CAUSAL REVERSAL AND IDENTIFICATION UNDER COEFFICIENT MODULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A coefficient that changes with the volatility state can reverse the direction preferred by causal discovery methods built for a fixed mechanism. We certify this reversal in a matched two-variable model: changing only the coefficient's state dependence makes both residual mutual information and the deployed DirectLiNGAM score prefer the wrong arrow. On a ten-variable benchmark, DirectLiNGAM, ICA-LiNGAM, VARLiNGAM and a constant-coefficient heteroskedastic-SEM score all fall below chance as positive modulation grows. Conditioning on the state does not settle the direction by itself: without a restriction on how the coefficient moves, a reverse model reproduces every state-conditional covariance. We then show how the same variation can identify a graph. When coefficients and relative noise variances are log-affine in an observed state, sinks have the smallest signed curvature of the log precision diagonal. This criterion recovers the DAG from three state-conditional covariances, provided coefficient growth does not exactly offset relative noise growth. We give a recursive estimator, HetCD-Ω, with an exact-state recovery guarantee and a matching fourth-order lower bound near this balance condition. On the same benchmark, a fixed-window implementation reaches 0.90 order accuracy under strong modulation of either sign, compared with 0.32 and 0.54 for DirectLiNGAM. These experiments also expose the limits of the repair: fixed windows lie outside the guarantee, one-signed modulation creates a variance-sorting shortcut, and an outcome-derived state proxy substantially reduces accuracy. A daily financial panel provides a consistency check only. The results specify when coefficient variation reverses a pooled criterion, when it identifies structure, and what state information that identification requires.

1 INTRODUCTION

Changes in noise variance can reveal causal structure that a stationary covariance cannot identify. Identification through heteroskedasticity and backShift use this variation while holding the structural coefficients fixed (Rigobon, 2003; Rothenhäusler et al., 2015). Non-Gaussian linear methods also rely on fixed coefficients, using independence between a cause and its structural error to orient an edge (Shimizu et al., 2006; Hyvärinen & Smith, 2013). What happens when the coefficient changes with the same state that changes the noise?

Such coupling is plausible when a common condition affects both transmission and uncertainty: a financial exposure may strengthen during volatile periods (Longin & Solnik, 2001; Ang & Chen, 2002), or a sensor's gain and noise may both depend on its operating state. We study a coefficient $b(v) = b_0 e^{\kappa v}$ driven by an exogenous scalar state v. In a matched example, switching on this dependence while preserving the noise process reverses both true residual mutual information and the population score used by DirectLiNGAM (Shimizu et al., 2011). This is a reversal of the preferred direction under misspecification, not merely a loss of identification.

Our main constructive result uses the variation that causes this failure. Suppose coefficients and relative noise variances are log-affine in the observed state. A sink's log precision diagonal is then affine after removing a term common to all nodes. A non-sink adds a strictly convex term whenever at least one outgoing edge is *unbalanced*: its coefficient growth does not exactly offset the child's relative noise growth. Comparing signed curvatures removes the unknown common term. Three

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.

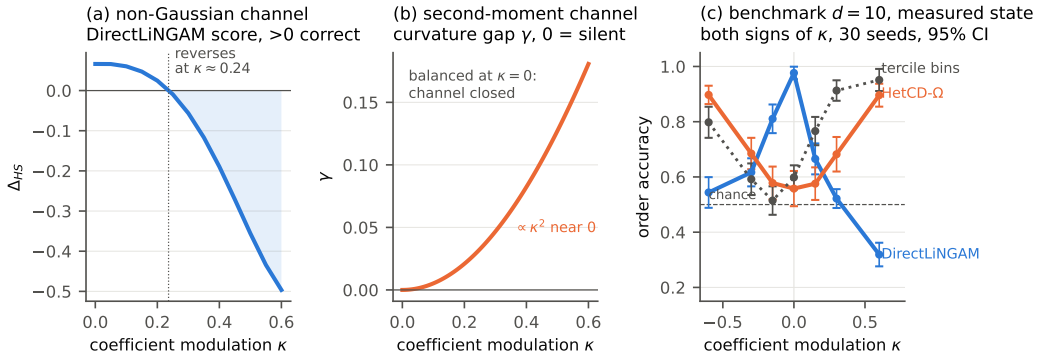

Figure 1: Coefficient modulation can reverse a pooled score while creating a conditional-covariance signal. (a) The DirectLiNGAM population score crosses zero near $\kappa = .24$ on the bivariate benchmark calibration (numerical quadrature; positive favours the true arrow). (b) With equal noise slopes, the sink–non-sink curvature gap is zero at $\kappa = 0$ and grows quadratically nearby. (c) On the ten-variable benchmark, fixed-window HetCD-Ω reaches .90 order accuracy at both $\kappa = \pm.6$, with a measured state; DirectLiNGAM reaches .32 and .54. Tercile bins behave differently under sign reversal. Error bars are 95% confidence intervals over 30 seeds.

conditional covariance matrices therefore identify a sink, its parents and coefficients; removing that sink and repeating recovers the graph. Our curvature criterion extends precision-based ordering (Ghoshal & Honorio, 2018; Zhou et al., 2023), exposing a balance obstruction and its statistical cost.

The resulting method, HetCD-Ω, has a recovery guarantee at exact state levels and a consistent shrinking-window version for an i.i.d. continuous state. Near balance, its sample requirement has the same fourth-order dependence on imbalance as our lower bound. The experiments use a separate fixed-window implementation. It improves as modulation grows, including under sign-reversed and mixed-sign controls, but requires a measured state and does not uniformly outperform the pooled baselines. These distinctions matter: averaging within state bins changes the covariance signal, and one-signed modulation creates a shortcut based on variance slopes.

Section 2 develops the matched reversal, Section 3 the graph criterion, Section 4 the estimator, and Sections 5–6 recovery, sorting shortcuts, state information and a financial consistency check.

2 Coefficient Modulation and Reversal

We begin with two variables, which isolates the effect of changing a coefficient.

Assumption 1 (modulated structural model)**.** A scalar state v_t follows a stationary unit-variance process. Mean-zero variables satisfy

$$x_t = s_x(v_t, u_{x,t})z_{x,t}, \qquad y_t = b(v_t)x_t + s_y(v_t, u_{y,t})z_{y,t}, \tag{1}$$

where $b(v) = b_0 e^{\kappa v}$, $s_x^2 = e^{Av+Cu_x}$ and $s_y^2 = e^{Av+Cu_y}$. The innovations z are i.i.d. and independent of the states; the idiosyncratic states u_x, u_y are independent of each other and of v. On a DAG, the extension is $X_t = B_t X_t + \varepsilon_t$ with $B_t = B \odot e^{Kv_t}$.

With Gaussian innovations and Gaussian idiosyncratic states, conditioning on v leaves independent, non-Gaussian scale-mixture errors when $C > 0$ (Proposition 10). Pooling over v generally makes those errors dependent, even at $\kappa = 0$. Thus the comparison below asks whether changing the coefficient reverses a pooled criterion under the *same* noise misspecification; it does not claim that the pooled constant-coefficient model meets all LiNGAM assumptions.

For standardized pooled variables $\tilde{X}, \tilde{Y}$ with correlation ρ, define residuals $R_Y = (\tilde{Y} - \rho\tilde{X})/\sqrt{1-\rho^2}$ and $R_X = (\tilde{X} - \rho\tilde{Y})/\sqrt{1-\rho^2}$. The entropy comparison is

$$\begin{aligned} D &= H(\tilde{Y}) + H(R_X) - H(\tilde{X}) - H(R_Y) \\ &= I(\tilde{Y}; R_X) - I(\tilde{X}; R_Y). \end{aligned} \tag{2}$$

Here H is differential entropy. A positive D favours $X \to Y$: the forward regression leaves the less dependent residual (Hyvärinen & Smith, 2013).

Theorem 1 (certified coefficient-only reversal)**.** *Let V, U_x, U_y be independent equiprobable signs and Z_x, Z_y independent standard Gaussians. Set $X = 2^{(V+U_x)/2} Z_x$ and $Y = 2^{mV} X + 2^{(V+U_y)/2} Z_y$. Then*

$$0.0355 < D\big|_{m=0} < 0.0575, \qquad -0.1037 < D\big|_{m=1} < -0.0812.$$

The arrow is $X \to Y$ in both models; only the coefficient's state dependence changes. The negative sign holds uniformly for $b(-1) \in [0.4999, 0.5001]$, $b(+1) \in [1.9999, 2.0001]$, and persists under sufficiently small smoothing of the discrete states.

Each variable in equation 2 is a finite Gaussian mixture. Its entropy is a one-dimensional integral enclosed by 128-bit ball arithmetic (Johansson, 2016); the construction and code are in Appendix B. Independent quadrature gives 0.0465 and -0.0925 (Appendix D).

Corollary 2 (the deployed ordering reverses)**.** *For the pairwise score of `lingam` 1.13.0 DirectLiNGAM, the population statistic is certified in $(0.069, 0.088)$ at $m = 0$ and $(-0.113, -0.096)$ at $m = 1$. Its sign determines the two-variable ordering, so under i.i.d. sampling the ordering eventually reverses almost surely.*

The packaged estimator placed X first in 20/20 constant-coefficient samples and 0/20 moving-coefficient samples at each $n \in \{500, 2000, 10000\}$. Modulation increases the child's scale-mixture kurtosis (Proposition 12), but a kurtosis crossing alone does not establish reversal; the certificate concerns the decision statistic itself. On the Gaussian-state benchmark calibration, numerical quadrature places the score crossing near $\kappa = .24$ (Figure 1a). This numerical threshold is distinct from the certified example above.

3 Identification from Conditional Covariances

The reversal suggests conditioning on the state. Conditional covariances alone, however, need a restriction on how mechanisms can change. We first show that restriction in two variables, then derive a graph criterion.

3.1 The Balance Condition in Two Variables

Write $q_x(v) = \mathrm{Var}(x \mid v)$ and $q_y(v) = \mathrm{Var}(s_y z_y \mid v)$. Allow different noise slopes, with ratio $r(v) = q_y(v)/q_x(v) = r_0 e^{\delta v}$. Then

$$\Sigma(v) = \begin{pmatrix} q_x & bq_x \\ bq_x & b^2 q_x + q_y \end{pmatrix}, \qquad c^\star(v) = \frac{\Sigma_{12}(v)}{\Sigma_{22}(v)} = \frac{b(v)}{b(v)^2 + r(v)}. \tag{3}$$

The reverse coefficient $c^\star(v)$, together with suitable noise variances, reproduces every positive definite $\Sigma(v)$ (Proposition 11). Unrestricted covariance matching therefore cannot distinguish the directions. Log-affinity changes this conclusion.

Theorem 3 (two-variable balance boundary)**.** *Suppose equation 3 holds on a state interval with $b(v) = b_0 e^{\kappa v}$, $b_0 \neq 0$, $r_0 > 0$. Let $\mathcal{F}$ be the signed log-affine functions and $p(v) = b(v)^2/(b(v)^2 + r(v))$. Then*

$$\frac{d^2}{dv^2} \log |c^\star(v)| = -(2\kappa - \delta)^2 p(v)(1 - p(v)). \tag{4}$$

Hence $c^\star \in \mathcal{F}$ if and only if $\delta = 2\kappa$. Off this boundary, no reverse model with a coefficient in $\mathcal{F}$ matches the covariance family.

The true coefficient has affine log magnitude; the reverse-implied coefficient has a curved log magnitude. They both satisfy the restriction only when relative noise growth exactly offsets coefficient growth. In particular, a constant coefficient and equal noise slopes have $\kappa = \delta = 0$ and lie on the boundary. The graph result below uses the same obstruction, expressed through precision diagonals.

3.2 A CURVATURE CRITERION FOR SINKS

Assumption 2 (observed-state graph model)**.** Let G be a DAG on $[d]$ and V an exogenous scalar state whose support contains $v_1 < v_2 < v_3$. Conditionally on $V = v$,

$$\begin{aligned} X &= B(v)X + \varepsilon, & B_{ij}(v) &= \beta_{ij} e^{\kappa_{ij} v}, \\ \mathbb{E}[\varepsilon \mid v] &= 0, & Q(v) &:= \mathrm{Cov}(\varepsilon \mid v) = g(v)\,\mathrm{diag}(d_i e^{a_i v}), \end{aligned} \tag{5}$$

where $\beta_{ij} \neq 0$ exactly when $j \to i$, $d_i > 0$, and $g(v) > 0$ is arbitrary. The errors are uncorrelated given the state; Gaussianity and independence are not required.

The factor g permits a shared, possibly nonlinear variance profile. Only *relative* noise variances must be log-affine. For an edge $j \to i$, define its imbalance as

$$\eta_{ij} = 2\kappa_{ij} - (a_i - a_j).$$

This is $2\kappa - \delta$ in Theorem 3.

Assumption 3 (separation)**.** Every edge has $\eta_{ij} \neq 0$.

A weaker condition suffices: in every ancestral subset, each non-sink has an unbalanced outgoing edge within that subset (Appendix A). An ancestral set contains the parents of all its members. Such sets arise when sinks are removed recursively.

For an ancestral set R, write $\Omega_R(v) = \Sigma_{RR}(v)^{-1}$ for the precision of its *marginal* covariance. The SEM factorization gives the following identity.

Lemma 4 (log-precision diagonals)**.** *Under Assumption 2, for every ancestral R and $j \in R$,*

$$\log \Omega_{R,jj}(v) + \log g(v) = -\log d_j - a_j v + \log\left(1 + \sum_{i \in \mathrm{ch}_R(j)} \frac{\beta_{ij}^2 d_j}{d_i} e^{\eta_{ij} v}\right). \tag{6}$$

The right-hand side is affine for a sink, and strictly convex for any node with an unbalanced outgoing edge in R.

For a sink, the sum is empty. For a non-sink, it is a log-sum-exp containing a constant and at least one nonzero exponential slope, which makes it strictly convex. The unknown $-\log g(v)$ affects all nodes equally. We can therefore compare curvature without estimating g.

With $t = (v_2 - v_1)/(v_3 - v_1)$, measure curvature by the signed Jensen gap

$$J_{R,j} = (1-t)\log \Omega_{R,jj}(v_1) + t \log \Omega_{R,jj}(v_3) - \log \Omega_{R,jj}(v_2). \tag{7}$$

An affine function has zero gap; a strictly convex function has positive gap. Every sink attains the same value J_g, the gap of $-\log g$, while every non-sink lies above it. Keeping the *sign* of the gap is essential: the common value need not be zero or positive.

Theorem 5 (graph identification from three covariances)**.** *Under Assumptions 2–3, $\{\Sigma(v_k)\}_{k=1}^3$ identifies G and $B(v_k)$. For every ancestral R:*

(i) the minimizers of $J_{R,j}$ are exactly the sinks of $G[R]$;
(ii) for a sink s, $\mathrm{pa}(s) = \{j \in R \setminus \{s\} : \Omega_{R,sj}(v_2) \neq 0\}$ and $B_{sj}(v_k) = -\Omega_{R,sj}(v_k)/\Omega_{R,ss}(v_k)$.

Any alternative (G', B', Q') satisfying Assumption 2, even without separation, and generating the same three covariances has $G' = G$, $B'(v_k) = B(v_k)$ and $Q'(v_k) = Q(v_k)$. Log-affinity then determines every nonzero coefficient function. The noise covariance is identified at the three anchors; arbitrary g precludes identifying it elsewhere from those anchors alone.

Why the graph is unique. The criterion finds a sink and its incident coefficients directly from the covariances. Removing a sink preserves the model on the remaining ancestral set. To exclude an alternative graph, the proof compares the common minimum gaps, establishes a shared sink, and inducts; the alternative may have its own common factor and balanced edges. This shared-sink argument follows Zhou et al. (2023, Theorem 1 and supplement A.1). The complete proof is in Appendix A. When all edges are balanced, every Cholesky order of the base covariance gives a model with the same conditional laws, so separation cannot simply be discarded (Proposition 8). This is a graph-level obstruction, not an impossibility claim for each individual balanced edge.

4 Estimation and Statistical Limits

Algorithm 1 replaces the population covariances by estimates. It removes sinks in reverse topological order and reads their parents from normalized precision rows. The direct implementation takes $O(d^4)$ operations. Rescaling individual variables changes each log precision by a constant, so it leaves both the ordering gaps and normalized pruning statistics unchanged.

Algorithm 1 HetCD-Ω: signed-curvature ordering and precision-row pruning

Input: Covariance estimates $\widehat{\Sigma}_k$ at $v_1 < v_2 < v_3$; threshold $\tau \geq 0$

1. $R \leftarrow [d]$, $A \leftarrow 0$, $\pi \leftarrow ()$, $t \leftarrow (v_2 - v_1)/(v_3 - v_1)$
2. **While** $R \neq \emptyset$:
   a. $\widehat{\Omega}_k \leftarrow \big((\widehat{\Sigma}_k)_{RR}\big)^{-1}$ for $k = 1, 2, 3$
   b. $\widehat{J}_j \leftarrow (1-t)\log \widehat{\Omega}_{1,jj} + t \log \widehat{\Omega}_{3,jj} - \log \widehat{\Omega}_{2,jj}$ for $j \in R$
   c. $s \leftarrow \arg\min_{j \in R} \widehat{J}_j$
   d. $A_{sj} \leftarrow \mathbf{1}\{|\widehat{\Omega}_{2,sj}|/\sqrt{\widehat{\Omega}_{2,ss}\widehat{\Omega}_{2,jj}} > \tau\}$ for $j \in R \setminus \{s\}$
   e. Prepend s to π; $R \leftarrow R \setminus \{s\}$
3. **Return** order π and adjacency A, where $A_{sj} = 1$ means $j \to s$

4.1 Recovery at Exact State Levels

The guarantee depends on two separations: the smallest excess curvature $\gamma = \min_{R,j \text{ non-sink}}(J_{R,j} - J_g)$, and the smallest nonzero normalized sink-row entry

$$\zeta = \min_{R,s \text{ sink},\, j \in \mathrm{pa}(s)} \frac{|\Omega_{R,sj}(v_2)|}{\sqrt{\Omega_{R,ss}(v_2)\Omega_{R,jj}(v_2)}}.$$

Both minima range over ancestral sets; an empty minimum is $+\infty$. They control ordering and edge detection, respectively.

Theorem 6 (finite-sample recovery)**.** *Under Assumptions 2–3, suppose for some $\rho \in (0,1)$,*

$$(1-\rho)\Sigma(v_k) \preceq \widehat{\Sigma}_k \preceq (1+\rho)\Sigma(v_k), \qquad k = 1, 2, 3. \tag{8}$$

If $\gamma > -4\log(1-\rho)$ and $\zeta > 4\rho/(1-\rho)$, Algorithm 1 with $\tau = 2\rho/(1-\rho)$ recovers G exactly. For independent Gaussian observations with known zero conditional mean at the three exact levels, recovery has probability at least $1-\alpha$ for a suitable threshold once the sample size per level satisfies

$$m \geq C\big(d + \log(1/\alpha)\big)\max(1, \gamma^{-2}, \zeta^{-2}),$$

with a universal constant C (explicit choices in Appendix A).

The relative covariance event controls every principal submatrix, including those selected adaptively. For fixed d, independent observations with finite fourth moments also give consistency with $\tau_m = m^{-1/4}$; this is an asymptotic threshold, not a confidence level.

4.2 THE COST OF APPROACHING BALANCE

In the two-variable model, the curvature gap is of order η^2 near $\eta = 2\kappa - \delta = 0$. With the other parameters fixed, the upper bound therefore requires order η^{-4} observations. The next result shows that this dependence cannot be improved uniformly within the model class.

Theorem 7 (same-class lower bound)**.** *For $d = 2$, any $\kappa, a, b, r, d_0 > 0$ and $\eta \neq 0$, there are forward ($X \to Y$) and reverse ($Y \to X$) models satisfying Assumptions 2–3, both with imbalance η and conditionally Gaussian errors, such that, as $\eta \to 0$,*

$$\mathrm{KL}(P_F(\cdot \mid v) \,\|\, P_R(\cdot \mid v)) = \frac{p^2(1-p)}{8}\eta^4 v^4 + O(|\eta|^5), \qquad p = \frac{b^2}{b^2 + r}.$$

For n independent observations with bounded state, if $n\eta_n^4 \to 0$, the minimax error for deciding the arrow tends to $1/2$. The result extends to any fixed d by adding isolated variables.

The match concerns the imbalance exponent at fixed dimension, not all constants or dimension dependence.

4.3 CONTINUOUS STATES: PROVEN AND EMPIRICAL VERSIONS

For an exactly observed continuous state, shrinking neighbourhoods around three quantiles inherit consistency under i.i.d. sampling and local regularity. Appendix A gives a construction with $m_n = n^{.8}$ observations per neighbourhood and threshold $n^{-1/10}$. The experiments instead use *fixed* equal-width windows, centred at state quantiles $.15, .5, .85$, with half-width $.5$ in standardized state units. We also report quantile terciles. Neither fixed partition is covered by the guarantee: the precision of an averaged covariance is not the precision at its centre, and within-window coefficient variation remains even for equal widths.

The window setting was selected among nine settings on the evaluation seeds. Its results are therefore exploratory. The benchmark also has a persistent state, beyond the i.i.d. sampling assumed in the guarantees. These are implementation and sampling gaps between the theorem and the experiments, which the controls below assess empirically.

5 EXPERIMENTS

The experiments address three questions: does reversal survive heterogeneous noise, does curvature recover useful orders and graphs, and can simple sorting explain that recovery?

Design and metrics. We use random ten-node DAGs with expected in-degree 1.5, $n = 3000$ observations and an AR(1) common state with $\phi = .97$. Coefficients have log-slope $\kappa \in \{0, .15, .3, .6\}$. Noise variances are $\ell_i \exp(a_i v_t + .5u_{i,t})$, where ℓ_i is log-uniform on $[1/4, 4]$ and a_i is uniform on $[.5, 1.5]$. All comparisons use 30 paired seeds. Order accuracy is the fraction of true edges whose parent precedes its child; .5 is chance. Structural Hamming distance (SHD) counts missing, extra and reversed edges once each. Directed F1 and exact-graph counts assess edge recovery.

We distinguish pooled observations, a measured state, and an outcome-derived proxy. HetCD-gated is a separate log-affine likelihood heuristic using EWMA variances (Appendix T); it is not Algorithm 1. Baseline implementations and parameters are in Appendix AI. The original heterogeneous-noise study and the selector were pre-registered; fixed windows, additional baselines, graph controls and the sign controls were added afterwards (Appendix F). No claim of confirmatory evaluation is made for those additions.

5.1 REVERSAL AND RECOVERY

DirectLiNGAM falls from .98 order accuracy at $\kappa = 0$ to .32 at $\kappa = .6$ (Table 1). VARLiNGAM (Hyvärinen et al., 2010) and ICA-LiNGAM (Shimizu et al., 2006) also fall below chance, to .32 and .24, and the constant-coefficient score HetSEM-const falls to .21 (Appendix F). DirectLiMIAM, designed for dependent noise scales (Mesters et al., 2026), reaches .87 but starts at .69 without

Table 1: Order accuracy under positive modulation: mean ± standard deviation over 30 paired seeds, $d = 10$, $n = 3000$. The methods use different information sets, shown explicitly. The location-scale score is our implementation; the DirectLiMIAM row uses its released moment-(2,3) variant. Fixed windows are exploratory. Full method, selector and sign-control results are in Appendix F.

Method	$\kappa = 0$	.15	.30	.60
Pooled observations				
DirectLiNGAM	.98±.06	.67±.16	.52±.10	.32±.12
HetSEM-const	.94±.06	.71±.17	.49±.15	.21±.12
DirectLiMIAM	.69±.12	.69±.14	.83±.14	.87±.08
Location-scale score	.57±.14	.72±.13	.85±.14	.98±.05
Measured state				
HetCD-Ω, fixed windows	.56±.18	.58±.16	.68±.17	.90±.12
HetCD-Ω, terciles	.60±.12	.77±.14	.91±.10	.95±.11
HetCD-gated	.89±.08	.86±.10	.77±.15	.61±.19
Estimated state: lagged PC1 of log-EWMA variances				
HetCD-Ω, fixed windows	.52±.20	.49±.19	.54±.22	.53±.23

Table 2: Graph recovery on the positive-modulation benchmark. Each entry is mean SHD / mean directed F1 / exact graphs out of 30. Threshold pruning uses $\tau = m^{-1/4}$, where m is the smallest window or bin count. Group BIC fits a separate constant coefficient per bin. These fixed-partition results are not covered by Theorem 6.

Order / pruner	$\kappa = 0$	.15	.30	.60
Measured state				
Windows / threshold	14.0/.43/0	13.5/.45/1	10.3/.58/2	4.6/.82/4
Terciles / threshold	13.3/.47/0	7.5/.69/0	3.5/.86/9	2.9/.90/14
Terciles / group BIC	15.3/.47/0	9.3/.66/0	10.3/.72/1	14.7/.67/0
Estimated state				
Windows / threshold	15.3/.39/0	15.8/.36/0	14.9/.42/1	17.5/.38/0

modulation. The pooled location-scale score reaches .98, the strongest positive-modulation order in this comparison.

Measured-state HetCD-Ω with fixed windows rises from .56 to .90. Its precision-row threshold pruner gives SHD 4.6, directed F1 .82 and 4/30 exact graphs at $\kappa = .6$ (Table 2). Terciles perform better in this cell, with SHD 2.9 and 14/30 exact graphs. Applying the same threshold pruner with measured-state terciles to the location-scale order gives SHD 2.1 and 12/30 exact graphs (Appendix F); the curvature estimator is therefore not uniformly best, even when graph recovery is the metric.

Pruning matters separately from ordering. Group BIC gives SHD 14.7 at $\kappa = .6$ on the tercile order and 13.2 even with the true order: a constant coefficient within each bin leaves dependence that the pruner interprets as extra edges. A control closer to the theorem uses conditionally Gaussian observations at three exact levels with a curved common factor ($d = 10$, $n = 12{,}000$). Here HetCD-Ω reaches .97 order accuracy and SHD 1.1; replacing signed gaps by unsigned curvature gives .32 accuracy (Appendix G). The common-factor cancellation is essential in this control.

Across the 1,876 benchmark edges, fixed-window accuracy rises from .52 for $|\eta| < .1$ to .96 for $|\eta| > 1.5$ (Figure 2, Appendix F). This is an empirical association with imbalance. The graph theorem does not imply an edge-by-edge guarantee, and these bins contain nearly balanced rather than exactly balanced edges.

Table 3: Sign controls at strong modulation: mean order accuracy over 30 seeds. Mixed signs are drawn independently per edge, $\kappa_{ij} = \pm.6$. Fixed-window accuracy is similar across the three settings; slope sorting reverses under negative modulation and is near chance with mixed signs. Full results are in Table 9.

