REGIME-SPECIFIC CAUSAL DISCOVERY FROM A VARIANCE-BASED MARKOV-SWITCHING-MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal causal discovery typically assumes that a single causal structure and set of mechanisms govern an entire time series. In many systems, however, the process repeatedly switches among unobserved regimes, and different regimes may have different causal mechanisms and lagged parent sets. Pooling observations across such regimes may therefore obscure genuine causal relations. We study the recovery of regime-specific lagged causal structure when the regimes themselves are latent and recurrent. We propose the Variance-based Markov-Switching-Model (VB-MSM), in which a shared latent Markov state selects regime-specific nonlinear conditional mechanisms. We measure the contribution of each candidate parent through its complete edge function, i.e., an edge is considered active when its contribution to the regime-conditional mean varies with its input. Thus our method is invariant of the specific parameterization. We show that removing a true parent changes the regime-conditional mean under common causal assumptions. VB-MSM combines nonlinear mechanism learning with exact forward–backward inference over the finite-state regime process in an EM-style update. On nonlinear switching benchmarks, it accurately recovers regime-specific graphs, latent regimes, and intervention effects. Comparisons with baseline methods like iMSM and FlowMSM+RHINO show its strengths across different mechanisms. Experiments with added noise also retain strong recovery at native and reduced sample lengths.

1 INTRODUCTION

Temporal causal discovery asks which variables directly affect later variables, recovering lagged parent sets under assumptions that connect observational dependence to intervention effects (Peters et al., 2013; Runge et al., 2019). Most formulations assume one graph and one family of mechanisms throughout the trajectory (Runge, 2020; Tank et al., 2022; Gong et al., 2023; Kong et al., 2025). This stationarity makes analyses simple, but can be restrictive for systems that move repeatedly among different regimes, e.g., operating conditions, weather patterns, or market environments (Michelangeli et al., 1995; Ma & Zhang, 2022; Brinkman et al., 2022; Hochman et al., 2021; Yang, 2023).

This paper studies such mechanism changes that are driven by recurrent unobserved categorical regimes (Balsells-Rodas et al., 2024b). At time t, a latent global state Z_t selects a lagged graph together with the corresponding structural equations mapping each variable's parents to its conditional distribution, and a regime may return after an absence. Neither the regime path nor the graphs are observed, and their recovery is coupled, because the regimes are distinguished by the conditional mechanisms they select, while estimating conditional mechanisms requires regime assignments.

A naive way is to pool all observations and fit a single graph, so that traditional methods for stationary series can apply. However, this can cancel coefficients of opposite sign, dilute an edge that is active in only one regime, and merge distinct parent sets into one inflated graph; see Figure 1 for illustration. Therefore, we try to use regime-specific information to recover regime-specific causal structures.

Our Variance-based Markov-Switching-Model (VB-MSM) lets a global state select additive nonlinear conditional means. A switching regression on fixed nonlinear features supplies the initial state partition, after which alternating updates fit the mechanisms and refine the state distribution. A structural penalty acts on complete edge functions, whose variation also supplies the graph score.

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.

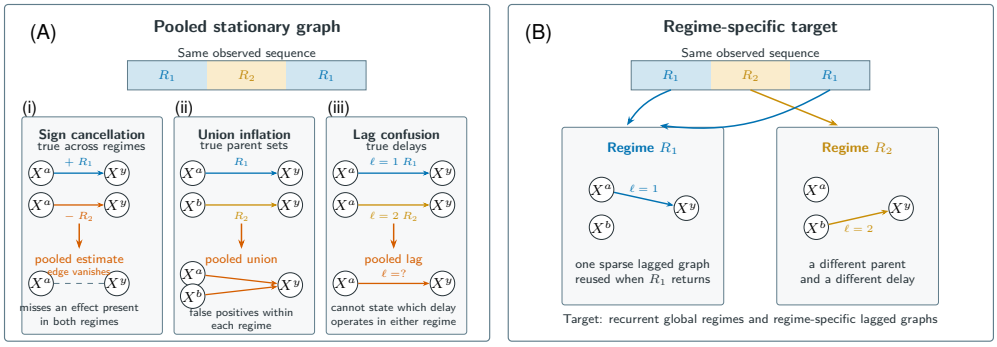

Figure 1: Pooling recurrent regimes can obscure regime-specific causal structures. Panel (A) shows (i) cancellation of opposite-sign coefficients, (ii) an inflated union of distinct parent sets, and (iii) ambiguity when an edge has different lags across regimes. Panel (B) shows the target of a recurrent regime path and its associated lagged graphs.

The target of our model is the complete contribution of a lagged variable to the conditional mean. We show that under explicit additive and support conditions, an edge function has positive variance exactly when deleting its input changes the regime-conditional mean. Such variation is therefore a signal for regime-specific causal support in the population level, and is hence our estimand.

We also compare VB-MSM with a matched zero-penalty control, with the nonlinear switching initializer it starts from, with a nonlinear MSM, and with FlowMSM combined with RHINO. CASTOR supplies a separate comparison through its native per-trajectory interface. Section 2 offers a review-level of their descriptions. Further experiments examine intervention contrasts and empirical innovations.

2 Related Work

Stationary temporal causal discovery. Classical methods define edges through structural equations, conditional independence, or penalized likelihood, among them VAR-LiNGAM, TiMINo, PCMCI$^+$, and DYNOTEARS (Hyvärinen et al., 2010; Peters et al., 2013; Runge, 2020; Pamfil et al., 2020). Neural Granger methods, TCDF, CR-VAE, and CausalFormer widen the predictive class (Tank et al., 2022; Nauta et al., 2019; Li et al., 2023; Kong et al., 2025), usually around a Granger-style target in which one variable's past improves prediction of another after conditioning on the remaining history. Our criterion rests on the same conditional-predictive foundation and on the causal assumptions in Condition 1. What differs is where the functions live, since we fit additive edge functions inside each regime-specific likelihood and read support from their variation. Amortized causal discovery and RHINO also place a graph inside a likelihood (Löwe et al., 2022; Gong et al., 2023), but assume mechanisms that stay fixed within a series.

Nonstationary causal discovery. CD-NOD and related state-space models allow mechanisms to change across time or domains (Huang et al., 2019; 2020). Regime-PCMCI alternates persistent regime assignment with conditional-independence discovery (Saggioro et al., 2020), while semi-stationary PCMCI assumes the changes recur periodically (Gao et al., 2023). CASTOR estimates temporal regime assignments and regime-wise DAGs together transductively (Rahmani & Frossard, 2025). In contrast, VB-MSM treats the regimes as Markov states and is generative.

Regime-dependent Granger causality. Markov-switching VARs state regime-dependent Granger noncausality directly (Psaradakis et al., 2005; Droumaguet et al., 2017). Our sparse VAR-HMM control is a penalized finite-state member of that family, and its state-specific lag coefficients define a linear support target. Corollary 1 shows its limitations. Even when the regimes are known, an edge that matters under intervention can be missing from the best regime-wise *linear coefficient* graph.

Recurrent global-state models. Identifiable high-order Markov switching models learn nonlinear recurrent regimes jointly and read graphs off thresholded derivatives of the fitted transition functions

Table 1: Closest interfaces from the literature.

Method	State object	Graph interface	Recurrent inference	New trajectories
High-order MSM	global	thresholded mechanism derivative	exact chain	generative
Sparse MS-VAR (VAR-HMM)	global	sparse linear lag coefficients	exact chain	generative
FlowMSM	global	downstream causal method	exact chain	generative
SDCI	variable-local	variational edge type	approximate recurrent	amortized
CASTOR	per-series regime	jointly fitted DAG	EM partition	transductive
VB-MSM	global	effective edge functions	exact chain	generative

(Balsells-Rodas et al., 2024b). FlowMSM uses normalizing flows for flexible regime-dependent densities and keeps regime detection separate from a later stationary causal-discovery stage such as DYNOTEARS or RHINO (Hulsman et al., 2026). VB-MSM instead fits additive edge functions inside the switching likelihood, and for smooth additive means the two readouts agree in population, since derivatives and variances share a support under the conditions in Appendix A. SDCI targets conditional summary graphs with per-variable local states and a variational graph model, requiring approximate recurrent inference in that larger state space (Balsells-Rodas et al., 2025). Multi-lag switching theory identifies distinct regimes and, under further assumptions, latent causal graphs (Balsells-Rodas et al., 2026a). ΩSDS allows recurrent regime transitions and optimizes an exact likelihood through an invertible observation model (Balsells-Rodas et al., 2026b). Our regime process is exogenous and our continuous variables observed, and a single global state keeps chain inference exact for a specified fitted model.

Additive switching models and functional support. Markov-switching generalized additive models fit state-specific nonlinear predictors by penalized likelihood (Langrock et al., 2017), and boosting extensions select variables within EM (Adam et al., 2022). Our penalty adapts the centered function norms of sparse additive modeling (Ravikumar et al., 2009). MOSAIC likewise uses additive functions to define latent-to-observed support and discusses variance scores for symmetric responses (Fan et al., 2026). Our subject is lagged support among observed variables, with the edge functions fitted inside one shared switching likelihood.

3 Latent-regime causal model

Sections 3.1 and 3.2 set out the causal target and the population identification result. Subsequent subsections describe our model.

3.1 Target and assumptions

Let $X_t = (X_t^1, \dots, X_t^p) \in \mathbb{R}^p$ and let

$$\mathcal{H}_t = (X_{t-1}, \dots, X_{t-L}) \in \mathbb{R}^{p \times L} \tag{1}$$

denote the candidate causal history. A global regime $Z_t \in \{1, \dots, K\}$ follows an exogenous first-order Markov chain with initial distribution π_0 and transition matrix Π, so that $p(Z_t \mid Z_{t-1}, \mathcal{H}_t) = \Pi_{Z_{t-1}, Z_t}$. We assume exogenous regime process, i.e., history does not trigger transition. We condition the fitted likelihood on $X_{1:L}$ and use π_0 for the first modeled state Z_{L+1}. When $Z_t = k$, every variable follows regime k's graph and mechanisms, so the state stands for one shared operating condition. Regime k carries a binary adjacency tensor $A^{(k)} \in \{0,1\}^{p \times p \times L}$ under the convention

$$A_{ij\ell}^{(k)} = 1 \quad \Longleftrightarrow \quad X_{t-\ell}^i \to X_t^j \text{ is a structural parent relation when } Z_t = k, \tag{2}$$

so that $A_{ij\ell}^{(k)} = \mathbf{1}\Big\{(i,\ell) \in \mathrm{Pa}_j^{(k)}\Big\}$. In the sense of regime-specific structural assignments, we have

$$X_t^j = f_j^{(k)}\Big(A_{:,j,:}^{(k)} \odot \mathcal{H}_t\Big) + \varepsilon_{t,j}^{(k)}, \qquad Z_t = k, \tag{3}$$

where $\odot$ is the Hadamard product and $\mathbb{E}\varepsilon_{t,j}^{(k)} = 0$. The mask $A_{:,j,:}^{(k)}$ decides which lagged entries may enter the mechanism for target j, so an entry with $A_{ij\ell}^{(k)} = 0$ cannot influence X_t^j.

We consider only lagged edges, so temporal order already fixes every direction. Whether those edges also admit a causal reading depends on the conditions below. Without them the output is a regime-specific predictive graph read off fitted functions, and nothing more.

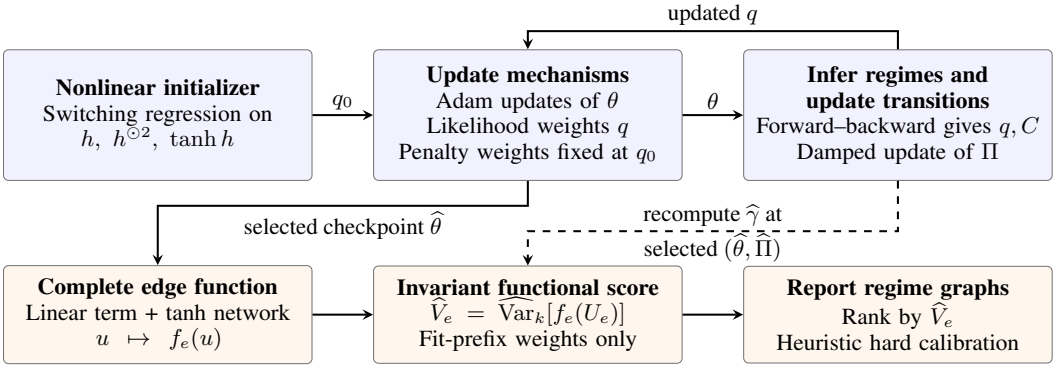

Figure 2: VB-MSM fitting and graph reporting. A nonlinear switching regression supplies q_0 for EM-style fitting. Each of eight cycles takes 100 Adam steps on (13), holding the likelihood weights q fixed within the cycle and the penalty weights at q_0 throughout fitting. Cycles 1–4 use $q = q_0$. After cycle 4 and each later mechanism update, exact forward–backward at the current (θ, Π) refreshes q and the expected transition counts C. The updated q feeds the next mechanism update. Graph reporting recomputes smoothing weights on the fit prefixes at the selected checkpoint $(\widehat{\theta}, \widehat{\Pi})$. The variation of each complete edge function gives an invariant support score, and a fixed mixture rule turns those scores into a hard graph.

3.2 Population interpretation

We identify regime-specific graph support given identification of the regime-indexed observation law, which the support theorem takes as a premise rather than proves. Let

$$m_j^{(k)}(h) = \mathbb{E}[X_t^j \mid \mathcal{H}_t = h, Z_t = k], \tag{4}$$

where h ranges over the values the history $\mathcal{H}_t$ can take.

