–67, 2016.

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

Caroline Uhler, G. Raskutti, Peter Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41, 07 2012. doi: 10.1214/12-AOS1080.

Jun Wang and Klaus Mueller. Visual causality analysis made practical. In *2017 IEEE Conference on visual analytics science and technology (VAST)*, pages 151–161. IEEE, 2017.

Y Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(7), 2008a.

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

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008c. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2008.08.001. URL https://www.sciencedirect.com/science/article/pii/S0004370208001008.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhenyu Zhu, Francesco Locatello, and Volkan Cevher. Sample complexity bounds for score-matching: Causal discovery and generative modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A. Related works

In this section we discuss works closely related to ours, in the context of observational causal discovery with and without latent variables.

Causal discovery with score-matching. Several methods for the causal discovery of fully observable models using the score have been recently proposed. Ghoshal and Honorio (2018) demonstrates the identifiability of the linear non-Gaussian model from the score, and it is complemented by Rolland et al. (2022), which shows the connection between score matching estimation of $\nabla \log p(X)$ and the inference of causal graphs underlying nonlinear additive noise models with Gaussian noise terms, also allowing for sample complexity bounds (Zhu et al., 2024). Montagna et al. (2023d) provides identifiability results in the nonlinear setting, without posing any restriction on the distribution of the noise terms. Montagna et al. (2023c) is the first to show that the Jacobian of the score provides information equivalent to conditional independence in the context of causal discovery, limited to the case of additive noise models. All of these studies make specialized assumptions to find theoretical guarantees of identifiability, whereas our paper provides a unifying view of causal discovery with the score function, which generalizes and expands the existing results.

Causal discovery with latent variables. Causal discovery with latent variables has been studied first in the context of *constraint-based* approaches with the FCI algorithm (Spirtes, 2001), which shows the identifiability of the equivalence class of a marginalized graph via conditional independence testing. There are several methods that have been proposed for specific structural or functional assumptions. E.g. assuming linearity and restrictions on possible graphs Silva et al. (2006) present a method based on Tetrad-constraints, while Chandrasekaran et al. (2010) use a maximum-likelihood-based approach for linear models and sparse graphs. A wide class of approaches builds on the assumption of non-Gaussian additive noise, going back to the work of Shimizu et al. (2006) and Hoyer et al. (2008) on cases without latent variables. Some examples include Janzing et al. (2009), who show how confounders can be identified in a bivariate setting with non-linear causal relationship, Adams et al. (2021), who use conditions on the structural coefficients, Wang and Drton (2023), who recover the causal structure from statistical moments or Dong et al. (2024), who impose constraints on the rank of cross-covariance matrices. The RCD and CAM-UV (Maeda and Shimizu, 2020, 2021) approaches demonstrate the inferrability of causal edges via testing for independent regression residuals. Like the aforementioned methods, both rely on strong assumptions on the causal mechanisms: their theoretical guarantees apply to models where the effects are generated by a linear (RCD) or nonlinear (CAM-UV) additive contribution of each cause.

Our work demonstrates that using the score function for causal discovery, one can unify and generalize several of these results, while being agnostic about the class of causal mechanisms of the observed variables, under the weaker requirement of additivity of the noise terms. Further, we show how the score can be utilized for causal discovery with latent variables in a non-parametric setting.

Appendix B. Useful results

In this section, we provide a collection of results and definitions relevant to the theory of this paper.

B.1. Definitions over graphs

Let $X = X_1, \ldots, X_d$ a set of random variables. A graph $\mathcal{G} = (X, E)$ consists of finitely many nodes or vertices X and edges E. We now provide additional definitions, separately for directed acyclic and mixed graphs.

Directed acyclic graph. In a *directed graph*, nodes can be connected by a *directed edge* ($\rightarrow$), and between each pair of nodes there is at most one directed edge. We say that X_1 is a *parent* of X_j if $X_i \rightarrow X_j \in E$, in which case we also say that X_j is a *child* of X_i. Two nodes are *adjacent* if they are connected by an edge. A *path* in $\mathcal{G}$ is a sequence of at least two distinct vertices $\pi = X_{i_1}, \ldots, X_{i_m}$ such that there is an edge between X_{i_k} and $X_{i_{k+1}}$ for each $k = 1, \ldots, m$. If $X_{i_k} \rightarrow X_{i_{k+1}}$ for every node in the path, we speak of a *directed path*, and call X_{i_1} an *ancestor* of X_{i_m}, X_{i_m} a *descendant* of X_{i_1}. Given the set $\mathrm{DE}_i^{\mathcal{G}}$ of descendants of a node X_i, we define the set of *non-descendants* of X_i as $\mathrm{ND}_i^{\mathcal{G}} = X \setminus (\mathrm{DE}_i^{\mathcal{G}} \cup \{X_i\})$. Given the path $\pi = X_{i_1}, \ldots, X_{i_m}$, we say that X_{i_k}, $k = 2, \ldots, m-1$, is a *collider on* π if $X_{i_{k-1}}, X_{i_{k+1}}$ are both parents of X_{i_k}, and we call the triplet $X_{i_{k-1}} \rightarrow X_{i_k} \leftarrow X_{i_{k+1}}$ a *v-structure*. A node without parents is called a *source node*. A node without children is called a *sink node*. A *directed acyclic graph* is a directed graph with no cycles.

Mixed graph. In a *mixed graph* nodes can be connected by a *directed edge* ($\rightarrow$) or a *bidirected edge* ($\leftrightarrow$), and between each pair of nodes there is at most one directed edge. Two vertices are said to be *adjacent* in a graph if there is an edge (of any kind) between them. The definitions of *parent*, *child*, *ancestor*, *descendant*, *path* provided for directed acyclic graph also apply in the case of mixed graphs. Additionally, X_i is a *spouse* of X_j (and vice-versa) if $X_i \leftrightarrow X_j \in E$. An *almost directed cycle* occurs when $X_i \leftrightarrow X_j \in E$ and X_i is an ancestor of X_j in $\mathcal{G}$. In the context of mixed graphs, given the path $\pi = X_{i_1}, \ldots, X_{i_m}$, we say that X_{i_k}, $k = 2, \ldots, m-1$, is a *collider on* π if the edges between $X_{i_{k-1}}, X_{i_k}$ and $X_{i_k}, X_{i_{k+1}}$ both have an arrowhead towards X_{i_k}. The triplet $X_{i_{k-1}}, X_{i_k}, X_{i_{k+1}}$ is a *v-structure*

For ease of reference from the main text, we separately define inducing paths and ancestors in directed acyclic graphs.

Definition 5 (Ancestor) *Consider a DAG $\mathcal{G}$ with set of nodes X, and X_i, X_j elements of X. We say that X_i is an ancestor of X_j if there is a directed path from X_i to X_j in the graph, as in $X_i \rightarrow \ldots \rightarrow X_j$.*

Definition 6 (Inducing path) *Consider a DAG $\mathcal{G}$ with set of nodes $X = V \cup U$, V, U disjoint subsets. We say that a path π with endpoints V_i, V_j is an inducing path relative to U if every non-endpoint V_k in the path and not in U is both a collider on π (i.e. $V_i \ldots \rightarrow V_k \leftarrow \ldots V_j$ appears) and an ancestor of V_i or V_j.*

Intuitively, an inducing path relative to U is a path between two variables in V and that cannot be separated by conditioning on any other node in V. This makes them natural candidates to encode dependencies between observable random variables V that can not be eliminated by conditioning on subsets of V, as they are due to the presence of inducing paths relative to U, with U unobserved random variables.

Example 1 (Examples of inducing paths) *Trivially, a single edge path is an inducing path relative to any set of vertices (as there are no colliders). As another example, let $U = \{U_1\}$, and $V =$*

$\{V_1, V_2, V_3, V_4\}$. Let there be a direct path $V_1 \to V_2 \to V_3 \to V_4$, and the path $\pi = V_1 \to V_2 \leftarrow U_1 \to V_4$. The path π is an inducing path relative to U (as V_2 is a collider on π and also an ancestor of V_4). As final example, let $U = \{U_1\}$, $V = \{V_1, V_2\}$: $V_1 \to U_1 \to V_2$ is an inducing path relative to U.

One natural way to encode inducing paths and ancestral relationships between variables is represented by maximal ancestral graphs.

Definition 7 (MAG) *A* maximal ancestral graph *(MAG) is a mixed graph such that:*

1. *there are no directed cycles and no almost directed cycles;*
2. *there are no inducing paths between two non-adjacent nodes.*

Next, we define conditional independence in the context of graphs.

Definition 8 (active paths and m-separation) *Let $\mathcal{M}$ be a mixed graph with nodes X. A path π in $\mathcal{M}$ between X_i, X_j elements of X is* active w.r.t. *$Z \subseteq X \setminus \{X_i, X_j\}$ if:*

1. *every non-collider on π is not in Z*
2. *every collider on π is an ancestors of a node in Z.*

X_i and X_j are said to be m-separated *by Z if there is no active path between X_i and X_j relative to Z. Two disjoint sets of variables W and Y are m-separated by Z if every variable in W is m-separated from every variable in Y by Z.*

If m-separation is applied to DAGs, it is called d-separation. An active path w.r.t. the empty set is simply called active.

The set of directed acyclic graphs that satisfy the same set of conditional independencies form an equivalence class, known as the *Markov equivalence class*.

Definition 9 (Markov equivalence class of a DAG) *Let $\mathcal{G}$ be a DAG with nodes X. We denote with $[\mathcal{G}]$ the Markov equivalence class of $\mathcal{G}$. A DAG $\tilde{\mathcal{G}}$ with nodes X is in $[\mathcal{G}]$ if the following conditions are satisfied for each pair X_i, X_j of distinct nodes in X:*

- *there is an edge between X_i, X_j in $\mathcal{G}$ if and only if there is an edge between X_i, X_j in $\tilde{\mathcal{G}}$;*
- *let $Z \subseteq X \setminus \{X_i, X_j\}$. Then $X_i \perp\!\!\!\perp^d_{\mathcal{G}} X_j | Z \iff X_i \perp\!\!\!\perp^d_{\tilde{\mathcal{G}}} X_j | Z$;*
- *let π be a path between X_i and X_j. X_k is a collider on π in $\mathcal{G}$ if and only if it is a collider on π in $\tilde{\mathcal{G}}$.*

In summary, graphs in the same equivalence class share the edges up to direction, the set of d-separations, and the set of colliders.

Just as for DAGs, there may be several MAGs that imply the same conditional independence statements. Denote the *Markov-equivalence class* of a MAG $\mathcal{M}$ with $[\mathcal{M}]$: this is represented by a partial mixed graph, the class of graphs that can contain four kinds of edges: $\to$, $\leftrightarrow$, $\circ\!\!-\!\!\circ$ and $\circ\!\!\to$, and hence three kinds of end marks for edges: arrowhead ($>$), tail ($-$) and circle ($\circ$).

Definition 10 (PAG, Definition 3 of Zhang (2008a)) *Let $[\mathcal{M}]$ be the Markov equivalence class of an arbitrary MAG $\mathcal{M}$. The partial ancestral graph (PAG) for $[\mathcal{M}]$, $P_{\mathcal{M}}$, is a partial mixed graph such that:*

- *$P_{\mathcal{M}}$ has the same adjacencies as $\mathcal{M}$ (and any member of $[\mathcal{M}]$) does;*
- *A mark of arrowhead is in $P_{\mathcal{M}}$ if and only if it is shared by all MAGs in $[\mathcal{M}]$; and*
- *A mark of tail is in $P_{\mathcal{M}}$ if and only if it is shared by all MAGs in $[\mathcal{M}]$.*

Intuitively, a PAG represents an equivalence class of MAGs by displaying all common edge marks shared by all members of the class and displaying circles for those marks that are not in common.

B.2. Equivalence between m-separation and d-separation

In this section, we provide a proof for Equation (6), stating the equivalence between m-separation and d-separation in a formal sense.

Lemma 11 (Adapted from Zhang (2008a)) *Let $\mathcal{G}$ be a DAG with nodes $X = V \cup U$, with V and U disjoint sets, and $\mathcal{M}_V^{\mathcal{G}}$ the marginalization of $\mathcal{G}$ onto V. For any $\{V_i, V_j\} \subseteq V$ and $V_Z \subseteq V \setminus \{V_i, V_j\}$, the following equivalence holds:*

$$V_i \perp\!\!\!\perp_{\mathcal{G}}^{d} V_j | V_Z \iff V_i \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} V_j | V_Z.$$

Proof The implication $V_i \perp\!\!\!\perp_{\mathcal{G}}^{d} V_j | V_Z \implies V_i \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} V_j | V_Z$ is a direct consequence of Lemma 18 from Spirtes and Richardson (1996), where we set $S = \emptyset$, since we do not consider selection bias. The implication $V_i \perp\!\!\!\perp_{\mathcal{G}}^{d} V_j | V_Z \impliedby V_i \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} V_j | V_Z$ follows from Lemma 17 by Spirtes and Richardson (1996), again with $S = \emptyset$. Note, that in their terminology "d-separation in MAGs" is what we call m-separation. ■

Next, we define the *faithfulness assumption*, a bridge between d-separation and probabilistic conditional independence.

