End-to-End Identifiable Recurrent Switching Dynamical Systems

Carles Balsells-Rodas **Zhengrui Xiang** **Xavier Sumba** **Yingzhen Li**
Imperial College London
{cb221, yingzhen.li}@imperial.ac.uk

Abstract

Learning identifiable representations in deep generative models remains a fundamental challenge, particularly for sequential data with regime-switching dynamics. Existing approaches establish identifiability under restrictive assumptions, such as stationarity or limited emission models, and typically rely on variational autoencoder (VAE) estimators, which introduce approximation gaps that limit the recovery of the latent structure. In this work, we address both the theoretical and practical limitations of this setting. First, we establish identifiability of a broad class of recurrent nonlinear switching dynamical systems under flexible assumptions, significantly extending prior results. Second, we introduce ΩSDS, a flow-based estimator that enables exact likelihood optimisation using expectation-maximisation. Through empirical validation on both synthetic and real-world data, our results demonstrate that ΩSDS achieves improved disentanglement compared to VAE-based estimators and more accurate forecasting of underlying dynamics.

1 Introduction

Many real-world time series exhibit nonstationary dynamics, where the underlying temporal mechanisms change over time. Examples include seasonal patterns in climate data [Saggioro et al., 2020], or behavioural modes in human motion [Tsuchida et al., 2019]. Regime-switching state-space models, including switching dynamical systems (SDSs; [Ghahramani and Hinton, 2000]) and Markov switching models (MSMs; [Lindgren, 1978, Poritz, 1982]), target these scenarios by introducing discrete regimes (or switches) which explicitly model dynamical changes. Recent sequential latent variable models incorporate regime-switching as structured priors [Linderman et al., 2017, Dong et al., 2020, Ansari et al., 2021, Liu et al., 2023, Geadah et al., 2024], improving flexibility while preserving interpretable continuous states and discrete switches.

Interpretability of deep generative models relies on identifiability, which ensures that the data distribution determines the latent structure up to acceptable equivalences, such as permutations of switches or latent variables. Without identifiability, observationally equivalent model solutions may represent contradictory latent dynamics, and therefore interpretability may reflect arbitrary parametrisations rather than the underlying data-generating mechanisms. For flexible nonlinear latent variable models, identifiability fails without additional structure [Hyvärinen and Pajunen, 1999]. Existing results impose such structures either on the latent distribution, for example through auxiliary variables or temporal dependence [Khemakhem et al., 2020, Hälvä and Hyvarinen, 2020, Klindt et al., 2021, Morioka et al., 2021, Yao et al., 2022], or the emission map [Kivva et al., 2022]. Switching dynamical systems introduce additional complexity, since the latent prior is a mixture over regime-dependent dynamics. Existing SDS identifiability results reflect this "no free lunch" principle between restricting emissions or latent dynamics: piecewise linear emissions allow flexible nonlinear mixture transitions [Balsells-Rodas et al., 2024], while general invertible emissions require more restrictive latent dynamics, such as stationary linear transitions [Hälvä et al., 2021].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison with identifiable regime-switching latent variable models. Full ident. excludes identifying K and denotes recovery of the latent representation, transitions, and switching process.

Method	Emission	Switching	Nonlin. dyn.	Identifies K	Full ident.	Exact LL
HMM-ICA [Hälvä and Hyvarinen, 2020]	Bijective	Autonomous	✗	✗	✗	✓
SNICA [Hälvä et al., 2021]	Invertible	Autonomous	✗	✗	✗	✗
NCTRL [Song et al., 2023]	Gaussian / obs.	Autonomous	✓	✓	✓	✗
iSDS [Balsells-Rodas et al., 2024]	PWL	Autonomous	✓	✓	✓	✗
CtrlNS [Song et al., 2024]	Invertible	Obs. dependent	✓	✗	✓	✗
SDCI [Balsells-Rodas et al., 2025]	N/A	Recurrent	✓	✓	✗	N/A
ΩSDS (ours)	Invertible	Recurrent	✓	✓	✓	✓

In this work, we push the above trade-off by establishing identifiability and estimation for recurrent switching dynamical systems (rSDS) with nonlinear transition dynamics and invertible emissions.

- We show that (i) temporal latents are identifiable up to affine transformations, (ii) recurrent Markov switching dynamics are identifiable up to regime permutations, and (iii) regime-dependent covariances reduce affine ambiguity to permutation and scaling. These results provide an end-to-end identifiability framework for recurrent deep SDSs.
- Motivated by the theory, we propose ΩSDS, a likelihood-based estimator that extracts temporal features using a flow-based emission and learns prior dynamics using expectation-maximisation. Unlike VAE-based estimators, ΩSDS optimises the exact likelihood, which avoids the additional approximation gap introduced by variational inference.
- Empirically, exact likelihood improves latent disentanglement on synthetic data, while recurrent switches support interpretable dynamics and strong forecasting performance on controlled physical systems and real-world dancing videos.

Related work SDSs are flexible models, but inference is generally challenging due to the interaction between continuous and discrete states [Ghahramani and Hinton, 2000]. Recent works develop inference for switching linear dynamical systems [Linderman et al., 2017], Gumbel-softmax relaxations [Becker-Ehmck et al., 2019], and infinite switching processes [Geadah et al., 2024]. Current nonlinear SDSs are mostly trained with variational objectives, including collapsed-switching inference [Dong et al., 2020], recurrent explicit-duration models [Ansari et al., 2021], and graph-based switching dynamics [Liu et al., 2023]. Flow-based likelihood estimation remains less explored [de Bézenac et al., 2020]. Identifiability for regime-switching latent variable models builds on nonlinear ICA and finite-mixture theory [Khemakhem et al., 2020, Yakowitz and Spragins, 1968, Allman et al., 2009]. HMM-ICA uses finite-state HMM identifiability to establish nonlinear ICA [Gassiat et al., 2016, Hälvä and Hyvarinen, 2020], while SNICA proves disentanglement under stationary temporal dependencies without identifying the mixture components [Hälvä et al., 2021]. Identifying latent transitions is harder, as temporal dependencies prevent a direct extension of Allman et al. [2009]. Recently, classical finite mixture model results [Yakowitz and Spragins, 1968] have been extended to non-recurrent MSMs, yielding identifiability under Gaussian observations [Song et al., 2023] and piecewise linear emissions [Balsells-Rodas et al., 2024]. Alternatively, Song et al. [2024] establishes results by assuming knowledge of the number of states K and observation-dependent regimes. Related causal-discovery methods allow recurrent switching but identify graph structures Balsells-Rodas et al. [2025] rather than the switching process. Table 1 summarises these distinctions.

2 Recurrent Switching Dynamical Systems

A recurrent Switching Dynamical System (rSDS), exemplified in Figure 1, models sequential data by introducing continuous latent states $\boldsymbol{z}_{1:T}$ and discrete switching variables $\boldsymbol{s}_{1:T}$. At each time step t, the observation $\boldsymbol{x}_t \in \mathbb{R}^n$ is generated through a (noisy) transformation of the continuous latent variable $\boldsymbol{z}_t \in \mathbb{R}^m$. While recent nonlinear SDS approaches use additive-noise emissions [Dong et al., 2020, Ansari et al., 2021], we assume a flow-based emission similarly used in Sorrenson et al. [2020]:

Assumption A1 (Emission model)**.** Observations $\boldsymbol{x}_t \in \mathbb{R}^n$ are generated via a $\mathcal{C}_1$ diffeomorphism $f_{\boldsymbol{\theta}_f} : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^n$ of latent variables $\boldsymbol{z}_t \in \mathbb{R}^m$ and i.i.d. noise $\boldsymbol{\varepsilon}_t \in \mathbb{R}^d$, with $d = n - m$.

$$\boldsymbol{x}_t = f_{\boldsymbol{\theta}_f}(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t), \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 I_d). \tag{1}$$

The latent dynamics $\boldsymbol{z}_{1:T}$ follow a recurrent Markov Switching Model (rMSM) [Ephraim and Roberts, 2005]. Therefore, at time t, the discrete switch $s_t \in \{1, \dots, K\}$ determines the transition dynamics of $\boldsymbol{z}_t$, while the switching process receives feedback from past latent variables $\boldsymbol{z}_{t-1}$. Under first-order Markov assumptions, the joint distribution of $(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{1:T})$ factorises as follows:

$$p_{\boldsymbol{\theta_z},\boldsymbol{\theta_s}}(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{1:T}) = p_{\boldsymbol{\theta_s}}(s_1) p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_1 | s_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t) p_{\boldsymbol{\theta_s}}(s_t \mid s_{t-1}, \boldsymbol{z}_{t-1}). \quad (2)$$

We make the following assumption on the latent transition dynamics $p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)$.

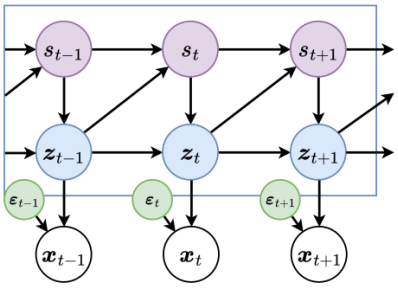

Figure 1: The considered rSDS model. Observations are generated through an invertible mapping of recurrent switching features and exogenous noises.

Assumption A2 (Analytic Gaussian autoregressive dynamics)**.** The latent variables $\boldsymbol{z}_t \in \mathbb{R}^m$ follow a regime-dependent Gaussian autoregressive process:

$$\boldsymbol{z}_t \mid s_t, \boldsymbol{z}_{t-1} \sim \mathcal{N}\Big(\boldsymbol{z}_t \mid \boldsymbol{m}_{s_t}(\boldsymbol{z}_{t-1}), \Sigma_{s_t}\Big), \quad \Sigma_{s_t} = \operatorname{diag}(\boldsymbol{\sigma}_{s_t}^2)$$

where each mean function $\boldsymbol{m}_k : \mathbb{R}^m \to \mathbb{R}^m$ is analytic, each covariance $\Sigma_k \succ 0$ is diagonal with $\boldsymbol{\sigma}_k^2$ denoting the diagonal, and both means $\boldsymbol{m}_k$ and/or covariances Σ_k are distinct across regimes.

This assumption is in the same format as in Balsells-Rodas et al. [2024], where results are established for non-recurrent switching, i.e., $p_{\boldsymbol{\theta_s}}(s_t \mid s_{t-1}, \boldsymbol{z}_{t-1}) = p_{\boldsymbol{\theta_s}}(s_t \mid s_{t-1})$. In contrast, we consider a common recurrent formulation in rSDS [Linderman et al., 2017, Dong et al., 2020, Ansari et al., 2021] to increase the flexibility of the switching process, where switches can adapt to latent distributional changes rather than following fixed Markov transition matrices.

Assumption A3 (Recurrent switching process)**.** The switches are dependent on previous latents:

$$p_{\boldsymbol{\theta_s}}(s_t = k \mid s_{t-1} = j, \boldsymbol{z}_{t-1}) = Q_{jk}(\boldsymbol{z}_{t-1}), \quad Q : \mathbb{R}^m \to [0,1]^{K \times K}, \quad \sum_{k=1}^{K} Q_{jk}(\boldsymbol{z}_{t-1}) = 1,$$

where Q is continuous almost everywhere in $\mathbb{R}^m$. We also assume each column of $Q(\boldsymbol{z})$ contains at least one positive entry for almost every $\boldsymbol{z} \in \mathbb{R}^m$ and $p(s_1 = k) > 0$ for all $k \in \{1, \dots, K\}$.

A.e. continuity of Q is a weak regularity assumption: it allows discontinuities on lower-dimensional boundaries, such as switching surfaces or collision boundaries, e.g., in bouncing-ball dynamics.

Identifiability desiderata in SDSs. Let $\boldsymbol{\Theta}$ be a family of parameters which parametrise an rSDS under Eqs. (1) and (2), where $\boldsymbol{\theta} := \{\boldsymbol{\theta}_f, \boldsymbol{\theta_z}, \boldsymbol{\theta_s}, m, K\}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We aim to establish identifiability for every parameter of $\boldsymbol{\theta}$: (i) the latent dimensionality m, (ii) the number of regimes K, (iii) the distribution of the latent variables $\boldsymbol{z}_t$, (iv) the autoregressive components $p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)$, and (v) the switching process $p_{\boldsymbol{\theta_s}}(s_t \mid s_{t-1}, \boldsymbol{z}_{t-1})$. Additionally, we denote $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T})$ and $p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_{1:T})$ as the observational and latent distributions induced by the model. Notably, our emission structure allows including the latent dimensionality m in the target equivalence class, which is non-standard in the literature [Song et al., 2023, Balsells-Rodas et al., 2024, Song et al., 2024].

Definition 2.1 (Latent identifiability up to affine transformations)**.** The latent variables $\boldsymbol{z}$ are identifiable up to affine transformations, if for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{1:T}), \forall \boldsymbol{x}_{1:T} \in \mathbb{R}^{Tn}$, we have $m = m'$ and there exists an invertible affine transformation $\psi : \mathbb{R}^m \to \mathbb{R}^m$ such that $p_{\boldsymbol{\theta_z}} = \psi^{(T)}_{\#} p_{\boldsymbol{\theta}'_{\boldsymbol{z}}}$, where $\psi^{(T)} : \mathbb{R}^{Tm} \to \mathbb{R}^{Tm}$ is defined by $\psi^{(T)}(\boldsymbol{z}_{1:T}) := (\psi(\boldsymbol{z}_1), \dots, \psi(\boldsymbol{z}_T))$.

Definition 2.2 (Transition identifiability up to permutations)**.** The transition model is identifiable up to permutations if for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that $p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_{1:T}) = p_{\boldsymbol{\theta}'_{\boldsymbol{z}}}(\boldsymbol{z}_{1:T})$, $\forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}$, we have $K = K'$ and there exists a permutation $\pi \in S_K$ such that for any $k \in \{1, \dots K\}$, $p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t = k) = p_{\boldsymbol{\theta}'_{\boldsymbol{z}}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t = \pi(k))$, and for any $k, j \in \{1, \dots, K\}$,

$$p_{\boldsymbol{\theta_s}}(s_t = k \mid s_{t-1} = j, \boldsymbol{z}_{t-1}) = p_{\boldsymbol{\theta}'_{\boldsymbol{s}}}(s_t = \pi(k) \mid s_{t-1} = \pi(j), \boldsymbol{z}_{t-1}).$$

Def. 2.1 concerns identification of the latent representation up to affine transformations, while Def. 2.2 targets the regime-dependent dynamics once the continuous latent space is fixed. Since the rSDS is composed of several distinct elements, establishing end-to-end identifiability is a challenging task.

3 Identifiability Theory for Recurrent Switching Dynamical Systems

In this section, we present our main identifiability results as follows:

i) **Latent variable identifiability.** We establish identifiability of the latent variable distribution $p_{\boldsymbol{\theta}_{\boldsymbol{z}}}(\boldsymbol{z}_{1:T})$ up to affine transformations under invertible emissions in Section 3.1.
ii) **rMSM identifiability.** We then establish identifiability of the transition dynamics up to permutations in Section 3.2.
iii) **Further disentangling latent variables.** Finally, we impose regime-dependent assumptions that reduce the affine ambiguity of latent variables to permutation and scaling in Section 3.3.

Importantly, each stage may be of independent interest, as we significantly strengthen previous results. For each result, we first state the additional assumptions required, followed by the corresponding theoretical result. We then provide a proof sketch and conclude with a discussion of its implications.

3.1 Latent Space Identifiability up to Affine Transformations

We first establish the affine identifiability of the continuous latent variables. Since we keep general invertible emissions, identifiability must come from the latent transition dynamics. The following assumptions ensure sufficient regime separation and local variability.

Assumption A4 (Sticky switching)**.** There exists a non-empty open set $\mathcal{Z} \subseteq \mathbb{R}^m$ such that

$$Q_{kk}(\boldsymbol{z}) \geq 1 - \epsilon(\boldsymbol{z}), \quad \epsilon(\boldsymbol{z}) \in [0, \tfrac{1}{2}),\ \forall k \in \{1, \dots, K\},\ \forall \boldsymbol{z} \in \mathcal{Z}, \tag{3}$$

where $\epsilon : \mathcal{Z} \to [0, \frac{1}{2})$ is continuous in $\mathcal{Z}$.

Assumption A5 (Transition separation)**.** Furthermore, there exist $\boldsymbol{z}^* \in \mathcal{Z}$, a constant $\ell^* > 0$, and an integer $T^* \geq 2$ such that, for any regime $i \in \{1, \dots, K\}$, there exists a latent history $\boldsymbol{z}_1^{(i)}(\boldsymbol{z}^*), \dots, \boldsymbol{z}_{T^*}^{(i)}(\boldsymbol{z}^*) \in \mathcal{Z},\ \boldsymbol{z}_{T^*}^{(i)}(\boldsymbol{z}^*) = \boldsymbol{z}^*$, for which

$$\frac{\mathcal{N}(\boldsymbol{z}_\tau^{(i)}(\boldsymbol{z}^*) | \boldsymbol{z}_{\tau-1}^{(i)}(\boldsymbol{z}^*), s_\tau = i)}{\max_{j \neq i} \mathcal{N}(\boldsymbol{z}_\tau^{(i)}(\boldsymbol{z}^*) | \boldsymbol{z}_{\tau-1}^{(i)}(\boldsymbol{z}^*), s_\tau = j)} > e^{\ell^*}, \quad \tau = 2, \dots, T^*, \quad \ell^* > \log \frac{1+\epsilon^*}{1-\epsilon^*}, \tag{4}$$

where $\epsilon^* = \max\{\epsilon(\boldsymbol{z}_1^{(i)}(\boldsymbol{z}^*)), \dots, \boldsymbol{z}_{T^*-1}^{(i)}(\boldsymbol{z}^*)\}$, and T^* is chosen large enough so that

$$p(s_{T^*} = i \mid \boldsymbol{z}_{1:T^*}^{(i)}(\boldsymbol{z}^*)) > \frac{1}{2}, \quad \forall i \in \{1, \dots, K\}. \tag{5}$$

Assumption A6 (Covariance separation)**.** For every regime $k \in \{1, \dots, K\}$, there exists a nonempty set $\mathcal{I}_k \subseteq \{1, \dots, m\}$ such that $\sigma_{k,i} \geq \sigma_{k',i}$, for any $i \in \mathcal{I}_k$ and any $k' \in \{1, \dots, K\} \setminus \{k\}$. Let $\mathcal{T}_k \subseteq \{1, \dots, K\} \setminus \{k\}$ denote the regimes tied with k in variance along all coordinates in $\mathcal{I}_k$:

$$\mathcal{T}_k := \Big\{ k' \in \{1, \dots, K\} \setminus \{k\} :\ \sigma_{k,i} = \sigma_{k',i},\ \forall i \in \mathcal{I}_k \Big\}. \tag{6}$$

If $\mathcal{T}_k \neq \emptyset$, we further assume that for every $\boldsymbol{z}_t \in \mathbb{R}^m$,

$$\boldsymbol{m}_{k,\mathcal{I}_k}(\boldsymbol{z}_t) \notin \operatorname{conv}(\{\boldsymbol{m}_{k',\mathcal{I}_k}(\boldsymbol{z}_t) : k' \in \mathcal{T}_k\}), \tag{7}$$

where $\boldsymbol{m}_{k,\mathcal{I}_k}(\cdot)$ denotes the subvector of $\boldsymbol{m}_k(\cdot)$ corresponding to coordinates in $\mathcal{I}_k$, and conv denotes the convex hull.

Assumption A7 (Locally nondegenerate regime)**.** There exist a regime $1 \leq k_0 \leq K$ and a point $\boldsymbol{z}^* \in \mathbb{R}^m$ such that the determinant of the Jacobian of $\boldsymbol{m}_{k_0}(\boldsymbol{z}^*)$ is nonzero: $\det J_{\boldsymbol{m}_{k_0}}(\boldsymbol{z}^*) \neq 0$.

Under these assumptions, we establish latent variable identifiability up to affine transformations, and provide further intuitions on the required conditions after a proof sketch.

Theorem 3.1. *Under assumptions (A1–A7), the latent representation of the rSDS (Eqs.* (1),(2)*) is identifiable up to affine transformations (Def. 2.1), if one of the following conditions holds:*
(i) the number of regimes K is known.
(ii) K is unknown, but $Q : \mathbb{R}^m \to [0,1]^{K\times K}$ is analytic.

See Appendix B.1 for proof. When K is known (e.g., controlled physical systems), identifiability reduces to aligning the mixture components, where we provide a proof sketch below. For unknown K (e.g., discovering behavioural patterns in videos), recovering the number of regimes requires additional structure, where condition (ii) in Thm. 3.1 provides a sufficient one.

Proof sketch (known K): Define the pushforward measure of $p(\boldsymbol{z}_{t+1}, \boldsymbol{\varepsilon}_{t+1} \mid \boldsymbol{z}_t, s_t = k)$ through $f_{\boldsymbol{\theta}_f}$:

$$\phi_k(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) := (f_{\boldsymbol{\theta}_f})_{\#}\big(\mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{m}_k(\boldsymbol{z}_t), \Sigma_k)\, p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{t+1})\big).$$

Considering all $k \in \{1, \dots, K\}$, define $\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) := (\phi_1(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t), \dots, \phi_K(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t))^\top$ as the pushforward measures of the K Gaussian components in vector form (define $\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ similarly).

I. Write the predictive distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t})$ as a mixture in terms of $\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t}) = \sum_{k=1}^{K} p(s_{t+1} = k \mid \boldsymbol{x}_{1:t}) \phi_k(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t). \tag{8}$$

Under (A4–A5) and using Lemma B.1, we can choose K distinct histories $(\boldsymbol{x}_1^{(k)}, \dots, \boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{x}_t)$ that share the same endpoint $\boldsymbol{x}_t$ but induce different dominant posterior distributions satisfying $p(s_{t+1} = k \mid \boldsymbol{x}_{1:t}) > \frac{1}{2}$. This allows the construction of a full-rank linear system in terms of $\boldsymbol{x}_t$. We then show there exists an invertible matrix $\Pi(\boldsymbol{x}_t)$ such that

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{1:T}) \quad \Rightarrow \quad \Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \Pi(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t). \tag{9}$$

II. Under (A6), Prop B.5, Prop B.4, and Lemma B.3, $\Pi(\boldsymbol{x}_t)$ is a constant permutation matrix.

III. Since each component in model $\boldsymbol{\theta}$ corresponds to a permuted component in model $\boldsymbol{\theta}'$, pick regime k_0 from (A7). Using an iVAE-style argument, we recover the latent dimensionality m and establish that the latent dynamics are identifiable up to affine transformations (Def. 2.1).

As discussed, establishing the above result requires restricting the flexibility of the nonlinear switching prior through (A4–A7). (A7) requires one regime for which the Jacobian of the mean transition is locally non-singular, so that variations in latent space are distinguishable. This condition is closely related to sufficient variability assumptions used in related ICA literature [Hyvarinen and Morioka, 2017, Khemakhem et al., 2020, Yao et al., 2022], but adapted to our setting where variability arises through the means of Gaussian transitions.

Understanding (A6): (A6) imposes a dominance condition on the regime-dependent variances, preventing any single regime from dominating across all directions. In shared-variance settings, variance dominance is not enough, and the regime-dependent means require further structure across the regimes and dimensions with equal variance. The convex hull condition in Eq. (7) is geometrically intuitive, but does not allow arbitrary nonlinear parametrisations. To enforce this, the following design is possible for a given regime $k \in [K]$:

$$\boldsymbol{m}_k(\boldsymbol{z}) = h(\boldsymbol{z}) + B(\boldsymbol{z})\boldsymbol{v}_k, \tag{10}$$

with invertible $B(\boldsymbol{z})$ for any $\boldsymbol{z} \in \mathbb{R}^m$ and $\boldsymbol{v}_k \in \mathbb{R}^m$. The condition holds if $\{\boldsymbol{v}_{k',\mathcal{I}_k} : k' \in \mathcal{T}_k \cup \{k\}\}$ are the vertices of a convex hull. Figure 2 exemplifies (A6), where $\boldsymbol{v}_k \in \{(1,1), (-1,-1), (-1,1), (1,-1)\}$ forms a convex shape, and $h(\boldsymbol{z}) := \boldsymbol{z}$, $B(\boldsymbol{z}) := \Delta t$.