Condition 1 (Regime-conditional causal support)**.** *(i) The regime-indexed law of $(\mathcal{H}_t, X_t)$ is identified from the observed process up to one common relabeling of the regimes, and every regime has positive occupancy. (ii) Given $(\mathcal{H}_t, Z_t)$, the assignments in (3) are causally sufficient, use a lag order at least as large as the true one, admit a stationary joint law for (X_t, Z_t) with finite second moments, and satisfy $\mathbb{E}[\varepsilon_{t,j}^{(k)} \mid \mathcal{H}_t, Z_t = k] = 0$ for every (j, k). (iii)* (Structural minimality) *If $A_{ij\ell}^{(k)} = 1$, then $m_j^{(k)}$ is not measurable with respect to $\mathcal{H}_{t,-(i,\ell)}$.*

Sufficient conditions for part (i) are available, e.g., distinguishable Gaussian regime kernels with analytic mean and covariance functions (Balsells-Rodas et al., 2024a;b; Hulsman et al., 2026), and is an assumption we make on our model. Stationarity in part (ii) fixes the population law so that the expectations are well-defined. Parts (ii) and (iii) connect conditional-mean dependence to structural support, through the common causal assumptions.

Theorem 1 (Identification of regime-specific lagged support)**.** *Under Condition 1, define*

$$\Delta_{ij\ell}^{(k)} = \mathbb{E}\left[\left\{m_j^{(k)}(\mathcal{H}_t) - \mathbb{E}[m_j^{(k)}(\mathcal{H}_t) \mid \mathcal{H}_{t,-(i,\ell)}, Z_t = k]\right\}^2 \middle| Z_t = k\right]. \tag{5}$$

Then $A_{ij\ell}^{(k)} = 1$ if and only if $\Delta_{ij\ell}^{(k)} > 0$. Consequently, the complete pool of graphs is identified from the observed law up to the same common relabeling of the regimes.

The quantity $\Delta_{ij\ell}^{(k)}$ compares the full-history prediction of X_t^j in regime k with the best prediction available once coordinate (i, ℓ) has been deleted. The mean squared difference between them is zero exactly when the deletion costs nothing, so a positive value identifies a lagged edge. The criterion therefore treats the dependence as a whole, regardless of its parameterized form. This criterion finds nonlinear edges that a linear coefficient graph could miss; see, e.g., Corollary 1 in the Appendix.

3.3 ADDITIVE SWITCHING LIKELIHOOD

Section 3.2 describes the true regime-conditional mean $m_j^{(k)}$ of (4). The model we fit is additive, so each lagged variable contributes through a function of that variable alone. For each candidate edge $e = (i, j, \ell)$ and regime k, we fit a scalar function

$$f_e^{(k)}(u) = \beta_e^{(k)} u + \sum_{r=1}^{H} \alpha_{er}^{(k)} \tanh\Big(w_{er}^{(k)} u + c_{er}^{(k)}\Big). \tag{6}$$

The argument u is one lagged value, and the linear coefficient $\beta_e^{(k)}$ together with the $H = 16$ tanh units carries the whole contribution of that single lagged variable. The conditional mean of target j at regime k is estimated by summing these contributions over sources and lags, namely

$$\mu_{tj}^{(k)} = \widehat{m}_j^{(k)}(\mathcal{H}_t) = b_j^{(k)} + \sum_{\ell=1}^{L} \sum_{i=1}^{p} f_{(i,j,\ell)}^{(k)}(X_{t-\ell}^i), \tag{7}$$

where $b_j^{(k)}$ is an intercept and $\mu_{tj}^{(k)}$ is the fitted mean for target j at time t under regime k. Every history coordinate enters through its own edge function. The model can therefore express nonlinear additive effects, but no interaction between two parents.

Similar to VAE literature, the fitted observation law is assumed Gaussian with diagonal covariance,

$$p_\theta(X_t \mid \mathcal{H}_t, Z_t = k) = \prod_{j=1}^{p} \mathcal{N}\Big(X_t^j; \mu_{tj}^{(k)}, (\sigma_j^{(k)})^2\Big), \tag{8}$$

where $\sigma_j^{(k)} > 0$ is a noise scale to be learned, specific to the regime and the target. These emissions and the Markov law for Z_t define the switching likelihood. All trajectories share the parameters, and each one has its own regime path as a realization.

3.4 FUNCTIONAL SUPPORT

The deletion gap $\Delta_e^{(k)}$ of Section 3.2 requires projecting onto functions of a reduced history, which the fitted model cannot report. We now give a quantity it does report, and state when the two agree.

We first introduce some notations. A history value is $h \in \mathbb{R}^{p \times L}$, with h_e its coordinate (i, ℓ) for the edge $e = (i, j, \ell)$, so h is a full history and h_e one number inside it. The matching random quantities are the realized history $\mathcal{H}_t$ and its coordinate $U_e = X_{t-\ell}^i$, so that $h = \mathcal{H}_t$ and $h_e = U_e$ once t is realized. Expectations are taken under Q_k, the regime-conditional law of $\mathcal{H}_t$.

Write $\mathcal{E}_j = \{(i, j, \ell) : 1 \le i \le p,\ 1 \le \ell \le L\}$ for the pL candidate edges into target j, parents and nonparents alike. We assume the population mean *additive* in regime k, i.e., it can decompose as

$$m_j^{(k)}(h) = b_j^{(k)} + \textstyle\sum_{e \in \mathcal{E}_j} f_e^{(k)}(h_e), \tag{9}$$

so no term depends on two coordinates at once. This is the functional form of the decoder in (7), which likewise sums over every source and lag. It holds automatically for the fitted $\widehat{m}_j^{(k)}$.

If $A_{ij\ell}^{(k)} = 0$ then (3) makes the mean invariant to coordinate (i, ℓ), so $f_e^{(k)}$ is constant there. Every candidate edge thus has a function defined before any graph is known, and the graph is read off from which functions are constant. Given (9), define the score $s_e^{(k)} = \sqrt{V_e^{(k)}}$ where

$$V_e^{(k)} = \mathrm{Var}_{Q_k}\{f_e^{(k)}(U_e)\} = \mathbb{E}\Big[\big\{f_e^{(k)}(U_e) - \bar{f}_e^{(k)}\big\}^2 \mid Z_t = k\Big], \quad \bar{f}_e^{(k)} = \mathbb{E}_{Q_k}[f_e^{(k)}(U_e)], \tag{10}$$

The score asks one question about each candidate edge $e \in \mathcal{E}_j$, namely, as the driver U_e moves across the values it takes in regime k, how much does this edge's contribution to the mean move? A nonparent has a constant function and contributes always the same amount, so it scores zero. Dependence of any shape makes the contribution vary and the score positive.

The score is a functional of the edge function rather than of the parameters that express it. Reparameterizing the network without changing the function $f_e^{(k)}$ leaves the score unchanged, as does

moving a constant between $f_e^{(k)}$ and the intercept $b_j^{(k)}$, since the variance in (10) is centered. For example, a deterministic gate admits $f_e(u) = \rho_e a_e(u) = r_e\{(\rho_e/r_e)a_e(u)\}$ for any $\rho_e, r_e \in (0,1)$ when the mechanism can be freely rescaled. Gate magnitudes and rankings can therefore change without changing the edge functions or likelihood, whereas the functional scores remain unchanged. This ambiguity concerns deterministic gates with freely rescalable mechanisms.

Variance is generally larger than the deletion gap, since part of an edge's variation may be predictable from another coordinate and deletion charges only for the rest. The two agree under further assumptions. If (9) holds and Q_k has a positive density on a product of intervals, then $V_e^{(k)} > 0$ picks out exactly the edges with $\Delta_e^{(k)} > 0$ (Proposition 1). The product support carries that argument, since it lets U_e vary while the rest of the history is held fixed, so an edge surviving deletion cannot have a constant function; see Appendix A.

Reporting (10) requires replacing Q_k, which is unavailable because the regime is never observed. Were the regime path known, one would estimate $V_e^{(k)}$ by the sample variance of $\widehat{f}_e^{(k)}$ over the times with $Z_t = k$. Since it is not, each time contributes to every regime in proportion to its posterior probability of being in that regime. Those probabilities are the *smoothing weights*

$$\widehat{\gamma}_{ntk} = p_{\widehat{\theta}}\big(Z_{nt} = k \mid X_{n,1:T_n^{\mathrm{fit}}}\big)\,, \tag{11}$$

the node marginals returned by forward–backward in the fitted chain. Three properties of (11) matter here. They are *smoothing* because each conditions on the whole observation block rather than only the past. They are *exact*, because the state of the regimes is finite, so forward–backward returns the marginals of the fitted posterior with no variational gap. They are computed on the *fit prefixes* alone, so T_n^{fit} ends at the fitting block and the graph is read off exclusively from training data. Each $\widehat{\gamma}_{ntk}$ is a soft regime assignment, nonnegative and summing to one over k, and $W_k = \sum_{n,t} \widehat{\gamma}_{ntk}$ is regime k's total posterior mass. With $u_{nte} = X^i_{n,t-\ell}$ being the realized driver value, we use

$$\widehat{V}_e^{(k)} = \frac{1}{W_k}\sum_{n,t} \widehat{\gamma}_{ntk}\big\{\widehat{f}_e^{(k)}(u_{nte}) - \overline{f}_e^{(k)}\big\}^2, \qquad \overline{f}_e^{(k)} = \frac{1}{W_k}\sum_{n,t} \widehat{\gamma}_{ntk}\widehat{f}_e^{(k)}(u_{nte}). \tag{12}$$

The sums run over the modeled fit-prefix times. It is the plug-in estimate for the conditional variance $V_e^{(k)}$, measuring how much the fitted edge function varies under the inferred regimes.

Two approximations separate it from the population quantity. The weights come from a fitted model rather than the true regime law, and the average is over the realized fit-prefix histories rather than Q_k. These weights come from the final fitted model, not from the initializer. The feature HMM of Section 3.5 supplies a separate posterior q_0, which anchors the penalty and is held fixed, whereas graph reporting recomputes (11) from the selected checkpoint, so reported scores use the regimes the model ends up inferring. Both are exact forward–backward marginals on the fit prefixes, each with its own emission model and transition matrix.

3.5 Functional regularization and fitting

VB-MSM penalizes variation in complete edge functions. Let q_0 be the fit-prefix feature-HMM posterior, weights of the form (11) computed from the initializer's emissions rather than the decoder's, written $\widehat{\gamma}^0_{ntk}$, and let $\widehat{V}^{(k)}_{e,0}$ denote (12) evaluated with $\widehat{\gamma}^0_{ntk}$ in place of $\widehat{\gamma}_{ntk}$. The subscript 0 marks the weights alone, and the edge functions inside are always the current ones, so $\widehat{V}^{(k)}_{e,0}$ varies with θ and carries the gradient that shrinks weak edges. Let $M = \sum_n (T_n^{\mathrm{fit}} - L)$ count the modeled time points across fit prefixes of lengths T_n^{fit}. For a current state distribution q, the structural loss is

$$\mathcal{L}_F(\theta; q) = -\frac{1}{M}\mathbb{E}_q[\log p_\theta(X \mid Z)] + \frac{\lambda}{K}\sum_{k,e}\left\{\sqrt{\widehat{V}^{(k)}_{e,0} + \epsilon} - \sqrt{\epsilon}\right\}, \qquad \epsilon = 10^{-12}. \tag{13}$$

The two terms of (13) play distinct roles. The first is the expected complete-data negative log likelihood, which under (8) reduces to a Gaussian loss on each target weighted by the current regime responsibilities. Dividing by M puts it on a per-observation scale so that λ means the same thing across trajectory counts and lengths.

The second term penalizes $\sqrt{\widehat{V}_e}$, the standardized-unit strength $s_e^{(k)}$ of (10), giving a group lasso acting on complete edge functions. This adapts the centered function-norm penalty of sparse additive modeling (Ravikumar et al., 2009). Dividing by K keeps the total comparable as the regime count changes, and adding ϵ makes $\sqrt{\cdot}$ differentiable at zero, with $\sqrt{\epsilon}$ subtracted so an inactive edge contributes exactly zero.

Fitting alternates two updates in the manner of EM. The structural parameters θ, namely every $\beta_e^{(k)}$, $\alpha_{er}^{(k)}$, $w_{er}^{(k)}$, $c_{er}^{(k)}$, each intercept $b_j^{(k)}$, and each $\log \sigma_j^{(k)}$, are updated by Adam on (13) with the responsibilities held fixed. The E-step recomputes the state distribution q by exact forward–backward at the current (θ, Π). A separate parameter update reestimates the transition matrix Π from expected transition counts, with damping and the Dirichlet pseudocounts of Appendix C. Only π_0 stays fixed as a pure initializer.

Concretely, the feature HMM supplies the initial regime distribution q_0, and the edge functions are fitted over eight cycles of 100 Adam steps each. The first four cycles use q_0 in the likelihood. Forward–backward refreshes these weights after cycle four and again after each later cycle. Appendix C gives the initialization, optimization coordinates, and transition updates. Filtering NLL on the checkpoint block picks the cycle and $\lambda \in \{0, 0.001, 0.01\}$, so neither the graph nor the state labels enter model selection. A $\lambda = 0$ control runs the same initialization and schedule. Both use generative state inference, and graph reporting uses the final model's fit-prefix weights rather than the fixed q_0 used by the penalty.

