On the Identifiability of Switching Dynamical Systems

Carles Balsells-Rodas [1] **Yixin Wang** [2] **Yingzhen Li** [1]

Abstract

The identifiability of latent variable models has received increasing attention due to its relevance in interpretability and out-of-distribution generalisation. In this work, we study the identifiability of Switching Dynamical Systems, taking an initial step toward extending identifiability analysis to sequential latent variable models. We first prove the identifiability of Markov Switching Models, which commonly serve as the prior distribution for the continuous latent variables in Switching Dynamical Systems. We present identification conditions for first-order Markov dependency structures, whose transition distribution is parametrised via non-linear Gaussians. We then establish the identifiability of the latent variables and non-linear mappings in Switching Dynamical Systems up to affine transformations, by leveraging identifiability analysis techniques from identifiable deep latent variable models. We finally develop estimation algorithms for identifiable Switching Dynamical Systems. Throughout empirical studies, we demonstrate the practicality of identifiable Switching Dynamical Systems for segmenting high-dimensional time series such as videos, and showcase the use of identifiable Markov Switching Models for regime-dependent causal discovery in climate data.

1. Introduction

State-space models (SSMs) are well-established sequence modelling techniques where their linear versions have been extensively studied (Lindgren, 1978; Poritz, 1982; Hamilton, 1989). Meanwhile, recurrent neural networks (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) have gained high popularity for sequence modelling thanks to their abilities in capturing non-linear and long-term dependencies. Nevertheless, significant progress has been made on fusing neural networks with SSMs, with Gu et al. (2022) as one of the latest examples. Many of these advances focus on building sequential latent variable models (LVMs) as flexible deep generative models (Chung et al., 2015; Li & Mandt, 2018; Babaeizadeh et al., 2018; Saxena et al., 2021), where 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; Smith et al., 2023). Despite the efforts in designing flexible SSM priors and developing stable training schemes, theoretical properties such as identifiability for sequential generative models are less studied, contrary to early literature on linear SSMs.

Identifiability, in general, establishes a one-to-one correspondence between the data likelihood and the model parameters (or latent variables), or an equivalence class of the latter. In causal discovery (Peters et al., 2017), identifiability refers to whether the underlying causal structure can be correctly pinpointed from infinite observational data. In independent component analysis (ICA, Comon (1994)), identifiability analysis focuses on both the latent sources and the mapping from the latents to the observed. While general non-linear ICA is ill-defined (Hyvärinen & Pajunen, 1999), recent results show that identifiability can be achieved using conditional priors (Khemakhem et al., 2020). For deep (non-temporal) latent variable models, the required access to auxiliary variables can be relaxed using a finite mixture prior (Kivva et al., 2022). Recent works have attempted to extend these results to sequential models using non-linear ICA (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Hälvä & Hyvarinen, 2020; Hälvä et al., 2021), or latent causal processes (Yao et al., 2022a;b; Lippe et al., 2023b;a).

In this work, we develop an identifiability analysis for Switching Dynamical Systems (SDSs) – a class of sequential LVMs with SSM priors that allow regime-switching behaviours, where their associated inference methods have been explored recently (Dong et al., 2020; Ansari et al., 2021). Our approach differs fundamentally from non-linear ICA since the latent variables are no longer independent. To address this challenge, we first construct the latent dynamical prior using Markov Switching Models (MSMs) (Hamilton, 1989) – an extension of Hidden Markov Models (HMMs) with autoregressive connections, for which we provide the first identifiability analysis of this model

[1]Imperial College London [2]University of Michigan. Correspondence to: Carles Balsells-Rodas <cb221@imperial.ac.uk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

family in non-linear transition settings. This also allows regime-dependent causal discovery (Saggioro et al., 2020) thanks to having access to the transition derivatives. We then extend the identifiability results to SDSs using recent identifiability analysis techniques for the non-temporal deep latent variable models (Kivva et al., 2022). Importantly, the identifiability conditions we provide are less restrictive, significantly broadening their applicability. In contrast,Yao et al. (2022a;b); Lippe et al. (2023a) require auxiliary information to handle regime-switching effects or distribution shifts, and Hälvä et al. (2021) imposes assumptions on the temporal correlations such as stationarity (Hyvarinen & Morioka, 2017). Moreover, unlike the majority of existing works (Hälvä et al., 2021; Yao et al., 2022a;b), our identifiability analysis does not require the injectivity of the decoder mapping. Below we summarise our main contributions in both theoretical and empirical forms:

- We present conditions in which first-order MSMs with non-linear Gaussian transitions are identifiable up to permutations (Section 3.1). We further provide the first analysis of identifiability conditions for non-parametric first-order MSMs in Appendix B.
- We further provide conditions for SDS's identifiability up to affine transformations of the latent variables and non-linear emission (Section 3.2).
- We demonstrate the effectiveness of identifiable SDSs on causal discovery and sequence modelling tasks. These include discovery of time-dependent causal structures in climate data (Section 6.2), and time-series segmentation in videos (Section 6.3).

2. Background

2.1. 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) piece-wise linear mapping to obtain the observation $\boldsymbol{x} \in \mathbb{R}^n$, $n \geq m$:

$$\boldsymbol{x} = f(\boldsymbol{z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

$$\boldsymbol{z} \sim p(\boldsymbol{z}) := \sum_{k=1}^{K} p(s=k)\mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{2}$$

Kivva et al. (2022) established that if the transformation f is *weakly injective* (see definition below), the prior distribution $p(\boldsymbol{z})$ is identifiable up to affine transformations[1] from the observations. If we further assume f is continuous and *injective*, both prior distribution $p(\boldsymbol{z})$ and non-linear mapping f are identifiable up to affine transformations. In Section 3.2 we extend these results to establish identifiability for SDSs using similar proof strategies.

