Journal of Machine Learning Research 27 (2026) 1-82 Submitted 11/25; Revised 8/26; Published 11/26

On the Identifiability of Regime-Switching Models with Multi-Lag Dependencies

Carles Balsells-Rodas CB221@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, UK

Toshiko Matsui T.MATSUI19@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, UK

Pedro A.M. Mediano P.MEDIANO@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, UK

Yixin Wang YIXINW@UMICH.EDU
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Yingzhen Li YINGZHEN.LI@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, UK

Editor: Aapo Hyvärinen

Abstract

Identifiability is central to the interpretability of deep latent variable models, ensuring parameterisations are uniquely determined by the data-generating distribution. However, it remains underexplored for deep regime-switching time series. We develop a general theoretical framework for multi-lag Regime-Switching Models (RSMs), encompassing Markov Switching Models (MSMs) and Switching Dynamical Systems (SDSs). For MSMs, we formulate the model as a temporally structured finite mixture and prove identifiability of both the number of regimes and the multi-lag transitions in a nonlinear-Gaussian setting. For SDSs, we establish identifiability of the latent variables up to permutation and scaling via temporal structure, which in turn yields conditions for identifiability of regime-dependent latent causal graphs (up to regime/node permutations). Our results hold in a fully unsupervised setting through architectural and noise assumptions that are directly enforceable via neural network design. We complement the theory with a flexible variational estimator that satisfies the assumptions and validate the results on synthetic benchmarks. Across real-world data sets from neuroscience, finance, and climate, identifiability leads to more trustworthy interpretability analysis, which is crucial for scientific discovery.

Keywords: identifiability, regime-switching models, Markov switching models, switching dynamical systems, time series, state-space models, finite mixture models, variational inference, nonlinear ICA, causal representation learning

1. Introduction

Identifiability in deep latent variable models has been a long-standing challenge over the recent years. It formalises when parameters are uniquely determined by the data-generating distribution. In practice, exact parameter recovery is often impossible, and instead, one assumes equivalence classes which are determined by the target of interest. For example, Causal Representation Learning (CRL; Schölkopf et al., 2021) targets the recovery of the

©2026 Carles Balsells-Rodas, Toshiko Matsui, Pedro A.M. Mediano, Yixin Wang and Yingzhen Li.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2778.html.

latent causal graph, typically up to permutation of nodes. In Independent Component Analysis (ICA; Comon, 1994), the equivalence class targets independent sources up to permutations and scaling, and often the functional form of the mixing function.

General nonlinear ICA is ill-posed without additional structure (Hyvärinen and Pajunen, 1999). Recent advances establish identifiability by combining deep generative models with auxiliary information (Khemakhem et al., 2020), or finite mixture models with piecewise linear mixing (Kivva et al., 2022). Other advances exploit temporal structure (Hyvarinen and Morioka, 2017; Hyvarinen et al., 2019; Hälvä and Hyvarinen, 2020; Hälvä et al., 2021), or latent causal processes (Yao et al., 2022a,b; Lippe et al., 2023b,a).

This paper focuses on identifiability for *Regime-Switching Models* (RSMs), including Markov Switching Models (MSMs) and Switching Dynamical Systems (SDSs), from the perspective of deep sequential generative models. RSMs describe time series whose behaviour depends on a discrete latent *regime* (state, or switch) that changes over time. They form a subset of *State-Space Models* (SSMs), a well-established class of sequence models (Lindgren, 1978; Poritz, 1982; Hamilton, 1989). With the increasing popularity of sequential latent variable models (Chung et al., 2015; Li and Mandt, 2018; Babaeizadeh et al., 2018; Saxena et al., 2021), SSMs have been incorporated as latent dynamical priors (Johnson et al., 2016; Linderman et al., 2016, 2017; Fraccaro et al., 2017; Dong et al., 2020; Ansari et al., 2021; Liu et al., 2023; Geadah et al., 2024). However, these works mainly focus on designing flexible priors and stable training schemes (Dong et al., 2020; Ansari et al., 2021; Liu et al., 2023; Zhao and Linderman, 2023).

To address these gaps, we develop a unified theoretical framework that establishes identifiability for RSMs with multi-lag dependencies. First, we show that multi-lag MSMs can be framed as finite mixture models with temporally-structured components, and prove identifiability of the number of regimes and the multi-lag transition functions in a nonlinear Gaussian setting. The analysis extends classical finite-mixture identifiability (Yakowitz and Spragins, 1968) to the multi-lag case, which is nontrivial due to lagged dependencies. Second, for SDSs, we prove identifiability of the latent variables up to permutation and scaling by extending Kivva et al. (2022) to the temporal setting. This yields identifiability of the regime-dependent latent causal graphs up to permutations of regimes and nodes. Our results hold in a fully unsupervised setting by imposing restrictions on the latent noise distributions and the functional class of the generative model. In contrast, other works that handle regime-switching require auxiliary information (Yao et al., 2022a,b; Lippe et al., 2023a) or stationarity (Hyvarinen and Morioka, 2017) in the temporal correlations (Hälvä et al., 2021). Recently, similar results with flexible assumptions have been established (Song et al., 2023, 2024; Balsells-Rodas et al., 2025).

The rest of the paper is organised as follows.[1] Section 2 discusses related work and positions our contributions. Section 3 reviews finite mixtures and latent variable models. Section 4 presents our main results: (i) parametric MSMs with nonlinear Gaussian transitions

1. This paper extends the conference version Balsells-Rodas et al. (2024) significantly by: (i) providing a more intuitive view of Markov Switching Models as finite mixtures and extending identifiability to higher temporal dependencies (with substantial new proofs and intuitions in the main text), (ii) going beyond affine identifiability of the latent variables for SDSs to establish permutation/scaling identifiability and causal structure identifiability up to node/regime permutation, and (iii) adding new real-world applications.

(nonparametric analysis is deferred to Appendix A), (ii) identifiability of SDS latents up to permutation and scaling, and (iii) identifiability of regime-dependent causal graphs up to regime/node permutation. Section 5 presents a flexible estimation approach which follows the assumptions for identifiability, and builds on Dong et al. (2020); Ansari et al. (2021). Section 6 validates our theoretical framework on synthetic data with baselines and ablations. Finally, Section 7 presents results on brain activity, financial time series, and climate data, illustrating interpretability gains enabled by identifiability.

Why identifiability matters in practice. Applications across neuroscience (Karniol-Tambour et al., 2024), climate (Saggioro et al., 2020), human motion (Dong et al., 2020; Liu et al., 2023), and animal behaviour (Ansari et al., 2021) demonstrate the promise of deep regime-switching models. However, without identifiability, interpretations of the learned latents (e.g., regimes, causal structures) may be spurious, as distinct or even contradictory parametrisations can achieve the same likelihood, thus invalidating scientific discovery. More broadly, identifiability prevents arbitrary latent reparametrisations (Locatello et al., 2019) and supports valid causal claims (Schölkopf et al., 2021). Our theoretical framework provides flexible conditions under which learned regimes, latents, and graphs are uniquely determined up to acceptable equivalences and therefore support trustworthy interpretations.

2. Related Work

MSMs (or autoregressive hidden Markov models) were first introduced for applications to speech analysis (Poritz, 1982; Ephraim and Roberts, 2005). In econometrics, Hamilton (1989) popularised MSMs for business cycles, with numerous extensions such as MS-GARCH (Haas et al., 2004). For overviews of asymptotic properties and additional applications, see Frühwirth-Schnatter (2006). SDSs extend MSMs with continuous latent states and regime-dependent dynamics. Classic developments include switching Kalman filters (Murphy, 1998) and variational learning for respiration data (Ghahramani and Hinton, 2000). Recently, regime switching has been used as a prior within deep latent-variable models, including SVAEs (Johnson et al., 2016; Zhao and Linderman, 2023), soft-switching Kalman Filters (Fraccaro et al., 2017), and recurrent SDSs (Linderman et al., 2016; Dong et al., 2020) with explicit duration models (Ansari et al., 2021) or graph-based interactions (Liu et al., 2023).

Finite-state HMMs enjoy strong identifiability under mild conditions (Allman et al., 2009; Gassiat et al., 2016, 2020). However, these strategies do not translate directly to MSMs due to explicit autoregression in continuous variables. While results exist for certain discrete MSMs (An et al., 2013), to our knowledge, a general nonparametric theory for multi-lag MSMs has been missing. Our MSM identifiability results address this gap by extending classical finite-mixture identifiability (Yakowitz and Spragins, 1968) to the multi-lag setting.

Modern identifiability for deep generative models often builds on nonlinear ICA with auxiliary variables or temporal structure. iVAE establishes identifiability with conditional priors from auxiliary information (Khemakhem et al., 2020), while Kivva et al. (2022) remove this requirement by using mixture priors with piecewise linear mixing. Temporal information can also act as auxiliary information, yielding identifiability through contrastive learning (Hyvarinen and Morioka, 2016; Hyvarinen et al., 2019). Recent work in CRL establishes temporal identifiability from variability or distribution changes in the latent dynamics (Yao et al., 2022b,a), and from temporal interventions (Lippe et al., 2023b,a). These approaches

Family	Method	Emission	Switching	Nonlin. dyn.	Identifies K	Full ident.
MSM	SDCI (Balsells-Rodas et al., 2025)	N/A	Recurrent	✓	✓	✗
	FANTOM (Rahmani and Frossard, 2025b)	N/A	Independent	✓	✓	✓
	iMSM (ours)	N/A	Autonomous	✓	✓	✓
SDS	HM-NICA (Hälvä and Hyvarinen, 2020)	Bijective	Autonomous	✗	✗	✗
	SNICA (Hälvä et al., 2021)	Invertible	Autonomous	✗	✗	✗
	NCTRL (Song et al., 2023)	Gaussian obs.	Autonomous	✓	✓	✓
	CtrlNS (Song et al., 2024)	Invertible	Obs.-dependent	✓	✗	✓
	iSDS (ours)	PWL	Autonomous	✓	✓	✓

Table 1: Comparison with related identifiable regime-switching works. Full ident. denotes recovery of both the transition mechanisms and switching process, excluding identification of the number of regimes K. PWL denotes piecewise linear.

primarily target disentangled representations, whereas our framework further identifies the underlying regime-switching mechanisms.

Several works introduce latent regime structure to relax stationarity or the need for auxiliary information. HM-NICA (Hälvä and Hyvarinen, 2020) replaces observed regimes with an identifiable stationary HMM prior (Gassiat et al., 2016); IIA-HMM (Morioka et al., 2021) combines recurrent nonlinear mixing with nonstationary innovations governed by a stationary HMM (Gassiat et al., 2016). SNICA (Hälvä et al., 2021) considers linear SDSs with weak stationarity (Hyvarinen and Morioka, 2017), but does not identify the number of switches or latent transitions. NCTRL (Song et al., 2023) establishes results for SDSs. Its strategy follows first-order MSM results presented in Balsells-Rodas et al. (2024), with Gaussianity imposed at the observation level. CtrlNS (Song et al., 2024) achieves identifiability via mechanism sparsity (Lachapelle et al., 2022) and direct observation-regime dependence, assuming a known number of regimes. Very recent works explicitly target regime-dependent causal discovery: CASTOR (Rahmani and Frossard, 2025a) and FANTOM (Rahmani and Frossard, 2025b) allow contemporaneous effects but assume i.i.d. regimes, and SDCI (Balsells-Rodas et al., 2025) allows flexible state dependencies, but does not identify the underlying switching process. Our results are complementary to these methodologies, as multi-lag MSMs can be viewed as a generalisation directly targeting regime-dependent transitions. Table 1 summarises the above distinctions.

Positioning. Compared with prior identifiability results for regime-switching and temporal CRL/ICA, our framework offers the following key advantages:

- **Multi-Lag MSM theory.** To our knowledge, we provide the first general identifiability results for multi-lag MSMs. As a theoretical extension, we also discuss how the analysis applies beyond standard parametric MSMs (see Remark 9 and Appendix A).
- **Assumptions at the function level.** Related works on temporal CRL (Yao et al., 2022b,a; Song et al., 2023, 2024) and ICA (Morioka et al., 2021; Hälvä et al., 2021) often rely on distributional or tail assumptions that are difficult to verify when designing models. Instead, we directly impose functional constraints that can be enforced via the neural parametrisation, making model specification more controllable.
- **Unknown number of regimes.** Methods such as HMM-ICA (Hälvä and Hyvarinen, 2020) and CtrlNS (Song et al., 2024) assume a known number of regimes. By extending classical finite-mixture theory (Yakowitz and Spragins, 1968), our results do not require

knowing the number of regimes. This matters in practice, as we can determine them by model selection (McLachlan and Peel, 2000), which is not applicable when the number of regimes is assumed known.

Our SDS identifiability result is restricted to piecewise linear emission functions. This class allows neural network parametrisations while maintaining flexible latent switching dynamics. However, it is more restrictive than the general nonlinear invertible emissions considered in other nonlinear ICA and temporal CRL approaches. Extending our theory to broader nonlinear emission families is left for future work.

3. Background

We first review classical theoretical results on the identifiability of finite mixture models, followed by recent identifiability results for non-temporal deep latent variable models. We then introduce regime-switching models, which are central to this work.

3.1 Finite Mixture Models

The identifiability of finite mixture models has been extensively studied in the statistical literature. Classical results include Teicher (1963), which studied identifiability for families with a finite number of elements, and Barndorff-Nielsen (1965), which established identifiability results for exponential families. Our theoretical framework presented in Section 4.1 extends finite mixture model results from Yakowitz and Spragins (1968), which show identifiability through linear independence of the mixing components. This characterisation will serve as the main tool for identifiability of MSMs.

Consider a distribution family with functions defined on $\boldsymbol{z} \in \mathbb{R}^m$,

$$\mathcal{F}_{\mathcal{A}} := \{F_a(\boldsymbol{z}) \mid a \in \mathcal{A}\}, \tag{1}$$

where $F_a(\boldsymbol{z})$ is an m-dimensional CDF. The index set $\mathcal{A}$ is assumed to satisfy that $F_a(\boldsymbol{z})$, as a function of $(\boldsymbol{z}, a)$, is measurable on $\mathbb{R}^m \times \mathcal{A}$. We introduce the notion of linear independence under finite mixtures of a family $\mathcal{F}_{\mathcal{A}}$.

Definition 1 *A family of functions $\mathcal{F}_{\mathcal{A}}$ (Eq. (1)) is said to contain linearly independent functions under finite mixtures if for any $\mathcal{A}_0 \subset \mathcal{A}$ such that $|\mathcal{A}_0| < +\infty$, the functions in $\{F_a(\boldsymbol{z}) \mid a \in \mathcal{A}_0\}$ are linearly independent.*

The above definition is a weaker requirement of linear independence on function families as it allows linear dependence from the linear combination of infinitely many other functions. Consider the following finite mixture distribution family:

$$\mathcal{H}_{\mathcal{A}} := \left\{ H(\boldsymbol{z}) = \sum_{i=1}^{K} c_i F_{a_i}(\boldsymbol{z}) \mid c_i > 0,\ a_i \in \mathcal{A},\ a_i \neq a_j, \forall i \neq j,\ \sum_{i=1}^{K} c_i = 1,\ K < +\infty \right\}, \tag{2}$$

which is defined from the *convex hull* of $\mathcal{F}_{\mathcal{A}}$ ($\mathrm{conv}(\mathcal{F}_{\mathcal{A}})$), or a linear combination of CDFs in $\mathcal{F}_{\mathcal{A}}$, with positive coefficients. Now we define the *identifiable finite mixture family* following Yakowitz and Spragins (1968).

Definition 2 *The finite mixture family $\mathcal{H}_{\mathcal{A}}$ is said to be identifiable up to permutations, when for any two finite mixtures $H(\boldsymbol{z}) = \sum_{i=1}^{K} c_i F_{a_i}(\boldsymbol{z})$ and $\tilde{H}(\boldsymbol{z}) = \sum_{i=1}^{\tilde{K}} \tilde{c}_i F_{\tilde{a}_i}(\boldsymbol{z})$, $H(\boldsymbol{z}) = \tilde{H}(\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathbb{R}^m$, if and only if $K = \tilde{K}$ and for each $1 \leq i \leq K$ there is some $1 \leq j \leq \tilde{K}$ such that $c_i = \tilde{c}_j$ and $F_{a_i}(\boldsymbol{z}) = F_{\tilde{a}_j}(\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathbb{R}^m$.*

Then, identifiability of finite mixture models is stated as follows.

Proposition 3 *(Yakowitz and Spragins, 1968) The finite mixture distribution family $\mathcal{H}_{\mathcal{A}}$ is identifiable up to permutations, if and only if functions in $\mathcal{F}_{\mathcal{A}}$ are linearly independent under the finite mixtures.*

For our purposes in this work, the above result can be generalised to PDFs defined on $\boldsymbol{z} \in \mathbb{R}^m$. Consider an equivalent family of PDFs:

$$\tilde{\mathcal{F}}_{\mathcal{A}} = \left\{ f_a(\boldsymbol{z}) = \frac{\partial^m}{\partial z_1 \dots \partial z_m} F_a(\boldsymbol{z}) \mid a \in \mathcal{A} \right\}. \tag{3}$$

Then, under the assumption that all F_a are absolutely continuous and differentiable almost everywhere, the family $\mathcal{F}_{\mathcal{A}}$ is linearly independent under finite mixtures if and only if the corresponding family $\tilde{\mathcal{F}}_{\mathcal{A}}$ is linearly independent under finite mixtures.

Proposition 4 *Consider the family $\tilde{\mathcal{F}}_{\mathcal{A}}$ in Eq. (3). Then the finite mixture distribution family is identifiable up to permutations if and only if functions in $\tilde{\mathcal{F}}_{\mathcal{A}}$ are linearly independent under finite mixtures.*

3.2 Identifiable Latent Variable Models

In the non-temporal case, many works explore identifiability of latent variable models (Khemakhem et al., 2020; Kivva et al., 2022). Specifically, consider a generative model where its latent variables $\boldsymbol{z} \in \mathbb{R}^m$ are drawn from a Gaussian mixture prior with K components ($K < +\infty$). Then $\boldsymbol{z}$ is transformed via a (noisy) piecewise linear mapping to obtain the observation $\boldsymbol{x} \in \mathbb{R}^n$, $n \geq m$:

$$\boldsymbol{x} = f(\boldsymbol{z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{z} \sim p(\boldsymbol{z}) := \sum_{k=1}^{K} p(s=k) \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{4}$$

Kivva et al. (2022) establish that if the transformation f is *weakly injective* (see definition below), the prior distribution $p(\boldsymbol{z})$ is identifiable up to affine transformations from the observations.[2] If we further assume f is continuous and *injective*, both the prior distribution $p(\boldsymbol{z})$ and the nonlinear mapping f are identifiable up to affine transformations. In Section 4.2 we extend these results to establish identifiability for SDSs using similar proof strategies. *Notation.* For $(\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_M^*) \in \mathbb{R}^{nM}$ and $M < +\infty$, we define $B(\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_M^*, \delta) := \{(\boldsymbol{x}_1, \dots, \boldsymbol{x}_M) \in \mathbb{R}^{nM} : ||(\boldsymbol{x}_1, \dots, \boldsymbol{x}_M) - (\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_M^*)|| < \delta\}$, where $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^{nM}$. Furthermore, we extend function notation to sets, i.e., for a set $\mathcal{S}$, $f(\mathcal{S}) := \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{S}\}$ and $f^{-1}(\mathcal{S}) := \{\boldsymbol{x} : f(\boldsymbol{x}) \in \mathcal{S}\}$.

Definition 5 *A mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ is said to be weakly injective if:*

2. See Def. 2.2 in Kivva et al. (2022), or the equivalent adapted to SDSs (Def. 12).

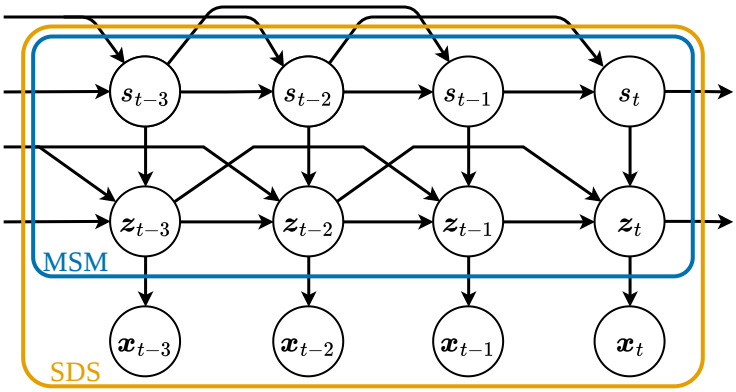


Figure 1: The generative models considered in this work. MSM (blue) treats $\boldsymbol{z}_t$ directly as observations, while SDS (orange) transforms $\boldsymbol{z}_t$ into observed $\boldsymbol{x}_t$. Adapted from Balsells-Rodas et al. (2024).

(i) there exists $\boldsymbol{x}_0 \in \mathbb{R}^n$ *and* $\delta > 0$ *s.t.* $|f^{-1}(\{\boldsymbol{x}\})| = 1$ *for every* $\boldsymbol{x} \in B(\boldsymbol{x}_0, \delta) \cap f(\mathbb{R}^m)$*, and*

(ii) $\{\boldsymbol{x} \in \mathbb{R}^n : |f^{-1}(\{\boldsymbol{x}\})| > 1\} \subseteq f(\mathbb{R}^m)$ *has measure zero with respect to the Lebesgue measure on* $f(\mathbb{R}^m)$*.*

Definition 6 *f is injective if* $|f^{-1}(\{\boldsymbol{x}\})| = 1, \forall \boldsymbol{x} \in f(\mathbb{R}^m)$*.*

3.3 Switching Dynamical Systems

A Switching Dynamical System (SDS), with an example graphical model illustrated in Figure 1, is a sequential latent variable model with its dynamics governed by both discrete and continuous latent states, $s_t \in \{1, \dots, K\}$, $\boldsymbol{z}_t \in \mathbb{R}^m$, respectively. At each time step t, the discrete state s_t determines the regime of the dynamical prior that the current continuous latent variable $\boldsymbol{z}_t$ should follow, and the observation $\boldsymbol{x}_t$ is generated from $\boldsymbol{z}_t$ using a (noisy) nonlinear transformation. This gives the following probabilistic model for the states and the observation variables:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{s}_{1:T}) = p_{\boldsymbol{\theta}}(\boldsymbol{s}_{1:T}) p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{s}_{1:T}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{z}_t), \tag{5}$$

where we use $\boldsymbol{\theta}$ to denote the model parameters. As $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T})$ is intractable, recent works (Dong et al., 2020; Ansari et al., 2021) have developed inference techniques based on variational inference (Kingma and Welling, 2014). These works consider an additional dependency of the switch s_t on $\boldsymbol{x}_t$ for more expressive segmentations, which is not considered in our theoretical analysis for simplicity.

The latent dynamic prior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$ is a Markov Switching Model (MSM), also referred to as an autoregressive HMM (Hamilton, 1989; Ephraim and Roberts, 2005). This type of prior uses the latent "switch" s_t to condition the distribution of $\boldsymbol{z}_t$ at each time-step, and the conditional dynamic model of $\boldsymbol{z}_{1:T}$ given $\boldsymbol{s}_{1:T}$ follows an autoregressive process, where the likelihood can be computed by marginalising $\boldsymbol{s}_{1:T}$. For simplicity, we consider $\boldsymbol{s}_{1:M} := s_M$

and $p_{\boldsymbol{\theta}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M \mid s_M)$ as the distribution of the initial M observations, and the MSM with lag M factorises as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) = \sum_{\boldsymbol{s}_{M:T}} p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T}) p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{s}_{M:T}),$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{s}_{M:T}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M \mid s_M) \prod_{t=M+1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t). \quad (6)$$

Note that the structure of the discrete latent state prior $p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T})$ is not specified, and the identifiability results presented in the next section do not require further assumptions herein. In experiments, we use a first-order Markov process for $p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T})$, described by a transition matrix $Q \in \mathbb{R}^{K \times K}$ such that $p_{\boldsymbol{\theta}}(s_t = j \mid s_{t-1} = i) = Q_{ij}$, and an initial distribution $p_{\boldsymbol{\theta}}(s_M = i) = \pi_i$.

4. Identifiability Analysis

In this section, we establish the identifiability of the SDS (Eq. (5)) presented in Section 3.3 under suitable assumptions. We start by extending finite mixture model identifiability results introduced in Section 3.1 to Markov switching models with finite lag M. Then, we leverage ideas from Kivva et al. (2022) which use a finite-mixture prior for non-temporal generative models; importantly, this theory relies on the use of identifiable finite mixture priors (up to mixture component permutations).

Formally, identifiability establishes an injective correspondence between probability distributions $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and model parameters $\boldsymbol{\theta}$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x}) \iff \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}. \quad (7)$$

Given that our models are parametrised by deep neural networks, explicit identifiability of parameters is not possible due to network symmetries (i.e., different parameters map to the same function). Instead, in this paper we refer to identifiability in terms of their functional form, and model mixture components.

Notation. In this section, we use M to denote the temporal lag, m to denote the dimensionality of continuous latent variables or the input dimensionality in MSMs, and n to denote the dimensionality of observations or the inputs to SDSs. We use the following notation for repeated Cartesian products of sets and pointwise products of function families:

$$\times^N \mathcal{A} := \underbrace{\mathcal{A} \times \cdots \times \mathcal{A}}_{N \text{ times}}, \quad \otimes^N \mathcal{P} := \underbrace{\mathcal{P} \otimes \cdots \otimes \mathcal{P}}_{N \text{ times}}, \quad (8)$$

where $\mathcal{A}$ is an index set and $\mathcal{P}$ is a family of functions. The functions in $\mathcal{P}$ may share overlapping variables, resulting in functions defined in the joint domain. For example, consider two families

$$\mathcal{P}_{1,\mathcal{A}_1} = \{p_{1,a_1}(\boldsymbol{z}_1, \boldsymbol{z}_2) : a_1 \in \mathcal{A}_1\}, \quad \mathcal{P}_{2,\mathcal{A}_2} = \{p_{2,a_2}(\boldsymbol{z}_2, \boldsymbol{z}_3) : a_2 \in \mathcal{A}_2\}, \quad (9)$$

with $(\boldsymbol{z}_1, \boldsymbol{z}_2), (\boldsymbol{z}_2, \boldsymbol{z}_3) \in \mathbb{R}^{2m}$. Then, their pointwise product is defined as:

$$\mathcal{P}_{1,\mathcal{A}_1} \otimes \mathcal{P}_{2,\mathcal{A}_2} := \Big\{ p_{1,a_1}(\boldsymbol{z}_1, \boldsymbol{z}_2) p_{2,a_2}(\boldsymbol{z}_2, \boldsymbol{z}_3) \mid (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2 \Big\}, \quad (10)$$

where functions are defined in $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) \in \mathbb{R}^{3m}$. We also write $[K] := \{1, \ldots, K\}$ for index sets from 1 to K.

4.1 Identifiable High-Order Markov Switching Models

4.1.1 Finite Mixture Model View of MSMs

For simplicity, we drop the subscript $\boldsymbol{\theta}$. The MSM $p(\boldsymbol{z}_{1:T})$ has an equivalent formulation as a finite mixture model, where from Eq. (6), the joint distribution given a sequence of switches $p(\boldsymbol{z}_{1:T} \mid \boldsymbol{s}_{M:T})$ is defined from the initial and transition distributions with lag M: $p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M \mid s_M)$ and $p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t)$ for $t > M$ respectively. Equivalently as in Yakowitz and Spragins (1968), we first define function families corresponding to initial and transition distributions.

$$\Pi_{\mathcal{A}}^{M} := \Big\{ p_a(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M) \mid a \in \mathcal{A} \Big\}, \quad \mathcal{P}_{\mathcal{B}}^{M} := \Big\{ p_b(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) \mid b \in \mathcal{B} \Big\}, \quad (11)$$

where $\mathcal{A}$ and $\mathcal{B}$ are index sets satisfying mild measure theoretic conditions (similarly as in Section 3.1). To exemplify, we can consider these families to be infinitely many parametrised functions where each $a \in \mathcal{A}$ corresponds to a different parameter configuration $\boldsymbol{\theta}_a$. We use different calligraphic symbols (e.g., $\mathcal{A}$, $\mathcal{B}$) because they index functions with different structures and output domains. For example, if the initial distribution family is a Gaussian family, the set $\mathcal{A}$ indexes elements in a finite-dimensional parameter space (means and covariances). In contrast, the set $\mathcal{B}$ used in the transition family indexes elements in a function space, the output of which defines the parameters of a conditional distribution. We provide details of parametrisations below.

We assume a general setting and consider an MSM whose initial distribution has K_0 regimes, where the transition distribution has K regimes. Therefore, the MSM discrete regimes satisfy $s_t \in [K]$ with $K < +\infty$, and $M < t \leq T$, with an initial regime $s_M \in [K_0]$, $K_0 < +\infty$. Given the above formulation, a Markov switching model considers K_0 initial distributions $\Pi_{\mathcal{A}_0}^{M}$ and K transition distributions $\mathcal{P}_{\mathcal{B}_0}^{M}$, where $|\mathcal{A}_0| = K_0$, $|\mathcal{B}_0| = K$, and $\mathcal{A}_0 \subset \mathcal{A}$, $\mathcal{B}_0 \subset \mathcal{B}$ respectively. Therefore, under these function families we have for each $k \in [K_0]$: $p(\boldsymbol{z}_{1:M} \mid s_M = k) = p_a(\boldsymbol{z}_{1:M})$, for some $a \in \mathcal{A}_0$; and for each $k \in [K]$: $p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t = k) = p_{b_t}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1:t-M})$ for some $b_t \in \mathcal{B}_0$ with $M < t \leq T$. In other words, the families $\Pi_{\mathcal{A}}^{M}$ and $\mathcal{P}_{\mathcal{B}}^{M}$ contain infinitely many functions, and each of the distributions in the MSM aligns with one of such functions.

We can view the joint distribution $p(\boldsymbol{z}_{1:T})$ as a mixture of paths, where each path is defined by a realisation of $\boldsymbol{s}_{M:T} \in [K_0] \times [K]^{T-M}$. Therefore, one can define a bijective *path indexing function* $\varphi : [C] \to [K_0] \times [K]^{T-M}$, with a total number of mixture paths $C = K_0 K^{T-M}$, such that each $i \in [C]$ uniquely retrieves a set of states $\boldsymbol{s}_{M:T} = \varphi(i)$. Then we can use $c_i = p(\boldsymbol{s}_{M:T} = \varphi(i))$ to represent the joint probability of the states $\boldsymbol{s}_{M:T} = \varphi(i)$ under p. Then for any M, T such that $M < T < +\infty$, we can construct the family of M-order

MSMs in finite mixture model sense:

$$\mathcal{M}^T(\Pi^M_{\mathcal{A}}, \mathcal{P}^M_{\mathcal{B}}) := \Big\{ \sum_{i=1}^{C} c_i p_{a^i}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b^i_t}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}) \mid c_i > 0,$$
$$p_{a^i} \in \Pi^M_{\mathcal{A}_0}, \quad p_{b^i_t} \in \mathcal{P}^M_{\mathcal{B}_0}, \quad M < t \leq T, \quad a^i \in \mathcal{A}_0 \subset \mathcal{A}, \quad b^i_t \in \mathcal{B}_0 \subset \mathcal{B}, \quad K_0, K < +\infty,$$
$$|\mathcal{A}_0| = K_0, \quad |\mathcal{B}_0| = K, \quad (a^i, b^i_{M+1:T}) \neq (a^j, b^j_{M+1:T}), \forall i \neq j, \quad \sum_{i=1}^{C} c_i = 1 \Big\}. \quad (12)$$

Since Eq. (12) requires $(a^i, b^i_{M+1:T}) \neq (a^j, b^j_{M+1:T})$ for any $i \neq j$, distinct mixture components correspond to distinct distribution trajectories. This prevents repeated components, and yields an injective mapping $\phi : [C] \to \mathcal{A}_0 \times (\times^{T-M} \mathcal{B}_0)$, such that $\phi(i) = (a^i, b^i_{M+1}, \dots, b^i_T)$. Combined with the path indexing function, we can use $\phi \circ \varphi^{-1}$ to uniquely map a set of states $\boldsymbol{s}_{M:T}$ to the indices.

The above notation illustrates how the MSM extends finite mixture models to temporal settings representing $p(\boldsymbol{z}_{1:T})$ as a finite mixture of $K_0 K^{T-M}$ trajectories, each composed of an initial distribution from $\Pi^M_{\mathcal{A}_0}$ and conditional distributions from $\mathcal{P}^M_{\mathcal{B}_0}$. However, this notion differs fundamentally from the i.i.d. mixture model setting in Yakowitz and Spragins (1968). To formalise this difference, we define the following *trajectory family*:

$$\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}} := \Pi^M_{\mathcal{A}} \otimes (\otimes^{T-M} \mathcal{P}^M_{\mathcal{B}}) = \Big\{ p_a(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b_t}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}) \mid$$
$$a \in \mathcal{A},\ b_t \in \mathcal{B},\ M < t \leq T \Big\}. \quad (13)$$

In Section 3.1, $\mathcal{H}_{\mathcal{A}}$ is defined as the *convex hull* $\operatorname{conv}(\mathcal{F}_{\mathcal{A}})$. In contrast, the MSM family defined in Eq. (12) forms a strict subset of $\operatorname{conv}(\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}})$, since the mixture components are restricted to finite subsets $\mathcal{A}_0 \subset \mathcal{A}$ and $\mathcal{B}_0 \subset \mathcal{B}$. On the other hand, mixtures over $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$ allow arbitrary selections from $\mathcal{B}$ at each time step, yielding a more expressive model family. This generalisation is crucial for our theoretical analysis below, where we prove identifiability of MSM by establishing linear independence under finite mixtures of $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$.

4.1.2 Identifiable Markov Switching Models

Having established the finite mixture model view of Markov switching models, we will use this notation of MSM in the rest of Section 4.1, as we will use finite mixture modelling techniques to establish its identifiability. In detail, we first define the identification of $\mathcal{M}^T(\Pi^M_{\mathcal{A}}, \mathcal{P}^M_{\mathcal{B}})$ as follows.

Definition 7 *We say the family of Markov switching models with order M and length T, where $M, T \in \mathbb{Z}^+$ and $M < T$, is* identifiable up to permutations*, if for any $p, \tilde{p} \in \mathcal{M}^T(\Pi^M_{\mathcal{A}}, \mathcal{P}^M_{\mathcal{B}})$ where*

$$p(\boldsymbol{z}_{1:T}) = \textstyle\sum_{i=1}^{C} c_i p_{a^i}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b^i_t}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}), \quad C = K_0 K^{T-M};\ \textit{and}$$

$$\tilde{p}(\boldsymbol{z}_{1:T}) = \sum_{i=1}^{\tilde{C}} \tilde{c}_i p_{\tilde{a}^i}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{\tilde{b}_t^i}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}), \quad \tilde{C} = \tilde{K}_0 \tilde{K}^{T-M},$$

with $p(\boldsymbol{z}_{1:T}) = \tilde{p}(\boldsymbol{z}_{1:T})\ \forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{mT}$*, we have* $C = \tilde{C},\ K_0 = \tilde{K}_0,\ K = \tilde{K}$*, and for each* $1 \le i \le C$ *there is some* $1 \le j \le \tilde{C}$ *such that:*

1. $c_i = \tilde{c}_j$;
2. *if* $b_{t_1}^i = b_{t_2}^i$ *for* $M < t_1, t_2 \le T$, $t_1 \ne t_2$, *then* $\tilde{b}_{t_1}^j = \tilde{b}_{t_2}^j$;
3. $p_{a^i}(\boldsymbol{z}_{1:M}) = p_{\tilde{a}^j}(\boldsymbol{z}_{1:M}),\ \forall \boldsymbol{z}_{1:M} \in \mathbb{R}^{mM}$;
4. $p_{b_t^i}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) = p_{\tilde{b}_t^j}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}),\ \forall (\boldsymbol{z}_t, \ldots, \boldsymbol{z}_{t-M}) \in \mathbb{R}^{m(M+1)}$.

The second requirement ensures that, for $t > M$, the switches are identified only up to a single global permutation of the labels, rather than allowing different permutations at different time indices, which is valid in the i.i.d. finite mixture case.

For the purpose of building deep generative models, we seek to define identifiable parametric families and defer the study of the nonparametric case in Appendix A. In particular we use a nonlinear Gaussian transition family as follows:

$$\mathcal{G}_{\mathcal{B}}^{M} = \Big\{ p_b(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) = \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b), \\ \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) \mid b \in \mathcal{B} \Big\}, \quad (14)$$

where $\boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) : \mathbb{R}^{mM} \to \mathbb{R}^m$ and $\boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m \times m}$ are nonlinear w.r.t. $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$ and denote the mean and covariance matrix, respectively. For each p_b to define a non-degenerate Gaussian density, we assume that $\boldsymbol{\Sigma}(\cdot, b)$ is positive definite for every $b \in \mathcal{B}$. We further require the following *unique indexing* assumption:

$$\forall b \ne b' \in \mathcal{B},\ \exists \mathcal{Z} \subset \mathbb{R}^{mM}, \mu(\mathcal{Z}) > 0,\ s.t.\ \forall (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) \in \mathcal{Z},\ \boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) \\ \ne \boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b') \text{ or } \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) \ne \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b'), \quad (15)$$

where μ denotes the Lebesgue measure on $\mathbb{R}^{mM}$. In other words, for such $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$, $p_b(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$ and $p_{b'}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$ are different Gaussian distributions on a set of nonzero measure. We also introduce a family of *initial distributions* with unique indexing:

$$\mathcal{I}_{\mathcal{A}}^{M} := \Big\{ p_a(\boldsymbol{z}_{1:M}) = \mathcal{N}(\boldsymbol{z}_{1:M} \mid \boldsymbol{\mu}_1(a), \boldsymbol{\Sigma}_1(a)) \ \mid \ a \in \mathcal{A} \Big\}, \quad (16)$$

$$\forall a \ne a' \in \mathcal{A} \Leftrightarrow \boldsymbol{\mu}_1(a) \ne \boldsymbol{\mu}_1(a') \text{ or } \boldsymbol{\Sigma}_1(a) \ne \boldsymbol{\Sigma}_1(a'), \quad (17)$$

where, similarly, $\boldsymbol{\Sigma}_1(a)$ is positive definite for every $a \in \mathcal{A}$. The above Gaussian families paired with unique indexing assumptions satisfy conditions that favour identifiability of MSMs.

