BEYOND ADDITIVITY: CAUSAL DISCOVERY IN LOCATION-SCALE NOISE MODELS WITH HIDDEN VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study causal discovery with hidden variables when causes modulate both the mean and the spread of their effects, a location-scale noise model; existing functional-model methods for hidden variables assume additive noise. We introduce LSNM-UV, a two-level location-scale model in which the hidden parents of a variable act inside the scale set by its observed parents, so that observed-parent effects can be divided out: every hidden effect collapses into an effective noise whose dependence structure encodes unobserved backdoor and causal paths. The resulting residual-independence patterns identify directed edges, non-edges and hidden-variable pairs, and identify bow-free acyclic directed mixed graphs. Under an explicit search-coverage condition, an oracle answering residual-independence queries at the population level makes the LSNM-UV search sound and complete. A thresholded HSIC statistic on fitted residuals answers every query on the oracle path correctly with probability tending to one under four explicit conditions on admissible population targets, targeting of independence witnesses, consistency of the empirical dependence statistic and a vanishing threshold, which gives graph consistency for a fixed number of variables; a fixed test level need not have this property. Controlled simulations diagnose the statistical conditions and their failure modes, an oracle-driven search shows that correct answers yield the correct graph, and a ten-node benchmark and the NetSim fMRI benchmark with hidden regions show the method at finite samples.

1 INTRODUCTION

Constraint-based causal discovery recovers structure from conditional independence up to Markov equivalence (Spirtes et al., 2000; Pearl, 2009); with latent confounding, FCI returns a partial ancestral graph and cannot distinguish members of the same class (Richardson and Spirtes, 2002; Zhang, 2008). Functional model assumptions break this symmetry, from LiNGAM (Shimizu et al., 2006) to additive noise models (Hoyer et al., 2009; Peters et al., 2014), causal additive models (Bühlmann et al., 2014), post-nonlinear models (Zhang and Hyvärinen, 2009), and more complex models (Kamimura and Pham, 2026), all under causal sufficiency. With hidden variables, the functional-model approaches have stayed with linear non-Gaussian or additive mechanisms: Hoyer et al. (2008) generalised LiNGAM through overcomplete ICA; CAM-UV (Maeda and Shimizu, 2021) uses residual independence to identify the parent–child relations that the additive model identifies while flagging pairs affected by unobserved backdoor paths (UBPs) or unobserved causal paths (UCPs), which we represent as bidirected edges of an acyclic directed mixed graph (ADMG) (Richardson, 2003); Pham et al. (2026) showed that some directions survive such paths, and Ashman et al. (2023) learn bow-free ADMGs under additive noise. In all of these the noise scale cannot depend on the parents. In many processes causes modulate the variance as well as the mean. Location-scale noise models (LSNMs), $x_i = f_i(\mathrm{PA}_i) + g_i(\mathrm{PA}_i)\epsilon_i$, allow this, and are identifiable for cause–effect pairs (Immer et al., 2023; Sun and Schulte, 2023) and for DAGs under heteroscedastic noise (Yin et al., 2024; Lin et al., 2025), but only under causal sufficiency (Lin et al. (2025) treat a single latent confounder only as a bivariate robustness case). A theory of observed causal structure under location-scale noise with hidden variables has been open.

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.

This paper gives one, and separates three statements that are usually run together: what the population identifies, what a search does with correct answers, and what a test on fitted residuals delivers. (i) *Model and reduction.* LSNM-UV places the hidden parents of x_i inside the scale set by its observed parents, so that the observed-parent location and scale can be divided out. Every hidden effect collapses into an effective noise η_i, and $\eta_i \perp\!\!\!\perp \eta_j$ iff no UBP or UCP connects x_i and x_j (Lemma 2). (ii) *Identifiability.* Residual-independence patterns with a common exclusion set for location and scale distinguish visible edges, visible non-edges and invisible pairs (Proposition 1) and identify bow-free ADMGs (Theorem 3). (iii) *Oracle search and consistency.* Under an explicit search-coverage condition, a residual-independence oracle makes the search sound and complete (Theorem 4). A thresholded HSIC statistic on fitted residuals answers every query on the oracle path correctly with probability tending to one under four explicit conditions on admissible population targets, targeting of independence witnesses, consistency of the empirical dependence statistic and a vanishing threshold (Theorem 5), giving graph consistency for fixed p (Corollary 6). A fixed test level is not covered by this asymptotic result. (iv) *Method and validation.* A location-scale residual estimator with inverse-scale weighting and a cross-fitted audit implements the search; controlled experiments diagnose the statistical conditions and their failure modes, an oracle-driven search shows that correct answers give the correct graph, and a ten-node benchmark and the NetSim fMRI benchmark with hidden regions compare with CAM-UV, FCI and BANG. Relative to CAM-UV, the scale channel, the separation of the oracle theorem from the finite-sample test, and the consistency theorem with its targeting condition are new.

2 PRELIMINARIES

Identifiability through oracles. Let $\mathcal{G}$ be a class of graphs, $\theta \in \Theta_G$ the parameters, functions and noise distributions of a model on G, and $P_{G,\theta}$ the distribution of the observed variables X.

Definition 2.1 (Identifiability; adapted from Peters et al., 2011)**.** A graph function T is identifiable if and only if $\forall G, G' \in \mathcal{G},\ \forall \theta' \in \Theta_{G'}, \forall \theta \in \Theta_G$:

$$P_{G,\theta} = P_{G',\theta'} \implies T(G) = T(G').$$

Definition 2.2 (Oracle identifiability)**.** T is oracle identifiable with respect to an oracle W if and only if $\forall G, G' \in \mathcal{G},\ \forall \theta' \in \Theta_{G'},\ \forall \theta \in \Theta_G$:

$$W(P_{G,\theta}, q) = W(P_{G',\theta'}, q)\ \forall q \implies T(G) = T(G').$$

Lemma 1. *Oracle identifiability implies identifiability.*

An *oracle* is a deterministic function $W(P_X, q)$ of the observed distribution and a query q; if its answers determine $T(G)$, then equal observed distributions give equal answers and equal $T(G)$, so T is identifiable (Lemma 1, proof in Appendix A.1). The paper builds such an oracle from residual-independence statements, at the level of pairs (Section 3.2) and of the individual queries the algorithm issues (Section 4).

Hidden paths and the target graph. Hidden variables act on observed pairs in two ways (Maeda and Shimizu, 2021; Pham et al., 2026).

Definition 2.3 (UCP and UBP)**.** There is an *unobserved causal path* (UCP) from x_j to x_i if x_j is an ancestor of an unobserved parent $u_k \notin X$ of x_i, so that $x_j \to \cdots \to u_k \to x_i$, with the final edge into x_i coming from an unobserved parent; a UCP exists *between* x_i and x_j if one exists in either direction. There is an *unobserved backdoor path* (UBP) between x_i and x_j if unobserved parents $u_k \notin X$ of x_i and $u_l \notin X$ of x_j share an ancestor v ($v = u_k$, $v = u_l$ or $u_k = v = u_l$ allowed), forming $x_i \leftarrow u_k \leftarrow \cdots \leftarrow v \to \cdots \to u_l \to x_j$.

An acyclic directed mixed graph (ADMG) on X has directed edges $x_j \to x_i$ and bidirected edges $x_i \leftrightarrow x_j$ and no directed cycle; we write it as (A, B) with $A(i,j) = 1$ iff $x_j \to x_i$ and $B(i,j) = 1$ iff there is a UBP or UCP between x_i and x_j (B symmetric, zero diagonals).

Definition 2.4 (Bow-free ADMG)**.** (A, B) is bow-free if no pair carries both a directed and a bidirected edge: $A(i,j)\,B(i,j) = 0$ for all $i \neq j$.

Figure 1 is the running example. (x_1, x_2) is bidirected through the hidden common cause u_1, the UBP $x_1 \leftarrow u_1 \to x_2$; (x_3, x_5) is bidirected through the hidden intermediate u_2, the UCP $x_5 \to u_2 \to x_3$,

and since x_5 is also a direct parent of x_3 this pair is a bow, excluded by Definition 2.4. (x_1, x_4) is not bidirected: every path between them passes through the observed x_3, so no UBP or UCP connects them; likewise (x_1, x_5), which meet only in their common child x_3.

3 The Causal Model

Let $X = \{x_i\}$ and $U = \{u_k\}$ denote the observed and unobserved variables, respectively, and let $G = (V, E)$ be a DAG on $V = X \cup U$, where $(v_j, v_i) \in E$ denotes $v_j \to v_i$. For each $v_i \in V$, let

$$K_i = \{x_j \in X : (x_j, v_i) \in E\}, \qquad Q_i = \{u_k \in U : (u_k, v_i) \in E\}$$

be its observed and unobserved parents. We consider the two-level location-scale noise model

$$v_i = f_i^1(K_i) + g_i^1(K_i) f_i^2(Q_i) + g_i^1(K_i) g_i^2(Q_i)\epsilon_i, \tag{1}$$

where $g_i^1, g_i^2 > 0$ and the exogenous noises $\{\epsilon_i\}$ are mutually independent and non-degenerate. Equation (1) is a subclass of the general location-scale model

$$v_i = F_i(K_i, Q_i) + G_i(K_i, Q_i)\epsilon_i, \tag{2}$$

obtained by restricting

$$F_i = f_i^1(K_i) + g_i^1(K_i) f_i^2(Q_i), \qquad G_i = g_i^1(K_i) g_i^2(Q_i).$$

Thus the observed parents define a common location–scale frame for the hidden-parent effects. In particular,

$$\frac{v_i - f_i^1(K_i)}{g_i^1(K_i)} = f_i^2(Q_i) + g_i^2(Q_i)\epsilon_i = \eta_i,$$

so the observed-parent contribution can be removed by a location–scale transformation.

When $Q_i = \varnothing$, the model reduces, after absorbing constants, to the standard causally sufficient location-scale model. When $g_i^1 \equiv g_i^2 \equiv 1$, it reduces to the additive-noise form of Ashman et al. (2023), and further to the causal additive model of Maeda and Shimizu (2021) when f_i^1 and f_i^2 are sums of univariate functions.

Assumption 1 (Causal Faithfulness Condition (CFC))**.** Any conditional independence on $V = X \cup U$ not entailed by d-separation on G does not hold (Spirtes et al., 2000; Pearl, 2009).

3.1 Effective-noise reduction and residual characterization

With $\eta_i \coloneqq f_i^2(Q_i) + g_i^2(Q_i)\epsilon_i$, Equation (1) becomes

$$v_i = f_i^1(K_i) + g_i^1(K_i)\eta_i, \tag{3}$$

a location-scale model on the observed variables with potentially dependent effective noises. Moreover, η_i is a deterministic function of $N_i = \{\epsilon_l : v_l \in \mathrm{an}(Q_i) \cup Q_i \cup \{x_i\}\}$, the exogenous noises of x_i, its hidden parents, and their ancestors.

To formalize residualization, let $S \subseteq X$ be an exclusion set with $x_i \in S$, and let $Z_{i,S} = \mathbf{x}_{X \setminus S}$ be the observed variables available when constructing a residual for x_i. Let $\mathcal{F}_i(S)$ and $\mathcal{G}_i(S)$ be classes of real-valued and positive-valued functions of $Z_{i,S}$, respectively. They contain the true functions f_i^1 and g_i^1 whenever $K_i \cap S = \varnothing$. Define

$$\mathcal{H}_i(S) = \left\{ \frac{x_i - f(Z_{i,S})}{g(Z_{i,S})} : f \in \mathcal{F}_i(S),\ g \in \mathcal{G}_i(S) \right\}, \qquad \mathcal{H} = \bigcup_{i,S} \mathcal{H}_i(S). \tag{4}$$

The same exclusion set is used for location and scale: an excluded variable may enter neither component. We write $h(i, S)$ for a generic element of $\mathcal{H}_i(S)$. We impose two no-cancellation conditions. The first requires every relevant exogenous noise, either entering through the latent structure or through an excluded observed parent, to remain detectable in the residual.

Assumption 2 (Residual irreducibility)**.** For every $h(i, S) \in \mathcal{H}_i(S)$:

(i) if $\epsilon_k \in N_i$, then $\epsilon_k \not\perp\!\!\!\perp h(i, S)$;

(ii) if $x_j \in K_i \cap S$, then $\epsilon_j \not\perp\!\!\!\perp h(i, S)$.

The second rules out cancellation between two residuals that retain a common exogenous noise.

Assumption 3 (Residual faithfulness)**.** For any $h_i, h_j \in \mathcal{H}$,

$$(\epsilon_k \not\perp\!\!\!\perp h_i) \wedge (\epsilon_k \not\perp\!\!\!\perp h_j) \quad \Longrightarrow \quad h_i \not\perp\!\!\!\perp h_j.$$

Thus, residual irreducibility ensures that relevant noise survives residualization, while residual faithfulness ensures that shared surviving noise induces residual dependence. Together they give the effective-noise characterization.

Lemma 2. *Under Assumptions 2 and 3,*

$$\eta_i \perp\!\!\!\perp \eta_j \quad \Longleftrightarrow \quad \textit{there is no UBP or UCP between } x_i \textit{ and } x_j.$$

Indeed, a UBP or UCP exists exactly when $N_i \cap N_j \neq \varnothing$. If $N_i \cap N_j = \varnothing$, η_i and η_j are functions of disjoint sets of mutually independent exogenous noises and are therefore independent. If $N_i \cap N_j \neq \varnothing$, Assumptions 2 and 3 imply dependence. Proofs are given in Appendix A.2.

Remark 1 (A violation)**.** The assumptions exclude exact cancellation phenomena. Let $u, \epsilon_1, \epsilon_2$ be independent standard normal variables and $x_1 = (u + \epsilon_1)/\sqrt{2}, \quad x_2 = (u + \epsilon_2)/\sqrt{2}$, with u hidden. Then $h_1 = x_1 - \frac{1}{2}x_2 \in \mathcal{H}_1(\{1\})$ and $h_2 = x_2$ both depend on u, yet $h_1 \perp\!\!\!\perp h_2$ by joint Gaussianity. Thus residual faithfulness fails, and the observed distribution cannot reveal the hidden backdoor structure by residual independence. We make no genericity claim that nonlinearity or non-Gaussianity alone guarantees these assumptions.

Proposition 1. *Under (1) and Assumptions 2 and 3,*

$$B(i,j) = 1 \iff \forall\, h_i \in \mathcal{H}_i(\{i\}),\ h_j \in \mathcal{H}_j(\{j\}) :\ h_i \not\perp\!\!\!\perp h_j, \tag{5}$$

$$A(i,j) = 1 \wedge B(i,j) = 0 \iff \Big[\forall\, h_i \in \mathcal{H}_i(\{i,j\}),\ h_j \in \mathcal{H}_j(\{j\}) :\ h_i \not\perp\!\!\!\perp h_j\Big] \tag{6}$$

$$\wedge \Big[\exists\, \tilde{h}_i \in \mathcal{H}_i(\{i\}),\ \tilde{h}_j \in \mathcal{H}_j(\{j\}) :\ \tilde{h}_i \perp\!\!\!\perp \tilde{h}_j\Big], \tag{7}$$

$$A(i,j) = A(j,i) = B(i,j) = 0 \iff \exists\, h_i \in \mathcal{H}_i(\{i,j\}),\ h_j \in \mathcal{H}_j(\{i,j\}) :\ h_i \perp\!\!\!\perp h_j. \tag{8}$$

For an invisible pair, a common noise in $N_i \cap N_j$ survives every admissible residual and therefore prevents independence. For a visible edge $x_j \to x_i$, excluding x_j forces the residual of x_i to retain ϵ_j, whereas allowing x_j permits the structural residual η_i; when $B(i,j) = 0$, this can be independent of η_j. For a visible non-edge, neither endpoint is an excluded parent of the other, so choosing the structural functions yields $\eta_i \perp\!\!\!\perp \eta_j$.

Example for Proposition 1. Consider Figure 1. Here u_1 is a hidden parent of x_1, x_2, giving the UBP $x_1 \leftarrow u_1 \rightarrow x_2$, while u_2 is a hidden descendant of x_5 and parent of x_3, giving the UCP $x_5 \rightarrow u_2 \rightarrow x_3$. The relevant structural equations are

$$x_1 = \eta_1, \qquad x_2 = \eta_2, \qquad x_3 = f_3^1(x_1, x_2, x_5) + g_3^1(x_1, x_2, x_5)\eta_3,$$

with noise sets $N_1 = \{\epsilon_{u_1}, \epsilon_1\}, \quad N_2 = \{\epsilon_{u_1}, \epsilon_2\}, \quad N_3 = \{\epsilon_5, \epsilon_{u_2}, \epsilon_3\}, \quad N_5 = \{\epsilon_5\}$.

Invisible pair: (x_1, x_2)**.** Since $\epsilon_{u_1} \in N_1 \cap N_2$, Assumption 2 implies that every $h_1 \in \mathcal{H}_1(\{1\})$ and $h_2 \in \mathcal{H}_2(\{2\})$ retains dependence on ϵ_{u_1}. Residual faithfulness therefore gives $h_1 \not\perp\!\!\!\perp h_2$ for all $h_1 \in \mathcal{H}_1(\{1\}),\ h_2 \in \mathcal{H}_2(\{2\})$, which is Equation (5).

Visible parent: $x_1 \rightarrow x_3$**, with** $B(3,1) = 0$**.** For Equation (6), take $S = \{3, 1\}$. Since the true parent x_1 is excluded from both the location and scale regressions, Assumption 2 implies that every $h_3 \in \mathcal{H}_3(\{3,1\})$ retains dependence on ϵ_1. Because $\epsilon_1 \in N_1$, every $h_1 \in \mathcal{H}_1(\{1\})$ also retains ϵ_1; residual faithfulness therefore gives $h_3 \not\perp\!\!\!\perp h_1$ for all such residuals.