[1] See Def. 2.2 in Kivva et al. (2022) or the equivalent adapted to SDSs (Def. 3.4).

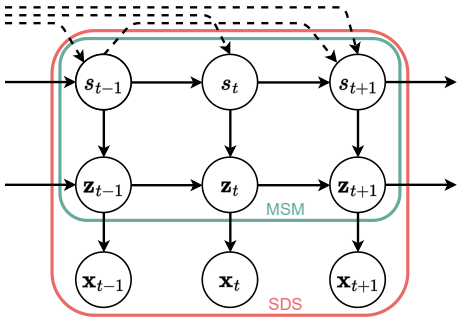


Figure 1: The generative model considered in this work, where the MSM is indicated in green and the SDS is indicated in red. The dashed arrows indicate additional dependencies which are accommodated by our theoretical results.

Definition 2.1. A mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ is said to be weakly injective if (i) there exsists $\boldsymbol{x}_0 \in \mathbb{R}^n$ 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 $f(\mathbb{R}^m)$.[2]. Here the notations are extended to sets, i.e., for a set $\mathcal{B}$, $f(\mathcal{B}) := \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{B}\}$ and $f^{-1}(\mathcal{B}) := \{\boldsymbol{x} : f(\boldsymbol{x}) \in \mathcal{B}\}$.

Definition 2.2. f is injective if $|f^{-1}(\{\boldsymbol{x}\})| = 1, \forall \boldsymbol{x} \in f(\mathbb{R}^m)$.

2.2. 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, $s_t \in \{1, \ldots, K\}, \boldsymbol{z}_t \in \mathbb{R}^m$, 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) non-linear transformation. This gives the following probabilistic model for the states and the observation variables:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}) = p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}) p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}|\mathbf{s}_{1:T}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{z}_t). \tag{3}$$

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 & Welling, 2014). These works consider an additional dependency on the switch s_t from $\boldsymbol{x}_t$ for more expressive segmentations, which is not considered in our theoretical analysis for simplicity.

[2] $B(\boldsymbol{x}_0, \delta) = \{\boldsymbol{x} \in \mathbb{R}^n : ||\boldsymbol{x} - \boldsymbol{x}_0|| < \delta\}$

For the latent dynamic prior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$, we consider a Markov Switching Model (MSM) which has also been referred to as autoregressive HMM (Ephraim & 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 $\mathbf{s}_{1:T}$ follows an autoregressive process. Under first-order Markov assumption for this conditional auto-regressive processes, this leads to the following probabilistic model for the prior:

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) = \sum_{\mathbf{s}_{1:T}} p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}) p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}|\mathbf{s}_{1:T}), \quad (4)$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}|\mathbf{s}_{1:T}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}_1|s_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t). \quad (5)$$

Note that the structure of the discrete latent state prior $p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T})$ is not specified, and the identifiability results presented in the next section do not require further assumptions herein. As illustrated in Figure 1 (solid lines), in experiments we use a first-order Markov process for $p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T})$, described by a transition matrix $Q \in \mathbb{R}^{K\times K}$ such that $p_{\boldsymbol{\theta}}(s_t = j|s_{t-1} = i) = Q_{ij}$, and an initial distribution $p_{\boldsymbol{\theta}}(s_1)$. In such case we also provide identifiability guarantees for the Q matrix and the initial distribution.

3. Identifiability Theory for MSMs & SDSs

This section establishes the identifiability of the SDS model (Eq. (3)) with MSM latent prior (Eq. (4)) under suitable assumptions. We address this challenge by leveraging ideas from Kivva et al. (2022) which uses finite mixture prior for static data generative models; importantly, this theory relies on the use of identifiable finite mixture priors (up to mixture component permutations). Inspired by this result, we first establish in Section 3.1 the identifiability of the Markov switching model $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$ as a finite mixture prior, which then allows us to extend the results of Kivva et al. (2022) to the temporal setting in Section 3.2 to prove identifiability of the SDS model. We drop the subscript $\boldsymbol{\theta}$ for simplicity.

3.1. Identifiable Markov Switching Models

The MSM $p(\boldsymbol{z}_{1:T})$ has an equivalent formulation as a finite mixture model. Suppose the discrete states satisfy $s_t \in \{1, ..., K\}$ with $K < +\infty$, then for any given $T < +\infty$, one can define a bijective *path indexing function* $\varphi : \{1, ..., K^T\} \to \{1, ..., K\}^T$ such that each $i \in \{1, ..., K^T\}$ uniquely retrieves a set of states $\mathbf{s}_{1:T} = \varphi(i)$. Then we can use $c_i = p(\mathbf{s}_{1:T} = \varphi(i))$ to represent the joint probability of the states $\mathbf{s}_{1:T} = \varphi(i)$ under p. Let us further define the family of initial and transition distributions for the continuous states $\boldsymbol{z}_t$:

$$\Pi_{\mathcal{A}} := \{p_a(\boldsymbol{z}_1)|a \in \mathcal{A}\}, \mathcal{P}_{\mathcal{A}} := \{p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})|a \in \mathcal{A}\}. \quad (6)$$

where $\mathcal{A}$ is an index set satisfying mild measure-theoretic conditions (Appendix A). Note $\mathcal{P}_{\mathcal{A}}$ assumes first-order Markov dynamics. Then, we can construct the family of first-order MSMs as

$$\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}}) := \Bigg\{ \sum_{i=1}^{K^T} c_i p_{a_1^i}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) \mid K < +\infty,\ p_{a_1^i} \in \Pi_{\mathcal{A}},\ p_{a_t^i} \in \mathcal{P}_{\mathcal{A}},\ t \geq 2,\ a_t^i \in \mathcal{A}, \quad a_{1:T}^i \neq a_{1:T}^j, \forall i \neq j, \quad \sum_{i=1}^{K^T} c_i = 1 \Bigg\}. \quad (7)$$

