SEAM: RECOVERING CAUSAL REVERSALS ACROSS RECURRING REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many systems switch between a few recurring states whose causal relations partly differ: some relations stay the same, while others reverse direction. From one multivariate time series with unknown state boundaries, we ask when the state changes, which causal relations change, and what an intervention would do in each state. We introduce SEAM (Shared Equations Across Mechanisms), which jointly learns the recurring states, a causal graph for each, and which structural equations the states share. Separate causal orders let an effect reverse; estimating the unchanged equations from both states improves precision. One description-length score decides between no change, a change in noise levels only, and a change in the causal mechanisms, and a resampling test reruns the entire search on surrogate series to assess the evidence for a change. We give conditions under which the decision is consistent, both for Gaussian VAR series without a change and for mechanism changes that no change in noise levels can mimic, and we show that a row-permutation variant of the test controls its level at every sample size for exchangeable observations without lags. On five-variable systems with a reversal, SEAM's error on the reversed effects is 0.016 at the larger effect size, compared with 0.406–1.000 for five regime-aware baselines, and over all 640 series it finds the changed equations with F1 0.951; on 1400 series without a change, it reports a mechanism change in 0.36% of them. On five public daily river series, SEAM recovers the seasons in all 3 four-year windows, and in 2 of them the fitted influence between the Columbia and the Willamette reverses with the season, as hydrological studies describe.

1 INTRODUCTION

During the Columbia River's spring flood, its high water pushes up the Willamette at Portland; in winter, Willamette floods push up the Columbia (Dempster & Lutz, 1968; Wherry et al., 2019). In free-flowing traffic a disturbance travels downstream, while in congestion it travels upstream (Lighthill & Whitham, 1955; Richards, 1956; Newell, 1993). Such systems switch between a few operating states, and their causal relations, even the direction of an effect, can change from one state to the next. An analyst wants to know when the system changes state, which causal relations change, and what an intervention would do in each state.

We study these questions for *one* multivariate time series whose state boundaries are unknown and whose states, which we call *regimes*, may recur. The boundaries and the causal graphs have to be learned together, because each helps to explain the other. A method with one causal order for all regimes cannot represent $a \to b$ in one regime and $b \to a$ in another: one of the two effects is forced to zero, so the method's normalized error on the reversed effects is at least $1/2$, even with the true regimes (Proposition 2). A model with a graph per regime, on the other hand, always fits at least as well as a single model, so a better fit alone cannot tell a change of mechanism from ordinary variation.

Existing methods address parts of this problem. Markov-switching VARs learn recurring regimes but model reduced-form dynamics, not causal directions (Hamilton, 1989; Krolzig, 1997). Methods that learn one graph (Runge, 2020; Pamfil et al., 2020; Hyvärinen et al., 2010; Gong et al., 2023), or let mechanisms change under a common orientation (Huang et al., 2020; Mooij et al., 2020; Peters et al., 2016; Squires et al., 2020; Gamella et al., 2022), cannot represent a reversal; MCD learns separate graphs but assigns each whole series to one of them (Varambally et al., 2024). Closer to our setting,

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.

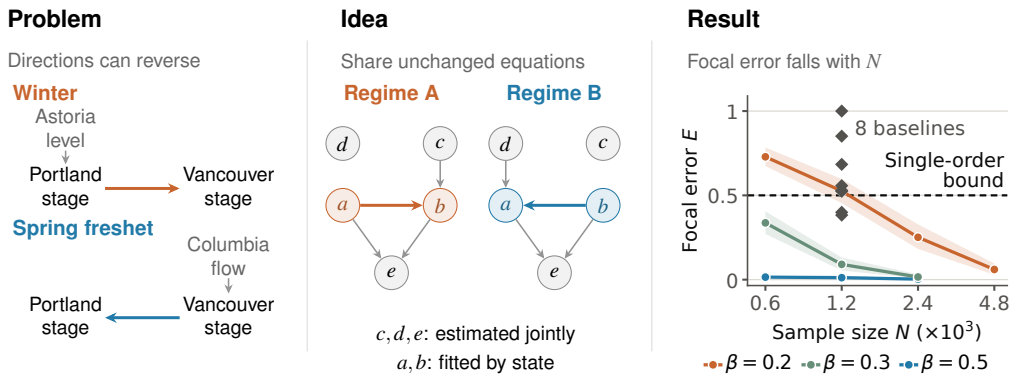


Figure 1: **Overview.** *Problem:* the influence between the two river stages reverses with the season. *Idea:* SEAM estimates the unchanged equations (c,d,e) jointly and the changed ones (a,b) per regime. *Result:* on the five-variable reversal design (Gaussian and river innovations), SEAM's error E falls with N, while any method with one causal order has $E \geq \frac{1}{2}$ (dashed; Proposition 2). 80 series per point, with 95% bootstrap bands; diamonds: the 8 baseline configurations at $N = 1200$, $\beta = 0.5$.

Regime-PCMCI learns regime-dependent lagged links (Saggioro et al., 2020), SPACETIME detects changes while keeping edge directions fixed (Mameche et al., 2025), and CASTOR, FANTOM and FlowMSM let same-time edges reverse across regimes (Rahmani & Frossard, 2025b;a; Hulsman et al., 2026). None of these five methods tests whether a change occurred: the number of regimes is an input, or is set by a window, an information criterion or an evidence bound, with no error guarantee (Appendix E).

SEAM (Shared Equations Across Mechanisms) gives each regime its own causal order, so effects can reverse; estimates each unchanged equation from both regimes, even when other equations and the causal order change; and tests whether anything changed. One score selects the labels, both graphs and the equations to share (Figure 1), and comparing the best one- and two-regime fits separates three outcomes: no change, a change in noise levels only, and a change in the structural equations.

Contributions.

- **Model and search** (Sections 2 and 3). One description-length score selects the regimes, a causal graph per regime and the equations to share; for fixed labels, the graph search is exact over all pairs of DAGs within the candidate parent sets.
- **A tested decision** (Section 3). SEAM reports no change, a scale change or a mechanism change, and it assesses the evidence with a resampling test that reruns the whole search on surrogate series.
- **Theory** (Section 4). Under the conditions of Theorem 1, the decision is consistent for Gaussian VAR series without a change and for separated mechanism changes, and a row-permutation variant of the test controls its level at every sample size for exchangeable rows without lags. We also bound the error of a single causal order, show why sharing needs sign alignment, and prove that, given the true regimes and structures, sharing does not increase the asymptotic variance of the effects.
- **Evidence** (Section 5). Experiments on prespecified designs with up to 20 variables, heavy tails and heteroscedastic noise measure change detection, localization of the changed equations, and graph and effect recovery against five regime-aware methods. On river data, SEAM finds the seasons and, in 2 of 3 windows, a reversal consistent with the documented one.

2 Problem Setup

Model. We observe one series $x_1, \ldots, x_N \in \mathbb{R}^d$. Each time point belongs to one of two regimes, $z_t \in \{0,1\}$, and the labels are unknown. In regime r the series follows a linear structural equation model (SEM) with p lags:

$$x_t = B_r^\top x_t + \sum_{\ell=1}^{p} A_{r,\ell}^\top x_{t-\ell} + \varepsilon_t, \qquad r = z_t. \tag{1}$$

The entry $B_r[i,j]$ is the coefficient of $x_{t,i}$ in the equation of $x_{t,j}$. The nonzero pattern of B_r is the *same-time graph* G_r, a directed acyclic graph (DAG).

Assumption 1 (Standing assumptions)**.** *(i) Regimes.* There are one or two regimes. With two, the labels are unknown and may alternate, as in $A|B|A|B$, in runs of at least $m = 5d(1+p)$ rows with at most eight switches, the limits used by the search (Appendix B). *(ii) Mechanisms.* Each regime follows (1) with B_r the coefficient matrix of a DAG. The two DAGs are unrestricted: they need not share a causal order, and an edge may point in opposite directions in the two. *(iii) Innovations.* ε_t is independent of the past and has independent coordinates with mean zero and regime-specific variances $\omega_{r,1},\dots,\omega_{r,d}$. No distributional family is assumed; the score uses a Gaussian working likelihood (Section 5.4). *(iv) Lags.* The same lag order $p \le 4$ holds in both regimes and is large enough that the one-regime residuals are serially uncorrelated.

Targets. We want to know whether there is a change ($K = 2$ regimes rather than $K = 1$), the labels z_t, the graphs G_0 and G_1, and the interventional effects in each regime,

$$\theta^{(r)}_{i\to j} \;=\; \big[(I - B_r)^{-1}\big]_{ij}, \tag{2}$$

the change in $x_{t,j}$ per unit intervention on $x_{t,i}$ in regime r, with the past held fixed. A *reversal* is a pair (a,b) with $a \to b$ in G_0 and $b \to a$ in G_1, so that $\theta^{(0)}_{a\to b} \neq 0 = \theta^{(1)}_{a\to b}$ and $\theta^{(1)}_{b\to a} \neq 0 = \theta^{(0)}_{b\to a}$. We call (a,b) the *focal pair* and its two nonzero effects, each of size β, the *focal effects*, and we score a fit by its normalized error on them:

$$E \;=\; \tfrac{1}{2}\Big[\big(\hat\theta^{(0)}_{a\to b} - \theta^{(0)}_{a\to b}\big)^2 + \big(\hat\theta^{(1)}_{b\to a} - \theta^{(1)}_{b\to a}\big)^2\Big]/\beta^2. \tag{3}$$

Each true regime is compared with the fitted regime that contains most of its time points; a one-regime fit is compared with both.

What can change. We compare the two regimes one structural equation at a time. An equation consists of a variable's same-time and lagged parents, its coefficients and its innovation variance. It is *shared* if it is the same in both regimes, and it is then estimated once, from the rows of both. It is *graded* if its parents and coefficients agree but its innovation variance differs, as under a soft intervention that makes the variable more or less noisy. It is *regime-specific* if its parents or coefficients differ. The regime-specific equations are our answer to "what changed": in Figure 1 they are the equations of a and b.

Three possible answers. SEAM decides between three nested explanations of the series: *no change* (one law, $\mathscr{H}_1$); a *scale change* (one mechanism at two noise levels, $\mathscr{H}_0 \supset \mathscr{H}_1$), in which every equation keeps its coefficients, in the data's coordinates, while the innovation variances may differ; and a *mechanism change* (two mechanisms, $\mathscr{H}_2$), in which some equation is regime-specific. Let $L_\star$ be the expected per-row Gaussian log-likelihood of the true model at the true labels, and $L_1 \le L_0 \le L_\star$ its best values within $\mathscr{H}_1$ and $\mathscr{H}_0$. The gaps $\Delta = L_\star - L_1 \ge \kappa = L_\star - L_0 \ge 0$ measure, in nats per row, how far the series is from one law and from one mechanism. Δ governs detection and κ attribution to a mechanism change (Theorem 1): to a Gaussian score, a mechanism change is visible exactly when $\kappa > 0$.

Identifiability, and a trap in sharing. Within one Gaussian regime, observational data identify G_r only up to its Markov equivalence class (MEC) (Verma & Pearl, 1990); colliders, such as those in Figure 1, orient some edges. Sharing equations across regimes adds information, but rewarding equal coefficients can also favor a wrong graph. Suppose that one variable x_2 is replaced by $-x_2$ in one regime, so that every coefficient linking x_2 to another variable changes sign. Every equation that contains x_2 then has different coefficients in the two regimes, while a wrong graph that makes x_2 a sink can show fewer differences and earn a better sharing score (Theorem 3).

SEAM avoids this trap by comparing equations up to a global sign alignment $D = \mathrm{diag}(\pm 1)$, which we call the *gauge*. The map $x \mapsto Dx$ preserves the graph and every coefficient's magnitude (Lemma 5); comparing equations up to it defines the *reflection quotient*. Alignment only decides which equations may be pooled: graphs, effects and reported coefficient changes stay in the original coordinates, so a coefficient whose sign flips is pooled after alignment and still reported as a mechanism change.

3 SEAM

SEAM alternates between fitting the two regime graphs for fixed labels and updating the labels for fixed regime models, keeping an update only if the score rises. It compares the best two-regime fit with the best one-regime fit and reruns the same search on surrogate series to test the difference (Figure 2); the report then separates a change of coefficients from a change of noise level. Algorithm 1 (Appendix B) gives the full procedure and its settings.

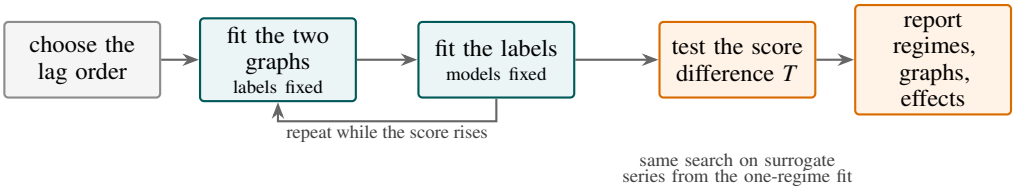


Figure 2: **How SEAM works.** Graph fitting and label updates alternate, and a refitted proposal is kept only when the score rises. The same search, run on surrogate series drawn from the one-regime fit, calibrates the score difference T. Teal: fitting steps. Orange: test and report.

The score. The score trades the fit against the complexity of the labels, the graphs and the sharing pattern. A *description* M fixes each variable's equation type (shared, graded or regime-specific), its parents in each regime, and the gauge D; coefficients and variances are profiled out. Each discrete choice is charged the length of a code that describes it. For labels z with $J(z)$ switches and a description M with $k(M)$ changed equations,

$$Q(z,M) \;=\; \ell(z,M) \;-\; \sum_{\text{parameters } \vartheta} \tfrac{1}{2}\log n_\vartheta \;-\; \sum_{\text{parent sets } P} |P|\log(d-1) \\ -\; \log\tbinom{d}{k(M)} \;-\; \log\tbinom{d}{|S(D)|} \;-\; J(z)\log N, \tag{4}$$

where ℓ is the Gaussian log-likelihood of the rows given their lagged past. A parameter costs $\frac{1}{2}\log n_\vartheta$, where n_ϑ is the number of rows that inform it: a shared coefficient is charged once on all rows, a regime-specific pair twice on fewer rows each. This is the Laplace approximation of the evidence under unit-information priors (Kass & Wasserman, 1995; Proposition 6). The other terms code each same-time parent set P, which k equations change (Scott & Berger, 2010), the set $S(D)$ of variables the gauge flips, and each switch. The parent-set charge depends only on the number of edges, so it never favors one member of a Markov equivalence class over another. With Q_2 the best two-regime score and Q_1 the best one-regime score, the evidence for a change is $T = Q_2 - Q_1$.

Lag order. The lag order p is chosen by BIC on the pooled VAR (Lütkepohl, 2005), as part of the one-regime model: too few lags would leave serial correlation that the test could mistake for regimes.

Graphs and labels. For fixed labels, maximizing Q over the two graphs is an integer program, as in integer-programming structure learning (Jaakkola et al., 2010; Cussens, 2011; Bartlett & Cussens, 2017). A binary variable selects, for variable j, an equation type b and a parent set P; assignment constraints give each variable one equation per regime, and cluster inequalities keep each regime's graph acyclic. The solution is optimal over *all* pairs of DAGs within the candidate parent sets (Proposition 7), and nothing ties the two graphs to a common order, so a reversal can be represented. For fixed regime models, each row's log-likelihood under each regime acts as an emission, and a two-state dynamic program finds the best labeling with runs of at least m rows and at most eight switches. The search starts from two data-driven labelings and keeps a refitted proposal only when Q increases (Appendices A.10 and B).

The test. From the one-regime fit $(\hat{A}, \hat{B})$, we form surrogate series by keeping the observed lags U and permuting whole residual rows: $X^* = (U\hat{A} + \hat{E}_\pi)(I - \hat{B})^{-1}$, the fixed-regressor bootstrap of Hansen (2000) applied to a structure-learning score, as in resampling tests for changepoints (Antoch & Hušková, 2001; Kirch, 2007). Rerunning the whole search on each surrogate gives a p-value $\hat{p}$. Theorem 1 gives the large-sample behavior of this test; its finite-sample guarantee is for a variant that permutes the observed rows themselves (Appendix A.7).

Decision and report. SEAM reports *no change* if $T \leq 0$, in which case no surrogate is drawn, or if $\hat{p} > \alpha$. Otherwise, on the selected labels, it re-solves the integer program for each candidate gauge, keeps the best fit, refines the labels while the score rises, and classifies each equation as shared, shared up to the gauge, graded or regime-specific. If some equation's coefficients differ between the regimes in the original coordinates, it reports a *mechanism change*, with the regimes, both graphs, the changed coefficients and the effects (2); otherwise it reports a *scale change*, with the one-regime graph and the ratio of innovation scales. T uses the identity gauge; sign alignment enters only the report. For $d \geq 10$, candidate parents are screened and T uses a local search over orders, also on the surrogates; the reported graphs are re-solved exactly within the screened family (Appendix B).

4 THEORY

Theorem 1 states when SEAM reports a mechanism change and identifies the population gap that decides it. Proposition 2, Theorem 3 and Proposition 4 explain the roles of separate causal orders, sign alignment and shared estimation. Proofs are in Appendix A, and Table 2 lists their assumptions.

Theorem 1 (Consistency of the report)**.** *Let Assumption 1 hold, with one regime in part (i) and two in parts (ii) and (iii).*

- *(i)* No mechanism change. *If the series is a stable Gaussian VAR(p) whose equations lie in the candidate family of Appendix B, and the dimension condition $S_\star < 1$ of Appendix A.8 holds (it does for every $d \leq 6$, $p \leq 4$), then with probability tending to one no two-regime description outscores the one-regime one, and* SEAM *reports no mechanism change. If the rows are exchangeable (no lag dependence, any innovation law), the row-permutation version of the test (Appendix A.7) reports a mechanism change with probability at most α, for every N.*
- *(ii)* Mechanism change. *If $\kappa > 0$ and the regularity, initialization, reporting and surrogate conditions of Appendix A.9 hold, then $T \geq N(\Delta - \xi) - o_p(N)$, where $\xi < \kappa$ is the per-row shortfall below $L_\star$ of a retained initialization, and* SEAM *reports a mechanism change with probability tending to one.*
- *(iii)* When $\kappa > 0$. $\kappa \geq \kappa_{\text{cov}}$, *an explicit function (28) of the regimes' conditional covariances, which is positive if the reversal has an* isolated anchor*: a variable that, given the past, is uncorrelated with every other variable in one regime but not in the other (Appendix A.9); the five-variable design of Section 5 has one. An unanchored reversal can have $\kappa = 0$; with Gaussian innovations it is then indistinguishable from a scale change under a common causal order, so no method that reports noise-level changes as one mechanism can report it.*

Part (i) holds because the switch charge bounds the largest gain of any labeling under the Gaussian null (Theorem 14): with probability tending to one, $T \leq 0$ and no surrogate is drawn. Part (ii) needs neither consistent labels nor a global optimum of the search; besides regularity, it uses score ascent from an initialization within κ per row of the population optimum (Proposition 21). Its leading term $N\Delta$ predicts that detection rates for all effect sizes and lengths fall on one curve (Figure 4). In the two-variable reversal, which has no anchor, $\Delta = \frac{1}{2}\log(1+\beta^4/4)$ while $\kappa = \Theta(\beta^6)$: for weak effects, attributing a change to the mechanism takes more data than detecting it.

Proposition 2 (Single-order floor)**.** *Let $a \to b$ in G_0 and $b \to a$ in G_1, with $|\theta^{(0)}_{a\to b}| = |\theta^{(1)}_{b\to a}| = \beta$. Any estimate whose fitted graphs admit a common causal order has $E \geq 1/2$ in (3). This includes one graph for all data, and pooled or multi-context methods with a fixed orientation. The bound holds for every sample size, even with the true labels.*

Theorem 3 (Reflection quotient)**.** *Beyond Assumption 1, let regime 0 be a Gaussian SEM without lags, faithful to a DAG G with connected skeleton. Let regime 1 be its reflection $x \mapsto D^\star x$ with $D^\star \notin \{\pm I\}$. The labels are known, and the regime shares are fixed in $(0,1)$. As $N \to \infty$:*

- *(a) If sharing is judged up to one global gauge, as in* SEAM*, every best-scoring description shares all d equations up to $\pm D^\star$, and its graph is in* MEC(G).
- *(b) If sharing is judged in data coordinates, best-scoring descriptions can leave* MEC(G) *and assign wrong effects. Take $G = \{0 \to 2, 1 \to 2, 0 \to 3, 2 \to 3\}$ with node 2 flipped. The wrong graph that makes node 2 a sink changes one equation instead of two. It has 13 parameters instead of 14, and it sets the effect of x_2 on x_3 to zero.*

The failure in (b) is a difference in dimension, not in fit, so no local penalty removes it. Over all 266 pairs of a connected four-node DAG and a nontrivial reflection, the data-coordinate optimum leaves the MEC in 22 cases and the quotient optimum in none.

Proposition 4 (Sharing efficiency)**.** *Fix the true regime labels, DAGs, lag supports, equation classes and gauge. Under the regularity conditions of Appendix A.6, the shared and graded fits of Section 3 have no greater asymptotic variance than separate-regime fits for any fixed linear contrast of the two complete effect matrices* (2)*; the regime-specific equations' asymptotic covariance is unchanged.*

5 EXPERIMENTS

Data. The evaluation followed a protocol fixed in advance, with separate seeds for development and evaluation (Appendix C). The reversal systems follow Figure 1: regime A has $a \to b \leftarrow c$, regime B has $b \to a \leftarrow d$, both regimes share the child e of a and b and stable lag-1 dynamics, and the focal coefficients have size β and random signs. The regimes follow the schedule $A|B$ or $A|B|A|B$. Innovations are either Gaussian, the law the score assumes, or resampled with replacement from the standardized residuals of recorded series, which are heavy-tailed and skewed: river discharge (CausalRivers; Stein et al., 2025), US macroeconomic series and an industrial testbed (SKAB; Katser & Kozitsin, 2020). Three quarters of the cells use recorded residuals. In half of the reversal cells, the innovations are heteroscedastic: each variable's noise scale is drawn log-uniformly from $[1/2, 2]$ and is the same in both regimes. For $d \in \{10, 20\}$, the focal pair sits inside a random DAG of expected degree two that both regimes share. The one-regime series are VAR(1) processes with persistence up to $\rho = 0.95$, under all four innovation laws, and VAR(2) processes under two of them, each at two lengths.

Baselines and metrics. We run five regime-aware methods with their authors' code (settings in Appendix C), and comparisons are paired on the series that both methods ran. Regime-PCMCI and FlowMSM require the number of regimes, and we give them the true number; a method told that there are two regimes cannot say that there is one. CASTOR and SPACETIME choose the number themselves. The FANTOM implementation sets the number of regimes to the series length over its window: 6 at the protocol's window, and 2 at a second window that we also report. SEAM given the true regimes shows what not knowing them costs. The prespecified primary endpoint is the focal error E of (3); a fit is a *focal success* when each true regime's matched fitted regime orients the focal pair correctly. Three further metrics were computed afterwards from the same frozen fits (Appendix F): *localization F1* for the set of equations declared changed, where an equation counts as changed when its parent set differs between the two fitted regimes, a rule that applies to every method; the *SHD* of each regime's same-time graph; and the *effect error* $\|\hat{\Theta}_r - \Theta_r\|_F / \|\Theta_r\|_F$ over the off-diagonal entries of the full effect matrix $\Theta_r = (I - B_r)^{-1}$, which counts spurious effects as well as missed ones. A run that fails or does not finish counts as $E = 1$ and as the empty output for the other metrics.

5.1 FALSE CHANGE REPORTS

SEAM rarely reported a mechanism change on series without one: 5 of 1400 one-regime series (0.36%; 95% interval [0.12, 0.83]%), 0 of the 400 with Gaussian innovations (0.00%, [0.00, 0.92]%) and 5 of the 1000 with recorded residuals (0.50%, [0.16, 1.16]%), and at most 1 in 50 in any cell. On the Gaussian series, the two-regime description never outscored the one-regime fit ($T \leq 0$ on all 400), as Theorem 1(i) predicts. With recorded residuals, $T > 0$ on 423 of 1000 series and the test rejected on 42 of them: 4.2% of the 1000 and 3.0% of all 1400, for a nominal level of 5%.

The two reporting conditions keep the final rate low. The test-only ablation, which drops both conditions and so also runs when $T \leq 0$, rejects on 72 series (5.14%); requiring $T > 0$ removes 30, and requiring a changed coefficient turns 37 of the other 42 into scale changes, since chance clusters of large innovations can resemble a change in noise level. On the reversal series, the conditions lower detection only from 0.972 to 0.967 and leave the effect error unchanged. On the 600 VAR(1) series with $N = 1200$ that the baselines ran, CASTOR, which chooses the number of regimes, returns more than one on 61%, and FlowMSM, given two, splits every series; its graphs and effect matrices are then further from the truth (SHD 0.91–0.96, effect error 0.143–0.279) than SEAM's (0.02 and 0.045; Table 27).

A change in noise level alone should be reported as a scale change. On 600 series with fixed graphs and coefficients in which one root variable's innovation variance rises from 1 to 4 (300 with one switch, 300 recurring), SEAM detects the change every time, reports a scale change in 578 (96.3%) and a mechanism change in 22 (3.7%; exact 95% interval 2.3–5.5%), and classifies the root's equation as graded in 584 series, with precision 96.1% (Appendix D.8).

5.2 Changed equations, graphs and effects

SEAM recovers the reversed effects with much lower error than every baseline. On the 320 reversal series with $\beta = 0.5$, its focal error is 0.016, compared with 0.015 given the true regimes; its focal success is 0.98 and its worst cell has $E = 0.037$ (Table 1). Every baseline configuration has mean error of at least 0.406, including Regime-PCMCI (1.000) and SPACETIME (0.684) on a 40-series subset (Table 24), and paired differences favor SEAM with intervals that exclude zero (Table 4).