Theorem 8 (Identifiability of parametric MSMs) *Consider the MSM family with order* M *defined in Equation (12) under nonlinear Gaussian families* $\mathcal{M}_{NL}^{T,M} := \mathcal{M}^T(\mathcal{I}_{\mathcal{A}}^{M}, \mathcal{G}_{\mathcal{B}}^{M})$ *with* $\mathcal{G}_{\mathcal{B}}^{M}$, $\mathcal{I}_{\mathcal{A}}^{M}$ *defined by Eqs. (14), (16) respectively. Then, the MSM is identifiable in terms of Def. 7 under the following assumptions:*

(m1) Unique indexing for $\mathcal{G}_{\mathcal{B}}^{M}$ and $\mathcal{I}_{\mathcal{A}}^{M}$: Eqs. (15) and (17) hold;

(m2) The mean and covariance in $\mathcal{G}_{\mathcal{B}}^{M}$, $\boldsymbol{m}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m}$ and $\boldsymbol{\Sigma}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m \times m}$, are analytic functions, for any $b \in \mathcal{B}$.

Proof sketch: See Appendix A for the full proof. The strategy can be summarised in the following steps. We defer additional intuitions and details to the end of this section, after a series of remarks.

1. Under the finite mixture model view, it suffices to show that the trajectory family, $\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M}$, is linearly independent under finite mixtures (Yakowitz and Spragins, 1968). Given that $\mathcal{M}^{T}(\Pi_{\mathcal{A}}^{M}, \mathcal{P}_{\mathcal{B}}^{M})$ is a subset of $\operatorname{conv}(\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M})$, identifiability of $\operatorname{conv}(\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M})$ implies the identifiability of $\mathcal{M}^{T}(\Pi_{\mathcal{A}}^{M}, \mathcal{P}_{\mathcal{B}}^{M})$.

2. To show linear independence under finite mixtures of $\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M}$, we begin with $T = 2M$ and construct a family of joint densities defined by nonparametric components $\mathcal{P}_{\mathcal{B}}^{M}$ and $\Pi_{\mathcal{A}}^{M}$:

$$\left\{ p_a(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{2M} p_{b_t}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}) \mid a \in \mathcal{A},\ b_t \in \mathcal{B},\ M < t \leq 2M \right\}.$$

   We show in Lemma 22 that this family is linearly independent under finite mixtures.

3. For $T > 2M$, we proceed by induction (Theorem 23). Under the induction hypothesis, we reuse the result from Lemma 22 to extend linear independence to longer trajectories.

4. We show that the Gaussian families $\mathcal{I}_{\mathcal{A}}^{M}$ and $\mathcal{G}_{\mathcal{B}}^{M}$ under Assumptions (m1) and (m2) satisfy the *nonparametric* case defined for $\Pi_{\mathcal{A}}^{M}$ and $\mathcal{P}_{\mathcal{B}}^{M}$ respectively.

Intuitively, "switches" can be considered as discontinuities in the dynamic model and are fully controlled by the discrete states. Assumption (m2) means $\boldsymbol{m}(\cdot, b)$ and $\boldsymbol{m}(\cdot, b')$ with $b \neq b'$ can only intersect at a set of points with zero Lebesgue measure (similarly for $\boldsymbol{\Sigma}(\cdot, b)$), and this smoothness property contributes to the identifiability for the continuous-state transitions $p_b(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}),\ b \in \mathcal{B}$. The following implications of Theorem 8, when combined, show the profound expressivity of identifiable MSMs:

- Assumption (m2) enables parametrising $p_b(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \dots \boldsymbol{z}_{t-M})$ using polynomials or neural networks (NNs) with analytic activations (e.g., Softplus). As mentioned, for NNs identifiability applies to the functional form only, since NN weights do not uniquely index NN functions.
- The identifiability result holds independently of the choice of $p(\boldsymbol{s}_{M:T})$, allowing e.g., nonstationarity and higher-order Markov dependencies.

Our experiments use a stationary first-order Markov prior for $p(\boldsymbol{s}_{M:T})$. Its transition parameters are identifiable up to permutations following Balsells-Rodas et al. (2024, Cor. C.1). We leave other cases to future work.

Remark 9 *Our strategy establishes identifiability by proving linear independence of the trajectory family $\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M}$ under finite mixtures (Theorem 23), which, together with Proposition 4, ensures identifiability in terms of Def. 2 for any switching model whose family lies in* $\operatorname{conv}(\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M})$. *Figure 2a illustrates the resulting hierarchy:*

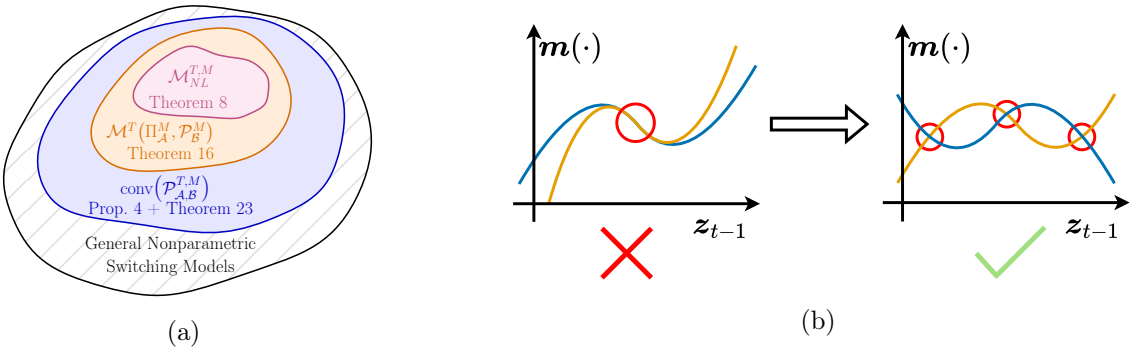


Figure 2: (a) Hierarchy of model classes covered by our analysis. The magenta region shows our main result for parametric Gaussian MSMs under analyticity and unique indexing (Sec. 4.1). The orange region shows the nonparametric result under abstract conditions on the transition families (App. A). The blue region shows the full scope of our proof technique, covering switching models in $\text{conv}(\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M})$ with linearly independent trajectory families. Moving outward increases generality at the cost of more abstract assumptions. (b) Intuition behind Lemma 17, where linear independence holds if, for any pair of functions (shown in blue and orange), the intersection in the conditioned-variable ($\boldsymbol{z}_{t-1}$) domain has measure zero. Adapted from Balsells-Rodas et al. (2024).

- ***Magenta region:*** *Main result (Thm. 8), covering parametric MSMs with Gaussian transitions.*
- ***Orange region:*** *Generalisation to nonparametric MSMs (Thm. 16), which also includes non-Gaussian transition families.*
- ***Blue region:*** *Full theoretical scope of our approach, which includes any switching model whose family lies within* $\text{conv}(\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M})$*, such as MSMs with time-varying components* $K(t)$*.*

4.1.3 Nonparametric Identifiability

The key to our proof technique is the establishment of step (2) in *Proof sketch*, which is proved through a new theoretical result: *Linear Independence on Product functions with M consecutive Overlapping variables* (LIPO-M; Lemma 22), and it can be of independent interest. Unlike Hidden Markov Models (HMMs), in MSMs, $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+1}$ are *not* conditionally independent given the discrete states (assuming $M = 1$ in this example). Therefore, seminal HMM identifiability results (Allman et al., 2009; Gassiat et al., 2016), which exploit conditional independence of $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+1}$ in HMMs and leverage Kruskal's "3-way arrays" technique (Kruskal, 1977), do not apply to nonparametric MSMs, rendering our development of new theory necessary.

Linear independence (first-order case). To understand how linear independence under finite mixtures on products with M consecutive variables is achieved, we start with $M = 1$ (LIPO-1), presented in Balsells-Rodas et al. (2024), which we provide in Appendix A.2.1 (Lemma 17) for completeness. Consider the first-order trajectory family on step (2), i.e.,

$T = 2$:

$$\Pi^1_{\mathcal{A}} \otimes \mathcal{P}^1_{\mathcal{B}} = \Big\{ p_a(\boldsymbol{z}_1) p_{b_2}(\boldsymbol{z}_2 \mid \boldsymbol{z}_1) \mid (a, b_2) \in \mathcal{A} \times \mathcal{B} \Big\}, \tag{18}$$

where the overlapping variable $\boldsymbol{z}_1$ challenges linear independence under finite mixtures. While LIPO-1 is established for abstract function families beyond probability distributions, we re-state an informal version of LIPO-1 for $\Pi^1_{\mathcal{A}} \otimes \mathcal{P}^1_{\mathcal{B}}$ for clarity below.

Lemma 10 (LIPO-1 (informal)) *Consider the first-order trajectory family $\Pi^1_{\mathcal{A}} \otimes \mathcal{P}^1_{\mathcal{B}}$ under the following assumptions:*

(a3) $\Pi^1_{\mathcal{A}}$ *is linearly independent under finite mixtures.*

(a4) There exists a nonzero measure set $\mathcal{Z} \subseteq \mathbb{R}^m$ such that the family $\mathcal{P}^1_{\mathcal{B}}$ remains linearly independent under finite mixtures for any $(\boldsymbol{z}_1, \boldsymbol{z}_2) \in \mathcal{Z} \times \mathbb{R}^m$.

(a5) When fixing $\boldsymbol{z}_1$, linear dependence in $\mathcal{P}^1_{\mathcal{B}}$ occurs only if the components contain repeated conditional distributions.

(a6) The functions in $\mathcal{P}^1_{\mathcal{B}}$ are continuous in $\boldsymbol{z}_1$.

Then the product family $\Pi^1_{\mathcal{A}} \otimes \mathcal{P}^1_{\mathcal{B}}$ is linearly independent under finite mixtures.

Figure 2b illustrates the intuition behind LIPO-1 (and LIPO-M). Linear independence of the product family is preserved provided that the conditional distributions coincide only on sets of measure zero in the overlapping variable domain. For $M = 1$ it suffices that linear independence holds on a nonzero-measure set of the overlap-variable space $\boldsymbol{z}_1$ (Assumption (a4)).

Proof sketch: See LIPO-1 (Lemma 17) for a general definition and the full proof. The key is to understand $\mathcal{P}^1_{\mathcal{B}}$ as a family of functions of two variables $(\boldsymbol{z}_1, \boldsymbol{z}_2)$, rather than conditional distributions. Under (a3), the overlapping variable $\boldsymbol{z}_1$ challenges linear independence under finite mixtures of $\mathcal{P}^1_{\mathcal{B}}$ when multiplied with $\Pi^1_{\mathcal{A}}$. Fixing the overlapping variable $\boldsymbol{z}_1 = \boldsymbol{\beta}$ reduces the problem to linear independence of the conditional family. By choosing $\mathcal{Z}$ from (a4) the linear independence equation can be written for any $\boldsymbol{\beta} \in \mathcal{Z}$ modulo zero measure sets:

$$\sum_{b \in \mathcal{B}_0} \Bigg(\underbrace{\sum_{a:(a,b)\in S_0} \gamma_{ab} p_a(\boldsymbol{z}_1 = \boldsymbol{\beta})}_{\neq 0 \quad \text{by (a2)}} \Bigg) p_b(\boldsymbol{z}_2 \mid \boldsymbol{z}_1 = \boldsymbol{\beta}) = 0, \quad \forall \boldsymbol{z}_2 \in \mathbb{R}^m,$$

where S_0 and $\mathcal{B}_0$, denote finite sets with indices for which linear dependence may occur, and γ_{ab} are the coefficients of the linear independence equation. This shows that any potential linear dependence must occur through coefficients that vary with the overlapping variable $\boldsymbol{z}_1$. Under (a5), any linear dependence of the conditional family arises through repeated functions. Since the coefficients depend on the overlapping variable $\boldsymbol{z}_1$, maintaining linear dependence would require the functions indexed by $\mathcal{B}_0$ to be partitioned into groups of repeated functions, where the partition changes with $\boldsymbol{z}_1$. Intuitively, changing partitions introduces abrupt changes that are incompatible with continuity (a6). Importantly, linear independence is preserved as long as it holds on a nonzero-measure subset of the overlap-variable domain.

Linear independence (general case). Our strategy to generalise to any lag $M < T$ focuses on extending the previous result to cover multiple overlapping variables. The challenge is that when the autoregressive order M increases, the number of overlapping variables between consecutive distributions also increases. For $M = 1$, each new factor only shares a single variable with the previous one, so overlaps remain "local" and can be controlled directly. For $M > 1$, however, a single variable may overlap with several different factors at once. This complicates the proof of linear independence under finite mixtures, as we no longer have a single overlap as in Eq. (18). To illustrate, for $M = 3$, the trajectory family at $T = 6$ has the following structure:

$$\Pi^3_{\mathcal{A}} \otimes (\otimes^3 \mathcal{P}^3_{\mathcal{B}}) = \Big\{ p_a(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) p_{b_4}(\boldsymbol{z}_4 \mid \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) p_{b_5}(\boldsymbol{z}_5 \mid \boldsymbol{z}_4, \boldsymbol{z}_3, \boldsymbol{z}_2) p_{b_6}(\boldsymbol{z}_6 \mid \boldsymbol{z}_5, \boldsymbol{z}_4, \boldsymbol{z}_3) \Big\}, \tag{19}$$

for every $a \in \mathcal{A}$ and $b_4, b_5, b_6 \in \mathcal{B}$, where we indicate overlapping variables with different colours. Compared to $M = 1$ in Eq. (18), when multiplying the initial and transition distribution, $p_a(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) p_{b_4}(\boldsymbol{z}_4 \mid \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$, the overlap $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$ has the same structure as in the $M = 1$ case, and linear independence for $T = 4$ can be established by LIPO-1. However at $T = 5$, the next factor $p_{b_5}(\boldsymbol{z}_5 \mid \boldsymbol{z}_4, \boldsymbol{z}_3, \boldsymbol{z}_2)$ introduces an overlap $(\boldsymbol{z}_4, \boldsymbol{z}_3, \boldsymbol{z}_2)$ that is no longer confined in a single pair of consecutive terms, as $(\boldsymbol{z}_3, \boldsymbol{z}_2)$ are present in $p_a(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$. This branching overlap violates assumption (a3) of LIPO-1, as the product family does not directly reduce to the single-overlap structure in Eq. (18). Similar arguments apply for $T = 6$, where $\boldsymbol{z}_3$ overlaps with all the functions.

To ensure linear independence under finite mixtures of the trajectory family for any lag M, we group variables to force overlaps only between consecutive distributions (similar to Eq. (18)). For instance, the example with $M = 3$ becomes:

$$\Pi^3_{\mathcal{A}} \otimes (\otimes^3 \mathcal{P}^3_{\mathcal{B}}) = \Big\{ p_a(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) p_{b_{4:6}}(\boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6 \mid \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) \mid a \in \mathcal{A}, b_{4:6} \in (\times^3 \mathcal{B}) \Big\}. \tag{20}$$

As in the case $M = 1$, linear independence under finite mixtures holds provided that the product $(\otimes^3 \mathcal{P}^3_{\mathcal{B}})$ satisfies assumptions (a4)–(a6) in LIPO-1. Therefore, the strategy is to impose these assumptions on $\mathcal{P}^M_{\mathcal{B}}$, and explore whether they extend to the product families $(\otimes^r \mathcal{P}^M_{\mathcal{B}})$, for all $1 < r \leq M$. In Appendix A.2.2 we prove that such an extension requires strengthening the nonparametric assumptions from (a4)–(a6) to (b4)–(b6) (Lemmas 20 and 21). While (b6) directly extends continuity with respect to the overlapping variable domain, (b4) is stronger, as to extend to product families, it requires the nonzero measure set $\mathcal{Z} \subseteq \mathbb{R}^{mM}$ in (a4) to have full measure. Furthermore, (b5) extends to product families $\otimes^M \mathcal{P}^M_{\mathcal{B}}$ with a weaker implication compared to (a5): linear dependence of a product family only guarantees that at least one factor in the product is linearly dependent. This prevents a direct application of the LIPO-1 argument and requires a recursive proof strategy in LIPO-M (Lemma 22). The proof proceeds through repeated applications of Lemma 21, which identifies factors responsible for linear dependence, and (b5) and (b6) are used to discard such factors one by one from 1 to M.

Linear independence via nested induction. To prove LIPO-M and extend the assumptions to product families (Lemmas 20 and 21), we use induction, with the base case being a single product of consecutive variables $(\mathcal{P}^M_{\mathcal{B}} \otimes \mathcal{P}^M_{\mathcal{B}})$. Therefore, the full proof of linear independence under finite mixtures of the trajectory family requires nested induction (see Theorem 23 in Appendix A.2.4). Figure 3 illustrates this idea:

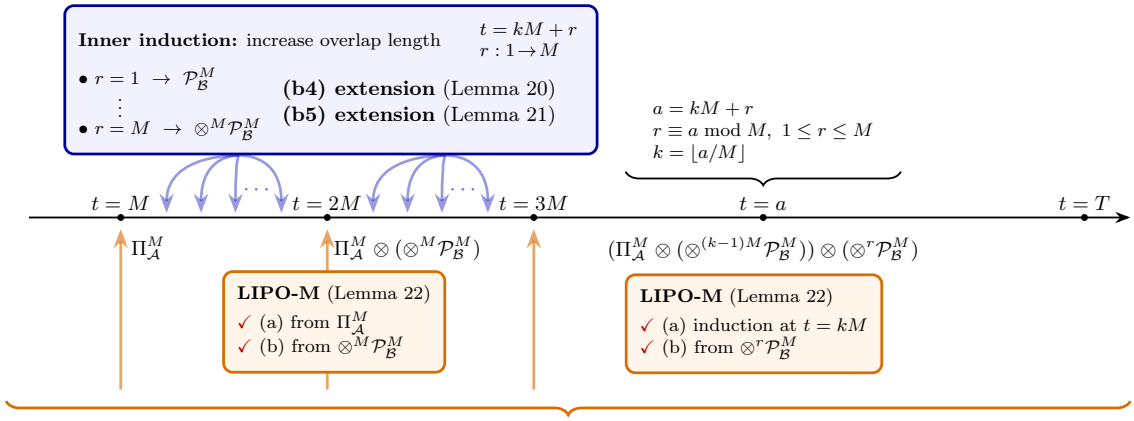


Figure 3: Linear independence under finite mixtures of $\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M}$ is proven by a nested induction strategy. The inner induction (blue box) constructs the product family $(\otimes^{r}\mathcal{P}_{\mathcal{B}}^{M},\ 1 \le r \le M)$, while the outer induction (orange boxes) increases the trajectory length. At $t = a$, LIPO-M shows linear independence up to the largest multiple of M, and linear independence of the additional r remaining steps is shown using LIPO-M on $(\otimes^{r}\mathcal{P}_{\mathcal{B}}^{M})$, where $r \equiv a \bmod M$.

- The inner induction constructs the product family $(\otimes^{r}\mathcal{P}_{\mathcal{B}}^{M})$ with $1 \le r \le M$ by increasing the number of consecutive transition distributions r.
- The outer induction increases the length of the trajectory family $\mathcal{P}_{\mathcal{A},\mathcal{B}}^{M,t}$, with $M < t \le T$, ensuring that overlaps are addressed as we introduce additional time steps.

A key technical point in Theorem 23 is that the induction hypothesis applies only when the trajectory length t is a multiple of M. Thus we distinguish two cases:

- If t is a multiple of M, we apply the induction hypothesis at $t - M$ and then use LIPO-M to combine the trajectory family at time $t - M$ with M consecutive products from $\mathcal{P}_{\mathcal{B}}^{M}$.
- If t is not a multiple of M, compute $r \equiv t \bmod M$. We apply the induction hypothesis at $t - r$ (a multiple of M), and then use LIPO-M to combine the trajectory family at time $t - r$ with the r additional products $(\otimes^{r}\mathcal{P}_{\mathcal{B}}^{M})$.

The same logic applies to the base cases $M < t \le 2M$, where LIPO-M is applied by combining the initial family $\Pi_{\mathcal{A}}^{M}$ with $t - M$ consecutive products from $\mathcal{P}_{\mathcal{B}}^{M}$.

Connection to Gaussian families. The last stage of the proof of Theorem 8 consists of connecting the nonparametric result to Assumptions (m1) and (m2). As mentioned, Yakowitz and Spragins (1968) show that multivariate Gaussian families are linearly independent under finite mixtures if and only if they satisfy unique indexing. This translates to the Gaussian transition family $\mathcal{G}_{\mathcal{B}}^{M}$ (Proposition 24), yielding (b1) and (b5). To establish (b4), we introduce zero-measure intersection of moments (d3) in Theorem 28, which guarantees that distinct components coincide only on zero-measure subsets of the overlapping variable domain. Finally, Assumption (m2) ensures continuity (b6) and that distinct functions differ almost everywhere,

which, paired with unique indexing (m1), implies (d3). This provides the connection between Gaussian families and the nonparametric analysis.

4.2 Identifiable Switching Dynamical Systems

Based on identifiable MSMs, our analysis of switching dynamical systems extends the setup from Kivva et al. (2022) (Section 3.2) to the temporal case. Assume the prior dynamic model $p(\boldsymbol{z}_{1:T})$ belongs to the nonlinear Gaussian MSM family $\mathcal{M}_{NL}^{T,M}$ specified by Thm. 8. At each time step t, the latent $\boldsymbol{z}_t \in \mathbb{R}^m$ generates observed data $\boldsymbol{x}_t \in \mathbb{R}^n$ via a piecewise linear transformation f. The generation process of such an SDS model can be expressed as follows:

$$\boldsymbol{x}_t = f(\boldsymbol{z}_t) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim p_{\boldsymbol{\varepsilon}}, \quad \boldsymbol{z}_{1:T} \sim p(\boldsymbol{z}_{1:T}) \in \mathcal{M}_{NL}^{T,M}. \tag{21}$$

Note that we can also write the decoding process as $\boldsymbol{x}_{1:T} = F(\boldsymbol{z}_{1:T}) + \boldsymbol{\varepsilon}_{1:T}$, where F is a factored transformation composed of f as defined below. This assumes that each observation $\boldsymbol{x}_t$ depends only on the instantaneous latent state $\boldsymbol{z}_t$ through the same emission function f, excluding, for instance, regime-dependent transformations f_{s_t}. Importantly, this notion allows F to inherit e.g., piecewise linearity and weak injectivity from f.

Definition 11 *We say that a function $F : \mathbb{R}^{mT} \to \mathbb{R}^{nT}$ is factored if it is composed of $f : \mathbb{R}^m \to \mathbb{R}^n$, such that for any $\boldsymbol{z}_{1:T} \in \mathbb{R}^{mT}$, $F(\boldsymbol{z}_{1:T}) = (f(\boldsymbol{z}_1), \ldots, f(\boldsymbol{z}_T))^\top$.*

The identifiability analysis of the SDS model (Eq. (21)) follows two steps (see Figure 4). (i) Following Kivva et al. (2022), we establish the identifiability for a prior $p(\boldsymbol{z}_{1:T}) \in \mathcal{M}_{NL}^{T,M}$ from $(F_{\#}p)(\boldsymbol{x}_{1:T})$ in the noiseless case. Here $F_{\#}p$ denotes the pushforward measure of p by F. (ii) When the noise $\boldsymbol{\varepsilon}_{1:T}$ distribution is known, $F_{\#}p$ is identifiable from $\left((F_{\#}p) * p_{\boldsymbol{\varepsilon}_{1:T}}\right)(\boldsymbol{x}_{1:T})$ using convolution tricks from Khemakhem et al. (2020).[3] In detail, we first define the notion of identifiability for $p \in \mathcal{M}^T(\Pi_{\mathcal{A}}^M, \mathcal{P}_{\mathcal{B}}^M)$ given noiseless observations.

Definition 12 *Given a family of factored transformations $\mathbb{F}$, for $F \in \mathbb{F}$ the prior $p \in \mathcal{M}^T(\Pi_{\mathcal{A}}^M, \mathcal{P}_{\mathcal{B}}^M)$ is said to be identifiable up to affine transformations, when for any $\tilde{F} \in \mathbb{F}$ and $\tilde{p} \in \mathcal{M}^T(\Pi_{\mathcal{A}}^M, \mathcal{P}_{\mathcal{B}}^M)$ s.t. $F_{\#}p = \tilde{F}_{\#}\tilde{p}$, there exists an invertible factored affine mapping $H : \mathbb{R}^{mT} \to \mathbb{R}^{mT}$ composed of $h : \mathbb{R}^m \to \mathbb{R}^m$, where $p = H_{\#}\tilde{p}$. Equivalently, identifiability up to permutation, scaling and translation refers to having $h(\boldsymbol{z}) = PD\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{z} \in \mathbb{R}^m$, where P and D are permutation and diagonal matrices respectively.*

We state the following identifiability result for nonlinear SDSs and defer the proof in Appendix B, which is mainly based on Kivva et al. (2022).

Theorem 13 (Identifiability of parametric SDSs) *Assume the observations $\boldsymbol{x}_{1:T}$ are generated from a latent variable model following Eq. (21), whose prior $p(\boldsymbol{z}_{1:T})$ belongs to $\mathcal{M}_{NL}^{T,M} = \mathcal{M}^T(\mathcal{I}_{\mathcal{A}}^M, \mathcal{G}_{\mathcal{B}}^M)$ and satisfies (m1 - m2). Assume:*

(f1) f is weakly injective (Def. 5) and piecewise linear;

and the following assumptions on the prior distribution $p(\boldsymbol{z}_{1:T})$.

3. In particular, identifiability in the noisy case holds under the assumption that the zero set of the characteristic function of $p_{\boldsymbol{\varepsilon}}$ has measure zero. We use $p_{\boldsymbol{\varepsilon}_{1:T}}$ to denote the distribution of $\boldsymbol{\varepsilon}_{1:T}$.

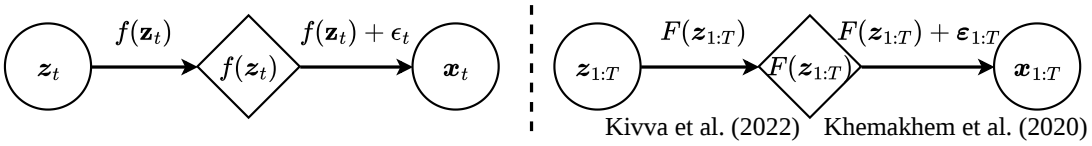


Figure 4: We assume $\boldsymbol{z}_t$ is transformed via f with noise $\boldsymbol{\varepsilon}_t$ at each time step t independently. We view this as a transformation on $\boldsymbol{z}_{1:T}$ via a factored F with noise $\boldsymbol{\varepsilon}_{1:T}$. Adapted from Balsells-Rodas et al. (2024).

(s1) At time $t > M$, the continuous latent variables $\boldsymbol{z}_t$ are mutually independent given $\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}$ and s_t. For any $i \neq j$, $z_t^{(i)} \perp\!\!\!\perp z_t^{(j)} \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t$;

(s2) There exists $k_1, k_2 \in [K]$ and a nonzero measure set $\mathcal{Z} \subset \mathbb{R}^{mM}$ such that the elements in $\{\boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, k_1)_{ii} / \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, k_2)_{ii} \mid i \in [m]\}$ are distinct for any $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) \in \mathcal{Z}$.

The prior distribution $p(\boldsymbol{z}_{1:T})$ is identifiable under the above assumptions as follows.

(i) If (f1) is satisfied, the prior $p(\boldsymbol{z}_{1:T})$ is identifiable up to affine transformations (Def. 12).

(ii) If (f1), (s1), and (s2) are satisfied, the prior $p(\boldsymbol{z}_{1:T})$ is identifiable up to permutation, scaling, and translation.

Assumption (s2) is a form of sufficient variability, related to conditions used for disentangling latent temporal processes (Yao et al., 2022b) and in interventional Gaussian models (Rajendran et al., 2024). However, its role in the proof differs from the above works: rather than establishing identifiability through full-rank systems, it resolves the initial affine ambiguity (Theorem 13) via the eigendecomposition of regime-dependent covariance ratios. The above result allows identifiable SDSs to employ e.g., Softplus networks for $p_{\boldsymbol{\theta}}(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t)$ with history-dependent or piecewise stationary noise, and certain (Leaky) ReLU networks for f (see Appendix H of Kivva et al., 2022). Again for neural networks, identifiability refers only to their functional form.

Beyond piecewise linear emissions. While our approach relaxes the injectivity assumption on the emission function, we still restrict f to be piecewise linear. Other works, such as Yao et al. (2022a), assume injective functions, where their results could be adapted to our framework. Specifically, Yao et al. (2022a) establish identifiability under the following conditions: (1) independence of latent variables (s1 in Theorem 13); (2) invertibility of the mixing function; and (3) access to $K = 2m + 1$ components satisfying certain linear independence conditions on vectors formed with derivatives of $\log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, k),\ k \in [K]$. However, this strategy either assumes access to the true component label k, or requires imposing assumptions directly on the observed conditional distributions $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-M}, k),\ k \in [K]$. For instance, Song et al. (2023) introduce Assumptions (m1) and (m2) directly on the observed distributions, which implicitly constrain both the latent dynamics and the emission function. This lack of separation between latent structure and observation model makes these assumptions difficult to verify or control in practice (e.g., if observed and latent distributions are Gaussian, then f must be affine). Moreover, without additional structural constraints, these strategies cannot fully exploit our identifiability results for MSMs, which are derived at the level of the latent dynamics. To illustrate, consider the corresponding SDS family

obtained by applying a factored transformation F to every trajectory component of the MSM family in Eq. (12):

$$\left\{ \sum_{i}^{C} c_i (F_{\#} p_i)(\boldsymbol{x}_{1:T}) \mid \; F \in \mathbb{F},\; \sum_{i}^{C} c_i = 1,\; c_i > 0,\; C = K_0 K^{T-M},\; K_0, K < +\infty, \right.$$
$$\left. p_i \in \mathcal{P}^{T,M}_{\mathcal{A}_0,\mathcal{B}_0} \subset \mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}},\; \mathcal{A}_0 \subset \mathcal{A},\; \mathcal{B}_0 \subset \mathcal{B},\; |\mathcal{A}_0| = K_0,\; |\mathcal{B}_0| = K \right\}. \quad (22)$$

Equation (22) therefore formalises SDSs in the finite-mixture sense. Identifiability here requires further analysis, since two different elements in the above family with $F \neq \tilde{F}$ could correspond to different MSMs with different components, $(K_0, K) \neq (\tilde{K}_0, \tilde{K})$, related via inverse CDF mappings: $p = \left(F^{-1} \circ \tilde{F}\right)_{\#} \tilde{p}$.

4.3 Regime-Dependent Causal Representation Learning

In regime-dependent time series (Saggioro et al., 2020), the SEM is time-dependent. However, it is *causally stationary* (Assaad et al., 2022) when conditioned on discrete latent variables $s_t \in [K]$ at each time step t. In the context of LVMs, the SEM operates on the latent space; and therefore, our framework is connected to causal representation learning (Schölkopf et al., 2021). When considering SDSs, the corresponding latent causal structure becomes a *regime-dependent causal graph* which we denote as $\mathbf{G}_{1:K} := \{\mathbf{G}_k\}_{k=1}^{K} \in \mathbb{G}^K$. It can be defined as a collection of causal graphs such that for each $k \in [K]$, $\mathbf{G}_k \in \mathbb{G}$ encodes the latent SEM at time t if $s_t = k$. To establish connections with the previous MSMs, we assume an *additive noise model* (Hoyer et al., 2008) with absence of instantaneous effects. Therefore, the corresponding regime-dependent SEM in latent space for variable $j \in [m]$ is expressed as follows:

$$z_t^{(j)} = f_{s_t}^{(j)} \left(\{ z_{t-\tau}^{(i)} \mid z_{t-\tau}^{(i)} \in \mathbf{Pa}_{s_t}^{(j)}(\tau), 1 \leq \tau \leq M \} \right) + \boldsymbol{\varepsilon}_t^{(j)}, \quad \boldsymbol{\varepsilon}_t^{(j)} \sim p_{\boldsymbol{\varepsilon}^{(j)}}, \quad (23)$$

where $\mathbf{Pa}_{s_t}^{(j)}(\tau)$ denotes the parents of variable j with lag τ, associated to $\mathbf{G}_k$ for $s_t = k$.

Definition 14 *We say that the regime-dependent graph of a SDS with order M is identifiable up to permutations if for any $p, \tilde{p} \in \mathcal{M}^T(\Pi_{\mathcal{A}}^M, \mathcal{P}_{\mathcal{B}}^M)$, and $F, \tilde{F} \in \mathbb{F}$, with respective regime-dependent graphs $\mathbf{G}_{1:K} \in (\times^K \mathbb{G})$ and $\tilde{\mathbf{G}}_{1:\tilde{K}} \in (\times^{\tilde{K}} \mathbb{G})$, such that $(F_{\#} p)(\boldsymbol{x}_{1:T}) = (\tilde{F}_{\#} \tilde{p})(\boldsymbol{x}_{1:T})$; $K = \tilde{K}$ and there exists a node permutation $\pi \in S_m$ (Def. 36) and a graph permutation $\sigma \in S_K$ such that $\mathbf{G}(k) = \pi(\tilde{\mathbf{G}}(\sigma(k)))$ for $k \in [K]$.*

The above node permutation (Def. 36) permutes the nodes of each graph in the set $\mathbf{G}_{1:K}$. Our definition aims to recover the inherent graph structure up to permutations of both continuous and discrete latent variables. We establish latent structure identifiability for regime-dependent time series as follows.

Corollary 15 (Latent structure identifiability) *Consider an SDS whose latent dynamics are given by a nonlinear Gaussian MSM with lag M, $\mathcal{M}_{NL}^{T,M}$, where observations are generated following Eq.* (21)*. Assume (m1) from Theorem 8 and (f1) from Theorem 13 are satisfied, and further assume:*

(m2) Analytic transition with heterogeneous noise: the transition distribution mean in* $\boldsymbol{m}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^m$ *is analytic, and* $\boldsymbol{\Sigma}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m \times m}$ *is diagonal, analytic, and satisfies (s2);*

(m3) Minimality: if $z^{(i)} \notin \mathbf{Pa}_k^{(j)}(\tau), 1 \leq \tau \leq M$*, then the* j*-th dimension of* $\boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, k)$ *is constant with respect to* $z_{t-\tau}^{(i)}$*.*

Then, the regime-dependent graph is identifiable up to permutations (Definition 14).

See Appendix C for the proof. The modified Assumption (m2*) groups (m2) from Theorem 8, together with (s1) and (s2) from Theorem 13. Under these conditions, together with (m1) and (f1), the continuous latent variable distribution is identifiable up to permutation and scaling by Theorem 13. This transformation preserves the structure of the Jacobian of the transition means, up to relabelling of the latent variables. Under minimality (m3), the Jacobian sparsity pattern determines the parent sets of each graph. Therefore, given the regime-dependent MSM transitions are identifiable up to relabelling of the regimes by Theorem 8, the regime-dependent graph is identifiable up to node and regime permutations. We note that under the above conditions, other structures in traditional causal time series analysis, such as the *full time graph* (Peters et al., 2017), are not identifiable in the context of high-order MSMs. This is because we have no access to the discrete latents $\boldsymbol{s}_{M:T}$, but rather their distribution up to permutations.

5. Estimation

We refer to our full approach, which combines the theoretical framework with the inference scheme below, as *identifiable Switching Dynamical System* (iSDS). For real-world applications (Section 7), we also consider MSMs. Together with our assumptions, we denote the corresponding method as the *identifiable Markov Switching Model* (iMSM).