Since Eq. (7) requires $a_{1:T}^i \neq a_{1:T}^j$ for any $i \neq j$, this also builds an injective mapping $\phi(i) = a_{1:T}^i$ from $i \in \{1, ..., K^T\}$ to $\mathcal{A}$. Combined with the path indexing function, this establishes an injective mapping $\phi \circ \varphi^{-1}$ to uniquely map a set of states $\mathbf{s}_{1:T}$ to the $a_{1:T}$ indices, and we can view $p_{a_1^i}(\boldsymbol{z}_1)$ and $p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ as equivalent notations of $p(\boldsymbol{z}_1|s_1)$ and $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$ respectively for $\mathbf{s}_{1:T} = \varphi(i)$. This notation shows that the MSM extends finite mixture models to temporal settings as a finite mixture of K^T trajectories composed by distributions in $\Pi_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{A}}$.

Having established the finite mixture model view of Markov switching models, we will use this notation of MSM in the rest of Section 3.1, as we will use finite mixture modelling techniques to establish its identifiability. In detail, we first define the identification of $\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ as follows.

Definition 3.1. The family of first-order MSMs $\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ is said to be identifiable up to permutations, when for $p_1, p_2 \in \mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ with

$$p_1(\boldsymbol{z}_{1:T}) = \textstyle\sum_{i=1}^{K^T} c_i p_{a_1^i}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}),$$
$$p_2(\boldsymbol{z}_{1:T}) = \textstyle\sum_{i=1}^{\hat{K}^T} \hat{c}_i p_{\hat{a}_1^i}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{\hat{a}_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}),$$

$p_1(\boldsymbol{z}_{1:T}) = p_2(\boldsymbol{z}_{1:T}), \forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}$, iff $K = \hat{K}$ and for each $1 \leq i \leq K^T$ there is some $1 \leq j \leq \hat{K}^T$ s.t.

1. $c_i = \hat{c}_j$;
2. if $a_{t_1}^i = a_{t_2}^i$ for $t_1, t_2 \geq 2$ and $t_1 \neq t_2$, then $\hat{a}_{t_1}^j = \hat{a}_{t_2}^j$;
3. $p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = p_{\hat{a}_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}), \forall t \geq 2$, $\boldsymbol{z}_t \in \mathbb{R}^m$;
4. $p_{a_1^i}(\boldsymbol{z}_1) = p_{\hat{a}_1^j}(\boldsymbol{z}_1), \forall \boldsymbol{z}_1 \in \mathbb{R}^m$.

The 2nd requirement eliminates the permutation equivalence of e.g., $\mathbf{s}_{1:4} = (1, 2, 3, 2)$; $\hat{\mathbf{s}}_{1:4} = (3, 1, 2, 3)$ which would be valid in the finite mixture case with vector indexing.

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

$$\mathcal{G}_{\mathcal{A}} = \{p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}(\boldsymbol{z}_{t-1}, a), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a)) \mid a \in \mathcal{A}, \boldsymbol{z}_t, \boldsymbol{z}_{t-1} \in \mathbb{R}^m\}, \quad (8)$$

where $\boldsymbol{m}(\boldsymbol{z}_{t-1}, a)$ and $\boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a)$ are non-linear w.r.t. $\boldsymbol{z}_{t-1}$ and denote the mean and covariance matrix, respectively. We further require the *unique indexing* assumption:

$$\forall a \neq a' \in \mathcal{A},\ \exists \boldsymbol{z}_{t-1} \in \mathbb{R}^d,\ s.t.\ \boldsymbol{m}(\boldsymbol{z}_{t-1}, a) \neq \boldsymbol{m}(\boldsymbol{z}_{t-1}, a') \text{ or } \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a) \neq \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a'). \quad (9)$$

In other words, for such $\boldsymbol{z}_{t-1}$, $p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ and $p_{a'}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ are two different Gaussian distributions. We also introduce a family of *initial distributions* with unique indexing:

$$\mathcal{I}_{\mathcal{A}} := \{p_a(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1; \boldsymbol{\mu}(a), \boldsymbol{\Sigma}_1(a)) \mid a \in \mathcal{A}\}, \quad (10)$$

$$a \neq a' \in \mathcal{A} \Leftrightarrow \boldsymbol{\mu}(a) \neq \boldsymbol{\mu}(a') \text{ or } \boldsymbol{\Sigma}_1(a) \neq \boldsymbol{\Sigma}_1(a'). \quad (11)$$

The above Gaussian families paired with unique indexing assumptions satisfy conditions which favour identifiability of first-order MSMs under non-linear Gaussian transitions.