For Equation (7), take $S = \{3\}$, so x_1 is available. Choosing the true structural functions gives

$$h_3^*(3, \{3\}) = \frac{x_3 - f_3^1(x_1, x_2, x_5)}{g_3^1(x_1, x_2, x_5)} = \eta_3, \qquad h_1^*(1, \{1\}) = \eta_1.$$

Since $N_3 \cap N_1 = \varnothing$, Lemma 2 gives $\eta_3 \perp\!\!\!\perp \eta_1$. Thus excluding the parent makes dependence unavoidable, whereas allowing it permits an independent residual pair. Examples for a visible non-edge (x_1, x_5) and an indirect observed path (x_1, x_4) are given in Appendix A.3.

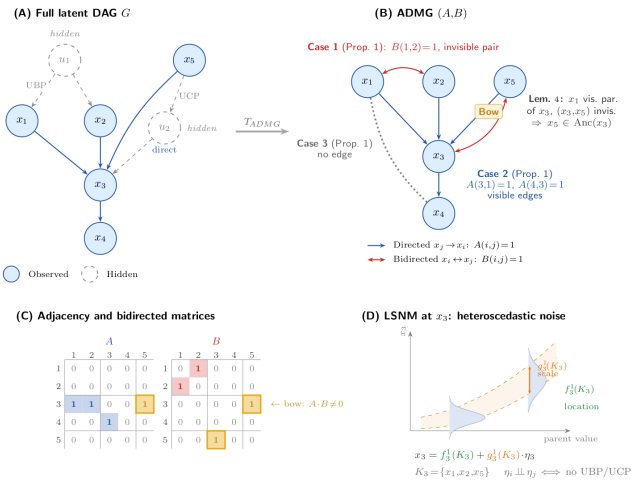


Figure 1: **(A)** Latent DAG with hidden u_1, u_2. **(B)** Projected ADMG with the three cases of Proposition 1. **(C)** Matrices A, B. **(D)** LSNM at x_3: K_3 sets location and scale; the residual η_3 carries the hidden-variable signal.

3.2 Bow-free ADMG identifiability

Define the pair oracle W_{RI} with query $q = (i, j)$: it returns $B(i, j) = 1$ if a UBP or UCP connects x_i and x_j, and otherwise $B(i, j) = 0$ together with $A(i, j)$ and $A(j, i)$. By Proposition 1 these are functions of P_X.

Theorem 3. *T_{ADMG} is identifiable over $\mathcal{G}_{full} = \{G : G$ a DAG and $T_{ADMG}(G)$ bow-free$\}$.*

Bow-freeness is required because when $B(i, j) = 1$, Proposition 1 identifies the pair as invisible but cannot determine whether a directed edge also exists between the same variables. Although this prevents full ADMG identification in the presence of bows, some ancestral information remains identifiable: under the conditions of Lemma 9, an observed variable x_k can still be identified as an ancestor of x_i (Appendix A.4.5). This yields ancestry rather than adjacency and is not used by the implementation.

4 Population oracle and oracle search

LSNM-UV (Algorithm 1, Appendix A.4) runs in two stages. Stage 1, LSNM-UV-Base (Algorithm 2, Appendix A.4), is the location-scale analogue of CAM-UV: Phase 1 finds parent candidates by sink tests on subsets of size at most d, Phase 2 prunes them, and Phase 3 marks pairs whose residuals cannot be made independent as invisible. Stage 2, CheckVisible (Algorithm 3), re-examines every invisible pair with larger regression sets. The output is a matrix with $A(i, j) = 1$ for $x_j \to x_i$, 0 for no edge, and NaN for a UBP or UCP.

4.1 The residual-independence oracle

Proposition 1 characterizes each pair (x_i, x_j) through residual-independence statements. At the pair level, let $W_{RI}(P_X, (i, j))$ denote this population answer: it classifies the pair as invisible, a visible directed edge, or a visible non-edge. For example, for a visible edge $x_j \to x_i$ with $B(i, j) = 0$, allowing x_j in the residualization of x_i permits the structural residual η_i, which can be independent of η_j, whereas excluding x_j makes dependence unavoidable.

The search does not query W_{RI} directly. Its primitive question is: *for these variables and these regression restrictions, can the residuals be made independent?* A residual query is $q = (i, J, S_i, (S_j)_{j\in J})$, where S_i, S_j are exclusion sets. Define the residual-independence oracle

$$\mathcal{O}_{\mathrm{RI}}(P_X, q) = \mathbf{1}\{\exists\, h_i \in \mathcal{H}_i(S_i),\ h_j \in \mathcal{H}_j(S_j)\ (j \in J) : h_i \perp\!\!\!\perp (h_j)_{j\in J}\}\,. \tag{9}$$

Thus $\mathcal{O}_{\mathrm{RI}}(q) = 1$ means that independence is achievable under the specified regression sets, whereas $\mathcal{O}_{\mathrm{RI}}(q) = 0$ means that dependence is unavoidable. For example, if $x_j \to x_i$ is visible, then

$$\mathcal{O}_{\mathrm{RI}}(i, j, \{i\}, \{j\}) = 1, \qquad \mathcal{O}_{\mathrm{RI}}(i, j, \{i, j\}, \{j\}) = 0 :$$

the first query allows x_j when residualizing x_i, while the second excludes it. Proposition 1 combines such oracle answers to recover W_{RI}.

$\mathcal{O}_{\mathrm{RI}}$ is a deterministic function of P_X and q, hence an oracle in the sense of Section 2. By Proposition 1 the pair answer $W_{RI}(P_X, (i, j))$ is a fixed Boolean combination of the answers to the four queries $(i, j, \{i\}, \{j\})$, $(i, j, \{i, j\}, \{j\})$, $(i, j, \{i\}, \{i, j\})$, $(i, j, \{i, j\}, \{i, j\})$, and W_{RI} determines $T_{ADMG}(G)$ on bow-free graphs (Theorem 3): residual queries determine pair classes, which determine the ADMG.

For an answer function $O : \mathcal{Q}_{p,d} \to \{0, 1\}$ let LSNM-UV(O) be Algorithm 1 with every decision "p-HSIC $> \alpha$" replaced by $O(q)$, and with Phase 1 examining the members of each subset in a fixed order (the implementation ranks them by p-value, Appendix A.5). Every branching of LSNM-UV(O) is a function of the answers received so far; $Q^*(P_X) \subseteq \mathcal{Q}_{p,d}$ is the finite set of queries issued by LSNM-UV($\mathcal{O}_{\mathrm{RI}}$). Soundness and completeness are defined in Appendix A.4 (Definition A.5).

Assumption 4 (Oracle search coverage)**.** For the oracle-driven execution of LSNM-UV-Base:

(i) every pair that the Base stage resolves as a visible directed edge or visible non-edge is correctly classified; and

(ii) for every pair (x_i, x_j) left unresolved for CHECKVISIBLE, its candidate pool Q_{ij} contains $(K_i \setminus \{x_j\}) \cup (K_j \setminus \{x_i\})$.

Theorem 4 (Oracle soundness and completeness)**.** *Suppose the target graph is bow-free, Assumptions 2 and 3 hold, and Assumption 4 holds. If* CHECKVISIBLE *searches all subsets of its candidate pool, then* LSNM-UV($\mathcal{O}_{\mathrm{RI}}$) *correctly classifies every observed pair as a visible directed edge, visible non-edge or invisible pair, and therefore returns* $T_{ADMG}(G)$.

The proof is in Appendix A.4.4. The theorem concerns the search logic; the implemented algorithm has data rather than the population oracle. Section 5 asks whether the data-based answers eventually coincide with those of the oracle.

5 STATISTICAL CONSISTENCY

In data, the population residual-independence oracle $\mathcal{O}_{\mathrm{RI}}$ is unavailable. For each query q, the algorithm fits residuals $\hat{h}_{i,q}$ and $\{\hat{h}_{j,q}\}_{j \in J}$, computes the empirical Hilbert–Schmidt independence criterion (Gretton et al., 2005) $\widehat{D}_q = \mathrm{HSIC}_n\left(\hat{h}_{i,q}, (\hat{h}_{j,q})_{j \in J}\right)$, and returns

$$\widehat{O}(q) = \mathbf{1}\{\widehat{D}_q \le \tau_n\}. \tag{10}$$

Thus $\widehat{O}(q) = 1$ is the empirical analogue of the oracle statement that an independent residual witness exists. Let $Q^*(P_X)$ denote the finite set of queries issued by the oracle-driven search LSNM-UV($\mathcal{O}_{\mathrm{RI}}$) and define

$$E_n = \left\{\widehat{O}(q) = \mathcal{O}_{\mathrm{RI}}(P_X, q) \text{ for every } q \in Q^*(P_X)\right\}. \tag{11}$$

On E_n, the empirical and oracle searches receive identical answers along the oracle path and therefore follow the same path and return the same graph. Hence graph consistency reduces to proving

$$P(E_n) \longrightarrow 1.$$

Simulation 7 (Appendix B.7) provides the empirical illustration of exactly this statement.

Population residual targets. Each query determines a regression problem whose population location and scale targets we denote by $f^*_{v,q}$ and $g^*_{v,q}$. Let

$$h^*_{v,q} = \frac{x_v - f^*_{v,q}}{g^*_{v,q}},$$

with the same invertible centering and scale normalization used by the implementation. These targets need not equal the structural functions. A sufficient condition under which $h^*_{v,q}$ is an affine transform of the effective noise is given in Proposition 2 (Appendix A.6).

Assumption 5 (Consistency conditions)**.** (C1) *(Admissible targets.)* For every $q \in Q^*(P_X)$ and endpoint v, the population target satisfies $h^*_{v,q} \in \mathcal{H}_v(S_v)$.

(C2) *(Targeting.)* For every $q \in Q^*(P_X)$ with $\mathcal{O}_{\mathrm{RI}}(P_X, q) = 1$, $h^*_{i,q} \perp\!\!\!\perp (h^*_{j,q})_{j\in J}$.

(C3) *(Statistic consistency.)* Let $D_q = \mathrm{HSIC}\big(h^*_{i,q}, (h^*_{j,q})_{j\in J}\big)$. We use characteristic kernels (Sriperumbudur et al., 2010), so $D_q = 0$ exactly when the target residuals are independent. For some deterministic sequence $a_n \to 0$,

$$\max_{q\in Q^*(P_X)} |\widehat{D}_q - D_q| = O_P(a_n).$$

(C4) *(Threshold.)* The threshold satisfies $\tau_n \to 0$, $\quad \frac{\tau_n}{a_n} \to \infty$.

Conditions C1–C2 concern whether the population targets of the regression procedure correspond to the residual witnesses used by the oracle. Condition C3 collects the statistical error introduced by residual estimation and finite-sample dependence estimation; Simulations 1 and 3–4 (Appendices B.1 and B.2) diagnose these components for the implemented estimator. Condition C4 separates this error from fixed positive population dependence; Simulation 5 (Appendix B.3) illustrates the resulting fixed-level versus vanishing-threshold tradeoff.

Theorem 5 (Query consistency)**.** *Under Conditions C1–C4,*

$$P(E_n) \longrightarrow 1.$$

If $\mathcal{O}_{\mathrm{RI}}(P_X, q) = 1$, Condition C2 gives $D_q = 0$, and C3–C4 imply $\widehat{D}_q = o_P(\tau_n)$. If $\mathcal{O}_{\mathrm{RI}}(P_X, q) = 0$, admissibility in C1 gives $D_q > 0$, while $\widehat{D}_q \to D_q$ and $\tau_n \to 0$. Thus every oracle-path query is answered correctly with probability tending to one, and, as in consistency proofs for the PC algorithm (Kalisch and Bühlmann, 2007), finiteness of $Q^*(P_X)$ gives $P(E_n) \to 1$; the full proof is in Appendix A.7.

Corollary 6 (Graph consistency)**.** *Under the assumptions of Theorems 4 and 5,*

$$P\Big(\widehat{G}_n = T_{ADMG}(G)\Big) \longrightarrow 1$$

on the bow-free target class.

Theorem 5 concerns a vanishing threshold. The practical algorithm instead uses the fixed operating level $\alpha = .01$, whose null decision error need not vanish. Simulation 5 (Appendix B.3) therefore compares this operating rule with a vanishing-threshold rule satisfying Condition C4.

6 Experiments

The experiments follow the same progression as the theory:

residual estimation $\longrightarrow$ residual-independence decisions $\longrightarrow$ pair classification $\longrightarrow$ graph recovery.

Controlled experiments are generated from (1); complete data-generating processes, replicate counts and diagnostic experiments are given in Appendix B. Throughout, n is the number of i.i.d. samples in a dataset, a data-based answer is the HSIC test on fitted residuals at the fixed level $\alpha = 0.01$, and the vanishing threshold of Condition C4 multiplies the critical value of that test by $\max\{1, n/400\}^{0.75}$.

We first examine the statistical conditions behind Theorem 5. On the single edge $X \to Y$, where the population location, scale shape and effective residual are known, all three estimation errors decrease with sample size (Simulation 1, Appendix B.1). We then compare the empirical dependence statistic $\widehat{D}_q$ with its population target D_q. With the correct regression sets supplied, the total error decreases with n; over the investigated range, the contribution from fitted residuals is smaller than ordinary finite-sample variation of the dependence statistic (Simulations 3–4, Appendix B.2). These experiments directly diagnose Condition C3 for the implemented estimator.

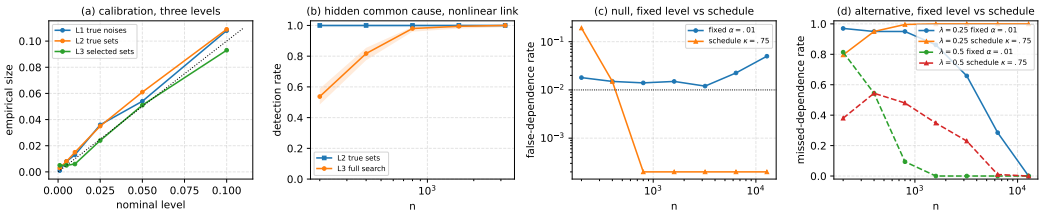


Figure 2: **From residuals to decisions.** (a) Rejection rate against nominal level, null fork, $n = 1600$, three information levels. (b) Detection rate of an identifiable nonlinear hidden common cause against n, true (L2) and selected (L3) parent sets. (c,d) False-dependence (null) and missed-dependence (weak couplings) rates against n: fixed level $\alpha = 0.01$ versus vanishing threshold.

Figure 2 (Simulations 2 and 5, Appendices B.4 and B.3) separates the remaining sources of decision error. L1 applies the independence test to the true effective noises, L2 uses fitted residuals with the true parent sets supplied, and L3 runs the complete search. Hence L1–L2 isolates residual estimation, while L2–L3 additionally includes regression-set selection. On the null fork, L1 and L2 are close to nominal calibration and L3 is conservative. For an identifiable nonlinear hidden common cause, L1 and L2 detect the dependence throughout, whereas L3 rises from 0.54 at $n = 200$ to 0.98 at $n = 800$ and essentially one thereafter. Finally, the fixed operating level $\alpha = .01$ retains a non-vanishing null error, while the vanishing statistic-scale threshold drives this error toward zero at the cost of power against weak alternatives at moderate n, illustrating Condition C4.

We next examine the complete residual-dependence signatures of Proposition 1 on elementary motifs (Simulation 6, Appendix B.6). Visible edges and non-edges are recovered near the ceiling imposed by the fixed-level test. Linear-Gaussian hidden mechanisms violate residual faithfulness and remain non-identifiable, whereas nonlinear hidden mechanisms become detectable at sample sizes determined by their weakest residual- dependence signal. Strengthening the departure from a location–scale conditional law correspondingly shifts hidden-pair recovery to smaller samples.

To separate search logic from statistical error, we also run the same search on a four-node graph with one UBP (Simulation 7, Appendix B.7), answering every residual query either exactly from the true graph or by the data-based test. The oracle-driven search recovers the graph throughout, while the agreement of the data-based answers with the oracle approaches one and the graph error approaches zero as n grows from 200 to 6400. A larger oracle experiment on the ten-node benchmark gives pair error below 0.01, showing that the remaining finite-sample gap is primarily statistical rather than a failure to combine correct oracle answers (Simulation 8, Appendix B.8).

The ten-node benchmark combines several observed parents, two UBPs and two UCPs in random bow-free graphs (Appendix B.8). Figure 3 compares LSNM-UV with CAM-UV, FCI and BANG on identical graphs and samples. LSNM-UV improves with sample size in both directed and bidirected recovery. At $n = 1000$ it attains directed F1 0.71, compared with 0.14 for CAM-UV and below 0.04 for FCI and BANG. For bidirected pairs it attains F1 0.31: its advantage comes primarily from precision (0.51), whereas CAM-UV and BANG obtain higher recall but substantially lower precision. Detailed oracle comparisons and robustness experiments are given in Appendix B.8.

6.1 NetSim fMRI benchmark

We finally evaluate on NetSim (Smith et al., 2011), simulated fMRI in which neural activity propagates through a known network and passes through a haemodynamic model with measurement noise. For each of the 26 acyclic configurations we withhold regions to create 100 hidden-variable datasets and run all four methods on identical samples (Appendix B.12). Performance varies substantially across configurations (Figure 15): CAM-UV has the higher directed F1 on average, while LSNM-UV is best in 8 of the 26. On sim19, a fast-sampling configuration (TR 0.25 s) with 100 ms neural lag (Figure 4), LSNM-UV obtains directed precision 0.76, recall 0.95 and F1 0.85, against F1 0.79, 0.11 and 0.53 for CAM-UV, FCI and BANG.

