LIKELIHOOD-BASED NONPARAMETRIC CAUSAL DISCOVERY UNDER LATENT CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning causal structure from observational data in the presence of latent confounding remains a fundamental challenge in causal discovery. Acyclic directed mixed graphs (ADMGs) provide a natural representation of this problem by indicating causal effects and latent confounding via directed and bidirected edges. A line of recent work has focused on score-based approaches that formulate ADMG learning as continuous optimization problems with acyclicity and further structural constraints enforced through differentiable functions. However, the methods that were developed either focus on linear structural equations or lack computational and statistical tractability due to structural equations involving explicitly introduced latent confounders with general non-linear effects. In this work, we seek to trade-off the advantages of these two perspectives and propose causal models that relate the observed variables through non-linear structural equations but summarize confounding in mere noise correlations. For the proposed model class, we establish structural identifiability of bow-free graphs under a residual faithfulness assumption. Moreover, given a known ancestral graph, we clarify that the model parameters are identifiable. With this insight, we develop a novel likelihood-based ADMG learning framework. Importantly, our approach does not introduce explicit latent variables and, thus, necessitates disentangling correlations among noise terms and linear parts of causal effects, which we address using a non-linearity regularizer. Experiments show that our method outperforms existing state-of-the-art baselines on synthetic datasets while remaining competitive on real-world datasets.

1 INTRODUCTION

Causal structure learning from observational data has wide applications, including computational biology (Zhang et al., 2023), finance (Ji et al., 2018), and deep learning (Yang et al., 2021). A large body of existing work focuses on causal structures represented by directed acyclic graphs (DAGs); compare (Spirtes & Zhang, 2019). Approaches that learn DAGs target the setting where all relevant variables are observed, an assumption that is often violated in practice. A fundamental challenge arises in the presence of latent confounders, i.e., there exist unobserved common causes that influence two or more observed variables. Such latent variables induce spurious dependencies between observed variables, which cannot be captured by a DAG over the observed variables alone. Consequently, ignoring latent confounding can lead to incorrect identification of causal relationships and hinder further causal inference (see e.g. Richardson & Spirtes, 2002).

Causal relationships in the presence of latent confounders can naturally be represented by acyclic directed mixed graphs (ADMGs). Traditional causal discovery methods that account for latent confounding, such as the fast causal inference (FCI) algorithm (Spirtes et al., 2000), recover a Markov equivalence class of ADMGs. However, in exploratory data analysis, it is also of interest to go beyond an equivalence class by leveraging additional assumptions encoded in restricted structural equation models (SEMs) to achieve structural identifiability (Peters et al., 2014; Maeda & Shimizu, 2021). Related work further restricts the considered ADMGs to ancestral or bow-free graphs (Wang & Drton, 2023). Formal definitions are provided in Section 2.1. Classically, structure learning algorithms search over combinatorial structures, which for score-based approaches with complex loss functions becomes computationally difficult when the number of variables becomes larger. Dong et al. (2026) consider structure identification under linear structural equation models with latent variables using a score-based greedy search method. As an alternative, recent work extends the idea of formulating continuous optimization problems, as in NOTEARS (Zheng et al., 2018), to ADMGs by encoding structural constraints, such as ancestrality and bow-freeness, as continuous differentiable functions (Bhattacharya et al., 2021). While this approach avoids discrete search, it is so far restricted to linear structural equations, and the resulting Markov equivalence class of graphs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

remains indistinguishable. Ng et al. (2024) explicitly model latent variables in linear structural equation models and also adopt an augmented Lagrangian approach for continuous optimization (Zheng et al., 2018), with general identifiability guarantees up to algebraic equivalence or Markov equivalence under an additional structural assumption. Ashman et al. (2023) considers non-linear direct causal relationships and introduces latent confounders with general non-linear effects explicitly. While this model achieves structural identifiability under a residual faithfulness condition, it does not study why residual faithfulness holds generically and leads to a complex learning objective and thus lacks computational and statistical tractability. In practice, we find that it is difficult to scale beyond small problem sizes. In contrast, Li et al. (2024) represent latent confounding implicitly through correlated Gaussian errors and establish structural identifiability under a sublinear growth condition on the structural functions. However, their method focuses only on recovering the directed causal structure and does not recover the bidirected edges representing latent confounding explicitly.

Contributions. In this work, we develop a novel nonparametric causal discovery framework under latent confounding, combining advances in differentiable constraints (Bhattacharya et al., 2021) and structural identifiability (Ashman et al., 2023; Maeda & Shimizu, 2021) to overcome the aforementioned limitations. Our contributions are summarized as follows:

- We study a nonparametric structural equation model in which direct causal relationships are modeled by non-linear functions, while latent confounders are represented via additive, correlated Gaussian noise terms. We define a residual faithfulness assumption and establish structural identifiability of bow-free graphs under this assumption without requiring the sublinear growth condition imposed by Li et al. (2024). Furthermore, we provide insights into why this assumption holds generically.
- Given a known graph structure, we clarify that the model parameters are identifiable when restricted to ancestral graphs.
- We introduce a non-linearity regularizer to enforce non-linear causal relationships and combine it with a differentiable acyclicity constraint (Bello et al., 2022) and constraints for ancestral and bow-free graphs (Bhattacharya et al., 2021). This leads to a likelihood-based optimization framework for ADMG discovery, inspired by the DAGMA algorithm (Bello et al., 2022), which employs a central path method to solve the constrained optimization problem.
- We empirically evaluate the proposed method on both synthetic and real-world datasets, demonstrating competitive performance compared to existing approaches. An anonymized implementation is available at `https://anonymous.4open.science/r/admg-code/`. The code will be publicly released upon acceptance.

Outline. Section 2 sets up the notation, introduces the nonparametric structural equation model with latent confounders, and reviews the definitions of ancestral and bow-free graphs. Section 3 introduces the residual faithfulness condition and explains why it holds generically. Under this condition, structural identifiability is established for bow-free graphs. Moreover, we obtain parameter identifiability for ancestral graphs. Section 4 reviews differentiable constraints that enforce ancestral and bow-free graphs, and introduces the resulting constrained optimization problem together with the proposed learning algorithm for sparse mixed graphs. In Section 5, we compare the performance of our method with state-of-the-art score-based baselines (Bhattacharya et al., 2021; Ashman et al., 2023) and the constraint-based FCI algorithm (Spirtes et al., 2000) in numerical experiments. Additional details on the experiments and proofs are provided in the Appendix.

2 Background

2.1 Ancestral and bow-free graphs

A mixed graph $G = (V, D, B)$ is a triple consisting of nodes $V = [d] := \{1, \dots, d\}$, and directed $D \subseteq V \times V$ and bidirected $B \subseteq V \times V$ edges. We also denote the directed edges $(v, w) \in D$ by $v \to w$ and the bidirected edges $(v, w) \in B$ by $v \leftrightarrow w$. Note that the bidirected edges have no orientation, that is, $v \leftrightarrow w \in B$ if and only if $w \leftrightarrow v \in B$. We assume that neither the directed part nor the bidirected part contain self-loops, that is, $v \to v \notin D$ and $v \leftrightarrow v \notin B$ for all $v \in V$. We denote by $\mathrm{pa}(v) = \{w \in V \mid w \to v \in D\}$ the parents of a node $v \in V$, and by $\mathrm{an}(v) = \{w \in V \mid \text{there exists a path } w \to \cdots \to v\}$ the ancestors of a node $v \in V$. For a graph $G = (V, D, B)$, we associate the directed adjacency matrix A_D with entries $(A_D)_{ij} = 1$ if $i \to j \in D$ and $(A_D)_{ij} = 0$ else. Similarly, we denote the bidirected adjacency matrix A_B with entries $(A_B)_{ij} = (A_B)_{ji} = 1$ if $i \leftrightarrow j \in B$, and $(A_B)_{ij} = 0$ else. The mixed graph $G = (V, D, B)$ is said to be acyclic if the directed path does not contain a cycle, that is, a path of the

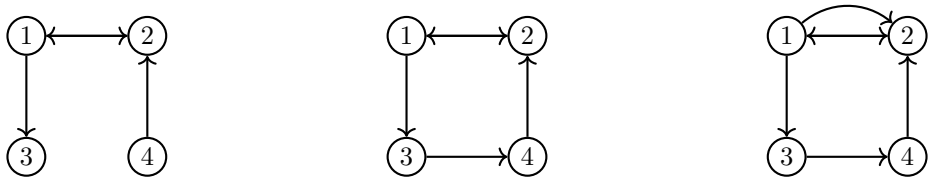


Figure 1: (a) An ADMG that is ancestral and bow-free. (b) An ADMG that is not ancestral but bow-free, since both the directed path $1 \to 3 \to 4 \to 2$ and the bidirected edge $1 \leftrightarrow 2$ exist. (c) An ADMG that is neither ancestral nor bow-free, since both $1 \to 2$ and $1 \leftrightarrow 2$ exist.

form $v \to \cdots \to v$. In this case, we also call G an acyclic directed mixed graph (ADMG). Moreover, an ADMG G is said to be ancestral if for any pair of nodes $v, w \in V$, a directed path $v \to \cdots \to w$ and bidirected edge $v \leftrightarrow w$ do not exist simultaneously in G. An ADMG is called bow-free if for any pair of nodes $v, w \in V$, the two edges $v \to w$ and $v \leftrightarrow w$ are not both in G. Note that ancestral graphs are a subset of bow-free graphs. Figure 1 illustrates ancestrality and bow-freeness.

2.2 Nonparametric structural equation models with latent confounding

Our approach is based on nonparametric structural equation models defined by ADMGs $G = (V, D, B)$, which we introduce now. Let SPD^d be the space of symmetric positive definite $d \times d$ matrices $\Sigma = (\sigma_{ij})$ and define $\mathrm{SPD}(G) \subseteq \mathrm{SPD}^d$ to be the subset of matrices such that $\sigma_{ij} = 0$ if $i \neq j$ and $i \leftrightarrow j \notin B$. Let $\mathbf{X} = (X_1, \ldots, X_d)^\top$ be a random vector indexed by the nodes $V = [d]$. We suppose that all variables are related by structural equations as

$$X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i, \tag{1}$$

where each f_i is a real-valued function and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d) \sim \mathcal{N}(0, \Sigma)$ is a noise vector with $\Sigma \in \mathrm{SPD}(G)$. Two noise terms $\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j$ are correlated if there is a bidirected edge $i \leftrightarrow j$, modeling the effect of unobserved confounding. Moreover, letting $\mathrm{pa}(i)$ denote the set of parents of i in G, the function f_i does not depend on X_k if $k \notin \mathrm{pa}(i)$, is nonconstant in each of its parent arguments, and is otherwise nonparametric, allowing for flexible, potentially non-linear dependencies among variables while encoding conditional independence structure through the graph G.

Formally, the structural equation model given by a graph $G = (V, D, B)$ is the set of distributions that can be generated via the mechanisms in Equation (1). To give a precise definition that allows to study structural identifiability, let $\mathcal{F}(G)$ be the set of functions $f = (f_i)_{i \in V}$ such that each $f_i : \mathbb{R}^{|\mathrm{pa}(i)|} \to \mathbb{R}$ depends only on the variables indexed by $\mathrm{pa}(i)$. Moreover, we assume that each f_i is measurable and twice differentiable for every $i \in V$. Furthermore, we define the mapping

$$h_f : \mathbb{R}^d \to \mathbb{R}^d, \quad \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \mapsto \begin{pmatrix} X_1 - f_1(\mathbf{X}_{\mathrm{pa}(1)}) \\ \vdots \\ X_d - f_d(\mathbf{X}_{\mathrm{pa}(d)}) \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_d \end{pmatrix}. \tag{2}$$

Since G is acyclic, after a topological ordering of the nodes, the map h_f is triangular. Therefore, it is bijective, with inverse h_f^{-1} obtained recursively by $X_i = \varepsilon_i + f_i(\mathbf{X}_{\mathrm{pa}(i)})$. Note that the inverse is continuous because each f_i is continuous.

Definition 2.1. *The structural equation model given by an ADMG $G = (V, D, B)$ is defined as the collection of joint distributions obtained as the pushforward measure of the probability measure of $\boldsymbol{\varepsilon}$ by the inverse of h_f, that is,*

$$\mathcal{M}(G) = \{\mathbb{P}^{\mathbf{X}} \in \mathcal{P}(\mathbb{R}^d) \mid \mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma) \text{ with } f \in \mathcal{F}(G) \text{ and } \Sigma \in \mathrm{SPD}(G)\},$$

where $\mathcal{P}(\mathbb{R}^d)$ is the collection of all distributions on $\mathbb{R}^d$.

In addition, for any distribution $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$ induced by some $f \in \mathcal{F}(G)$ and $\Sigma \in \mathrm{SPD}(G)$, we assume that, for each $i \in V$, the derivatives of f_i are square-integrable with respect to $\mathbb{P}^{\mathbf{X}}$ and the Hessian of f_i is integrable with respect to $\mathbb{P}^{\mathbf{X}}$.

3 IDENTIFIABILITY

We first establish structural identifiability for the models defined in the previous section. Specifically, given a distribution $\mathbb{P}^{\mathbf{X}}$ assumed to lie in some model $\mathcal{M}(G)$ for an unknown graph G, we establish conditions under which $\mathbb{P}^{\mathbf{X}}$ belongs to exactly one such model, allowing unique recovery of G. We impose two types of assumptions. First, we restrict to bow-free graphs, assuming $\mathbb{P}^{\mathbf{X}}$ lies in some $\mathcal{M}(G)$ with G bow-free. However, the models $\mathcal{M}(G)$ and $\mathcal{M}(G')$ for two distinct graphs G, G' may still have a small overlap. To address this, we further restrict to subsets $\tilde{\mathcal{M}}(G) \subseteq \mathcal{M}(G)$ obtained from functionals that satisfy a residual faithfulness condition, which ensures that $\tilde{\mathcal{M}}(G) \cap \tilde{\mathcal{M}}(G') = \emptyset$. Faithfulness conditions are imposed in any structure recovery result we are aware of; see e.g. Maeda & Shimizu (2021) and Ashman et al. (2023). Importantly, we argue that this condition is mild: most distributions in a model $\mathcal{M}(G)$ satisfy residual faithfulness conditions, so the excluded set $\mathcal{M}(G) \setminus \tilde{\mathcal{M}}(G)$ is indeed small.