Method	$\kappa = .6$	$\kappa = -.6$	Mixed $\pm.6$
DirectLiNGAM	.32	.54	.43
DirectLiMIAM	.87	.70	.79
Location-scale score	.98	.82	.79
HetCD-Ω, fixed windows, measured	.90	.90	.88
HetCD-Ω, terciles, measured	.95	.80	.86
Slope sorting, measured	.93	.07	.53

5.2 Sorting controls and the sign of modulation

Under one-signed modulation, descendants tend to have steeper variance profiles. Sorting variables by the OLS slope of $\log X_j^2$ on the state, without estimating any covariance or graph, gives .93 order accuracy at $\kappa = .6$. This shortcut resembles variance sortability (Reisach et al., 2021), but uses a variance's slope across states instead of its level. A separate minimum-precision-diagonal control scores .79 on raw inputs and .03 after standardization at $\kappa = .6$; HetCD-Ω gives identical orders under either scaling (Appendix F).

Tercile averaging is also sensitive to steep variance profiles. For a log-linear variance, averaging over a bin adds approximately half the squared slope times the within-bin state variance to its log. Tail terciles have more state variation than the middle tercile, which can favour a steeper variable as a sink. Equal-width windows reduce this distortion but do not remove within-window coefficient variation. The sign controls test how much the observed performance depends on the shortcut.

At $\kappa = -.6$, slope sorting reverses to .07, tercile accuracy falls to .80, and fixed-window accuracy remains .90 (Table 3). With independent modulation signs across edges, slope sorting is near chance and the two curvature implementations reach .88 and .86. The pooled location-scale score remains competitive at .82 and .79 on these controls. Its decline from .98 is consistent with sensitivity to the shortcut; it does not establish that sorting explains all of its decisions. DirectLiNGAM still reverses on the mixed-sign generator (.43).

5.3 Method selection and the closest varying-coefficient baseline

The observable spread $S = \max_j J_j - \min_j J_j$ measures whether any excess curvature is present at the first recursion step. It is not the minimum separation required for graph recovery. We evaluated two pre-registered selection rules, using either state permutations or a bootstrap threshold calibrated on separate seeds. With tercile HetCD-Ω and HetCD-gated as components, the permutation rule gives accuracies $.89/.87/.85/.95$ across the four positive-κ cells and selects curvature on $3/13/63/100\%$ of seeds. This is a useful empirical selection rule on this generator, not a calibrated test: state permutation breaks exchangeability even at $\kappa = 0$, and the components are exposed to slope sorting. Only some pre-registered criteria were met (Appendix H).

CHOD (Zhou et al., 2022) is the closest varying-coefficient baseline. Its covariate-independent noise covariance differs from our relative-noise model. Equal relative noise slopes reduce to its setting after a common determinant normalization; unequal slopes need not. Appendix I gives a three-node example outside every constant-noise DAG representation. Our native and adapted CHOD runs used 384 chains plus longer probes, but failed the mixing diagnostics (split $\hat{R}$ up to 21). The appendix reports descriptive outputs and diagnostics; these runs do not support an accuracy ranking.

6 State information and external evidence

The state must be measured on a coordinate where the log-affine laws hold. An affine change of coordinate preserves the criterion; an arbitrary monotone change need not. A certified example changes the gap's sign under $u = v^3$ (Appendix A). Thus state ranks alone are insufficient.

Estimated and mislabelled states. Replacing the measured state by a lagged principal component of log-EWMA variances (RiskMetrics, 1996) reduces fixed-window order accuracy to .49–.54 across the nonnegative, one-signed benchmark cells, despite a correlation of .87 with the true state. At $\kappa = .6$, SHD becomes 17.5 (Tables 1–2). Negative and mixed-sign cells reach .59 and .67; the proxy does not fail equally in every setting. Tercile accuracy remains .85 at positive $\kappa = .6$, alongside proxy-based slope-sorting accuracy of .94. High proxy correlation or high tercile accuracy therefore does not establish recovery of the conditional-covariance signal.

A separate result concerns conditioning the residual-MI criterion, rather than HetCD-Ω. In the matched example of Theorem 1, exact labels restore the true direction, but independently flipping 5% of the labels makes the criterion reverse again (Theorem 9, Appendix B). The threshold is example-specific: the child's variance differs sixteen-fold between states, so small contamination strongly affects the low-variance group. Known-channel correction restores 19/20 decisions in the reported experiment at a 21-fold variance cost (Appendix AA).

Financial consistency check. On a daily 2003–2026 panel, monetary-policy surprises provide reference directions from the two-year yield to seven other rate-related variables (Gürkaynak et al., 2005; Nakamura & Steinsson, 2018; Bauer & Swanson, 2023). With lagged log MOVE or log VIX as observed states, DirectLiNGAM and HetCD-gated place the two-year yield first and agree with all seven directions. Fixed-window HetCD-Ω agrees with 4/7 under MOVE and 2/7 under VIX. The permutation selection rule chooses the fallback under both states. All seven edges share a parent, so this is one ordering event, not seven independent confirmations. The panel does not demonstrate an observed reversal or validate the graph model. Oil-supply-instrumented comparisons involve same-day feedback outside our assumptions and are reported separately (Appendices AF–AH).

7 Related work

Heterogeneity and changing mechanisms. Identification through heteroskedasticity and back-Shift hold the structural matrix fixed while shock variances change (Sentana & Fiorentini, 2001; Rigobon, 2003; Lanne et al., 2010; Rothenhäusler et al., 2015; Lewis, 2021). CD-NOD uses independent changes in causal modules; coordinated dependence on one state violates that orientation premise (Huang et al., 2020). Huang et al. (2019) model jointly evolving coefficients and variances through state-space dynamics. CHOD (Zhou et al., 2022) permits covariate-dependent coefficients with covariate-independent noise. Our restriction is log-affinity of coefficients and relative noise variances, with an unrestricted common variance factor. Location-scale models instead allow the conditional mean and scale to depend on parents (Immer et al., 2023; Strobl & Lasko, 2023; Xia et al., 2026); the parent-dependent scale alone does not make them the exogenous-state model studied here.

Precision orderings and misspecification. Ghoshal & Honorio (2018) remove terminal vertices using precision diagonals under variance-gap conditions; related orderings use conditional variances or Hessians (Chen et al., 2019; Gao et al., 2020; Park, 2020; Rolland et al., 2022). Zhou et al. (2023) use covariate-dependent precision diagonals and a shared-sink induction under covariate-independent noise. We build on that proof template, replacing precision levels by signed curvature across states. Mesters et al. (2026) establish LiNGAM reversal from dependent noise scales with fixed coefficients and propose DirectLiMIAM. Our matched construction holds the noise process unchanged across conditions and changes the coefficient; it certifies both residual MI and the deployed score. Score-based inversion under other misspecification is studied by Montagna et al. (2023). Our sorting controls extend the concern of Reisach et al. (2021) to state-dependent variance slopes.

8 Conclusion

Coefficient modulation reverses the arrow a pooled non-Gaussian criterion prefers, for residual mutual information and the deployed DirectLiNGAM score alike. Read through state-conditional covariances, the same movement identifies the graph: under log-affine coefficients and relative noise variances, sinks minimize the signed curvature of the log precision diagonal, and three conditional covariances recover the DAG. Identification fails on the balance surface and weakens near it at a fourth-order rate the estimator matches.

AI USE STATEMENT

In this work we used generative AI tools, in the form of large language model based assistants, for research ideation and execution, for retrieval and discovery of related work, for proving mathematical claims, to draft sections of the paper, and to aid and polish the writing. Concretely: the assistants proposed and discussed candidate constructions and experimental designs, wrote and ran much of the analysis code and the computer-assisted certificate programs under our direction, drafted proofs of the results in Sections 2–4 and of the label-error result and of the appendix propositions, searched for and summarized related work, and drafted and edited the text of every section. We have reviewed all AI-assisted work. The proofs were checked numerically where a numerical check exists (the interval-arithmetic certificates, the population-recovery checks of Appendix A, the cross-check of Appendix D) and reviewed by the authors; the code was read, executed by the authors and its outputs compared with the numbers in the paper; related work surfaced by the assistants was read in the original before being cited, and the literature was also surveyed manually; drafted text was rewritten and checked against the results it describes. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is methodological. The empirical material consists of synthetic data, public daily financial market series (Federal Reserve Economic Data, exchange-traded fund prices, published monetary-policy and oil-supply shock series) a public physical-measurement dataset, and, in one appendix check, the public EuroBaVar beat-to-beat recordings of 23 volunteers as distributed by its authors; no new human-participant data were collected and no personal or private records are involved. We do not redistribute the market or instrument data: the supplementary material instead contains a script that rebuilds the daily panel from its original providers, to be used under those providers' terms, together with checksums of the panel analysed in the paper. The main foreseeable misuse of the results is the one the paper warns against: reading a confidently oriented arrow from a non-Gaussian or a heteroskedasticity-based method as a causal finding in a regime where the structural coefficient co-moves with volatility, for example in financial risk management, where the paper shows the arrow can be systematically reversed. The paper's contribution is to make that failure and its repair explicit, together with the conditions under which the repair itself fails. The authors declare no competing interests beyond their employment.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code for every number in the paper, organised by section: the interval-arithmetic certificate programs behind Theorems 1 and 9 and Corollary 2 (`certificates/`), with an independent quadrature cross-check (`crosscheck/`, Appendix D); the generator, the estimators (HetCD-Ω with both window partitions, HetCD-gated, the selector rules, the controls) and the benchmark scripts, evaluation seeds, pre-registration documents and frozen outputs behind Tables 1–3 and Figures 1 and 2 (`benchmark/`, `core/`, Appendices E, F–G); the executed CHOD comparison, including the R/C++ sources, the variant kernels and the mixing diagnostics (`chod/`, Appendix I); the label-error and state-information studies (`label_error/`, `state_information/`, Appendices Z–AD); and the real-data pipeline, which rebuilds the daily market panel from its original providers, checks the result against checksums of the panel used here, and documents where to obtain the instrument series, together with the light-tunnel protocol (`real_data/`, Appendices AF–AH). Assumptions and complete proofs are in Appendices B and A. Appendix AI lists package versions and the provenance of every baseline, and Appendix AJ the pre-registered designs. The appendix names the output file behind each table and figure, and the scripts regenerate those files from the seeds given.

REFERENCES

A. Ang, J. Chen. Asymmetric correlations of equity portfolios. *J. Financial Econ.* 63(3):443–494, 2002.

O. E. Barndorff-Nielsen, J. Kent, M. Sørensen. Normal variance-mean mixtures and z distributions. *Int. Stat. Rev.* 50(2):145–159, 1982.

M. D. Bauer, E. T. Swanson. A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual* 37:87–155, 2023.

W. Chen, M. Drton, Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika* 106(4):973–980, 2019.

P. Comon. Independent component analysis, a new concept? *Signal Processing* 36(3):287–314, 1994.

G. Darmois. Analyse générale des liaisons stochastiques: étude particulière de l'analyse factorielle linéaire. *Rev. Inst. Int. Stat.* 21(1/2):2–8, 1953.

K. J. Forbes, R. Rigobon. No contagion, only interdependence: measuring stock market comovements. *J. Finance* 57(5):2223–2261, 2002.

J. L. Gamella, J. Peters, P. Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence* 7:107–118, 2025.

M. Gao, Y. Ding, B. Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *NeurIPS*, 2020.

A. Ghoshal, J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. *AISTATS*, PMLR 84:1466–1475, 2018.

W. Gong, J. Jennings, C. Zhang, N. Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *ICLR*, 2023.

R. S. Gürkaynak, B. Sack, E. T. Swanson. Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *Int. J. Central Banking* 1(1):55–93, 2005.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054, 1982.

A. Harvey, E. Ruiz, N. Shephard. Multivariate stochastic variance models. *Rev. Econ. Stud.* 61(2):247–264, 1994.

B. Huang, K. Zhang, M. Gong, C. Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. *ICML*, PMLR 97:2901–2910, 2019.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.* 21(89):1–53, 2020.

Y. Huang. Corrected score with sizable covariate measurement error: pathology and remedy. *Statistica Sinica* 24(1):357–374, 2014. doi:10.5705/ss.2012.181.

R. Hulsman, C. Balsells-Rodas, S. Magliacane. Identifiable Markov switching models with instantaneous effects and exponential families. *ICML*, 2026. arXiv:2606.02231.

A. Hyvärinen, S. M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.* 14:111–152, 2013.

A. Hyvärinen, K. Zhang, S. Shimizu, P. O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *J. Mach. Learn. Res.* 11:1709–1731, 2010.

A. Immer, C. Schultheiss, J. E. Vogt, B. Schölkopf, P. Bühlmann, A. Marx. On the identifiability and estimation of causal location-scale noise models. *ICML*, PMLR 202:14316–14332, 2023.

F. Javed. The kurtosis of normal variance-mean mixtures. *arXiv:2606.22951*, 2026.

F. Johansson. Arb: efficient arbitrary-precision midpoint-radius interval arithmetic. arXiv:1611.02831, 2016.

D. R. Känzig. The macroeconomic effects of oil supply news: evidence from OPEC announcements. *Amer. Econ. Rev.* 111(4):1092–1125, 2021.

M. Lanne, H. Lütkepohl, K. Maciejowska. Structural vector autoregressions with Markov switching. *J. Econ. Dyn. Control* 34(2):121–131, 2010.

D. Laude, J.-L. Elghozi, A. Girard, et al. Comparison of various techniques used to estimate spontaneous baroreflex sensitivity (the EuroBaVar study). *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 286(1):R226–R231, 2004.

D. J. Lewis. Identifying shocks via time-varying volatility. *Rev. Econ. Stud.* 88(6):3086–3124, 2021.

F. Longin, B. Solnik. Extreme correlation of international equity markets. *J. Finance* 56(2):649–676, 2001.

G. Mesters, A. Ribot, A. Seigal, P. Zwiernik. Causal discovery under mean independence and linearity. *arXiv:2605.04381*, 2026.

F. Montagna, A. A. Mastakouri, E. Eulig, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, F. Locatello. Assumption violations in causal discovery and the robustness of score matching. *NeurIPS*, 2023.

M. J. Morrissey, D. Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* 55(2):338–344, 1999.

E. Nakamura, J. Steinsson. High-frequency identification of monetary non-neutrality: the information effect. *Quart. J. Econ.* 133(3):1283–1330, 2018.

T. Nakamura. Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* 77(1):127–137, 1990.

R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, B. Aragam. DYNOTEARS: structure learning from time-series data. *AISTATS*, PMLR 108:1595–1605, 2020.

G. Park. Identifiability of additive noise models using conditional variances. *J. Mach. Learn. Res.* 21(75):1–34, 2020.

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

J. Peters, P. Bühlmann, N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B* 78(5):947–1012, 2016.

A. Rahmani, P. Frossard. Causal temporal regime structure learning. *AISTATS*, PMLR 258:4546–4554, 2025.

A. Rahmani, P. Frossard. Flow-based non-stationary temporal regime causal structure learning. *NeurIPS*, 2025. arXiv:2506.17065.

A. Reisach, C. Seiler, S. Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *NeurIPS*, 2021.

R. Rigobon. Identification through heteroskedasticity. *Rev. Econ. Stat.* 85(4):777–792, 2003.

J. P. Morgan/Reuters. *RiskMetrics Technical Document*, 4th ed., 1996.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. *ICML*, PMLR 162:18741–18753, 2022.

D. Rothenhäusler, C. Heinze, J. Peters, N. Meinshausen. backShift: learning causal cyclic graphs from unknown shift interventions. *NeurIPS*, 2015.

E. Sentana, G. Fiorentini. Identification, estimation and testing of conditionally heteroskedastic factor models. *J. Econometrics* 102(2):143–164, 2001.

S. Shimizu, A. Hyvärinen. Discovery of linear non-Gaussian acyclic models in the presence of latent classes. *ICONIP 2007*, LNCS 4984:752–761, 2008.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7:2003–2030, 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.* 12:1225–1248, 2011.

V. P. Skitovich. On a property of the normal distribution. *Dokl. Akad. Nauk SSSR* 89(2):217–219, 1953.

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

E. V. Strobl, T. A. Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *J. Comput. Sci.* 72:102099, 2023.

N. Tanaka, S. Shimizu, T. Washio. A Bayesian estimation approach to analyze non-Gaussian data-generating processes with latent classes. arXiv:1408.0337, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar, G. Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. arXiv:1011.3027.

Y. S. Wang, M. Kolar, M. Drton. Confidence sets for causal orderings. arXiv:2305.14506v2, 2024.

X. Xia, L. Chen, Y. Hu, C. Li. Learning a directed acyclic graph with additive heteroscedastic errors. *arXiv:2605.26515*, 2026.

X. Zheng, B. Aragam, P. Ravikumar, E. P. Xing. DAGs with NO TEARS: continuous optimization for structure learning. *NeurIPS*, 2018.

F. Zhou, K. He, Y. Ni. Causal discovery with heterogeneous observational data. *UAI*, PMLR 180:2383–2393, 2022.

F. Zhou, K. He, Y. Ni. Individualized causal discovery with latent trajectory embedded Bayesian networks. *Biometrics* 79(4):3191–3202, 2023.

APPENDIX GUIDE

The appendix preserves the complete experimental record. The first group supports the main claims; later sections give additional controls, alternative estimators and exploratory studies.

Question	Where to look
Why is the graph identified?	Appendix A: graph proof, finite-sample recovery, lower bound and state coordinate.
Why does the pooled score reverse?	Appendices B–D: certificates, thresholds and independent numerical checks.
How was recovery evaluated?	Appendices E–H: generator, full benchmark, Gaussian controls and selector.
What happened with CHOD?	Appendix I: model comparison, source audit and mixing diagnostics.
How broad is the evidence?	Appendices J–S: additional generators and controls; Appendices T–Y: the separate likelihood heuristic.
What if state information is wrong?	Appendices Z–AD: state information, correction, confidence sets and related checks.
What do the real data show?	Appendices AF–AH: external evidence and observed-state analyses.
How can the record be checked?	Appendices AI and AJ: implementation and pre-registration.

A PROOFS FOR IDENTIFICATION AND ESTIMATION

A.1 PROOF OF LEMMA 4

Fix v and an ancestral set R. Because R is closed under parents and G is a DAG, the equations of X_R do not involve X_{R^c}, so $X_R = B_R(v)X_R + \varepsilon_R$ with B_R the restriction of B and $Q_R = \mathrm{Cov}(\varepsilon_R \mid v)$ diagonal. Hence $\Sigma_R = (I - B_R)^{-1}Q_R(I - B_R)^{-\top}$ and $\Omega_R = (I - B_R)^{\top}Q_R^{-1}(I - B_R)$, whose (j, j) entry is $\sum_{k\in R}(I - B_R)^2_{kj}/Q_{kk} = Q_{jj}^{-1} + \sum_{i\in \mathrm{ch}_R(j)} B_{ij}^2/Q_{ii}$. Substituting $B_{ij} = \beta_{ij}e^{\kappa_{ij}v}$ and $Q_{ii} = g(v)d_ie^{a_iv}$,

$$\begin{aligned}\Omega_{R,jj}(v) &= g(v)^{-1}\Big[d_j^{-1}e^{-a_jv} + \sum_{i\in \mathrm{ch}_R(j)} \beta_{ij}^2 d_i^{-1}e^{(2\kappa_{ij}-a_i)v}\Big] \\ &= g(v)^{-1}d_j^{-1}e^{-a_jv}\Big[1 + \sum_{i\in \mathrm{ch}_R(j)} \tfrac{\beta_{ij}^2 d_j}{d_i}e^{\eta_{ij}v}\Big],\end{aligned}$$

which is equation 6. If j is a sink of $G[R]$ the bracket is 1 and the right-hand side is affine. Otherwise $\phi(v) = \log\left(w_0 + \sum_i w_i e^{\eta_{ij}v}\right)$ with positive weights is a log-sum-exp of affine functions; its second derivative is the variance of the slopes $\{0\} \cup \{\eta_{ij}\}$ under the softmax weights, which is strictly positive whenever at least one $\eta_{ij} \neq 0$. □

A.2 PROOF OF THEOREM 5

(i) By Lemma 4, for a sink s of $G[R]$ the function $v \mapsto \log \Omega_{R,ss}(v) + \log g(v)$ is affine, so its Jensen gap vanishes and $J_{R,s} = J_g$, which does not depend on s. For a non-sink j, strict convexity gives $J_{R,j} > J_g$. *(ii)* For a sink s of $G[R]$, $(I - B_R)_{ks} = 0$ for $k \neq s$, so $\Omega_{R,sj} = (I - B_R)_{ss}(I - B_R)_{sj}/Q_{ss} = -B_{sj}/Q_{ss}$ and $\Omega_{R,ss} = 1/Q_{ss}$; the ratio is $B_{sj}(v)$, nonzero iff $j \in \mathrm{pa}(s)$.

Recursion. Removing a sink s from an ancestral R leaves an ancestral set, and $X_{R\setminus\{s\}}$ satisfies Assumption 2 with $G[R \setminus \{s\}]$ and the same ε; separation is edgewise and is inherited. Applying (i)–(ii) at each step recovers a topological order and every coefficient at each v_k.

Uniqueness. Let (G', B', Q') satisfy Assumption 2 (possibly with balanced edges and a different common factor g') and generate the same $\Sigma(v_k)$, $k = 1, 2, 3$. Apply Lemma 4 to the alternative: every node of G' has $J_{[d],j} \geq J_{g'}$ with equality at its sinks (convexity, not strict convexity, is all that

is used). The observed gaps are the same under both models, so $\min_j J_{[d],j} = J_{g'}$ and, by (i) for the true model, the minimizer set equals the set of true sinks. Hence every sink s of G' is a sink of G. By (ii), applied to each model, both models read the same parent set and the same coefficients for s from the same precision row. Remove s: the marginal covariances of $X_{[d]\setminus\{s\}}$ agree, both reduced models satisfy Assumption 2 on $G[\cdot]$ and $G'[\cdot]$ respectively, and separation persists for the true one. Induction on d shows that every variable receives the same incoming edges and coefficients in both models, so $G' = G$ and $B' = B$ at the observed states. Nothing was assumed about g'. □

Weaker separation. The edgewise condition can be replaced by requiring an unbalanced outgoing edge for every non-sink in every ancestral subset. Lemma 4 then separates sinks from non-sinks on each set reached by the recursion. Removing a sink leaves an ancestral subset, so the same condition applies at the next step. A node whose children have all been removed is now a sink and no longer needs an outgoing edge. The shared-sink uniqueness argument uses this condition only for the true graph; the alternative still needs no separation.

Proposition 8 (balance is a genuine ambiguity)**.** *If every edge is balanced ($\eta_{ij} \equiv 0$), then $\Sigma(v) = g(v)D(v)\Sigma_0 D(v)$ with $D(v) = \mathrm{diag}(e^{a_i v/2})$, and for conditionally Gaussian errors every DAG obtained from a Cholesky factorization of Σ_0 in another variable order yields a model satisfying Assumption 2 with the same conditional laws. In particular constant coefficients with a common noise slope are not identified from second moments.*

A.3 Proof of Proposition 8

If $\eta_{ij} = 0$ on every edge then $\kappa_{ij} = (a_i - a_j)/2$, so $B(v) = D(v)B_0D(v)^{-1}$ with $B_0 = (\beta_{ij})$ and $D(v) = \mathrm{diag}(e^{a_i v/2})$. Then $(I - B(v))^{-1} = D(v)(I - B_0)^{-1}D(v)^{-1}$ and $\Sigma(v) = g(v)D(v)(I - B_0)^{-1}\,\mathrm{diag}(d_i)(I - B_0)^{-\top}D(v) = g(v)D(v)\Sigma_0 D(v)$. For any permutation σ, the Cholesky factorization of Σ_0 in the order σ gives $\Sigma_0 = (I - \tilde{B}_0)^{-1}\tilde{Q}_0(I - \tilde{B}_0)^{-\top}$ with $\tilde{B}_0$ strictly lower-triangular in that order and $\tilde{Q}_0$ diagonal; conjugating by $D(v)$ produces $\tilde{B}(v) = D\tilde{B}_0D^{-1}$, whose entries are $\tilde{\beta}_{ij}e^{(a_i - a_j)v/2}$ (log-affine, balanced) and $\tilde{Q}(v) = g(v)\,\mathrm{diag}(\tilde{d}_i e^{a_i v})$. This is a model satisfying Assumption 2 with the DAG of $\tilde{B}_0$, and for Gaussian errors the conditional law is determined by $\Sigma(v)$, so the laws coincide. □

A.4 Proof of Theorem 6

Step 1: the event controls every submatrix. Loewner order is preserved under passing to principal submatrices and reversed under inversion, so equation 8 gives, for every R and k, $(1 + \rho)^{-1}\Omega_R(v_k) \preceq \widehat{\Omega}_{R,k} \preceq (1 - \rho)^{-1}\Omega_R(v_k)$. In particular each diagonal entry satisfies $|\log \widehat{\Omega}_{R,k,jj} - \log \Omega_{R,jj}(v_k)| \le r := -\log(1 - \rho)$.

Step 2: ordering. The gap equation 7 has weights $(1 - t), t, -1$ whose absolute values sum to 2, so $|\widehat{J}_{R,j} - J_{R,j}| \le 2r$ for every $j \in R$. If $\gamma > 4r$, then for every non-sink j and sink s of $G[R]$, $\widehat{J}_{R,j} \ge J_{R,j} - 2r \ge J_g + \gamma - 2r > J_g + 2r \ge \widehat{J}_{R,s}$, so the empirical minimizer is a true sink. Because the same event controls every R, this holds at every step of the recursion regardless of which sink was chosen earlier; no union bound over the 2^d subsets is needed.

Step 3: pruning. Let $\Omega = \Omega_R(v_2)$, $\widehat{\Omega} = \widehat{\Omega}_{R,2}$, $q = \rho/(1 - \rho)$. From Step 1, $M := \Omega^{-1/2}(\widehat{\Omega} - \Omega)\Omega^{-1/2}$ has spectrum in $[-\rho/(1 + \rho), q] \subset [-q, q]$, so $|\widehat{\Omega}_{sj} - \Omega_{sj}| = |e_s^\top \Omega^{1/2} M \Omega^{1/2} e_j| \le q\sqrt{\Omega_{ss}\Omega_{jj}}$. Write $P = \sqrt{\Omega_{ss}\Omega_{jj}}$ and $\widehat{P}$ likewise; Step 1 gives $P/\widehat{P} \in [1 - \rho, 1 + \rho]$. With $c_{sj} = -\Omega_{sj}/P$ (so $|c_{sj}| \le 1$) and $\widehat{c}_{sj} = -\widehat{\Omega}_{sj}/\widehat{P}$,

$$|\widehat{c}_{sj} - c_{sj}| \le \frac{|\widehat{\Omega}_{sj} - \Omega_{sj}|}{\widehat{P}} + |c_{sj}|\left|\frac{P}{\widehat{P}} - 1\right| \le q(1 + \rho) + \rho = \frac{2\rho}{1 - \rho} = \tau.$$

For a sink s, $c_{sj} = 0$ iff $j \notin \mathrm{pa}(s)$ (Theorem 5(ii)), so thresholding $|\widehat{c}_{sj}|$ at τ admits no false edge, and recovers every true edge if $\zeta > 2\tau = 4\rho/(1 - \rho)$.