Bidirected recovery is weak for all methods, particularly the residual-based ones (pooled LSNM-UV bidirected F1 0.02): most true hidden pairs are either classified as directed before the hidden-pair

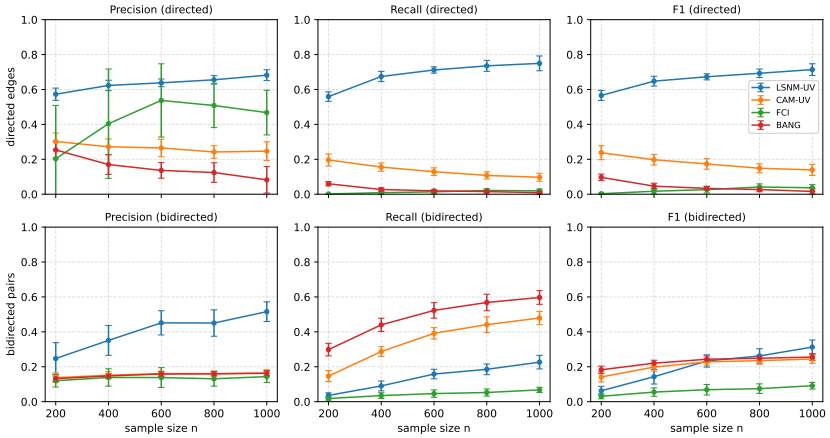


Figure 3: **Ten-node benchmark.** Precision, recall and F1 for directed edges (top) and bidirected pairs (bottom) as sample size increases. Each graph has observed directed structure, two UBPs and two UCPs; all methods run on the same 400 bow-free graphs and samples per n; error bars are ± 1 s.d. over 16 batches of 25 graphs.

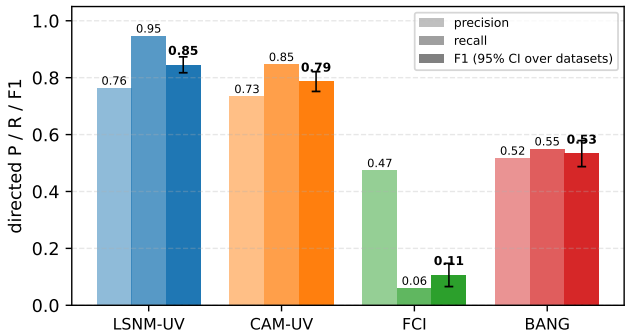


Figure 4: **NetSim sim19** (fast sampling, TR 0.25 s, with 100 ms neural lag), 100 hidden-region datasets: directed precision, recall and F1 for the four methods; error bars show 95% intervals over datasets.

stage or reach the residual test without being detected (Appendix B.12), consistent with weak residual-dependence margins, as in the motif experiments (Simulation 6).

7 Discussion and Limitations

The paper separates a population statement, an algorithmic one and a statistical one, each with its own assumption and each assumption visible in an experiment; the limitations follow from the assumptions. The common-frame condition is essential: if a hidden shift bypasses g_i^1, the structural residual still depends on K_i (Remark 2). Residual faithfulness excludes hidden pairs whose conditional laws are themselves location-scale families (Remark 1); such pairs are absorbed as directed edges by every location-scale method, and a small margin $\delta(P_X)$ means a large sample before a pair becomes visible, which is what limits bidirected recall. Bow-freeness is needed for the full ADMG. The consistency theorem takes the estimator's convergence as a condition and needs the targets to be witnesses (C2). The sufficient targeting condition of Proposition 2 can fail when an included regressor depends on the effective noise; characterizing targeting directly at the decision level is a natural next step. In practice the search uses a fixed level, not covered by the theorem, although the schedule the theorem covers already beats it from $n \approx 1600$ on the benchmark; and on recordings the common-frame assumption cannot be verified from the data alone, so the output is a set of hypotheses about direct effects and hidden confounding, with the NetSim hidden pairs at the floor for every residual-based method as the limitation of residual faithfulness in practice.

AI USE STATEMENT

In this work, we used generative AI tools for assist in generating parts of the simulation and analysis code, which the authors verified and tested for correctness. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims. Additionally, we used generative AI tools for draft parts of a research paper, summarize or analyse existing literature, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All results are produced by the frozen implementation described in Appendix A.5; the generators of every experiment are specified in Appendix B, every reported number is read from logged runs with stated seeds and replicate counts, and the code and run logs will be released with the camera-ready version.

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 11th International Conference on Learning Representations (ICLR)*, 2023.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Arthur Gretton, Olivier Bousquet, Alexander Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory (ALT)*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20 (NeurIPS)*, pages 585–592, 2008.

Patrik O. Hoyer, Shohei Shimizu, Antti J. Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pages 689–696, 2009.

Alexander Immer, Christoph Schultheiss, Julia E. Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 14316–14332, 2023.

Ryo Kamimura and Thong Pham. Causal Discovery via Transformed Low-Rank Quantile Surfaces. arXiv:2609.16931, 2026.

Yingyu Lin, Yuxing Huang, Wenqin Liu, Haoran Deng, Ignavier Ng, Kun Zhang, Mingming Gong, Yian Ma, and Biwei Huang. A skewness-based criterion for addressing heteroscedastic noise in causal discovery. In *The 13th International Conference on Learning Representations (ICLR)*, 2025.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *PMLR*, pages 97–106, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 589–598, 2011.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Thong Pham, Takashi Nicholas Maeda, and Shohei Shimizu. Causal additive models with unobserved causal paths and backdoor paths. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *PMLR*, pages 1585–1593, 2026.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003. doi: 10.1111/1467-9469.00323.

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

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

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

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13 (2):689–705, 1985.

Xiangyu Sun and Oliver Schulte. Cause-effect inference in location-scale noise models: Maximum likelihood vs. independence testing. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 5447–5483, 2023.

Y. Samuel Wang, Mladen Kolar, and Mathias Drton. Confidence sets for causal orderings. *Journal of the American Statistical Association*, 121(553):690–703, 2026. doi: 10.1080/01621459.2025.2542552.

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

Naiyu Yin, Tian Gao, Yue Yu, and Qiang Ji. Effective causal discovery under identifiable heteroscedastic noise model. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 38(15):16486–16494, 2024. doi: 10.1609/aaai.v38i15.29586.

David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 15639–15654, 2023.

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

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 647–655, 2009.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 IDENTIFIABILITY VIA ORACLES

Proof of Lemma 1. Suppose T is oracle identifiable with respect to W, and let $P_{G,\theta} = P_{G',\theta'}$. Since W is a deterministic function of P_X, we have $W(P_{G,\theta}, q) = W(P_{G',\theta'}, q)$ for all q. By oracle identifiability, $T(G) = T(G')$. □

A.2 THEORETICAL PROOFS

Notation. Throughout, $h(i, S)$ is a residual in $\mathcal{H}_i(S)$ (Equation (4)). The exclusion set S is common to the location and the scale function, so an excluded variable enters neither. Because $\mathcal{F}_i(S)$ and $\mathcal{G}_i(S)$ contain f_i^1 and g_i^1 whenever $K_i \cap S = \varnothing$, the structural residual $\eta_i = (x_i - f_i^1(K_i))/g_i^1(K_i)$ is admissible in $\mathcal{H}_i(S)$ for every such S.

Remark 2 (Why the common frame matters)**.** If the hidden shift does not share the observed-parent scale, for example

$$x_i = f_i^1(K_i) + f_i^2(Q_i) + g_i^1(K_i)g_i^2(Q_i)\epsilon_i,$$

then the natural structural residual becomes

$$\frac{x_i - f_i^1(K_i)}{g_i^1(K_i)} = \frac{f_i^2(Q_i)}{g_i^1(K_i)} + g_i^2(Q_i)\epsilon_i,$$

which still depends on K_i. The common-frame restriction in (1) is precisely what removes this interaction.

A.2.1 PROOF OF LEMMA 2

Proof. Recall that

$$\eta_i = f_i^2(Q_i) + g_i^2(Q_i)\epsilon_i$$

is a deterministic function of

$$N_i = \{\epsilon_\ell : v_\ell \in \mathrm{an}(Q_i) \cup Q_i \cup \{x_i\}\}.$$

First observe that

$$N_i \cap N_j \neq \varnothing \quad \Longleftrightarrow \quad \text{a UBP or UCP connects } x_i \text{ and } x_j.$$

A UBP places the noise of a common ancestor of hidden parents of the two endpoints in both noise sets. A UCP from x_j to x_i places ϵ_j in N_i as well as N_j, and symmetrically in the opposite direction. Conversely, a shared noise belonging to one endpoint corresponds to a UCP from that endpoint, while any other shared noise is associated with a common ancestor of hidden parents of both endpoints and hence a UBP.

Suppose first that no UBP or UCP connects x_i and x_j. Then $N_i \cap N_j = \varnothing$. Since the exogenous noises are mutually independent, N_i and N_j are independent collections. Because η_i and η_j are functions of these disjoint collections,

$$\eta_i \perp\!\!\!\perp \eta_j.$$

Conversely, suppose that a UBP or UCP connects x_i and x_j. Then choose

$$\epsilon_k \in N_i \cap N_j.$$

The structural residuals

$$\eta_i = \frac{x_i - f_i^1(K_i)}{g_i^1(K_i)} \in \mathcal{H}_i(\{i\}), \qquad \eta_j = \frac{x_j - f_j^1(K_j)}{g_j^1(K_j)} \in \mathcal{H}_j(\{j\})$$

are admissible. Residual irreducibility gives

$$\epsilon_k \not\perp\!\!\!\perp \eta_i, \qquad \epsilon_k \not\perp\!\!\!\perp \eta_j,$$

and residual faithfulness therefore gives

$$\eta_i \not\perp\!\!\!\perp \eta_j.$$

Hence

$$\eta_i \perp\!\!\!\perp \eta_j \quad \Longleftrightarrow \quad \text{no UBP or UCP connects } x_i \text{ and } x_j.$$

□

A.2.2 Proof of Proposition 1

Consequences of residual irreducibility. Assumption 2 immediately gives two properties used repeatedly below. For every admissible $h_i \in \mathcal{H}_i(S)$,

$$\epsilon_k \in N_i \Longrightarrow \epsilon_k \not\perp\!\!\!\perp h_i, \tag{12}$$

$$x_j \in K_i \cap S \Longrightarrow \epsilon_j \not\perp\!\!\!\perp h_i. \tag{13}$$

In particular, since $\epsilon_i \in N_i$, every admissible residual of x_i retains dependence on its own exogenous noise ϵ_i. No separate "irremovability" lemma is required.

Proof of Proposition 1. We treat the three pair classes separately.

Invisible pair. Suppose first that $B(i,j) = 1$. By Lemma 2, there exists

$$\epsilon_k \in N_i \cap N_j.$$

For every $h_i \in \mathcal{H}_i(\{i\})$ and $h_j \in \mathcal{H}_j(\{j\})$, by (12),

$$\epsilon_k \not\perp\!\!\!\perp h_i, \qquad \epsilon_k \not\perp\!\!\!\perp h_j.$$

Residual faithfulness therefore implies

$$h_i \not\perp\!\!\!\perp h_j$$

for every admissible pair, proving (5).

Conversely, if $B(i,j) = 0$, choose the structural residuals

$$h_i = \eta_i = \frac{x_i - f_i^1(K_i)}{g_i^1(K_i)} \in \mathcal{H}_i(\{i\}), \qquad h_j = \eta_j \in \mathcal{H}_j(\{j\}).$$

Lemma 2 gives $\eta_i \perp\!\!\!\perp \eta_j$, so universal dependence cannot hold. This proves the first equivalence.

Visible edge. Suppose $A(i,j) = 1$ and $B(i,j) = 0$, i.e. $x_j \to x_i$ is visible. For every $h_i \in \mathcal{H}_i(\{i,j\})$, the true parent x_j is excluded. Residual irreducibility therefore gives

$$\epsilon_j \not\perp\!\!\!\perp h_i.$$

Moreover $\epsilon_j \in N_j$, so for every $h_j \in \mathcal{H}_j(\{j\})$,

$$\epsilon_j \not\perp\!\!\!\perp h_j.$$

Residual faithfulness yields

$$h_i \not\perp\!\!\!\perp h_j$$

for every such pair, proving (6).

When x_j is allowed in the residualization of x_i, choose the structural residuals $h_i = \eta_i$ and $h_j = \eta_j$. Because $B(i,j) = 0$, Lemma 2 gives

$$\eta_i \perp\!\!\!\perp \eta_j,$$

proving (7).

Conversely, suppose (6) and (7) hold. The existence of an independent pair in (7) rules out $B(i,j) = 1$ by the first part of the proposition. Hence $B(i,j) = 0$.

If $A(i,j) = 0$, then $x_j \notin K_i$, so the structural residual η_i is admissible in $\mathcal{H}_i(\{i,j\})$; likewise $\eta_j \in \mathcal{H}_j(\{j\})$. Since $B(i,j) = 0$, Lemma 2 gives $\eta_i \perp\!\!\!\perp \eta_j$, contradicting the universal dependence in (6). Therefore $A(i,j) = 1$.

Visible non-edge. Suppose

$$A(i,j) = A(j,i) = B(i,j) = 0.$$

Neither endpoint is a parent of the other, so the structural residuals are admissible with both endpoints excluded:

$$\eta_i \in \mathcal{H}_i(\{i,j\}), \qquad \eta_j \in \mathcal{H}_j(\{i,j\}).$$

Since $B(i,j) = 0$, Lemma 2 gives

$$\eta_i \perp\!\!\!\perp \eta_j,$$

proving existence of an independent residual pair.

Conversely, suppose there exist

$$h_i \in \mathcal{H}_i(\{i,j\}), \qquad h_j \in \mathcal{H}_j(\{i,j\})$$

with $h_i \perp\!\!\!\perp h_j$.

If $B(i,j) = 1$, choose $\epsilon_k \in N_i \cap N_j$. Residual irreducibility and residual faithfulness would imply $h_i \not\perp\!\!\!\perp h_j$, a contradiction. Hence $B(i,j) = 0$.

If $A(i,j) = 1$, then $x_j \in K_i \cap \{i,j\}$, so residual irreducibility gives

$$\epsilon_j \not\perp\!\!\!\perp h_i.$$

Since $\epsilon_j \in N_j$, residual irreducibility also gives

$$\epsilon_j \not\perp\!\!\!\perp h_j.$$

Residual faithfulness again contradicts $h_i \perp\!\!\!\perp h_j$. Thus $A(i,j) = 0$. The same argument with i and j exchanged gives $A(j,i) = 0$.

This proves all three characterizations. □

A.2.3 Proof of Theorem 3

Proof of Theorem 3. Define the oracle W_{RI} with query $q = (i,j)$ as follows.

When there is a UBP or UCP between x_i and x_j, Proposition 1, Case 1, gives that (5) holds; the oracle returns $B(i,j) = 1$.

When there is no UBP or UCP between x_i and x_j, Case 1 gives that (5) does not hold; the oracle returns $B(i,j) = 0$. It additionally returns the value of $A(i,j)$: if $x_j \in K_i$, Case 2 gives that (6) and (7) hold, so the oracle returns $A(i,j) = 1$; otherwise Case 3 gives that (8) holds, so the oracle returns $A(i,j) = 0$.

These three cases are mutually exclusive. For a bow-free ADMG they are also exhaustive: the bow-free condition $A(i,j)\,B(i,j) = 0$ rules out $A(i,j) = 1$ and $B(i,j) = 1$ simultaneously, so every pair (x_i, x_j) falls into exactly one case. Since each case is determined by dependence and independence tests on the observed distribution, the oracle W_{RI} is a function of the observed distribution, and together mutual exclusivity and exhaustiveness give identifiability of the full ADMG.

□

A.3 Worked examples of Proposition 1 and Lemma 9

The examples complete the example of Section 3.1 on the graph of Figure 1, whose structural equations (Equation (3)) are

$$\begin{aligned}
&x_1 = \eta_1, && \eta_1 = f_1^2(u_1) + g_1^2(u_1)\,\epsilon_1 && (K_1 = \varnothing,\ Q_1 = \{u_1\})\\
&x_2 = \eta_2, && \eta_2 = f_2^2(u_1) + g_2^2(u_1)\,\epsilon_2 && (K_2 = \varnothing,\ Q_2 = \{u_1\})\\
&x_5 = \epsilon_5 && && (K_5 = Q_5 = \varnothing)\\
&x_3 = f_3^1(K_3) + g_3^1(K_3)\,\eta_3, && \eta_3 = f_3^2(u_2) + g_3^2(u_2)\,\epsilon_3 && (K_3 = \{x_1, x_2, x_5\},\ Q_3 = \{u_2\})\\
&x_4 = f_4^1(x_3) + g_4^1(x_3)\,\epsilon_4 && && (K_4 = \{x_3\},\ Q_4 = \varnothing)
\end{aligned}$$

with noise sets $N_1 = \{\epsilon_{u_1}, \epsilon_1\}$, $N_2 = \{\epsilon_{u_1}, \epsilon_2\}$, $N_5 = \{\epsilon_5\}$, $N_3 = \{\epsilon_5, \epsilon_{u_2}, \epsilon_3\}$ and $N_4 = \{\epsilon_4\}$. The invisible pair (x_1, x_2) and the visible edge $x_1 \to x_3$ are treated in the main text.

Visible non-edge: (x_1, x_5)**,** $A(1,5) = A(5,1) = B(1,5) = 0$**.** With $S = \{1,5\}$ the regression set is $\{x_2, x_3, x_4\}$. Since $K_1 = K_5 = \varnothing$, the structural residuals $\eta_1 = x_1 \in \mathcal{H}_1(\{1,5\})$ and $\eta_5 = x_5 = \epsilon_5 \in \mathcal{H}_5(\{1,5\})$ are admissible. Since $N_1 \cap N_5 = \varnothing$, Lemma 2 gives $\eta_1 \perp\!\!\!\perp \eta_5$, which is Equation (8): the pair is a visible non-edge although x_1 and x_5 share the child x_3.