Table 1: **Reversal systems with five variables** ($d = 5$, $N = 1200$, $\beta = 0.5$; 2 schedules $\times$ 4 innovation laws $\times$ {unit, heteroscedastic} scales $\times$ 20 series). The focal error E (3) is the primary endpoint; the other metrics are defined in Section 5. Brackets are 95% bootstrap intervals over series (seed clusters). FANTOM ran on the unit-scale cells only (Appendix C). The test-only ablation, FANTOM at window 600 and the $\beta = 0.3$ block are in Appendix F.

method	series	focal error E	localization F1	SHD	effect error
SEAM	320	0.016 [0.009, 0.023]	1.00 [1.00, 1.00]	0.05 [0.03, 0.09]	0.080 [0.070, 0.094]
SEAM, true regimes	320	0.015 [0.009, 0.023]	1.00 [1.00, 1.00]	0.05 [0.03, 0.08]	0.079 [0.069, 0.093]
CASTOR ($\tau = 0.1$)	320	0.692 [0.630, 0.760]	0.25 [0.21, 0.29]	2.32 [2.21, 2.43]	0.881 [0.851, 0.911]
CASTOR (authors)	320	0.951 [0.906, 0.998]	0.16 [0.13, 0.19]	3.21 [3.13, 3.28]	1.001 [0.979, 1.022]
FlowMSM + DYNOTEARS	320	0.537 [0.481, 0.601]	0.40 [0.37, 0.44]	3.57 [3.43, 3.72]	0.809 [0.774, 0.845]
FlowMSM + VARLiNGAM	320	0.520 [0.462, 0.588]	0.48 [0.43, 0.53]	2.31 [2.09, 2.53]	0.587 [0.540, 0.634]
FANTOM	80	0.406 [0.328, 0.486]	0.58 [0.58, 0.59]	3.37 [3.16, 3.60]	0.824 [0.774, 0.874]

SEAM also identifies the changed equations (localization F1 0.998) and recovers the same-time graphs and full effect matrices (Table 1); its advantage over every baseline holds for all three metrics (Tables 21 and 24). Over all 640 series of the block, including $\beta = 0.3$, it reaches localization F1 0.951, directed-edge F1 0.971 and effect error 0.117 (Appendix F). With $d = 10$ and $d = 20$ variables, SEAM finds the changed equations with F1 1.000 and recovers the graphs with SHD 0.06 and 0.32 and the effect matrices with error 0.039 and 0.046; windowed CASTOR and FlowMSM reach SHD 1.6–15.6 and effect error 0.31–0.57 (Table 22).

Focal error stays small across noise types: 0.013 with Gaussian innovations, 0.016 with recorded residuals, 0.022 with heteroscedastic innovations and 0.009 with unit scales. At $\beta = 0.3$, focal error rises to 0.133 and localization F1 falls to 0.904. Figure 8 locates the loss: with the true regimes, localization is essentially perfect at both effect sizes; with learned regimes, the loss is concentrated in the 7% of weak-effect series reported as one regime. On the 298 series where a change is detected, localization F1 is 0.970 (0.973 with the true regimes) and effect error 0.131 (0.126).

5.3 The benefit of sharing

Sharing improves the estimate of the full effect matrix, and the gain comes from the unchanged equations (Figure 3a). Given the true regimes but learning the graphs and equation classes, sharing reduces effect error by 23% across 640 series, with gains in every stratum (Table 25). The coefficient error of shared equations falls on 619/640 series, while that of changed equations is nearly unchanged, as Proposition 4 predicts. The focal effects belong to changed equations and do not improve: their paired error difference is $+0.006$ $[-0.001, +0.014]$, and $+0.016$ with heteroscedastic innovations.

Sign alignment keeps sharing safe under reflections (Figure 3b). On the reflection design with the true regimes, 94% of fitted regime graphs lie in the true MEC, compared with 4% when sharing ignores sign alignment, and effect error falls from 0.548 to 0.062. Without sharing, the graphs remain good but effect error is higher (0.098; Table 26). Both the aligned and the no-sharing fits localize the sign changes, scored as coefficient changes in the original coordinates, with F1 0.97.

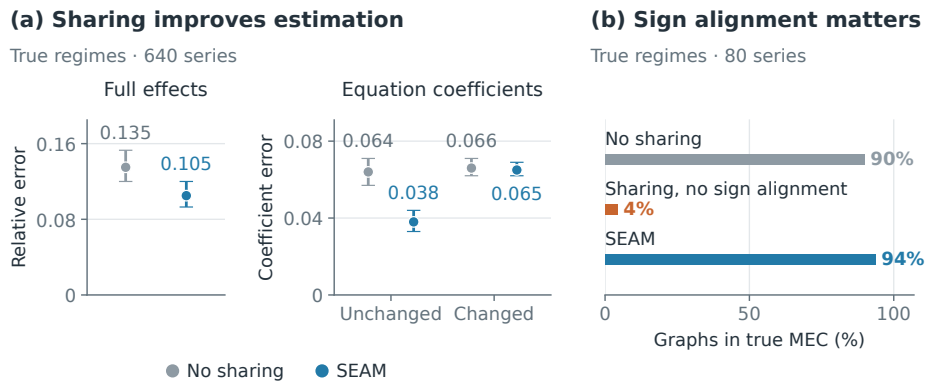


Figure 3: **Sharing improves estimation; sign alignment preserves graph recovery.** Both panels use the true regimes and learn the graphs and equation classes. **(a)** Errors with and without sharing on 640 reversal series; bars are 95% seed-cluster bootstrap intervals. **(b)** Fraction of regime graphs in the true Markov equivalence class on 80 series of the reflection design of Theorem 3, averaged over both regimes.

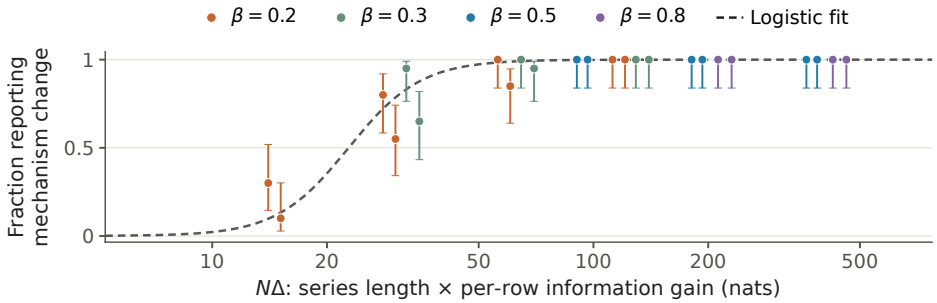


Figure 4: **Detection follows $N\Delta$ under Gaussian innovations.** Each point is a cell of 20 series with unit innovation scales (480 series in total); bars are 95% Wilson intervals. Horizontal positions are geometric means of $N\Delta$. The dashed logistic fit is evaluated at $N = 1200$.

5.4 Power and Robustness

Under Gaussian innovations, detection rates for all effect sizes and lengths fall on one curve in $N\Delta$, as Theorem 1(ii) predicts (Figure 4): the logistic coefficient of $\log(N\Delta)$ is 4.59, and separate effect-size terms add nothing ($p = 0.70$).

Under t_3 innovations, the Gaussian statistic loses power on 40 reversal series. A Student-t variant of the statistic, not part of SEAM because it costs 5–10 times more (Appendix D.7), raises detection from 0.60 to 1.00 and localization F1 from 0.575 to 1.000, lowers effect error from 0.321 to 0.091, and reports a change on 1 of 100 one-regime t_3 series. On the benchmarks of CASTOR, Regime-PCMCI and SPACETIME, SEAM recovers the regimes at least as well as each method does on its own data (Appendix D.5); on Regime-PCMCI's tutorial system, whose coefficients are known, SEAM's report names exactly the two equations whose coefficients change, on 19 of 20 series. CASTOR's generator, with dense and unrelated regime graphs, marks the limit of SEAM's advantage: SEAM matches CASTOR's effect error (0.364 vs. 0.351) and localizes changes with F1 0.968 vs. 0.943, but recovers the graphs less well (SHD 5.38 vs. 1.90).

5.5 Seasonal Reversals in the Columbia–Willamette System

We return to the Columbia–Willamette reversal of Section 1 (Dempster & Lutz, 1968; Wherry et al., 2019) with five public daily series: USGS stages at Portland (Willamette) and Vancouver (Columbia), discharges at Salem (Willamette) and The Dalles (Columbia), and the NOAA water level at Astoria, near the Columbia's mouth. Without tuning, SEAM reports a mechanism change in each of the 3

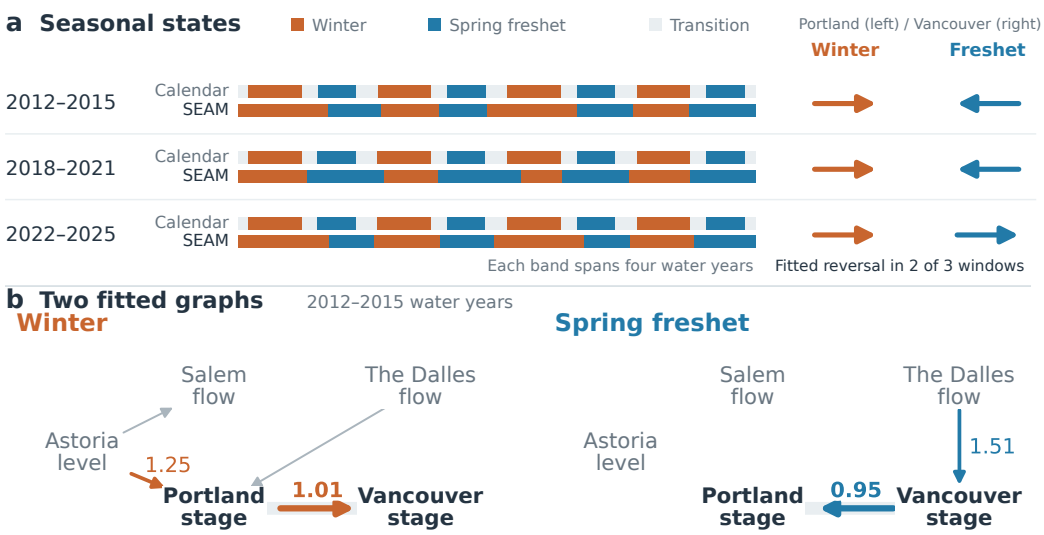


Figure 5: **Fitted seasonal reversals in the Columbia–Willamette system. (a)** Calendar seasons and SEAM's fitted regimes in the three four-year windows, and the fitted direction of the Portland–Vancouver edge in each season. **(b)** The fitted same-day graphs for water years (WY) 2012–2015; edges with $|\text{coefficient}| < 0.05$ are omitted. The edge reverses with the season in two of the three windows, as hydrological studies describe; Figure 7 shows all three windows (WY2016–2017 have gaps in the Portland record).

non-overlapping four-year windows of water years (October to September), each at $p = 0.025$, the smallest value its 39 surrogates allow. The fitted regimes match the seasons (Figure 5): balanced accuracy is 0.80–0.90 against a label based on the ratio of the two rivers' discharges and 0.94–0.95 against the calendar.

In 2 of the 3 windows, the fitted graphs reverse as hydrological studies describe: in the freshet regime the Columbia's flow drives the Vancouver stage, which drives Portland; in winter the Astoria water level drives Portland, which drives Vancouver. The third window has the smallest freshets in the record. There, the freshet regime has no edge from the Columbia's flow, and both seasons keep the Portland-to-Vancouver direction. No evaluated CASTOR, Regime-PCMCI or SPACETIME configuration reports this reversal in any window (Appendix D.9).

6 Limitations and Conclusion

Limitations. The model is linear with two regimes; segmenting three or more is not implemented. Same-time orientations come from what a Gaussian likelihood can distinguish, so a reversal between two otherwise unconnected variables is not identifiable. We analyze one real system (Section 5.5).

Conclusion. SEAM answers a monitoring question: does a time series reflect one causal mechanism, or recurring regimes with different effects, and if so, which equations changed? It learns the regimes and their graphs together, estimates the unchanged equations from both regimes, and tests the decision to report a change. On the evaluated reversals it identifies the changed equations and recovers each regime's graph and effect matrix, and on series without a change it rarely reports one.

Reproducibility Statement

The supplementary archive contains the SEAM implementation (`seam_v3/`; Appendix B), the data generators, the baseline wrappers, the evaluation code and every stored record. `reproduce.py` regenerates every number, table and figure from the records and re-runs a sample of fits from their seeds. Every data set is regenerated from its seed, and re-executing the frozen estimator with the stored permutation p-values reproduces all 7,980 stored SEAM fits (every arm except the no-sharing

ablation; Appendix F). Appendix C gives the protocol, the seed assignments and the departures from the protocol.

ETHICS STATEMENT

The study uses public, non-personal data only: river stage, discharge and water-level records of the U.S. Geological Survey and NOAA, and residuals of the CausalRivers, statsmodels macroeconomic and SKAB series, used under their licenses, which the supplementary archive reproduces with attributions. No human subjects or personal data are involved. A fitted reversal is a statement within the model class of Section 2 and should be checked against domain knowledge before it informs decisions.

AI USE STATEMENT

In this work, we used generative AI tools (a large language model assistant) to help develop the theoretical framework, formulate the mathematical claims, provide ingredients for their proofs and assist in writing the proofs; to refine hypotheses and to design, and give feedback on, the method and the experiments; to implement the method, the baseline wrappers and the data generators, and to generate the synthetic data sets with them; to clean and reformat the recorded series; and to interpret the results. We have not used generative AI tools for translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to write and edit code, suggest experimental parameters, run and monitor the experiments on cloud machines, create the figures, search for and summarize related literature, format references, suggest the structure of the paper, and draft and edit its text. We have reviewed all AI-assisted work: each stated result was checked by the authors step by step, the counterexample of Theorem 3 by exhaustive enumeration, the dimension condition of Theorem 1 by exact arithmetic and its covariance certificate on the confirmatory motif by numerical minimization over every order and gauge; unit tests check the implementation, including the integer program against brute-force enumeration; the confirmatory run used frozen code under a prespecified protocol whose departures are listed (Appendix C); every stored fit re-executes from its seed (Appendix F); and every reference was checked against its published record. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jaromír Antoch and Marie Hušková. Permutation tests in change point analysis. *Statistics & Probability Letters*, 53(1):37–46, 2001.

Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

Mark Bartlett and James Cussens. Integer linear programming for the Bayesian network structure learning problem. *Artificial Intelligence*, 244:258–271, 2017.

Julian Besag and Peter Clifford. Sequential Monte Carlo p-values. *Biometrika*, 78(2):301–304, 1991.

James Cussens. Bayesian network learning with cutting planes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

George R. Dempster, Jr. and Gale A. Lutz. Water-discharge determinations for the tidal reach of the Willamette River from Ross Island Bridge to Mile 10.3, Portland, Oregon. Water-Supply Paper 1586-H, U.S. Geological Survey, 1968.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems 21*, 2008.

Nir Friedman. The Bayesian structural EM algorithm. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

Juan L. Gamella, Armeen Taeb, Christina Heinze-Deml, and Peter Bühlmann. Characterization and greedy learning of Gaussian structural causal models under unknown interventions, 2022. arXiv:2211.14897.

Charles J. Geyer. On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 22(4): 1993–2010, 1994.

Alexander J. Gibberd and James D. B. Nelson. Regularized estimation of piecewise constant Gaussian graphical models: the group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634, 2017.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *International Conference on Learning Representations (ICLR)*, 2023.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

Bruce E. Hansen. Testing for structural change in conditional models. *Journal of Econometrics*, 97 (1):93–115, 2000.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Roel Hulsman, Carles Balsells-Rodas, and Sara Magliacane. Identifiable Markov switching models with instantaneous effects and exponential families. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2606.02231.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11: 1709–1731, 2010.

Tommi Jaakkola, David Sontag, Amir Globerson, and Marina Meila. Learning Bayesian network structure using LP relaxations. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Robert E. Kass and Larry Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431): 928–934, 1995.

Iurii D. Katser and Vyacheslav O. Kozitsin. SKAB – Skoltech Anomaly Benchmark. Kaggle, 2020.

Claudia Kirch. Block permutation principles for the change analysis of dependent data. *Journal of Statistical Planning and Inference*, 137(7):2453–2474, 2007.

Mladen Kolar and Eric P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6: 2069–2106, 2012.

Hans-Martin Krolzig. *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*, volume 454 of *Lecture Notes in Economics and Mathematical Systems*. Springer, 1997.

Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

Markku Lanne, Helmut Lütkepohl, and Katarzyna Maciejowska. Structural vector autoregressions with Markov switching. *Journal of Economic Dynamics and Control*, 34(2):121–131, 2010.

M. J. Lighthill and G. B. Whitham. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London A*, 229(1178):317–345, 1955.

Scott W. Linderman, Matthew J. Johnson, Andrew C. Miller, Ryan P. Adams, David M. Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

Sarah Mameche, Lénaïg Cornanguer, Urmi Ninad, and Jilles Vreeken. SPACETIME: Causal discovery from non-stationary time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19405–19413, 2025.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

G. F. Newell. A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B*, 27(4):281–287, 1993.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605, 2020.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems 35*, 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Martin Rabel and Jakob Runge. Context-specific causal graph discovery with unobserved contexts: Non-stationarity, regimes and spatio-temporal patterns, 2025. arXiv:2511.21537.

Abdellah Rahmani and Pascal Frossard. Flow based approach for dynamic temporal causal models with non-Gaussian or heteroscedastic noises. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 8300–8349, 2025a.

Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4546–4554, 2025b.

Paul I. Richards. Shock waves on the highway. *Operations Research*, 4(1):42–51, 1956.

Roberto Rigobon. Identification through heteroskedasticity. *Review of Economics and Statistics*, 85 (4):777–792, 2003.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019.

Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. Causal structure discovery from distributions arising from mixtures of DAGs. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *Journal of the American Statistical Association*, 117 (537):251–264, 2022.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11):113115, 2020.

James G. Scott and James O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619, 2010.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers – scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations (ICLR)*, 2025.

Sumanth Varambally, Yian Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49171–49202, 2024.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1990.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems 31*, 2018.

Susan A. Wherry, Tamara M. Wood, Hans R. Moritz, and Keith B. Duffy. Assessment of Columbia and Willamette River flood stage on the Columbia Corridor Levee System at Portland, Oregon, in a future climate. Scientific Investigations Report 2018–5161, U.S. Geological Survey, 2019.

Lei Zan, Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Beyond stationarity in time series: Discovering causal structures and latent regimes via Markov blankets, 2026. arXiv:2609.05150; AALTD workshop at ECML PKDD 2026.

A PROOFS

Throughout, regime r has $n_r = \pi_r N$ rows, and H_r denotes its mean raw second-moment matrix of (u_t, x_t), where u_t collects the lags. An equation of type b for variable j, with same-time parents P and lag set L, has a local cost $c_j(b, P)$. The cost is the negative profiled log-likelihood of the rows it covers plus its terms in (4).

Table 2: **Objects, assumptions and scope of the guarantees.** All results also assume Assumption 1.

Procedure or claim	Proved result and assumptions	Where it is tested
Implemented fixed-U test, one-mechanism null Theorem 1(i), Theorem 14	Stable Gaussian VAR(p) with positive-definite innovation covariance; equations in the candidate family of Appendix B; a one-regime fit within $o_p(\log N)$ of its optimum; $S_\star < 1$ (Table 3). Then $\Pr(T > 0) \to 0$, $\Pr(\hat{p} \le \alpha) \to 0$ for every fixed $\alpha < 1$, and $\Pr(\text{report two mechanisms}) \to 0$. An asymptotic result, not finite-sample level-α calibration of the surrogates, which hold the observed lags U fixed.	Section 5.1: $T \le 0$ on all 400 Gaussian series. With recorded residuals, which the Gaussian assumption does not cover, $T > 0$ on 423 of 1000 and the test rejected on 42. The screened search of the $d \ge 10$ experiments is outside this theorem.
Observed-row permutation test Theorem 11	An alternative test that reorders the observed rows and recomputes T; it is not the fixed-U procedure that Algorithm 1 selects. If the rows are exchangeable (no lag dependence, any law), $\Pr(\hat{p} \le \alpha) \le \alpha$ for every α and every N; the final report of two mechanisms is a subset of rejection (Remark 1).	Every test in Section 5 uses the fixed-U surrogates; this test is not run there. The two coincide only for an empty lag design, which the lag-order step never selects (Appendix A.7).
Separated two-mechanism alternative Theorem 1(ii), Theorem 18	$\kappa > 0$ alone is not enough. Also required: the conditions of Lemma 17 and a one-regime score of at most $NL_1 + o_p(N)$; a retained initialization with gap $\xi < \kappa$; a reporting step that loses at most $o_p(N)$ (it does for $d < 10$ and is assumed for $d \ge 10$); surrogate statistics whose positive parts are $o_p(N)$; and an attainable p-value, $(R+1)^{-1} \le \alpha$. Then $T \ge N(\Delta - \xi) - o_p(N)$, and two mechanisms are reported with probability tending to one.	Detection rates are measured in Sections 5.2 and 5.4 and Figure 4. The theorem needs neither consistent labels nor a global optimum of the search; it concerns the report, not the consistency of labels, graphs or effects.
Reversal identifiability; single-order limitation Theorem 1(iii), Proposition 2	$\kappa \ge \kappa_{\text{cov}}$ (28). An isolated anchor, a variable whose conditional covariance with every other variable is zero in one regime and not in the other, is a sufficient condition for a positive covariance gap (29), and the five-variable motif satisfies it; an unanchored reversal can have $\kappa = 0$. Proposition 2: $E \ge 1/2$ for every estimate whose fitted reversal graphs share one causal order, for every N, even with the true labels.	The motif of Section 5 meets the anchor condition; the river data of Section 5.5 are analyzed empirically.
Reflection quotient; sharing efficiency Theorem 3, Proposition 4	Theorem 3: regime 0 a Gaussian SEM without lags, faithful to a DAG with connected skeleton; regime 1 its exact reflection $x \mapsto D^\star x$; known labels and fixed regime shares; $N \to \infty$. Proposition 4: fixed true labels, DAGs, lag supports, equation classes and gauge, with the regularity of Appendix A.6; shared and graded fits have no greater asymptotic variance than separate fits for any fixed linear contrast of the two effect matrices.	The enumeration after Theorem 3 covers connected four-node DAGs. Section 5.3 measures sharing with true labels and learned graphs (Figure 3).

Counts are events in the recorded runs. Of the 1400 one-regime series of Section 5.1, the test rejected on 42 (3.0%), and 5 ended in a final report of two mechanisms (0.36%). The variance-only study of Appendix D.8 (578 of 600 series reported as scale changes, 22 as mechanism changes) tests attribution rather than detection.

A.1 Reflections

Lemma 5 (Reflections preserve the model class)**.** *Let $x = B^\top x + A^\top u + e$ with independent, sign-symmetric innovations independent of u. Let $D = \mathrm{diag}(d_1, \ldots, d_d)$ with $d_i \in \{\pm 1\}$. Put $D_u = D$ blockwise if u consists of lags of x, and $D_u = I$ if u is exogenous. Then $x' = Dx$ satisfies*

$$x' = (DBD)^\top x' + (D_u A D)^\top u' + e', \qquad u' = D_u u, \quad e' = De.$$

The innovations are independent with the same marginal laws. The graph, every coefficient magnitude and every innovation law are unchanged.

Proof. $x'_j = d_j x_j = \sum_i (d_i B_{ij} d_j) x'_i + \sum_l ((D_u)_l A_{lj} d_j) u'_l + d_j e_j$, because $d_i^2 = (D_u)_l^2 = 1$. Each $d_j e_j$ has the law of e_j by symmetry, and independence is preserved. □

A.2 The penalty is a block-local Laplace code

Proposition 6. *Fix one equation block with m coefficients and variance parameters that are either common, informed by n_B rows, or regime-specific, informed by n_r rows. Assume positive-definite limiting moments, $n_r/N \to \pi_r > 0$, and a smooth positive prior. Then*

$$-\log m_{block} = -\hat{\ell}_{block} + \tfrac{m}{2} \log n_B + \sum_{variances} \tfrac{1}{2} \log n_{(var)} + R_{block} + o_p(1).$$

Under the unit-information normal prior of the block (Kass & Wasserman, 1995), $R_{block} = o_p(1)$, and the sum over blocks is the continuous part of (4). A BIC with $\log N$ for every parameter corresponds to priors that are N/n times more diffuse. Relative to the local code, it charges a regime-specific equation with m coefficients and one variance an extra $\frac{1}{2}(m+1) \log \frac{1}{\pi(1-\pi)}$ nats over a shared one, where π is the share of regime 0, and so it shares too much.

Proof. This is Laplace's method. For a Gaussian regression, the information matrix of a block is block-diagonal in coefficients and variances at the maximum likelihood estimate. For a graded block the cross terms are $O_p(\sqrt{n})$ against $O(n)$ diagonal blocks, which changes the log-determinant by $O_p(n^{-1})$. The coefficient information is $\sum_r n_r H_{r,PP} / \sigma_r^2 = n_B \bar{I}_\beta$, and each variance has information $n_{(\mathrm{var})}/(2\sigma^4)$. Hence $\log\det I = m \log n_B + \sum \log n_{(\mathrm{var})} + \log\det \bar{I}$. For the unit-information prior, $-\log g(\hat{\theta}) = \frac{\dim}{2} \log 2\pi - \frac{1}{2} \log\det \bar{I} + o_p(1)$, and the constants cancel. □

A.3 The structure step is exact

Proposition 7 (Exact structure step)**.** *Fix the labels, the candidate parent sets and the terms of (4). The integer points of the program in Section 3 correspond one-to-one to descriptions whose two graphs are DAGs, and the objective equals $-Q$. Its optimum is therefore the best description over* all *pairs of DAGs. Solving with any subset of the cycle cuts, and adding the cuts that the solution violates, terminates at this optimum.*

Proof. A digraph is acyclic if and only if every cluster C of at least two nodes contains a node with no parent in C (Jaakkola et al., 2010). Equivalently,

$$\sum_{j \in C} \mathbf{1}[\mathrm{pa}(j) \cap C \neq \emptyset] \leq |C| - 1.$$

The program has one binary variable $x[j, b, P]$ for each variable j, equation type b and candidate parent set P. It imposes two families of constraints:

- *assignment*: $\sum_{b \ni r} \sum_P x[j, b, P] = 1$ for each regime r and variable j;
- *acyclicity*: $\sum_{j \in C} \sum_{b \ni r} \sum_{P \cap C \neq \emptyset} x[j, b, P] \leq |C| - 1$ for each regime r and cluster C.

Here $b \ni r$ means that the equation type b covers regime r. A shared or graded equation covers both regimes, a regime-specific one covers its own.

Fix an integer point. Assignment selects, for each (j, r), exactly one active variable whose type covers r. This defines the parent sets $\mathrm{pa}_r(j)$. The acyclicity constraints of regime r then say that G_r is acyclic.

Conversely, every description with acyclic G_0, G_1 sets exactly its own variables to one and satisfies all constraints. The objective is additive over the chosen equations; the multiplicity code enters through a one-hot count of non-shared equations. It therefore equals $-Q$ up to the label-dependent constant $J(z)\log N$.