Step 4: the Gaussian event. Write $x_i = \Sigma(v_k)^{1/2} z_i$ with $z_i \sim N(0, I_d)$, $i \le m$, and Z the $m \times d$ matrix of the z_i. Then $\Sigma(v_k)^{-1/2}\widehat{\Sigma}_k\Sigma(v_k)^{-1/2} = Z^\top Z/m$. By Gaussian singular-value

concentration (Vershynin, 2012, Cor. 5.35), with probability at least $1 - 2e^{-u^2/2}$ all singular values of $Z/\sqrt{m}$ lie in $[1-a, 1+a]$, $a = (\sqrt{d}+u)/\sqrt{m}$; squaring, the eigenvalues of $Z^\top Z/m$ lie in $[(1-a)^2, (1+a)^2] \subseteq [1-\rho, 1+\rho]$ with $\rho = 2a + a^2$. A union bound over the three states with $u = \sqrt{2\log(6/\alpha)}$ gives equation 8 with probability $\geq 1-\alpha$. Solving $\rho \leq \rho_0$ for m gives $m \geq (\sqrt{d}+u)^2/(\sqrt{1+\rho_0}-1)^2$; for $\rho_0 \leq 1/2$ the denominator is $\asymp \rho_0^2$, and $\rho_0 \asymp \min(1, \gamma, \zeta)$ is what Steps 2–3 require, giving the displayed order. □

Conservativeness. For the chain $X_1 \to X_2 \to X_3$ with $\beta = (.7, -.8)$, $\kappa \equiv .6$, $a \equiv 1$, $d = (.7, 1.5, 2)$, levels $(-1, .1, 1.4)$ and $\alpha = .05$, the bound requires $m = 204{,}259$ per state; at that m, 200 Wishart draws all satisfied equation 8 and recovered the graph. The benchmark of Section 5 recovers $d = 10$ graphs at $n = 3000$ total. The bound certifies; it does not predict sample sizes.

Consistency under weaker moments. At fixed d with three exact state levels, finite conditional fourth moments give $\|\widehat{\Sigma}_k - \Sigma(v_k)\| = O_p(m^{-1/2})$, hence $\widehat{J}_{R,j} \to J_{R,j}$ and the order is eventually correct with probability $\to 1$ when $\gamma > 0$. With $\tau_m = m^{-1/4}$, true normalized entries ($\geq \zeta$) exceed τ_m and null entries ($O_p(m^{-1/2})$) fall below it eventually, so $\widehat{G} \to G$. Finite second moments alone give order consistency but not the rate needed for the deterministic threshold.

Continuous exactly observed state. Let (X, V) be i.i.d., d fixed, $\mathbb{E}[X \mid v] = 0$, V with a continuous density bounded below near its $.2, .5, .8$ quantiles $c_1 < c_2 < c_3$, $\Sigma(\cdot)$ locally Lipschitz and positive definite there, and conditional fourth moments locally bounded. Estimate $\Sigma(c_k)$ by the average of $XX^\top$ over the $m_n = \lceil n^{.8} \rceil$ observations whose state is nearest to the empirical k-th quantile. The bias is $O_p(m_n/n) = O_p(n^{-.2})$ (window width), the variance term $O_p(m_n^{-1/2}) = O_p(n^{-.4})$, and the anchor error $O_p(n^{-1/2})$; the total is $O_p(n^{-1/5})$, the same for the interpolation weight t. Continuity of the recursion in its inputs gives order consistency, and thresholding at $\tau_n = n^{-1/10}$ gives graph consistency. These exponents are conservative and not optimized; the estimator is weak at $n \leq 3000$ (Section 5). Selecting windows by V alone, before looking at X, is what makes the argument non-circular; a state computed from X is not covered.

A.5 Proof of Theorem 7

Let $b, r, d_0 > 0$, $p = b^2/(b^2+r)$, $C = b^2 + r$, $h = \eta v$, and $g(\cdot) > 0$ arbitrary. Forward model ($X \to Y$): $b_F(v) = be^{\kappa v}$, $q_X(v) = gd_0 e^{av}$, $q_Y(v) = gd_0 re^{(a+2\kappa-\eta)v}$; its imbalance is $2\kappa - [(a+2\kappa-\eta) - a] = \eta$. Reverse model ($Y \to X$): $c_R(v) = \frac{b}{C} e^{[-\kappa+(1-p)\eta]v}$, $q_{Y,R}(v) = gd_0 C e^{[a+2\kappa-(1-p)\eta]v}$, $q_{X,R}(v) = g\frac{d_0 r}{C} e^{(a-p\eta)v}$; its imbalance is $2[-\kappa+(1-p)\eta] - [(a-p\eta) - (a+2\kappa-(1-p)\eta)] = \eta$. Both models therefore satisfy Assumptions 2–3. Direct multiplication gives, with $R(h) = pe^{(1-p)h} + (1-p)e^{-ph}$,

$$\Sigma_{R,XY} = \Sigma_{F,XY}, \quad \det \Sigma_R = \det \Sigma_F, \quad \Sigma_{R,XX} = R(h)\,\Sigma_{F,XX}, \quad \Sigma_{R,YY} = \Sigma_{F,YY}/R(h).$$

For two centered Gaussians with equal determinants, $\mathrm{KL}(P_F \| P_R) = \frac{1}{2}\big(\mathrm{tr}(\Sigma_R^{-1}\Sigma_F) - 2\big)$, and the identities above reduce it to $\frac{1}{2}\big(1 + \frac{b^2}{r}e^h\big)\big[R(h) + R(h)^{-1} - 2\big]$. Since $R(h) = 1 + \frac{p(1-p)}{2}h^2 + O(h^3)$, $R + R^{-1} - 2 = \frac{p^2(1-p)^2}{4}h^4 + O(h^5)$ and $\frac{1}{2}(1 + b^2/r) \cdot \frac{p^2(1-p)^2}{4} = \frac{p^2(1-p)}{8}$ because $(1+b^2/r)(1-p) = 1$; this is the displayed expansion, uniform for bounded v. For n independent observations the KL tensorizes, Pinsker gives total variation $\leq \sqrt{n\,\mathrm{KL}/2} \to 0$ when $n\eta_n^4 \to 0$, and the two-point testing identity puts the optimal equal-prior error at $\frac{1}{2}(1 - \mathrm{TV}) \to \frac{1}{2}$. Adjoining $d-2$ independent isolated variables with identical laws under both hypotheses leaves the KL unchanged. □

Matching. In the same subfamily each non-sink excess is $J_{R,j} - J_g = c_{R,j}\eta^2 + o(\eta^2)$ with $c_{R,j} > 0$ (second-order expansion of the log-sum-exp in Lemma 4), so $\gamma \asymp \eta^2$ and Theorem 6 requires $m \asymp \eta^{-4}$ at fixed d, the power of Theorem 7. The two-variable Gaussian lower bound of Proposition 14 allowed an unrestricted reverse variance law; the pair above lies inside the restricted class, which is what makes the comparison with Theorem 6 legitimate.

State coordinate (interval-certified). For $X \to Y$ with $b(v) = .7e^{.6v}$, $q_X = e^v$, $q_Y = e^{1.3v}$ and levels $\{-1, 0, 2\}$ ($\eta = .9$), the source-minus-sink signed gap is in $[0.18144968, 0.18144969]$;

replacing the levels by $u = v^3$ (same three conditional covariances, interpolation weight $1/9$ instead of $1/3$) gives $[-0.28425377, -0.28425376]$. Both enclosures are 128-bit ball arithmetic (`core/graph_theory/state_coordinate_check.py`). The criterion needs the state on an affine image of the coordinate in which the laws are log-affine; monotone relabelling is not enough.

Numerical check. Population recovery by Algorithm 1 from exact covariances was verified on 258 random DAGs with arbitrary common factors, heterogeneous slopes and signed coefficients, and the $d = 2$ case of the KL expansion of Theorem 7 was checked to 10^{-10} (`core/graph_theory/theory_checks.py`).

B Proofs

B.1 Statements deferred from the main text

Theorem 9 (exact labels repair; 5% label errors reverse again)**.** *In the model of Theorem 1 with $m = 1$, observe a label L equal to V with probability $1 - p$ and to $-V$ otherwise, independently of everything else. The label-conditional comparison $D(p) = I(Y; R_X \mid L) - I(X; R_Y \mid L)$ with within-label least-squares residuals is certified in $D(0) \in (.0181, .0217)$, $D(.01) \in (.0134, .0170)$, $D(.05) \in (-.0129, -.0090)$, $D(.1) \in (-.0448, -.0408)$; the constant-coefficient control is certified positive at $p = .05$.*

Proposition 10 (conditioning on the common state is not Gaussian conditioning)**.** *Under Assumption 1 with $C > 0$ and Gaussian innovations and idiosyncratic states, $x \mid v$ is a normal scale mixture with excess kurtosis $3(e^{C^2} - 1)$, and the two structural errors are independent given v. The within-state law therefore identifies direction in the LiNGAM sense; only conditioning on (v, u_x, u_y) yields a Gaussian law without direction information.*

Proposition 11 (covariance ambiguity without restriction)**.** *Every positive definite $\Sigma(v)$ is reproduced by the reverse regression representation with coefficient $c^\star(v) = \Sigma_{12}/\Sigma_{22} = b/(b^2 + r)$, root variance Σ_{22} and residual variance $\det \Sigma/\Sigma_{22}$. No functional of $\{\Sigma(v)\}_v$ distinguishes the directions if the reverse coefficient and variances are unrestricted.*

Proposition 12 (kurtosis inflation)**.** *Let $W_x = s_x^2$, $W_y = b(v)^2 s_x^2 + s_y^2$ and $\mathrm{CV}^2(W) = \mathrm{Var}(W)/\mathbb{E}[W]^2$. Under Assumption 1 with Gaussian innovations, identically distributed idiosyncratic states u_x, u_y and the exponential moments $\mathbb{E}e^{2Av+2Cu}$ finite: (a) both marginals are normal scale mixtures with excess kurtoses $3\mathrm{CV}^2(W_x)$ and $3\mathrm{CV}^2(W_y)$, the former free of κ; (b) $\Delta(\kappa) = \mathrm{Kurt}(y) - \mathrm{Kurt}(x)$ has $\Delta(0) \le 0$ with equality iff $C = 0$; (c) with Gaussian common and idiosyncratic states and $\kappa > -A/2$, $\mathrm{CV}^2(b(v)^2 s_x^2)$ is strictly increasing in κ, $\Delta(\kappa) \to \infty$ as $\kappa \to \infty$, and Δ has the closed form given in the proof (for bounded states such as the sign states of Theorem 1 the gap converges to a finite limit instead). A kurtosis crossing is a mechanism, not a proof that a given ordering score reverses; the score-level statements are Theorem 1 and Corollary 2. Above the root $\kappa^\star$ of Δ (0.071 at the benchmark calibration) the child is the more leptokurtic variable.*

Proposition 13 (compact-family Gaussian quasi-likelihood separation)**.** *Let the support of v contain an interval, $\mathbb{E}|v| < \infty$, $\mathbb{E}|\log q_x(v)| < \infty$. Restrict candidate coefficients in both directions to a compact set of signed log-affine functions containing the true forward coefficient, and profile over arbitrary positive conditional variances. If $\delta \neq 2\kappa$, the minimum expected Gaussian negative quasi-log-likelihood is strictly smaller in the forward direction, even when observations given v are not Gaussian. The profiled reverse excess loss is $D(c) = \frac{1}{2}\mathbb{E}\log\{1 + [(b^2+r)c - b]^2/r\}$, which vanishes iff $c = c^\star$ a.s.*

Proposition 14 (quartic weak separation near the boundary)**.** *With bounded state support, $\mathrm{Var}(v) > 0$, $b_0 > 0$ (a sign change of y reduces $b_0 < 0$ to this case), $\eta = 2\kappa - \delta$, $p = b_0^2/(b_0^2 + r_0)$ and Πv^2 the L^2 projection of v^2 on $\mathrm{span}\{1, v\}$, the minimum profiled gap over a compact log-affine family $\{c_\theta = e^{\theta^\top (1,v)} : \theta \in K\}$ whose parameter set K contains the balanced reverse parameter $\theta_0 = (\log[b_0/(b_0^2+r_0)], -\kappa)$ in its interior satisfies $D_{\min}(\eta) = \frac{b_0^2}{8r_0} p^2(1-p)^2\,\mathbb{E}[(v^2 - \Pi v^2)^2]\,\eta^4 + O(|\eta|^5)$, with a unique interior minimizer for small $|\eta|$. The leading coefficient is positive iff v has at least three support points; with two the gap is exactly zero for small $|\eta|$. For conditionally Gaussian errors a reverse model achieving $D_{\min}$ has KL divergence $D_{\min}(\eta)$ from the forward law, so with $n\eta_n^4 \to 0$ the optimal equal-prior error for telling the two directions apart tends to $1/2$.*

For $v \sim \mathrm{Uniform}[-a, a]$ the projection residual has variance $4a^4/45$, so halving the state range cuts the leading gap sixteen-fold.

B.2 Proof of Proposition 10 (common-state conditioning)

Independence of the structural errors given v follows from independence of (u_x, z_x) and (u_y, z_y). With $x = e^{(Av+Cu_x)/2} z_x$ and $u_x, z_x \sim N(0,1)$ independent, $\mathbb{E}[x^2 \mid v] = e^{Av}\mathbb{E}e^{Cu_x} = e^{Av+C^2/2}$ and $\mathbb{E}[x^4 \mid v] = 3e^{2Av}\mathbb{E}e^{2Cu_x} = 3e^{2Av+2C^2}$, so the conditional excess kurtosis is $3e^{2C^2}/e^{C^2} - 3 = 3(e^{C^2} - 1)$, which is 0.852 at $C = 0.5$ and positive for every $C > 0$. At fixed v the pair (x, y) is a linear model $y = b(v)x + e_y$ with x and e_y independent and both non-Gaussian; by the Darmois–Skitovich theorem (Darmois, 1953; Skitovich, 1953; Comon, 1994) a linear mixture of independent sources has a unique independent-source representation unless the sources are Gaussian, which is the identifiability condition of LiNGAM (Shimizu et al., 2006). Conditioning on (v, u_x, u_y) instead makes both errors Gaussian and the representation non-unique; that is the statement that was previously mis-attributed to conditioning on v. □

B.3 Proof of Proposition 11 (covariance ambiguity)

Fix v and write (S_{11}, S_{12}, S_{22}) for the entries of $\Sigma(v)$. The reverse one-way model with y exogenous has covariance $\begin{pmatrix} c^2\sigma_y^2 + \sigma_x^2 & c\sigma_y^2 \\ c\sigma_y^2 & \sigma_y^2 \end{pmatrix}$. Matching termwise: $\sigma_y^2 = S_{22}$, $c = S_{12}/S_{22}$, $\sigma_x^2 = S_{11} - S_{12}^2/S_{22} = \det\Sigma(v)/S_{22}$, all positive because $\Sigma(v)$ is the covariance of two non-degenerate variables. The construction is carried out independently for every v. For equation 3 explicitly, $c^\star(v) = bq_x/(b^2q_x + q_y) = b/(b^2 + r)$. Nothing here constrains c or the variances to any family, which is why the ambiguity is removed by such a constraint (Theorem 3). □

B.4 Proof of Theorem 3 (noise-slope boundary)

From Proposition 11, $\log|c^\star(v)| = \log|b_0| + \kappa v - \log\left(b_0^2 e^{2\kappa v} + r_0 e^{\delta v}\right)$. The last term is a log-sum-exp of two exponentials with rates 2κ and δ; its second derivative is the variance of the rate under the softmax weights $p(v) = b_0^2 e^{2\kappa v}/(b_0^2 e^{2\kappa v} + r_0 e^{\delta v})$ and $1 - p(v)$, i.e. $(2\kappa - \delta)^2 p(1-p)$, which gives equation 4. It vanishes identically on an interval iff $2\kappa = \delta$, since $p(1-p) > 0$. If $\delta = 2\kappa$ the sum is $(b_0^2 + r_0)e^{2\kappa v}$ and $c^\star = b_0(b_0^2 + r_0)^{-1}e^{-\kappa v} \in \mathcal{F}$. If $\delta \neq 2\kappa$, $\log|c^\star|$ is strictly concave on the interval and coincides with no affine function on it; because covariance matching fixes the reverse coefficient uniquely to Σ_{12}/Σ_{22} (proof above), no reverse model with coefficient in $\mathcal{F}$ reproduces the whole family. □

B.5 Proof of Proposition 13 (quasi-likelihood separation)

For a candidate reverse coefficient c, the conditional Gaussian negative log-likelihood of x given y and v with variance function $\sigma^2(v)$ has expectation $\frac{1}{2}\mathbb{E}[\log\sigma^2(v) + \mathbb{E}((x - cy)^2 \mid v)/\sigma^2(v)]$ up to constants; minimizing over σ^2 pointwise gives $\sigma^2 = \mathbb{E}((x - cy)^2 \mid v)$ and the profiled loss $\frac{1}{2}\mathbb{E}\log\mathbb{E}((x - cy)^2 \mid v)$. Relative to the unconstrained optimum $c = c^\star$, the excess is

$$D(c) = \tfrac{1}{2}\,\mathbb{E}\log\frac{\mathbb{E}((x - cy)^2 \mid v)}{\mathbb{E}((x - c^\star y)^2 \mid v)} = \tfrac{1}{2}\,\mathbb{E}\log\left\{1 + \frac{\Sigma_{22}^2}{\det\Sigma}(c - c^\star)^2\right\}$$
$$= \tfrac{1}{2}\,\mathbb{E}\log\left\{1 + \frac{[(b^2 + r)c - b]^2}{r}\right\},$$

using $\mathbb{E}((x - cy)^2 \mid v) = \det\Sigma/\Sigma_{22} + \Sigma_{22}(c - c^\star)^2$ and equation 3. $D \geq 0$ with equality iff $c = c^\star$ a.s. Over the compact candidate set the exponential forms of b, r and c bound the integrand by $K_0 + K_1|v|$, so by dominated convergence and $\mathbb{E}|v| < \infty$, D is continuous in the candidate parameters and its minimum is attained; a zero minimum would force $c = c^\star$ a.s. and by continuity on the support interval, which Theorem 3 rules out for $\delta \neq 2\kappa$. The forward family contains b and attains zero excess. The finiteness of the common baseline loss uses $\mathbb{E}|\log q_x(v)| < \infty$. Because the expected Gaussian loss depends only on conditional second moments, no Gaussianity of the observations is used; for scale mixtures the comparison is of excess cross-entropies, not of KL divergences to a Gaussian model. □

B.6 Proof of Proposition 14 (quartic gap)

Let $\ell_\eta = \log c^\star_\eta$, $z = (1, v)^\top$, $\theta_0 = (\log[b_0/(b_0^2 + r_0)], -\kappa)^\top$, $t = \theta^\top z - \ell_\eta$ and $w_0 = b_0^2/r_0$. Writing $r = r_0 e^{(2\kappa-\eta)v}$ the exact profiled loss of a reverse candidate $c_\theta = e^{\theta^\top z}$ is $D_\eta(\theta) = \frac{1}{2}\mathbb{E}\log\{1 + w_0 e^{\eta v}(e^t - 1)^2\}$. At $\eta = 0$ its unique zero is θ_0 since $\mathrm{Var}(v) > 0$; the Hessian there is $w_0\mathbb{E}zz^\top \succ 0$. Compactness and uniform continuity force every global minimizer to θ_0 as $\eta \to 0$, and the implicit function theorem with local strict convexity gives a unique smooth minimizing branch $\hat\theta(\eta)$ for small $|\eta|$; bounded support justifies differentiation under the expectation. Uniformly on the support, $\ell_\eta(v) = \theta_0^\top z + (1-p)\eta v - \frac{1}{2}p(1-p)\eta^2 v^2 + O(|\eta|^3)$. Expanding the stationarity condition, the first-order term is absorbed by the slope adjustment $(1-p)\eta$ and the second-order term projects $-\frac{1}{2}p(1-p)v^2$ onto $\mathrm{span}\{1, v\}$, so at the minimizer $t = \frac{1}{2}p(1-p)\eta^2(v^2 - \Pi v^2) + O(|\eta|^3)$ and $D_\eta = \frac{w_0}{2}\mathbb{E}[t^2] + O(|\eta|^5)$, which is the displayed expansion. The projection residual $\mathbb{E}[(v^2 - \Pi v^2)^2]$ vanishes iff v is supported on at most two roots of a quadratic; with two support points the affine interpolant of ℓ_η lies in the compact set for small $|\eta|$ and gives exactly zero loss. For conditionally Gaussian errors, the reverse model built from $\hat\theta(\eta)$ with root variance $\mathrm{Var}(y \mid v)$ and residual variance $\mathbb{E}((x - c_{\hat\theta}y)^2 \mid v)$ has KL divergence exactly $D_{\min}(\eta)$ from the forward law; for n i.i.d. samples with $n\eta_n^4 \to 0$ the product-law KL and hence total variation vanish, so the optimal equal-prior test error tends to $1/2$. Explicitly, writing $\hat\theta(\eta) = \theta_0 + u_1\eta + u_2\eta^2 + O(|\eta|^3)$, the order-$\eta$ stationarity condition $\mathbb{E}[z(u_1^\top z - (1-p)v)] = 0$ gives $u_1 = (0, 1-p)^\top$, so $t = O(\eta^2)$ uniformly, and the order-η^2 condition $\mathbb{E}[z(u_2^\top z + \frac{1}{2}p(1-p)v^2)] = 0$ gives $u_2 = -\frac{1}{2}p(1-p)(\mathbb{E}zz^\top)^{-1}\mathbb{E}[zv^2]$. Numerical checks on symmetric, asymmetric and discrete states agree with the expansion (for $v \sim \mathrm{Uniform}[-2, 2]$, $(b_0, r_0) = (0.7, 1)$ the fourth-order coefficient is 0.0042434651; `certificates/quartic_rate_check.py`). □

B.7 Proof of Theorem 1 (true residual mutual information)

Identity. The maps $(\tilde X, \tilde Y) \mapsto (\tilde Y, R_X)$ and $(\tilde X, \tilde Y) \mapsto (\tilde X, R_Y)$ are linear with the same absolute Jacobian $1/\sqrt{1-\rho^2}$, so $H(\tilde Y, R_X) = H(\tilde X, R_Y)$ and $D = I(\tilde Y; R_X) - I(\tilde X; R_Y)$ (Hyvärinen & Smith, 2013, Section 2.4). *Densities.* Conditioned on the three signs, (X, Y) is bivariate Gaussian with covariance components $(a, c, d) = (2^{v+u}, 2^{mv}a, 2^{2mv}a + 2^{v+w})$, all exactly rational; every linear form in equation 2 is therefore a symmetric mixture of eight centred Gaussians with rational variances, standardized by a rational total. *Enclosure.* For such a density f, f decreases on $[0, \infty)$, so on a grid cell the interval hull of the endpoint values of f encloses f on the cell and interval arithmetic on $-p\log p$ times the cell width encloses the cell's contribution even where $-p\log p$ is not monotone. Tails beyond L use $-\log f(x) \le A + x^2/(2v^*)$ with v^* the widest component, $A = \log 8 + \frac{1}{2}\log(2\pi v^*)$, and Mills' bound. With 8192 cells and 128-bit Arb arithmetic this gives the stated enclosures; a 4096-cell run gives the same strict signs (`certificates/entropy/entropy_certificate.py`). A second construction using exact Gaussian cell probability masses times endpoint log-density enclosures (`certificates/mass/mass_certificate.py`) gives $(0.04441, 0.04862)$ and $(-0.09469, -0.09023)$, nested within the first. *Neighbourhood.* Propagating interval coefficients through all covariance, regression and variance calculations with a deliberately loose uniform tail bound encloses D in $(-0.1201, -0.0648)$ for $b(-1) \in [0.4999, 0.5001]$, $b(+1) \in [1.9999, 2.0001]$; radius 0.001 was not certified by these bounds, which is a limitation of the enclosure and not evidence that the sign changes. Small bounded smoothing of the states preserves both strict signs by continuity and dominated convergence; no radius is certified. This is a computer-assisted proof conditional on the code implementing the stated operations and on Arb's validated arithmetic, not a proof-assistant formalization.

B.8 Proof of Corollary 2 (the deployed approximation score)

The pairwise score of `lingam` 1.13.0 (Hyvärinen & Smith, 2013) is

$$\begin{aligned} H(u) &= \tfrac{1}{2}(1 + \log 2\pi) - k_1\big(\mathbb{E}\log\cosh u - \gamma\big)^2 - k_2\big(\mathbb{E}[u\, e^{-u^2/2}]\big)^2, \\ \Delta_{\mathrm{HS}} &= H(\tilde Y) + H(R_X) - H(\tilde X) - H(R_Y), \end{aligned} \tag{9}$$

with $k_1 = 79.047$, $k_2 = 7.4129$, $\gamma = 0.37457$. Each variable here is a symmetric mixture, so the odd expectation vanishes exactly. The certificate follows the construction above

with $F(\sigma) = 2\int_0^\infty \varphi(z)\log\cosh(\sigma z)\,dz$ bracketed by monotone Riemann sums on $[0, 8]$ ($N = 8192$) and the tail bound $\log\cosh(\sigma z) \le \sigma z$, $\int_L^\infty z\varphi = \varphi(L)$, in 128-bit ball arithmetic (`certificates/deployed_score/certificate.py`); an independent reconstruction with Gaussian interval masses gives $(0.067873, 0.088949)$ and $(-0.114588, -0.095049)$ (`certificates/recheck/recheck_certificate.py`). For two variables and no prior knowledge, DirectLiNGAM's candidate objectives are $-\min(0, \Delta_{\mathrm{HS}})^2$ and $-\min(0, -\Delta_{\mathrm{HS}})^2$, so a strictly positive Δ_{HS} selects X first and a strictly negative one selects Y first; there is no intervening optimizer. Under i.i.d. sampling, sample moments and OLS coefficients converge almost surely, $\log\cosh$ is Lipschitz and $ue^{-u^2/2}$ has a bounded derivative, so the sample score converges to its strictly signed limit. This is not a guarantee for floating-point software at extreme inputs, nor a statement about adaptive-lasso edge selection. The installed module hash and the verification that an independent stable implementation matches the package's internal score to 1.6×10^{-15} are in `certificates/entropy/implementation_audit.json`.