3.6 Hard-graph reporting

For threshold-free AUPRC we use $\widehat{V}_e$ in squared standardized response units. Hard reporting needs a cutoff, obtained by fitting a one-component and a two-component Gaussian mixture to $\log \max(\widehat{V}_e, 10^{-12})$ across all candidate edges, with 20 mixture starts, seed zero, and covariance regularization 10^{-6}. Two components are accepted only when $\mathrm{BIC}(M_1) - \mathrm{BIC}(M_2) \geq 10$. The mixture determines how many edges to retain. We count the candidate edges whose scores have at least a 50% probability of belonging to the higher-score group. We then retain that number of edges with the largest variance scores. These calibration settings can be shared by whatever method uses functional scores, and see no true graph, true edge count, response mask, or the test likelihood.

4 Experiments

4.1 Protocol

The main benchmark is a centered-even nonlinear process with 48 trajectories of length 300, $p = 8$, $K = 3$, lag order one, and mean dwell 35. Four autoregressive drivers supply two regime-specific parents to each of four responses, and we evaluate recovery at fixed K. Its best linear approximation is near zero, so linear switching is misspecified here while the nonlinear causal information survives.

Each trajectory is split chronologically into 50, 15, and 35 percent, for fitting, checkpoint selection, and testing. All 48 trajectories contribute fit prefixes, and preprocessing draws on the fit prefix alone. The filtering state may cross block boundaries, but fitting and checkpoint selection never see later observations. The pooled switching models use fixed architectures and the true regime count. On the same 100 systems we compare VB-MSM, its matched $\lambda = 0$ control, the nonlinear feature HMM, and that HMM's raw-feature ablation. A paired mechanism experiment changes only the standardized response transform, to linear, monotone tanh, or mixed additive effects, and the literature comparison runs iMSM and FlowMSM with RHINO on 20 paired systems per mechanism. Appendix F adds a 100-system iMSM study, budget checks, and a full-trajectory CASTOR comparison. The empirical-innovation study uses 20 systems per condition with six variables and two regimes.

On the test block we report one-step RMSE from causal forward filtering and regime NMI for the exact generative Viterbi path, computed offline with the parameters fixed. Graph recovery is scored separately, because the model learns K support tensors rather than a graph per time point. Within each regime we compute response-support AUPRC (average precision) and hard F1, then average across regimes. Functional scores pass through the truth-blind rule of Section 3.6, and Appendix F states the external thresholds. Methods are paired within each system, and means and standard

Table 2: Centered-even mechanisms, with mean ± standard error over 100 paired systems. All graph entries use effective-function variances and the same fixed mixture rule. The zero-penalty control shares the VB-MSM initialization and training schedule.

Estimator	AUPRC	Hard F1	Regime NMI	RMSE
VB-MSM	$\mathbf{0.9975 \pm 0.0016}$	$\mathbf{0.9945 \pm 0.0026}$	$\mathbf{0.9708 \pm 0.0081}$	$\mathbf{0.7171 \pm 0.0026}$
VB-MSM, $\lambda = 0$	0.9974 ± 0.0017	0.9867 ± 0.0101	0.9706 ± 0.0081	0.7191 ± 0.0026
Feature HMM, nonlinear	0.9904 ± 0.0033	0.9786 ± 0.0054	0.8292 ± 0.0132	0.8317 ± 0.0022
Feature HMM, raw only	0.6800 ± 0.0118	0.0528 ± 0.0155	0.0565 ± 0.0077	0.9651 ± 0.0007

errors treat independent systems as the sampling units. See the Appendix for further details and reproducibility information.

4.2 Functional support and regime recovery

Table 2 compares the estimators on the centered-even process. VB-MSM reaches hard F1 0.9945 ± 0.0026, regime NMI 0.9708 ± 0.0081, and filtering RMSE 0.7171 ± 0.0026. The nonlinear initializer has AUPRC 0.9904 ± 0.0033 and F1 0.9786 ± 0.0054 from its complete edge functions. Fitting the neural mechanisms raises NMI by 0.1416 ± 0.0101 and lowers RMSE by 0.1145 ± 0.0020, both in all 100 systems. Dropping the squared and tanh initializer features sends NMI to 0.0565 ± 0.0077.

The matched $\lambda = 0$ control recovers nearly the same support and regimes. AUPRC differs by 0.00011 ± 0.00016 and is exactly equal in 97 systems, while F1 differs by 0.0078 ± 0.0080 and is identical in 98. The large recovery gains belong to the fitting procedure as a whole. Functional regularization delivers a drop in RMSE of 0.00200 ± 0.00005, better in 99 systems and tied in one. Checkpoint NLL selects $\lambda = 0.01$ in 98 systems.

4.3 Intervention contrasts

We evaluate all 96 driver-to-response contrasts for VB-MSM, its zero-penalty control, and the feature HMM, using $\mathrm{do}(U = -1.5)$ or $\mathrm{do}(U = 1.5)$ against $\mathrm{do}(U = 0)$ in raw data units. Estimates use complete fitted conditional means without hard-graph pruning, with all other history inputs fixed at zero. Nonzero-effect NMAE divides mean absolute estimation error by mean absolute true effect within each system.

On centered-even mechanisms VB-MSM has NMAE 0.0399 ± 0.0098 against 0.4441 ± 0.0056 for the feature HMM, a paired reduction of 0.4042 ± 0.0063. The $\lambda = 0$ control has NMAE 0.0397 ± 0.0098, so the penalty adds 0.00022 ± 0.00006 to nonzero-effect error while cutting 0.00089 ± 0.00009 from zero-effect estimates.

4.4 Paired mechanism comparisons

Table 3 keeps the structural systems fixed and varies only the response functions. On linear mechanisms the feature HMM forecasts slightly better and estimates nonzero intervention effects more accurately. On tanh and mixed mechanisms all three estimators recover response support perfectly in every system. For the mixed process VB-MSM raises regime NMI by 0.0120 ± 0.0009 and lowers RMSE by 0.0344 ± 0.0008 against the feature HMM, with effect NMAE 0.0126 against 0.2397. The tanh errors are 0.0146 and 0.0509.

Across the 300 fits, checkpoint NLL selects $\lambda = 0.01$. It lowers RMSE in every condition while support and regime recovery change little.

4.5 Nonlinear switching comparisons

Table 4 compares VB-MSM with iMSM and FlowMSM with RHINO on 20 paired systems per mechanism, iMSM using three starts and a 2,000-epoch limit. Appendix F gives the 100-system iMSM study, budget checks, and CASTOR results.

Table 3: Paired response-shape comparisons over 100 systems per condition. Entries are means ± standard errors. Full metrics appear in Appendix E.

Mechanism	Estimator	Hard F1	NMI	RMSE	Effect NMAE
Linear	Feature HMM	$\mathbf{0.9971 \pm 0.0029}$	$\mathbf{0.9929 \pm 0.0027}$	$\mathbf{0.7067 \pm 0.0008}$	$\mathbf{0.0128 \pm 0.0045}$
	VB-MSM, $\lambda = 0$	0.9969 ± 0.0031	$\mathbf{0.9929 \pm 0.0027}$	0.7095 ± 0.0008	0.0177 ± 0.0045
	VB-MSM	0.9969 ± 0.0031	0.9928 ± 0.0027	0.7078 ± 0.0008	0.0178 ± 0.0045
Tanh	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9982 \pm 0.0002}$	0.7068 ± 0.0005	0.0509 ± 0.0002
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	0.9981 ± 0.0002	0.7083 ± 0.0005	$\mathbf{0.0146 \pm 0.0002}$
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	0.9981 ± 0.0002	$\mathbf{0.7066 \pm 0.0005}$	$\mathbf{0.0146 \pm 0.0002}$
Mixed	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	0.9864 ± 0.0009	0.7424 ± 0.0010	0.2397 ± 0.0034
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	0.9983 ± 0.0002	0.7099 ± 0.0005	$\mathbf{0.0125 \pm 0.0001}$
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9984 \pm 0.0002}$	$\mathbf{0.7080 \pm 0.0005}$	0.0126 ± 0.0002

Table 4: Nonlinear comparisons on 20 paired systems per mechanism, with mean ± standard error. iMSM uses three starts and a 2,000-epoch limit. AUPRC and F1 concern response edges. Higher is better except for RMSE. Dashes denote unevaluated metrics. CASTOR uses a different fitting interface and appears separately in Table 9.

Condition	Estimator	AUPRC	Hard F1	NMI	RMSE
Centered-even	VB-MSM	0.9950 ± 0.0038	$\mathbf{0.9890 \pm 0.0077}$	0.9676 ± 0.0201	$\mathbf{0.7177 \pm 0.0052}$
Centered-even	iMSM	$\mathbf{1.0000 \pm 0.0000}$	0.7502 ± 0.0164	0.9921 ± 0.0011	0.7430 ± 0.0017
Centered-even	FlowMSM + RHINO	0.9735 ± 0.0052	0.9218 ± 0.0103	$\mathbf{0.9933 \pm 0.0012}$	—
Linear	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9949 \pm 0.0007}$	$\mathbf{0.7075 \pm 0.0012}$
Linear	iMSM	0.9985 ± 0.0015	0.9833 ± 0.0167	0.9715 ± 0.0229	0.7135 ± 0.0057
Linear	FlowMSM + RHINO	0.9891 ± 0.0034	0.8888 ± 0.0121	0.9941 ± 0.0007	—
Mixed	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9980 \pm 0.0004}$	$\mathbf{0.7086 \pm 0.0011}$
Mixed	iMSM	$\mathbf{1.0000 \pm 0.0000}$	0.9925 ± 0.0035	0.9973 ± 0.0005	0.7208 ± 0.0015
Mixed	FlowMSM + RHINO	0.9982 ± 0.0018	0.9142 ± 0.0105	0.9962 ± 0.0005	—

On the first 20 centered-even systems, iMSM ranks support perfectly and has higher mean regime NMI, while VB-MSM has higher mean hard F1 and lower mean RMSE (Table 7).

Against FlowMSM with RHINO, VB-MSM has the higher mean hard F1 on all three synthetic mechanisms, the centered-even gain being 0.0672 ± 0.0113, while FlowMSM has the higher mean regime NMI on centered-even mechanisms. Varying only the readout leaves response-edge ranking almost unchanged but moves hard F1 from 0.9945 to 0.0860, since the derivative fails the BIC margin in 91 of 100 centered-even systems and recovers 0.8814 under iMSM's cutoff (Appendix H).

Empirical innovations. With empirical innovations, VB-MSM recovers the response graph perfectly at native and half length, with RMSE close to iMSM's. Five-percent gross shocks reverse this: VB-MSM has F1 0.1011 and NMI 0.0136, against 0.8058 and 0.8118 for iMSM and 0.9563 and 0.9671 for FlowMSM with RHINO. Appendix J gives details.

5 Conclusion and Limitations

We introduced VB-MSM, a method that uses variance-based edge-function scores and exact forward–backward inference to recover regime-specific lagged causal structures from latent recurrent regimes. The method achieves near-perfect support recovery on centered-even benchmarks and consistently outperforms iMSM and FlowMSM with RHINO on hard F1 for centered-even and mixed mechanisms. It recovers the true support on linear, tanh, and mixed mechanisms, and estimates intervention effects substantially more accurately than the feature-HMM baseline.

The main limitations concern causal assumptions, empirical scope, and finite-sample guarantees. Causal interpretation requires an identified regime-indexed law, an exogenous regime process, sufficient lags, causal sufficiency, and structural minimality. Additionally, functional-variance equivalence assumes additive means and positive density on a Cartesian product of intervals. When these conditions fail, the fitted score remains predictive, but its causal interpretation is not guaranteed. Empirically, we study small $L = 1$ systems with specified K. Unknown-K selection, independent-trajectory generalization, and broader scaling remain open. Gross shocks sharply degrade recovery, and neither neural optimization nor mixture calibration has a finite-sample consistency guarantee.

AI USE STATEMENT

We used generative AI tools to assist with code review, experiment orchestration, manuscript drafting, and language editing. We take responsibility for the final content, claims, and artifacts produced with this assistance.

REPRODUCIBILITY STATEMENT

The Appendix gives the data-generating processes, chronological split, model configurations, optimization schedules, and graph-calibration rules. Source and environment manifests accompany the seed-level results. Run manifests record every attempted fit and its outcome. Functional graph scores use fitting-prefix observations and posterior weights. Comparator readouts follow their stated protocols, including CASTOR's complete-trajectory fitting. State-label and support truth enter only evaluation. Separate scripts reproduce the nonlinear initializer controls, VB-MSM fits, literature comparisons, and intervention contrasts. Numerical checks verify the reported aggregates and the invariance of functional scores. Code, configurations, and result records accompany the paper.

REFERENCES

Timo Adam, Andreas Mayr, and Thomas Kneib. Gradient boosting in Markov-switching generalized additive models for location, scale, and shape. *Econometrics and Statistics*, 22:3–16, 2022.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2639–2672. PMLR, 2024a.

Carles Balsells-Rodas, Yixin Wang, Pedro A. M. Mediano, and Yingzhen Li. Identifying nonstationary causal structures with high-order Markov switching models. *arXiv preprint arXiv:2406.17698*, 2024b.

Carles Balsells-Rodas, Xavier Sumba, Tanmayee Narendra, Ruibo Tu, Gabriele Schweikert, Hedvig Kjellström, and Yingzhen Li. Causal discovery from conditionally stationary time series. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2715–2741. PMLR, 2025.

Carles Balsells-Rodas, Toshiko Matsui, Pedro A. M. Mediano, Yixin Wang, and Yingzhen Li. On the identifiability of regime-switching models with multi-lag dependencies. *arXiv preprint arXiv:2601.03325*, 2026a.