Theorem 3.2. *Define the following first-order Markov switching model family under the non-linear Gaussian families, $\mathcal{M}_{NL}^T = \mathcal{M}^T(\mathcal{I}_{\mathcal{A}}, \mathcal{G}_{\mathcal{A}})$ with $\mathcal{G}_{\mathcal{A}}$, $\mathcal{I}_{\mathcal{A}}$ defined by Eqs. (8), (10) respectively. Then, the MSM is identifiable in terms of Def. 3.1 under the following assumptions:*

(a1) Unique indexing for $\mathcal{G}_{\mathcal{A}}$ and $\mathcal{I}_{\mathcal{A}}$: Eqs. (9), (11) hold;

(a2) For any $a \in \mathcal{A}$, the mean and covariance in $\mathcal{G}_{\mathcal{A}}$, $\boldsymbol{m}(\cdot, a) : \mathbb{R}^m \to \mathbb{R}^m$ and $\boldsymbol{\Sigma}(\cdot, a) : \mathbb{R}^m \to \mathbb{R}^{m \times m}$, are analytic functions.

The proof strategy can be summarised in 4 steps.

(1) Under the finite mixture model view, it suffices to show $\{p_{a^i_{1:T}}(\boldsymbol{z}_{1:T}) \mid a^i_{1:T} \in \mathcal{A} \times \cdots \times \mathcal{A}\}$ contains linearly independent functions (Yakowitz & Spragins, 1968).

(2) Since $p_{a^i_{1:T}}(\boldsymbol{z}_{1:T})$ is first-order Markov, we just need to find conditions for the linear independence of $\{p_{a^i_{1:2}}(\boldsymbol{z}_{1:2})\}$, and prove $T \geq 3$ cases by induction.

(3) For *non-parametric* $\{p_{a^i_{1:2}}(\boldsymbol{z}_{1:2})\}$ case, we specify conditions on $\mathcal{P}_{\mathcal{A}} = \{p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})\}$ and $\Pi_{\mathcal{A}} = \{p_a(\boldsymbol{z}_1)\}$ to ensure linear independence of $\{p_{a^i_{1:2}}(\boldsymbol{z}_{1:2})\}$.

(4) We show the MSM family with (non-linear) Gaussian transitions $\mathcal{M}_{NL}^T = \mathcal{M}^T(\mathcal{I}_{\mathcal{A}}, \mathcal{G}_{\mathcal{A}})$ is a special case of (3) when assuming (a1 - a2).

Intuitively, "switches" can be identified as discontinuities in the dynamic model and are fully controlled by the discrete states (assumptions a1, a2). Also assumption (a2) means $\boldsymbol{m}(\cdot, a)$ and $\boldsymbol{m}(\cdot, a')$ with $a \neq a'$ can only intercept at a set of points with zero Lebesgue measure (similarly for $\boldsymbol{\Sigma}(\cdot, a)$), and this smoothness property contributes to the identifiability for the continuous-state transitions $p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$.

The central part of our proof strategy involves showing linear independence in products of consecutive variables; e.g.,

$$\Pi_{\mathcal{A}} \otimes \mathcal{P}_{\mathcal{A}} \otimes \mathcal{P}_{\mathcal{A}} = \left\{ p_{a_1}(\boldsymbol{z}_1) p_{a_2}(\boldsymbol{z}_2|\boldsymbol{z}_1) p_{a_3}(\boldsymbol{z}_3|\boldsymbol{z}_2) \right\}, \quad (12)$$

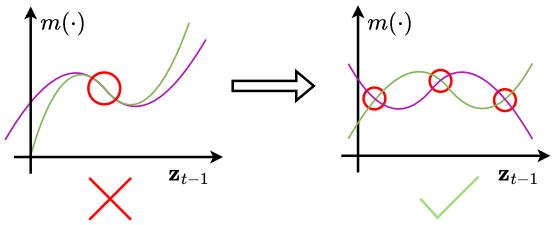


Figure 2: Illustration of the intuition behind Lemma B.3, where linear independence holds if, for any pair of functions (shown in green and purple), the intersection in the domain of the conditioned variable ($\boldsymbol{z}_{t-1}$) is zero-measured.

where the overlapping variable (indicated with different colours) challenges linear independence for any consecutive product of linearly independent functions. Figure 2 illustrates the intuition behind our main linear independence result, Lemma B.3, showing that linear independence of the product function can be achieved, as long as for the function family of each component, linear dependence only happens in zero-measure sets. Our result only requires the above to hold in at least a non-zero measured set of the overlapping variable space, and it connects to non-linear Gaussians via assumption (d3) in Theorem B.8. However, the parametric assumption (a2) is stronger, as it ensures the non-zero measure intersection is satisfied almost everywhere.

The following indications of Thm. 3.2, when combined, show the profound expressivity of identifiable MSMs:

- Assumption (a2) enables parameterising $p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ with e.g., polynomials and neural networks (NNs) with analytic activations (e.g. SoftPlus). 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 $p(\mathbf{s}_{1:T})$, allowing e.g., non-stationarity and higher-order Markov dependencies.

Our experiments consider $p(\mathbf{s}_{1:T})$ to be stationary and first-order Markov, where its state transitions are also identifiable (Appendix C). We leave other cases to future work.

Remark. Key to the full proof is the establishment of step (3), which is proved via a new theoretical result on linear independence (Lemma B.3), 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. 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 non-parametric MSMs, rendering our development of new theory (Lemma B.3) necessary.

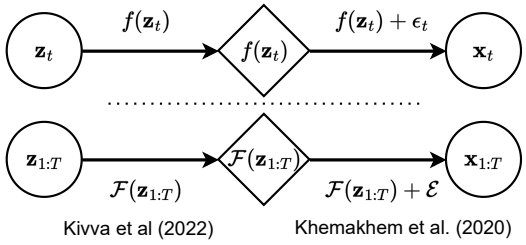


Figure 3: We assume $\boldsymbol{z}_t$ is transformed via f with noise $\boldsymbol{\epsilon}_t$ at each time-step t independently. We view this as a transformation on $\boldsymbol{z}_{1:T}$ via a factored $\mathcal{F}$ with noise $\mathcal{E}$.