B.9 Proof of Theorem 9 (label-conditional mutual information)

Given $L = l$, the eight sign configurations of (V, U_x, U_y) split into the four with $V = l$, each of weight $(1-p)/4$, and the four with $V = -l$, each of weight $p/4$; within-label least-squares coefficients $\beta_l = \mathrm{Cov}(X, Y \mid l)/\mathrm{Var}(X \mid l)$ and $\alpha_l = \mathrm{Cov}(X, Y \mid l)/\mathrm{Var}(Y \mid l)$ are rational in p. The raw triangular maps have unit Jacobian and $\mathrm{Var}(Y \mid l)\mathrm{Var}(R_X \mid l) = \mathrm{Var}(X \mid l)\mathrm{Var}(R_Y \mid l) = \det\mathrm{Cov}((X, Y) \mid l)$, so the standardized entropy difference within each label equals the conditional mutual-information difference, and $D(p)$ averages the two labels. Each standardized variable is again a symmetric finite Gaussian mixture, and the mass construction above (grid $1/1024$, 128-bit precision) gives the stated rational enclosures (`certificates/mass/mass_certificate.json`). The symmetric, nondifferential label mechanism is essential: it is what makes the within-label laws explicit mixtures. At $m = 1$, $\mathrm{Var}(Y \mid V = -1) = 25/32$ and $\mathrm{Var}(Y \mid V = +1) = 25/2$, a sixteen-fold ratio; with $p = 1/20$ the high-state share of the low-label group's outcome second moment is $(25/2)(1/20)/[(25/2)(1/20) + (25/32)(19/20)] = 16/35$.

B.10 Proof of Proposition 12 (kurtosis inflation)

(a) Conditional on the states, $x = \sqrt{W_x}\,z_x$ and $y = \sqrt{W_y}\,\tilde{z}$ with $\tilde{z} = (bs_x z_x + s_y z_y)/\sqrt{W_y} \sim N(0, 1)$ independent of the states (Gaussian innovations). Hence each marginal is a normal scale mixture $\sqrt{W}\,G$ with $G \sim N(0, 1) \perp W$, for which $\mathbb{E}[(\sqrt{W}G)^2] = \mathbb{E}W$ and $\mathbb{E}[(\sqrt{W}G)^4] = 3\mathbb{E}[W^2]$, giving excess kurtosis $3\mathbb{E}[W^2]/\mathbb{E}[W]^2 - 3 = 3\,\mathrm{CV}^2(W)$. This is the $\lambda = 0$ case of the fourth-cumulant formula for normal variance-mean mixtures (Javed, 2026, Theorem 1 with κ_3-, κ_4-terms absent): $K_4 = \kappa_2(W)\,\Gamma(\Omega)$ with $\kappa_2(W) = \mathrm{Var}(W)$. Since $W_x = s_x^2$ does not involve b, $\mathrm{Kurt}(x)$ is free of κ.

(b) At $\kappa = 0$, $W_y = b_0^2 s_x^2 + s_y^2 = b_0^2 X + Y$ where $X = e^{Av + Cu_x}$ and $Y = e^{Av + Cu_y}$ have the same marginal law (same mean m and variance σ^2) and correlation $\rho = \mathrm{Cov}(X, Y)/\sigma^2 \in [0, 1]$. Then

$$\mathrm{CV}^2(b_0^2 X + Y) = \frac{\sigma^2(b_0^4 + 1 + 2b_0^2\rho)}{m^2(b_0^2+1)^2} = \mathrm{CV}^2(X)\cdot\frac{b_0^4 + 1 + 2b_0^2\rho}{(b_0^2+1)^2} \le \mathrm{CV}^2(X) = \mathrm{CV}^2(W_x),$$

with equality iff $\rho = 1$, i.e. iff $C = 0$. Hence $\Delta(0) = 3[\mathrm{CV}^2(W_y) - \mathrm{CV}^2(W_x)] \le 0$ with equality iff $C = 0$.

(c) Let $U_\kappa = b_0^2 e^{2\kappa v} s_x^2 = b_0^2 e^{(2\kappa + A)v + Cu_x}$. Let $R(\kappa) := \mathbb{E}[U_\kappa^2]/\mathbb{E}[U_\kappa]^2 = 1 + \mathrm{CV}^2(U_\kappa)$ and differentiate using $\partial_\kappa U_\kappa = 2vU_\kappa$:

$$\frac{d}{d\kappa}\log R(\kappa) = \frac{4\mathbb{E}[vU_\kappa^2]}{\mathbb{E}[U_\kappa^2]} - \frac{4\mathbb{E}[vU_\kappa]}{\mathbb{E}[U_\kappa]} = 4\big(\mathbb{E}_{U_\kappa^2}[v] - \mathbb{E}_{U_\kappa}[v]\big),$$

where $\mathbb{E}_W[\cdot]$ denotes expectation under the probability measure with density proportional to W. The likelihood ratio between the two tilted measures is $\propto U_\kappa$, which is a strictly increasing function of v for $2\kappa + A > 0$; a measure tilted by an increasing function stochastically dominates, so $\mathbb{E}_{U_\kappa^2}[v] > \mathbb{E}_{U_\kappa}[v]$ unless U_κ is a.s. constant in v. Hence R, and therefore $\mathrm{CV}^2(U_\kappa)$, is strictly increasing in

κ on $\kappa > -A/2$. As $\kappa \to \infty$, $W_y/U_\kappa \to 1$ in the relevant moments (the noise term s_y^2 is κ-free) while $\mathrm{CV}^2(U_\kappa) \to \infty$ (for lognormal U_κ, $\mathrm{CV}^2 = e^{(2\kappa+A)^2+C^2} - 1$), so $\Delta(\kappa) \to \infty$.

(d) With $v, u_x, u_y \sim N(0,1)$ independent and $\mathbb{E}e^{a\xi} = e^{a^2/2}$ for $\xi \sim N(0,1)$: $\mathbb{E}[W_x] = e^{(A^2+C^2)/2}$, $\mathbb{E}[W_x^2] = e^{2(A^2+C^2)}$, giving $\mathrm{Kurt}(x) = 3(e^{A^2+C^2} - 1)$. For the child, $\mathrm{Var}(y) = \mathbb{E}[W_y] = b_0^2\mathbb{E}e^{(g+A)v+Cu_x} + \mathbb{E}e^{Av+Cu_y} = b_0^2 P_1 + m_1$, and $\mathbb{E}[y^4] = 3\mathbb{E}[W_y^2] = 3\mathbb{E}[b_0^4 e^{2(g+A)v+2Cu_x} + 2b_0^2 e^{(g+2A)v+Cu_x+Cu_y} + e^{2Av+2Cu_y}] = 3b_0^4 Q_1 + 6b_0^2 Q_2 + 3m_2$ with the stated exponents. Δ follows. Monotonicity of the full $\Delta(\kappa)$ (not only of its parent-channel component) was verified numerically on the grid $s_x/s_y \in [0.3, 3]$, $\kappa \in [0, 1]$; the root $\kappa^\star$ is computed by bisection (Table 4). □

Non-Gaussian innovations . With $\mathbb{E}z^4 = 3 + \kappa_z$, $\mathbb{E}[x^4 \mid \text{states}] = (3+\kappa_z)s_x^4$ and $\mathbb{E}[y^4 \mid \text{states}] = (3+\kappa_z)(b^4 s_x^4 + s_y^4) + 6b^2 s_x^2 s_y^2 = 3W_y^2 + \kappa_z(b^4 s_x^4 + s_y^4)$, since $(b^2 s_x^2 + s_y^2)^2 = b^4 s_x^4 + 2b^2 s_x^2 s_y^2 + s_y^4$. The parent's fourth moment is scaled by the full factor $(1 + \kappa_z/3)$ while the child's cross term is not, which lowers the child's relative kurtosis and moves $\kappa^\star$ to the right.

B.11 DAG PATH ALGEBRA, THE CURVATURE BOUND, AND WHY NOTHING MORE IS CLAIMED

Write $X_t = A_t\varepsilon_t$ with $A_t = (I - B_t)^{-1} = \sum_{k\geq 0} B_t^k$, which is a finite sum because the coefficient matrix of a DAG is nilpotent (its spectral radius is zero; the rejection-sampling on spectral radius in the generator therefore never binds, and large path products can still inflate variance). Since $B_t = B \odot e^{Kv_t}$, $[B_t^k]_{ij}$ is a sum over directed paths of length k from j to i of $\prod_{(a,b)\in p} B_{ab} e^{K_{ab}v} = c_p e^{S_p v}$, so $A_{ij}(v) = \sum_{p:j\to i} c_p e^{S_p v}$. If all paths between a pair share $S_p \equiv S$ and $\sum_p c_p \neq 0$, this is $(\sum_p c_p)e^{Sv} \in \mathcal{F}$. For same-sign c_p, $\log|A_{ij}(v)| = \text{const} + \log\sum_p |c_p| e^{S_p v}$ and, as in the proof of Theorem 3, $\frac{d^2}{dv^2}\log|A_{ij}(v)| = \mathrm{Var}_{w(v)}(S_p)$ with $w_p(v) \propto |c_p|e^{S_p v}$, at most $(S_{\max} - S_{\min})^2/4$ (Popoviciu). With mixed signs, near-cancellation can invalidate the bound for the signed sum. Removing a source r by $X_j' := X_j - A_{jr}(v)X_r$ leaves $\sum_{k\neq r} A_{jk}(v)\varepsilon_k$, the reduced form of the modulated model on $G \setminus \{r\}$ with the same innovations. That is all that is proved. It is not a recovery guarantee, for two reasons. First, the marginal of a source r and a descendant j has residual $\eta_j = \sum_{k\neq r} A_{jk}(v)\varepsilon_k$ whose conditional variance has its own state profile, so the marginal pair has its own noise-ratio slope δ_{jr}, and Theorem 3 separates that marginal only if $\delta_{jr} \neq 2S_{jr}$; nothing in the DAG structure guarantees this. Second, a greedy ordering needs its score argmax to be a source at every step, a substantive assumption that held on 97%/90%/71% of greedy steps at $\kappa = 0.15/0.30/0.45$ in our measurements (Appendix U) and is not implied by the algebra. Finally, the earlier claim that an alternative-order Cholesky factorization of $\Sigma(v)$ yields a reversed structure on the same skeleton is false: for independent unit-variance X, Z, ε and $Y = X + Z + \varepsilon$, the collider $X \to Y \leftarrow Z$ reordered with Y first gives $\mathbb{E}[X \mid Y, Z] = \frac{1}{2}Y - \frac{1}{2}Z$, a nonzero X–Z edge that the original graph does not have. A fork and a collider imply different conditional independences and cannot share second-moment representations without additional edges.

C THRESHOLDS AND COMPARATIVE STATICS

Table 4: Kurtosis-crossing threshold $\kappa^\star$ (root of Δ; closed form in the proof of Proposition 12). Baseline $(A, C, b_0) = (1, 0.5, 0.7)$.

calibration	change	$\kappa^\star$
baseline	·	0.071
stronger common-vol share	$A = 1.6$	0.045
larger coefficient	$b_0 = 1.0$	0.052
more idiosyncratic vol	$C = 0.95$	0.168

Part (a) of Proposition 12 is an instance of the fourth cumulant of normal variance-mean mixtures with zero mean shift (Barndorff-Nielsen et al., 1982; Javed, 2026). The exact population

values of the kurtosis gap from the closed form are $\Delta = -1.02, 2.00, 12.70, 59.07, 288.3$ at $\kappa = 0, .15, .3, .45, .6$; a 4×10^5 Monte-Carlo estimate ($-0.3/2.1/8.8/28.8/85.3$) underestimates the tail by up to $3\times$ at $\kappa = 0.6$, as expected when fourth moments of heavy-tailed variables are estimated from a persistent state. The crossing arrives earlier when volatility is more strongly shared and the link is stronger (the crisis regime) and later with more idiosyncratic volatility. Heavier innovations delay it: with $\mathbb{E}z^4 = 3+\kappa_z$, $\mathbb{E}[x^4 \mid \text{states}] = (3+\kappa_z)s_x^4$ while $\mathbb{E}[y^4 \mid \text{states}] = 3W_y^2+\kappa_z(b^4 s_x^4+s_y^4)$, so the innovation kurtosis scales the parent's fourth moment fully but leaves the child's cross term $6b^2 s_x^2 s_y^2$ untouched, lowering the child's relative inflation. This is why the $t(4.5)$ cells of Table 13 cross later. The population score equation 9 (Appendix B) on the Gaussian-state benchmark, integrated by Gauss–Hermite quadrature over (v, u_x, u_y) and the Gaussian innovation with $\log\cosh$ evaluated as $\mathrm{logaddexp}(a, -a) - \log 2$ (the earlier Monte-Carlo evaluation, six seeds of $n = 6 \times 10^5$, gave a crossing at 0.236; its justification "because the functionals are bounded" was wrong, since $\log\cosh$ is integrable here but unbounded): Code `core/generator/checks.py`,

Table 5: Quadrature refinement of the population score at $(A, C, b_0) = (1, 0.5, 0.7)$. The root stabilizes near 0.238; this is numerical evidence of a crossing, not a certified threshold or a monotonicity theorem.

nodes per coordinate	estimated crossing	$\Delta_{\mathrm{HS}}(0)$	$\Delta_{\mathrm{HS}}(0.3)$	$\Delta_{\mathrm{HS}}(0.6)$
20	0.237893	0.065887	−0.056059	−0.493805
32	0.237857	0.066259	−0.056428	−0.496419
48	0.237799	0.066362	−0.056597	−0.497531
64	0.237771	0.066389	−0.056661	−0.497920

data `core/generator/results/quadrature.csv`.

D Independent numerical cross-check of the certified quantities

We re-implemented the matched example from its statement alone, without importing the certificate code: the same eight-component Gaussian-mixture densities, standardized, with each entropy computed by adaptive quadrature (`scipy.integrate.quad`, relative tolerance 10^{-12}, integration split at multiples of the widest component's standard deviation). This is numerical evidence, not a certificate; its value is that a second code path lands inside every enclosure. Sweeping the coefficient exponent m continuously, the pooled difference is monotone on the grid and crosses zero at $m^* = 0.5031$, i.e. a coefficient ratio $4^{m^*} \approx 2.0$ between the two states. With exact labels the conditional difference stays positive for every $m \in [0, 1]$ (from 0.0275 at $m = 0$ to 0.0199 at $m = 1$), so exact labels always repair this example; the label-error crossing is $p^* = 3.43\%$ at $m = 1$ and grows as m decreases: $m = 0.52$: 24.4%, $m = 0.60$: 12.2%, $m = 0.68$: 8.3%, $m = 0.76$: 6.3%, $m = 0.84$: 5.0%, $m = 0.92$: 4.1%, $m = 1.00$: 3.4%. These crossings are numerical, not certified; they show that the 5% in Theorem 9 is not a universal threshold.

Table 6: Our quadrature versus the two certified enclosures (`crosscheck/E3_true_mi_crosscheck.py`).

quantity	quadrature (ours)	density-height certificate	cell-mass certificate
pooled m=0	+0.04651	(0.03548, 0.05754)	(0.04441, 0.04862)
pooled m=1	-0.09246	(-0.10374, -0.08117)	(-0.09469, -0.09023)
conditional m=1, p=0	+0.01988	–	(0.01808, 0.02167)
conditional m=1, p=0.01	+0.01519	–	(0.01336, 0.01702)
conditional m=1, p=0.05	-0.01093	–	(-0.01286, -0.00901)
conditional m=1, p=0.10	-0.04278	–	(-0.04478, -0.04077)
conditional m=0, p=0.05	+0.03208	–	(0.03024, 0.03392)

Table 7: Full positive-modulation ordering results, mean ± standard deviation over 30 seeds. The selector combines tercile curvature with HetCD-gated; its selection rate is shown separately. Sign controls are in Table 9. Pre-registration covers only the methods specified below; later additions are exploratory.

order accuracy	$\kappa = 0$	.15	.30	.60
pooled (no state)				
DirectLiNGAM	**.98**±.06	.67±.16	.52±.10	.32±.12
HetSEM-const	.94±.06	.71±.17	.49±.15	.21±.12
DirectLiMIAM	.69±.12	.69±.14	.83±.14	.87±.08
location-scale score	.57±.14	.72±.13	.85±.14	**.98**±.05
measured state				
HetCD-Ω, fixed windows	.56±.18	.58±.16	.68±.17	**.90**±.12
HetCD-Ω, tercile bins	.60±.12	.77±.14	**.91**±.10	.95±.11
HetCD-gated	.89±.08	.86±.10	.77±.15	.61±.19
Permutation selector	.89±.08	**.87**±.10	.85±.16	.95±.11
Curvature selected	.03	.13	.63	1.00
estimated state (lagged PC1 of log-EWMA variances)				
HetCD-Ω (windows)	.52±.20	.49±.19	.54±.22	.53±.23
HetCD-Ω, tercile bins	.53±.17	.66±.21	.73±.22	.85±.14
controls that use the state and no graph				
Slope sort (measured / proxy)	.50 / .49	.73 / .73	.85 / .84	.93 / .94

E BENCHMARK GENERATOR

Common state: $v_t = \phi v_{t-1} + \sqrt{1-\phi^2}\,\xi_t$, $\xi_t \sim N(0,1)$, $\phi = 0.97$, $v_0 \sim N(0,1)$. Idiosyncratic states $g_{i,t}$: independent AR(1) processes with the same law. Noise variances follow a multivariate stochastic-volatility specification (Harvey et al., 1994), $H_{i,t} = \exp(\sigma_c v_t + \sigma_{\mathrm{id}} g_{i,t})$ with $\sigma_c = A = 1$, $\sigma_{\mathrm{id}} = C = 0.5$; innovations $\varepsilon_{i,t} = \sqrt{H_{i,t}}\, z_{i,t}$ with z Gaussian or $t(4.5)/\sqrt{4.5/2.5}$. Structure: a random DAG on d nodes with expected in-degree 1.5 and weights B (magnitude 0.7 at the baseline for $d = 2$; drawn as in `make_dag` for $d \geq 10$); $B_t = B \odot e^{\kappa v_t}$, with the draw rejected if $\max_t \rho(B_t) \geq 1$. Data $X_t = (I - B_t)^{-1}\varepsilon_t$, $n = 3000$. The multivariate model is the DAG analogue of equation 1–equation 3.

Seed conventions, stated once: Design 1 (Appendix J) uses generator seeds 0–19; the 40-seed head-to-head uses seed indices 0–39 passed to `simulate(seed)`, which adds an offset of 1000 (the earlier "seeds 1000–1009" wording referred to the first ten of these); the temporal-adapter sweep of Appendix J uses seed indices 0–9 of the same generator at $d = 10$, $n = 3000$, $\kappa \in \{0, .2, .35, .5\}$.

F HETEROGENEOUS-NOISE GENERATOR: PRE-REGISTRATION AND FULL RESULTS

Pre-registration (frozen 25 September 2026). The generator of Section 5 adds to Appendix E per-node noise levels ℓ_i, $\log \ell_i \sim U(\log \frac{1}{4}, \log 4)$, and per-node noise slopes $a_i \sim U(.5, 1.5)$, drawn once per seed after the DAG. Calibration seeds 98000–98029 at $\kappa = 0$ (for the calibrated selector rule); evaluation seeds 99000–99029 at $\kappa \in \{0, .15, .3, .6\}$, paired across κ and methods. Four questions were fixed: whether the non-Gaussian reversal persists (DirectLiNGAM below .5 at $\kappa = .6$); whether the single-state precision diagonal loses its equal-variance advantage; whether HetCD-Ω keeps its strong-modulation accuracy ($\geq .9$ at $\kappa = .6$) and the selector tracks the better component; and how accuracy depends on the edge imbalance $|\eta_{ij}|$, which is saved per edge. All four outcomes are in Section 5. Failed fits were to be retained as rows with missing metrics; there were none. The pre-registration covers this run only (DirectLiNGAM, DirectLiMIAM, HetCD-gated, the tercile-bin curvature order, the min-precision control, the selector rules, all with the three-bin group-BIC pruner). The other studies in this appendix—the additional pooled scores, the location-scale score, the slope-sorting control, the equal-width windows and their setting, the negative-κ and

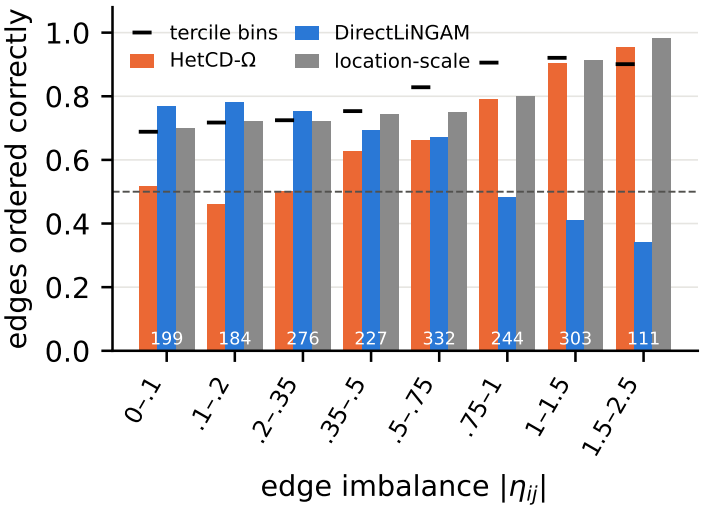

Figure 2: Edge-level accuracy against imbalance magnitude on the heterogeneous-noise benchmark, pooling the four nonnegative modulation levels. Counts appear at the base; small ticks show the tercile-bin estimator. These are empirical edge-level summaries, not a guarantee for individual edges.

mixed-sign generators, the threshold pruner—was added afterwards, in some cases after seeing the pre-registered results, and is exploratory.

Additional pooled scores. ICA-LiNGAM, VARLiNGAM (one lag, instantaneous matrix) and HetSEM-const were run on the same seeds after the main study (`benchmark/T6_out/T6d_field_sweep.csv`): ICA-LiNGAM .67±.20, .42±.16, .39±.18, .24±.16 and VARLiNGAM .97±.07, .67±.16, .52±.10, .32±.13 at $\kappa = 0, .15, .3, .6$ (HetSEM-const in Table 1). HetSEM-const fits a pooled OLS coefficient per pair, scores a heteroskedastic-SEM likelihood with per-observation EWMA variances in each direction, and removes the best source greedily; it is the constant-coefficient version of HetCD-gated and the closest thing to a Rigobon-type per-observation read of the second-moment channel. The location-scale score, added after review, is our implementation of the RESQUE-style direction score of the equal-variance study (Appendix J): inputs standardized; for a candidate cause c and remaining variable j, residualize j within eight quantile bins of c (bin mean and standard deviation), and score $|\mathrm{corr}(r, c)| + |\mathrm{corr}(r^2, c)|$; remove greedily the variable with the smallest summed score over the remaining set, regressing it out of the rest. Order accuracy .57±.14, .72±.13, .85±.14, .98±.05 at $\kappa = 0, .15, .3, .6$ (seeds 99000–99029; `benchmark/T6_out/T6e_lsnm.csv`). Its accuracy on the equal-variance generator was .87/.93/.97 at $\kappa = .15/.3/.6$. What it reads is examined below.

Edge-level accuracy by imbalance. Figure 2 pools the 1,876 true edges across the four $\kappa \geq 0$ levels. Bin edges $\{0, .1, .2, .35, .5, .75, 1, 1.5, 2.5\}$; counts 199, 184, 276, 227, 332, 244, 303, 111. HetCD-Ω (windows): .52, .46, .50, .63, .66, .79, .90, .96. Tercile bins: .69, .72, .73, .75, .83, .91, .92, .90. DirectLiNGAM: .77, .78, .75, .69, .67, .48, .41, .34. Location-scale: .70, .72, .72, .74, .75, .80, .91, .98. On this generator the window estimator's accuracy rises with $|\eta|$ from chance in the smallest bin; these are nearly balanced, not exactly balanced, edges, and the graph guarantee neither covers nor excludes recovering an individual weakly separated edge from the rest of the graph. The tercile bins are above chance in the same bins, which the slope-sorting cue described below accounts for (`benchmark/T6_out/T6_edge_accuracy_vs_eta.csv`).

Standardization control. The min-precision-diagonal recursion at the middle state bin gives .76/.76/.76/.79 on raw inputs and .42/.27/.17/.03 on per-variable standardized inputs (measured state), and .77/.79/.84/.96 raw and .44/.30/.19/.04 standardized with the estimated state. HetCD-Ω gives identical orders on raw and standardized inputs in every cell.

Graph pruners. Three pruners were applied to the tercile-bin HetCD-Ω order (the window order with the threshold pruner is in Table 2 and `benchmark/T6_out/T6i_windows_bench.csv`): the theorem's normalized precision-row threshold at $\tau = m^{-1/4}$ (Table 2); the common three-bin group-BIC pruner used in the pre-registered comparison; and the profiled log-affine pruner of the equal-variance study, which fits varying parent coefficients and a log-linear residual variance by bounded multistart optimization (`benchmark/T6b_standardized_and_logaffine.py`; slow, and not rerun on the final seeds). The threshold estimator returns 16.0 edges on average at $\kappa = .6$ against 15.6 true (mean over the 120 evaluation graphs) and 29.6 for the BIC pruner. Both partitions estimate window-averaged covariances, which differ from the pointwise covariances of Theorem 5 even in population; the window, tercile-bin and exact-level versions are reported under separate names throughout.