Indirect path: (x_1, x_4), $A(4,1) = A(1,4) = B(1,4) = 0$. Despite the causal chain $x_1 \to x_3 \to x_4$, there is no direct edge. With $S = \{1,4\}$ the structural residuals are admissible, $\eta_1 = x_1$ and $(x_4 - f_4^1(x_3))/g_4^1(x_3) = \epsilon_4$, and $N_1 \cap N_4 = \varnothing$ gives $\eta_1 \perp\!\!\!\perp \epsilon_4$, so Equation (8) holds. The regression of x_4 on its observed parent x_3 absorbs the entire $x_1 \to x_3 \to x_4$ signal.

Bow: (x_3, x_5), $A(3,5) = B(3,5) = 1$. Here x_5 is a direct parent of x_3 and also reaches it through the UCP $x_5 \to u_2 \to x_3$, so $\epsilon_5 \in N_3 \cap N_5$. By (12), every admissible residual of x_3 and every admissible residual of x_5 retains ϵ_5, whatever the exclusion sets, and residual faithfulness makes every such pair dependent. Hence Equation (5) holds while neither (7) nor (8) can hold: residual independence classifies the pair as invisible and cannot reveal the directed edge. Lemma 9 still yields ancestral information. Take x_1 as the visible parent of x_3 ($A(3,1) = 1$, $B(3,1) = 0$) and (x_3, x_5) as the invisible pair. Excluding x_5 from the residualization of x_3 leaves ϵ_5 in every $h(3, \{3,5\})$ by (13); when the dependence in Equation (20) holds, the lemma identifies x_5 as an ancestor of x_3. Bow-freeness (Definition 2.4) remains necessary for full ADMG recovery (Theorem 3).

A.4 ALGORITHM

By Theorem 3, T_{ADMG} is identifiable on bow-free graphs; when $B(i,j) = 1$ residual independence cannot reveal whether $A(i,j) = 1$, and bow-freeness forces $A(i,j) = 0$. This subsection describes a combinatorial search that classifies every pair (x_i, x_j) as a visible directed edge, a visible non-edge or an invisible pair. The pair-level characterization of Proposition 1 holds for any DAG, but the search guarantee (Theorem 4) is stated for bow-free target graphs under the search-coverage condition of Assumption 4: the argument that the Base stage collects the parents it needs (the invariant in the proof of Lemma 7) uses bow-freeness, and we make no claim beyond that class. On a bow-free graph the invisible pairs are exactly the bidirected edges, so the three pair classes determine the full ADMG.

Algorithm 1 LSNM-UV

Require: Data matrix $\mathbf{X}$ for p observed variables, max parents d, significance level α
Ensure: Adjacency matrix A

```
Stage 1: A ← LSNM-UV-Base(X, d, α)                    [Algorithm 2]
Stage 2: Re-examine invisible pairs
S ← {(i, j) | A(i, j) = A(j, i) = NaN}
for each (i, j) ∈ S with i < j do
    CHECKVISIBLE(i, j)                                 [Algorithm 3]
end for
return A
```

Remark 3 (Comparison with CAM-UV-X of Pham et al. (2026))**.** The CAM-UV-X algorithm of Pham et al. (2026) includes three additional stages beyond the base algorithm and CHECKVISIBLE: CHECKONPATH (which tests invisibility with respect to subsets $X \setminus \{x_k\}$), CHECKCI (which uses conditional independence among original variables), and CHECKPARENTINVI (which propagates parentship information). CHECKONPATH and CHECKPARENTINVI rely on the residual independence characterization holding for arbitrary subsets $X' \subseteq X$ (Lemmas 1–3 of Pham et al. (2026), the subset versions of Lemmas 1–3 of Maeda and Shimizu (2021)); CHECKCI needs invisibility only with respect to X (Pham et al. (2026), Appendix B).

In the CAM-UV model Maeda and Shimizu (2021); Pham et al. (2026), reclassifying an observed parent x_k as unobserved moves its additive term from the observed-parent sum to the unobserved-parent sum, so the model on $X' = X \setminus \{x_k\}$ is again a CAM-UV model (the closure property of Pham et al. (2026), Appendix B). Therefore, Lemmas 1–3 can be restated and reproved for X' using the same arguments.

In the LSNM-UV model, removing an observed parent x_k from the observed set does *not* preserve the LSNM structure. If x_k appears in the scale function g_i^1 of some x_i, the resulting model for x_i on $X' = X \setminus \{x_k\}$ has a noise structure that depends on the hidden x_k, which is no longer an LSNM. Consequently, Proposition 1 cannot be applied on subsets $X' \subsetneq X$, and the *language* of subset visibility used in Pham et al. (2026) does not transfer.

However, the *core argument* behind CHECKONPATH, identifying which observed variables lie on invisible paths , does not require subset visibility. It can be stated directly in terms of the location-scale residuals $h(i, S)$ on the full data X, using the residual irreducibility and faithfulness assumptions that underlie Proposition 1. We formalize this in Lemma 9, which provides the LSNM analogue of Lemma 4 of Pham et al. (2026) without invoking subset visibility.

Let us start by defining visibility/invisibility for our LSNM:

Definition A.1 (Visible parent)**.** x_j is a *visible parent* of x_i if and only if $x_j \in K_i$ and there is no UBP/UCP between x_i and x_j. Equivalently (by Proposition 1, Case 2):

$$\forall\, h_1, h_2 \in \mathcal{H} :\ h_1(i, \{i, j\}) \not\perp\!\!\!\perp h_2(j, \{j\}), \tag{14}$$

$$\exists\, h_1, h_2 \in \mathcal{H} :\ h_1(i, \{i\}) \perp\!\!\!\perp h_2(j, \{j\}). \tag{15}$$

Definition A.2 (Visible non-edge)**.** (x_i, x_j) is a *visible non-edge* if and only if $A(i, j) = 0$, $A(j, i) = 0$, and there is no UBP/UCP between x_i and x_j. Equivalently (by Proposition 1, Case 3):

$$\exists\, h_1, h_2 \in \mathcal{H} :\ h_1(i, \{i, j\}) \perp\!\!\!\perp h_2(j, \{i, j\}). \tag{16}$$

Definition A.3 (Invisible pair)**.** (x_i, x_j) is an *invisible pair* if and only if there is a UBP/UCP between x_i and x_j. Equivalently (by Proposition 1, Case 1):

$$\forall\, h_1, h_2 \in \mathcal{H} :\ h_1(i, \{i\}) \not\perp\!\!\!\perp h_2(j, \{j\}). \tag{17}$$

A.4.1 LSNM-UV-BASE

Stage 1: LSNM-UV-Base. This is the LSNM analogue of CAM-UV Maeda and Shimizu (2021): Phases 1–2 adapt their Algorithm 1 and Phase 3 adapts their Algorithm 2, replacing additive residuals by location-scale residuals. The algorithm has three phases: Phase 1 extracts parent candidates, Phase 2 prunes false candidates, and Phase 3 detects invisible pairs.

Definition A.4 (Sink)**.** Let $K \subseteq X$ be a set of observed variables and $x_i \in K$. We call x_i a *sink* of K if no variable $x_j \in K \setminus \{x_i\}$ is a descendant of x_i in G.

Throughout, $h(i, S) \in \mathcal{H}_i(S)$ denotes a location-scale residual with the common exclusion set S of Equation (4): an excluded variable enters neither the location nor the scale function. To regress x_i on a desired set $R \subseteq X \setminus \{x_i\}$ we take $S = X \setminus R$, so that $x_i \in S$; the location-scale regression of Appendix A.5 fits f and g on the same regression set R.

When we write "p-HSIC" below, we mean the p-value of the gamma independence test based on the Hilbert–Schmidt Independence Criterion.

Lemma 7 (Sink identification in LSNM-UV)**.** *Let $K \subseteq X$ with $x_i \in K$, and let $M_l \subseteq X \setminus K$ be a set of previously identified parent candidates for each x_l. If the following two conditions hold, then x_i is a sink of K:*

$$\exists\, h_i \in \mathcal{H} :\ \forall\, x_j \in K \setminus \{x_i\},\ \exists\, h_j \in \mathcal{H} :$$
$$h_i\Big(i,\ X \setminus (M_i \cup K \setminus \{x_i\})\Big) \ \perp\!\!\!\perp\ h_j\Big(j,\ X \setminus M_j\Big), \tag{18}$$

$$\forall\, x_j \in K \setminus \{x_i\},\ \forall\, h_i', h_j' \in \mathcal{H} :$$
$$h_i'\Big(i,\ X \setminus M_i\Big) \ \not\perp\!\!\!\perp\ h_j'\Big(j,\ X \setminus M_j\Big). \tag{19}$$

Proof of Lemma 7. We prove the contrapositive: if x_i is not a sink of K, then at least one of (18) or (19) fails.

Since x_i is not a sink, there exists $x_j \in K \setminus \{x_i\}$ that is a descendant of x_i in G. The following three cases are exhaustive: either x_i is a direct parent of x_j (Case A), or $A(j, i) = 0$ with $B(i, j) = 1$ (Case B), or $A(j, i) = 0$ with $B(i, j) = 0$ (Case C).

Case A: x_i is a direct parent of x_j ($A(j, i) = 1$). We show (18) fails. Since $M_j \subseteq X \setminus K$ and $x_i \in K$, we have $x_i \notin M_j$. This follows directly from the lemma hypothesis $M_j \subseteq X \setminus K$: since

$x_i \in K$, M_j and K are disjoint, so $x_i \notin M_j$. Hence $x_i \in K_j \cap (X \setminus M_j)$: the true parent x_i is excluded from every $h_j(j, X \setminus M_j)$, and (13) gives

$$\epsilon_i \not\perp\!\!\!\perp h_j(j, X \setminus M_j) \quad \text{for all admissible } h_j.$$

Since $\epsilon_i \in N_i$, (12) gives $\epsilon_i \not\perp\!\!\!\perp h_i\big(i,\ X \setminus (M_i \cup K \setminus \{x_i\})\big)$ for all admissible h_i. Since both residuals retain the same exogenous noise ϵ_i, Assumption 3 gives:

$$h_i\big(i,\ X \setminus (M_i \cup K \setminus \{x_i\})\big) \not\perp\!\!\!\perp h_j(j,\ X \setminus M_j) \quad \text{for all } h_i, h_j \in \mathcal{H}.$$

No h_i can achieve independence with h_j, so (18) fails.

Case B: $B(i,j) = 1$ **and** $A(j,i) = 0$ **(invisible pair).** We show (18) fails. Since $B(i,j) = 1$, there exists shared exogenous noise $\epsilon_l \in N_i \cap N_j$ (Lemma 2). By (12), this noise is retained by every admissible residual of x_i and of x_j, whatever the exclusion sets:

$$\epsilon_l \not\perp\!\!\!\perp h_i(i, S_1) \quad \text{and} \quad \epsilon_l \not\perp\!\!\!\perp h_j(j, S_2) \quad \text{for all } h_i, h_j \in \mathcal{H} \text{ and all valid } S_1, S_2.$$

By Assumption 3, $h_i \not\perp\!\!\!\perp h_j$ for all $h_i, h_j \in \mathcal{H}$, and (18) fails.

Case C: $A(j,i) = 0$ **and** $B(i,j) = 0$ **(indirect descendant, no confounding).** We show (19) fails.

We first establish the following invariant, which the algorithm maintains under bow-freeness:

$$K_\ell \subseteq M_\ell \quad \text{for every } x_\ell \in K.$$

Parents of x_ℓ outside K are covered by the standard inductive hypothesis $K_\ell \setminus K \subseteq M_\ell$. For a parent $x_k \in K_\ell \cap K$, bow-freeness gives $A(\ell,k)\,B(\ell,k) = 0$; since $A(\ell,k) = 1$, we have $B(\ell,k) = 0$, so (x_ℓ, x_k) is a visible pair. The pair $\{x_\ell, x_k\}$ was processed in an earlier iteration with $t = 2$ (the reset rule $t \leftarrow 2$ in Algorithm LSNM-UV-Base, ensures pair-level subsets are revisited whenever a new parent is added). Since $A(\ell,k) = 1$ and $B(\ell,k) = 0$, by Case A the sink test for $\{x_\ell, x_k\}$ correctly identifies x_ℓ as the sink, adding x_k to M_ℓ. Hence $K_\ell \subseteq M_\ell$.

By acyclicity, x_j being a descendant of x_i rules out $x_j \in K_i$. Symmetrically, $A(j,i) = 0$ gives $x_i \notin K_j$. Combined with the invariant above, $K_i \subseteq M_i$ and $K_j \subseteq M_j$.

Choose h_i' using the true structural functions $f = f_i^1$, $g = g_i^1$, with exclusion set $X \setminus M_i$ (so the regression uses M_i). Since $K_i \subseteq M_i$, both f_i^1 and g_i^1 have access to all true parents of x_i; the structural residual is admissible because $K_i \cap (X \setminus M_i) = \varnothing$, and Equation (3) gives:

$$h_i'(i,\ X \setminus M_i) = \frac{x_i - f_i^1(K_i)}{g_i^1(K_i)} = \eta_i.$$

Symmetrically, $h_j'(j,\ X \setminus M_j) = \eta_j$. Since $B(i,j) = 0$, Lemma 2 gives $\eta_i \perp\!\!\!\perp \eta_j$. Therefore $\exists\, h_i', h_j' \in \mathcal{H}$ with

$$h_i'(i,\ X \setminus M_i) \perp\!\!\!\perp h_j'(j,\ X \setminus M_j),$$

and (19) fails.

□

Algorithm 2 LSNM-UV-Base

Require: Data matrix $\mathbf{X}$ for p observed variables, max subset size d, significance level α
Ensure: Adjacency matrix A, parent sets $\{M_1, \ldots, M_p\}$

Phase 1: Extract parent candidates
for $i = 1$ to p **do**
  $M_i \leftarrow \varnothing$
end for
$t \leftarrow 2$
while $t \leq d$ **do**
  noChange $\leftarrow$ **True**
  for each $K \subseteq X$ with $|K| = t$ **do**
    Step A: Find the most endogenous variable in K
    for each $x_i \in K$ **do**
      $\text{score}(x_i) \leftarrow \text{p-HSIC}\Big(h_i\big(i,\ X \setminus (M_i \cup K \setminus \{x_i\})\big),\ \big\{h_j(j,\ X \setminus M_j)\big\}_{x_j \in K \setminus \{x_i\}}\Big)$
    end for
    $x_b \leftarrow \arg\max_{x_i \in K} \text{score}(x_i)$
    Step B: Verify x_b **is a sink (Lemma 7)**
    $e \leftarrow \text{p-HSIC}\Big(h_b\big(b,\ X \setminus (M_b \cup K \setminus \{x_b\})\big),\ \big\{h_j(j,\ X \setminus M_j)\big\}_{x_j \in K \setminus \{x_b\}}\Big)$
    $d_{\max} \leftarrow \max_{x_j \in K \setminus \{x_b\}} \text{p-HSIC}\Big(h_b(b,\ X \setminus M_b),\ h_j(j,\ X \setminus M_j)\Big)$
    if $e > \alpha$ **and** $d_{\max} < \alpha$ **then**
      $M_b \leftarrow M_b \cup (K \setminus \{x_b\})$ ▷ x_b is sink; others are parent candidates
      noChange $\leftarrow$ **False**
    end if
  end for
  if *noChange* **then**
    $t \leftarrow t + 1$
  else
    $t \leftarrow 2$ ▷ Restart: new parents may unlock new sinks
  end if
end while

Phase 2: Prune false parent candidates
for $i = 1$ to p **do**
  for each $x_j \in M_i$ **do**
    if $\text{p-HSIC}\Big(h_i\big(i,\ X \setminus (M_i \setminus \{x_j\})\big),\ h_j(j,\ X \setminus M_j)\Big) > \alpha$ **then**
      $M_i \leftarrow M_i \setminus \{x_j\}$ ▷ x_j not a direct parent
    end if
  end for
end for

Phase 3: Build adjacency matrix and detect invisible pairs
Initialize $A(i, j) \leftarrow 0$ for all i, j
for $i = 1$ to p **do**
  for each $x_j \in M_i$ **do**
    $A(i, j) \leftarrow 1$ ▷ x_j is a parent of x_i
  end for
end for
for each pair (i, j) with $i < j,\ x_j \notin M_i,\ x_i \notin M_j$ **do**
  if $\text{p-HSIC}\Big(h_i(i,\ X \setminus M_i),\ h_j(j,\ X \setminus M_j)\Big) < \alpha$ **then**
    $A(i, j) \leftarrow \text{NaN};\ A(j, i) \leftarrow \text{NaN}$ ▷ Invisible pair: UBP/UCP exists
  end if
end for

return $A, \{M_1, \ldots, M_p\}$

- **Lines 9–13 (Step A):** We start with all subsets K of size t contained in X. To find a *sink* for the subset K, we first find the most endogenous element in K. For each element x_i in K,

we compute a score: after regressing x_i on its parent candidates M_i and all other members of the subset K, is x_i's residual jointly independent of the residuals of all other members of K, each regressed on their own parent candidates? A high score (high p-value) means independence, and hence x_b is chosen as the element achieving the maximum score. When a sink is confirmed in Step B, the algorithm resets $t \leftarrow 2$ (line 25): the newly discovered parents in M_b may unlock sinks in other subsets that previously failed the test, so the search restarts from pairs.

- **Lines 14–20 (Step B):** We check if x_b is indeed a sink. Condition (18) (independence with K included) is checked in line 15, and condition (19) (dependence without K) is checked in line 16. In line 17, if $e > \alpha$ (independence) and $d_{\max} < \alpha$ (dependence) are both verified, x_b is confirmed as a sink and $K \setminus \{x_b\}$ is added to M_b.
- **Lines 28–35 (Phase 2):** The pruning test checks whether $h_i(i, X \setminus (M_i \setminus \{x_j\})) \perp\!\!\!\perp h_j(j, X \setminus M_j)$. The exclusion set $X \setminus (M_i \setminus \{x_j\})$ contains x_j, so x_j enters neither the location nor the scale function of h_i. If x_j is a true parent of x_i, (13) gives $\epsilon_j \not\perp\!\!\!\perp h_i$ and, since $\epsilon_j \in N_j$, (12) gives $\epsilon_j \not\perp\!\!\!\perp h_j$; residual faithfulness (Assumption 3) then gives $h_i \not\perp\!\!\!\perp h_j$ for all admissible h_i, h_j, so true parents are never pruned.