Definition 12 (Faithfulness) *Let X be generated according to the structural causal model (1), with causal graph $\mathcal{G}$. We say that the density p of X entailed by the generative SCM is faithful to the graph $\mathcal{G}$ if $X_i \perp\!\!\!\perp X_j | X_Z \implies X_i \perp\!\!\!\perp_{\mathcal{G}}^{d} X_j | X_Z$ for all i, j and $X_Z \subseteq X$.*

B.3. Additive noise model identifiability

We study the identifiability of the additive noise model, reporting results from Peters et al. (2014). We start with a formal definition of identifiability in the context of causal discovery.

Definition 13 (Identifiable causal model) *Let $(X, N, \mathcal{F}, p_N)$ be an SCM with underlying graph $\mathcal{G}$ and p_X joint density function of the variables of X. We say that the model is identifiable from observational data if the distribution p_X can not be generated by a structural causal model with graph $\tilde{\mathcal{G}} \neq \mathcal{G}$.*

First, we consider the case of models of two random variables

$$X_2 \coloneqq f(X_1) + N, \quad X_1 \perp\!\!\!\perp N. \tag{19}$$

Condition 1 (Condition 19 of Peters et al. (2014)) *Consider an additive noise model with structural equations (19). The triple (f, p_{X_1}, p_N) does not solve the following differential equation for all pairs x_1, x_2 with $f'(x_2)\nu''(x_2 - f(x_1)) \neq 0$:*

$$\xi''' = \xi'' \left(\frac{f''}{f'} - \frac{\nu''' f'}{\nu''} \right) + \frac{\nu''' \nu' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'} - 2\nu'' f'' f' + \nu' f''', \tag{20}$$

Here, $\xi \coloneqq \log p_{X_1}$, $\nu \coloneqq \log p_N$, the logarithms of the strictly positive densities. The arguments $x_2 - f(x_1)$, x_1, and x_1 of ν, ξ and f respectively, have been removed to improve readability.

Next, we show that a structural causal model satisfying Condition 1 is identifiable, as in Definition 13

Theorem 14 (Theorem 20 of Peters et al. (2014)) *Let p_{X_1,X_2} the joint distribution of a pair of random variables generated according to the model of Equation (19) that satisfies Condition 1, with graph $\mathcal{G}$. Then, $\mathcal{G}$ is identifiable from the joint distribution.*

Finally, we show an important fact, holding for identifiable bivariate models, which is that the score $\frac{\partial}{\partial X_1} \log p(x_1, x_2)$ is non-constant in x_1.

Lemma 15 (Sufficient variability of the score) *Let p_{X_1,X_2} the joint distribution of a pair of random variables generated according to a structural causal model that satisfies Condition 1, with graph $\mathcal{G}$. Then:*

$$\frac{\partial}{\partial X_1}(\xi'(x_1) - f'(x_1)\nu'(x_2 - f(x_1))) \neq 0,$$

for all pairs (x_1, x_2).

Proof By contradiction, assume that there exists (x_1, x_2) such that $\frac{\partial}{\partial X_1}(\xi'(x_1) - f'(x_1)\nu'(x_2 - f(x_1))) = 0$. Then:

$$\frac{\partial}{\partial X_1} \left(\frac{\frac{\partial^2}{\partial X_1^2}\pi(x_1, x_2)}{\frac{\partial^2}{\partial X_1 \partial X_2}\pi(x_1, x_2)} \right) = 0,$$

where $\pi(x_1, x_2) = \log p(x_1, x_2)$. By explicitly computing all the partial derivatives of the above equation, we obtain that equation 20 is satisfied, which violates Condition 1. ■

These results guaranteeing the identifiability of the bivariate additive noise model can be generalized to the multivariable case, with a set of random variables $X = \{X_1, \ldots, X_k\}$ that satisfy:

$$X_i \coloneqq f_i(X_{\mathrm{PA}_i^{\mathcal{G}}}) + N_i, i = 1, \ldots, k, \tag{21}$$

where $\mathcal{G}$ is the resulting causal graph directed and acyclic. The intuition is that, rather than studying the multivariate model as a whole, we need to ensure that Condition 1 is satisfied for each pair of nodes, adding restrictions on their marginal conditional distribution.

Definition 16 (Definition 27 of Peters et al. (2014)) *Consider an additive noise model with structural equations (21). We call this SCM a restricted additive noise model if for all $X_j \in X$, $X_i \in X_{\mathrm{PA}_j^{\mathcal{G}}}$, and all sets $X_S \subseteq X$, $S \subset \mathbb{N}$, with $X_{\mathrm{PA}_j^{\mathcal{G}}} \setminus \{X_i\} \subseteq X_S \subseteq X_{\mathrm{ND}_j}^{\mathcal{G}} \setminus \{X_i, X_j\}$, there is a value x_S with $p(x_S) > 0$, such that the triplet*

$$(f_j(x_{\mathrm{PA}_j^{\mathcal{G}} \setminus \{i\}}, \cdot), p_{X_i|X_S=x_S}, p_{N_j})$$

satisfies Condition 1. Here, $f_j(x_{\mathrm{PA}_j^{\mathcal{G}} \setminus \{i\}}, \cdot)$ denotes the mechanism function $x_i \mapsto f_j(x_{\mathrm{PA}_j^{\mathcal{G}}})$. Additionally, we require the noise variables to have positive densities and the functions f_j to be continuous and three times continuously differentiable.

Then, for a restricted additive noise model, we can identify the graph from the distribution.

Theorem 17 (Theorem 28 of Peters et al. (2014)) *Let X be generated by a restricted additive noise model with graph $\mathcal{G}$, and assume that the causal mechanisms f_j are not constant in any of the input arguments, i.e. for $X_i \in X_{\mathrm{PA}_j^{\mathcal{G}}}$, there exist $x_i \neq x_i'$ such that $f_j(x_{\mathrm{PA}_j^{\mathcal{G}} \setminus \{i\}}, x_i) \neq f_j(x_{\mathrm{PA}_j^{\mathcal{G}} \setminus \{i\}}, x_i')$. Then, $\mathcal{G}$ is identifiable.*

B.4. Other auxiliary results

We state one crucial result that we require for the proof of Proposition 4.

Lemma 18 *Let $V_j \in V$, and $Z \subset \mathbb{N}$ such that $V_Z = V_{\mathrm{PA}_j^{\mathcal{G}}} \cup \{V_j\}$. Assume that $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp \tilde{N}_j$, $\tilde{N}_j$ as defined in Equation (14). Then, the score of V_j with respect to density $p(V_Z)$ satisfies:*

$$\partial_{V_j} \log p(V_Z) = \partial_{\tilde{N}_j} \log p(\tilde{N}_j).$$

Proof By Bayes' rule, we have that $p(V_Z) = p(V_j|V_{\mathrm{PA}_j^{\mathcal{G}}})p(V_{\mathrm{PA}_j^{\mathcal{G}}})$, such that the log-likelihood can be written as:

$$\log p(V_Z) = \log p(V_j|V_{\mathrm{PA}_j^{\mathcal{G}}}) + \log p(V_{\mathrm{PA}_j^{\mathcal{G}}}).$$

Taking the partial derivative w.r.t V_j, Equation (14) implies

$$\partial_{V_j} \log p(V_Z) = \partial_{V_j} \log p(V_j|V_{\mathrm{PA}_j^{\mathcal{G}}}).$$

Note that given $V_j := f_j(V_{\mathrm{PA}_j^{\mathcal{G}}}) + \tilde{N}_j$ and the independence $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp \tilde{N}_j$, using the change of variable formula for invertible transforms on the density $p(V_j|V_{\mathrm{PA}_j^{\mathcal{G}}})$, we find that $p(V_j|V_{\mathrm{PA}_j^{\mathcal{G}}}) = p(\tilde{N}_j)$. Moreover, the chain rule of derivatives and $\partial_{V_j}\tilde{N}_j = 1$ imply $\partial_{V_j} \log p(\tilde{N}_j) = \partial_{\tilde{N}_j} \log p(\tilde{N}_j)$, such that the claim follows. ■

Appendix C. Proofs of theoretical results

C.1. Proof of Proposition 2

Proof *(Proof of Proposition 2)* Observe that

$$\frac{\partial^2}{\partial V_i \partial V_j} \log p(v_Z) = 0 \iff V_i \perp\!\!\!\perp_{\mathcal{G}}^{d} V_j | V_Z \setminus \{V_i, V_j\} \iff V_i \perp\!\!\!\perp_{\mathcal{M}_V^{\mathcal{G}}}^{m} V_j | V_Z \setminus \{V_i, V_j\},$$

where the first equivalence holds by a combination of the faithfulness assumption with the global Markov property, as explicit in Equation (3), and the second due to Lemma 11. Then, the claim is proven. ■

C.2. Proof of Proposition 3

Proof *(Proof of Proposition 3)* The forward direction is immediate from Equation (9) and $R_j = N_j$, when X_j is a sink (Equation (11)). Thus, we focus on the backward direction. Given

$$\mathbb{E}\left[\left(\mathbb{E}\left[\partial_{X_j} \log p(X) \mid R_j\right] - \partial_{X_j} \log p(X)\right)^2\right] = 0,$$

we want to show that X_j has no children, which we prove by contradiction.

Let us introduce a function $q : \mathbb{R} \to \mathbb{R}$ such that:

$$\mathbb{E}\left[\partial_{X_j} \log p(X) \mid R_j = r_j\right] = q(r_j),$$

and $s_j : \mathbb{R}^{|X|} \to \mathbb{R}$,

$$s_j(x) = \partial_{X_j} \log p(x).$$

The mean squared error equal to zero implies that $s_j(X)$ is a constant, once R_j is observed. Formally, under the assumption of $p(x) > 0$ for each $x \in \mathbb{R}^k$, this implies that

$$p(s_j(x) \neq q(R_j) | R_j = r_j) = 0, \forall x \in \mathbb{R}^k.$$

By contradiction, we assume that X_j is not a sink, and want to show that $s_j(X)$ is not constant in X, given R_j fixed. Let X_i such that $X_j \in X_{\mathrm{PA}_i^{\mathcal{G}}}$. Being the structural causal model identifiable, there is no model with distribution p_X whose graph has a backward edge $X_i \to X_j$: thus, the Markov factorization of Equation (2) is unique and implies:

$$\partial_{X_j} \log p(X) = \partial_{N_j} \log p(N_j) - \sum_{k \in \mathrm{CH}_j^{\mathcal{G}}} \partial_{X_j} h_k(X_{\mathrm{PA}_k}) \partial N_k \log p(N_k).$$

We note that, by definition of residual in Equation (10), $R_j = r_j$ fixes the following distance:

$$R_j = N_j - \mathbb{E}[N_j | X_{\setminus X_j}].$$

Hence, conditioning on R_j doesn't restrict the support of X: given $R_j = r_j$, for any $x_{\setminus X_j}$ (value of the vector of elements in $X \setminus \{X_j\}$), $\exists n_j$ with $p(n_j > 0)$ (by the hypothesis of strictly positive densities of the noise terms) that satisfies

$$r_j = n_j - \mathbb{E}[N_j | x_{\setminus X_j}].$$

Next, we condition on all the parents of X_i, except for X_j, to reduce our problem to the simpler bivariate case. Let $S \subset \mathbb{N}$ and $X_S \subseteq X$ such that $X_{\mathrm{PA}_i^{\mathcal{G}}} \setminus \{X_j\} \subseteq X_S \subseteq X_{\mathrm{ND}_i^{\mathcal{G}}} \setminus \{X_i, X_j\}$, and consider x_S such that $p(x_S > 0)$. Let $X_{\mathrm{PA}_i^{\mathcal{G}}} = x_{\mathrm{PA}_i^{\mathcal{G}}}$ hold under $X_S = x_S$. We define $X_{j_{|x_s}} := X_j | (X_S = x_S)$, and similarly $X_{|x_s} := X | (X_S = x_S)$. Being the SCM a restricted

additive noise model, by Definition 16, the triplet $(g_i, p_{X_j|_{x_s}}, p_{N_i})$ satisfies Condition 1, where $g_i(x_j) = h_i(x_{\mathrm{PA}_i^{\mathcal{G}} \setminus \{X_j\}}, x_j)$. Consider $X_i = x_i$, and the pair of values (x_j, x_j^*) such that $x_j \neq x_j^*$ and

$$\nu''_{N_i}(x_i - g_i(x_j))g'_i(x_j) \neq 0,$$
$$\nu''_{N_i}(x_i - g_i(x_j^*))g'_i(x_j^*) \neq 0,$$

where we resort to the usual notation $\nu_{N_i} \coloneqq \log p_{N_i}$. By Lemma 15, (x_i, x_j) and (x_i, x_j^*) satisfy:

$$\partial_{X_j}(\xi'(x_j) - \nu'_{N_i}(x_i - g_i(x_j))g'_i(x_j)) \neq 0,$$
$$\partial_{X_j}(\xi'(x_j^*) - \nu'_{N_i}(x_i - g_i(x_j^*))g'_i(x_j^*)) \neq 0,$$

where $\xi \coloneqq \log p_{X_j|_{x_s}}$. Thus, we can fix x_j and x_j^* (which are arbitrarily chosen) such that

$$\partial_{X_j}(\xi'(x_j) - \nu'_{N_i}(x_i - g_i(x_j))g'_i(x_j)) - \partial_{X_j}(\xi'(x_j^*) - \nu'_{N_i}(x_i - g_i(x_j^*))g'_i(x_j^*)) \neq 0. \tag{22}$$

Fixing $X_{|x_S, x_j} = x$ and $X_{|x_S, x_j^*} = x^*$, where the two values differ only in their j-th component, we find the following difference:

$$s_j(x) - s_j(x^*) = \partial_{X_j}(\xi'(x_j) - \nu'_{N_i}(x_i - g_i(x_j))g'_i(x_j)) - \partial_{X_j}(\xi'(x_j^*) - \nu'_{N_i}(x_i - g_i(x_j^*))g'_i(x_j^*)),$$

which is different from 0 by Equation (22). This contradicts the fact that the score s_j is constant once R_j is fixed, which proves our claim. ■

C.3. Proof of Proposition 4

We concentrate on the statement of Equation (17), given that the proof of Equation (18) follows a similar template. We separately analyze their backward and forward directions. The following proofs use several ideas from the demonstration of Proposition 3.