5.1 Variational Inference

We adopt the collapsed variational inference setup of Dong et al. (2020). Parameters are learned by maximising the evidence lower bound (ELBO) (Kingma and Welling, 2014). The approximate posterior over latent variables $\{\boldsymbol{z}_{1:T}, \boldsymbol{s}_{M:T}\}$ uses a factorisation which incorporates the exact posterior of the discrete latent states conditioned on the continuous latent variables:

$$q_{\boldsymbol{\phi},\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{M:T} \mid \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t \mid \boldsymbol{x}_t) p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T} \mid \boldsymbol{z}_{1:T}). \tag{24}$$

The variational posterior over $\boldsymbol{z}_{1:T}$ factorises across time. This mirrors our injectivity assumptions on the latent variable transformation, and simplifies the estimation. Prior work (Ansari et al., 2021; Dong et al., 2020) constructs $q_{\boldsymbol{\phi},\boldsymbol{\theta}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})$ via a bidirectional RNN on $\boldsymbol{x}_{1:T}$, followed by a forward RNN. With the exact posterior $p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T} \mid \boldsymbol{z}_{1:T})$, we can marginalise $\boldsymbol{s}_{M:T}$ from the ELBO (see Appendix D for details):

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) \geq \mathcal{L}_{\text{ELBO}} := \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})} \Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T} \mid \boldsymbol{z}_{1:T}) \Big] - KL\Big(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) \Big). \tag{25}$$

We estimate expectations with Monte Carlo samples $\tilde{\boldsymbol{z}}_{1:T} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})$ and apply the reparametrisation trick for backpropagation (Kingma and Welling, 2014), enabling stochastic gradient ascent. The gradient is then given by:

$$\nabla \mathcal{L}_{\text{ELBO}} \approx \nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T} \mid \tilde{\boldsymbol{z}}_{1:T}) - \nabla \log q_{\boldsymbol{\phi}}(\tilde{\boldsymbol{z}}_{1:T} \mid \boldsymbol{x}_{1:T}) + \nabla \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{1:T}). \tag{26}$$

Following Dong et al. (2020), the prior gradient admits an efficient decomposition, similar to standard HMM inference (Bishop, 2006):

$$\nabla \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{1:T}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T} \mid \tilde{\boldsymbol{z}}_{1:T})} \Big[\nabla \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{1:T}, \boldsymbol{s}_{M:T}) \Big] \tag{27}$$

$$= \sum_{t=M+1}^{T} \sum_{k=1}^{K} \gamma_{t,k}(\tilde{\boldsymbol{z}}_{1:T}) \nabla \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_t \mid \tilde{\boldsymbol{z}}_{t-1}, \ldots, \tilde{\boldsymbol{z}}_{t-M}, s_t = k) \tag{28}$$

$$+ \sum_{k=1}^{K} \gamma_{M,k}(\tilde{\boldsymbol{z}}_{1:T}) \nabla \log \pi_k + \sum_{t=M+2}^{T} \sum_{k=1}^{K} \sum_{k'=1}^{K} \xi_{t,k,k'}(\tilde{\boldsymbol{z}}_{1:T}) \nabla \log Q_{k',k}, \tag{29}$$

where $\gamma_{t,k}(\tilde{\boldsymbol{z}}_{1:T}) = p_{\boldsymbol{\theta}}(s_t = k \mid \tilde{\boldsymbol{z}}_{1:T})$ and $\xi_{t,k,k'}(\tilde{\boldsymbol{z}}_{1:T}) = p_{\boldsymbol{\theta}}(s_t = k, s_{t-1} = k' \mid \tilde{\boldsymbol{z}}_{1:T})$ are computed with forward-backward update rules under a first-order Markov chain over $\boldsymbol{s}_{M:T}$. See Appendix D for details. A recent alternative (Ansari et al., 2021) computes $p(\boldsymbol{z}_{1:T})$ using a forward posterior only, but this requires backpropagation through the entire sequence, considerably increasing computational cost.

While maximising the ELBO provides a practical estimation approach, it does not guarantee statistical consistency in general. Khemakhem et al. (2020) show that the ELBO recovers the MLE when the variational family can represent the exact posterior and the objective is globally optimised. However, many variational estimators, including iSDS, assume Gaussian posteriors. This offers a stable optimisation route, but may prevent estimator consistency when the true posterior lies outside of the variational family. In particular, the true posterior in SDSs is obtained after marginalising over the K^{T-M+1} possible regime paths $\boldsymbol{s}_{M:T}$ (Ghahramani and Hinton, 2000), which implies that it may be highly non-Gaussian or multimodal. Tighter variational objectives based on importance sampling or sequential Monte Carlo, such as IWAE (Burda et al., 2016) or FIVO (Maddison et al., 2017), provide alternatives that can account for the multimodality of the continuous latent posterior. However, tighter bounds do not guarantee statistical consistency and may introduce additional optimisation challenges, such as increased gradient variance. Therefore, our choice reflects a compromise between posterior expressiveness and optimisation stability. A formal consistency analysis is left for future work.

5.2 Initialisation

Our work uses the same setup as in Dong et al. (2020), where the approximate posterior combines the exact posterior of the latent switches given the latent states $p_{\boldsymbol{\theta}}(\boldsymbol{s}_{1:T} \mid \boldsymbol{z}_{1:T})$. They observe that the proposed approach is prone to state collapse and introduce additional losses with annealing schedules. Ansari et al. (2021) instead introduce temperature annealing on discrete logits together with a forward posterior. Such schedules are data-dependent and require careful hyperparameter tuning. In our experience, a strong initialisation of $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$ is key to stabilising training and mitigating state collapse. This is particularly important

in the nonlinear setting, where randomly initialised emission and transition functions may produce uninformative latent representations, causing the inferred discrete states to collapse and prevent learning regime-dependent dynamics.

We first project $\boldsymbol{x}_{1:T}$ into a lower-dimensional space using Principal Component Analysis (PCA) and fit an MSM to the projected time series following Eq. (29). The PCA projection provides an approximation of the latent representation, from which we obtain preliminary estimates of the regime assignments and nonlinear transition functions. The learned π, Q, and regime-specific parameters are then used to initialise $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$. After initialisation, all model components are jointly learned by maximising the ELBO. Similar initialisation strategies have proven effective in prior work (Linderman et al., 2016) and serve as a foundation for improved training stability in discrete latent variable models. The iMSM follows the same likelihood-based optimisation, using the gradient decomposition in Eq. (29).

5.3 Sparsity Regularisation

To recover the underlying causal structure, we promote sparsity in the regime-specific mean transitions $\boldsymbol{m}(\cdot, k)$ via an element-wise ℓ_1 penalty on their Jacobians $J_{\boldsymbol{m}_k}$ for $k \in [K]$. The regularised objective is

$$\mathcal{L}_{\text{REG}} := \mathcal{L}_{\text{ELBO}} - \eta \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{mM} \left| \left(J_{\boldsymbol{m}_k}\right)_{ij} \right|, \tag{30}$$

where m denotes the latent dimension. We provide an empirical cost analysis of the Jacobian regularisation in Appendix D.1, showing that the regularised objective does not introduce substantial overhead in our training scheme. Related work has used Jacobian-based criteria to learn causal structure (Reizinger et al., 2023), while Brady et al. (2025) introduce interaction regularisation to promote compositional structure. In our setting, we regularise the Jacobians of the regime-dependent transition functions to learn sparse regime-specific causal graphs.

6. Synthetic Experiments

In this section, we present experiments on synthetic data, where access to the ground-truth structure enables us to empirically verify the theoretical results. Specifically, we (i) compare the ability of iSDS to recover identifiable latent variables against recent baselines; (ii) investigate the sample and sequence length efficiency; and (iii) test whether the assumptions in Theorem 8 are necessary for identifiability in both MSMs and SDSs. The experiments in this section focus primarily on iSDS. Additional iMSM results with non-diagonal covariance matrices are provided in Appendix E.4.

Data Generation. We generate data mostly following Balsells-Rodas et al. (2024), with additional scenarios for the noise distribution. The regimes follow a first-order Markov chain, with transition matrix Q that assigns 0.9 probability to remain in the same regime, and 0.1 probability to transition to the next one. Continuous latent variables $\boldsymbol{z}_{1:M}$ are initialised from a K-component Gaussian Mixture Model with uniform categorical prior. For $t > M$, they evolve according to a Gaussian transition model with K regimes and regime-specific parameters. For each $k \in [K]$, the mean function $\boldsymbol{m}(\cdot, k)$ is given by a two-layer MLP with cosine activations, where each layer is a locally linear map (Zheng et al., 2018), constrained

by a randomly generated causal graph $\mathbf{G}_k$. Covariance matrices $\mathbf{\Sigma}(\cdot, k)$ are diagonal, and we consider three cases:

- Constant noise: fixed, $\mathbf{\Sigma}(\cdot, k) = \mathbf{\Sigma}$.
- Heterogeneous noise: regime-dependent, $\mathbf{\Sigma}(\cdot, k) = \mathbf{\Sigma}(k)$.
- History-dependent noise: diagonal entries proportional to the absolute value of $\boldsymbol{m}(\cdot, k)$.

The latter two satisfy Assumption (s2), i.e., distinct diagonal ratios across at least one pair of regimes. Observations are generated by passing latent variables through a two-layer MLP with Leaky ReLU activations. Unless otherwise stated, we generate $N = 10000$ training samples and 1000 evaluation samples, each of length $T = 100$.

Metrics. To evaluate the learned solutions against ground truth, we use the following metrics (details in Appendix E.3):

- **Regime estimation:** the component with highest posterior probability is selected, and the F_1 score is computed after regime permutation alignment.
- **Continuous latent variable alignment:** we report both weak and strong mean correlation coefficients (MCC). The strong MCC, also defined in Khemakhem et al. (2020), is a standard measure in ICA and evaluates alignment up to scaling and permutation. In contrast, weak MCC (Kivva et al., 2022) allows affine transformations.
- **Causal structure recovery:** the F_1 score between ground-truth and estimated causal graphs. This is obtained by thresholding the Jacobian after regime and continuous latent variable permutation alignment.
- **Transition functions:** the averaged L_2 distance and R^2 score between ground-truth and estimated means $\boldsymbol{m}(\cdot, k)$ after regime permutation and latent variable affine alignment.

6.1 Baseline Comparison

We compare iSDS against five representative baselines in identifiable representation learning:

- **iVAE*** (Khemakhem et al., 2020) where ground-truth discrete regimes are provided as auxiliary information (a stronger baseline).
- **SNICA** (Hälvä et al., 2021), based on linear switching dynamical systems, where each latent variable follows its own SDS.
- **TDLR** (Yao et al., 2022a), which assumes non-Gaussianity of temporal latent variables but no switches.
- **NCTRL** (Song et al., 2023), based on SDSs, but assuming Gaussianity on the observations, and certain non-Gaussianity on the latent variables.
- **CtrlNS** (Song et al., 2024), which assumes the number of switches is known and directly depends on the observations.

To compare iSDS with other baselines, we explore different data settings, A to F, all with $n = 10$ observed dimensions and the following characteristics:

A) $m = 3$, $K = 3$, $M = 1$, constant noise.

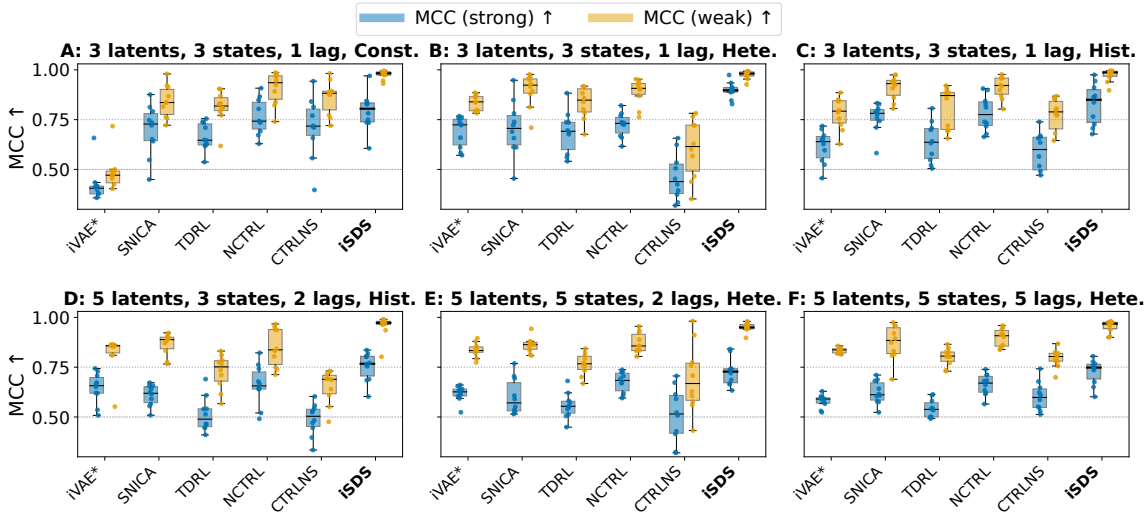
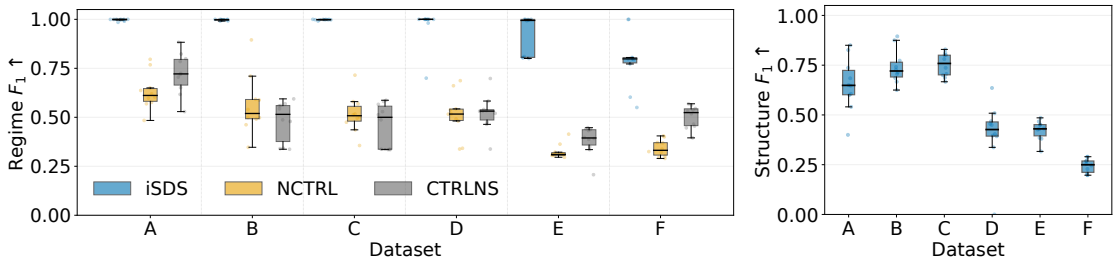


Figure 5: Weak and strong MCC scores across data settings A–F (10 data set seeds each).



(a) (b)

Figure 6: (a) Regime estimation and (b) Causal structure recovery across data settings A-F (10 data set seeds each).

B) $m = 3$, $K = 3$, $M = 1$, heterogeneous noise.

C) $m = 3$, $K = 3$, $M = 1$, history-dependent noise.

D) $m = 5$, $K = 3$, $M = 2$, history-dependent noise.

E) $m = 5$, $K = 5$, $M = 2$, heterogeneous noise.

F) $m = 5$, $K = 5$, $M = 5$, heterogeneous noise.

Figure 5 shows the weak and strong MCC scores for each configuration (A-F) across 10 data set seeds. We generally observe that iSDS outperforms the baselines in terms of weak MCC, achieving near-perfect affine alignment in all settings. This is followed by SNICA and NCTRL, as expected since both are SDS-based approaches. Regarding strong MCC, iSDS again outperforms all baselines, although performance drops for settings A, and D-F. In A-C the only difference lies in noise distribution. Lower strong MCC in A is expected, since identifiability up to scaling and permutation is not guaranteed under constant covariance matrices. In D-F, the lower MCC scores can be attributed to increased complexity from higher lags, dimensions, regimes, and nonstationary noise. For regime estimation, Figure 6a compares iSDS with NCTRL and CtrlNS as they extract regimes in a comparable way. iSDS consistently achieves superior F_1 scores, where performance decreases with more regimes (E and F). NCTRL performs significantly worse, despite being based on SDSs. This can

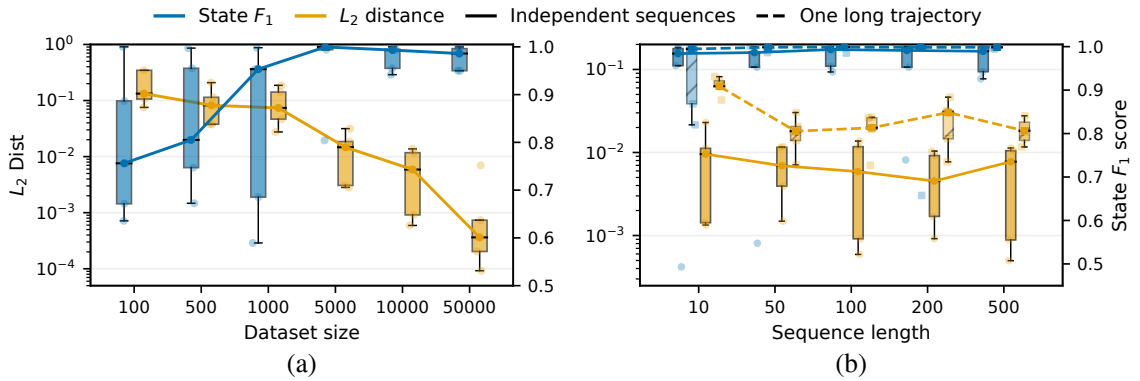


Figure 7: Data efficiency analysis of iSDS by varying (a) data set size, or (b) sequence length.

be explained by the fact that NCTRL estimates regimes directly from observations under Gaussianity assumptions, which is a model mismatch as the generated observations are generally non-Gaussian. Finally, Figure 6b reports causal structure recovery. In A-C, the F_1 score improves with increasing nonstationarity of the noise, consistent with our theoretical results. In contrast, settings D-F show reduced performance in line with the previously reported MCC scores.

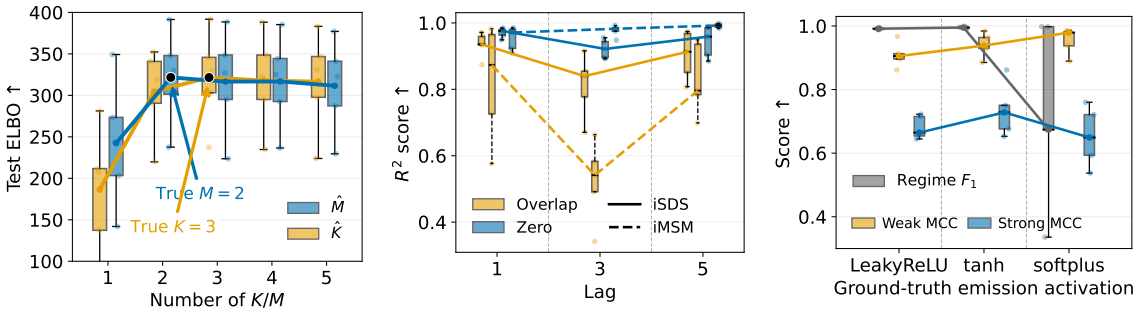
6.2 Data Efficiency

To assess the sample and sequence efficiency of iSDS, we generate data with $n = m = 5$ latent and observed dimensions, $M = 2$ lags, $K = 3$ regimes, and constant noise. We evaluate the regime F_1 score, and the average L_2 distance of mean functions $\boldsymbol{m}(\cdot)$ after affine alignment. Figure 7 shows two complementary settings, where we vary data set size (a) and sequence length (b), respectively. In panel (a) the sequence length is fixed to $T = 100$, and in panel (b), we compare using 10k independent samples of varying length against windows subsampled from one long sample (of length 10k).

We observe that large sample sizes (panel (a)) substantially improve regime recovery and mean-function estimation. In contrast, once enough sequences are available (panel (b)), increasing sequence length yields comparatively small gains, and accurate regime recovery is possible even from short sequences. Moreover, subsampling from long trajectories can be effective for recovering regimes, but may degrade estimation of the transition functions. Overall, these results suggest that iSDS benefits more from additional independent trajectories than from longer individual trajectories.

6.3 Model Selection

Our theoretical framework establishes that for both iMSM and iSDS, the number of regimes K is identifiable. Assuming consistency holds for iSDS, our results facilitate learning the ground truth K by model selection (McLachlan and Peel, 2000). Similar arguments apply for selecting M, where for a smaller lag the model becomes misspecified, and for larger M, extra lags may be unused. We show this experimentally, and leave more thorough theoretical analyses for future work. To select K and M we use the elbow method (Thorndike, 1953),



(a) (b) (c)

Figure 8: (a) Model selection results with iSDS for five data set seeds. Test ELBO for models with increasing number of switches ($\hat{K}$; in yellow), and lags ($\hat{M}$; in blue). Black points indicate ground-truth number of switches and lags ($K = 3$ and $M = 2$, respectively). (b–c) Ablation experiments on (b) intersection assumptions with iMSM and iSDS for 5 data set seeds in each setting, and (c) nonlinear activation of the emission function in iSDS. For (b), we use average R^2 of mean functions with zero measure intersections (Zero), and nonzero measure intersections (Overlap) across lags.

which is a standard heuristic for finite mixture models. For iSDS, we use the ELBO on a held-out data set, rather than the log-likelihood which could be possible for iMSM. Figure 8a showcases model selection for iSDS with $M = 2$ lags and $K = 3$ where we increase the estimated regimes $\hat{K}$ while maintaining the ground-truth lag M (yellow line), and increase the estimated lag $\hat{M}$ while maintaining the ground-truth number of regimes (blue line). In both cases, we observe that the highest test ELBO is reached at the ground-truth M and K, and the performance plateaus for higher choices of $\hat{M}$ and $\hat{K}$.

6.4 Ablations

We perform two ablations to empirically assess whether the sufficient conditions used in our identifiability theory are necessary in practice. First, Section 4.1 establishes identifiability for parametric MSMs with analytic functions and unique indexing. For $M = 1$, a sufficient condition is the existence of an open set where regime-specific mean (and covariance) functions do not intersect. However, for general M, we require this set to have full measure. These are sufficient, but not proven necessary, conditions. We assess necessity experimentally by designing two settings for both iMSM and iSDS: (i) Zero, where the ground-truth mean functions intersect at most on a zero measure set; and (ii) Overlap, where functions coincide on a set with nonzero measure but differ elsewhere. We consider $n = m = 5$, $K = 3$, constant noise, and lags $M \in \{1, 3, 5\}$. In the Overlap setting, approximately 15% of the latent samples are generated using the same transition function. See Appendix E.1 for further details. Second, Section 4.2 establishes identifiability for SDSs under piecewise linearity. To examine whether identifiable representations are restricted to this function class, we additionally generate observations with $n = m = 5$, $K = 3$, $M = 2$, using deeper

neural networks with 5 hidden layers and either `LeakyReLU` or smooth activations (`tanh` and `softplus`).

Figure 8b reports the average R^2 of the mean functions $\boldsymbol{m}(\cdot)$ after regime permutation alignment (with an additional affine alignment for iSDS). Using R^2 rather than L_2 facilitates comparisons across lags with different input/output scales. As expected, Zero outperforms Overlap. However, Overlap shows strong fits with high R^2 scores, far from indicating model failure. For iSDS, the affine alignment can cause the R^2 scores to be higher. However, for iMSM we observe strong fits for $M = 5$. This suggests that the stronger full measure non-intersection is sufficient but may not be necessary in practice, as otherwise we would obtain much lower R^2 scores (e.g., $R^2 < 0.5$). Therefore, the decrease in performance in Overlap can be attributed to estimation challenges from the overlap region, where tracking the regimes is more difficult, rather than to a fundamental lack of identifiability.

In Figure 8c, we report regime F_1, strong MCC, and weak MCC scores. While we establish identifiability under piecewise linear assumptions, using smooth activations for data generation still enables iSDS to achieve strong recovery of the regime structure, despite some instability across seeds when using `softplus`. The high weak MCC scores also suggest strong recovery of the continuous latents up to affine ambiguity, while the strong MCC scores are comparable across the different activation functions considered. This suggests that piecewise linearity may not be necessary for identifiability in regime-switching models and indicates interesting future work directions for extending the theoretical results to more general invertible emissions.

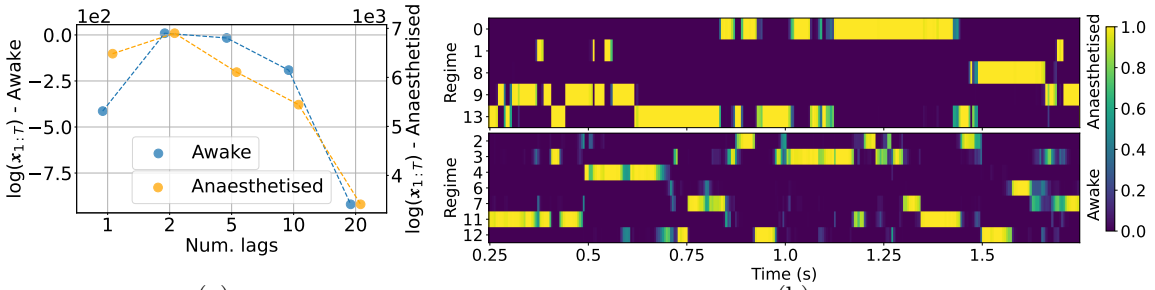
7. Applications to Real-World Data

Having established the validity of our theoretical results under controlled conditions, we now turn to real-world data, where ground-truth structure is generally unavailable. In this section, we therefore focus on assessing the applicability of the theory developed in Section 4 in a fully unsupervised setting. We study data sets from diverse scientific domains and investigate whether assumptions on smoothness and nonstationarity (i.e., Assumptions (m2) and (s2)) facilitate learning structured latent representations. Importantly, our identifiability-based framework establishes that learned representations are one-to-one with respect to observations, up to acceptable equivalences (e.g., scaling, permutation). This uniqueness property allows a more trustworthy interpretation of the recovered structures, allowing us to relate them to established domain knowledge. Overall, our aim in this section is to demonstrate how the proposed iMSM and iSDS can support scientific discovery in real-world data.

7.1 Brain Activity Data

We first illustrate iMSM on electrocorticography (ECoG) recordings from the NeuroTycho database,[4] originally presented by Yanagawa et al. (2013). Signals from 128 electrodes were recorded in a macaque monkey under two conditions: normal wakefulness (Awake) and propofol-induced loss of consciousness (Anaesthetised). Following prior work (Mediano et al., 2023), we expect the awake condition to exhibit more complex dynamics, reflected in more frequent transitions between latent regimes.

4. The NeuroTycho database is available at `http://neurotycho.org/`.



(a) (b)

Figure 9: (a) Test log-likelihood of ECoG data using different lags and $K = 15$ regimes. (b) Example posteriors over regimes for Awake and Anaesthetised conditions. Posteriors are smoothed with a length-3 uniform kernel to suppress frame-level artifacts.

The raw ECoG is sampled at 1kHz and comprises Awake and Anaesthetised segments, each lasting 929.8 and 650.65 seconds respectively. We select 21 electrodes approximately covering the visual cortex of the brain. We first apply a second-order Butterworth notch filter at 50Hz to eliminate line noise. Then, we downsample the data to 200Hz, standardise each channel independently, and chunk each sequence into epochs of 2 seconds ($T = 400$). This yields 464 Awake and 325 Anaesthetised segments, with 50 from each condition held out for testing.

In both data settings, we set $K = 15$ and parametrise transition functions with cosine activations (model selection details in Appendix E.5). Figure 9a reports test log-likelihoods for iMSM as a function of lag order M. As expected, Awake is consistently harder to model (lower log-likelihoods), indicating more complex dynamics. iMSM with $M = 2$ achieves the best performance in both settings. Notably, the Anaesthetised condition remains competitive at $M = 1$ but degrades rapidly for $M > 2$, consistent with simpler dynamics. In contrast, Awake benefits significantly from higher lag orders.

Figure 9b illustrates posterior regime occupancy for 2 s epochs in each condition (smoothed by a length-3 uniform kernel). We only show regimes with more than 1% occupancy in the epoch. Anaesthetised shows prolonged stays within a smaller subset of regimes, whereas Awake occupies a larger fraction of regimes with more frequent transitions. Table 2 quantifies these differences: relative to Anaesthetised, Awake uses more regimes overall and per sample, exhibits higher transition frequency (Hz), and has shorter regime durations on average.

Crucially, iMSM provides regime identifiability guarantees to ensure the differences in regime usage reflect condition-specific structure supported by the data. This highlights the importance of considering iMSM or iSDS for neuroscience time series such as ECoG, complementing existing methodologies (Mediano et al., 2023).

7.2 Financial Data

Financial time series are a direct application area for regime-switching models. Early developments in MSMs (Hamilton, 1989) focused on modelling market-cycle regimes and laid the foundations for regime-switching models in finance. In this section, we study US sector-based stock price data and use iSDS to extract structured latent representations and

Condition	Transition freq. (Hz)	Total usage	Usage per sample	Max. duration (s)
Awake	22.05 ± 1.06	53.33%	$44\% \pm 1.47\%$	0.22 ± 0.02
Anaesthetised	17.71 ± 1.29	33.33%	$26.40\% \pm 1.48\%$	0.35 ± 0.04

Table 2: Regime statistics on test visual-cortex ECoG data with $M = 2$ and $K = 15$. "Total usage" is the fraction of regimes occupied at least once across the test set. "Usage per sample" is the mean fraction of regimes occupied within a single 2 s epoch. "Max duration" is the mean of the per-epoch maximum dwell time. See Appendix E.5 for details.

regimes. Related applications are considered with both MSMs (Alizadeh and Nomikos, 2004; Cavaliere et al., 2014; Mahmoudi and Ghaneei, 2022) and SDSs (Azzouzi and Nabney, 1999; Carvalho and Lopes, 2007; Fox et al., 2011). Empirically, researchers typically work with *log returns* or *log price(-dividend)*.

Log returns $r_t = \log\left(\frac{p_t}{p_{t-1}}\right)$ are often treated as stationary (Cont, 2001), and regime-switching mainly focuses on regime-dependent means and/or volatilities. This includes MSMs combined with conditional heteroskedasticity (GARCH; Bollerslev, 1986), where the variance depends on latent variables and/or past returns (Ang and Bekaert, 2002; Carvalho and Lopes, 2007). Such models often interpret states as bull/bear or tranquil/turbulent and are widely used for risk assessment and allocation (Guidolin and Timmermann, 2007).

Because prices are typically nonstationary, regime-switching on log price data is often modelled through present-value frameworks that separate company fundamentals (e.g., earnings) from speculative components (Balke and Wohar, 2009; Chan and Santi, 2021). These models capture longer-horizon regimes, such as persistent bear or bull phases.

Data modality and identifiability. We analyse both data modalities (log returns and log prices) to illustrate our theory in a real-world setting. Our SDS results imply that regimes are identifiable under suitable modelling conditions (Theorem 13). For continuous latents, identifiability up to permutation and scaling requires certain heterogeneous or history-dependent noise (Theorem 13; Corollary 15). Specifically, Assumption (s2) requires that for a pair of regimes k_1, k_2, the ratios of latent variances across dimensions $i \in [m]$ must be distinct. This condition fails if regimes only introduce a common rescaling of all latent variables. For instance, if $\mathbf{\Sigma}(k) = c_k \mathbf{\Sigma}_0$, then the ratio between regime-dependent variances is constant across dimensions, and therefore violates (s2).

This motivates our comparison between log returns and log prices. Log returns $r_t = \log p_t - \log p_{t-1}$ remove accumulated price-level effects and focus on one-step relative changes. They can still satisfy (s2) if regimes induce sufficient dimension-specific variance. However, when returns mainly capture market-volatility or risk, the regime variance effect may be closer to a global rescaling, making (s2) harder to meet. In contrast, log prices retain accumulated trends, which may evolve differently across sectors over market phases. This can induce more heterogeneous latent variable variances, making (s2) more plausible for log prices. We therefore evaluate both settings to assess which modality yields more structured representations in iSDS.

Data set and model setting. We source 100 U.S. stock prices from Bloomberg Terminal (Bloomberg L.P., 2025), grouped into 10 sectors, and retain daily prices for the past ten years (Jan. 2015–May 2025). This yields a sequence of 2630 time steps. See Appendix E.6

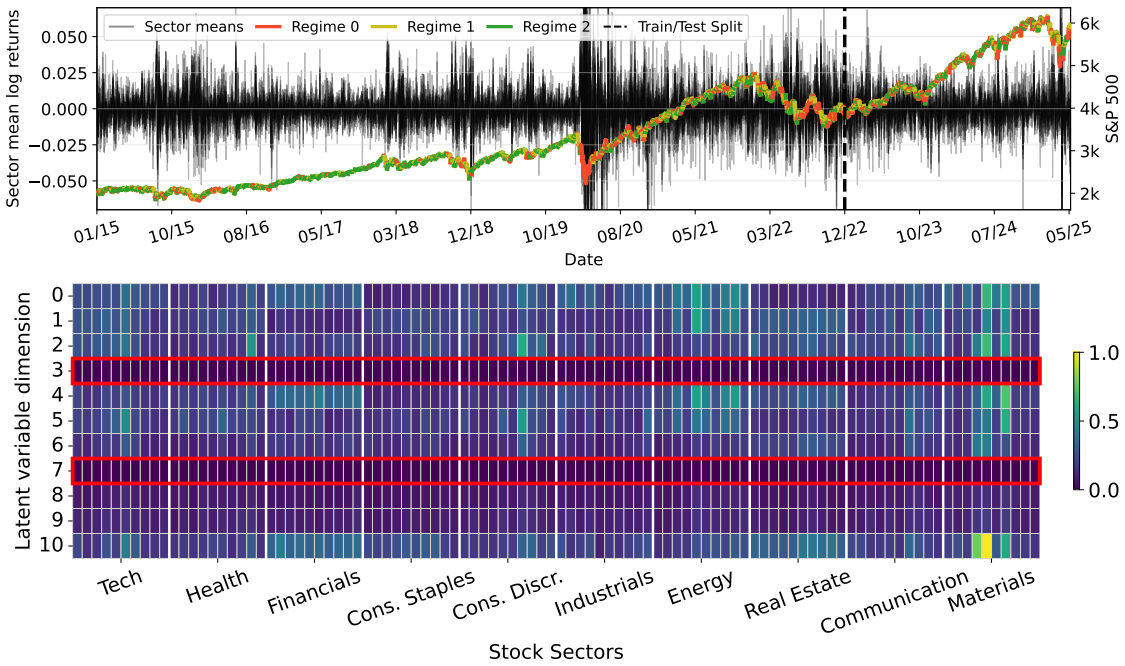


Figure 10: Top: Regimes inferred from **log returns**, visualized on the S&P 500 with sector-mean log returns in black; the dashed line marks the train/test split. Bottom: Mean absolute decoder Jacobian for **log returns**, grouped by sector. Red rectangles highlight weak areas.

for the company list and sector mapping. For training, we draw random subsequences of $T = 200$, from the first 2000 time steps, and we leave the remaining 630 for testing. We train iSDS with 11 latent variables and 3 regimes: ten sector latents (one per sector) plus a global latent, and three market regimes intended to capture, e.g., bull/bear/sideways or high/mid/low volatility, similar to related works on financial time series (Chan and Santi, 2021). We set $M = 5$ and use heterogeneous covariances, i.e., $\mathbf{\Sigma}(\cdot, k) = \mathbf{\Sigma}(k)$ for $k \in [K]$. See Appendix E.2 for additional training details.

Results on log returns. Figure 10 shows the training/test segmentation overlaid on the S&P 500 for context. When trained on log returns, iSDS exhibits regimes capturing short-horizon volatility adaptation, where green, yellow, and red colours correspond to low, medium, and high volatility, respectively. We observe a prolonged low-volatility period from late 2016 to early 2018 and a significantly long high-volatility regime in March 2020 (COVID-19). From 2020 onward, the segmentation consists mainly of medium and high-volatility regimes, consistent with the increased variability of the S&P 500.

To assess latent structure, Figure 10 bottom displays the decoder Jacobian (latents $\rightarrow$ assets), grouped by sector. The pattern shows little sector-specific structure: multiple latent variables mapping to the same sector (e.g., significant concentration around the "Materials" sector), while some latents (e.g., z_3 and z_7) are weakly used. This is consistent with the identifiability discussion above: the regimes learned from log returns primarily capture market volatility, suggesting that regime changes act mostly as a common variance rescaling rather than inducing distinct variance ratios across sectors.

Results on log prices. Figure 11 reports the corresponding segmentation for log prices. Here, iSDS extracts longer-horizon regimes: a green state aligned with bull periods, and a

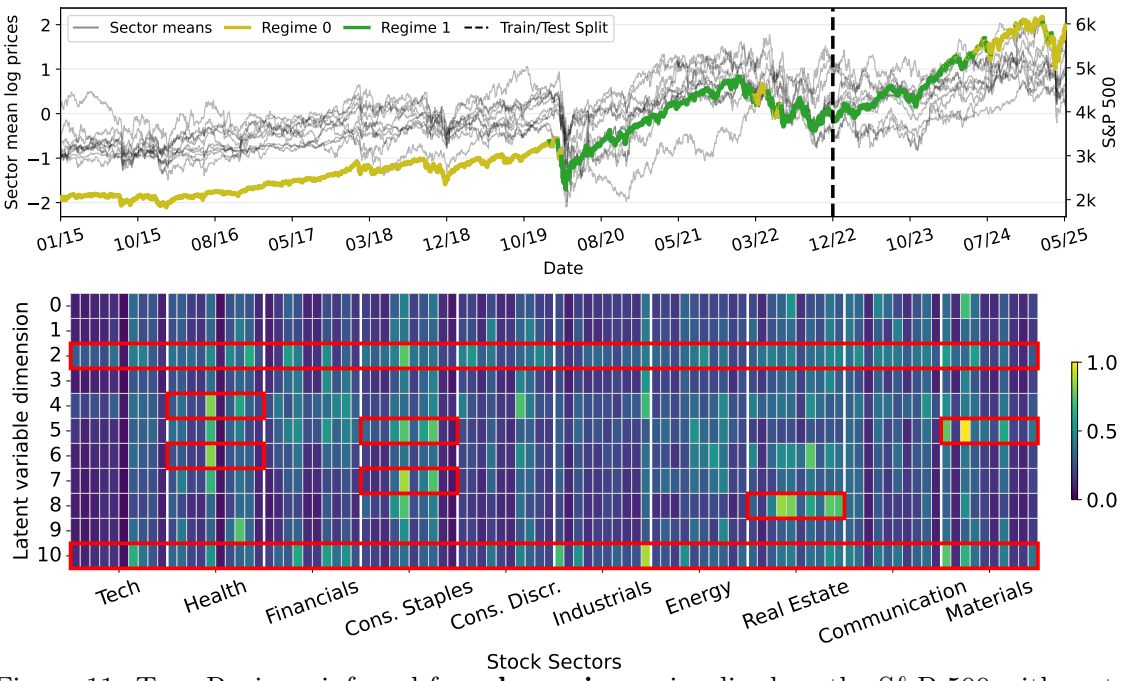


Figure 11: Top: Regimes inferred from **log prices**, visualized on the S&P 500 with sector-mean log prices in black; the dashed line marks the train/test split. Bottom: Mean absolute decoder Jacobian values for **log prices**, grouped by sector. Red rectangles highlight latent variables that act globally, or rather focus on specific sectors.

yellow state aligned with lateral/bear phases. Compared to the returns model, the inferred regimes are more persistent and track macroeconomic phases, consistent with the literature on present-value frameworks (Chan and Santi, 2021).