Windows against tercile bins. Added after the pre-registered run; exploratory. Two distortions separate a window-averaged covariance from the pointwise covariance of Theorem 5. First, for a variable whose conditional variance is $\ell_j e^{a_j v}$, the variance pooled over a window W is $\ell_j \mathbb{E}[e^{a_j v} \mid v \in W] \approx \ell_j e^{a_j \bar{v}_W}(1 + a_j^2 \operatorname{Var}(v \mid W)/2)$, so the pooled log variance exceeds the pointwise value by about $a_j^2 \operatorname{Var}(v \mid W)/2$; with quantile terciles of a near-Gaussian state the tail terciles have four to five times the conditional state variance of the middle one, so for a variable whose marginal variance is steep in the state the term enters $\widehat{J}_j$ with weights $(1 - t, t, -1)$ and lowers its gap, which pushes it toward the sink position whether or not it is one. Windows of equal width make this first term closer to constant across windows (equal only to first order and only for a locally flat state density), and this is the intuition behind the choice. Second, and not removed by any choice of widths, the precision of an averaged covariance is not the average precision: in the two-variable model, $(\Sigma_W^{-1})_{YY} = [\mathbb{E}_W q_Y + \mathbb{E}_W(b^2 q_X) - (\mathbb{E}_W b q_X)^2 / \mathbb{E}_W q_X]^{-1}$, and the last two terms equal $\mathbb{E}_W q_X$ times the q_X-weighted within-window variance of b, which is positive whenever the coefficient moves inside the window. A true sink's pooled precision therefore carries coefficient variation as well as its noise law. The equal-width windows are thus an empirically investigated way of reducing some of the binning distortion, not a derivation that the sorting cue is removed; the sign-reversed and mixed-sign controls in Table 9 are the evidence for their robustness. Table 8 shows the two partitions in the two-variable model $Y = be^{\kappa v} X + \varepsilon_Y$ with noise slopes (a_X, a_Y), $\sigma_c = 1$, $b = .8$, $n = 3000$, 30 seeds: tercile bins are at 1.00 or 0.00 on the balanced cells, depending only on which variable is steeper, and at .50 on an unbalanced edge whose cause is the steeper variable; the equal-width windows are at .33–.43 on the balanced cells (30 seeds each, so not a demonstration of exact chance), .87–.97 on the four cells with $|\eta| = 1.2$, and .53 on the $\eta = .6$ cell. The window setting (centres at the .15/.5/.85 quantiles, half-width .5 in the units of the supplied state, which is the population standard deviation for the simulated unit-variance state and the sample standard deviation for the standardized proxies and real states; about 660 observations per window at $n = 3000$) was selected among nine settings on the evaluation seeds (`benchmark/T6i_window_sweep.py`, `benchmark/T6_out/T6i_window_sweep.txt`): order accuracy at $\kappa = .6$ ranged from .78 to .90 across settings and no setting reached the tercile-bin value, at $n = 3000$ or at $n = 4500$, where the windows hold the same 1000 observations as a tercile (`benchmark/T6i_samplesize_check.py`: windows .90, bins .95 at $\kappa = .6$; windows .91, bins .83 at $\kappa = -.6$).

Table 8: Two-variable cells (LiNGAM denotes DirectLiNGAM; fraction of 30 seeds with X placed first). $\eta = 2\kappa - (a_Y - a_X)$.

cell	(κ, a_X, a_Y)	LiNGAM	terciles	windows	loc.-scale
modulated, balanced ($\eta = 0$)	$(.6, 1, 2.2)$	.40	1.00	.37	.73
modulated, half-balanced ($\eta = .6$)	$(.6, 1, 1.6)$	.00	1.00	.53	1.00
modulated, unbalanced ($\eta = 1.2$)	$(.6, 1, 1)$	.00	1.00	.97	1.00
negative, balanced ($\eta = 0$)	$(-.6, 2.2, 1)$	.87	.00	.43	.00
negative, unbalanced ($\eta = -1.2$)	$(-.6, 1, 1)$	.03	.87	.97	.33
unmodulated, balanced	$(0, 1, 1)$	.97	.47	.40	.13
unmodulated, balanced, $b = .3$	$(0, 1, 1)$	1.00	.43	.33	.10
unmodulated, effect steeper ($\eta = -1.2$)	$(0, 1, 2.2)$	1.00	1.00	.87	.27
unmodulated, cause steeper ($\eta = 1.2$)	$(0, 2.2, 1)$	1.00	.50	.97	.00

Slope-sorting control. For each variable regress $\log X_j^2$ on the (measured or estimated) state and order by the slope, flattest first (`benchmark/T6k_slope_sort.py`). One-signed generator, measured state: .07, .18, .26, .50, .73, .85, .93 at $\kappa = -.6, -.3, -.15, 0, .15, .3, .6$; estimated state: .43, .17, .26, .49, .73, .84, .94. Ordering by the true noise slopes a_j gives .51 at every κ: the cue is the modulation-induced steepening of descendants' variances, not the noise slopes. On the mixed-sign generator (`benchmark/core_v6.simulate_hetero_mixed`, independent signs $\kappa_{ij} = \pm\kappa$, everything else as pre-registered) the control is at .50–.54 with either state.

Sign of the modulation and the mixed-sign generator: full results. Table 9 gives the methods of Table 7 except the selector on the one-signed generator at $\kappa < 0$ (`benchmark/T6_out/T6n_negative_full.csv`, `T6j`, `T6i`) and on the mixed-sign generator at $\kappa > 0$ (`benchmark/T6_out/T6m_mixed_sign_full.csv`). Mean $|\eta_{ij}|$ on the mixed-sign generator is .40, .60 and 1.18 at $\kappa = .15, .3, .6$.

Table 9: Negative and mixed-sign modulation, all methods, order accuracy (30 seeds; SHD of the threshold pruner in brackets). Left block: one-signed generator, $\kappa < 0$. Right block: mixed-sign generator, $\kappa_{ij} = \pm\kappa$.

	$\kappa = -.6$	$-.3$	± 0	$\pm .15$	$\pm .3$	$\pm .6$
slope-sort (control)	.07	.18	.50	.53	.54	.53
DirectLiNGAM	.54	.62	.98	.82	.63	.43
ICA-LiNGAM	.39	.47	.67	.51	.46	.37
VARLiNGAM	.53	.62	.97	.83	.65	.46
HetSEM-const	.42	.66	.94	.85	.61	.37
DirectLiMIAM	.70	.39	.69	.63	.67	.79
location-scale score	.82	.51	.57	.62	.70	.79
HetCD-gated, measured	.62	.71	.89	.87	.78	.69
Windows, measured	.90 [4.5]	.68 [10.7]	.56 [14.0]	.58 [12.6]	.69 [9.9]	.88 [4.3]
Terciles, measured	.80 [8.0]	.59 [14.5]	.60 [13.3]	.63 [12.6]	.76 [8.9]	.86 [5.8]
HetCD-gated, proxy	.54	.80	.89	.92	.85	.56
Windows, proxy	.59 [12.8]	.57 [14.1]	.52 [15.3]	.51 [14.8]	.57 [13.1]	.67 [13.6]
Terciles, proxy	.65 [13.8]	.55 [14.9]	.53 [14.5]	.61 [13.0]	.72 [9.7]	.79 [10.9]
Slope sort, proxy	.42	.17	.49	.53	.55	.57

The location-scale score: what it reads. On conditionally Gaussian data drawn at three exact state levels from the population covariances of the Gaussian control (Appendix G; $d = 10$, $n = 12{,}000$, 30 seeds; `benchmark/T6_out/T6g_ls_gaussian.csv`) the score reaches .57, .65, .73 at $\kappa = .15, .3, .6$ against .66, .77, .97 for HetCD-Ω at the same exact levels. Pooled, these data are a three-component Gaussian mixture, so they are not free of non-Gaussian information (a coordinate whose conditional variance varies across the three states has positive pooled excess kurtosis); what this cell removes is the continuous scale mixture of the benchmark, and the score has little to read without it. On the benchmark its .98 comes with one-signed modulation and falls to .82 at $\kappa = -.6$ and .79 on the mixed-sign generator. In the two-variable cells (Table 8) it is reversed without modulation (.10–.27 with the true arrow) and un-reversed by positive modulation: its failure is the mirror image of LiNGAM's. Its order handed to the theorem's threshold pruner with the measured state gives SHD 2.1 and 12 exact graphs at $\kappa = .6$ on the one-signed generator (`benchmark/T6_out/T6f_ls_graph.csv`); to the pooled BIC pruner, SHD 20.1. Edge-level accuracy by $|\eta|$ is in Figure 2.

G Gaussian controls and the matched observed-state study

Gaussian controls. To test the covariance mechanism apart from non-Gaussianity and equal variances, a separate study generates conditionally Gaussian data at three exact states $\{-1, 0, 1\}$ with unequal noise levels (log-uniform on $[.2, 5]$), unequal slopes ($a_i \sim U(.7, 1.3)$), and a curved common factor $g(v) = e^{.4v^2}$; sample covariances are drawn as exact Wishart sufficient statistics. Table 10: at $d = 10$, $n = 12{,}000$, $\kappa = .6$ the signed-curvature order reaches .97 and the theorem's

threshold estimator SHD 1.1 with 21/30 exact graphs; the unsigned (absolute) curvature criterion, which does not cancel the common factor, is at .32, and the single-state minimum diagonal at .80. Weak modulation remains hard: .63 order accuracy at $\kappa = .15$. A perturbation audit of 100 graph models and 641 ancestral submatrices found every case satisfying the sufficient conditions of Theorem 6 recovered.

Table 10: Gaussian controls, three exact states, heterogeneous noise levels and slopes, curved common factor, 30 seeds; order accuracy and SHD.

	$\kappa = .15$		.30		.60	
	order	SHD	order	SHD	order	SHD
$d = 10$, $n = 12{,}000$, *three exact states, heterogeneous noise slopes, curved common g*						
HetCD-Ω, theorem threshold	0.63	11.9	0.77	7.5	0.97	1.1
HetCD-Ω, group-BIC pruner	0.63	14.6	0.77	9.2	0.97	1.6
min precision diagonal (one state)	0.80	6.6	0.80	6.6	0.80	6.6
unsigned curvature (ablation)	0.37	23.1	0.36	24.8	0.32	27.3
$d = 20$, $n = 12{,}000$, *three exact states, heterogeneous noise slopes, curved common g*						
HetCD-Ω, theorem threshold	0.62	31.3	0.74	20.2	0.96	2.8
HetCD-Ω, group-BIC pruner	0.62	45.1	0.74	29.5	0.96	4.5
min precision diagonal (one state)	0.75	23.2	0.75	23.1	0.75	23.4
unsigned curvature (ablation)	0.40	72.5	0.34	81.2	0.28	92.0

Matched observed-state study ($d = 5$). The study that feeds the CHOD comparison (Appendix I) has 120 Gaussian datasets, $d = 5$, $n = 1500$, $\kappa \in \{0, .3, .6\}$, homogeneous ($a_i \equiv 1$) or heterogeneous ($a_i \sim U(.7, 1.3)$) slopes, three exact states or a continuous uniform state, and ten paired seeds per cell; every method receives X and the exactly observed state; inputs are scaled by pooled root mean square without centering. At $\kappa = .6$ with heterogeneous slopes and three exact states the threshold estimator has mean SHD .8 and 7/10 exact graphs; with a continuous state the shrinking-window version has SHD 4.1 and no exact graphs, and the fixed-bin group-BIC procedure SHD 2.9. Population scalar normalization of the observations (dividing by $\det \Sigma(v)^{1/(2d)}$) leaves the signed-gap minimizers and the normalized precision entries unchanged, which is why the determinant-normalized rows in the archive differ from the raw ones only through finite-sample cross-fitting.

H THE SELECTOR: PRE-REGISTRATION, RULES AND OUTCOMES

Statistic. For standardized X and a supplied state s: three quantile bins of s; within-bin covariances $\widehat{\Sigma}_k$ and bin-mean states $v_1 < v_2 < v_3$; Jensen gaps J_j as in equation 7 over all d variables at the first recursion step; spread $S = \max_j J_j - \min_j J_j$. For pointwise conditional covariances S is zero when every edge is balanced and equals the largest non-sink excess otherwise; with three fixed bins the identity is approximate, and the $\kappa = 0$ false-selection rates below are the empirical check. The spread rather than a sink margin is used because two true sinks legitimately tie.

Rules (fixed before evaluation). *Permutation score:* permute s across observations 200 times, recompute S, select the curvature order when fewer than 5% of permuted values exceed the observed one. *Calibrated ratio:* within-bin observation bootstrap (200 replicates, bins and anchors fixed) gives a dispersion $\widehat{\mathrm{se}}(S)$; select when $S/\widehat{\mathrm{se}}(S)$ exceeds the 95th percentile of the same ratio on 30 separate $\kappa = 0$ seeds. Each rule is paired with two fallbacks, HetCD-gated and pooled DirectLiNGAM, giving four named methods. Two acceptance criteria were fixed: (1) at least one selector reaches mean order accuracy $\geq .95$ at every κ with the exact state; (2) that selector is within .03 of the better of its two components at every κ.

What the rules are not. After review we describe both as empirical selection rules, not tests. Permuting the state destroys the dependence between X and v that exists even at $\kappa = 0$ (through the common volatility and the AR state), so the permuted values are not a null distribution and the 5% level is not a size; when the state is a function of X (the estimated proxy) permutation has no null interpretation at all. The bootstrap quantity holds the partition fixed and resamples independently

within bins, so it is a conditional dispersion, not a standard error of the unconditional statistic. Reporting the first of four pre-specified rules in the main text is a choice made after seeing all four; all four are below.

Table 11: All selector variants, heterogeneous-noise generator, order accuracy (selection rate of the curvature order in parentheses).

	$\kappa = 0$	.15	.30	.60
measured state				
permutation score / HetCD-gated	.89 (.03)	.87 (.13)	.85 (.63)	.95 (1.0)
permutation score / DirectLiNGAM	.97 (.03)	.70 (.13)	.77 (.63)	.95 (1.0)
calibrated ratio / HetCD-gated	.86 (.13)	.85 (.30)	.88 (.80)	.95 (1.0)
calibrated ratio / DirectLiNGAM	.94 (.13)	.72 (.30)	.84 (.80)	.95 (1.0)
best of two per seed (HetCD-Ω, HetCD-gated)	.89	.89	.93	.95
estimated state				
permutation score / HetCD-gated	.86 (.07)	.91 (.03)	.78 (.30)	.59 (.30)
permutation score / DirectLiNGAM	.94 (.07)	.67 (.03)	.58 (.30)	.46 (.30)
calibrated ratio / HetCD-gated	.89 (.03)	.92 (.03)	.84 (.23)	.69 (.57)
calibrated ratio / DirectLiNGAM	.98 (.03)	.68 (.03)	.60 (.23)	.63 (.57)
best of two per seed	.89	.92	.90	.85

Outcomes. Criterion (1) failed: no single selector meets the .95 criterion at every κ on this generator (the permutation rule with the DirectLiNGAM fallback reaches .97 at $\kappa = 0$ but .70 at .15), because no component pair does. Criterion (2) held for the permutation-score/HetCD-gated rule at $\kappa \in \{0, .15, .6\}$ and failed at $\kappa = .3$ (.85 against .93). On the equal-variance generator (Appendix J) the same rule scored .99/.96/.92/1.00 and met criterion (2) everywhere. The $\kappa = 0$ false-selection rate of the permutation rule is 3% with the measured state on both generators; with the estimated state it is 7% here and 20% on the equal-variance generator. Paired seed-bootstrap intervals for every selector-minus-component difference are in the archive.

I CHOD: MODEL CLASSES AND AN EXECUTED COMPARISON THAT DID NOT CONVERGE

Determinant normalization. For a DAG, $\det(I - B(v)) = 1$, so $\det \Sigma(v) = \det Q(v) = g(v)^d \prod_i d_i \, e^{(\sum_i a_i)v}$. Dividing every observation by $\det \Sigma(v)^{1/(2d)}$, which divides the covariance by $\det \Sigma(v)^{1/d}$, leaves $B(v)$ unchanged and turns the noise covariance into $\mathrm{diag}[d_i(\prod_\ell d_\ell)^{-1/d} e^{(a_i - \bar{a})v}]$. An arbitrary common factor can therefore be removed without knowing the graph, and if every a_i is equal the normalized noise covariance is constant in v: the equal-slope case of Assumption 2 lies inside CHOD's constant-noise class (Zhou et al., 2022). "We allow an arbitrary common volatility" is therefore not by itself a distinction from CHOD. The identity also yields an observable common baseline for the Jensen gaps, $J_g = -d^{-1} J[\log \det \Sigma]$, so that $E_j = J[\log \Omega_{jj}] + d^{-1} J[\log \det \Sigma]$ is zero at sinks and positive at separated non-sinks.

Unequal slopes lie outside. For the chain $X_1 \to X_2 \to X_3$ with $b_{21} = b_{32} = .7e^{.2v}$ and $Q(v) = g(v)\,\mathrm{diag}(1, e^v, e^{2v})$, both imbalances are $-.6$ and Theorem 5 identifies the graph. After determinant normalization the three marginal variances are e^{-v}, $.49e^{-.6v} + 1$ and $.2401e^{-.2v} + .49e^{.4v} + e^v$; all three change between $v = 0$ and $v = 1$ (by $-.632$, $-.221$, $+1.916$, interval-certified). Every constant-noise DAG has a root whose marginal variance is constant, even with correlated constant errors, so no such DAG represents this family; and any common scalar normalization producing constant errors would have to make the determinant constant, hence coincide with determinant normalization up to a constant. This excludes constant-noise DAG representations; CHOD also allows confounding and some cycles, so the two classes overlap without nesting, and the example is a distinction, not a refutation.

Execution. The authors' R/C++ implementation (repository commit `0635917`) was run on the matched $d = 5$ study above under R 4.3.3. A source audit found that the Gaussian quadratic used

in the edge updates applies covariance eigenvectors in an inconsistent orientation; in compiled form it disagrees with a direct matrix inverse by up to 34.3% relative error on random fixtures, and a three-line correction agrees to 2.6×10^{-15}. The default spline basis omits its first function and is identically zero at the left boundary, forcing every coefficient to zero there; on three exact states it has rank 2 of 10 columns, the full basis rank 3 of 11. Four variants were therefore run, all separately labelled: *native* (unchanged); *spectral* (the quadratic correction only); *spectral-full* (correction plus full basis); *matched-diagonal* (correction, full basis, and independent inverse-gamma noise-variance updates matching our diagonal-noise assumption in place of the inverse-Wishart update). Only the first is author code. Inputs: raw and cross-fitted determinant-normalized observations; no true graph, coefficient, variance or population normalizer enters any fit.

Outcome. Smoke test: 16/16 chains. Pilot: 24 datasets $\times$ two normalizations $\times$ {native, matched-diagonal} $\times$ four chains at the authors' budget (2000 iterations, 1000 burn-in, thinning 5), 384/384 completed, about 20 s each. Matched-diagonal chains are frozen: median one graph move per chain after burn-in, 85 of 192 chains with none, between-chain edge-inclusion disagreement 1.0 on every dataset, split-$\hat{R}$ 2.9–21.4 (median 5.0), bulk ESS about 4. Native chains move constantly (median 550 moves) but wander over near-empty graphs: edge inclusion .05–.30 on every pair, 3.0–3.2 edges of about 9 in the kept draws, $\hat{R}$ 1.02–3.13. A 10,000-iteration probe on one dataset changed neither (one move; 3,548 moves with inclusion .10–.33). Descriptively the returned graphs have SHD 7.75 (matched) and 8.33 (native) against 1.42 for the threshold estimator at three exact states and 3.40 for the fixed-bin BIC procedure over the same 24 datasets, but with two datasets per cell and no mixing these numbers are not method rankings, and we do not report them as such. The 1,920-chain confirmation run was not started. Possible causes we could not resolve here include the input scale relative to the kernel's priors and the eight unidentified spline coefficients per edge on three states; we make no claim about CHOD's identification theory from this.

J THE EQUAL-VARIANCE GENERATOR AND WHY IT WAS REPLACED

The generator of Appendix E as used in earlier versions of this work draws every node's noise variance from the same law, $H_{i,t} = \exp(\sigma_c v_t + \sigma_{\mathrm{id}} u_{i,t})$ with common $\sigma_c, \sigma_{\mathrm{id}}$. At any state value the noise variances are therefore equal in distribution across nodes. Under that condition a single-state minimum-precision-diagonal recursion (Peters & Bühlmann, 2014; Ghoshal & Honorio, 2018) is identified for any coefficient law, and on this generator it scores 1.00 order accuracy at every κ and with both the exact and the estimated state (Table 12, from the selector study of Appendix H run on this generator). A benchmark that a 2018 equal-variance method solves at every modulation level cannot support the claim that modulation defeats existing methods; it shows only that modulation defeats methods that do not exploit equal variances. All main-text results therefore use the heterogeneous generator of Section 5. The results on the equal-variance generator remain informative about the reversal itself and about the curvature estimator, and we keep them here.

On this generator the state-aware log-affine pruner (`core/graph_theory/evaluate.py`) applied to the HetCD-Ω order gives SHD 3.63 (8/30 exact graphs) at $\kappa = .3$ and 1.87 (25/30 exact) at $\kappa = .6$ with the exact state, against 5.77 and 14.87 for the HetCD-gated order and 10.27 and 5.40 for DirectLiMIAM's; given the true order it recovers every graph exactly at every κ. With the estimated state at $\kappa = .6$ the same pruner gives SHD 20.2 for the HetCD-Ω order and 17.5 for DirectLiMIAM's, and 15.3 for the true order. The full-field phase diagram of Appendix K (PC (Spirtes et al., 2000), CD-NOD, backShift, the invariance test) and the family, heterogeneous-edge and scale sweeps of Appendices L–N were run on this generator.

K EXTENDED PHASE DIAGRAM

Additional sweeps over the state persistence $\phi \in \{0.90, 0.97\}$ and sample size confirm the same pattern for DirectLiNGAM and the Wald-invariance score (`S2_robustness.py`, `s2_robustness.csv`). The bivariate cell ($d = 2$) is the clean case: with a single edge, DirectLiNGAM is at 0.0 (perfectly reversed) from $\kappa = 0.45$.

Table 12: Equal-variance generator, $d = 10$, $n = 3000$, 30 fresh seeds (95000–95029), order accuracy. Rows marked * are from the selector study on this generator (seeds 97000–97029).

	$\kappa = 0$	.15	.30	.60
pooled				
DirectLiNGAM	.98±.05	.71±.19	.49±.15	.28±.12
DirectLiMIAM (moment 2,3)	.57±.14	.63±.16	.80±.15	.92±.08
exact state				
HetCD-Ω (signed curvature)	.49±.18	.70±.17	.91±.09	.98±.05
HetCD-Ω, shrinking-window covariances	.50±.17	.53±.17	.58±.18	.84±.12
HetCD-gated (symmetric)	.98±.05	.95±.08	.87±.14	.73±.18
stratified DirectLiNGAM (terciles)	.95±.07	.91±.10	.82±.09	.69±.14
min precision diagonal, one state*	.99±.03	.99±.02	1.00±.02	1.00±.00
selector (permutation score) / HetCD-gated*	.99±.02	.96±.07	.92±.11	1.00±.02
estimated state				
HetCD-Ω	.48±.15	.55±.15	.72±.17	.84±.14
HetCD-gated (symmetric)	.98±.05	.99±.04	.92±.15	.62±.25
min precision diagonal, one state*	.99±.03	.99±.03	1.00±.00	1.00±.00

Table 13: Reversal across methods on the equal-variance generator, remaining cells (mean±sd over 20 seeds). backShift-style, slope-constancy and Wald-invariance are our own scores (Appendix AI).

cell	method	κ=0	.15	.30	.45	.60
d=2, Gauss.	DirectLiNGAM	.950±.22	.900±.31	.300±.47	.000±.00	.000±.00
	VARLiNGAM	.950±.22	.900±.31	.300±.47	.050±.22	.000±.00
	ICA-LiNGAM	.900±.31	.850±.37	.700±.47	.350±.49	.300±.47
	slope-constancy	1.00±.00	1.00±.00	.700±.47	.150±.37	.000±.00
	backShift-style (ours)	.600±.50	.600±.50	.450±.51	.100±.31	.100±.31
	Wald-invariance	1.00±.00	1.00±.00	.950±.22	.300±.47	.200±.41
d=10, $t(4.5)$	DirectLiNGAM	1.00±.00	.983±.08	.778±.13	.627±.16	.464±.15
	VARLiNGAM	1.00±.00	.990±.05	.785±.14	.622±.15	.471±.13
	ICA-LiNGAM	1.00±.00	.962±.12	.736±.25	.503±.23	.353±.23
	slope-constancy	.876±.08	.755±.15	.498±.19	.300±.16	.184±.12
	backShift-style (ours)	.689±.15	.533±.16	.341±.19	.256±.14	.174±.11
	PC	.727±.21	.773±.19	.753±.18	.651±.21	.644±.26
	CD-NOD	.721±.20	.766±.19	.747±.17	.642±.20	.623±.27
	Wald-invariance	.872±.08	.741±.18	.542±.17	.431±.12	.449±.12
d=20, Gauss.	DirectLiNGAM	.987±.03	.735±.12	.506±.11	.395±.11	.337±.08
	VARLiNGAM	.985±.03	.748±.10	.490±.11	.418±.10	.316±.07
	ICA-LiNGAM	.434±.09	.427±.09	.367±.13	.334±.15	.230±.12
	slope-constancy	.877±.06	.762±.11	.539±.13	.351±.14	.290±.07
	backShift-style (ours)	.573±.09	.492±.11	.353±.11	.267±.09	.229±.08
	PC	.758±.11	.776±.15	.716±.14	.668±.18	.665±.15
	CD-NOD	.752±.11	.766±.16	.715±.14	.668±.18	.666±.15
	Wald-invariance	.927±.04	.810±.10	.641±.14	.521±.09	.545±.11

L REVERSAL ACROSS NINE MODULATION FAMILIES

The bivariate model of Assumption 1 is kept (common AR(1) state, Gaussian innovations with common and idiosyncratic volatility, x exogenous) but $b(v)$ is drawn from nine families with randomized parameters: exp-affine (positive control), linear, quadratic, tanh, sigmoid, $|v|$, step, sine, and a random smooth function (random Fourier features); 120 configurations each, 1080 in total. Every configuration is placed on the paper's own axis $\hat{\kappa}$ (the reduced-form affine slope of $\log|\hat{b}|$ on the estimated state) so families are comparable. Two questions are asked of each configuration: do the field methods reverse (DirectLiNGAM, HetSEM-const), and does recovery survive (HetCD-affine,

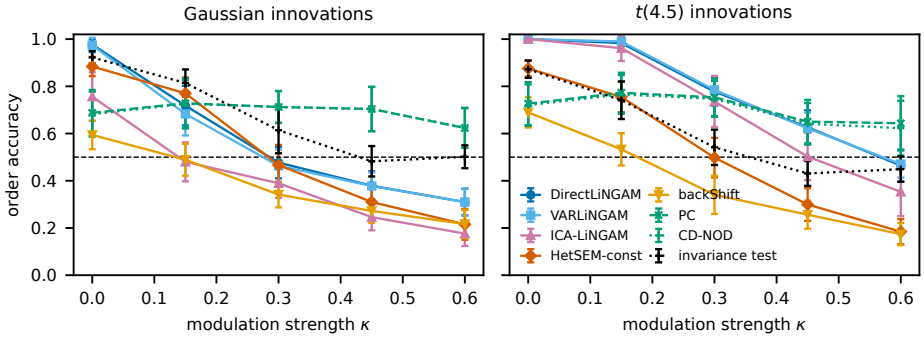

Figure 3: Phase diagram on the equal-variance generator at $d = 10$ (mean and 95% CI over 20 seeds). Left: Gaussian innovations; right: $t(4.5)$. Each of the plotted ordering scores (DirectLiNGAM, ICA-LiNGAM, VARLiNGAM, the slope-constancy and backShift-style scores of Appendix AI) crosses the chance line as modulation grows; heavier innovations delay the crossing; constraint-based methods do not reverse. The legend labels "backShift", "HetSEM-const" and "invariance test" of this archived figure denote our backShift-style heuristic, the slope-constancy score and the Wald-invariance score respectively (Appendix AI).