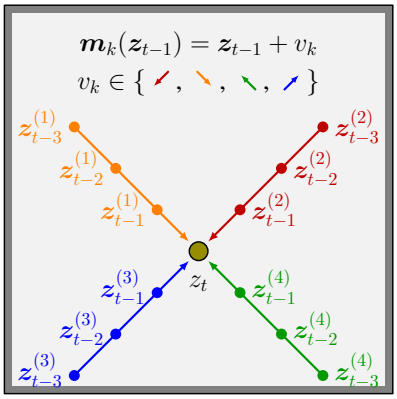

Figure 2: Bouncing ball example ($K = 4$), each corresponding to a fixed velocity $\boldsymbol{v}_k$ in the 2D latent space. Histories $\boldsymbol{z}_{t-3:t-1}^{(k)}$ associated with different velocities remain separated, illustrating the convex hull and likelihood separation intuitions.

Understanding (A4–A5): The intuition behind these conditions is that we have histories $\boldsymbol{z}_{1:t}^{(k)}$ such that the posterior distribution $p(s_t|\boldsymbol{z}_{1:t}^{(k)})$ concentrates on a certain regime k, and by Lemma B.1, this is possible through sticky switches (A4), and likelihood

separation across regimes (A5). While (A4) is a common modelling choice in regime-switching models [Fox et al., 2008, Linderman et al., 2017, Glaser et al., 2020], (A5) is analogous to signal-strength conditions used in change-point detection. For example, Yu et al. [2023] assumes nonzero jumps between changes. In our case, (A5) imposes consistent regime-dependent likelihood margins along a history. Figure 2 further illustrates this notion, where the latent variable transitions of a bouncing ball are separated by a positive margin thanks to having fixed-velocity dynamics driven by $\boldsymbol{v}_k \in \mathbb{R}^2$. Observations of an SDS in this example would correspond to videos where each frame is rendered from the 2D position of the ball in the box. A possible limitation to (A5) is the difficulty in maintaining large enough margins during multiple transitions such that the posterior domination is satisfied. However, stickiness plays a key role in weakening the required margin. We expand this discussion in Appendix B.2, where we show that by Corollary B.2 the sufficient history length T^* can be determined from the initial posterior $p(s_1|\boldsymbol{z}_1)$, the margin ℓ^*, and stickiness ϵ^*. Specifically for uniform initial posterior over K regimes, $p(s_1 = k|\boldsymbol{z}_1) = \frac{1}{K},\ k \in \{1, \dots, K\}$, posterior dominance is possible after one transition ($T^* = 2$) if $\ell^* > \log \frac{K-1+\epsilon^*}{1-\epsilon^*}$. Therefore, (A4–A5) can be interpreted as a design choice on a trade-off between transition separation and switching stickiness. For more details, see Appendix B.2. Below, we exemplify nonlinear dynamics.

A visual example for (A5): Consider $m = 1$, $K = 3$, equal variances $\sigma^2 = 0.1$, and cosine transitions:

$$m_k(z) = \cos\left(z + \frac{\pi(k-1)}{2}\right),\ k = 1, 2, 3.$$

The transition means are shown in Figure 3. Choosing the common endpoint $z_2 = 1$, an aligned set of predecessors is $z_1^{(1)} = 0$, $z_1^{(2)} = \frac{\pi}{2}$, and $z_1^{(3)} = -\frac{\pi}{2}$, for which $m_k(z_1^{(k)}) = 1$, $k = 1, 2, 3$. As illustrated in the figure, the minimum likelihood margin is $\ell^* = 5$. Since (A5) requires existence, this single construction is sufficient. Under a uniform initial posterior, the margin is large enough that any $\epsilon^* < \frac{1}{2}$ satisfies Lemma B.1 with $T^* = 2$. In contrast, choosing a less separated endpoint $z_2 = 0.5$ yields a margin $\ell^* = 0.67$, where no $\epsilon^* > 0$ can satisfy $\ell^* > \log \frac{K-1+\epsilon^*}{1-\epsilon^*}$ to obtain one-step dominance. This illustrates that stronger local separation permits shorter histories, while weaker separation must be compensated either by stronger stickiness or longer histories.

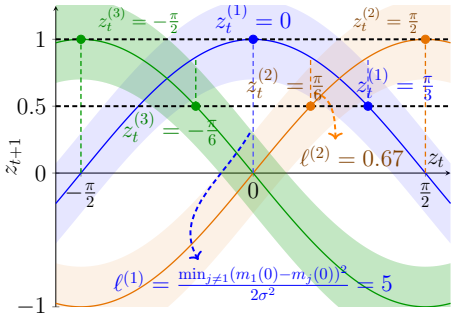

Figure 3: Cosine transition means for the $K = 3$ example. For $\boldsymbol{z}_{t+1} = 1, 0.5$ we find positive margins $\ell^* = 5, 0.67$, respectively.

3.2 Identifying Recurrent Prior Dynamics

Given that the Gaussian family is closed under affine transformations, we continue our analysis by working directly on latent dynamics driven by rMSMs. We present the following identifiability result.

Theorem 3.2. *The rMSM is identifiable up to permutations under the following conditions:*

(i) Under Assumption (A2), the continuous latent transitions $p_{\boldsymbol{\theta}_z}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)$ are identifiable.

(ii) Under Assumptions (A2-A5) and further assuming $Q : \mathbb{R}^m \to [0, 1]^{K\times K}$ is analytic, the rMSM is identifiable following Definition 2.2.

Proof sketch. See Appendix B.4 for the full proof. We prove rMSM identifiability in two steps.

(i) Thm. B.6: By Yakowitz and Spragins [1968] and (A2), the transitions $p_{\boldsymbol{\theta}_z}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)$ are identifiable up to a permutation indexed by $\boldsymbol{z}_{t-1}$, which is constant from analyticity by (A2).

(ii) Thm. B.7: As in Step **I** of Theorem 3.1, we construct a system of equations in terms of $Q(\boldsymbol{z}_{t-1})$ by stacking K histories, such that under (A4–A5) the system has a unique solution.

Our proof strategy for (i) extends Balsells-Rodas et al. [2025], by generalising the transitions to analytic functions, and makes the above result compatible with identifying latent components of the rSDS by combining with the previous affine identifiability results (Def. 2.1). I.e., there exists a permutation $\pi \in S_k$, and an affine transformation $\psi(\boldsymbol{z}) = A\boldsymbol{z} + \boldsymbol{b}$, such that:

$$\boldsymbol{m}_k(\boldsymbol{z}) = A\boldsymbol{m}'_{\pi(k)}(A^{-1}(\boldsymbol{z} - \boldsymbol{b})) + \boldsymbol{b},\ Q_{jk}(\boldsymbol{z}) = Q'_{\pi(j)\pi(k)}(A^{-1}(\boldsymbol{z} - \boldsymbol{b})),\ k = 1 : K,\ \boldsymbol{z} \in \mathbb{R}^m. \tag{11}$$

For other rSDS designs, Thm. 3.2 is independent of Section 3.1 provided we establish identifiability up to affine transformations (Def 2.1). For example, emission piecewise linearity [Kivva et al., 2022] could replace (A4-A7), and Theorem 3.2 would still be valid to identify the transitions $\boldsymbol{m}_k$ as in Eq. (11). However, for full identifiability including $Q(\boldsymbol{z}_{t-1})$, further restrictions would be necessary.

3.3 Disentanglement: Resolving the Affine Indeterminacy

Our last stage consists of resolving the structure of the affine mapping $\psi(\boldsymbol{z}) = A\boldsymbol{z} + \boldsymbol{b}$ in Eq. (11) to establish identifiable disentangled representations and permit nonlinear ICA [Hyvärinen and Pajunen, 1999]. This identifiability, presented in Thm. 3.3, requires the following additional assumption.

Assumption A8. For a pair of regimes $k_1 \neq k_2 \in \{1, \dots, K\}$, with diagonal variances $\operatorname{diag}(\boldsymbol{\sigma}_{k_1}^2), \operatorname{diag}(\boldsymbol{\sigma}_{k_2}^2)$, respectively, define the ratio vector $\boldsymbol{r}_{k_1,k_2}$ between variances:

$$\boldsymbol{r}_{k_1,k_2} := \left(\frac{\sigma_{k_1,1}}{\sigma_{k_2,1}}, \quad \dots, \quad \frac{\sigma_{k_1,m}}{\sigma_{k_2,m}} \right) = \frac{\boldsymbol{\sigma}_{k_1}}{\boldsymbol{\sigma}_{k_2}} \in \mathbb{R}^m. \tag{12}$$

For every pair of regimes $k_1 \neq k_2,\ k_1, k_2 \in \{1, \dots, K\}$, stack ratios $\boldsymbol{r}_{k_1,k_2} \in \mathbb{R}^m$:

$$R \in \mathbb{R}^{\binom{K}{2} \times m}, \quad R_{(k_1,k_2),i} = \boldsymbol{r}_{k_1,k_2,i}, \quad i \in \{1, \dots, m\}. \tag{13}$$

That is, each row of R is a ratio vector formed by combining regime pairs. We assume that R has no two identical columns, i.e., for $i \neq j \in \{1, \dots, m\}$, there exists (a, b) such that $R_{(a,b),i} \neq R_{(a,b),j}$.

Theorem 3.3. *Let $K \geq 2$. Let $\boldsymbol{z}'_t = A\boldsymbol{z}_t + \boldsymbol{b}$, where $A \in \mathbb{R}^{m \times m}$ is invertible and $\boldsymbol{b} \in \mathbb{R}^m$. Under Assumption (A8), from $\boldsymbol{z}'_t$, we can recover an invertible matrix $A' \in \mathbb{R}^{m \times m}$ such that $(A')^{-1}A = PD$, where P is a permutation matrix, and D is diagonal.*

See Appendix B.5 for the proof, which uses similar eigen-decomposition tricks introduced in the strategy of Kivva et al. [2022]. We provide examples on how Eq. (13) is constructed, and partial disentanglement when (A8) is not satisfied. Thm. 3.3 shows that in regime-dependent settings, increasing the number of regimes with diverse variances strengthens identifiable latent space disentanglement.

Remark 3.4. For mixture models, Kivva et al. [2022, Assumption (P2)] enables disentanglement via regime-dependent variances, which extends to the temporal case as follows. For K regimes with diagonal covariances, i.e., $z_{t,i} \perp\!\!\!\perp z_{t,j} \mid s_t, \boldsymbol{z}_{t-1}$ for all $i \neq j \in \{1, \dots, m\}$, there exists one pair of regimes $k_1 \neq k_2 \in \{1, \dots, K\}$ such that the ratio vector $\boldsymbol{r}_{k_1,k_2}$ has distinct elements. However, this condition does not scale well to high dimensions, since ratios with close values may challenge disentanglement in practical estimation. We note (A8) is weaker than Assumption (P2) of Kivva et al. [2022], as a single ratio vector with distinct elements forms R with distinct columns. Therefore, (A8) works as a scalable generalisation by fully exploiting the mixture setting of rSDS.

4 ΩSDS: Exact Likelihood Learning for SDSs

The identifiability results suggest a natural estimation approach: use an invertible emission to distinguish temporal features from exogenous noise, and perform exact inference on the resulting latent rMSM. We propose ΩSDS, a likelihood-based estimator that follows this principle by combining an invertible flow-based emission model with latent-space inference based on rMSMs. For simplicity, we drop the subscript $\boldsymbol{\theta}$ in the model equations. The flow defines a generative emission map from the latent dynamical variables and exogenous noise to observations, while its inverse acts as an encoder: $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = f^{-1}(\boldsymbol{x}_t)$, $1 \leq t \leq T$. Using the change-of-variables formula, the marginal log-likelihood of an observed trajectory is expressed as follows.

$$\log p(\boldsymbol{x}_{1:T}) = \sum_{t=1}^{T} \log \left| \det J_{f^{-1}}(\boldsymbol{x}_t) \right| + \log p(\boldsymbol{z}_{1:T}) + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{1:T}), \quad p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t) = \mathcal{N}(\boldsymbol{\varepsilon}_t; 0, \sigma_{\boldsymbol{\varepsilon}}^2 I) \tag{14}$$

where $\sigma_{\boldsymbol{\varepsilon}}$ is left as a hyper-parameter, and $J_{f^{-1}}$ denotes the Jacobian of the inverse transformation. In our experiments, we parametrise f using a multi-layer TarFlow [Zhai et al., 2025], $f = (f^{(1)} \circ \cdots \circ f^{(L)})$, where each layer is a masked autoregressive flow where the conditioning is implemented using causal transformers. To encourage a decomposition between dynamical variables and exogenous noise, we replace the permutation blocks in TarFlow with invertible LU mixing layers, as introduced

in RealNVP [Dinh et al., 2017]. For video experiments, we instead use a multi-scale flow architecture, inspired by Glow [Kingma and Dhariwal, 2018]. See Appendix C.2 for details.

The latent sequence likelihood $p(\boldsymbol{z}_{1:T})$ is defined by the rMSM prior with discrete regimes $\boldsymbol{s}_{1:T}$ in Eq. (2). We optimise the likelihood in Eq. (14) by gradient ascent. The gradient of the latent log-likelihood $\log p(\boldsymbol{z}_{1:T})$ can be written as an expectation under the exact switch posterior, $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{z}_{1:T}) = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{s}_{1:T}|\boldsymbol{z}_{1:T})}[\log p(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{1:T})]$, which yields the following [Dong et al., 2020]:

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{z}_{1:T}) = \sum_{k=1}^{K} \gamma_{1,k}(\boldsymbol{z}_{1:T}) \nabla_{\boldsymbol{\theta}} \log p(s_1 = k) + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{t,k,l}(\boldsymbol{z}_{1:T}) \nabla_{\boldsymbol{\theta}} \log Q_{lk}(\boldsymbol{z}_{t-1})$$
$$+ \sum_{k=1}^{K} \gamma_{1,k}(\boldsymbol{z}_{1:T}) \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{z}_1 | s_1 = k) + \sum_{t=2}^{T} \sum_{k=1}^{K} \gamma_{t,k}(\boldsymbol{z}_{1:T}) \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, s_t = k),$$

where $\gamma_{t,k}(\boldsymbol{z}_{1:T}) = p(s_t = k \mid \boldsymbol{z}_{1:T})$ and $\xi_{t,k,l}(\boldsymbol{z}_{1:T}) = p(s_t = k, s_{t-1} = l \mid \boldsymbol{z}_{1:T})$ are the marginal and two-step posterior distributions given $\boldsymbol{z}_{1:T}$. These quantities are computed using standard forward-backward recursions [Bishop, 2006], where details are deferred to Appendix C.1.

Training regime-switching models with nonlinear transitions can suffer from state collapse, where only a subset of regimes is used [Dong et al., 2020]. Rather than relying on annealing schedules, we initialise the rMSM parameters from PCA features and use a PCA-alignment auxiliary loss early in training. This stabilises the flow decomposition and provides a reliable early latent representation for regime learning. PCA-based initialisation is standard in SDSs [Linderman et al., 2017].

Consistency. Unlike VAE-based approaches, ΩSDS optimises the exact marginal likelihood. By invertibility of the emission, $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = f_{\boldsymbol{\theta}_f}^{-1}(\boldsymbol{x}_t)$, the latent $p_{\boldsymbol{\theta}_z}(\boldsymbol{z}_{1:T})$ is an MSM likelihood with covariate-dependent transitions $Q(\boldsymbol{z}_{t-1})$, aligning with the setting of MLE consistency results of Pouzo et al. [2022]. Therefore, under the usual assumptions for MSM and MLE consistency, including correct specification, global optimisation, positivity of the transition probabilities, ergodicity, and likelihood regularity, the resulting MLE favours consistency up to the equivalence class characterised in Defs. 2.1, 2.2, and Theorem 3.3 [Wald, 1949, Pouzo et al., 2022]. These regularity conditions are standard in likelihood-based consistency analyses and are additional to our identifiability assumptions, but compatible with them. However, the latent path in ΩSDS depends on the flow parameters $\boldsymbol{z}_t = f_{\boldsymbol{\theta}_f}^{-1}(\boldsymbol{x}_t)_{1:m}$, and the flow, transition means, covariances, and switching process are estimated jointly. Therefore, a formal consistency statement would require further analysis, which we leave as future work. This contrasts with VAE-based SDSs, which optimise an ELBO and require sufficient coverage of the variational family [Khemakhem et al., 2020]. Collapsed-switching approaches marginalise $\boldsymbol{s}_{1:T}$ conditionally on $\boldsymbol{z}_{1:T}$, but rely on an amortised variational posterior $q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$ which may introduce approximation gaps [Balsells-Rodas et al., 2024]. We further discuss this empirically in Appendix C.3.

5 Experiments

We evaluate ΩSDS through synthetic data with controlled ground-truth factors, 2D bouncing ball videos, and real-world dancing videos. Code for our implementation and experiments can be found in `https://github.com/charlio23/omega-SDS`. Appendix D.1 discusses the alignment between our assumptions and the synthetic datasets.

5.1 Synthetic Data

To assess the advantages of ΩSDS over VAE-based alternatives, we follow the settings in iSDS [Balsells-Rodas et al., 2024], using a non-recurrent Q, piecewise-linear emissions, and two dimensionality cases: noiseless observations ($n = m$), and noise-augmented observations ($n = 5m$). We compare ΩSDS with iSDS, NCTRL [Song et al., 2023], CtrlNS [Song et al., 2024], TDRL [Yao et al., 2022], and iVAE [Khemakhem et al., 2020] (with ground-truth regimes as auxiliary variables). Appendix D.4 provides data-generation details, ablations on Assumption (A8), and identification results for the number of regimes K and recurrent $Q(\boldsymbol{z})$.

Figure 4 reports switch recovery F_1 in the noiseless setting and latent recovery MCC in both settings. As expected, ΩSDS recovers regimes comparably to iSDS, since the data is compatible

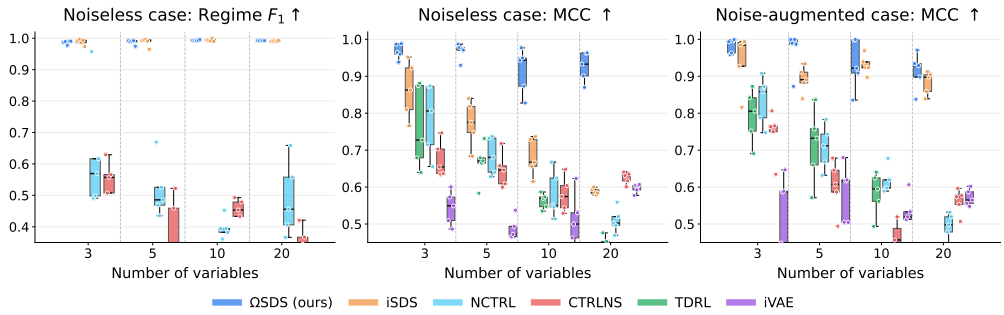

Figure 4: Synthetic data results under increasing dimensionality. Left: Regime F_1 score (ΩSDS and iSDS achieve near perfect recovery). Middle and right: MCC of noiseless and noise-augmented settings. Higher is better for both F_1 and MCC. Reported results use 5 different dataset seeds.

Table 2: Regime estimation and long-rollout quality. Regime F_1 is computed on observed input sequences ($T = 64$) and on predicted regimes from $t = 65$ to $t = 100$. LPIPS is computed on 5-frame-context rollouts. Higher is better for F_1; lower is better for LPIPS. We highlight the **best** and second-best methods.

Method	Input F_1 ↑	Pred. F_1 ↑	LPIPS@20 ↓	LPIPS@500 ↓
VRNN	–	–	0.014	0.028
KVAE	0.105	0.105	0.053	0.324
iSDS	0.819	0.345	0.111	0.106
REDSDS	0.655	0.580	0.073	0.135
SNLDS	0.998	0.913	0.007	0.059
ΩSDS (A)	0.867	0.362	0.110	0.092
ΩSDS (R)	**1.000**	**0.932**	**0.003**	**0.027**

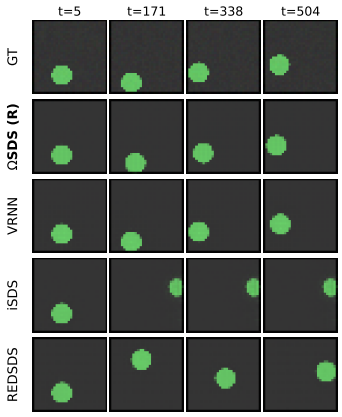

Figure 5: Long forecast example.

with a non-recurrent SDS. The main difference appears in latent recovery. In the noiseless setting, ΩSDS maintains consistently high MCC scores across dimensions, while VAE-based approaches degrade with increasing latent dimensionality. This suggests that ΩSDS provides a more scalable approach to disentanglement. In the noise-augmented setting, VAE-based methods improve, possibly due to the additional observed constraints. However, ΩSDS remains competitive in both settings. Overall, these results show that ΩSDS provides a scalable alternative to VAE-based SDS methods for disentanglement, even in non-recurrent settings where existing baselines are directly applicable.

5.2 Video Experiments

2D Bouncing Ball Videos. We synthesise videos of a ball bouncing inside a 2D square box, starting from random positions and one of four velocity modes, $k \in \{\nearrow, \nwarrow, \searrow, \swarrow\}$. This benchmark suits ΩSDS and showcases our theoretical results, where the discrete regimes determine the direction of motion, and switches depend on the continuous states through wall collisions. In contrast, previous identifiable SDS models (e.g., iSDS; Balsells-Rodas et al., 2024) cannot incorporate recurrent feedback from continuous states into the switching process. We train on $10k$ videos of length $T = 64$, with noisy coloured 32×32 frames. See Appendix D.5 for details on data generation and training.

We compare recurrent and non-recurrent variants, ΩSDS (R) and ΩSDS (A), against iSDS and other baselines that are not designed around identifiability: VRNN, KVAE, SNLDS, and REDSDS. All switching methods use $K = 4$, while VRNN is a no-switching recurrent baseline. Table 2 reports regime F_1 scores on held-out observed sequences of length $T = 64$, and on predictions from $t = 65$ to $t = 100$. We also report LPIPS errors for long-horizon forecasts at 20 and 500 steps, using a context of 5 observed frames. Figure 5 shows a representative 500-step rollout. The recurrent model, ΩSDS (R), achieves the strongest regime estimation and prediction performance, while also obtaining the best long-horizon perceptual quality. SNLDS is the closest competitor, since it is a VAE-based design with rMSMs in latent space. Crucially, both recurrent switching methods are competitive with VRNN, a memory-based sequence model with unidentifiable latent dynamics. The comparison between ΩSDS (R), ΩSDS (A), and iSDS highlights the role of recurrent switches. While non-recurrent switching can recover regimes on observed sequences, it struggles to predict future

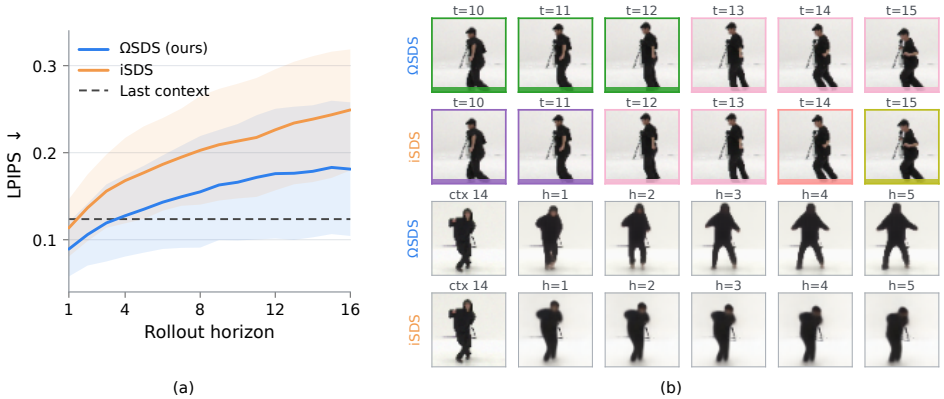

Figure 6: AIST dance video results. (a) Mean LPIPS error of predicted frames from 15 context frames across 100 held-out sequences, with shaded regions indicating one standard deviation. For reference, "last context" displays the average LPIPS error in terms of the final context frame. (b) Segmentation and forecasting visualisations, where colours indicate distinct inferred regimes.

dynamics. Additional evaluation and visualisations are provided in Appendix D.5. Overall, these results show that enforcing identifiable latent dynamics does not sacrifice predictive performance, while providing an interpretable decomposition of complex dynamics into simpler behavioural modes.