Carles Balsells-Rodas, Zhengrui Xiang, Xavier Sumba, and Yingzhen Li. End-to-end identifiable and consistent recurrent switching dynamical systems. *arXiv preprint arXiv:2605.06315*, 2026b.

Braden A. W. Brinkman, Han Yan, Arianna Maffei, Il Memming Park, Alfredo Fontanini, Jin Wang, and Giancarlo La Camera. Metastable dynamics of neural circuits and networks. *Applied Physics Reviews*, 9(1):011313, 2022.

Matthieu Droumaguet, Anders Warne, and Tomasz Woźniak. Granger causality and regime inference in Markov switching VAR models with bayesian methods. *Journal of Applied Econometrics*, 32 (4):802–818, 2017.

Shicheng Fan, Nour Elhendawy, Jianle Sun, Ke Fang, Kun Zhang, Yihang Wang, and Lu Cheng. MOSAIC: Module discovery via sparse additive identifiable causal learning for scientific time series. *arXiv preprint arXiv:2605.05524*, 2026.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46624–46657. Curran Associates, Inc., 2023.

Élisabeth Gassiat, Alice Cleynen, and Stephane Robin. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1):61–71, 2016.

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

Assaf Hochman, Gabriele Messori, Julian F. Quinting, Joaquim G. Pinto, and Christian M. Grams. Do Atlantic-European weather regimes physically exist? *Geophysical Research Letters*, 48(20): e2021GL095574, 2021.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2901–2910. PMLR, 2019.

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 *Forty-third International Conference on Machine Learning*, 2026.

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.

Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. CausalFormer: An interpretable Transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):102–115, 2025.

D. R. Kothawale and M. Rajeevan. Monthly, seasonal and annual rainfall time series for All-India, homogeneous regions and meteorological subdivisions: 1871–2016. Technical Report Research Report No. RR-138, Indian Institute of Tropical Meteorology, August 2017.

Roland Langrock, Thomas Kneib, Richard Glennie, and Théo Michelot. Markov-switching generalized additive models. *Statistics and Computing*, 27:259–270, 2017.

Hongming Li, Shujian Yu, and Jose Principe. Causal recurrent variational autoencoder for medical time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8562–8570, 2023.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 509–525. PMLR, 2022.

Jie Ma and Jinkai Zhang. Progress of process monitoring for the multi-mode process: A review. *Applied Sciences*, 12(14):7207, 2022.

Paul-Antoine Michelangeli, Robert Vautard, and Bernard Legras. Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52(8):1237–1256, 1995.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Zacharias Psaradakis, Morten O. Ravn, and Martin Sola. Markov switching causality and the money–output relationship. *Journal of Applied Econometrics*, 20(5):665–683, 2005.

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

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, 108(D14):4407, 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*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397. PMLR, 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.

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.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural Granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022.

Wei Yang. Business cycles, regime shifts, and return predictability. *Journal of Financial and Quantitative Analysis*, 58(7):3058–3084, 2023.

Appendix

Notations and Preliminaries. We collect the notation and concepts that are used in the paper but not defined in the main text or elsewhere in this appendix. Trajectories are indexed by $n = 1, \dots, N$ and have full length T_n; the fit prefix of trajectory n runs over times $1, \dots, T_n^{\mathrm{fit}}$, and the synthetic benchmark uses full trajectories of length 300. A hat marks fitted quantities, e.g., the selected structural parameters $\widehat{\theta}$ and transition matrix $\widehat{\Pi}$, and the fitted scores $\widehat{V}_e^{(k)}$. We write $x \pm s$ for a mean x with standard error s. The operator $\mathrm{do}(U = u)$ denotes a hard intervention that sets a variable U to the value u, leaving the regime process, the remaining history, and the innovation law unchanged.

The evaluation abbreviations are as follows. AUPRC is the area under the precision–recall curve (average precision, AP). F1, or hard F1, is the F1 score of the thresholded hard graph. NMI is the normalized mutual information between estimated and true regime labels. RMSE is the root-mean-squared error of one-step forward-filtering forecasts. NLL is the negative log-likelihood. NMAE and MAE are the normalized and mean absolute error of intervention-effect estimates. For hidden-state inference, *filtering* conditions on observations up to and including the present, *smoothing* conditions on a whole observation block, and the *Viterbi* path is the maximum-a-posteriori state sequence. A regime's *mean dwell* is its expected number of consecutive time steps.

Reproducibility All code and data necessary to reproduce the results in the paper are available at `https://anonymous.4open.science/r/VB-MSM-TempCausDis-anonymized/`.

A Proofs

Condition 2 (A nonlinear edge with zero linear coefficient)**.** *Fix a regime k, candidate parent $U = X_{t-1}^i$, target X_t^j, and remaining history $H^- = \mathcal{H}_{t,-(i,1)}$. Conditional on $Z_t = k$, suppose*

$$(U, H^-) \overset{d}{=} (-U, H^-), \qquad X_t^j = g_k(H^-) + a_k r(U) + \varepsilon_{t,j}^{(k)}, \tag{14}$$

where $a_k \neq 0$, r is even, all displayed random variables and functions have finite second moments, and

$$\mathbb{E}[U^2 \mid Z_t = k] > 0, \qquad \mathbb{E}\big[\mathrm{Var}\{r(U) \mid H^-, Z_t = k\} \mid Z_t = k\big] > 0. \tag{15}$$

Assume $\mathbb{E}[\varepsilon_{t,j}^{(k)} \mid U, H^-, Z_t = k] = 0$. A structural intervention on U leaves Z_t, H^-, and the target innovation law unchanged. There are intervention values u, v with $r(u) \neq r(v)$.

The reflection assumption allows dependence between U and H^-. In the lag-one centered-even benchmark, reflecting the complete trajectory of driver i preserves the centered Gaussian autoregression. The other drivers and regime path are unchanged, and every response uses an even transform of its driver parents. The resulting lag-one history therefore has the required joint symmetry. Nondegenerate Gaussian innovations give the joint history a positive density, which prevents a nonconstant $r(U)$ from being determined by the other coordinates. The argument specifically concerns the centered-even mechanism.

Corollary 1 (Separation from regime-wise linear coefficient graphs)**.** *Under Conditions 1 and 2,*

$$\Delta_{ij1}^{(k)} = a_k^2 \mathbb{E}\big[\mathrm{Var}\{r(U) \mid H^-, Z_t = k\} \mid Z_t = k\big] > 0, \tag{16}$$

$$\mathbb{E}[X_t^j \mid \mathrm{do}(U = u), H^- = h^-, Z_t = k] - \mathbb{E}[X_t^j \mid \mathrm{do}(U = v), H^- = h^-, Z_t = k] = a_k\{r(u) - r(v)\}. \tag{17}$$

The coefficient on U in a best population linear predictor from an intercept and (U, H^-) is zero. Thus a graph defined by population linear coefficient support omits this intervention-relevant edge even with observed regimes.

Proof of Theorem 1. By Condition 1(i), identification of the switching law gives the regime-indexed distribution of $(\mathcal{H}_t, X_t^j)$ up to a common relabeling of the regimes. The conditional mean $m_j^{(k)}$ is therefore identified as an element of L^2, and square integrability in part (ii) places it there. Its conditional expectation given $\mathcal{H}_{t,-(i,\ell)}$ is the unique L^2 projection onto the functions measurable

with respect to the reduced history. The quantity $\Delta_{ij\ell}^{(k)}$ in (5) is the squared L^2 distance between $m_j^{(k)}$ and that projection, so it is identified, and it vanishes exactly when $m_j^{(k)}$ is itself measurable with respect to the reduced history.

It remains to match this to the support. If $A_{ij\ell}^{(k)} = 0$, the assignment (3) excludes $X_{t-\ell}^i$ from $f_j^{(k)}$. Since $\mathbb{E}[\varepsilon_{t,j}^{(k)} \mid \mathcal{H}_t, Z_t = k] = 0$ by part (ii), the conditional mean equals $f_j^{(k)}$ evaluated at the masked history, which is a function of the reduced history alone, and hence $\Delta_{ij\ell}^{(k)} = 0$. If $A_{ij\ell}^{(k)} = 1$, structural minimality in part (iii) states precisely that no such measurability holds, so the distance to the projection is strictly positive. Positive regime occupancy, also from part (i), makes every regime-indexed quantity above well defined. The recovered supports inherit the relabeling carried by the switching law, and no further ambiguity is introduced. □

Proof of Corollary 1. All expectations in this proof are conditional on $Z_t = k$. Innovation mean independence gives $m_j^{(k)}(U, H^-) = g_k(H^-) + a_k r(U)$. Its projection after removing U is $g_k(H^-) + a_k \mathbb{E}[r(U) \mid H^-, Z_t = k]$. Subtracting and squaring gives (16). The support theorem then identifies the edge. Under the stated intervention assumptions, replacing U by u or v changes only $a_k r(U)$ in the structural assignment, proving (17).

Joint reflection implies $\mathbb{E}[U \mid H^-, Z_t = k] = 0$. Hence U has zero mean and is orthogonal to the intercept, every coordinate of H^-, and $g_k(H^-)$. Evenness of r implies $\mathbb{E}[U r(U) \mid Z_t = k] = 0$. Innovation mean independence also gives $\mathbb{E}[U \varepsilon_{t,j}^{(k)} \mid Z_t = k] = 0$. Therefore $\mathbb{E}[U X_t^j \mid Z_t = k] = 0$. The normal equation for the coefficient of U reduces to its coefficient times $\mathbb{E}[U^2 \mid Z_t = k] = 0$. Positive second moment makes that coefficient zero, even if coefficients on the remaining regressors are nonunique.

If U is additionally independent of H^-, the gap reduces to $a_k^2 \operatorname{Var}(r(U) \mid Z_t = k)$. A decoder-representable example for $U \sim \mathcal{N}(0, 1)$ and $c \neq 0$ is

$$r(u) = \tanh(u + c) + \tanh(c - u) - \mathbb{E}[\tanh(U + c) + \tanh(c - U)]. \tag{18}$$

This function is even and nonconstant. Two hidden tanh units represent its variable part, with the target intercept absorbing the constant. □

Proposition 1 (Functional support for additive means)**.** *Assume Condition 1. Suppose the true regime-conditional mean has the additive representation (9) over all candidate edges $e \in \mathcal{E}_j$, with square-integrable summands. Assume Q_k has a Lebesgue density positive almost everywhere on a Cartesian product of nonempty intervals and zero outside that product. Then, for every $e \in \mathcal{E}_j$,*

$$A_e^{(k)} = 1 \quad \Longleftrightarrow \quad \Delta_e^{(k)} > 0 \quad \Longleftrightarrow \quad V_e^{(k)} > 0. \tag{19}$$

The variance $V_e^{(k)}$ is determined by the conditional mean and Q_k, independently of the choice of additive representation.

Proof. Additivity and the definition of conditional expectation give

$$\Delta_e^{(k)} = \mathbb{E}_{Q_k}\Big[\operatorname{Var}\{f_e^{(k)}(U_e) \mid H^{-e}, Z_t = k\}\Big] \leq V_e^{(k)}. \tag{20}$$

Thus $V_e^{(k)} = 0$ implies $\Delta_e^{(k)} = 0$. Conversely, a zero deletion gap implies that $f_e^{(k)}(U_e)$ equals a measurable function of H^{-e} almost surely. Positivity of the joint density transfers this equality to Lebesgue-almost every point in the product support. Fubini's theorem then implies that $f_e^{(k)}$ is constant almost everywhere on its coordinate interval, so its variance is zero. Theorem 1 gives equivalence with structural support. Applying the same argument to the difference between two additive representations shows that their edge summands differ only by constants, which do not affect variance. □

The proposition supplies a population target under explicit additivity and support assumptions. It does not establish consistency of the fitted decoder, the regime posterior, or a finite-sample hard-graph calibrator. In particular, empirical mixture calibration of functional scores remains a reporting rule whose performance requires evaluation.

The next lemma is the deterministic step that carries an error in the fitted edge function into an error in its score. It is stated for an arbitrary probability measure, so it applies to the empirical weighting $\widehat{P}_k$ implied by (12) even though that measure and the functions were obtained from the same observations.

Lemma 1 (Stability of the variance score)**.** *Let P be a probability measure and let f, g be square integrable. Then*

$$\left|\sqrt{\mathrm{Var}_P(g)} - \sqrt{\mathrm{Var}_P(f)}\right| \le \inf_{c\in\mathbb{R}} \|g - f - c\|_{L^2(P)}. \tag{21}$$

If the right side is at most a and $\mathrm{Var}_P(f) \le v$, then $|\mathrm{Var}_P(g) - \mathrm{Var}_P(f)| \le 2\sqrt{v}\,a + a^2$. If f is constant P-almost surely, then $\mathrm{Var}_P(g) \le a^2$.

Proof. The centering operator $C_P h = h - Ph$ is the orthogonal projection in $L^2(P)$ onto mean-zero functions, so its operator norm is at most one. The reverse triangle inequality gives $|\,\|C_P g\|_2 - \|C_P f\|_2| \le \|C_P(g-f)\|_2 = \inf_c \|g - f - c\|_2$, which is (21). The second claim follows from $|x^2 - y^2| \le 2ya + a^2$ whenever $x, y \ge 0$ and $|x - y| \le a$, and the third from setting $y = 0$. □

Two consequences matter for reporting. The infimum over c means that a constant offset between the fitted and true edge function never contributes, matching the invariance of the population score. The final claim is the sharper one for absent edges, where the estimated score is bounded by the *square* of the function error, so a nonparent stays near zero even when the fit is only moderately accurate. Neither statement establishes that the fitting procedure makes that error small, which remains open alongside the calibration question above.