For the lazy cuts, suppose a solution is optimal for a subset of the acyclicity constraints and has acyclic graphs. It is feasible for the full program, hence optimal. Otherwise, a strongly connected component of a cyclic regime graph violates the constraint for its node set. The loop terminates because there are finitely many constraints. □

A.4 Proof of Proposition 2 (single-order floor)

Suppose the fitted graphs $\hat{G}_0, \hat{G}_1$ admit a common topological order τ, and say a precedes b in τ. Then b is not an ancestor of a in $\hat{G}_1$. Every directed path from b to a is absent, so $\hat{\theta}^{(1)}_{b\to a} = [(I-\hat{B}_1)^{-1}]_{ba} = 0$. The second term of (3) equals $(\theta^{(1)}_{b\to a})^2/\beta^2 = 1$, so $E \geq 1/2$. If b precedes a, the same argument applies to the first term. A single graph for all data, or any method with one orientation across contexts, is a special case. The argument never uses the sample size or the labels. □

A.5 Proof of Theorem 3 (reflection quotient)

Consistency principle. Let $\ell(M)$ be the expected negative log-likelihood per row of the best fit within description M. Let $\dim(M) = D_p + D_s$ count the pooled and the regime-specific parameters; a graded equation pools its coefficients. By Proposition 6, the cost of M (minus its score (4)) is

$$N\ell(M) + \tfrac{1}{2}\dim(M)\log N + O(1).$$

The discrete codes are $O(1)$. A description that is not exact, meaning $\ell(M)$ exceeds its minimum, loses at rate N. Among exact descriptions, those of minimal dimension win, and the $O(1)$ terms break ties.

(a) Let M be exact. Its regime-0 graph G'_0 is an I-map of P_0 with $|E(G'_0)| + d$ parameters. Faithfulness implies $|E(G'_0)| \geq |E(G)|$, with equality if and only if $G'_0 \in \mathrm{MEC}(G)$:

- every adjacency of G is present in every I-map;
- an I-map with the skeleton of G has exactly its v-structures, since otherwise it would imply a separation that P_0 does not satisfy.

An equation that is not pooled carries at least one extra parameter, its regime-1 variance. Hence

$$\dim(M) \geq |E(G)| + d + \#\{\text{non-pooled equations}\},$$

with equality only if $G'_0 \in \mathrm{MEC}(G)$.

The description "G with all equations pooled up to $D^\star$" is exact by Lemma 5 and attains $|E(G)| + d$. So minimal descriptions pool every equation and have their graph in $\mathrm{MEC}(G)$. If every equation is pooled up to a gauge g, exactness requires $g\Sigma_0 g = D^\star\Sigma_0 D^\star$. If $gD^\star \neq \pm I$, its sign classes split the nodes into two nonempty parts. The connected skeleton has an edge u–v between the parts, and faithfulness gives $\Sigma_{0,uv} \neq 0$, a contradiction. So $g = \pm D^\star$.

(b) The generating description pools the equations of 0 and 1, which costs $1+1$ parameters. It keeps 2 (parents $0,1$) and 3 (parents $0,2$) regime-specific, at $3+3$ each: 14 parameters in total. In the alternative, x_3 is regressed on (x_0, x_1), with coefficients $(\beta_{03} + \beta_{02}\beta_{23}, \beta_{12}\beta_{23})$ and residual variance $\beta_{23}^2\omega_2 + \omega_3$. Every term involving node 2 contains d_2 twice, so this equation is identical in both regimes and can be pooled (3 parameters). With $0,1$ pooled $(1+1)$ and 2 regressed on (x_0, x_1, x_3) regime-specific $(4+4)$, the total is 13. The alternative is exact. It has the v-structure $0 \to 3 \leftarrow 1$, which G lacks, and x_2 has no descendants.

(c) The enumeration is `experiments/v2theory/theorem_r_search.py` in the code release. It takes the exact optimum over all pairs of orderings.

Candidate gauges. $S(g) = \log\det(\pi_0\Sigma_0 + \pi_1 g\Sigma_1 g)$ is minimized over sign patterns exactly at $g = \pm D^\star$, by strict concavity of $\log\det$ and the argument of (a). So the reporting step's candidate list (Appendix B) contains $\pm D^\star$ with probability tending to one. □

A.6 Proof of Proposition 4 (sharing efficiency)

The argument is the classical one for equality-constrained estimation (Geyer, 1994), specialized to SEAM's structural fits: two regimes with their own causal orders, pooled, graded and regime-specific equations in one fixed gauge, and the complete effect maps $\Theta_r = (I - B_r)^{-1}$. Its content here is the martingale calculation of the score covariance under lags and non-Gaussian innovations, the exact projection form of the existing profile fits, and the propagation to the effects.

Setting and conditions. Index by $a = (r, j)$ the equation of variable j in regime r, and write it as $x_{t,j} = v_{t,a}^\top b_{0,a} + \varepsilon_{t,j}$ for $z_t = r$, where $v_{t,a}$ stacks its same-time and lagged parents. For a DAG, $\det(I - B_r) = 1$, so the Gaussian working log-likelihood is the sum of the equation log-likelihoods. Let b stack every b_a, one copy per regime; empty blocks are allowed.

(i) *Fixed correct description.* The lag order, both DAGs, the lag supports, the equation classes and the gauge are fixed and correct; the regimes may have different topological orders. The true labels are deterministic with $n_r/N \to \pi_r \in (0,1)$. The correct equalities among coefficients are $Rb = 0$ with R of full row rank and $Rb_0 = 0$; the gauge contributes signed equalities to R. Without constraints the correction below is zero.

(ii) *Innovations.* With respect to the filtration of past observations and innovations, ε_t is independent of the past with independent coordinates, conditional mean zero and conditional variances $\omega_{rj} > 0$ depending only on the active regime; innovations and regressors have uniformly bounded $(4+\delta)$-th moments for some $\delta > 0$, and the initial conditions are independent of later innovations.

(iii) *Regressor limits.* For every nonempty block, $n_r^{-1} \sum_{t:z_t=r} v_{t,a} v_{t,a}^\top \to_p G_a \succ 0$. Write $H = \mathrm{blockdiag}_a(\pi_r G_a/\omega_{rj})$ and $V = H^{-1}$. Stable reduced-form dynamics in each regime with a bounded number of switches, positive regime proportions and finite-moment initial conditions satisfy (ii) and (iii).

(iv) *Fits.* The estimates are the exact local profiles of Appendix B: pooled least squares for a shared equation, which also equates its two innovation variances; separate least squares for a regime-specific equation; and common coefficients with variances profiled over their ratio for a graded equation. The discrete charges of (4) are constant once labels and description are fixed, so these are the fits the score selects. Columns are scaled by their root mean squares before fitting; the equivariance to column scaling maps the fits back, and every statement below is in the original coordinates.

Theorem 8 (Coefficient and complete-effect efficiency)**.** *Under (i)–(iv), let $\hat{b}_U$ be the separate-regime least-squares estimate and $\hat{b}_S$ the sharing estimate under the correct description. Then the profiled variances are consistent, $\sqrt{N}(\hat{b}_U - b_0) \Rightarrow \mathcal{N}(0, V)$ and $\sqrt{N}(\hat{b}_S - b_0) \Rightarrow \mathcal{N}(0, V_S)$ with*

$$V_S = V - VR^\top (RVR^\top)^{-1} RV \preceq V. \tag{5}$$

Let $h(b)$ stack $\mathrm{vec}(\Theta_0)$ and $\mathrm{vec}(\Theta_1)$, or their off-diagonal entries, and $J = Dh(b_0)$. Then $\sqrt{N}\{h(\hat{b}_A) - h(b_0)\} \Rightarrow \mathcal{N}(0, JV_A J^\top)$ for $A \in \{U, S\}$, $JV_S J^\top \preceq JVJ^\top$, and for a contrast $a^\top h$ the asymptotic variance decreases by $\|(RVR^\top)^{-1/2} RVJ^\top a\|^2$, which is positive exactly when $RVJ^\top a \neq 0$.

Proof. Separate estimation. At the truth, the stacked score is $s_t = (\mathbf{1}\{z_t = r\} v_{t,a} \varepsilon_{t,j}/\omega_{rj})_a$. Each $\varepsilon_{t,j}$ is independent of its parents and of the past, so $\mathbb{E}[s_t \mid \mathscr{F}_{t-1}] = 0$. Distinct equations at one time point have zero conditional cross-covariance: order the active DAG topologically and take j later than k; in $v_{t,(r,j)} \varepsilon_{t,j} \varepsilon_{t,k} v_{t,(r,k)}^\top$ every factor except $\varepsilon_{t,j}$ is measurable with respect to the past and the innovations preceding j, so integrating $\varepsilon_{t,j}$ gives zero; cross-regime blocks vanish because one regime is active. The conditional covariance of s_t is therefore block-diagonal with blocks $\mathbf{1}\{z_t = r\}\mathbb{E}[v_{t,a} v_{t,a}^\top \mid \mathscr{F}_{t-1}]/\omega_{rj}$; by the martingale law of large numbers and (iii) its average converges to H, the moment bound gives the conditional Lindeberg condition, and the martingale central limit theorem gives $N^{-1/2} \sum_t s_t \Rightarrow \mathcal{N}(0, H)$. The normal equations give $\sqrt{N}(\hat{b}_U - b_0) = H_N^{-1} N^{-1/2} \sum_t s_t$ with $H_N \to_p H$, hence the first limit.

Consistency of the sharing fit. Let $L_{a,N} = N^{-1} \sum_{t:z_t=r} v_{t,a} v_{t,a}^\top \to_p \pi_r G_a$ and $c_{a,N} = N^{-1} \sum_{t:z_t=r} (x_{t,j} - v_{t,a}^\top \hat{b}_{U,a})^2 \to_p \pi_r \omega_{rj}$, so that $\mathrm{RSS}_{a,N}(b_a)/N = c_{a,N} + (b_a - \hat{b}_{U,a})^\top L_{a,N} (b_a - \hat{b}_{U,a})$ exactly. Group the blocks by fitted variance: the two regimes of a shared equation form one group, every other

block is its own group; by correctness all blocks of a group have the same true variance ω_g. With $n_g = \sum_{a \in g} n_{r(a)}$ and $\rho_{g,N} = n_g/N \to \rho_g > 0$, the profiled negative log-likelihood per row is, up to constants, $f_N(b) = \frac{1}{2}\sum_g \rho_{g,N} \log\{\mathrm{RSS}_{g,N}(b)/n_g\}$ on $Rb = 0$, and it converges uniformly on compacts to $f(b) = \frac{1}{2}\sum_g \rho_g \log\{\omega_g + \rho_g^{-1}\sum_{a \in g} \pi_{r(a)}(b_a - b_{0,a})^\top G_a (b_a - b_{0,a})\}$, whose constrained minimizer is unique at b_0. With probability tending to one every nonempty $L_{a,N}$ has eigenvalues bounded below, every $c_{a,N}$ and $\rho_{g,N}$ is bounded below and every $\hat{b}_{U,a}$ is bounded, so a diverging block makes its group's term diverge while the others stay bounded below; the minimizers are therefore stochastically bounded, and uniform convergence gives $\hat{b}_S \to_p b_0$ and $\hat{\omega}_{S,g} = \mathrm{RSS}_{g,N}(\hat{b}_S)/n_g \to_p \omega_g$ for shared, graded and regime-specific groups alike.

Projection. At the fitted variances define $\hat{H}_S = \mathrm{blockdiag}_a\big((N\hat{\omega}_{S,rj})^{-1}\sum_{t:z_t=r} v_{t,a}v_{t,a}^\top\big)$. With these weights held fixed, the joint maximizer minimizes a weighted residual sum of squares subject to $Rb = 0$; its unconstrained minimizer is still $\hat{b}_U$, because the weight is constant within each block, and the constrained normal equations give the exact identity

$$\hat{b}_S = \hat{b}_U - \hat{H}_S^{-1}R^\top(R\hat{H}_S^{-1}R^\top)^{-1}R\hat{b}_U,$$

which for shared equations is their pooled least squares and for graded equations their weighted least squares at the profiled ratio. Since $\hat{H}_S \to_p H$ and $Rb_0 = 0$, $\sqrt{N}(\hat{b}_S - b_0) = P\sqrt{N}(\hat{b}_U - b_0) + o_p(1)$ with $P = I - VR^\top(RVR^\top)^{-1}R$, and $PVP^\top = V_S$. Only consistent weights were used, so the identity holds for non-Gaussian innovations under (ii).

Effects. Differentiating $(I - B_r)\Theta_r = I$ gives $d\,\mathrm{vec}(\Theta_r) = (\Theta_r^\top \otimes \Theta_r)\,d\,\mathrm{vec}(B_r)$; composing with the coordinate map from b to B_r, and with the off-diagonal selector if used, gives J, and the delta method gives the limits. Congruence preserves the semidefinite order, and $a^\top J(V - V_S)J^\top a = \|(RVR^\top)^{-1/2}RVJ^\top a\|^2$. □

Corollary 9 (Where the gain sits)**.** *V is block-diagonal and R has nonzero columns only on shared and graded blocks, so V_S agrees with V on every regime-specific block: sharing leaves the regime-specific equations' asymptotic covariance unchanged. A contrast $a^\top h$ whose Jacobian $J^\top a$ is supported on regime-specific blocks alone has $RVJ^\top a = 0$ and gains nothing; in the reversal designs of Section 5 the focal effects $\theta^{(0)}_{a\to b}$ and $\theta^{(1)}_{b\to a}$ are direct coefficients of regime-specific equations, so Proposition 4 predicts no gain on the focal endpoint and a gain on the shared equations and on every effect that passes through them. For one shared equation, the sharing fit's asymptotic covariance in its gauge coordinates is $(\pi_0 G_{0j}/\omega_{0j} + \pi_1 G_{1j}/\omega_{1j})^{-1}$, no larger than either separate covariance $\omega_{rj}G_{rj}^{-1}/\pi_r$, and half of it with equal moments, variances and regime shares; for a graded equation the profiled ratio supplies the inverse-variance weights.*

Corollary 10 (Quadratic risk)**.** *For fixed $W \succeq 0$, if the scaled squared errors are uniformly integrable, $N\mathbb{E}[\{h(\hat{b}_A) - h(b_0)\}^\top W\{h(\hat{b}_A) - h(b_0)\}] \to \mathrm{tr}(WJV_AJ^\top)$, so the first-order risk reduction is $\mathrm{tr}[WJVR^\top(RVR^\top)^{-1}RVJ^\top] \geq 0$; $W = I$ gives the Frobenius risk of the effect matrices and the off-diagonal selector its off-diagonal part. Uniform integrability is needed for this expectation, not for Theorem 8. The theorem fixes the correct description; Section 5.3 supplies the true labels but learns the graphs and equation classes, and its measured reduction is finite-sample evidence, not this limit.*

A.7 Finite-sample validity of a row-permutation test (Theorem 1(i), second clause)

The finite-sample clause of Theorem 1(i) concerns a test whose surrogates are formed differently from those of Algorithm 1. This section defines that test, proves its validity, and states how it relates to the implemented one.

The row-permutation test. Let T be a fixed measurable function of the ordered rows $X = (x_1, \ldots, x_N)$, with any internal randomness drawn from fixed seeds. The case of interest is the statistic of steps 2–9 of Algorithm 1 for the model without lags ($p = 0$, an empty lag design), every heuristic of the search included; it is such a function whenever the time limit of its integer program does not bind, since a binding limit makes the result depend on the machine. Nothing below uses its form. Draw $\sigma_1, \ldots, \sigma_R$ independently and uniformly from the permutations of $\{1, \ldots, N\}$, independently of X, and let $X^*_s = (x_{\sigma_s(1)}, \ldots, x_{\sigma_s(N)})$ be the observed rows in a new order; nothing is refitted to

form it. Put $T_0 = T(X)$ and $T_s = T(X_s^*)$. The test returns $\hat{p} = 1$ if $T_0 \le 0$ and otherwise the sequential value (6) below, with the h and R of Appendix B, and it reports two mechanisms when $\hat{p} \le \alpha$, $T_0 > 0$ and some coefficient of the reported description differs between its regimes, as in Algorithm 1.

Theorem 11 (Finite-sample validity of the row-permutation test)**.** *If the rows $x_1, \dots, x_N$ are exchangeable, for instance independent and identically distributed with any law, so that there is no lag dependence, the row-permutation test satisfies* $\Pr(\hat{p} \le \alpha) \le \alpha$ *for every* $\alpha \in [0,1]$ *and every N, and it reports two mechanisms with probability at most* α.

Remark 1 (The report rule keeps the level). Whenever the test has level α, as the row-permutation test does under Theorem 11, the reporting rule inherits that level, since its rejection region is a subset of the test's; the two extra conditions can only cost power (Section 5.1).

The p-value is computed sequentially (Appendix B): surrogates are drawn one at a time, up to a cap R, and drawing stops at the h-th exceedance. Write $B_t = \#\{s \le t : T_s \ge T_0\}$ for $t \le R$ (surrogates never drawn are defined on the same probability space, as a device of the proof), and $S = \min(\inf\{t : B_t = h\}, R)$ for the number actually drawn. The returned value is

$$\hat{p} = \begin{cases} h/S & \text{if } B_S = h, \\ (1 + B_R)/(R+1) & \text{if } B_S < h \text{ (so } S = R), \end{cases} \tag{6}$$

the truncated sequential estimator of Besag & Clifford (1991); the implemented test applies the same rule to its own surrogates. Let $p_R = (1 + B_R)/(R+1)$ denote the p-value that drawing all R surrogates would give.

Exchangeability. Condition on the multiset of rows, and let $Y = (y_1, \dots, y_N)$ list it in a fixed order, so that $(Y \circ \tau)_t = y_{\tau(t)}$ for a permutation τ. With exchangeable rows, $X = Y \circ \tau$ with τ uniform and independent of $\sigma_1, \dots, \sigma_R$, and $X_s^* = Y \circ (\tau\sigma_s)$. Then $\tau, \tau\sigma_1, \dots, \tau\sigma_R$ are i.i.d. uniform, so $(T_0, T_1, \dots, T_R)$ is exchangeable. Nothing here uses how T is computed, only that one fixed function is applied to the data and to every reordering. When $T_0 \le 0$ the test returns $\hat{p} = 1$ without drawing, which can only increase a valid p-value, so it suffices to prove $\Pr(\hat{p} \le \alpha) \le \alpha$ for (6); a report of two mechanisms requires $\hat{p} \le \alpha$, so its probability obeys the same bound.

Lemma 12 (Sequential rank calibration)**.** *If $(T_0, \dots, T_R)$ is exchangeable, then $\hat{p}$ in (6), with ties counted as exceedances, satisfies* $\Pr(\hat{p} \le \alpha) \le \alpha$ *for every fixed* $\alpha \in [0,1]$.

Proof. Suppose first that there are no ties. For every $m \le R$ the descending rank $1 + B_m$ of T_0 among $T_0, \dots, T_m$ is uniform on $\{1, \dots, m+1\}$.

Case $\alpha < h/R$. The first branch of (6) cannot reject, since $h/S \ge h/R > \alpha$. Nor can a full-budget rejection contain h exceedances: on $B_R \ge h$, $p_R \ge (h+1)/(R+1) \ge h/R > \alpha$. So $\{\hat{p} \le \alpha\} = \{p_R \le \alpha\}$ and $\Pr(\hat{p} \le \alpha) = \lfloor \alpha(R+1) \rfloor / (R+1) \le \alpha$.

Case $h/R \le \alpha \le 1$. Put $m = \lceil h/\alpha \rceil$, so $h \le m \le R$. Every no-hit outcome rejects, since then $\hat{p} \le h/(R+1) \le \alpha$; a hit outcome rejects exactly when the h-th exceedance occurs at draw m or later. Together, $\{\hat{p} \le \alpha\} = \{B_{m-1} \le h-1\}$: the observed statistic is among the largest h of $T_0, \dots, T_{m-1}$. By exchangeability this has probability $h/m \le \alpha$. (For $m = 1$ read $B_0 = 0$; this covers $h = 1$, $\alpha = 1$.)

Ties. Attach independent uniform tie-breakers $V_0, \dots, V_R$ and order the pairs (T_s, V_s) lexicographically. The augmented vector is exchangeable and tie-free, so the argument above applies to the p-value $\tilde{p}$ computed from it. Counting every original tie as an exceedance raises each comparison indicator relative to this ordering, and (6) is nondecreasing under coordinatewise increases of the indicators: if both paths reach h, the raised path reaches it no later and returns a larger h/S; if neither does, its final count is larger; if only the raised path reaches h, its value is at least h/R while the other is at most $h/(R+1)$. Hence $\hat{p} \ge \tilde{p}$ pointwise and $\Pr(\hat{p} \le \alpha) \le \Pr(\tilde{p} \le \alpha) \le \alpha$. □

The same monotonicity covers the implementation's convention for a surrogate on which the search fails (Appendix B): its statistic is set to $+\infty$, an exceedance, which again raises indicators and so $\hat{p}$.

With Lemma 12, this proves Theorem 11. □

Corollary 13 (The stopped and full-budget decisions agree)**.** *Couple the stopped and full-budget procedures on the same surrogate stream. Then for every data set,* $\mathbf{1}\{\hat{p} \le \alpha\} = \mathbf{1}\{p_R \le \alpha\}$ *whenever* $\alpha < h/R$.

Proof. Without a hit, $\hat{p} = p_R$. With a hit, $B_R \geq h$, so $p_R \geq (h+1)/(R+1) > \alpha$ and $\hat{p} \geq h/R > \alpha$: neither rejects. □

The paper's settings have $h/R = 10/39$ and $5/19$, both above $\alpha = 0.05$, so the sequential stopping changes no decision in Section 5: at $R = 39$ the test rejects exactly when $B_{39} \leq 1$, and at $R = 19$ exactly when $B_{19} = 0$. Corollary 13 assumes nothing about the null or about how the surrogates are formed, so it holds for the implemented surrogates too. The reporting rule's two further conditions ($T > 0$ and a coefficient that differs between the regimes) restrict the rejection region, so they preserve this equivalence and the level of any valid test (Remark 1).

Relation to the implemented test. Algorithm 1 does not run this test. It selects a lag order $p \in \{1, \ldots, 4\}$, holds the observed lag matrix U fixed, and draws $X^* = (U\hat{A} + \hat{E}_\pi)(I - \hat{B})^{-1}$ from the one-regime fit $(\hat{A}, \hat{B})$, where $\hat{E} = X(I - \hat{B}) - U\hat{A}$ are its residuals and $\hat{E}_\pi$ their rows in the order π (the fixed-regressor bootstrap of Hansen (2000); Appendix B). These surrogates are not reorderings of the observed rows, their statistics need not be exchangeable with $T(X)$, and Theorem 11 gives them no finite-sample level. The two procedures coincide only for an empty lag design: then $\hat{E} = X(I - \hat{B})$, so $X^* = \Pi X$ for the permutation matrix Π of π, which is the case the code treats as an exact row permutation; Algorithm 1 never selects it. For the implemented test, the level statement is the asymptotic one of Theorem 14 under a Gaussian VAR null: $T \leq 0$ with probability tending to one, so no surrogate is drawn and the probability of rejecting tends to zero. Its rejection rates on other one-regime series are measured in Section 5.1.

A.8 Null consistency with lags (Theorem 1(i), first clause)

Theorem 14 (Null consistency with lags)**.** *Let the rows follow a stable Gaussian VAR(p) with positive-definite innovation covariance whose equations lie in the candidate family of Appendix B (every VAR(p) does when $d - 1 + dp \leq 16$), and let the one-regime score be within $o_p(\log N)$ of its optimum over that family, as it is for $d < 10$ when the integer program returns within its time limit (Proposition 7). Under the dimension condition $S_\star < 1$ of (7), the probability that any admissible two-regime description scores above the one-regime description tends to zero. Hence* $\Pr(T > 0) \to 0$, $\Pr(\hat{p} \leq \alpha) \to 0$ *for every $\alpha < 1$, and the probability of reporting two mechanisms tends to zero.*

Throughout this section the rows follow a stable Gaussian VAR(p): $x_t = C_0^\top u_t + \varepsilon_t$ with $u_t = (x_{t-1}, \ldots, x_{t-p}) \in \mathbb{R}^{dp}$ and $\varepsilon_t \sim N_d(0, \Sigma_0)$ i.i.d., $\Sigma_0 \succ 0$, so that $\Gamma = \mathbb{E}[u_t u_t^\top] \succ 0$. Write $k = d(1 + p)$ and $m = 5k$ for the minimum run of Appendix B. Scores are the Gaussian score (4) in the identity gauge; Student-t emissions only generate proposals (Appendix A.10), and the per-column standardization adds the same constant to every description, so the argument is written in the original units. For fixed d and p the codes for parent sets, lag sets, equation classes and gauge are bounded, and each parameter costs $\frac{1}{2}\log n_\vartheta$ with n_ϑ the number of rows informing it.

The family and the dimension condition. A description gives each regime a *component*, a DAG with a lag support per equation, from the finite candidate family $\mathscr{F}$ of Appendix B: every parent set is a candidate, and every lag support is when $d - 1 + dp \leq 16$ (at most two lag coefficients per equation otherwise). For $M \in \mathscr{F}$ let $r_j(M)$ be the number of regression coefficients of equation j, its same-time parents plus its lag coefficients, and define

$$S_\star = \max_{M \in \mathscr{F}} \sum_{j=1}^{d} \frac{r_j(M)}{m - r_j(M)}, \qquad b_\star = \frac{1}{1 + S_\star}. \tag{7}$$

The dimension condition of Theorem 14 is $S_\star < 1$, that is $b_\star > \frac{1}{2}$. The maximum is attained by a complete DAG with the largest lag supports, whose equation sizes are $r_j = \ell + j - 1$ with $\ell = dp$ when $d - 1 + dp \leq 16$ and $\ell = 2$ otherwise; Table 3 lists $S_\star$ for $m = 5d(1 + p)$ (exact arithmetic; for $d = 6$, $p = 1$ the value is $25775333/25827165$). It is below one for every $d \leq 6$ and $p \leq 4$, for $d = 9$, and for $d = 7, 8$ with $p \geq 2$. The unscreened cases $d = 7, 8$ with $p = 1$ and the screened search at $d \geq 10$ are outside the theorem.

Table 3: The dimension condition of Theorem 14: $S_\star$ of (7) for the candidate family and minimum run of Appendix B. Entries below one satisfy the condition; bold entries do not.

d	$p=1$	$p=2$	$p=3$	$p=4$
1	0.111	0.154	0.176	0.190
2	0.288	0.354	0.389	0.410
3	0.465	0.554	0.601	0.630
4	0.643	0.755	0.813	0.146
5	0.820	0.955	0.209	0.166
6	0.998	0.318	0.235	0.186
7	**1.176**	0.353	0.261	0.207
8	**1.354**	0.388	0.287	0.227
9	0.652	0.422	0.312	0.248

The theorem's other two conditions are that the family contains the true equations, which holds for every VAR(p) when $d-1+dp \le 16$ and otherwise when each true equation has at most two lag coefficients, and that the one-regime score Q_1 is within $e_N = o_p(\log N)$ of its optimum $Q_1^\star$ over the one-regime family, as it is when the integer program returns within its time limit (Proposition 7). Let D_0 be the parameter dimension of the smallest correct one-regime description and ℓ_0 the log-likelihood of all rows at the true parameter.

Two likelihood bounds. Let $\mathscr{I}$ be a deterministic set of $n \ge m$ rows, any union of intervals, and $M \in \mathscr{F}$. Let $\hat{\ell}_{\mathscr{I},M}$ be the log-likelihood of the rows in $\mathscr{I}$ maximized over the parameters of M, each equation profiled on its own, and $\ell_{0,\mathscr{I}}$ the log-likelihood of those rows at the true parameter.