Dancing Videos. To showcase our identifiability results in real-world scenarios, we explore the dance video dataset of Tsuchida et al. [2019] and focus on segmenting and forecasting sequences from the *Middle Hip Hop* genre. See Appendix D.6 for details, along with additional visualisations. Figure 6 shows results of ΩSDS in contrast to the non-recurrent identifiable approach iSDS [Balsells-Rodas et al., 2024]. We (a) report LPIPS as a function of rollout time, and (b) compare posterior regime estimation (top two rows) and rollout predictions with 15 context frames (bottom two rows). ΩSDS improves rollout prediction and yields more interpretable dynamics by decomposing dance patterns, such as "move up" or "move down" in the shown example. The generation quality of ΩSDS is substantially better than iSDS, thanks to the use of recurrent dynamics and flow-based emissions. Although ΩSDS does not predict the exact future dance sequence, it produces more plausible movements and therefore achieves superior forecasting performance.

6 Conclusions

We presented identifiability results for recurrent switching nonlinear dynamical systems by imposing restrictions on the latent dynamics, recovering the latent representation up to acceptable equivalences under general invertible emissions. We also introduced ΩSDS, a flow-based estimator that uses exact-likelihood optimisation, avoiding the approximation gap introduced by VAE-based estimators. Empirically, our method improves disentanglement over variational alternatives and shows strong forecasting performance due to recurrent switching.

Limitations. Our results require constraints on both the emission and latent dynamics. While standard, invertibility excludes information-losing effects such as occlusions. On the other hand, the smoothness in (A2), the likelihood-separation condition (A5), and the geometric condition in the equal-variance case (A6) restrict the class of admissible nonlinear dynamics, preventing arbitrary parametrisations. Still, likelihood separation can be balanced with sticky switching (A4). Furthermore, while ΩSDS avoids the variational approximation gap, consistency remains open, since consistency of rMSMs alone requires further assumptions from covariate-dependent MSMs [Pouzo et al., 2022], and the joint flow-rMSM parameter estimation requires further analysis. Experimentally, the AIST dataset does not provide ground-truth regimes, preventing quantitative evaluation of the learned switching dynamics. For such real-world scenarios where the latent space is unknown, assumption verification is a fundamental limitation of identifiability theory, which restricts interpretability of the learned structures. Therefore, to extend these guarantees to real-world applications, future work will focus on weaker separation conditions, robustness to model misspecification, and extensions of likelihood-based models to broader nonlinear ICA and causal representation learning settings.

Acknowledgements

C. B.-R., X. S., and Z. X. thank the Department of Computing at Imperial College London for PhD scholarships. We also thank reviewers of NeurIPS 2026 for pointing out minor errors in proofs for a few lemmas during the submission phase.

References

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Abdul Fatir Ansari, Konstantinos Benidis, Richard Kurle, Ali Caner Turkmen, Harold Soh, Alexander J Smola, Bernie Wang, and Tim Januschowski. Deep explicit duration switching models for time series. *Advances in Neural Information Processing Systems*, 34:29949–29961, 2021.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Eew3yUQQtE.

Carles Balsells-Rodas, Xavier Sumba, Tanmayee Narendra, Ruibo Tu, Gabriele Schweikert, Hedvig Kjellstrom, and Yingzhen Li. Causal discovery from conditionally stationary time series. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2715–2741. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/balsells-rodas25a.html.

Philip Becker-Ehmck, Jan Peters, and Patrick Van Der Smagt. Switching linear dynamics for variational bayes filtering. In *International conference on machine learning*, pages 553–562. PMLR, 2019.

Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

Emmanuel de Bézenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, and Tim Januschowski. Normalizing kalman filters for multivariate time series analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2995–3007. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f47cef5e38c952f94c5d61726027439-Paper.pdf.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Zhe Dong, Bryan Seybold, Kevin Murphy, and Hung Bui. Collapsed amortized variational inference for switching nonlinear dynamical systems. In *International Conference on Machine Learning*, pages 2638–2647. PMLR, 2020.

Yariv Ephraim and William JJ Roberts. Revisiting autoregressive hidden markov modeling of speech signals. *IEEE Signal processing letters*, 12(2):166–169, 2005.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An hdp-hmm for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319, 2008.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *Advances in neural information processing systems*, 30, 2017.

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.

Victor Geadah, Jonathan W Pillow, et al. Parsing neural dynamics with infinite recurrent switching linear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024.

Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000.

Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in neural information processing systems*, 33:14867–14878, 2020.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pages 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkE3y85ee`.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

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

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=EbIDjBynYJ8`.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pages 914–922. PMLR, 2017.

Georg Lindgren. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, pages 81–91, 1978.

Yongtuo Liu, Sara Magliacane, Miltiadis Kofinas, and Efstratios Gavves. Graph switching dynamical systems. In *International Conference on Machine Learning*, pages 21867–21883. PMLR, 2023.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvarinen. Independent innovation analysis for nonlinear vector autoregressive process. In *International conference on artificial intelligence and statistics*, pages 1549–1557. PMLR, 2021.

A Poritz. Linear predictive hidden markov models and the speech signal. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1291–1294. IEEE, 1982.

Demian Pouzo, Zacharias Psaradakis, and Martin Sola. Maximum likelihood estimation in markov regime-switching models with covariate-dependent transition probabilities. *Econometrica*, 90(4): 1681–1710, 2022.

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

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8092–8113. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/19a567abaec3990cb40d7a013556fecd-Paper-Conference.pdf.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J709rtAUD1.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygeHgSFDH.

Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

H. Tong and K. S. Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(3):245–292, 1980. ISSN 00359246.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, Netherlands, November 2019.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949. ISSN 00034851. URL http://www.jstor.org/stable/2236315.

Chai Wah Wu. On bounds of extremal eigenvalues of irreducible and m-reducible matrices. *Linear algebra and its applications*, 402:29–45, 2005.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Vi-sZWNA_Ue.

Yi Yu, Oscar Hernan Madrid Padilla, Daren Wang, and Alessandro Rinaldo. A note on online change point detection. *Sequential Analysis*, 42(4):438–471, 2023.

Shuangfei Zhai, Ruixiang ZHANG, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Ángel Bautista, Navdeep Jaitly, and Joshua M. Susskind. Normalizing flows are capable generative models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2uheUFcFsM.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Appendix for "End-to-End Identifiable Recurrent Switching Dynamical Systems"

Contents

A Unidentifiability of Non-Temporal Mixture Generative Models **15**

B Proofs and Extended Discussions on Identifiability **16**

B.1 Proof of Theorem 3.1 16

B.2 Discussion on Assumptions (A4–A5) 20

B.3 Discussion on Assumption (A6) 23

B.4 Proof of Theorem 3.2 25

B.5 Extended discussion on disentanglement and Proof of Theorem 3.3 29

C Method Details and Discussions **31**

C.1 Forward-backward Inference 31

C.2 ΩSDS implementation details 32

C.3 Likelihood-based vs VAE-based estimation 32

D Experimental Details **33**

D.1 Assumption to Dataset Correspondence 33

D.2 Metrics 34

D.3 Baseline specifications 35

D.4 Synthetic Data 35

D.5 Bouncing Ball 38

D.6 AIST Dance Videos 38

E Broader Impacts **41**

A Unidentifiability of Non-Temporal Mixture Generative Models

Khemakhem et al. [2020] illustrates identifiability for unconditional priors. This notion also extends to models with mixture priors, where the latent variables are sampled from a mixture distribution:

$$\boldsymbol{x} = f_{\boldsymbol{\theta}_f}(\boldsymbol{z}), \quad \boldsymbol{z} \sim \sum_{k=1}^{K} p(s=k)p(\boldsymbol{z} \mid s=k). \tag{15}$$

In this setting, even the number of components K is unidentifiable from the observed data, which leads to an unidentifiable latent representation. To see this, consider two models with transformations $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}'}$, and latent distributions with K and K' mixture components, respectively, such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^n$. We can write the observed density as a pushforward measure of $\boldsymbol{z}$ through $f_{\boldsymbol{\theta}_f}$ (or $f_{\boldsymbol{\theta}'_f}$):

$$(f_{\boldsymbol{\theta}\#}p_{\boldsymbol{\theta}_z})(\boldsymbol{x}) = (f_{\boldsymbol{\theta}'\#}p_{\boldsymbol{\theta}'_z})(\boldsymbol{x}), \tag{16}$$

where the composition $f_{\boldsymbol{\theta}_f}^{-1} \circ f_{\boldsymbol{\theta}'_f}$ can transform one latent distribution into another. Since a general invertible transformation can map any distribution to a uniform via its CDF, the composition is flexible enough to arbitrarily create or destroy mixture components. This makes K unidentifiable under the following construction:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left((f_{\boldsymbol{\theta}}^{-1} \circ f_{\boldsymbol{\theta}'})_{\#}p_{\boldsymbol{\theta}'_z}\right)(\boldsymbol{x}), \quad \text{with } (f_{\boldsymbol{\theta}}^{-1} \circ f_{\boldsymbol{\theta}'}) = F_{p_{\boldsymbol{\theta}'_z}}^{-1} \circ F_{p_{\boldsymbol{\theta}_z}}, \tag{17}$$

where $F^{-1}_{p_{\boldsymbol{\theta}'_z}}$ and $F_{\boldsymbol{\theta}_z}$ are the inverse CDF and the CDF of $p_{\boldsymbol{\theta}'_z}$ and $p_{\boldsymbol{\theta}_z}$, respectively. To solve the above problem, Kivva et al. [2022] proves identifiability of mixture priors under the assumption of piecewise linear $f_{\boldsymbol{\theta}}$. This imposes restrictions on $f_{\boldsymbol{\theta}}$, but enables identifiability while maintaining a flexible GMM prior. Comparing to Khemakhem et al. [2020], which instead constrains the prior, this highlights a "no free lunch" principle for identifiability: without restricting the generative mapping or the latent distribution, the model is fundamentally unidentifiable.

B Proofs and Extended Discussions on Identifiability

B.1 Proof of Theorem 3.1

Proof. **Notation.** Consider two models $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with emission maps $f_{\boldsymbol{\theta}_f}$ and $f_{\boldsymbol{\theta}'_f}$, respectively. For simplicity, write $f := f_{\boldsymbol{\theta}_f},\ f' := f_{\boldsymbol{\theta}'_f}$. For the first model, define $\boldsymbol{w}_t = (\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = f^{-1}(\boldsymbol{x}_t)$, $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m}$. For the second model, define $\boldsymbol{w}'_t = (\boldsymbol{z}'_t, \boldsymbol{\varepsilon}'_t) = f'^{-1}(\boldsymbol{x}_t)$, $\boldsymbol{z}'_t = f'^{-1}(\boldsymbol{x}_t)_{1:m'}$. Denote by $g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t) := \mathcal{N}(\boldsymbol{z}_{t+1}; \boldsymbol{m}_k(\boldsymbol{z}_t), \Sigma_k)$ the $k-$th latent regime transition density, and $\tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{z}_t) = g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t) p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{t+1})$. Since $\boldsymbol{\varepsilon}_{t+1}$ is independent of $\boldsymbol{w}_t$, we write $\tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{z}_t)$ instead of $\tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{w}_t)$. Define the pushforward measure of $\tilde{g}_k$ to observation space as $\phi_k(\cdot \mid \boldsymbol{x}_t) = f_{\#}\tilde{g}_k(\cdot \mid \boldsymbol{z}_t)$ with $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m}$, and define $g'_{k'}$, $\tilde{g}'_{k'}$, and $\phi'_{k'}$ analogously for the second model, using f' and $p'_{\boldsymbol{\varepsilon}}$.

Step I. For two parametrisations $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{1:T})$, their marginal densities on $\boldsymbol{x}_{1:t}$ are equal for every $t \leq T$. Using conditional densities implies

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t}), \tag{18}$$

for every $1 \leq t < T$ and almost any $\boldsymbol{x}_{1:t} \in \mathbb{R}^{tn}$.

Now consider model $\boldsymbol{\theta}$, and fix a history $\boldsymbol{h} = \boldsymbol{x}_{1:t-1}$. Since f is invertible, conditioning on $(\boldsymbol{x}_t, \boldsymbol{h})$ is equivalent to conditioning on $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t, \boldsymbol{z}_{1:t-1}, \boldsymbol{\varepsilon}_{1:t-1})$ through f^{-1}. Since the temporal dynamics are independent of the exogenous noise, the conditional distribution of $\boldsymbol{w}_t$ depends on the history only through $(\boldsymbol{z}_t, \tilde{\boldsymbol{h}})$, where $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m}$ and $\tilde{\boldsymbol{h}} = \boldsymbol{z}_{1:t-1}$. Therefore,

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t, \boldsymbol{h}) &= \left|\det J_{f^{-1}}(\boldsymbol{x}_{t+1})\right| \, p_{\boldsymbol{\theta}}(\boldsymbol{w}_{t+1} \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}) && (19) \\ &= \left|\det J_{f^{-1}}(\boldsymbol{x}_{t+1})\right| \sum_{k=1}^{K} p_{\boldsymbol{\theta}}(s_{t+1} = k \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}) \, \tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{z}_t) && (20) \\ &= \sum_{k=1}^{K} p_{\boldsymbol{\theta}}(s_{t+1} = k \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}) \, \phi_k(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t). && (21) \end{aligned}$$

Denote $\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) := (\phi_1(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t), \ldots, \phi_K(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t))^\top$, and let $\boldsymbol{H} = (\tilde{\boldsymbol{h}}^{(1)}, \ldots, \tilde{\boldsymbol{h}}^{(K)})$ be a collection of K observed histories, where $\boldsymbol{h}^{(i)} = \boldsymbol{x}^{(i)}_{1:t-1}$. For model $\boldsymbol{\theta}$, denote $\tilde{\boldsymbol{h}}^{(i)}$ the i-th latent history under f^{-1}, and define:

$$\boldsymbol{P}_{\boldsymbol{\theta},\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) := \begin{pmatrix} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{h}^{(1)}) \\ \vdots \\ p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{h}^{(K)}) \end{pmatrix}. \tag{22}$$

Also define

$$A_{\boldsymbol{H}}(\boldsymbol{z}_t) := \begin{pmatrix} p_{\boldsymbol{\theta}}(s_{t+1} = 1 \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(1)}) & \ldots & p_{\boldsymbol{\theta}}(s_{t+1} = K \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(1)}) \\ \vdots & \ddots & \vdots \\ p_{\boldsymbol{\theta}}(s_{t+1} = 1 \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(K)}) & \ldots & p_{\boldsymbol{\theta}}(s_{t+1} = K \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(K)}) \end{pmatrix}. \tag{23}$$

Then, we can write the observed conditional densities as mixtures over shared one-step observational components $\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$, with mixture weights indexed by $\boldsymbol{H}$:

$$\boldsymbol{P}_{\boldsymbol{\theta},\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = A_{\boldsymbol{H}}(\boldsymbol{z}_t)\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t). \tag{24}$$

For two parametrisations $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ where $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{1:t})$, under the same set of histories $\boldsymbol{H}$, we define $\boldsymbol{P}_{\boldsymbol{\theta}',\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ and $A'_{\boldsymbol{H}}(\boldsymbol{z}'_t)$ analogously to $\boldsymbol{P}_{\boldsymbol{\theta},\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ and $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$,

respectively. Note we use the same observed histories $\boldsymbol{H}$ for both models, which are model-dependent through f^{-1} and $(f')^{-1}$. Then, under $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = p_{\boldsymbol{\theta}'}(\boldsymbol{x}_{1:T})$, we have

$$A_{\boldsymbol{H}}(\boldsymbol{z}_t)\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \boldsymbol{P}_{\boldsymbol{\theta},\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \boldsymbol{P}_{\boldsymbol{\theta}',\boldsymbol{H}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = A'_{\boldsymbol{H}}(\boldsymbol{z}'_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t). \tag{25}$$

We note that, for the same collection of K observed histories $\boldsymbol{H}$ and a model $\boldsymbol{\theta}'$ with K' regimes, the matrix $A'_{\boldsymbol{H}}$ has size $K \times K'$, while $\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ has length K'.

We now show that $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible and continuous on a nonzero measure set in $\mathbb{R}^m$. By first-order Markov switching dynamics, row i of $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is the one-step predictive distribution of s_{t+1} given the latent history $\tilde{\boldsymbol{h}}^i$ ending at $\boldsymbol{z}_t$. Therefore, $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ factorises as follows:

$$A_{\boldsymbol{H}}(\boldsymbol{z}_t)_{ik} = p_{\boldsymbol{\theta}}(s_{t+1} = k \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(i)}) = \sum_{j=1}^{K} p_{\boldsymbol{\theta}}(s_t = j \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(i)}) Q_{jk}(\boldsymbol{z}_t), \tag{26}$$

$$A_{\boldsymbol{H}}(\boldsymbol{z}_t) = F_{\boldsymbol{H}}(\boldsymbol{z}_t)Q(\boldsymbol{z}_t), \tag{27}$$

where row i of $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is the filtering posterior $(p_{\boldsymbol{\theta}}(s_t = 1 \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(i)}), \ldots, p_{\boldsymbol{\theta}}(s_t = K \mid \boldsymbol{z}_t, \tilde{\boldsymbol{h}}^{(i)}))$ ending at $\boldsymbol{z}_t$. We note that exogenous noise terms disappear from the above equation due to conditioning.

From (A4–A5), there exists a nonzero measure set $\mathcal{Z}$ on which we can choose $\tilde{\boldsymbol{h}}^{(i)} = (\boldsymbol{z}_1^{(i)}, \ldots, \boldsymbol{z}_{t-1}^{(i)})$, with $\boldsymbol{z}_\tau^{(i)} \in \mathcal{Z}$, such that Lemma B.1 gives $p(s_t = i|\boldsymbol{z}_t^*, \tilde{\boldsymbol{h}}^{(i)}) > \frac{1}{2}$ for large enough $t > 1$ and some $\boldsymbol{z}_t^* \in \mathcal{Z}$. Using Lemma B.1 for $i \in \{1, \ldots, K\}$, we can order histories such that $F_{\boldsymbol{H}}(\boldsymbol{z}_t^*)_{ii} > \frac{1}{2}$, i.e., history i dominates the filtering distribution on regime i. Since each row of $F_{\boldsymbol{H}}(\boldsymbol{z}_t^*)$ sums to one, this implies strict row diagonal dominance, and therefore, $F_{\boldsymbol{H}}(\boldsymbol{z}_t^*)$ is invertible by the Levy–Desplanques theorem [Wu, 2005]. Similarly by (A4), $Q_{kk}(\boldsymbol{z}_t) > 1/2$ for $\boldsymbol{z}_t \in \mathcal{Z}$, which means $Q(\boldsymbol{z}_t)$ is also strictly row diagonally dominant and invertible. Therefore, $A_{\boldsymbol{H}}(\boldsymbol{z}_t^*)$ is invertible. Now we extend invertibility to a nonzero measure set $\mathcal{U} \subset \mathbb{R}^m$. For each fixed history $\tilde{\boldsymbol{h}}^{(i)}$ with a common endpoint $\boldsymbol{z}_t^*$, define

$$M_i(\boldsymbol{z}_t^*) = \log g_i(\boldsymbol{z}_t^* \mid \boldsymbol{z}_{t-1}^{(i)}) - \max_{j \neq i} \log g_j(\boldsymbol{z}_t^* \mid \boldsymbol{z}_{t-1}^{(i)}). \tag{28}$$

The functions in M_i are continuous given Gaussianity, and since $M_i(\boldsymbol{z}_t^*) > \ell^*$, there exists a neighbourhood $\mathcal{U}_i \subset \mathcal{Z}$ of $\boldsymbol{z}_t^*$ such that

$$M_i(\boldsymbol{z}_t) > \ell^*, \quad \forall \boldsymbol{z}_t \in \mathcal{U}_i,\ \forall i \in \{1, \ldots, K\}. \tag{29}$$

Let $\mathcal{U} = \cap_{i=1}^K \mathcal{U}_i$, which is a non-empty set since each $\mathcal{U}_i$ is a neighbourhood of the same endpoint $\boldsymbol{z}_t^*$. Then from Eq. (29), all K histories satisfy the same conditions of Lemma B.1 for every $\boldsymbol{z}_t \in \mathcal{U}$. Therefore, the same strict diagonal dominance argument applies, and $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible for any $\boldsymbol{z}_t \in \mathcal{U}$. The filtering equation involves finite sums with Gaussians $g_k(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1})$, which implies $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is continuous on $\mathcal{U}$. Since $Q(\boldsymbol{z}_t)$ is continuous on $\mathcal{Z}$, and $\mathcal{U} \subset \mathcal{Z}$, it follows that $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible and continuous for any $\boldsymbol{z}_t \in \mathcal{U}$, where $\mathcal{U}$ has nonzero measure.

Case (ii): Unknown K and analytic $Q(\boldsymbol{z}_t)$. The previous argument gives a nonzero measure set $\mathcal{U} \subset \mathbb{R}^m$ where $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible. This implies $\det F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is not identically zero. For fixed histories $\boldsymbol{H}$, the filtering probabilities are finite sums and products of Gaussians, and therefore $\det F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is analytic. From Mityagin [2015], its zero set has measure zero, and this implies $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible almost everywhere. When assuming $Q(\boldsymbol{z}_t)$ is analytic, the same argument applies. By (A4), $Q(\boldsymbol{z}_t)$ is strictly row diagonally dominant, and hence invertible on the non-empty open set $\mathcal{Z}$. Therefore, $\det Q(\boldsymbol{z}_t)$ is analytic and not identically zero, which implies $Q(\boldsymbol{z}_t)$ is invertible almost everywhere. Given the above, we have that $A_{\boldsymbol{H}}(\boldsymbol{z}_t) = F_{\boldsymbol{H}}(\boldsymbol{z}_t)Q(\boldsymbol{z}_t)$ is invertible for almost any $\boldsymbol{z}_t \in \mathbb{R}^m$.

Step II. On the nonzero set $\mathcal{U}$ where $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible, the identity from Eq. (25) gives

$$\Phi(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t) = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t), \qquad \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) := A_{\boldsymbol{H}}(\boldsymbol{z}_t)^{-1} A'_{\boldsymbol{H}}(\boldsymbol{z}'_t), \tag{30}$$

where $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m}$ and $\boldsymbol{z}'_t = f'^{-1}(\boldsymbol{x}_t)_{1:m'}$. To clarify, for model $\boldsymbol{\theta}'$ with K' regimes, $A'_{\boldsymbol{H}}(\boldsymbol{z}'_t) \in \mathbb{R}^{K \times K'}$ and $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) \in \mathbb{R}^{K \times K'}$. Integrating over $\boldsymbol{x}_{t+1}$ gives:

$$\int \Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) d\boldsymbol{x}_{t+1} = \int \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) d\boldsymbol{x}_{t+1} \implies \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\mathbf{1} = \mathbf{1}. \tag{31}$$

We can also show $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is full row rank using linear independence properties of the Gaussian family. We start by writing a linear independence equation for the rows of $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ in vector form:

$$c^\top \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) = 0, \qquad c \in \mathbb{R}^K. \tag{32}$$

where $c = 0$ if and only if the rows of $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ are linearly independent. Multiplying by $\Phi'(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)$ and using (30), we obtain, for every $\boldsymbol{x}_{t+1} \in \mathbb{R}^n$,

$$\begin{aligned} 0 = \big(c^\top \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\big)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = c^\top \big(\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\big) \\ = c^\top \Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \sum_{k=1}^K c_k \phi_k(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t), \end{aligned}$$

where since the Gaussian family is linearly independent [Yakowitz and Spragins, 1968] and f is injective, their pushforwards $\{\phi_1(\cdot \mid \boldsymbol{x}_t), \ldots, \phi_K(\cdot \mid \boldsymbol{x}_t)\}$ are also linearly independent. Thus $c = 0$, which implies $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is full row rank. If model $\boldsymbol{\theta}$ has K regimes, then $K \le K'$. At this stage, the proof branches according to (i) known K, and (ii) unknown K, but analytic $Q(\boldsymbol{z}_{t-1})$.