3.2. Identifiable Switching Dynamical Systems

Based on identifiable MSMs, our analysis of Switching Dynamical Systems extends the setup from Kivva et al. (2022) (Section 2.1) to the temporal case. Assume the prior dynamic model $p(\boldsymbol{z}_{1:T})$ belongs to the non-linear Gaussian MSM family $\mathcal{M}_{NL}^T$ specified by Thm. 3.2. 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 piece-wise linear transformation f. The generation process of such SDS model can be expressed as follows:

$$\boldsymbol{x}_t = f(\boldsymbol{z}_t) + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{13}$$

$$\boldsymbol{z}_{1:T} \sim p(\boldsymbol{z}_{1:T}) \in \mathcal{M}_{NL}^T. \tag{14}$$

Note that we can also write the decoding process as $\boldsymbol{x}_{1:T} = \mathcal{F}(\boldsymbol{z}_{1:T}) + \mathcal{E}$ with $\mathcal{E} = (\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_T)^\top$, where $\mathcal{F}$ is a factored transformation composed by f as defined below. Importantly, this notion allows $\mathcal{F}$ to inherit e.g., piece-wise linearity and (weakly) injectivity properties from f.

Definition 3.3. We say that a function $\mathcal{F} : \mathbb{R}^{mT} \to \mathbb{R}^{nT}$ is factored if it is composed by $f : \mathbb{R}^m \to \mathbb{R}^n$, such that for any $\boldsymbol{z}_{1:T} \in \mathbb{R}^{mT}$, $\mathcal{F}(\boldsymbol{z}_{1:T}) = (f(\boldsymbol{z}_1), \ldots, f(\boldsymbol{z}_T))^\top$.

The identifiability analysis of the SDS model (Eq. (13)) follows two steps (see Figure 3). (i) Following Kivva et al. (2022), we establish the identifiability for a prior $p(\boldsymbol{z}_{1:T}) \in \mathcal{M}_{NL}^T$ from $(\mathcal{F}_\# p)(\boldsymbol{x}_{1:T})$ in the noiseless case. Here $\mathcal{F}_\# p$ denotes the pushforward measure of p by $\mathcal{F}$. (ii) When the noise $\mathcal{E}$ distribution is known, $\mathcal{F}_\# p$ is identifiable from $(\mathcal{F}_\# p) * \mathcal{E}$ using convolution tricks from Khemakhem et al. (2020). In detail, we first define the notion of identifiability for $p(\boldsymbol{z}_{1:T}) \in \mathcal{M}^T(\Pi_\mathcal{A}, \mathcal{P}_\mathcal{A})$ given noiseless observations.

Definition 3.4. Given a family of factored transformations $\mathbb{F}$, for $\mathcal{F} \in \mathbb{F}$ the prior $p \in \mathcal{M}^T(\Pi_\mathcal{A}, \mathcal{P}_\mathcal{A})$ is said to be identifiable up to affine transformations, when for any $\mathcal{F}' \in \mathbb{F}$ and $p' \in \mathcal{M}^T(\Pi_\mathcal{A}, \mathcal{P}_\mathcal{A})$ s.t. $\mathcal{F}_\# p = \mathcal{F}'_\# p'$, there exists an invertible factored affine mapping $\mathcal{H} : \mathbb{R}^{mT} \to \mathbb{R}^{mT}$ composed by $h : \mathbb{R}^m \to \mathbb{R}^m$, where $p \equiv \mathcal{H}_\# p'$.

For the factored $\mathcal{F}$, *identifiability up to affine transformations* extends from Def. 2.2.1 in Kivva et al. (2022), which is defined on f. I.e., for factored mappings $\mathcal{F}$, $\mathcal{F}'$ composed by f, f', respectively, if $f = (f' \circ h)$, where $h : \mathbb{R}^m \to \mathbb{R}^m$ is an invertible affine transformation, then $\mathcal{F} = (\mathcal{F}' \circ \mathcal{H})$ with $\mathcal{H}$ a factored mapping composed by h. Now we can state the following identifiability result for (non-linear) SDS (proof in Appendix D).

Theorem 3.5. *Assume the observations $\boldsymbol{x}_{1:T}$ are generated from a latent variable model following Eq. (13), whose prior $p(\boldsymbol{z}_{1:T})$ belongs to $\mathcal{M}_{NL}^T = \mathcal{M}^T(\mathcal{I}_\mathcal{A}, \mathcal{G}_\mathcal{A})$ and satisfies (a1 - a2), and f is a piece-wise linear mapping. Then:*

(i) If f, which composes $\mathcal{F}$, is weakly injective (Def. 2.1), the prior $p(\boldsymbol{z}_{1:T})$ is identifiable up to affine transformations as defined in Def. 3.4.

(ii) If f, which composes $\mathcal{F}$, is continuous and injective (Def. 2.2), both prior $p(\boldsymbol{z}_{1:T})$ and f are identifiable up to affine transformations (Def. 3.4).

Thm. 3.5 allows identifiable SDSs to employ e.g., SoftPlus networks for $p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$, and certain (Leaky) ReLU networks for f (see Kivva et al. (2022)). Again, for neural networks identifiability refers only to their functional form.