Relation to derivative support. Suppose each $f_e^{(k)}$ is continuously differentiable on its support interval and define the finite score

$$D_e^{(k)} = \mathbb{E}_{Q_k}\left[\left|\frac{\partial m_j^{(k)}(\mathcal{H}_t)}{\partial h_e}\right|\right] = \mathbb{E}_{Q_k}|f_e^{(k)\prime}(U_e)|. \tag{22}$$

Under the conditions of Proposition 1, $D_e^{(k)} > 0$ if and only if $V_e^{(k)} > 0$. Zero derivative expectation and positive density imply that the continuous derivative is zero throughout the interval interior, so the function is constant. Conversely, a constant function has zero derivative. The variance score therefore has the same population support as derivative-based readouts (Balsells-Rodas et al., 2024b), although their magnitudes, rankings, and finite-sample behavior can differ. Both scores depend on the represented functions rather than their neural parameters.

Population equivalence does not make the two interchangeable once the functions are estimated, because they have different continuity properties. Let U be uniform on $[0, 2\pi]$ and let $f_n(u) = n^{-1}\sin(nu)$. Then $\|f_n\|_{L^2} = 1/(\sqrt{2}n) \to 0$ and $\mathrm{Var}\{f_n(U)\} = 1/(2n^2) \to 0$, while $\mathbb{E}|f_n'(U)| = 2/\pi$ and $\mathbb{E}[f_n'(U)^2] = 1/2$ for every n. Convergence of fitted functions in L^2 therefore controls the variance score but leaves the derivative score bounded away from zero. By Lemma 1 the variance readout inherits whatever accuracy the mechanism fit achieves in prediction norm, which is the norm the likelihood controls, whereas a derivative readout additionally requires derivative convergence or an explicit smoothness bound. Every f_n here is nonconstant, so the example does not contradict the population equivalence above, and it is not a claim that the fitted width-16 networks oscillate in this way.

B Complete generative specification

For N independent trajectories, the conditional joint model is

$$\begin{aligned} &p_\theta(X_{L+1:T}, Z_{L+1:T}, \Pi \mid X_{1:L}) \\ &\quad = p(\Pi)\prod_{n=1}^{N}\Bigg[\pi_0(Z_{n,L+1})\prod_{t=L+2}^{T}\Pi_{Z_{n,t-1},Z_{nt}} \\ &\qquad\qquad \times \prod_{t=L+1}^{T}\prod_{j=1}^{p} p_{\theta,j}(X_{nt}^j \mid \mathcal{H}_{nt}, Z_{nt})\Bigg]. \end{aligned} \tag{23}$$

Here the displayed X and Z blocks include all trajectories. For one trajectory, define $p_{\pi_0,\Pi}(Z_{L+1:T}) = \pi_0(Z_{L+1}) \prod_{t=L+2}^{T} \Pi_{Z_{t-1},Z_t}$. Generative forward–backward evaluates

$$p_\theta(Z_{L+1:T} \mid X_{1:T}, \Pi) \propto p_{\pi_0,\Pi}(Z_{L+1:T}) \exp\left\{ \sum_{t=L+1}^{T} \log p_\theta(X_t \mid \mathcal{H}_t, Z_t) \right\}. \tag{24}$$

Causal filtering uses the same initial-time convention.

Each transition row has a Dirichlet prior with an augmented diagonal, favoring persistence without forcing off-diagonal probabilities to zero, so regimes can still be left and revisited. Observation variances depend on regime and target. All model parameters are shared across trajectories, while the regime paths are trajectory specific.

C Initialization and training details

Proposition 2 (Exact finite-chain update)**.** *Let $\zeta_{1:m}$ be a finite-state path with positive finite initial, transition, and emission potentials. Write their product as*

$$\widetilde{p}(\zeta_{1:m}) = a(\zeta_1) \prod_{s=2}^{m} b_s(\zeta_{s-1}, \zeta_s) \prod_{s=1}^{m} c_s(\zeta_s). \tag{25}$$

Among all path distributions q, the objective $\mathbb{E}_q[\log \widetilde{p}(\zeta_{1:m})] + \mathbb{H}[q]$ has the unique maximizer $q = \widetilde{p}/\sum_\zeta \widetilde{p}(\zeta)$. Forward–backward computes its marginals exactly.

Proof. Let $C = \sum_\zeta \widetilde{p}(\zeta)$, which is finite and positive. The objective equals $\log C - \mathrm{KL}(q \| \widetilde{p}/C)$. The divergence is nonnegative and vanishes only at the stated distribution. Its finite-chain factorization permits forward–backward evaluation. □

For VB-MSM, $m = T - L$ and $\zeta_s = Z_{L+s}$. The emissions in (8) and the fitted Markov law supply the potentials.

Feature initializer. For the flattened lag vector $h \in \mathbb{R}^{pL}$ the fixed features are $(h, h^{\odot 2}, \tanh h)$, standardized coordinatewise on the fit prefix. A switching ridge regression predicts X_t with regime-specific diagonal Gaussian noise. Initialization standardizes feature–response outer products, fits a 12-component PCA on fit-prefix signatures, applies a centered nine-point moving average restricted to that prefix, and clusters with K-means using 20 starts. Twelve Markov EM iterations use ridge penalty 0.01, transition pseudocount 0.01, and variance floor 10^{-5}. A final weighted regression and forward–backward pass supply the initial state distribution q_0. The squared features are suited to even mechanisms, although they do not contain the benchmark transform $\tanh\{1.5(u^2 - 1)\}$. The initializer is an important part of the estimator and its standalone performance is reported separately.

VB-MSM fitting. For each $\lambda \in \{0, 0.001, 0.01\}$, VB-MSM starts from the same random weights and feature-HMM state law. Optimization uses scaled coordinates $B_e = 2\beta_e$ and $v_{er} = 2\alpha_{er}$, with the regime index suppressed. Thus (6) is evaluated as $f_e(u) = \frac{1}{2}[B_e u + \sum_r v_{er} \tanh(w_{er} u + c_{er})]$. The stored B_e and v_{er} are independent $\mathcal{N}(0, 0.05^2)$ draws, so the initial coefficients β_e and α_{er} have standard deviation 0.025. Input weights are $\mathcal{N}(0, 1)$ and hidden biases are uniform on $[-2, 2]$. The intercept $b_j^{(k)}$ is the sum of two stored target offsets, both initialized at zero. Log noise scales also start at zero. The feature-HMM regression weights and noise variances are not transferred.

Eight cycles each take 100 full-prefix Adam steps on (13), with learning rate 0.005 and default Adam moments $(0.9, 0.999)$ in the stored coordinates (B, v, w, c), intercepts, and log noise scales. The first four structural cycles use q_0. After cycle four and each later structural cycle, exact forward–backward on the fit prefixes supplies the next state weights and expected transition counts. The functional penalty always uses the fixed q_0 weights. Every cycle updates Π using the pseudocounts and geometric damping formula below, while π_0 stays fixed. At the end of each cycle, causal filtering evaluates NLL on the checkpoint interval. The lowest-NLL cycle is retained for each λ, and the lowest of those losses selects λ, with ties favoring the smaller value. Selection uses no support, state-label, or test information. The $\lambda = 0$ candidate also supplies the matched unpenalized control. Final-model smoothing on the fit prefixes supplies the weights (11) used for graph reporting. The implementation

evaluates (12) in double precision and subtracts the first modeled history's edge value from every term before accumulating the weighted mean. That shift leaves the centered deviations unchanged, since the same constant enters $\widehat{f}_e^{(k)}$ and $\overline{f}_e^{(k)}$, but it keeps an exactly constant edge function at exactly zero rather than at the difference of two large uncentered moments.

Writing C_{ab} for expected transition counts on the fit prefixes, the transition update is

$$\widehat{\Pi}_{ab} = \frac{C_{ab} + 0.01 + \mathbf{1}\{a = b\}}{\sum_{b'}(C_{ab'} + 0.01 + \mathbf{1}\{a = b'\})}, \qquad \Pi_{ab}^{\text{new}} = \frac{\sqrt{\Pi_{ab}^{\text{old}}\widehat{\Pi}_{ab}}}{\sum_{b'}\sqrt{\Pi_{ab'}^{\text{old}}\widehat{\Pi}_{ab'}}}. \tag{26}$$

The pseudocounts correspond to Dirichlet concentration 1.01 in each entry plus one on the diagonal.

Computational cost. Let $M = \sum_n (T_n^{\text{fit}} - L)$ and $d = pL$. One structural pass costs $O(MKpdH)$ and evaluates Kpd scalar networks. Forward–backward costs $O(MK^2)$. Storing node and pair marginals costs $O(MK + MK^2)$, and a fully vectorized structural pass can store $O(MKpdH)$ intermediate activations. The structural parameter count is $O(KpdH)$. The implementation processes full prefixes, so activation memory, rather than the chain recursion alone, limits scaling. For $p = 8$, $K = 3$, $L = 1$, and $H = 16$, VB-MSM has 9,480 trainable structural parameters, excluding 12 Markov-law parameters. Its eight cycles use 100 structural steps each. Fitting also includes the initializer and checkpoint evaluations. The feature HMM has 624 regression and variance parameters with its nonlinear bank, or 240 with raw features, excluding the Markov law and preprocessing statistics. These operation counts do not substitute for a high-dimensional scaling study.

The 100-system VB-MSM experiments use three NVIDIA RTX 3090 GPUs with 24 GiB each and one worker per GPU. The host has Intel Xeon Gold 5218R processors at 2.10 GHz and 80 logical CPU cores. Feature controls use four CPU workers, with numerical libraries limited to one thread per worker. The environment uses PyTorch 2.1.0 with CUDA 12.1, NumPy 1.26.4, SciPy 1.13.1, and scikit-learn 1.5.2.

The 20-system studies run VB-MSM, iMSM, and FlowMSM with RHINO on the same GPU hardware, using up to three concurrent workers per device. CASTOR uses up to 24 CPU workers, with eight trajectory fits per system running concurrently. Compute budgets differ across methods.

D SYNTHETIC DATA-GENERATING PROCESS

Each independent replication draws a graph, coefficients, regime paths, and innovations, then generates 48 trajectories with $T = 300$, three latent states, and mean dwell time 35. The eight observed variables comprise four drivers and four responses. Each driver follows a stable autoregression with regime-invariant support, providing recurring input without carrying regime information, while each response has two driver parents whose identities change across regimes. The process remains stable throughout.

Let U denote a standardized driver value. The response contribution is the bounded centered-even feature

$$h(u) = \frac{\tanh\{c(u^2 - 1)\} - \mu_c}{s_c}, \qquad c = 1.5, \tag{27}$$

where μ_c and s_c are the mean and standard deviation of $\tanh\{c(V^2 - 1)\}$ for $V \sim \mathcal{N}(0, 1)$, computed by 128-point Gauss–Hermite quadrature. Driver autoregressions have coefficient 0.35 and unit stationary variance. Each response sums its two parent contributions with independently drawn signs and coefficient magnitudes uniform on $[0.6, 0.9]$, plus Gaussian noise with standard deviation 0.25. We discard a 500-step burn-in. The transform is nonconstant and even, so it carries nonlinear conditional-mean information while its population linear projection is zero under the symmetric driver distribution.

The paired shape experiment uses either u, standardized $\tanh(1.5u)$, the centered-even transform above, or a mixture of these three. Nonlinear transforms are centered and scaled by their standard-normal moments using the same quadrature rule. In the mixed condition, the sum of the zero-based regime, lag, source, and target indices modulo three assigns the response-edge mechanism in the order linear, tanh, centered-even. Changing the mechanism preserves each seed's graph, coefficients,

Table 5: Functional graph and state recovery over 100 paired systems per shape. Entries are mean ± standard error. Graph scores and hard calibration use complete edge functions for every estimator. The raw-feature control omits squared and tanh features.

Mechanism	Estimator	AUPRC	Hard F1	NMI	RMSE
Centered-even	Feature HMM	0.9904 ± 0.0033	0.9786 ± 0.0054	0.8292 ± 0.0132	0.8317 ± 0.0022
	VB-MSM, $\lambda = 0$	0.9974 ± 0.0017	0.9867 ± 0.0101	0.9706 ± 0.0081	0.7191 ± 0.0026
	VB-MSM	$\mathbf{0.9975 \pm 0.0016}$	$\mathbf{0.9945 \pm 0.0026}$	$\mathbf{0.9708 \pm 0.0081}$	$\mathbf{0.7171 \pm 0.0026}$
	Feature HMM, raw	0.6800 ± 0.0118	0.0528 ± 0.0155	0.0565 ± 0.0077	0.9651 ± 0.0007
Linear	Feature HMM	0.9977 ± 0.0023	$\mathbf{0.9971 \pm 0.0029}$	0.9929 ± 0.0027	0.7067 ± 0.0008
	VB-MSM, $\lambda = 0$	0.9977 ± 0.0023	0.9969 ± 0.0031	0.9929 ± 0.0027	0.7095 ± 0.0008
	VB-MSM	0.9978 ± 0.0022	0.9969 ± 0.0031	0.9928 ± 0.0027	0.7078 ± 0.0008
	Feature HMM, raw	$\mathbf{1.0000 \pm 0.0000}$	0.9900 ± 0.0100	$\mathbf{0.9957 \pm 0.0003}$	$\mathbf{0.7040 \pm 0.0005}$
Tanh	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9982 \pm 0.0002}$	0.7068 ± 0.0005
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	0.9981 ± 0.0002	0.7083 ± 0.0005
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	0.9981 ± 0.0002	$\mathbf{0.7066 \pm 0.0005}$
	Feature HMM, raw	0.9977 ± 0.0023	0.9975 ± 0.0025	0.9854 ± 0.0036	0.7436 ± 0.0007
Mixed	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	0.9864 ± 0.0009	0.7424 ± 0.0010
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	0.9983 ± 0.0002	0.7099 ± 0.0005
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9984 \pm 0.0002}$	$\mathbf{0.7080 \pm 0.0005}$
	Feature HMM, raw	0.8251 ± 0.0061	0.7771 ± 0.0096	0.9537 ± 0.0069	0.8209 ± 0.0023

regime paths, driver observations, and innovation draws. All fitted methods use the same splits and settings across the four shapes.