DirectLiMIAM). Code `legacy/G1_modulation_zoo.py`, data `legacy/g1_zoo_biv.csv`, figure `fig_g1_zoo.pdf`.

Table 14: Bivariate zoo, amplitude-strong configurations (std of $\log|b| \geq 0.30$). "Reverses" is the fraction of configurations oriented backwards; "recovers" the fraction oriented correctly.

	n	DLiNGAM rev.	HetSEM rev.	HetCD rec.	DLiMIAM rec.
monotone families	381	.74	.67	.71	.34
non-monotone families	301	.71	.62	.56	.44

Pooled over all nine families and binned on $\hat{\kappa}$ (.05/.15/.25/.35/.47/.66/1.06), DirectLiNGAM orients .87/.65/.31/.22/.25/.19/.03 and HetCD-affine .95/.88/.83/.72/.52/.46/.39; DirectLiMIAM is .62/.44/.26/.34/.50/.43/.24. Two caveats. DirectLiMIAM is not a clean tail winner in the bivariate zoo: pooled over families including non-monotone ones, its mean-independence assumption degrades too, so "the escape owns the tail" is a multivariate, monotone-family statement (Appendix N). And step and random-smooth families rarely reach large $\hat{\kappa}$ because their reduced-form affine slope is small; on the amplitude axis they still reverse, which shows that $\hat{\kappa}$ under-measures non-monotone modulation while the reversal occurs regardless.

At $d = 10$ (Figure 5; `legacy/g1mv_d10.csv`), monotone families give DirectLiNGAM .997/.815/.580/.404, HetCD-gated 1.000/.997/.903/.741 and DirectLiMIAM .503/.581/.619/.647 at $\kappa = 0/.2/.35/.5$; non-monotone families give .994/.709/.472/.319, .997/.930/.707/.435 and .564/.702/.740/.748. The same pattern holds at $d = 6$.

M Heterogeneous and sparse modulation

Per-edge generator $B_{ij,t} = B_{ij}\exp(\kappa_{ij}h_{ij}(v_t))$ on a modulated subset of edges and B_{ij} elsewhere; $\rho \in \{0.25, 0.5, 1\}$ is the modulated fraction, $\kappa_{ij} \sim \kappa_{\text{scale}}\,U(0.5, 1.5)$ (mean $|\kappa| \approx 0.31$ at $\kappa_{\text{scale}} = 0.3$, ≈ 0.61 at 0.6), optional per-edge shapes $h_{ij} \in \{\text{affine}, \text{step}, \tanh\}$; common scalar state kept; $d = 10$, 10 seeds; accuracy split by whether an edge modulates. Code `legacy/G2_hetero_modulation.py`; data `g2_hetero_*.csv`.

At $\rho = 1$ the generator reduces to the homogeneous benchmark and reproduces its $\kappa \approx 0.6$ numbers (DirectLiNGAM .24, HetCD .45, DirectLiMIAM .91), a consistency check. With mixed per-edge shapes ($\kappa_{\text{scale}} = 0.6$, $\rho = 0.5$) the field reverses on modulated edges (.364), HetCD-gated matches

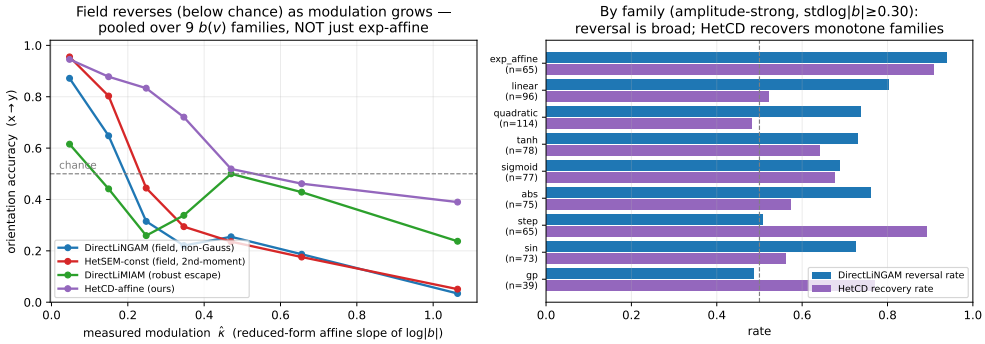

Figure 4: Bivariate zoo. Left: orientation accuracy versus measured modulation $\hat{\kappa}$, pooled over nine families. Right: per-family reversal rate of DirectLiNGAM and recovery rate of HetCD among amplitude-strong configurations.

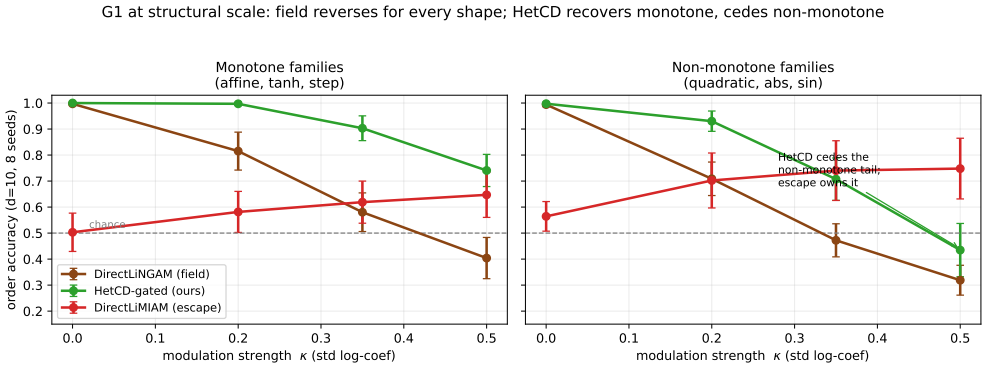

Figure 5: Family zoo at $d = 10$ (order accuracy, 8 seeds). The common modulation factor is family-shaped, $B_t = B \odot \exp(\kappa\, h(v_t))$, with h standardized so that κ is the std of the log-coefficient across families. Left: monotone shapes (affine, tanh, step); right: non-monotone (quadratic, $|v|$, sine). Every pooled score reverses for every shape; HetCD (legacy gate) recovers the monotone shapes throughout and loses the non-monotone tail to DirectLiMIAM.

DirectLiMIAM there (.69 vs .69) and wins the constant edges (.88 vs .58). The escape's clean-edge tax is thus visible within a single graph.

N Scale: $d = 20$ and $d = 40$

Order accuracy, 5 seeds, $n = 2200$ ($d = 20$) and $n = 2000$ ($d = 40$); solver non-convergence is recorded as missing rather than fabricated (two of forty gated runs at $d = 20$, $\kappa \geq 0.35$). Code `legacy/G10_scale.py`; data `legacy/g10_scale.csv`. HetCD-gated is $O(d^4)$ in the worst case and was skipped at $d = 40$ for cost; the marginal variant is shown. The gated-over-marginal gap widens with κ (.65 vs .47 at $\kappa = 0.5$, $d = 20$), so per-edge de-confounding matters more as modulation grows. All numbers are order accuracy; SHD and edge-F1 require a skeleton-estimation pass with false-positive control that we have not run, and a lagged-plus-instantaneous temporal generator is likewise not run.

O Temporal and regime learners: full record

The four temporal adapters call released code (Table 38), but an audit of the archived adapters found that none is a full regime-learning run: the "FlowMSM" (Hulsman et al., 2026) adapter imports `causal_discovery/varlingam.py` and calls the VARLiNGAM wrapper di-

Table 15: Order accuracy on modulated versus constant edges within one graph ($d = 10$, 10 seeds).

		modulated edges, $\rho =$			constant edges, $\rho =$	
κ_{scale}	method	.25	.50	1.0	.25	.50
0.3	DirectLiNGAM	.467	.545	.591	–	–
	HetCD-gated	**1.000**	**.912**	**.843**	1.00	.95
	DirectLiMIAM	.667	.739	.821	.58–.71	
0.6	DirectLiNGAM	.233	.264	.242	.795	.710
	HetCD-gated	.73	.57	.45	.93	.82
	DirectLiMIAM	**.88**	**.86**	**.91**	.58–.71	

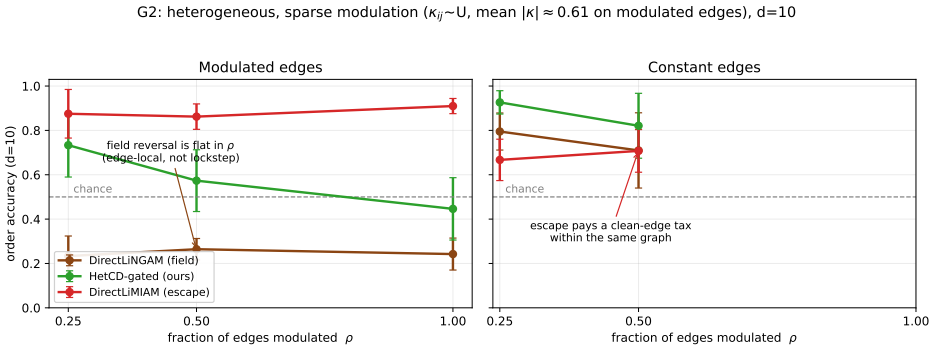

Figure 6: Heterogeneous, sparse modulation at $d = 10$ ($\kappa_{\text{scale}} = 0.6$). Left: accuracy on the modulated edges is flat in ρ for the field (edge-local reversal). Right: on the constant edges of the same graph the escape pays a clean-edge tax.

rectly (no regime inference); the "CASTOR" (Rahmani & Frossard, 2025a) adapter selects `CASTOR_linear_1_regime`, its single-regime linear component, whose engine is DYNOTEARS (Pamfil et al., 2020); Rhino (Gong et al., 2023) runs through FlowMSM's packaged component rather than the `causica` release; FANTOM's (Rahmani & Frossard, 2025b) single-window adapter was numerically unstable. The archived scoring forces orientation on the true skeleton and breaks equal-weight ties toward a column index, and raw prediction matrices were not retained, so coverage-aware rescoring is not possible from the CSVs. The numbers below are therefore component stress tests on recorded configurations; they do not establish failure of the FlowMSM or CASTOR frameworks. Each adapter is run on three bivariate conditions (Appendix P) and on the $d = 10$ sweep, 10 seeds, GPU harness `legacy/colab_G3_temporal_ABC.py`. Data: `legacy/g3_temporal_ABC_results-pt1.csv` (raw, persistent state), `_iidstate.csv` (raw, i.i.d. state), `_std.csv` (standardized), `-fantom.csv`; reference rows from `legacy/g3_temporal_sweep_reference_rows.csv` and `legacy/s4_seeds40_symmetric_gate.csv`. The chain of evidence: (1) with the state drawn i.i.d. CASTOR and Rhino are unchanged, so their apparent immunity is not a temporal-precedence channel; (2) reading the CASTOR code, `CASTOR_linear_1_regime` calls DYNOTEARS, a least-squares NOTEARS-family learner, the class Reisach et al. (2021) showed exploits varsortability on raw simulated data; (3) varsortability of the raw generator (fraction of true edges whose child has the larger marginal variance) rises from .90 to 1.00 with κ (10 seeds; Appendix J); (4) standardizing the inputs collapses CASTOR to chance and removes Rhino's rise, while FlowMSM and DirectLiNGAM, which are scale-invariant, are unchanged. Rhino is run through FlowMSM's packaged backend rather than the `causica` release. FANTOM's adapter was byte-verified against the authors' supplementary code and run at their configuration (`max_it=3`); it is below chance on the noise-only condition A, so it is numerically unstable in this single-regime contemporaneous use and we draw no conclusion from it.

Table 16: Scaling of the reversal and the recovery.

$d = 20, \kappa =$	0	.20	.35	.50
DirectLiNGAM	.903	.552	.478	.419
HetCD-gated	**.968**	**.867**	.787	.651
HetCD-marginal	.967	.821	.680	.470
DirectLiMIAM	.512	.699	**.830**	**.888**
$d = 40, \kappa =$	0			.40
DirectLiNGAM	.936			.438
HetCD-marginal	**.981**			.565
DirectLiMIAM	.490			**.920**

Table 17: What the archived adapters execute and what may be concluded.

archived label	executed component	permitted interpretation
FlowMSM	direct call of its VARLiNGAM wrapper	VARLiNGAM backend stress test; no regime inference
CASTOR	`CASTOR_linear_1_regime` (DYNOTEARS engine)	single-regime linear component; no regime learning
Rhino	FlowMSM-packaged Rhino component	component result; native-setting reproduction pending
FANTOM	single-window adapter; unstable	no comparative conclusion

P NOISE HETEROSKEDASTICITY VERSUS COEFFICIENT HETEROSKEDASTICITY

One bivariate generator, three conditions, 100 seeds each (code `legacy/G3_abc_distinction.py`, data `legacy/g3_abc_biv.csv`): (A) common-noise heteroskedasticity, constant coefficient; (B) idiosyncratic volatility only, modulated coefficient; (C) both (the benchmark). Conditions A and B separate interdependence from a changed transmission coefficient, the distinction of Forbes & Rigobon (2002). A and C are matched on the observed tercile log-variance gap (≈ 2.2 vs 2.5). Condition B sharpens the role of the estimated state $\hat{v}$: with it HetCD also reverses there, because the coefficient co-moves with a state that leaves no footprint in the volatility ($|\mathrm{corr}(\hat{v}, v)| = 0.08$–$0.33$), while with the oracle state it is perfect at every κ. In C the state does imprint on the volatility ($|\mathrm{corr}| \approx 0.72$, which is what volatility clustering is) and the proxy works. The recovery's precondition is state–volatility coupling, a testable property of the data.

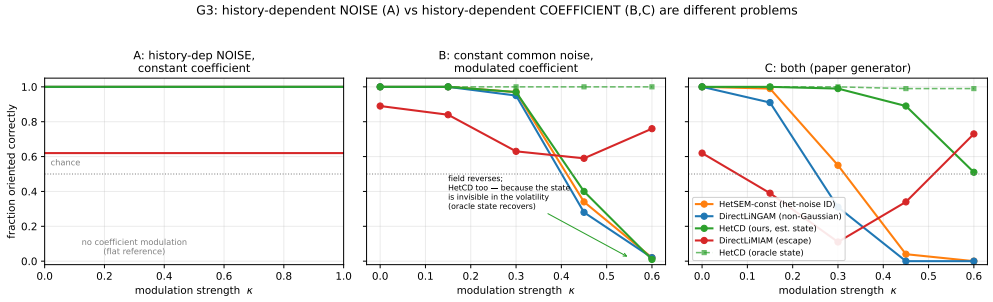

Figure 7: Conditions A, B, C. Moving the noise (A) is identifying; moving the coefficient (B, C) is reversing. HetCD needs the state to be recoverable from the volatility (B, estimated versus oracle state).

Table 18: Archived adapters: bivariate conditions A/B/C (fraction of 10 seeds oriented correctly) and the $d = 10$ sweep, by input treatment.

Method	A	B	C	$\kappa = 0$	.20	.35	.50
raw, persistent state							
DYNOTEARS*	1.0	1.0	1.0	.944	.982	.981	.938
Rhino	1.0	1.0	1.0	.568	.611	.680	.784
VARLiNGAM*	1.0	.40	.00	.993	.711	.514	.313
FANTOM (authors' config)	.20	.40	.50	.611	.581	.414	.350
raw, i.i.d. state ($\phi = 0$)							
DYNOTEARS*	1.0	1.0	1.0	.911	.983	.968	.937
Rhino	1.0	1.0	1.0	.592	.670	.760	.798
standardized							
DYNOTEARS*	.90	.90	.80	.480	.403	.419	.379
Rhino	.90	1.0	.80	.418	.463	.439	.521
VARLiNGAM*	1.0	.40	.00	.993	.711	.477	.265

Table 19: Fraction oriented correctly under the three conditions.

cond.	κ	HetSEM-const	DLiNGAM	HetCD ($\hat{v}$)	HetCD (v)	DLiMIAM
A	–	1.00	1.00	1.00	1.00	.62
B	.30	.97	.95	.97	1.00	.63
B	.60	.02	.02	.01	1.00	.76
C	.30	.55	.31	.99	1.00	.11
C	.60	.00	.00	.51	.99	.73

Q State quality and coupling on the benchmark

Design (frozen before running). Benchmark generator at $d = 10$, $n = 3000$, except that the coefficient is driven by $w = \rho v + \sqrt{1-\rho^2}\, v'$, with v' an independent AR(1) state drawn from a separate random stream so that graph, volatility state and noise are identical across ρ; $\rho = 1$ reproduces the benchmark exactly (verified to machine precision). $\kappa \in \{0.30, 0.45\}$, $\rho \in \{1, 0.5, 0\}$, seed indices 0–19. States supplied: the volatility proxy $\hat{v}$ (lag-one PC1 of log-EWMA variances, as deployed), the exact coefficient state w, and w plus independent $N(0, 0.5^2)$ or $N(0, 1)$ noise (correlation with w about 0.89 and 0.71). Methods: pooled DirectLiNGAM; DirectLiNGAM stratified on terciles of the supplied state with per-stratum orders averaged by rank (our implementation); HetCD with the symmetric gate given the supplied state. Metric: order accuracy. Code `crosscheck/E1_state_quality.py`, data `crosscheck/E1_state_quality.csv`. Four patterns. (i) Decoupling weakens the pooled reversal (DirectLiNGAM $0.41 \to 0.59$ at $\kappa = 0.45$ as ρ goes from 1 to 0) without removing it. (ii) Stratification is only as good as its state: with the exact state it gains 0.30–0.36 over pooling, with correlation-0.71 noise it gains 0.01–0.05 (exact versus noisy, paired Wilcoxon $p \leq 0.002$ in every cell). (iii) HetCD with the volatility proxy gains 0.33–0.37 over pooling when $\rho = 1$, 0.23–0.31 at $\rho = 0.5$ and nothing at $\rho = 0$ (-0.02, $+0.01$, not significant); given the exact state at $\rho = 0$ it gains 0.08–0.16 over $\hat{v}$ ($p = 0.06$, 0.006). (iv) HetCD is not monotone in state quality: at $\rho \leq 0.5$ a noisier state sometimes helps, because a weaker state lowers the gate and part of HetCD's signal comes from observation-level variance estimates rather than from the state. We do not read (iv) as a feature; it says HetCD's use of the state is not the principled conditioning of Theorem 9.

R Full-graph metrics with an estimated skeleton

Each method returns an order; the skeleton is estimated from the pooled standardized data given that order with the adaptive lasso used inside `lingam` (OLS weights, LassoLarsIC with BIC); DirectLiNGAM uses its own adjacency matrix, which is the same procedure. The "true order" row applies the pruning to the correct order, a reference for how much of the error

Table 20: Order accuracy by coefficient–volatility coupling ρ and supplied state (20 seeds; per-cell standard deviations 0.08–0.19 are in the CSV). $|\mathrm{corr}(\hat{v}, w)|$ measures how visible the coefficient state is in volatility. Here $\epsilon_\sigma \sim N(0, \sigma^2)$; "Pooled is DirectLiNGAM without the state.

			Stratified DirectLiNGAM				
κ	ρ	$\lvert\mathrm{corr}(\hat{v}, w)\rvert$	Pooled	$\hat{v}$	w	$w + \epsilon_{.5}$	$w + \epsilon_1$
0.30	0.0	0.20	0.79	0.70	0.90	0.82	0.76
0.30	0.5	0.49	0.62	0.68	0.85	0.79	0.67
0.30	1.0	0.74	0.54	0.63	0.85	0.68	0.58
0.45	0.0	0.29	0.59	0.62	0.84	0.76	0.62
0.45	0.5	0.52	0.48	0.51	0.76	0.71	0.52
0.45	1.0	0.72	0.41	0.46	0.77	0.55	0.42
			HetCD with the symmetric gate				
κ	ρ	$\lvert\mathrm{corr}(\hat{v}, w)\rvert$	Pooled	$\hat{v}$	w	$w + \epsilon_{.5}$	$w + \epsilon_1$
0.30	0.0	0.20	0.79	0.77	0.85	0.87	0.93
0.30	0.5	0.49	0.62	0.86	0.86	0.93	0.94
0.30	1.0	0.74	0.54	0.92	0.89	0.89	0.86
0.45	0.0	0.29	0.59	0.60	0.76	0.78	0.83
0.45	0.5	0.52	0.48	0.78	0.80	0.78	0.73
0.45	1.0	0.72	0.41	0.74	0.75	0.72	0.59

is in the pruning step rather than the order. SHD counts missing, extra and reversed edges once each. Benchmark generator, $d = 10$, 20 seeds, 15.05 true edges on average. Code `crosscheck/E2_graph_metrics.py`. HetCD's SHD is lower than DirectLiNGAM's at every

Table 21: SHD (mean±sd), directed-edge and skeleton F1, and number of estimated edges.

	κ=0	.15	.30	.45	.60
DirectLiNGAM	2.7±4.2	15.9±8.4	25.3±8.6	30.2±7.3	34.4±5.5
directed F1 / skeleton F1	0.92 / 0.94	0.56 / 0.73	0.34 / 0.62	0.24 / 0.57	0.15 / 0.55
estimated edges	17.2	26.2	32.5	35.4	37.8
DirectLiMIAM	19.8±9.1	17.1±7.9	17.2±6.1	18.3±3.9	20.6±5.1
directed F1 / skeleton F1	0.41 / 0.71	0.49 / 0.73	0.55 / 0.68	0.56 / 0.64	0.55 / 0.60
estimated edges	27.3	26.6	28.8	30.7	33.0
HetCD (symmetric gate)	1.6±2.2	5.8±5.2	14.2±7.6	22.4±8.0	28.4±6.8
directed F1 / skeleton F1	0.95 / 0.96	0.84 / 0.87	0.67 / 0.72	0.47 / 0.63	0.34 / 0.56
estimated edges	16.5	20.1	27.6	33.0	36.0
true order, same pruning	1.4±2.1	4.2±4.5	12.8±5.9	17.1±4.5	19.0±5.1
directed F1 / skeleton F1	0.96 / 0.96	0.89 / 0.89	0.72 / 0.72	0.63 / 0.63	0.60 / 0.60
estimated edges	16.5	19.3	27.6	31.1	32.5

$\kappa > 0$ (paired Wilcoxon $p \leq 0.0006$) and lower than DirectLiMIAM's for $\kappa \leq 0.15$ ($p < 10^{-4}$); DirectLiMIAM's is lower at $\kappa \geq 0.45$ ($p \leq 0.008$). The true-order row isolates a second failure: pooled pruning produces about twice as many edges as exist at $\kappa \geq 0.45$ (32.6 versus 15.1 at $\kappa = 0.6$), because a single pooled coefficient cannot absorb a modulated one and the remaining dependence is fitted by extra parents. Full-graph recovery under modulation therefore needs state-aware pruning as well as a correct order.

S Family generality at the population level

The log-affine family is an identifying assumption of Theorem 5 and of the two-variable results; the experiments below ask how the estimator behaves when the generator leaves the family, not whether identification survives there. Refitting HetCD (legacy gate) with a degree-2 log-coefficient does not recover the quadratic generator ($0.16/0.07/0.07$ at $\kappa = .3/.45/.6$, below the degree-1 fit's $0.41/0.26/0.16$; a $K = 20$ bin sweep tops out at 0.30) and regresses in-family accuracy from 1.00

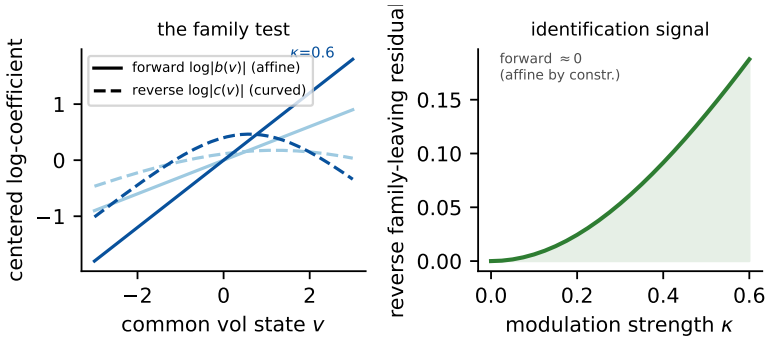

Figure 8: Identification through coefficient heteroskedasticity (Theorem 3, benchmark $\delta = 0$). Left: the forward log-coefficient is affine in the state; the reverse-implied one is curved. Right: the family-leaving residual of the reverse direction grows with modulation; this is the signal the estimator reads.

Table 22: Family-leaving residual of $\log|\text{coef}|$ under a quadratic true coefficient $\log b(v) = \kappa(v + 0.6(v^2 - 1))$ (Gauss–Hermite, $v \sim N(0,1)$, $b_0 = 0.7$). Fit within the quadratic family, the forward is lawful and the reverse-implied coefficient still leaves it.

κ	0	.15	.30	.45	.60
forward, quadratic fit	0	0	0	0	0
reverse, quadratic fit	0	.035	.110	.194	.280

to 0.56 at $\kappa = 0.3$. On the market panel the log-affine law fits reduced-form coefficient movement about as well as it fits the in-family synthetic (median bin-scatter residual 0.12 over 752 pairs against an in-family reference of ≈ 0.13; 53% of pairs at or below it, 73% within $1.5\times$), with curvature present on a substantial minority (a degree-2 fit improves the residual by more than 20% on 45% of pairs), which is what the trust diagnostic flags.