Lemma 15 (Short-segment tail)**.** *There is a C, depending only on the null law, d, p and $\mathscr{F}$, such that for every $M \in \mathscr{F}$, every deterministic $\mathscr{I}$ with $|\mathscr{I}| = n \ge m$ and every $x \ge 1$,*

$$\Pr\big(\hat{\ell}_{\mathscr{I},M} - \ell_{0,\mathscr{I}} \ge x\big) \le C(1+n+x)^C e^{-b_\star x}. \tag{8}$$

The constant does not depend on where the rows of $\mathscr{I}$ lie or how they are spaced, and M need not contain the true law. For a labeling z with groups $\mathscr{I}_0, \mathscr{I}_1$ of sizes n_0, n_1, let $\hat{\ell}_{\mathrm{sat}}(z)$ be the log-likelihood maximized over a separate unrestricted regression of x_t on u_t, with its own covariance, in each group, and $\Lambda_z = \hat{\ell}_{\mathrm{sat}}(z) - \ell_0$; this dominates the likelihood gain of every two-regime description at z.

Lemma 16 (Moderate deviations of large groups)**.** *Fix $A > 0$ and $t \in (0,1)$. There are C, c, C_0, depending only on the null law, d, p, A and t, such that uniformly over labelings with $n_{\min} = \min(n_0, n_1) \ge C_0 \log N$ and over $1 \le x \le A \log N$,*

$$\Pr(\Lambda_z \ge x) \le Ce^{-tx} + Ce^{-cn_{\min}}. \tag{9}$$

C_0 may be enlarged so that the second term is $o(N^{-8})$, which survives a union bound over the $O(N^8)$ admissible labelings.

The lemmas are proved after the theorem.

Proof of Theorem 14. Fix an admissible (z, M) with $1 \le J = J(z) \le 8$ and both regimes used; call the group with more rows the majority and the other the minority, of size $n_{\min}$. The likelihood of a description is at most the sum of its two components' likelihoods maximized separately, since removing the sharing constraints enlarges the feasible set. The smallest correct one-regime description at the true parameter gives

$$Q_1 \;\ge\; \ell_0 - \tfrac{D_0}{2}\log N - C - e_N. \tag{10}$$

The number of admissible labelings with J switches whose minority has r runs and n rows is at most $C_J N^{J-r} n^{r-1}$ (a minority run at an end of the series has a fixed start, and the run lengths form a composition of n), and alternation forces $r \ge J/2$, so

$$\#\{z : J(z) = J,\ n_{\min} < C_0 \log N\} \le C_J N^{J/2} (\log N)^{C_J}, \qquad \#\{z : J(z) = J,\ n_{\min} \le N^{1/4}\} \le C_J N^{5J/8}. \tag{11}$$

Incorrect majority component. If the majority component cannot represent the true equations, its best likelihood on the majority rows is below the true one by at least cN with probability tending to one, uniformly over majorities: the family is finite and the set of laws a component represents is closed, so the Kullback–Leibler distance from the true conditional law to the component is bounded away from zero, and the moment matrices of any union of at most nine intervals holding at least $N/2$ rows concentrate uniformly (step 1 of the proof of Lemma 16 with $n \geq N/2$). The minority's gain over the true law is at most $A \log N$ simultaneously over all admissible labelings with probability tending to one, by Lemma 15 at $x = A \log N$ for $n_{\min} < C_0 \log N$ and Lemma 16 otherwise, with the counts (11) and A large. Codes differ by $O(\log N)$. No such description beats Q_1.

Correct majority component. Its equations have at least D_0 parameter slots, a of them shared and b_{maj} specific to the majority, $a + b_{\mathrm{maj}} \geq D_0$. An equation that is not fully shared, a graded one included since it has its own minority variance, gives the minority a specific slot, $b_{\min} \geq 1$. As $n_{\mathrm{maj}} \geq N/2$, the parameter charge exceeds $\frac{D_0}{2} \log N$ by at least

$$\tfrac{1}{2}\big(a \log N + b_{\mathrm{maj}} \log n_{\mathrm{maj}} + b_{\min} \log n_{\min}\big) - \tfrac{D_0}{2} \log N \;\geq\; \tfrac{1}{2} \log n_{\min} - C. \tag{12}$$

A description whose equations are all fully shared has the parameters of a one-regime description and pays $J \log N$ for its switches, so it scores below $Q_1^\star - J \log N$ and, by (10), below Q_1 with probability tending to one. For the rest, three ranges of $n_{\min}$ are treated in turn.

Small minorities, $m \leq n_{\min} < C_0 \log N$. The score of all N rows at the true parameter is $O_p(\sqrt{N})$ and its largest row is $O_p(\log N)$, so removing any $C_0 \log N$ rows changes it by $O_p((\log N)^2)$; the majority's unrestricted profile gain over the true law is therefore $O_p(1)$ uniformly over all such removals. With (10) and (12), dropping the nonnegative $\frac{1}{2} \log n_{\min}$,

$$Q(z, M) - Q_1 \;\leq\; \hat{\ell}_{\mathscr{I}, M_{\min}} - \ell_{0,\mathscr{I}} - J \log N + O_p(1) + e_N + C, \tag{13}$$

with $\mathscr{I}$ the minority and $M_{\min}$ its component. For fixed $\eta > 0$ the remainder is at most $\eta \log N$ with probability tending to one. Lemma 15 at $x = (J - \eta) \log N$ and a union bound over (11) and over $\mathscr{F}$ bound the probability that some description in this range beats Q_1 by $C(\log N)^C N^{J/2 - b_\star(J-\eta)}$, which tends to zero for small η because $b_\star > \frac{1}{2}$.

Intermediate minorities, $C_0 \log N \leq n_{\min} \leq N^{1/4}$. Lemma 16 with $t = 3/4$ at $x = (J - \eta) \log N$, applied to the saturated gain that dominates the description's likelihood gain, and the count (11) give the union bound $CN^{5J/8 - (3/4)(J-\eta)} = CN^{-J/8 + 3\eta/4} \to 0$; the second term of (9) contributes $o(N^{-8}) N^{5J/8}$.

Large minorities, $n_{\min} \geq N^{1/4}$. Now (12) adds at least $\frac{1}{8} \log N - C$ to the switch charge $J \log N$, there are at most $C_J N^J$ labelings, and Lemma 16 with $t = 0.99$ gives the exponent $J - 0.99(J + \frac{1}{8}) < 0$ for $1 \leq J \leq 8$. The inequalities are strict, so $e_N = o_p(\log N)$ is absorbed by a small η.

Together, $\Pr(\exists (z, M) : Q(z, M) > Q_1) \to 0$. On the complement $Q_2 \leq Q_1$, so $T \leq 0$, Algorithm 1 returns $\hat{p} = 1$ without drawing a surrogate, and neither a rejection nor a two-mechanism report occurs. □

Lag order. Algorithm 1 chooses p from $\{1, \ldots, 4\}$ by BIC on the pooled regression. Let $p^\dagger$ be the smallest order that represents the true conditional mean. Smaller orders lose a likelihood of order N; larger ones gain $O_p(1)$ and pay a charge of order $\log N$. The selector therefore returns $p^\dagger$ with probability tending to one, and the theorem applies at $p^\dagger$ with an $o(1)$ added to the probability.

Proof of Lemma 15. *Step 1: two Gaussian facts.* Let $w_t = (u_t^\top, x_t^\top)^\top \in \mathbb{R}^k$. For a unit vector v, the scalar series $y_t(v) = v^\top w_t$ is a polynomial filter of order p of the VAR: $y_t(v) = \sum_{i=1}^d P_i(L) x_{t,i}$ with L the lag operator and scalar polynomials P_i of degree at most p whose coefficient vectors together have unit norm, so some P_i has coefficient norm at least $d^{-1/2}$. The spectral density of a stable VAR with $\Sigma_0 \succ 0$ is bounded below by $c_0 I$, so the spectral density f_v of $y(v)$ satisfies $f_v(\omega) \geq c_0 |P_i(e^{i\omega})|^2$, and

$$\exp \int \log f_v \, \tfrac{d\omega}{2\pi} \;\geq\; c_0 \, \mathrm{M}(P_i)^2 \;\geq\; c_0 4^{-p} \|\mathrm{coef}(P_i)\|^2 \;\geq\; c_0 4^{-p}/d =: a,$$

where $\mathrm{M}(P) = \exp \int \log |P(e^{i\omega})| \, d\omega / 2\pi$ is the Mahler measure, by Jensen's formula the leading coefficient times $\prod \max(1, |\rho|)$ over the roots ρ, which bounds every coefficient (an elementary

symmetric function of the roots) up to the factor 2^p. The variance of $y_t(v)$ given any finite set of earlier values of $y(v)$ is the error variance of a linear predictor $A(L)y_t$ with $A(0)=1$, which is $\int |A|^2 f_v \geq \exp \int \log f_v \geq a$ by the arithmetic–geometric mean inequality and $\int \log|A| \geq \log|A(0)| = 0$. Hence for every deterministic set $\mathscr{I}$ of n times the covariance matrix $K_{v,\mathscr{I}}$ of $(y_t(v))_{t\in\mathscr{I}}$, whose determinant is the product of the successive conditional variances, has $\det K_{v,\mathscr{I}} \geq a^n$, however the times are spaced. The density of that Gaussian vector is therefore at most $(2\pi a)^{-n/2}$, and the volume of the ball of radius $\sqrt{nu}$ in $\mathbb{R}^n$ gives

$$\Pr\Big(n^{-1}\sum_{t\in\mathscr{I}} (v^\top w_t)^2 \leq u\Big) \leq (Cu)^{n/2} \qquad (u>0), \tag{14}$$

uniformly in v and $\mathscr{I}$, and for any coordinate subset of w_t by embedding. Second, stability bounds the covariance operator of $(w_t)_{t\in\mathscr{I}}$ uniformly, so the Gram matrix $H_{\mathscr{I}} = n^{-1}\sum_{t\in\mathscr{I}} w_t w_t^\top$ satisfies

$$\Pr(\operatorname{tr} H_{\mathscr{I}} > \mu) \leq \exp(Cn - cn\mu) \tag{15}$$

by the moment generating function of a Gaussian quadratic form.

Step 2: equationwise Gram bounds. Equation j of M uses the vector $w_{j,t}$ of its r_j regressors and its response, of dimension r_j+1, with Gram matrix $H_j = n^{-1}\sum_{t\in\mathscr{I}} w_{j,t} w_{j,t}^\top$, a principal submatrix of $H_{\mathscr{I}}$. On $\{\operatorname{tr} H_{\mathscr{I}} \leq \mu\}$ take an η-net of the unit sphere of $\mathbb{R}^{r_j+1}$ with $\eta = \frac{1}{2}\sqrt{\varepsilon_j/\mu}$; it has at most $C\eta^{-r_j}$ points, the sphere having dimension r_j. If $\lambda_{\min}(H_j) \leq \varepsilon_j$, some unit v has $v^\top H_j v \leq \varepsilon_j$, and the nearest net point v' has $v'^\top H_j v' \leq \varepsilon_j + 2\sqrt{\mu\varepsilon_j}\,\eta + \mu\eta^2 = \frac{9}{4}\varepsilon_j$. By (14),

$$\Pr\big(\lambda_{\min}(H_j) \leq \varepsilon_j,\ \operatorname{tr} H_{\mathscr{I}} \leq \mu\big) \leq C C_1^{n/2} \mu^{r_j/2} \varepsilon_j^{(n-r_j)/2}. \tag{16}$$

For $b \in (0,1]$, fixed below, set

$$\mu = C_2(1+x/n), \qquad \varepsilon_j = c_0\,\mu^{-r_j/(n-r_j)} \exp\big(-2bx/(n-r_j)\big). \tag{17}$$

Then (16) equals $CC_1^{n/2} c_0^{(n-r_j)/2} e^{-bx}$, and since $n \geq m \geq 2k > 2r_j$ a small c_0, fixed independently of n, $\mathscr{I}$ and x, makes it at most e^{-bx}; a large C_2 makes (15) at most e^{-2x}. So

$$G = \{\operatorname{tr} H_{\mathscr{I}} \leq \mu,\ \lambda_{\min}(H_j) \geq \varepsilon_j\ \forall j\} \quad \text{has} \quad \Pr(G^c) \leq (d+1)e^{-bx}. \tag{18}$$

Step 3: precision coordinates. Write the conditional density of equation j as $f_j(x_j \mid \mathrm{pa}_j, u) = (2\pi)^{-1/2} a_j \exp\{-\frac{1}{2}(a_j x_j - \gamma_j^\top v_j)^2\}$ with $a_j > 0$ the root precision, v_j its r_j regressors and $\gamma_j = a_j\beta_j$, and put $\theta_j = (-\gamma_j^\top, a_j)^\top$ in the order of $w_{j,t}$. Up to a constant,

$$\ell_j(\theta_j) = n\log a_j - \tfrac{n}{2}\theta_j^\top H_j \theta_j, \tag{19}$$

strictly concave on $a_j > 0$ when $H_j \succ 0$. On G its maximizer $\hat\theta_j$ is interior and satisfies $H_j\hat\theta_j = e_a/\hat a_j$, with e_a the unit vector of the response coordinate, and $\hat\theta_j^\top H_j \hat\theta_j = 1$; hence $\mu^{-1/2} \leq \|\hat\theta_j\| \leq \varepsilon_j^{-1/2}$, and $\hat a_j \geq \mu^{-1/2}$ because the fitted residual mean square is at most the response's, which is at most μ. The product of the d equation densities in a topological order is a normalized density of x_t given u_t for every parameter value, stable or not.

Step 4: a deterministic grid. For each equation take a net of the unit sphere of $\mathbb{R}^{r_j+1}$ of angular radius $\delta_j = c\sqrt{\varepsilon_j/(n\mu)}$, with at most $C\delta_j^{-r_j}$ points, and a geometric grid of radii on $[\mu^{-1/2}, \varepsilon_j^{-1/2}]$ with log-spacing $c/\sqrt{n}$, with at most $C\sqrt{n}\{1+\log(\mu/\varepsilon_j)\}$ points; keep the grid parameters whose response coordinate is positive. The grid depends on n, x and M only. Write $\hat\theta_j = \rho\hat v$ and pick a net direction v within δ_j of $\hat v$; let $e = \rho(v-\hat v)$, so $\|e\| \leq c/\sqrt{n\mu}$ and $w := e^\top H_j e \leq c^2/n$. Let e_{resp} be the response coordinate of e and $\zeta = e_{\mathrm{resp}}/\hat a_j = \hat\theta_j^\top H_j e$, so $|\zeta| \leq \sqrt{w} \leq c/\sqrt{n}$ by Cauchy–Schwarz in the H_j inner product; in particular $\hat a_j + e_{\mathrm{resp}} > 0$. Let $s\rho$ be the nearest grid radius, $|\log s| \leq c/\sqrt{n}$, and $\theta' = s(\hat\theta_j + e)$. Substituting in (19),

$$\ell_j(\theta') - \ell_j(\hat\theta_j) = n\big\{\log s - \tfrac{s^2-1}{2}\big\} + n\big\{\log(1+\zeta) - s^2\zeta\big\} - \tfrac{n}{2}s^2 w, \tag{20}$$

and each term is at least $-Cc^2$ uniformly in n, μ and ε_j (for the middle one write $\log(1+\zeta)-s^2\zeta = [\log(1+\zeta)-\zeta]+(1-s^2)\zeta$). A fixed small c, depending on d only, makes the total loss over the d equations at most one. So on G the grid contains a parameter within one likelihood unit of $\hat{\ell}_{\mathscr{I},M}$, and its size is

$$K \le C\prod_{j=1}^{d}\Big[\sqrt{n}\{1+\log(\mu/\varepsilon_j)\}\Big(\frac{n\mu}{\varepsilon_j}\Big)^{r_j/2}\Big] \le C(1+n+x)^C \exp\Big\{bx\sum_j \frac{r_j}{n-r_j}\Big\}, \tag{21}$$

by (17): $n\mu = C_2(n+x)$, $r_j/(n-r_j)$ is bounded, and $\log(\mu/\varepsilon_j) \le C(1+\log\mu + x/n)$.

Step 5: chronological likelihood ratios. For a fixed grid parameter θ, $L(\theta) = \prod_{t\in\mathscr{I}} f_\theta(x_t \mid u_t)/f_0(x_t \mid u_t)$ has expectation one under the null: taking the conditional expectation given the past at the latest $t \in \mathscr{I}$ and proceeding backwards, each factor integrates to one because u_t is a function of the past and $f_\theta(\cdot \mid u_t)$ is a normalized density. Nothing is conditioned on future lags. By Markov's inequality $\Pr(L(\theta) \ge e^{x-1}) \le e^{1-x}$, so with (18) and the grid,

$$\begin{aligned}\Pr\big(\hat{\ell}_{\mathscr{I},M} - \ell_{0,\mathscr{I}} \ge x\big) &\le (d+1)e^{-bx} + Ke^{1-x} \\ &\le (d+1)e^{-bx} + C(1+n+x)^C \exp\Big\{-x\Big(1 - b\sum_j \frac{r_j}{n-r_j}\Big)\Big\}.\end{aligned}$$

Choosing $b = \big(1+\sum_j r_j/(n-r_j)\big)^{-1}$ makes both exponents b, and $b \ge b_\star$ for $n \ge m$ because $r/(n-r)$ decreases in n. A union over the finite family $\mathscr{F}$ gives (8). □

Proof of Lemma 16. *Step 1: moment concentration.* By stability the covariance operator of any finite stack of the w_t is bounded, and the stack is a bounded linear image of the i.i.d. innovations. For a deterministic set of n rows, each centered entry of $n^{-1}\sum u_t u_t^\top$, $n^{-1}\sum u_t \varepsilon_t^\top$ and $n^{-1}\sum \varepsilon_t \varepsilon_t^\top$ is a centered Gaussian quadratic form whose matrix has operator norm $O(1/n)$ and squared Frobenius norm $O(1/n)$, uniformly in the set (selection matrices and lag shifts have operator norm at most one). Gaussian quadratic-form concentration gives, for a fixed small $\delta > 0$, an event $G_z(\delta)$ on which both groups' moment matrices are within δ of Γ, 0 and Σ_0 in operator norm, with

$$\Pr(G_z(\delta)^c) \le Ce^{-c\delta n_{\min}}. \tag{22}$$

Step 2: localization. On $G_z(\delta)$ the unrestricted maximizers in group r, $\hat{C}_r - C_0 = (U_r^\top U_r)^{-1}U_r^\top E_r$ and $\hat{\Sigma}_r = n_r^{-1}\{E_r^\top E_r - E_r^\top U_r(U_r^\top U_r)^{-1}U_r^\top E_r\}$, lie within $O(\delta)$ of the truth, and along the segment from the true parameter to the maximizer the negative Hessian of the log-likelihood in the coordinates $\theta = (\operatorname{vec} C, \operatorname{vech}\Sigma)$ is within $C\delta$ of the expected information H_z in H_z-relative operator norm, where H_z is block diagonal with blocks $n_r H_0$, $H_0 \succ 0$. Let S_z be the score at the true parameter, stacked over the two groups. The quadratic upper bound along the segment gives, for small δ,

$$\Lambda_z \le \frac{1}{2(1-C\delta)} S_z^\top H_z^{-1} S_z. \tag{23}$$

Step 3: the score covariance. The row scores are martingale differences for the natural filtration: the coefficient part is $u_t\varepsilon_t^\top$, linear in the current innovation with a predictable multiplier, and the covariance part is a centered quadratic form in ε_t; the two parts are uncorrelated because Gaussian odd moments vanish, and scores at distinct times are uncorrelated by the martingale property. Hence, without conditioning on the lags, $\operatorname{Var}(S_z) = H_z$ exactly. Put $Z_z = H_z^{-1/2}S_z$, of fixed dimension $q = 2\{d^2p + d(d+1)/2\}$. Every unit projection $v^\top Z_z$ is a centered Gaussian quadratic form $g^\top A_{z,v} g - \operatorname{tr} A_{z,v}$ in the innovations with $2\operatorname{tr}(A_{z,v}^2) = 1$ and $\|A_{z,v}\|_{\mathrm{op}} \le Cn_{\min}^{-1/2}$, by the bounded-filter argument of step 1 with the $n_r^{-1/2}$ normalization.

Step 4: directional tails. For a centered Gaussian quadratic form Ξ with variance one and operator bound a, $\log \mathbb{E}e^{\lambda\Xi} \le \lambda^2/\{2(1-2a|\lambda|)\}$ for $2a|\lambda| < 1$ (diagonalize and use $-\frac{1}{2}\log(1-y) - \frac{y}{2} \le y^2/\{4(1-|y|)\}$), and $\lambda = u/(1+2au)$ gives

$$\Pr(\Xi \ge u) \le \exp\Big\{-\frac{u^2}{2(1+2au)}\Big\}. \tag{24}$$

Take a ρ-net of the unit sphere of $\mathbb{R}^q$, of at most $(1+2/\rho)^q$ points; $\|Z_z\| \geq v$ forces some net direction to have projection at least $(1-\rho)v$. By (23), $\Lambda_z \geq x$ on $G_z(\delta)$ forces $\|Z_z\|^2 \geq 2(1-C\delta)x$, so (24) gives

$$\Pr(\Lambda_z \geq x,\ G_z(\delta)) \leq (1+2/\rho)^q \exp\Big\{-\frac{(1-\rho)^2(1-C\delta)\,x}{1+C'\sqrt{x/n_{\min}}}\Big\}. \tag{25}$$

For $x \leq A\log N$ and $n_{\min} \geq C_0 \log N$, $\sqrt{x/n_{\min}} \leq \sqrt{A/C_0}$; choosing δ and ρ small and then C_0 large makes the exponent at least tx. With (22) this is (9), and enlarging C_0 further makes $Ce^{-c_\delta C_0 \log N} = o(N^{-8})$. □

A.9 Detection and mechanism reporting (Theorem 1(ii)–(iii))

This section proves parts (ii) and (iii) of Theorem 1 in three steps: a bound on the score of *every* description in $\mathscr{H}_0$, at *every* admissible labeling (Lemma 17); the detection and reporting statement (Theorem 18), which needs only the score ascent of Proposition 21, not consistent labels or a global optimum; and a second-moment certificate that the mechanism gap κ is positive (Lemma 20), with the reversal motif of Section 5 as an instance and the unanchored counterexample.

Population quantities. Fix d, the lag order p and the switch cap K. The true labels z^0 have a bounded number of runs and proportions $n_r/N \to \pi_r > 0$. Within each true regime the conditional mean and covariance of x_t given u_t are represented by the linear SEM family; the innovations need not be Gaussian. Let $L_\star$ be the limiting expected Gaussian log-likelihood per row at the true two-regime conditional moments, L_1 the best value for one regime ($\mathscr{H}_1$), and L_0 the best value at the true labels over the class $\mathscr{M}_0 = \mathscr{H}_0$ of Section 2, the descriptions in which *no coefficient differs between the regimes* in the data's coordinates: shared and graded equations, whose innovation variances may differ. The gaps are those of Section 2,

$$\Delta = L_\star - L_1, \qquad \kappa = L_\star - L_0. \tag{26}$$

Duplicating a one-regime model gives a member of $\mathscr{M}_0$, so $0 \leq \kappa \leq \Delta$: Δ measures how different the two conditional laws are, κ how far they are from an explanation with common coefficients and different noise levels. (The bounds below also hold, with the same proofs, for the larger class of descriptions shared up to a gauge, whose gap is at most κ.) Write $Q_N(z,M) = \hat{\ell}_N(z,M) - \mathrm{pen}_N(z,M)$ for (4); for fixed d, p, K the charges are nonnegative and uniformly $O(\log N)$.

Conditions.

(C1) *Within-run laws.* On every interval I inside a true run of regime r, uniformly over such intervals and over parameters θ in a compact set with variances bounded away from zero, $N^{-1}\big|\sum_{t\in I}\{f_s(x_t,u_t;M,\theta) - q_{rs}(M,\theta)\}\big| = o_p(1)$, where f_s is the fitted Gaussian row log-likelihood of fitted regime s and q_{rs} its expectation under true regime r; the joint second-moment matrices of (u_t,x_t) are positive definite.

(C2) *Profile regularity.* For every $\varepsilon > 0$, profiles with both fitted regime fractions at least ε localize to a compact parameter set with probability tending to one; the one-regime profile obeys its law of large numbers, and removing an adaptively chosen vanishing fraction of rows changes its maximized log-likelihood by $o_p(N)$ (for saturated regression this follows from a finite $(2+\delta)$ empirical moment and a positive-definite joint moment limit, by the Hölder bound in Lemma 19).

(C3) *Small-group control.* Let I range over admissible fitted groups (unions of at most $K+1$ intervals) and $\ell_{\mathrm{sat}}(I)$ be the unrestricted Gaussian regression profile on I with the frozen lag design. Then $\lim_{\varepsilon\downarrow 0} \limsup_N \Pr\{\sup_{0<|I|\leq \varepsilon N} |\ell_{\mathrm{sat}}(I)|/N > \eta\} = 0$ for every $\eta > 0$.

The fixed minimum run of Appendix B is unchanged; (C3) only rules out a vanishing group with a nearly singular residual covariance contributing order N to the likelihood. These are conditions on the likelihood, not on the success of the equation classifier.

Lemma 17 (All-label bound)**.** *Under (C1)–(C3),* $\sup_{z\in\mathscr{Z}_N, M\in\mathscr{M}_0} Q_N(z,M) \leq NL_0 + o_p(N)$*, where* $\mathscr{Z}_N$ *is the admissible labeling set of Appendix A.10.*

Proof. Consider first fitted regime fractions at least $\varepsilon > 0$. Intersecting the boundedly many true runs with the boundedly many fitted runs and applying (C1) on each piece, uniformly over the localized

profiles of (C2),

$$N^{-1}\ell_N(z,M;\theta)=\sum_{r,s}\pi_{rs,N}(z)\,q_{rs}(M,\theta)+o_p(1)\le\sum_r\pi_{r,N}\max_s q_{rs}(M,\theta)+o_p(1).$$

Choosing the maximizing fitted label for each true regime gives one of four assignments. A bijective assignment is a description in $\mathscr{M}_0$ at the true labels after relabeling; a constant assignment duplicates one fitted component in both true regimes, which is also in $\mathscr{M}_0$ (with the identity gauge). Every such population value is at most L_0, and profiling over θ preserves the bound. If one fitted group's fraction vanishes, its contribution is at most its saturated profile, which is $o_p(N)$ by (C3); the remaining group is a one-regime regression on all but a vanishing fraction of rows, whose normalized score tends to $L_1\le L_0$ by the deletion property in (C2). Let $N\to\infty$ and then $\varepsilon\downarrow 0$. Subtracting the nonnegative charges keeps the bound. □

The search path. Let $Q_{\mathrm{seed},N}$ be the score of an initialization retained by the search, $Q_{2,N}=Q_2$ the best retained two-regime score, $Q_{1,N}=Q_1$, and $Q_{\mathrm{rep},N}$ the score of the description the reporting step returns at the selected labels (defined on every data set, whether or not an earlier gate stops the algorithm). By Proposition 21(b,c), $Q_{2,N}\ge Q_{\mathrm{seed},N}$.