(i) For known K, $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) \in \mathbb{R}^{K \times K}$ is square and, since it has full row rank, it is invertible.

(ii) Unknown K but analytic Q: By the analytic extension argument on the last paragraph of Step **I**, $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible almost everywhere. We now apply the same argument to model $\boldsymbol{\theta}'$ using (A4–A5) and Lemma B.1. Therefore, we choose a different collection of K' observed histories $\boldsymbol{H}'$ such that $A'_{\boldsymbol{H}'}(\boldsymbol{z}'_t)$ is invertible almost everywhere. Given f and f' are invertible, we can choose the same $\boldsymbol{x}_t$ such that both $A'_{\boldsymbol{H}'}(\boldsymbol{z}'_t)$ and $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ are invertible. For this $\boldsymbol{x}_t$, we have

$$\Phi'(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t) = \Pi'_{\boldsymbol{H}'}(\boldsymbol{x}_t)\Phi(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t), \qquad \Pi'_{\boldsymbol{H}'}(\boldsymbol{x}_t) := \big(A'_{\boldsymbol{H}'}(\boldsymbol{z}'_t)\big)^{-1} A_{\boldsymbol{H}'}(\boldsymbol{z}_t), \tag{33}$$

where $\Pi'_{\boldsymbol{H}'}(\boldsymbol{x}_t) \in \mathbb{R}^{K' \times K}$. Under the same linear independence arguments on model $\boldsymbol{\theta}'$, $\Pi'_{\boldsymbol{H}'}(\boldsymbol{x}_t) \in \mathbb{R}^{K' \times K}$ is full-row rank. Therefore, $K' \le K$. Combined with the previous inequality $K \le K'$, this yields $K = K'$ and hence $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is square and invertible.

Now we show $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is a permutation. Under (A2) and (A6) for model $\boldsymbol{\theta}'$, by Proposition B.4 and Lemma B.3, each element in $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is nonnegative. Similarly, using the same result on model $\boldsymbol{\theta}$, and using

$$\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t), \tag{34}$$

each element in $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}$ is also nonnegative. Furthermore, integrating over $\boldsymbol{x}_{t+1}$ again gives $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}\mathbf{1} = \mathbf{1}$. Therefore, we know

$$\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) \ge 0, \quad \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1} \ge 0, \quad \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\mathbf{1} = \mathbf{1}. \tag{35}$$

We prove $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ has at most one positive entry by contradiction. Take any row i, and assume it has two positive entries in columns $j \neq j'$

$$\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij} > 0, \quad \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij'} > 0. \tag{36}$$

Given $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1} = I_K$, for some $k \neq i$ we have:

$$0 = \big(\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}\big)_{ik} = \sum_{j=1}^K \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij}\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{jk}. \tag{37}$$

Given both matrices have nonnegative entries, for $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij} > 0$ and $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij'} > 0$, we have $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{jk} = 0$ and $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{j'k} = 0$ for every $k \neq i$. Therefore rows j and j' of $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}$ can only have nonnegative entries on column i. Since $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}$ is invertible, neither row can be zero, therefore

$$\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{ji} > 0, \quad \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{j'i} > 0. \tag{38}$$

Now consider $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) = I_K$:

$$0 = \big(\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\big)_{jj'} = \sum_{k=1}^K \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{jk}\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{kj'} = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)^{-1}_{ji}\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij'}. \tag{39}$$

But $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ji}^{-1} > 0$ implies $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ij'} = 0$, which is a contradiction. Therefore, every row of $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ has at most one positive entry, and since it is invertible, every column has also at most one positive entry. Finally, the condition $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\mathbf{1} = \mathbf{1}$, forces $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ to be a permutation matrix, as each nonzero entry must be 1.

Finally, we show $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is locally constant by continuity. Denote by $\mathcal{V} \subseteq \mathbb{R}^n$ a nonzero measure set of points $\boldsymbol{x}_t$ such that $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m} \in \mathcal{U}$, where $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ is invertible on $\mathcal{U}$, and $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) = A_{\boldsymbol{H}}(\boldsymbol{z}_t)^{-1} A'_{\boldsymbol{H}}(\boldsymbol{z}'_t)$ with $\boldsymbol{z}_t = f^{-1}(\boldsymbol{x}_t)_{1:m}$ and $\boldsymbol{z}'_t = f'^{-1}(\boldsymbol{x}_t)_{1:m'}$. Since $F_{\boldsymbol{H}}(\boldsymbol{z}_t)$ and $F'_{\boldsymbol{H}}(\boldsymbol{z}'_t)$ are continuous, and each Q and Q' are continuous almost everywhere from (A3), both $A_{\boldsymbol{H}}(\boldsymbol{z}_t)$ and $A'_{\boldsymbol{H}}(\boldsymbol{z}'_t)$ are continuous. In case (ii), continuity follows directly from the analyticity of Q and Q'.

Therefore, $\Pi_{\boldsymbol{H}}(\boldsymbol{x})$ is also continuous and a permutation for any $\boldsymbol{x}_t \in \mathcal{V}$. This implies there exists a nonzero measure subset $\mathcal{V}_0 \subseteq \mathcal{V}$ on which $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ is constant, as otherwise any change in the permutation would induce a discontinuity in $\mathcal{V}_0 \subseteq \mathcal{V}$. Therefore, for every $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t)$ such that $f(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) \in \mathcal{V}_0$, there exists a fixed permutation $\pi \in S_K$, such that for any $k \in \{1, \dots, K\}$:

$$\phi_k(\boldsymbol{x}_{t+1} \mid f(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t)) = \phi'_{\pi(k)}(\boldsymbol{x}_{t+1} \mid f(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t)). \tag{40}$$

Step III. Now we finish the proof using an iVAE-style argument, with derivatives instead of auxiliary variables [Khemakhem et al., 2020]. Given that f is a $\mathcal{C}^1$ diffeomorphism, there exists a latent reparametrisation such that:

$$\boldsymbol{w}'_t = (\boldsymbol{z}'_t, \boldsymbol{\varepsilon}'_t) = (f'^{-1} \circ f)(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = \psi(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = \psi(\boldsymbol{w}_t), \tag{41}$$

with $\psi : \mathbb{R}^{m+d} \to \mathbb{R}^{m'+d'}$ invertible. By Step **II**, there exists a nonzero measure set $\mathcal{V}_0 \subseteq \mathbb{R}^n$ and a permutation $\pi \in S_K$ such that, for every $k \in \{1, \dots, K\}$, and every $\boldsymbol{w}_t = (\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t)$ with $f(\boldsymbol{w}_t) \in \mathcal{V}_0$, the reparametrisation $\boldsymbol{w}'_t = \psi(\boldsymbol{w}_t)$ gives the following relation:

$$\tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{w}_t) = |\det J_\psi(\boldsymbol{w}_{t+1})| \tilde{g}'_{\pi(k)}(\psi(\boldsymbol{w}_{t+1}) \mid \psi(\boldsymbol{w}_t)) \tag{42}$$

Taking logarithms and differentiating first with respect to $\boldsymbol{w}_t$ followed by $\boldsymbol{w}_{t+1}$, the Jacobian term vanishes because it depends only on $\boldsymbol{w}_{t+1}$. Thus,

$$H_k(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) = J_\psi(\boldsymbol{w}_{t+1})^\top H'_{\pi(k)}\big(\psi(\boldsymbol{w}_{t+1}), \psi(\boldsymbol{w}_t)\big) J_\psi(\boldsymbol{w}_t), \tag{43}$$

where

$$H_k(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) := \nabla_{\boldsymbol{w}_{t+1}} \nabla_{\boldsymbol{w}_t}^\top \log \tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{w}_t), \tag{44}$$

$$H'_{k'}(\boldsymbol{w}'_{t+1}, \boldsymbol{w}'_t) := \nabla_{\boldsymbol{w}'_{t+1}} \nabla_{\boldsymbol{w}'_t}^\top \log \tilde{g}'_{k'}(\boldsymbol{w}'_{t+1} \mid \boldsymbol{w}'_t). \tag{45}$$

Since $\log \tilde{g}_k(\boldsymbol{w}_{t+1} \mid \boldsymbol{w}_t) = \log g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t) + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{t+1})$, and $\log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{t+1})$ does not depend on $\boldsymbol{w}_t$, it contributes zero to the mixed derivative. Therefore

$$H_k(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) = \begin{pmatrix} \Sigma_k^{-1} J_{\boldsymbol{m}_k}(\boldsymbol{z}_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad H'_{k'}(\boldsymbol{w}'_{t+1}, \boldsymbol{w}'_t) = \begin{pmatrix} \Sigma_{k'}'^{-1} J_{\boldsymbol{m}'_{k'}}(\boldsymbol{z}'_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{46}$$

We first identify the latent dimension m. Given Eq. (46),

$$\operatorname{rank}(H_k(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t)) = \operatorname{rank}(J_{\boldsymbol{m}_k}(\boldsymbol{z}_t)) \leq m, \ \operatorname{rank}(H'_{k'}(\boldsymbol{w}'_{t+1}, \boldsymbol{w}'_t)) = \operatorname{rank}(J_{\boldsymbol{m}'_{k'}}(\boldsymbol{z}'_t)) \leq m'.$$

Under (A7), there exist $k_0 \in \{1, \dots, K\}$ and $\boldsymbol{z}^* \in \mathbb{R}^m$ such that $\det J_{\boldsymbol{m}_{k_0}}(\boldsymbol{z}^*) \neq 0$. Since $\boldsymbol{m}_{k_0}$ is an analytic function, the function $\det J_{\boldsymbol{m}_{k_0}}$ is analytic and not identically zero. Therefore, its zero set has measure zero [Mityagin, 2015]. Therefore, for almost every $\boldsymbol{w}_t = (\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) \in f^{-1}(\mathcal{V}_0)$, $\det J_{\boldsymbol{m}_{k_0}}(\boldsymbol{z}_t) \neq 0$, and $\operatorname{rank}(J_{\boldsymbol{m}_{k_0}}(\boldsymbol{z}_t)) = m$. Fix $\boldsymbol{w}_t \in f^{-1}(\mathcal{V}_0)$. Since J_ψ is invertible, Eq. (43) preserves rank, and therefore

$$m = \operatorname{rank}\big(H_{k_0}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t)\big) = \operatorname{rank}\big(H'_{\pi(k_0)}(\boldsymbol{w}'_{t+1}, \boldsymbol{w}'_t)\big) \leq m'.$$

Equivalently, using (A7) on model $\boldsymbol{\theta}'$ yields $m' \leq m$. Therefore, $m = m'$, which implies $d = d'$.

We now show that the temporal component of the reparametrisation ψ depends only on temporal features $\boldsymbol{z}_t$. Write the reparametrisation as follows

$$\psi(\boldsymbol{z}, \boldsymbol{\varepsilon}) = \big(\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon}), \psi_{\boldsymbol{\varepsilon}}(\boldsymbol{z}, \boldsymbol{\varepsilon})\big), \quad \psi_{\boldsymbol{z}} : \mathbb{R}^{m+d} \to \mathbb{R}^m, \quad \psi_{\boldsymbol{\varepsilon}} : \mathbb{R}^{m+d} \to \mathbb{R}^d$$

For $k = k_0$, the null space (or kernel) of matrix H_{k_0} is $\ker H_{k_0}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) = \{\mathbf{0}\} \times \mathbb{R}^d$ and similarly for $H'_{\pi(k_0)}$. We now choose any $\boldsymbol{v} \in \ker H_{k_0}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t)$. From Eq. (43) and using properties of the null space, we have

$$0 = H_{k_0}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t)\boldsymbol{v} = J_\psi(\boldsymbol{w}_{t+1})^\top H'_{\pi(k_0)}\big(\psi(\boldsymbol{w}_{t+1}), \psi(\boldsymbol{w}_t)\big) J_\psi(\boldsymbol{w}_t)\boldsymbol{v}, \tag{47}$$

where since $J_\psi(\boldsymbol{w}_{t+1})^\top$ is invertible (from invertibility of ψ), the vector it multiplies must be zero:

$$H'_{\pi(k_0)}\big(\psi(\boldsymbol{w}_{t+1}), \psi(\boldsymbol{w}_t)\big) J_\psi(\boldsymbol{w}_t)\boldsymbol{v} = 0. \tag{48}$$

Therefore, $J_\psi(\boldsymbol{w}_t)\boldsymbol{v} \in \ker H'_{\pi(k_0)}(\psi(\boldsymbol{w}_{t+1}), \psi(\boldsymbol{w}_t))$. We know $\boldsymbol{v} = (\mathbf{0}, \boldsymbol{u})^\top$, with arbitrary $\boldsymbol{u} \in \mathbb{R}^d$. Using the Eq. 47, for every $\boldsymbol{u} \in \mathbb{R}^d$, there exists some $\boldsymbol{u}' \in \mathbb{R}^d$ such that:

$$\begin{pmatrix} \nabla_{\boldsymbol{z}}\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) & \nabla_{\boldsymbol{\varepsilon}}\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) \\ \nabla_{\boldsymbol{z}}\psi_{\boldsymbol{\varepsilon}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) & \nabla_{\boldsymbol{\varepsilon}}\psi_{\boldsymbol{\varepsilon}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{u}' \end{pmatrix} \implies \nabla_{\boldsymbol{\varepsilon}}\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon})\boldsymbol{u} = \mathbf{0}. \tag{49}$$

Therefore $\nabla_{\boldsymbol{\varepsilon}}\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) = \mathbf{0}$, which implies $\psi_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\varepsilon}) = \psi_{\boldsymbol{z}}(\boldsymbol{z})$. We note from Eq. (49), $J_\psi(\boldsymbol{w}_t)$ is block lower triangular, and since ψ is invertible, the diagonal block $J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z})$ is invertible. Now, take the top-left $m \times m$ block of Eq. (43). For any regime k, we obtain

$$\Sigma_k^{-1} J_{\boldsymbol{m}_k}(\boldsymbol{z}_t) = J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z}_{t+1})^\top \Sigma'^{-1}_{\pi(k)} J_{\boldsymbol{m}'_{\pi(k)}}(\psi_{\boldsymbol{z}}(\boldsymbol{z}_t)) J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z}_t). \tag{50}$$

Now, we use the regime k'_0 in model $\boldsymbol{\theta}'$ that satisfies (A7), and set $\hat{k} = \pi^{-1}(k'_0)$ for convenience. Using similar arguments on analyticity of $\boldsymbol{m}'_{k'_0}$, $J_{\boldsymbol{m}'_{k'_0}}(\psi_{\boldsymbol{z}}(\boldsymbol{z}_t))$ is invertible almost everywhere. The matrices on the right-hand side are invertible almost everywhere. Rearranging yields,

$$J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z}_{t+1})^\top = \Sigma_{\hat{k}}^{-1} J_{\boldsymbol{m}_{\hat{k}}}(\boldsymbol{z}_t) J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z}_t)^{-1} J_{\boldsymbol{m}'_{k'_0}}(\psi_{\boldsymbol{z}}(\boldsymbol{z}_t))^{-1} \Sigma'_{k'_0},$$

where the right-hand side does not depend on $\boldsymbol{z}_{t+1}$. Therefore, $J_{\psi_{\boldsymbol{z}}}(\boldsymbol{z}_{t+1})$ is constant in $\boldsymbol{z}_{t+1}$, for $\boldsymbol{z}_t$ and $\boldsymbol{\varepsilon}_t$ such that $\boldsymbol{w}_t = (\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) \in f^{-1}(\mathcal{V}_0)$. Since the Gaussian transition density has full support on $\mathbb{R}^m$, this identity extends to all $\boldsymbol{z}_{t+1} \in \mathbb{R}^m$. Therefore $\psi_{\boldsymbol{z}}(\boldsymbol{z}_t)$ is affine:

$$\boldsymbol{z}'_t = \psi_{\boldsymbol{z}}(\boldsymbol{z}_t) = A\boldsymbol{z}_t + \boldsymbol{b}, \quad A \in \mathbb{R}^{m \times m}, \quad \boldsymbol{b}, \boldsymbol{z}_t \in \mathbb{R}^m, \tag{51}$$

Here A denotes the affine reparametrisation matrix and should not be confused with the history-dependent matrix $A_{\boldsymbol{H}}$ used above. Since $\psi_{\boldsymbol{z}}(\boldsymbol{z}_t)$ is invertible, A is invertible.

□

B.2 Discussion on Assumptions (A4–A5)

As shown in Step **I** of the proof of Theorem 3.1, Assumptions (A4–A5) are crucial to show invertibility of the linear system obtained by stacking K histories. Below, we provide a concentration result showing that histories aligned with a given regime induce filtering posteriors dominated by that regime.

Lemma B.1. *Consider a regime path $(k_1^*, \ldots, k_t^*)$, and latent variables $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t)$, such that for all $\tau = 2, \ldots, t$,*

$$Q_{k^*_{\tau-1}, k^*_\tau}(\boldsymbol{z}_{\tau-1}) = \max_{k' \in \{1,\ldots,K\}} Q_{k^*_{\tau-1}, k'}(\boldsymbol{z}_{\tau-1}) \geq 1 - \epsilon^*, \tag{52}$$

for some $\epsilon^* \in [0, \frac{1}{2})$. Denote $g_k(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}_k(\boldsymbol{z}_{t-1}), \Sigma_k)$ and assume

$$\frac{g_{k^*_\tau}(\boldsymbol{z}_\tau | \boldsymbol{z}_{\tau-1})}{\max_{k' \neq k^*_\tau} g_{k'}(\boldsymbol{z}_\tau | \boldsymbol{z}_{\tau-1})} \geq e^{\ell^*}, \tag{53}$$

for some $\ell^* > 0$. If

$$\ell^* > \log \frac{1 + \epsilon^*}{1 - \epsilon^*}, \tag{54}$$

then, there exists $T^* < +\infty$, such that for all $t \geq T^*$, we have

$$p(s_t = k_t^* | \boldsymbol{z}_{1:t}) > \frac{1}{2}. \tag{55}$$

Proof. For $\tau = 2, \ldots, t$, define the predictive and filtering posteriors respectively as follows:

$$\pi_\tau(k) = p(s_\tau = k|\boldsymbol{z}_{1:\tau-1}), \quad \alpha_\tau(k) = p(s_\tau = k|\boldsymbol{z}_{1:\tau}). \tag{56}$$

Using Bayes' rule:

$$\alpha_\tau(k^*_\tau) = \frac{\pi_\tau(k^*_\tau) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1})}{\pi_\tau(k^*_\tau) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}) + \sum_{k\neq k^*_\tau}^K \pi_\tau(k) g_k(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1})}, \tag{57}$$

where we find an upper bound to the denominator.

$$\sum_{k\neq k^*_\tau}^K \pi_\tau(k) g_k(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}) \leq \sum_{k\neq k^*_\tau}^K \pi_\tau(k) \max_{k\neq k^*_\tau} g_k(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}) \leq (1-\pi_\tau(k^*_\tau)) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}) e^{-\ell^*}. \tag{58}$$

Let R_τ be the posterior odds at time τ, defined as follows:

$$R_\tau := \frac{1-\alpha_\tau(k^*_\tau)}{\alpha_\tau(k^*_\tau)}. \tag{59}$$

Given $\alpha_\tau(k^*_\tau)$ and $1-\alpha_\tau(k^*_\tau)$ share the denominator, R_τ has the following simpler form:

$$R_\tau = \frac{\sum_{k\neq k^*_\tau}^K \pi_\tau(k) g_k(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1})}{\pi_\tau(k^*_\tau) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1})} \leq \frac{(1-\pi_\tau(k^*_\tau)) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}) e^{-\ell^*}}{\pi_\tau(k^*_\tau) g_{k^*_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1})} = e^{-\ell^*} \frac{1-\pi_\tau(k^*_\tau)}{\pi_\tau(k^*_\tau)} \tag{60}$$

Now we want to find bounds for both $1-\pi_\tau(k^*_\tau)$ and $\pi_\tau(k^*_\tau)$. Consider the predictive posterior on the aligned state at time τ:

$$\pi_\tau(k^*_\tau) = \sum_{k=1}^K \alpha_{\tau-1}(k) Q_{k,k^*_\tau}(\boldsymbol{z}_{\tau-1}) \geq \alpha_{\tau-1}(k^*_{\tau-1}) Q_{k^*_{\tau-1},k^*_\tau}(\boldsymbol{z}_{\tau-1}) \geq \alpha_{\tau-1}(k^*_{\tau-1})(1-\epsilon^*). \tag{61}$$

Then, for $1-\pi_\tau(k^*_\tau)$:

$$1-\pi_\tau(k^*_\tau) \leq 1-\alpha_{\tau-1}(k^*_{\tau-1})(1-\epsilon^*) = 1-\alpha_{\tau-1}(k^*_{\tau-1}) + \alpha_{\tau-1}(k^*_{\tau-1})\epsilon^*, \tag{62}$$

and we can find a recursion for R_τ:

$$R_\tau \leq e^{-\ell^*} \frac{1-\alpha_{\tau-1}(k^*_{\tau-1}) + \alpha_{\tau-1}(k^*_{\tau-1})\epsilon^*}{\alpha_{\tau-1}(k^*_{\tau-1})(1-\epsilon^*)} = e^{-\ell^*} \left(\frac{R_{\tau-1}}{1-\epsilon^*} + \frac{\epsilon^*}{1-\epsilon^*} \right) \tag{63}$$

$$\leq R_{\tau-1} \frac{e^{-\ell^*}}{1-\epsilon^*} + \frac{e^{-\ell^*}\epsilon^*}{1-\epsilon^*} = aR_{\tau-1} + b, \tag{64}$$

where we define

$$a := \frac{e^{-\ell^*}}{1-\epsilon^*}, \quad b := \frac{e^{-\ell^*}\epsilon^*}{1-\epsilon^*}. \tag{65}$$

Therefore, for $R_1 > 0$, we know:

$$R_t \leq \left(\prod_{\tau=2}^t a \right) R_1 + \sum_{\tau=2}^t b \prod_{i=\tau+1}^t a = a^{t-1} R_1 + b \sum_{r=0}^{t-2} a^r. \tag{66}$$

We notice the sum on the right is a geometric sum, which simplifies as follows:

$$R_t \leq a^{t-1} R_1 + b \frac{1-a^{t-1}}{1-a}. \tag{67}$$

Now, given Eq. (54), we obtain $a < 1$:

$$\ell^* > \log \frac{1+\epsilon^*}{1-\epsilon^*} \implies \ell^* > \log \frac{1}{1-\epsilon^*} \implies a = \frac{e^{-\ell^*}}{1-\epsilon^*} < 1. \tag{68}$$

Therefore, the above expression admits a floor R_∞ defined as

$$R_\infty = \frac{b}{1-a} = \frac{e^{-\ell^*}\epsilon^*}{1-\epsilon^*-e^{-\ell^*}}. \tag{69}$$

Then,

$$R_t \le a^{t-1}R_1 + (1-a^{t-1})R_\infty = a^{t-1}(R_1 - R_\infty) + R_\infty. \tag{70}$$

Moreover,

$$\ell^* > \log\frac{1+\epsilon^*}{1-\epsilon^*} \implies e^{-\ell^*}(1+\epsilon^*) < 1-\epsilon^* \implies \frac{e^{-\ell^*}\epsilon^*}{1-\epsilon^*-e^{-\ell^*}} < 1 \implies R_\infty < 1. \tag{71}$$

Therefore, Eq. (54) implies $R_\infty < 1$. Since $a < 1$, the term $a^{t-1}(R_1 - R_\infty)$ vanishes as $t \to \infty$. Given $R_\infty < 1$, there exists $T^* < +\infty$ such that for all $t \ge T^*$, $R_t < 1$. Finally,

$$R_t < 1 \implies \frac{1-\alpha_t(k_t^*)}{\alpha_t(k_t^*)} < 1 \implies \alpha_t(k_t^*) > \frac{1}{2}. \tag{72}$$

Therefore, we have $p(s_t = k_t^* | \boldsymbol{z}_{1:t}) = \alpha_t(k_t^*) > \frac{1}{2}$ for $t \ge T^*$. □

The concentration result above shows that, given initial posterior odds R_1, a likelihood margin ℓ^*, and stickiness ϵ^*, there exists a finite time length T^* after which the posterior is dominated by the aligned regime. The sufficient length T^* is determined by R_1, ℓ^* and ϵ^*. Below, we provide the minimum length required for posterior dominance.