The decoder Jacobian (Fig. 11 bottom) shows a significantly clearer structure than in the returns case (cf. Fig. 10 bottom). Several latents align with sectors or sector pairs (e.g., z_5: Materials/Consumer Staples; z_5, z_7: Consumer Staples; z_8: Real Estate; z_4, z_6: Health Care). Other latents (z_2 and z_{10}) behave as global factors. Some sectors (e.g., Technology and Industrials) spread across multiple latents, which is plausible given their cyclicality and sensitivity to market conditions. In general, the log price model yields a more structured, sector-aware representation, consistent with our identifiability discussion, where log prices retain accumulated sector-specific trends, making dimension-specific variances more plausible. The lack of stronger sector-specific disentanglement is also consistent with our synthetic results, where iSDS did not achieve high strong MCC scores.

Summary. Comparing both modalities, iSDS on log returns discovers short-time volatility regimes with weak sector disentanglement, while log prices yield more persistent macro regimes with clearer sector-aligned latents. The stronger structured representation in the log price modality is consistent with the nonstationarity of the data and with our heterogeneity requirements for identifiable representations. In Appendix E.6, we provide additional visualizations of the learned latent structures for both data modalities.

7.3 Climate Data

Climate data is characterised by strong seasonal patterns and regime-dependent approaches have recently been explored in this context (Saggioro et al., 2020). Modelling Earth systems is crucial to understand climatological dynamics and to detect extreme events. In this section, we investigate high-dimensional spatio-temporal climate data to demonstrate the applicability of our theoretical framework to this scientific domain.

We consider the Normalised Difference Vegetation Index (NDVI) over Africa, following a setup similar to that of Varando et al. (2022). The NDVI measures vegetation greenness as the normalised ratio between near-infrared and red reflectance, and is widely used as a proxy for vegetation productivity and land-surface phenology (Pettorelli, 2013). We use the MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006 product (available at `https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod13c1-006`) (Didan, 2015) together with the ENSO 3.4 climate index, both provided in preprocessed form by Varando et al. (2022). The data set follows preprocessing similar to that described in their work: NDVI is reprojected to a 25km Equal-Area Scalable Earth (EASE) grid,[5] retaining the pixels corresponding to Africa (bounding box $[-20,\ 55]$, $[-37,\ 38]$ in [latitude, longitude]). For each year, the product provides NDVI values from day 1 to 353 every 16 days. We select the period 2001-2020. Missing values are filled temporally (per pixel) using linear interpolation. The resulting data set is a video tensor with 460 time steps and 289×356 pixels. The ENSO 3.4 index is sampled to match the 16-day temporal resolution. Since the index is available daily, a 16-day sliding window is used, followed by $16\times$ downsampling. With these data sets, we expect our iMSM and iSDS to extract regimes corresponding to seasonal dynamics and, similarly to Granger-PCA (Varando et al., 2022), investigate whether the learned latent variables align with ENSO 3.4.

Model setting. We select the 360 most recent time steps, and reserve the first 100 for testing and model selection. We train iMSM following Varando et al. (2022), where we use PCA on the land pixels (a total of 56196 pixels) and retain 10 principal components. For iSDS, we use random 64×64 patches from land pixels, and use a CNN-MLP encoder-decoder. Here we consider 20 latent variables to account for additional spatial features on the continent shape. Given the inherent seasonality in climate, we assume known $K = 4$ (one regime per season). Through model selection, we set heterogeneous covariances, GELU activations in the transitions, $M = 1$ for iMSM, and $M = 2$ for iSDS. See Appendices E.2 and E.7 for additional training and model selection details, respectively.

Segmentation results. Figure 12 shows segmentation results of both models. Because iMSM is trained on 10 principal components, the extracted regimes act globally across the continent. We observe three coherent seasonal regimes recurring each year: Mar-May, Jun-Aug, and Sep-Feb. In contrast, iSDS yields region-dependent regime structures: one regime is predominant over the Sahara and Congo (patches 1 and 4 in Figure 12A); while two/three regimes alternate over Ethiopia, the Sahel, and South Africa (patches 2, 3, and 5 in Figure 12B). Although the spatial regime structure differs across regions, the underlying regime effects are shared in iSDS: for example, South Africa's segmentation is approximately

5. Information on EASE grid reprojections can be found here `https://nsidc.org/data/user-resources/help-center/guide-ease-grids`.

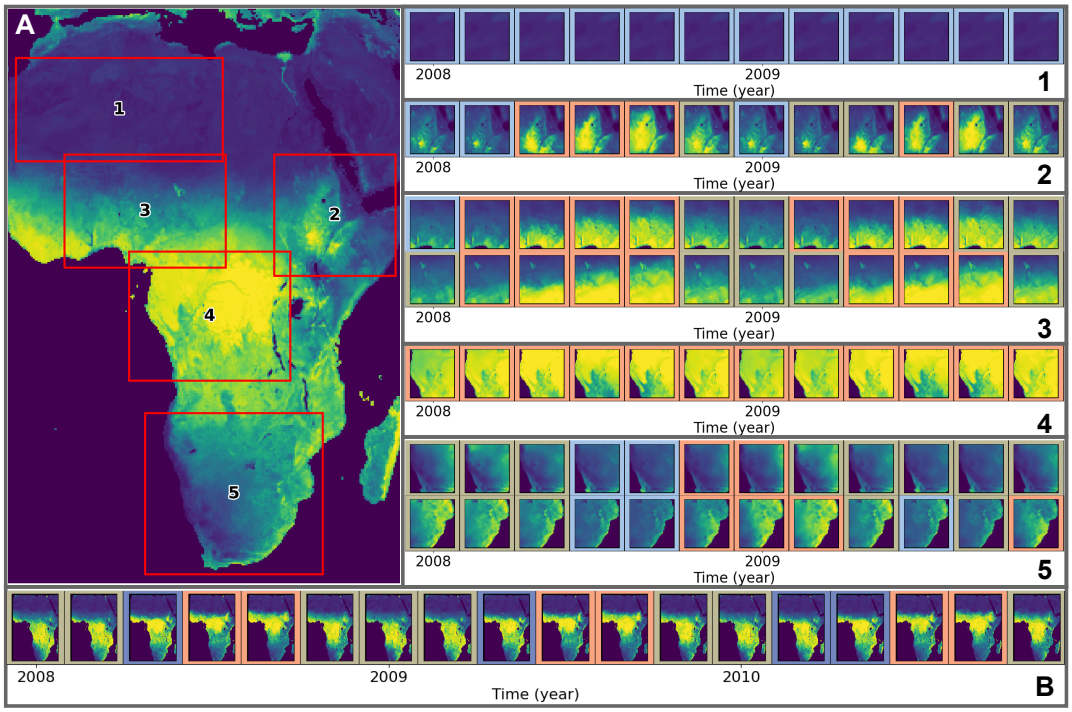


Figure 12: **A.** iSDS segmentations of NDVI for five regions across Africa: (1) Sahara, (2) Ethiopia, (3) Sahel, (4) Congo, and (5) South Africa. **B.** Continent-wide iMSM segmentation. Brighter pixels indicate higher NDVI values, and coloured edges denote inferred regimes.

the opposite of the Sahel/Ethiopia segmentations. Visualisations of global and regional year-based structures are provided in Appendix E.7.

Correlation with ENSO. To quantitatively assess the learned regime-dependent dynamics, we relate the learned latents to the preprocessed ENSO 3.4 time series. For iMSM (global), we split the time series into seasonal slices determined by the regimes and compute lagged Pearson correlations between each PCA component and ENSO 3.4 over lags ranging from 0-240 days (16-day steps; ENSO leads), and also report the "All times" baseline (i.e., no splits). For iSDS (regional), we sample 10 64×64 land patches per region, compute latent-ENSO correlation matrices (latents $\times$ lags), and combine them with Fisher z-averaging. For each season/region, we report the peak correlation strength and its lag. Full correlation matrices, and additional visualisations are provided in Appendix E.7.

The results are summarised in Tables 3 and 4. For iMSM (cf. Table 3), we observe that regime-dependent splits yield clearer peaks than the "All times" baseline. This indicates that aggregating across seasons attenuates the lagged effects of ENSO on the time series. Our results show that iMSM segments the time series into regimes where the effects of ENSO can be perceived with different lags. For iSDS (cf. Table 4), the regional responses are heterogeneous. We observe that Sahara, Congo and Sahel are not significantly aligned with ENSO. On the other hand, South Africa shows slightly higher correlation with ENSO in the South African Summer compared to Winter. Furthermore, Ethiopia shows a very high correlation with ENSO during the months of January-March. In climate analysis, researchers study effects of phenomena such as ENSO across lag scales (Rouault and Tomety, 2022), and

Season	Correlation	Lag (days)
All times	0.38	80
Mar–May	0.49	208
Jun–Sep	0.37	80
Oct–Feb	**0.53**	**80**

Table 3: Per-season and global ENSO alignment with iMSM PCs. Peak correlation and lag with ENSO leading.

Region	Season	Correlation	Lag (days)
Sahara	All times	0.15	48
Ethiopia	Jan–Mar	**0.56**	**144**
	Apr–Aug	0.27	100
	Aug–Dec	0.26	0
Sahel	Mar–Aug	0.18	112
	Sep–Feb	0.19	0
Congo	All times	0.12	144
South Africa	Mar–Jun & Sep-Nov	0.25	64
	Jun–Sep	0.22	128
	Dec–Feb	0.28	192

Table 4: Per-season ENSO alignment with iSDS latents. For each region, we report averaged peak correlation and lag with ENSO leading.

within different regions, including the Sahel (Nicholson, 2013) and Ethiopia (Bekele-Biratu et al., 2018). In general, iMSM and iSDS can provide noteworthy insights into external drivers such as ENSO, by decoupling seasonal and regional effects.

Summary. Climate data yield a strong seasonal structure in NDVI over Africa. iMSM captures the global scales, while iSDS reveals region-specific regime structures. Although interpretability of the latent variables is complex in climate time series, our regime identifiability results still enable interpretable regime-conditioned analyses. Extending iSDS with recent identifiable spatiotemporal frameworks could further disentangle regional effects and improve latent space interpretability (Wang et al., 2025).

8. Conclusions

We developed a unified identifiability framework for regime-switching models with multi-lag dependencies. For Markov switching models, we framed the model as a temporal finite mixture and proved identifiability of both the number of regimes and the regime-specific multi-lag transitions in a nonlinear Gaussian setting, with a thorough nonparametric analysis provided in Appendix A. For switching dynamical systems, we established identifiability of latent variables up to permutation and scaling and derived conditions under which regime-dependent latent causal graphs are identifiable up to regime and node permutations. Importantly, the guarantees hold in a fully unsupervised setting via flexible architectural (function class) and noise assumptions. Empirically, we validated the theory on synthetic data and demonstrated iMSM and iSDS across scientific domains, showing that identifiability yields trustworthy interpretations by ensuring unique representations up to acceptable equivalences.

While we provide strong identifiability results, our iSDS guarantees assume piecewise linear generative functions and specific noise classes, which may not hold universally. Our ablations show that smooth emissions still yield strong latent recovery, suggesting that piecewise linearity may be a limitation of the current proof. Moreover, verifying whether these assumptions align with real-world data is inherently challenging, as we may have limited knowledge of the underlying mechanisms that generate data. This highlights the importance of developing identifiability results under the most flexible assumptions possible. On the

estimation side, we rely on variational inference and do not prove statistical consistency. Empirically, we observe strong synthetic performance, with recovery of regime-dependent causal structures becoming more challenging as dimensionality increases. Although K is identifiable under our assumptions, in practice both K and M must be selected from finite data. Thus, reliable model selection remains tied to the quality of the estimator. In our real-world experiments, allowing a large number of regimes can help mitigate state collapse, while selecting an unnecessarily large maximum lag increases the risk of overfitting.

Our framework is complementary to existing methodologies and suggests several directions for future work. Some examples include incorporating contemporaneous effects, continuous-time dynamics, and infinite mixture processes, as well as going beyond piecewise linearity. These extensions would broaden the applicability of identifiable regime-switching models for scientific discovery. Furthermore, our identifiability results extend beyond standard MSMs with time-homogeneous Markov switching to models with time-dependent regime distributions, offering room for experimental exploration.

Acknowledgments

We thank Gherardo Varando and Gustau Camps-Valls (Image Processing Laboratory, Universitat de València) for early discussions on the climate application and for sharing preprocessed data sets used in Section 7.3. C. B.-R. was supported by the Department of Computing at Imperial College London through the Doctoral Teaching Programme in Computing (PhD scholarship) at the time of submission, and also thanks colleagues at Imperial College London for helpful discussions on the theoretical results and contents of this paper. Y. W. is supported in part by the Office of Naval Research under grant N00014-23-1-2590, the National Science Foundation under grant No. 2231174, No. 2310831, No. 2428059, No. 2435696, No. 2440954, and a Michigan Institute for Data Science Propelling Original Data Science (PODS) grant. The authors declare no competing interests.

Appendix A. Proof of Theorem 8

We organise the proof strategy into 4 steps.

1. We show the requirement for identifiability is linear independence of the trajectory family (Appendix A.1).

2. We provide linear independence results for products of nonparametric functions. We first start with $M = 1$ (Appendix A.2.1), and we then strengthen the assumptions (Appendix A.2.2) to allow products of M functions (Appendix A.2.3).

3. We prove linear independence of the trajectory family for nonparametric transitions of order M using induction (Appendix A.2.4).

4. We show the parametric assumptions (m1), (m2) satisfy linear independence of the trajectory family (Appendix A.3).

A.1 Linear Independence Renders Identifiability

Theorem 16 *If the functions in $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$ are linearly independent under finite mixtures, then the distribution family $\mathcal{M}^T(\Pi^M_{\mathcal{A}}, \mathcal{P}^M_{\mathcal{B}})$ is identifiable as defined in Def 7.*

Proof By Proposition 4, linear independence of $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$ under finite mixtures implies identifiability up to permutation in the sense of Definition 2. Then, for $p(\boldsymbol{z}_{1:T})$ and $\tilde{p}(\boldsymbol{z}_{1:T})$ from Definition 7, we have $C = \tilde{C}$, and for every $1 \leq i \leq C$, there exists $1 \leq j \leq \tilde{C}$ such that $c_i = \tilde{c}_j$ and:

$$p_{a^i}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b^i_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M}) = p_{\tilde{a}^j}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{\tilde{b}^j_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M}),\ \forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}. \quad (31)$$

Given that we have conditional PDFs, if the joint distributions are equal on $\boldsymbol{z}_{1:T}$, then the distributions on $\boldsymbol{z}_{1:T-1}$ are also equal:

$$p_{a^i}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T-1} p_{b^i_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M}) = p_{\tilde{a}^j}(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T-1} p_{\tilde{b}^j_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M}),\ \forall \boldsymbol{z}_{1:T-1} \in \mathbb{R}^{(T-1)m}. \quad (32)$$

Therefore, $p_{b^i_T}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1},\dots,\boldsymbol{z}_{T-M}) = p_{\tilde{b}^j_T}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1},\dots,\boldsymbol{z}_{T-M})$ for all $\boldsymbol{z}_t,\dots,\boldsymbol{z}_{t-M} \in \mathbb{R}^m$. We can follow the same reasoning for other time indices to have $p_{b^i_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M}) = p_{\tilde{b}^j_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M})$ for all $t > M, \boldsymbol{z}_t,\dots,\boldsymbol{z}_{t-M} \in \mathbb{R}^m$. Similar logic applies to the initial distribution, where we have $p_{a^i}(\boldsymbol{z}_{1:M}) = p_{\tilde{a}^j}(\boldsymbol{z}_{1:M})$ for all $\boldsymbol{z}_1,\dots,\boldsymbol{z}_M \in \mathbb{R}^m$; and we have $K_0 = \tilde{K}_0$, $K = \tilde{K}$.

Finally, if there exists $t_1 \neq t_2$ such that $b^i_{t_1} = b^i_{t_2}$ but $\tilde{b}^j_{t_1} \neq \tilde{b}^j_{t_2}$, we have for any $\boldsymbol{\alpha} \in \mathbb{R}^{Mm}, \boldsymbol{\beta} \in \mathbb{R}^m$:

$$\begin{aligned} p_{\tilde{b}^j_{t_1}}(\boldsymbol{z}_{t_1} = \boldsymbol{\beta}|\boldsymbol{z}_{t_1-1:t_1-M} = \boldsymbol{\alpha}) &= p_{b^i_{t_1}}(\boldsymbol{z}_{t_1} = \boldsymbol{\beta}|\boldsymbol{z}_{t_1-1:t_1-M} = \boldsymbol{\alpha}) \\ &= p_{b^i_{t_2}}(\boldsymbol{z}_{t_2} = \boldsymbol{\beta}|\boldsymbol{z}_{t_2-1:t_2-M} = \boldsymbol{\alpha}) \\ &= p_{\tilde{b}^j_{t_2}}(\boldsymbol{z}_{t_2} = \boldsymbol{\beta}|\boldsymbol{z}_{t_2-1:t_2-M} = \boldsymbol{\alpha}), \end{aligned}$$

which implies linear dependence under finite mixtures of $\mathcal{P}^M_{\mathcal{B}}$. We note this contradicts the assumption of linear independence under finite mixtures of the trajectory family $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$: such assumption says we should have

$$\sum_i \sum_j \gamma_{ij} p_{b^i_{1:T-1}}(\boldsymbol{z}_{1:T-1}) p_{b^j_T}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1:T-M}) = 0, \quad \forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}, \quad (33)$$

if and only if $\gamma_{ij} = 0$ for all $1 \leq i \leq C/K, 1 \leq j \leq K$. However, given linear dependence on $\mathcal{P}^M_{\mathcal{B}}$, we can rearrange the terms and set $\gamma_{i,j} = \gamma_j \neq 0, 1 \leq j \leq K$ such that the above

equation is satisfied. Therefore a contradiction is reached and we have $b^i_{t_1} = b^i_{t_2} \Leftrightarrow \tilde{b}^j_{t_1} = \tilde{b}^j_{t_2}$ for all $M < t_1, t_2 \leq T$. ■

The next step is to show conditions under which the trajectory family $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$ is linearly independent under finite mixtures.

A.2 Linear Independence of Nonparametric Product Families

To establish linear independence of the nonparametric trajectory family, we analyze consecutive products of conditional distributions. For instance, when $M = 1$, considering two consecutive variables gives the product

$$\{p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, s_t)p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, s_{t+1})\}$$

with one overlapping variable $\boldsymbol{z}_t$. Increasing M increases the number of overlaps when considering $M + 1$ consecutive variables. The following product

$$\left\{\prod_{\ell=0}^{M} p(\boldsymbol{z}_{t+\ell} \mid \boldsymbol{z}_{t-1+\ell}, \ldots, \boldsymbol{z}_{t-M+\ell}, s_{t+\ell}) : (s_t, \ldots, s_{t+M}) \in [K]^{M+1}\right\}, \tag{34}$$

contains two types of overlapping variables for $M > 1$:

1. M variables that appear both as observed variables and as conditioning variables, namely $\boldsymbol{z}_{t+M-1}, \ldots, \boldsymbol{z}_t$ in Eq. (34).

2. $M - 1$ variables that appear only as conditioning variables in multiple factors, namely $\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M+1}$ in Eq. (34).

For $M > 1$, the additional overlaps between the conditioning variables $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M+1})$ complicate the analysis, as these prevent a direct extension of the case $M = 1$ presented in Balsells-Rodas et al. (2024).

A.2.1 Preliminaries

We recall the central result for the case $M = 1$ from Balsells-Rodas et al. (2024), which we name ***L**inear **I**ndependence on **P**roduct functions with* 1 *consecutive **O**verlap* (LIPO-1). It provides conditions under which linear independence is preserved for products with a single overlapping variable.

Lemma 17 (LIPO-1) *Consider two scalar real-valued function families $\mathcal{U}_I := \{u_i(\boldsymbol{y}, \boldsymbol{x}) | i \in I\}$ and $\mathcal{V}_J := \{v_j(\boldsymbol{z}, \boldsymbol{y}) | j \in J\}$ with $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathbb{R}^{d_y}$ and $\boldsymbol{z} \in \mathbb{R}^{d_z}$. We further make the following assumptions:*

(a1) Positive function values: $u_i(\boldsymbol{y}, \boldsymbol{x}) > 0$ for all $i \in I, (\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{R}^{d_y} \times \mathcal{X}$. Similar positive function values assumption applies to $\mathcal{V}_J$: $v_j(\boldsymbol{z}, \boldsymbol{y}) > 0$ for all $j \in J, (\boldsymbol{z}, \boldsymbol{y}) \in \mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$.

(a2) Unique indexing: for $\mathcal{U}_I$, $i \neq i' \in I \Leftrightarrow \exists\, \boldsymbol{x}, \boldsymbol{y}$ s.t. $u_i(\boldsymbol{y}, \boldsymbol{x}) \neq u_{i'}(\boldsymbol{y}, \boldsymbol{x})$. Similar unique indexing assumption applies to $\mathcal{V}_J$;

(a3) Linear independence under finite mixtures on specific nonzero measure subsets for $\mathcal{U}_I$: for any nonzero measure subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $\mathcal{U}_I$ contains linearly independent functions under finite mixtures on $(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}$.

(a4) Linear independence under finite mixtures on specific nonzero measure subsets for $\mathcal{V}_J$: there exists a nonzero measure subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$, such that for any nonzero measure subsets $\mathcal{Y}' \subset \mathcal{Y}$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$, $\mathcal{V}_J$ contains linearly independent functions under finite mixtures on $(\boldsymbol{z}, \boldsymbol{y}) \in \mathcal{Z} \times \mathcal{Y}'$;

(a5) Linear dependence under finite mixtures for subsets of functions in $\mathcal{V}_J$ implies repeating functions: for any $\boldsymbol{\beta} \in \mathbb{R}^{d_y}$, any nonzero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$ and any subset $J_0 \subset J$ such that $|J_0| < +\infty$, $\{v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) | j \in J_0\}$ contains linearly dependent functions on $\boldsymbol{z} \in \mathcal{Z}$ only if $\exists\ j \neq j' \in J_0$ such that $v_j(\boldsymbol{z}, \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \boldsymbol{\beta})$ for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$.

(a6) Continuity for $\mathcal{V}_J$: for any $j \in J$, $v_j(\boldsymbol{z}, \boldsymbol{y})$ is continuous in $\boldsymbol{y} \in \mathbb{R}^{d_y}$.

Then for any nonzero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$, $\mathcal{U}_I \otimes \mathcal{V}_J := \{v_j(\boldsymbol{z}, \boldsymbol{y}) u_i(\boldsymbol{y}, \boldsymbol{x}) | i \in I, j \in J\}$ contains linearly independent functions under finite mixtures defined on $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}$.

Proof Assume this sufficiency statement is false, then there exists a nonzero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$, $S_0 \subset I \times J$ with $|S_0| < +\infty$ and a set of nonzero values $\{\gamma_{ij} \in \mathbb{R} | (i,j) \in S_0\}$, such that

$$\sum_{(i,j) \in S_0} \gamma_{ij} v_j(\boldsymbol{z}, \boldsymbol{y}) u_i(\boldsymbol{y}, \boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}. \tag{35}$$

Note that the choices of S_0 and γ_{ij} are independent of any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ values, but might be dependent on $\mathcal{Z}$. By assumptions (a1), the index set S_0 contains at least 2 different indices (i,j) and (i',j'). In particular, S_0 contains at least 2 different indices (i,j) and (i',j') with $j \neq j'$, otherwise we can extract the common term $v_j(\boldsymbol{z}, \boldsymbol{y})$ out:

$$\sum_{(i,j) \in S_0} \gamma_{ij} v_j(\boldsymbol{z}, \boldsymbol{y}) u_i(\boldsymbol{y}, \boldsymbol{x}) = v_j(\boldsymbol{z}, \boldsymbol{y}) \left(\sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y}, \boldsymbol{x}) \right) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z},$$

and as there exist at least 2 different indices (i',j) and (i,j) in S_0, we have at least one $i' \neq i$, and the above equation contradicts assumptions (a1) - (a3).

Now define $J_0 = \{j \in J | \exists (i,j) \in S_0\}$ the set of all possible j indices that appear in S_0, and from $|S_0| < +\infty$ we have $|J_0| < +\infty$ as well. We rewrite the linear combination equation (Eq. (35)) for any $\boldsymbol{\beta} \in \mathbb{R}^{d_y}$ as

$$\sum_{j \in J_0} \left(\sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) \right) v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{36}$$

Now under assumption (a5), we can split the index set J_0 into subsets indexed by $k \in K(\boldsymbol{\beta})$ as follows, such that within each index subset $J_k(\boldsymbol{\beta})$ the functions with the corresponding indices are equal:

$$\begin{aligned} J_0 = \cup_{k \in K(\boldsymbol{\beta})} J_k(\boldsymbol{\beta}), \quad J_k(\boldsymbol{\beta}) \cap J_{k'}(\boldsymbol{\beta}) = \emptyset, \forall k \neq k' \in K(\boldsymbol{\beta}), \\ j \neq j' \in J_k(\boldsymbol{\beta}) \quad \Leftrightarrow \quad v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}), \quad \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned} \tag{37}$$

Write $\mathcal{Y} \subset \mathbb{R}^{d_y}$ the nonzero measure subset defined by assumption (a4). Then we can rewrite Eq. (36) for any $\boldsymbol{\beta} \in \mathcal{Y}$ as

$$\sum_{k \in K(\boldsymbol{\beta})} \left(\sum_{j \in J_k(\boldsymbol{\beta})} \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) \right) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{38}$$

Recall from Eq. (37) that $v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta})$ and $v_{j'}(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta})$ are the same functions on $\boldsymbol{z} \in \mathcal{Z}$ iff. $j \neq j'$ are in the same index set $J_k(\boldsymbol{\beta})$. This means if Eq. (35) holds, then for any $\boldsymbol{\beta} \in \mathcal{Y}$, under assumptions (a1) and (a5),

$$\sum_{j \in J_k(\boldsymbol{\beta})} \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad k \in K(\boldsymbol{\beta}). \tag{39}$$

Define $C(\boldsymbol{\beta}) = \min_k |J_k(\boldsymbol{\beta})|$ the minimum cardinality count for j indices in the $J_k(\boldsymbol{\beta})$ subsets.

Now define $D(\mathcal{Y})$ as the set of density points of $\mathcal{Y}$ under Lebesgue measure $\mu(\cdot)$ (with $B(\boldsymbol{\beta}, \epsilon)$ denoting the ℓ_2-norm-based ϵ-ball around a point $\boldsymbol{\beta} \in \mathbb{R}^{d_y}$):

$$D(\mathcal{Y}) := \left\{ \boldsymbol{\beta} \in \mathcal{Y} | \lim_{\epsilon \downarrow 0} \frac{\mu(\mathcal{Y} \cap B(\boldsymbol{\beta}, \epsilon))}{\mu(B(\boldsymbol{\beta}, \epsilon))} = 1 \right\}.$$

Since $\mu(\mathcal{Y}) > 0$, Lebesgue density theorem (see e.g., Mattila, 1995) guarantees that $D(\mathcal{Y}) \neq \emptyset$ and $\mu(D(\mathcal{Y})) = \mu(\mathcal{Y})$. Then choose $\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\beta} \in D(\mathcal{Y})} C(\boldsymbol{\beta})$:

1. We have $C(\boldsymbol{\beta}^*) < |J_0|$ and $|K(\boldsymbol{\beta}^*)| \geq 2$. Otherwise for all $j \neq j' \in J_0$ we have $v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta})$ for all $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{\beta} \in \mathcal{Y}$, so that they are linearly dependent on $(\boldsymbol{z}, \boldsymbol{y}) \in \mathcal{Z} \times \mathcal{Y}$, a contradiction to assumption (a4) by setting $\mathcal{Y}' = D(\mathcal{Y})$.

2. Now assume $|J_1(\boldsymbol{\beta})| = C(\boldsymbol{\beta})$ w.l.o.g. for all $\boldsymbol{\beta} \in \mathcal{Y}$ including $\boldsymbol{\beta}^*$. From assumption (a5), we know that for any $j \in J_1(\boldsymbol{\beta}^*)$ and $j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*)$, the set $\{\boldsymbol{z} \in \mathcal{Z} | v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}^*) = v_{j'}(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}^*)\}$ cannot have nonzero measure. Then as $|J_0| < +\infty$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$ has nonzero measure, there exist $\boldsymbol{z}_0 \in \mathcal{Z}$ and $\delta > 0$ such that

   $$|v_j(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta}^*) - v_{j'}(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta}^*)| \geq \delta, \quad \forall j \in J_1(\boldsymbol{\beta}^*), \forall j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*).$$

   Under assumption (a6), there exists $\epsilon_j > 0$ such that we can construct an ϵ-ball $B(\boldsymbol{\beta}^*, \epsilon_j)$ using ℓ_2-norm, such that

   $$|v_j(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta}^*) - v_j(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta})| \leq \delta/3, \quad \forall \boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon_j).$$

   By definition of $\boldsymbol{\beta}^* \in D(\mathcal{Y})$ as a density point of $\mathcal{Y}$, and since $\mu(\mathcal{Y} \setminus D(\mathcal{Y})) = 0$, there exists $\epsilon' > 0$ such that $\mu(B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})) > 0$ for all $0 < \epsilon < \epsilon'$. Choosing a suitable $0 < \epsilon < \min(\epsilon', \min_{j \in J_0} \epsilon_j)$ (note that $\min_{j \in J_0} \epsilon_j > 0$ as $|J_0| < +\infty$) and constructing an ϵ-ball $B(\boldsymbol{\beta}^*, \epsilon)$, we have for any $j \in J_1(\boldsymbol{\beta}^*)$ and $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$, there exists $k \in K(\boldsymbol{\beta})$ such that $j \in J_k(\boldsymbol{\beta})$ and for all $j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*) \Rightarrow j' \in J_0 \backslash J_k(\boldsymbol{\beta})$ due to

   $$|v_j(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta}) - v_{j'}(\boldsymbol{z} = \boldsymbol{z}_0, \boldsymbol{y} = \boldsymbol{\beta})| \geq \delta/3, \quad \forall \boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y}).$$

So this means $J_k(\boldsymbol{\beta}) \subset J_1(\boldsymbol{\beta}^*)$ and therefore $C(\boldsymbol{\beta}) = |J_1(\boldsymbol{\beta})| \leq |J_k(\boldsymbol{\beta})| \leq |J_1(\boldsymbol{\beta}^*)|$. Now by definition of $\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\beta} \in D(\mathcal{Y})} C(\boldsymbol{\beta})$ and $|J_1(\boldsymbol{\beta}^*)| = C(\boldsymbol{\beta}^*)$, we have $|J_1(\boldsymbol{\beta})| = |J_k(\boldsymbol{\beta})| = |J_1(\boldsymbol{\beta}^*)|$, therefore $J_k(\boldsymbol{\beta}) = J_1(\boldsymbol{\beta}^*)$. Here we can set $k = 1$ w.l.o.g. for all $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$: if $k \neq 1$ we can then swap the indices 1 and k in the index set $K(\boldsymbol{\beta})$. Therefore $J_1(\boldsymbol{\beta}) = J_1(\boldsymbol{\beta}^*)$ for all $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$.

3. One can show that $|J_1(\boldsymbol{\beta}^*)| = 1$, otherwise by definition of the split (Eq. (37)) and the above point, there exists $j \neq j' \in J_1(\boldsymbol{\beta}^*)$ such that $v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta})$ for all $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$, a contradiction to assumption (a4) by setting $\mathcal{Y}' = B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$. Now assume that $j \in J_1(\boldsymbol{\beta}^*)$ is the only index in the subset, then the fact proved in the above point that $J_1(\boldsymbol{\beta}) = J_1(\boldsymbol{\beta}^*)$ for all $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$ means

$$\sum_{i:(i,j)\in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \forall \boldsymbol{\beta} \in B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y}),$$

again a contradiction to assumption (a3) by considering the nonzero measure set $B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y})$.

The above 3 points indicate that Eq. (39) cannot hold for all $\boldsymbol{\beta} \in D(\mathcal{Y})$ (and therefore for all $\boldsymbol{\beta} \in \mathcal{Y}$) under assumptions (a3) - (a6), therefore a contradiction is reached. ■

The proof shows that under (a1)–(a4), linear dependence could in principle occur for every value of the overlapping variable $\mathbf{y}$, which then requires additional assumptions (a5)–(a6).

Why does LIPO-1 not solve the general case $M > 1$? As noted in Section 4.1, LIPO-1 applies directly when $M = 1$ and, more generally, when $T = M + 1$. For given initial and transition families, we set

$$u_i(\boldsymbol{y}, \boldsymbol{x}) := p_a(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M), \quad \boldsymbol{x} := \emptyset, \quad \boldsymbol{y} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M), \quad i = a,$$

and

$$v_j(\boldsymbol{z}, \boldsymbol{y}) := p_b(\boldsymbol{z}_{M+1} \mid \boldsymbol{z}_M, \ldots, \boldsymbol{z}_1), \quad \boldsymbol{z} := \boldsymbol{z}_{M+1}, \quad \boldsymbol{y} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M), \quad j = b,$$

so that the overlapping variable is $\boldsymbol{y} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M)$ and the assumptions (a1)–(a6) are satisfied. For $T > M + 1$, this approach fails because the variable $\boldsymbol{z}$ in one step must become $\boldsymbol{y}$ in the next step to apply LIPO-1. For example, with $T = M + 2$, we require $v'(\boldsymbol{z}', \boldsymbol{y}')$ to satisfy (a4-a6). Therefore we define

$$v'_j(\boldsymbol{z}', \boldsymbol{y}') := p_b(\boldsymbol{z}_{M+2} \mid \boldsymbol{z}_{M+1}, \ldots, \boldsymbol{z}_2), \quad \boldsymbol{y}' := (\boldsymbol{z}_2, \ldots, \boldsymbol{z}_{M+1}), \quad \boldsymbol{z}' := \boldsymbol{z}_{M+2}.$$

However, from the previous step we have $\boldsymbol{z} := \boldsymbol{z}_{M+1}$, and the only $u'_i(\boldsymbol{y}, \boldsymbol{x})$ that preserves (a3) defines

$$u'_i(\boldsymbol{y}', \boldsymbol{x}') := p_a(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M) p_b(\boldsymbol{z}_{M+1} \mid \boldsymbol{z}_M, \ldots, \boldsymbol{z}_1), \quad \boldsymbol{x}' := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M), \quad \boldsymbol{y}' := \boldsymbol{z}_{M+1}.$$

We note this $\boldsymbol{y}'$ is incompatible with the required $(\boldsymbol{z}_2, \ldots, \boldsymbol{z}_{M+1})$ in $v'_j(\boldsymbol{z}', \boldsymbol{y}')$, therefore LIPO-1 cannot be used. To ensure $\boldsymbol{y}' = \boldsymbol{z}$ we can define $v_j(\boldsymbol{z}, \boldsymbol{y})$ as a product over M consecutive transitions:

$$v_j(\boldsymbol{z}, \boldsymbol{y}) := \prod_{\ell=1}^{M} p_{b_\ell}(\boldsymbol{z}_{t+\ell} \mid \boldsymbol{z}_{t-M+\ell:t-1+\ell}),$$
$$\boldsymbol{y} := (\boldsymbol{z}_{t-M+1}, \ldots, \boldsymbol{z}_t),\ \boldsymbol{z} := (\boldsymbol{z}_{t+1}, \ldots, \boldsymbol{z}_{t+M}),\ j := (b_1, \ldots, b_M).$$

With this definition, the overlap variables now satisfy $\boldsymbol{y}' = \boldsymbol{z}$, so the product family has the structure required to apply LIPO-1 sequentially. The remaining task is to verify that this product family still satisfies assumptions (a1)–(a2) and (a4)–(a6).

A.2.2 Extending Assumptions on Product Families

Notation. In the main text and above, we write the coordinate variables in ascending order. However, for better readability, in this section we write variables in descending order, $\boldsymbol{z}_M, \ldots, \boldsymbol{z}_1$ and $\boldsymbol{y}_M, \ldots \boldsymbol{y}_1$, to mirror the product family construction. We also use the following letters to denote distinct notions in our proofs: r denotes the number of consecutive products, ℓ denotes the product factor index, and q denotes the class label in the split (used later in LIPO-M, Lemma 22). We also remove d_y, d_z (used in LIPO-1), and instead we use d to denote the dimensionality of a single variable $\boldsymbol{y}_i, i \in [M]$ or $\boldsymbol{z}$.

Below we explore assumptions on nonparametric families with $M+1$ variables (aligned with some transition distribution $p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, s_t)$, such that the product of M consecutive variables satisfies (a1), (a2), and (a4-a6). This would allow us to use LIPO-1 (or an equivalent extension) to prove linear independence under finite mixtures of the trajectory family $\mathcal{P}^{T,M}_{\mathcal{A},\mathcal{B}}$. We note assumption (a2) is redundant, as it holds for $\mathcal{U}_I$ and $\mathcal{V}_J$ if (a3) and (a4) hold, respectively. Therefore, we remove (a2) from our theoretical analysis for simplicity. Given a family $\mathcal{V}_J = \{v_j(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) | j \in J\}$ with variables $(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1)$ defined on $\times^{M+1}\mathbb{R}^d$, we provide the assumption modifications following the enumeration presented in LIPO-1 (Lemma 17). We keep (a3) unchanged as we will not modify the family $\mathcal{U}_I$.