T HetCD: algorithm, gate repair, and the 40-seed comparison

Algorithm 2 HetCD causal ordering (symmetric gate)

Input: data $X \in \mathbb{R}^{n\times d}$ (standardized), estimated state s, $m_{\text{ref}} = 0.34$, $K = 6$

1. rem $\leftarrow \{1..d\}$; $\pi \leftarrow [\,]$
2. **While** $|\text{rem}| > 1$:
   a. for each $c \in$ rem: $\text{score}(c) \leftarrow \sum_{j\in\text{rem}\setminus c}\big[\ell_w(c \to j) - \ell_w(j \to c)\big]$, with $w = \text{clip}\big((\max\{|\hat\kappa_{c\to j}|, |\hat\kappa_{j\to c}|\}/m_{\text{ref}})^2, 0, 1\big)$ and ℓ_w the EWMA heteroskedastic-SEM log-likelihood of the w-scaled partial residuals of the pair on rem $\setminus \{c, j\}$, with $\hat b \in \mathcal{F}$ fitted by K-bin log-affine regression
   b. $c^\star \leftarrow \arg\max_c \text{score}(c)$; append $c^\star$ to π
   c. for each $j \in \text{rem} \setminus c^\star$: $X_j \leftarrow X_j - \hat A_{jc^\star}(s) X_{c^\star}$ (remove the fitted log-affine total effect)
   d. rem $\leftarrow$ rem $\setminus\, c^\star$
3. **Return** π followed by the last index

Score. $\ell(c \to j) = -\frac{1}{2}\sum_t[\log h_{c,t} + c_t^2/h_{c,t} + \log h_{r,t} + r_t^2/h_{r,t}]$ with $r_t = j_t - \hat b(s_t)c_t$ and $h_{\cdot,t}$ EWMA variances ($\lambda = 0.94$) that include the current squared observation; it is not a predictive likelihood. $\hat b(v) = \text{sgn}(\text{med}_k \hat b_k)\exp(\hat a_0 + \hat a_1 v)$ from $K = 6$ quantile bins of s (minimum 25 observations per bin); $\hat\kappa = \hat a_1$. The state s is the lag-one first principal component of the log-EWMA variances of all columns (Assumption 1's $\hat v$); PCA loadings use the full batch.

The gate bug and its repair. The legacy implementation computed the gate slope only in the direction from the smaller to the larger column index, so relabelling columns could change the recovered order. In 36 permutation trials ($d = 6$, $n = 1500$, 12 seeds per $\kappa \in \{0, 0.3, 0.6\}$) the legacy ordering changed under column permutation in 14; the symmetric gate, which uses $\max$ of the two directional slopes for the unordered pair, changed in 0. This guarantees pair-gate invariance under swaps; whole-order equivariance additionally needs unique greedy maxima. All other scoring and peeling choices are unchanged. On ten historical seeds the symmetric gate gives $1.00/1.00/.946/.763/.594$ against legacy $1.00/1.00/.845/.693/.512$ at $\kappa = 0/.15/.3/.45/.6$; on twelve fresh seeds at $d = 6$ the differences are $+0.000/+0.011/+0.052$ with paired 95% intervals $[0, 0]$, $[-0.091, 0.114]$, $[-0.059, 0.163]$. Code `core/generator/permutation_check.py`, `core/generator/reproduce_and_repair.py`. The de-confounding by partial residualization on the remaining variables may condition on descendants and has no general causal justification; it is an empirical device whose value is measured in Appendix U. Data `legacy/s4_seeds40.csv`, `legacy/s4_seeds40_symmetric_gate.csv`.

Table 23: Symmetric versus legacy gate on the 40-seed head-to-head grid ($d = 10$, $n = 3000$, seed indices 0–39); paired Wilcoxon over seeds. W/L counts wins/losses of symmetric versus legacy; Δ_L compares symmetric HetCD with DirectLiMIAM.

κ	legacy	sym.	Δ	p	W/L	LiMIAM	Δ_L	p
0	.985	.985	0	–	0/0	.526	$+.459$	$< 10^{-4}$
.15	.975	.975	0	–	1/1	.666	$+.310$	$< 10^{-4}$
.30	.868	.925	$+.057$	.008	11/4	.832	$+.093$	.001
.45	.672	.748	$+.075$	.020	22/10	.906	$-.158$	$< 10^{-4}$
.60	.519	.591	$+.072$	.005	24/10	.916	$-.325$	$< 10^{-4}$

U De-confounding ablation

Table 24: All HetCD variants at $d = 10$ (mean±sd over 10 seeds). "Unconditional" residualizes every pair on all remaining variables regardless of modulation; "capped" limits the conditioning set; "oracle" uses the true lagged state.

variant	κ=0	.30	.45	.60
marginal ($w \equiv 0$)	1.00±.00	.885±.16	.563±.26	.497±.26
unconditional de-conf., $\hat{v}$	.871±.14	.863±.12	.729±.16	.575±.22
unconditional de-conf., capped, $\hat{v}$	.878±.10	.867±.09	.687±.21	.584±.25
unconditional de-conf., oracle state	.866±.12	.807±.14	.790±.12	.679±.16
gated ($w = \mathrm{clip}((\hat{m}/.34)^2)$), $\hat{v}$	1.00±.00	.845±.21	.693±.23	.512±.26
gated, oracle state	1.00±.00	.924±.12	.772±.18	.681±.23
DirectLiMIAM (reference)	.534±.17	.783±.17	.894±.09	.907±.07

The gate signal itself separates regimes: the median per-pair $|\hat{m}|$ over true edges is 0.09, 0.16, 0.27, 0.41, 0.50 at $\kappa = 0, .15, .3, .45, .6$ (`S4_gate_signal.py`), so w is near 0 on clean data and saturates near $\kappa \approx 0.34$ by construction.

Mechanism (bivariate). With 200 replications of the bivariate model, HetSEM with a constant b orients $1.00/0.48/0.01$ at $\kappa = 0/0.3/0.6$; the same likelihood with $b \in \mathcal{F}$ in $\hat{v}$ orients $1.00/0.995/0.61$; with the oracle state $1.00/1.00/1.00$. The only change between the first two rows is the family restriction.

De-confounding in isolation. Pairwise orientation on the true edges with the oracle state, so neither ordering nor state estimation contaminates the measurement (10 seeds, 14.9 edges per graph): marginal $.915/.785/.734$ and de-confounded $.925/.860/.830$ at $\kappa = .30/.45/.60$. Most of the collapse is present before ordering, confirming cross-parent contamination; de-confounding recovers it in proportion to κ; but the recoverable pairwise ceiling (0.83) is below DirectLiMIAM's 0.91.

Where the accuracy is lost at $\kappa = 0.6$. Three losses stack: pairwise degradation with oracle state and de-confounding (0.83), ordering (0.68 full-order with the oracle state), and state estimation (0.51 with $\hat{v}$). Levers on the last two are bounded by the oracle ceiling $0.68 < 0.91$.

Cost of the two multivariate approximations (oracle state, 20 seeds). Pairwise orientation on source–descendant pairs split by slope balance: balanced .91/.87/.81 versus unbalanced (31% of pairs) .90/.87/.71 at $\kappa = .15/.30/.45$; the two separate only where the method already loses to DirectLiMIAM. Source-detectability rate (fraction of greedy steps whose score-argmax is a true source of the reduced DAG): .97/.90/.71; this is a measured property of the algorithm on this generator, not a lemma.

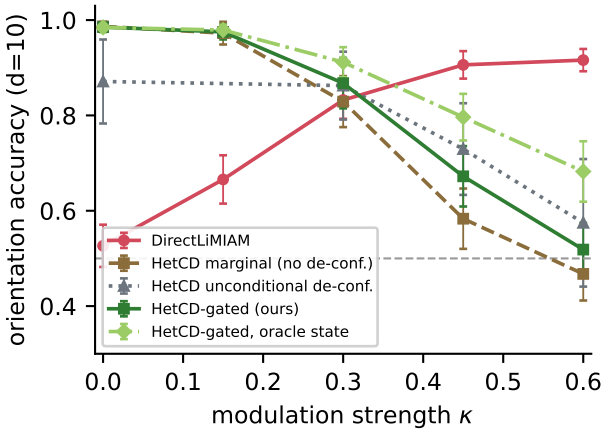

Figure 9: De-confounding ablation ($d = 10$, 10 seeds). Unconditional de-confounding damages clean data; gating keeps the clean-data ceiling and recovers the $\kappa = 0.45$ lift; the oracle state bounds what better state estimation could add.

V SENSITIVITY TO THE TUNING CONSTANTS

Table 25: HetCD-gated over the gate reference m_{ref} (at $K = 6$) and bin count K (at $m_{\text{ref}} = .34$); mean±sd over 10 seeds. Headline setting in bold.

κ	m_{ref}=.28	**.34**	.40	K=4	K=8
0	1.00±.00	**1.00**±.00	1.00±.00	1.00±.00	1.00±.00
.15	1.00±.00	**1.00**±.00	1.00±.00	.971±.09	1.00±.00
.30	.904±.11	**.845**±.21	.904±.14	.851±.18	.928±.11
.45	.694±.22	**.693**±.23	.701±.21	.637±.31	.741±.19
.60	.545±.24	**.512**±.26	.606±.21	.397±.25	.601±.20

A larger m_{ref} helps at $\kappa = 0.6$ (+0.09) and $K = 8$ improves every κ (+0.03 to +0.09), all within one seed-sd; the headline uses the pre-registered constants and is therefore conservative. The gate signal separates regimes on its own: the median per-pair $|\hat{\kappa}|$ over true edges is .09/.16/.27/.41/.50 at $\kappa = 0/.15/.3/.45/.6$.

W OUT-OF-MODEL STRESS

A two-factor volatility state (coefficient driven by one factor, noise by both) makes the scalar $\hat{v}$ a lossy proxy and HetCD-gated tracks the single-factor numbers within a seed-sd. A step coefficient is out of the family but monotone; the sign-of-median plus affine fit tracks the jump and HetCD is unharmed while DirectLiMIAM drops to ≈ 0.57. A quadratic log-coefficient is where HetCD fails, reversing at $\kappa = 0.6$, while DirectLiMIAM is unaffected; the family-leaving residual is $\approx 3\times$

Table 26: Order accuracy at $d = 10$ (10 seeds), HetCD-gated / DirectLiMIAM / DirectLiNGAM. "affine" is the in-family baseline; "step" a tercile regime switch (out of family, monotone); "quad" $\log b(v) = \kappa(v + 0.6(v^2 - 1))$ (curved). Last column: mean family-leaving residual of the affine fit.

state	coef	κ=.30	.45	.60	diag. r
1-factor	affine	.85/.78/.53	.69/.89/.38	.51/.91/.25	.13
	step	.93/.56/.79	.83/.57/.59	.62/.57/.45	.15
	quad	.56/.90/.26	.33/.88/.19	.25/.87/.12	.52
2-factor	affine	.90/.81/.58	.70/.90/.42	.43/.89/.34	.13
	step	.90/.58/.66	.66/.58/.40	.47/.51/.37	.15
	quad	.46/.87/.35	.28/.85/.29	.21/.86/.17	.56

its in-family value there (detection 26–35% per edge at a 5% false-positive threshold) and at the false-positive floor under the harmless step (2–4%). Appendix Y gives the better-calibrated version of this diagnostic.

X Alternative baselines

Table 27: Baselines that could plausibly already solve the problem ($d = 10$, order accuracy, mean±sd over 10 seeds).

method	κ=0	.15	.30	.45	.60
DirectLiNGAM	.99±.02	.73±.18	.53±.07	.38±.14	.25±.10
DirectLiMIAM	.53±.17	.68±.13	.78±.17	.89±.09	.91±.07
state-conditioned DirectLiNGAM (per $\hat{v}$ tercile)	.99±.03	.89±.10	.65±.22	.51±.17	.28±.14
location-scale / quantile-invariance score	.55±.09	.64±.12	.87±.13	.93±.11	.97±.06
router on $\hat{\kappa}$ (DirectLiNGAM / DirectLiMIAM)	.99±.02	.73±.18	.56±.13	.82±.18	.91±.07
oracle two-way router	.99±.02	.79±.11	.80±.13	.89±.09	.91±.07
HetCD-gated	**1.00**±.00	**1.00**±.00	**.85**±.21	.69±.23	.51±.26

The location-scale score is our implementation of a RESQUE-style parent-scale heteroscedastic-noise method. We expected it to reverse; it does not, because the modulated scale it must detect co-moves with the common state rather than with the parent, so the parent-conditional residual stays approximately invariant. Whether the published RESQUE behaves the same is open; the reversal we prove is specific to methods that read the parent's own second or higher moments. CD-NOD's orientation phase is not run here (no working implementation in our environment); its constraint-based immunity is documented on PC/CD-NOD skeletons in Appendix J.

Y The trust diagnostic, recomputed

The archived script `legacy/G9_diagnostic_abstention.py` fit the affine goodness of fit in the column $0 \to 1$ direction, which the generator always makes the true direction, although its comment says it fits HetCD's chosen direction. The audit (`core/generator/checks.py`) rebuilt all 1,080 zoo configurations, reproduced HetCD's orientation decisions exactly, and recomputed R^2_{aff} in the *selected* direction, which is computable without the truth. On the 682 amplitude-strong configurations the selected-direction AUC for predicting a correct HetCD orientation is 0.799; the fixed true-direction feature has 0.588. Two other test-time statistics were tried and failed: monotonicity of the bin scatter (AUC 0.40) and sign-flip/curvature flags (0.55–0.62; an escape-if-curved policy never beats always-HetCD). A 3,000-replicate cluster bootstrap over shared configuration seeds gives a fixed-policy improvement interval of $[-0.013, 0.044]$ for the router over always-HetCD, including zero. The families informed development, seeds are shared across families and the amplitude filter uses genera-

Table 28: Selected-direction diagnostic and routing on the bivariate zoo (682 amplitude-strong configurations). Routing thresholds are chosen on eight families and evaluated on the ninth, repeated over all nine.

subset	n	HetCD	DirectLiMIAM
misspecified ($R^2_{\text{aff}} < 0.5$)	221	.38	**.50**
well-specified ($R^2_{\text{aff}} \geq 0.5$)	461	**.77**	.33
policy		**accuracy**	
always DirectLiMIAM (archived outcomes)		.384	
always HetCD		.642	
leave-one-family-out router		.658	
oracle best-of-two		.784	

tor truth, so this is a retrospective exercise, not prospective validation. The soft partial-residualization gate of Algorithm 2 and this hard router are different algorithms; one does not implement the other. Data `core/generator/results/diagnostic_recomputed.csv`, `core/generator/results/routing_family_holdout.csv`.

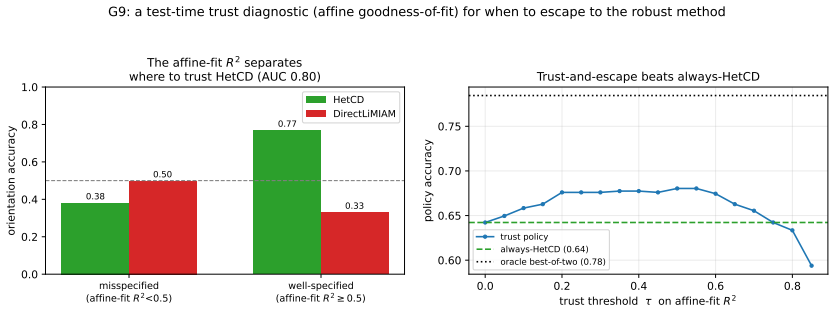

Figure 10: Archived (true-direction) version of the diagnostic figure, retained for the record; the selected-direction numbers in Table 28 supersede it.

Z Observed-state recovery, interval feasibility, and proxy-state failure

These experiments (`state_information/qml_recovery`, `gmm`, `moment_diagnostics`, `interval_method`; `certificates/deployed_score`, `recheck`) use a bivariate generator with $v \sim \text{Uniform}[-2, 2]$ i.i.d., independent Gaussian idiosyncratic states and innovations, $q_y/q_x = e^{\delta v}$ and $b = 0.7e^{\kappa v}$; designs were frozen before running; the observed current v is supplied to the recovery scores (an oracle-information study, not an end-to-end latent-state evaluation).

Cross-fitted quasi-likelihood (QML). Three folds; training-only state-quantile bins (their number is $\min(30, \max(4, \lfloor n_{\text{train}}^{1/3} \rfloor))$) estimate source and residual second moments; in each direction a signed log-affine coefficient minimizes the training bin-profiled residual log variance (two starts per sign, compact bounds); held-out losses include log-variance and squared-error terms. This approximates the population objective of Proposition 13 but is not a proved consistent estimator. At $\kappa = 0.6$ the QML margin in the true direction, 0.009272 (SE 0.000372), matches a numerically minimized population gap of 0.009276; at $\kappa = 0.3$ the population gap is 0.000569 and QML is at chance, with a paired accuracy difference from HetCD of -0.45 (seed-bootstrap $[-0.65, -0.25]$). Recovery at moderate movement is not established by the covariance criterion; the legacy HetCD score's advantage there comes from information beyond the pooled covariance family (its EWMA variances carry idiosyncratic-history heteroskedasticity), consistent with Proposition 11. At $\kappa = \delta = 0$ the covariance gap is zero and the forced-choice accuracy falls to 0.15: the criterion cannot use the conditional non-Gaussian information that Proposition 10 says exists. **State error.** On separate seeds at

Table 29: Fresh bivariate orientation accuracy at $n = 600/2400/9600$ (values in this order), 20 seeds per cell; QML and the legacy bivariate HetCD score receive the true current state; DirectLiNGAM 1.13.0 and DirectLiMIAM are rerun on every dataset. "Constant" is $\kappa = \delta = 0$; "Gaussian ambiguity" is $\delta = 2\kappa$ with $C = 0$; "mixture covariance tie" is $\delta = 2\kappa$ with $C = 0.5$. Ambiguity-row accuracies measure agreement with an arbitrary generating arrow.

condition	QML	HetCD	DirectLiNGAM	DirectLiMIAM
constant	.45/.40/.15	.70/.95/1.00	.80/1.00/1.00	.65/.75/.75
moving (κ=.3)	.50/.45/.55	1.00/1.00/1.00	.65/.70/.70	.40/.10/.00
strong (κ=.6)	.90/.95/1.00	1.00/1.00/1.00	.00/.00/.00	.20/.00/.00
Gaussian ambiguity	.55/.60/.65	.65/.40/.60	.35/.05/.00	.65/.75/.85
mixture covariance tie	.45/.40/.35	.65/.85/1.00	.60/.55/.35	.80/.80/.80

$n = 2400$, $\kappa = 0.6$, QML accuracy is .95 with the exact state, .20 with $v + N(0, 1)$ and .75 with an independent irrelevant Gaussian state (Wilson intervals $[.764, .991]$, $[.081, .416]$, $[.531, .888]$); the effects of state corruption are not monotone.

Interval feasibility with a guarantee. At $K \geq 3$ observed state levels, fit separate slope t-intervals (Bonferroni, level α); retain a direction iff one signed log-affine coefficient curve passes through every interval (two-variable linear feasibility per sign); report only a unique survivor; abstain otherwise. With independent, nondegenerate conditional-Gaussian samples the wrong-only report probability is at most $\alpha/2$ (2.5% unconditional at $\alpha = 0.05$; not a bound on wrong/reported), and $\alpha_n = 1/n$ gives pointwise correct-orientation consistency away from ambiguity for fixed levels and growing samples per level (the argument is standard regression intervals, a union bound and confidence-set inversion, credited as conventional). In a frozen design of 24 cells $\times$ 200 fresh seeds (three state levels, $\eta \in \{0, .3, .6, 1.2\}$, $n \in \{2400, 4800, 9600\}$, $C \in \{0, 0.5\}$) no wrong report occurred at $\eta > 0$; at $\eta = 0$ the rule abstained (0/200 reports); at $\eta = 0.3$, $n = 9600$ it reported once (correctly); at $\eta = 0.6$, $n = 9600$ it reported 118/200; at $\eta = 1.2$ it reported 200/200 (Wilson coverage $[98.1\%, 100\%]$) and 197/200 under the $C = 0.5$ mixture, which is outside the theorem. **Noisy labels break it.** With nearest-state labelling from $v + N(0, 1)$ (21.1% wrong labels) at $\eta = 1.2$, $n = 9600$, the rule reports 197/200 and all 197 are wrong (three reject both). With proxy SD 0.5 (3.0% wrong labels): 106 correct reports, 94 both-rejected. The population mixture covariance $\Sigma_{\text{obs},j} = \sum_i P(V = i \mid \text{obs} = j)\Sigma_i$ explains the geometry: with three equally spaced labels a log-affine coefficient has zero second difference of log slopes, and at $\eta = 1.2$, proxy SD 1, the second difference is -0.822 forward versus -0.026 reverse, so label mixing makes the wrong family much closer to log-affine.

Model checking as a safeguard. A two-step GMM (Hansen, 1982) on the four moments $\mathbb{E}[X\{Y - \operatorname{sgn} e^{a_0 + a_1 S} X\}h_j(S)] = 0$, h_j Legendre polynomials of $\tanh(S/2)$, compared with χ^2_2 at frozen $\alpha = 0.05$; report a direction iff it is not rejected and the other is. With the exact state at $n = 9600$ the

Table 30: Frozen GMM rule on fresh seeds ($\alpha = 0.05$; 100 datasets per row except the last two, 50 each). Cover: fraction receiving an arrow; wrong/all: all datasets as denominator; QML forces a direction.

condition	n	cover	wrong/all	both fail	QML error
constant	2400 / 9600	.07 / .04	.04 / .01	.00 / .00	.64 / .69
moving (κ=.3)	2400 / 9600	.24 / .62	.05 / .00	.00 / .01	.50 / .38
strong (κ=.6)	2400 / 9600	.92 / .94	.00 / .00	.04 / .03	.01 / .00
Gaussian tie	2400 / 9600	.07 / .06	.04 / .01	.00 / .00	.48 / .52
mixture covariance tie	2400 / 9600	.06 / .04	.03 / .02	.00 / .00	.61 / .69
state: oracle/noisy/irrelevant	2400	.86/.69/.07	.00/.68/.02	.07/.22/.00	.00/.85/.31
quadratic/sine coefficient	9600	.00/.00	.00/.00	.92/.92	.00/.72

rule reports 62 arrows at moderate and 94 at strong modulation with zero wrong (Wilson selective-error interval $[0, .058]$ for the former). With the noisy proxy at $n = 2400$, 69 of 100 datasets receive an arrow and 68 are wrong (unconditional error .68, $[.583, .763]$; selective error .986, $[.922, .997]$);

the true-direction moment null is rejected in 96 cases, because after conditioning on the proxy that null no longer represents the structural direction. Held-out quadratic and sine coefficient families produce no arrows (46/50 both-rejected). Across 1,400 evaluations 449 optimizer starts returned non-success and cellwise numerical abstention was 1–8%; all outcomes are retained.

Measurement-error-corrected moments. With $S = V + \epsilon$, $\epsilon \sim N(0, t)$ independent and t known, the corrected polynomials $H_j(s;t)$ ($H_0 = 1$, $H_1 = s$, $H_{j+1} = sH_j - jtH_{j-1}$) satisfy $\mathbb{E}[e^{aS-a^2t/2}H_j(S - at;t) \mid V] = e^{aV}V^j$ (Nakamura, 1990), so latent-state moments are recoverable from observable ones without a Gaussian prior on V. Substituting four corrected moments into the same GMM rule: at $\kappa = 0.6$, $n = 9600$ with correct calibration the rule reports 52 arrows, 49 right and 3 wrong (selective error .058, [.020, .156]), a paired reduction in unconditional wrong-arrow frequency of .16 ([.09, .24]) relative to no correction; but at $n = 2400$ it gives 26 wrong among 37, at $\kappa = 0.3$, $n = 9600$ 11 wrong among 20, and incorrect calibration causes severe errors; the true-direction nominal p-value falls below .05 in 39%/32%/12% of the three correctly calibrated cells. A separate diagnostic that supplies the true forward coefficient and no optimizer rejects the true model in 24 of 100 cases at nominal 5% ([16.7%, 33.2%]), and corrected Fourier moments still reject 23, so the miscalibration is not an optimization artifact (Huang, 2014). We do not adopt this candidate.

AA LABEL-ERROR CORRECTION AND ITS VARIANCE COST

Code: `label_error/calibrated_scores/`. Frozen design: the matched example at $n \in \{2000, 10000\}$, 20 seeds, $m \in \{0, 1\}$, symmetric label error $p \in \{0, .01, .05, .1, .2, .4\}$; methods: exact-state oracle, naive conditioning on the noisy label, inverse-channel correction (Morrissey & Spiegelman, 1999), with signed channel weights, of the within-label means, covariances and the two nonlinear entropy features of the packaged pairwise score ($\mathbb{E}\log\cosh u$ and $\mathbb{E}[u\,e^{-u^2/2}]$), with the true p (known) or p estimated from 500 or 5000 independent calibration pairs; the 19/20 figure below concerns this empirical entropy-approximation score, not a direct estimate of the mutual information of Theorem 9. 2400 evaluation rows on 80 paired base datasets; 66 nonpositive-covariance numerical abstentions are retained and counted. **Why correction is expensive.** With the symmetric

Table 31: Correct/wrong/abstain out of 20 (selected cells; all cells in `summary.csv`).

m	n	p	exact state	naive	known p	500 cal.	5000 cal.
1	2000	0.05	20/0/0	2/18/0	13/7/0	9/11/0	10/10/0
1	2000	0.20	20/0/0	0/20/0	8/11/1	5/12/3	8/10/2
1	2000	0.40	20/0/0	0/20/0	3/7/10	5/5/10	3/7/10
1	10000	0.05	20/0/0	0/20/0	19/1/0	16/4/0	19/1/0
1	10000	0.20	20/0/0	0/20/0	12/8/0	5/15/0	11/9/0
1	10000	0.40	20/0/0	0/20/0	3/15/2	3/12/5	2/16/2
0	2000	0.05	20/0/0	20/0/0	20/0/0	20/0/0	20/0/0
0	2000	0.20	20/0/0	20/0/0	18/2/0	18/2/0	17/3/0
0	2000	0.40	20/0/0	20/0/0	10/5/5	6/6/8	9/6/5
0	10000	0.05	20/0/0	20/0/0	20/0/0	20/0/0	20/0/0
0	10000	0.20	20/0/0	20/0/0	20/0/0	20/0/0	20/0/0
0	10000	0.40	20/0/0	20/0/0	15/5/0	13/4/3	15/5/0

channel M, corrected within-state moments are M^{-1} applied to observed within-label moments. For the low-state outcome variance at $m = 1$ the observed low-label moments mix state variances $25/32$ and $25/2$, whose fourth moments differ by a factor of 256; the delta-method asymptotic variance of the corrected estimator, divided by that of the exact-label estimator, is about 21 at $p = 0.05$ and about 152 at $p = 0.2$. This is an estimator-specific ratio, not a lower bound for every procedure. The correction also needs calibration data that unsupervised discovery lacks, and its assumptions (symmetric, nondifferential, known class count) are strong.