Remark. The MSM family is closed under factored affine transformations (Prop. C.2) and the proved MSM identifiability result (Thm. 3.2) is up to permutation equivalence. This means for $p_1, p_2 \in \mathcal{M}_{NL}^T$, one can show that $p_1 = \mathcal{H}_\# p_2$ with an invertible factored affine transformation $\mathcal{H}$ composed by $h(\boldsymbol{z}) = A\boldsymbol{z} + \mathbf{b}, \boldsymbol{z} \in \mathbb{R}^m$. It indicates the following relations for $\boldsymbol{m}_i(\cdot, a)$ and $\boldsymbol{\Sigma}_i(\cdot, a)$, $i = 1, 2$, with $\sigma(\cdot)$ a permutation over indices $a \in \mathcal{A}$:

$$\boldsymbol{m}_1(\boldsymbol{z}, a) = A\boldsymbol{m}_2\left(A^{-1}(\boldsymbol{z} - \mathbf{b}), \sigma(a)\right) + \mathbf{b}, \tag{15}$$

$$\boldsymbol{\Sigma}_1(\boldsymbol{z}, a) = A\boldsymbol{\Sigma}_2\Big(A^{-1}\left(\boldsymbol{z} - \mathbf{b}\right), \sigma(a)\Big)A^T. \tag{16}$$

3.3. On Identifiability Conditions for Sequential LVMs

Key to the design of identifiable models is the specification of constraints to eliminate unwanted equivalences or symmetries. For deep sequential latent variable models, these conditions are required for (i) the temporal dependencies of the latent dynamic model $p(\boldsymbol{z}_{1:T})$, and (ii) the emission model $p(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})$ to translate the latent variables to observations. We compare our proposed identifiability conditions with existing work, with flexible/restrictive conditions highlighted in different colours.

- Identifiable SDS (ours): An MSM dynamic model $p(\boldsymbol{z}_{1:T})$ with smoothness conditions on $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$ but no restrictions on $p(s_{1:T})$. The emission model is restricted to use (weakly) injective mappings.
- HMM-ICA (Hälvä & Hyvarinen, 2020): Latent and stationary HMM priors following Gassiat et al. (2016), which requires stationarity and bijective emissions.
- SNICA (Hälvä et al., 2021): Includes SDS under general conditions on latent temporal dependencies (*weak*

stationarity) but restricts the emission to use injective mappings. Identifiability does not cover transitions.

- IIA-HMM (Morioka et al., 2021): Introduces a recurrent structure in the emissions with bijectivity of the demixing function. Allows non-stationary innovations using regime-switching sources from a stationary HMM Gassiat et al. (2016).
- Mechanism Sparsity (Lachapelle et al., 2022; 2024): A temporal model $p(\boldsymbol{z}_{1:T}|\boldsymbol{u}_{1:T})$ with auxiliary observations $\boldsymbol{u}_{1:T}$ and sparse interactions within $\boldsymbol{z}_{1:T}$. The emission model is restricted to use diffeomorphisms.

Overall, often relaxed constraints for one part of the latent dynamic model need to be compensated with restrictive conditions on the other components. The above works share a limitation on the use of at least (weakly) injective mappings for the emission model, which is a common issue in identifiable DGM research since Khemakhem et al. (2020) and thus requires substantial future work to address.

4. Model Parameter Estimation

We consider two modelling choices for N sequences $\mathcal{D} = \{\boldsymbol{x}_{1:T}\}$ of length T: (a) the MSM model (Eq. (4) with $\boldsymbol{z}_{1:T}$ replaced by $\boldsymbol{x}_{1:T}$) and (b) the SDS model with MSM latent prior (Eqs. (3),(4)). Although our theory imposes no restrictions on the discrete latent state distribution, the methods are implemented with a first-order stationary Markov chain, i.e., $p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}) = p_{\boldsymbol{\theta}}(s_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(s_t|s_{t-1})$.

Markov Switching Models We use expectation maximisation (EM) for efficient estimation of mixture models (Bishop, 2006). Below we discuss the M-step update of the transition distribution parametrised by neural networks; more details (including polynomial parametrisations) can be found in Appendix E. When considering analytic neural networks, we take a Generalised EM (GEM) (Dempster et al., 1977) approach where a gradient ascent step is performed

$$\boldsymbol{\theta}^{\text{new}} \leftarrow \boldsymbol{\theta}^{\text{old}} + \eta \sum_{t=2}^{T}\sum_{k=1}^{K} \gamma_{t,k} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, s_t = k), \quad (17)$$

where $\gamma_{t,k} = p_{\boldsymbol{\theta}}(s_t = k|\boldsymbol{x}_{1:T})$ and the update rule can be computed using back-propagation. In the equation we indicate the update rule for a single sample, but note that we use mini-batch stochastic gradient ascent when N is large. Convergence is guaranteed for this approach to a local maximum of the likelihood (Bengio & Frasconi, 1994).