(a1) → (b1) $v_j(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) > 0$ for all $j \in J, (\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) \in (\times^{M+1}\mathbb{R}^d)$.

(a4) → (b4) There exists a full-measure set $\mathcal{Y} \subset (\times^M \mathbb{R}^d)$ (i.e., $\mu((\times^M \mathbb{R}^d) \setminus \mathcal{Y}) = 0$) such that for all nonzero-measure sets $\mathcal{Y}' \subset \mathcal{Y}$, and $\mathcal{Z} \subset \mathbb{R}^d$, $\mathcal{V}_J$ contains linearly independent functions under finite mixtures on $(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) \in \mathcal{Z} \times \mathcal{Y}'$;

(a5) → (b5) For any $(\boldsymbol{\beta}_\ell, \ldots, \boldsymbol{\beta}_1) \in (\times^\ell \mathbb{R}^d)$, any ℓ with $1 \le \ell \le M$, any nonzero measure subsets $\mathcal{Z} \subset \mathbb{R}^d$, $\mathcal{Y} \subset (\times^{M-\ell}\mathbb{R}^d)$, and any subset $J_0 \subset J$ such that $|J_0| < +\infty$: $\{v_j(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_{\ell+1}, \boldsymbol{y}_\ell = \boldsymbol{\beta}_\ell, \ldots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) | j \in J_0\}$ contains linearly dependent functions on $(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_{\ell+1}) \in \mathcal{Z} \times \mathcal{Y}$ only if $\exists\ j \neq j' \in J_0$ such that $v_j(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_{\ell+1}, \boldsymbol{y}_\ell = \boldsymbol{\beta}_\ell, \ldots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) = v_{j'}(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_{\ell+1}, \boldsymbol{y}_\ell = \boldsymbol{\beta}_\ell, \ldots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1)$ for all $(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_{\ell+1}) \in \times^{M-\ell+1}\mathbb{R}^d$.

(a6) → (b6) For any $j \in J$, $v_j(\boldsymbol{z}, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1)$ is continuous in $(\boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) \in (\times^M \mathbb{R}^d)$.

Contrary to the modifications of (a1) and (a6), we note that the extension of (a4) and (a5) to M consecutive products requires them to be strengthened respectively to (b4) and

(b5). For (b4), the nonzero measure subset $\mathcal{Y}$ must now have full measure. This differs from the case $M = 1$ where $\mathcal{Y}$ can have any nonzero measure, and the intuition behind this requirement can be explained from the appearance of conditioned-conditioned overlaps for $M > 1$ (See Eq. (34)). For (b5), linear dependence only implied by repeating functions for fixed $\boldsymbol{y}$ is strengthened, and now, this must be satisfied for repeating functions for certain fixed subgroups of variables from $\boldsymbol{y}_M, \dots, \boldsymbol{y}_1$.

We provide the following definitions which will be used in the results below.

Definition 18 (Projection onto coordinates) *Let $\mathcal{A} \subseteq (\times^r \mathbb{R}^d)$ and $I = \{i_1, \dots, i_\ell\} \subset [r]$, with $1 \leq \ell \leq r$. The projection of $\mathcal{A}$ onto the coordinates in I is*

$$\pi_I(\mathcal{A}) := \{(\boldsymbol{x}_i)_{i \in I} \in (\times^\ell \mathbb{R}^d) \mid (\boldsymbol{x}_1, \dots, \boldsymbol{x}_r) \in \mathcal{A}\} \subseteq (\times^\ell \mathbb{R}^d).$$

Definition 19 (Conditioned projection) *Let $\mathcal{A} \subseteq (\times^r \mathbb{R}^d)$ with variables $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_r)$. Given $I = \{i_1, \dots, i_\ell\} \subset [r]$ and assignments $(\tilde{\boldsymbol{x}}_{i_1}, \dots, \tilde{\boldsymbol{x}}_{i_\ell})$, define*

$$\mathcal{A}\big|_{(\boldsymbol{x}_{i_1} = \tilde{\boldsymbol{x}}_{i_1}, \dots, \boldsymbol{x}_{i_\ell} = \tilde{\boldsymbol{x}}_{i_\ell})} := \Big\{(\boldsymbol{x}_j)_{j \notin I} \in (\times^{r-\ell} \mathbb{R}^d) : \\ (\boldsymbol{y}_1, \dots, \boldsymbol{y}_r) \in \mathcal{A},\ \boldsymbol{y}_i = \tilde{\boldsymbol{x}}_i\ \forall i \in I,\ \boldsymbol{y}_j = \boldsymbol{x}_j\ \forall j \notin I\Big\}.$$

Assumption (b4) extends to consecutive product function families as follows.

Lemma 20 (Extension of (b4)) *Assume a function family $\mathcal{V}_J = \{v_j(\boldsymbol{z}, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \mid j \in J\}$ defined on $(\boldsymbol{z}, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \in \times^{M+1} \mathbb{R}^d$, such that it satisfies (b1), (b4), (b5), and (b6).*

Then, for any r with $2 \leq r \leq M$, there exists a full-measure set $\mathcal{Y} \subset (\times^M \mathbb{R}^d)$ such that for all nonzero-measure sets $\mathcal{Y}' \subset \mathcal{Y}$, and $\mathcal{Z}_r \subset (\times^r \mathbb{R}^d)$, the product family

$$\otimes^r \mathcal{V}_J := \{v_{j_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_r) \dots v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) | (j_r, \dots, j_1) \in \times^r J\},$$

is linearly independent under finite mixtures on $(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \in \mathcal{Z}_r \times \mathcal{Y}'$.

Proof We prove the above statement by induction, where we start with the base case ($r = 2$) as follows. The proof strategy for the base case largely follows the LIPO-1 proof.

Case $r = 2$. Assume the statement is false. Then, for every set $\mathcal{Y} \subset (\times^M \mathbb{R}^d)$ with $\mu(\mathcal{Y}) = \mu(\times^M \mathbb{R}^d)$, there exist nonzero measure sets $\mathcal{Z}_2 \subset (\mathbb{R}^d \times \mathbb{R}^d)$ and $\mathcal{Y}' \subset \mathcal{Y}$ such that $\mathcal{V}_J \otimes \mathcal{V}_J$ contains linearly dependent functions. We fix $\mathcal{Y}$ from (b4) and consider such a choice of $\mathcal{Y}'$ and $\mathcal{Z}_2$. This means there exists an index set $S_0 \subset (J \times J), |S_0| < +\infty$ and the corresponding nonzero coefficients $\{\gamma_{j_1 j_2} \in \mathbb{R}, (j_1, j_2) \in S_0\}$, such that the following linear dependence condition holds:

$$\sum_{(j_1, j_2) \in S_0} \gamma_{j_1 j_2} v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_2) v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) = 0, \\ \forall (\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \in (\mathcal{Z}_2 \times \mathcal{Y}'), \quad (40)$$

The choices of S_0 and $\gamma_{j_1 j_2}$ values might depend on the choice of $\mathcal{Z}_2$ and $\mathcal{Y}'$, and $\mathcal{Y}'$ depends on the choice of full-measure $\mathcal{Y}$.

From (b1), the set S_0 contains at least two different indices (j_1, j_2) and (j_1', j_2'), with $j_2 \neq j_2'$. To see this, assume we have $j_2 = j_2'$ for all possible $(j_1, j_2), (j_1', j_2') \in S_0$. Then we can factor out $v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_2)$ and rewrite the linear dependence equation as

$$v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_2) \sum_{j_1:(j_1,j_2)\in S_0} \gamma_{j_1 j_2} v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) = 0, \\ \forall (\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) \in (\mathcal{Z}_2 \times \mathcal{Y}'), \quad (41)$$

which contradicts (b4). To see this, since $\mathcal{Z}_2$ has nonzero measure, its projection $\pi_{\{1\}}(\mathcal{Z}_2)$ has nonzero measure. For any $\boldsymbol{z}_1 \in \pi_{\{1\}}(\mathcal{Z}_2)$ there exists $\boldsymbol{z}_2$ such that $(\boldsymbol{z}_2, \boldsymbol{z}_1) \in \mathcal{Z}_2$. Therefore the remaining factor in Eq. (41) contradicts (b4) due to linear dependence in $\pi_{\{1\}}(\mathcal{Z}_2) \times \mathcal{Y}'$.

Now we define $J_0 := \{j_2 \in J | \exists (j_1, j_2) \in S_0\}$, $|J_0| < +\infty$. For any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in (\pi_{\{1\}}(\mathcal{Z}_2) \times \pi_{\{M,\ldots,2\}}(\mathcal{Y}'))$, where π denotes set projections defined in Def. 18, Eq. (40) becomes

$$\sum_{j_2 \in J_0} \left(\sum_{j_1:(j_1,j_2)\in S_0} \gamma_{j_1 j_2} v_{j_1}(\boldsymbol{z}_1 = \boldsymbol{\beta}_M, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1, \boldsymbol{y}_1) \right) \\ v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) = 0, \\ \forall (\boldsymbol{z}_2, \boldsymbol{y}_1) \in (\mathcal{Z}' \times \mathcal{Y}_1'). \quad (42)$$

Here, both $\mathcal{Z}' = \mathcal{Z}_2|_{(\boldsymbol{z}_1 = \boldsymbol{\beta}_M)}$ and $\mathcal{Y}_1' = \mathcal{Y}'|_{(\boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)}$ are conditioned projections (Def. 19) of $\mathcal{Z}_2$ and $\mathcal{Y}'$ respectively and depend on $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$.

Given $\mathcal{Z}_2$ has nonzero measure, there exists a nonzero measure subset $\widetilde{\mathcal{Z}}_1 \subseteq \pi_{\{1\}}(\mathcal{Z}_2)$ such that for any $\boldsymbol{\beta}_M \in \widetilde{\mathcal{Z}}_1$, $\mu\left(\mathcal{Z}_2|_{(\boldsymbol{z}_1 = \boldsymbol{\beta}_M)}\right) > 0$, with $\mu(\cdot)$ denoting the Lebesgue measure. Similarly for $\mathcal{Y}'$, there exists a nonzero measure subset $\widetilde{\mathcal{Y}}_{M:2} \subseteq \pi_{\{M,\ldots,2\}}(\mathcal{Y}')$ such that for any $(\boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{\beta}_1) \in \widetilde{\mathcal{Y}}_{M:2}$, $\mu\left(\mathcal{Y}'|_{(\boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)}\right) > 0$. Otherwise, the conditioned projections would have zero measure for almost every conditioning value, which would imply that $\mathcal{Z}_2$ or $\mathcal{Y}'$ has measure zero. Therefore, we define $\mathcal{D} := \widetilde{\mathcal{Z}}_1 \times \widetilde{\mathcal{Y}}_{M:2}$, which has nonzero measure, and, by construction, for any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{D}$, the conditioned projections $\mathcal{Z}'$ and $\mathcal{Y}_1'$ also have nonzero measure.

Now, for any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{D}$ and similarly as in LIPO-1 (Lemma 17), we split J_0 into subsets indexed by $k \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$:

$$J_0 = \cup_{k \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)} J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1), \; J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \cap J_{k'}(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = \emptyset, \\ \forall k \neq k' \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1), \quad (43)$$

such that two distinct indices $j, j' \in J_0$ belong to the same subset $J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$ if and only if their corresponding indexed functions are equal:

$$j, j' \in J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1), \quad j \neq j' \quad \Leftrightarrow \\ v_j(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) = v_{j'}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1), \quad \forall \boldsymbol{z}_2 \in \mathbb{R}^d. \quad (44)$$

Then we can rewrite the linear dependence condition for any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{D}$ as

$$\sum_{k \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)} \Bigg(\sum_{j_2 \in J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)} \sum_{j_1 : (j_1, j_2) \in S_0} \gamma_{j_1 j_2} v_{j_1}(\boldsymbol{z}_1 = \boldsymbol{\beta}_M, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1, \boldsymbol{y}_1) \\ v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) \Bigg) = 0, \quad \forall (\boldsymbol{z}_2, \boldsymbol{y}_1) \in \mathcal{Z}' \times \mathcal{Y}'_1 \quad (45)$$

From Eq. 44, $v_j(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)$ and $v_{j'}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)$ are the same functions for all $\boldsymbol{z}_2 \in \mathbb{R}^d$ iff. $j \neq j'$ are in the same index set $J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$. Under assumption (b5), representatives from distinct index sets are linearly independent on $\boldsymbol{z}_2 \in \mathcal{Z}'$. Otherwise, given $\mathcal{Z}'$ has nonzero measure for any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{D}$, assumption (b5) would imply equality for all $\boldsymbol{z}_2 \in \mathbb{R}^d$, contradicting Eq. 44. Therefore, for any $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{D}$,

$$\sum_{j_2 \in J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)} \sum_{j_1 : (j_1, j_2) \in S_0} \gamma_{j_1 j_2} v_{j_1}(\boldsymbol{z}_1 = \boldsymbol{\beta}_M, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1, \boldsymbol{y}_1) = 0, \\ \forall \boldsymbol{y}_1 \in \mathcal{Y}'_1, \quad \forall k \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1). \quad (46)$$

Define $C(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = \min_k |J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)|$ as the minimum cardinality count of the subsets $J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$. Again similarly as in LIPO-1 (Lemma 17), let $D(\widetilde{\mathcal{Z}}_1)$ and $D(\widetilde{\mathcal{Y}}_{M:2})$ denote the sets of Lebesgue density points of $\widetilde{\mathcal{Z}}_1$ and $\widetilde{\mathcal{Y}}_{M:2}$, respectively. By the Lebesgue density theorem, both sets have the same nonzero measure as the corresponding original sets. Choose $(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1) \in \arg\min_{(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2})} C(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$:

1. We have $C(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1) < |J_0|$ and $|K(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1)| \geq 2$. Otherwise for all $j \neq j' \in J_0$ we have $v_j(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) = v_{j'}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)$ for all $\boldsymbol{z}_2 \in \mathbb{R}^d$ and $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2})$, so that they are linearly dependent on $(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M, \ldots, \boldsymbol{y}_2) \in \mathcal{Z}' \times \left[\left(D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2}) \right) \cap \mathcal{Y} \right]$ for any nonzero measure $\mathcal{Z}' \subseteq \mathbb{R}^d$, a contradiction to assumption (b4) by setting $\mathcal{Y}' = \left(D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2}) \right) \cap \mathcal{Y}$, which has nonzero measure since $\mathcal{Y}$ has full measure.

2. Now assume $|J_1(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)| = C(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$ w.l.o.g. for all $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2})$, including $(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1)$. From assumption (b5), we know that for any $j \in J_1(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1)$ and $j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1)$, the set

   $$\left\{ \boldsymbol{z}_2 \in \mathcal{Z}_2|_{(\boldsymbol{z}_1 = \boldsymbol{\beta}^*_M)} \mid v_j(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}^*_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}^*_1) = v_{j'}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}^*_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}^*_1) \right\}$$

   cannot have nonzero measure. Then, as $|J_0| < +\infty$ and $\mathcal{Z}_2|_{(\boldsymbol{z}_1 = \boldsymbol{\beta}^*_M)} \subset \mathbb{R}^d$ has nonzero measure, there exist $\boldsymbol{z}_0 \in \mathcal{Z}_2|_{(\boldsymbol{z}_1 = \boldsymbol{\beta}^*_M)}$ and $\delta > 0$ such that

   $$|v_j(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}^*_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}^*_1) - v_{j'}(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}^*_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}^*_1)| \geq \delta, \\ \forall j \in J_1(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1), \forall j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*_M, \ldots, \boldsymbol{\beta}^*_1) \quad (47)$$

Under assumption (b6), there exists $\epsilon_j > 0$ such that we can construct an ϵ-ball $B(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*, \epsilon_j)$ using ℓ_2-norm, such that

$$|v_j(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}_M^*, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1^*) - v_j(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)| \leq \delta/3, \\ \forall (\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in B(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*, \epsilon_j). \quad (48)$$

By definition of $(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*) \in D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2})$ and since the sets of density points have the same measure as their corresponding original sets, there exists $\epsilon' > 0$ such that $\mu\left(B(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*, \epsilon) \cap \left[D(\widetilde{\mathcal{Z}}_1) \times D(\widetilde{\mathcal{Y}}_{M:2})\right]\right) > 0$ for all $0 < \epsilon < \epsilon'$. Choosing a suitable $0 < \epsilon < \min(\epsilon', \min_{j \in J_0} \epsilon_j)$ (note that $\min_{j \in J_0} \epsilon_j > 0$ as $|J_0| < +\infty$), define

$$\mathcal{B} := \left[B(\boldsymbol{\beta}_M^*, \epsilon/2) \cap D(\widetilde{\mathcal{Z}}_1)\right] \times \left[B((\boldsymbol{\beta}_{M-1}^*, \ldots, \boldsymbol{\beta}_1^*), \epsilon/2) \cap D(\widetilde{\mathcal{Y}}_{M:2})\right],$$

where $\mathcal{B}$ has nonzero measure and $\mathcal{B} \subseteq B((\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*), \epsilon) \cap \mathcal{D}$. Then, for any $j \in J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ and $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}$, there exists $k \in K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$ such that $j \in J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$ and for all $j' \in J_0 \backslash J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ we have $j' \notin J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$ due to

$$|v_j(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) - v_{j'}(\boldsymbol{z}_2 = \boldsymbol{z}_0, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)| \geq \delta/3, \\ \forall (\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}. \quad (49)$$

So this means $J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \subset J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ and therefore,

$$C(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = |J_1(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)| \leq |J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)| \leq |J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)|.$$

By definition of $(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ and $|J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)| = C(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$, we have

$$|J_1(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)| = |J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)| = |J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)|,$$

therefore $J_k(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$. Here we can set $k = 1$ wlog for all $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}$: if $k \neq 1$ we can then swap the indices 1 and k in the index set $K(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1)$. Therefore, $J_1(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ for all $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}$.

3. One can show that $|J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)| = 1$, otherwise by definition of the splits and the above point, there exist $j \neq j' \in J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ such that $v_j(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1) = v_{j'}(\boldsymbol{z}_2, \boldsymbol{z}_1 = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)$ for all $\boldsymbol{z}_2 \in \mathbb{R}^d$ and $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}$, a contradiction to assumption (b4) by setting $\mathcal{Y}' := \mathcal{B} \cap \mathcal{Y}$, which has nonzero measure. Now assume that $j_2 \in J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ is the only index in the subset. Then the fact proved in the above point that $J_1(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) = J_1(\boldsymbol{\beta}_M^*, \ldots, \boldsymbol{\beta}_1^*)$ for all $(\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}$ means

$$\sum_{j_1:(j_1,j_2) \in S_0} \gamma_{j_1 j_2} v_{j_1}(\boldsymbol{z}_1 = \boldsymbol{\beta}_M, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1, \boldsymbol{y}_1) = 0, \\ \forall \boldsymbol{y}_1 \in \mathcal{Y}'|_{(\boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)}, \quad \forall (\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}. \quad (50)$$

By construction of $\mathcal{B}$, we may write it as $\mathcal{B} = \mathcal{B}_Z \times \mathcal{B}_Y$, where both $\mathcal{B}_Z$ and $\mathcal{B}_Y$ have nonzero measure. For any $(\boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{\beta}_1) \in \mathcal{B}_Y$, the conditioned projection $\mathcal{Y}'|_{(\boldsymbol{y}_M = \boldsymbol{\beta}_{M-1}, \ldots, \boldsymbol{y}_2 = \boldsymbol{\beta}_1)}$ has nonzero measure by construction of $\widetilde{\mathcal{Y}}_{M:2}$. Therefore, the above equation contradicts (b4) by restricting $\mathcal{Y}'$ to the nonzero measure set $\mathcal{Y}^* := \{(\boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) \in \mathcal{Y}' \mid (\boldsymbol{y}_M, \ldots, \boldsymbol{y}_2) \in \mathcal{B}_Y\}$, which gives linear dependence on $\mathcal{B}_Z \times \mathcal{Y}^*$.

The above 3 points indicate that for the above choice of $\mathcal{Z}_2$ and $\mathcal{Y}$, we find linear dependence of $\mathcal{V}_J$ on a product of nonzero measure sets. This is a direct contradiction of assumption (b4), where linear independence under finite mixtures must hold for any nonzero measure sets $\mathcal{Z}' \subseteq \mathbb{R}^d$ and $\mathcal{Y}' \subseteq \mathcal{Y}$, with $\mathcal{Y}$ having full measure.

We note the assumption of $\mathcal{Y}$ having full measure in (b4), as opposed to (a4) is crucial here. Otherwise, the resulting set $\mathcal{B}$ could lie outside of the choice of $\mathcal{Y}$, as it is constructed from projections of $\mathcal{Z}_2 \subset (\mathbb{R}^d \times \mathbb{R}^d)$ and $\mathcal{Y}' \subset \mathcal{Y}$. Therefore, we would have $\mu(\mathcal{B} \cap \mathcal{Y}) = 0$ and the above contradiction would not hold.

Case $r > 2$. Now assume the statement holds for $r-1$, and again prove the case by contradiction. Assume the statement is false for $r > 2$. Then, for every set $\mathcal{Y} \subset (\times^M \mathbb{R}^d)$ with $\mu(\mathcal{Y}) = \mu(\times^M \mathbb{R}^d)$, there exists nonzero measure sets $\mathcal{Z}_r \subset (\times^r \mathbb{R}^d)$ and $\mathcal{Y}' \subset \mathcal{Y}$ such that the family $\otimes^r \mathcal{V}_J$ contains linearly dependent functions. As before, we fix $\mathcal{Y}$ from (b4) and consider such a choice of $\mathcal{Y}'$ and $\mathcal{Z}_r$. Then, there exists $S_0 \subset (\times^r J)$, $|S_0| < +\infty$ such that

$$\sum_{(j_r,\dots j_1)\in S_0} \gamma_{j_r,\dots,j_1} v_{j_r}(\boldsymbol{z}_r,\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_r)\dots v_{j_1}(\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) = 0,$$
$$\forall(\boldsymbol{z}_r,\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) \in (\mathcal{Z}_r \times \mathcal{Y}'), \quad (51)$$

with $\{\gamma_{j_r,\dots,j_1} \in \mathbb{R}, (j_r,\dots,j_1) \in S_0\}$ a set of nonzero coefficients which, as before, might depend on the choice of $\mathcal{Z}_r$ and $\mathcal{Y}' \subset \mathcal{Y}$. The previous equality can be arranged as follows

$$\sum_{(j_r,\dots j_1)\in S_0} \gamma_{j_r,\dots,j_1} v_{j_r}(\boldsymbol{z}_r,\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_r) v_{j_{r-1},\dots,j_1}(\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) = 0,$$
$$\forall(\boldsymbol{z}_r,\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) \in (\mathcal{Z}_r \times \mathcal{Y}'), \quad (52)$$

where $v_{j_{r-1},\dots,j_1} := v_{j_{r-1}}(\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_{r-1})\dots v_{j_1}(\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1)$ for $j_i \in J, 1 \leq i < r$ denotes the functions in the family $\otimes^{r-1}\mathcal{V}_J$ and satisfies the following:

(b1) Positive function values: $v_{j_{r-1},\dots,j_1}(\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) > 0$ for all $(j_{r-1},\dots,j_1) \in (\times^{r-1} J), (\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) \in (\times^{M+r-1}\mathbb{R}^d)$;

and from the induction hypothesis:

(b4*) There exists a full-measure set $\mathcal{Y} \subset (\times^M \mathbb{R}^d)$ such that for every nonzero-measure set $\mathcal{Y}' \subset \mathcal{Y}$, and $\mathcal{Z}_{r-1} \subset (\times^{r-1}\mathbb{R}^d)$, the family $\otimes^{r-1}\mathcal{V}_J$ contains linearly independent functions under finite mixtures in $(\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_1) \in \mathcal{Z}_{r-1} \times \mathcal{Y}'$.

The strategy here is to reduce this case to the above base case ($r = 2$), where importantly we need to show:

(1) S_0 contains at least two $(j_r,\dots,j_1)$ and $(j'_r,\dots,j'_1)$ with $j_r \neq j'_r$; and

(2) the overlapping variables between the product of families, $(\boldsymbol{z}_{r-1},\dots,\boldsymbol{z}_1,\boldsymbol{y}_M,\dots,\boldsymbol{y}_r) \in (\pi_{\{r-1,\dots,1\}}(\mathcal{Z}_r) \times \pi_{\{M,\dots,r\}}(\mathcal{Y}'))$ do not cause linear dependence in nonzero measure sets for any choice of $\mathcal{Z}_r$ and $\mathcal{Y}' \subset \mathcal{Y}$.

We can simply see (1) holds from (b1) and (b4*) by following the same logic from Eq. (41), where the linear dependence relation contradicts (b4*) if we have $j_r = j'_r$. Therefore, we define $J_0 := \{j_r \in J | \exists (j_r, \dots, j_1) \in S_0\}$, $|J_0| < +\infty$. Then, linear dependence can occur for any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in (\pi_{\{r-1,\dots,1\}}(\mathcal{Z}_r) \times \pi_{\{M,\dots,r\}}(\mathcal{Y}')) \subset (\times^M \mathbb{R}^d)$, where π denotes set projections defined in Def. 18.

$$\sum_{j_r \in J_0} v_{j_r}(\boldsymbol{z}_r, \boldsymbol{z}_{r-1} = \boldsymbol{\beta}_M, \dots, \boldsymbol{z}_1 = \boldsymbol{\beta}_{M-r+2}, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-r+1} \dots, \boldsymbol{y}_r = \boldsymbol{\beta}_1)$$
$$\Bigg(\sum_{(j_{r-1},\dots,j_1):(j_r,\dots,j_1) \in S_0} \gamma_{j_r,\dots,j_1} v_{j_{r-1},\dots,j_1}(\boldsymbol{z}_{r-1} = \boldsymbol{\beta}_M, \dots, \boldsymbol{z}_1 = \boldsymbol{\beta}_{M-r+2}, \boldsymbol{y}_M = \boldsymbol{\beta}_{M-r+1},$$
$$\dots, \boldsymbol{y}_r = \boldsymbol{\beta}_1, \boldsymbol{y}_{r-1}, \dots, \boldsymbol{y}_1) \Bigg) = 0,$$
$$\forall (\boldsymbol{z}_r, \boldsymbol{y}_{r-1}, \dots, \boldsymbol{y}_1) \in \Big(\mathcal{Z}_r|_{(\boldsymbol{z}_{r-1}=\boldsymbol{\beta}_M,\dots,\boldsymbol{z}_1=\boldsymbol{\beta}_{M-r+2})} \times \mathcal{Y}'|_{(\boldsymbol{y}_M=\boldsymbol{\beta}_{M-r+1},\dots,\boldsymbol{y}_r=\boldsymbol{\beta}_1)} \Big), \quad (53)$$

where both $\mathcal{Z}_r|_{(\boldsymbol{z}_{r-1}=\boldsymbol{\beta}_M,\dots,\boldsymbol{z}_1=\boldsymbol{\beta}_{M-r+2})}$ and $\mathcal{Y}'|_{(\boldsymbol{y}_M=\boldsymbol{\beta}_{M-r+1},\dots,\boldsymbol{y}_r=\boldsymbol{\beta}_1)}$ are conditioned projections (Def. 19) from $\mathcal{Z}_r$ and $\mathcal{Y}'$ respectively and are dependent on $\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1$. As in the base case, we can restrict the projections to nonzero measure subsets such that the corresponding conditioned projections have nonzero measure. That is, there exist nonzero measure subsets $\widetilde{\mathcal{Z}}_{r-1:1} \subseteq \pi_{\{r-1,\dots,1\}}(\mathcal{Z}_r)$ and $\widetilde{\mathcal{Y}}_{M:r} \subseteq \pi_{\{M,\dots,r\}}(\mathcal{Y}')$, and we define $\mathcal{D} := \widetilde{\mathcal{Z}}_{r-1:1} \times \widetilde{\mathcal{Y}}_{M:r}$ such that the conditioned projections have nonzero measure for any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in \mathcal{D}$. Then, we can establish the following equivalence between variables and sets (given the choice of $\mathcal{Z}_r$ and $\mathcal{Y}' \subset \mathcal{Y}$) with respect to the base case ($r = 2$).

- $\boldsymbol{z}_r$ is equivalent to $\boldsymbol{z}_2$, and therefore $\mathcal{Z}_r|_{(\boldsymbol{z}_{r-1}=\boldsymbol{\beta}_M,\dots,\boldsymbol{z}_1=\boldsymbol{\beta}_{M-r+2})}$ is equivalent to $\mathcal{Z}_2|_{(\boldsymbol{z}_1=\boldsymbol{\beta}_M)}$.
- $\boldsymbol{z}_{r-1}, \dots, \boldsymbol{z}_1$ are equivalent to $\boldsymbol{z}_1$, and therefore $\pi_{\{r-1,\dots,1\}}(\mathcal{Z}_r)$ is equivalent to $\pi_{\{1\}}(\mathcal{Z}_2)$.
- $\boldsymbol{y}_M, \dots, \boldsymbol{y}_r$ are equivalent to $\boldsymbol{y}_M, \dots, \boldsymbol{y}_2$, and therefore $\pi_{\{M,\dots,r\}}(\mathcal{Y}')$ is equivalent to $\pi_{\{M,\dots,2\}}(\mathcal{Y}')$.
- $\boldsymbol{y}_{r-1}, \dots, \boldsymbol{y}_1$ are equivalent to $\boldsymbol{y}_1$, and therefore $\mathcal{Y}'|_{(\boldsymbol{y}_M=\boldsymbol{\beta}_{M-r+1},\dots,\boldsymbol{y}_r=\boldsymbol{\beta}_1)}$ is equivalent to $\mathcal{Y}'|_{(\boldsymbol{y}_M=\boldsymbol{\beta}_{M-1},\dots,\boldsymbol{y}_2=\boldsymbol{\beta}_1)}$.

Given the above equivalences, the remaining arguments follow directly from the base case $r = 2$. For any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in \mathcal{D}$ we split J_0 into subsets indexed by $K(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1)$ such that the functions in each subset are equal, similarly as in Eqs. (43)–(44). Defining $C(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) := \min_k |J_k(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1)|$ as in the case $r = 2$, we choose

$$(\boldsymbol{\beta}_M^*, \dots, \boldsymbol{\beta}_1^*) \in \arg \min_{(\boldsymbol{\beta}_M,\dots,\boldsymbol{\beta}_1) \in D(\widetilde{\mathcal{Z}}_{r-1:1}) \times D(\widetilde{\mathcal{Y}}_{M:r})} C(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1).$$

Then, assumptions (b4)–(b6), and the same density-point argument shown in the base case establish linear dependence of the elements $v_{j_{r-1},\dots,j_1}(\cdot)$ on a product of nonzero measure sets, analogously to Eq. (50).

The linear dependence of $\otimes^{r-1}\mathcal{V}_J$ on a product of nonzero measure sets contradicts the induction hypothesis (b4*), as $\mathcal{Y}$ has full measure. Hence, Eq. (52) cannot hold for such a

choice of $\mathcal{Y}'$ and $\mathcal{Z}_r$, and a contradiction is reached. ■

Regarding (b5), we observe that linear dependence on M products of functions given fixed subsets of variables does not imply repeating functions anymore. Instead, we show that at least one component of the product family is linearly independent under finite mixtures.

Lemma 21 (Extension of (b5)) *Assume a function family $\mathcal{V}_J = \{v_j(\boldsymbol{z}, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \mid j \in J\}$ with variables $\boldsymbol{z}, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1$ defined on $\times^{M+1}\mathbb{R}^d$, such that it satisfies (b1), (b4), (b5), (b6).*

Then, for any r with $2 \leq r \leq M$, any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in (\times^M \mathbb{R}^d)$, any nonzero measure $\mathcal{Z}_r \subset (\times^r \mathbb{R}^d)$, and $S_0 \subset (\times^r J)$ such that $|S_0| < +\infty$ and the family

$$\Big\{ v_{j_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_r = \boldsymbol{\beta}_r) \dots \\ v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) | (j_r, \dots, j_1) \in S_0 \Big\}$$

contains linearly dependent functions only if $\exists \ell \leq r$ such that

$$\Big\{ v_{j_\ell}(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_\ell = \boldsymbol{\beta}_\ell) | j_\ell \in J_\ell \Big\}, \\ \text{with } J_\ell := \big\{ j_\ell \in J | \exists (j_r, \dots, j_1) \in S_0 \big\}$$

contains linearly dependent functions on $(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1) \in \pi_{\{\ell, \dots, 1\}}(\mathcal{Z}_r)$.

Proof We prove the statement by induction starting from the base case ($r = 2$) as follows, where we utilise the contraposition of LIPO-1 (Lemma 17) on failure of assumptions (a3) or (a4) to establish the result. The result establishes linear independence of product domains, but we will use linear independence on arbitrary nonzero measure joint sets. However, the contraposition also applies, since such a set contains a nonzero measure set of shared variable assignments where the corresponding conditioned projection sets (Def. 19) have nonzero measure, allowing the same contraposition argument.

Case $r = 2$. Fix $\mathcal{Z}_2 \subset \mathbb{R}^d \times \mathbb{R}^d$, and assume that for any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in (\times^M \mathbb{R}^d)$, and $S_0 \subset J \times J$, the family

$$\Big\{ v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_2 = \boldsymbol{\beta}_2) v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) | (j_2, j_1) \in S_0 \Big\}$$

contains linearly dependent functions. From (b1, b4-b6), we can connect these assumptions to the ones defined in Lemma 17 by setting $u_i(\boldsymbol{y} = \boldsymbol{z}_1, \boldsymbol{x} = \emptyset) = v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1)$, $i = j_1$, and $v_j(\boldsymbol{z} = \boldsymbol{z}_2, \boldsymbol{y} = \boldsymbol{z}_1) = v_{j_2}(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_2 = \boldsymbol{\beta}_2)$, $j = j_2$, where (a1), (a5) and (a6) are satisfied. Therefore, by contraposition of Lemma 17, since (a1), (a5), and (a6) are satisfied while the product family contains linearly dependent families, we know that either one of the families contains linearly dependent functions (from the failure of either (a3) or (a4) in Lemma 17).

Case $r > 2$. Fix $\mathcal{Z}_r \subset (\times^r \mathbb{R}^d)$, and assume that for any $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_1) \in (\times^M \mathbb{R}^d)$, and $S_0 \subset (\times^r J)$, the family

$$\Big\{ v_{j_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_r = \boldsymbol{\beta}_r) \dots \\ v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) | (j_r, \dots, j_1) \in S_0 \Big\}$$

contains linearly dependent functions. As before, from (b1, b4-b6) and by setting

$$u_i(\boldsymbol{y} = (\boldsymbol{z}_{r-1}, \dots, \boldsymbol{z}_1), \boldsymbol{x} = \emptyset) = v_{j_{r-1}}(\boldsymbol{z}_{r-1}, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_{r-1} = \boldsymbol{\beta}_{r-1}) \dots \\ v_{j_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1), \quad i = (j_{r-1}, \dots, j_1),$$

and $v_j(\boldsymbol{z} = \boldsymbol{z}_r, \boldsymbol{y} = (\boldsymbol{z}_{r-1}, \dots, \boldsymbol{z}_1)) = v_{j_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_r = \boldsymbol{\beta}_r), \quad j = j_r$, we can establish correspondence to assumptions on Lemma 17. We know (a1) and (a6) are satisfied directly from (b1) and (b6). (a5) is satisfied by setting $\ell = M + 1 - r$ in (b5). Then, by contraposition of Lemma 17, since the product contains linearly dependent functions, either (a3) or (a4) must fail. This implies either one of the families must contain linearly dependent functions. Therefore, either

$$\Big\{v_{j_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_r = \boldsymbol{\beta}_r) | j_r \in J_r\Big\}, \text{ with } J_r = \big\{j_r \in J | \exists (j_r, \dots, j_1) \in S_0\big\}$$

contains linearly dependent functions on some nonzero measure subset of $\mathcal{Z}_r$, or the product family of the first $r-1$ factors contains linearly dependent functions on some nonzero measure subset $\widetilde{\mathcal{Z}}_{r-1} \subset \pi_{\{r-1,\dots,1\}}(\mathcal{Z}_r)$. From the induction hypothesis there is some $\ell \leq r-1$ such that

$$\Big\{v_{j_\ell}(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \dots, \boldsymbol{y}_\ell = \boldsymbol{\beta}_\ell) | j_\ell \in J_\ell\Big\}, \\ \text{with } J_\ell = \big\{j_\ell \in J | \exists (j_r, \dots, j_1) \in S_0\big\}$$

contains linearly dependent functions on $(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1) \in \pi_{\{\ell,\dots,1\}}(\widetilde{\mathcal{Z}}_{r-1})$.

Finally, by (b5), linear dependence on any nonzero measure subset implies repeating functions on the whole domain of the variables $(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1)$. Therefore, in the case $\ell = r$, the corresponding component family is linearly dependent on $\mathcal{Z}_r$, and for $\ell \leq r-1$, it is linearly dependent on $\pi_{\{\ell,\dots,1\}}(\mathcal{Z}_r)$.

■

A.2.3 Linear Independence on Products with M Overlapping Variables

Given that assumption (b5) does not extend similarly as (b4), we generalise LIPO-1 to LIPO-M (Linear Independence on Product functions with M consecutive Overlapping variables) using Lemma 21 to show linear independence under finite mixtures when $\mathcal{V}_J$ is defined as a product of M functions with consecutive variables.