Proof [Proof of Equation (17), backward direction] Given $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp_{\mathcal{G}}^{d} U^j \wedge V_i \in V_{\mathrm{PA}_j^{\mathcal{G}}}$, we want to show that there exists $V_Z \subseteq V$, $\{V_i, V_j\} \subseteq V_Z$, such that:

$$\mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 = 0.$$

Let $V_Z = V_{\mathrm{PA}_j^{\mathcal{G}}} \cup \{V_i, V_j\}$. We will show that this is indeed the right choice for V_Z. By Lemma 18, the score of V_j is:

$$\partial_{V_j} \log p(V_Z) = \partial_{\tilde{N}_j} \log p(\tilde{N}_j).$$

Further, by Equation (16) we know that

$$R_j(V_Z) = \tilde{N}_j + c,$$

where $c = -\mathbb{E}[\tilde{N}_j]$ is a constant. It follows that the least square estimator of the score of V_j from $R_j(V_Z)$ satisfies the following equation:

$$\mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)] = \mathbb{E}[\partial_{V_j} \log p(\tilde{N}_j)|\tilde{N}_j] = \partial_{V_j} \log p(\tilde{N}_j),$$

where the first equality holds because $\mathbb{E}[\cdot|\tilde{N}_j] = \mathbb{E}[\cdot|\tilde{N}_j + c]$. Then, we find

$$\mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 = \mathbb{E}[\partial_{V_j} \log p(\tilde{N}_j) - \partial_{V_j} \log p(\tilde{N}_j)]^2 = 0,$$

which is exactly our claim. ■

Proof [Proof of Equation (17), forward direction] Given that there is $V_Z \subseteq V$, $\{V_i, V_j\} \subseteq V_Z$, such that $\mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 = 0$, we want to show that $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp^d_{\mathcal{G}} U^j \wedge V_i \in V_{\mathrm{PA}_j^{\mathcal{G}}}$. We prove the contrapositive statement, hence, assuming that V_i is connected to V_j in the marginal MAG and that $V_{\mathrm{PA}_j^{\mathcal{G}}} \not\perp\!\!\!\perp^d_{\mathcal{G}} U^j \vee V_i \notin V_{\mathrm{PA}_j^{\mathcal{G}}}$, we want to show that for each $V_Z \subseteq V$ with $\{V_i, V_j\} \subseteq V_Z$, the following holds:

$$\mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 \neq 0. \tag{23}$$

Let us introduce $q : \mathbb{R} \to \mathbb{R}$ such that:

$$\mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z) = r_j] = q(r_j),$$

and further define:

$$s_j(V_Z) = \partial_{V_j} \log p(V_Z).$$

Having the mean squared error in Equation (23) equals zero implies that $s_j(V_Z)$ is a constant, once $R_j(V_Z)$ is observed, meaning that $p(s_j(V_Z) \neq q(R_j(V_Z))|R_j(V_Z)) = 0$. Thus, the goal of the proof is to show that there are uncountable values of V_Z such that the score is not a constant once $R_j(V_Z)$ is fixed. To do that, we assume that conditioning on $R_j(V_Z)$ doesn't restrict the support of V_Z, meaning that the support of V_Z and the support of $V_Z|R_j(v_z)$ are the same. We derive the consequences of this assumption (and later prove that it holds): consider v_Z, $v_Z^* \in \mathrm{supp}(V_Z|R_j(V_Z) = r_j)$ (support of $V_Z|R_j(V_Z) = r_j$), taken from the set of uncountable values such that the score s_j function is not a constant, meaning that $s_j(v_Z) \neq s_j(v_Z^*)$. Given that different v_Z and v_Z^* are selected from an uncountable subset of the support, we conclude that the conditional score $s_j|(R_j(V_Z) = r_j) := \partial_{V_j} \log p(V_Z|R_j(V_Z) = r_j)$ is not a constant for at least an uncountable set of points, such that the claim follows.

To conclude the proof, we need to show that for each r_j, $\mathrm{supp}(V_Z|R_j(V_Z) = r_j) = \mathrm{supp}(V_Z)$. First, we consider the case where $V_{\mathrm{PA}_j^{\mathcal{G}}} \not\subset V_Z$. By Equation (15), $R_j(V_Z) = r_j$ fixes the distance

$$r_j = \tilde{N}_j + \left(f_j(V_{\mathrm{PA}_j^{\mathcal{G}}}) - \mathbb{E}[f(V_{\mathrm{PA}_j^{\mathcal{G}}}) + \tilde{N}_j|V_{Z\setminus\{j\}}]\right) =: \tilde{N}_j + h(V_{Z\setminus\{j\}}, V_{\mathrm{PA}_j^{\mathcal{G}}}), \tag{24}$$

h newly defined function. By assumption of positive density of the noise N_j on the support $\mathbb{R}$, for each $v_Z \in \mathrm{supp}(V_Z)$, there is $\tilde{n}_j \in \mathbb{R}$ such that $p_{\tilde{N}_j}(n) > 0$ and

$$r_j = \tilde{n}_j + \left(f_j(V_{\mathrm{PA}_j^{\mathcal{G}}}) - \mathbb{E}[V_{\mathrm{PA}_j^{\mathcal{G}}} + \tilde{N}_j|v_{Z\setminus\{j\}}]\right)$$

is true (in fact, note that by fixing r_j we have a system of one equation and two unknowns, which has a solution for each value taken by $h(V_Z, V_{\mathrm{PA}_j^{\mathcal{G}}})$). Then, the claim is proven for $V_{\mathrm{PA}_j^{\mathcal{G}}} \not\subset V_Z$.

We now consider $V_{\mathrm{PA}_j^{\mathcal{G}}} \subset V_Z$, and further separate the proof in subcases. Note that Equation (24) becomes:

$$r_j = \tilde{N}_j - \mathbb{E}[\tilde{N}_j | V_{Z\setminus\{j\}}] \tag{25}$$

To finalize our proof, it sufficient to show that $\mathbb{E}[\tilde{N}_j|V_{Z\setminus\{j\}}] \neq \mathbb{E}[\tilde{N}_j]$: in fact, if this is the case, we have that for each $v_Z \in \mathrm{supp}(V_Z)$ there is $\tilde{n}_j$ such that $p(\tilde{n}_j > 0)$ and Equation (25) is satisfied - which in turns implies our claim that conditioning on $R_j(V_Z)$ does not limit the support of V_Z.

Case 1. Consider $U^j \not\perp\!\!\!\perp_{\mathcal{G}}^d V_{\mathrm{PA}_j^{\mathcal{G}}}$. Together with $V_{\mathrm{PA}_j^{\mathcal{G}}} \in V_Z$, it implies that $\mathbb{E}[\tilde{N}_j|V_{Z\setminus\{j\}}] \neq \mathbb{E}[\tilde{N}_j]$. proving our claim.

Case 2. Consider $V_i \notin V_{\mathrm{PA}_j^{\mathcal{G}}}$. We know that V_i adjacent to V_j in the MAG $\mathcal{M}_{V_Z}^{\mathcal{G}}$, meaning there is an inducing path through U between the two variables. As we shall see, this implies that $V_i \not\perp\!\!\!\perp_{\mathcal{G}}^d U^j | V_{Z\setminus\{i,j\}}$, which gives $\mathbb{E}[\tilde{N}_j|V_{Z\setminus\{j\}}] \neq \mathbb{E}[\tilde{N}_j]$.

(i) For $V_i \in V_{\mathrm{AN}_j^{\mathcal{G}}}$, there must be a direct path $\pi = V_i \to \ldots \to X_{j-1} \to V_j$ where either $X_{j-1} \in U^j$, or $X_{j-1} = V_k \in V$ and is descendant of some $\tilde{U} \in U^j$, i.e. there is a path $V_i \to \ldots \to V_k \leftarrow \tilde{U}$: if this was not the case, then $V_i \perp\!\!\!\perp_{\mathcal{G}}^d V_j | V_k$, but this contradicts the definition of inducing path.

(ii) For $V_i \in V_{\mathrm{DE}_j^{\mathcal{G}}}$, given $V_i \not\perp\!\!\!\perp_{\mathcal{G}}^d V_j | V_{Z\setminus\{i,j\}}$, clearly V_i, U^j not d-separated.

(iii) For $V_i \leftrightarrow V_j$, there is at least one latent common cause $\tilde{U}$ such that there exists $\pi = V_i \leftarrow \ldots \leftarrow \tilde{U} \to \ldots \to V_j$. Given that no observable variable can block every active path between V_i, V_j, then no observed variables can block all paths of the form of π (i.e. with a latent common cause), hence, there is one such path where all variables are hidden, meaning that $V_i \not\perp\!\!\!\perp_{\mathcal{G}}^d U^j | V_{Z\setminus\{i,j\}}$.

■

The proof of Equation (18) follows from minor adjustments of the above demonstration, hence we omit it.

Appendix D. Miscellanea

D.1. The connection between NoGAM on linear models and existing literature

In this section, we elaborate on our comment that the findings of Proposition 3, relative to the identifiability of the causal order with the score even with linear mechanisms, *are not surprising, in light of previous literature*. While Montagna et al. (2023d) limits themselves to nonlinear additive noise models, previously Ghoshal and Honorio (2018) showed that the causal order can be identified by the *precision matrix* of the data, namely the inverse of the covariance matrix. In case of X generated by a linear additive noise model, the score of $X = x$ satisfies $\nabla \log p(x) = \Theta x$, where Θ denotes the precision matrix. Despite Montagna et al. (2023d) and Ghoshal and Honorio (2018) differ in how they exploit the score for estimation of the causal order, the novelty of our result in Proposition 3 is showing that findings in Montagna et al. (2023d) are less restrictive than they prove in their original paper, limiting their theoretical guarantees with the assumption of nonlinear mechanisms. Instead, an immediate generalization of the estimation of the causal order from the precision matrix in the setting of nonlinear ANMs is beyond the scope of this paper.

$X \longrightarrow Y \longrightarrow Z$ (a) Edges that FCI cannot orient

$X \longrightarrow Y \longleftarrow Z$ (b) Edges that FCI can only orient partially

Figure 3: Examples where the orientations of Proposition 4 is more informative than FCI.

D.2. On the difference of identifiability guarantees between Proposition 4 and FCI

As we have discussed before, the semantics of the edges that Proposition 4 can orient is different from edges in a MAG or PAG. The condition $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp_{\mathcal{G}}^{d} U^j \wedge V_i \in V_{\mathrm{PA}_j^{\mathcal{G}}}$ in Proposition 4 means that the node V_i is a direct parent of V_j in the original underlying DAG $\mathcal{G}$. On the other hand, in a MAG or PAG, a directed egde $V_i \to V_j$ indicates an ancestral relationship, meaning that in the underlying DAG there is a directed path from V_i to V_j. So in this sense, the information conveyed by a directed edge differs. But further, since Proposition 4 and FCI rely on different assumptions, also the kinds of structure that can be identified differ. We will illustrate this in the following examples.

First, it is important to note that all orientation rules given by Zhang (2008b) boil down to the presence of colliders in the resulting PAG. Therefore, in the absence of these, FCI cannot orient any edges in contrast to the orientation rule presented in Proposition 4.

Example 2 (Direct edges that FCI cannot detect) *Suppose the underlying distribution is generated by the DAG in Figure 3(a). Since there is no unshielded triplet and thus, no collider structure that could be detected by FCI, the PAG that FCI will output reads $X \circ\!\!-\!\!\circ Y \circ\!\!-\!\!\circ Z$. Since X is a direct parent of Y and Y has no unobserved parents, Proposition 4 can orient this edge and thus, Proposition 4 will indicate the directed edge. Similarly, for $Y \to Z$.*

Example 3 (Partially directed edges) *Now suppose the underlying distribution is generated by the DAG in Figure 3(b) and no variable is unobserved. In this case there is a collider that can be detected by FCI. Still, the edges can only be oriented partially, i.e. the output reads $X \circ\!\!\to Y \leftarrow\!\!\circ Z$. Particularly, this means there could still be arrowheads towards X and Z indicating, that they are connected by a confounder and X (or Z) is not an ancestor of Y. Again, Proposition 4 can orient these edges. This entails the assertion that X and Z are indeed direct parents of Y.*

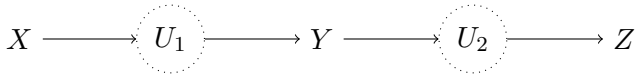
On the other hand, the following examples show how the orientation rules of FCI can still be applied in the presence of hidden confounders and mediators, which is not the case for Proposition 4.