We first introduce the necessary notation to define residual faithfulness. For a vector $\mathbf{x} \in \mathbb{R}^k$ and a subset $S \subseteq [k]$, we extend a given function $f : \mathbb{R}^{|S|} \to \mathbb{R}$ to another function $\tilde{f} : \mathbb{R}^k \to \mathbb{R}$ trivially via $\tilde{f}(\mathbf{x}) := f(\mathbf{x}_S)$. In this case, with a slight abuse of notation, we also write $f(\mathbf{x})$ for $\tilde{f}(\mathbf{x})$. We write x_j for the j-th coordinate and $\mathbf{x}_{-j}$ for the vector obtained by removing the j-th coordinate. Similarly, for a function $h : \mathbb{R}^k \to \mathbb{R}^l$, we write $h_j(\boldsymbol{\varepsilon})$ for the j-th coordinate of $h(\boldsymbol{\varepsilon})$ and $h_{-j}(\boldsymbol{\varepsilon})$ for the vector obtained by removing the j-th coordinate. For any function $f : \mathbb{R}^d \to \mathbb{R}$, define

$$\mathrm{Dep}(f) := \{i \in V \mid \text{the function } x_i \mapsto f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d) \text{ is non-constant}\}.$$

For a distribution $\mathbb{P}^{\mathbf{X}}$ on $\mathbb{R}^d$, define

$$\mathcal{H}(\mathbb{P}^{\mathbf{X}}) := \left\{ (h, \Sigma) \in C(\mathbb{R}^d; \mathbb{R}^d) \times \mathrm{SPD}^d \;\middle|\; \begin{array}{l} h \text{ is a bijection with continuous inverse,} \\ \text{and } h_* \mathcal{N}(0, \Sigma) = \mathbb{P}^{\mathbf{X}} \end{array} \right\}.$$

Said differently, if a random vector $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ and $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$, then we have $\mathbf{X} \overset{d}{=} h(\boldsymbol{\varepsilon})$ for every $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. A function $g : \mathbb{R} \to \mathbb{R}$ is non-linear if for all $a, b \in \mathbb{R}$, there exists $x \in \mathbb{R}$ such that $g(x) \neq ax + b$. Define

$$\mathcal{G} := \left\{ g \in C(\mathbb{R}^{d-1}, \mathbb{R}) \;\middle|\; \begin{array}{l} \text{the function } x_j \mapsto g(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_{d-1}) \\ \text{is either constant or non-linear for every } j \in [d-1] \end{array} \right\}.$$

To establish structural identifiability, inspired by (Ashman et al., 2023), we analyze the dependence structure of the residuals obtained from non-linear regressions of X_i onto $\mathbf{X}_{-i}$ for pairs of indices $i, j \in V$ with $i \neq j$ under different graph structures. A key difficulty is that the choice of regression functions may artificially remove dependencies between residuals, even when they share common or dependent noise components. Therefore, it is natural to restrict to pairs of regression functions that do not eliminate such intrinsic dependencies induced by the noise terms.

Definition 3.1. *Let $\mathbb{P}^{\mathbf{X}}$ be a distribution on $\mathbb{R}^d$, and let $\mathbf{X}$ be a random vector with distribution $\mathbb{P}^{\mathbf{X}}$. For any $i, j \in V$ with $i \neq j$, we say that a tuple (g_1, g_2) of functions $g_1, g_2 : \mathbb{R}^{d-1} \to \mathbb{R}$ satisfies the (i, j)-residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$ if the following holds for any $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$: Whenever there are $k, l \in V$ such that*

$$k \in \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}) - g_1(h_{-i}(\boldsymbol{\varepsilon}))) \text{ and } l \in \mathrm{Dep}(h_j(\boldsymbol{\varepsilon}) - g_2(h_{-j}(\boldsymbol{\varepsilon}))),$$

with $\mathrm{Cov}[\varepsilon_k, \varepsilon_l] \neq 0$, it holds that $(X_i - g_1(\mathbf{X}_{-i})) \not\perp\!\!\!\perp (X_j - g_2(\mathbf{X}_{-j}))$. Consequently, we define

$$\mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) := \left\{ (g_1, g_2) \in \mathcal{G} \;\middle|\; (g_1, g_2) \text{ satisfies the} (i, j)\text{-residual faithfulness condition w.r.t. } \mathbb{P}^{\mathbf{X}} \right\}.$$

Said differently, a tuple of functions (g_1, g_2) satisfies the (i, j)-residual faithfulness condition with respect to a distribution $\mathbb{P}^{\mathbf{X}}$ of a random vector $\mathbf{X}$ if the following holds for any $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$: First, note that $\mathbf{X} \overset{d}{=} h(\boldsymbol{\varepsilon})$ for $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Now, suppose that there exist $k, l \in V$ such that $(X_i - g_1(\mathbf{X}_{-i}))$ depends on ε_k and $(X_j - g_2(\mathbf{X}_{-j}))$ depends on ε_l, with $\mathrm{Cov}[\varepsilon_k, \varepsilon_l] \neq 0$. Then the residuals $(X_i - g_1(\mathbf{X}_{-i}))$ and $(X_j - g_2(\mathbf{X}_{-j}))$ must be dependent.

Assumption 3.1. *(Residual faithfulness assumption). Let $G = (V, D, B)$ be a bow-free mixed graph, and let $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$. Define*

$$\tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}}) := \{f \in \mathcal{F}(G) \mid (f_i, f_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) \text{ for all } i, j \in V \text{ with } i \neq j\}.$$

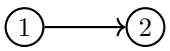
We define further

$$\tilde{\mathcal{M}}(G) = \left\{ \mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G) \mid \text{there is } f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}}) \text{ and } \Sigma \in \mathit{SPD}(G) \text{ s.t. } \mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma) \right\}.$$

Said differently, $\tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$ contains the non-linear functions $f \in \mathcal{F}(G)$ such that (f_i, f_j) satisfies the (i,j)-residual faithfulness condition for $\mathbb{P}^{\mathbf{X}}$ for all $i, j \in V$ with $i \neq j$. Moreover, $\tilde{\mathcal{M}}(G)$ is the set of distributions $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$ such that any $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ can be generated by functions in $\tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$.

In the following, we provide an example of bivariate mixed graph together with a pair (g_1, g_2) that satisfies the $(1,2)$-residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$ for any $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$.

Example 3.1. *Let G be the following graph with two nodes:*



Consider $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$, defined in reference to $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in SPD(G)$. The graph implies $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$ since there is no bidirected edge between 1 and 2. Assume the functional relationships $X_1 = \varepsilon_1$ and $X_2 = X_1^3 + \varepsilon_2$. Now, take $g_1 = 0$ and $g_2(X_1) = X_1^2$. Then $X_1 - g_1(X_2) = \varepsilon_1$ and $X_2 - g_2(X_1) = \varepsilon_1^3 - \varepsilon_1^2 + \varepsilon_2$. In particular, for $k = l = 1$, we have that $k \in \mathrm{Dep}(X_1 - g_1(X_2))$ and $l \in \mathrm{Dep}(X_2 - g_2(X_1))$ since both $X_1 - g_1(X_2)$ and $X_2 - g_2(X_1)$ depend on ε_1. Hence, to verify that (g_1, g_2) satisfies the $(1,2)$ residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$, we have to check that $(X_1 - g_1(X_2)) \not\perp\!\!\!\perp (X_2 - g_2(X_1))$. This follows from the fact that

$$\begin{aligned}\mathrm{Cov}[X_1 - g_1(X_2), X_2 - g_2(X_1)] &= \mathrm{Cov}\big[\varepsilon_1, \varepsilon_1^3\big] - \mathrm{Cov}\big[\varepsilon_1, \varepsilon_1^2\big] + \mathrm{Cov}[\varepsilon_1, \varepsilon_2] \\ &= \mathbb{E}[\varepsilon_1^4] - \mathbb{E}[\varepsilon_1]\mathbb{E}[\varepsilon_1^3] - \mathbb{E}[\varepsilon_1^3] + \mathbb{E}[\varepsilon_1]\mathbb{E}[\varepsilon_1^2] = \mathbb{E}[\varepsilon_1^4] > 0.\end{aligned}$$

Additional examples, including the multivariate case, are provided in Appendix A. The following Proposition shows that the residual faithfulness condition is satisfied for generic functions under a fixed distribution $\mathbb{P}^{\mathbf{X}}$ with bivariate $\mathbf{X}$.

Proposition 3.1. *Let $(X_1, X_2) \sim \mathbb{P}^{\mathbf{X}}$ be a fixed distribution with probability density $p_{\mathbf{X}}$. Consider arbitrary $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Assume further that h is a C^2-diffeomorphism. Let $g_1, g_2 \in C^2(\mathbb{R})$, and set $U := X_1 - g_1(X_2)$ and $V := X_2 - g_2(X_1)$. Define the map*

$$T : \mathbb{R}^2 \to \mathbb{R}^2,\ (x_1, x_2) \mapsto (u, v) := (x_1 - g_1(x_2), x_2 - g_2(x_1)),$$

and the set $S := \{(x_1, x_2) \in \mathbb{R}^2 \mid 1 - g_1'(x_2)g_2'(x_1) \neq 0\}$. For any fixed $(x_1^0, x_2^0) \in S$, let $(u^0, v^0) = T(x_1^0, x_2^0)$ be a point in the image. There exist open neighborhoods $N_{(x_1^0, x_2^0)}, M_{(u^0, v^0)} \subset \mathbb{R}^2$ such that for any $(u, v) \in M_{(u^0, v^0)}$, the preimage $(\varepsilon_1, \varepsilon_2) := (T \circ h)^{-1}(u, v)$ is well-defined. Moreover, assume that there exists $k, l \in [2]$ such that

$$k \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad l \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon}))),$$

and $\mathrm{Cov}[\varepsilon_k, \varepsilon_l] \neq 0$. If $U \perp\!\!\!\perp V$, then T must satisfy the following partial differential equation:

$$\begin{aligned}0 = \frac{1}{2\sigma_{12}\sigma_{21} - 2\sigma_{11}\sigma_{22}} \big\{ & 2\sigma_{22}\varepsilon_{1u}\varepsilon_{1v} - (\sigma_{12} + \sigma_{21})\varepsilon_{2v}\varepsilon_{1u} - (\sigma_{12} + \sigma_{21})\varepsilon_{1v}\varepsilon_{2u} + 2\sigma_{11}\varepsilon_{2u}\varepsilon_{2v} \\ & + 2\sigma_{22}\varepsilon_1\varepsilon_{1uv} - (\sigma_{12} + \sigma_{21})(\varepsilon_2\varepsilon_{1uv} + \varepsilon_1\varepsilon_{2uv}) + 2\sigma_{11}\varepsilon_2\varepsilon_{2uv}\big\} \\ & + \frac{\partial^2}{\partial u \partial v} \log |\det J_{(T \circ h)^{-1}}(u, v)|,\end{aligned} \tag{3}$$

where σ_{ij} denotes the ij-th entry of Σ, and $\varepsilon_{iu} := \frac{\partial}{\partial u}\varepsilon_i$ and $\varepsilon_{iuv} := \frac{\partial^2}{\partial u \partial v}\varepsilon_i$ for $i = 1, 2$. Moreover, at least one summand in Equation (3) is not identically zero on $M_{(u^0, v^0)}$.

Proposition 3.1 states that independence of the residuals $X_1 - g_1(X_2)$ and $X_2 - g_2(X_1)$ can only occur if the functions g_1 and g_2 satisfy a restrictive differential equation. Note that the elements of S are the points with nonzero Jacobian, ensuring local invertibility of the map T. Since g_1, g_2 are non-linear, their derivatives g_1', g_2' are non-constant and thus, for each $x_1 \in \mathbb{R}$, it holds that $(x_1, x_2) \in S$ for all but countably many $x_2 \in \mathbb{R}$. Moreover, if we fix u, then Equation (3) leads to a different ordinary differential equation in v. The residual faithfulness condition is violated only when the solution spaces of all ordinary differential equations overlap. While Proposition 3.1 clarifies that the residual faithfulness condition is generically satisfied in the bivariate case, Proposition B.1 in the Appendix illustrates this in the multivariate case.

Now, we return to our goal of showing structure identifiability of bow-free mixed graphs.

Definition 3.2 (Structure identifiability)**.** *For a bow-free mixed graph G_1 and a joint distribution $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_1)$, we say that G_1 is structurally identifiable from $\mathbb{P}^{\mathbf{X}}$ if there exists no other bow-free mixed graph $G_2 \neq G_1$ such that $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_2)$.*

In a bow-free mixed graph, if we consider two nodes $i, j \in V$, then we are in one of three cases: Either there is a bidirected edge between i and j, there is a directed edge between i and j, or there is no edge between i and j. The following lemma differentiates between the three cases by considering residuals from nonparametric regressions of X_i and X_j on all other variables in the graph.

Lemma 3.1 (Case 1). *Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \in B$ if and only if, for all $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$, it holds that $(X_i - g_i(\mathbf{X}_{-i})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-j}))$.*

Lemma 3.2 (Case 2). *Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \notin B$ if and only if there exists $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$ such that $(X_i - g_i(\mathbf{X}_{-(i,j)})) \perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)}))$.*

Note that, while functions $g \in \mathcal{G}$ are defined on the domain $\mathbb{R}^{d-1}$, we plug in $\mathbf{X}_{-(i,j)} \in \mathbb{R}^{d-2}$ by setting the remaining argument to zero, i.e., we abbreviate $g(\mathbf{X}_{-(i,j)}) = g(\mathbf{X}_{-(i,j)}, 0)$.

Lemma 3.3 (Case 3). *Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $j \to i \in D$ and $i \leftrightarrow j \notin B$ if and only if the following two conditions hold:*

$$\forall (g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) : (X_i - g_i(\mathbf{X}_{-(i,j)})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-j})),$$
$$\exists (g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) : (X_i - g_i(\mathbf{X}_{-i})) \perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})).$$

Combining the three lemmas gives our main theorem.

Theorem 3.1 (Structural identifiability of bow-free mixed graphs). *Let G be a bow-free mixed graph and let $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then G is structurally identifiable from $\mathbb{P}^{\mathbf{X}}$.*

Moreover, when restricting to ancestral graphs, we obtain that the structural functions f and the covariance matrix Σ are uniquely determined by the density $p^{\mathbf{X}}_{(f,\Sigma)}$ of a distribution $\mathbb{P}^{\mathbf{X}} \in M(G)$.