  Conversely, if x_j is not a direct parent, then $M_i \setminus \{x_j\}$ still contains all true parents K_i, so the structural residual is admissible for the exclusion set $X \setminus (M_i \setminus \{x_j\})$. This gives $h_i(i, X \setminus (M_i \setminus \{x_j\})) = \eta_i$. Similarly, $h_j(j, X \setminus M_j) = \eta_j$. Note that any $x_j \in M_i$ that is not a true parent must form a visible non-edge with x_i (invisible pair partners cannot enter M_i because the sink test, which requires joint independence, always fails when an invisible partner is in K). Since the pair is visible (no UBP/UCP), **Lemma 2** gives $\eta_i \perp\!\!\!\perp \eta_j$. Non-parents are correctly pruned.
- **Lines 36–47 (Phase 3):** For each non-adjacent pair (x_i, x_j), the test $h_i(i, X \setminus M_i) \not\perp\!\!\!\perp h_j(j, X \setminus M_j)$ detects invisible pairs. When $M_i = K_i$ (the true parents), the true structural functions give $h_i(i, X \setminus M_i) = \eta_i$, and **Lemma 2** gives $\eta_i \not\perp\!\!\!\perp \eta_j$ if and only if a UBP/UCP exists. This corresponds to **Proposition 1**, Case 1 (bidirected edge).

  However, when $M_i \subsetneq K_i$ (some true parents were not identified in Phase 1, e.g., because invisible pairs blocked sink identification), the residual $h_i(i, X \setminus M_i)$ does not equal η_i and may show dependence with h_j even when no UBP/UCP exists. This causes visible pairs to be incorrectly marked as invisible. This is a known limitation, discussed further below.

A.4.2 Limitations of LSNM-UV-Base

LSNM-UV-Base is *incomplete* in identifying visible pairs and *unsound* in identifying invisible pairs. Soundness and completeness can be defined as follows.

Definition A.5 (Soundness and completeness)**.**

- The algorithm is *sound for visible edges* if every visible edge in the output is also a visible edge in the ground truth.
- The algorithm is *complete for visible edges* if every visible edge in the ground truth also appears as a visible edge in the output.
- Analogous definitions apply to visible non-edges and invisible pairs.

The source of these limitations is that Phase 1 may fail to identify all parents of x_i. Specifically, the sink identification in Lemma 7 requires that (18) holds: including $K \setminus \{x_i\}$ in the regression achieves independence. If some $x_j \in K$ has a UBP/UCP with x_i (i.e., the pair is invisible), the shared hidden noise persists in the residuals regardless of the regression set, so (18) fails. The sink test does not pass, and the invisible parent is never added to M_i.

Remark 4. These limitations are the same as those identified for CAM-UV (Maeda and Shimizu, 2021) in the additive setting by Pham et al. (2026), and are not specific to the LSNM extension. They arise because Phase 1 cannot identify invisible parents, not because of any issue with the location-scale structure.

A.4.3 CheckVisible

Stage 2: checkVisible. For each pair (x_i, x_j) marked as invisible (NaN) by LSNM-UV-Base, this procedure tests whether it can be re-classified as a visible edge or visible non-edge. The key insight

is that LSNM-UV-Base searches over regression sets of the form M_i (parent candidates found in Phase 1), which may be incomplete. CHECKVISIBLE searches over a larger set Q that includes all parents identified by LSNM-UV-Base *plus* all variables whose relationship to x_i or x_j is unclear (marked NaN). This broader search can discover independence-achieving regression sets that Phase 1 missed.

For this procedure we use the following lemma, which certifies from a single independent residual pair that a pair is not invisible and rules out one direction of parentship.

Lemma 8 (One-sided independence)**.** *Suppose there exist admissible residuals of x_i and x_j that are independent when x_j may be used in residualizing x_i, while x_i is excluded from the residualization of x_j. Then*

$$B(i,j) = 0 \qquad \textit{and} \qquad A(j,i) = 0.$$

That is, the pair is not invisible and x_i is not a parent of x_j. The symmetric statement holds with i and j exchanged.

Proof. Independence first rules out $B(i,j) = 1$: otherwise a noise in $N_i \cap N_j$ would survive in both residuals by residual irreducibility, and residual faithfulness would force dependence.

Now suppose $x_i \to x_j$. Since x_i is excluded from the residualization of x_j, residual irreducibility implies that this residual retains ϵ_i. Every admissible residual of x_i also retains ϵ_i, because $\epsilon_i \in N_i$. Residual faithfulness would again force dependence, contradicting the assumed independence. Hence $A(j,i) = 0$. □

Applying Lemma 8 in both directions, a pair with witnesses of both kinds is not invisible and neither endpoint is a parent of the other, so it is a visible non-edge.

Algorithm 3 CHECKVISIBLE LSNM-UV

Require: Indices i and j

$P_i \leftarrow \{v \mid A(i,v) = 1\}$; $P_j \leftarrow \{v \mid A(j,v) = 1\}$
$Q \leftarrow \big(P_i \cup P_j \cup \{k \mid A(i,k) = \text{NaN or } A(j,k) = \text{NaN}\}\big) \setminus \{i,j\}$
iNotParent $\leftarrow$ **False**; *jNotParent* $\leftarrow$ **False**
for each $M \subseteq Q$ and $N \subseteq Q$ **do**
  $e \leftarrow$ p-HSIC$\Big(h_i\big(i,\ X \setminus M\big),\ h_j\big(j,\ X \setminus N\big)\Big)$
  ▷ Test visible non-edge: Prop 1, Case 3
  if $e > \alpha$ **then**
    $A(i,j) \leftarrow 0$; $A(j,i) \leftarrow 0$; **return**
  else
    $a_1 \leftarrow$ p-HSIC$\Big(h_i\big(i,\ X \setminus (M \cup \{x_j\})\big),\ h_j\big(j,\ X \setminus N\big)\Big)$
    $a_2 \leftarrow$ p-HSIC$\Big(h_i\big(i,\ X \setminus M\big),\ h_j\big(j,\ X \setminus (N \cup \{x_i\})\big)\Big)$
    if $a_1 > \alpha$ **then**
      iNotParent $\leftarrow$ **True** ▷ x_i is not a parent of x_j
    end if
    if $a_2 > \alpha$ **then**
      jNotParent $\leftarrow$ **True** ▷ x_j is not a parent of x_i
    end if
    if *iNotParent* $\wedge$ *jNotParent* **then**
      $A(i,j) \leftarrow 0$; $A(j,i) \leftarrow 0$; **return** ▷ Visible non-edge
    end if
  end if
end for
if *iNotParent* **then**
  $A(i,j) \leftarrow 1$; $A(j,i) \leftarrow 0$ ▷ $x_j \to x_i$ is a visible edge
end if
if *jNotParent* **then**
  $A(j,i) \leftarrow 1$; $A(i,j) \leftarrow 0$ ▷ $x_i \to x_j$ is a visible edge
end if

For a pair (x_i, x_j), the candidate pool $Q = Q_{ij}$ (line 2) consists of all parents identified by LSNM-UV-Base together with all variables whose relationship to x_i or x_j is invisible, excluding x_i and x_j themselves. For $M, N \subseteq Q$ the loop issues three queries:

- **Lines 5–8** (e): both endpoints are excluded from each other's residualization. An independent pair is a witness for a visible non-edge (Proposition 1, Equation (8); the argument applies to any exclusion sets containing both endpoints).
- **Line 10** (a_1): x_j is added to the regression set of x_i, while x_i stays excluded from that of x_j. An independent pair shows that the pair is not invisible and that x_i is not a parent of x_j (Lemma 8).
- **Line 11** (a_2): the symmetric query; an independent pair shows that the pair is not invisible and that x_j is not a parent of x_i.
- **Line 18**: once witnesses of both kinds have been found, neither endpoint is a parent of the other and the pair is a visible non-edge.
- **Lines 23–28**: if the loop ends with only *iNotParent* set, no e witness exists for any $M, N \subseteq Q$ and the pair is not invisible, so it is the visible edge $x_j \to x_i$; symmetrically for *jNotParent*. If neither flag is set, the pair stays invisible.

A.4.4 Soundness and completeness of LSNM-UV

The proof uses only exact answers of the residual-independence oracle $\mathcal{O}_{\mathrm{RI}}$ (Equation (9)); no test level enters the argument.

Proof of Theorem 4. Pairs already resolved by LSNM-UV-Base are correct by Assumption 4, so it remains to consider a pair (x_i, x_j) passed to CHECKVISIBLE. Let $Q = Q_{ij}$ be its candidate pool (line 2 of Algorithm 3). By search coverage,

$$(K_i \setminus \{x_j\}) \cup (K_j \setminus \{x_i\}) \subseteq Q.$$

For $M, N \subseteq Q$, write $e(M, N) = 1$ when the oracle reports that independence is achievable after residualizing x_i on M and x_j on N. Let $a_1(M, N) = 1$ denote the corresponding query after additionally allowing x_j in the residualization of x_i, and let $a_2(M, N) = 1$ denote the symmetric query after additionally allowing x_i in the residualization of x_j. That is, $e(M, N) = \mathcal{O}_{\mathrm{RI}}(i, j, X \setminus M, X \setminus N)$, $a_1(M, N) = \mathcal{O}_{\mathrm{RI}}(i, j, X \setminus (M \cup \{x_j\}), X \setminus N)$ and $a_2(M, N) = \mathcal{O}_{\mathrm{RI}}(i, j, X \setminus M, X \setminus (N \cup \{x_i\}))$.

We consider the three possible pair classes.

Visible non-edge. Suppose

$$A(i, j) = A(j, i) = B(i, j) = 0.$$

Take

$$M = K_i, \qquad N = K_j.$$

By search coverage, $M, N \subseteq Q$. Since neither endpoint is a parent of the other, these sets are admissible in the ordinary e query. Choosing the structural location and scale functions yields

$$h_i = \eta_i, \qquad h_j = \eta_j.$$

Because $B(i, j) = 0$, Lemma 2 gives

$$\eta_i \perp\!\!\!\perp \eta_j.$$

Hence $e(M, N) = 1$. Exhaustive CHECKVISIBLE therefore finds an independence witness and classifies the pair as a visible non-edge (unless it has already returned the same classification at line 18).

Conversely, if CHECKVISIBLE finds $e(M, N) = 1$ for any $M, N \subseteq Q$, then an admissible pair of residuals is independent while both endpoints are excluded from one another's regressions. By Proposition 1, the pair is a visible non-edge (the converse part of its proof for visible non-edges uses only that both exclusion sets contain both endpoints, so it applies to $\mathcal{H}_i(X \setminus M)$ and $\mathcal{H}_j(X \setminus N)$). Likewise, if the loop returns at line 18 because an a_1 and an a_2 witness were both found, Lemma 8 applied in both directions shows that the pair is not invisible and that neither endpoint is a parent of the other. Thus this classification is sound.

Visible edge. Suppose $x_j \to x_i$ is visible:

$$A(i,j) = 1, \qquad B(i,j) = 0.$$

For every $M, N \subseteq Q$, the ordinary query $e(M,N)$ excludes x_j from the residualization of x_i. By Proposition 1, such residuals are necessarily dependent (the argument for (6) applies to any exclusion set of x_i containing x_j). Hence

$$e(M,N) = 0 \qquad \text{for all } M, N \subseteq Q.$$

Now choose

$$M = K_i \setminus \{x_j\}, \qquad N = K_j.$$

Search coverage gives $M, N \subseteq Q$ (acyclicity gives $x_i \notin K_j$). In the a_1 query, x_j is added back to the regression of x_i, so choosing the structural functions gives

$$h_i = \eta_i, \qquad h_j = \eta_j.$$

Since $B(i,j) = 0$,

$$\eta_i \perp\!\!\!\perp \eta_j,$$

and therefore

$$a_1(M,N) = 1.$$

By contrast, every a_2 query still excludes the true parent x_j from the residualization of x_i. Residual irreducibility gives

$$\epsilon_j \not\perp\!\!\!\perp h_i.$$

Moreover $\epsilon_j \in N_j$, so every admissible residual of x_j retains ϵ_j. Residual faithfulness therefore implies dependence, and

$$a_2(M,N) = 0 \qquad \text{for all } M, N \subseteq Q.$$

Thus exhaustive CHECKVISIBLE identifies the direction $x_j \to x_i$ (lines 23–28).

The argument for $x_i \to x_j$ is symmetric.

Conversely, if $e(M,N) = 0$ for every searched pair of sets and an a_1 witness exists but no a_2 witness exists, Lemma 8 rules out both an invisible pair and $x_i \to x_j$, and $e \equiv 0$ rules out a visible non-edge by the first case, leaving $x_j \to x_i$ as the unique pair class. Hence the directed classification is sound.

Invisible pair. Suppose $B(i,j) = 1$. Let

$$\epsilon_k \in N_i \cap N_j.$$

For every residual appearing in an e, a_1, or a_2 query, by (12),

$$\epsilon_k \not\perp\!\!\!\perp h_i, \qquad \epsilon_k \not\perp\!\!\!\perp h_j.$$

Residual faithfulness therefore implies dependence for every such query. Consequently

$$e(M,N) = a_1(M,N) = a_2(M,N) = 0$$

for all $M, N \subseteq Q$, so CHECKVISIBLE leaves the pair invisible.

Conversely, the preceding two cases show that every visible non-edge and every visible directed edge has an oracle independence witness that exhaustive CHECKVISIBLE must find. Therefore a pair left invisible cannot belong to either visible class and must satisfy $B(i,j) = 1$.

Thus every pair is classified correctly. On the bow-free target class, an invisible pair has no simultaneous directed edge, so these pair classifications uniquely determine $T_{ADMG}(G)$. □

A.4.5 IDENTIFYING ANCESTORS ON INVISIBLE PATHS

Lemma 4 of Pham et al. (2026) provides conditions under which observed variables can be identified as lying on invisible paths (UBPs/UCPs) between a visible parent pair. That result is stated using the notion of visibility with respect to subsets $X' \subseteq X$, which relies on the additive model being closed under marginalization of observed variables (see Remark 3). We now provide an LSNM analogue that avoids subset visibility entirely, stating all conditions directly in terms of location-scale residuals $h(i,S)$ on the full data.

Lemma 9 (Identifying ancestors on invisible paths)**.** *Consider distinct $x_i, x_j, x_k \in X$. Suppose:*

1. *x_j is a visible parent of x_i (Definition A.1): $A(i,j) = 1$ and $B(i,j) = 0$.*
2. *The pair (x_i, x_k) is invisible (Definition A.3): $\forall\, h_1, h_2 \in \mathcal{H}$: $h_1(i, \{i\}) \not\perp\!\!\!\perp h_2(k, \{k\})$.*

Under Assumptions 1 and 3, suppose further that the following holds:

$$\forall\, h_1, h_2 \in \mathcal{H}: \quad h_1(i, \{i,k\}) \not\perp\!\!\!\perp h_2(j, \{j,k\}). \tag{20}$$

Then x_k is an ancestor of x_i.

Proof of Lemma 9. We prove the contrapositive: if $x_k \notin \mathrm{Anc}(x_i)$, then independence is achievable in (20).

Step 1: $x_k \notin K_i$ and $x_k \notin K_j$. Since $x_k \notin \mathrm{Anc}(x_i)$ and every parent is an ancestor, x_k is not a parent of x_i, so $x_k \notin K_i$.

Since $x_j \to x_i$ is a directed edge, every ancestor of x_j is also an ancestor of x_i (otherwise the graph would contain a cycle). Therefore $x_k \notin \mathrm{Anc}(x_i)$ implies $x_k \notin \mathrm{Anc}(x_j)$, and in particular $x_k \notin K_j$.

Hence $K_i \subseteq X \setminus \{x_i, x_k\}$ and $K_j \subseteq X \setminus \{x_j, x_k\}$.

Step 2: Constructing independent residuals without x_k. Choose h_1^* and h_2^* using the true structural functions $f = f_i^1,\ g = g_i^1$ and $f = f_j^1,\ g = g_j^1$ respectively.

Since $K_i \subseteq X \setminus \{x_i, x_k\}$, the regression for x_i on K_i is a valid choice with exclusion set $S_1 = \{i,k\}$ (both f_i^1 and g_i^1 use only variables in $K_i \subseteq X \setminus S_1$). This gives:

$$h_1^*(i, \{i,k\}) \;=\; \frac{x_i - f_i^1(K_i)}{g_i^1(K_i)} \;=\; \eta_i.$$

Since $K_j \subseteq X \setminus \{x_j, x_k\}$, the same argument with $S_2 = \{j,k\}$ gives:

$$h_2^*(j, \{j,k\}) \;=\; \frac{x_j - f_j^1(K_j)}{g_j^1(K_j)} \;=\; \eta_j.$$

Step 3: Independence via noise-set disjointness (Lemma 2). Since $B(i,j) = 0$ (Premise 1), there is no UBP/UCP between x_i and x_j. By the proof of Lemma 2, the noise sets

$$N_i = \{\epsilon_l : v_l \in \mathrm{an}(Q_i) \cup Q_i \cup \{x_i\}\}, \qquad N_j = \{\epsilon_l : v_l \in \mathrm{an}(Q_j) \cup Q_j \cup \{x_j\}\}$$

satisfy $N_i \cap N_j = \varnothing$: any shared ϵ_l would imply a UBP or UCP between x_i and x_j (first part of that proof), contradicting $B(i,j) = 0$. Since the exogenous noises ϵ_l are mutually independent, $\eta_i = \phi_i(N_i)$ and $\eta_j = \phi_j(N_j)$ are functions of non-overlapping independent noise terms, giving $\eta_i \perp\!\!\!\perp \eta_j$.