Theorem 18 (Detection and mechanism reporting)**.** *Assume Lemma 17, that $Q_{1,N}\le NL_1+o_p(N)$, and:*

(i) some retained initialization has $Q_{\mathrm{seed},N}\ge N(L_\star-\xi)-o_p(N)$ for a fixed $0\le\xi<\kappa$;
(ii) the reporting step loses at most $o_p(N)$: $Q_{\mathrm{rep},N}\ge Q_{2,N}-o_p(N)$;
(iii) the surrogate condition: for the R surrogate statistics, $N^{-1}\max_{s\le R}(T^*_{N,s})_+\to_p 0$, and $(R+1)^{-1}\le\alpha$ (sufficient conditions follow the proof).

Then $T_N=Q_{2,N}-Q_{1,N}\ge N(\Delta-\xi)-o_p(N)$, the reported description exceeds every member of $\mathscr{M}_0$ at every admissible labeling by at least $N(\kappa-\xi)-o_p(N)$, and SEAM *reports two mechanisms with probability tending to one.*

Proof. From (i) and the ascent, $Q_{2,N}\ge N(L_\star-\xi)-o_p(N)$; subtracting the one-regime bound gives the first claim, which is positive and linear in N since $\Delta\ge\kappa>\xi$. Combining (ii) with Lemma 17 gives the second claim, positive with probability tending to one; so some coefficient of the reported description differs between the regimes in the data's coordinates, since otherwise its fitted law would lie in $\mathscr{M}_0$ and contradict the bound. Choose $c\in(0,\Delta-\xi)$. By (iii), with probability tending to one $T_N>cN>\max_{s\le R}T^*_{N,s}$, so no surrogate exceeds T_N: no early stop occurs (stopping needs $h\ge 1$ exceedances), all R surrogates are drawn, and (6) returns $\hat p=1/(R+1)\le\alpha$. The gates $T>0$, $\hat p\le\alpha$ and "some coefficient differs" all pass. □

Neither consistent labels nor a global optimum of the three-round search is assumed. With $\xi=0$ the bound is $T_N\ge N\Delta-o_p(N)$; a remainder of order $\log N$ does not follow from a population gap, since likelihood-ratio fluctuations are of order $\sqrt{N}$. Condition (ii) holds with no loss when the reporting fit can retain the incumbent description. For $d<10$ the reporting step of Appendix B does: it re-solves the integer program exactly for every candidate gauge, the identity included, on the incumbent's candidate sets, accepts a label refinement only when the score rises, and then classifies the equations of a fixed pair of graphs, where the incumbent classification is feasible. For $d\ge 10$ it re-screens the candidates on the selected labels, which need not keep the incumbent's, and (ii) is assumed.

Lemma 19 (A sufficient seed condition)**.** *Let $\tilde z$ be an initial labeling with, after a global exchange of regime names if needed, $m_N=\#\{t:\tilde z_t\ne z^0_t\}=o_p(N)$. Suppose a correct description (M^0,θ^0) with $O(\log N)$ code length remains a candidate at $\tilde z$, the fit retained there scores at least $\ell_N(\tilde z,M^0,\theta^0)-O(\log N)$, the true coefficients and inverse variances are bounded, and $N^{-1}\sum_t(1+\|x_t\|^2+\|u_t\|^2)^{1+\delta/2}=O_p(1)$ for some $\delta>0$. Then (i) holds with $\xi=0$.*

Proof. The difference of two fixed Gaussian row log-likelihoods is a quadratic polynomial in (x_t,u_t), so its absolute value is at most CW_t with $W_t=1+\|x_t\|^2+\|u_t\|^2$, and only mislabeled rows contribute. By Hölder's inequality,

$$N^{-1}\left|\ell_N(\tilde z,M^0,\theta^0)-\ell_N(z^0,M^0,\theta^0)\right|\le C\,(m_N/N)^{\delta/(2+\delta)}\left(N^{-1}\textstyle\sum_t W_t^{1+\delta/2}\right)^{2/(2+\delta)}=o_p(1),$$

and $N^{-1}\ell_N(z^0, M^0, \theta^0) \to_p L_\star$ by (C1). The retained-score condition and the sublinear code length finish the proof. □

For $d < 10$ the structure step is exact over every candidate parent set and lag support (Proposition 7), which supplies the retained-score condition whenever the true description is in the candidate family, whose lag supports are capped at two coefficients per equation when $d-1+dp > 16$; for the screened search at $d \geq 10$ it must hold for the candidates actually retained.

The surrogate condition (iii). A fixed-design surrogate is $X^* = (U\hat{A} + \hat{E}_\pi)(I-\hat{B})^{-1}$, where $\hat{M} \in \mathscr{F}$ is the fitted one-regime description with coefficients $(\hat{A}, \hat{B})$ and residuals $\hat{E} = X(I-\hat{B}) - U\hat{A}$. Its rows are $x_t^* = C^\top u_t + \eta_t$ with $C = \hat{A}(I-\hat{B})^{-1}$ and $\eta_t = (I-\hat{B})^{-\top}\hat{e}_{\pi(t)}$: a conditional mean linear in the full lag vector, plus rows of a fixed matrix drawn by a uniform permutation independent of U. On a surrogate, let $\ell^*_{\mathrm{sat},1}$ be the one-regime unrestricted Gaussian regression likelihood on the observed U (all dp lag coefficients, free covariance), $\ell^*_{\mathrm{sat},2}(z)$ its two-regime version, and, for a one-regime description $M^\dagger \in \mathscr{F}$, $\delta^*(M^\dagger) = \ell^*_{\mathrm{sat},1} - \hat{\ell}^*(M^\dagger) \geq 0$ the shortfall of its profiled likelihood. Every two-regime SEM likelihood is at most its saturated version and the charges are nonnegative, while the one-regime search, exact over $\mathscr{F}$ up to e_N, scores at least $\hat{\ell}^*(M^\dagger)$ minus an $O(\log N)$ charge. Hence, for every $M^\dagger \in \mathscr{F}$,

$$T_N^* \leq \sup_z \{\ell^*_{\mathrm{sat},2}(z) - \ell^*_{\mathrm{sat},1}\} + \delta^*(M^\dagger) + C\log N + e_N. \tag{27}$$

The first term is $o_p(N)$: on every block of rows the saturated regression of a surrogate has the same limit, C and the covariance of η, because the rows of η are drawn independently of U, and uniform block moment laws and (C2)–(C3) for the surrogate samples turn this into an $o_p(N)$ two-regime excess. The second term is $o_p(N)$ for a suitable $M^\dagger$ in two cases.

(a) *The complete DAG with every lag coefficient is a candidate.* It reparametrizes the saturated regression, so $\delta^* = 0$. It belongs to $\mathscr{F}$ when $d < 10$ and $d-1+dp \leq 16$ (Appendix A.8). Otherwise every equation has at most two lag coefficients, as at $d = 5$, $p = 3$, the order selected on every river window of Section 5.5, and this candidate is not available.

(b) *The fitted residuals are asymptotically uncorrelated.* Take $M^\dagger = \hat{M}$. At its fitted coefficients its structural residuals on the surrogate are exactly the rows $\hat{e}_{\pi(t)}$, whose second-moment matrix $\hat{S} = N^{-1}\hat{E}^\top\hat{E}$ does not depend on π, so $\hat{\ell}^*(\hat{M}) \geq -\frac{N}{2}\sum_j\{\log(2\pi\hat{S}_{jj}) + 1\}$. The saturated residual covariance is $(I-\hat{B})^{-\top}\hat{S}(I-\hat{B})^{-1}$ less the projection of η on U, which is $o_p(1)$ because the rows of η are drawn independently of U and the lags have mean zero under (1); with $\det(I-\hat{B}) = 1$,

$$\delta^*(\hat{M}) \leq \frac{N}{2}\Big(\sum_j \log\hat{S}_{jj} - \log\det\hat{S}\Big) + o_p(N),$$

and by Hadamard's inequality the bracket tends to zero exactly when the correlation matrix of $\hat{E}$ tends to the identity.

In either case (iii) follows from (C2)–(C3) for the surrogates. Otherwise the bound (27) can leave a term of order N, and (iii) is a condition of Theorem 1(ii) rather than a consequence of the others.

When is $\kappa > 0$? Let the true conditional moments be $\mathbb{E}[x_t \mid u_t, z_t = r] = \mu_r(u_t)$ and $\mathrm{Var}(x_t \mid u_t, z_t = r) = \Sigma_r \succ 0$. Enlarge $\mathscr{M}_0$ by allowing arbitrary regime-specific conditional means, every common complete order and every sign gauge; this can only raise its score. For an order σ and gauge D, let $\Sigma_0^D = \Sigma_0$, $\Sigma_1^D = D\Sigma_1 D$, let P_j be the variables preceding position j, and define the sequential regression quantities $C_{r,j} = (\Sigma_r^D)_{P_jP_j}$, $b_{r,j} = C_{r,j}^{-1}(\Sigma_r^D)_{P_j,\sigma(j)}$ and $\omega_{r,j} = (\Sigma_r^D)_{\sigma(j)\sigma(j)} - (\Sigma_r^D)_{\sigma(j)P_j}b_{r,j}$. Set

$$\kappa_{\mathrm{cov}} = \min_{\sigma,D}\ \frac{1}{2}\sum_{j=2}^{d}\ \inf_{b\in\mathbb{R}^{j-1}}\ \sum_{r=0}^{1}\pi_r\log\Big(1 + \frac{(b-b_{r,j})^\top C_{r,j}(b-b_{r,j})}{\omega_{r,j}}\Big). \tag{28}$$

Lemma 20 (Covariance certificate)**.** *$\kappa \geq \kappa_{\mathrm{cov}}$, and $\kappa_{\mathrm{cov}} > 0$ if and only if for every (σ, D) some j has $b_{0,j} \neq b_{1,j}$. If some (σ, D, j) has $\|b_{0,j} - b_{1,j}\| = \delta > 0$, then its summand is at least $\frac{\pi_{\min}}{2}\log(1 + a\delta^2/4)$ with $a = \min_r \lambda_{\min}(C_{r,j})/\omega_{r,j}$.*

Proof. For a common complete order and common coefficient b in equation j, the expected residual variance in regime r after optimizing the conditional mean is $\omega_{r,j}+(b-b_{r,j})^\top C_{r,j}(b-b_{r,j})$; optimizing the free regime-specific variance makes the expected Gaussian log-likelihood a constant minus half its logarithm. Regime-specific coefficients factorize each Σ_r exactly, and subtracting that unrestricted value gives the summand in (28); equations are optimized independently, so gaps add. A sparse common DAG embeds in a complete order with zero coefficients and a shared-variance equation in its graded version, so minimizing over the finite set of orders and gauges bounds $L_\star - L_0$ from below. Each summand is nonnegative, zero when the two coefficient vectors agree, and, when they differ, continuous and coercive in b with a positive minimum because $C_{r,j} \succ 0$, $\omega_{r,j} > 0$ and $\pi_r > 0$. For the explicit bound, every b is at distance at least $\delta/2$ from one of the two vectors; keep that term and drop the rest. □

The minimum in (28) ranges over all sign gauges, so the bound holds for $\mathscr{H}_0$ (the identity gauge) and for descriptions shared up to any gauge alike.

Examples. *Two-variable reversal, unit innovation variances.* With $x_b = \beta x_a + \varepsilon_b$ in regime 0 and $x_a = \beta x_b + \varepsilon_a$ in regime 1, the two regression slopes in order (a,b) have magnitudes $|\beta|$ and $|\beta|/(1+\beta^2)$, which a sign gauge cannot equalize, and the normalized regressor variances are 1 and $(1+\beta^2)^2$; the other order exchanges the regimes. Lemma 20 gives $\kappa \geq \frac{\pi_{\min}}{2}\log\big(1+\beta^6/(4(1+\beta^2)^2)\big) > 0$. For balanced regimes the best one-regime covariance is $(\Sigma_0+\Sigma_1)/2$ and $\Delta = \frac{1}{2}\log(1+\beta^4/4)$, while the common slope midway between the two gives $\kappa = O(\beta^6)$: separating two mechanisms from a scale change is a smaller population signal than separating two laws from one.

Anchored reversal. An isolated anchor is a variable c whose conditional covariance with every other variable is zero in one regime and nonzero with at least one variable in the other. Relabel the regimes so that $(\Sigma_1)_{-c,c} = 0$ and $q_c = \|(\Sigma_0)_{-c,c}\|_2 > 0$, where $-c$ denotes all variables except c. Set $m_\Sigma = \min_r \lambda_{\min}(\Sigma_r)$, $M_\Sigma = \max_r \lambda_{\max}(\Sigma_r)$ and $\pi_{\min} = \min_r \pi_r$. Then

$$\kappa \;\geq\; \kappa_{\mathrm{cov}} \;\geq\; \tfrac{\pi_{\min}}{2}\log\Big(1+\frac{m_\Sigma^2 q_c^2}{8M_\Sigma^4}\Big) > 0. \tag{29}$$

To prove this, fix any order and sign gauge, and put c at position k. In this order let H_r be the strictly lower matrix of sequential regression coefficients from Σ_r^D, $T = I - H_0$, and $\Omega = T\Sigma_0 T^\top$ the diagonal covariance of the regression residuals. Each residual variance minimizes the variance of a linear contrast whose response coefficient is one, so it lies in $[m_\Sigma, M_\Sigma]$. Hence $\|T^{-1}\|_2 \leq \sqrt{M_\Sigma/m_\Sigma}$ from $\Sigma_0 = T^{-1}\Omega T^{-\top}$. Isolation is preserved by every sign gauge and makes all coefficients incident to c in H_1 zero. With $\mathbf{e}_k$ the kth coordinate vector, set $u = T\mathbf{e}_k - \mathbf{e}_k$, $w = T^\top \mathbf{e}_k - \mathbf{e}_k$ and $\nu_k = \Omega_{kk}$. Thus

$$\sum_{j=2}^{d} \|b_{0,j} - b_{1,j}\|_2^2 = \|H_0 - H_1\|_F^2 \geq \|u\|_2^2 + \|w\|_2^2.$$

The identity $T\Sigma_0 T^\top \mathbf{e}_k = \nu_k \mathbf{e}_k$ gives

$$\Sigma_0 \mathbf{e}_k - \nu_k \mathbf{e}_k = -\Sigma_0 w - \nu_k T^{-1} u.$$

Restricting to coordinates other than k and taking norms yields

$$q_c \leq M_\Sigma \|w\|_2 + M_\Sigma\sqrt{M_\Sigma/m_\Sigma}\,\|u\|_2 \leq \sqrt{2M_\Sigma^3/m_\Sigma}\,\big(\|u\|_2^2 + \|w\|_2^2\big)^{1/2}.$$

Consequently $\sum_j \|b_{0,j} - b_{1,j}\|_2^2 \geq q_c^2 m_\Sigma/(2M_\Sigma^3)$. In Lemma 20, each equation has $a \geq m_\Sigma/M_\Sigma$. Sum its bounds over equations and use $\sum_j \log(1+x_j) \geq \log(1+\sum_j x_j)$ for $x_j \geq 0$ to obtain (29), uniformly over orders and gauges.

For the centered conditional variables in the five-variable reversal motif, $b = \beta a + \gamma c + \varepsilon_b$ in regime 0, while c is isolated in regime 1. Exogeneity and independent innovations give $(\Sigma_0)_{cb} = \gamma\omega_{0,c}$, so $q_c \geq |\gamma|\omega_{0,c} > 0$ whenever $\gamma \neq 0$. The anchor $d \to a$ gives the symmetric result after exchanging the regimes. Adding the shared child e of a and b leaves the marginal covariance $(\Sigma_0)_{cb}$ and regime-1 isolation of c unchanged. Hence (29) covers this motif with the shared child and any positive innovation variances; no faithfulness assumption is required.

An unanchored failure. In the two-variable reversal with $0 < |\beta| < 1$ and innovation variances $\omega_a = (1-\beta^2)\omega_b$ in both regimes, both covariance matrices are also generated by the common orientation $a \to b$ with slope β and the two variance pairs exchanged. An all-graded description then fits both regimes exactly and $\kappa_{\text{cov}} = 0$; under Gaussian innovations without lags the two explanations are the same observational law, so no method that treats all graded descriptions as one mechanism can report a reversal consistently there. The anchor is identification information, not a convenience of the proof. □

A.10 THE LABEL REFINEMENT: EXACT PROPOSALS AND SCORE ASCENT

Fix the rows, the lag order p, the minimum run $m = 5d(1+p)$ and the switch cap $K = 8$ of Appendix B, and let $\mathscr{Z}$ be the labelings whose runs have at least m rows, that use both regimes and that switch at most K times. Write (4) as $Q(z,M) = \hat{\ell}(z,M) - P_M(z) - C(M) - \lambda J(z)$ with $\lambda = \log N$, where $\hat{\ell}$ is the profiled log-likelihood, $P_M(z) = \frac{1}{2}\sum_\vartheta \log n_\vartheta(z)$ the parameter charge and $C(M)$ the remaining code lengths. At an incumbent (z_k, M_k) the refinement freezes its per-row emissions $e^{(k)}_{t,r}$, Gaussian or Student-t_4, and passes $F_k(z) = \sum_t e^{(k)}_{t,z_t} - \lambda J(z)$ to the dynamic program; the proposal z'_k is refitted to (z'_k, M'_k) and compared with the incumbent on the complete score.

Proposition 21 (Exact proposals and accepted-score ascent)**.** *(a) The constrained two-state dynamic program returns a global maximizer of F_k over $\mathscr{Z}$; if the incumbent labeling is feasible, $F_k(z'_k) \geq F_k(z_k)$.*

(b) If a proposal is retained only when $Q(z'_k, M'_k) > Q(z_k, M_k)$ and the incumbent is kept otherwise, the incumbent's score is nondecreasing and every accepted refinement increases it strictly, for Gaussian or Student-t proposal emissions and for any candidate generation or graph search, screened or local.

(c) If every possible incumbent lies in a finite set of (labeling, description) pairs with one deterministic finite profiled score each, no retained pair recurs after an accepted refinement, so an uncapped loop that stops at a round without acceptance makes finitely many accepted refinements; the implemented three-round cap terminates directly, and the best-of-starts score is at least each initialization's score.

Proof. (a) Let $E_r(t) = \sum_{u \leq t} e_{u,r}$ and let $H(t,j,r)$ be the best frozen-emission score of the first t rows over labelings with exactly j switches, ending in state r, every run at least m rows; infeasible entries are $-\infty$. Then $H(t,0,r) = E_r(t)$ for $t \geq m$ and, for $j \geq 1$, $H(t,j,r) = \max_{jm \leq s \leq t-m}\{H(s, j-1, 1-r) + E_r(t) - E_r(s) - \lambda\}$; the value returned is $\max_r \max_{1 \leq j \leq \min(K, \lfloor N/m \rfloor - 1)} H(N,j,r)$, and $j \geq 1$ makes both states occur. Every feasible labeling decomposes uniquely at its last switch into a feasible prefix and a final run of at least m rows charged one switch, and every finite transition builds a feasible labeling, so induction on j proves the recurrence and traceback returns a maximizer. The incumbent lies in the maximization set. (b) Acceptance increases Q by definition and rejection leaves it unchanged; the argument never identifies F_k with Q, so it does not depend on the emission family or on the optimality of the graph search. (c) A recurring pair after an accepted refinement would receive two different scores. The graphs, equation classes, lag supports, gauges and $\mathscr{Z}$ are finite, and a label-dependent parent screen only selects subsets of this finite universe; the round cap bounds the loop without this assumption, and the maximum over runs keeps each run's initial score as a lower bound. □

Proposition 22 (Gaussian gain certificate)**.** *Let $\hat{\eta}_k$ attain $\hat{\ell}(z_k, M_k)$ and let the frozen emissions be the Gaussian row log-likelihoods at $\hat{\eta}_k$, so that $F_k(z) = \ell(z, M_k, \hat{\eta}_k) - \lambda J(z)$. If M_k is kept and its parameter domain is feasible at z'_k, then $Q(z'_k, M_k) - Q(z_k, M_k) \geq F_k(z'_k) - F_k(z_k) - [P_{M_k}(z'_k) - P_{M_k}(z_k)]$; if the refitted candidate satisfies $Q(z'_k, M'_k) \geq Q(z'_k, M_k)$, the same bound holds for it. A frozen-emission gain larger than the change in parameter charge therefore guarantees acceptance, and when the regime row counts are unchanged any positive gain does. For a description with pooled or regime-specific parameters, $P_M(z') - P_M(z) = \sum_r a_r(M) \log\{n_r(z')/n_r(z)\}$, where $a_r(M)$ is half the number of parameters estimated from regime r alone, graded variances included.*

Proof. Profiling at z'_k may keep $\hat{\eta}_k$, so $\hat{\ell}(z'_k, M_k) \geq \ell(z'_k, M_k, \hat{\eta}_k)$, with equality at z_k; subtracting, $C(M_k)$ cancels and the switch terms are inside F_k. The candidate version follows from the assumed

domination, which holds when the graph search is exact over a candidate set containing M_k and can fail when a new screen removes it. □

A.11 Codes and search moves

Proposition 23 (Multiplicity code)**.** $\log\binom{d}{k}$ *is, up to a constant, the code of the changed set under a beta-binomial prior with a uniform prior on the change probability (Scott & Berger, 2010). For a graded split of an unchanged equation, the local charge is* $\frac{1}{2}\log(n_0 n_1/N)$ *and the null gain is* $\frac{1}{2}\chi^2_1$*. Declaring one more changed equation when k are changed therefore has false-split probability* $\Pr\left(\chi^2_1 > \log(n_0 n_1/N) + 2\log\frac{d-k}{k+1}\right)$.

Proof. With $p \sim U(0,1)$ and changes i.i.d. Bernoulli(p), a given set of size k has probability $\int_0^1 p^k(1-p)^{d-k}dp = 1/((d+1)\binom{d}{k})$. The rest is the χ^2_1 tail. □

Proposition 24 (Bounded evidence under Student-t emissions)**.** *Let two fitted regimes differ only in scale, with* $s_{1j} = \rho_j s_{0j}$*. With Student-*t_ν *emissions, each equation's contribution to the per-row log ratio is monotone in* r_j^2*. It runs from* $-\frac{1}{2}\log\rho_j$ *to* $\frac{\nu}{2}\log\rho_j$*, so* $|\ell(r)| \le C := \frac{\nu}{2}\sum_j |\log\rho_j|$ *uniformly in the residual r. Gaussian emissions are unbounded. A relabeling DP with switch charge* $\log N$ *therefore needs a run of at least* $\lceil 2\log N/C\rceil$ *rows to switch under t emissions, but a single extreme row can suffice under Gaussian ones.*

Proof. Write the t term as $\frac{\nu}{2}\log\rho_j - \frac{\nu+1}{2}\log\frac{\nu s_{1j}+r^2}{\nu s_{0j}+r^2}$. The last logarithm is monotone in r^2, from $\log\rho_j$ to 0. □

Relabeling moves change only which labelings are proposed; a step is kept when Q rises. The moves are part of the procedure that is repeated on every surrogate, so T remains a fixed function of the data, as Theorem 11 requires.

Proposition 25 (Rank features are robust)**.** *Summarize a tile of b rows by the Spearman correlations of its columns. The features are unchanged by any strictly increasing transformation of a column within the tile, and in particular by rescaling the tile. Replacing one row by arbitrary values moves every feature by at most* $12/(b+1)$.

Proof. Ranks are invariant under increasing maps. Spearman's coefficient is $\sum_i r_i s_i/V$, with centered ranks r_i, s_i and $V = b(b^2-1)/12$. Moving one row changes its rank by some m and shifts m other ranks by one in the opposite direction. Hence $|\sum_i (r'_i - r_i)s'_i| \le m(2b-1-m)/2 \le b(b-1)/2$, and likewise for the other column. So $|\Delta\rho| \le b(b-1)/V = 12/(b+1)$. □

B Algorithm details and settings

Every setting below is fixed in `seam_v3/seam.py`, and the reporting rule in `seam_v3/decide.py`. A unit test guards the configuration, and a second test checks that the entry point calls the estimator exactly as the development runs did.

Algorithm 1 gives the procedure in full. Each step is the one described in Section 3, and the settings below fix every choice. The lines after T is formed are the reporting rule of Section 3, in the order in which the implementation applies it: a two-regime description that does not score higher than the one-regime description is discarded before any surrogate is drawn; the test is run only when $T > 0$; and the equations of the accepted description are classified before the report is chosen, so that two regimes are reported only when some coefficient differs between the regimes in the original coordinates. A graded equation (same coefficients, different innovation variance) does not count as a change.

Algorithm 1 SEAM (all settings fixed; `seam_v3`)

Require: series $x_{1:N}$, level α

$p \leftarrow \arg\min_{p \le 4} \mathrm{BIC}(\text{pooled VAR}(p))$; $U \leftarrow [x_{t-1}, \ldots, x_{t-p}]$
$Q_1 \leftarrow$ exact one-regime optimum of (4)
for each initial labeling z (principal axis; rank correlation) **do**
  repeat
    $M \leftarrow$ exact integer program on z ▷ all pairs of DAGs; identity gauge
    $z' \leftarrow$ two-state DP on the emissions of M (Gaussian or Student-t)
    accept (z', M') only if Q increases
  until no accepted step (at most 3 rounds)
$Q_2 \leftarrow$ best Q found; $T \leftarrow Q_2 - Q_1$
if $T \le 0$ **then return** one-regime graph ▷ no gain; $\hat{p} = 1$, no surrogates drawn
$\hat{p} \leftarrow$ permutation p-value of T over fixed-design surrogates (same code, search included)
if $\hat{p} > \alpha$ **then return** one-regime graph
for each candidate gauge g (the identity and the sign patterns ranked below) **do**
  re-solve the integer program on $\hat{z}$ with sharing up to g ▷ both graphs re-optimized
$\hat{M} \leftarrow$ the best of these; refine $\hat{z}$ with it, re-solving every gauge, while Q increases
with the two graphs of $\hat{M}$ fixed, classify each equation: shared, shared up to the gauge, graded or regime-specific
classify coefficient changes in the original coordinates
if no coefficient of $\hat{M}$ differs between the regimes **then return** one-regime graph and the estimated ratio of innovation scales ▷ scale change
else return regimes $\hat{z}$, the two graphs, the changed coefficients, effects (2) ▷ two mechanisms

Preprocessing and lag order. Columns are scaled to unit root mean square; the method is equivariant to column scaling. The lag order $p \in \{1, \ldots, 4\}$ minimizes the Schwarz criterion

$$n \log \det \hat{\Sigma}_p + p\, d^2 \log n$$

of the pooled VAR. All orders are fitted on the same rows, without an intercept, and ties go to the smaller order. The design uses $u_t = (x_{t-1}, \ldots, x_{t-p})$.