Corollary B.2. *Let $\boldsymbol{z}_{1:t}$ be a history from Lemma B.1, with associated regime path $(k_1^*, \dots, k_t^*)$. Define,*

$$a := \frac{e^{-\ell^*}}{1-\epsilon^*}, \qquad R_\infty := \frac{e^{-\ell^*}\epsilon^*}{1-\epsilon^*-e^{-\ell^*}}, \qquad R_1 = \frac{1-p(s_1 = k^* \mid \boldsymbol{z}_1)}{p(s_1 = k^* \mid \boldsymbol{z}_1)}, \tag{73}$$

Assume $R_1 > R_\infty$. Then any time index $t < +\infty$ which satisfies the equation below ensures $p(s_t = k_t^* | \boldsymbol{z}_{1:t}) > \frac{1}{2}$

$$t \;\ge\; 2 + \left\lfloor \log\!\left(\frac{1-R_\infty}{R_1 - R_\infty}\right)\frac{1}{\log a} \right\rfloor. \tag{74}$$

Proof. By Lemma B.1, $R_t \le a^{t-1}(R_1 - R_\infty) + R_\infty$. For $R_t < 1$, we have

$$a^{t-1}(R_1 - R_\infty) + R_\infty < 1 \implies a^{t-1}(R_1 - R_\infty) < 1 - R_\infty.$$

Since $R_1 > R_\infty$ and $R_\infty < 1$, dividing both sides gives

$$a^{t-1} < \frac{1-R_\infty}{R_1 - R_\infty}.$$

Because $0 < a < 1$, we have $\log a < 0$, so taking logarithms reverses the inequality:

$$t - 1 > \frac{\log\left(\frac{1-R_\infty}{R_1-R_\infty}\right)}{\log a} \implies t \;\ge\; 2 + \left\lfloor \log\!\left(\frac{1-R_\infty}{R_1 - R_\infty}\right)\frac{1}{\log a} \right\rfloor.$$

□

In Section 3.1, we provide a simple example where posterior dominance is possible after one transition. In the example, we assume uniform initial posterior $p(s_1 = k|\boldsymbol{z}_1) = \frac{1}{K},\ k \in \{1, \dots, K\}$. Therefore, the initial posterior odds are $R_1 = K - 1$. Substituting R_1, together with R_∞ and a into Corollary B.2 yields the one-step dominance condition

$$\ell^* > \log\frac{K-1+\epsilon^*}{1-\epsilon^*}.$$

(A5) under nonlinear Gaussian transitions (A2) To further illustrate the intuition behind Assumption A5, we contextualise it under the nonlinear Gaussian transitions in Assumption A2. In this setting, the likelihood margin $\ell^\star$ admits an explicit form.

For simplicity, assume equal isotropic transition variances across regimes, i.e. $\Sigma_k = \sigma^2 I$ for all $k \in \{1, \dots, K\}$. Then, for two regimes $i, k \in \{1, \dots, K\}$, the log-likelihood margin is

$$\ell_{i,k}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = \frac{1}{2\sigma^2}\left(\|\boldsymbol{z}_t - \boldsymbol{m}_k(\boldsymbol{z}_{t-1})\|^2 - \|\boldsymbol{z}_t - \boldsymbol{m}_i(\boldsymbol{z}_{t-1})\|^2\right). \tag{75}$$

Therefore, if a history is aligned with regime i, i.e. $\boldsymbol{z}_t = \boldsymbol{m}_i(\boldsymbol{z}_{t-1})$, the margin is reduced to a scaled distance between the functions.

$$\ell_{i,k}(\boldsymbol{z}_{t-1}) = \frac{1}{2\sigma^2} \left\| \boldsymbol{m}_i(\boldsymbol{z}_{t-1}) - \boldsymbol{m}_k(\boldsymbol{z}_{t-1}) \right\|^2. \tag{76}$$

This simplified expression further shows that while (A5) imposes certain history-alignment conditions, well-separated regime-specific transition functions naturally support posterior dominance. Notably, in high-dimensional settings, each well-separated coordinate contributes positively to the likelihood margin.

B.3 Discussion on Assumption (A6)

Recall Step **II** in the proof of Theorem 3.1, where we have a matrix $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ such that

$$\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$$

In this context, Assumption (A6) is used to show that the matrix $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ has nonnegative entries. First, we establish a general sufficient condition. The condition requires that each component asymptotically dominates all other components along some direction.

Lemma B.3. *Assume that for a fixed $\boldsymbol{x}_t$, we have*

$$\Phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)\Phi'(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t), \tag{77}$$

where $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t) \in \mathbb{R}^{K\times K}$, and Φ and Φ' are the one-step observed component densities defined in Theorem 3.1. Let $\boldsymbol{z}'_t = f'^{-1}(\boldsymbol{x}_t)_{1:m'}$. Assume that, for every regime $k \in \{1,\dots,K\}$, there exists a direction $\boldsymbol{d}'_k(\boldsymbol{z}'_t) \in \mathbb{R}^{m'}$ and a fixed $\boldsymbol{\varepsilon}'_{t+1} \in \mathbb{R}^{d'}$, such that with $\boldsymbol{x}^{(k)}_{t+1}(a) = f'\big(a\boldsymbol{d}'_k(\boldsymbol{z}'_t), \boldsymbol{\varepsilon}'_{t+1}\big)$, we have, for every $k' \neq k$,

$$\lim_{a\to\infty} \frac{\phi'_{k'}(\boldsymbol{x}^{(k)}_{t+1}(a) \mid \boldsymbol{x}_t)}{\phi'_k(\boldsymbol{x}^{(k)}_{t+1}(a) \mid \boldsymbol{x}_t)} = 0. \tag{78}$$

Then, $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ has nonnegative entries.

Proof. Take row i of Eq. (77) and divide by $\phi'_k(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)$:

$$\frac{\phi_i(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{\phi'_k(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)} = \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ik} + \sum_{k'\neq k}^{K} \Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ik'} \frac{\phi'_{k'}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{\phi'_k(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)} \tag{79}$$

Evaluate at $\boldsymbol{x}_{t+1} = \boldsymbol{x}^{(k)}_{t+1}(a) = f'\big(a\boldsymbol{d}'_k(\boldsymbol{z}'_t), \boldsymbol{\varepsilon}'_{t+1}\big)$, and take the limit $a \to \infty$. The ratio terms in the sum vanish:

$$\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ik} = \lim_{a\to\infty} \left(\frac{\phi_i(\boldsymbol{x}_{t+1}(a)|\boldsymbol{x}_t)}{\phi'_k(\boldsymbol{x}_{t+1}(a) \mid \boldsymbol{x}_t)} \right) \geq 0. \tag{80}$$

Therefore, $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)_{ik} \geq 0$. Since we choose arbitrary $i, k \in \{1,\dots,K\}$, all the entries in $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ are nonnegative. □

The lemma above establishes nonnegativity of $\Pi_{\boldsymbol{H}}(\boldsymbol{x}_t)$ under an asymptotic dominance condition on the observed components. We show that Assumption (A6) is sufficient for this dominance condition, and thus for applying Lemma B.3. First, we connect the dominance condition in Lemma B.3 to latent Gaussian transitions under Assumptions (A1–A2)

Proposition B.4. *Under Assumptions (A1–A2), fix $\boldsymbol{x}_t$ and write $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = f^{-1}(\boldsymbol{x}_t)$. A sufficient condition for the dominance condition in Lemma B.3, i.e., for regime $k \in \{1,\dots,K\}$ there exists $\boldsymbol{d}_k(\boldsymbol{z}_t) \in \mathbb{R}^m$ such that for any $k' \neq k$, Eq. (78) holds, is the following. For every regime $k \in \{1,\dots,K\}$, there exists a direction $\boldsymbol{d}_k(\boldsymbol{z}_t) \in \mathbb{R}^m$ such that, for every $k' \neq k$, either:*

(i) $\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) > 0$, *or*

(ii) $\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) = 0$ *and* $\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_k^{-1}\boldsymbol{m}_k(\boldsymbol{z}_t) - \Sigma_{k'}^{-1}\boldsymbol{m}_{k'}(\boldsymbol{z}_t)\right) > 0$.

Proof. We first connect the observed component dominance to the Gaussian component dominance. Fix $\boldsymbol{x}_t$ and write $(\boldsymbol{z}_t, \boldsymbol{\varepsilon}_t) = f^{-1}(\boldsymbol{x}_t)$. Given $k \in \{1,\dots,K\}$ and a direction $\boldsymbol{d}_k(\boldsymbol{z}_t)$ and $\boldsymbol{\varepsilon}_{t+1}$, write

$$\boldsymbol{x}^{(k)}_{t+1}(a) = f\big(a\boldsymbol{d}_k(\boldsymbol{z}_t), \boldsymbol{\varepsilon}_{t+1}\big).$$

By the change-of-variable rule, we have

$$\phi_k(\boldsymbol{x}_{t+1}^{(k)}(a) \mid \boldsymbol{x}_t) = \left|\det J_{f^{-1}(\boldsymbol{x}_{t+1}^{(k)}(a))}\right| g_k\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big) p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_{t+1}). \tag{81}$$

For any $k' \neq k \in \{1, \ldots, K\}$,

$$\frac{\phi_{k'}(\boldsymbol{x}_{t+1}^{(k)}(a) \mid \boldsymbol{x}_t)}{\phi_k(\boldsymbol{x}_{t+1}^{(k)}(a) \mid \boldsymbol{x}_t)} = \frac{g_{k'}\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)}{g_k\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)}. \tag{82}$$

Therefore, we continue by establishing dominance over Gaussian components. Consider the log ratio between two Gaussians $g_{k'}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)$ and $g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)$, and consider the terms dependent on $\boldsymbol{z}_{t+1}$.

$$\begin{aligned}\log \frac{g_{k'}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)}{g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)} = &-\frac{1}{2}\big(\boldsymbol{z}_{t+1} - \boldsymbol{m}_{k'}(\boldsymbol{z}_t)\big)^\top \Sigma_{k'}^{-1}\big(\boldsymbol{z}_{t+1} - \boldsymbol{m}_{k'}(\boldsymbol{z}_t)\big) \\ &+ \frac{1}{2}\big(\boldsymbol{z}_{t+1} - \boldsymbol{m}_k(\boldsymbol{z}_t)\big)^\top \Sigma_k^{-1}\big(\boldsymbol{z}_{t+1} - \boldsymbol{m}_k(\boldsymbol{z}_t)\big) + \mathcal{O}(1).\end{aligned} \tag{83}$$

Group quadratic and linear terms together.

$$\begin{aligned}\log \frac{g_{k'}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)}{g_k(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)} = &-\frac{1}{2}\boldsymbol{z}_{t+1}^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{z}_{t+1} \\ &- \boldsymbol{z}_{t+1}^\top \left(\Sigma_k^{-1}\boldsymbol{m}_k(\boldsymbol{z}_t) - \Sigma_{k'}^{-1}\boldsymbol{m}_{k'}(\boldsymbol{z}_t)\right) + \mathcal{O}(1).\end{aligned} \tag{84}$$

Now we take a ray with direction $\boldsymbol{d}_k(\boldsymbol{z}_t) \in \mathbb{R}^m$:$\boldsymbol{z}_{t+1} = a\boldsymbol{d}_k(\boldsymbol{z}_t)$, $a \in \mathbb{R}$. Since $\boldsymbol{x}_t$, and therefore $\boldsymbol{z}_t$, is fixed, $\boldsymbol{d}_k(\boldsymbol{z}_t)$ is fixed along this limit. We obtain

$$\begin{aligned}\log \frac{g_{k'}\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)}{g_k\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)} = &-\frac{a^2}{2}\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) \\ &- a\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_k^{-1}\boldsymbol{m}_k(\boldsymbol{z}_t) - \Sigma_{k'}^{-1}\boldsymbol{m}_{k'}(\boldsymbol{z}_t)\right) + \mathcal{O}(1).\end{aligned} \tag{85}$$

Since Eq. (78) converges to 0 for $a \to \infty$, we need the above equation to converge to $-\infty$. Under condition (i), the coefficient of the quadratic term is strictly positive, so the log ratio converges to $-\infty$. Under condition (ii), the quadratic term vanishes and the coefficient of the linear term is strictly positive, so the log ratio again converges to $-\infty$. Therefore,

$$\lim_{a \to \infty} \frac{g_{k'}\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)}{g_k\big(a\boldsymbol{d}_k(\boldsymbol{z}_t) \mid \boldsymbol{z}_t\big)} = 0. \tag{86}$$

Since the same direction $\boldsymbol{d}_k(\boldsymbol{z}_t)$ satisfies one of conditions (i)–(ii) for every $k' \neq k$, all ratios vanish simultaneously along the same ray. Equation (82) then gives Eq. (78). □

We finally connect dominance under nonlinear Gaussian transitions to Assumption (A6) through the following corollary. Recall that Assumption (A2) assumes diagonal covariances, denoted by $\Sigma_k = \mathrm{diag}(\boldsymbol{\sigma}_k^2)$, for each $k \in \{1, \ldots, K\}$.

Proposition B.5. *Under Assumptions (A2) and (A6), for every $\boldsymbol{z}_t \in \mathbb{R}^m$ and $k \in \{1, \ldots, K\}$, there exists a direction $\boldsymbol{d}_k(\boldsymbol{z}_t) \in \mathbb{R}^m$ such that, for every $k' \neq k$, either condition (i) or condition (ii) of Proposition B.4 holds.*

Proof. Fix $\boldsymbol{z}_t$ and regime k. First, consider the nonempty coordinate set $\mathcal{I}_k$ and suppose $\mathcal{T}_k = \emptyset$. Then, for every $k' \neq k$, there exists at least one coordinate $i \in \mathcal{I}_k$ such that $\sigma_{k,i} > \sigma_{k',i}$. We can simply choose $\boldsymbol{d}_k(\boldsymbol{z}_t)$ independently of $\boldsymbol{z}_t$:

$$\boldsymbol{d}_k(\boldsymbol{z}_t) = \sum_{i \in \mathcal{I}_k} \mathbf{e}_i, \tag{87}$$

where $\mathbf{e}_i$ is the canonical vector of the $i-$th dimension. Since the covariance matrices are diagonal, substituting this choice into condition (i) of Proposition B.4 gives, for any $k' \neq k$:

$$\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) = \sum_{i \in \mathcal{I}_k} \left[\left(\frac{1}{\sigma_{k',i}}\right)^2 - \left(\frac{1}{\sigma_{k,i}}\right)^2\right] > 0. \tag{88}$$

Every term in the sum is nonnegative because $\sigma_{k,i} \geq \sigma_{k',i}$ for every $i \in \mathcal{I}_k$, and at least one term is strictly positive because $k' \notin \mathcal{T}_k$. Therefore, condition (i) of Proposition B.4 holds for any $k' \neq k$.

Now suppose $\mathcal{T}_k \neq \emptyset$ and denote $C_k(\boldsymbol{z}_t) := \operatorname{conv}(\{\boldsymbol{m}_{k',\mathcal{I}_k}(\boldsymbol{z}_t) : k' \in \mathcal{T}_k\})$. Since $C_k(\boldsymbol{z}_t)$ is compact and convex from Eq. (7) in Assumption A6, $\boldsymbol{m}_{k,\mathcal{I}_k}(\boldsymbol{z}_t) \notin C_k(\boldsymbol{z}_t)$. The strict separation theorem [Bertsekas, 2009, Proposition 1.5.3] shows that there exists a nonzero vector $\boldsymbol{u}_k(\boldsymbol{z}_t) \in \mathbb{R}^{|\mathcal{I}_k|}$ and a scalar $b \in \mathbb{R}$ such that

$$\boldsymbol{u}_k(\boldsymbol{z}_t)^\top \boldsymbol{y} < b < \boldsymbol{u}_k(\boldsymbol{z}_t)^\top \boldsymbol{m}_{k,\mathcal{I}_k}(\boldsymbol{z}_t), \qquad \forall \boldsymbol{y} \in C_k(\boldsymbol{z}_t). \tag{89}$$

Since $\boldsymbol{m}_{k',\mathcal{I}_k}(\boldsymbol{z}_t) \in C_k(\boldsymbol{z}_t)$ for every $k' \in \mathcal{T}_k$, using $\boldsymbol{m}_{k',\mathcal{I}_k}(\boldsymbol{z}_t)$ on Eq. (89) yields

$$\boldsymbol{u}_k(\boldsymbol{z}_t)^\top \left(\boldsymbol{m}_{k,\mathcal{I}_k}(\boldsymbol{z}_t) - \boldsymbol{m}_{k',\mathcal{I}_k}(\boldsymbol{z}_t)\right) > 0, \qquad \forall k' \in \mathcal{T}_k. \tag{90}$$

Since the above inequality is strict and $\mathcal{I}_k$ is finite, we can perturb $\boldsymbol{u}_k(\boldsymbol{z}_t)$ such that all its entries are nonzero while preserving Eq. (90). We can now connect Assumption (A6) with condition (ii) of Proposition B.4. Choose $\boldsymbol{d}_k(\boldsymbol{z}_t)$ as follows.

$$d_{k,i}(\boldsymbol{z}_t) = \begin{cases} \sigma_{k,i}^2 u_{k,i}(\boldsymbol{z}_t), & i \in \mathcal{I}_k, \\ 0, & i \notin \mathcal{I}_k. \end{cases} \tag{91}$$

Consider $k' \in \mathcal{T}_k$, where by definition, $\sigma_{k,i} = \sigma_{k',i}$ for any $i \in \mathcal{I}_k$. Since $\boldsymbol{d}_{k,i}(\boldsymbol{z}_t) = 0$ if $i \notin \mathcal{I}_k$, the quadratic equation in condition (ii) from Proposition B.4 holds:

$$\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) = \sum_{i \in \mathcal{I}_k} \left[\left(\frac{d_{k,i}}{\sigma_{k',i}}\right)^2 - \left(\frac{d_{k,i}}{\sigma_{k,i}}\right)^2 \right] = 0. \tag{92}$$

Substituting our choice of $\boldsymbol{d}_k(\boldsymbol{z}_t)$ in the linear part of condition (ii) gives:

$$\begin{aligned} &\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_k^{-1} \boldsymbol{m}_k(\boldsymbol{z}_t) - \Sigma_{k'}^{-1} \boldsymbol{m}_{k'}(\boldsymbol{z}_t)\right) && (93) \\ &= \sum_{i \in \mathcal{I}_k} \sigma_{k,i}^2 u_{k,i}(\boldsymbol{z}_t) \left(\frac{m_{k,i}(\boldsymbol{z}_t)}{\sigma_{k,i}^2} - \frac{m_{k',i}(\boldsymbol{z}_t)}{\sigma_{k',i}^2} \right) && (94) \\ &= \sum_{i \in \mathcal{I}_k} u_{k,i}(\boldsymbol{z}_t) \Big(m_{k,i}(\boldsymbol{z}_t) - m_{k',i}(\boldsymbol{z}_t) \Big) && (95) \\ &= \boldsymbol{u}_k(\boldsymbol{z}_t)^\top \Big(m_{k,\mathcal{I}_k}(\boldsymbol{z}_t) - m_{k',\mathcal{I}_k}(\boldsymbol{z}_t) \Big) > 0 && (96) \end{aligned}$$

where the variances cancel out because they are equal across $\mathcal{I}_k$. The resulting inequality follows from Eq. (90). Therefore, condition (ii) holds for every $k' \in \mathcal{T}_k$. Finally, consider $k' \notin \mathcal{T}_k$. This case is similar to the case $\mathcal{T}_k = \emptyset$, where there is at least one $i \in \mathcal{I}_k$ such that $\sigma_{k,i} > \sigma_{k',i}$. Since every coordinate $d_{k,i}(\boldsymbol{z}_t)$ for $i \in \mathcal{I}_k$ is nonzero, condition (i) of Proposition B.4 is satisfied:

$$\boldsymbol{d}_k(\boldsymbol{z}_t)^\top \left(\Sigma_{k'}^{-1} - \Sigma_k^{-1}\right) \boldsymbol{d}_k(\boldsymbol{z}_t) = \sum_{i \in \mathcal{I}_k} d_{k,i}^2 \left[\left(\frac{1}{\sigma_{k',i}}\right)^2 - \left(\frac{1}{\sigma_{k,i}}\right)^2 \right] > 0, \tag{97}$$

The inequality follows the same argument as Eq. (88). Therefore, under Assumption (A6), the conditions of Proposition B.4 hold. □

B.4 Proof of Theorem 3.2

Consider our latent variable model, which is an MSM with recurrent state transitions:

$$p_{\boldsymbol{\theta_z},\boldsymbol{\theta_s}}(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t | \boldsymbol{z}_{t-L_z:t-1}, s_t) p_{\boldsymbol{\theta_s}}(s_t | s_{t-1}, \boldsymbol{z}_{t-1}) \tag{98}$$

where for the first part of Theorem 3.2, we consider latent transitions with arbitrary lag L_z. This broadens the applicability of our theoretical results. Under this setting, and w.l.o.g., one can adapt (A2) to a multi-lag setting where the rMSM considers variables $\boldsymbol{z}_t$ which are Gaussian, such that the means and covariances are dependent on s_t and $\boldsymbol{z}_{t-L_z:t-1}$ as follows:

$$p_{\boldsymbol{\theta_z}}(\boldsymbol{z}_t | \boldsymbol{z}_{t-L_z:t-1}, s_t) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}_{s_t}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-L_z}), \Sigma_{s_t}) \tag{99}$$

where $\boldsymbol{m}_k : \mathbb{R}^{mL_z} \to \mathbb{R}^m$ are analytic functions, distinct for all $k \in \{1, \ldots, K\}$. We establish L_z-lagged regime-dependent transition identifiability up to permutations below.

B.4.1 Identifiability of autoregressive components (Theorem 3.2 (i))

Theorem B.6. *Under assumption (A2), the Markov Switching Model is identifiable with respect to its autoregressive components $p_{\boldsymbol{\theta}_z}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)$ up to permutation.*

Proof. For notation simplicity, we drop the subscript $\boldsymbol{\theta}$, and consider two MSM distributions $p(\boldsymbol{z}_{1:T})$ and $\hat{p}(\boldsymbol{z}_{1:T})$. Assume there exist two MSMs satisfying $p(\boldsymbol{z}_{1:T}) = \hat{p}(\boldsymbol{z}_{1:T})$, with K and $\hat{K}$ regimes, respectively. Since wlog. $p(\boldsymbol{z}_{1:t}) = p(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1})p(\boldsymbol{z}_{1:t-1})$, the fact that $p(\boldsymbol{z}_{1:T}) = \hat{p}(\boldsymbol{z}_{1:T})$, and $p(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1})$ is a probability distribution imply that

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1}) = \hat{p}(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1}), \quad \forall t = 2, ..., T. \tag{100}$$

Now, since the model assumes a conditional Markov structure controlled by the current state s_t (98), we can show that the conditional distribution is a finite mixture distribution:

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1}) = \sum_{k=1}^{K} p(s_t = k|\boldsymbol{z}_{1:t-1}) p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k). \tag{101}$$

Since $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k)$ is Gaussian (99), Proposition 2 in Yakowitz and Spragins [1968] shows Gaussians with distinct parameters are identifiable up to permutations. Therefore, we have $K = \hat{K}$, and for almost any given $\boldsymbol{z}_{t-L_z:t-1} = \boldsymbol{\alpha} \in \mathbb{R}^{mL_z}$ (modulo some zero-measure sets) we have the following result: there exists a permutation indexed by $\boldsymbol{\alpha}$, $\pi : \mathbb{R}^{mL_z} \to S_K$ such that $\hat{k} = \pi(\boldsymbol{\alpha}, k)$ for any $k \in \{1, ..., K\}$

$$\begin{aligned} p(s_t = k|\boldsymbol{z}_{1:t-L_z-1}, \boldsymbol{z}_{t-L_z:t-1} = \boldsymbol{\alpha}) &= \hat{p}\big(s_t = \pi(\boldsymbol{\alpha}, k)|\boldsymbol{z}_{1:t-L_z-1}, \boldsymbol{z}_{t-L_z:t-1} = \boldsymbol{\alpha}\big), \\ p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1} = \boldsymbol{\alpha}, s_t = k) &= \hat{p}\big(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1} = \boldsymbol{\alpha}, s_t = \pi(\boldsymbol{\alpha}, k)\big). \end{aligned} \tag{102}$$

To clarify, Proposition 2 in Yakowitz and Spragins [1968] establishes that multivariate Gaussian distributions are identifiable if and only if the means or covariances are distinct. Wlog., assume covariances are identical, the distribution is unidentifiable on the set where pairs of mean functions coincide. For any model, we can determine this set:

$$\mathcal{C}_{k,k'} := \Big\{\boldsymbol{\alpha} \in \mathbb{R}^{mL_z} : \boldsymbol{m}_k(\boldsymbol{\alpha}) = \boldsymbol{m}_{k'}(\boldsymbol{\alpha})\Big\}, \quad \mathcal{C} := \bigcup_{k \neq k'}^{K} \mathcal{C}_{k,k'}. \tag{103}$$

where for $\hat{p}$ we can determine a similar collision set $\hat{\mathcal{C}}$ finding similar $\hat{\mathcal{C}}_{k,k'}$ for any $k \neq k' \in \{1, \ldots, \hat{K}\}$. Given that the zero set of an analytic function has Lebesgue measure zero Mityagin [2015], $\mathcal{C}_{k,k'}$ has Lebesgue measure zero. Therefore $\mathcal{C}$ and $\hat{\mathcal{C}}$ have Lebesgue measure zero since $K < +\infty$. Therefore, for any pair of models, Equation (102) holds for any $\boldsymbol{\alpha} \in \mathcal{Z}$, where $\mathcal{Z} = \mathbb{R}^{mL_z} \setminus (\mathcal{C} \cup \hat{\mathcal{C}})$, and $\mu(\mathcal{Z}) = \mu(\mathbb{R}^{mL_z})$.