Graph metrics score all eight lagged sources against each of the four response targets in every regime, giving 96 candidate edges and 24 true edges. Every method uses this same scoring mask. The driver autoregressive edges stay in the fitted graph and are included in the truth-blind calibration, so the calibration faces the full problem rather than a pre-filtered one. Each regime has eight true driver-to-response edges.

E Complete VB-MSM results

Tables 5 and 6 report all four response shapes on the same 100 paired systems. The raw-feature HMM is a standalone initializer control. It is not followed by neural fitting in this experiment. Across the four shapes, all 1,200 neural penalty candidates and 800 feature-HMM fits completed without failures. The 100-system neural batches, each including three penalty candidates, took 17.0, 20.7, 16.5, and 16.6 minutes for centered-even, linear, tanh, and mixed responses, respectively, using the three GPU workers described in Appendix C. The corresponding full-bank and raw-feature control batches took 3.6, 2.7, 3.6, and 3.3 minutes using four CPU workers. These batch times describe the reported small-system setting. Each system uses one shared neural initialization across penalty values. Variation across optimizer starts on fixed data is not separately measured.

The function penalty has different effects on nonzero and zero contrasts. On centered-even mechanisms it slightly increases nonzero-effect NMAE while reducing spurious null effects, as reported in the main text. The same tradeoff occurs on mixed mechanisms. On tanh mechanisms, penalized and unpenalized nonzero-effect NMAE are nearly equal. The feature HMM has smaller null-effect error there despite its larger nonzero-effect error. A single support or magnitude metric therefore does not capture every estimation tradeoff.

F Nonlinear switching comparators

High-order nonlinear MSM. We use the public neural MSM implementation associated with Balsells-Rodas et al. (2024b), denoted iMSM. Its locally connected conditional-mean networks have two hidden layers of width 16 and softplus activations, following the high-order synthetic study. The model retains full regime covariance matrices, an initial Gaussian distribution, and the native generalized EM updates. Its conditional means allow parent interactions. It has 14,256 trainable parameters, compared with 9,480 structural parameters for VB-MSM.

Table 6: One-step intervention contrasts over 100 paired systems per shape, with mean ± standard error. Effect AP ranks driver-to-response candidates by their mean absolute estimated effect across the two contrasts. NMAE and sign accuracy use nonzero-effect queries. Null-effect MAE is the mean absolute estimate on queries whose true effect is zero, in raw response units. All estimates use complete fitted means without hard-graph pruning.

Mechanism	Estimator	Effect AP	Effect NMAE	Sign accuracy	Null-effect MAE
Centered-even	Feature HMM	0.9940 ± 0.0020	0.4441 ± 0.0056	0.9825 ± 0.0048	0.0560 ± 0.0036
	VB-MSM, $\lambda = 0$	$\mathbf{0.9980 \pm 0.0013}$	$\mathbf{0.0397 \pm 0.0098}$	$\mathbf{0.9875 \pm 0.0043}$	0.0305 ± 0.0053
	VB-MSM	0.9979 ± 0.0014	0.0399 ± 0.0098	$\mathbf{0.9875 \pm 0.0043}$	$\mathbf{0.0296 \pm 0.0052}$
Linear	Feature HMM	0.9969 ± 0.0031	$\mathbf{0.0128 \pm 0.0045}$	0.9983 ± 0.0017	$\mathbf{0.0128 \pm 0.0034}$
	VB-MSM, $\lambda = 0$	$\mathbf{0.9971 \pm 0.0029}$	0.0177 ± 0.0045	$\mathbf{0.9985 \pm 0.0015}$	0.0181 ± 0.0033
	VB-MSM	$\mathbf{0.9971 \pm 0.0029}$	0.0178 ± 0.0045	0.9983 ± 0.0017	0.0172 ± 0.0033
Tanh	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	0.0509 ± 0.0002	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0095 \pm 0.0001}$
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0146 \pm 0.0002}$	$\mathbf{1.0000 \pm 0.0000}$	0.0147 ± 0.0002
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0146 \pm 0.0002}$	$\mathbf{1.0000 \pm 0.0000}$	0.0139 ± 0.0002
Mixed	Feature HMM	$\mathbf{1.0000 \pm 0.0000}$	0.2397 ± 0.0034	$\mathbf{1.0000 \pm 0.0000}$	0.0161 ± 0.0003
	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0125 \pm 0.0001}$	$\mathbf{1.0000 \pm 0.0000}$	0.0151 ± 0.0002
	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	0.0126 ± 0.0002	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0144 \pm 0.0002}$

The 100-system comparison uses three independent random starts. Adam uses learning rate 0.007, no weight penalty, and at most 500 epochs with batch size 1,000. The 48 whole trajectories form one batch, preserving their boundaries. The native training-likelihood stopping rule uses patience four and at most two learning-rate halvings. After each structural update, causal filtering NLL on the checkpoint interval selects a candidate snapshot, preferring the earlier checkpoint within each restart when scores tie. The same criterion chooses among restarts, with ties resolved by the smaller restart index. The native initial state probabilities are uniform. These random starts differ from the feature-HMM state initialization used by VB-MSM. The candidate counts coincide in this 100-system comparison, but the optimization budgets and model classes differ. The selected checkpoint is at epoch 500 in 61 of the 100 centered-even systems and 25 of the 100 mixed systems.

We vary the epoch limit and number of random starts on the first 20 centered-even and first 20 mixed systems. One comparison increases the maximum to 2,000 epochs while retaining the same three starts. All checkpoints remain eligible, including those before epoch 500. The other keeps the 500-epoch limit and uses ten starts, matching the default count in the public FlowMSM iMSM wrapper. Its candidate set contains the same three starts plus seven additional starts. The native stopping rule and checkpoint selection criterion are unchanged. Both comparisons use the same paired VB-MSM reference. They measure sensitivity to training duration and initialization search without establishing convergence to a global optimum.

The 20-system comparison across centered-even, linear, and mixed mechanisms uses three starts and the 2,000-epoch limit throughout. The empirical-innovation comparison uses the same settings with six variables and two regimes, giving 6,312 trainable parameters plus the separately updated Markov law. We also retain the best eligible checkpoint from epochs 1–500 of each plasmode fit as a nested budget control. It requires no additional fit and uses the same checkpoint selection rule. Across the 360 starts in these six conditions, 359 stop under the native rule before epoch 2,000. One linear-system start reaches the limit, and no selected checkpoint is at epoch 2,000.

Graph extraction uses the mean absolute conditional-mean Jacobian from Equation (58) of Balsells-Rodas et al. (2024b), evaluated on fit-prefix histories. Each history is grouped by the smoothed marginal-argmax regime at its target time. The hard graph retains scores strictly above the published threshold 0.05. All Kp^2 lagged candidates are scored before applying the response-target evaluation mask. Primary regime and graph results use the common exact Viterbi label alignment. Marginal-argmax results are retained separately. Forecasts use causal filtering. Intervention evaluation on centered-even and mixed mechanisms uses the same 96 raw-scale contrasts as VB-MSM. Because iMSM permits interactions, its contrasts refer to the specified zero background for the other history inputs.

On the first 20 centered-even systems, increasing the iMSM epoch limit from 500 to 2,000 reduces effect NMAE by 0.0196 ± 0.0029 and test NLL by 0.0659 ± 0.0103. AUPRC is already one at 500 epochs and remains one. Hard F1 decreases by 0.0505 ± 0.0110, and RMSE increases by

Table 7: iMSM budget comparisons on the same first 20 systems per mechanism, with mean ± standard error. Numbers after iMSM give random starts × maximum epochs per start. Every condition uses the same paired VB-MSM reference. Higher is better for AUPRC, F1, and NMI. Lower is better for RMSE and nonzero-effect NMAE.

Shape	Estimator	AUPRC	Hard F1	NMI	RMSE	Effect NMAE
Centered-even	VB-MSM	0.9950 ± 0.0038	$\mathbf{0.9890 \pm 0.0077}$	0.9676 ± 0.0201	$\mathbf{0.7177 \pm 0.0052}$	0.0532 ± 0.0311
Centered-even	iMSM, 3×500	$\mathbf{1.0000 \pm 0.0000}$	0.8007 ± 0.0178	$\mathbf{0.9923 \pm 0.0012}$	0.7385 ± 0.0014	0.0652 ± 0.0042
Centered-even	iMSM, $3 \times 2{,}000$	$\mathbf{1.0000 \pm 0.0000}$	0.7502 ± 0.0164	0.9921 ± 0.0011	0.7430 ± 0.0017	$\mathbf{0.0456 \pm 0.0030}$
Centered-even	iMSM, 10×500	$\mathbf{1.0000 \pm 0.0000}$	0.8311 ± 0.0157	$\mathbf{0.9923 \pm 0.0012}$	0.7372 ± 0.0014	0.0601 ± 0.0027
Mixed	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9980 \pm 0.0004}$	$\mathbf{0.7086 \pm 0.0011}$	$\mathbf{0.0130 \pm 0.0003}$
Mixed	iMSM, 3×500	$\mathbf{1.0000 \pm 0.0000}$	0.9915 ± 0.0035	0.9971 ± 0.0005	0.7201 ± 0.0015	0.0336 ± 0.0009
Mixed	iMSM, $3 \times 2{,}000$	$\mathbf{1.0000 \pm 0.0000}$	0.9925 ± 0.0035	0.9973 ± 0.0005	0.7208 ± 0.0015	0.0324 ± 0.0011
Mixed	iMSM, 10×500	$\mathbf{1.0000 \pm 0.0000}$	0.9952 ± 0.0027	0.9972 ± 0.0005	0.7194 ± 0.0014	0.0328 ± 0.0011

Table 8: Nonlinear comparison over 100 paired systems per mechanism, with mean ± standard error. iMSM uses three starts and a 500-epoch limit. AUPRC and hard F1 concern response edges. RMSE uses causal filtering.

Condition	Estimator	AUPRC	Hard F1	NMI	RMSE
Centered-even	VB-MSM	$\mathbf{0.9975 \pm 0.0016}$	$\mathbf{0.9945 \pm 0.0026}$	$\mathbf{0.9708 \pm 0.0081}$	$\mathbf{0.7171 \pm 0.0026}$
Centered-even	iMSM	0.9781 ± 0.0069	0.7746 ± 0.0115	0.9488 ± 0.0119	0.7466 ± 0.0025
Mixed	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9984 \pm 0.0002}$	$\mathbf{0.7080 \pm 0.0005}$
Mixed	iMSM	$\mathbf{1.0000 \pm 0.0000}$	0.9920 ± 0.0017	0.9975 ± 0.0002	0.7199 ± 0.0007

0.0045 ± 0.0008. The higher training budget therefore improves density and effect estimation without improving every endpoint. All 120 candidates in the centered-even and mixed budget comparison stop under the native training-likelihood rule before epoch 2,000.

On this centered-even cohort, VB-MSM has effect NMAE 0.0532 ± 0.0311, compared with 0.0456 ± 0.0030 for the longer iMSM fit. The paired difference is 0.0076 ± 0.0298, despite lower VB-MSM error in 18 of the 20 systems. Its paired hard-F1 advantage is 0.2388 ± 0.0124, while its AUPRC is lower in two systems and tied in 18. The longer iMSM fit has recall one in every system, so false positives under its fixed threshold account for its hard-F1 errors. These results distinguish hard-graph reporting from the support information in derivative rankings. On mixed mechanisms, the longer iMSM fit and VB-MSM both have perfect AUPRC and almost identical mean NMI. VB-MSM has lower RMSE and effect NMAE in all 20 systems, with paired reductions 0.0122 ± 0.0008 and 0.0194 ± 0.0011.

Increasing from three to ten starts at 500 epochs selects an additional start in 13 of the 20 centered-even systems and 16 of the 20 mixed systems. On centered-even mechanisms, it increases iMSM hard F1 by 0.0304 ± 0.0159 and reduces effect NMAE by 0.0051 ± 0.0024. AUPRC remains one, and mean NMI changes by 0.0000 ± 0.0005. Against this ten-start fit, VB-MSM has higher hard F1 in all 20 systems and lower RMSE in 18, with paired differences 0.1579 ± 0.0155 and -0.0195 ± 0.0056. On mixed mechanisms, its RMSE and effect NMAE are lower in every system. The selected ten-start checkpoint remains at epoch 500 in 15 centered-even and four mixed systems. Table 7 reports both budget checks, which vary one factor at a time.

Results over 100 systems. Table 8 compares VB-MSM with iMSM on 100 systems per mechanism, using three starts and a 500-epoch limit for iMSM. On centered-even systems, VB-MSM exceeds iMSM by 0.0194 ± 0.0068 AUPRC and 0.2199 ± 0.0114 hard F1. It reduces RMSE by 0.0295 ± 0.0035 and effect NMAE by 0.0705 ± 0.0140. The NMI difference is smaller, 0.0220 ± 0.0139. On mixed systems, both methods rank support perfectly, while VB-MSM has lower RMSE and effect error in every system. Architectures, initialization, and graph thresholds differ, so these comparisons do not isolate the functional penalty.