Therefore $h_1^*(i, \{i,k\}) \perp\!\!\!\perp h_2^*(j, \{j,k\})$, contradicting Condition (20). By contrapositive, $x_k \in \mathrm{Anc}(x_i)$. □

Lemma 10 (Identifying ancestors on invisible paths for sets)**.** *Consider distinct $x_i, x_j \in X$ and a set $\{x_{k_1}, \ldots, x_{k_m}\} \subset X \setminus \{x_i, x_j\}$. Suppose:*

1. *x_j is a visible parent of x_i (Definition A.1): $A(i,j) = 1$ and $B(i,j) = 0$.*
2. *For each $q = 1, \ldots, m$, the pair (x_i, x_{k_q}) is invisible (Definition A.3): $\forall\, h_1, h_2 \in \mathcal{H}$: $h_1(i, \{i\}) \not\perp\!\!\!\perp h_2(k_q, \{k_q\})$.*

Under Assumptions 1 and 3, suppose further that for each $q = 1, \ldots, m$ the following holds:

$$\forall\, h_1, h_2 \in \mathcal{H}: \quad h_1(i, \{i, k_q\}) \not\perp\!\!\!\perp h_2(j, \{j, k_q\}). \tag{21}$$

Then each x_{k_q} is an ancestor of x_i.

Proof. Fix any $q \in \{1, \ldots, m\}$. We prove the contrapositive: if $x_{k_q} \notin \mathrm{Anc}(x_i)$, then independence is achievable in (21). Proof same as before.

□

A.5 FINITE-SAMPLE IMPLEMENTATION

Location-scale residuals. For a response x_i and regression set R, we estimate both its conditional location and scale. We first fit a penalized additive cubic-spline regression $\hat{f}_i^{(0)}$ of x_i on R (Stone, 1985). From the residuals $r_i = x_i - \hat{f}_i^{(0)}(R)$, we fit the log-scale by regressing $\log(r_i^2 + 10^{-8})$ on R and set

$$\hat{g}_i^{(0)} = \exp\Big\{ \tfrac{1}{2} \widehat{E[\log(r_i^2 + 10^{-8}) \mid R]} \Big\}.$$

Indeed, if $r_i = g_i(R)\epsilon_i$, then $E[\log r_i^2 \mid R] = 2\log g_i(R) + E\log \epsilon_i^2$, so this identifies the scale up to a multiplicative constant, which does not affect independence.

Location error is amplified where the conditional scale is small:

$$\hat{h}_i - h_i^* = \frac{f_i^* - \hat{f}_i}{\hat{g}_i} + h_i^* \frac{g_i^* - \hat{g}_i}{\hat{g}_i}.$$

We therefore refit the location using inverse-scale weights $w_s = \max\{\hat{g}_i^{(0)}(R_s), c_w\}^{-2}$, recompute the scale, and form

$$\hat{h}_i = \frac{x_i - \hat{f}_i(R)}{\hat{g}_i(R)}.$$

The final residual is standardized to zero mean and unit variance. Appendix B.1 reports the accuracy of this estimator.

Residual-independence decisions. Each query is evaluated using HSIC with Gaussian kernels, median-heuristic bandwidths and the gamma approximation to its null distribution (Gretton et al., 2008). The practical search uses the fixed operating level $\alpha = 0.01$; this is distinct from the vanishing-threshold sequence required by Theorem 5. Phase 1 uses the residual tests to rank and verify sink candidates, while CHECKVISIBLE searches small regression subsets of its candidate pool rather than the exhaustive subset search used in the oracle theorem. Two safeguards complete the rule: CHECKVISIBLE anchors its regression sets, every candidate set having the form $P_i \cup s$ with P_i the Phase 1 parents of x_i and $|s| \leq 2$, so the witness search may add candidates but never discards found parents; and every acceptance inside CHECKVISIBLE requires the accepting independence p-value to exceed a witness bar of 0.10, p-values in $(0.01, 0.10]$ counting as abstention rather than as evidence of independence. Appendices B.3, B.4 and B.11 report the behaviour of these decisions.

After Stage 2, final pair classifications are re-examined on held-out data. The sample is split in two halves; residual models are fitted on one half and tested on the other, and the roles are then reversed. A directed edge is reclassified as invisible when both held-out tests reject independence at $\alpha_{\text{audit}} = 0.001$, and an invisible pair is reclassified as a non-edge when both accept. The audit is applied once, after the search; Appendix B.8 reports its effect on the benchmark.

Frozen configuration. All experiments use one configuration: test level $\alpha = 0.01$; audit level $\alpha_{\text{audit}} = 0.001$; witness bar 0.10; anchored CHECKVISIBLE sets $P_i \cup s$ with $|s| \leq 2$; at most $d = 3$ explanatory variables in the Phase 1 sink tests; weight floor $c_w = 0.05$; smoothing parameters chosen by generalised cross-validation, with the weighted location and scale fits iterated three times.

A.6 A SUFFICIENT CONDITION FOR TARGETING

Proposition 2 (Targeting). *Let $R \subseteq X \setminus \{x_i\}$ satisfy $K_i \subseteq R$ and $\mathbf{x}_R \perp\!\!\!\perp \eta_i$. Let the population location target be the conditional mean and let the scale functional $\sigma(\cdot \mid \mathbf{x}_R)$ be positively homogeneous. If*

$$c_1 = E\eta_i, \qquad c_2 = \sigma(\eta_i - c_1) > 0,$$

then

$$f^* = f_i^1 + c_1 g_i^1, \qquad g^* = c_2 g_i^1, \qquad h^* = \frac{\eta_i - c_1}{c_2}.$$

Consequently, when both endpoints of a query satisfy these conditions, their target residuals are independent if and only if their effective noises are independent.

Proof. Since

$$x_i = f_i^1(K_i) + g_i^1(K_i)\eta_i$$

and $\eta_i \perp\!\!\!\perp \mathbf{x}_R$,

$$E[x_i \mid \mathbf{x}_R] = f_i^1(K_i) + g_i^1(K_i)E\eta_i = f_i^1 + c_1 g_i^1.$$

Hence the location residual is

$$x_i - f^* = g_i^1(K_i)(\eta_i - c_1).$$

Positive homogeneity of the scale functional gives

$$g^* = \sigma(x_i - f^* \mid \mathbf{x}_R) = g_i^1(K_i)\sigma(\eta_i - c_1) = c_2 g_i^1(K_i),$$

and therefore

$$h^* = \frac{x_i - f^*}{g^*} = \frac{\eta_i - c_1}{c_2}.$$

Since this is an invertible affine transformation of η_i, it preserves independence. □

A.7 Proof of Theorem 5

Proof. Fix $q \in Q^*(P_X)$. If $\mathcal{O}_{\mathrm{RI}}(P_X, q) = 1$, then Condition C2 gives independence of the target residuals and hence $D_q = 0$. By Conditions C3–C4,

$$\widehat{D}_q = O_P(a_n) = o_P(\tau_n),$$

so $P\{\widehat{O}(q) \neq 1\} \to 0$.

If $\mathcal{O}_{\mathrm{RI}}(P_X, q) = 0$, Condition C1 places the target residuals in the admissible classes. By definition of $\mathcal{O}_{\mathrm{RI}}$ they are therefore dependent, so $D_q > 0$. Condition C3 gives $\widehat{D}_q \xrightarrow{P} D_q$, while Condition C4 gives $\tau_n \to 0$. Hence $P\{\widehat{O}(q) \neq 0\} \to 0$.

Finally,

$$P(E_n^c) \leq \sum_{q \in Q^*(P_X)} P\{\widehat{O}(q) \neq \mathcal{O}_{\mathrm{RI}}(P_X, q)\} \longrightarrow 0,$$

because $Q^*(P_X)$ is finite. Thus $P(E_n) \to 1$. □

B Experimental details and additional results

All controlled experiments in this section share one fixed function family: with $s(x) = (1 + e^{-x})^{-1}$,

$$f_1(x) = 0.8x + 0.5\sin x, \qquad g_1(x) = 0.05 + 0.45\, s(x), \qquad f_2(x) = 0.6x + 0.4\cos(1.5x), \qquad g_2(x) = 0.10 + 0.40\, s(-0.7x)$$

all exogenous noises are independent $N(0,1)$, residuals are those of the implementation (Appendix A.5), and every test is the HSIC test at the operating level $\alpha = 0.01$ unless stated otherwise. The ten-node benchmark experiments use the random generator of Appendix B.8. The sequence follows the theory: Simulation 1 asks whether the residual estimator improves with n; Simulations 3–4 ask whether the dependence statistic approaches its population target, diagnosing Condition C3; Simulation 5 illustrates the threshold Condition C4; Simulation 2 decomposes decision error into test, estimation and search; Simulation 6 asks whether individual decisions combine into the population signatures of Proposition 1; Simulation 7 asks whether correct query decisions combine into the correct graph; and Simulation 8 with the benchmark asks how the whole procedure performs on nontrivial graphs.

B.1 Residual estimation (Simulation 1)

Purpose. Condition C3 is stated directly for the final empirical dependence statistic, but one source of its error is residual estimation. We therefore first examine the concrete location–scale estimator in isolation.

Design. We generate

$$X = \epsilon_X, \qquad Y = f_1(X) + g_1(X)\epsilon_Y,$$

with independent standard-normal noises and supply the true parent set $\{X\}$ to the estimator. Because f_1, g_1 and ϵ_Y are known to the simulator, location, scale-shape and residual errors can be evaluated directly on held-out data. We use $n = 200, \ldots, 6400$ and $R = 1000$ replicates per sample size.

Results. Location, constant-aligned scale and standardized-residual errors all decrease steadily with n, with approximately common empirical log–log slope -0.46. Error is larger where $g_1(X) < 0.10$ and decays at a similar rate, at a level 3.3 to 3.9 times higher; this is the amplification that the inverse-scale weighting of Appendix A.5 targets. On true visible pairs of the ten-node benchmark (Appendix B.8), the weighted estimator raises the non-rejection rate of the fitted-null residual test at $\alpha = 0.01$ from 0.81 (unweighted) to 0.91 (fit size 12,800, test size 1600). The raw scale differs by a global multiplicative constant, which is irrelevant after residual standardization.

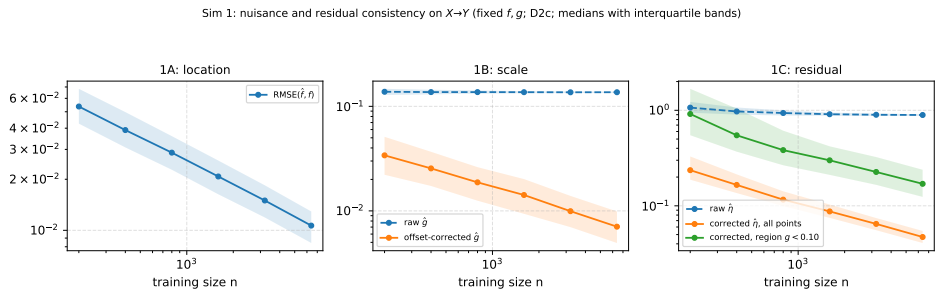


Figure 5: Simulation 1: recovery of the location, scale shape and standardized residual on the single edge $X \to Y$.

B.2 Statistic consistency (Simulations 3–4)

Purpose. Condition C3 requires $\widehat{D}_q - D_q \to 0$. We examine this condition directly for the implemented residual estimator and diagnose the two finite-sample sources of discrepancy.

Design. Simulations 3–4 reuse the replicates of Simulation 2 (Appendix B.4): the null fork $X \leftarrow Z \to Y$ with $X = f_1(Z) + g_1(Z)\epsilon_X$ and $Y = f_2(Z) + g_2(Z)\epsilon_Y$, and the Gaussian-coupled hidden common cause of Appendix B.5 ($A \to X$, $B \to Y$, $X = f_1(A) + g_1(A)\eta_X$, $Y = f_2(B) + g_2(B)\eta_Y$, $\eta_X = (\lambda U + \epsilon_X)/\sqrt{1+\lambda^2}$ and likewise η_Y, U hidden, $\lambda \in \{0.1, 0.25, 0.5, 1\}$), with residuals fitted on the true parent sets and $n = 200, \ldots, 6400$. The population reference D_{ref} is computed from 10^6 draws of the true effective noises by averaging HSIC over independent blocks of size $m \in \{1000, 2000, 4000\}$ and extrapolating linearly in $1/m$.

Results. For each replicate we compute the dependence statistic using both the fitted residuals and the true effective residuals, and compare both with a high-accuracy population reference D_{ref}. We report

$$E_{\mathrm{fit}} = |\widehat{D}_n^{\mathrm{fit}} - \widehat{D}_n^{\mathrm{true}}|, \qquad E_{\mathrm{samp}} = |\widehat{D}_n^{\mathrm{true}} - D_{\mathrm{ref}}|,$$

and

$$E_{\mathrm{total}} = |\widehat{D}_n^{\mathrm{fit}} - D_{\mathrm{ref}}|.$$

Across both the null fork and a dependent hidden-common-cause design, E_{total} decreases with sample size. The residual-fitting component is smaller than the finite-sample HSIC component throughout the investigated range and decreases faster. These results empirically diagnose Condition C3; they are not used as a proof of it.

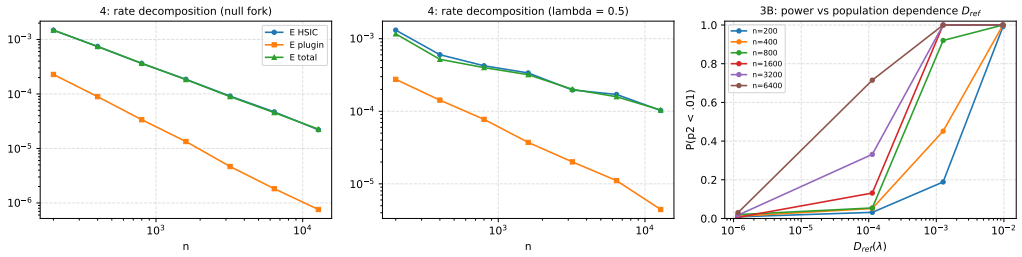


Figure 6: Simulations 3–4: residual-fitting, finite-sample and total error of the empirical dependence statistic, together with power against the population dependence D_{ref}.

B.3 Fixed and vanishing thresholds (Simulation 5)

Purpose. Condition C4 requires a threshold that tends to zero while remaining larger than the estimation error in Condition C3. The implementation instead uses the practical fixed level $\alpha = .01$. We compare the two regimes on identical replicates.

Design. The null fork of Simulation 2 and, as alternatives, its identifiable nonlinear hidden common cause and the Gaussian-coupled hidden common cause with $\lambda \in \{0.25, 0.5\}$ (Appendices B.4 and B.5), all at Level 2 (fitted residuals with the true parent sets), for $n = 200, \ldots, 12{,}800$. For each query we record $S_n = n\,\mathrm{HSIC}_n$ and the gamma null quantile $t_{\alpha,n}$, and compare the fixed rule $S_n > t_{0.01,n}$ with the statistic-scale threshold $S_n > t_{0.01,n} \max\{1, n/n_0\}^{\kappa}$, $n_0 = 400$, $\kappa = 0.75$. The floor at n_0 keeps the schedule from being more liberal than the fixed rule for $n < n_0$. The figure also shows the fixed level 0.05 and the p-value schedule $\alpha_n = 0.01 \min\{1, n_0/n\}$.

Results. Under the null, the fixed-level rule's false-dependence rate remains of order the nominal level (0.012 to 0.023 for $n \le 6400$), whereas the statistic-scale threshold coincides with it for $n \le n_0$ and makes no false-dependence decision from $n = 800$ on. Against the weak alternatives the threshold is more conservative at moderate n: for $\lambda = 0.5$ it misses 0.48, 0.35 and 0.23 of the dependent pairs at $n = 800$, 1600 and 3200, against at most 0.09 for the fixed level, and catches up only at $n = 6400$ to 12,800; for $\lambda = 0.25$ it misses almost all of them up to $n = 12{,}800$. The nonlinear hidden common cause is detected by both rules at every n. On the ten-node benchmark the same schedule matches or beats the fixed level from $n = 1600$ on (Appendix B.8). Thus the experiment illustrates the asymptotic-versus-finite-sample tradeoff encoded by Condition C4.

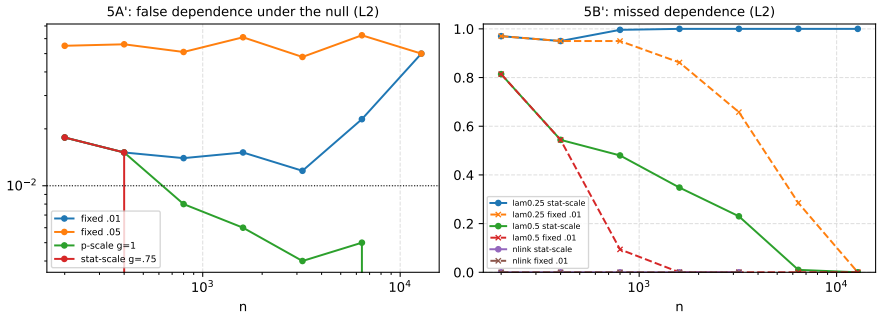


Figure 7: Simulation 5: fixed-level and vanishing-threshold decisions at Level 2. Left: false-dependence rate under the null fork against n. Right: missed-dependence rate under the Gaussian couplings $\lambda \in \{0.25, 0.5\}$ and the nonlinear hidden link against n.