Switching Dynamical Systems We adopt variational inference as presented in Ansari et al. (2021). The parameters are learned by maximising the evidence lower bound (ELBO) (Kingma & Welling, 2014), and the proposed approximate posterior over the latent variables $\{\boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}\}$ incorporates the exact posterior of the discrete latent states given the continuous latent variables:

$$q_{\phi,\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}|\boldsymbol{x}_{1:T}) = q_\phi(\boldsymbol{z}_1|\boldsymbol{x}_{1:T}) \prod_{t=2}^{T} q_\phi(\boldsymbol{z}_t|\boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:T}) p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}|\boldsymbol{z}_{1:T}). \quad (18)$$

As in Ansari et al. (2021); Dong et al. (2020), the variational posterior over the continuous variables $q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$ simulates a smoothing process by first using a bi-directional RNN on $\boldsymbol{x}_{1:T}$, and then a forward RNN on the resulting embeddings. By introducing an exact posterior, the discrete latent variables can be marginalised from the ELBO objective (see Appendix E.2 for details),

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})}\Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})\Big] - KL\Big(q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})\Big). \quad (19)$$

We use Monte Carlo estimation with samples $\boldsymbol{z}_{1:T} \sim q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$ and the reparametrization trick for back-propagation (Kingma & Welling, 2014), and jointly learn the parameters using stochastic gradient ascent on the ELBO objective. The prior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$ can be computed exactly using the forward algorithm (Bishop, 2006) with messages $\{\alpha_{t,k}(\boldsymbol{z}_{1:t}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:t}, s_t = k)\}$ by marginalising out $\mathbf{s}_{1:T}$:

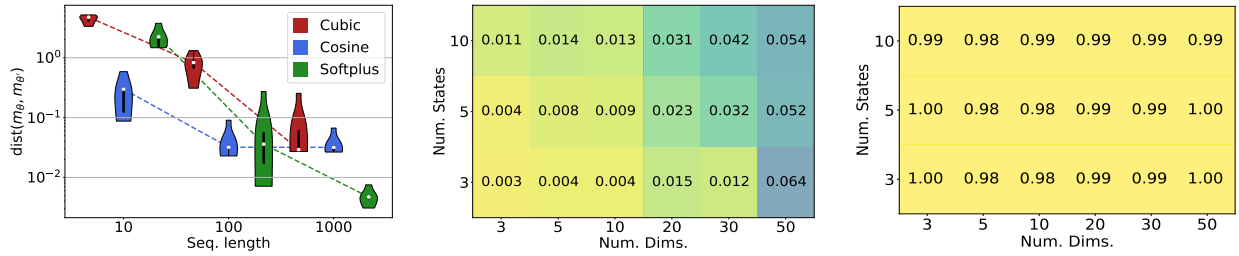
$$p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) = \sum_{k=1}^{K} \alpha_{T,k}(\boldsymbol{z}_{1:T}), \quad (20)$$

$$\alpha_{t,k}(\boldsymbol{z}_{1:t}) = \sum_{k'=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t = k) \times p_{\boldsymbol{\theta}}(s_t = k|s_{t-1} = k')\alpha_{t-1,k'}(\boldsymbol{z}_{1:t-1}). \quad (21)$$

An alternative approach for SDS inference is presented in Dong et al. (2020), although Ansari et al. (2021) outlines some disadvantages (see Appendix E.2 for a discussion). Note that estimating parameters with ELBO maximisation has no consistency guarantees in general, and we leave additional analyses regarding consistency for future work.

5. Related Work

Sequential generative models based on non-linear SDSs have gained interest over the recent years. Notable works include structured VAEs (Johnson et al., 2016), soft-switching Kalman Filters (Fraccaro et al., 2017), or recurrent SDSs (Linderman et al., 2016; Dong et al., 2020) which can include explicit duration models on the switches (Ansari et al., 2021). However, identifiability in such highly non-linear scenarios is rarely studied. A similar situation is found for MSMs, which were first introduced by Poritz (1982) and have been studied decades ago for speech analysis (Poritz,



(a) Function forms (b) L_2 distance (c) Avg. F_1 score

Figure 4: Synthetic experiment results on MSMs. (a) L_2 distance error using different transition functions with varying T. (b) L_2 distance error and (c) averaged F_1 score of non-linear data (cosine activations) with increasing states and dimensions.

1982; Ephraim & Roberts, 2005) and economics (Hamilton, 1989). Although studies have been conducted on maximum likelihood estimation for some first-order and stationary MSMs and their asymptotic properties (Frühwirth-Schnatter & Frèuhwirth-Schnatter, 2006), identifiability for general high-order autoregressive MSMs has not been proved. The main complication arises from the explicit dependency on the observed variables, which poses a great challenge to prove linear independence of $p(\boldsymbol{z}_{1:T}|\boldsymbol{s}_{1:T})$. To our knowledge, identifiability in MSMs has been explicitly studied only in An et al. (2013) which assumes discrete $\boldsymbol{z}_t$ states and uses the joint probability of four consecutive observations.

Regarding non-linear ICA for time series, identifiability results extend from Khemakhem et al. (2020) where past values are used as auxiliary information (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Klindt et al., 2021). Specifically, Yao et al. (2022a;b) recover temporal latent causal variables and allow non-stationarity via distribution shifts or time-dependent effects across observed regimes. Another line of work leverages intervention targets (Lippe et al., 2023b) or unknown binary interactions with regime information (Lippe et al., 2023a). Again these identifiability results require at least injectivity for the decoder function – see Section 3.3 for further discussions on this limitation.

6. Experiments

We evaluate the identifiable MSMs and SDSs with three experiments: (1) simulation studies with ground truth available for verification of the identifiability results; (2) regime-dependent causal discovery in climate data with identifiable MSMs; and (3) segmentation of high-dimensional sequences of salsa dancing using MSMs and SDSs. Additional results are also presented in Appendix F. Code for inference and data generation can be found in `https://github.com/charlio23/identifiable-SDS`.