FlowMSM with RHINO. The nonlinear two-stage comparison combines FlowMSM with the public RHINO backend (Hulsman et al., 2026; Gong et al., 2023) on 20 systems per condition for centered-even, linear, and mixed mechanisms, and for the three empirical-innovation conditions in Appendix J. FlowMSM uses ten random restarts, at most 500 epochs per restart, batch size 1,000, learning rate 0.007, weight decay 10^{-4}, gradient clipping at 3, a 20-epoch warmup, and early-stopping patience ten. Checkpoint NLL selects the fitted model and restart. The initial neural spline flow has one transform with four bins and hidden width 8. Its transition flow has two transforms with eight bins

and hidden width 32. Both use one hidden layer and GELU activations. FlowMSM has 14,732 parameters for the eight-variable, three-regime synthetic systems and 11,052 for the six-variable, two-regime plasmode.

Our adapter uses the same row-stochastic transition convention in the forward and backward recursions. With $\Pi_{ab} = P(Z_t = b \mid Z_{t-1} = a)$, the prediction step is $\widetilde{\alpha}_t(b) = \sum_a \alpha_{t-1}(a)\Pi_{ab}$. Exact enumeration of asymmetric three-state chains checks likelihoods, smoothed marginals, and pair marginals under this convention. We use the released flow architecture, loss, and optimization settings. The native checkpoint likelihood includes the single preceding observation as an initial emission, while RHINO uses it only as lagged context. Fit and checkpoint state assignments use native marginal decoding separately on their own observation blocks. Neither uses later test observations. Final state NMI and graph alignment use full-trajectory Viterbi decoding with fixed fitted parameters.

For each estimated regime, RHINO receives two-observation fragments containing a target and its original lag-one history. Fragments never cross trajectory boundaries. Every fit-prefix target is included once, and the chronological checkpoint interval supplies a separate validation set in place of the public wrapper's internal 80/20 split. Instantaneous edges are disabled for the common lagged search space. The native flow likelihood, nonlinear causal decoder, spline hypernetwork, and optimization objective are retained. Each regime uses at most 500 epochs, batch size 1,000, sparsity factor 5, graph temperature 0.25, and the native mixed precision. The native held-out objective, checked every ten epochs, selects the checkpoint. Each regime model has 27,763 parameters with eight variables and 26,398 with six variables.

RHINO posterior edge probabilities supply AUPRC, and probabilities at least 0.5 define its hard graph. The FlowMSM Viterbi path aligns the regime-specific graphs with the truth only during evaluation. Every condition uses the same 20 systems as its VB-MSM and iMSM references. This comparison does not equate architectures or fitting budgets, and it does not evaluate RHINO intervention magnitudes or forecasting.

On centered-even systems, VB-MSM has paired AUPRC and hard F1 gains of 0.0215 ± 0.0063 and 0.0672 ± 0.0113 over FlowMSM with RHINO. Hard F1 is higher in 17 systems, equal in two, and lower in one. The mean NMI difference is -0.0257 ± 0.0202. On linear and mixed systems, the hard F1 gains are 0.1112 ± 0.0121 and 0.0858 ± 0.0105, respectively, and are positive in every system. AUPRC ties in 13 linear and 19 mixed systems. These results distinguish edge ranking from hard selection under each method's graph rule.

CASTOR. We use the nonlinear CASTOR implementation (Rahmani & Frossard, 2025) distributed with FlowMSM (Hulsman et al., 2026) on the same 20 synthetic systems per mechanism. Its native interface fits separate mechanism and regime parameters for each of the 48 complete trajectories. The last 35 percent of each trajectory supplies evaluation labels, although CASTOR has observed these data during fitting. All estimators use the same pooled fit-prefix standardizer. The other estimators also pool fit prefixes when estimating their mechanism and regime parameters. CASTOR's results therefore describe a separate transductive procedure with different access to observations and different parameter sharing.

CASTOR starts with three temporal segments, using a window of 100 observations, and re-estimates assignments during EM. CASTOR removes unoccupied regimes. Each fit takes six outer EM iterations with width-10 sigmoid mechanism networks and $\lambda_1 = \lambda_2 = 0.01$. In each cycle, the time-weight model takes 200 initial updates and up to six further batches of 100 updates while its loss is at least 0.7, using the bounded loop supplied with FlowMSM. The inner augmented-Lagrangian solver uses at most 200 iterations, acyclicity tolerance 10^{-8}, and penalty cap 10^{16} with the native L-BFGS-B settings. Six outer iterations follow the released nonlinear example and do not establish convergence. The model retains its instantaneous edges and exclusion of lagged self-edges during fitting. Evaluation scores only the common lagged response-edge mask.

Each regime network stores 3,048 parameters, including 320 weights fixed at zero by the diagonal bounds. The initial three-regime time-weight model has six parameters. The lag input at time t is X_{t-1}, with a zero-padded first row. Unthresholded input-layer group norms supply continuous lag scores, and the native rule retains scores strictly above 0.3. Each surviving state retains its corresponding graph. We align each trajectory's estimated regimes using its test-suffix labels. Unmatched true regimes receive zero scores and empty graphs. NMI is computed within each trajectory because the fitted state labels are trajectory-specific. Each system metric averages all 48

Table 9: CASTOR through its native per-trajectory interface. Each fit observes the complete trajectory. Graph alignment and NMI use the final 35 percent. Entries are means ± standard errors over 20 systems per mechanism, after averaging the 48 separate fits within each system. Other methods pool fitting prefixes, so this comparison does not equate parameter sharing or access to observations.

Condition	Estimator	AUPRC	Hard F1	NMI
Centered-even	CASTOR	0.2631 ± 0.0010	0.0232 ± 0.0017	0.0672 ± 0.0078
Linear	CASTOR	0.5755 ± 0.0062	0.4528 ± 0.0079	0.2286 ± 0.0150
Mixed	CASTOR	0.4893 ± 0.0052	0.3388 ± 0.0069	0.1386 ± 0.0121

Table 10: Secondary CASTOR evaluation using complete trajectories for label alignment and NMI, with the fitted models and graph scores from Table 9. Entries are means ± standard errors across 20 systems per mechanism. Full-trajectory NMI uses a different observation interval from test-suffix NMI.

Condition	Estimator	Full-alignment AUPRC	Full-alignment F1	Full-trajectory NMI
Centered-even	CASTOR	0.2639 ± 0.0010	0.0237 ± 0.0018	0.0573 ± 0.0053
Linear	CASTOR	0.5978 ± 0.0057	0.4769 ± 0.0077	0.2753 ± 0.0156
Mixed	CASTOR	0.5060 ± 0.0057	0.3556 ± 0.0077	0.1541 ± 0.0100

trajectories, and standard errors use the 20 independent systems. We report neither forecasting nor intervention metrics for this transductive interface.

A short suffix may omit regimes, which can leave the graph-label alignment ambiguous. As a secondary transductive diagnostic, we also align graphs using each complete trajectory and report full-trajectory NMI separately. Graph scores, thresholds, and masks stay fixed. We count maximizing label assignments for both intervals to record any ambiguity that remains. Full-trajectory NMI concerns a different observation interval from the other methods' test-suffix NMI.

Under this protocol, CASTOR has mean response-edge F1 of 0.0232, 0.4528, and 0.3388 on centered-even, linear, and mixed systems, respectively (Table 9). All 2,880 trajectory fits return results, but 2,505 retain fewer than three regimes. All estimated lag graphs are empty in 446 centered-even trajectory fits. Of 47,000 inner mechanism fits, 41,250 meet the acyclicity tolerance and 5,750 stop at the penalty cap. These results describe the specified fitting schedule on individual recurrent trajectories, with different parameter sharing and observation access from the pooled estimators.

Complete-trajectory alignment reduces the number of fits with multiple optimal label assignments from 1,283 to 220. Mean hard F1 increases by 0.0005 ± 0.0005, 0.0240 ± 0.0026, and 0.0168 ± 0.0021 for the three mechanisms, respectively. The alignment interval therefore affects the reported graph-recovery metrics, although recovery remains limited (Table 10). Full-trajectory NMI is 0.0573, 0.2753, and 0.1541, respectively. These NMI values evaluate a longer interval and do not measure an effect of label alignment.

In the 100-system comparison, the centered-even effect NMAE is 0.0399 ± 0.0098 for VB-MSM and 0.1104 ± 0.0121 for iMSM. On mixed mechanisms it is 0.0126 ± 0.0002 and 0.0334 ± 0.0005, respectively. The paired NLL reductions are 1.3257 ± 0.0665 and 0.5109 ± 0.0107.

Table 11: Additional metrics on the 100 paired systems per mechanism in Table 8, with mean ± standard error. NLL uses causal filtering. Effect NMAE evaluates nonzero raw-scale intervention contrasts. Null-effect MAE is the mean absolute estimate for zero-effect queries. Lower is better throughout. Dashes denote unevaluated metrics.

Condition	Estimator	NLL	Effect NMAE	Null-effect MAE
Centered-even	VB-MSM	$\mathbf{5.6469 \pm 0.0574}$	$\mathbf{0.0399 \pm 0.0098}$	$\mathbf{0.0296 \pm 0.0052}$
Centered-even	iMSM	6.9726 ± 0.0405	0.1104 ± 0.0121	0.0738 ± 0.0104
Mixed	VB-MSM	$\mathbf{5.4404 \pm 0.0113}$	$\mathbf{0.0126 \pm 0.0002}$	$\mathbf{0.0144 \pm 0.0002}$
Mixed	iMSM	5.9513 ± 0.0138	0.0334 ± 0.0005	0.0219 ± 0.0003

G REDUNDANT STATES AND THE REGIME COUNT

Negative log likelihood scores the observed predictive distribution, whereas regime count describes the latent representation. They align only when model-class regularity excludes redundant representations. This is a standard issue in hidden-state models, whose identification results impose restrictions such as linearly independent emission distributions (Gassiat et al., 2016). We give the elementary duplicated-kernel construction to specify the boundary of our fixed-K analysis.

Proposition 3 (Duplicated regimes preserve the observed law)**.** *Let a one-state temporal model have conditional density $g(x_t \mid h_t)$. For every integer $R \geq 2$, every initial distribution on R states, and every R-state transition matrix, there is an R-state switching model with exactly the same observed distribution, obtained by assigning the same g to every hidden state. The unrestricted observed-data likelihood therefore does not identify the regime count.*

Proof. For any observed trajectory and its initial history, the duplicated model satisfies

$$p_R(x_{L+1:T} \mid x_{1:L}) = \sum_{z_{L+1:T}} p(z_{L+1:T}) \prod_{t=L+1}^{T} g(x_t \mid h_t) \tag{28}$$

$$= \prod_{t=L+1}^{T} g(x_t \mid h_t) \sum_{z_{L+1:T}} p(z_{L+1:T}) \tag{29}$$

$$= p_1(x_{L+1:T} \mid x_{1:L}). \tag{30}$$

The second line follows because the emission factor is independent of z, so it leaves the sum, and the third because the state path distribution sums to one. The equality holds trajectory by trajectory, not merely in expectation, so the population likelihood cannot separate the one-state representation from any duplicated-state representation. □

The model therefore conditions on a specified K. This classical redundant-state construction concerns the unrestricted representation, not the minimal number of distinct regimes. Distinct-component restrictions can identify the regime count (Balsells-Rodas et al., 2024b; 2026a). The duplicated-kernel example does not satisfy those restrictions. Selection of a minimal count is separate from the fixed-K support question.

H VARIANCE AGAINST DERIVATIVE READOUTS

This appendix reports the readout ablation summarized in Section 4.5. Only the score changes. Every fitted mechanism, smoothing weight, regime alignment, preprocessing step, and calibration setting is the one used elsewhere in the paper, and no parameter is refitted. The comparison therefore isolates the readout, unlike a comparison against a separate method, which would also vary the architecture, initialization, and fitting budget.

Scores and protocol. Alongside $\widehat{V}_e^{(k)}$ of (12) we compute

$$\widehat{D}_e^{(k)} = \sum_{n,t} \widetilde{\gamma}_{ntk} \left| \widehat{f}_e^{(k)\prime}(u_{nte}) \right|, \tag{31}$$

where $\widetilde{\gamma}$ denotes the saved smoothing weights normalized within each regime. The derivative is taken analytically from the complete fitted edge function, including its linear term, so cancellation between the linear and tanh paths is retained exactly as in the variance score. Both readouts use soft weights. This is an ablation of our own reporting interface and is not a rerun of iMSM, whose mechanism class and native graph rule differ. The evaluation covers 919 distinct saved checkpoints across 460 system-condition datasets, comprising the four response shapes with 100 systems each and the three empirical-innovation conditions with 20 systems each, and includes both the checkpoint-selected fit and its matched zero-penalty control. Conditions share paired seeds and structural draws, so uncertainty is computed within each condition from its independent systems and the conditions are not pooled as independent replications.

Table 12: Variance and mean absolute derivative readouts on identical fitted models, for checkpoint-selected fits, with mean ± standard error over systems. The response mask scores driver-to-response candidates, and the all-candidate mask additionally scores every fitted lagged entry. Both readouts pass through the rule of Section 3.6.