Now we show that $\pi(\boldsymbol{\alpha}, k)$ is a constant almost everywhere, i.e., $\pi(\boldsymbol{\alpha}, k) = \pi(k)$. First, define the pairwise distance between the mean functions and the minimum distance:

$$d_{k,k'}(\boldsymbol{\alpha}) := ||\boldsymbol{m}_k(\boldsymbol{\alpha}) - \boldsymbol{m}_{k'}(\boldsymbol{\alpha})||, \quad \delta_0(\boldsymbol{\alpha}) := \min_{k \neq k'} d_{k,k'}(\boldsymbol{\alpha}). \tag{104}$$

Choose a non-collision point $\boldsymbol{\alpha}_0 \in \mathcal{Z}$, where we know $\delta_0(\boldsymbol{\alpha}_0) > 0$, and $\boldsymbol{m}_k(\boldsymbol{\alpha}_0) = \hat{\boldsymbol{m}}_{\hat{k}}(\boldsymbol{\alpha}_0)$ with $\hat{k} = \pi(\boldsymbol{\alpha}_0, k)$ for any $k \in \{1, \ldots, K\}$ (from Eq. (102)). Define $\hat{\delta}_0(\boldsymbol{\alpha}_0) > 0$ similarly for $\hat{\boldsymbol{m}}_{\hat{k}}$, $\hat{k} \in \{1, \ldots, K\}$. Now set $\delta := \frac{1}{4}\min\{\delta_0(\boldsymbol{\alpha}_0), \hat{\delta}_0(\boldsymbol{\alpha}_0)\}$. By continuity, for all $k, \hat{k} \in \{1, \ldots, K\}$, there exists $\epsilon > 0$ such that we can construct an ϵ-ball $B(\boldsymbol{\alpha}_0, \epsilon)$ using ℓ_2-norm, where

$$||\boldsymbol{m}_k(\boldsymbol{\alpha}_0) - \boldsymbol{m}_k(\boldsymbol{\alpha})|| < \delta, \quad ||\hat{\boldsymbol{m}}_{\hat{k}}(\boldsymbol{\alpha}_0) - \hat{\boldsymbol{m}}_{\hat{k}}(\boldsymbol{\alpha})|| < \delta, \quad \forall \boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0, \epsilon). \tag{105}$$

Now for any $\boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0, \epsilon)$ and $k \neq k' \in \{1, \ldots, K\}$, we can compute a lower bound for $d_{k,k'}(\boldsymbol{\alpha})$:

$$\begin{aligned} d_{k,k'}(\boldsymbol{\alpha}) &= ||\boldsymbol{m}_k(\boldsymbol{\alpha}) - \boldsymbol{m}_{k'}(\boldsymbol{\alpha})|| \\ &\geq ||\boldsymbol{m}_k(\boldsymbol{\alpha}_0) - \boldsymbol{m}_{k'}(\boldsymbol{\alpha}_0)|| - ||\boldsymbol{m}_k(\boldsymbol{\alpha}_0) - \boldsymbol{m}_k(\boldsymbol{\alpha})|| - ||\boldsymbol{m}_{k'}(\boldsymbol{\alpha}_0) - \boldsymbol{m}_{k'}(\boldsymbol{\alpha})|| \\ &> \delta_0(\boldsymbol{\alpha}_0) - 2\delta \geq \delta_0(\boldsymbol{\alpha}_0)/2. \end{aligned}$$

And similarly, we can establish $\hat{d}_{k,k'}(\boldsymbol{\alpha}) > \hat{\delta}_0(\boldsymbol{\alpha}_0)/2$ for any $\boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0, \epsilon)$ with $\epsilon > 0$.

Assume $\pi(\boldsymbol{\alpha}, k)$ is non-constant. Then $\exists \boldsymbol{\alpha}_1 \in B(\boldsymbol{\alpha}_0, \epsilon)$ such that:

$$\hat{k}_1 = \pi(\boldsymbol{\alpha}_1, k) \neq \pi(\boldsymbol{\alpha}_0, k) = \hat{k}_0 \tag{106}$$

From Eq. (102), we have $\boldsymbol{m}_k(\boldsymbol{\alpha}_0) = \hat{\boldsymbol{m}}_{\hat{k}_0}(\boldsymbol{\alpha}_0)$ and $\boldsymbol{m}_k(\boldsymbol{\alpha}_1) = \hat{\boldsymbol{m}}_{\hat{k}_1}(\boldsymbol{\alpha}_1)$ and continuity gives

$$||\hat{\boldsymbol{m}}_{\hat{k}_1}(\boldsymbol{\alpha}_1) - \hat{\boldsymbol{m}}_{\hat{k}_0}(\boldsymbol{\alpha}_1)|| \leq ||\hat{\boldsymbol{m}}_{\hat{k}_1}(\boldsymbol{\alpha}_1) - \boldsymbol{m}_k(\boldsymbol{\alpha}_0)|| + ||\boldsymbol{m}_k(\boldsymbol{\alpha}_0) - \hat{\boldsymbol{m}}_{\hat{k}_0}(\boldsymbol{\alpha}_1)|| < 2\delta \leq \hat{\delta}_0(\boldsymbol{\alpha}_0)/2,$$

but because $\hat{k}_0 \neq \hat{k}_1$, we also have

$$||\hat{\boldsymbol{m}}_{\hat{k}_1}(\boldsymbol{\alpha}_1) - \hat{\boldsymbol{m}}_{\hat{k}_0}(\boldsymbol{\alpha}_1)|| = \hat{d}_{\hat{k}_0,\hat{k}_1}(\boldsymbol{\alpha}_1) > \hat{\delta}_0(\boldsymbol{\alpha}_0)/2,$$

and a contradiction is reached. Therefore, for any $\boldsymbol{\alpha}_0 \in \mathcal{Z}$, there exists an open subset $\mathcal{U} \subset \mathcal{Z}$ where $\pi(\boldsymbol{\alpha}, k) = \pi_{\mathcal{U}}(k)$ for all $\boldsymbol{\alpha} \in \mathcal{U}$

$$\boldsymbol{m}_k(\boldsymbol{\alpha}) = \hat{\boldsymbol{m}}_{\pi_{\mathcal{U}}(k)}(\boldsymbol{\alpha}), \quad \forall \boldsymbol{\alpha} \in \mathcal{U}, \tag{107}$$

Since the mean functions on both models are analytic and distinct, equality on a nonempty open set $\mathcal{U}$ implies equality everywhere on $\mathbb{R}^{mL_z}$ by the identity theorem for vector-valued real analytic functions. This implies two distinct permutations $\pi_{\mathcal{U}_1} \neq \pi_{\mathcal{U}_2}$ cannot occur on disjoint open subsets $\mathcal{U}_1 \subset \mathcal{Z}$, $\mathcal{U}_2 \subset \mathcal{Z}$. Otherwise, $\exists k \in \{1, \ldots, K\}$ such that $\pi_{\mathcal{U}_1}(k) \neq \pi_{\mathcal{U}_2}(k)$, and hence forcing $\hat{\boldsymbol{m}}_{\pi_{\mathcal{U}_1}(k)} \equiv \boldsymbol{m}_k \equiv \hat{\boldsymbol{m}}_{\pi_{\mathcal{U}_2}(k)}$ contradicting the assumption of distinct means on $\hat{\boldsymbol{m}}_k,\ k \in \{1, \ldots, K\}$.

In summary, there exists a global permutation $\pi \in S_K$ such that $\hat{k} = \pi(k),\ k \in 1, \ldots, K$ and the following equivalence holds almost everywhere for $\boldsymbol{z}_{t-L_z:t-1} \in \mathbb{R}^{mL_z}$:

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k) = \hat{p}\big(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = \pi(k)\big), \quad \forall \boldsymbol{z}_{t-L_z:t-1} \in \mathcal{Z}, \quad \mu(\mathcal{Z}) = \mu(\mathbb{R}^{mL_z}), \tag{108}$$

and from Gaussian assumptions on $\boldsymbol{z}_t$, and analyticity on $\boldsymbol{m}_k$,

$$\boldsymbol{m}_k \equiv \hat{\boldsymbol{m}}_{\pi(k)}, \quad \Sigma_k \equiv \hat{\Sigma}_{\pi(k)}. \tag{109}$$

□

B.4.2 Identifiability of switching process (Theorem 3.2 (ii))

Below we present the identifiability of the switching process up to permutation by further assuming (A4–A5). Therefore, we fall back to the single-lagged case with $L_z = 1$ for simplicity.

Theorem B.7. *The MSM is identifiable up to state permutation equivalence, further assuming (A4–A5) and that the transition matrix $p(s_t|s_{t-1}, \boldsymbol{z}_{t-1}) = Q(\boldsymbol{z}_{t-1})$ is an analytic function in $\boldsymbol{z}_{t-1}$.*

Proof. From Theorem 3.2, we have established that the number of discrete states K is identifiable, and the transitions $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$ and one-step predictive filtering posteriors $p(s_t|\boldsymbol{z}_{1:t-1})$ are identifiable up to permutation equivalence. This identifiability result extends to all time steps $t > 1$, and due to regime stationarity of MSMs, this permutation equivalence does not change over time.

Now w.l.o.g., fix a particular ordering of the discrete state. Note that for $t > 1$, $p(s_t = k'|\boldsymbol{z}_{1:t})$ is also identifiable for all $k' \in \{1, ..., K\}$, as $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t = k')$ and $p(s_t = k'|\boldsymbol{z}_{1:t-1})$ are identifiable as well (from Theorem 3.2) and then by applying Bayes' rule:

$$p(s_t = k'|\boldsymbol{z}_{1:t}) \propto p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t = k')p(s_t = k'|\boldsymbol{z}_{1:t-1}). \tag{110}$$

Now fix $t \geq T^*$ and choose $\boldsymbol{\beta} \in \mathcal{Z}$ as given by (A4–A5). For each regime $k \in \{1, \ldots, K\}$, (A4–A5) and Lemma B.1 provide a history $\boldsymbol{z}_{1:t}^{(k)} = (\boldsymbol{\alpha}_k, \boldsymbol{\beta})$ such that $p(s_t = k \mid \boldsymbol{z}_{1:t-1} = \boldsymbol{\alpha}_k, \boldsymbol{z}_t = \boldsymbol{\beta}) > \frac{1}{2}$.

Define the matrices

$$A(\boldsymbol{z}) = \begin{pmatrix} a_1(\boldsymbol{z}) \\ \vdots \\ a_K(\boldsymbol{z}) \end{pmatrix}, \qquad a_k(\boldsymbol{z}) := p(s_t \mid \boldsymbol{z}_{1:t-1} = \boldsymbol{\alpha}_k, \boldsymbol{z}_t = \boldsymbol{z}), \tag{111}$$

and

$$B(\boldsymbol{z}) = \begin{pmatrix} b_1(\boldsymbol{z}) \\ \vdots \\ b_K(\boldsymbol{z}) \end{pmatrix}, \qquad b_k(\boldsymbol{z}) := p(s_{t+1} \mid \boldsymbol{z}_{1:t-1} = \boldsymbol{\alpha}_k, \boldsymbol{z}_t = \boldsymbol{z}). \tag{112}$$

For every $\boldsymbol{z} \in \mathbb{R}^m$, the recurrent transition model implies $B(\boldsymbol{z}) = A(\boldsymbol{z})Q(\boldsymbol{z})$. At $\boldsymbol{z} = \boldsymbol{\beta}$, the matrix $A(\boldsymbol{\beta})$ is strictly diagonally dominant given Lemma B.1, and therefore $A(\boldsymbol{\beta})$ is invertible by the Levy-Desplanques theorem [Wu, 2005]. Furthermore, $A(\boldsymbol{z})$ is also continuous, as each row is formed by a weighted average of continuous functions in $\boldsymbol{z}$. Therefore, there exists a nonzero measure set $\mathcal{U}_{\boldsymbol{\beta}}$, such that $A(\boldsymbol{z})$ is invertible for any $\boldsymbol{z} \in \mathcal{U}_{\boldsymbol{\beta}}$. This implies

$$Q(\boldsymbol{z}) = A(\boldsymbol{z})^{-1}B(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{U}_{\boldsymbol{\beta}}, \tag{113}$$

i.e., $Q(\boldsymbol{z})$ is identifiable on $\mathcal{U}_{\boldsymbol{\beta}}$.

The above statement implies any other transition matrix $Q'(\boldsymbol{z})$ inducing the same distribution $p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_{1:t-1}, \boldsymbol{z}_t = \boldsymbol{z})$ satisfies $Q(\boldsymbol{z}) = Q'(\boldsymbol{z})$ for any $\boldsymbol{z} \in \mathcal{U}_{\boldsymbol{\beta}}$. Since both $Q(\boldsymbol{z})$, and $Q'(\boldsymbol{z})$ are analytic and coincide on the nonempty open set $\mathcal{U}_{\boldsymbol{\beta}}$, the identity theorem for real-analytic functions implies $Q(\boldsymbol{z}) = Q'(\boldsymbol{z})$ for any $\boldsymbol{z} \in \mathbb{R}^m$. Therefore, Q is identifiable up to state permutation equivalence.

□

B.4.3 Alternative assumptions on the switching process

We additionally establish permutation identifiability of more restrictive switching processes, such as autonomous switching $p_{\boldsymbol{\theta}_s}(s_t \mid s_{t-1})$ and observation-dependent switching from the MSM view, $p_{\boldsymbol{\theta}_s}(s_t \mid \boldsymbol{z}_t)$, where $\boldsymbol{z}_t$ denotes an MSM observed variable. The latter connects to self-exciting threshold autoregressive models (SETAR; Tong and Lim, 1980). Establishing such results does not involve additional assumptions beyond (A2). Therefore, for completeness, we consider again an arbitrary lag L_z.

Corollary B.8. *Under Assumption (A2), the switch transitions are identifiable up to state permutation equivalence if the transition matrix is observation-dependent with respect to MSM observations, i.e.,* $p_{\boldsymbol{\theta}_s}(s_t|s_{t-1}, \boldsymbol{z}_{t-1}) := p_{\boldsymbol{\theta}_s}(s_t|\boldsymbol{z}_{t-1})$.

Proof. Drop the subscript $p_{\boldsymbol{\theta}_s}$ for simplicity. From Theorem B.6, the number of discrete states K is identifiable, and the one-step filtering posteriors $p(s_t|\boldsymbol{z}_{1:t-1})$ are identifiable. In an observation-dependent model, s_t is conditionally independent of $\boldsymbol{z}_{1:t-2}$ given $\boldsymbol{z}_{t-1}$. Therefore,

$$p(s_t|\boldsymbol{z}_{1:t-1}) = p(s_t|\boldsymbol{z}_{t-1}) \tag{114}$$

which implies the transitions are identifiable up to permutations. □

Theorem B.9. *Under assumption (A2), the switch transitions are identifiable up to state permutation equivalence if the transition matrix is autonomous:* $p_{\boldsymbol{\theta}_s}(s_t|s_{t-1}, \boldsymbol{z}_{t-1}) := p_{\boldsymbol{\theta}_s}(s_t|s_{t-1})$.

Proof. Again, drop the subscript $p_{\boldsymbol{\theta}_s}$ for simplicity. From Theorem B.6, we have established that the number of discrete states K is identifiable, and the transitions $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$ and one-step predictive filtering posteriors $p(s_t|\boldsymbol{z}_{1:t-1})$ are identifiable up to permutation equivalence. This identifiability result extends to all time steps $t > 1$, and due to regime stationarity of MSMs, this permutation equivalence does not change over time.

Now fix a particular ordering of the discrete state w.l.o.g.. Note that for $t > 1$, $p(s_t = k'|\boldsymbol{z}_{1:t})$ is also identifiable for all $k' \in \{1, ..., K\}$, as $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k')$ and $p(s_t = k'|\boldsymbol{z}_{1:t-1})$ are identifiable as well (from Theorem 3.2) and then by applying Bayes' rule:

$$p(s_t = k'|\boldsymbol{z}_{1:t}) \propto p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k')p(s_t = k'|\boldsymbol{z}_{1:t-1}). \tag{115}$$

Under an autonomous first-order Markov state transition model:

$$p(s_{t+1} = k|\boldsymbol{z}_{1:t}) = \sum_{k'=1}^{K} p(s_t = k'|\boldsymbol{z}_{1:t})p(s_{t+1} = k|s_t = k'). \tag{116}$$

Denote the $K \times K$ transition matrix $p(s_{t+1}|s_t)$ as Q whose kth column is $Q_k = [p(s_{t+1} = k|s_t = 1), ..., p(s_{t+1} = k|s_t = K)]$. To determine Q, we can construct a linear system from Eq. (115) by using different values of $\boldsymbol{z}_{1:t}$. Fix $\boldsymbol{z}_{1:t-1}$ and select K different points $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ for $\boldsymbol{z}_t$. Now, denote A, B matrices:

$$A = (a_1, \ldots, a_K), \quad a_k = p(s_t \mid \boldsymbol{z}_{1:t-1}, \boldsymbol{z}_t = \boldsymbol{\beta}_k), \tag{117}$$

$$B = (b_1, \ldots, b_K)^\top, \quad b_k = p(s_{t+1} \mid \boldsymbol{z}_{1:t-1}, \boldsymbol{z}_t = \boldsymbol{\beta}_k), \tag{118}$$

which yields the following linear system:

$$B = AQ. \tag{119}$$

Using Eq. (115) and given $p(s_t = k|\boldsymbol{z}_{1:t-1}) > 0$, the filtering posterior admits the following form:

$$p(s_t = k|\boldsymbol{z}_{1:t}) = \frac{w_k p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k)}{\sum_{k'} w_{k'} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-L_z:t-1}, s_t = k')}, \quad w_k > 0. \tag{120}$$

And we can write A as follows.

$$A = \mathrm{diag}(w) G D^{-1}, \tag{121}$$

where

$$G_{kj} = p(\boldsymbol{z}_t = \boldsymbol{\beta}_j|\boldsymbol{z}_{t-L_z:t-1}, s_t = k), \quad D_{jj} = \sum_{k=1}^{K} w_k p(\boldsymbol{z}_t = \boldsymbol{\beta}_j|\boldsymbol{z}_{t-L_z:t-1}, s_t = k). \tag{122}$$

From $w_k > 0$ and $p(\boldsymbol{z}_t = \boldsymbol{\beta}_j|\boldsymbol{z}_{t-L_z:t-1}, s_t = k) > 0$, D is invertible. Furthermore, G is invertible. To see why, Gaussian distributions with distinct means or covariances are linearly independent [Yakowitz and Spragins, 1968]. Therefore, we can select $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$, such that G is invertible. Since both A and B are identifiable and A is invertible, $Q = A^{-1}B$. Therefore, the autonomous transition matrix is identifiable up to state permutations. □

B.5 Extended discussion on disentanglement and Proof of Theorem 3.3

We begin with a simple example illustrating Kivva et al. [2022, Assumption (P2)].

Example. Consider $m = 3,\ K = 2$ with $\boldsymbol{\sigma}_1 = (1, \frac{1}{2}, \frac{1}{4})$, and $\boldsymbol{\sigma}_2 = (1, \frac{1}{4}, \frac{1}{2})$. We note the above condition holds given $\boldsymbol{\sigma}_1/\boldsymbol{\sigma}_2 = (1, 2, \frac{1}{2})$.

The intuition of Theorem 3.3 starts by considering the case where the ratio vector has some, but not all, distinct entries. For a pair of regimes where Kivva et al. [2022, Assumption (P2)] is violated, we have dimension groups with equal ratio (from labels $\{1, \ldots, m\}$), such that we obtain identifiability across groups, but cannot resolve intra-group affine ambiguity. Therefore, with additional regimes, such that each regime resolves different groups, latent variable disentanglement could be obtained.

Example. Assume $m = 3$, but now $K = 3$. Assume for $k = 1, 2$, the ratio vector is:

$$\begin{aligned} &\sigma_{1,1} = 1, \quad \sigma_{1,2} = 0.25, \quad \sigma_{1,3} = 0.25, \\ &\sigma_{2,1} = 1, \quad \sigma_{2,2} = 0.25, \quad \sigma_{2,3} = 0.5, \\ &\boldsymbol{\sigma}_1/\boldsymbol{\sigma}_2 = (1, \quad 1, \quad 0.5). \end{aligned}$$

Intuitively, using a similar Kivva-style SVD-based proof, we can distinguish the first and second groups. Therefore, the affine ambiguity is reduced, as the equivalence introduces a block-diagonal matrix. $\boldsymbol{z}' = PB\boldsymbol{z} + \boldsymbol{b}$, where for this example the block diagonal matrix is as follows:

$$B = \begin{pmatrix} b_1 & b_2 & 0 \\ b_3 & b_4 & 0 \\ 0 & 0 & b_5 \end{pmatrix}$$

Introducing additional regimes will introduce additional restrictions to the equivalence, and therefore, if we obtain another pair, e.g., $k = 2, 3$, such that the relation for that becomes

$$B = \begin{pmatrix} b'_1 & 0 & 0 \\ 0 & b'_2 & b'_3 \\ 0 & b'_4 & b'_5 \end{pmatrix}$$

we will obtain a diagonal term in the equivalence.

B.5.1 Block-permutation disentanglement

We formalise the above example by proving block-permutation identifiability. Assume we have a non-temporal GMM for $\boldsymbol{z} \in \mathbb{R}^m$, expressed as follows:

$$\boldsymbol{z} \sim \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{\mu}_k, diag(\boldsymbol{\sigma}_k^2)). \tag{123}$$

Lemma B.10. *Let $K \geq 2$ and $c_k > 0$ for all $k \in \{1, \ldots, K\}$. Let $\boldsymbol{z}' = A\boldsymbol{z} + \boldsymbol{b}$, where $A \in \mathbb{R}^{m \times m}$ is invertible and $\boldsymbol{b} \in \mathbb{R}^m$. Assume there exist $k_1 \neq k_2 \in \{1, \ldots, K\}$ such that the ratio vector $\boldsymbol{r}_{k_1,k_2}$ induces a partition $\{1, \ldots, m\} = I_1 \cup \cdots \cup I_L$ where for any $\ell \in \{1, \ldots, L\}$, $L \leq m$,*

$$i, j \in I_\ell \iff r_{k_1,k_2,i} = r_{k_1,k_2,j}. \tag{124}$$

and for $\ell \neq \ell'$, $I_\ell \cap I'_\ell = \emptyset$. That is, each I_ℓ contains exactly those dimension labels with the same variance ratio. Then from $\boldsymbol{z}'$, we can recover an invertible matrix $A' \in \mathbb{R}^{m \times m}$ such that $(A')^{-1}A = P_1 B P_2$, where P_1, P_2 are permutation matrices, and B is a block diagonal matrix defined as follows:

$$B = \text{blockdiag}(B_1, \ldots, B_L) \tag{125}$$

where each $B_\ell \in \mathbb{R}^{|I_\ell| \times |I_\ell|}$, $\ell \in \{1, \ldots, L\}$ is invertible.