Notation: We use the following block notation $\boldsymbol{z}_{\ell:1}$, $\boldsymbol{\beta}_{M:\ell}$, $\boldsymbol{y}_{M:\ell}$ to denote tuples of variables $(\boldsymbol{z}_\ell, \dots, \boldsymbol{z}_1)$, $(\boldsymbol{\beta}_M, \dots, \boldsymbol{\beta}_\ell)$, $(\boldsymbol{y}_M, \dots, \boldsymbol{y}_\ell)$, respectively.

Lemma 22 (LIPO-M) *Assume two families $\mathcal{U}_I = \{u_i(\boldsymbol{y}, \boldsymbol{x}) | i \in I\}$ and $\mathcal{V}_J = \{v_j(\boldsymbol{z}, \boldsymbol{y}) | j \in J\}$, with $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathbb{R}^{d_y}$, and $\boldsymbol{z} \in \mathbb{R}^{d_z}$, with $d_y = d \cdot M$, $d_z = d \cdot r$, and $1 \leq r \leq M$. We further assume:*

(a) The family $\mathcal{U}_I$ satisfies assumptions (a1) and (a3).

(b) Each element in the family $\mathcal{V}_J$ is defined as a product of r functions, all of which belong to the same family $\mathcal{W}_K = \{w_k(\boldsymbol{z}, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) | k \in K\}$, with $\boldsymbol{z}_1 \in \mathbb{R}^d$, $(\boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \in (\times^M \mathbb{R}^d)$, and is expressed as follows

$$\mathcal{V}_J := \otimes^r \mathcal{W}_K = \{w_{k_r}(\boldsymbol{z}_r, \dots, \boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_r) \dots w_{k_1}(\boldsymbol{z}_1, \boldsymbol{y}_M, \dots, \boldsymbol{y}_1) \mid (k_r, \dots, k_1) \in (\times^r K)\}$$

where $J := (\times^r K)$, and $\boldsymbol{z} := (\boldsymbol{z}_r, \dots, \boldsymbol{z}_1)$, $\boldsymbol{y} := (\boldsymbol{y}_M, \dots, \boldsymbol{y}_1)$. The family $\mathcal{W}_K$ satisfies assumptions (b1), (b4), (b5), and (b6).

Then for any nonzero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$, $\mathcal{U}_I \otimes \mathcal{V}_J := \{u_i(\boldsymbol{y}, \boldsymbol{x}) v_j(\boldsymbol{z}, \boldsymbol{y}) | i \in I, j \in J\}$ contains linearly independent functions under finite mixtures defined on $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}$.

Proof Assume this sufficiency statement is false, then there exist a nonzero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$, $S_0 \subset I \times J$ with $|S_0| < +\infty$ and a set of nonzero values $\{\gamma_{ij} \in \mathbb{R} | (i,j) \in S_0\}$, such that

$$\sum_{(i,j) \in S_0} \gamma_{ij} v_j(\boldsymbol{z}, \boldsymbol{y}) u_i(\boldsymbol{y}, \boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}. \tag{54}$$

For convenience, we denote the elements of J as indices despite being a tuple. Note that the choices of S_0 and γ_{ij} are independent of any $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ values, but might be dependent on $\mathcal{Z}$. We wish to follow the logic presented in LIPO-1 (Lemma 17) under (a1-a6). Below we establish some connections. First, we know that the family $\mathcal{V}_J$ as defined in (b) is a product of r elements with $1 \leq r \leq M$, all of which satisfy (b1) and (b4-b5). Lemma 20 shows that there exists some full-measure $\mathcal{Y}$, such that for any subset $\mathcal{Z} \times \mathcal{Y}' \subset \mathbb{R}^{d_z} \times \mathcal{Y}$, the family $\mathcal{V}_J$ is linearly independent under finite mixtures. This is stronger than (a4) in Lemma 17 because (b4) provides a full-measure $\mathcal{Y}$, while (a4) only requires a nonzero measure $\mathcal{Y}$. Furthermore, it also satisfies (a1) as it is defined as a product of r functions that satisfy (b1), with $1 \leq r \leq M$. Therefore, we follow the same arguments as in Lemma 17, which show that the index set S_0 contains at least 2 different indices (i,j) and (i,j') with $j \neq j'$. Moreover, define $J_0 = \{j \in J | \exists (i,j) \in S_0\}$ the set of all possible j indices that appear in S_0, where $|J_0| < +\infty$. Rewrite Eq. (54) as

$$\sum_{j \in J_0} \left(\sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) \right) v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) = 0, \qquad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}.$$

For fixed $j \in J_0$, assumption (a3) implies that $\{\boldsymbol{\beta} \in \mathbb{R}^{d_y} | \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{X}\}$ has measure zero. Write $\mathcal{Y} \subset \mathbb{R}^{d_y}$ the full-measure set from (b4). For almost every $\boldsymbol{\beta} \in \mathcal{Y}$, the above equation implies that $\{v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) | j \in J_0\}$ contains linearly dependent functions on $\boldsymbol{z} \in \mathcal{Z}$.

We will now write J, S_0, and J_0 in tuple format, as we need to expand the product family $\mathcal{V}_J$. Therefore, we now regard $j = (k_r, \dots, k_1)$, and re-define $J_0 = \{(k_r, \dots, k_1) \in J | \exists (i, k_r, \dots, k_1) \in S_0\}$. From (b4-b6), Lemma 21 shows that, for almost every $\boldsymbol{\beta} \in \mathcal{Y}$, linear dependence of the product of r functions ($1 \leq r \leq M$) that composes the family $\mathcal{V}_J$ implies linearly dependent functions in at least one of the components. Since $\ell \in \{1, \dots, r\}$

takes finitely many values, there exists some $\ell \leq r$ and a nonzero measure subset $\mathcal{Y}' \subseteq \mathcal{Y}$ such that $\{w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) \mid k_\ell \in K_0^{(\ell)}\}$, with $\boldsymbol{y}_{M:\ell} := (\boldsymbol{y}_M, \ldots, \boldsymbol{y}_\ell)$, $\boldsymbol{\beta}_{M:\ell} := (\boldsymbol{\beta}_M, \ldots, \boldsymbol{\beta}_\ell)$, and $K_0^{(\ell)} := \{k_\ell \in K | \exists (i, k_r, \ldots, k_1) \in S_0\}$, contains linearly dependent functions on $(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1) \in \pi_{\{\ell,\ldots,1\}}(\mathcal{Z})$ for every $\boldsymbol{\beta} \in \mathcal{Y}'$.

Under assumption (b5), for each $\boldsymbol{\beta} \in \mathcal{Y}'$ we can split the index set $K_0^{(\ell)}$ into subsets indexed by $q \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ as follows, such that within each index subset $K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ the functions with the corresponding indices are equal:

$$\begin{aligned} K_0^{(\ell)} = \cup_{q \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell})} K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell}), \quad K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) \cap K_{q'}^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = \emptyset, \forall q \neq q' \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell}), \\ k_\ell \neq k'_\ell \in K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) \quad \Leftrightarrow \quad w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) = \\ w_{k'_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}), \forall (\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1) \in \pi_{\{\ell,\ldots,1\}}(\mathcal{Z}). \end{aligned} \tag{55}$$

For any $\boldsymbol{\beta} \in \mathcal{Y}'$, analogously define $K_0^{(m)} := \{k_m \in K \mid \exists (i, k_r, \ldots, k_1) \in S_0\}$ for each $m \in \{1, \ldots, r\}$, and split each $K_0^{(m)}$ similarly as in Eq. (55), according to equality of the function components $w_{k_m}(\boldsymbol{z}_m, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:m} = \boldsymbol{\beta}_{M:m})$. Then we can rewrite Eq. (54) as

$$\begin{aligned} \sum_{(q_r,\ldots,q_1) \in \times_{m=1}^r L^{(m)}(\boldsymbol{\beta}_{M:m})} \Bigg(\sum_{(k_r,\ldots,k_1) \in \times_{m=1}^r K_{q_m}^{(m)}(\boldsymbol{\beta}_{M:m})} \sum_{i:(i,k_r,\ldots,k_1) \in S_0} \gamma_{i,k_r,\ldots,k_1} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) \\ w_{k_r}(\boldsymbol{z}_r, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M) \ldots w_{k_1}(\boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M, \ldots, \boldsymbol{y}_1 = \boldsymbol{\beta}_1) \Bigg) = 0, \\ \forall (\boldsymbol{x}, \boldsymbol{z}_r, \ldots, \boldsymbol{z}_1) \in \mathcal{X} \times \mathcal{Z}. \end{aligned} \tag{56}$$

Note that we use $\gamma_{i,k_r,\ldots,k_1} = \gamma_{ij}$ for any $(i, j) \in S_0$ and $(k_r, \ldots, k_1) \in J_0$, such that $j = (k_r, \ldots, k_1)$. We now wish to factor out the common $w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})$ within each $K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$, analogously to LIPO-1 (Lemma 17). This cannot be done directly since the remaining factors have overlapping variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell$. For each $m \in \{1, \ldots, r\}$, (b5) implies that representatives from distinct subsets $K_{q_m}^{(m)}(\boldsymbol{\beta}_{M:m})$ are linearly independent in each component. Therefore, by contraposition of Lemma 21, the corresponding product family is linearly independent, which implies that in the above equation, the coefficients associated to each tuple $(q_1, \ldots, q_r)$ are zero. For any fixed $q \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ and by summing the zero coefficients over all tuples $(q_1, \ldots, q_r)$ satisfying $q_\ell = q$, we can factor out $w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})$ from the above equation analogously to LIPO-1:

$$\begin{aligned} \sum_{k_\ell \in K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})} \sum_{(i,k_r,\ldots,k_{\ell+1},k_{\ell-1},\ldots,k_1):(i,k_r,\ldots,k_1) \in S_0} \gamma_{i,k_r,\ldots,k_1} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) w_{k_r}(\boldsymbol{z}_r, \ldots, \boldsymbol{z}_1, \\ \boldsymbol{y}_M = \boldsymbol{\beta}_M) \ldots w_{k_{\ell+1}}(\boldsymbol{z}_{\ell+1}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell+1} = \boldsymbol{\beta}_{M:\ell+1}) w_{k_{\ell-1}}(\boldsymbol{z}_{\ell-1}, \ldots, \boldsymbol{z}_1, \\ \boldsymbol{y}_{M:\ell-1} = \boldsymbol{\beta}_{M:\ell-1}) \ldots w_{k_1}(\boldsymbol{z}_1, \boldsymbol{y}_{M:1} = \boldsymbol{\beta}_{M:1}) = 0, \\ \forall (\boldsymbol{x}, \boldsymbol{z}_r, \ldots, \boldsymbol{z}_1) \in \mathcal{X} \times \mathcal{Z}, \ q \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell}). \end{aligned} \tag{57}$$

Define $C^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = \min_q |K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})|$ the minimum cardinality count for k_ℓ indices in the $K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ subsets. We follow analogous density point arguments from LIPO-1 (Lemma 17).

Therefore, let $D(\mathcal{Y}')$ denote the set of density points of $\mathcal{Y}'$ under Lebesgue measure $\mu(\cdot)$. Since $\mu(\mathcal{Y}') > 0$, the Lebesgue density theorem guarantees that $\mu(D(\mathcal{Y}')) = \mu(\mathcal{Y}')$. We then choose $\boldsymbol{\beta}^*_{M:\ell} \in \arg\min_{\boldsymbol{\beta}_{M:\ell} \in \pi_{\{M,\dots,\ell\}} D(\mathcal{Y}')} C^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$, where there exists some $\boldsymbol{\beta}^* \in D(\mathcal{Y}')$ where the corresponding projection onto $\{M,\dots,\ell\}$ is $\boldsymbol{\beta}^*_{M:\ell}$. We fix any such $\boldsymbol{\beta}^*$:

1. We have $C^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell}) < |K_0^{(\ell)}|$ and $|L^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})| \geq 2$. Otherwise, for all $k_\ell \neq k'_\ell \in K_0^{(\ell)}$ we have $w_{k_\ell}(\boldsymbol{z}_\ell,\dots,\boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) = w_{k'_\ell}(\boldsymbol{z}_\ell,\dots,\boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})$ for all $(\boldsymbol{z}_\ell,\dots,\boldsymbol{z}_1) \in \pi_{\{\ell,\dots,1\}}(\mathcal{Z})$ and $\boldsymbol{\beta}_{M:\ell} \in \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$, so that they are linearly dependent on $(\boldsymbol{z}_\ell,\dots,\boldsymbol{z}_1,\boldsymbol{y}_{M:\ell}) \in \pi_{\{\ell,\dots,1\}}(\mathcal{Z}) \times \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$, a contradiction to assumption (b4). Note that $\pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$ has nonzero measure since $D(\mathcal{Y}')$ also has nonzero measure.

2. Now assume $|K_1^{(\ell)}(\boldsymbol{\beta}_{M:\ell})| = C^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ w.l.o.g. for all $\boldsymbol{\beta}_{M:\ell} \in \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$, including $\boldsymbol{\beta}^*_{M:\ell}$. From assumption (b5), we know that for any $k_\ell \in K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ and $k'_\ell \in K_0^{(\ell)} \backslash K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$, the set $\{\boldsymbol{z}_{\ell:1} \in \pi_{\{\ell,\dots,1\}}(\mathcal{Z}) \mid w_{k_\ell}(\boldsymbol{z}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}^*_{M:\ell}) = w_{k'_\ell}(\boldsymbol{z}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}^*_{M:\ell})\}$ cannot have nonzero measure. Then as $|K_0^{(\ell)}| < +\infty$ and $\pi_{\{\ell,\dots,1\}}(\mathcal{Z}) \subset (\times^\ell \mathbb{R}^d)$ has nonzero measure, there exist $(\boldsymbol{z}^{(0)}_{\ell:1}) \in \pi_{\{\ell,\dots,1\}}(\mathcal{Z})$ and $\delta > 0$ such that

$$\left| w_{k_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}^*_{M:\ell}) - w_{k'_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}^*_{M:\ell}) \right| \geq \delta, \\ \forall k_\ell \in K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell}), \quad \forall k'_\ell \in K_0^{(\ell)} \backslash K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell}). \quad (58)$$

Under assumption (b6), for each $k_\ell \in K_0^{(\ell)}$ there exists $\epsilon_{k_\ell} > 0$ such that

$$|w_{k_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}^*_{M:\ell}) - w_{k_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})| \leq \delta/3, \\ \forall \boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon_{k_\ell}). \quad (59)$$

By definition of $\boldsymbol{\beta}^* \in D(\mathcal{Y}')$ as a density point of $\mathcal{Y}'$, and since $\mu(\mathcal{Y}' \setminus D(\mathcal{Y}')) = 0$, there exists $\epsilon_0 > 0$ such that $\mu\left(B(\boldsymbol{\beta}^*, \epsilon) \cap D(\mathcal{Y}')\right) > 0$ for all sufficiently small $0 < \epsilon < \epsilon_0$. Choosing $0 < \epsilon'_\ell < \min(\epsilon_0, \min_{k_\ell \in K_0^{(\ell)}} \epsilon_{k_\ell})$ (note that $\min_{k_\ell \in K_0^{(\ell)}} \epsilon_{k_\ell} > 0$ as $|K_0^{(\ell)}| < +\infty$), we have for any $k_\ell \in K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ and $\boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \cap \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$, there exists $q \in L^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ such that $k_\ell \in K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ and, for all $k'_\ell \in K_0^{(\ell)} \backslash K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$, we have $k'_\ell \in K_0^{(\ell)} \backslash K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$, due to

$$\left| w_{k_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) - w_{k'_\ell}(\boldsymbol{z}_{\ell:1} = \boldsymbol{z}^{(0)}_{\ell:1}, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) \right| \geq \delta/3, \\ \forall \boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \cap \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}')). \quad (60)$$

Therefore $K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) \subseteq K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$, and $C^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = |K_1^{(\ell)}(\boldsymbol{\beta}_{M:\ell})| \leq |K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})| \leq |K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})|$. Now by definition of $\boldsymbol{\beta}^*_{M:\ell} \in \arg\min_{\boldsymbol{\beta}_{M:\ell} \in \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))} C^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$ and $|K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})| = C^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$, we have $|K_1^{(\ell)}(\boldsymbol{\beta}_{M:\ell})| = |K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell})| = |K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})|$, therefore $K_q^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$. Here we can set $q = 1$ w.l.o.g. for all $\boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \cap \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$: if $q \neq 1$, we can swap the indices 1 and q in the index set $L^{(\ell)}(\boldsymbol{\beta}_{M:\ell})$. Therefore, $K_1^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ for all $\boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \cap \pi_{\{M,\dots,\ell\}}(D(\mathcal{Y}'))$.

3. One can show that $|K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})| = 1$, otherwise by definition of the split (Eq. (55)) and the above point, there exists $k_\ell \neq k'_\ell \in K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ such that $w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell}) = w_{k'_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})$ for all $(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1) \in \pi_{\{\ell,\ldots,1\}}(\mathcal{Z})$ and $\boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \cap \pi_{\{M,\ldots,\ell\}}(D(\mathcal{Y}'))$, again a contradiction to assumption (b4).

Given the above 3 points and assuming $k_\ell \in K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ is the only index in the subset, then the fact proved in the above point that $K_1^{(\ell)}(\boldsymbol{\beta}_{M:\ell}) = K_1^{(\ell)}(\boldsymbol{\beta}^*_{M:\ell})$ means, for any $\boldsymbol{\beta} \in D(\mathcal{Y}')$ such that $\boldsymbol{\beta}_{M:\ell} \in B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell)$:

$$\sum_{(i,k_r,\ldots,k_{\ell+1},k_{\ell-1},\ldots,k_1):(i,k_r,\ldots,k_1)\in S_0} \gamma_{i,k_r,\ldots,k_1} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) w_{k_r}(\boldsymbol{z}_r, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M) \cdot$$
$$\ldots \cdot w_{k_{\ell+1}}(\boldsymbol{z}_{\ell+1}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell+1} = \boldsymbol{\beta}_{M:\ell+1}) w_{k_{\ell-1}}(\boldsymbol{z}_{\ell-1}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell-1} = \boldsymbol{\beta}_{M:\ell-1}) \cdot$$
$$\ldots \cdot w_{k_1}(\boldsymbol{z}_1, \boldsymbol{y}_{M:1} = \boldsymbol{\beta}_{M:1}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}_r, \ldots, \boldsymbol{z}_1) \in \mathcal{X} \times \mathcal{Z}; \quad (61)$$

where the term $w_{k_\ell}(\boldsymbol{z}_\ell, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell} = \boldsymbol{\beta}_{M:\ell})$ can be factored out from the above equation thanks to (b1). Given that earlier we established linear dependence of $\{v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) | j \in J_0\}$ for each $\boldsymbol{\beta} \in \mathcal{Y}'$, we define $\mathcal{Y}_\ell := \big(B(\boldsymbol{\beta}^*_{M:\ell}, \epsilon'_\ell) \times \pi_{\{\ell-1,\ldots,1\}}(D(\mathcal{Y}'))\big) \cap D(\mathcal{Y}')$. Note that $\mu(\mathcal{Y}_\ell) > 0$, since $B(\boldsymbol{\beta}^*, \epsilon'_\ell) \cap D(\mathcal{Y}') \subseteq \mathcal{Y}_\ell$. Then, from Eq. (61), for each $\boldsymbol{\beta}_{M:1} \in \mathcal{Y}_\ell$ the family

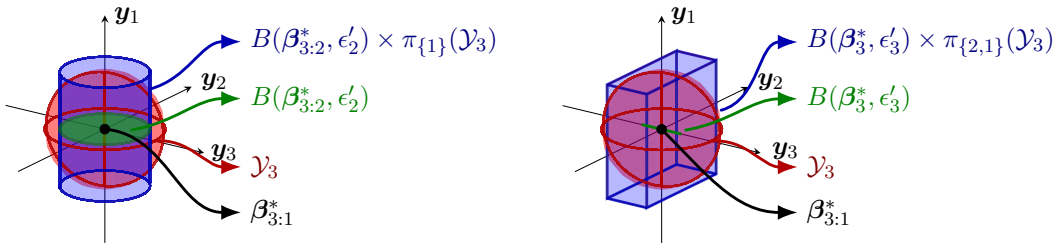
$$\Big\{ w_{k_r}(\boldsymbol{z}_r, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_M = \boldsymbol{\beta}_M) \ldots w_{k_{\ell+1}}(\boldsymbol{z}_{\ell+1}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell+1} = \boldsymbol{\beta}_{M:\ell+1}) \cdot$$
$$w_{k_{\ell-1}}(\boldsymbol{z}_{\ell-1}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell-1} = \boldsymbol{\beta}_{M:\ell-1}) \ldots w_{k_1}(\boldsymbol{z}_1, \boldsymbol{y}_{M:1} = \boldsymbol{\beta}_{M:1}) |,$$
$$(k_r, \ldots, k_{\ell+1}, k_{\ell-1}, \ldots, k_1) \in \{k_r, \ldots, k_{\ell+1}, k_{\ell-1}, \ldots, k_1) | \exists (i, k_r, \ldots, k_1) \in S_0\} \Big\}$$

contains linearly dependent functions. The above family is a product of $r-1$ functions, all of which belong to the same family $\mathcal{W}_K$ which satisfies (b1), (b4), (b5) and (b6). Therefore, we can use Lemma 21 again. Then, we know for each $\boldsymbol{\beta}_{M:1} \in \mathcal{Y}_\ell$, there is at least one $\ell' \leq r$ with $\ell' \neq \ell$ such that $\{w_{k_{\ell'}}(\boldsymbol{z}_{\ell'}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell'} = \boldsymbol{\beta}_{M:\ell'}) | k_{\ell'} \in K_0^{(\ell')}\}$, with $K_0^{(\ell')} = \{k_{\ell'} \in K | \exists (i, k_r, \ldots, k_1) \in S_0\}$ with the previously selected k_ℓ fixed, contains linear dependent functions under finite mixtures. As before, since $\ell' \neq \ell$ takes finitely many values, there exists a nonzero-measure subset $\mathcal{Y}'_\ell \subseteq \mathcal{Y}_\ell$ for which the same $\ell' \leq r$ satisfies the above for any $\boldsymbol{\beta} \in \mathcal{Y}'_\ell$.

As before, we create splits as in Eq. (55), choose

$$\boldsymbol{\beta}^*_{M:\ell'} \in \arg \min_{\boldsymbol{\beta}_{M:\ell'} \in \pi_{\{M,\ldots,\ell'\}}(D(\mathcal{Y}'_\ell))} C^{(\ell')}(\boldsymbol{\beta}_{M:\ell'}),$$

where there exists some $\boldsymbol{\beta}^* \in D(\mathcal{Y}'_\ell)$ whose corresponding projection onto $\{M, \ldots, \ell'\}$ is $\boldsymbol{\beta}^*_{M:\ell'}$, and we fix any such $\boldsymbol{\beta}^*$. We then construct $B(\boldsymbol{\beta}^*_{M:\ell'}, \epsilon'_{\ell'}) \cap \pi_{\{M,\ldots,\ell'\}}(D(\mathcal{Y}'_\ell))$ with $\epsilon'_\ell > \epsilon'_{\ell'} > 0$, such that by (b4), (b6) and Eq. (55), the corresponding split satisfies $|K_1^{(\ell')}(\boldsymbol{\beta}_{M:\ell'})| = 1$ for all $\boldsymbol{\beta}_{M:\ell'} \in B(\boldsymbol{\beta}^*_{M:\ell'}, \epsilon'_{\ell'}) \cap \pi_{\{M,\ldots,\ell'\}}(D(\mathcal{Y}'_\ell))$ (we again assume $|K_1^{(\ell')}(\boldsymbol{\beta}_{M:\ell'})| = C^{(\ell')}(\boldsymbol{\beta}_{M:\ell'})$ w.l.o.g.). Therefore we can factor out $w_{k_{\ell'}}(\boldsymbol{z}_{\ell'}, \ldots, \boldsymbol{z}_1, \boldsymbol{y}_{M:\ell'} = \boldsymbol{\beta}_{M:\ell'})$ corresponding to $K_1^{(\ell')}(\boldsymbol{\beta}_{M:\ell'})$ from Eq. (61) for any $\boldsymbol{\beta} \in D(\mathcal{Y}'_\ell)$ such that $\boldsymbol{\beta}_{M:\ell'} \in B(\boldsymbol{\beta}^*_{M:\ell'}, \epsilon'_{\ell'})$.



(a) $\ell' = 2$ (b) $\ell' = 3$

Figure 13: Illustrations of nonzero intersection of subsets to construct $\mathcal{Y}_{\ell'}$ with $M = 3,\ \ell = 1$.

Now we construct a new nonzero set of admissible $\boldsymbol{\beta}$-values such that we can factor out the ℓ'-th factor of the product family $\{v_j(\cdot)\}$. We also need to ensure that the new set $\mathcal{Y}_{\ell'}$ is a subset of $\mathcal{Y}_\ell$, so that all previously factored terms remain valid. Define

$$\mathcal{Y}_{\ell'} := D(\mathcal{Y}'_\ell) \cap \Big(B(\boldsymbol{\beta}^*_{M:\ell'}, \epsilon'_{\ell'}) \times \pi_{\{\ell'-1,\dots,1\}}(D(\mathcal{Y}'_\ell)) \Big), \tag{62}$$

where, by construction, $\mathcal{Y}_{\ell'} \subseteq D(\mathcal{Y}'_\ell) \subseteq \mathcal{Y}'_\ell \subseteq \mathcal{Y}_\ell$. The set also has nonzero measure, since $B(\boldsymbol{\beta}^*, \epsilon'_{\ell'}) \cap D(\mathcal{Y}'_\ell) \subseteq \mathcal{Y}_{\ell'}$ and $\boldsymbol{\beta}^* \in D(\mathcal{Y}'_\ell)$. Figure 13 illustrates this for $M = 3$ and $\ell = 1$, where the red sphere depicts the admissible set. We observe that the product between $B(\boldsymbol{\beta}^*_{M:\ell'}, \epsilon'_{\ell'})$ and the projection $\pi_{\{\ell'-1,\dots,1\}}(D(\mathcal{Y}'_\ell))$ results in a cylinder (or a cube if $\ell' = 1$), and their intersection with $D(\mathcal{Y}'_\ell)$ is nonzero measure. For illustrative purposes only, this example assumes $\boldsymbol{\beta}_{M:\ell}$ across $\ell \in [M]$ are aligned: i.e., for $\ell \neq \ell'$ and $\boldsymbol{\beta}^{*(\ell)}_{M:\ell}$, $\boldsymbol{\beta}^{*(\ell')}_{M:\ell'}$, we have $\boldsymbol{\beta}^{*(\ell)}_{M:\ell'} = \boldsymbol{\beta}^{*(\ell')}_{M:\ell'}$. However, the nonzero measure intersection does not rely on alignment: it already follows from the density-point construction above.

Therefore, for each $\boldsymbol{\beta} \in \mathcal{Y}_{\ell'}$, the family $\{v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) | j \in J_0\}$ is still linearly dependent under finite mixtures, where two terms in the product v_j have only one element (w_{k_ℓ} and $w_{k_{\ell'}}$). By repeating this procedure over all remaining indices, we obtain a nested sequence: $\mathcal{Y}_{final} \subseteq \cdots \subseteq \mathcal{Y}_{\ell'} \subseteq \mathcal{Y}_\ell$ where each constructed set has nonzero measure and, for each $\boldsymbol{\beta} \in \mathcal{Y}_{final}$, every factor in $\{v_j(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{\beta}) | j \in J_0\}$ has been reduced to a single element. Then, Eq. (61) reduces to

$$\sum_{i:(i,k_r,\dots,k_1)\in S_0} \gamma_{i,k_r,\dots,k_1} u_i(\boldsymbol{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{\beta}) \in \mathcal{X} \times \mathcal{Y}_{final}. \tag{63}$$

$\mathcal{Y}_{final}$ is a nonzero measure set on which the family $\{u_i(\cdot)\}$ is linearly dependent, contradicting (a3). Therefore, a contradiction is reached. ■

A.2.4 Linear Independence of Nonparametric Trajectory Families

In LIPO-M, we state results for generic per-variable dimension d, so that $\boldsymbol{y} \in \mathbb{R}^{dM}$ and $\boldsymbol{z} \in \mathbb{R}^{dr}$ with $1 \leq r \leq M$. In the nonparametric trajectory setting we set $d = m$ because $\boldsymbol{z}_t \in \mathbb{R}^m$. Below we present linear independence under finite mixtures for nonparametric trajectory families.

Theorem 23 *Define the following trajectory family for a trajectory* $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \mathbb{R}^{Tm}$*:*

$$\mathcal{P}_{\mathcal{A},\mathcal{B}}^{T,M} = \Bigg\{ p_{a,b_{M+1:T}}(\boldsymbol{z}_{1:T}) = p_a(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b_t}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}), \\ p_a \in \Pi_{\mathcal{A}}^M, \quad p_{b_t} \in \mathcal{P}_{\mathcal{B}}^M, \quad t = M+1, ..., T \Bigg\}, \quad (64)$$

and assume $\Pi_{\mathcal{A}}^M$ *and* $\mathcal{P}_{\mathcal{B}}^M$ *satisfy the following assumptions,*

(c1) $\Pi_{\mathcal{A}}^M$ *satisfies (a1), (a3), and*

(c2) $\mathcal{P}_{\mathcal{B}}^M$ *satisfies (b4-b6).*

Then the following statement holds: For any $T > M$ *such that* $r = M$ *if* $(T \mod M) = 0$*, or* $r = (T \mod M)$*, otherwise; and any subset* $\mathcal{Z}_r \subset (\times^r \mathbb{R}^m)$*, the trajectory family contains linearly independent distributions under finite mixtures for* $(\boldsymbol{z}_{1:T-r}, \boldsymbol{z}_{T-r+1:T}) \in \mathbb{R}^{(T-r)m} \times \mathcal{Z}_r$*.*

Proof We first prove the statement for any $T > M$ such that $T \mod M = 0$ by induction as follows.

$\underline{T = 2M}$: The result can be proved using Lemma 22 by setting in the proof,

$$u_i(\boldsymbol{y} = \boldsymbol{z}_{1:M}, \boldsymbol{x} = \boldsymbol{z}_0) = p_a(\boldsymbol{z}_{1:M}), i = a,$$

and

$$v_j(\boldsymbol{z} = \boldsymbol{z}_{M+1:2M}, \boldsymbol{y} = \boldsymbol{z}_{1:M}) = \prod_{\ell=1}^{M} p_{b_{M+\ell}}(\boldsymbol{z}_{M+\ell} | \boldsymbol{z}_{\ell:M-1+\ell}), j = b_{M+1:2M},$$

where j is a tuple index, and $\boldsymbol{z}_{\ell:M-1+\ell} := (\boldsymbol{z}_\ell, \boldsymbol{z}_{\ell+1}, \ldots, \boldsymbol{z}_{M-1+\ell})$ is a sliding window of length M. We observe that the Lemma holds using assumptions (c1-c2) directly as $v_j(\boldsymbol{z} = \boldsymbol{z}_{M+1:2M}, \boldsymbol{y} = \boldsymbol{z}_{1:M})$ is a product of M functions that satisfies (b) in LIPO-M (Lemma 22).

$\underline{T = rM, r \in \mathbb{Z}^+, r > 2}$: Assume the statement holds for the trajectory family when $T = \tau - M$. Note that we can write $p_{a,b_{M+1:\tau}}(\boldsymbol{z}_{1:\tau})$ as

$$p_{a,b_{M+1:\tau}}(\boldsymbol{z}_{1:\tau}) = p_{a,b_{M+1:\tau-M}}(\boldsymbol{z}_{1:\tau-M}) \prod_{\ell=1}^{M} p_{b_{\tau-M+\ell}}(\boldsymbol{z}_{\tau-M+\ell} | \boldsymbol{z}_{\tau-2M+\ell:\tau-M-1+\ell}). \quad (65)$$

Then the statement for $T = \tau$ can be proved using LIPO-M (Lemma 22) by setting

$$u_i(\boldsymbol{y} = \boldsymbol{z}_{\tau-2M+1:\tau-M}, \boldsymbol{x} = \boldsymbol{z}_{1:\tau-2M}) = p_{a,b_{M+1:\tau-M}}(\boldsymbol{z}_{1:\tau-M}), i = a, b_{M+1:\tau-M},$$

and

$$v_j(\boldsymbol{z} = \boldsymbol{z}_{\tau-M+1:\tau}, \boldsymbol{y} = \boldsymbol{z}_{\tau-2M+1:\tau-M}) = \\ \prod_{\ell=1}^{M} p_{b_{\tau-M+\ell}}(\boldsymbol{z}_{\tau-M+\ell} | \boldsymbol{z}_{\tau-2M+\ell:\tau-M-1+\ell}), j = b_{\tau-M+1:\tau}. \quad (66)$$

Note that both i and j denote tuple indices. The family spanned with $p_{a,b_{M+1:\tau-M}}(\boldsymbol{z}_{1:\tau-M}), i = a, b_{M+1:\tau-M}$ satisfies (a1) from (c1), and from the induction hypothesis it is linearly independent under finite mixtures, which is exactly (a3). The family in Eq. (66) is a product of M functions where (b1) and (b4-b6) are satisfied from (c2), which imply (b) in Lemma 22.

Given the above result we proceed to prove the case where $T \mod M \neq 0$.

$T \mod M \neq 0$: Let r be the remainder of T/M: $T \mod M = r$. From the previous result we know the statement holds for $T - r$, i.e., when T is a multiple of M. We can write $p_{a,b_{M+1:T}}(\boldsymbol{z}_{1:T})$ as follows

$$p_{a,b_{M+1:T}}(\boldsymbol{z}_{1:T}) = p_{a,b_{M+1:T-r}}(\boldsymbol{z}_{1:T-r}) \prod_{\ell=1}^{r} p_{b_{T-r+\ell}}(\boldsymbol{z}_{T-r+\ell} | \boldsymbol{z}_{T-r+\ell-M:T-r+\ell-1}). \tag{67}$$

We can again prove the statement with Lemma 22 by setting

$$u_i(\boldsymbol{y} = \boldsymbol{z}_{T-r-M+1:T-r}, \boldsymbol{x} = \boldsymbol{z}_{1:T-r-M}) = p_{a,b_{M+1:T-r}}(\boldsymbol{z}_{1:T-r}), i = a, b_{M+1:T-r},$$

and

$$v_j(\boldsymbol{z} = \boldsymbol{z}_{T-r+1:T}, \boldsymbol{y} = \boldsymbol{z}_{T-r-M+1:T-r}) = \prod_{\ell=1}^{r} p_{b_{T-r+\ell}}(\boldsymbol{z}_{T-r+\ell} | \boldsymbol{z}_{T-r+\ell-M:T-r+\ell-1}), j = b_{T-r+1:T}. \tag{68}$$

The family spanned with $p_{a,b_{M+1:T-r}}(\boldsymbol{z}_{1:T-r}), i = a, b_{M+1:T-r}$ satisfies (a1) from (c1). (a3) is satisfied as follows:

- For $M < T < 2M$, we have $T - r = M$ and (a3) holds directly from (c1).
- For $T > 2M$, we have $(T-r) \mod M = 0$ and (a3) holds from the induction hypothesis at $T - r$.

The family in Eq. (68) is a product of r function families which satisfy (b1) and (b4-b6) from (c2). This implies (b) in Lemma 22.

We showed that the conditions in LIPO-M are satisfied for all possible cases of T, and therefore by LIPO-M, the resulting trajectory family is linearly independent under finite mixtures. ■

A.3 Linear Independence Under Finite Mixtures with Assumptions (m1) and (m2)

Recall assumptions (m1) and (m2):

(m1) Unique indexing for $\mathcal{G}_{\mathcal{B}}^{M}$ and $\mathcal{I}_{\mathcal{A}}^{M}$: Eqs. (15) and (17) hold;

(m2) The mean and covariance in $\mathcal{G}_{\mathcal{B}}^{M}$, $\boldsymbol{m}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m}$ and $\boldsymbol{\Sigma}(\cdot, b) : \mathbb{R}^{mM} \to \mathbb{R}^{m \times m}$, are analytic functions, for any $b \in \mathcal{B}$.

Below we explore whether these parametric conditions ensure linear independence under finite mixtures of the trajectory family (and thus, identifiability). The unique indexing assumption of the Gaussian transition family (Eq. (15)) implies linear independence as shown below.

Proposition 24 *Functions in $\mathcal{G}_{\mathcal{B}}$ are linearly independent under finite mixtures on variables $(\boldsymbol{z}_t, \ldots, \boldsymbol{z}_{t-M})$ if the unique indexing assumption (Eq. (15)) holds.*

Proof Assume the statement is false and w.l.o.g. $M = 1$. Then there exists $\mathcal{B}_0 \subset \mathcal{B}$ with $|\mathcal{B}_0| < +\infty$ and a set of nonzero values $\{\gamma_b | b \in \mathcal{B}_0\}$, such that

$$\sum_{b \in \mathcal{B}_0} \gamma_b \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}(\boldsymbol{z}_{t-1}, b), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, b)) = 0, \quad \forall \boldsymbol{z}_t, \boldsymbol{z}_{t-1} \in \mathbb{R}^m.$$

In particular, this equality holds for any $\boldsymbol{z}_{t-1} \in \mathbb{R}^m$, meaning that a weighted sum of Gaussian distributions (defined on $\boldsymbol{z}_t$) equals to zero. Note that Yakowitz and Spragins (1968) proved that multivariate Gaussian distributions with different means and/or covariances are linearly independent. Therefore the equality above implies for any $\boldsymbol{z}_{t-1}$

$$\boldsymbol{m}(\boldsymbol{z}_{t-1}, b) = \boldsymbol{m}(\boldsymbol{z}_{t-1}, b') \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, b) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, b') \quad \forall b, b' \in \mathcal{B}_0, b \neq b',$$

a contradiction to the unique indexing assumption. ■

The transition and initial Gaussian distribution families defined in Eqs. (14) and (16) respectively align with assumptions (a) and (b) in LIPO-M as follows.