Condition	Response AUPRC		Response hard F1		All-candidate AUPRC	
	$\widehat{V}$	$\widehat{D}$	$\widehat{V}$	$\widehat{D}$	$\widehat{V}$	$\widehat{D}$
Centered-even	0.9975 ± 0.0016	0.9976 ± 0.0018	0.9945 ± 0.0026	0.0860 ± 0.0275	0.9966 ± 0.0018	0.9931 ± 0.0025
Linear	0.9978 ± 0.0022	0.9978 ± 0.0022	0.9969 ± 0.0031	0.8671 ± 0.0338	0.9978 ± 0.0022	0.9977 ± 0.0023
Tanh	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.8498 ± 0.0359	1.0000 ± 0.0000	1.0000 ± 0.0000
Mixed	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.3135 ± 0.0460	1.0000 ± 0.0000	0.9991 ± 0.0002
Native innovations	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9762 ± 0.0038
Half-length	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9614 ± 0.0056
Five-percent shocks	0.8763 ± 0.0453	0.8506 ± 0.0444	0.1011 ± 0.0464	0.1011 ± 0.0464	0.8721 ± 0.0255	0.7251 ± 0.0237

Table 13: Calibration diagnostics for the same fits. The middle columns count systems in which the mixture rule of Section 3.6 rejects two components and therefore reports an empty graph. The right columns apply the fixed cutoff $\widehat{D} > 0.05$ to the soft-weighted derivative instead, with mean ± standard error.

Condition	Systems	Two components rejected		Fixed cutoff $\widehat{D} > 0.05$	
		$\widehat{V}$	$\widehat{D}$	AUPRC	Hard F1
Centered-even	100	0	91	0.9976 ± 0.0018	0.8814 ± 0.0087
Linear	100	0	13	0.9978 ± 0.0022	0.9811 ± 0.0041
Tanh	100	0	15	1.0000 ± 0.0000	0.9783 ± 0.0021
Mixed	100	0	68	1.0000 ± 0.0000	0.9497 ± 0.0031
Native innovations	20	0	20	1.0000 ± 0.0000	0.8765 ± 0.0140
Half-length	20	0	20	1.0000 ± 0.0000	0.6988 ± 0.0144
Five-percent shocks	20	16	16	0.8506 ± 0.0444	0.5289 ± 0.0063

Ranking against hard reporting. Table 12 separates two questions that the headline F1 gap conflates. On the response mask the two readouts rank edges almost identically. The paired centered-even AUPRC difference is -0.0001 ± 0.0004, equal in 96 of 100 systems, and both readouts are exactly one on tanh and mixed mechanisms and under native and half-length innovations. Their hard graphs nevertheless differ sharply. The cause is the acceptance rule rather than the ranking. Table 13 shows that the derivative scores fail the BIC margin of Section 3.6 in 91 of 100 centered-even systems and in every empirical-innovation system at native and half length, returning an empty graph, whereas the variance scores accept two components in all of those fits. For selected centered-even fits the median BIC improvement is about 100.76 for variance and -1.64 for the derivative.

A fixed cutoff removes most of the gap. Applying iMSM's published cutoff 0.05 to the soft-weighted derivative raises centered-even F1 from 0.0860 to 0.8814 while leaving AUPRC unchanged at 0.9976. The primary F1 gap is therefore a property of the score and the calibrator together, not evidence that derivative scores cannot detect these edges. We report it as such. The cutoff is transferred from a different procedure and is not calibrated for these fitted models, and no threshold was swept or chosen against the truth. Squaring the derivative changes no selected edge in any of the 919 checkpoints, so the gap is also not an artifact of comparing a variance with an unsquared magnitude. The mean squared derivative $\widehat{J}_e^{(k)} = \sum_{n,t} \widetilde{\gamma}_{ntk} \widehat{f}_e^{(k)\prime}(u_{nte})^2$ likewise leaves it in place, with mean response F1 of 0.1100, 0.5070, 0.6196, and 0.2073 on the four response shapes.

All-candidate ranking. Scoring every lagged candidate rather than response edges alone does separate the rankings, since the response mask is near ceiling for both readouts. The paired all-candidate AUPRC gain for variance is 0.0035 ± 0.0010 on centered-even mechanisms and 0.0009 ± 0.0002 on mixed mechanisms, and 0.0238 ± 0.0038, 0.0386 ± 0.0056, and 0.1470 ± 0.0151 under native, half-length, and shock innovations. The shock gain is positive in all 20 systems. These gains concern the expanded target set. They do not alter the shock-condition conclusion of Section 4.5, where both readouts have hard F1 0.1011 and the variance calibrator also rejects two components in 16 of 20 systems. Variance scoring does not repair that failure. In the other direction, variance

has all-candidate AUPRC one but hard F1 0.9231 under native and half-length innovations, so its calibrator can select false edges despite a perfect ranking.

Numerical checks. The evaluation reconstructs the archived variance scores from each checkpoint and matches them exactly in double precision, and every reconstructed hard graph matches its archived graph. Central finite differences verify the analytic derivatives on five input histories for every regime, target, and candidate source in every checkpoint, with maximum discrepancy about 7.94×10^{-9}. An independent implementation of average precision, precision, recall, F1, and structural Hamming distance reproduces 27,570 macro metric values from the saved arrays, with maximum discrepancy 2.22×10^{-16}. These checks confirm the numerical calculation and its correspondence to the saved results. They do not establish statistical validity of the mixture rule.

I NONLINEAR INITIALIZER CONTROLS

The initializer control fits a switching regression on $(h, h^{\odot 2}, \tanh h)$. Its raw-only ablation removes squared and tanh features together. Both use the same fitting prefixes, K, ridge penalty, and EM settings. Contributions from the linear, squared, and tanh terms for one source are combined before taking their variance. Neither scoring nor calibration uses the response mask. The mask and true labels enter only the final performance calculation. The 200 centered-even feature-control fits took 214 seconds using four CPU workers.

J EMPIRICAL-INNOVATION PLASMODE DETAILS

We evaluate a known structural graph under empirical innovations from monthly ENSO and rainfall records. The process has six variables, two regimes, lag order one, and 48 trajectories per system. Three autoregressive drivers feed three responses, each with one parent per regime. The mean dwell is 35, and a 500-step burn-in is discarded. Native trajectories have length 300. A half-length condition uses 150 observations, and a shock condition independently multiplies each innovation vector by five with probability 0.05. Methods are paired within each condition. Regime paths and innovation draws are not held fixed across the length and shock conditions.

The theoretical mean-independence and positive-density conditions are not established for this block-resampled process, so the experiment assesses recovery of a prescribed structural graph under empirical noise. The common support mask contains 18 driver-to-response candidates across the two regimes, with six true edges. Preprocessing, fitting, and checkpoint selection use the chronological blocks specified in the main protocol.

Residual reservoir. We join monthly HadISST Niño3.4 sea-surface-temperature anomalies (Rayner et al., 2003) to the All-India rainfall series in IITM Research Report 138 (Kothawale & Rajeevan, 2017), obtaining 1,752 complete months from January 1871 through December 2016. The report stores rainfall in tenths of millimetres. Our preparation script converts units and verifies the published annual checksums. We standardize the two series, fit a pooled VAR(1) with an intercept, and standardize its paired residuals. Independent draws of 12-month blocks from this reservoir provide innovations for each variable pair in the plasmode. Replications share the prescribed graph and empirical residual reservoir, while independently generated regime paths and innovation-block draws vary across replications. All model standardization and fitting use the generated trajectories' parameter-fitting prefixes.

The rainfall parser isolates Table 1 of IITM RR-138, requires exactly 17 numerical entries per year (12 months and five summaries), verifies every annual total within the report's independent rounding tolerance, and requires all years 1871–2016 with 1,752 time stamps, with no missing values, no smoothing, and no imputation.

J.1 VB-MSM AND NONLINEAR COMPARATORS

The comparison uses 20 systems per condition. VB-MSM uses the width-16 additive decoder, feature-HMM initialization, eight cycles of 100 structural updates, four anchored cycles, and three penalty values from Section 3.5. It has 3,564 structural parameters and six entries in its initial distribution and transition matrix. The matched zero-penalty control shares its initialization and

Table 14: Empirical-innovation comparison, with mean ± standard error over 20 paired systems per condition. AUPRC and hard F1 use the response-edge mask, and NMI uses the test-suffix Viterbi states. RMSE and NLL use causal filtering. Dashes denote unevaluated metrics.

Condition	Estimator	AUPRC	Hard F1	NMI	RMSE	NLL
Native	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9961 \pm 0.0007}$	$\mathbf{0.6521 \pm 0.0011}$	$\mathbf{4.3585 \pm 0.0072}$
Native	iMSM	$\mathbf{1.0000 \pm 0.0000}$	0.9964 ± 0.0036	0.9951 ± 0.0008	0.6522 ± 0.0011	4.5133 ± 0.0120
Native	FlowMSM + RHINO	$\mathbf{1.0000 \pm 0.0000}$	0.9518 ± 0.0130	0.9938 ± 0.0007	—	—
Half length	VB-MSM	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9960 \pm 0.0010}$	$\mathbf{0.6619 \pm 0.0018}$	$\mathbf{4.4509 \pm 0.0121}$
Half length	iMSM	$\mathbf{1.0000 \pm 0.0000}$	0.8682 ± 0.0142	0.9941 ± 0.0012	0.6711 ± 0.0020	4.7778 ± 0.0193
Half length	FlowMSM + RHINO	$\mathbf{1.0000 \pm 0.0000}$	0.9509 ± 0.0131	0.9911 ± 0.0013	—	—
5% shocks	VB-MSM	0.8763 ± 0.0453	0.1011 ± 0.0464	0.0139 ± 0.0042	0.8261 ± 0.0049	6.3601 ± 0.0256
5% shocks	iMSM	0.9846 ± 0.0154	0.8058 ± 0.0235	0.8147 ± 0.0429	$\mathbf{0.7557 \pm 0.0055}$	$\mathbf{5.5244 \pm 0.0462}$
5% shocks	FlowMSM + RHINO	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9563 \pm 0.0120}$	$\mathbf{0.9666 \pm 0.0023}$	—	—

Table 15: Controls on the same 20 empirical-innovation systems per condition as Table 14, with mean ± standard error. The zero-penalty model shares the VB-MSM initialization and fitting schedule. The nested iMSM control selects among eligible checkpoints through epoch 500 of the same three starts used by the 2,000-epoch fit.

Condition	Estimator	AUPRC	Hard F1	NMI	RMSE	NLL
Native	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9962 \pm 0.0007}$	$\mathbf{0.6528 \pm 0.0011}$	$\mathbf{4.3640 \pm 0.0072}$
Native	iMSM, eligible epochs ≤ 500	$\mathbf{1.0000 \pm 0.0000}$	0.9929 ± 0.0049	0.9943 ± 0.0008	0.6535 ± 0.0011	4.6248 ± 0.0137
Half length	VB-MSM, $\lambda = 0$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9957 \pm 0.0010}$	$\mathbf{0.6631 \pm 0.0018}$	$\mathbf{4.4612 \pm 0.0120}$
Half length	iMSM, eligible epochs ≤ 500	$\mathbf{1.0000 \pm 0.0000}$	0.9164 ± 0.0192	0.9953 ± 0.0010	0.6676 ± 0.0019	4.8168 ± 0.0181
5% shocks	VB-MSM, $\lambda = 0$	0.8817 ± 0.0428	0.1125 ± 0.0518	0.0137 ± 0.0042	0.8267 ± 0.0049	6.4161 ± 0.0521
5% shocks	iMSM, eligible epochs ≤ 500	$\mathbf{0.9846 \pm 0.0154}$	$\mathbf{0.8329 \pm 0.0259}$	$\mathbf{0.8143 \pm 0.0428}$	$\mathbf{0.7513 \pm 0.0058}$	$\mathbf{5.5651 \pm 0.0447}$

fitting schedule. iMSM uses three starts and a 2,000-epoch limit. FlowMSM with RHINO uses the settings in Appendix F. Each method retains its own graph score and hard threshold. Ground-truth states and graphs enter only evaluation.

Under native residuals, the paired VB-MSM minus iMSM RMSE difference is -0.00014 ± 0.00045. VB-MSM has lower test NLL in all 20 systems, with mean reduction 0.1548 ± 0.0104. For half-length records, it has higher hard F1 and lower RMSE in 19 systems each. Under gross shocks, iMSM has higher hard F1 and lower RMSE and NLL in every system, and higher NMI in 19. The paired iMSM gains are 0.7047 ± 0.0615 in hard F1 and 0.8009 ± 0.0427 in NMI.

FlowMSM with RHINO ranks response support perfectly in all 60 empirical-innovation systems. Its native-length and half-length hard F1 are 0.9518 ± 0.0130 and 0.9509 ± 0.0131, respectively. Under shocks, its paired gains over VB-MSM are 0.8551 ± 0.0449 in hard F1 and 0.9527 ± 0.0055 in NMI, with both metrics higher in all 20 systems.

The matched zero-penalty fit also recovers the response graph perfectly in every native and half-length system. Functional regularization reduces mean RMSE by 0.00072 ± 0.00003 and 0.00123 ± 0.00007, respectively. Under gross shocks, the zero-penalty model has mean hard F1 0.1125 and NMI 0.0137. Regularization reduces mean shock-condition RMSE by 0.00056 ± 0.00016, without repairing graph and regime recovery.

iMSM's advantage under shocks is also present among its eligible checkpoints through epoch 500. Their mean hard F1, NMI, and RMSE are 0.8329, 0.8143, and 0.7513, respectively. Extending checkpoint eligibility to 2,000 epochs improves mean test NLL in all three conditions. Mean hard F1 and RMSE nevertheless worsen for half-length and shock records, while AUPRC is unchanged. The different behavior of these metrics persists even when the data, starts, and fitting algorithm are held fixed.