AB CAUTIOUS DECISIONS: CONFIDENCE SETS AND SEARCH RESOLUTION

Code: `state_information/` (`native_amica`, `centered_witness`, `bounded_betting`) and `label_error/` (`recovery_abstention_bridge`, `resolution_ablation`). The procedure retains a direction unless classwise independence of regressor and residual can be rejected for every coefficient in a searched domain, using inverse-channel-weighted witnesses with a known, invertible, nondifferential label channel, an independent pilot for centring, and bounded betting statistics with explicit remainder bounds; it reports a unique direction only if the other is excluded. Under i.i.d. sampling, a valid direction in the searched domain and exact arithmetic, an internal argument bounds the per-run unconditional probability of a wrong unique report by 0.025; this is not a bound on the error rate among reported arrows, and the code is floating point. The generic statistical machinery (test inversion, empirical Bernstein and betting bounds) is prior work (Wang et al., 2024). Earlier abstentions on exact labels

Table 32: Confidence procedure on the reversal example at $n = 30000$ (three fresh shared seeds per cell; known channel; 45-second budget). "Coarse"/"fine": terminal search resolution 10^{-3} versus 2^{-16}, the only change. Gaussian controls are genuinely ambiguous under the tested class model, so retaining both there is correct.

model	label error	Correct only	Both	Unresolved	Wrong only
Gaussian control	0.05	0/3 / 0/3	3/3 / 3/3	0 / 0	0 / 0
reversal example	0	0/3 / 2/3	3/3 / 1/3	0 / 0	0 / 0
reversal example	0.05	0/3 / 1/3	3/3 / 1/3	0 / 1	0 / 0
reversal example	0.20	0/3 / 0/3	3/3 / 2/3	0 / 1	0 / 0

were partly computational: all four retained reverse-state intervals had rejected midpoints and were excluded at finer resolution. The remaining abstention and time-limit stops are unresolved recovery problems. Three seeds per cell cannot validate a rare-error rate. On clean single-class controls with uniform or skewed errors at $n = 100000$ the procedure oriented 3/3 each and retained both on 3/3 Gaussian controls; at $n = 30000$ and 12000 it mostly abstained (`centered_witness`, `bounded_betting`).

AC A CALIBRATED FINITE-MIXTURE LIKELIHOOD REFERENCE

Code: `state_information/` (`mixture_likelihood`, `non_gaussian_states`, `gaussian_equivalence`, `abstention`). With three known state values and a known (or separately calibrated) label-error channel, the full joint Gaussian-mixture likelihood of $(X, Y,$ noisy label$)$ with a log-affine coefficient, fitted by multistart optimization, oriented 100/100 strong-modulation datasets in each of two noise conditions (Wilson $[.963, 1]$) and 42/50 moderate ones in each, from noisy labels with standard deviation 1 at $n = 9600$, where DirectLiNGAM and DirectLiMIAM (which receive no labels) oriented none; estimating the channel from 30 or 300 labelled calibration observations per state kept 50/50 strong and 41/50 moderate. It is a conventional finite-mixture method, not a new algorithm; latent-class LiNGAM models predate it (Shimizu & Hyvärinen, 2008; Tanaka et al., 2014). It fails under coefficient misspecification: with correct calibration but a quadratic log-coefficient, or a coefficient built so the reverse regression is log-affine, it orients 0/30 in each, and a naive held-out margin/SE threshold reports 17 and 26 arrows, all wrong. A flexible conditional non-Gaussian variant (exponential-power errors, free coefficient per class) recovered 20/20 on uniform errors where the scale-mixture model got 4/20, but weak non-Gaussianity gave 7–8/20 and centred exponential errors 3/10 calibrated and 1/10 blind; the pinned official AMICA baseline produced no usable orientations in its runs (`state_information/native_amica`). No variant is a general replacement.

AD MARGINAL TRANSFORMS THAT DO NOT REPAIR THE REVERSAL

Because the certified reversal operates through the pooled marginals' heavier tails, some transforms partially undo it, never to ceiling and at the cost of the identifying signal; own-volatility devolatilization removes the signal entirely. Gaussianization also left the two covariance-based scores tested

Table 33: Marginal transforms applied before DirectLiNGAM ($d = 10$, 20 seeds), and the same transforms before two covariance-based scores.

	κ=0	.30	.60
DirectLiNGAM, raw	.977±.05	.477±.15	.310±.13
own-vol devolatilization (finance standard)	.826±.18	.484±.18	.486±.17
common-factor devolatilization	.986±.04	.732±.18	.587±.18
rank-gaussianization	.952±.08	.739±.14	.676±.16
conditional on state	.876±.10	.759±.17	.588±.18
backShift-style heuristic (ours) raw / gaussianized	.594 / .619	.341 / .552	.217 / .464
slope-constancy score (ours) raw / gaussianized	.884 / .909	.467 / .774	.215 / .427

here at chance; Proposition 11 concerns the state-conditional covariances themselves and does not speak to every marginal transform, so this is an observation about the tested preprocessing, not an impossibility result.

AE Invariant causal prediction

Table 34: ICP (Peters et al., 2016) with volatility-tercile environments, $d = 10$ (mean±sd over 20 seeds), and rank-gaussianization applied before two of our covariance-based scores (Appendix AI).

	κ=0	.30	.60
ICP accuracy (committed edges)	.689±.23	.277±.21	.457±.39
ICP abstention rate	.715±.13	.640±.10	.850±.10
backShift-style heuristic (ours) raw / gaussianized	.594 / .619	.341 / .552	.217 / .464
slope-constancy score (ours) raw / gaussianized	.884 / .909	.467 / .774	.215 / .427

Even at $\kappa = 0$ ICP abstains heavily because heteroskedastic noise already breaks noise-invariance; the load-bearing observations are the wrong commitments at $\kappa = 0.3$ and the abstention at $\kappa = 0.6$. Gaussianization lifts backShift only to chance, as Proposition 11 implies.

AF Real-data details

EuroBaVar (Laude et al., 2004)**.** 23 subjects, two postures (lying, standing), beat-to-beat [RRI, SBP, DBP, MBP]; only the beat-to-beat files are used. The edge is the baroreflex arm $\text{SBP}_t \to \text{RRI}_{t+1}$ and posture is the modulating state. For each subject the coefficient is estimated in each posture and $\hat{\kappa}$ is the log-affine slope on the posture indicator. Median $|\hat{\kappa}| = 0.82$, range 0.05–3.23. Within-posture discovery recovers SBP→RRI on 0.30 (lying) and 0.35 (standing) of subjects; only $4/23$ subjects have a clean SBP→RRI baseline in both postures, and pooling postures reverses 0 of them. The high-$|\hat{\kappa}|$ tercile is lowest (0.12), directionally consistent with modulation degrading orientation, but from an already below-chance base, so we do not lean on it. Code `legacy/G4b_eurobavar.py`; data `legacy/g4b_eurobavar.csv`.

Light-tunnel pilot. From the causal-chamber light-tunnel data (Gamella et al., 2025), dataset `lt_test_v1`, we fixed before download the pair red-light setting → first infrared sensor (`ir_1`) at `flag=0`, `diode=2`, across four exposure settings; the first 250 measurements of each exposure estimate state means (development) and the last 250 are the test set; no state or sensor was chosen by score outcome. Test-set linear slopes are $24.11, 47.98, 95.72, 191.90$ with $R^2 \in [0.976, 0.996]$. The raw pooled $\log\cosh$ score is $+0.009591$ (favouring the known input direction); after centering by development-set state means it is -0.003907; within-state scores are $-0.003884, +0.003843, -0.001363, -0.005891$. These use the $\log\cosh$ functional only, not the full entropy approximation for asymmetric distributions, and point-score signs are not statistically established directional errors. The pilot supports real state-dependent measurement gain and does not establish coefficient-induced reversal: several within-state preferences are already

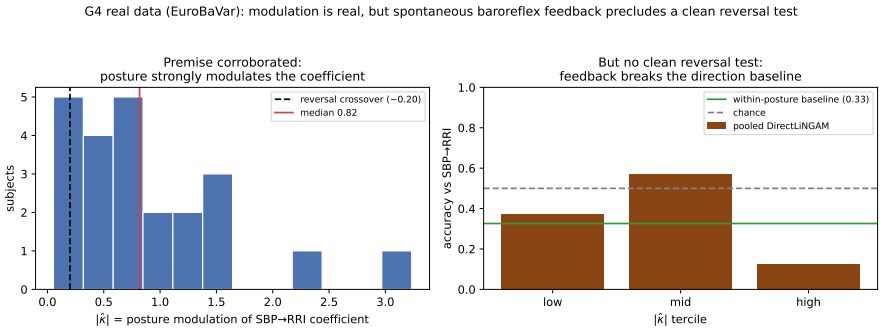

Figure 11: EuroBaVar. Left: posture strongly modulates the SBP→RRI coefficient (median $|\hat{\kappa}| = 0.82$). Right: spontaneous baroreflex feedback breaks the within-posture direction baseline, so there is no clean reversal to exhibit.

opposite and raw pooling favours the true direction. Sequential exposure blocks, drift, a nonlinear response and changing means limit attribution. A stronger external test needs matched constant- versus moving-gain conditions and independent acquisitions. Data, protocol and attribution: `real_data/light_tunnel/`. The physiology dataset PRCP (PhysioNet) was considered and set aside: baroreflex arcs are modelled with feedback in both directions, so it is not a one-way ground-truth pair without a justified lag model.

AG MARKET-DATA DETAILS

Series and transformations. Log returns (`ld`): spy, qqq, iwm, tlt, agg, lqd, hyg, eem, efa, xlf, xle, xlk, xlv, xli, xly, xlp, xlu, xlb, dxy, eurusd, usdjpy, wti, gold, copper. First differences (`d`): dgs2, dgs10, dgs30, t10y2y, t5yie, t10yie, t5yifr, dfii5, dfii10, ig_oas, hy_oas, bbb_oas, baa10y, vix, vix3m, move, skew, nfci, stlfsi4. The 33-series liquid core used for the modulation sweep is the union of the return series and dgs2, dgs10, t5yie, t10yie, ig_oas, hy_oas, vix, vix3m, move.

Common volatility state. For each series, EWMA variance with $\lambda = 0.94$ initialized at the first-250-day variance; log-variances standardized cross-sectionally; first principal component (SVD of the centred matrix), sign-aligned with the cross-sectional mean log-vol, lagged one day, standardized. PC1 explains 58.7% of log-vol variance. The tercile-mean gap of the standardized state is ≈ 2.0, so a $3\times$ coefficient swing across terciles corresponds to $\hat{\kappa} \approx \ln 3/2.0 \approx 0.56$.

Instrumented oracle. Rate family: FOMC-day monetary-policy surprises (Gürkaynak et al., 2005; Nakamura & Steinsson, 2018; Bauer & Swanson, 2023) instrument dgs2 in a local-projection IV; an edge dgs2→asset enters the oracle with the 2SLS sign when $|t| > 1.96$ (8 edges: agg, lqd, dxy, usdjpy, dgs10, t10yie, dfii5, dfii10). Oil family: daily oil-supply-news surprises on OPEC announcement dates (Känzig, 2021) instrument wti (18 edges: spy, tlt, agg, lqd, eem, efa, xlf, xle, xlb, usdjpy, copper, dgs2, dgs10, dgs30, t5yie, t10yie, hy_oas, vix3m). Each pairwise method is run on the aligned pair (5,586–5,734 days); multivariate methods are run on the full panel and the edge direction read from the recovered order. Agreement is the fraction of edges whose recovered direction matches the instrumented one; intervals are Wilson 95%.

HetCD on the oracle edges (per edge). The directional statistic $\ell(p \to c) - \ell(c \to p)$ is positive on all eight rate edges (range 90–916 in log-likelihood units) and negative on 14 of 18 oil edges; the four oil agreements are agg, lqd, t5yie, t10yie (`s4_real_oracle_hetcd.csv`). The per-edge modulation $\hat{m}$ has median $|\hat{m}| = 0.079$ on rates and 0.194 on oil; the single largest value (1.62, wti→usdjpy) is a bin-scatter outlier that the family fit flags.

HetCD-gated (deployed ordering) on the rate subsystem. Running the full gated method (Algorithm 2, $K = 6$, MREF $= 0.34$) on the nine rate variables {dgs2, dgs10, dfii5, dfii10, t10yie, agg, lqd, dxy, usdjpy} with the panel common vol state ($n = 5{,}540$ complete days) returns the

order [dgs2, dgs10, dfii10, lqd, dxy, agg, dfii5, t10yie, usdjpy]: dgs2 is ranked *source* (position 1 of 9), so all eight instrumented edges dgs2→ $\{\cdot\}$ come out parent-before-child (8/8; the edge into t10yie is deterministic and is not counted in the main text, Appendix AH). Two robustness checks: (i) re-estimating the common state from the nine rate series' own log-EWMA vols gives [dgs2, dgs10, dfii10, lqd, agg, dxy, dfii5, usdjpy, t10yie], again dgs2 first (8/8); (ii) running the ordering over the full 35-series liquid panel ($n = 4{,}560$) still ranks dgs2 first of 35, so the eight rate edges remain 8/8. The marginal (ungated) ordering agrees in all three. This confirms the 8/8 rate result under the deployed estimator, not only under the pairwise score (`S4_real_gated_rates.py`, `s4_real_gated_rates.csv`).

Modulation estimator $\hat{\kappa}$. For an ordered pair (p, c): standardize both, bin on the state into six quantile bins, OLS of c on p within each bin, log-affine least squares of $\log|\hat{b}_k|$ on the bin-mean state; $\hat{\kappa} = \hat{a}_1$. A tercile cross-check $\log(\hat{b}_{\text{hi}}/\hat{b}_{\text{lo}})/(\bar{s}_{\text{hi}} - \bar{s}_{\text{lo}})$ agrees in rank. Pairs with $|\text{corr}| < 0.15$ are dropped (752 of 1,056 retained). The full vector is in `s4_realkappa_pairs.csv`.

IV-by-regime robustness (Section 6). The structural test regresses S&P returns on the 2y-yield change by 2SLS within lagged-VIX volatility terciles on 194 FOMC days. With the Bauer–Swanson orthogonalized surprise `bs_orth` the first stage is strong in the calm tercile ($F = 52$) and weak elsewhere ($F \leq 11$); the Swanson fed-funds factor `sw_ff` ($F \leq 5$ in all terciles) and a Kuttner-style front fed-funds-futures surprise ($F \leq 2$) are weaker still and give point estimates with intervals spanning zero. Pooling four instruments (`bs_orth`, `sw_ff`, `sw_fg`, `brw_shock`) in an over-identified 2SLS lifts the stressed first stage only to $F = 10$ (vs. 48 calm) and gives an IV slope ratio stressed/calm of ≈ 1.2, far from the reduced-form "3×" and still borderline-identified. We conclude we cannot document structural coefficient modulation on this channel; the failure mode is characterized theoretically and on the benchmark, and its presence in markets is left open. Across all three instruments the reduced-form OLS slope falls from calm to stressed (ratio 0.24), so on this identified channel the reduced-form does not "rise in crises," and the IV cannot pin the structural slope in stress. Data and code: `S4_iv_regime.py`, `s4_iv_regime.csv`; misspecification and baseline experiments: `S4_misspec.py`/`s4_misspec2.csv`, `S4_baselines.py`/`s4_baselines.csv`; 40-seed head-to-head: `S4_seeds40.py`/`s4_seeds40.csv`.

Table 35: **Superseded historical output (earlier version).** Market-data direction agreement with 26 instrument-identified edges as scored in the previous version: the rate count of eight includes the deterministic breakeven edge later removed (Appendix AH, seven edges), the edges share parents and a panel so no binomial interval is appropriate and none is shown, and PC and NOTEARS commit on only 7 edges. Kept so that a reader of the earlier version can locate the change; the current check is Table 37.

method	all (26)	rates (8)	oil (18)
DirectLiNGAM (pairwise)	.35	.88	.11
DirectLiNGAM (multivariate)	.58	.88	.44
VARLiNGAM	.65	.88	.56
backShift-style heuristic (ours)	.69	.75	.67
slope-constancy score (ours)	.50	1.00	.28
PC (7 committed edges)	.57	.67	.50
NOTEARS (Zheng et al., 2018) (7 committed edges)	.43	.50	.33
Wald-invariance score	.50	1.00	.28
HetCD (pairwise score)	.46	1.00	.22
HetCD-gated (ordering)	n/a	1.00	n/a
median $\lvert\hat{\kappa}\rvert$ (HetCD)	.19	.08	.19

Frozen physical protocol (not run). A matched contrast on the causal-chamber light tunnel (Gamella et al., 2025) was specified before any acquisition: a randomized light input as cause, an infrared sensor as effect, the sensor's gain manipulated physically by a second polarizer at three angles, and four interleaved arms (gain fixed or moving, input scale fixed or moving) of 300 rows each with twelve independently started acquisitions per arm, with the pooled log-cosh score's cluster-bootstrap interval as the primary gate. The protocol is included with the code

Table 36: IV-by-regime on 194 FOMC days: structural slope of S&P returns on the 2-year yield by 2SLS within volatility terciles, instrument `bs_orth` (Bauer–Swanson orthogonalized surprise). Brackets are 90% bootstrap CIs.

tercile	n	OLS slope	IV slope [90% CI]	first-stage F
calm	62	-5.06	-6.85 [-9.6, -4.6]	51.8
mid	52	-1.79	-10.3 [-69, 8.5]	11.0
stressed	52	-1.19	-12.0 [-34, 3.8]	8.5

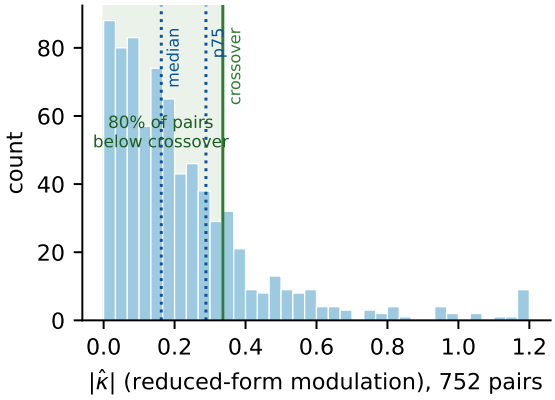

Figure 12: Distribution of estimated modulation $|\hat{\kappa}|$ across 752 market pairs; 80% lie below the HetCD-gated/DirectLiMIAM accuracy crossover on the equal-variance benchmark (0.34) and 60% below the real-calibrated one (0.20). These crossovers are estimator comparisons, distinct from the $\kappa \approx .24$ at which the DirectLiNGAM population score crosses zero.

(`real_data/light_tunnel/external_protocol.txt`); no data were collected, and the paper makes no claim that depends on it.

AH Real data with an observed volatility state

The rate-subsystem check of Section 6 uses the eight FOMC-instrumented edges from the two-year yield (dgs2) to {agg, lqd, dxy, usdjpy, dgs10, t10yie, dfii5, dfii10}, first differences for yields and log returns for prices, June 2003 to April 2026. The ten-year breakeven t10yie is by construction dgs10 minus dfii10 (correlation 1.000 with that difference in the panel), so the nine-variable correlation matrix has a zero eigenvalue; precision-based methods cannot run on it and regression-based methods rely silently on a ridge. It is dropped, and the instrumented edge into it is not counted; earlier versions of this work reported 8/8 including that deterministic edge. Three states are supplied: the lagged log MOVE index ($n = 5485$; correlation .87 with the panel PCA proxy), the lagged log VIX ($n = 5540$; correlation .60), and the PCA proxy itself. Table 37 gives the orders. The curvature spread S and its permutation score are 1.05 (.66), .59 (.97) and .95 (.70): under none of the three states does the spread separate from its permuted values, so the selector defers to the non-Gaussian order. The per-variable tercile-bin Jensen gaps place dfii5, agg and lqd as the most sink-like and dxy as the most source-like variable under every state; the instrument says dgs2 $\rightarrow$ dxy, which the tercile order gets wrong under MOVE and right under VIX. The tercile-bin 7/7 under VIX is close to what slope-sorting alone gives (6/7) and is not reproduced by the window estimator (2/7), so we do not read it as recovered structure.

AI Implementation details and compute

Baselines. DirectLiNGAM (Shimizu et al., 2011): `lingam.DirectLiNGAM` (pairwise likelihood-ratio score). ICA-LiNGAM (Shimizu et al., 2006) and VARLiNGAM (Hyvärinen et al.,

Table 37: Rate subsystem with observed states: instrumented edges oriented parent-before-child (of 7) and the rank of dgs2 in the recovered order (of 8).

Method	Lagged log MOVE	Lagged log VIX	PCA proxy
HetCD-Ω	4/7 (4)	2/7 (6)	2/7 (6)
terciles	4/7 (4)	7/7 (1)	6/7 (2)
slope-sort	0/7 (8)	6/7 (2)	0/7 (8)
gated	7/7 (1)	7/7 (1)	7/7 (1)
min diag.	2/7 (6)	2/7 (6)	2/7 (6)
DLiNGAM	7/7 (1)	7/7 (1)	7/7 (1)
DLiMIAM	4/7 (4)	4/7 (4)	4/7 (4)
S (perm.)	1.05 (.66)	.59 (.97)	.95 (.70)

2010): the `lingam` package (VARLiNGAM with one lag; the instantaneous matrix is scored). DirectLiMIAM: the authors' implementation with `score='moment', moment_powers=(2,3)`, on standardized data; its `pwling` score reproduces DirectLiNGAM exactly ($0.993/0.529/0.253$ at $\kappa = 0/.3/.6$), which is a useful check that the two implementations share a data path. "backShift-style" (this appendix only): our own bivariate Python heuristic inspired by backShift (Rothenhäusler et al., 2015), which forms covariance differences across volatility-tercile environments, solves the 2×2 generalized eigenproblem of two differenced matrices and reads the orientation off the recovered connectivity; it does not call the R package, and no result in this paper comes from the official backShift implementation. PC (Spirtes et al., 2000) and CD-NOD (Huang et al., 2020): `causal-learn` with Fisher-z tests at $\alpha = 0.05$ (CD-NOD with the time index as the surrogate). Constant-coefficient heteroskedastic SEM (*HetSEM-const*, Table 1 and Appendix F): the directional score of Appendix T evaluated with $\hat{b}(v)$ replaced by a single pooled OLS coefficient, so that all state dependence is carried by the per-observation EWMA variances; this is the Rigobon identification-through-noise-heteroskedasticity signal, scored in each direction and differenced. The rows labelled *slope-constancy score* in the equal-variance-generator tables (Appendices J–K, formerly "HetSEM-const" there) are a different, earlier score: it groups observations by the relative EWMA variance of the pair and compares the coefficient of variation of the forward and reverse regression slopes across groups. *Wald-invariance score* (formerly "invariance"): a feasible-GLS Wald statistic for state dependence of the coefficient in each direction; the direction with the smaller statistic is chosen, so the score always commits and does not abstain. PC and CD-NOD are scored on committed edges only (the fraction of true edges that the method orients and orients correctly, among those it orients); unoriented or missing edges are not counted and their coverage was not saved, so these rows are not comparable with the all-edge order accuracy of the ordering methods.

HetCD. EWMA $\lambda = 0.94$ (RiskMetrics, 1996) for both the state and the per-observation variances; $K = 6$ bins with a minimum of 25 observations per bin; $m_{\text{ref}} = 0.34$. The order search evaluates every ordered pair among remaining variables at each step; at $d = 10$, $n = 3000$ a full ordering takes ≈ 2 s on a single CPU core (the sensitivity sweep of 50 orderings ran in under two minutes), and the $d = 10$ head-to-head over 5 values of κ and 10 seeds runs in minutes. The core experiments use no GPU; the archived temporal-adapter harness (Appendix O) requested a GPU for Rhino.

Population decision variable. The quadrature evaluation of equation 9 and the packaged-estimator bracketing ($x \to y$ for $\kappa \leq 0.21$, $y \to x$ for $\kappa \geq 0.27$ on the same draws; `S4_lr_reversal.py`) are in Appendix C.

Statistics. Tables report the mean and sample sd over seeds; figures show 95% normal CIs of the mean. Paired comparisons use the two-sided Wilcoxon signed-rank test on per-seed accuracies; "all-tied" cases (identical accuracies on every seed) are reported as such rather than as a p-value.

AJ Pre-registration record

Before the head-to-head and before the market-data modulation sweep, the kill-test was fixed as: (i) hold order accuracy ≥ 0.9 at $\kappa = 0$; (ii) beat official DirectLiMIAM (moment score) at $\kappa = 0.3$; (iii) beat it at $\kappa = 0.6$; on the benchmark generator (Appendix E) with $d = 10$, $n = 3000$, seeds 1000–1009, and order accuracy as the metric (`S4_killtest_d10.py` header). Outcome: (i) met

Table 38: Baseline provenance. Released implementations are used where one exists; the rows marked "ours" are our own implementations and are named as such in the tables. Table 17 states what each temporal adapter executes.

method	implementation	inputs
DirectLiNGAM	`lingam` 1.13.0	raw in Table 1 (internal standardization); standardized in the appendix tables
ICA-LiNGAM, VAR-LiNGAM	`lingam` 1.13.0	standardized
DirectLiMIAM	authors' `limiam` repository, moment score	raw in Table 1; standardized in the appendix tables
backShift-style covariance-difference heuristic	ours (Python); not the R package	standardized, tercile environments
slope-constancy score, Wald-invariance score	ours (Python)	standardized
HetSEM-const	ours (Python); EWMA likelihood score	raw
PC, CD-NOD	`causal-learn`, Fisher-z, $\alpha = 0.05$; scored on committed edges	standardized
CASTOR adapter	official repo, `CASTOR_linear_1_regime` (DYNOTEARS)	raw, standardized
FlowMSM adapter	official repo; calls the VAR-LiNGAM backend directly	raw, standardized
Rhino adapter	FlowMSM's packaged backend (not `causica`)	raw, standardized
FANTOM	authors' supplementary code, byte-verified	raw
HetCD (legacy, symmetric gate)	ours; `core/generator/core.py`	standardized

(1.00); (ii) not met as a significant win, point estimate ahead ($+0.06$) with $p = 0.63$; (iii) not met (0.51 vs. 0.91, $p = 0.008$). The gate constant $\text{MREF} = 0.34$ was fixed from the crossover measured on the marginal method before the gated method was built; $K = 6$ was fixed from the bivariate kill-test. Neither was tuned afterwards; Table 25 reports what tuning would have bought. With the column-order repair of the gate (Appendix T), which is an implementation correction and not a tuning change, criterion (ii) is met on the 40-seed grid ($+0.093$, $p = 0.001$) and (iii) still fails.