Theorem 3.2 (Parameter identifiability of ancestral mixed graphs). *Let $G = (V, D, B)$ be a mixed graph and let $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$. If G is ancestral, then the structural functions f and the covariance matrix Σ are identifiable. More precisely, let $p^{\mathbf{X}}_{(f,\Sigma)}$ denote the probability density corresponding to the distribution $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. If $p^{\mathbf{X}}_{(f,\Sigma)} = p^{\mathbf{X}}_{(\tilde{f},\tilde{\Sigma})}$, then $f = \tilde{f}$ almost surely and $\Sigma = \tilde{\Sigma}$.*

4 Optimization

We present a practical framework for Gaussian likelihood-based nonparametric ADMG discovery, abbreviated by GLL-ADMG. Throughout this section, we use i to index data points $\mathbb{X} = \{\mathbf{x}^i\}_{i=1}^n$ sampled independently from a distribution $\mathbb{P}^{\mathbf{X}}$. As in Zheng et al. (2020), we use neural networks to approximate the functions f_j that define a structural equation model. Specifically, for each f_j, consider a multilayer perceptron (MLP) with h hidden layers and a single activation function $\phi : \mathbb{R} \to \mathbb{R}$, applied elementwise, defined as $f_j^{\theta_j}(\mathbf{x}) = \phi(A^{(h)}\phi(\cdots A^{(2)}\phi(A^{(1)}\mathbf{x})))$, where $\mathbf{x} \in \mathbb{R}^d$ and $\theta_j = (A_j^{(1)}, \ldots, A_j^{(h)})$ is a tuple of parameter matrices of suitable sizes. Collect all parameters in $\theta = (\theta_1, \ldots, \theta_d)$ and define $f^\theta : \mathbb{R}^d \to \mathbb{R}^d$ by $f^\theta(x) = (f_1^{\theta_1}(x), \ldots, f_d^{\theta_d}(x))^\top$. Recall that we assume that the derivatives of each function f_j are square-integrable with respect to the observed distribution $\mathbb{P}^{\mathbf{X}}$. We denote the partial derivative of f_j with respect to its k-th coordinate by $\partial_k f_j$. As in Waxman et al. (2024), we define the weighted directed adjacency matrix $W_1(f) \in \mathbb{R}^{d \times d}$ with entries $W_1(f)_{kj} := \|\partial_k f_j\|_{L^2(\mathbb{P}^{\mathbf{X}})}$, and consequently $W_1(\theta)_{kj} := \left\|\partial_k f_j^\theta\right\|_{L^2(\mathbb{P}^{\mathbf{X}})}$. Furthermore, for a symmetric matrix Σ, we define the weighted bidirected adjacency matrix $W_2(\Sigma) := \Sigma - \mathbb{D}(\Sigma)$, where $\mathbb{D}(\Sigma)$ denotes the diagonal matrix containing the diagonal elements of Σ.

4.1 Differentiable constraints

In Theorem 3.1 we established that bow-free graphs are structure identifiable, and we obtained in Theorem 3.2 that we can additionally identify the parameters when restricting to ancestral graphs. In our estimation approach, we combine the log-determinant constraint developed by Bello et al. (2022)

that ensures acyclicity with the constraints for ancestrality and bow-freeness introduced by Bhattacharya et al. (2021). This gives the differentiable constraints

$$h^s(A_D, A_B) = -\log\det(sI_d - A_D) + d\log(s) + \text{sum}(e^{A_D} \circ A_B) = 0$$

for acyclicity and ancestrality, and

$$h^s(A_D, A_B) = -\log\det(sI_d - A_D) + d\log(s) + \text{sum}(A_D \circ A_B) = 0$$

for acyclicity and bow-freeness. Here, $s > 0$ is a hyperparameter, $A \circ B$ denotes the Hadamard matrix productfor two matrices A, B of the same dimension, and $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$ denotes the matrix exponential for a matrix A. Recall that A_D and A_B denote the directed and bidirected adjacency matrices of the graph $G = (V, D, B)$, respectively. In practice, we replace the binary adjacency matrices A_D and A_B with $W_1(\theta) \circ W_1(\theta)$ and $W_2(\Sigma) \circ W_2(\Sigma)$, respectively, to ensure positive entries. Note that, by Bello et al. (2022), the acyclicity constraint is defined on the domain $\mathbb{W}^s = \{W \in \mathbb{R}^{d\times d} \mid s > \rho(W \circ W)\}$, for any $s > 0$, where $\rho(W \circ W)$ denotes the spectral radius of $W \circ W$. Consequently, we obtain the constraint $h^s(\theta, \Sigma) := h^s(W_1(\theta) \circ W_1(\theta), W_2(\Sigma) \circ W_2(\Sigma))$.

4.2 Regularizers

To learn sparse graphs, we impose regularizers on both the directed and bidirected adjacency matrices. Recall that we assume that the derivatives of each function f_j are square-integrable. Similar to Waxman et al. (2024), we consider the L_1 sparsity regularizer for the weighted directed adjacency matrix $W_1(f)$ given by $\Omega_1(f) = \sum_{k,j=1}^{d} \|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})}$. Based on the data $\mathbb{X} = \{\mathbf{x}^i\}_{i=1}^n$ and the MLP approximation f^θ, we use the empirical estimate

$$\Omega_1^{\mathbb{X}}(f^\theta) = \sum_{k,j=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} |\partial_k f_j^\theta(\mathbf{x}^i)|\right).$$

In addition, we introduce a sparsity regularizer for the weighted bidirected adjacency matrix $W_2(\Sigma)$:

$$\Omega_2(\Sigma) = \frac{1}{2}\|W_2(\Sigma)\|_F := \frac{1}{2} \sum_{k,j=1}^{d} |W_2(\Sigma)|_{kj}.$$

Another ingredient to get structure identifiability in Theorem 3.1 is that we assume that each function f_j is non-linear in the parents. Inspired by Peebles et al. (2020), we enforce this via the regularizer

$$\mathcal{R}_{nl}^m(f) := \sum_{j,k=1}^{d} \|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})} \max\{0, m - \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}\},$$

where $H(f_j)$ is the Hessian with entries $H_{kl}(f_j)(x) = \frac{\partial^2 f_j(x)}{\partial x_k \partial x_l}$, and $m > 0$ is a hyperparameter. The intuition for the regularizer is as follows. If $k \in \text{pa}(j)$, then the term $\|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})}$ should be nonzero because f_j is non-constant in x_k. In this case, if the function is linear, then $\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} = 0$ and $\mathcal{R}_{nl}^m(f) > 0$. Hence, requiring that $\mathcal{R}_{nl}^m(f) = 0$ forces that f_j is non-linear in its parents. This is important as otherwise error correlations could be subsumed in linear functions f_j and vice versa.

Lemma 4.1. *For m small enough, more precisely,*

$$0 < m \leq \inf\{\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} \mid j, k \in V,\ \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} > 0\},$$

we have $\mathcal{R}_{nl}^m(f) > 0$ if and only if there exist $j, k \in V$ such that $k \in \text{Dep}(f_j(\mathbf{x}))$ and the map $x_k \mapsto f_j(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_{d-1})$ is linear.

In Lemma 4.1 we use the convention that $\inf \emptyset = \infty$. Now, based on the data $\mathbb{X} = \{\mathbf{x}^i\}_{i=1}^n$, we construct an empirical estimator of $\mathcal{R}_{nl}^m(f)$. Define $\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}^{\mathbb{X}} = \frac{1}{n}\sum_{i=1}^n |H_{kk}^{(j)}(\mathbf{x}^i)|$ and $\|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})}^{\mathbb{X}} = \frac{1}{n}\sum_{i=1}^n |\partial_k f_j(\mathbf{x}^i)|$. Then, the estimator is given by

$$\mathcal{R}_{nl}^m(f^\theta)^{\mathbb{X}} = \sum_{j,k=1}^{d} \left\|\partial_k f_j^\theta\right\|_{L_1(\mathbb{P}^{\mathbf{X}})}^{\mathbb{X}} \max\{0, m - \left\|H_{kk}(f_j^\theta)\right\|_{L_1(\mathbb{P}^{\mathbf{X}})}^{\mathbb{X}}\}.$$

4.3 OVERALL LEARNING PROBLEM

We consider the negative log-likelihood loss function derived from Gaussianity of the noise:

$$\mathcal{L}(\theta, \Sigma \mid \mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} [(\mathbf{x}^i - f^\theta(\mathbf{x}^i))^T \Sigma^{-1} (\mathbf{x}^i - f^\theta(\mathbf{x}^i))] + \log\det(\Sigma).$$

The overall learning problem is then given by

$$\min_{\theta, \Sigma} \mathcal{L}(\theta, \Sigma \mid \mathbb{X}) + \lambda_1 \Omega_1^{\mathbb{X}}(f^\theta) + \lambda_2 \Omega_2(\Sigma) + \lambda_{nl} \mathcal{R}_{nl}^m(f^\theta)^{\mathbb{X}} \text{ s.t. } h^s(\theta, \Sigma) = 0, \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_{nl}, m$ and s are positive hyperparameters. We propose the GLL-ADMG algorithm that uses the ADAM optimizer (Kingma & Ba, 2015) to iteratively solve the constrained optimization problem (4). It is inspired by the method proposed by Bello et al. (2022) for models that do not capture latent confounding. During optimization, we apply the Log-Cholesky parameterization to the covariance matrix Σ to ensure that it is symmetric positive definite (Pinheiro & Bates, 1996). We state the precise procedure in Algorithm 1 in the Appendix. After obtaining minimizers $\tilde{\theta}, \tilde{\Sigma}$, we compute the empricial version of the weighted directed adjacency matrix by $W_1^{\mathbb{X}}(\tilde{\theta})_{kj} = \left(\frac{1}{n} \sum_{i=1}^{n} (\partial_k f_j^{\tilde{\theta}}(\mathbf{x}^i))^2\right)^{1/2}$. Then, for small thresholds ω_1, ω_2, we obtain the sparse matrices $\widehat{W}_1 = W_1^{\mathbb{X}}(\tilde{\theta}) \circ \mathbf{1}(W_1^{\mathbb{X}}(\tilde{\theta}) > \omega_1)$ and $\widehat{\Sigma} = \tilde{\Sigma} \circ \mathbf{1}(\tilde{\Sigma} > \omega_2)$.

5 EXPERIMENTS

We study the empirical performance of our method GLL-ADMG on synthetic and real-world datasets. We evaluate our model under both ancestral and bow-free constraints, denoted by GLL-ADMG-A and GLL-ADMG-BF, respectively. We compare our approach against three ADMG learning baselines: N-ADMG (Ashman et al., 2023), FCI (Spirtes et al., 2000), and DCD (Bhattacharya et al., 2021), including its ancestral (DCD-A) and bow-free (DCD-BF) variants.

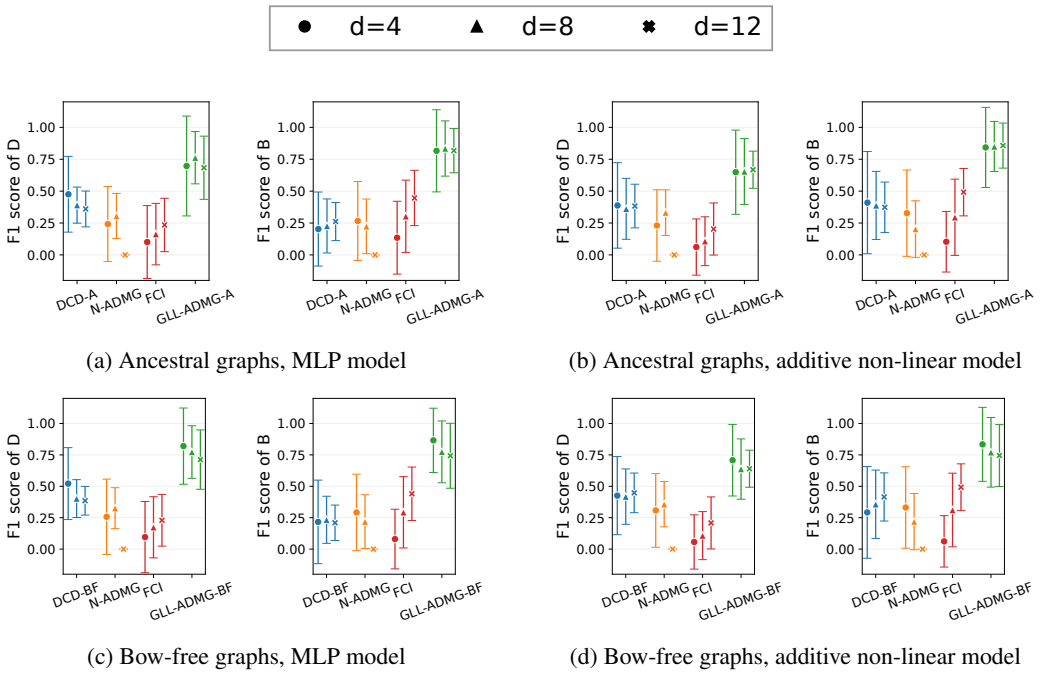
5.1 SYNTHETIC DATASETS

We generate ancestral and bow-free Erdős–Rényi-type ADMGs by sampling admissible directed and bidirected edges independently with probability $1/(d-1)$, retaining graphs containing both edge types. This yields graphs whose expected number of edges is bounded by d, containing both directed and bidirected components. For each graph G, we simulate data from a structural equation model $X_i = f_i(\mathbf{X}_{\text{pa}(i)}) + \varepsilon_i$ for all $i \in V$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \text{SPD}(G)$. We consider two data generation mechanisms for the functions f_i: a one-hidden-layer MLP model with randomly initialized weights, and an additive non-linear model, with causal relationships randomly picked from the following common non-linear functions: $g(x) = \exp(x), g(x) = \tanh(x), g(x) = \sin(x)$. Further simulation details are provided in Appendix E. Following prior work (Ashman et al., 2023; Bhattacharya et al., 2021), we consider graphs with $d \in \{4, 8, 12\}$. For each d, we randomly generate 10 ancestral and 10 bow-free ER mixed graphs. For each graph, we sample 10 datasets with 2000 samples under both the MLP and additive non-linear data generation mechanisms, yielding 100 simulations per setting. The maximum running time for each algorithm is set to 8 hours. In our experiments, N-ADMG frequently fails to complete within the allotted runtime or encounters numerical instability, particularly for $d = 12$, and partially for $d = 8$. Consequently, for $d = 8$, performance is computed using only the successful runs. For $d = 12$, we report an F1 score as 0. Appendix E reports detailed failure rates and running times, together with results from additional experiments involving non-Gaussian errors and non-bow-free graphs, which demonstrate the robustness of our method to model misspecification.