Example 4 (Unoriented edges) *Suppose the underlying distribution is generated by the DAG in Figure 4(a) and the variable U_1 and U_2 are unobserved. Like in Example 2, FCI outputs*

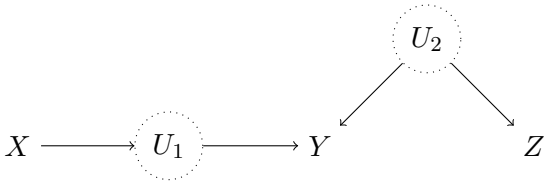
$$X \circ\!\!-\!\!\circ Y \circ\!\!-\!\!\circ Z.$$

But here Proposition 4 cannot orient the given edges either, due to the unobserved mediators.

Example 5 (Collider with unobserved nodes) *Suppose the underlying distribution is generated by the DAG in Figure 4(b) and the variable U_1 and U_2 are unobserved. Similarly to Example 3, FCI outputs the partialy oriented collider $X \circ\!\!\to Y \leftarrow\!\!\circ Z$. But here Proposition 4 cannot orient the given edges, due to the unobserved mediators.*



(*a*) Edges that cannot be oriented by FCI or Proposition 4.



(*b*) Edges that Proposition 4 cannot orient but FCI partially.

Figure 4: Examples where the output of FCI is more informative than Proposition 4 due to hidden variables.

The relevance of hidden confounders for our orientation criterion begs the question how the edges detected by Proposition 4 are related to *visible* edges (Zhang, 2008a). Intuitively, a visible edge is an edge that cannot be confounded. But again, the rules of FCI and our new rule are in no strict hierarchy. First note, that e.g. Example 2 already shows that not every edge that is orientable by Proposition 4 is visible. But there is also no inclusion the other way around, as the following examples show.

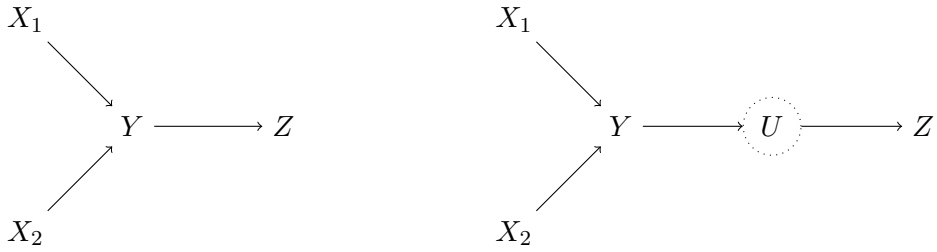
Example 6 (Collider with unobserved nodes) *Assume the underlying distribution is generated by the DAG in Figure 5(a). The edges $X_1 \circ\!\!\rightarrow Y$ and $X_2 \circ\!\!\rightarrow Y$ and $Y \rightarrow Z$ can be oriented by FCI. Further, $Y \rightarrow Z$ is visible, due to the missing link between e.g. X_1 and Z together with the arrowhead into Y on the edge between X_1 and Y. And again, Proposition 4 can orient $Y \rightarrow Z$ as it is a direct, unconfounded link.*

Example 7 (Collider with unobserved nodes) *Assume the underlying distribution is generated by the DAG in Figure 5(b) with unobserved node U. Like in Example 6, $Y \rightarrow Z$ is orientable by FCI and is visible. But due to the unobserved mediator, the link is not direct anymore and Proposition 4 cannot orient $Y \rightarrow Z$.*

Summing up these examples, there is no strict inclusion or hierarchy between edges identified by FCI and by Proposition 4.

D.3. On the difference of identifiability guarantees between Proposition 4 and CAM-UV

As we have noted before, the CAM-UV relies on stronger structural assumptions than our proposed criterion in Proposition 4. In this section we want to elaborate on the subtle differences in the graphical structures that can be recovered via these criteria. In Section E.2 we highlight further differences that are not due to Proposition 4. First note, that Proposition 4 cannot orient an edge whenever CAM-UV also cannot. To this end, recall that Maeda and Shimizu (2021) define an *unobserved backdoor path* between nodes $V_i \in V$ and $V_j \in V$ as a path $V_i \leftarrow U_k \leftarrow \cdots \leftarrow X_l \rightarrow \cdots, \rightarrow U_m \rightarrow V_j$, where $U_i, U_m \in U$ and all other nodes in X. Similarly, they define an *unobserved causal path* (UCP) between V_i and V_j to be a path $V_i \rightarrow \cdots \rightarrow U_k \rightarrow V_j$, where



(*a*) $Y \to Z$ is visible and orientable by Proposition 4.

(*b*) The edge $Y \to Z$ in the respective PAG is visible and cannot be oriented by Proposition 4.

Figure 5: Visible edges are different from the edges that can be oriented via Proposition 4.

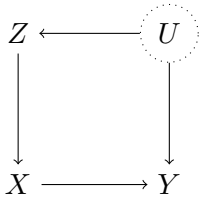


Figure 6: In this graph we have $U \not\perp\!\!\!\perp_{\mathcal{G}}^{d} X$ and Proposition 4 cannot orient the edge $X \to Y$. But CAM-UV can orient this edge, as there is no unobserved confounding path or unobserved backdoor path between X and Y.

$U_k \in U$ and all other nodes are in X. In their work, they can orient an edge iff there is no UCP or UBP between V_i and V_j.

Lemma 19 *Let $V_i, V_j \in V$ such that there is an edge between them in $\mathcal{M}_V^{\mathcal{G}}$ and there is a UBP or UCP between them. Then we have $\mathbb{E}\left[\left(\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]\right)^2\right] \neq 0$ for all $V_Z \subseteq V$ with $\{V_i, V_j\} \in V_Z$.*

Proof Suppose $V_i \in \mathrm{PA}_j^{\mathcal{G}}$ and there is a UBP between V_i and V_j. Then there is an unobserved parent U_m of V_j that is connected to V_i via this backdoor path, so we get $V_{\mathrm{PA}_j^{\mathcal{G}}} \not\perp\!\!\!\perp_{\mathcal{G}}^{d} U^j$. Now suppose there is an UCP from V_i to V_j. If $V_i \in \mathrm{PA}_j^{\mathcal{G}}$, we have again $V_{\mathrm{PA}_j^{\mathcal{G}}} \not\perp\!\!\!\perp_{\mathcal{G}}^{d} U^j$, since there is an unobserved parent U_k of V_j that is a descedant of V_i. The rest follows from Proposition 4. ■

On the other hand, there can be cases where CAM-UV recovers a direct edge, while Proposition 4 does not indicate one.

Example 8 *Suppose the underlying distribution is generated by the DAG in Figure 6 with U being unobserved. There is no UCPs, as the path from X to Y contains no unobserved nodes and no UBP, since X has an observed parent on the only backdoor path between X and Y, namely Z. Therefore, CAM-UV can orient this edge. Yet, Proposition 4 cannot, as we have $U \not\perp\!\!\!\perp_{\mathcal{G}}^{d} X$.*

D.4. Kernel ridge regression estimation of the residuals in AdaScore

In AdaScore, the residuals of Equation (15), which we recall to be defined as

$$R_k(V_Z) := V_k - \mathbb{E}[V_k \mid V_{Z\setminus\{k\}}]$$

for $V_Z \subseteq V$ and $V_k \in V$, are estimated via kernel-ridge regression (KRR). In particular, notice that for any pair of random variables X, Y, the least squares estimator is $\mathbb{E}[Y|X]$. It is thus immediate to see that, for $V_Z = V_{\mathrm{PA}_j^{\mathcal{G}}} \cup \{V_j\}$, $V_{\mathrm{PA}_j^{\mathcal{G}}} \perp\!\!\!\perp_{\mathcal{G}}^{d} U^j$ (the unobserved parents of V_j), given $\hat{V}_k := \mathbb{E}[V_k|V_Z]$ (the unregularized least squares estimator), $R_j(V_Z) = \tilde{N}_j - \mathbb{E}[\tilde{N}_j] = V_k - \hat{V}_k$. The least squares algorithm is thus the right choice to estimate the required residuals. The use of feature maps (as in KRR) is justified by the potential nonlinear mechanisms in structural equations, while the use of regularization is necessary to avoid overfitting and vanishing residuals.

Appendix E. Algorithm

E.1. Detailed description of our algorithm

In Proposition 2 we have seen that score-matching can detect m-separations and therefore the skeleton of the PAG describing the data. If one is willing to make the assumptions required for Proposition 4 it could be desirable to use this to orient edges, as Proposition 4 offers additional causal insights that we have discussed in Appendix D.2. Therefore, one could simply find the skeleton of the PAG using the fast adjacency search (Spirtes et al., 2000) and then orient the edges by applying Proposition 4 on every subset of the neighbourhood of every node. This would yield a very costly algorithm. But if we make the assumptions required to orient edges with Proposition 4 we can do a bit better. In Algorithm 2 we present an algorithm that still has the same worst case runtime but runs polynomially in the best case. The main intuition is that we iteratively remove irrelevant nodes in the spirit of the original SCORE algorithm (Rolland et al., 2022). To this end, we first check if there is any unconfounded sink (i.e. a sink that is not connected to any edge which cannot be oriented using Proposition 4) if we consider the set of all remaining variables. If there is one, we can orient its parents and ignore it afterwards. If there is no such node, we need to fall back to the procedure proposed above, i.e. we need to check the condition of Proposition 4 on all subsets of the neighbourhood of a node, until we find no node with a direct outgoing edge. In Proposition 20 we show that this way we do not fail to orient an edge or fail to remove any adjacency. In the following discussion, we will use the notation

$$\delta_i(V_Z) := \mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2,$$

for the second residual from Proposition 4 and also

$$\delta_{i,j}(V_Z) := \frac{\partial^2}{\partial V_i \partial V_j} \log p(V_Z)$$

for the cross-partial derivative, where $V_i, V_j \in V$ and $Z \subseteq V$.

Proposition 20 (Correctness of algorithm) *Let $X = V \cup U$ with $V \cap U = \emptyset$ be generated by a restricted additive noise model of the form from Equation (4). Let $\mathcal{G}$ be the causal DAG of X and $\mathcal{M}_V^{\mathcal{G}}$ be the marginal MAG of $\mathcal{G}$. Then Algorithm 2 outputs a directed edge from $V_i \in V$ to $V_j \in V$ iff there is a direct edge in $\mathcal{G}_X$ between them and $\mathrm{PA}_j^{\mathcal{G}} \perp\!\!\!\perp_{\mathcal{G}}^{d} U^j$. Further, the output of Algorithm 2 has the same skeleton as $\mathcal{M}_V^{\mathcal{G}}$.*

Algorithm 2: AdaScore Algorithm

Procedure `AdaScore` $(p, V_1, \ldots, V_d)$ **:**

$S \leftarrow \{V_1, \ldots, V_d\}, E \leftarrow \{\}$; ▷ Init remaining nodes and edges

for $V_i \in S$ **do**

$B_i \leftarrow \{V_1, \ldots, V_d\}$; ▷ Neighbourhoods

end

while $S \neq \emptyset$ **do** ▷ While nodes remain

if $\exists V_i \in S : \delta_i(V_S) = 0$ **then** ▷ Unconfounded sink

$S \leftarrow S \setminus \{V_i\}$ $E \leftarrow E \cup \{V_j \rightarrow V_i : \delta_{i,j}(V_S) \neq 0\}$; ▷ Add edges like DAS

end

else

$V_i \leftarrow \min_{V_j \in S} \delta_j(V_S)$

for $V_j \in \{V_k \in B_i : \min_{S' \subseteq B_i} \delta_{i,k}(V_{S' \cup \{V_i, V_k\}}) = 0\}$ **do**

$B_i \leftarrow B_i \setminus \{V_j\}$ ▷ Prune neighbourhoods

$B_j \leftarrow B_j \setminus \{V_i\}$

end

for $V_j \in B_i$ **do** ▷ Orient edges in B_i

$m_i = \min_{S' \subseteq B_i} \delta_i(V_{S' \cup \{V_i, V_j\}})$

$m_j = \min_{S' \subseteq B_j} \delta_j(V_{S' \cup \{V_i, V_j\}})$

if $m_i = 0 \wedge m_j \neq 0$ **then**

$E \leftarrow E \cup \{V_j \rightarrow V_i\}$

end

else if $m_i \neq 0 \wedge m_j = 0$ **then**

$E \leftarrow E \cup \{V_i \rightarrow V_j\}$

end

else

$E \leftarrow E \cup \{V_i - V_j\}$

end

end

if $\exists V_j \in B_i : (V_i \rightarrow V_j) \in E$ **then**

continue with V_j

end

else

$S \leftarrow S \setminus \{V_i\}$; ▷ Remove V_i

break

end

end

end

for $V_i - V_j \in E$ **do** ▷ Prune undirected edges

if $\min_{S' \subseteq B_i} \delta_{i,j}(V_{S' \cup \{V_i, V_j\}}) = 0 \vee \min_{S' \subseteq B_j} \delta_{i,j}(V_{S' \cup \{V_i, V_j\}}) = 0$ **then**

$E \leftarrow E \setminus \{V_i - V_j\}$

end

end

return E

Proof Note, that in this proof we will refer to edges that can be oriented w.r.t. a set $S \subseteq V$, where we mean applying Proposition 4 to the observed and unobserved nodes implied by the partitioning $X = S \cup (X \setminus S)$, instead of $X = V \cup U$. We also implicitly change the definitions of observed and unobserved parents PA_i and U^i for a node V_i w.r.t. S. We define that a node is an *unconfounded sink* w.r.t. to a set S iff it has no children in S and is not incident to an edge that cannot be oriented via Proposition 4 w.r.t. S.