B.4 Three levels of information (Simulation 2)

Graphs. The null fork $X \leftarrow Z \rightarrow Y$ with $X = f_1(Z) + g_1(Z)\epsilon_X$, $Y = f_2(Z) + g_2(Z)\epsilon_Y$, whose effective noises are independent by construction; and the identifiable nonlinear hidden common cause

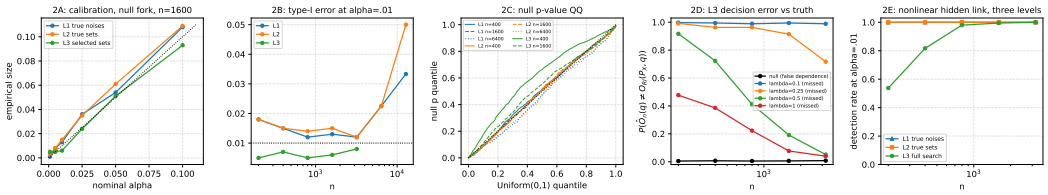


Figure 8: Simulation 2. Calibration at three information levels, type-I error against n, null p-value QQ, decision error of the complete pipeline against the generating truth, and the identifiable nonlinear-link alternative at all three levels.

with observed parents $A \to X$, $B \to Y$, $X = f_1(A) + g_1(A)\eta_X$, $Y = f_2(B) + g_2(B)\eta_Y$, and coupled noises $\eta_X = (f_1(U) + g_1(U)\epsilon_X)/1.147$, $\eta_Y = (f_2(U) + g_2(U)\epsilon_Y)/0.720$ (U hidden; the constants are population standard deviations). $R = 1000$ per size up to $n = 3200$ and 400 at 6400 for the null; $R = 350$ for the alternative.

Levels. L1 applies the test to the true effective noises; L2 to fitted residuals with the true parent sets supplied; L3 runs the complete search and tests the residuals of the selected sets. Hence L1→L2 isolates residual estimation and L2→L3 adds regression-set selection; this is the decomposition behind panels (a,b) of Figure 2.

Result (Figure 8). On the null, L1 and L2 are close to the nominal level at every level and size (empirical size 0.012 to 0.023 at $\alpha = 0.01$) and their null p-values are uniform on the QQ scale; L3 is conservative (0.005 to 0.008), never liberal, and selects both true parent sets in 66% of replicates. The decision error of the complete pipeline against the generating truth combines the type-I error under the null with one minus the L3 power under the alternatives. On the nonlinear hidden common cause, L1 and L2 detect the dependence at every n while L3 rises from 0.54 at $n = 200$ to 0.98 at 800 and 1.00 at 3200, with the correct parent sets selected in 0.45 to 0.93 of replicates as n grows: the gap between L2 and L3 is the cost of search, and it closes.

B.5 Signal strength and required sample size

Graph. $A \to X$, $B \to Y$, $X \leftarrow U \to Y$ with $X = f_1(A) + g_1(A)\eta_X$, $Y = f_2(B) + g_2(B)\eta_Y$ and Gaussian-coupled effective noises $\eta_X = (\lambda U + \epsilon_X)/\sqrt{1+\lambda^2}$ and likewise η_Y, $\lambda \in \{0.1, 0.25, 0.5, 1\}$; $R = 350$ per cell up to $n = 3200$ and 200 at 6400.

Purpose and caveat. This is a power and detectability diagnostic, not evidence for population identifiability: the linear Gaussian coupling admits a rival directed location-scale representation, so the L3 curves report the behaviour of the implemented search on this family rather than an identifiability claim. The L1 and L2 curves are the clean object, the power of the residual test as a function of signal strength.

Result (Figure 9). Larger dependence signal means smaller required sample size: power reaches one at a λ-dependent n, L1 and L2 are nearly identical throughout (estimation is not the bottleneck), and $\lambda = 0.1$, whose population dependence is indistinguishable from zero at our reference precision, correctly stays at the test level for every n.

B.6 Residual-independence signatures on motifs (Simulation 6)

Motifs and checks. The four elementary motifs, each with the complete set of residual checks the proposition prescribes: visible non-edge $Z \to X$, $Z \to Y$ with $X = f_1(Z) + g_1(Z)\epsilon_X$, $Y = f_2(Z) + g_2(Z)\epsilon_Y$ (raw dependence; independence of the residuals given Z); visible edge $X \to Y$ with $Y = f_1(X) + g_1(X)\epsilon_Y$ (independence of the residual given the parent; dependence without it); hidden common cause $X \leftarrow U \to Y$ and hidden intermediate $X \to U \to Y$ (raw dependence and dependence of the residual in both directions). The plotted quantity is the probability that the entire signature is correct at $\alpha = 0.01$; $R = 1000$ per cell up to $n = 3200$ and 300 at 6400.

Three hidden regimes. The hidden mechanisms are instantiated in three regimes: a linear Gaussian boundary ($X = (\frac{1}{2}U + \epsilon_X)/\sqrt{1.25}$ and likewise Y; for the intermediate, $U = 0.7X + \epsilon_U$ and

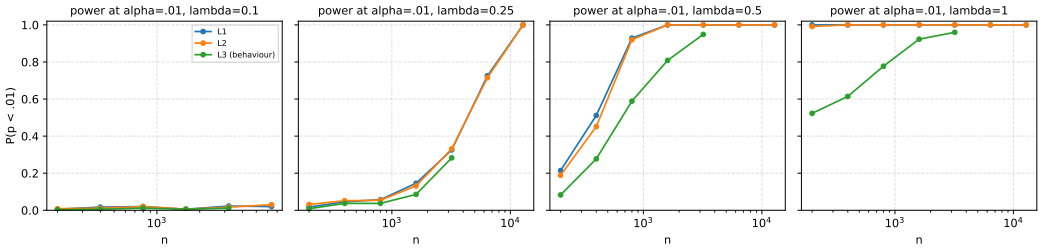


Figure 9: Signal-strength diagnostic: power at $\alpha = 0.01$ of the three information levels against n, one panel per coupling strength λ. The L3 curves are behaviour of the implemented search on this family (see text).

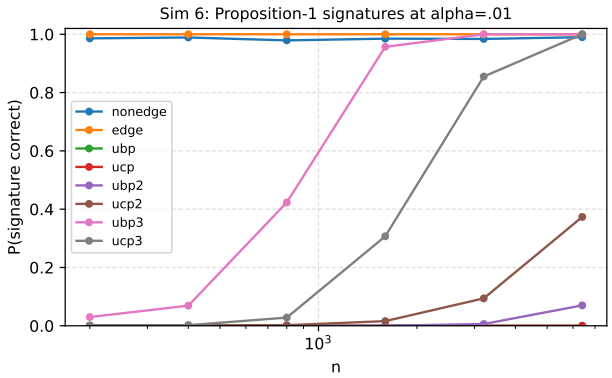


Figure 10: Simulation 6. Probability that the complete Proposition-1 signature is correct at $\alpha = 0.01$, for the visible motifs and the three hidden regimes.

$Y = f_1(U) + g_1(U)\epsilon_Y$), for which the conditional law of each variable given the other is itself a location-scale family, so residual faithfulness fails and the pair is non-identifiable; an ordinary nonlinear regime built from the plan family ($X = f_1(U) + g_1(U)\epsilon_X$, $Y = f_2(U) + g_2(U)\epsilon_Y$; relay $U = f_2(X) + g_2(X)\epsilon_U$, $Y = f_1(U) + g_1(U)\epsilon_Y$); and a deliberately large-margin regime with softplus locations and steep tanh scales ($f(u) = \log(1 + e^{2u})$ and $1.2\log(1 + e^{-2u})$; $g(u) = 1.05 + 0.95\tanh(2u)$ and $1.15 + 1.05\tanh(-2u)$), whose standardized conditional shapes vary strongly.

Result (Figure 10). Visible motifs sit at the ceiling imposed by the fixed-level test (0.98 to 1.00 throughout). The Gaussian boundary stays at zero at every n, with the p-values of its failing checks drifting toward the uniform distribution: there is no population signal, an empirical illustration of the residual faithfulness assumption. The ordinary nonlinear signatures emerge late (0.07 and 0.37 at $n = 6400$, governed by the weakest of the required checks), and the large-margin signatures approach one much earlier (common cause 0.42, 0.96 and 1.00 at $n = 800$, 1600 and 3200; intermediate 1.00 by 6400). Detection is governed by the margin of the weakest residual-dependence signal.

B.7 Oracle search versus empirical search (Simulation 7)

Graph and data. $X_1 \to X_3 \to X_4$, $X_2 \leftarrow U \to X_3$ with U hidden: $X_2 = f_2(U) + g_2(U)\epsilon_2$, $\eta_3 = (f_1(U) + g_1(U)\epsilon_3)/1.148$, $X_3 = f_1(X_1) + g_1(X_1)\eta_3$, $X_4 = f_2(X_3) + g_2(X_3)\epsilon_4$. The identical search runs twice on each sample: once with every residual-independence answer taken from the true graph, once from the data. $R = 300$ per size up to $n = 1600$, 150 at 3200 and 60 at 6400; oracle $R = 50$ per size.

Result (Figure 11). The oracle-driven search recovers the exact graph in every replicate at every n, consistent with the oracle-search conclusion of Theorem 4 on this instance. The empirical run issues about twenty queries; its mean agreement with the oracle rises from 0.72 at $n = 200$ through 0.93 at

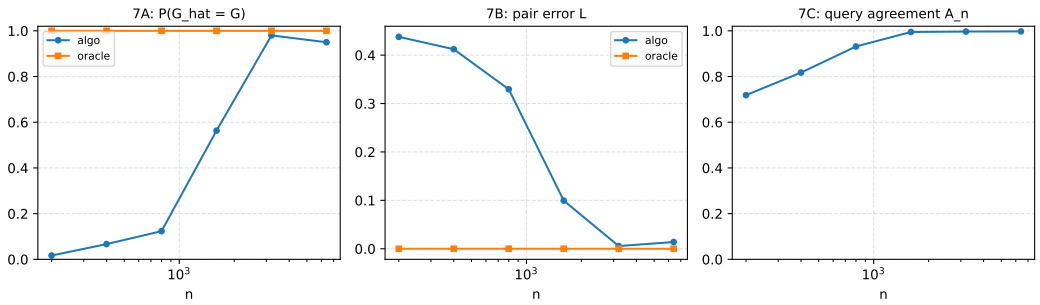


Figure 11: Simulation 7. Exact recovery, pair-classification error and mean query agreement of the empirical search, against the oracle-driven search on the same samples.

800 to 0.997 from 1600 on, its pair-classification error falls from 0.44 to 0.006, and exact recovery rises from 0.02 to 0.98 at $n = 3200$. Accurate query answers combine into an accurate graph, the chain behind Corollary 6, observed on one graph.

B.8 Ten-node benchmark and oracle diagnostics (Simulation 8)

The ten-node benchmark of Figure 3, Simulation 8 and Appendices B.9–B.11 use the random generator below. All randomness of a trial is controlled by a single integer seed; Table 2 lists the hyperparameters.

Graph. *Observed skeleton.* Direct edges among the p observed variables $\{x_0, \ldots, x_{p-1}\}$ are drawn from an Erdős–Rényi DAG with edge probability 0.3, following Maeda and Shimizu (2021) Section 5.1 exactly. For each ordered pair (j, i) with $j < i$, the edge $x_j \to x_i$ is included independently with probability 0.3. Using index order as the topological order guarantees acyclicity.

Hidden common causes (UBPs). For each of the n_{cc} hidden common causes u_k:

1. Select a pair (x_a, x_b) of observed variables that have no direct edge between them in either direction.
2. Add directed edges $u_k \to x_a$ and $u_k \to x_b$. The node u_k is a root (no parents of its own).

When u_k is marginalised out, the projected ADMG acquires the bidirected edge $x_a \leftrightarrow x_b$, representing an Unobserved Backdoor Path (UBP). Because (x_a, x_b) was chosen to have no direct edge, the bow-free condition is satisfied for this pair.

Hidden intermediates (UCPs). For each of the n_{int} hidden intermediates y_k:

1. Select an existing direct observed edge $x_j \to x_i$.
2. *Remove* the direct edge $x_j \to x_i$.
3. Add $x_j \to y_k$ and $y_k \to x_i$.

When y_k is marginalised out, the projected ADMG acquires the bidirected edge $x_j \leftrightarrow x_i$, representing an Unobserved Causal Path (UCP). Removing the direct edge keeps this pair bow-free.

Functions. All variables are generated in topological order of the full latent DAG, so that every parent of v_i has already been assigned values when v_i is drawn.

Two-level generation. For each variable x_i, let K_i denote its observed direct parents and Q_i its hidden direct parents. We implement Equation (1) in two layers:

$$\eta_i = f_i^2(Q_i) + g_i^2(Q_i)\,\epsilon_i, \tag{22}$$

$$x_i = f_i^1(K_i) + g_i^1(K_i)\,\eta_i, \tag{23}$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Layer (22) aggregates the influence of all hidden parents into a single noise term η_i. Layer (23) depends only on K_i, so hidden parents enter x_i exclusively through η_i. Variables with no parents are pure noise: $x_i = \epsilon_i$.

Nonlinear functional form. Each structural component is assigned a nonlinear function family from Table 1. Our experiments use: f^1 = RBF, g^1 = tanh, f^2 = softplus, g^2 = logarithmic. For each parent v_j, the assigned family provides a randomly parameterised building block $\varphi(v_j)$ whose parameters are drawn independently per parent, per variable, and separately for location and scale.

Table 1: Nonlinear function families used as building blocks $\varphi(v_j)$. All parameters are drawn independently per parent and per variable.

Family	Form $\varphi(v_j)$	Parameters
RBF	$a\exp\big(-b(v_j-c)^2\big)+d$	$a{\sim}\mathrm{U}(1,3){\cdot}\{\pm1\}$, $b{\sim}\mathrm{U}(0.1,1)$, $c{\sim}\mathrm{U}(-2,2)$, $d{\sim}\mathrm{U}(-1,1)$
Tanh	$a\tanh(bv_j+c)+d$	$a{\sim}\mathrm{U}(1,3){\cdot}\{\pm1\}$, $b{\sim}\mathrm{U}(0.5,2)$, $c{\sim}\mathrm{U}(-2,2)$, $d{\sim}\mathrm{U}(-1,1)$
Softplus	$a\log\big(1+e^{bv_j+c}\big)+d$	$a{\sim}\mathrm{U}(0.5,2){\cdot}\{\pm1\}$, $b{\sim}\mathrm{U}(0.5,2)$, $c{\sim}\mathrm{U}(-2,2)$, $d{\sim}\mathrm{U}(-1,1)$
Logarithmic	$a\,\mathrm{sign}(z)\log(\lvert z\rvert{+}1)+d$	$z{=}bv_j{+}c$; $a{\sim}\mathrm{U}(1,3){\cdot}\{\pm1\}$, $b{\sim}\mathrm{U}(0.5,2)$, $c{\sim}\mathrm{U}(-2,2)$, $d{\sim}\mathrm{U}(-1,1)$

The location and scale functions are additive over parents:

$$f(K_i) = \sum_{x_j \in K_i} \varphi_f(x_j), \tag{24}$$

$$g(K_i) = \exp\left(\frac{\sum_{x_j \in K_i} \varphi_g(x_j)}{\mathrm{sd}\left(\sum_{x_j} \varphi_g\right)} \right), \tag{25}$$

and analogously for f_i^2, g_i^2 with Q_i. The log-scale sum in (25) is normalised by its empirical standard deviation before exponentiation to prevent numerical overflow; the result is clipped to $[0.1,\ 10]$. Each variable is standardised to zero mean and unit empirical variance; η_i is likewise standardised before being used in Layer 1.

The key structural difference from Maeda and Shimizu (2021) is the *heteroscedastic* scale $g_i^1(K_i)$: in CAM-UV the noise is additive and homoscedastic ($g_i^1 \equiv 1$), whereas here the noise variance of x_i varies with the values of its observed parents.

Ground truth. After all data are generated the hidden variable columns are discarded. The ground-truth adjacency matrix A records the surviving direct edges among observed variables, $A(i,j) = 1$ iff $x_j \to x_i$. The bidirected matrix B is computed over *all* observed pairs from the noise sets of Lemma 2: $B(i,j) = 1$ iff $N_i \cap N_j \neq \varnothing$, which captures not only the planted UBP and UCP endpoints but every pair connected through hidden ancestors. Because such indirect hidden connections can also create bows, bow-freeness is not automatic: candidate graphs whose projection contains a pair with $A(i,j)B(i,j) = 1$ are rejected and redrawn before data generation, so all accepted graphs lie in the bow-free target class. A uniformly random permutation π is applied to the columns of the data matrix X before it is returned. The ground-truth matrices A and B are permuted consistently via $A' = A[\pi,\pi]$, $B' = B[\pi,\pi]$. This ensures the algorithm cannot exploit the coincidence that variable indices equal the topological order of the generating DAG.

Methods. All methods receive the same $n \times p$ data matrix X (with permuted columns) and significance level $\alpha = 0.01$.

- **LSNM-UV** uses the frozen implementation of Appendix A.5 with the parameter values given there. Benchmark-specific settings: a limit of 3600 s per run, which no run of any method reached. The output is a $p \times p$ matrix with $\mathrm{mat}[i,j] = 1$ for $x_j \to x_i$, $\mathrm{mat}[i,j] = \mathrm{NaN}$ for an invisible (bidirected) pair and $\mathrm{mat}[i,j] = 0$ for no edge. We extract $\hat{A}$ by setting NaN entries to 0, and $\hat{B}(i,j) = 1$ whenever $\mathrm{mat}[i,j]$ or $\mathrm{mat}[j,i]$ is NaN.