Figure 2 compares the performance of GLL-ADMG with the baselines when we restrict to ancestral and bow-free graphs. Overall, GLL-ADMG consistently outperforms the baselines across all datasets, highlighting its effectiveness relative to competing methods. Compared to Ashman et al. (2023), we attribute the discrepancy between the performance of DCD and N-ADMG in our experiments to differences in the graph and data generation setup. In particular, we focus on practically relevant sparse graphs, whereas Ashman et al. (2023) consider relatively dense graphs with node, edge, and confounder counts $(d, e, m) \in \{(4, 6, 2), (8, 20, 6), (12, 50, 10)\}$.

5.2 REAL DATASET

Finally, we apply our GLL-ADMG algorithm to the protein expression dataset in Sachs et al. (2005), one of the most widely used benchmarks in causal discovery (Brouillard et al., 2025). Since the



(a) Ancestral graphs, MLP model

(b) Ancestral graphs, additive non-linear model

(c) Bow-free graphs, MLP model

(d) Bow-free graphs, additive non-linear model

Figure 2: Causal discovery results on synthetic datasets generated using MLP and additive non-linear models with Gaussian exogenous noise, under ancestral and bow-free graph restrictions. Each subfigure reports F1 scores for directed (D) and bidirected (B) edges. Error bars represent standard errors over 100 simulations.

ground-truth graph is a directed acyclic graph (DAG), we introduce latent confounding by treating one observed variable as unobserved. In particular, the variable PKC is the only root node acting as a common cause of multiple downstream variables. We therefore remove PKC and introduce bidirected edges between every pair of its observed children. This yields an ADMG over the remaining variables. The ground-truth DAG and the resulting ADMG are shown in the Appendix E. The resulting ground-truth ADMG is neither ancestral nor bow-free, and thus, it does not satisfy the assumptions we impose for identifiability. The observed data are obtained by removing the PKC column from the original dataset, and we only consider a subsample of 2000 to reduce computational cost.

Due to the numerical instability of N-ADMG, we compare GLL-ADMG-A and DCD-A with FCI. For directed edges, baseline methods achieve higher F1 scores: FCI and DCD-A both obtain 0.24, compared to 0.15 for GLL-ADMG-A. However, for bidirected edges, all baselines fail to identify any edges (F1 $= 0$), whereas our method achieves an F1 score of 0.36 with a precision of 1, demonstrating its ability to capture latent confounding. Moreover, the ground-truth ADMG contains 10 nodes with 12 directed and 10 bidirected edges, making it denser than the synthetic datasets. For a fair comparison, we keep the same hyperparameters as in the synthetic setting, which partially explains the lower directed F1 score of GLL-ADMG.

6 Conclusion

We study nonparametric structural equation models under latent confounding and establish structural identifiability of bow-free graphs under a mild residual faithfulness assumption, as well as parameter identifiability for ancestral graphs. Building on these results, we develop GLL-ADMG, a log-likelihood-based framework for learning mixed graphs that jointly enforces acyclicity and structural constraints. GLL-ADMG outperforms the baselines in our simulations and identifies latent structure missed by the baselines in the real-data example. Future work includes generalizing structure identifiability beyond bow-freeness and parameter identifiability beyond ancestral graphs, developing computationally efficient alternatives to the Hessian-based nonlinearity regularizer, and allowing more flexible error models while preserving identifiability and tractability.

AI USE STATEMENT

In this work, we used generative AI tools for editing to improve readability. We have not used generative AI tools for any other task such as generating synthetic data sets, helping developing conceptual frameworks, formulating mathematical claims, assisting in any form in the proofs, proposing or refining hypotheses, designing or providing feedback on the methodology or on the experiments, implementing our method, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Assisting with translation is not applicable to this work. We have reviewed all AI-assisted work. All edits assisted by generative AI were manually reviewed by the authors to ensure that they did not alter the intended meaning, technical claims, or scientific content of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our work. The proposed method and theoretical results are described in detail in Sections 2, 3, and 4, with complete mathematical proofs provided in Appendix B. The experimental setup, including data generation procedures, algorithm design, and performance evaluation, is described in Section 5 and Appendix C, with further details on the numerical experiments and hyperparameter choices provided in Appendix E. Additional implementation details, including computational resources, are provided in Appendix E. The source code for reproducing the experiments is available through the anonymous GitHub repository linked in Section 1.

REFERENCES

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural ADMG learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 2314–2322. PMLR, 2021.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad K, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. In *Causal Learning and Reasoning, Lausanne, Switzerland, 7-9 May 2025*, Proceedings of Machine Learning Research, pp. 834–873. PMLR, 2025.

Xinshuai Dong, Ignavier Ng, Haoyue Dai, Jiaqi Sun, Xiangchen Song, Peter Spirtes, and Kun Zhang. Score-based greedy search for structure identification of partially observed causal models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qiang Ji, Elie Bouri, Rangan Gupta, and David Roubaud. Network causality structures among bitcoin and other financial assets: A directed acyclic graph approach. *The Quarterly Review of Economics and Finance*, 70:203–213, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Chunlin Li, Xiaotong Shen, and Wei Pan. Nonlinear causal discovery with confounders. *J. Amer. Statist. Assoc.*, 119(546):1205–1214, 2024.

Yurou Liang, Oleksandr Zadorozhnyi, and Mathias Drton. Kernel-based differentiable learning of non-parametric directed acyclic graphical models. In *International Conference on Probabilistic Graphical Models, De Lindenberg, Nijmegen, the Netherlands, 11-13 September 2024*, Proceedings of Machine Learning Research, pp. 253–272. PMLR, 2024.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, Proceedings of Machine Learning Research, pp. 97–106. AUAI Press, 2021.

Ignavier Ng, Xinshuai Dong, Haoyue Dai, Biwei Huang, Peter Spirtes, and Kun Zhang. Score-based causal discovery of latent variable causal models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37531–37552. PMLR, 2024.

William S. Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, Lecture Notes in Computer Science, pp. 581–597. Springer, 2020.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

José C. Pinheiro and Douglas M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

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

Peter Spirtes and Kun Zhang. Search for causal models. In *Handbook of graphical models*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pp. 439–469. CRC Press, Boca Raton, FL, 2019.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research (JMLR)*, 24:Paper No. [271], 61, 2023.

Daniel Waxman, Kurt Butler, and Petar M Djurić. Dagma-DCE: Interpretable, non-parametric differentiable causal discovery. *IEEE Open Journal of Signal Processing*, 2024.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9593–9602. Computer Vision Foundation / IEEE, 2021.

Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P. Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5 (10):1066–1075, 2023.

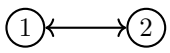
Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31:9492–9503, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

A EXAMPLES OF THE RESIDUAL FAITHFULNESS CONDITION

To illustrate the residual faithfulness condition, we consider several mixed graphs G together with a corresponding distribution $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$. For each graph, we give a simple example of a pair (g_1, g_2) that satisfies the (i,j)-residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$ for each $i \neq j$.

Example A.1. *1. Let G be the following graph with two nodes.*



Consider $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in SPD(G)$. The graph implies $\varepsilon_1 \not\perp\!\!\!\perp \varepsilon_2$ since there is a bidirected edge between 1 and 2. For this example, we take $g_1 = g_2 = 0$. Then, we have $(X_1 - g_1(X_2)) = \varepsilon_1$, $(X_2 - g_2(X_1)) = \varepsilon_2$. In particular, we have that $1 \in \mathrm{Dep}(X_1 - g_1(X_2))$ and $2 \in \mathrm{Dep}(X_2 - g_2(X_1))$. Since $\varepsilon_1 \not\perp\!\!\!\perp \varepsilon_2$, it follows that

$$(X_1 - g_1(X_2)) \not\perp\!\!\!\perp (X_2 - g_2(X_1)).$$

That is, (g_1, g_2) satisfies the $(1,2)$-residual faithfulness condition w.r.t. $\mathbb{P}^{\mathbf{X}}$.

2. Let G be the following graph with three nodes.



Consider $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in SPD(G)$. The graph implies $\varepsilon_1 \perp\!\!\!\perp (\varepsilon_2, \varepsilon_3)$, and $\varepsilon_2 \not\perp\!\!\!\perp \varepsilon_3$. Assume the functional relationships $X_1 = \varepsilon_1, X_2 = X_1^2 + \varepsilon_2$, and $X_3 = \varepsilon_3$. For this example, we take $g_1 = 0, g_2(X_1, X_3) = X_1^2 - X_1$, and $g_3 = 0$. Then, we have $X_1 - g_1(\mathbf{X}_{-1}) = \varepsilon_1, X_2 - g_2(\mathbf{X}_{-2}) = (X_1^2 + \varepsilon_2) - (X_1^2 - X_1) = \varepsilon_1 + \varepsilon_2, X_3 - g_3(\mathbf{X}_{-3}) = \varepsilon_3$. In particular, we have that $1 \in \mathrm{Dep}(X_1 - g_1(\mathbf{X}_{-1}))$ and $1 \in \mathrm{Dep}(X_2 - g_2(\mathbf{X}_{-2}))$. Moreover, $2 \in \mathrm{Dep}(X_2 - g_2(\mathbf{X}_{-2}))$ and $3 \in \mathrm{Dep}(X_3 - g_3(\mathbf{X}_{-3}))$. Furthermore, it holds that

$$\begin{aligned} &Cov[X_1 - g_1(\mathbf{X}_{-1}), X_2 - g_2(\mathbf{X}_{-2})] = Cov[\varepsilon_1, \varepsilon_1] \neq 0 \\ &Cov[X_2 - g_2(\mathbf{X}_{-2}), X_3 - g_3(\mathbf{X}_{-3})] = Cov[\varepsilon_2, \varepsilon_3] \neq 0. \end{aligned}$$

Therefore, we obtain

$$X_1 - g_1(\mathbf{X}_{-1}) \not\perp\!\!\!\perp X_2 - g_2(\mathbf{X}_{-2}), \quad X_2 - g_2(\mathbf{X}_{-2}) \not\perp\!\!\!\perp X_3 - g_3(\mathbf{X}_{-3}).$$

That is, (g_1, g_2) satisfies the $(1,2)$-residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$, and (g_2, g_3) satisfies $(2,3)$-residual faithfulness condition with respect to $\mathbb{P}^{\mathbf{X}}$.

B PROOFS

To prove the Proposition 3.1, we use the following lemma.

Lemma B.1. *Let $g_1, g_2 \in C^2(\mathbb{R})$. Define the map*

$$T : \mathbb{R}^2 \to \mathbb{R}^2,\ (x_1, x_2) \mapsto (u, v) := (x_1 - g_1(x_2), x_2 - g_2(x_1)),$$

and the set

$$S := \{(x_1, x_2) \in \mathbb{R}^2 \mid 1 - g_1'(x_2) g_2'(x_1) \neq 0\}.$$

Then, for any fixed $(x_1^0, x_2^0) \in S$, let $(u^0, v^0) = T(x_1^0, x_2^0)$, there exists open neighborhoods $N_{(x_1^0, x_2^0)}, M_{(u^0, v^0)} \subset \mathbb{R}^2$ such that $T : N_{(x_1^0, x_2^0)} \to M_{(u^0, v^0)}$ is a C^2-diffeomorphism.

Proof. The Jacobian of T is

$$J_T(x_1, x_2) = \begin{pmatrix} 1 & -g_1'(x_2) \\ -g_2'(x_1) & 1 \end{pmatrix}.$$

Hence, we have

$$\det(J_T(x_1, x_2)) = 1 - g_1'(x_2) g_2'(x_1).$$

Let $(x_1^0, x_2^0) \in S$, then $\det(J_T(x_1^0, x_2^0)) \neq 0$. Since $g_1, g_2 \in C^2(\mathbb{R})$, it follows that $T \in C^2(\mathbb{R}^2, \mathbb{R}^2)$. Therefore, by the inverse function theorem, letting $(u^0, v^0) = T(x_1^0, x_2^0)$, there exists open neighborhoods $N_{(x_1^0, x_2^0)}, M_{(u^0, v^0)} \subset \mathbb{R}^2$ such that $T : N_{(x_1^0, x_2^0)} \to M_{(u^0, v^0)}$ is a C^2-diffeomorphism. □

Proposition 3.1. Let $(X_1, X_2) \sim \mathbb{P}^{\mathbf{X}}$ be a fixed distribution with probability density $p_{\mathbf{X}}$. Consider arbitrary $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Assume further that h is a C^2-diffeomorphism. Let $g_1, g_2 \in C^2(\mathbb{R})$, and denote $U := X_1 - g_1(X_2)$ and $V := X_2 - g_2(X_1)$. Define the map

$$T : \mathbb{R}^2 \to \mathbb{R}^2,\ (x_1, x_2) \mapsto (u, v) := (x_1 - g_1(x_2), x_2 - g_2(x_1)),$$

and the set $S := \{(x_1, x_2) \in \mathbb{R}^2 \mid 1 - g_1'(x_2)g_2'(x_1) \neq 0\}$. For any fixed $(x_1^0, x_2^0) \in S$, let $(u^0, v^0) = T(x_1^0, x_2^0)$ be a point in the image. There exist open neighborhoods $N_{(x_1^0, x_2^0)}, M_{(u^0, v^0)} \subset \mathbb{R}^2$ such that for any $(u, v) \in M_{(u^0, v^0)}$, the preimage $(\varepsilon_1, \varepsilon_2) := (T \circ h)^{-1}(u, v)$ is well-defined. Moreover, assume that there exists $k, l \in [2]$ such that

$$k \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad l \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon}))),$$

and $\mathrm{Cov}[\varepsilon_k, \varepsilon_l] \neq 0$. If $U \perp\!\!\!\perp V$, then T must satisfy the following partial differential equation:

$$\begin{aligned} 0 = \frac{1}{2\sigma_{12}\sigma_{21} - 2\sigma_{11}\sigma_{22}} \big\{ & 2\sigma_{22}\varepsilon_{1u}\varepsilon_{1v} - (\sigma_{12} + \sigma_{21})\varepsilon_{2v}\varepsilon_{1u} - (\sigma_{12} + \sigma_{21})\varepsilon_{1v}\varepsilon_{2u} + 2\sigma_{11}\varepsilon_{2u}\varepsilon_{2v} \\ & + 2\sigma_{22}\varepsilon_1\varepsilon_{1uv} - (\sigma_{12} + \sigma_{21})(\varepsilon_2\varepsilon_{1uv} + \varepsilon_1\varepsilon_{2uv}) + 2\sigma_{11}\varepsilon_2\varepsilon_{2uv} \big\} \\ & + \frac{\partial^2}{\partial u \partial v} \log |\det J_{(T \circ h)^{-1}}(u, v)|, \end{aligned} \tag{3}$$

where σ_{ij} denotes the ij-th entry of Σ, and $\varepsilon_{iu} := \frac{\partial}{\partial u}\varepsilon_i$ and $\varepsilon_{iuv} := \frac{\partial^2}{\partial u \partial v}\varepsilon_i$ for $i = 1, 2$. Moreover, at least one summand in Equation (3) is not identically zero on $M_{(u^0, v^0)}$.