We prove the statement by induction over the steps of the algorithm. Let S be the set of remaining nodes in an arbitrary step of the algorithm. Our induction hypothesis is that for $V_i, V_j \in S$ and $V_k \in B_i$ (where B_i is defined as in Algorithm 2) we have

1. V_i is an unconfounded sink w.r.t. to some set $S' \subseteq S$ iff V_i is an unconfounded sink w.r.t. some $S'' \subseteq V$
2. if there is no $S' \subseteq V \setminus \{V_i, V_j\}$ such that $V_i \perp\!\!\!\perp V_j \mid S'$ then $V_j \in B_i$

Clearly, this holds in the initial step as $S = V$.

Suppose we find $\delta_i(V_S) = 0$ for $V_i \in S$ and $|S| > 1$. By Proposition 4, we know that all nodes that are adjacent in the underlying MAG and are in S must be parents of V_i. This means, all nodes that are not separable from V_i must be direct parents of V_i, which are, by our induction hypothesis 2), the nodes in B_i. Since V_i does not have children in S, it also suffices to check $V_i \perp\!\!\!\perp V_j | S \setminus \{V_i, V_j\}$ for $V_j \in S$ (instead of conditioning on all subsets of S) to determine whether V_j is in B_i. So we can already add these direct edges to the output. If, on the other hand, V_i is not adjacent to a node in S, we have $V_i \perp\!\!\!\perp V_j | S \setminus \{V_i, V_j\}$ for $V_j \in B_i$, so we add precisely the correct set of parents. Since V_i is not a parent of any of the nodes in $S \setminus \{V_i\}$, V_i cannot be in the set of unobserved parents of a node in $S \setminus \{V_i\}$ after it's removal and conditioning on V_i cannot block an open path. Thus, the induction hypothesis still holds in the next step.

Suppose now there is no unconfounded sink and we explore V_i. By our induction hypothesis 2), B_i contains the parents of V_i and by Proposition 4 it suffices to only look at subsets of B_i to orient direct edges, as we only orient edges that also exist in the MAG. And also due to the induction hypothesis 2) B_i contains all nodes that are not separable from V_i. So by adding undirected edges to all nodes in B_i can only add too many edges but not miss some.

Now it remains to show that the induction hypothesis holds if we set S to $S \setminus \{V_i\}$ in the previous case. For 1) we need to show that V_i cannot prevent the orientation of an edge w.r.t. $S \setminus \{V_i\}$. Suppose there are $V_k, V_l \in S \setminus \{V_i\}$ such that they are not separable and $\mathrm{PA}_l \not\perp\!\!\!\perp^d_{\mathcal{G}} U^l$. If $V_i \notin U^l$, then the edge between V_k and V_l could not have been oriented w.r.t. to S already. So suppose $V_i \in U^l$, i.e. there is a direct edge from V_i to V_l in the original DAG. We would not remove V_i from S if this edge was orientable, so there must a hidden confounder or mediator between V_i and V_l, i.e. we have $\mathrm{PA}_l \not\perp\!\!\!\perp^d_{\mathcal{G}} U^l$ w.r.t. to S. But then we also have $V_k \not\perp\!\!\!\perp^d_{\mathcal{G}} U^l$ w.r.t. to S, due to the edge between V_i and V_k. So in this case, Proposition 4 wouldn't allow us to direct the edge w.r.t. $S \setminus \{V_i\}$ as we again have $\mathrm{PA}_l \not\perp\!\!\!\perp^d_{\mathcal{G}} U^l$. So by removing V_i we do not render any edges unorientable. For 2) it suffices to note that we only remove nodes from B_i if we found an independence.

For $|S| < 2$, the algorithm enters the final pruning stage. From the discussion above it is clear, that we already have the correct result, up to potentially too many undirected edges. In the final step we certainly remove all these edges $V_i - V_j$, as we check m-separation for all subsets of the neighbourhoods $\mathrm{Adj}(V_i)$ and $\mathrm{Adj}(V_j)$, which are supersets of the true neighbourhoods. ■

E.2. On the output of AdaScore and CAM-UV

The algorithm we described in Algorithm 2 outputs (if desired by the user) a mixed graph with the skeleton of the underlying MAG and direct edges that indicate a direct causal influence. The CAM-UV algorithm similarly outputs direct parental relationships and a set of pairs of nodes with what Maeda and Shimizu (2021) call *unobserved backdoor paths* (UBP) and *unobserved causal paths* (UCP) between them. Yet, they do not investigate whether their algorithm *only* outputs said pairs. In fact, the following example shows that CAM-UV can also add pairs to this set that have neither a UBP nor a UCP between them.

Example 9 *In the following we will extensively reference Algorithm 1 and Algorithm 2 from Maeda and Shimizu (2021). Suppose we apply CAM-UV to a sufficiently large sample from a distribution that fulfills CAM-UV's assumptions w.r.t. to the graph G shown in Figure 7, where U is unobserved. First note, that there is neither a UBP nor a UCP w.r.t. to U between I and Y. Further, CAM-UV will not add X to the set M_Y of non-descendants of Y, due to the UBP between them. Similarly, I is not added to M_Y, as*

$$Y - f(X) - g(I) \not\perp\!\!\!\perp \{X, I\} \quad \text{and} \quad Y - h(I) \not\perp\!\!\!\perp I$$

for any functions f, g and h. Hence, Algorithm 2 in (Maeda and Shimizu, 2021) then finds

$$I \not\perp\!\!\!\perp Y$$

and adds them to the set of pairs with UBP and UCP between them.

On the other hand, the path $I \to X \leftarrow U \to Y$ is an inducing path w.r.t. to U. Therefore it is well-defined that AdaScore adds an (undirected) edge here.

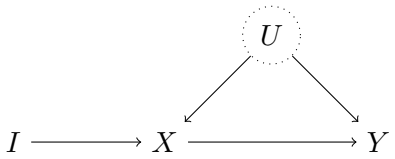


Figure 7: The CAM-UV algorithm will add I and Y to the set of nodes with unobserved backdoor path or unobserved causal path between them, though there is no such path between them.

This begs the question whether CAM-UV then also outputs the skeleton of the underlying MAG like AdaScore. The answer is negative, as the following example shows.

Example 10 *Suppose we apply CAM-UV to a sufficiently large sample from a distribution that fulfills CAM-UV's assumptions w.r.t. to the graph G shown in Figure 7, where now U_1 and U_2 are unobserved. Again, there is neither a UBP nor a UCP w.r.t. to U_1, U_2 between Z and Y. Similarly to Example 9, CAM-UV will not add X to the parent sets of Z and Y due to the UBP between X and Y and the UCP between Z and X. Therefore, Algorithm 2 in (Maeda and Shimizu, 2021) will find*

$$Z \not\perp\!\!\!\perp Y,$$

and add the pair (Z, Y) to the nodes with UBP or UCP between them.

In contrast, Z and Y are not connected by an inducing path, as X separates them. Therefore, the underlying MAG (and thus the output of AdaScore) does not contain an edge between Z and Y.

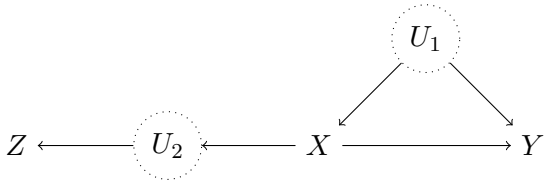


Figure 8: CAM-UV will add Z and Y to the set of nodes with unobserved backdoor path or unobserved causal path between them, though there is neither such path nor an inducing between them.

E.3. Finite sample version of AdaScore

All theoretical results in the paper have assumed that we know the density of our data. Obviously, in practice, we have to deal with a finite sample instead. Especially, in Proposition 2 and Proposition 4 we derived criteria that compare random variables with zero. Clearly, this condition is almost never met in practice. Therefore, we need to find ways to reasonably set thresholds for these random quantities.

First note, that we use the Stein gradient estimator (Li and Turner, 2017) to estimate the score function, given by

$$\hat{\mathbf{G}}^{\text{Stein}} = -(\mathbf{K} + \eta \mathbf{I})^{-1} \langle \nabla, \mathbf{K} \rangle, \tag{26}$$

where $\mathbf{v}_Z^i$ is the i-th of $m \in \mathbb{N}$ samples of the observable variables in the set Z, $\mathbf{K}$ is the matrix with $\mathbf{K}_{ij} = \mathcal{K}(\mathbf{v}^i, \mathbf{v}^j)$, we use $\langle \nabla, \mathbf{K} \rangle$ to denote a matrix with $\langle \nabla, \mathbf{K} \rangle_{ij} = \sum_{k=1}^m \nabla_{v_j^k} \mathcal{K}(\mathbf{v}_Z^i, \mathbf{v}_Z^k)$ and $\mathcal{K}$ is the RBF kernel in our case. This means especially that for a node V_i we get a vector of estimates

$$\left(\partial V_i \widehat{\log p}(\mathbf{v}_Z^k) \right)_{k=1,\dots,m} \tag{27}$$

i.e. an estimate of the score for every one of the m samples. Analogously, we get a $m \times d \times d$ tensor for the estimates of $\frac{\partial^2}{\partial V_i \partial V_j} \log p(v)$ and m empirical estimates of the residual $\delta_i(V_Z)$.

In Proposition 2 we showed that

$$\frac{\partial^2}{\partial V_i \partial V_j} \log p(v_Z) = 0 \iff X_i \perp\!\!\!\perp^m_{\mathcal{M}_V^{\mathcal{G}}} V_j | V_Z \setminus \{V_i, V_j\}.$$

In the finite sample version, we use a one-sample t-test on the vector of estimated cross-partial derivatives with the null hypothesis that the means is zero. Due to the central limit theorem, the sample mean follows approximately a Gaussian distribution, regardless of the true distribution of the observations. In the pruning steps, we do not conduct a test for every possible subset of the neighbourhood. Instead, we pick the subset with the minimal mean absolute value of the estimated

cross-partial derivatives and conduct the t-test for this set. I.e. if we consider pruning the edge $V_i - V_j$) we pick the subset

$$Z = \min_{Z' \subseteq B_j \cup B_i} \text{mean } |\widehat{\delta_{i,j}(V_{Z'})}|.$$

We then conduct the t-test with the empirical estimates of $\delta_{i,j}(V_Z)$.

For Proposition 4, we first prove the following connection to the independence of fitted additive noise models.

Proposition 21 (Zero MSE implies ANM) *Let Z be a set of variable indices such that $V_j \in V_Z$. Then with $R_j(V_Z) := V_j - \mathbb{E}[V_j|V_{\setminus V_j}]$ we get*

$$\mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 = 0 \implies V_j = \mathbb{E}[V_j|V_{\setminus V_j}] + R_j(V_Z),$$

with $V_{\setminus V_j} \perp\!\!\!\perp R_j(V_Z)$. In other words, if the score can be estimated with zero MSE, then V_j can be described by an additive noise model in $V_{\setminus V_j}$ with independent noise.

Proof If the MSE of the final regression is zero, we have by definition

$$0 = \mathbb{E}[\partial_{V_j} \log p(V_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(V_Z)]]^2 \tag{28}$$

$$= \int \left(\partial_{V_j} \log p(v_Z) - \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(v_Z)]\right)^2 dv_Z, \tag{29}$$

which implies that $\partial_{V_j} \log p(v_Z) = \mathbb{E}[\partial_{V_j} \log p(V_Z)|R_j(v_Z)]$ almost everywhere, since this is an integral over non-negative values. Denote $q(r) = \mathbb{E}[\partial_{V_j} \log p(V_Z)|r]$. By the fundamental theorem of calculus we have

$$\begin{aligned}
\log p(v_Z) &= \int \partial_{V_j} \log p(v_Z) dv_j + c(v_{\setminus V_j}) \\
p(v_Z) &= \exp\left(\int \partial_{V_j} \log p(v_Z) dv_j + c(v_{\setminus V_j})\right) \\
&= \exp\left(\int q(R_j(v_Z)) dv_j + c(v_{\setminus V_j})\right) \\
&= \exp\left(\int q(v_j - \mathbb{E}[V_j|v_{\setminus V_j}]) dv_j + c(v_{\setminus V_j})\right) \\
&= \exp\left(Q(R_j(v_Z)) + c(v_{\setminus V_j})\right) \\
&= \exp\left(Q(R_j(v_Z))\right) \cdot \exp\left(c(v_{\setminus V_j})\right),
\end{aligned}$$

for some function Q and some $c(v_{\setminus V_j})$ which is constant in V_j (but may depend on $V_{\setminus V_j}$). Since V_j is deterministic in $V_{\setminus V_j}$ and $R_j(V_Z)$ by definition, we get

$$p(R_j(v_Z), v_{\setminus V_j}) = p(v_Z) = \exp\left(Q(R_j(v_Z))\right) \cdot \exp\left(c(v_{\setminus V_j})\right)$$

and further, since this joint distribution can be factorized into two functions of $V_{\setminus V_j}$ and $R_j(V_Z)$ respectively, we get $R_j(V_Z) \perp\!\!\!\perp V_{\setminus V_j}$ Now we have established that we can write

$$V_j = \mathbb{E}[V_j|V_{\setminus V_j}] + R_j(V_Z),$$

and $V_{\setminus V_j} \perp\!\!\!\perp R_j(V_Z)$, or in other words, that V_j can be described by an additive noise model in $V_{\setminus V_j}$. ■

This means, we can use a classical independence test on the regression residual $R_j(V_Z)$ (that we had to estimate anyway) and the regressors $V_{\setminus V_j}$. If such a test rejects, this also rejects that $\delta_j(V_Z) = 0$. We used the kernel independence test proposed by (Zhang et al., 2012) to this end.