Remark B.11. In particular, if all ratios are distinct, i.e., $|I_\ell| = 1$ for all $\ell \in \{1, \ldots, L\}$ (equivalently $L = m$), B is diagonal. This reduces to Kivva et al. [2022].

Remark B.12. If coordinates are ordered by blocks, $P_2 = I$, and $(A')^{-1}A = P_1 B$.

Proof. Let $\boldsymbol{z}$ be GMM-distributed with diagonal covariances $\Sigma_k = \text{diag}(\boldsymbol{\sigma}_k^2)$, $k \in \{1, \ldots, K\}$. Let $\hat{\Sigma}_k$ denote the covariance matrix of the k-th component of $\boldsymbol{z}'$. Under the affine transformation $\boldsymbol{z}' = A\boldsymbol{z} + \boldsymbol{b}$, we have

$$\hat{\Sigma}_k = A\Sigma_k A^\top, \tag{126}$$

where each $\hat{\Sigma}_k$, $k \in \{1, \ldots, K\}$ is symmetric positive definite. Fix two regimes k_1 and k_2 (wlog. $k_1 = 1, k_2 = 2$) and consider the matrix $M := \hat{\Sigma}_2^{-1}\hat{\Sigma}_1$:

$$M = \hat{\Sigma}_2^{-1}\hat{\Sigma}_1 = A^{-\top}\Sigma_2^{-1}A^{-1}A\Sigma_1 A^\top = A^{-\top}D^2 A^\top, \tag{127}$$

where $D = diag(\boldsymbol{r}_{(1,2)})$. Given Eq (127), M is similar to the diagonal D^2, and hence is diagonalisable. From the eigen-decomposition of M, we may find invertible U such that $M = UD^2U^{-1}$. Define $A' := U^{-\top}$. Then, $M = (A')^{-\top}D^2(A')^\top$. Combining with Eq. (127) yields (after transposing)

$$AD^2A^{-1} = A'D^2(A')^{-1} \implies SD^2 = D^2S, \tag{128}$$

with $S = (A')^{-1}A$. Given D^2 is diagonal, we have

$$(d_j^2 - d_i^2)S_{ij} = 0, \quad i, j \in \{1, \ldots, m\}. \tag{129}$$

This implies $S_{ij} = 0$ whenever $d_i \neq d_j$, and no restriction appears if $d_i = d_j$. Given $d_i = r_{1,2,i}$, R is structured from the partition $\{I_1, \ldots, I_L\}$, since $I_\ell = \{i \in \{1, \ldots, m\} : d_i = d_\ell\}$, $\ell \in \{1, \ldots, L\}$. Therefore, S is block-diagonal with respect to the partition up to coordinate re-ordering. That is, there exists a permutation matrix P, such that:

$$S = (A')^{-1}A = P^\top BP, \quad B = \text{blockdiag}(B_1, \ldots, B_L), \tag{130}$$

where each block $B_\ell \in \mathbb{R}^{|I_\ell| \times |I_\ell|}, \ell \in \{1, \ldots, L\}$ is invertible. Invertibility of each B_ℓ follows from invertibility of S. Finally, note that the eigendecomposition of M is not unique, as we can reorder eigenvectors. Therefore, we may redefine $\tilde{A}' := A'\tilde{P}$. Then, $\tilde{S} = (A'\tilde{P})^{-1}A = \tilde{P}^\top S$. Since we have shown $S = P^\top BP$, it follows that

$$\tilde{S} = (\tilde{P}^\top P^\top)BP. \tag{131}$$

Since the choice of $\tilde{P}$ relabels coordinates, we therefore can obtain invertible A' (replacing $\tilde{A}'$ by A') such that

$$(A')^{-1}A = P_1 B P_2, \tag{132}$$

with P_1, P_2 permutation matrices and B block-diagonal with invertible blocks. □

B.5.2 Proof of Theorem 3.3

Lemma B.10 shows block-permutation identifiability for a single pair of regimes. When considering all possible pairs, we can construct the ratio matrix R presented under (A8), and the proof of Theorem 3.3 aggregates all regime pairs together to strengthen the block-permutation identifiability result into a single permutation of latent variables.

Proof. The proof follows largely from Lemma B.10, where the previous commutativity requirement now induces increasing restrictions for every row of the ratio matrix R. For each pair of regimes $i = (i_1, i_2)$ with $i_1 < i_2,\ i_1, i_2 \in \{1, \dots, K\}$, define $M_i := \hat{\Sigma}_{i_2}^{-1}\hat{\Sigma}_{i_1}$. Using $\hat{\Sigma}_k = A\Sigma_k A^\top$,

$$M_i = A^{-\top}\Sigma_{i_2}^{-1}A^{-1}A\Sigma_{i_1}A^\top = A^{-\top}D_i^2A^\top. \tag{133}$$

where D_i is diagonal, and $D_i = \operatorname{diag}(R_{i,:}) = \operatorname{diag}(\boldsymbol{r}_{(i_1,i_2)})$ (D_i is constructed from the i-th row of R). Therefore, all M_i are diagonalisable under $A^{-\top}$, and furthermore, any two $M_i, M_j, i \neq j$ commute:

$$M_iM_j = A^{-\top}D_i^2A^\top A^{-\top}D_j^2A^\top = A^{-\top}D_i^2D_j^2A^\top = A^{-\top}D_j^2D_i^2A^\top = M_jM_i. \tag{134}$$

Therefore, all M_i are simultaneously diagonalisable under the same basis [Horn and Johnson, 2012, p. 64]. That is, there exists an invertible matrix U such that $M_i = UD_i^2U^{-1}$ for all $i \in \{1, \dots, \binom{K}{2}\}$. Define $A' := U^{-\top}$, then

$$M_i = (A')^{-\top}D_i^2(A')^\top. \tag{135}$$

Define $S := (A')^{-1}A$. Combining with Eq. (133) (and transposing) yields $SD_i^2 = D_i^2S$. And since each D_i^2 is diagonal, we have

$$(d_{i,k}^2 - d_{i,j}^2)S_{kj} = 0, \quad k, j \in \{1, \dots, m\}, \quad \forall i \in \left\{1, \dots, \binom{K}{2}\right\}. \tag{136}$$

This implies $S_{kj} = 0$ whenever $d_{i,k} \neq d_{i,j}$. Given $d_{i,j} = R_{ij}$, and no two columns of R are equal, this implies S is a diagonal matrix. As in Lemma B.10, the eigendecomposition is defined up to permutation of eigenvectors. Replacing A' for $A'P$, with P a permutation matrix, transforms S into $P^\top S$. Therefore, we can write S up to relabelling:

$$S = (A')^{-1}A = PD, \tag{137}$$

with P a permutation matrix and D diagonal. □

C Method Details and Discussions

C.1 Forward-backward Inference

Below we provide further details on computing the posterior distributions $\gamma_{t,k}(\boldsymbol{z}_{1:T}) = p(s_t = k \mid \boldsymbol{z}_{1:T})$ and $\xi_{t,k,l}(\boldsymbol{z}_{1:T}) = p(s_t = k, s_{t-1} = l \mid \boldsymbol{z}_{1:T})$. Given a sequence $\boldsymbol{z}_{1:T}$ with K regimes, for $k \in \{1, \dots, K\}$, the forward algorithm computes messages $\alpha_{t,k}(\boldsymbol{z}_{1:t}) = p(\boldsymbol{z}_{1:t}, s_t = k)$ for $t = 2, \dots, T$:

$$\alpha_{t,k}(\boldsymbol{z}_{1:t}) = p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t = k) \times \sum_{l=1}^{K} Q_{lk}(\boldsymbol{z}_{t-1})\alpha_{t-1,l}(\boldsymbol{z}_{1:t-1}),$$
$$\alpha_{1,k}(\boldsymbol{z}_1) = p(\boldsymbol{z}_1 \mid s_1 = k)p(s_1 = k), \tag{138}$$

and the backward algorithm computes messages $\beta_{t,k}(\boldsymbol{z}_{t+1:T}) = p(\boldsymbol{z}_{t+1:T} \mid \boldsymbol{z}_t, s_t = k)$ for $t = T-1, \dots, 1$:

$$\beta_{t,k}(\boldsymbol{z}_{t+1:T}) = \sum_{l=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, s_{t+1} = l) \times Q_{kl}(\boldsymbol{z}_t)\beta_{t+1,l}(\boldsymbol{z}_{t+1:T}), \tag{139}$$

with $\beta_{T,k} = 1$ for all $k \in \{1, \dots, K\}$. The posteriors $\gamma_{t,k}(\boldsymbol{z}_{1:T}) = p(s_t = k \mid \boldsymbol{z}_{1:T})$ and $\xi_{t,k,l}(\boldsymbol{z}_{1:T}) = p(s_t = k, s_{t-1} = l \mid \boldsymbol{z}_{1:T})$ are then computed as follows

$$\gamma_{t,k}(\boldsymbol{z}_{1:T}) = \frac{\alpha_{t,k}(\boldsymbol{z}_{1:t})\beta_{t,k}(\boldsymbol{z}_{t+1:T})}{\sum_{j=1}^{K}\alpha_{t,j}(\boldsymbol{z}_{1:t})\beta_{t,j}(\boldsymbol{z}_{t+1:T})},$$
$$\xi_{t,k,l}(\boldsymbol{z}_{1:T}) = \frac{\alpha_{t-1,l}(\boldsymbol{z}_{1:t-1})\beta_{t,k}(\boldsymbol{z}_{t+1:T})p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t = k)Q_{lk}(\boldsymbol{z}_{t-1})}{\sum_{j=1}^{K}\alpha_{t,j}(\boldsymbol{z}_{1:t})\beta_{t,j}(\boldsymbol{z}_{t+1:T})} \tag{140}$$

To avoid numerical issues, we normalise Equations (138) and (139) at each time step t with respect to s_t, and run computations in log-space using log-sum-exp tricks.

C.2 ΩSDS implementation details

Synthetic data. As discussed in Section 4, for the synthetic experiments we parametrise the emission map using TarFlow [Zhai et al., 2025]. TarFlow provides an expressive invertible transformation that unmixes the observed features without requiring deep implementations. In contrast, affine couplings [Dinh et al., 2017] require many coupling layers to perform nonlinear unmixing, which we found to destabilise training in higher-dimensional settings.

TarFlow was originally designed for image data. Considering the n-dimensional input vectors used in our synthetic dataset, we use the corresponding non-spatial configuration with channel size 1 and patch size 1. All the synthetic data experiments use 4 TarFlow blocks, where the transformer architecture of each block uses 64 hidden dimensions, 2 layers, and 4 heads. However, instead of the permutation operation used in the original TarFlow architecture, we use invertible LU linear mixing layers, as in RealNVP Dinh et al. [2017]. We found this modification important in practice, as keeping the original permutation operation led to training failure with the rMSM failing to recover regime information. The output of the final flow block is split into rMSM features $\boldsymbol{z}_t$ and exogenous noise $\boldsymbol{\varepsilon}_t$.

Video data. To train videos, we instead use a multi-scale Glow-style architecture [Kingma and Dhariwal, 2018], which provides scalability in high-dimensional image inputs. The architecture consists of invertible 1×1 convolutions, squeeze operations, split operations, and coupling layers. Following the original Glow design, the flow is organised into B blocks, each containing L coupling layers. At each block, the squeeze operation downsamples the spatial dimension by a factor of 2, and rearranges the extra dimensions into additional channels. The split operation then discards half of the features from subsequent flow operations, but retains them as part of the latent representation.

This hierarchical structure provides a decomposition that we use to distinguish temporal features from exogenous noise. For an input image $\boldsymbol{x}_t$ and B blocks, the inverse flow computes:

$$f^{-1}(\boldsymbol{x}_t) = \big(\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(B-1)}, \boldsymbol{z}_t\big),$$

where the rMSM features represent the deepest features of the flow, and the remaining features are aggregated as exogenous noise $\boldsymbol{\varepsilon}_t = \mathrm{concat}(\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(B-1)})$. The marginal likelihood is then computed following Eq. 14. For training, both the bouncing-ball and dancing-video experiments use 7 coupling layers per block with 128 channels. We use 4 blocks for bouncing balls, and 5 blocks for dancing videos. We found that using deeper architectures made the model more prone to state collapse.

rMSM implementation details Our latent rMSM is implemented similarly to Balsells-Rodas et al. [2024]. For synthetic data, the transitions consist of two-layer MLPs with 32 hidden dimensions and a nonlinear activation that matches the one used in the data-generation process (see Section D.4). For video experiments, we use three-layer MLPs with 128 hidden dimensions and GeLU activations. In this setting, the recurrent switching process Q is parametrised using a two-layer MLP.

C.3 Likelihood-based vs VAE-based estimation

To further motivate likelihood-based models in identifiable deep generative models, we extend the discussion in Section 5 by comparing the exact likelihood objective of ΩSDS with VAE-based alternatives. Our goal is to understand how the form of a variational approximation can affect structure learning in contrast to exact inference under the same generative modelling assumptions.

We consider synthetic data with $n = m = 10$ dimensions and $K = 4$, where further data-generation and training details are discussed in Appendix D.4. Table 3 summarises the results in terms of regime F_1 score and latent recovery measured by MCC. We consider three variational approximations under the collapsed-switching factorisation $q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})\, p(\boldsymbol{s}_{1:T} \mid \boldsymbol{z}_{1:T})$ used in Balsells-Rodas et al. [2024], together with one fully amortised posterior of the form $q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})\, q(\boldsymbol{s}_{1:T} \mid \boldsymbol{x}_{1:T})$. The latter uses bidirectional RNNs to encode both continuous and discrete latent variables, and $\boldsymbol{s}_{1:T}$ is sampled using Gumbel-softmax relaxation [Jang et al., 2017]. For the collapsed-switching family, we consider the following choices for the continuous posterior: (1) a factored encoder $q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) = \prod_t q(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T})$ implemented with a shared two-layer MLP, (2) the same factored encoder under IWAE [Burda et al., 2016] using 10 importance samples, and (3) an encoder which

Table 3: Comparison of different posterior parameterisations on the synthetic data.

Posterior	MCC	Regime F_1
$\prod_t q(\boldsymbol{z}_t \mid \boldsymbol{x}_t)\, p(\boldsymbol{s}_{1:T} \mid \boldsymbol{z}_{1:T})$	0.621	0.995
$\prod_t q(\boldsymbol{z}_t \mid \boldsymbol{x}_t)\, p(\boldsymbol{s}_{1:T} \mid \boldsymbol{z}_{1:T})$ + IWAE	0.621	0.998
$q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})\, p(\boldsymbol{s}_{1:T} \mid \boldsymbol{z}_{1:T})$ (BiRNN+RNN)	0.622	0.990
$q(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})\, q(\boldsymbol{s}_{1:T} \mid \boldsymbol{x}_{1:T})$	0.629	0.757
ΩSDS	0.977	0.993

uses a bi-directional RNN, followed by a forward causal RNN, as proposed in Dong et al. [2020], Ansari et al. [2021].

The results show that VAE-based estimators can recover the regime information accurately, but produce substantially worse continuous latent representations. Furthermore, the choice of variational family is critical, as the more expressive amortised posterior fails in both structure learning and regime estimation. In comparison, ΩSDS achieves strong performance in both regime estimation and structure learning. These results suggest that identifiability alone is not sufficient for reliable latent recovery, since the estimation method plays a crucial role in uncovering the underlying dynamics. By directly optimising the exact likelihood, ΩSDS avoids estimation bottlenecks that can arise from the variational posterior design. We note this result is more pronounced in settings with many latent dimensions and a comparable number of observed dimensions. In our experiments, VAE-based disentanglement improves in lower-dimensional settings ($m = 3$ or $m = 5$), or when the observed dimension is much larger than the latent dimensions, e.g., $n = 5m$ (see Section 5.1).

D Experimental Details

D.1 Assumption to Dataset Correspondence

Below, we discuss the correspondence between the assumptions that enable our theoretical results and the synthetic scenarios explored in our work. For real-world data, establishing such correspondence without domain knowledge is an inherent limitation.

(A1) The synthetic data satisfy the invertible emission assumption by construction, including the noise-augmented setting. For bouncing-ball videos, the ball is rendered in pixel space from points in the box, but invertibility of the image renderer is difficult to ensure, since nearby points can yield the same pixel representation.

(A2) The synthetic data use neural networks with cosine activations as transition means and regime-dependent covariances. Bouncing-ball dynamics are linear when conditioned on the regimes, as shown in Eq. (10).

(A3) Synthetic data assume autonomous switching, which is a special case of the recurrent model, while bouncing-ball switches depend on box collisions.

(A4) For synthetic data, we set the diagonal entries of the transition matrix Q to 0.9. For bouncing balls, the transition $Q(\boldsymbol{z})$ is close to a diagonal in the interior of the box and changes at collision boundaries.

(A5) For likelihood separation on synthetic data, we use causally-governed dynamics, where differences in regime-dependent graphs encourage likelihood separation as discussed in Appendix B.2. Below, we numerically verify the required transition-separation conditions for the generated synthetic datasets. For bouncing ball dynamics, distinct velocities and low variance yield a log-likelihood margin of $2(\Delta t)^2/\sigma^2$ using Eq. 10. High switching stickiness in the interior gives $\epsilon^* \approx 0$. Therefore, on points sufficiently far from the box edges, we can construct $K = 4$ histories with a common endpoint. As discussed in Section 3.1, Figure 2 illustrates transition separation.

(A6) In the synthetic data, we enforce covariance dominance through data generation. The bouncing-ball dynamics showcase the equal-variance case, where the distinct velocities $\boldsymbol{v}_k$ form the vertices of a convex hull in $\mathbb{R}^2$, as discussed in Section 3.1.

Table 4: Numerical (A5) verification, where $\min_s \Delta_{\min}^{(s)}$ denotes the minimum margin found across the 5 random dataset seeds.

m	K	$\min_s \Delta_{\min}^{(s)}$
3	3	3.80
5	3	7.83
10	4	11.93
20	5	53.27

(A7) The synthetic data use full-rank transition structures, and the regime-dependent linear dynamics in the bouncing-ball data are full-rank in $\mathbb{R}^2$.

(A8) Similarly to (A6), we enforce this condition in the synthetic data. For bouncing-ball videos, this assumption does not hold since we assume equal variances.

Since the synthetic data give us access to the underlying dynamics, we can numerically verify whether the likelihood separation conditions given by (A4)–(A5) are satisfied in practice. We assess this following the discussion in Appendix B.2, where Corollary B.2 shows how likelihood separation and switching stickiness determine the required history length. Since we consider switch transitions with $Q_{kk} = 0.9,\ k = 1, \dots, K$, we have $\epsilon^* = 0.1$, and the required likelihood-separation threshold from (A5) is

$$\Delta_{\text{sep}} = \log\left(\frac{1+\epsilon^*}{1-\epsilon^*}\right) = \log(1.1/0.9) \approx 0.2007. \tag{141}$$

Therefore, given an endpoint $\boldsymbol{z}^*$ and a predecessor point $\boldsymbol{z}_1$, we can compute the likelihood margin exactly

$$\Delta_k(\boldsymbol{z}_1, \boldsymbol{z}^*) := \log p(\boldsymbol{z}^* \mid \boldsymbol{z}_1, s_2 = k) - \max_{k' \in \{1,\dots,K\}\setminus\{k\}} \log p(\boldsymbol{z}^* \mid \boldsymbol{z}_1, s_2 = k'), \tag{142}$$

which is given by the sampled regime-dependent means $\boldsymbol{m}_k$ and variances $\boldsymbol{\sigma}_k^2$. Since (A5) requires the existence of K histories, we fix the common endpoint to $\boldsymbol{z}^* = 0$ and consider two-step histories $(\boldsymbol{z}_1^{(k)}, 0)$ for each $k \in \{1, \dots, K\}$. Then, we find a predecessor $\boldsymbol{z}_1^{(k)}$ that gives a large value of $\Delta_k(\boldsymbol{z}_1^{(k)}, 0)$ while maintaining posterior dominance. That is, we compute $\boldsymbol{z}_1^{(k)} = \arg\max_{\boldsymbol{z}_1} \Delta_k(\boldsymbol{z}_1, 0)$ subject to $p(s_2 = k \mid \boldsymbol{z}_1, 0) > \frac{1}{2}$. After repeating this process for each regime, the smallest likelihood margin is $\Delta_{\min} = \min_{k \in \{1,\dots,K\}} \Delta_k(\boldsymbol{z}_1^{(k)}, 0)$. Table 4 reports the numerical verification for distinct dimension and regime-number configurations, where $\min_s \Delta_{\min}^{(s)}$ denotes the minimum transition separation across the 5 random data seeds. For each dataset configuration, the selected histories satisfy $\min_s \Delta_{\min}^{(s)} > \Delta_{\text{sep}}$, confirming that (A5) holds for our synthetic data.

D.2 Metrics

Regime F_1 score We evaluate the performance of our method to extract regimes by computing the F_1 score in terms of estimated switches $\hat{\boldsymbol{s}}_{1:T}$. We compute each $\hat{s}_t$ by selecting the regime label with highest probability weight according to the marginal posterior conditioned on $\boldsymbol{z}_{1:T}$: i.e., $\hat{s}_t = \arg\max_{k \in \{1,\dots,K\}} p(s_t = k \mid \boldsymbol{z}_{1:T})$.

MCC Mean correlation coefficient (MCC) is a standard metric to evaluate identifiable generative models [Khemakhem et al., 2020]. It evaluates whether the learned representation agrees with the ground truth up to permutation and component-wise nonlinearity. Therefore, it fits our linear scaling and permutation equivalence class for the continuous latent variables.

LPIPS LPIPS [Zhang et al., 2018] measures perceptual similarity between two images using deep visual features extracted from a pretrained network. Compared with pixel-wise errors such as MSE, LPIPS better reflects human-perceived visual differences, making it suitable for evaluating the perceptual quality of video rollouts.

D.3 Baseline specifications

Throughout our empirical evaluation, we compare our method against baselines for regime estimation, disentanglement, and forecasting. Below, we briefly describe each baseline and provide implementation details for reproducibility.

Synthetic data. For the synthetic experiments, we consider the following baselines:

- **iSDS** [Balsells-Rodas et al., 2024], based on non-recurrent SDSs with piecewise linear emissions.
- **CtrlNS** [Song et al., 2024], based on observation-dependent switching with a known number of regimes K.
- **NCTRL** [Song et al., 2023], which assumes non-Gaussian temporal latent variables and Gaussian observations.
- **TDRL** [Yao et al., 2022], which assumes a stationary non-Gaussian temporal representation.
- **iVAE** [Khemakhem et al., 2020], a non-temporal identifiable generative model. We provide ground-truth regime information as auxiliary information to obtain a stronger baseline.

For all synthetic-data baselines, public code is available and used in our experiments. Our synthetic dataset can be directly imported into the model training pipeline, and we use the default training configuration. Since SDS training can fail due to poor local optima, we train each model with 3 random seeds for each dataset and select the checkpoint with the best validation loss, corresponding to the ELBO for all cases.