Proof. By Lemma B.1, the map $T : N_{(x_1^0, x_2^0)} \to M_{(u^0, v^0)}$ is a C^2-diffeomorphism. Since h is also a C^2-diffeomorphism, the composition $T \circ h$ is a local C^2-diffeomorphism on $h^{-1}(N_{(x_1^0, x_2^0)})$. Thus, for any $(u, v) \in M_{(u^0, v^0)}$, the preimage $(\varepsilon_1, \varepsilon_2) := (T \circ h)^{-1}(u, v)$ is well-defined. The first and second derivatives $\varepsilon_{iu}, \varepsilon_{iuv}$ are also well-defined for $i = 1, 2$.

By the density transformation formula, it holds that for all $(u, v) \in M_{(u^0, v^0)}$, we obtain

$$p_{(U,V)}(u, v) = p_{(\varepsilon_1, \varepsilon_2)}((T \circ h)^{-1}(u, v)) |\det J_{(T \circ h)^{-1}}(u, v)|.$$

If $U \perp\!\!\!\perp V$, then we have

$$\frac{\partial^2}{\partial u \partial v} \log p_{(U,V)}(u, v) = \frac{\partial^2}{\partial u \partial v}(\log p_U(u) + \log p_V(v)) = 0.$$

Therefore, we obtain

$$0 = \frac{\partial^2}{\partial u \partial v}\pi_{(\varepsilon_1, \varepsilon_2)}((T \circ h)^{-1}(u, v)) + \frac{\partial^2}{\partial u \partial v} \log |\det J_{(T \circ h)^{-1}}(u, v)|.$$

where $\pi_{(\varepsilon_1, \varepsilon_2)}$ denotes the log-density of $(\varepsilon_1, \varepsilon_2)$. Since

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix},$$

we obtain that

$$\pi_{(\varepsilon_1, \varepsilon_2)}(\varepsilon_1, \varepsilon_2) = -\frac{1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21})}(\varepsilon_1^2\sigma_{22} - (\sigma_{12} + \sigma_{21})\varepsilon_1\varepsilon_2 + \varepsilon_2^2\sigma_{11}) + c$$

with some constant $c \in \mathbb{R}$. Consequently, it holds that

$$\begin{aligned} 0 = \frac{1}{2\sigma_{12}\sigma_{21} - 2\sigma_{11}\sigma_{22}} \big\{ & 2\sigma_{22}\varepsilon_{1u}\varepsilon_{1v} \qquad \text{(I)} \\ & - (\sigma_{12} + \sigma_{21})\varepsilon_{2v}\varepsilon_{1u} \qquad \text{(II)} \\ & - (\sigma_{12} + \sigma_{21})\varepsilon_{1v}\varepsilon_{2u} \qquad \text{(III)} \\ & + 2\sigma_{11}\varepsilon_{2u}\varepsilon_{2v} \qquad \text{(IV)} \\ & + 2\sigma_{22}\varepsilon_1\varepsilon_{1uv} - (\sigma_{12} + \sigma_{21})(\varepsilon_2\varepsilon_{1uv} + \varepsilon_1\varepsilon_{2uv}) + 2\sigma_{11}\varepsilon_2\varepsilon_{2uv} \big\} \\ & + \frac{\partial^2}{\partial u \partial v} \log |\det J_{(T \circ h)^{-1}}(u, v)|. \end{aligned} \tag{3}$$

Moreover, it holds that

$$\begin{pmatrix} \varepsilon_{1u} & \varepsilon_{1v} \\ \varepsilon_{2u} & \varepsilon_{2v} \end{pmatrix} := J_{(T\circ h)^{-1}}(u,v) = J^{-1}_{T\circ h}(\varepsilon_1,\varepsilon_2) = \frac{1}{\det(J_{T\circ h})}\begin{pmatrix} v_{\varepsilon_2} & -u_{\varepsilon_2} \\ -v_{\varepsilon_1} & u_{\varepsilon_1} \end{pmatrix}.$$

Consider the case

$$1 \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad 1 \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon}))),$$

then we have $u_{\varepsilon_1}, v_{\varepsilon_1} \not\equiv 0$, which implies that $\varepsilon_{2u}, \varepsilon_{2v} \not\equiv 0$. Consequently, term (IV) in (3) is not identically zero.

For the case

$$1 \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad 2 \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon})))$$

and $Cov[\varepsilon_1, \varepsilon_2] \neq 0$, then we have $u_{\varepsilon_1}, v_{\varepsilon_2} \not\equiv 0$, which implies that $\varepsilon_{1u}, \varepsilon_{2v} \not\equiv 0$. Consequently, term (II) in (3) is not identically zero. The remaining cases,

$$2 \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad 1 \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon})))$$

and

$$2 \in \mathrm{Dep}(h_1(\boldsymbol{\varepsilon}) - g_1(h_2(\boldsymbol{\varepsilon}))), \quad 2 \in \mathrm{Dep}(h_2(\boldsymbol{\varepsilon}) - g_2(h_1(\boldsymbol{\varepsilon})))$$

follow analogously. □

Proposition B.1. *Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ be a fixed distribution, and let $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$ be such that $\mathbf{X} \overset{d}{=} h(\boldsymbol{\varepsilon})$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For any $i, j \in V$ with $i \neq j$, if there exist $k, l \in V$ such that*

$$k \in \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}) - g_1(h_{-i}(\boldsymbol{\varepsilon}))), \quad l \in \mathrm{Dep}(h_j(\boldsymbol{\varepsilon}) - g_2(h_{-j}(\boldsymbol{\varepsilon}))),$$

and $Cov[\varepsilon_k, \varepsilon_l] \neq 0$, then there exist subsets $K_1, K_2, I, J \subset V$ and non-constant functions $a(\cdot), b(\cdot)$ such that

$$X_i - g_1(\mathbf{X}_{-i}) = a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I) \quad X_j - g_2(\mathbf{X}_{-j}) = b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J),$$

with $Cov[\boldsymbol{\varepsilon}_I, \boldsymbol{\varepsilon}_J] = 0$ and $Cov[\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_{K_2}] \neq 0$. Let $K := K_1 \cup K_2$, and denote

$$m_a(\boldsymbol{\varepsilon}_K) := \mathbb{E}[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I)|\boldsymbol{\varepsilon}_K], \quad m_b(\boldsymbol{\varepsilon}_K) := \mathbb{E}[b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K].$$

Let $L^2(\Omega, \mathcal{A}, \mathbb{P}) := \{X : \Omega \to \mathbb{R} \mid \mathbb{E}[X^2] < \infty\}$, with inner product $\langle X, Y\rangle_{\mathbb{P}} := \mathbb{E}[XY]$. Then

$$Cov[X_i - g_1(\mathbf{X}_{-i}), X_j - g_2(\mathbf{X}_{-j})] = 0$$

if and only if $m_a(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_a(\boldsymbol{\varepsilon}_K)]$ and $m_b(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_b(\boldsymbol{\varepsilon}_K)]$ are orthogonal in $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

Proof. Since $\boldsymbol{\varepsilon}_I \perp\!\!\!\perp \boldsymbol{\varepsilon}_J$, we obtain that

$$a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I) \perp\!\!\!\perp b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K$$

and hence

$$Cov[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I), b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K] = 0.$$

By the law of total covariance, it holds that

$$\begin{aligned} &Cov[X_i - g_1(\mathbf{X}_{-i}), X_j - g_2(\mathbf{X}_{-j})] \\ &= Cov[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I), b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)] \\ &= \mathbb{E}[Cov[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I), b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K]] + Cov[\mathbb{E}[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I)|\boldsymbol{\varepsilon}_K], \mathbb{E}[b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K]] \\ &= Cov[\mathbb{E}[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I)|\boldsymbol{\varepsilon}_K], \mathbb{E}[b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K]]. \end{aligned}$$

Denote

$$m_a(\boldsymbol{\varepsilon}_K) := \mathbb{E}[a(\boldsymbol{\varepsilon}_{K_1}, \boldsymbol{\varepsilon}_I)|\boldsymbol{\varepsilon}_K], \quad m_b(\boldsymbol{\varepsilon}_K) := \mathbb{E}[b(\boldsymbol{\varepsilon}_{K_2}, \boldsymbol{\varepsilon}_J)|\boldsymbol{\varepsilon}_K],$$

then we have

$$\begin{aligned} Cov[X_i - g_1(\mathbf{X}_{-i}), X_j - g_2(\mathbf{X}_{-j})] &= Cov[m_a(\boldsymbol{\varepsilon}_K), m_b(\boldsymbol{\varepsilon}_K)] \\ &= \mathbb{E}[(m_a(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_a(\boldsymbol{\varepsilon}_K)])(m_b(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_b(\boldsymbol{\varepsilon}_K)])] \\ &= \langle (m_a(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_a(\boldsymbol{\varepsilon}_K)]), (m_b(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_b(\boldsymbol{\varepsilon}_K)])\rangle_{\mathbb{P}}. \end{aligned}$$

Therefore, we have

$$Cov[X_i - g_1(\mathbf{X}_{-i}), X_j - g_2(\mathbf{X}_{-j})] = 0$$

if and only if $m_a(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_a(\boldsymbol{\varepsilon}_K)]$ and $m_b(\boldsymbol{\varepsilon}_K) - \mathbb{E}[m_b(\boldsymbol{\varepsilon}_K)]$ are orthogonal in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. □

Proposition B.1 implies that for any $i, j \in V$ with $i \neq j$, generic non-linear functions satisfy the (i, j)-residual faithfulness condition with respect to the distribution $\mathbb{P}^{\mathbf{X}}$, except for very specific cases. One special example satisfying the orthogonality is given by

$$X_i - g_1(\mathbf{X}_{-i}) = \varepsilon_k, \quad X_j - g_2(\mathbf{X}_{-j}) = \varepsilon_k^2 - 1$$

where $\varepsilon_k \sim (0, 1)$. Then we obtain

$$Cov[X_i - g_1(\mathbf{X}_{-i}), X_j - g_2(\mathbf{X}_{-j})] = \mathbb{E}[\varepsilon_k(\varepsilon_k^2 - 1)] - \mathbb{E}[\varepsilon_k]\mathbb{E}[\varepsilon_k^2 - 1] = \mathbb{E}[\varepsilon_k^3] - \mathbb{E}[\varepsilon_k] = 0 - 0 = 0.$$

Since zero covariance is only a necessary condition for independence, the class of functions that violate the residual faithfulness condition is highly restricted.

For the proofs of Lemmas 3.1 to 3.3, corresponding to the three different cases, we first establish the following lemmas. The second lemma will serve as the key technical tool..

Lemma B.2. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that there exist functions $u, v : \mathbb{R} \to \mathbb{R}$ with*

$$g(x + y) = u(x) + v(y) \quad \textit{for all } x, y \in \mathbb{R}.$$

If g is continuous, then g is linear, i.e., there exist some $a, b \in \mathbb{R}$ such that $g(x) = ax + b$ for all $x \in \mathbb{R}$.

Proof. First, we set $y = 0$ in $g(x+y) = u(x)+v(y)$, then we yield $u(x) = g(x) - v(0)$. Similarly, setting $x = 0$ gives $v(y) = g(y) - u(0)$. Substituting into the original equation yields

$$g(x + y) = g(x) + g(y) - \big(v(0) + u(0)\big).$$

Evaluating at $x = y = 0$, we find

$$g(0) = g(0) + g(0) - \big(v(0) + u(0)\big).$$

Hence, we obtain $v(0) + u(0) = g(0)$. Therefore, it holds that

$$g(x + y) = g(x) + g(y) - g(0).$$

Define $G(x) := g(x) - g(0)$. Then we obtain

$$G(x + y) = g(x + y) - g(0) = g(x) + g(y) - 2g(0) = G(x) + G(y).$$

Thus G satisfies Cauchy's functional equation. Since g is continuous, then G is continuous. Hence, we have $G(x) = ax$ for some constant $a \in \mathbb{R}$. Therefore, we conclude that

$$g(x) = G(x) + g(0) = ax + b,$$

where $b = g(0)$. □

Lemma B.3. *Let $G = (V, D, B)$ be an acyclic mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$, and let $(h, \Sigma) \in \mathcal{H}(\mathbb{P}^{\mathbf{X}})$ such that $\mathbf{X} \overset{d}{=} h(\boldsymbol{\varepsilon})$ for $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in SPD(G)$. Let $s(X_i)$ denote an arbitrary non-constant function that depends on X_i. Then the residual of any non-linear regression of $s(X_i)$ onto $\mathbf{X}_{-i}$ depends on ε_i. That is,*

$$\forall g \in \mathcal{G}, i \in \mathrm{Dep}(s(h_i(\boldsymbol{\varepsilon})) - g(h_{-i}(\boldsymbol{\varepsilon}))).$$

Proof. Assume that

$$i \notin \mathrm{Dep}(s(h_i(\boldsymbol{\varepsilon})) - g(h_{-i}(\boldsymbol{\varepsilon}))).$$

Since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$, there exist $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$ and $\Sigma \in \mathrm{SPD}(G)$ such that $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, then we have

$$h_i(\boldsymbol{\varepsilon}) \overset{d}{=} X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i = c(\boldsymbol{\varepsilon}_{\mathrm{anc}(i) \cup \{i\}}),$$

where $c(\cdot)$ denotes some real-valued function. Since $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$, it follows that $j \in \mathrm{Dep}(c(\boldsymbol{\varepsilon}_{\mathrm{anc}(i) \cup \{i\}}))$ for all $j \in \mathrm{anc}(i) \cup \{i\}$. Note that $i \in \mathrm{Dep}(s(h_i(\boldsymbol{\varepsilon})))$. In order for the term $g(h_{-i}(\boldsymbol{\varepsilon}))$ to cancel the dependence on ε_i in $s(h_i(\boldsymbol{\varepsilon}))$, it must itself depend on ε_i. Thus, $h_{-i}(\boldsymbol{\varepsilon})$ must depend on at least one descendant of node i which denoted by k, and $k \in \mathrm{Dep}(g(\mathbf{x}_{-i}))$. Note that $k \notin \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}))$ due to the acyclicity of G.