- **CAM-UV** (Maeda and Shimizu, 2021). `lingam.CAMUV` (v1.12.2) with default HSIC independence test and `num_explanatory_vals` $= 3$. Output parsing identical to LSNM-UV.
- **FCI** (Spirtes et al., 2000; Zhang, 2008). From the `causal-learn` package (v0.1.4.5) with Fisher-z test at $\alpha = 0.01$. We extract only *definite* edges (no circle marks): directed $x_j \to x_i$ when $G[j,i] = -1$ and $G[i,j] = 1$; bidirected $x_i \leftrightarrow x_j$ when $G[i,j] = G[j,i] = 1$.
- **BANG** (Wang and Drton, 2023). Source adapted from `https://github.com/ysamwang/ngBap` with minor fixes (a syntax repair and `lm.fit` in place of `RcppArmadillo::fastLm`), called via `rpy2` (v3.5.11, R v4.3.1). Run as `bang(X, K=3, level=0.01, restrict=1, testType="dhsic")` with the `dHSIC` R package (v2.2). $\hat{A}(i,j) = 1$ wherever `dEdge`$[i,j] \neq 0$; $\hat{B}(i,j) = 1$ from off-diagonal nonzero entries of `bEdge`.

Metrics and error bars. For directed edges, a true positive requires both position and direction to be correct: $\hat{A}(i,j) = A(i,j) = 1$. NaN entries are treated as 0 for evaluation. Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, Recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, F1 $= 2 \cdot \mathrm{Prec} \cdot \mathrm{Rec}/(\mathrm{Prec}+\mathrm{Rec})$. For bidirected edges, the same definitions apply using $\hat{B}$ and B, evaluated on upper-triangle pairs only. The 400 trials are split into 16 independent batches of 25 trials each. Reported values are the mean across batch-level means; error bars show ± 1 standard deviation across the batch means.

Table 2: Hyperparameters of the data-generation procedure.

Parameter	Value	Description
p	10	Number of observed variables
n_{cc}	2	Hidden common causes (one UBP each)
n_{int}	2	Hidden intermediates (one UCP each)
ER edge probability	0.3	Matches Maeda and Shimizu (2021) Sec. 5.1
Function families	RBF / tanh / softplus / logarithmic	f^1 / g^1 / f^2 / g^2
n	200–1000	Sample sizes of Figure 3 (steps of 200)
Trials per setting	400	Independent random seeds

Simulation 8 runs. Three runs on identical graphs and samples, $n \in \{200, 400, 800, 1600, 3200\}$: the implemented algorithm at the fixed level ($R = 400$ per size up to 800, 300 at 1600, 100 at 3200); the statistic-scale schedules $t_{\alpha,n}(n/n_0)^{\kappa}$ with $n_0 = 400$ and $\kappa \in \{0.5, 0.75\}$ ($R = 80$; these runs use the schedule without the floor at n_0, which differs from $\max\{1, n/n_0\}^{\kappa}$ only at $n = 200$); and the search driven by exact oracle answers ($R = 400$ at every size).

Result (Figure 12). This experiment explains why the algorithm falls short of the oracle rather than comparing methods. The exact-oracle search has pair error 0.004 to 0.006 and exact recovery 0.97 to 0.98 at every n: with correct answers the search recovers the graph, and the residual 2 to 3% are graphs whose in-degree exceeds the search depth. These oracle failures are outside the search-width condition of Theorem 4 rather than failures of the residual characterization. The implemented run's pair error falls from 0.34 to 0.24 over the grid, so the oracle gap is statistical and shrinks with n; directed and bidirected recovery improve accordingly (bidirected F1 0.06 to 0.48). The scheduled thresholds track the fixed level and overtake it from $n = 1600$ on (directed F1 0.78 against 0.75 at $n = 1600$ and 0.81 against 0.75 at $n = 3200$ for $\kappa = 0.75$), consistent with Simulation 5.

Cross-fitted audit. With the true parent sets on 200 generated graphs at $n = 1000$, testing fitted residuals on held-out rows detected bidirected pairs as often as testing on the fitting rows at equal test size (0.608 against 0.603 at $\alpha = 0.01$), but rejected independence on directed pairs, whose residual query is a population null, far more often (0.257 against 0.036); requiring both folds to reject lowered this to 0.107. The audit therefore uses the stricter level $\alpha_{\mathrm{audit}} = 0.001$ and requires both folds to agree. On the benchmark at $n = 1000$ (400 graphs) it converted 449 of 5109 checked directed edges into bidirected pairs: 139 true bidirected pairs, 176 true directed edges (4.1% of the 4243 true directed edges), 106 reversed edges and 28 non-edges. It converted 26 of 599 checked bidirected pairs into non-edges.

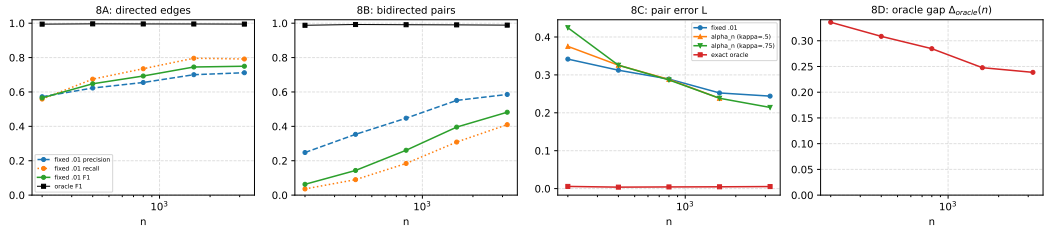


Figure 12: Simulation 8. Directed and bidirected precision, recall and F1 with the exact-oracle reference, pair-classification error of the fixed level, the scheduled thresholds and the oracle, and the oracle gap.

Table 3: Ten-node benchmark at $n = 1000$: pooled precision, recall and F1 over $R = 400$ graphs; all methods at $\alpha = 0.01$ on identical data.

	Directed			Bidirected		
Method	P	R	F1	P	R	F1
LSNM-UV	0.682	0.749	**0.714**	0.513	0.224	**0.312**
CAM-UV	0.248	0.098	0.140	0.164	0.480	0.244
FCI	0.458	0.019	0.037	0.143	0.067	0.091
BANG	0.089	0.009	0.016	0.162	0.596	0.255

B.9 Benchmark values at $n = 1000$

Table 3 gives the pooled precision, recall and F1 of Figure 3 at $n = 1000$, over $R = 400$ identical graphs and samples for all methods.

B.10 Robustness to the exogenous-noise distribution

We repeat the ten-node benchmark with uniform exogenous noise. Figure 13 shows the same precision, recall and F1 decomposition as the Gaussian benchmark. Bidirected recovery improves substantially under the non-Gaussian noise, consistent with stronger residual-shape information. Across the three noise families the bidirected F1 of LSNM-UV at $n = 1000$ orders as Gaussian 0.312 $<$ exponential 0.371 $<$ uniform 0.434.

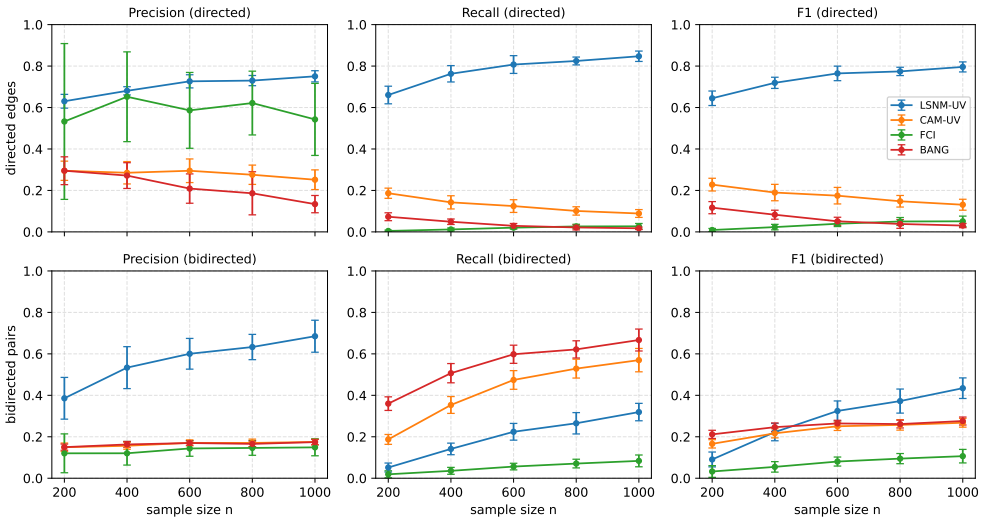


Figure 13: **Ten-node benchmark with uniform exogenous noise.** Directed (top) and bidirected (bottom) precision, recall and F1 for LSNM-UV, CAM-UV, FCI and BANG.

B.11 Calibration on benchmark queries

Simulation 2's null fork is an easy problem; this experiment asks whether the three-level distinction survives in the multi-parent, search-selected situations of the ten-node benchmark. For the final configuration we record the residual queries on true directed pairs, whose query is a population null (top row of Figure 14), and on true bidirected pairs (bottom row), and evaluate them at Level 1 (true noises), Level 2 (fitted residuals with the true parent sets) and Level 3 (the parent sets selected by the algorithm), up to $n = 3200$. L1–L2 assess test calibration; L3 reports end-to-end rejection behavior after graph-dependent set selection. A wrongly selected set can make the residual query itself no longer population-null, so the Level-3 rate on directed pairs is a selected-set rejection rate rather than a type-I error.

At Level 1 the fixed level stays close to nominal (0.014 to 0.016). At Level 2 it becomes increasingly liberal on these multi-parent queries (0.04 at $n = 200$ to 0.11 at $n = 3200$), unlike the single-parent null fork of Appendix B.4: the estimation error of Condition C3 is visible here. The statistic-scale threshold with $\kappa = 1$ brings the Level-2 rate below the nominal level from $n = 1000$ on, at a cost in power (0.50 against 0.82 at $n = 3200$). The Level-3 selected-set rejection rate of the fixed level lies between 0.13 and 0.26.

B.12 NetSim fMRI benchmark

Data. NetSim (Smith et al., 2011) provides 28 simulated fMRI configurations. Each propagates neural activity along a network of 5, 10, 15 or 50 brain regions, passes it through a nonlinear (balloon) haemodynamic model and adds thermal measurement noise. Each configuration contains 50 subjects that share one network topology (connection strengths and haemodynamic parameters vary slightly across subjects), with 50 to 5000 scans per subject (200 in the standard configurations). Configuration 13 adds negative backward connections to half of the forward connections, creating two-cycles, and configuration 14 reverses one connection to create a cycle; both lie outside the acyclic model class, and we use the remaining 26. Because NetSim observations arise from simulated time series, this is an out-of-assumption robustness benchmark rather than a verification of the i.i.d. conditions of Theorem 5.

Hidden-variable instances. For each configuration we binarise the subject-averaged connection matrix (an edge wherever the average weight is non-zero), standardise every region within subject and pool the 50 subjects. Each of 100 draws keeps a random subset of regions (4 of 5, 7 of 10, 10 of 15, 10 of 50) and hides the rest. The ground truth on the kept regions is the projected ADMG:

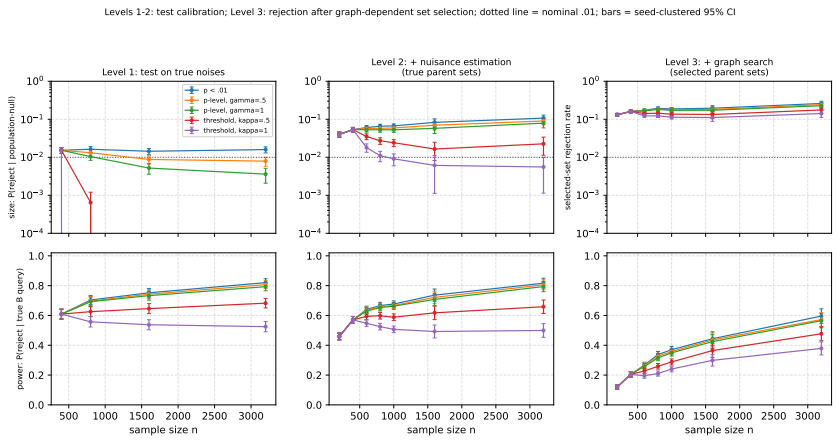


Figure 14: Rejection rates on ten-node benchmark queries at three information levels, up to $n = 3200$. Top: population-null queries on true directed pairs (Levels 1–2: test size; Level 3: selected-set rejection rate). Bottom: power on true bidirected pairs. The fixed level is compared with p-value and statistic-scale schedules floored at $n_0 = 400$; bars are seed-clustered 95% intervals.

$A(i, j) = 1$ for a direct connection $x_j \to x_i$ between kept regions, and $B(i, j) = 1$ iff the noise sets of Lemma 2, with every hidden region treated as an unobserved variable, intersect. Draws whose projection contains a bow are redrawn; every configuration admits bow-free subsets. From the pooled series 1000 rows are sampled without replacement, and the four methods run on the identical $1000 \times p$ matrix at $\alpha = 0.01$ with a limit of 3600 seconds per run (10,400 runs, none of which failed or reached the limit).

Directed recovery across configurations. Figure 15(a) shows directed F1 per configuration. LSNM-UV has the best directed F1 in 8 of the 26 configurations (15, 19, 20, 22, 23, 24, 27 and 28). Its largest margins occur where the acquisition strains the additive approximation: with stronger connections (configuration 15, mean strength 0.9 instead of 0.4) it attains 0.58 against 0.12, 0.17 and 0.04 for CAM-UV, FCI and BANG, and with nonstationary connection strengths (configuration 22) 0.67 against 0.22, 0.03 and 0.14; Figure 15(b) places these configurations far above the diagonal. Pooled over all configurations CAM-UV leads on directed F1 (0.62 against 0.54 for LSNM-UV, 0.26 for BANG and 0.15 for FCI), as expected on signals close to linear and additive; where it leads, LSNM-UV trails in both precision and recall (Figure 16(a)). On configuration 19 (fast sampling, TR 0.25 s, with 100 ms neural lag; Figure 4) the paired per-dataset difference in directed F1 between LSNM-UV and CAM-UV is 0.071, with 95% interval $[0.034, 0.107]$.

Bidirected pairs. Pooled bidirected precision, recall and F1 are 0.10, 0.01 and 0.02 for LSNM-UV, 0.10, 0.03 and 0.04 for CAM-UV, 0.08, 0.04 and 0.06 for BANG, and 0.50, 0.10 and 0.17 for FCI (Figure 15(c)). The loss channels of LSNM-UV (Figure 16(b)) locate its floor: of the 5207 true bidirected pairs, 38% are absorbed as a directed edge before Phase 3 and 60% reach the Phase 3 residual test and are not rejected; about 2% are detected. The pairs are tested and do not reject, consistent with residual-dependence margins below the detection floor of the test (Remark 1); FCI's partial correlations detect the part of the hidden dependence that its linear model covers.

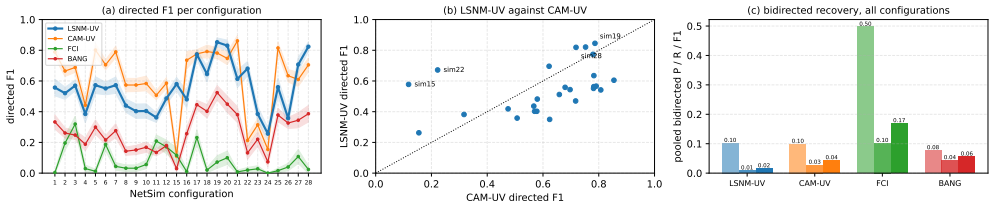


Figure 15: NetSim benchmark, 26 acyclic configurations, 100 hidden-region datasets each. (a) Directed F1 per configuration with 95% intervals over datasets. (b) LSNM-UV against CAM-UV directed F1, one point per configuration. (c) Pooled bidirected precision, recall and F1.

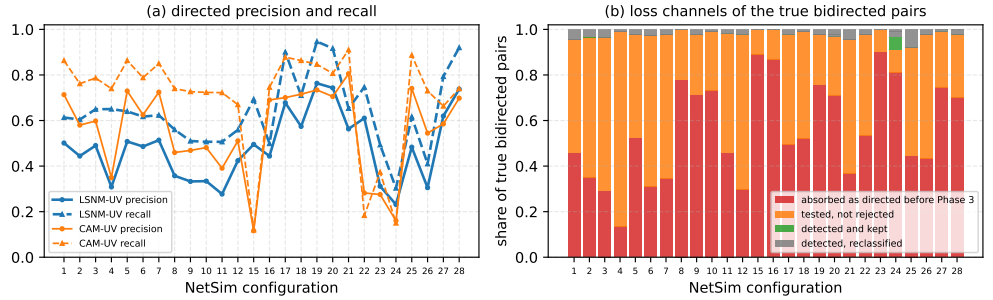


Figure 16: NetSim diagnostics. (a) Directed precision and recall per configuration for LSNM-UV and CAM-UV. (b) Loss channels of the true bidirected pairs of LSNM-UV per configuration: absorbed as a directed edge before Phase 3, tested and not rejected, detected and kept, and detected but reclassified in Stage 2.

C Broader Impacts

This work is theoretical and methodological, with indirect societal impact. It may improve causal discovery from observational data in settings with latent variables and non-additive, heteroscedastic noise, which are common in areas such as biomedicine, economics, education, and environmental science. By relaxing additive-noise assumptions, the proposed approach can support more reliable scientific hypothesis generation and guide follow-up experiments.

The main risk is misuse or overinterpretation. The guarantees rely on assumptions such as acyclicity, faithfulness-type conditions, location-scale structure, and, for full ADMG recovery, bow-freeness. In finite samples, errors in regression or independence testing may lead to incorrect causal graphs. Such errors could be harmful in high-stakes domains if inferred causal relations are treated as definitive evidence for interventions or decisions. Missed latent confounding is a particular concern.

Accordingly, this method should be used as an exploratory tool, not as a standalone basis for high-stakes decisions. Practical use should include uncertainty quantification, sensitivity analysis, domain expertise, and, where possible, validation with interventional or prospective data. The experiments use synthetic data and do not involve human subjects or private information; the computational footprint is modest.