Bouncing Balls For bouncing balls, we compare against VAE-based baselines, which use the same encoder-decoder representation for fairness. The encoder consists of a 5-layer CNN with 64 channels, padding 1, and kernel size 3, with LeakyReLU activations, and strides alternating between 2 and 1. A 2-layer MLP with LeakyReLU activations and 64 hidden units is used as the posterior distribution head. The decoder uses a similar architecture in reverse, starting with a 2-layer MLP with 64 hidden dimensions and LeakyReLU activations, followed by transposed convolutions under the same configuration. We use the following baselines:

- **iSDS** [Balsells-Rodas et al., 2024], described above. Notably, the encoder-decoder architecture matches the original implementation.
- **REDSDS** [Ansari et al., 2021], which proposes recurrent explicit-duration models and uses the same collapsed-switching as in Dong et al. [2020]. To improve segmentation results, we explore several annealing schedules implemented in their publicly available code to prevent state collapse and select the best result on validation ELBO.
- **SNLDS** [Dong et al., 2020], which combines collapsed-switching with recurrent switching dynamics. This is the closest VAE-based design to ΩSDS (R). We use the publicly available implementation from REDSDS. For both SNLDS and REDSDS, we adapt the implementations to image data, which was not originally supported.
- **KVAE** [Fraccaro et al., 2017], where a soft-switching Kalman filter is used as the prior dynamics. We use our own implementation. To encourage regime learning, we use a warm-up phase where the prior dynamics weights are detached for 1k iterations, as suggested in the original paper.
- **VRNN** [Chung et al., 2015], a recurrent latent variable model for sequential data. We use our own implementation.

Dancing Videos. For dancing videos, we compare with iSDS [Balsells-Rodas et al., 2024]. Since they use the same dataset in their original work, we use the same architecture as described in their implementation details.

D.4 Synthetic Data

For synthetic evaluation, we generate data using a similar procedure described in Balsells-Rodas et al. [2024], with several modifications. Since we aim to compare our likelihood-based formulation against VAE-based baselines, we consider a non-recurrent dataset to align with previous approaches. For each dataset, we sample $10{,}000$ train sequences and $1{,}000$ test sequences, both with length $T = 100$.

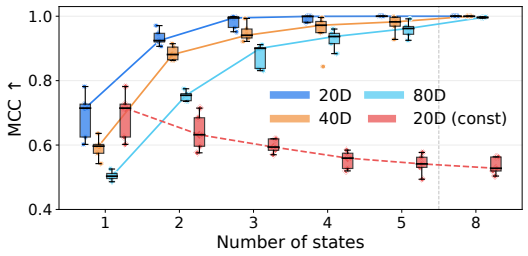

Figure 7: Ablation on (A8). Strong MCC on linear data for increasing regimes and dimensions with regime-dependent variances, with "const" using constant variances across regimes. Results are reported using 5 data seeds.

The latent regime dynamics evolve in terms of a first-order Markov chain, where Q is set to maintain the same regime with probability 90% and transition to the next one with probability 10%, and the initial distribution is uniform over the regimes. The initial latent distributions are sampled from random Gaussian components, and the mean transitions $\boldsymbol{m}_k$ are parametrised using locally connected networks [Zheng et al., 2018] where the sparsity is set to allow 3 interactions per element on average. The networks consist of two-layer MLPs with 16 hidden units and cosine activations. We modify data generation in terms of Balsells-Rodas et al. [2024], where we use regime-dependent variances and sample scales $\sigma_k \sim \mathcal{U}(0.001, 0.5)$. To avoid ambiguity in disentanglement during learning, we verify the variance ratios for pairs of regimes as described in Section 3.3, and resample if the ratios are too similar across dimensions and regime pairs. Specifically, for $m = 3, 5, 10, 20$, we resample if no pair of regimes has dimension-wise ratio values that differ by more than $0.35,\ 0.35,\ 0.10,\ 0.05$, respectively. For noise-augmented settings, we concatenate the latent variables with noise sampled from $\mathcal{N}(0, 0.01)$ at each time step. Observations are then generated using random two-layer LeakyReLU networks with a hidden dimension equal to the observation dimension.

In our experiments, we increase the number of regimes K with the latent dimension, setting $K = 3, 3, 4, 5$ for $m = 3, 5, 10, 20$. Sparsity of the transitions is also adjusted with dimension as described above. For each setting, we generate 5 data seeds and train our models on NVIDIA GeForce RTX 2080 Ti GPUs. We use the Adam optimiser [Kingma and Ba, 2015] with a learning rate of $5 \cdot 10^{-3}$ for the first 100 epochs, after which we reduce it to $5 \cdot 10^{-4}$. In all our settings, we detach the transition matrix Q for the first 10 epochs. In the noise-augmented settings, we additionally use PCA-based initialisations to encourage better initial MSM parametrisations and prevent state collapse. We train ΩSDS for 1000 epochs with batch size 64, which takes roughly 3 hours per seed for $m = 3$ and 16 hours for $m = 20$. The increased training time is due to the causal connectivity in TarFlow, as it considers a single-vector sequence. Future work could explore using a similar patch-based architecture for vector-valued inputs to improve scalability.

D.4.1 Ablation on (A8)

In Section 3.3, we introduce (A8) as a generalisation of Kivva et al. [2022, Assumption P2] for non-temporal mixture generative models. As discussed, Kivva et al. [2022, Assumption P2] establishes disentanglement by considering differences in variance ratios from a single pair of regimes. Instead, (A8) relaxes this requirement by exploiting all possible $\binom{K}{2}$ regime pairs. This generalisation is useful when the latent dimensionality increases, since dimension-wise ratio differences within a single pair of regimes may become smaller.

We assess the scalability of (A8) and experiment with ΩSDS on noiseless linear emissions and the same MSM settings described above. We consider increasing latent dimensions $m \in \{20, 40, 80\}$, increasing regimes $K \in \{1, 2, 3, 4, 5, 8\}$, and randomly sampled regime-dependent variances for each regime. We also include a control setting with constant variances across regimes and $m = 20$. Figure 7 reports the MCC of the learned latent features by ΩSDS in each setting, across 5 data seeds. We observe that increasing dimensions makes disentanglement more challenging, but the results improve when introducing additional regimes with regime-dependent variances. However, maintaining constant variances results in worse disentanglement. This shows that the regime-dependent heterogeneity condition introduced by (A8) is a scalable solution for learning disentangled representations.

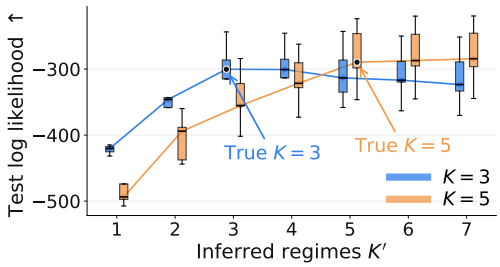

Figure 8: Model selection results with ground-truth numbers of regimes $K = 3$ and $K = 5$, with test log-likelihood across 5 data seeds.

Table 5: Results with recurrent $Q(\boldsymbol{z})$. Values are mean ± 95% CI across 5 data seeds.

Method	MCC ↑	Regime F_1 ↑	Q off-diag MAE ↓	Q diag MAE ↓
iSDS	0.722 ± 0.114	0.790 ± 0.211	0.101 ± 0.097	0.198 ± 0.197
ΩSDS	$\mathbf{0.946 \pm 0.052}$	$\mathbf{0.986 \pm 0.012}$	$\mathbf{0.022 \pm 0.010}$	$\mathbf{0.039 \pm 0.019}$

D.4.2 Selecting the number of regimes

For scenarios where the number of regimes K is unknown, it needs to be selected a priori for ΩSDS learning. Theorem 3.1 (ii) establishes identifiability of K when Q is analytic. This implies that given a generated dataset with pre-defined K, we cannot have another model with K' such that $K' \neq K$ with the same data likelihood. This connects to model selection in clustering methods such as the elbow method [Thorndike, 1953], where the selection rule considers the test log-likelihood plateau for increasing number of regimes K. This approach is similar to Balsells-Rodas et al. [2025], where identifiability of K allows for such model selection heuristics. In our synthetic experiments, the non-recurrent Q already satisfies analyticity since it is constant with respect to $\boldsymbol{z}$. Figure 8 shows model selection results for $K' = 1, \ldots, 7$ for noiseless linear emissions with $m = n = 5$, where we consider scenarios with $K = 3$ and $K = 5$, and 5 data seeds. The results show that the test log likelihood of ΩSDS plateaus when $K' = K$, and stays lower for $K' < K$. Furthermore, inspection of the regime posteriors on test data shows that the total usage of redundant states stays below 1% across all seeds for $K' = 6, 7$ when $K = 5$, and $K' = 4$ when $K = 3$. When likelihood decreases substantially, the average total usage of redundant states increases from 1.75% at $K' = 5$ to 4.75% at $K' = 7$ for $K = 3$. Therefore the elbow selection rule coincides with the observed likelihood plateau at the ground-truth number of regimes.

D.4.3 Recurrent switching processes

We explore recurrent switching in terms of our synthetic dataset, by maintaining the same regime-dependent dynamics on the continuous latent variables, and modifying the non-recurrent transition matrix Q used in Section 5.1. We parametrise $Q(\boldsymbol{z})$ as follows:

$$Q(\boldsymbol{z})_j = \text{softmax}(\boldsymbol{u}_j(\boldsymbol{z})), \quad u_{jk}(\boldsymbol{z}) = \begin{cases} \alpha, & j = k, \\ \boldsymbol{\beta}_{jk}^\top \tanh(\boldsymbol{z}), & j \neq k, \end{cases} \tag{143}$$

where $Q(\boldsymbol{z})_j$ denotes the j-th row of $Q(\boldsymbol{z})$ and u_{jk} denotes the corresponding unnormalised logit. Each vector $\boldsymbol{\beta}_{jk} \in \mathbb{R}^m$ is sampled from $\mathcal{N}(\mathbf{0}, I_m)$ and L_1-normalised to ensure $|\boldsymbol{\beta}_{jk}^\top \tanh(\boldsymbol{z})| \leq 1$ for stability. α is chosen to ensure stickiness. Similarly to the non-recurrent case, we enforce $Q_{jj}(\boldsymbol{z}) = 0.9$ at $\boldsymbol{z} = \mathbf{0}$:

$$\alpha = \log\left(\frac{(K-1)0.9}{1-0.9}\right), \tag{144}$$

which gives $\alpha \approx 2.89$ for $K = 3$. Since each off-diagonal logit lies in $[-1, 1]$, for $K = 3$ the diagonal entries are bounded by: $0.768 \leq Q_{jj}(\boldsymbol{z}) \leq 0.961$.

The above choices ensure that the switching process introduces feedback from the continuous latents, while maintaining stability and stickiness across different time steps and data seeds. The remaining

Table 6: Regime estimation and long-rollout quality reported in Table 2 with 95% CI for completeness.

Method	Input $F_1 \uparrow$	Pred. $F_1 \uparrow$	LPIPS@20 $\downarrow$	LPIPS@500 $\downarrow$
VRNN	–	–	0.014 ± 0.016	0.028 ± 0.021
KVAE	0.105 ± 0.020	0.105 ± 0.029	0.053 ± 0.029	0.324 ± 0.014
iSDS	0.819 ± 0.027	0.345 ± 0.032	0.111 ± 0.034	0.106 ± 0.022
REDSDS	0.655 ± 0.025	0.580 ± 0.051	0.073 ± 0.067	0.135 ± 0.051
SNLDS	0.998 ± 0.001	0.913 ± 0.007	0.007 ± 0.003	0.059 ± 0.017
ΩSDS (auto)	0.867 ± 0.023	0.362 ± 0.030	0.110 ± 0.034	0.092 ± 0.015
ΩSDS (recurrent)	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.932 \pm 0.004}$	$\mathbf{0.003 \pm 0.002}$	$\mathbf{0.027 \pm 0.026}$

configurations follow the non-recurrent case, including the nonlinear emissions and regime-dependent dynamics. We consider $m = 5$, $K = 3$, and noiseless emissions across 5 data seeds. Table 5 shows results for ΩSDS and iSDS, where we report MCC, regime F_1 score, and mean absolute error (MAE) for both diagonal and off-diagonal transition entries, after permutation alignment. We observe that ΩSDS achieves high MCC and regime F_1, and accurately captures the recurrent transition parametrisation $Q(\boldsymbol{z})$. In contrast, the non-recurrent iSDS performs worse on both regime and continuous latent recovery because its fixed transition matrix is misspecified.

D.5 Bouncing Ball

As described in Section 5.2, we generate videos of a bouncing ball with a resolution of 32×32. All included methods are trained for 30k steps with a batch size of 64 and a video length of $T = 64$, and we use 100 held-out samples for evaluation. In this setting, we use the Adam optimiser [Kingma and Ba, 2015] with a learning rate of 10^{-4} for the flow architecture and 10^{-3} for the rMSM architecture. We do not use PCA-based initialisation as described in Section 5. Instead, we freeze the weights of the switching network Q for 1k iterations. Training ΩSDS takes roughly 2 hours on 4 NVIDIA RTX 5080 GPUs. We use uniform dequantisation before centering the pixels, following Kingma and Dhariwal [2018]. Specifically, each normalised pixel value is mapped from a discrete bin to a continuous value by adding uniform noise within its quantisation bin. Unless otherwise stated, we use 32 dequantisation bins and fixed-seed noise for evaluation to ensure deterministic comparisons across methods. The added dequantisation noise relaxes the discrete 8-bit pixel values into a continuous input distribution while leaving the perceptual appearance of the frames essentially unchanged. Because this small pixel-level noise can disproportionately increase image-space MSE, we report rollout frame quality using the perceptual LPIPS metric instead.

We show representative regime segmentations across baselines in Figure 9, and visualisations of rollout predictions for both short and long horizons in Figures 10 and 11. These examples illustrate that ΩSDS correctly identifies the distinct velocity patterns of the ball and produces accurate forecasts. Importantly, we distinguish the non-recurrent setting, represented by iSDS and ΩSDS (A), from the recurrent setting, represented by ΩSDS (R) and SNLDS. Notably, we observe that KVAE makes the ball disappear as time passes, while no-switching baselines move the ball or split it across boundaries. The improved results in the recurrent case motivate the development of our identifiability theory for recurrent dynamics.

For quantitative comparison, Figure 12 shows LPIPS as a function of rollout time on held-out data across baselines. We observe that both recurrent-switching methods SNLDS and ΩSDS (R) maintain low perceptual error, as well as the non-switching VRNN baseline, while other switching-based methods with weaker assumptions fail to capture the long-term underlying dynamics.

D.6 AIST Dance Videos

We conduct experiments on the real-world AIST dance video dataset Tsuchida et al. [2019]. The dataset contains $12{,}670$ sequences of varying lengths, spanning 10 dance genres, with $1{,}267$ sequences per genre, and includes variations in actors and camera viewpoints. We focus on segmenting sequences from the *Middle Hip Hop* genre, reserving 100 sequences for testing. Each sequence is preprocessed by: (i) temporally subsampling the video by a factor of 4, (ii) cropping each frame around the dancer, (iii) resizing frames to 64×64.

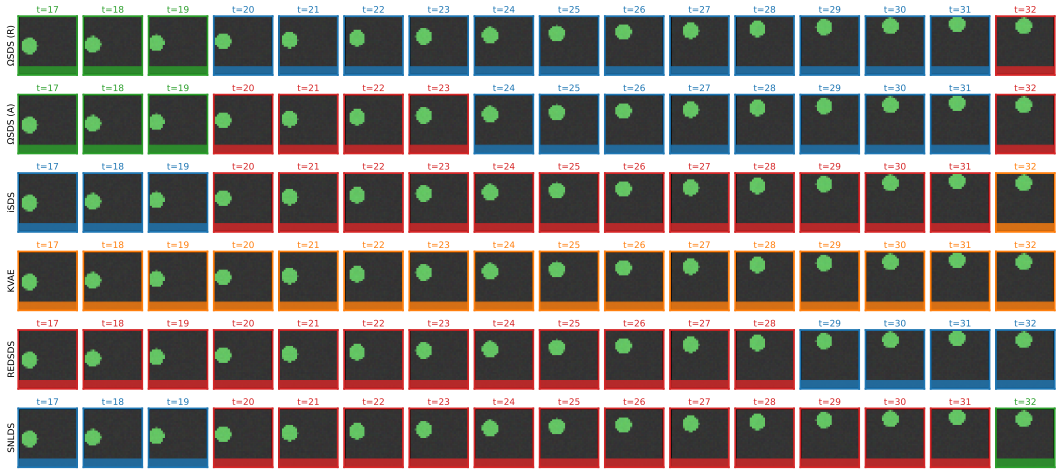

Figure 9: Posterior regimes detected by different models. Regimes are recovered up to permutations.

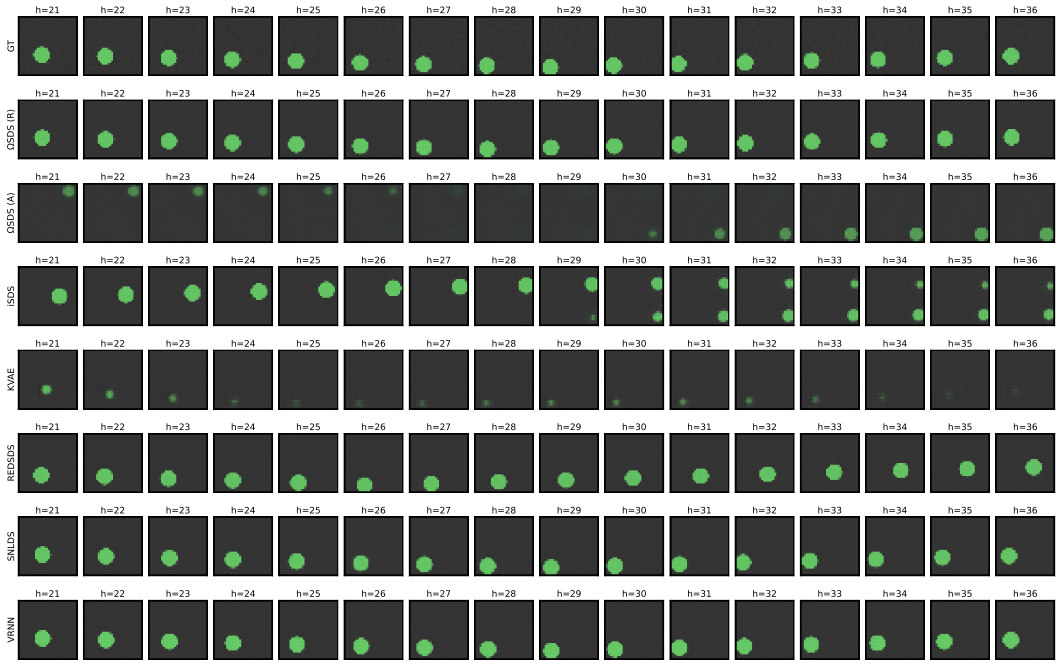

Figure 10: Comparison of short predictive rollouts with different methods. The given context length is 5.

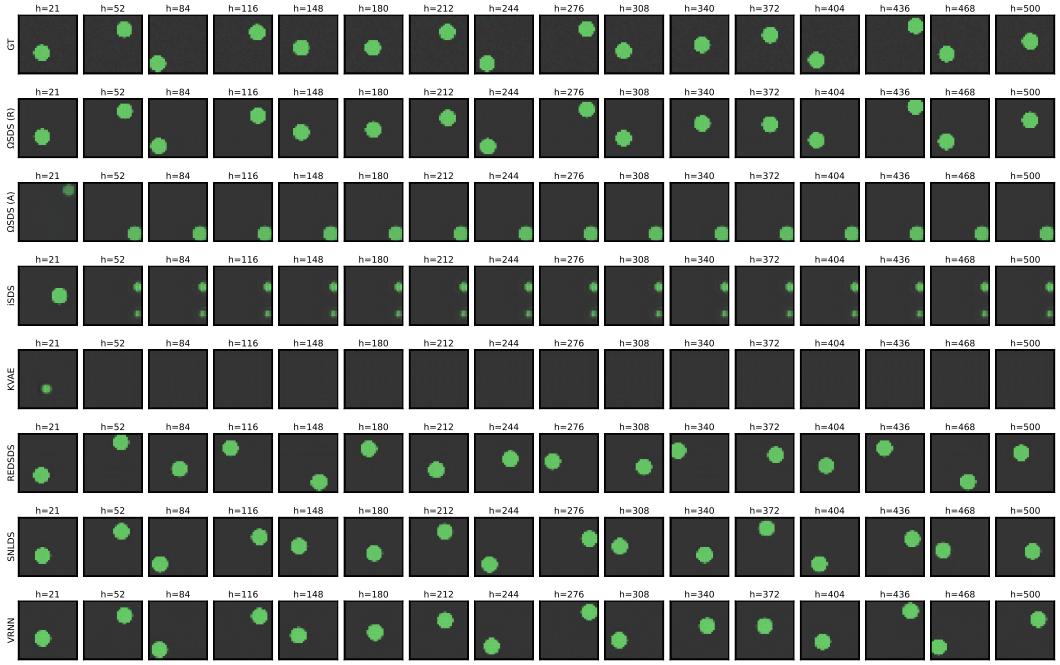

Figure 11: Comparison of long predictive rollouts with different methods. The given context length is 5.

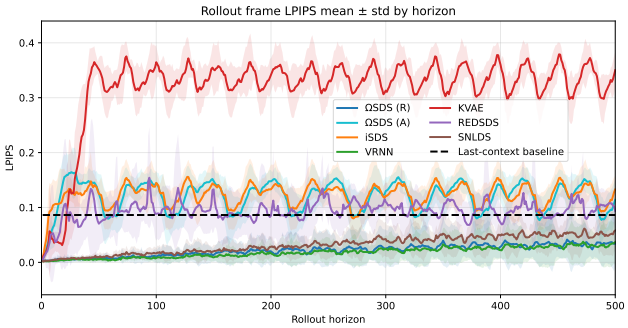

Figure 12: Plot of rollout frame LPIPS, context length = 5.

Since video lengths vary, during training we randomly choose 64 consecutive frames from each sequence. We use the same dequantisation setting as in the bouncing ball experiment (see Section D.5). We train the model under the same conditions as described in the bouncing ball experiments, and set $K = 16$. We train both ΩSDS and iSDS for 150k steps with a batch size 64, which takes roughly one day on 4 NVIDIA RTX 5080 GPUs.

Training ΩSDS on real-world videos is challenging since the number of regimes K is unknown and the underlying dynamics may not be Markovian. Despite these difficulties, ΩSDS learns meaningful regime-dependent dynamics. We show qualitative regime prediction results in Figure 13. Overall, ΩSDS produces more stable regime inference, although it suffers from state collapse in some examples (bottom). In contrast, iSDS estimates regimes with very short durations, suggesting that it fails to uncover informative regimes.

We also show rollout predictions for both ΩSDS and iSDS in Figure 14. The examples illustrate the benefit of adding recurrent feedback to the switching process, which helps the model capture future dance patterns. For example, in the top rollout sample, ΩSDS predicts expressive body-motion patterns, while iSDS preserves a stationary pose.



Figure 13: Posterior regimes of AIST dance videos.

E Broader Impacts

This work contributes to identifiable and interpretable representation learning for sequential data with regime-switching dynamics. By improving the recovery of latent states and switching processes, the proposed framework may support more transparent modelling of complex time series, with potential applications such as climate data, motion analysis, behavioural modelling, and neuroscience. Such interpretability can be beneficial when downstream decisions require understanding of the underlying dynamics rather than relying only on opaque predictions. However, applying these models to real-world human behavioural data requires caution, and inferred regimes without expert validation should not be interpreted as definitive labels.



Figure 14: MAP predictive rollout results of ΩSDS and iSDS given 15 context frames.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claim (identifiability) made in the abstract and introduction is included in Section 3 and 4, supported by empirical results in the experiment section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a discussion on limitations in Section 6

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions and proof sketch are introduced accordingly in Section 2 and 3. The corresponding proof is included in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Section D. Code is available in the following repository: `https://github.com/charlio23/omega-SDS`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Part of the dataset we used is publicly available (AIST Dance DB [Tsuchida et al., 2019]), and code for data preprocessing, generation and training can be found in `https://github.com/charlio23/omega-SDS`, where model and dataset seeds keep results reproducible from code. For specific checkpoints and pre-processed data, please contact the corresponding authors.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: In Section 5, Appendix C.2 and Appendix D we have provided all necessary experimental settings to understand the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For synthetic experiments, we run our experiments on multiple dataset seeds to verify the robustness of our empirical evidence. For video datasets, we report the error bar within the validation set for the experiments included in Section 5 and Section D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the approximate time to complete the training in Section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: -

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide discussions about broader impacts in Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available codebases for experiments with baselines and provide citations accordingly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.