Since $g \in \mathcal{G}$ and $k \in \mathrm{Dep}(g(\mathbf{x}_{-i}))$, g is non-linear in k-th argument. Lemma B.2 implies that for all non-constant functions $u(\cdot), v(\cdot)$, there exist $x_k, y_k \in \mathbb{R}$ such that

$$g(x_k + y_k, \mathbf{x}_{-k}) \neq u(x_k, \mathbf{x}_{-k}) + v(y_k, \mathbf{x}_{-k}).$$

Therefore, it holds that

$$\begin{aligned} g(h_{-i}(\boldsymbol{\varepsilon})) = g(\mathbf{X}_{-i}) &= g(f_k(\mathbf{X}_{\mathrm{pa}(k)\setminus i}, X_i) + \varepsilon_k, \mathbf{X}_{-i\setminus k}) \\ &= g(f_k(\mathbf{X}_{\mathrm{pa}(k)\setminus i}, c(\boldsymbol{\varepsilon}_{\mathrm{anc}(i)\cup\{i\}})) + \varepsilon_k, \mathbf{X}_{-i\setminus k}) \\ &\neq u(\varepsilon_k, \boldsymbol{\varepsilon}_{-(k,i)}) + v(\varepsilon_i, \boldsymbol{\varepsilon}_{-(k,i)}), \end{aligned}$$

for all non-constant functions $u(\cdot), v(\cdot)$. Hence, we have

$$i \in \mathrm{Dep}(v(\varepsilon_i, \boldsymbol{\varepsilon}_{-(k,i)}) - g(h_{-i}(\boldsymbol{\varepsilon}))).$$

Since $s(h_i(\boldsymbol{\varepsilon})) = v(\varepsilon_i, \boldsymbol{\varepsilon}_{-(k,i)})$ for some non-constant function $v(\cdot)$, it follows that

$$i \in \mathrm{Dep}(s(h_i(\boldsymbol{\varepsilon})) - g(h_{-i}(\boldsymbol{\varepsilon}))).$$

□

Now we prove Lemmas 3.1 to 3.3, corresponding to the three different cases, which form the key basis for structural identifiability.

Lemma 3.1 (Case 1). Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \in B$ if and only if, for all $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$, it holds that

$$(X_i - g_i(\mathbf{X}_{-i})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-j})) \tag{5}$$

Proof. Since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$, there exist $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$ and $\Sigma \in \mathrm{SPD}(G)$ such that $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, and let h denote h_f^{-1}, then we have

$$h_i(\boldsymbol{\varepsilon}) \overset{d}{=} X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i.$$

By Lemma B.3, we conclude that,

$$i \in \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}) - g_i(h_{-i}(\boldsymbol{\varepsilon}))).$$

That is, $g_i(\mathbf{X}_{-i})$ cannot cancel out term ε_i from X_i. In the same way, $g_j(\mathbf{X}_{-j})$ cannot cancel out term ε_j from X_j. Since $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$ satisfies the (i, j)-residual faithfulness condition w.r.t. $\mathbb{P}^{\mathbf{X}}$, it holds that if

$$\varepsilon_i \not\perp\!\!\!\perp \varepsilon_j, \tag{6}$$

we obtain Equation (5). For the other direction, suppose that Equation (5) holds for all $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$. Since $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$, we have $(f_i, f_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$. Taking

$$g_i(\mathbf{X}_{-i}) = f_i(\mathbf{X}_{\mathrm{pa}(i)}), \quad g_j(\mathbf{X}_{-j}) = f_j(\mathbf{X}_{\mathrm{pa}(j)}),$$

we obtain Equation (6). Equation (6) is equivalent to $i \leftrightarrow j \in B$. That is, there exists an unobserved confounder between nodes i and j. Since the graph is bow-free, there is no direct cause between nodes i and j. □

Lemma 3.2 (Case 2). Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \notin B$ if and only if there exists $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$ such that

$$(X_i - g_i(\mathbf{X}_{-(i,j)})) \perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})). \tag{7}$$

Proof. Since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$, there exist $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$ and $\Sigma \in \mathrm{SPD}(G)$ such that $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, and let h denote h_f^{-1}, then we have

$$h_i(\boldsymbol{\varepsilon}) \overset{d}{=} X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i.$$

We start with Equation (7) and show that $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \notin B$. If Equation (7) holds, then Equation (5) does not. By Lemma 3.1, there is no unobserved confounder between nodes i and j. We prove that there is no direct causal relationship between nodes i and j by contradiction. Assume that Equation (7) holds for some $g_i, g_j \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$, and that node j is a direct cause of node i. Then X_i is a function of X_j. By Lemma B.3, we conclude that

$$j \in \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}) - g_i(h_{-(i,j)}(\boldsymbol{\varepsilon}))).$$

Similarly, we have

$$j \in \mathrm{Dep}(h_j(\boldsymbol{\varepsilon}) - g_j(h_{-(i,j)}(\boldsymbol{\varepsilon}))).$$

Therefore, it holds that

$$(X_i - g_i(\mathbf{X}_{-(i,j)})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})),$$

since $(g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}})$ satisfies the (i,j)-residual faithfulness condition w.r.t. $\mathbb{P}^{\mathbf{X}}$. This contradicts Equation (7). The same arguments apply when i is a direct cause of j. Therefore, there is no direct causal relationship between nodes i and j.

For the other direction, suppose that $i \to j \notin D$, $j \to i \notin D$ and $i \leftrightarrow j \notin B$. Define

$$g_i(\mathbf{X}_{-(i,j)}) = f_i(\mathbf{X}_{\mathrm{pa}(i)}), \quad g_j(\mathbf{X}_{-(i,j)}) = f_j(\mathbf{X}_{\mathrm{pa}(j)}).$$

Then we obtain

$$X_i - g_i(\mathbf{X}_{-(i,j)}) = \varepsilon_i, \quad X_j - g_j(\mathbf{X}_{-(i,j)}) = \varepsilon_j.$$

Since $i \leftrightarrow j \notin B$, we have $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$. Therefore, we conclude that

$$(X_i - g_i(\mathbf{X}_{-(i,j)})) \perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})).$$

□

Lemma 3.3 (Case 3). Let $G = (V, D, B)$ be a bow-free mixed graph. Let $\mathbf{X} \sim \mathbb{P}^{\mathbf{X}}$ with $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then, for any $i, j \in V$ with $i \neq j$, we have $j \to i \in D$ and $i \leftrightarrow j \notin B$ if and only if the following two conditions hold:

$$\forall (g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) : (X_i - g_i(\mathbf{X}_{-(i,j)})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-j})), \tag{8}$$

$$\exists (g_i, g_j) \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) : (X_i - g_i(\mathbf{X}_{-i})) \perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})). \tag{9}$$

Proof. We start with Equation (8) and (9) and then show that $j \to i \in D$ and $i \leftrightarrow j \notin B$. If Equation (9) holds, then Equation (5) does not. By Lemma 3.1, there is no latent confounder between X_i and X_j. If Equation (8) holds, then Equation (7) does not. By Lemma 3.2, there is a direct causal relationship between nodes i and j.

Since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$, there exist $f \in \tilde{\mathcal{F}}(G, \mathbb{P}^{\mathbf{X}})$ and $\Sigma \in \mathrm{SPD}(G)$ such that $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, and let h denote h_f^{-1}, then we have

$$h_i(\boldsymbol{\varepsilon}) \overset{d}{=} X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i.$$

Assume that node j is a direct cause of node i. Define

$$g_i(\mathbf{X}_{-i}) = f_i(\mathbf{X}_{\mathrm{pa}(i)}), \quad g_j(\mathbf{X}_{-(i,j)}) = f_j(\mathbf{X}_{\mathrm{pa}(j)}).$$

Then we obtain

$$X_i - g_i(\mathbf{X}_{-i}) = \varepsilon_i, \quad X_j - g_j(\mathbf{X}_{-(i,j)}) = \varepsilon_j.$$

Since there is no latent confounder between nodes i and j, we have $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$. Therefore, Equation (9) holds.

Now assume that node i is a direct cause of node j. By Lemma B.3, we obtain

$$\forall g_i \in \mathcal{G} : \ i \in \mathrm{Dep}(h_i(\boldsymbol{\varepsilon}) - g_i(h_{-i}(\boldsymbol{\varepsilon}))).$$

Similarly, since X_j is a function of X_i, we have

$$\forall g_j \in \mathcal{G} : \ i \in \mathrm{Dep}(h_j(\boldsymbol{\varepsilon}) - g_j(h_{-(i,j)}(\boldsymbol{\varepsilon}))).$$

Consequently, it holds that

$$\forall g_i, g_j \in \mathcal{G}_{ij}(\mathbb{P}^{\mathbf{X}}) : \ (X_i - g_i(\mathbf{X}_{-i})) \not\perp\!\!\!\perp (X_j - g_j(\mathbf{X}_{-(i,j)})).$$

This contradicts Equation (9). Thus, if Equation (9) holds, then node j is a direct cause of node i.

For the other direction, suppose that $j \to i \in D$ and $i \leftrightarrow j \notin B$. Then, as shown above, Equation (9) holds. If Equation (8) does not hold, then Equation (5) does not hold. By Lemma 3.1, there is a latent confounder between node i and node j, which contradicts the assumption $i \leftrightarrow j \notin B$. □

Theorem 3.1 (Structural identifiability of bow-free mixed graphs). Let G be a bow-free mixed graph and let $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G)$. Then G is structurally identifiable from $\mathbb{P}^{\mathbf{X}}$.

Proof. For a bow-free mixed graph G_1 and a joint distribution $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_1)$, suppose that there exists a different bow-free mixed graph $G_2 \neq G_1$ such that $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_2)$. Since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_1)$, there exist $f^{G_1} \in \tilde{\mathcal{F}}(G_1, \mathbb{P}^{\mathbf{X}})$ and $\Sigma_1 \in \mathrm{SPD}(G_1)$ such that $\mathbb{P}^{\mathbf{X}} = (h^{-1}_{f^{G_1}})_* \mathcal{N}(0, \Sigma_1)$. Let $\boldsymbol{\varepsilon}^{G_1} \sim \mathcal{N}(0, \Sigma_1)$, then we have

$$X_i = f_i^{G_1}(\mathbf{X}_{\mathrm{pa}_x^{G_1}(i)}) + \varepsilon_i^{G_1}.$$

Similarly, since $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_2)$, there exist $f^{G_2} \in \tilde{\mathcal{F}}(G_2, \mathbb{P}^{\mathbf{X}})$ and $\Sigma_2 \in \mathrm{SPD}(G_2)$ such that, for $\boldsymbol{\varepsilon}^{G_2} \sim \mathcal{N}(0, \Sigma_2)$, we have

$$X_i = f_i^{G_2}(\mathbf{X}_{\mathrm{pa}_x^{G_2}(i)}) + \varepsilon_i^{G_2}.$$

Since $G_1 \neq G_2$, the two graphs differ by at least one edge. By Lemma 3.1, 3.2, 3.3, each case leads to mutually exclusive conditional independence or dependence constraints. Hence, we obtain

$$\tilde{\mathcal{M}}(G_1) \cap \tilde{\mathcal{M}}(G_2) = \emptyset,$$

which contradicts the assumption that $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_1)$ and $\mathbb{P}^{\mathbf{X}} \in \tilde{\mathcal{M}}(G_2)$. □

Theorem 3.2 (Parameter identifiability of ancestral mixed graphs). Let $G = (V, D, B)$ be a mixed graph and let $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$. If G is ancestral, then the structural functions f and the covariance matrix Σ are identifiable. More precisely, let $p^{\mathbf{X}}_{(f,\Sigma)}$ denote the probability density corresponding to the distribution $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. If $p^{\mathbf{X}}_{(f,\Sigma)} = p^{\mathbf{X}}_{(\tilde{f},\tilde{\Sigma})}$, then $f = \tilde{f}$ almost surely and $\Sigma = \tilde{\Sigma}$.

Proof. Since $\mathbb{P}^{\mathbf{X}} \in \mathcal{M}(G)$, there exist $f \in \mathcal{F}(G)$ and $\Sigma \in \mathrm{SPD}(G)$ such that $\mathbb{P}^{\mathbf{X}} = (h_f^{-1})_* \mathcal{N}(0, \Sigma)$. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, then we have

$$X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i.$$

Let g_i be the composition of the above functions applied iteratively until the inputs have no ancestors. Then we obtain

$$X_i = g_i(\boldsymbol{\varepsilon}_{\mathrm{an}(i)}) + \varepsilon_i.$$

Since G is ancestral, there is no bidirected edge between any node and any of its ancestors. Hence, ε_i is independent of $(\varepsilon_j)_{j \in \mathrm{an}(i)}$ for all $i \in V$. The conditional expectation of X_i given its parents can be computed from the probability density $p^{\mathbf{X}}_{(f,\Sigma)}$ as

$$\mathbb{E}[X_i \mid \mathbf{X}_{\mathrm{pa}(i)}] = \int x_i \cdot p^{\mathbf{X}}_{(f,\Sigma)}(x_i \mid \mathbf{x}_{\mathrm{pa}(i)}) dx_i \tag{10}$$

almost surely, where

$$p^{\mathbf{X}}_{(f,\Sigma)}(x_i \mid \mathbf{x}_{\mathrm{pa}(i)}) = \frac{p^{\mathbf{X}}_{(f,\Sigma)}(x_i, \mathbf{x}_{\mathrm{pa}(i)})}{\int p^{\mathbf{X}}_{(f,\Sigma)}(x_i, \mathbf{x}_{\mathrm{pa}(i)}) dx_i}$$

and $p^{\mathbf{X}}_{(f,\Sigma)}(x_i, \mathbf{x}_{\mathrm{pa}(i)}) = \int p^{\mathbf{X}}_{(f,\Sigma)}(x_1, \ldots, x_d) d\mathbf{x}_{-(i,\mathrm{pa}(i))}$ for all $i \in V$.