Equation types and costs. For each variable and candidate parent set, the local costs are computed from per-regime second-moment matrices.

- *Shared equation:* pooled least squares.
- *Regime-specific equation:* least squares within each regime.
- *Graded equation:* common coefficients with regime-specific variances. Its exact profile likelihood reduces to a one-dimensional search over the variance ratio ρ with weighted least squares at each ρ.

Lag sets are selected inside each local cost by the same charges.

Initial labelings.

- *Principal-axis split.* The series is cut into tiles of $b = \max(m, \lceil \sqrt{N} \rceil)$ rows. Each tile is summarized by its whitened second moments, and tiles are split at the midpoint of the first principal axis of these summaries. Provisional regressions per group then give emissions for the exact two-state DP.
- *Rank split.* The same procedure on the upper triangle of each tile's Spearman correlation matrix. These features are unchanged by monotone transformations within a tile. A single row moves any feature by at most $12/(b+1)$ (Proposition 25).

Each labeling is refined with Gaussian emissions and with Student-t_4 emissions (Proposition 24), for up to three rounds. A step is kept only if Q rises (Proposition 21).

Segmentation. The two-state DP, exact for fixed emissions (Appendix A.10), enforces three constraints:

- a minimum run of $m = 5d(1+p)$ rows;

- at most 8 switches;
- a charge of $\log N$ per switch.

Both regimes must occur.

Integer program. The program is solved with HiGHS, single-threaded, with lazily added cluster cuts and dominance pruning of parent sets.

- $d < 10$*:* every subset of the other variables is a candidate parent set.
- $d \geq 10$*, statistic:* candidates are screened to the 8 strongest by the maximal absolute partial correlation over the pooled and provisional-regime moments, with in-degree at most 3. The statistic uses a local search over variable orders.
- $d \geq 10$*, reported description:* solved exactly on the selected labels, with candidates re-screened on those labels and no in-degree cap.

Gauge. The statistic T, on the data and on every surrogate, is computed in the identity gauge: its equation types are shared, graded and regime-specific in the data's coordinates, and no sign alignment enters the test. The gauge enters only the reporting step, on the selected labels. The candidate gauges are the identity and the three sign patterns g that minimize the saturated criterion $\log\det(\pi_0\hat{\Sigma}_0 + \pi_1 g\hat{\Sigma}_1 g)$, ranked exhaustively for $d \leq 12$; for $d \geq 10$, one pattern is used besides the identity. For each candidate gauge the integer program is re-solved, both graphs included, with the second regime written in g's coordinates, so that its shared and graded equations are pooled up to g; the lag block follows the signs of x when $p = 1$ and is left unreflected, like an exogenous input, when $p \geq 2$. The best-scoring gauge fit is kept; label refinements under this model re-solve every gauge again and are kept when the score rises. Afterwards, with the two graphs of that fit fixed, each equation is classified as

- shared exactly;
- shared up to the gauge;
- graded;
- regime-specific,

an exact choice once the graphs are fixed. The mechanism-change report compares coefficients in the original coordinates; an equation pooled after sign alignment can still be reported as changed.

Calibration. The one-regime model is fitted by the same program. Surrogates are

$$X^* = (U\hat{A} + \hat{E}_\pi)(I - \hat{B})^{-1},$$

with U the observed lags. T is recomputed with identical code and seeds.

- $d < 10$*:* at most $R = 39$ surrogates, drawn one at a time and stopped at the $h = 10$th exceedance $T(X_s^*) \geq T(X)$ (Besag & Clifford, 1991). With $S \leq R$ the number drawn and B_S the exceedances among them, the returned value is $\hat{p} = h/S$ if the h-th exceedance was reached and $\hat{p} = (1 + B_R)/(R+1)$ otherwise (6). Ties count as exceedances, and so does a surrogate on which the search fails (its statistic is set to $+\infty$).
- $d \geq 10$*:* $R = 19$ and $h = 5$.

The smallest attainable p-value is $1/(R+1)$, which is 0.025 and 0.05 respectively, and the sequential value rejects at $\alpha = 0.05$ exactly when the full budget would, whatever the surrogates (Corollary 13). The finite-sample level of Theorem 11 belongs to a test that reorders the observed rows instead; the two coincide only for an empty lag design, which the lag-order step never selects (Appendix A.7).

Cost. The integer program has at most $d \cdot 4 \cdot |\mathscr{P}|$ binary variables, where $|\mathscr{P}|$ is the number of candidate parent sets per variable. There is no exponential state space over orders. Table 16 reports wall-clock times from the confirmatory run.

C Confirmatory protocol

Protocol. The estimator, designs, baseline settings, endpoints and analysis script were fixed before evaluation, using development and evaluation seeds from disjoint ranges. The decision rule of

Section 3 was developed on the development data. The supplementary archive provides the protocol, seed assignments, execution records and verification script.

Blocks reported here.

Block	Cells	Reps	Arms
C1 reversals	$d = 5$, $N = 1200$; 2 schedules $\times$ 4 noises $\times$ {unit, heteroscedastic} innovation scales $\times$ $\beta \in \{0.3, 0.5\}$	20	SEAM, test-only ablation, ceiling, CASTOR$\times$2, FlowMSM$\times$2
C1b	the same at $\beta = 0.5$, Gaussian and Causal-Rivers noise	10	Regime-PCMCI[a], SPACETIME[a]
C2 scale	$d \in \{10, 20\} \times N \in \{2400, 4800\}\times$ 2 schedules $\times$ 2 noises, $\beta = 0.5$	10	SEAM, test only, CASTOR (windowed), FlowMSM$\times$2
C3 nulls	VAR(1) $\rho \in \{0.5, 0.8, 0.95\} \times N \in \{1200, 4800\}\times$ 4 noises; VAR(2) $\times$ 2 lengths $\times$ 2 noises	50	SEAM, test only, CASTOR[b], FlowMSM$\times$2[b]
C4 theory	power grid $\beta \times N$ from $N = 600$ to 4800; reflection counterexample $N = 600 \ldots 38400$	20	SEAM, test only; ceiling, no quotient, no sharing
C5 benchmarks	CASTOR's generator $d \in \{5, 10\}$; the Regime-PCMCI and SPACETIME tutorials	20	SEAM and the benchmark's own method
C6 sharing	the C1 data sets, true labels	20	sharing, no sharing, ceiling
C7 FANTOM	the C1 primary cell ($\beta = 0.5$, unit scales, 4 noises $\times$ 2 schedules)	10	FANTOM
C9 stress	t_3 innovations: reversals and one-regime series	20/50	SEAM, Student-t statistic, test only, CASTOR

[a] run time: about an hour a series for SPACETIME and ten minutes for Regime-PCMCI at $d = 5$. [b] the VAR(1) cells with $N = 1200$.

Baseline settings. Baselines use their authors' code; the settings and evaluated variants are listed below.

- **CASTOR.** Window 200, minimum regime 100 rows, edge threshold 0.4 and 5 EM iterations; we also run threshold 0.1, which suited our data better. Above $N = 1200$ the window is $N/6$ and the minimum regime half the window: the authors' window over-splits long series, and a minimum regime as long as the window crashes the code.
- **Regime-PCMCI.** tigramite 5.2.10.1 with its defaults (ParCorr, $\tau_{\max} = 1$), given the true number of regimes and the true number of transitions plus one, as in its tutorial.
- **SPACETIME.** The package defaults: linear score, regime clustering, $\tau_{\max} = 1$.
- **FlowMSM.** The authors' defaults (10 restarts, up to 500 epochs, early stopping, initial regime persistence 20), with the model's regime count set to 2 — the true number, as Regime-PCMCI is given it — and both causal-discovery backends, DYNOTEARS and VARLiNGAM.
- **FANTOM.** The authors' code as it is vendored in the FlowMSM repository, with window and minimum regime length matched to CASTOR's (200 and 100), lag 1, and *the authors' training budget of 2000 inner epochs* rather than the 300 the vendored configuration uses to speed training up. At 300 epochs its error on our development series was 0.740; at 2000 it was 0.122, and we report the stronger setting. In this vendored implementation, `FANTOM.fit` takes `model_nr_regimes = T // window` and fits one regime per window, so the number of regimes is the series length divided by the window rather than a selected output. This describes the evaluated implementation. At 200 on a 1201-row series that is 6; the authors' own default of 10 would give 120. Because this was not what the protocol intended when it listed FANTOM among the methods given the true number, we also ran it at a window of 600, which yields two, and report both. At window 600 it returns 2 regimes on all 80 series and scores $E = 0.570$ with focal success 0.25, against $E = 0.406$ and 0.00 at window 200 (Table 1): given the number of regimes it orients the focal pair on a quarter of the series but places the boundary less well.
- **Refitting.** FlowMSM and FANTOM return binary graphs, so their coefficients are refit by least squares on the rows they assign to each regime, restricted to their own graph — the treatment SPACETIME and Regime-PCMCI already get.
- **Wall-clock budget.** Two hours per data set and method, three for FANTOM, and five minutes for CASTOR (twenty at $d \geq 10$). CASTOR's running time is bimodal: across 3,236 runs it returns in a

median of 26 seconds or it does not return at all, and no run has ever finished between 260 seconds and the original two-hour cap; five minutes is three times the slowest completed run at $d = 5$ and twelve times the median. A run that exceeds its budget, or crashes, counts as a failure with $E = 1$ and no decision, and is tabulated separately.

Endpoints and analysis. The primary endpoint is the focal error E of (3), which the protocol calls the label-matched focal effect error: each true regime is compared with the fitted regime that contains most of its time points. Per cell we report the mean E with a 95% bootstrap interval over data sets, focal success with Wilson intervals, and false alarms with Clopper–Pearson intervals; comparisons are paired on the same data sets, with a bootstrap interval for the mean difference and the count of wins. A run reported as a change of scale rather than a change of mechanism is recorded as such and is not counted as a false alarm; the count is reported.

The prespecified expectations for the reported comparisons included: a false-alarm interval containing or below 0.05 with no cell above $6/50$; mean $E \leq 0.10$ in every $\beta = 0.5$ reversal cell with a paired advantage over each baseline; and against the test-only ablation, detection no more than two percentage points lower and E no more than 2% higher. For detection we fit

$$\operatorname{logit}\Pr(K = 2) = a + b\log(N\Delta) + c\log N + \textstyle\sum_\beta \gamma_\beta$$

per noise law, with Δ computed per seed from 40,000-row series at the true labels.

Prespecified statements for the reported comparisons. Each statement below was prespecified with the quantity that measures it; the right-hand column is computed by the analysis script from the records.

prespecified statement	what the records show
reversed effects recovered where a single order cannot	worst $\beta = 0.5$ cell $E = 0.037$, paired advantage over all 8 baseline configurations
it scales to $d = 10$ and $d = 20$	worst cell $E = 0.004$
the test controls false alarms	mechanism reports: 0.36% [0.12, 0.83], worst cell 1/50
the two report conditions do not cost detection	detection 0.967 against 0.972
no reversal data set is called a scale change	1 of 800
half of the test-only alarms come back as scale changes	37 of 72 relabelled; 30 more removed by the gain condition, 5 left

The statements keep their prespecified wording. In the false-alarm row, 0.36% counts final reports of two mechanisms, not test rejections: the test rejected on 42 of the 1400 one-regime series (3.0%), and 37 of these were reported as scale changes (Section 5.1).

Departures from the protocol. Departures are recorded with their time in the protocol file; the ones that bear on the reported results are these.

- **CASTOR's budget** was cut from two hours to five minutes a series (twenty at $d \geq 10$) during the run, in two steps, for the reason given above; capped runs are reported as failures to terminate, and the counts are in Table 16. No completed run was affected: none has ever finished in the window the change removes.
- **The stress block was narrowed before it ran.** The two tanh cells were dropped: the mild one was already fine without the two conditions, and the strong one tests misspecification of the linear model, a limit the paper states in words rather than measures again. The t_3 cells were kept and the Student-t statistic was added to them.
- **FANTOM's replicates** were reduced from 20 to 10 of the primary cell, because a series costs 49.9 minutes at the authors' training budget.
- **FlowMSM's replicates** were halved during the run, for machine time and before its results were examined: it ran the first 10 replicates of each C1 cell and 25 of each C3 cell and kept the 63 C1 series it had already fitted. The removed series, 257 in C1 and 300 in C3 with both backends, were fitted after the run, once its results on the halved plan had been examined, with the same code, seeds and settings, so FlowMSM is reported on the complete design; refitting 16 of its original series in the same way reproduced every stored field but the run time. In the original job its focal errors at $\beta = 0.5$ were 0.517 and 0.513 over 192 series, against 0.537 and 0.520 over 320 on the complete design. Its run times (Table 16) are those of the original job.

- **The sharing arms of C1** were run on a separate machine rather than on the cloud machines, with the same code, plan and seeds, because the estimator's package was not in the machines' bundle at launch.
- **Packaging and scheduling faults** — a module missing from the bundle, a worker count read from the wrong source, a dispatch that did not know the new arm's name, and the CASTOR budget above — cost machine time and were fixed by restarting from the records already written; the records produced by the packaging fault, which were failures of packaging and not of the method, were discarded and those tasks recomputed.

D ADDITIONAL RESULTS

D.1 ORDER REVERSALS: PAIRED COMPARISON AND PER-CELL RESULTS

Table 4: **Paired comparison on the same series** (every C1 and C1b cell in which the baseline was run). Mean advantage in effect error (baseline minus SEAM; positive favors SEAM) with a 95% bootstrap interval over series, and the number of series on which SEAM is better.

baseline	n	mean E advantage	95% CI	SEAM better on
FANTOM	80	0.399	[0.320, 0.480]	78/80
FlowMSM + VARLiNGAM	640	0.524	[0.471, 0.579]	596/640
FANTOM (window 600)	80	0.564	[0.478, 0.651]	80/80
FlowMSM + DYNOTEARS	640	0.662	[0.594, 0.736]	607/640
SPACETIME	40	0.678	[0.619, 0.741]	40/40
CASTOR ($\tau = 0.1$)	640	0.784	[0.698, 0.880]	633/640
Regime-PCMCI	40	0.994	[0.992, 0.996]	40/40
CASTOR (authors)	640	1.042	[0.963, 1.133]	639/640

Table 5: **Per cell,** $\beta = 0.3$ ($d = 5$, $N = 1200$, 20 series a cell): SEAM as reported, the test-only ablation, and SEAM given the true regimes. "scales" is whether the innovations have unit scale or per-variable scales drawn from $[1/2, 2]$ (heteroscedastic; the same scales in both regimes).

scales	innovations	schedule	method	n	E	focal
unit	gaussian	recurrent	SEAM, test only	20	0.069	0.90
unit	gaussian	recurrent	SEAM	20	0.069	0.90
unit	gaussian	recurrent	SEAM, true regimes	20	0.045	0.95
unit	gaussian	switch	SEAM, test only	20	0.014	1.00
unit	gaussian	switch	SEAM	20	0.014	1.00
unit	gaussian	switch	SEAM, true regimes	20	0.013	1.00
unit	macro	recurrent	SEAM, test only	20	0.119	0.80
unit	macro	recurrent	SEAM	20	0.119	0.80
unit	macro	recurrent	SEAM, true regimes	20	0.068	0.90
unit	macro	switch	SEAM, test only	20	0.049	0.95
unit	macro	switch	SEAM	20	0.049	0.95
unit	macro	switch	SEAM, true regimes	20	0.048	0.95
unit	rivers	recurrent	SEAM, test only	20	0.192	0.65
unit	rivers	recurrent	SEAM	20	0.192	0.65
unit	rivers	recurrent	SEAM, true regimes	20	0.014	1.00
unit	rivers	switch	SEAM, test only	20	0.089	0.85
unit	rivers	switch	SEAM	20	0.089	0.85
unit	rivers	switch	SEAM, true regimes	20	0.062	0.90
unit	skab	recurrent	SEAM, test only	20	0.039	0.95
unit	skab	recurrent	SEAM	20	0.064	0.90
unit	skab	recurrent	SEAM, true regimes	20	0.038	0.95
unit	skab	switch	SEAM, test only	20	0.044	0.95
unit	skab	switch	SEAM	20	0.044	0.95
unit	skab	switch	SEAM, true regimes	20	0.069	0.95
hetero	gaussian	recurrent	SEAM, test only	20	0.190	0.65
hetero	gaussian	recurrent	SEAM	20	0.191	0.65
hetero	gaussian	recurrent	SEAM, true regimes	20	0.192	0.65
hetero	gaussian	switch	SEAM, test only	20	0.137	0.75
hetero	gaussian	switch	SEAM	20	0.137	0.75
hetero	gaussian	switch	SEAM, true regimes	20	0.137	0.75
hetero	macro	recurrent	SEAM, test only	20	0.265	0.50
hetero	macro	recurrent	SEAM	20	0.265	0.50
hetero	macro	recurrent	SEAM, true regimes	20	0.217	0.60
hetero	macro	switch	SEAM, test only	20	0.175	0.70
hetero	macro	switch	SEAM	20	0.175	0.70
hetero	macro	switch	SEAM, true regimes	20	0.173	0.70
hetero	rivers	recurrent	SEAM, test only	20	0.166	0.70
hetero	rivers	recurrent	SEAM	20	0.166	0.70
hetero	rivers	recurrent	SEAM, true regimes	20	0.115	0.80
hetero	rivers	switch	SEAM, test only	20	0.169	0.70
hetero	rivers	switch	SEAM	20	0.169	0.70
hetero	rivers	switch	SEAM, true regimes	20	0.165	0.70
hetero	skab	recurrent	SEAM, test only	20	0.123	0.80
hetero	skab	recurrent	SEAM	20	0.123	0.80
hetero	skab	recurrent	SEAM, true regimes	20	0.093	0.85
hetero	skab	switch	SEAM, test only	20	0.268	0.50
hetero	skab	switch	SEAM	20	0.268	0.50
hetero	skab	switch	SEAM, true regimes	20	0.219	0.60

Table 6: **Per cell,** $\beta = 0.5$; columns as in Table 5.

scales	innovations	schedule	method	n	E	focal
unit	gaussian	recurrent	SEAM, test only	20	0.007	1.00
unit	gaussian	recurrent	SEAM	20	0.007	1.00
unit	gaussian	recurrent	SEAM, true regimes	20	0.007	1.00
unit	gaussian	switch	SEAM, test only	20	0.005	1.00
unit	gaussian	switch	SEAM	20	0.005	1.00
unit	gaussian	switch	SEAM, true regimes	20	0.004	1.00
unit	macro	recurrent	SEAM, test only	20	0.007	1.00
unit	macro	recurrent	SEAM	20	0.007	1.00
unit	macro	recurrent	SEAM, true regimes	20	0.007	1.00
unit	macro	switch	SEAM, test only	20	0.009	1.00
unit	macro	switch	SEAM	20	0.009	1.00
unit	macro	switch	SEAM, true regimes	20	0.009	1.00
unit	rivers	recurrent	SEAM, test only	20	0.030	0.95
unit	rivers	recurrent	SEAM	20	0.030	0.95
unit	rivers	recurrent	SEAM, true regimes	20	0.005	1.00
unit	rivers	switch	SEAM, test only	20	0.005	1.00
unit	rivers	switch	SEAM	20	0.005	1.00
unit	rivers	switch	SEAM, true regimes	20	0.005	1.00
unit	skab	recurrent	SEAM, test only	20	0.005	1.00
unit	skab	recurrent	SEAM	20	0.005	1.00
unit	skab	recurrent	SEAM, true regimes	20	0.005	1.00
unit	skab	switch	SEAM, test only	20	0.007	1.00
unit	skab	switch	SEAM	20	0.007	1.00
unit	skab	switch	SEAM, true regimes	20	0.007	1.00
hetero	gaussian	recurrent	SEAM, test only	20	0.034	0.95
hetero	gaussian	recurrent	SEAM	20	0.034	0.95
hetero	gaussian	recurrent	SEAM, true regimes	20	0.034	0.95
hetero	gaussian	switch	SEAM, test only	20	0.007	1.00
hetero	gaussian	switch	SEAM	20	0.007	1.00
hetero	gaussian	switch	SEAM, true regimes	20	0.031	0.95
hetero	macro	recurrent	SEAM, test only	20	0.014	1.00
hetero	macro	recurrent	SEAM	20	0.014	1.00
hetero	macro	recurrent	SEAM, true regimes	20	0.038	0.95
hetero	macro	switch	SEAM, test only	20	0.037	0.95
hetero	macro	switch	SEAM	20	0.037	0.95
hetero	macro	switch	SEAM, true regimes	20	0.037	0.95
hetero	rivers	recurrent	SEAM, test only	20	0.011	1.00
hetero	rivers	recurrent	SEAM	20	0.011	1.00
hetero	rivers	recurrent	SEAM, true regimes	20	0.011	1.00
hetero	rivers	switch	SEAM, test only	20	0.032	0.95
hetero	rivers	switch	SEAM	20	0.032	0.95
hetero	rivers	switch	SEAM, true regimes	20	0.032	0.95
hetero	skab	recurrent	SEAM, test only	20	0.007	1.00
hetero	skab	recurrent	SEAM	20	0.007	1.00
hetero	skab	recurrent	SEAM, true regimes	20	0.007	1.00
hetero	skab	switch	SEAM, test only	20	0.032	0.95
hetero	skab	switch	SEAM	20	0.032	0.95
hetero	skab	switch	SEAM, true regimes	20	0.007	1.00

Table 7: **Larger systems** ($\beta = 0.5$, $N \in \{2400, 4800\}$, 2 schedules $\times$ 2 innovation laws $\times$ 10 series a cell). The focal reversal is embedded in a random sparse DAG shared by both regimes.

size	method	n	E	focal
$d = 10$	SEAM	80	0.002	1.00
$d = 10$	SEAM, test only	80	0.002	1.00
$d = 10$	FlowMSM + DYNOTEARS	80	0.361	0.59
$d = 10$	FlowMSM + VARLiNGAM	80	0.493	0.38
$d = 20$	SEAM	80	0.002	1.00
$d = 20$	SEAM, test only	80	0.002	1.00
$d = 20$	FlowMSM + DYNOTEARS	80	0.687	0.10
$d = 20$	FlowMSM + VARLiNGAM	80	0.727	0.06

Table 8: **Sharing, given the true regimes** ($d = 5$, $N = 1200$, 320 series with heteroscedastic innovations and 320 with unit scales). The sharing estimator of Section 3 compared with separate estimation in each regime. The columns report the focal error E and focal success; complete-effect and coefficient errors are in Table 25.

innovations	estimator	n	E	focal
heteroscedastic	sharing (Section 3)	320	0.094	0.84
heteroscedastic	no sharing	320	0.078	0.87
unit scale	sharing (Section 3)	320	0.025	0.97
unit scale	no sharing	320	0.029	0.97

D.2 False alarms per cell

Table 9: **One-regime series.** Series on which each method's own answer has more than one regime, out of 1400 with no change: VAR(1) with persistence up to 0.95 and VAR(2), two lengths, four innovation laws, 50 series a cell. SEAM reports a change only when its test rejects at $\alpha = 5\%$; CASTOR chooses the number itself; FlowMSM is given it. The baselines were run on the 600 VAR(1) series with $N = 1200$ (Appendix C).

method	series	reported a change	rate	95% CI
SEAM	1400	5	0.4%	[0.1, 0.8]
SEAM, test only	1400	72	5.1%	[4.0, 6.4]
CASTOR ($\tau = 0.1$)	600	365	60.8%	[56.8, 64.8]
FlowMSM + DYNOTEARS	600	600	100.0%	[99.4, 100.0]
FlowMSM + VARLiNGAM	600	600	100.0%	[99.4, 100.0]

By innovation law, the two-regime description outscored the one-regime fit ($T > 0$) on 0 of the 400 Gaussian series, 364 of 400 with river residuals, 40 of 300 with macroeconomic residuals and 19 of 300 with SKAB residuals; the test then rejected on 0, 24, 12 and 6 of them, and the mechanism condition left 0, 2, 3 and 0 reports of two mechanisms. The Gaussian column is what Theorem 1(i) predicts; the heavy-tailed, skewed residuals of the recorded series are where the test does the work.

Table 10: **One-regime series, per cell**: series in which a change of mechanism was reported, out of 50 a cell, under the decision rule of Section 3 and under the test-only ablation that drops its positive-gain and mechanism conditions. VAR(2) series use Gaussian and CausalRivers innovations only. A series reported as a change of innovation scale is not counted here; those counts are in Section 5.1.

null	ρ	N	innovations	method	n	changes	rate
var	0.50	1200	gaussian	SEAM, test only	50	3	6.0%
var	0.50	1200	gaussian	SEAM	50	0	0.0%
var	0.50	1200	macro	SEAM, test only	50	3	6.0%
var	0.50	1200	macro	SEAM	50	1	2.0%
var	0.50	1200	rivers	SEAM, test only	50	3	6.0%
var	0.50	1200	rivers	SEAM	50	0	0.0%
var	0.50	1200	skab	SEAM, test only	50	2	4.0%
var	0.50	1200	skab	SEAM	50	0	0.0%
var	0.50	4800	gaussian	SEAM, test only	50	2	4.0%
var	0.50	4800	gaussian	SEAM	50	0	0.0%
var	0.50	4800	macro	SEAM, test only	50	2	4.0%
var	0.50	4800	macro	SEAM	50	0	0.0%
var	0.50	4800	rivers	SEAM, test only	50	4	8.0%
var	0.50	4800	rivers	SEAM	50	1	2.0%
var	0.50	4800	skab	SEAM, test only	50	1	2.0%
var	0.50	4800	skab	SEAM	50	0	0.0%
var	0.80	1200	gaussian	SEAM, test only	50	1	2.0%
var	0.80	1200	gaussian	SEAM	50	0	0.0%
var	0.80	1200	macro	SEAM, test only	50	2	4.0%
var	0.80	1200	macro	SEAM	50	0	0.0%
var	0.80	1200	rivers	SEAM, test only	50	3	6.0%
var	0.80	1200	rivers	SEAM	50	0	0.0%
var	0.80	1200	skab	SEAM, test only	50	1	2.0%
var	0.80	1200	skab	SEAM	50	0	0.0%
var	0.80	4800	gaussian	SEAM, test only	50	2	4.0%
var	0.80	4800	gaussian	SEAM	50	0	0.0%
var	0.80	4800	macro	SEAM, test only	50	1	2.0%
var	0.80	4800	macro	SEAM	50	0	0.0%
var	0.80	4800	rivers	SEAM, test only	50	3	6.0%
var	0.80	4800	rivers	SEAM	50	0	0.0%
var	0.80	4800	skab	SEAM, test only	50	1	2.0%
var	0.80	4800	skab	SEAM	50	0	0.0%
var	0.95	1200	gaussian	SEAM, test only	50	5	10.0%
var	0.95	1200	gaussian	SEAM	50	0	0.0%
var	0.95	1200	macro	SEAM, test only	50	3	6.0%
var	0.95	1200	macro	SEAM	50	1	2.0%
var	0.95	1200	rivers	SEAM, test only	50	1	2.0%
var	0.95	1200	rivers	SEAM	50	0	0.0%
var	0.95	1200	skab	SEAM, test only	50	5	10.0%
var	0.95	1200	skab	SEAM	50	0	0.0%
var	0.95	4800	gaussian	SEAM, test only	50	4	8.0%
var	0.95	4800	gaussian	SEAM	50	0	0.0%
var	0.95	4800	macro	SEAM, test only	50	2	4.0%
var	0.95	4800	macro	SEAM	50	1	2.0%
var	0.95	4800	rivers	SEAM, test only	50	5	10.0%
var	0.95	4800	rivers	SEAM	50	1	2.0%
var	0.95	4800	skab	SEAM, test only	50	3	6.0%
var	0.95	4800	skab	SEAM	50	0	0.0%
var2	–	1200	gaussian	SEAM, test only	50	2	4.0%
var2	–	1200	gaussian	SEAM	50	0	0.0%
var2	–	1200	rivers	SEAM, test only	50	4	8.0%
var2	–	1200	rivers	SEAM	50	0	0.0%
var2	–	4800	gaussian	SEAM, test only	50	3	6.0%
var2	–	4800	gaussian	SEAM	50	0	0.0%
var2	–	4800	rivers	SEAM, test only	50	1	2.0%
var2	–	4800	rivers	SEAM	50	0	0.0%

D.3 Single-graph methods given the true segments (development check)

This check is *not* part of the confirmatory study. It was run during development, on development seeds, with one setting per method. It tests Proposition 2 on methods that model changing mechanisms but return one graph: GnIES (Gamella et al., 2022), UT-IGSP (Squires et al., 2020) and CD-NOD (Huang et al., 2020).