As candidate sink for set $Z \subseteq V$, we pick the node $V_i = \min_i \text{mean}(\delta_i(V_Z))$.

In the case where we use Proposition 4 to *orient* edges, we only need to decide whether a previously undirected edge $V_i - V_j$ needs to be oriented one way, the other way, or not at all. Again, here the issue lies in the fact that we need to iterate over possible sets of parents of the nodes. We pick the subset

$$Z_j = \min_{Z' \subseteq B_j} \text{mean}(\widehat{\delta_j(V_{Z'})},$$

i.e. the set with the lowest MSE. We then conduct the test with the estimates of $\delta_i(V_{Z_j})$ and $\delta_j(V_{Z_j})$ to check if the edge is pointing to V_j. If there is a directed edge between them, one of the residuals will be independent from the regressors and the other one won't.

Just like Montagna et al. (2023d) we use a cross-validation scheme to generate the residuals, in order to prevent overfitting. We split the dataset into several equally sized, disjoint subsamples. For every residual we fit the regression on all subsamples that don't contain the respective target.

E.4. Complexity

Proposition 22 *Complexity Let n be the number of samples and d the number of observable nodes. Algorithm 2 runs in*

$$\Omega\left((d^2 - d) \cdot (r(n,d) + s(n,d))\right) \quad \textit{and} \quad \mathcal{O}\left(d^2 \cdot 2^d (r(n,d) + s(n,d))\right),$$

where $r(n,d)$ is the time required to solve a regression problem and $s(n,d)$ is the time for calculating the score. With e.g. kernel-ridge regression and the Stein-estimator, both run in $\mathcal{O}(n^3)$.

Proof Algorithm 2 runs its main loop d times. It first checks for the existence of an unconfounded sink, which involves solving $2d$ regression problems (including cross-validation prediction) and calculating the score, adding up to $(2d^2 - d)$ regressions and d score evaluations. In the worst case, we detect no unconfounded sink and iterate through all subsets of the neighbourhood of a node (which is in the worst case of size $d - 1$) and for all other nodes in the neighbourhood we solve $2d$ regression problems and evaluate the score. For each subset we calculate two regression functions, the score and calculate the entries in the Hessian of the log-density, i.e. $d \cdot 2^d$ regressions, $d \cdot 2^{d-1}$ scores and additionally 2^{d-1} Hessians. If we are unlucky, this node has a directed outgoing edge and we continue with this node (with the same size of nodes). This can happen $d - 1$ times. So we get $(d^2 - d) \cdot 2^d$ regressions and $(d^2 - d) \cdot 2^{d-1}$ scores and Hessians. In the final pruning step we calculate for every previously undirected edge (of which there can be $(d^2 - d)/2$) a Hessian for all subsets of the neighbourhoods, which can again be 2^{d-1} subsets. Using the pruning procedure from CAM for the directed edges we also spend at most $\mathcal{O}(nd^3)$ steps.

In the best case, we always find an unconfounded sink. Then our algorithm reduces to NoGAM. ■

Appendix F. Experimental details

In this section, we present the details of our experiments in terms of synthetic data generation and algorithms hyperparameters.

F.1. Synthetic data generation

In this work, we rely on synthetic data to benchmark AdaScore's finite samples performance. For each dataset, we first sample the ground truth graph and then generate the observations according to the causal graph.

Erdös-Renyi graphs. The ground truth graphs are generated according to the Erdös-Renyi model. It allows specifying the number of nodes and the probability of connecting each pair of nodes). In ER graphs, a pair of nodes has the same probability of being connected. To introduce hidden variables we randomly drop columns from the data matrices. Whenever the resulting graph contains no hidden confounder (or in the case of nonlinear data also hidden mediators) we reject the choice of hidden variables and sample again. We similarly reject choices without hidden variables or hidden mediators for the experiments in Figure 2. We always compare against the graph that AdaScore would output in the limit of infinite data.

Nonlinear causal mechanisms. Nonlinear causal mechanisms are parametrized by a neural network with random weights. We create a fully connected neural network with one hidden layer with 10 units, Parametric ReLU activation function, followed by one normalizing layer before the final fully connected layer. The weights of the neural network are sampled from a standard Gaussian distribution. This strategy for synthetic data generation is commonly adopted in the literature (Montagna et al., 2023a,b; Ke et al., 2023; Brouillard et al., 2020; Lippe et al., 2022).

Linear causal mechanisms. For the linear mechanisms, we define a simple linear regression model predicting the effects from their causes and noise terms, weighted by randomly sampled coefficients. Coefficients are generated as samples from a Uniform distribution supported in the range $[-3, -0.5] \cup [0.5, 3]$. We don't use too small coefficients to avoid trivial cases of *close to unfaithful* datasets (Uhler et al., 2012).

Noise terms distribution. The noise terms are sampled from a Uniform distribution supported between -2 and 2.

Finally, we remark that we standardize the data by their empirical standard deviation. This is known to remove shortcuts that allow finding a correct causal order sorting variables by their marginal variance, as in *varsortability*, described in Reisach et al. (2021), or sorting variables by the magnitude of their score $|\partial_{X_i} \log p(X)|$, a phenomenon known as *scoresortability* analyzed by Montagna et al. (2023b).

F.2. AdaScore hyperparameters

For AdaScore, we set the α level for the required hypothesis testing at 0.05. For the CAM-pruning step, the level is instead set at 0.001, the default value of the dodidscover Python implementation of the method, and commonly found in all papers using CAM-pruning for edge selection (Rolland et al., 2022; Montagna et al., 2023c,d; Bühlmann et al., 2014). For the remaining parameters. The regression hyperparameters for the estimation of the residuals are found via cross-validation during

Table 2: p-values for stochastic ordering: sparse linear fully observable model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.61060	0.44153	0.76942	0.79324	0.39970	0.56901	0.23885	0.22248
nogam	0.60030	0.12113	0.08691	0.03375	0.41007	0.88947	0.92142	0.96828
rcd	0.61060	0.08267	0.34901	0.68977	0.39970	0.92142	0.66085	0.32940
lingam	0.70844	0.79324	0.99854	1.00000	0.30083	0.21454	0.00177	0.00000
random	0.00000	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000

inference: tuning is done minimizing the generalization error on the estimated residuals, without using the performance on the causal graph ground truth. Finally, for the score matching estimation, the regularization coefficients are set to 0.001. In the respective experiments, we provided the fact that there are no latent confounders to AdaScore.

F.3. Random baseline

In our synthetic experiments we also considered a random baseline. Since we did not want the sparsity of the ground truth graph to influence the performance of the random baseline, we chose the following approach: we use the Erdös-Renyi model, described above, to generate a new graph with the same edge probability and the same number of nodes (including hidden variables) as the ground truth graph. We then project this graph onto the actually observed nodes by generating the PAG skeleton over these nodes and directing edges iff they are identifiable by Proposition 4 (using graphical criteria).

F.4. Compute resources

All experiments have been run on an AWS EC2 instance of type `m5.12xlarge`. These machines contain Intel Xeon Platinum 8000 processors with 3.1 GHz and 48 virtual cores as well as 192 GB RAM. All experiments can be run within a day.

Appendix G. Additional Experiments

In this section, we provide additional experimental results. All synthetic data has been generated as described in Appendix F.1.

G.1. Hypothesis Tests

Additionally to the plots in Figure 1 (and also Figure 11) we conducted hypothesis tests to see whether the results of AdaScore are stochastically greater or less than the results of the other algorithms. I.e. we tested the null-hypothesis that $P(A > B) = P(B < A)$, where A denotes the SHD of AdaScore w.r.t. the ground truth graph and B the SHD of one of the baseline algorithms. In Tables 2 to 9 we provided p-values for the exact Mann-Whitney U test with alternatives that AdaScore is less than the given baseline or greater, respectively.

Table 3: p-values for stochastic ordering: sparse nonlinear fully observable model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	1.00000	1.00000	0.99953	0.98249	0.00000	0.00000	0.00052	0.01876
nogam	1.00000	0.99995	0.99892	0.95951	0.00000	0.00006	0.00120	0.04554
rcd	0.07464	0.13831	0.00019	0.00001	0.93279	0.87331	0.99985	0.99999
lingam	0.97206	0.99564	0.70844	0.65099	0.03172	0.00516	0.30083	0.35898
random	0.60030	0.02009	0.00005	0.00001	0.42051	0.98249	0.99995	0.99999

Table 4: p-values for stochastic ordering: sparse linear latent variables model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.18443	0.60030	0.74417	0.79324	0.82267	0.41007	0.26455	0.22248
nogam	0.18443	0.23058	0.10060	0.07858	0.82267	0.77752	0.90412	0.92536
rcd	0.31976	0.62082	0.41007	0.50533	0.69917	0.38940	0.60030	0.50533
lingam	0.25583	0.76942	0.97702	0.99600	0.76115	0.23885	0.02620	0.00436
random	0.00001	0.00000	0.00000	0.00000	0.99999	1.00000	1.00000	1.00000

Table 5: p-values for stochastic ordering: sparse nonlinear latent variables model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.99854	0.97206	0.87887	0.76115	0.00161	0.03172	0.13242	0.25583
nogam	0.34901	0.52665	0.42051	0.28242	0.67060	0.48400	0.58993	0.72659
rcd	0.96625	0.51600	0.22248	0.00146	0.03589	0.50533	0.79324	0.99868
lingam	0.12113	0.13831	0.33915	0.01054	0.88425	0.87331	0.68024	0.99095
random	0.00010	0.00001	0.00000	0.00000	0.99991	0.99999	1.00000	1.00000

Table 6: p-values for stochastic ordering: dense linear fully observable model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.91309	0.28242	0.55847	0.99823	0.09132	0.73545	0.46271	0.00213
nogam	0.87331	0.12113	0.01632	0.00367	0.13831	0.88947	0.98479	0.99693
rcd	0.81557	0.07464	0.82961	0.96828	0.19171	0.92915	0.17733	0.03375
lingam	0.91309	0.98585	1.00000	1.00000	0.09132	0.01521	0.00000	0.00000
random	0.00000	0.00000	0.00108	0.00234	1.00000	1.00000	0.99912	0.99806

Table 7: p-values for stochastic ordering: dense nonlinear fully observable model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	1.00000	0.99999	0.99095	0.79324	0.00000	0.00001	0.01054	0.21454
nogam	1.00000	1.00000	1.00000	0.96625	0.00000	0.00000	0.00000	0.03813
rcd	0.16363	0.02298	0.00000	0.00000	0.84937	0.97991	1.00000	1.00000
lingam	0.94892	0.81557	0.76115	0.48400	0.05713	0.19171	0.24726	0.53729
random	0.97545	0.28242	0.00010	0.00000	0.02620	0.73545	0.99992	1.00000

Table 8: p-values for stochastic ordering: dense linear latent variables model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.24726	0.11053	0.90868	0.99902	0.76115	0.89452	0.10060	0.00120
nogam	0.05404	0.18443	0.09132	0.00977	0.94892	0.82961	0.91309	0.99163
rcd	0.48400	0.13242	0.99912	0.99928	0.52665	0.87887	0.00108	0.00088
lingam	0.19171	0.90412	0.99948	0.99987	0.81557	0.10060	0.00058	0.00015
random	0.01316	0.00000	0.00027	0.25583	0.98777	1.00000	0.99976	0.75274

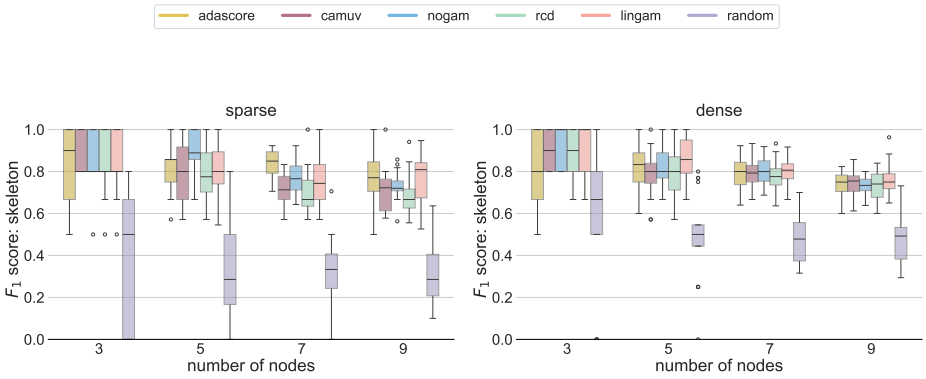
Table 9: p-values for stochastic ordering: dense nonlinear latent variables model

	less				greater			
	3	5	7	9	3	5	7	9
camuv	0.48400	0.89940	0.97991	0.97702	0.53729	0.10548	0.02298	0.02620
nogam	0.00052	0.16363	0.37918	0.10060	0.99953	0.84937	0.64102	0.90412
rcd	0.71758	0.32940	0.98368	0.36903	0.29156	0.68977	0.01751	0.65099
lingam	0.00016	0.00108	0.06371	0.00257	0.99987	0.99902	0.93965	0.99766
random	0.00072	0.00177	0.13831	0.00003	0.99935	0.99854	0.87331	0.99997

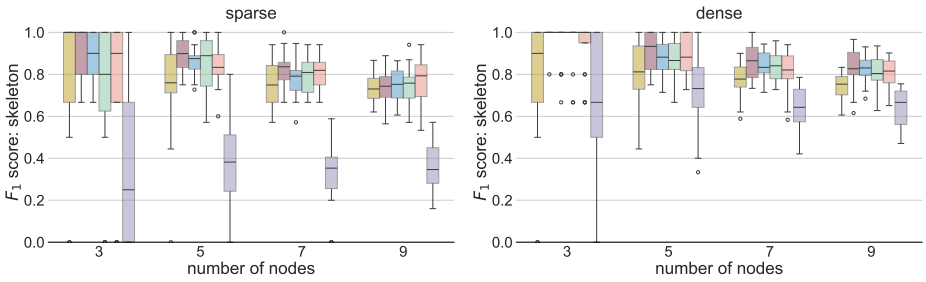
G.2. Non-additive mechanisms

In Figure 1 we have demonstrated the performance of our proposed method on data generated by linear SCMs and non-linear SCMs with additive noise. But Proposition 2 also holds for *any* faithful distribution generated by an acyclic model. Thus, we employed as mechanism a neural network-based approach similar to the non-linear mechanism described in Appendix F. Instead of adding the noise term, we feed it as additional input into the neural network. Results in this setting are reported in Figure 9. As neither the mixed graph mode of AdaScore nor any of the baseline algorithms has theoretical guarantees for the orientation of edges in this scenario, we report the F_1-score (popular in classification problems) w.r.t. to the existence of an edge, regardless of orientation. Our experiments show that AdaScore can, in general, correctly recover the graph's skeleton in all the scenarios, with an F_1 score median between 1 and ~ 0.75, respectively for small and large numbers of nodes.