Moreover, it holds that

$$\begin{aligned}\mathbb{E}[X_i \mid \mathbf{X}_{\mathrm{pa}(i)}] &= \mathbb{E}[f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i \mid \mathbf{X}_{\mathrm{pa}(i)}] \\ &= f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \mathbb{E}[\varepsilon_i \mid g_{\mathrm{pa}(i)}(\boldsymbol{\varepsilon}_{\mathrm{an}(\mathrm{pa}(i))}) + \boldsymbol{\varepsilon}_{\mathrm{pa}(i)}].\end{aligned}$$

Since $\{\mathrm{an}(\mathrm{pa}(i)), \mathrm{pa}(i)\} \subset \mathrm{an}(i)$, ε_i is independent of $(\varepsilon_j)_{j \in \mathrm{an}(\mathrm{pa}(i)) \cup \mathrm{pa}(i)}$. Therefore, we have

$$\mathbb{E}[\varepsilon_i \mid g_{\mathrm{pa}(i)}(\boldsymbol{\varepsilon}_{\mathrm{an}(\mathrm{pa}(i))}) + \boldsymbol{\varepsilon}_{\mathrm{pa}(i)}] = 0,$$

and hence it follows that

$$f_i(\mathbf{X}_{\mathrm{pa}(i)}) = \mathbb{E}[X_i \mid \mathbf{X}_{\mathrm{pa}(i)}],$$

which can be computed by (10). Note that

$$h_f : \mathbb{R}^d \to \mathbb{R}^d, \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \mapsto \begin{pmatrix} X_1 - f_1(\mathbf{X}_{\text{pa}(1)}) \\ \vdots \\ X_d - f_d(\mathbf{X}_{\text{pa}(d)}) \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_d \end{pmatrix}$$

is invertible. Then, by the density transformation formula, the probability density of $\boldsymbol{\varepsilon}$ is

$$p^{\boldsymbol{\varepsilon}}_{(f,\Sigma)}(e_1, \ldots, e_d) = p^{\mathbf{X}}_{(f,\Sigma)}(h_f^{-1}(e_1, \ldots, e_d)) \cdot |\det(J_{h_f^{-1}})|, \tag{11}$$

where $J_{h_f^{-1}}$ denotes the Jacobian matrix of h_f^{-1}. Therefore, we obtain

$$\mathbb{E}[\varepsilon_i \varepsilon_j] = \int e_i e_j p^{\boldsymbol{\varepsilon}}_{(f,\Sigma)}(e_i, e_j) de_i de_j,$$

where the marginal probability density $p^{\boldsymbol{\varepsilon}}_{(f,\Sigma)}(e_i, e_j)$ is obtained from the joint probability density in Equation (11). Consequently, we conclude

$$\sigma_{ij} = Cov[\varepsilon_i, \varepsilon_j] = \mathbb{E}[\varepsilon_i \varepsilon_j],$$

which is uniquely determined. □

Lemma 4.1 For m small enough, more precisely,

$$0 < m \leq \inf\{\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} \mid j, k \in V,\ \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} > 0\},$$

we have $\mathcal{R}^m_{nl}(f) > 0$ if and only if there exist $j, k \in V$ such that $k \in \text{Dep}(f_j(\mathbf{x}))$ and the map $x_k \mapsto f_j(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_{d-1})$ is linear.

Proof. If there exist $j, k \in V$ such that $k \in \text{Dep}(f_j(\mathbf{x}))$, then $\|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})} > 0$. If the map $x_k \mapsto f_j(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_{d-1})$ is linear, then $\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} = 0$. Hence, we have

$$\max\{0, m - \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}\} = m > 0.$$

Therefore, it holds that

$$\mathcal{R}^m_{nl}(f) = \sum_{j,k=1}^{d} \|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})} \max\{0, m - \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}\} > 0.$$

Conversely, if for all $j, k \in V$, either $k \notin \text{Dep}(f_j(\mathbf{x}))$ or the map

$$x_k \mapsto f_j(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_{d-1})$$

is non-linear, then either $\|\partial_k f_j\|_{L_1(\mathbb{P}^{\mathbf{X}})} = 0$ or $\|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})} > 0$. In the latter case, since $m \leq \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}$, we have

$$\max\{0, m - \|H_{kk}(f_j)\|_{L_1(\mathbb{P}^{\mathbf{X}})}\} = 0$$

and therefore $\mathcal{R}^m_{nl}(f) = 0$. □

C ALGORITHM

We present the precise procedure of the GLL-ADMG algorithm in Algorithm 1. Equation (12) in Algorithm 1 indicates that, starting from $\theta^{(t)}$ and $\Sigma^{(t)}$, the next iterates $\theta^{(t+1)}$ and $\Sigma^{(t+1)}$ are obtained using the ADAM optimizer (Kingma & Ba, 2015). During optimization, we apply the Log-Cholesky parameterization to the covariance matrix Σ to ensure that it remains symmetric positive definite (Pinheiro & Bates, 1996). For more details, we refer to Appendix D.

Algorithm 1: GLL-ADMG

Input: Data $\mathbb{X}$, initial step $\mu^{(0)}$, decay $\alpha \in (0,1)$, sparsity parameters λ_1, λ_2, non-linearity parameter λ_{nl}, hyperparameter $s, m > 0$, number of iterations T, and thresholds ω_1, ω_2.

Output: $\widehat{W}_1$, the estimated weighted adjacency matrix and $\widehat{\Sigma}$, the estimated covariance matrix.

Initialize $\theta^{(0)}$ such that $W_1^{\mathbb{X}}(\theta^{(0)}) \in \mathbb{W}^s$.

for $t \leftarrow 0$ **to** $T-1$ **do**

$$\theta^{(t+1)}, \Sigma^{(t+1)} = \arg\min_{\theta,\Sigma} \ \mu^{(t)} \Big\{ \mathcal{L}(\theta^{(t)}, \Sigma^{(t)} \mid \mathbb{X}) + \lambda_1 \Omega_1^{\mathbb{X}}(f^{\theta^{(t)}}) + \lambda_2 \Omega_2(\Sigma^{(t)}) + \lambda_{nl} \mathcal{R}_{nl}^m (f^{\theta^{(t)}})^{\mathbb{X}} \Big\} + h^s(\theta^{(t)}, \Sigma^{(t)}). \tag{12}$$

Set $\mu^{(t+1)} = \alpha \mu^{(t)}$.

Threshold matrices $\widehat{W}_1 = W_1^{\mathbb{X}}(\theta^{(T)}) \circ \mathbf{1}\big(W_1^{\mathbb{X}}(\theta^{(T)}) > \omega_1\big), \quad \widehat{\Sigma} = \Sigma^{(T)} \circ \mathbf{1}\big(\Sigma^{(T)} > \omega_2\big)$.

D PARAMETERIZATION OF THE COVARIANCE MATRIX

During the optimization process, the covariance matrix Σ needs to be symmetric and positive definite. By the Cholesky decomposition, there exists a unique lower triangular matrix $L \in \mathbb{R}^{d\times d}$ with positive diagonal entries such that $\Sigma = LL^T$. However, since L is constrained to have positive diagonal entries, it cannot be treated as a fully unconstrained parameter during optimization. Therefore, we employ the following Log-Cholesky parameterization for Σ (Pinheiro & Bates, 1996), which maps an arbitrary lower triangular matrix $M \in \mathbb{R}^{d\times d}$ to a symmetric positive definite matrix Σ.

Lemma D.1. *Let $LT_d \subset \mathbb{R}^{d\times d}$ denote the set of lower triangular matrices. There exists a bijective map*

$$\Sigma(\cdot) : LT_d \to SPD^d, \quad M \mapsto \Sigma(M) = [e^{\mathbb{D}(M)} + \lfloor M \rfloor][e^{\mathbb{D}(M)} + \lfloor M \rfloor]^T,$$

where $\mathbb{D}(M)$ denotes the diagonal matrix containing the diagonal elements of M, and $\lfloor M \rfloor$ denotes the strictly lower triangular part of the matrix M.

Moreover, the inverse map is

$$M(\cdot) : SPD^d \to LT_d, \quad \Sigma \mapsto M(\Sigma) = \lfloor L(\Sigma) \rfloor + \log \mathbb{D}(L(\Sigma)),$$

where $L(\Sigma)$ denotes the Cholesky factor of Σ, and $\log \mathbb{D}(L(\Sigma))$ denotes the matrix logarithm of $\mathbb{D}(L(\Sigma))$.

Proof. Let

$$L(\Sigma(M)) := e^{\mathbb{D}(M)} + \lfloor M \rfloor.$$

Then $L(\Sigma(M))$ is lower triangular with positive diagonal entries. Hence, $L(\Sigma(M))L(\Sigma(M))^T$ is symmetric positive definite.

Let $M_1, M_2 \in \mathrm{LT}_d$ satisfy $\Sigma(M_1) = \Sigma(M_2)$. By uniqueness of the Cholesky decomposition, the Cholesky factor $L(\Sigma(M_1)) = L(\Sigma(M_2))$. Therefore, we have

$$\mathbb{D}(M_1) = \mathbb{D}(M_2), \quad \lfloor M_1 \rfloor = \lfloor M_2 \rfloor.$$

Since M_1, M_2 are lower triangular, we conclude that $M_1 = M_2$. Thus, $\Sigma(\cdot)$ is injective.

For any $\Sigma \in \mathrm{SPD}^d$, there exists a Cholesky factor $L(\Sigma)$ such that $\Sigma = L(\Sigma)L(\Sigma)^T$. Define

$$M = \mathbb{D}(M) + \lfloor M \rfloor = \log \mathbb{D}(L(\Sigma)) + \lfloor L(\Sigma) \rfloor,$$

and hence it holds that

$$\Sigma(M) = [e^{\mathbb{D}(M)} + \lfloor M \rfloor][e^{\mathbb{D}(M)} + \lfloor M \rfloor]^T = L(\Sigma)L(\Sigma)^T = \Sigma.$$

Thus, $\Sigma(\cdot)$ is surjective, and the inverse map is given as in the statement. □

E EXPERIMENTAL DETAILS AND HYPERPARAMETER SELECTION

Simulations. First, we generate both ER-type ancestral and bow-free mixed graphs randomly. For a graph with d nodes, we first sample the directed component by fixing a topological ordering and including each admissible directed edge independently with probability $p_D = 1/(d-1)$. This yields an expected number of directed edges equal to $d/2$. We then sample bidirected edges independently with the same probability $p_B = 1/(d-1)$. For ancestral graphs, bidirected edges are added subject to the ancestrality constraint that no bidirected edge is introduced between an ancestor-descendant pair. Similarly, for bow-free graphs, bidirected edges are added subject to the constraint that no bidirected edge is introduced between a parent–child pair. Consequently, the expected number of bidirected edges is at most $d/2$ and depends on the sampled directed component. This choice keeps the expected number of edges bounded by d. We retain only those graphs that contain both directed and bidirected edges. For each graph G, we simulate data according to the SEM $X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i$ for all $i \in V$, evaluated in topological order, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathrm{SPD}(G)$. To evaluate performance under different data generation mechanisms, we consider two models for the functions f_i for all $i \in V$:

(1) **MLP model.** Following the procedure of Zheng et al. (2020), each variable is generated by a one-hidden-layer multilayer perceptron applied to its parents, with 100 hidden units. The network parameters are sampled uniformly from $(-2, -0.5) \cup (0.5, 2)$.

(2) **Additive non-linear model.** Similarly to the procedure of Liang et al. (2024), each variable is generated by an additive model, i.e., a linear combination of non-linear transformations of its parents. The non-linear transformations are randomly selected from $g(x) = \exp(x), g(x) = \tanh(x), g(x) = \sin(x)$. The corresponding weights are sampled uniformly from $(-3, -2) \cup (2, 3)$.

For the experiments in Section 5.1, the covariance matrix $\Sigma \in \mathrm{SPD}(G)$ of $\boldsymbol{\varepsilon}$ has unit diagonal entries, and off-diagonal entries are sampled uniformly from $(-0.4, -0.8) \cup (0.4, 0.8)$. Besides Gaussian exogenous noise, we additionally consider heavy-tailed and skewed exogenous noise with heterogeneous variances. To retain the same off-diagonal covariance as in Section 5.1, we first generate an auxiliary matrix $\widetilde{\Sigma} \in \mathrm{SPD}(G)$ following the same procedure as before. Note that $\widetilde{\Sigma}$ is used only to specify the off-diagonal covariances and is not the covariance matrix of the exogenous noise. Each observed variable is generated according to the SEM $X_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \varepsilon_i$, where the functions $(f_i)_{i \in V}$ are generated by either MLP or additive non-linear models as before. We construct the exogenous noise as

$$\varepsilon_i = b_i \widetilde{\varepsilon}_i + \sum_{j>i} \sqrt{|\widetilde{\Sigma}_{ij}|}\, U_{ij} + \sum_{j<i} \mathrm{sgn}(\widetilde{\Sigma}_{ji}) \sqrt{|\widetilde{\Sigma}_{ji}|}\, U_{ji},$$

where $(U_{ij})_{i,j \in V}$ are mutually independent standard Gaussian confounders, and the variables $\widetilde{\varepsilon}_i$ are independent across $i \in V$ and independent of all Gaussian confounders. By construction, the actual covariance matrix $\Sigma := \mathrm{Cov}(\boldsymbol{\varepsilon})$ satisfies $\Sigma_{ij} = \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \widetilde{\Sigma}_{ij}$ for all $i, j \in V$ with $i \neq j$, while its diagonal entries are determined by

$$\Sigma_{ii} = \mathrm{Var}(\varepsilon_i) = b_i^2 \,\mathrm{Var}(\widetilde{\varepsilon}_i) + \sum_{j \neq i} |\widetilde{\Sigma}_{ij}|.$$

Thus, $\widetilde{\Sigma}$ controls the off-diagonal covariances of the exogenous noise, whereas its marginal variances are heterogeneous.