- **Design.** A static order-reversal design: $d = 5$, $N = 2400$, schedule $A|B|A|B$, $\beta = 0.5$. Four real-noise donors (Gaussian, CausalRivers, macroeconomic, SKAB) $\times$ 2 motifs $\times$ 5 replicates give 40 data sets.
- **Help given to the baselines.** Each baseline was given the true segments. Its graph was then fitted with regime-specific least-squares coefficients and variances.

Method (true segments)	mean E	median E	data sets with $E < 0.5$
GnIES	0.550	0.504	0 of 40
UT-IGSP	0.502	0.501	0 of 40
CD-NOD (C = true label)	0.568	0.502	0 of 40

The smallest error of any baseline on any data set was 0.500, as Proposition 2 requires for one graph. Better segmentation cannot fix this: the reversal is outside these methods' model class. SEAM's results on the confirmatory reversal designs are in Table 1.

D.4 Downstream use of the fitted regimes (earlier round)

These two tasks were run in the earlier round of this protocol, on that round's series and with the test-only report rule, and are reported here as they were reported there; like Appendix D.3, they are not part of the confirmatory study. They were prespecified in an addendum to that protocol and use the saved fits of that round's reversal block (E1: $d = 5$, $N \in \{1200, 4800\}$, $\beta \in \{0.3, 0.5\}$), refitting nothing.

- **Decide.** Four questions per series: in each regime, does an intervention on a move b, and in which direction; and the same for b on a. A question is answered "yes" if the matched fitted effect satisfies $|\hat{\theta}| \geq 0.1\beta$, and a "yes" counts as correct only with the right sign. A method without same-time effects answers exactly 2 of the 4 correctly. A single causal order answers at most 3, and one graph with common coefficients at most 2.
- **Recognize.** For each true regime, the generator draws an independent future sequence with fresh innovations. Its first m rows are assigned to the fitted regime with the larger conditional likelihood. A fit with $K = 1$ scores $1/2$.

Table 11: **Using the learned regimes** ($d = 5$; the series of Table 1, earlier round). "Decide": fraction of the four regime-specific intervention questions answered correctly. "Recog.": fraction of new sequences assigned to the right regime from their first 20 rows. A failed run scores 0 on "decide" and $1/2$ on "recog.".

	$N = 1200$				$N = 4800$			
	$\beta = 0.3$		$\beta = 0.5$		$\beta = 0.3$		$\beta = 0.5$	
Method	decide	recog.	decide	recog.	decide	recog.	decide	recog.
SEAM	0.94	0.89	1.00	0.98	1.00	0.93	1.00	0.98
SEAM, true regimes	0.98	0.92	1.00	0.98	1.00	0.93	1.00	0.98
CASTOR 0.4	0.50	0.49	0.53	0.52	0.50	0.50	0.53	0.52
CASTOR 0.1	0.52	0.52	0.59	0.57	0.51	0.51	0.57	0.58
Regime-PCMCI	–	–	0.46	0.57	–	–	–	–
SPACETIME	–	–	0.51	0.65	–	–	–	–

D.5 Other authors' benchmarks, heavy tails, and the reflection quotient

On CASTOR's own generator, with dense graphs and equal innovation variances, CASTOR recovers the graphs better (SHD 1.90 against 5.38, with its one run that did not terminate scored as the empty

graph; 1.56 on its 39 completed runs) and SEAM recovers the regimes at least as well (ARI 1.00 against 0.92). On Regime-PCMCI's tutorial system SEAM recovers the regimes without being told their number (ARI 0.95) against 0.86 for Regime-PCMCI given it, and on SPACETIME's tutorial both place the change within a few steps of the truth (0.9 and 2.2). Both tutorial generators fix their structural coefficients, so the fits can also be scored beyond their regimes (Table 13). Regime-PCMCI's system changes two equations: the effect $x_0 \to x_1$ reverses its sign and $x_0 \to x_2$ disappears. SEAM's report identifies exactly these two equations as changed on 19 of 20 series, and its coefficient errors in the unchanged and the changed equations are 0.043 and 0.048, against 0.056 and 0.069 for Regime-PCMCI's graphs refitted by least squares. A sign reversal keeps the parent set, so the parent-set rule registers only the second change; under it the two methods score 0.67 and 0.72. SPACETIME's system changes one coefficient and one innovation scale in the same equation; SEAM's report identifies exactly that equation on 16 of 20 series and is one law on 3, with coefficient errors 0.055 and 0.209.

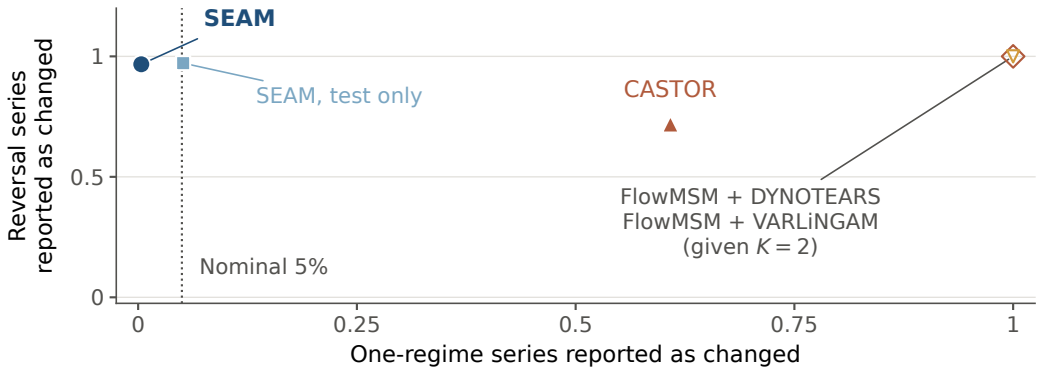


Figure 6: **Change reports on one-regime and reversal series.** Each point shows how often a method reports a change on one-regime series (horizontal) and order reversals (vertical). The vertical line marks 5%. FlowMSM is given $K = 2$ and returns two regimes on every series. The $N\Delta$ detection curves are in Figure 4.

On these data, a logistic fit of detection on $\log(N\Delta)$ has coefficient 4.59 under Gaussian innovations, and separate terms for the effect size add nothing to it ($p = 0.70$), which is the collapse that the leading term of Theorem 1 predicts and Figure 4 shows. On the reflection counterexample of Theorem 3, the quotient's rate of recovering the correct equivalence class rises with N, to 1.00 at the largest length, while the estimator without the quotient stays at 0.00; that failure does not shrink with N, because it is a difference in dimension and not in fit.

Table 12: **The baselines' own benchmarks.** CASTOR's generator (two regimes with independent dense graphs, unit weights, equal noise variances), the Regime-PCMCI tutorial system and the SPACETIME tutorial. ARI is of the regime labels, over completed runs; SHD is per regime, with a run that did not terminate scored as the empty graph as in Appendix F (CASTOR: 39 of 40 completed, on which its SHD is 1.56); "changepoint error" is the distance in steps from the true change.

benchmark	method	n	ARI	SHD	changepoint error
CASTOR's generator	SEAM	40	1.00	5.38	–
CASTOR's generator	SEAM, test only	40	1.00	5.38	–
CASTOR's generator	CASTOR (authors)	40	0.92	1.90	–
Regime-PCMCI tutorial	SEAM	20	0.95	–	–
Regime-PCMCI tutorial	Regime-PCMCI	20	0.86	–	–
SPACETIME tutorial	SEAM	20	–	–	0.9
SPACETIME tutorial	SPACETIME	20	–	–	2.2

Table 13: **Coefficients of the tutorial systems** (20 series each). Both generators are purely lagged with independent innovations, so their lag matrices are the structural coefficients. Localization F1 under the parent-set and coefficient rules; lag-graph F1; same-time SHD against the true same-time graph, which is empty; mean per-equation error of the direct coefficients of the unchanged and of the changed equations; RMSE of the effect matrix. The focal error E and the relative effect error are undefined for these systems, which have no same-time effects. Regime-PCMCI's coefficients are refitted by least squares in each regime and so differ in every equation, which the coefficient rule reads as every equation changed. SPACETIME's output on its tutorial is one graph and the change points, so its graph is scored.

		localization F1		lag	same-time	coefficient error		effect
method	n	parent sets	coefficients	F1	SHD	unchanged	changed	RMSE
Regime-PCMCI tutorial								
SEAM	20	0.67	0.99	0.99	0.05	0.043	0.048	0.003
Regime-PCMCI	20	0.72	0.80	0.98	0.00	0.056	0.069	0.000
SPACETIME tutorial								
SEAM	20	0.85	0.83	0.95	0.03	0.055	0.209	0.006
SPACETIME	20	–	–	1.00	0.00	–	–	–

Table 14: **Heavy tails** (t_3 innovations, $d = 5$, $N = 1200$): the Gaussian statistic, the Student-t statistic of Appendix D.7, the test-only ablation, and CASTOR.

method	n	E	focal	worst cell
SEAM	40	0.266	0.57	0.343
SEAM, Student-t	40	0.006	1.00	0.006
SEAM, test only	40	0.266	0.57	0.343
CASTOR ($\tau = 0.1$)	40	0.601	0.05	0.675

Table 15: **The reflection counterexample of Theorem 3**, true regimes: how often each estimator puts both regime graphs in the correct Markov equivalence class (a joint criterion; Figure 3 and Table 26 count individual regime graphs), and the structural Hamming distance per regime. At the largest sample size both SEAM and separate-regime estimation place both graphs in the correct class on every series; sharing without sign alignment does so on at most one series in twenty at every sample size.

N	estimator	n	correct equivalence class	SHD
600	SEAM, true regimes	20	0.85	0.50
600	no sharing	20	0.75	0.72
600	sharing without the quotient	20	0.05	2.60
2400	SEAM, true regimes	20	0.95	0.15
2400	no sharing	20	0.80	0.53
2400	sharing without the quotient	20	0.05	2.30
9600	SEAM, true regimes	20	0.95	0.25
9600	no sharing	20	0.90	0.17
9600	sharing without the quotient	20	0.05	2.10
38400	SEAM, true regimes	20	1.00	0.00
38400	no sharing	20	1.00	0.00
38400	sharing without the quotient	20	0.00	2.17

D.6 Run times

Table 16 gives the running times of every method in the two reversal blocks. On a one-regime series, the full decision rule is much cheaper than the test-only ablation: a median of 7 seconds against 151. The reason is the positive-gain condition: when the two-regime description does not beat the

one-regime description, the decision is settled before the permutation loop starts, and SEAM skips the 39 surrogate fits it would otherwise run.

Table 16: Wall-clock minutes per series on one core in the two reversal blocks (confirmatory run; SEAM includes the calibration). "Did not terminate" counts runs stopped by the method's wall budget (Appendix C). FlowMSM's rows are its original job; the series it fitted after the run (Appendix C) ran on another machine type.

block	method	n	median minutes	slowest	did not terminate	failed
$d = 5$	SPACETIME	40	66.2	89.9	0	0
$d = 5$	FANTOM	80	49.9	52.7	0	0
$d = 5$	FANTOM (window 600)	80	27.2	29.1	0	0
$d = 5$	Regime-PCMCI	40	9.6	14.1	0	3
$d = 5$	SEAM, test only	640	2.6	5.6	0	0
$d = 5$	SEAM	640	2.5	6.6	0	0
$d = 5$	FlowMSM + DYNOTEARS	383	1.2	1.8	0	0
$d = 5$	FlowMSM + VARLiNGAM	383	1.1	1.7	0	0
$d = 5$	CASTOR ($\tau = 0.1$)	640	0.5	1.2	56	0
$d = 5$	CASTOR (authors)	640	0.4	1.2	82	0
$d = 10, 20$	SEAM, test only	160	12.4	64.4	0	0
$d = 10, 20$	SEAM	160	11.2	59.9	0	0
$d = 10, 20$	FlowMSM + DYNOTEARS	160	6.1	18.3	0	0
$d = 10, 20$	FlowMSM + VARLiNGAM	160	5.7	17.9	0	0
$d = 10, 20$	CASTOR (window 800)	80	1.6	5.4	0	0
$d = 10, 20$	CASTOR (window 400)	80	1.3	4.4	2	0

D.7 A STUDENT-t STATISTIC

Theorem 11 requires only that the statistic be a fixed function of the data, so the row-permutation test keeps its finite-sample level when the Gaussian score in T is replaced by a Student-t likelihood; with the fixed-design surrogates of Algorithm 1, both statistics are judged by their measured rejection rates. We fit it by structural EM: each E-step reweights rows, and each M-step solves the same integer program on weighted moments (Friedman, 1998; Lange et al., 1989). The tail index is chosen by profile likelihood on a grid, and a large index reduces the statistic to the Gaussian one. This variant is not part of SEAM: it adds a second likelihood and costs 5–10 times more time. Table 14 compares the two statistics on the t_3 designs. The t statistic restores detection under t_3 innovations (error 0.006 against 0.266), and under the decision rule of Section 3 it reports a change on 1 of 100 one-regime t_3 series. In the earlier round, with the test alone, the same statistic was the one that lost the level; the positive-gain and mechanism conditions are what make it usable.

D.8 A VARIANCE-ONLY CHANGE BETWEEN REGIMES

The scale-change report of Section 3 is exercised here on a designed variance-only change, in a study specified in writing before any fit and reported with every record in the supplement. Five variables, one same-time edge $a \to b$ with coefficient 1, self-lag 0.4 in every equation, Gaussian innovations, $N = 1200$ rows; between the two regimes only the innovation variance of the root a changes, from 1 to 4, so the graph, the coefficients and the causal order are the same in both. Two schedules, one switch and a recurring one, 600 independent series in all, each fitted by SEAM unchanged (Table 17). Every series completed; the lag order was 1 in all 600. The change is detected on every series ($p = 0.025$ throughout) and attributed to the noise level on 578 of 600 (96.3%), with all same-time and lag coefficients shared across regimes; a mechanism change is declared on 22 (3.7%, exact interval 2.3–5.5%; 4.3% and 3.0% by schedule); the change is never missed. Label error is at most 2.1% in the 22 series with a false mechanism report. In 21 of these series the selected description assigns regime-specific lag coefficients to one equation (to two equations in 1 of them); in 16 the equation is that of the root a. The estimated between-regime lag-coefficient differences range from 0.04 to 0.28, although the true coefficient is common. The remaining series assigns regime-specific coefficients of opposite sign to one same-time edge and one lag coefficient. The attribution to scale or mechanism is the description-length choice between graded and regime-specific equations (Section 3).

Table 17: **A variance-only change between regimes.** Five variables, one same-time edge $a \to b$ (coefficient 1), self-lag 0.4, Gaussian innovations, $N = 1200$; only the innovation variance of the root a changes, from 1 to 4, between the two regimes. Counts of SEAM's three reports per schedule (every planned series completed) with exact 95% intervals for the mechanism-change rate. *No change*: no distributional change reported (the one-law report); *scale change*: a change reported with common same-time and lag coefficients; *mechanism change*: at least one equation assigned regime-specific coefficients.

schedule	n	no change	scale change	mechanism change	95% interval
one switch ($A_{600}B_{600}$)	300	0	287	13 (4.3%)	2.3–7.3%
recurring ($A_{300}B_{300}A_{300}B_{300}$)	300	0	291	9 (3.0%)	1.4–5.6%
both	600	0	578	22 (3.7%)	2.3–5.5%

The same choice attributes the change to the right equation. Scored equation by equation against the generator (an evaluation added after the fits; its rules are in the supplement), the root's equation is graded in 584 of 600 series (97.3%, exact interval 95.7–98.5%; 97.7% and 97.0% by schedule) and regime-specific in the other 16; 24 of the 2,400 unchanged equations are graded (precision 96.1%), and all five equations are classified correctly in 554 series (Table 18).

Table 18: **The graded class, equation by equation** (variance-only change, 600 series, five equations each; the root's equation is truly graded and the other four are unchanged). Classes are read in the data's coordinates: an equation pooled only after sign alignment counts as regime-specific ("specific"); read modulo the sign gauge, the root's equation is graded in 586 of 600 series. Recall has an exact 95% interval over series; precision is the fraction of equations classified graded that are the root's.

		root equation (truly graded)			unchanged equations		graded	all five
schedule	series	graded	specific	recall	graded	specific	precision	correct
one switch	300	293	7	97.7% [95.3, 99.1]	13	8	95.8%	274
recurring	300	291	9	97.0% [94.4, 98.6]	11	0	96.4%	280
both	600	584	16	97.3% [95.7, 98.5]	24	8	96.1%	554

D.9 A REAL SYSTEM: THE COLUMBIA–WILLAMETTE BACKWATER

Data. Daily means of the 15-minute gage height at USGS 14211720 (Willamette at Portland, PDX) and 14144700 (Columbia at Vancouver, VAN); USGS daily discharge at 14191000 (Willamette at Salem, SAL) and 14105700 (Columbia at The Dalles, TDA), in logs; the daily mean of the NOAA hourly water level at Astoria (9439040, AST). Gaps of at most three days are interpolated; a window must have no longer gap and at least 95% coverage in every water year. The windows are the admissible four-water-year spans tiled without overlap (WY2012–2015, WY2018–2021, WY2022–2025; WY2016–2017 are excluded because the Portland stage record has gaps of 4 and 9 days in WY2017) and, in a second pass specified after the first was seen, the 6 non-overlapping two-year spans. The labels used to read the fitted regimes are never given to SEAM: a flow-ratio label (the standardized log discharge at The Dalles minus that at Salem, above $+0.5$ freshet, below -0.5 winter, else unlabeled) and the calendar (15 May–31 August freshet; November–March winter). The check was specified in writing before any fit, and every window is reported.

Result. Figure 7 shows the 3 four-year windows. In 9 of 9 windows SEAM reports two mechanisms ($p = 0.025$) and the fitted regimes are the seasons (balanced accuracy 0.77–0.92 against the flow-ratio label, 0.92–1.00 against the calendar). The freshet law has the Columbia's flow at The Dalles driving a stage in 6 of 9 windows (coefficient 1.28–1.78), and in none of the three windows that cover WY2022–2025, whose June–July discharges are the smallest of the record; the winter law has the Astoria water level driving the Portland stage in 8 of 9 (0.99–1.33), the exception being the four-year window of those same years. The Portland–Vancouver edge, between two stages that move almost one-to-one at daily resolution (coefficient 0.95–1.05 in every law), reverses between the seasons, consistent with the documented account, in 4 of the 6 anchored windows; in the others the freshet law routes the Columbia's flow to Portland and keeps Portland $\to$ Vancouver, so that the

flow reaches the Portland stage first rather than the Vancouver stage. The edge reverses in 0 of the 3 unanchored windows. Without the Astoria water-level series, in a sensitivity run specified afterwards, the regimes are still the seasons (0.76–0.90), but the edge reverses as documented in 0 of the 3 four-year windows, and in WY2022–2025 it reverses the other way: the documented reversal found in 2 of these windows depends on the Astoria series, the winter law's driver of the Portland stage. The baselines of Section 5, run on the same 9 windows in their protocol configurations (36 runs), report the reversal in 0: CASTOR returns one regime in 9 windows of 9 at its protocol threshold and 8 at the lower one; Regime-PCMCI's regimes match the seasons in 2 of 8 completed runs and its graphs are lagged only; SPACETIME's regimes match the seasons in 4 of 9 and its single DAG keeps one orientation of the Portland–Vancouver edge in every regime. The search seed does not matter: seeds 1–4 reproduce every fitted law and decision of seed 0 in 36 of 36 runs.

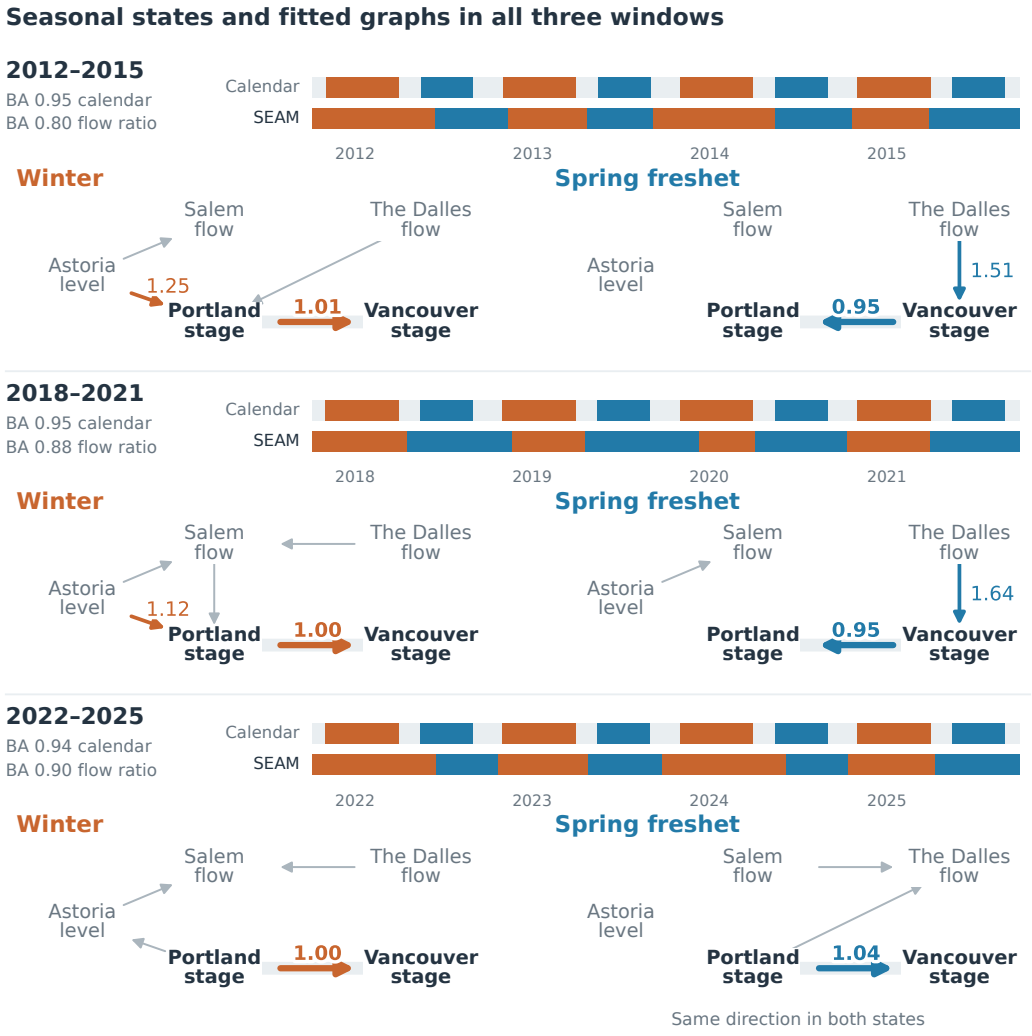


Figure 7: **The Columbia–Willamette backwater, all 3 four-year windows.** Each row shows paired Calendar and SEAM bands and the two fitted same-day graphs. Blue denotes the spring freshet, orange winter, and gray calendar transition periods; fitted regimes are matched to the seasons by the flow-ratio label. The graphs show the Portland and Vancouver stages, the Salem and The Dalles discharges, and the daily mean water level at Astoria. The Portland–Vancouver edge is highlighted; nonfocal edges are shown for absolute coefficients of at least 0.05, with only key coefficients labeled.

E Related work in more detail

Section 1 places prior work by the limitation SEAM addresses; this appendix adds the detail, and Table 19 summarizes what each method can be asked.

Table 19: **What each method can be asked:** whether it models instantaneous edges, lets an edge's direction differ between regimes, selects the number of regimes rather than requiring it, and tests whether any change occurred (scope of SEAM's guarantee: Section 4). †: the evaluated FANTOM implementation sets its count from the series length and window size (Appendix C).

method	instantaneous edges	direction may change	selects number of regimes	tests for no change
Regime-PCMCI (Saggioro et al., 2020)	—	—	—	—
SPACETIME (Mameche et al., 2025)	✓	—	✓	—
CASTOR (Rahmani & Frossard, 2025b)	✓	✓	✓	—
FANTOM (Rahmani & Frossard, 2025a)	✓	✓	†	—
FlowMSM (Hulsman et al., 2026)	✓	✓	—	—
SEAM (this paper)	✓	✓	✓	✓

One graph per series. PCMCI and PCMCI+ (Runge et al., 2019; Runge, 2020), non-Gaussian structural VARs (Hyvärinen et al., 2010), DYNOTEARS (Pamfil et al., 2020) and Rhino (Gong et al., 2023) learn one lagged and same-time graph. By Proposition 2, each has effect error at least $1/2$ on a reversal, even when given the true regimes.

Changing mechanisms, fixed orientation. CD-NOD (Huang et al., 2020), joint causal inference (Mooij et al., 2020), invariance methods (Peters et al., 2016; Pfister et al., 2019) and methods for unknown interventions (Squires et al., 2020; Gamella et al., 2022) model which mechanisms change across contexts, but under one orientation, so Proposition 2 applies to them too (Appendix D.3 checks it on three). Mixtures of DAGs (Saeed et al., 2020) recover a union graph or a common order, and MCD (Varambally et al., 2024) assigns each whole series to one causal model.