Proposition 25 *The conditional Gaussian distribution family $\mathcal{G}_{\mathcal{B}}$ (Eq. (14)), under the unique indexing assumption (Eq. (15)), satisfies assumptions (b1), (b5), and (b6) in Lemmas 20, 21, and 22, if we define $\mathcal{V}_J := \mathcal{G}_{\mathcal{B}}, \boldsymbol{z} := \boldsymbol{z}_t$ and $(\boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) := (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$.*

Proposition 26 *The initial Gaussian distribution family $\mathcal{I}_{\mathcal{A}}$ (Eq. (16)), under the unique indexing assumption (Eq. (17)), satisfies assumptions (a1), (a3) in Lemma 22, if we define $\mathcal{U}_I := \mathcal{I}_{\mathcal{A}}, \boldsymbol{y} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M)$ and $\boldsymbol{x} = \emptyset$.*

To see why (b5) holds under the Gaussian transition family $\mathcal{G}_{\mathcal{B}}$, we note from Prop. 24 that linear dependence occurs only if the unique indexing assumption does not hold. Therefore, we can fix a subset of the $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$ variables, such that the resulting mean and covariance functions violate the unique indexing assumption, which would imply linear dependence on the resulting function family. To verify (b4), we require the following zero-measure intersection of moments result.

Proposition 27 *Assume Gaussian family transitions $\mathcal{G}_{\mathcal{B}}$ under unique indexing defined by Eq. (15), with zero-measure intersection of moments: For any $b, b' \in \mathcal{B}$ with $b \neq b'$, the set*

$$\mathcal{X}_{b,b'} := \Big\{ (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}) \in \mathbb{R}^{mM} \mid \boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) = \boldsymbol{m}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b'), \\ \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M}, b') \Big\}$$

has zero measure, and (b4) holds if $\mathcal{V}_J := \mathcal{G}_{\mathcal{B}}$, $\boldsymbol{z} := \boldsymbol{z}_t$, and $(\boldsymbol{y}_M, \ldots, \boldsymbol{y}_1) := (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$.

Proof We prove the statement by contradiction. Define $\mathcal{V}_J := \mathcal{G}_\mathcal{B}$, $\boldsymbol{z} := \boldsymbol{z}_t$, and $(\boldsymbol{y}_M, \dots, \boldsymbol{y}_1) := (\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M})$ from (b4); and w.l.o.g. set $\mathcal{Y} := \mathbb{R}^{mM}$. If (b4) does not hold, then there exist nonzero measure sets $\mathcal{Y}' \subset \mathcal{Y}$ and $\mathcal{Z} \subset \mathbb{R}^m$ such that $\mathcal{G}_\mathcal{B}$ contains linearly dependent functions. From Proposition 24, this means there exists an index set $\mathcal{B}_0 \subset \mathcal{B}$ with $|\mathcal{B}_0| < +\infty$ such that the unique indexing assumption (Eq. (15)) does not hold in $\mathcal{Y}'$. From the intersection of moments assumption, we know that this set is $\mathcal{Y}' = \bigcup_{b \neq b' \in \mathcal{B}_0} \mathcal{X}_{b,b'}$, which has zero measure. This contradicts the requirement for a nonzero measure set $\mathcal{Y}' \subset \mathcal{Y}$ in (b4). ■

Finally, we present linear independence under finite mixtures of the trajectory family in the parametric case.

Theorem 28 *Define the following trajectory family under the nonlinear Gaussian model*

$$\left\{ p_{a,b_{M+1:T}}(\boldsymbol{z}_{1:T}) = p_a(\boldsymbol{z}_{1:M}) \prod_{t=M+1}^{T} p_{b_t}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}), \right.$$
$$\left. p_a \in \mathcal{I}_\mathcal{A}^M, \quad p_{b_t} \in \mathcal{G}_\mathcal{B}^M, \quad t = M+1, ..., T \right\}, \quad (69)$$

with $\mathcal{G}_\mathcal{B}^M$, $\mathcal{I}_\mathcal{A}^M$ defined by Eqs. (14), (16) respectively. Assume:

(d1) Unique indexing for $\mathcal{G}_\mathcal{B}^M$ and $\mathcal{I}_\mathcal{A}^M$: Eqs. (15), (17) hold;

(d2) The functions in $\mathcal{G}_\mathcal{B}$ are continuous with respect to $(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}) \in \mathbb{R}^{mM}$;

(d3) Zero-measure intersection of moments: For any $b, b' \in \mathcal{B}$ with $b \neq b'$, the set

$$\mathcal{X}_{b,b'} := \left\{ (\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}) \in \mathbb{R}^{mM} \mid \boldsymbol{m}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, b) = \boldsymbol{m}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, b'), \right.$$
$$\left. \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, b) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, b') \right\}$$

has zero measure.

Then, the trajectory family contains linearly independent distributions under finite mixtures for $(\boldsymbol{z}_{1:T-r}, \boldsymbol{z}_{T-r+1}, \dots, \boldsymbol{z}_T) \in \mathbb{R}^{(T-r)m} \times (\times^r \mathbb{R}^m)$, where $r = M$ if $T \mod M = 0$, or $r = T \mod M$ otherwise.

Proof Note that (d2) and (b6) are equivalent. Assumptions (a1), (a3), (b1), and (b5) are satisfied due to Propositions 25, 26. Assumption (b4) holds due to assumption (d3) via Prop 27. Then, the statement holds by Theorem 23. ■

A.4 Conclusion

Below we conclude the proof of Theorem 8.

Proof We first clarify the alignment from assumptions (m1-m2) to (d1-d3) in Theorem 28. (d1) and (m1) are equivalent. (d2) is satisfied from (m2) as analytic functions are $\mathcal{C}^\infty$, and therefore continuous. From Mityagin (2015), we know the zero set of an analytic function is measure zero. Therefore, (m2) implies that the intersections of the means or covariances for any $b, b' \in \mathcal{B}$, with $b \neq b'$ are also measure zero, which implies (d3). Therefore, from Theorem 28, the trajectory family under the Gaussian transition family $\mathcal{G}_\mathcal{B}$ is linearly independent under finite mixtures. By Theorem 16, linear independence under finite mixtures of the trajectory family implies identifiability of the MSM in terms of Def. 7. ■

Appendix B. Proof of SDS Identifiability

We extend the results from Kivva et al. (2022) and Balsells-Rodas et al. (2024) with our multi-lag MSM family as a prior distribution for $\boldsymbol{z}_{1:T}$. The strategy requires finding some open set where the transformations F and G from two identically distributed SDSs are invertible. Then, identifiability can be shown thanks to piecewise linearity of F and G, and the multi-lag MSM having an analytic density and being closed under affine transformations.

B.1 Preliminaries

We need to introduce some definitions and results that will be used in the proof. These have been previously defined in Kivva et al. (2022).

Definition 29 *Let $D_0 \subseteq D \subseteq \mathbb{R}^n$ be open sets. Let $f_0 : D_0 \to \mathbb{R}$. We say that an analytic function $f : D \to \mathbb{R}$ is an analytic continuation of f_0 onto D if $f(\boldsymbol{x}) = f_0(\boldsymbol{x})$ for every $\boldsymbol{x} \in D_0$.*

Definition 30 *Let $\boldsymbol{x}_0 \in \mathbb{R}^m$ and $\delta > 0$. Let $p : B(\boldsymbol{x}_0, \delta) \to \mathbb{R}$. Define*

$$Ext(p) : \mathbb{R}^m \to \mathbb{R}$$

to be the unique analytic continuation of p on the entire space $\mathbb{R}^m$ if such a continuation exists, and to be 0 otherwise.

Definition 31 *Let $D_0 \subset D$ and $p : D \to \mathbb{R}$ be a function. We define $p|_{D_0} : D_0 \to \mathbb{R}$ to be a restriction of p to D_0, namely a function that satisfies $p|_{D_0}(\boldsymbol{x}) = p(\boldsymbol{x})$ for every $\boldsymbol{x} \in D_0$.*

Definition 32 *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a piecewise affine function. We say that a point $\boldsymbol{x} \in f(\mathbb{R}^m) \subseteq \mathbb{R}^n$ is generic with respect to f if the pre-image $f^{-1}(\{\boldsymbol{x}\})$ is finite and there exists $\delta > 0$, such that $f : B(\boldsymbol{z}, \delta) \to \mathbb{R}^n$ is affine for every $\boldsymbol{z} \in f^{-1}(\{\boldsymbol{x}\})$.*

We also need to show that the points in the pre-image of a piecewise factored mapping F can be computed using the multi-lag MSM prior under assumptions (m1-m2), which is then used to show that there exists a nonzero measure set where both F and G are

invertible. We follow Lemma D.6 from Balsells-Rodas et al. (2024), where we observe that including additional lags does not alter the statement. We also adapt Corollary D.7 and Theorem D.8 from Balsells-Rodas et al. (2024) to multi-lag MSMs, where the proofs require minor clarifications.

Lemma 33 (adapted from Balsells-Rodas et al., 2024) *Consider a random variable $\boldsymbol{z}_{1:T}$ distributed according to the multi-lag MSM family $\mathcal{M}_{NL}^{T,M}$ with assumptions (m1-m2). Let us consider $f : \mathbb{R}^m \to \mathbb{R}^m$, a piecewise affine mapping which generates the random variable $\boldsymbol{x}_{1:T} = F(\boldsymbol{z}_{1:T})$ as $\boldsymbol{x}_t = f(\boldsymbol{z}_t), 1 \leq t \leq T$. Also, consider $\boldsymbol{x}^{(0)} \in \mathbb{R}^m$ a generic point with respect to f. Then, $\boldsymbol{x}_{1:T}^{(0)} = \{\boldsymbol{x}^{(0)}, \dots, \boldsymbol{x}^{(0)}\} \in \mathbb{R}^{Tm}$ is also a generic point with respect to F and the number of points in the pre-image $F^{-1}(\{\boldsymbol{x}_{1:T}^{(0)}\})$ can be computed as*

$$\left|F^{-1}\left(\left\{\boldsymbol{x}_{1:T}^{(0)}\right\}\right)\right| = \lim_{\delta \to 0} \int_{\boldsymbol{x}_{1:T} \in \mathbb{R}^{Tm}} Ext\left(p|_{B(\boldsymbol{x}_{1:T}^{(0)}, \delta)}\right) d\boldsymbol{x}_{1:T}$$

Proof If $\boldsymbol{x}^{(0)} \in \mathbb{R}^m$ is a generic point with respect to f, $\boldsymbol{x}_{1:T}^{(0)}$ is also a generic point with respect to F since the pre-image is $F^{-1}(\{\boldsymbol{x}_{1:T}^{(0)}\})$ now larger but still finite. In other words, $F^{-1}(\{\boldsymbol{x}_{1:T}^{(0)}\})$ is the Cartesian product $\times^T \mathcal{Z}$, where $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_\ell\}$ are the points in the pre-image $f^{-1}(\{\boldsymbol{x}^{(0)}\})$. Considering this, we have well defined affine mappings $G_{i_1,\dots,i_T} : B(\boldsymbol{z}_{i_1}, \dots, \boldsymbol{z}_{i_T}, \epsilon) \to \mathbb{R}^{Tm}$, $i_t \in [\ell]$ for $1 \leq t \leq T$, such that $G_{i_1,\dots,i_T} = F(\boldsymbol{z}_{1:T}), \forall \boldsymbol{z}_{1:T} \in B(\boldsymbol{z}_{i_1}, \dots, \boldsymbol{z}_{i_T}, \epsilon)$. This affine mapping $G_{i_1,\dots,i_T}$ is factored: $g_{i_t}(\boldsymbol{z}_t) = f(\boldsymbol{z}_t),\ \forall \boldsymbol{z}_t \in B(\boldsymbol{z}_i, \epsilon)$,

$$G_{i_1,\dots,i_T} = \begin{pmatrix} A_{i_1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & A_{i_T} \end{pmatrix} \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_T \end{pmatrix} + \begin{pmatrix} b_{i_1} \\ \vdots \\ b_{i_T} \end{pmatrix},$$

with invertible $A_{i_1}, \dots, A_{i_T}$. Let $\delta_0 > 0$ such that

$$B(\boldsymbol{x}_{1:T}^{(0)}, \delta_0) \subseteq \bigcap_{i_1,\dots,i_T}^{\ell} G_{i_1,\dots,i_T}\big(B(\boldsymbol{z}_{i_1}, \dots, \boldsymbol{z}_{i_T}, \epsilon)\big).$$

We can compute the likelihood for every $\boldsymbol{x}_{1:T} \in B(\boldsymbol{x}_{1:T}^{(0)}, \delta)$ with $0 < \delta < \delta_0$ using Prop. C.2 from Balsells-Rodas et al. (2024) where the MSM is closed under factored affine transformations. We can assume $M = 1$ w.l.o.g., since for lag $M > 1$, the MSM is still closed under factored affine transformations,

$$p|_{B(\boldsymbol{x}_{1:T}^{(0)}, \delta)} = \sum_{i_1,\dots,i_T}^{\ell} \sum_{j_1,\dots,j_T}^{K} p\left(s_{1:T} = \{j_1, \dots, j_T\}\right) \mathcal{N}\left(\boldsymbol{x}_1; A_{i_1}\boldsymbol{\mu}(j_1) + \boldsymbol{b}_{i_1}, A_{i_1}\boldsymbol{\Sigma}_1(j_1)A_{i_1}^T\right)$$
$$\prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{x}_t; A_{i_t}\boldsymbol{m}\left(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \boldsymbol{b}_{i_{t-1}}\right), j_t\right) + \boldsymbol{b}_{i_t}, A_{i_t}\boldsymbol{\Sigma}\Big(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \boldsymbol{b}_{i_{t-1}}\right), j_t\Big)A_{i_t}^T\right),$$

where we use $\boldsymbol{x}'_{t-1} = A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \boldsymbol{b}_{i_{t-1}}\right)$, and for higher lags $M > 1$ we will have $(\boldsymbol{x}'_{t-1}, \dots, \boldsymbol{x}'_{t-M}) = \left(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \boldsymbol{b}_{i_{t-1}}\right), \dots, A_{i_{t-M}}^{-1}\left(\boldsymbol{x}_{t-M} - \boldsymbol{b}_{i_{t-M}}\right)\right)$ in the arguments of

$\boldsymbol{m}(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$ in the equation above. The previous density is an analytic function which is defined on an open neighbourhood of $\boldsymbol{x}_{1:T}^{(0)}$. Then from the identity theorem of analytic functions the resulting density defines the analytic extension of $p|_{B(\boldsymbol{x}_{1:T}^{(0)},\delta)}$ on $\mathbb{R}^{Tm}$. Then, we have

$$
\begin{aligned}
&\int_{\boldsymbol{x}_{1:T}\in\mathbb{R}^{Tm}} \operatorname{Ext}\left(p|_{B(\boldsymbol{x}_{1:T}^{(0)},\delta)}\right) d\boldsymbol{x}_{1:T} \\
&\quad= \int_{\boldsymbol{x}_{1:T}\in\mathbb{R}^{Tm}} \sum_{i_1,\ldots,i_T}^{\ell} \sum_{j_1,\ldots,j_T}^{K} p\left(s_{1:T}=\{j_1,\ldots,j_T\}\right)\mathcal{N}\left(\boldsymbol{x}_1; A_{i_1}\boldsymbol{\mu}(j_1)+\boldsymbol{b}_{i_1}, A_{i_1}\boldsymbol{\Sigma}_1(j_1)A_{i_1}^T\right) \\
&\qquad \prod_{t=2}^{T}\mathcal{N}\Big(\boldsymbol{x}_t; A_{i_t}\boldsymbol{m}\left(\boldsymbol{x}'_{t-1}, j_t\right)+\boldsymbol{b}_{i_t}, A_{i_t}\boldsymbol{\Sigma}\left(\boldsymbol{x}'_{t-1}, j_t\right)A_{i_t}^T\Big)\, d\boldsymbol{x}_{1:T} \\
&\quad= \sum_{i_1,\ldots,i_T}^{\ell} \int_{\boldsymbol{x}_{1:T}\in\mathbb{R}^{Tm}} \sum_{j_1,\ldots,j_T}^{K} p\left(s_{1:T}=\{j_1,\ldots,j_T\}\right)\mathcal{N}\left(\boldsymbol{x}_1; A_{i_1}\boldsymbol{\mu}(j_1)+\boldsymbol{b}_{i_1}, A_{i_1}\boldsymbol{\Sigma}_1(j_1)A_{i_1}^T\right) \\
&\qquad \prod_{t=2}^{T}\mathcal{N}\Big(\boldsymbol{x}_t; A_{i_t}\boldsymbol{m}\left(\boldsymbol{x}'_{t-1}, j_t\right)+\boldsymbol{b}_{i_t}, A_{i_t}\boldsymbol{\Sigma}\left(\boldsymbol{x}'_{t-1}, j_t\right)A_{i_t}^T\Big)\, d\boldsymbol{x}_{1:T} \\
&\quad= \sum_{i_1,\ldots,i_T}^{\ell} 1 = \ell^T = \left|F^{-1}(\{\boldsymbol{x}_{1:T}^{(0)}\})\right|.
\end{aligned}
$$

■

As mentioned, this leads to having an open set where both F and G mappings are invertible.

Corollary 34 (adapted from Balsells-Rodas et al., 2024) *Let F, $G:\mathbb{R}^{Tm}\to\mathbb{R}^{Tn}$ be factored piecewise affine mappings, with $\boldsymbol{x}_t := f(\boldsymbol{z}_t)$ and $\boldsymbol{x}'_t := g(\boldsymbol{z}'_t)$, for $1\le t\le T$. Assume f and g are weakly injective (Def. 5). Let $\boldsymbol{z}_{1:T}$ and $\boldsymbol{z}'_{1:T}$ be distributed according to the identifiable multi-lag MSM family $\mathcal{M}_{NL}^{T,M}$ with assumptions (m1-m2). Assume $F(\boldsymbol{z}_{1:T})$ and $G(\boldsymbol{z}'_{1:T})$ are identically distributed. Assume that for $\boldsymbol{x}_0\in\mathbb{R}^n$ and $\delta>0$, f is invertible on $B(\boldsymbol{x}_0,2\delta)\cap f(\mathbb{R}^m)$.*

Then, for $\boldsymbol{x}_{1:T}^{(0)}=\{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_0\}\in\mathbb{R}^{Tn}$ there exists $\boldsymbol{x}_{1:T}^{(1)}\in B(\boldsymbol{x}_{1:T}^{(0)},\delta)$ and $\delta_1>0$ such that both F and G are invertible on $B(\boldsymbol{x}_{1:T}^{(1)},\delta_1)\cap F(\mathbb{R}^{Tm})$.

Proof The proof is identical to that of Balsells-Rodas et al. (2024), replacing Lemma D.6 there with Lemma 33 (multi-lag version), which shows that the points in the pre-image of G can be computed thanks to $\boldsymbol{z}'_{1:T}$ following a multi-lag MSM under assumptions (m1-m2) as almost every point in the pre-image of G is generic, leading to invertibility of both F and G for a nonzero measure set. ■

Theorem 35 (adapted from Balsells-Rodas et al., 2024) *Let F, $G:\mathbb{R}^{Tm}\to\mathbb{R}^{Tn}$ be factored piecewise affine mappings, with $\boldsymbol{x}_t := f(\boldsymbol{z}_t)$ and $\boldsymbol{x}'_t := g(\boldsymbol{z}'_t)$, for $1\le t\le T$. Let*

$\boldsymbol{z}_{1:T}$ and $\boldsymbol{z}'_{1:T}$ be distributed according to the identifiable multi-lag MSM family $\mathcal{M}_{NL}^{T,M}$ under assumptions (m1-m2). Assume $F(\boldsymbol{z}_{1:T})$ and $G(\boldsymbol{z}'_{1:T})$ are identically distributed, and that there exists $\boldsymbol{x}_{1:T}^{(0)} \in \mathbb{R}^{Tn}$ and $\delta > 0$ such that F and G are invertible on $B(\boldsymbol{x}_{1:T}^{(0)}, \delta) \cap F(\mathbb{R}^{Tm})$. Then there exists an invertible factored affine transformation H such that $H(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$.

Proof Since a multi-lag MSM is closed under affine transformations (Balsells-Rodas et al., 2024, Prop C.2), and the distributions of the multi-lag MSM family $\mathcal{M}_{NL}^{T,M}$ under assumptions (m1-m2) are analytic functions, the proof is identical to Balsells-Rodas et al. (2024). ■

Now we have all the elements to prove Theorem 13 (i).

B.2 Proof of Theorem 13 (i)

Proof It follows from Balsells-Rodas et al. (2024). Assume there exists another model that generates the same distribution from Eq. (21), with a prior $p' \in \mathcal{M}_{NL}^{T,M}$ under assumptions (m1-m2), and nonlinear emission F', composed of a weakly injective, piecewise linear f': i.e., $(F \# p)(\boldsymbol{x}_{1:T}) = (F' \# p')(\boldsymbol{x}_{1:T})$.

From weakly-injectivity, at least for some $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\delta > 0$, f is invertible on $B(\boldsymbol{x}_0, 2\delta) \cap f(\mathbb{R}^m)$. This satisfies the preconditions from Corollary 34, which implies there exists $\boldsymbol{x}_{1:T}^{(1)} \in B(\boldsymbol{x}_{1:T}^{(0)}, \delta)$ and $\delta_1 > 0$ such that both F and F' are invertible on $B(\boldsymbol{x}_{1:T}^{(1)}, \delta_1) \cap F(\mathbb{R}^{Tm})$. Thus, by Theorem 35, there exists an affine transformation H such that $H(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$, i.e., $p \in \mathcal{M}_{NL}^{T,M}$ is identifiable up to affine transformations. ■

B.3 Proof of Theorem 13 (ii)

Proof It follows from Theorem 13 (i) and Theorem E.1 from Kivva et al. (2022), which is applicable for any $t \in \{M+1, \dots, T\}$ as $\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \dots \boldsymbol{z}_{t-M}$ is a GMM. Note that $t \in \{1, \dots, M\}$ is already covered by Kivva et al. (2022) as it reduces to the initial distribution case. From Theorem 13 (i) we have identifiability up to affine transformations: i.e., we have some $\boldsymbol{z}'_{1:T}$ such that $\boldsymbol{z}'_t = A\boldsymbol{z}_t + \boldsymbol{b}, t \in \{1, \dots, T\}$. Given (s1) and (s2), for any $t \in \{M+1, \dots, T\}$, the conditions of Theorem E.1 in Kivva et al. (2022) are satisfied in a nonzero measure set in the space of $(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M})$. Therefore, we can find $\tilde{A}$ such that $\tilde{\boldsymbol{z}}_t = \tilde{A}^{-1}\boldsymbol{z}'_t = \tilde{P}\tilde{D}\boldsymbol{z}_t + \tilde{\boldsymbol{b}}$, where $\tilde{P}$ and $\tilde{D}$ are permutation and scaling matrices, respectively. ■

Appendix C. Proof of Corollary 15

We first provide the definition of a node permutation in the context of temporal graph structures.

Definition 36 *Let $\mathbf{G} = (V, E_{1:M})$ be a graph, where, as defined in Section 4.3, $\mathbf{G} \in \mathbb{G}$ has $V := [m]$ as nodes and contains information about the temporal edges with maximum lag M. A node permutation $\pi \in S_m$ on $\mathbf{G}$ results in a permuted graph $\tilde{\mathbf{G}} = \pi(\mathbf{G})$ with the same set*

of vertices, but with a permuted set of edges, for each $1 \leq \tau \leq M$, as follows:

$$\tilde{\mathbf{G}} = (\tilde{V}, \tilde{E}_{1:M}), \quad \tilde{V} = \{\pi(1), \dots, \pi(m)\}, \quad \tilde{E}_\tau = \Big\{ \big(\pi(i), \pi(j)\big) \mid \big(i, j\big) \in E_\tau \Big\} \tag{70}$$

Proof Assume two SDSs $p(\boldsymbol{x}_{1:T}), \tilde{p}(\boldsymbol{x}_{1:T})$ with corresponding regime-dependent graphs $\mathbf{G}_{1:K}$, $\tilde{\mathbf{G}}_{1:\tilde{K}}$.

Given assumption (m1) and (m2*), we recover $p(\boldsymbol{z}_{1:T})$ up to permutation and scaling from Theorem 13. Combined with Theorem 8, we have identifiability up to permutations from Def. 7, which implies $K = \tilde{K}$. We focus on condition 4 where for each $1 \leq i \leq C$, there is some $1 \leq j \leq \tilde{C}$ such that:

$$p_{b_t^i}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1:t-M}) = \tilde{p}_{\tilde{b}_t^j}(\boldsymbol{z}'_t | \boldsymbol{z}'_{t-1:t-M}), \quad \forall (\boldsymbol{z}'_t, \dots, \boldsymbol{z}'_{t-M}) \in \mathbb{R}^{m(M+1)}. \tag{71}$$

Then, at each time step t there exists a permutation σ such that for each $k \in [K]$, $p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1:t-M}, s_t = k) = \tilde{p}(\boldsymbol{z}'_t | \boldsymbol{z}'_{t-1:t-M}, s_t = \sigma(k)), \quad \forall (\boldsymbol{z}_t, \dots, \boldsymbol{z}_{t-M}) \in \mathbb{R}^{m(M+1)}$. We know from condition 2 that σ is constant for $t > M$. Since we know the distributions are Gaussian, from Gaussian identifiability (Yakowitz and Spragins, 1968) we have

$$\boldsymbol{m}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, k) = PD\tilde{\boldsymbol{m}}(\boldsymbol{z}'_{t-1}, \dots, \boldsymbol{z}'_{t-M}, \sigma(k)) + \boldsymbol{b} \quad k \in [K], \tag{72}$$

where each $\boldsymbol{z}'_t = D^{-1}P^T\boldsymbol{z}_t + \tilde{\boldsymbol{b}}$ since the MSM is closed under affine transformations (Balsells-Rodas et al., 2024, Prop. C.2). Now, given that the functions $\boldsymbol{m}(\cdot, k), k \in [K]$ are analytic, the Jacobian will preserve similar permutation equivalences. W.l.o.g. we fix $\tau \in [M]$, and $k \in [K]$ and denote the Jacobian of the transition w.r.t. $\boldsymbol{z}_{t-\tau}$ as $J_{\boldsymbol{m}(\cdot,k),\tau}$. We compute the derivative w.r.t. $\boldsymbol{z}_{t-\tau}$ on Eq. (72):

$$\frac{\partial \boldsymbol{m}(\boldsymbol{z}_{t-1}, \dots, \boldsymbol{z}_{t-M}, k)}{\partial \boldsymbol{z}_{t-\tau}} = PD \frac{\partial \tilde{\boldsymbol{m}}(\boldsymbol{z}'_{t-1}, \dots, \boldsymbol{z}'_{t-M}, \sigma(k))}{\partial \boldsymbol{z}'_{t-\tau}} \frac{\partial \boldsymbol{z}'_{t-\tau}}{\partial \boldsymbol{z}_{t-\tau}}, \tag{73}$$

$$J_{\boldsymbol{m}(\cdot,k),\tau} = PDJ_{\tilde{\boldsymbol{m}}(\cdot,\sigma(k)),\tau}D^{-1}P^T. \tag{74}$$

Given the minimality assumption (m3), we can directly map $J_{\boldsymbol{m}(\cdot,k),\tau}$ and $J_{\tilde{\boldsymbol{m}}(\cdot,\sigma(k)),\tau}$ to some adjacency matrix via thresholding entries with absolute value greater than 0. Given D is diagonal, the structure of $DJ_{\tilde{\boldsymbol{m}}(\cdot,\sigma(k)),\tau}D^{-1}$ is preserved. The resulting set of graph adjacency matrices $A_{k,\tau}, \tilde{A}_{k,\tau}$ for regime k and lag τ relate as follows

$$A_{k,\tau} = P\tilde{A}_{\sigma(k),\tau}P^T, \tag{75}$$

where the permutation matrix P is associated to some permutation $\pi \in S_m$, and reflects a permutation of the rows and columns of $A_{k,\tau}$ following π. Therefore, the above operation is equivalent to a node permutation as defined in Def. 36, and we have $\mathbf{G}_k = \pi(\tilde{\mathbf{G}}_{\sigma(k)})$, for each $k \in [K]$. ■

Appendix D. Estimation Details

The ELBO presented in Eq. (25) is derived as follows:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) &= \log \int \sum_{\boldsymbol{s}_{M:T}} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{s}_{M:T}) d\boldsymbol{z}_{1:T} \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{M:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}|\boldsymbol{s}_{M:T}) p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T})}{q_{\boldsymbol{\phi},\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}, \boldsymbol{s}_{M:T}|\boldsymbol{x}_{1:T})} \right] \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} + \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T}|\boldsymbol{z}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}|\boldsymbol{s}_{M:T}) p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T})}{p_{\boldsymbol{\theta}}(\boldsymbol{s}_{M:T}|\boldsymbol{z}_{1:T})} \right] \right] \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} + \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) \right] \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} \Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T}) \Big] - KL\Big(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) \Big).
\end{aligned}$$

Given samples $\tilde{\boldsymbol{z}} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$, the posterior distributions $\gamma_{t,k}(\tilde{\boldsymbol{z}}_{1:T})$ and $\xi_{t,k,k'}(\tilde{\boldsymbol{z}}_{1:T})$ used in Eq. (29) can be computed using forward and backward messages (Bishop, 2006): $\{\alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t}) = p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{1:t}, s_t = k)\}$, and $\{\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T}) = p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{t+1:T}|s_t = k)\}$, respectively:

$$\gamma_{t,k}(\tilde{\boldsymbol{z}}_{1:T}) = \frac{\alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t})\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T})}{\sum_{k=1}^{K} \alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t})\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T})}, \tag{76}$$

$$\xi_{t,k,k'}(\tilde{\boldsymbol{z}}_{1:T}) = \frac{\alpha_{t-1,k'}(\tilde{\boldsymbol{z}}_{1:t-1})\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T}) p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_t|\tilde{\boldsymbol{z}}_{t-1}, \ldots, \tilde{\boldsymbol{z}}_{t-M}, s_t = k) p_{\boldsymbol{\theta}}(s_t = k \mid s_{t-1} = k')}{\sum_{k=1}^{K} \alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t})\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T})} \tag{77}$$

where $\alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t})$ and $\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T})$ are computed as follows:

$$\alpha_{t,k}(\tilde{\boldsymbol{z}}_{1:t}) = p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_t|\tilde{\boldsymbol{z}}_{t-1:t-M}, s_t = k) \times \sum_{k'=1}^{K} p_{\boldsymbol{\theta}}(s_t = k|s_{t-1} = k')\alpha_{t-1,k'}(\tilde{\boldsymbol{z}}_{1:t-1}), \tag{78}$$

$$\beta_{t,k}(\tilde{\boldsymbol{z}}_{t+1:T}) = \sum_{k'=1}^{K} p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{t+1}|\tilde{\boldsymbol{z}}_{t:t+1-M}, s_{t+1} = k') \times p_{\boldsymbol{\theta}}(s_{t+1} = k'|s_t = k)\beta_{t+1,k'}(\tilde{\boldsymbol{z}}_{t+1:T}). \tag{79}$$

D.1 Cost of Jacobian Regularisation

We measure the wall-clock time of the Jacobian-regularised objective in Equation (30), used to encourage learning regime-dependent latent causal graphs. To quantify the cost of the regularisation term, we use paired GPU runs with identical model initialisations and input batches, comparing training steps with and without the Jacobian penalty. Table 5 reports forward and forward/backward times for $K = 10$ regimes, lag $M = 2$, sequence length $T = 100$, and mini-batch size $B = 100$, across different dimensions $n = m \in \{5, 10, 20, 50\}$. The Jacobian penalty is implemented using vectorised automatic differentiation through `torch.func.jacrev` and `torch.vmap`, allowing the Jacobians to be evaluated in parallel across the batch and time axes. This implementation results in a moderate overhead of approximately 1.2–1.4× for a full training step.

	Forward			Forward + backward		
$n = m$	No reg. (s)	Jac. reg. (s)	Ratio	No reg. (s)	Jac. reg. (s)	Ratio
5	0.045 ± 0.001	0.067 ± 0.001	$1.489\times$	0.158 ± 0.002	0.192 ± 0.002	$1.216\times$
10	0.050 ± 0.003	0.072 ± 0.003	$1.417\times$	0.156 ± 0.004	0.193 ± 0.005	$1.234\times$
20	0.044 ± 0.001	0.066 ± 0.001	$1.495\times$	0.154 ± 0.004	0.188 ± 0.003	$1.219\times$
50	0.047 ± 0.002	0.074 ± 0.002	$1.574\times$	0.140 ± 0.004	0.195 ± 0.003	$1.390\times$

Table 5: Wall-clock overhead (in seconds) of transition-Jacobian regularisation with 95% CI. Timings are paired forward/backward training-step measurements. Settings: $B = 100$, $T = 100$, $M = 2$, $K = 10$; mean and standard deviation are reported across 50 repeats.

Appendix E. Experiment Details

The code for reproducing our experiments can be found here: `https://github.com/charlio23/identifiable-SDS`.

E.1 Synthetic Data Generation

We provide additional details of the data generation process. The initial Gaussian components are sampled from $\mathcal{N}(0, 0.7^2\mathbf{I})$. For the mean transition functions $\boldsymbol{m}(\cdot, k)$, we already indicated the use of locally connected networks (Zheng et al., 2018) where network dependencies follow a regime-dependent causal structure. The networks consist of two-layer MLPs with 16 hidden units and cosine activations. The covariance matrix $\boldsymbol{\Sigma}(\cdot, k)$ is designed as follows depending on the setting:

- Constant noise: fixed, $\boldsymbol{\Sigma} = 0.01\mathbf{I}$.
- Heterogeneous noise: $\boldsymbol{\Sigma}(k) = \text{diag}(\boldsymbol{\sigma}_k^2)$, where for each $k \in [K]$ and $i \in [m]$, $\sigma_{k,i}^2 \sim U(0.005, 0.08)$, such that assumption (s2) holds. That is, we check whether there are at least two regimes $k_1, k_2 \in [K]$ with distinct diagonal ratios across dimensions, and re-sample otherwise.
- History-dependent noise: $\boldsymbol{\Sigma}(\cdot, k) = \text{diag}(C_k \cdot \text{sigmoid}(\boldsymbol{m}(\cdot, k)))$, where $C_k \sim U(0.05, 0.1)$.

For iSDS, we further generate observations using two-layer Leaky ReLU networks with 8 hidden units. Regarding ablations, the assumptions we explore use the following networks for the mean transitions:

- Overlap: piecewise analytic functions with cosine activations, where we force the same transition function across states if the L_2 norm of the conditioned trajectory at time t $(\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-M})$ is between 3 and 5.
- Zero: Random networks with cosine activations.

In this experiment, we generate 5 data sets on each setting. For the Overlap setting, Table 6 reports the proportion of samples that share the same transition function, together with the interval used for each lag. The intervals were chosen to yield an overlap proportion of $\sim 15\%$.

Lag	Overlap interval	Overlap proportion
1	$[3, 5]$	0.1245 ± 0.0381
3	$[3, 4]$	0.1671 ± 0.0216
5	$[3, 4]$	0.1407 ± 0.0528

Table 6: Estimated proportion of samples in the overlap region. Values report mean ± 95% confidence interval across 5 data seeds.

E.2 Training Specifications

The experiments are implemented in PyTorch (Paszke et al., 2019) and performed on NVIDIA RTX 2080 Ti GPUs, except for the experiments on climate data with SDSs, where we used NVIDIA GeForce RTX 3090 GPUs. In all our experiments, we use Adam optimiser (Kingma and Ba, 2015) with a learning rate of $7 \cdot 10^{-3}$ for iMSM, and $5 \cdot 10^{-4}$ for iSDS. Other specific configurations depend on the data set and are indicated below. Similar to related approaches (Hälvä and Hyvarinen, 2020), we use random restarts to achieve better parameter estimates: 3 for iMSM initialisation (see below), and 5 for the full models.

To train iSDS, we generally follow a 4-stage procedure.

- **Initialisation (iMSM):** We initialise the latent dynamical prior by training iMSM on the observations after dimensionality reduction via PCA. To train iMSM, we use a batch size of 100, set the learning rate to $7 \cdot 10^{-3}$, and decrease it by a factor of 0.5 when the likelihood plateaus, up to twice. We use 3 random restarts, and select the best model on training likelihood to initialise iSDS.
- **Encoder-decoder pre-training:** We find an initialisation for the encoder and decoder by freezing the gradients on the dynamical prior.
- **Warmup phase:** iSDS is trained jointly but we keep the switch parameters π, Q frozen. This helps us prevent state collapse.
- **Final phase:** All the networks are fine-tuned.

Related works (Dong et al., 2020; Ansari et al., 2021) introduce annealing to force uniform regime posterior assignments and ensure the model uses all the switches for inference. Instead, we find that combining initialisation from iMSM and random restarts prevents state collapse. Below we describe the specific configurations and network architecture choices for each data set.