On the other hand, if we let AdaScore output a PAG, we can apply the ordinary orientation rules of FCI (Spirtes et al., 2000). In Figure 10 we compare the vanilla FCI algorithm with kernel independence test (Zhang et al., 2012) against AdaScore on dense and sparse ground truth graphs.



(*a*) Fully observable model



(*b*) Latent variables model

Figure 9: Empirical results for non-additive causal mechanisms on sparse and dense graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the F_1 score w.r.t. the existence of edges (the higher, the better).

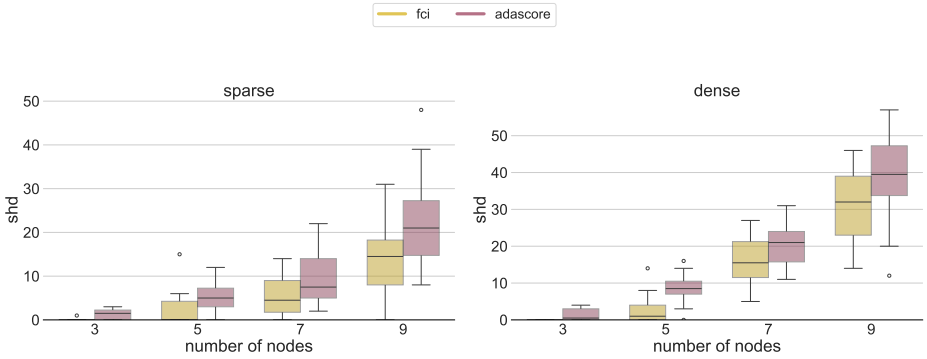


Figure 10: Results for the vanilla FCI algorithm with kernel independence test versus PAG output of AdaScore. We used non-additive causal mechanisms on sparse (left) and dense graphs (right) with different numbers of nodes. We report the SHD.

G.3. Dense graphs

In this section, we present the experiments on dense Erdös-Renyi graphs where each pair of nodes is connected by an edge with probability 0.3. The results are illustrated in Figure 11. For dense graphs, recovery results are similar to the sparse case, with AdaScore generally providing comparable performance to the other methods.

G.4. F_1 scores

The following plots show the F_1 score as an additional metric for the previously discussed experiments. Since the F_1 score is only applicable to binary decisions, we calculate it with respect to the binary classification of whether there is an identifiable direct edge in the ground truth graph or not in Figures 12 and 13, or whether there is an edge that is not identifiable via Proposition 4 respectively in Figure 14.

G.5. Increasing number of samples

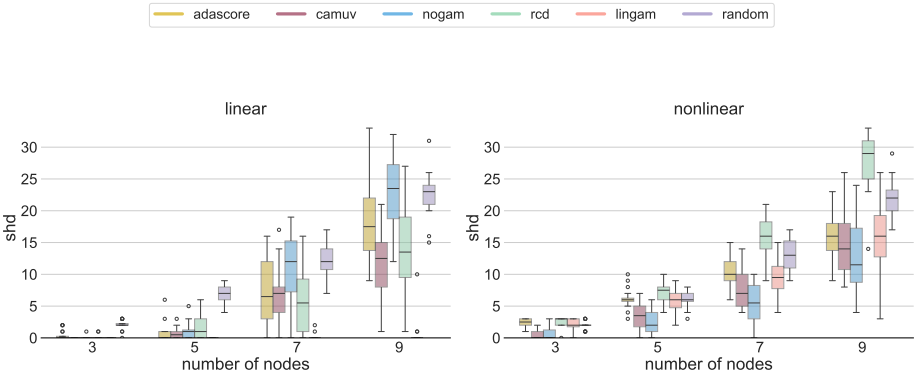
In the following series of plots we demonstrate the scaling behaviour of our method w.r.t. to the number of samples. Figure 16 shows results with edge probability 0.5 and Figure 15 with 0.3. All graphs contain seven observable nodes. As before we observe that AdaScore performs comparably to other methods.

G.6. Runtimes

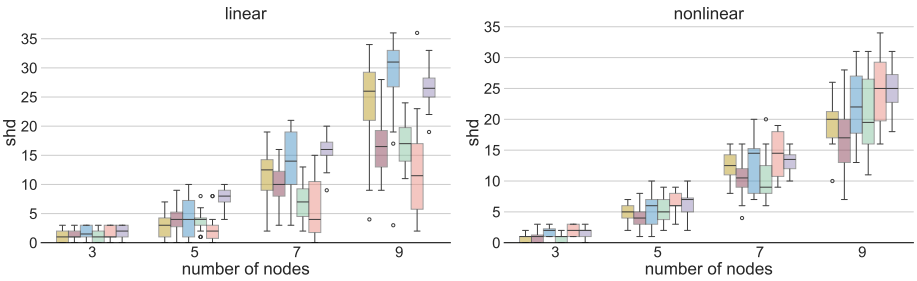
In Figures 17 to 20 we have plotted the runtimes of the benchmarked methods in different settings.

G.7. Limitations

In this section, we remark the limitations of our empirical study. It is well known that causal discovery lacks meaningful, multivariate benchmark datasets with known ground truth. For this reason, it is common to rely on synthetically generated datasets. We believe that results on synthetic

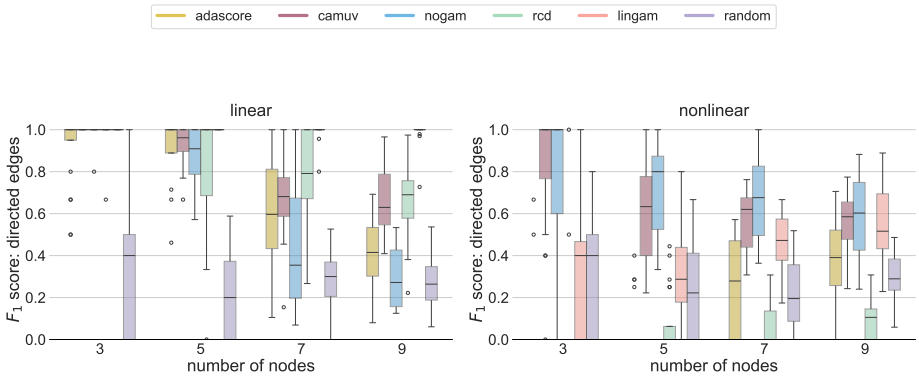


(*a*) Fully observable model

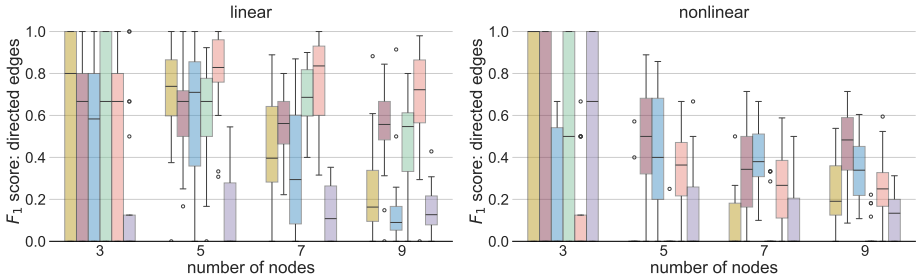


(*b*) Latent variables model

Figure 11: Empirical results on dense graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the SHD accuracy (lower is better).



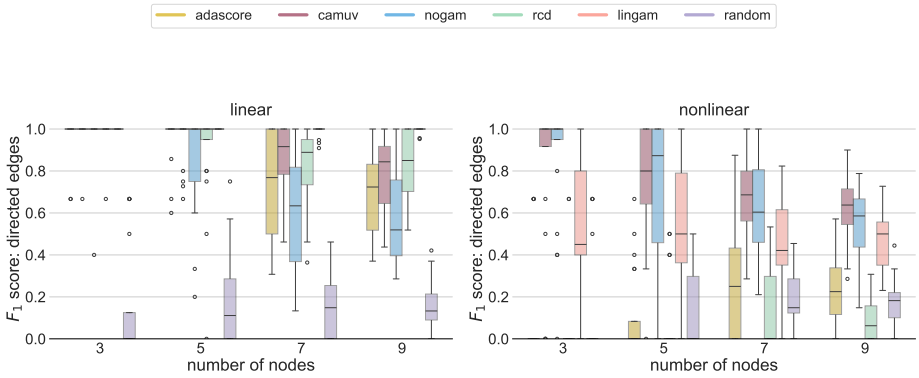
(*a*) Fully observable model



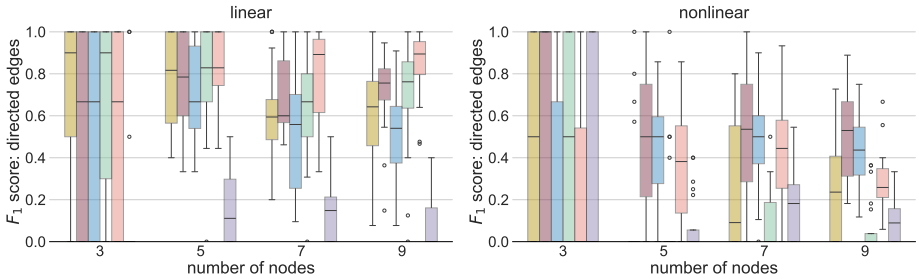
(*b*) Latent variables model

Figure 12: Empirical results on dense graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the F_1 score w.r.t. to the binary decision of whether there is an identifiable direct edge or not (the higher, the better).

graphs should be taken with care, as there is no strong reason to believe that they should mirror the benchmarked algorithms' behaviors in real-world settings, where often there is no prior knowledge about the structural causal model underlying available observations.

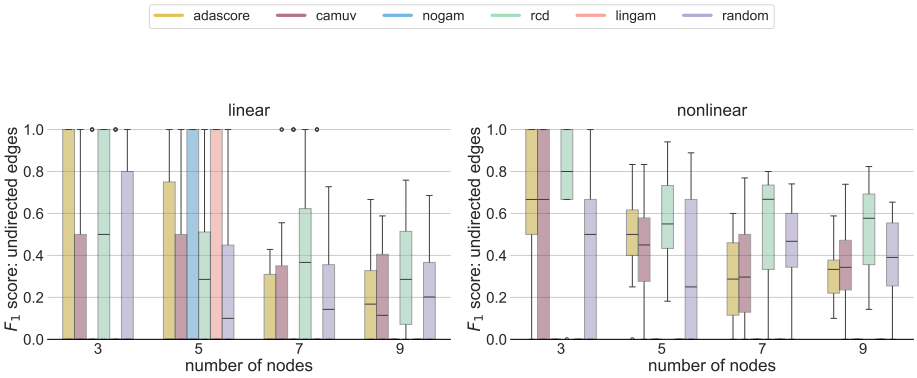


(*a*) Fully observable model

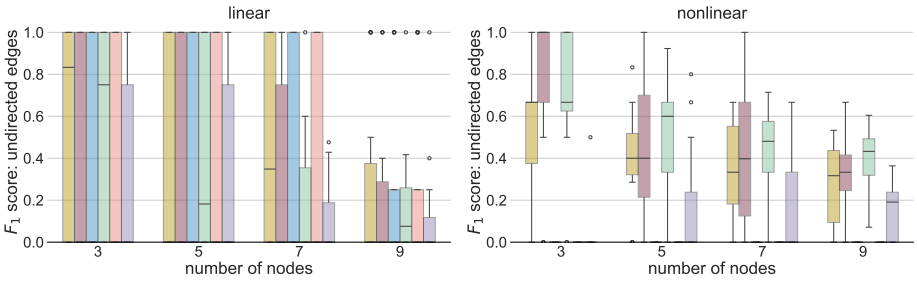


(*b*) Latent variables model

Figure 13: Empirical results on sparse graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the F_1 score w.r.t. to the binary decision of whether there is an identifiable direct edge or not (the higher, the better).