Recurring regimes and structural breaks without a causal graph. Markov-switching VARs (Hamilton, 1989; Krolzig, 1997) and switching linear dynamical systems (Fox et al., 2008; Linderman et al., 2017) learn recurring regimes with unknown labels, but each regime is a reduced form: they fit the conditional mean and covariance of the rows and orient nothing, so an intervention query has no answer in them. Identification through heteroskedasticity (Rigobon, 2003; Lanne et al., 2010) recovers a fixed structural matrix from the change of covariance between regimes, under the assumption that the structure does not change, which is the opposite of SEAM's question. Structural-break methods for VARs (Bai & Perron, 1998; Safikhani & Shojaie, 2022) and piecewise-constant graphical models (Kolar & Xing, 2012; Gibberd & Nelson, 2017) locate changes of the reduced form or of the undirected precision matrix, without regimes that recur and without directions. Across contexts with known labels, the sparse-mechanism-shift score (Perry et al., 2022) and difference-DAG estimation (Wang et al., 2018) ask which mechanisms changed, with a common order in the first and a difference of undirected structure in the second; neither segments a series or tests for the existence of a change.

Regime-dependent graphs in one series. Regime-PCMCI (Saggioro et al., 2020) alternates regime assignment with PCMCI; CASTOR (Rahmani & Frossard, 2025b) learns a DAG per regime, and their number, by EM; SPACETIME (Mameche et al., 2025) combines changepoints with minimum-description-length structure learning. Recent methods also let an edge change direction with the regime: FANTOM (Rahmani & Frossard, 2025a) under non-Gaussian or heteroscedastic noise, FlowMSM (Hulsman et al., 2026) with identifiability results for Markov switching models, and RCBNB-MB (Zan et al., 2026). Rabel & Runge (2025) study context-specific graphs with unobserved contexts. What these methods share is that none of them tests whether a change occurred: each fixes the number of regimes by an input, a window, an information criterion or an evidence bound, and attaches no error guarantee to it (Table 19). We run five of them as baselines.

Tests for change and exact search. Permutation tests for changepoints are classical (Antoch & Hušková, 2001; Kirch, 2007), and Hansen (2000) introduced the fixed-regressor bootstrap; SEAM applies the permutation principle to its complete structure-learning score, search included (Section 3). Its structure step builds on integer programming for Bayesian networks (Jaakkola et al., 2010; Cussens, 2011; Bartlett & Cussens, 2017). Its penalty builds on the unit-information view of BIC (Kass & Wasserman, 1995) and on multiplicity-adjusted codes (Scott & Berger, 2010).

F LOCALIZATION, GRAPHS AND EFFECT MATRICES FROM THE FROZEN FITS

The localization F1, SHD and effect error of Section 5 were computed after the confirmatory run, from the laws (B_k, A_k, s_k^2) and training labels stored in the records, against the truth regenerated from each series' seed (`experiments/posthoc/rescore_metrics.py`, `paper/make_rescore_tables.py`). The prespecified endpoints of Appendix C are unchanged. The focal error recomputed from the stored laws, which are rounded to six decimals, differs from the stored E by at most $1.02 \cdot 10^{-5}$ on each of the 9,057 completed runs with a focal pair.

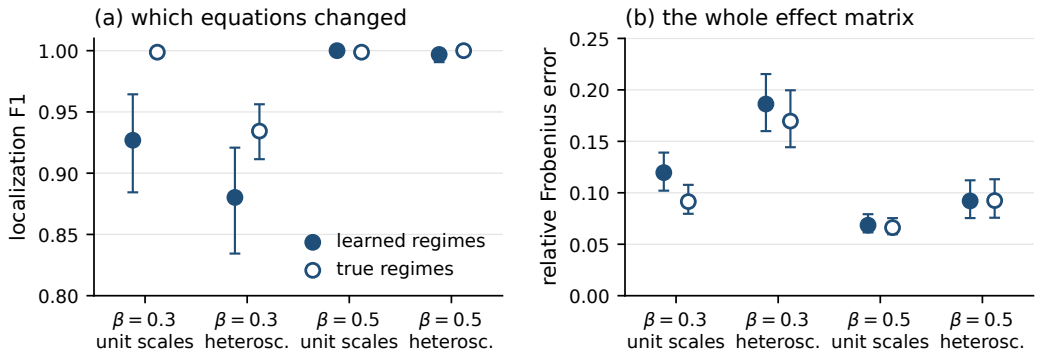


Figure 8: **Where estimating the regimes costs accuracy** ($d = 5$, 160 series per group). SEAM with learned regimes (filled) against SEAM given the true regimes (open), with 95% bootstrap intervals over series. **(a)** Localization F1. **(b)** Relative error of the complete effect matrix.

Matching and failed runs. Regimes are matched by the rule of the primary endpoint: each true regime is compared with the fitted regime that holds most of its training rows, and a one-regime output (no change, or a scale change) is compared with both true regimes and declares no equation changed. A failed run (148 records: CASTOR runs that did not terminate, Regime-PCMCI errors) scores as the empty output: no equation declared changed, the empty graph, $\hat{\Theta} = I$, and $E = 1$ as in Appendix C. On its completed runs alone, CASTOR at $\tau = 0.1$ (295 of 320 in the block of Table 1) scores $E = 0.666$, localization F1 0.27, SHD 2.18 and effect error 0.871; at the authors' settings (270 completed) $E = 0.942$, F1 0.19, SHD 3.06, effect error 1.001. FANTOM completed every run.

Changed equations. An *equation* is variable j's column of B and A. Under the *parent-set rule*, used for every cross-method comparison, the equation is declared changed when its same-time or lagged parent set differs between the two matched laws; under the *coefficient rule*, SEAM's own definition, when any coefficient differs (equal coefficients with different innovation variance is graded and does not count as a change); the two rules give SEAM F1 0.951 and 0.947 on the $d = 5$ block. A separate gauge-aligned diagnostic asks which equations can be shared after sign alignment. It evaluates pooling rather than the original-coordinate change report: a physical coefficient sign change can be correctly reported while the aligned equation is pooled. The reversal blocks have the same truly changed equations under both conventions; the reflection design distinguishes them. SPACETIME also stores its own per-variable change output; under it SPACETIME, which keeps one graph across regimes and therefore never changes a parent set, scores F1 0.43 against 0.00 under the parent-set rule. Regime-PCMCI returns one graph per regime, so its own change output is the parent-set rule itself (F1 0.21).

Graphs and effects. SHD counts a reversed edge once; directed-edge precision and recall count it as one false positive and one false negative; skeleton metrics ignore orientation, and "in MEC" is the fraction of fitted graphs in the true graph's Markov equivalence class, the limit of what one Gaussian regime identifies on its own. $\Theta_r = (I - B_r)^{-1}$ is in the model's row convention; its true off-diagonal block is nonzero on every series, so the relative error is always defined, and no fitted $I - B$ was singular.

Intervals. Intervals are 95% bootstrap intervals over series; series of different cells of one block share an innovation seed (the same draws under another β or scale setting), so pooled intervals resample seed clusters, and paired comparisons use the series both arms ran.

Reproducible and complete records. Every stored SEAM fit is reproducible from its seed: re-executing the frozen estimator on the regenerated series, with the stored p-value standing in for the permutation test where a fit ran one, reproduces all 7,980 of them (the fits of every arm except the no-sharing arm, which shares no equation by construction and was not re-executed), every law to $2 \cdot 10^{-6}$, every labeling, and every stored statistic T and decision. The quantities the records do not keep are taken from these re-executions: the gauge with the scores of the candidate gauges, the class of every equation, and the lag coefficients beyond lag one. FlowMSM's records of its first 300 one-regime series kept the number of regimes but not the fitted graphs. Refitted from their seeds, these fits reproduce every stored number of regimes, switch count and regime-detection loss; with the series fitted under the complete design (Appendix C), which store their laws, all 600 one-regime series of each backend are scored (Table 27). Three evaluations use these records beyond the prespecified endpoints: lag graphs at every lag (below), the graded class on the variance-only series (Table 18), and the coefficients of the Regime-PCMCI and SPACETIME tutorial systems (Table 13).

Lags beyond the first. SEAM selected a lag order above one on every VAR(2) one-regime series and on 15 of the 40 series of CASTOR's generator, and nowhere else. Scored at every lag, with the true edges of a lag the fit lacks counted as missed, it finds every true lag-2 edge of the VAR(2) series, with 0.12 spurious lag-2 edges per regime (all-lag F1 0.990 against 0.992 at lag one); on CASTOR's generator, which has no lag-2 effects, its all-lag F1 is 0.752 against 0.770 at lag one (Table 20). Every other fit in Table 20 has one lag, where the two scores coincide.

Table 20: **Lag graphs at every lag.** $p \geq 2$: fits with a lag order above one. F1 at lag one, at lag two and over all lags, and SHD over all lags, per series and averaged, with 95% bootstrap intervals over seed clusters for the all-lag scores; lag-2 false positives and false negatives per regime. Fitted and true coefficients are aligned by variable, direction and lag, self-lags included; the true edges of a lag a fit does not have count as false negatives, and a failed run as the empty output.

			F1			SHD	lag 2, per regime	
method	n	$p \geq 2$	lag 1	lag 2	all lags	all lags	FP	FN
VAR(2) one-regime series (C3)								
SEAM	200	200	0.992	0.989	0.990 [0.987, 0.993]	0.245 [0.175, 0.315]	0.12	0.00
SEAM, test only	200	200	0.992	0.989	0.990 [0.987, 0.993]	0.240 [0.170, 0.315]	0.12	0.00
VAR(1) one-regime series (C3)								
SEAM	1200	0	0.994	1.000	0.994 [0.993, 0.995]	0.092 [0.074, 0.109]	0.00	0.00
SEAM, test only	1200	0	0.994	1.000	0.994 [0.993, 0.995]	0.092 [0.073, 0.111]	0.00	0.00
CASTOR ($\tau = 0.1$)	600	0	0.927	1.000	0.927 [0.915, 0.938]	0.843 [0.740, 0.952]	0.00	0.00
FlowMSM + DYNOTEARS	600	0	0.872	1.000	0.872 [0.865, 0.879]	2.207 [2.067, 2.347]	0.00	0.00
FlowMSM + VARLiNGAM	600	0	0.912	1.000	0.912 [0.903, 0.922]	1.620 [1.422, 1.815]	0.00	0.00
CASTOR's generator (C5)								
SEAM	40	15	0.770	0.812	0.752 [0.712, 0.793]	3.587 [2.900, 4.338]	0.46	0.00
SEAM, test only	40	15	0.770	0.812	0.752 [0.711, 0.794]	3.587 [2.875, 4.325]	0.46	0.00
CASTOR (authors)	40	0	0.887	1.000	0.887 [0.818, 0.941]	1.712 [0.750, 3.225]	0.00	0.00

Table 21: **Paired advantage of SEAM on shared series** ($d = 5$, $\beta = 0.5$): SEAM's localization F1 minus the baseline's, and the baseline's SHD and effect error minus SEAM's, with 95% bootstrap intervals over the series both ran.

baseline	shared series	Δ localization F1	Δ SHD	Δ effect error
CASTOR ($\tau = 0.1$)	320	+0.751 [+0.712, +0.787]	+2.266 [+2.148, +2.380]	+0.801 [+0.768, +0.833]
CASTOR (authors)	320	+0.837 [+0.803, +0.869]	+3.155 [+3.072, +3.234]	+0.921 [+0.896, +0.945]
FlowMSM + DYNOTEARS	320	+0.593 [+0.557, +0.632]	+3.514 [+3.364, +3.658]	+0.729 [+0.692, +0.766]
FlowMSM + VARLiNGAM	320	+0.516 [+0.465, +0.561]	+2.258 [+2.031, +2.494]	+0.507 [+0.458, +0.560]
FANTOM	80	+0.417 [+0.410, +0.423]	+3.344 [+3.106, +3.569]	+0.760 [+0.711, +0.812]
FANTOM (window 600)	80	+0.695 [+0.631, +0.754]	+4.500 [+4.150, +4.838]	+0.807 [+0.758, +0.856]

Table 22: **Larger systems** ($d \in \{10, 20\}$, $N \in \{2400, 4800\}$, $\beta = 0.5$): the metrics of Table 1 with the directed-edge F1. CASTOR's window is $N/6$. The largest true in-degree exceeds the screening cap of three used by the statistic on 8% ($d = 10$) and 55% ($d = 20$) of the series; the reported graphs are re-solved without the cap (Appendix B).

d	method	series	focal error E	localization F1	SHD	edge F1	effect error
10	SEAM	80	0.002	1.00	0.06	1.00	0.039
10	SEAM, test only	80	0.002	1.00	0.06	1.00	0.039
10	CASTOR (windowed, $N = 2400$)	40	0.399	0.21	1.86	0.86	0.462
10	CASTOR (windowed, $N = 4800$)	40	0.464	0.13	1.57	0.89	0.415
10	FlowMSM + DYNOTEARS	80	0.361	0.37	3.74	0.80	0.361
10	FlowMSM + VARLiNGAM	80	0.493	0.40	7.67	0.62	0.573
20	SEAM	80	0.002	1.00	0.32	0.99	0.046
20	SEAM, test only	80	0.002	1.00	0.32	0.99	0.046
20	CASTOR (windowed, $N = 2400$)	40	0.515	0.15	3.15	0.88	0.375
20	CASTOR (windowed, $N = 4800$)	40	0.436	0.13	1.68	0.94	0.313
20	FlowMSM + DYNOTEARS	80	0.687	0.07	8.43	0.79	0.408
20	FlowMSM + VARLiNGAM	80	0.727	0.09	15.61	0.65	0.563

Table 23: **Weak effects** ($d = 5$, $\beta = 0.3$): the metrics of Table 1.

method	series	focal error E	localization F1	SHD	effect error
SEAM	320	0.133	0.90	0.29	0.153
SEAM, test only	320	0.132	0.91	0.29	0.152
SEAM, true regimes	320	0.104	0.97	0.19	0.131
CASTOR ($\tau = 0.1$)	320	1.025	0.15	2.46	0.897
CASTOR (authors)	320	1.281	0.06	3.18	1.006
FlowMSM + DYNOTEARS	320	0.936	0.26	3.62	0.868
FlowMSM + VARLiNGAM	320	0.677	0.31	2.21	0.571

Table 24: **The C1b subset** ($d = 5$, $N = 1200$, $\beta = 0.5$, unit scales, Gaussian and river innovations, both schedules, 10 series each): Regime-PCMCI and SPACETIME ran on these 40 series, and SEAM's runs on the same series are shown with them.

method	series	focal error E	localization F1	SHD	effect error
SEAM	40	0.006	1.00	0.01	0.062
SEAM, true regimes	40	0.006	1.00	0.03	0.062
Regime-PCMCI	40	1.000	0.21	4.00	1.000
SPACETIME	40	0.684	0.00	3.50	0.871

Table 25: **Sharing, with the true regimes** ($d = 5$, 640 series, both β and both scale settings). Effect error and RMSE of the complete effect matrix; root-mean-square coefficient error of the truly shared and the truly changed equations; localization F1 under the parent-set rule; the prespecified focal error. Below, paired differences of the sharing estimator against no sharing (negative favors sharing), overall and by stratum.

estimator	series	effect error	effect RMSE	coef. error, shared eq.	coef. error, changed eq.	localization F1	focal E
no sharing	640	0.135	0.029	0.064	0.066	0.92	0.053
sharing (Section 3)	640	0.105	0.022	0.038	0.065	0.98	0.060

series	Δ effect error	Δ coef. error, shared	Δ coef. error, changed	Δ focal E
all 640 series	-0.030 [-0.045, -0.018]	-0.026 [-0.032, -0.020]	-0.001 [-0.004, +0.001]	+0.006 [-0.001, +0.014]
unit innovation scales	-0.038 [-0.054, -0.023]	-0.026 [-0.033, -0.020]	-0.003 [-0.008, -0.000]	-0.003 [-0.013, +0.005]
heteroscedastic innovations	-0.023 [-0.040, -0.006]	-0.026 [-0.034, -0.019]	+0.001 [-0.003, +0.005]	+0.016 [+0.004, +0.027]
$\beta = 0.3$	-0.030 [-0.048, -0.013]	-0.027 [-0.033, -0.021]	+0.001 [-0.003, +0.004]	+0.017 [+0.004, +0.030]
$\beta = 0.5$	-0.031 [-0.044, -0.019]	-0.025 [-0.032, -0.020]	-0.003 [-0.006, -0.000]	-0.005 [-0.011, +0.000]

Table 26: **The reflection design of Theorem 3, with the true regimes** (80 series). Regime B is regime A with selected variable signs flipped. The gauge-aligned diagnostic counts equations whose fitted coefficients differ after sign alignment; all equations in this design can be shared under that alignment. The last column evaluates coefficient changes in the original coordinates, where sign changes count as genuine changes.

estimator	series	graphs in the true MEC	SHD	effect error	equations differing after alignment	F1, observed coordinates
SEAM, true regimes	80	94%	0.23	0.062	0.00	0.97
sharing without the quotient	80	4%	2.29	0.548	1.14	0.67
no sharing	80	90%	0.36	0.098	2.15	0.97

Table 27: **One-regime series**: the decision as reported, before any matching (change reported; scale change reported; FlowMSM is given two regimes), with SEAM on all series and on the VAR(1) series with $N = 1200$ that the baselines ran; the number of series on which the stored fit has two or more laws ($K \geq 2$; a scale-change report stores one pooled law); on those series, the mean number of equations whose parent sets differ between any two fitted laws (support-changed) and whose coefficients differ (coefficient-changed), counted across the fitted laws themselves before any truth matching; and the graph and effect metrics of the law that holds most rows.

block	method	series	change reported	scale change	$K \geq 2$	support-changed eq. ($K \geq 2$)	coef.-changed eq. ($K \geq 2$)	SHD	effect error
C3	SEAM	1400	0.36%	2.6%	5	0.60	1.00	0.01	0.033
C3	SEAM, test only	1400	5.14%	0.0%	72	0.32	0.35	0.01	0.033
C3	SEAM, the baselines' series	600	0.33%	2.5%	2	0.50	1.00	0.02	0.045
C3	CASTOR ($\tau = 0.1$)	600	60.83%	0.0%	365	2.79	4.98	0.47	0.306
C3	FlowMSM + DYNOTEARS	600	100.00%	0.0%	600	3.76	5.00	0.91	0.143
C3	FlowMSM + VARLiNGAM	600	100.00%	0.0%	600	1.93	5.00	0.96	0.279
C9_null_t3	SEAM	100	2.00%	5.0%	2	1.00	1.00	0.03	0.045
C9_null_t3	SEAM, Student-t	100	1.00%	10.0%	1	1.00	1.00	0.04	0.045
C9_null_t3	SEAM, test only	100	7.00%	0.0%	7	0.29	0.29	0.03	0.045

Table 28: **Graph and equation metrics by block and method.** Localization precision, recall and F1 under the parent-set rule, computed per series and then averaged, for every method alike, and F1 under the coefficient rule; directed-edge precision and recall; skeleton F1; fraction of fitted graphs in the true Markov equivalence class; lag-graph F1; RMSE of the off-diagonal effect matrix. Failed runs are scored as the empty output.

block	method	n	failed %	P support	R support	F1 support	F1 coef.	edge P	edge R	skeleton F1	in MEC %	lag F1	effect RMSE
C1	SEAM	640	0	0.96	0.95	0.95	0.95	0.97	0.97	0.99	89	0.98	0.025
C1	SEAM, test only	640	0	0.96	0.95	0.95	0.95	0.97	0.97	0.99	89	0.98	0.024
C1	SEAM, true regimes	640	0	1.00	0.98	0.98	0.98	0.98	0.98	0.99	91	0.98	0.022
C1	sharing (Section 3)	640	0	1.00	0.98	0.98	0.98	0.98	0.98	0.99	91	0.98	0.022
C1	no sharing	640	0	0.87	0.99	0.92	0.57	0.97	0.97	0.99	91	0.98	0.029
C1	CASTOR (authors)	640	13	0.10	0.14	0.11	0.11	0.79	0.21	0.40	0	0.17	0.219
C1	CASTOR ($\tau = 0.1$)	640	9	0.15	0.29	0.20	0.18	0.67	0.50	0.73	16	0.74	0.194
C1	FlowMSM + DYNOTEARS	640	0	0.25	0.52	0.33	0.31	0.48	0.71	0.76	2	0.75	0.182
C1	FlowMSM + VARLiNGAM	640	0	0.34	0.51	0.40	0.31	0.63	0.74	0.84	17	0.88	0.127
C1	FANTOM	80	0	0.41	1.00	0.58	0.57	0.41	0.52	0.82	2	0.57	0.198
C1	FANTOM (window 600)	80	0	0.21	0.53	0.30	0.30	0.32	0.49	0.75	0	0.57	0.210
C1b	Regime-PCMCI	40	8	0.19	0.26	0.21	0.17	1.00	0.00	0.00	0	0.71	0.243
C1b	SPACETIME	40	0	0.00	0.00	0.00	0.49	0.41	0.53	0.79	0	0.84	0.213
C2	SEAM	160	0	1.00	1.00	1.00	1.00	0.99	1.00	1.00	89	0.97	0.007
C2	SEAM, test only	160	0	1.00	1.00	1.00	1.00	0.99	1.00	1.00	89	0.97	0.007
C2	CASTOR (windowed, $N = 2400$)	80	2	0.13	0.35	0.18	0.10	0.91	0.86	0.93	38	0.91	0.065
C2	CASTOR (windowed, $N = 4800$)	80	0	0.09	0.28	0.13	0.07	0.92	0.91	0.96	44	0.97	0.057
C2	FlowMSM + DYNOTEARS	160	0	0.16	0.45	0.22	0.13	0.72	0.90	0.87	8	0.83	0.058
C2	FlowMSM + VARLiNGAM	160	0	0.19	0.43	0.24	0.13	0.58	0.72	0.82	10	0.79	0.087
C4_power	SEAM	800	0	0.76	0.75	0.75	0.76	0.92	0.92	0.95	74	0.98	0.032
C4_power	SEAM, test only	800	0	0.77	0.76	0.76	0.77	0.93	0.92	0.95	74	0.98	0.032
C4_theoremR	SEAM, true regimes	80	0	0.00	0.00	0.00	0.97	0.96	0.96	0.98	94	–	0.039
C4_theoremR	no sharing	80	0	0.06	0.09	0.07	0.97	0.94	0.94	0.97	90	–	0.062
C4_theoremR	sharing without the quotient	80	0	0.06	0.04	0.05	0.67	0.56	0.67	0.86	4	–	0.317
C5_castorbench	SEAM	40	0	0.96	0.98	0.97	0.97	0.71	0.81	0.85	40	0.77	0.255
C5_castorbench	SEAM, test only	40	0	0.96	0.98	0.97	0.97	0.71	0.81	0.85	40	0.77	0.255
C5_castorbench	CASTOR (authors)	40	2	0.95	0.94	0.94	0.94	0.93	0.93	0.92	71	0.89	0.244
C9_t3	SEAM	40	0	0.57	0.57	0.57	0.57	0.83	0.90	0.92	56	0.96	0.077
C9_t3	SEAM, test only	40	0	0.57	0.57	0.57	0.57	0.83	0.90	0.92	56	0.96	0.077
C9_t3	SEAM, Student-t	40	0	1.00	1.00	1.00	1.00	0.98	0.99	1.00	98	0.99	0.022
C9_t3	CASTOR ($\tau = 0.1$)	40	0	0.13	0.28	0.18	0.16	0.64	0.61	0.84	29	0.86	0.188

Table 29: **Per cell, SEAM** ($d = 5$, 20 series a cell): localization F1 and effect error with learned and with true regimes.

β	scales	innovations	schedule	n	loc. F1	loc. F1, true regimes	SHD	effect error	effect error, true regimes
0.3	unit	gaussian	recurrent	20	0.95	1.00	0.10	0.113	0.091
0.3	unit	gaussian	switch	20	1.00	1.00	0.00	0.068	0.066
0.3	unit	macro	recurrent	20	0.95	1.00	0.28	0.134	0.101
0.3	unit	macro	switch	20	1.00	1.00	0.10	0.096	0.094
0.3	unit	rivers	recurrent	20	0.67	0.99	0.55	0.191	0.073
0.3	unit	rivers	switch	20	0.90	1.00	0.23	0.135	0.104
0.3	unit	skab	recurrent	20	0.95	1.00	0.15	0.102	0.081
0.3	unit	skab	switch	20	1.00	1.00	0.07	0.118	0.121
0.3	hetero	gaussian	recurrent	20	0.82	0.93	0.42	0.190	0.183
0.3	hetero	gaussian	switch	20	0.93	0.93	0.25	0.144	0.144
0.3	hetero	macro	recurrent	20	0.82	0.91	0.62	0.226	0.196
0.3	hetero	macro	switch	20	0.93	0.93	0.35	0.150	0.149
0.3	hetero	rivers	recurrent	20	0.88	0.95	0.35	0.164	0.135
0.3	hetero	rivers	switch	20	0.90	0.93	0.42	0.182	0.175
0.3	hetero	skab	recurrent	20	0.90	0.98	0.25	0.196	0.167
0.3	hetero	skab	switch	20	0.85	0.90	0.53	0.239	0.209
0.5	unit	gaussian	recurrent	20	1.00	1.00	0.00	0.065	0.064
0.5	unit	gaussian	switch	20	1.00	1.00	0.00	0.054	0.054
0.5	unit	macro	recurrent	20	1.00	1.00	0.03	0.065	0.064
0.5	unit	macro	switch	20	1.00	1.00	0.07	0.070	0.070
0.5	unit	rivers	recurrent	20	1.00	0.99	0.05	0.072	0.060
0.5	unit	rivers	switch	20	1.00	1.00	0.05	0.069	0.066
0.5	unit	skab	recurrent	20	1.00	1.00	0.05	0.060	0.058
0.5	unit	skab	switch	20	1.00	1.00	0.05	0.094	0.094
0.5	hetero	gaussian	recurrent	20	1.00	1.00	0.03	0.082	0.080
0.5	hetero	gaussian	switch	20	1.00	1.00	0.00	0.063	0.074
0.5	hetero	macro	recurrent	20	1.00	1.00	0.05	0.074	0.084
0.5	hetero	macro	switch	20	1.00	1.00	0.10	0.085	0.084
0.5	hetero	rivers	recurrent	20	1.00	1.00	0.00	0.072	0.073
0.5	hetero	rivers	switch	20	1.00	1.00	0.23	0.123	0.122
0.5	hetero	skab	recurrent	20	1.00	1.00	0.05	0.118	0.117
0.5	hetero	skab	switch	20	0.97	1.00	0.10	0.119	0.106