Synthetic data. For training, we use a batch size of 100, initialise the latent iMSM as described above, pre-train the encoder and decoder for 40 epochs, and warmup for 10 additional epochs. We then unfreeze all the gradients and train for 700 epochs with a learning rate of $5 \cdot 10^{-4}$ and decay of 0.8 every 200 epochs. The transition means and covariances follow the architecture from the data generation process. The encoder and decoder are two-layer MLPs with 128 hidden units and Leaky ReLU activations. We regularise the Jacobian transitions with $\eta = 0.05$ in the loss function Eq. (30).

Brain activity data. Training follows the initialisation phase of iSDS, and the transitions consist of two-layer MLPs with 32 hidden units with GELU activations.

Financial data. Training largely follows the synthetic data procedure with minor modifications, and we use the same setting for log prices and log returns. In the initialisation phase, we use the whole time series to encourage better regime learning, with a learning rate of $7 \cdot 10^{-3}$ which is reduced by a factor of 0.5 if likelihood plateaus for 10 iterations. iMSM training stops after the model plateaus 3 times. The rest of the phases follow the procedure described in the synthetic data with subsampled time series of length $T = 200$. We train for 100k iterations, with a learning rate of $5 \cdot 10^{-4}$ and a decay of 0.8 every 40k iterations. Pre-train and warmup phases last for 2k and 8k iterations respectively. The architecture consists of an encoder-decoder built from two-layer MLPs with 256 hidden dimensions and Leaky ReLU activations, and mean transitions of two-layer MLPs with 32 hidden dimensions and GELU activations, and regime-dependent diagonal covariances. Here we consider $\eta = 0$ in the loss function Eq. (30).

Climate data. Training follows the financial time-series procedure, where initialisation trains iMSM on the whole time series with $T = 360$ time steps. We later use the iMSM trained for the interpretability analysis presented in Section 7.3. To train the rest of iSDS, we subsample sequences of length $T = 100$. We pre-train the encoder-decoder for 250k iterations, with a learning rate of $5 \cdot 10^{-4}$ and a decay of 0.8 every 100k iterations. We then warm up iSDS for 4000 iterations, and we run the final phase for 120k iterations, with a decay of 0.8 every 20k iterations. For the network architecture, the transitions are the same as those used for the financial time series, and we use CNNs for inference and generation. For inference, we use a 9-layer CNN with increasing channels, kernel size 3, and padding 1, Leaky ReLU activations, and we alternate between using stride 2 and 1. We start with 64 channels and double the channels every two layers, until we have 512 channels. We then run a three-layer MLP with Leaky ReLU activations and 256 hidden dimensions. For generation, we use a similar network, starting with a two-layer MLP with Leaky ReLU activations and 256 hidden dimensions, and use transposed convolutions instead of convolutions (with the same configuration as before). We also consider $\eta = 0$ in this setting.

E.3 Evaluation Metrics

In Section 6 we outline how each metric is computed. Here we provide additional details for the L_2 distance and R^2 score on the regime-specific mean functions $\boldsymbol{m}(\cdot, k)$, and for recovering the latent regime-dependent causal structure.

E.3.1 Mean Function Metrics

Let $f, g : \mathbb{R}^{mM} \to \mathbb{R}^m$ and define the squared L_2 distance as follows:

$$d^2_{L_2}(f, g) := \int_{\boldsymbol{x} \in \mathbb{R}^{mM}} \left|\left| f(\boldsymbol{x}) - g(\boldsymbol{x}) \right|\right|^2 d\boldsymbol{x} \approx \frac{1}{I} \sum_{i=1}^{I} \left|\left| f\left(\boldsymbol{x}^{(i)}\right) - g\left(\boldsymbol{x}^{(i)}\right) \right|\right|^2, \tag{80}$$

where $\boldsymbol{x}^{(i)}$ denotes a sample from held-out data. Because models are identifiable up to a permutation of the regimes and an affine transformation of the continuous latents (or permutation and scaling), we first align the regimes by choosing the permutation σ that maximises the F_1 score of regime estimation using posterior assignments from $p(\boldsymbol{s}_{M:T} | \boldsymbol{z}_{1:T})$. Recall that under an affine transformation, the multi-lag MSM preserves the following

equivalence between mean functions:

$$\boldsymbol{m}(\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M},k) = A\hat{\boldsymbol{m}}\Big(A^{-1}(\boldsymbol{z}_{t-1}-\boldsymbol{b}),\dots,A^{-1}(\boldsymbol{z}_{t-M}-\boldsymbol{b}),\sigma(k)\Big)+\boldsymbol{b}, \tag{81}$$

where A, $\boldsymbol{b}$ denote the affine transformation parameters. To evaluate functions within the appropriate regime domain, we partition the held-out trajectories into K sets $\{\mathcal{Z}_k\}_{k\in[K]}$, where each set $\mathcal{Z}_k$ contains variables such that for each $M<t\leq T$, $\{\boldsymbol{z}_{t-1}^{(i)},\dots,\boldsymbol{z}_{t-M}^{(i)}\}\in\mathcal{Z}_k$ if $k=\arg\max p(s_t^{(i)}|\boldsymbol{z}_{1:T}^{(i)})$, where index i denotes a sample in the held-out data set. We use 1000 held-out trajectories. We then compute the error as follows:

$$L_2 \text{ err iSDS} := \min_{A,\boldsymbol{b}} \frac{1}{K}\sum_{k=1}^{K} d_{L_2}^2\Big(\boldsymbol{m}\big(\cdot,k\big), A\hat{\boldsymbol{m}}\big(\cdot,k\big)+\boldsymbol{b}\Big), \tag{82}$$

where inputs to $\boldsymbol{m}(\cdot,k)$ are drawn from $\mathcal{Z}_k$, and inputs to $\hat{\boldsymbol{m}}(\cdot,k)$ are mapped according to Eq. (81).

For R^2 scores, we follow a similar procedure, and compute a variance-normalised L_2 discrepancy:

$$d_{R^2}(f,g) := 1-\frac{d_{L_2}^2(f,g)}{Var(f)}, \quad Var(f)=\frac{1}{I}\sum_i \Big|\Big|f\big(\boldsymbol{x}^{(i)}\big)-\frac{1}{I}\sum_j f\big(\boldsymbol{x}^{(j)}\big)\Big|\Big|^2. \tag{83}$$

In ablations, we observe low F_1 regime scores due to the introduced function overlaps. Therefore, for iMSMs we choose the regime permutation which maximises average R^2:

$$R^2 \text{ err iMSM} := \max_{\sigma\in S_K}\frac{1}{K}\sum_{k=1}^{K} d_{R^2}\Big(\boldsymbol{m}\big(\cdot,k\big),\hat{\boldsymbol{m}}\big(\cdot,\sigma(k)\big)\Big). \tag{84}$$

For iSDS, we follow the same procedure, but first align the latents via an affine map:

$$R^2 \text{ err iSDS} := \max_{\sigma\in S_K}\frac{1}{K}\sum_{k=1}^{K} d_{R^2}\Big(\boldsymbol{m}\big(\cdot,k\big),A\hat{\boldsymbol{m}}\big(\cdot,\sigma(k)\big)+\boldsymbol{b}\Big), \tag{85}$$

with inputs to $\hat{\boldsymbol{m}}\big(\cdot\big)$ rescaled as in Eq. (81). Here we evaluate on the same 1000 held-out trajectories; unlike the L_2 computation, we do not pre-sort samples by regimes in the ablations (due to the low F_1 score).

E.3.2 Causal Structure Recovery

Let $\boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}(\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M})$ denote the Jacobian of $\hat{\boldsymbol{m}}(\cdot,k)$ at $(\boldsymbol{z}_{t-1},\dots,\boldsymbol{z}_{t-M})$. We estimate the regime-dependent latent causal graph $\hat{\mathbf{G}}_{1:K}$ by thresholding the mean absolute Jacobian over samples assigned to regime k. Using the same partition $\mathcal{Z}_k$ with size $|\mathcal{Z}_k|=N_k$, we define:

$$\hat{\mathbf{G}}_k := \mathbf{1}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\Big|\boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}\left(\boldsymbol{z}_{t-1}^{(i)},\dots,\boldsymbol{z}_{t-M}^{(i)}\right)\Big|>\tau\right), \quad (\boldsymbol{z}_{t-1}^{(i)},\dots,\boldsymbol{z}_{t-M}^{(i)})\in\mathcal{Z}_k, \tag{86}$$

where we use $\tau = 0.05$ in our experiments, and $\mathbf{1}(\cdot)$ denotes the indicator function, which equals 1 if the argument is true and 0 otherwise. For lag M, $\hat{\mathbf{G}}_k$ has shape $m \times (mM)$ (rows: current variables; columns: lagged variables across all M lags). Regimes are aligned by the permutation that maximises the regime F_1 score (as above). For latent alignment, we decompose the affine map fitted for the L_2 distance as $A \approx DP$, with P a permutation matrix and D diagonal. P is obtained by selecting maximal-absolute entries per row. For each lag $\tau \in [M]$ with $\hat{E}_{k,\tau}$ denoting the lag-τ adjacency matrix of $\hat{\mathbf{G}}_k$, we permute with $P\hat{E}_{\sigma(k),\tau}P^T$ for latent variable alignment. Finally, we report the average F_1 score between the aligned estimate and the ground-truth graphs $\mathbf{G}_{1:K}$, averaged over regimes.

E.4 Additional Synthetic Results

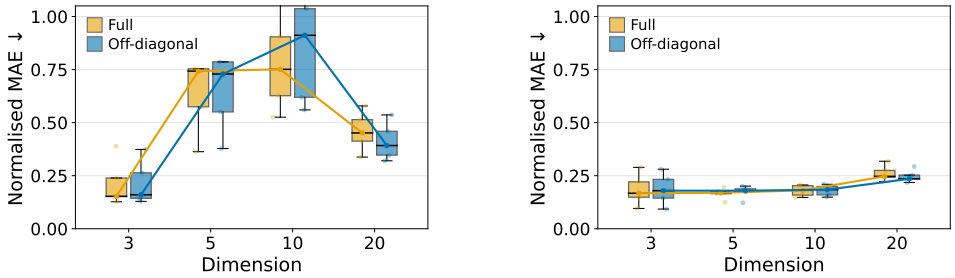
In the main experiments, we showcase structure and regime learning of iSDS, all of which assume diagonal covariances in the underlying MSM structure. For completeness, we evaluate full-covariance MSMs, and sample data with $K = 3$ regimes, lag $M = 2$, sequence length $T = 100$, with a data set of 1000 sequences. We consider MSM dimensions $m \in \{3, 5, 10, 20\}$ and increase the sparsity to allow $3 - -4$ interactions per dimension across all lags: that is, $\{0.3, 0.7, 0.85, 0.9\}$, respectively. For each setting, we use 5 different data seeds, and the training scheme follows the specifications in Appendix E.2, but using 10 random restarts instead. We consider two covariance settings: heterogeneous noise and history-dependent covariance. We follow the same data generation process as for iSDS (Appendix E.1), with modifications to generate full covariances. Specifically, we replace the diagonal covariances by a parametrisation similar to GARCH models (Bollerslev, 1990), where each regime has a random correlation matrix $\boldsymbol{R}_k$, and the covariance is obtained using a regime and time-dependent scaling $\boldsymbol{D}_{k,t}\boldsymbol{R}_k\boldsymbol{D}_{k,t}$ for some diagonal matrix $\boldsymbol{D}_{k,t}$ at time t. Here, $\boldsymbol{R}_k$ is a symmetric positive definite matrix with unit diagonal. We sample each setting as follows:

- Heterogeneous noise: $\boldsymbol{\Sigma}(k) = \operatorname{diag}(\boldsymbol{\sigma}_k)\boldsymbol{R}_k \operatorname{diag}(\boldsymbol{\sigma}_k) + 10^{-4}\mathbf{I}$, where $\sigma^2_{k,i} \sim U(0.01, 0.12)$.
- History-dependent noise: We modify the original setting by first sampling coordinate-wise scales: $\boldsymbol{v}(\boldsymbol{z}_{t-1:t-M}, k) = \boldsymbol{C}_k \odot \operatorname{sigmoid}(\boldsymbol{g}_k(\boldsymbol{z}_{t-1:t-M}))$, where $C_{k,i} \sim U(0.08, 0.16)$, and $\boldsymbol{g}_k$ is a random function independent of $\boldsymbol{m}(\cdot, k)$ which follows the same specifications. The covariance is then computed as:

$$\boldsymbol{\Sigma}(\boldsymbol{z}_{t-1:t-M}, k) = \operatorname{diag}\Big(\sqrt{\boldsymbol{v}(\boldsymbol{z}_{t-1:t-M}, k)}\Big)\, \boldsymbol{R}_k \operatorname{diag}\Big(\sqrt{\boldsymbol{v}(\boldsymbol{z}_{t-1:t-M}, k)}\Big) + 10^{-5}\mathbf{I}.$$

Note that in this setting we use larger upper bounds compared to the iSDS covariances, in order to make covariance estimation more challenging.

In Figure 14, we report normalised MAE for the full covariance matrix and for off-diagonal entries, where the MAE is divided by the mean absolute magnitude of the corresponding ground-truth covariance entries. Across all settings, the model maintains near-perfect regime recovery, with an average regime F_1 score of 0.994 ± 0.001. For covariance recovery, we observe that covariances in the history-dependent setting are easier to estimate than those in the fixed heterogeneous setting. Although iMSM performs well in the lowest-dimensional case, the normalised covariance error increases at intermediate dimensions. This suggests that accurately estimating unconstrained full covariance matrices remains challenging in



(a) Heterogeneous. (b) History-dependent.

Figure 14: Full covariance recovery for iMSM under (a) heterogeneous and (b) history-dependent covariance settings. We report normalised MAE for the full covariance and off-diagonal entries, computed as the MAE divided by the mean absolute magnitude of the ground-truth entries.

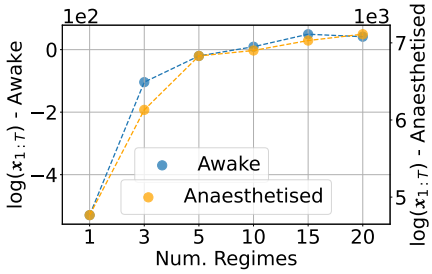


Figure 15: Test log-likelihood for different numbers of regimes with lag 2 for Awake and Anaesthetised conditions.

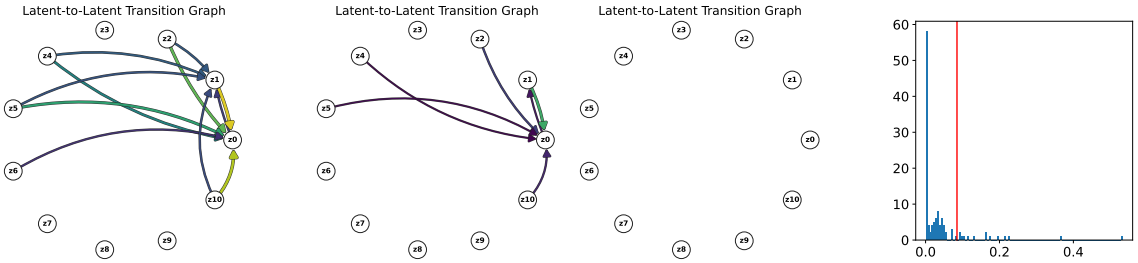
finite samples. In contrast, the history-dependent setting imposes additional covariance structure through history-dependent scales, which stabilises covariance estimation.

E.5 Brain Activity Data

Figure 15 reports test log-likelihoods for the Awake and Anaesthetised conditions as the number of regimes K increases. Performance improves with larger K and plateaus around $K \approx 10$. We adopt an overcomplete setting with $K = 15$ in the main experiments, and expect the models to use a subset of the regimes in practice. As reported in Table 2 (Section 7.1), the total regime usage is 53.33% for Awake and 33.33% for Anaesthetised, corresponding to 8 and 5 active regimes. Allowing a larger candidate set of regimes helps prevent state collapse, as unused regimes remain inactive, and test likelihood does not degrade.

Metrics from Table 2. Each sequence lasts 2 s, with $T = 400$ time steps. We compute all metrics from the smoothed posterior over regimes (uniform kernel of length 3).

- **Transition frequency (Hz):** Regime changes per second, averaged across epochs.
- **Total usage:** Fraction of regimes whose occupancy exceeds 1% of time points in at least one test epoch.



(a) Regime 0. (b) Regime 1. (c) Regime 2. (d) Edge histogram.

Figure 16: (a–c) Latent transition graphs by regimes and (d) edge intensities (log returns).

- **Usage per sample:** Mean (over epochs) fraction of regimes whose per-epoch occupancy exceeds 1% of time points.
- **Max duration (s):** Maximum regime dwell time, averaged across epochs.

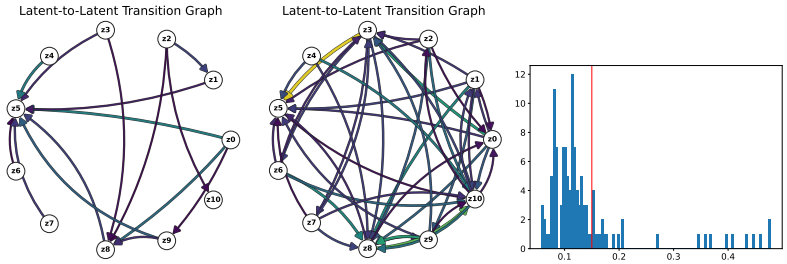
E.6 Financial Data

Below we list the tickers of the 100 stock-sector assignments used in our experiments, organised into 10 sectors. DIS and GOOGL appear in two sector groups due to exposure to multiple sectors.

- **Tech:** AAPL, MSFT, GOOGL, META, AMZN, NVDA, INTC, CSCO, IBM, ORCL
- **Health:** JNJ, PFE, MRK, ABBV, BMY, LLY, AMGN, GILD, ILMN, UNH
- **Financials:** JPM, BAC, WFC, C, GS, MS, USB, PNC, TFC, COF
- **Consumer Staples:** KO, PEP, PG, CL, KMB, MDLZ, GIS, K, WMT, COST
- **Consumer Discr.:** NKE, SBUX, MCD, DIS, BKNG, TJX, GAP, KMX, ROST, YUM
- **Industrials:** BA, CAT, MMM, GE, HON, LMT, RTX, UNP, CSX, DAL
- **Energy:** XOM, CVX, COP, PSX, DVN, OXY, KMI, EOG, HES, VLO
- **Real Estate:** SPG, PLD, PSA, AVB, EQR, WELL, VTR, DLR, O, ESS
- **Communication:** VZ, T, TMUS, CMCSA, CHTR, DIS, NFLX, PARA, EA, GOOGL
- **Materials:** DD, SHW, LYB, NEM, FCX, NUE, AA, MLM, CRH, CE

We also examine latent transition structure across regimes for log returns (Fig. 16), and the edge-intensity distribution (Fig. 16d). The histogram is computed using the average absolute off-diagonal Jacobian entries for each regime, and it is concentrated near zero. This indicates weak latent interactions. For interpretability, we show the higher-intensity subset (right of the red threshold). In Regimes 0 and 1, z_0 and z_1 dominate the latent dynamics (consistent with global market factors). Furthermore, the high-volatility regime (Regime 0) exhibits denser interactions than the medium and low-volatility regimes (Regimes 1 and 2), consistent with the need for stronger dynamical adjustments during turbulent periods.

For log price data, Figure 17 shows regime-specific latent transition graphs and the corresponding edge-intensity histogram (Fig. 17c). The edge distribution is centred near 0.1



(a) Regime 0. (b) Regime 1. (c) Edge histogram.

Figure 17: (a-b) Latent transition graphs by regime and (c) edge intensities (log price).

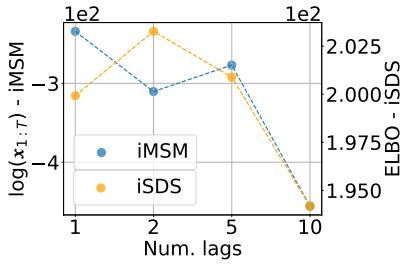


Figure 18: Test log-likelihood (blue) and ELBO (orange) for iMSM and iSDS respectively with increasing M. Quantities are reported per observation size for better comparison (i.e., 10 for iMSM, 64×64 for iSDS).

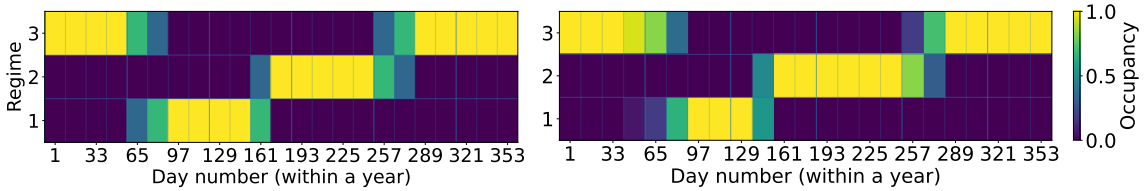
with a long tail of stronger interactions. The latent dynamics differ across regimes, with the lateral/bear regime (Regime 0) exhibiting simpler and weaker dynamics compared to the bull regime (Regime 1). The two global factors (z_2 and $z_1 0$) are involved differently across regimes. Notably, interactions are concentrated on z_5 and z_8, which are aligned with Consumer Staples, Real Estate, and Materials. This suggests these sectors require relatively complex, regime-dependent adjustments (e.g., Consumer Staples as a defensive sector; Real Estate as rate-sensitive).

E.7 Climate Data

Below we discuss details for our climate data experiments presented in Section 7.3.

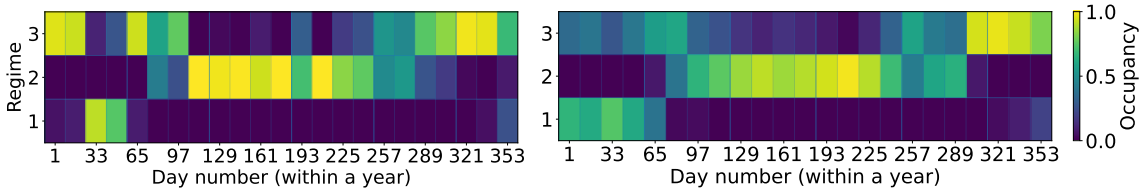
E.7.1 Model Selection

The data settings differ between iMSM and iSDS. For iMSM, we follow Varando et al. (2022) and project land pixels onto 10 principal components. For iSDS, we train on random 64×64 land patches with subsequences of length $T = 100$. We fix $K = 4$ regimes and vary the number of lags M. Figure 18 reports held-out performance as a function of M: test log-likelihood and ELBO (higher is better in both cases). To make scales comparable, we report both metrics per observation (i.e., per PC for iMSM and per pixel for iSDS). We observe that iMSM performs best at $M = 1$, while iSDS benefits from higher orders, with $M = 2$ performing best. We note that iMSM operates on globally mixed signals coming from very high-dimensional data (land pixels), and therefore increasing M can quickly overfit



(a) Jan. 2001 – Dec. 2004 (held-out). (b) Jan. 2005 – Dec. 2020.

Figure 19: Year-based structures from land pixels using iMSM.



(a) Jan. 2001 – Dec. 2004 (held-out). (b) Jan. 2005 – Dec. 2020.

Figure 20: Year-based structures on Ethiopia uncovered by iSDS.

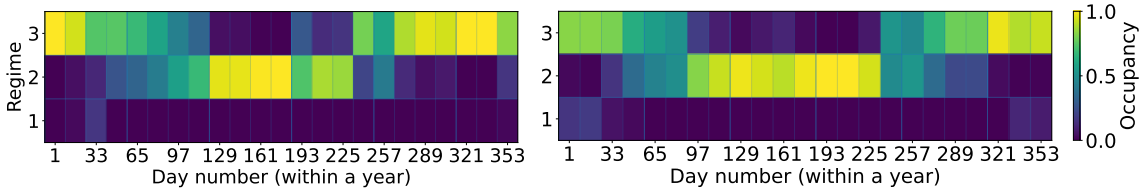
unseen sequences. In contrast, iSDS models local dynamics with shared transitions, and parameter sharing regularises additional lag parameters.

E.7.2 Regime Calendars

In Section 7.3, Tables 3 and 4 report ENSO correlations by season for iMSM and iSDS respectively. We visualise the within-year regime occupancy of iMSM in Figure 19. As mentioned, the dynamics are structured yearly within 3 periods: from March to May, from June to September, and from October to February. The structures remain consistent for unobserved years. For iSDS, we observe 5 types of recurrent year-based patterns depending on location. Figures 20–22 show Ethiopia, Sahel, and South Africa. We omit calendars for Sahara and Congo, as they predominantly use a single regime. Across regions, the reported months in Table 4 agree with the visual occupancy, which remains consistent for unseen years.

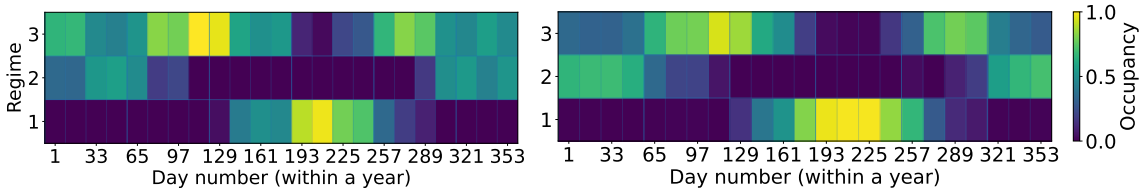
E.7.3 ENSO Matrices

Here we detail how Tables 3 and 4 are computed. We use all available data and evaluate lagged correlations with ENSO on a 0–240-day grid in 16-day steps (ENSO leads).



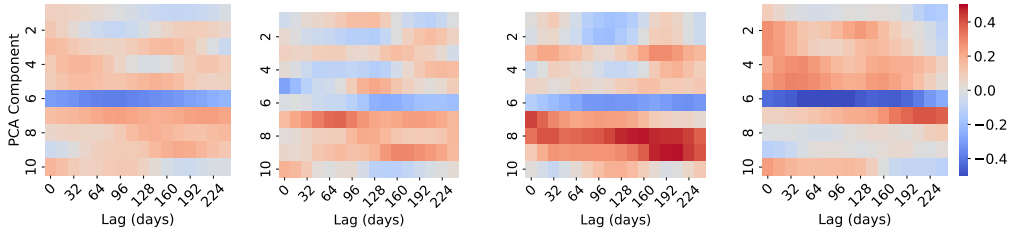
(a) Jan. 2001 – Dec. 2004 (held-out). (b) Jan. 2005 – Dec. 2020.

Figure 21: Year-based structures on Sahel uncovered by iSDS.



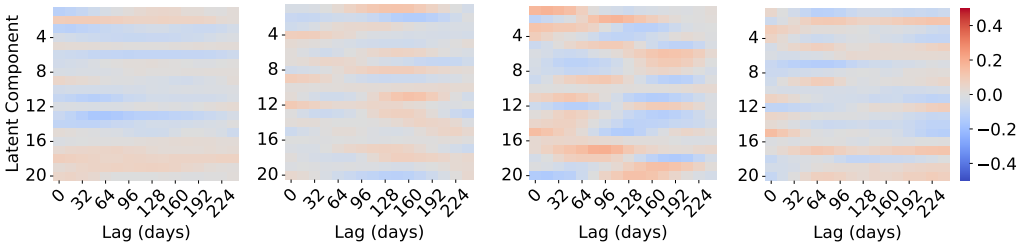
(a) Jan. 2001 – Dec. 2004 (held-out). (b) Jan. 2005 – Dec. 2020.

Figure 22: Year-based structures on South Africa uncovered by iSDS.

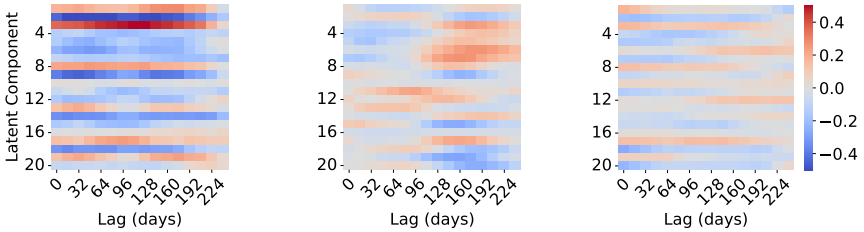


(a) All times. (b) Jun.–Sep. (c) Mar.–May. (d) Sep.–Feb.

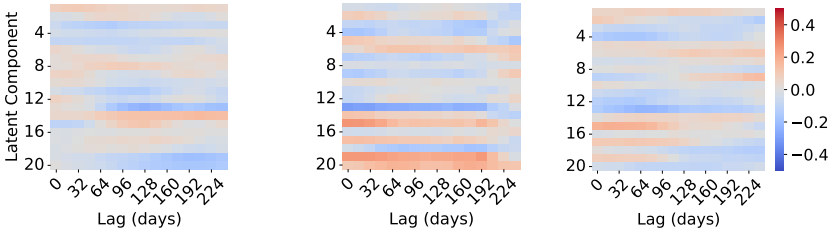
Figure 23: Correlation matrices of PCs from land pixels with respect to lagged ENSO (ENSO leads) with (a) all data, and (b–c) season-based data according to iMSM regime posteriors.



(a) Sahara, all times. (b) Congo, all times. (c) Sahel, Mar–Aug. (d) Sahel, Sep–Feb.



(e) Ethiopia, Jan–Mar. (f) Ethiopia, Apr–Aug. (g) Ethiopia, Aug–Dec.



(h) S. Africa, Jun–Sep. (i) S. Africa, Dec–Feb. (j) S. Africa, Mar–Jun & Sep–Nov.

Figure 24: Correlation matrices of iSDS latent variables from different locations with respect to lagged ENSO (ENSO leads) on (a) Sahara, (b) Congo, (c–d) Sahel, (e–g) Ethiopia, and (h–j) South Africa. For region-based correlation matrices, we sort the latent variables according to the corresponding regime posterior.

- **iMSM:** For each principal component (PC) we form a correlation matrix of size 10×16 (components, lags) for "all times" and seasonal subsets obtained from iMSM regime posteriors. We visualise the iMSM correlation matrices in Figure 23, and Table 3 reports, for each slice, the peak absolute correlation and its lag (days).
- **iSDS:** For each region, we sample 10 patches and sort the latent variables according to the iSDS regime posteriors. Then, for each season we compute similar per-patch correlation matrices of size 20×16 (latent variables, lags), and aggregate patches using Fisher z-averaging (Fisher, 1915). Table 4 reports peak correlations and lag (days) for each region and season, and we visualise the regional correlation matrices in Figure 24.

References

Amir Alizadeh and Nikos Nomikos. A Markov regime switching approach for hedging stock indices. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 24 (7):649–674, 2004.

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.

Yonghong An, Yingyao Hu, Johns Hopkins, and Matt Shum. Identifiability and inference of hidden Markov models. Technical report, Technical report, 2013.

Andrew Ang and Geert Bekaert. International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4):1137–1187, 2002.

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.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Mehdi Azzouzi and Ian T Nabney. Modelling financial time series with switching state space models. In *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr)(IEEE Cat. No. 99TH8408)*, pages 240–249. IEEE, 1999.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rk49Mg-CW`.

Nathan S Balke and Mark E Wohar. Market fundamentals versus rational bubbles in stock prices: a Bayesian perspective. *Journal of Applied Econometrics*, 24(1):35–75, 2009.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=j88QAtutwW.

Ole Barndorff-Nielsen. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, 12(1):115–121, 1965.

Endalkachew Bekele-Biratu, Wassila M. Thiaw, and Diriba Korecha. Sub-seasonal variability of the belg rains in ethiopia. *International Journal of Climatology*, 38(7):2940–2953, 2018.

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

Bloomberg L.P. Historical closing prices (px_last) for u.s. equities—tickers as listed in this paper. Bloomberg Terminal, Imperial College London, 2025. Data range: 2015-01-01 to 2025-05-31; Accessed 2025-11-09.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Tim Bollerslev. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The review of economics and statistics*, pages 498–505, 1990.

Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *International Conference on Learning Representations*, volume 2025, pages 46606–46660, 2025.

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

Carlos M Carvalho and Hedibert F Lopes. Simulation-based sequential analysis of Markov switching stochastic volatility models. *Computational Statistics & Data Analysis*, 51(9): 4526–4542, 2007.

Giuseppe Cavaliere, Michele Costa, and Luca De Angelis. Investigating stock market behavior using a multivariate Markov-switching approach. In *Advances in Latent Variables: Methods, Models and Applications*, pages 185–196. Springer, 2014.

Joshua CC Chan and Caterina Santi. Speculative bubbles in present-value models: a Bayesian Markov-switching state space approach. *Journal of Economic Dynamics and Control*, 127: 104101, 2021.

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.

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

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

Kamel Didan. MOD13C1 MODIS/Terra vegetation indices 16-day L3 global 0.05deg CMG V006, 2015.

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.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Emily Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on signal processing*, 59 (4):1569–1585, 2011.

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.

Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*, volume 425. Springer, 2006.

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.

Elisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space. *The Journal of Machine Learning Research*, 21(1):4589–4628, 2020.

Victor Geadah, International Brain Laboratory, and Jonathan W. Pillow. Parsing neural dynamics with infinite recurrent switching linear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=YIls9HEa52`.

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

Massimo Guidolin and Allan Timmermann. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544, 2007.

Markus Haas, Stefan Mittnik, and Marc S Paolella. A new approach to Markov-switching garch models. *Journal of financial Econometrics*, 2(4):493–530, 2004.

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.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21, 2008.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*, 29, 2016.

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.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, , Sandeep R Datta, and Ryan P Adams. Composing graphical models with neural networks for structured representations and fast inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Orren Karniol-Tambour, David M. Zoltowski, E. Mika Diamanti, Lucas Pinto, Carlos D Brody, David W. Tank, and Jonathan W. Pillow. Modeling state-dependent communication between brain regions with switching nonlinear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=WQwV7Y8qwa`.

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.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

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.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18 (2):95–138, 1977.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, 2018.

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.

Scott W Linderman, Andrew C Miller, Ryan P Adams, David M Blei, Liam Paninski, and Matthew J Johnson. Recurrent switching linear dynamical systems. *arXiv preprint arXiv:1610.08466*, 2016.

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

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1263–1273. PMLR, 31 Jul–04 Aug 2023a. URL https://proceedings.mlr.press/v216/lippe23a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=itZ6ggvMnzS.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Mohammadreza Mahmoudi and Hana Ghaneei. Detection of structural regimes and analyzing the impact of crude oil market on canadian stock market: Markov regime-switching approach. *Studies in Economics and Finance*, 39(4):722–734, 2022.

Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1995.

Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

Pedro A.M. Mediano, Fernando E. Rosas, Andrea I. Luppi, Valdas Noreika, Anil K. Seth, Robin L. Carhart-Harris, Lionel Barnett, and Daniel Bor. Spectrally and temporally resolved estimation of neural signal diversity. *eLife*, March 2023. doi: doi.org/10.7554/eLife.88683.1. URL `http://dx.doi.org/10.7554/eLife.88683.1`.

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.

Kevin P Murphy. Switching kalman filters, 1998.

Sharon E Nicholson. The west african sahel: A review of recent studies on the rainfall regime and its interannual variability. *International Scholarly Research Notices*, 2013(1):453521, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Nathalie Pettorelli. The normalized difference vegetation index. *The Normalized Difference Vegetation Index*, pages 1–208, 10 2013. doi: 10.1093/acprof:osobl/9780199693160.001.0001.

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.

Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025a.

Abdellah Rahmani and Pascal Frossard. Flow based approach for dynamic temporal causal models with non-gaussian or heteroscedastic noises. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=BgwtOwFSvY.

Goutham Rajendran, Patrik Reizinger, Wieland Brendel, and Pradeep Kumar Ravikumar. An interventional perspective on identifiability in gaussian lti systems with independent component analysis. In *Causal Learning and Reasoning*, pages 41–70. PMLR, 2024.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ICA. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2Yo9xqR6Ab.

M Rouault and FS Tomety. Impact of el niño–southern oscillation on the benguela upwelling. *Journal of Physical Oceanography*, 52(10):2573–2587, 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.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. In *Neural Information Processing Systems*, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

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. *Advances in Neural Information Processing Systems*, 36: 8092–8113, 2023.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. *Advances in Neural Information Processing Systems*, 37:77098–77131, 2024.

Henry Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

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

Gherardo Varando, Miguel-Ángel Fernández-Torres, Jordi Muñoz-Marí, and Gustau Camps-Valls. Learning causal representations with granger pca. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Kun Wang, Sumanth Varambally, Duncan Watson-Parris, Yian Ma, and Rose Yu. Discovering latent causal graphs from spatiotemporal data. In *Forty-second International Conference on Machine Learning*, 2025.

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

Toru Yanagawa, Zenas C. Chao, Naomi Hasegawa, and Naotaka Fujii. Large-scale information flow in conscious and unconscious states: an ecog study in monkeys. *PLOS ONE*, 8(11):null, 11 2013. doi: 10.1371/journal.pone.0080845. URL `https://doi.org/10.1371/journal.pone.0080845`.

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*, 2022a. URL `https://openreview.net/forum?id=Vi-sZWNA_Ue`.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b. URL `https://openreview.net/forum?id=RDlLMjLJXdq`.

Yixiu Zhao and Scott Linderman. Revisiting structured variational autoencoders. In *International Conference on Machine Learning*, pages 42046–42057. PMLR, 2023.

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.