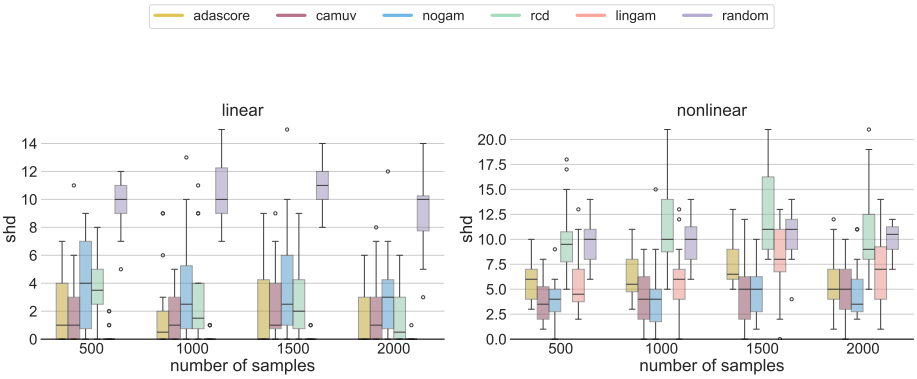


(*a*) Dense model

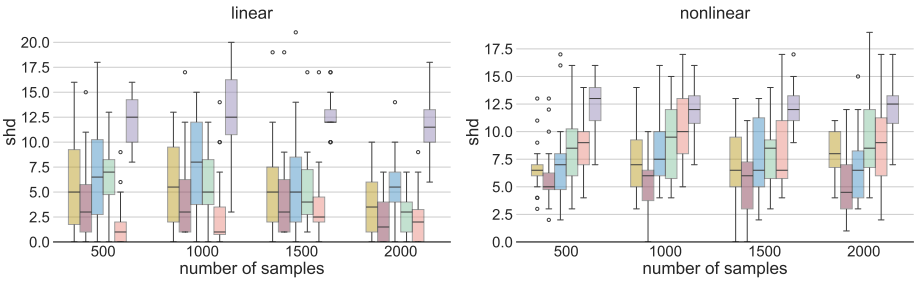


(*b*) Sparse model

Figure 14: Empirical results on sparse graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models. We report the F_1 score w.r.t. to the binary decision of whether there is an unidentifiable edge or not (the higher, the better).

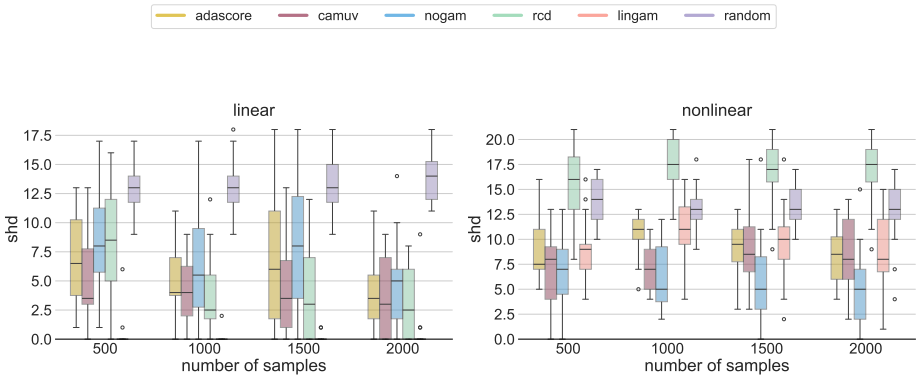


(*a*) Fully observable model

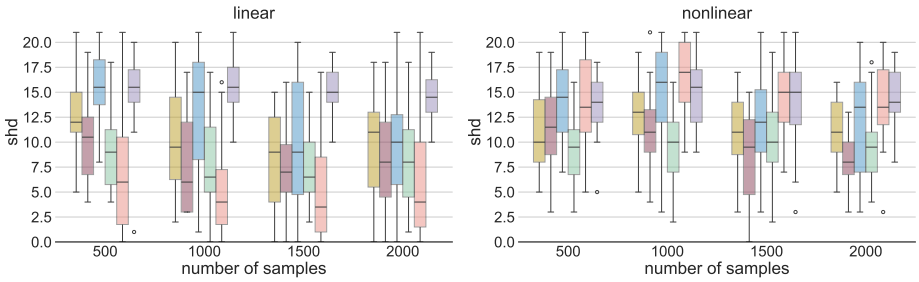


(*b*) Latent variables model

Figure 15: Empirical results on sparse graphs with different numbers of samples and seven nodes, on fully observable (no hidden variables) and latent variable models. We report the SHD accuracy (the lower, the better).

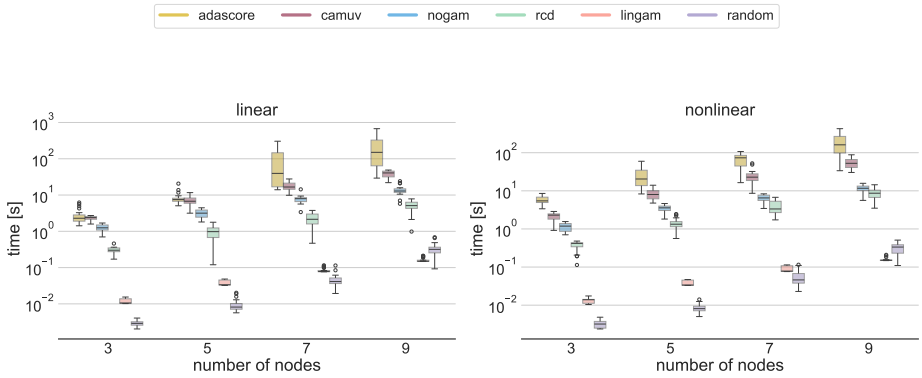


(*a*) Fully observable model

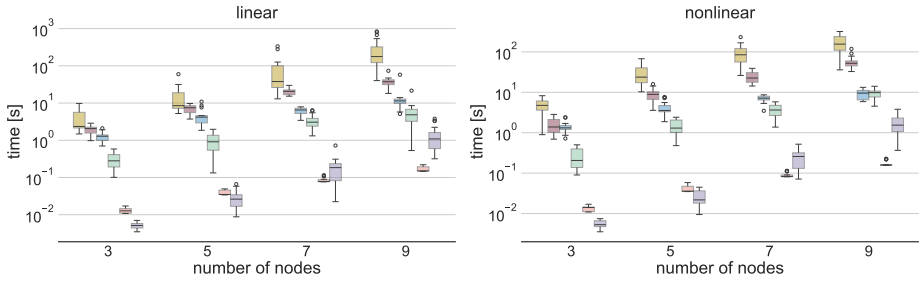


(*b*) Latent variables model

Figure 16: Empirical results on dense graphs with different numbers of samples and seven nodes, on fully observable (no hidden variables) and latent variable models. We report the SHD accuracy (the lower, the better).

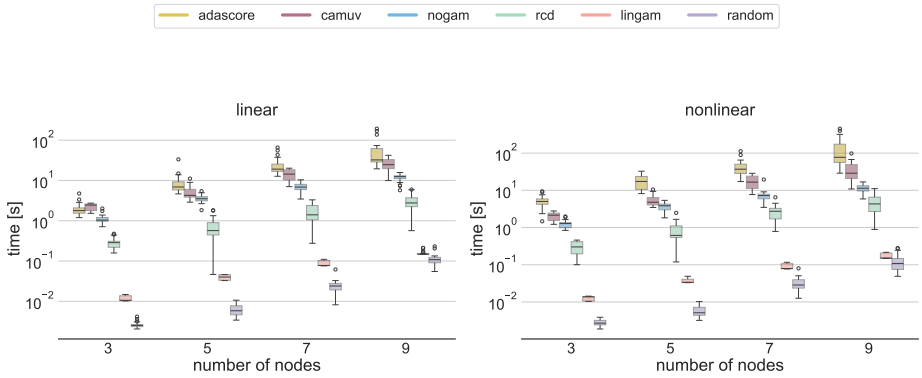


(*a*) Fully observable model

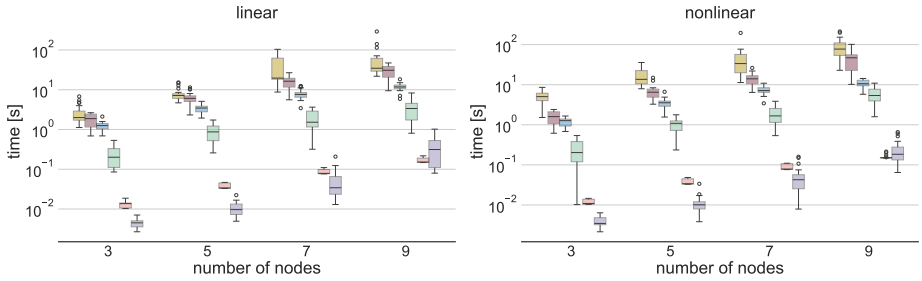


(*b*) Latent variables model

Figure 17: Runtime in seconds on dense graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models.

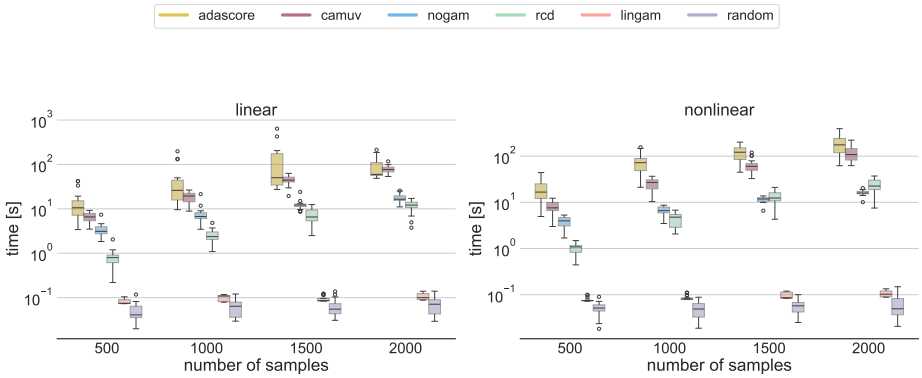


(*a*) Fully observable model

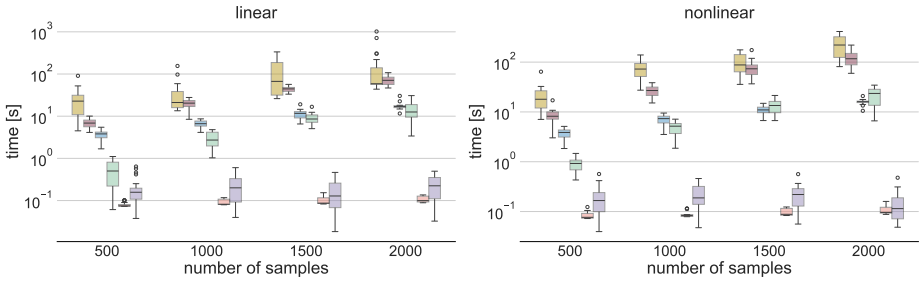


(*b*) Latent variables model

Figure 18: Runtime in seconds on sparse graphs with different numbers of nodes, on fully observable (no hidden variables) and latent variable models.

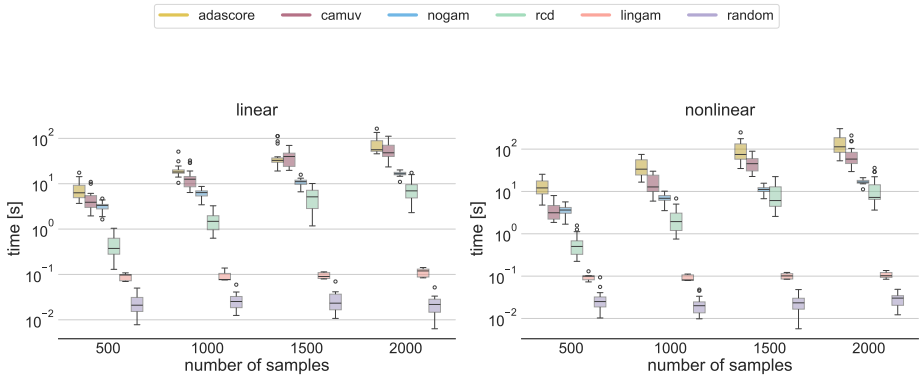


(*a*) Fully observable model

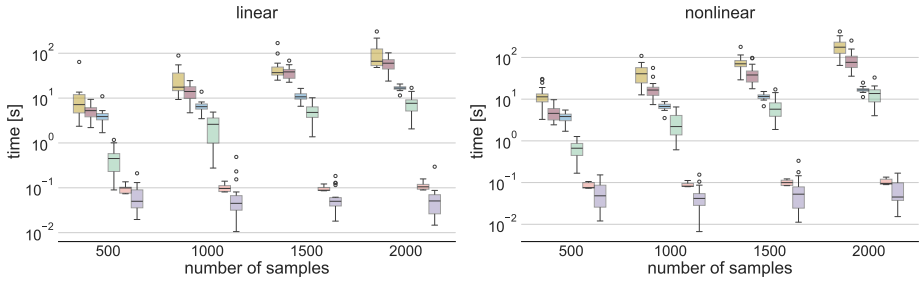


(*b*) Latent variables model

Figure 19: Runtime in seconds on dense graphs with different numbers of samples, on fully observable (no hidden variables) and latent variable models.



(*a*) Fully observable model



(*b*) Latent variables model

Figure 20: Runtime in seconds on sparse graphs with different numbers of samples, on fully observable (no hidden variables) and latent variable models.