We consider two distributions for $\widetilde{\varepsilon}$:

(1) Heavy-tailed noise: Let

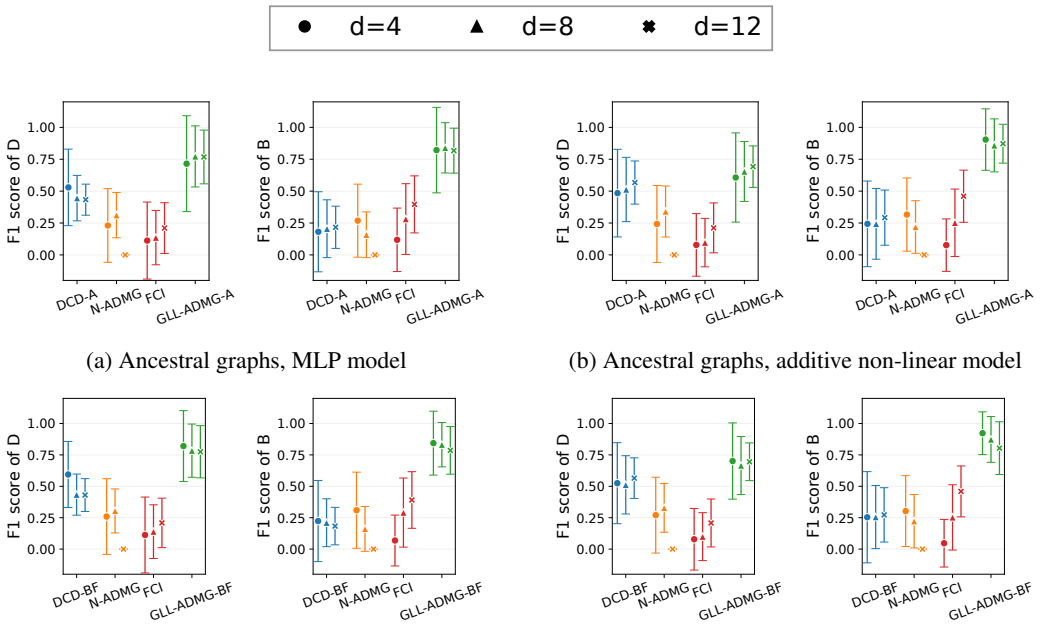
$$\widetilde{\varepsilon}_i \overset{\mathrm{iid}}{\sim} t_5, \qquad b_i \overset{\mathrm{iid}}{\sim} \mathrm{Uniform}(0.5, 1.0).$$

Since $\mathrm{Var}(\widetilde{\varepsilon}_i) = \frac{5}{3}$, the resulting marginal variances are

$$\Sigma_{ii} = \frac{5}{3} b_i^2 + \sum_{j \neq i} |\widetilde{\Sigma}_{ij}|.$$

(2) Skewed noise: Let

$$\widetilde{\varepsilon}_i = Y_i - 2, \qquad Y_i \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(2, 1), \qquad b_i \overset{\mathrm{iid}}{\sim} \mathrm{Uniform}(0.8, 1.0).$$



(a) Ancestral graphs, MLP model

(b) Ancestral graphs, additive non-linear model

(c) Bow-free graphs, MLP model

(d) Bow-free graphs, additive non-linear model

Figure 3: Causal discovery results on synthetic datasets generated using MLP and additive non-linear models with heavy-tailed exogenous noise, under ancestral and bow-free graph restrictions. Each subfigure reports F1 scores for directed (D) and bidirected (B) edges. Error bars represent standard errors over 100 simulations.

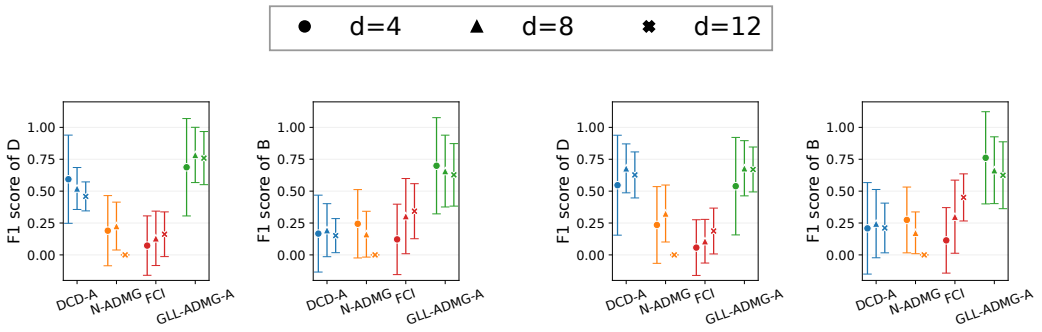
The centering ensures that $\mathbb{E}[\widetilde{\varepsilon}_i] = 0$ while preserving the positive skewness of the Gamma distribution. Since $\mathrm{Var}(\widetilde{\varepsilon}_i) = 2$, the resulting marginal variances are

$$\Sigma_{ii} = 2b_i^2 + \sum_{j \neq i} |\widetilde{\Sigma}_{ij}|.$$

In both settings, the exogenous noise preserves the prescribed off-diagonal covariance while exhibiting heterogeneous marginal variances.

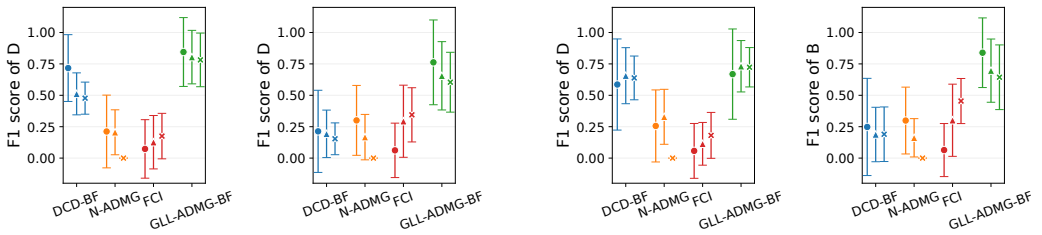
Figure 3 compares GLL-ADMG with the baselines on data generated with heavy-tailed exogenous noise under ancestral and bow-free graph restrictions, whereas Figure 4 presents the corresponding results for skewed exogenous noise. Figure 5 reports the results for Gaussian exogenous noise under non-bow-free graph structures. The non-bow-free graphs are generated by independently sampling directed and bidirected edges, each with probability $1/(d-1)$. If the resulting graph contains no bow, we randomly select a pair of nodes connected by a directed edge and add a bidirected edge between them, ensuring the presence of at least one bow. Across these model-misspecification settings, GLL-ADMG consistently outperforms the baselines, demonstrating its robustness to deviations from Gaussianity and bow-freeness.

Hyperparameters. We run GLL-ADMG, where each f_i is approximated by an MLP with one hidden layer of 10 units. Inspired by Waxman et al. (2024), we additionally pretrain GLL-ADMG by performing a single iteration of the DAG learning algorithm DAGMA (Bello et al., 2022). Although DAGMA and GLL-ADMG are designed for different tasks, they share the same functional approximation structure for each f_i and both aim to recover the directed edges. Therefore, a single iteration of DAGMA provides a good initialization and avoids costly backward passes through the Jacobian and Hessian. We set the hyperparameter $s = 1$ as suggested by Bello et al. (2022), and set the sparsity regularization parameters to $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$. The non-linearity regularization parameter is set to $\lambda_{nl} = 8$, with the corresponding hyperparameter $m = 0.1$. The remaining hyperparameters of GLL-ADMG and its pretraining are set to the default values in Waxman et al. (2024). Besides, in the optimization process of GLL-ADMG, we set the initial step as $\mu^{(0)} = 1$ and set the number of iterations to $T = 5$. The learning rate for the MLP parameters θ is set to $\mathrm{lr} = 0.001$, and the learning rate for the covariance matrix Σ is set to $0.3\,\mathrm{lr}$. For the baselines, we



(a) Ancestral graphs, MLP model

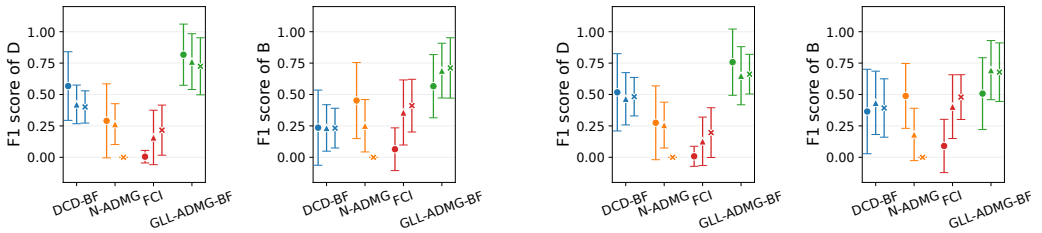
(b) Ancestral graphs, additive non-linear model



(c) Bow-free graphs, MLP model

(d) Bow-free graphs, additive non-linear model

Figure 4: Causal discovery results on synthetic datasets generated using MLP and additive non-linear models with skewed exogenous noise, under ancestral and bow-free graph restrictions. Each subfigure reports F1 scores for directed (D) and bidirected (B) edges. Error bars represent standard errors over 100 simulations.



(a) Non-bow-free graphs, MLP model

(b) Non-bow-free graphs, additive non-linear model

Figure 5: Causal discovery results on synthetic datasets generated using MLP and additive non-linear models with Gaussian exogenous noise, under non-bow-free graphs. Each subfigure reports F1 scores for directed (D) and bidirected (B) edges. Error bars represent standard errors over 100 simulations.

use their default settings. Specifically, the DCD algorithm is run 5 times and the solution with the best BIC score is selected, as prescribed in its original setup. Other methods do not include such a procedure by default, and we therefore do not apply it to them. For computational efficiency, we set the relative function-value tolerance of the L-BFGS-B optimizer in the DCD algorithm to 10^{-6} for each optimization call. This tolerance is applied within each dual-ascent update of the augmented Lagrangian, where the optimizer is invoked multiple times. For DCD and GLL-ADMG, the thresholds are set to $\omega_1 = 0.3$ and $\omega_2 = 0.2$ for directed and bidirected adjacency matrices, respectively. N-ADMG generates adjacency matrices via sampling from the learned model. Following the original setup of the algorithm, 100 samples are drawn and the mean estimate is reported. The N-ADMG implementation is no longer publicly maintained. We therefore use an earlier available version, with minor modifications for compatibility, and follow the available implementation as closely as possible. Due to the lack of an official reference implementation and dataset, the results for N-ADMG should be interpreted with caution.

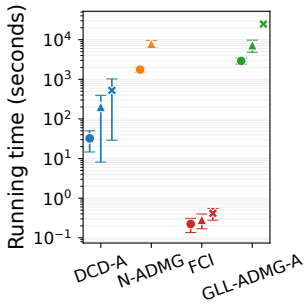


Figure 6: Running times on CPUs for datasets generated from the additive non-linear model with Gaussian exogenous noise, under the ancestral graph restriction. Error bars represent standard errors over 100 simulations.

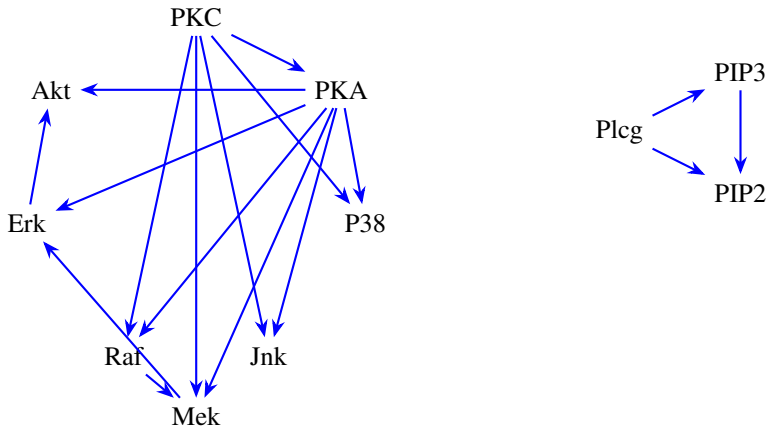


Figure 7: Ground-truth DAG of the real dataset.

Failure rate of N-ADMG The failure rate of N-ADMG is reported in Figure 9 for data generated across ancestral and bow-free graph settings for both MLP and additive non-linear models.

Computational cost We compare the computational costs of the algorithms using their running times on datasets generated from the additive non-linear model under the ancestral-graph restriction (see Figure 6). FCI and DCD are substantially faster than N-ADMG and GLL-ADMG. Note that N-ADMG frequently fails to complete within the allotted runtime or encounters numerical instability, particularly for $d = 12$, and partially for $d = 8$. Its running time is therefore computed using only successful runs and is not reported for $d = 12$. The higher computational cost of GLL-ADMG is primarily due to the Hessian evaluations required by the nonlinearity regularizer.

Ground-truth graph of the real dataset The ground-truth DAG of the protein expression dataset in Sachs et al. (2005) is shown in Figure 7. Figure 8 shows the corresponding mixed graph obtained by treating PKC as unobserved and introducing bidirected edges between every pair of its observed children.

Computational resources Experiments were conducted on a server equipped with dual Intel Xeon Gold 6242R CPUs (80 logical cores in total) and 128 GB RAM.

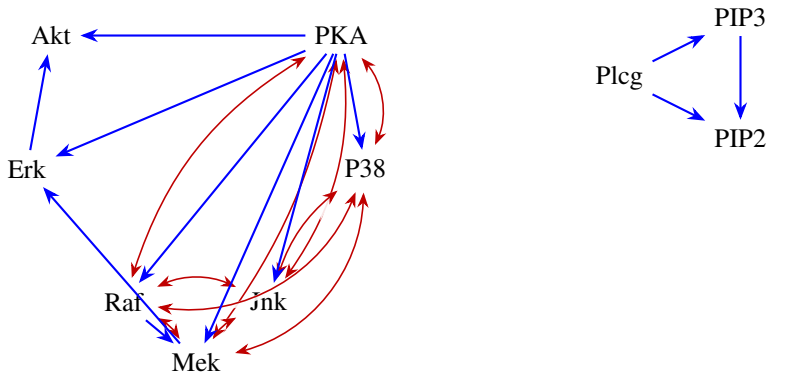
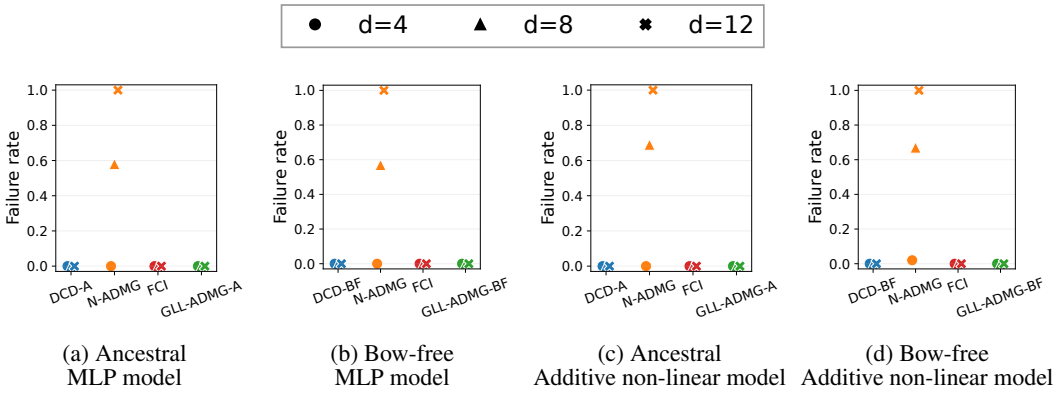


Figure 8: Ground-truth mixed graph obtained by removing PKC. Blue arrows denote directed edges, and red bidirected edges denote latent confounding.



(a) Ancestral MLP model (b) Bow-free MLP model (c) Ancestral Additive non-linear model (d) Bow-free Additive non-linear model

Figure 9: Failure rates across graph classes and data generation mechanisms with Gaussian exogenous noise.