6.1. Synthetic Experiments

MSMs We generate data using ground-truth MSMs and evaluate the estimated functions upon them. We parametrise the transition distribution means using random cubic polynomials or networks with cosine/SoftPlus activations; and use fixed transition covariances. When increasing dimensions, we use locally connected networks (Zheng et al., 2018) to construct regime-dependent causal structures for the grouth-truth MSMs, which encourages sparsity and stable data generation. The estimation error is computed with L_2 distance between grounth-truth and estimated transition mean functions. We also estimate the causal structure via thresholding the Jacobian of the estimated transition function, and compute the F_1 score to evaluate structure estimation accuracy. Both evaluation metrics are averaged over K components after accounting for permutation equivalence. See Appendices F.1 to F.4 for experimental details.

Figure 4a shows increasing the sequence length generally reduces the L_2 estimation error. The polynomials are estimated with higher error for short sequences, which could be caused by the high-frequency components from the cubic terms. Meanwhile the smoothness of the softplus networks allows the MSM to achieve consistently better parameter estimates. Regarding scalability, Figure 4b shows low estimation errors when increasing the number of dimensions and components. For structure estimation acurracy, Figure 4c shows that the MSM with non-linear transitions is able to maintain high F_1 scores, despite the differences in L_2 distance when increasing dimensions and states (4b). Although the approach is restricted by first-order Markov assumptions, the synthetic setting shows promising directions for high-dimensional regime-dependent causal discovery.

SDSs We use data from 2D MSMs with cosine activations for the transition means, fixed covariances, and $K = 3$ components, and a random Leaky ReLU network to generate observations (with no additive noise). For evaluation, we generate 1000 test samples and compute the F_1 score of the state posterior with respect the ground truth component. We also compute the L_2 distance of $\boldsymbol{m}(\boldsymbol{z}, \cdot)$ using Eq. (15). Again both metrics are computed after accounting for permutation and affine transform equivalence (Appendix F.1). Results in Table 6 show that the method obtains high F_1 scores as well as a low L_2 distance between the estimated and ground-truth transition mean functions. We further apply identifiable SDSs to synthetic videos generated by

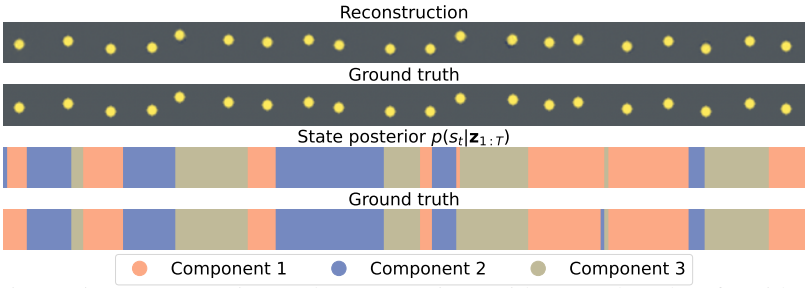


Figure 5: Reconstruction and segmentation (with ground truth) of a video generated from 2D latent variables sampled from a MSM (frame size 32×32).

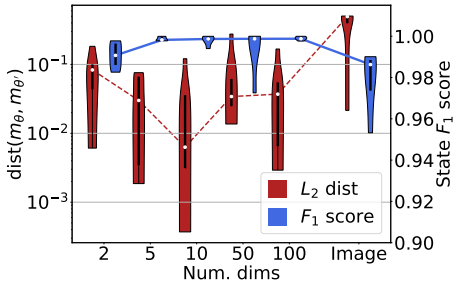
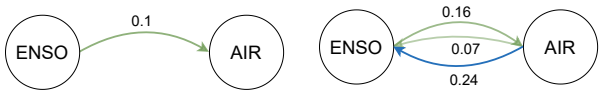


Figure 6: Synthetic experiment on SDSs for increasing $\boldsymbol{x}_t$ dimension.

Table 1: (weak) MCC scores (mean and 95-CI on 5 datasets) of iSDS in comparison to other nonlinear ICA baselines.

Model	MCC-strong	MCC-weak
iVAE*	0.642 ± 0.104	0.770 ± 0.059
Δ-SNICA	0.940 ± 0.041	0.979 ± 0.043
iSDS (ours)	0.936 ± 0.041	0.992 ± 0.008



(a) linear effects (b) non-linear effects

Figure 7: Regime-dependent graphs generated assuming (a) linear and (b) non-linear effects. Green and blue lines indicate effects in summer and winter months respectively.

treating the 2D latents as positions of a moving ball, and show a reconstruction with the corresponding segmentation in Figure 5. This high-dimensional setting increases the difficulty of accurate estimation as the reconstruction term of the ELBO out-weights the KL term for learning the latent MSM (Appendix F.3). This results in an increased L_2 distance from the ground truth MSM. Still, the estimated model achieves high-fidelity reconstructions (with an averaged pixel MSE of $8.89 \cdot 10^{-5}$), and accurate structure estimation as indicated by the F_1 score.

In Table 1 we compare our identifiable SDS (iSDS) with iVAE (Khemakhem et al., 2020), and Δ-SNICA (Hälvä et al., 2021) in the nonlinear ICA domain. For iVAE, which requires auxiliary observations, we use the ground-truth discrete states to create a strong non-temporal baseline (iVAE*). Using 50-dimensional test observations generated from the 2D latent MSMs, we compute the mean coefficient correlation (MCC-strong) as described in Khemakhem et al. (2020). Since our results achieve identifiability up to affine transformations, we also report the MCC scores after affine alignment of the latent variables (MCC-weak), as used in Kivva et al. (2022). Our approach achieves high MCC scores in both strong and weak settings. For Δ-SNICA, we observe successful disentanglement despite the linear prior dynamics differing from the groundtruth MSM structure based on cosine networks. This is due to the mean-field variational inference design from Johnson et al. (2016) and overparametrised linear SDSs, which allow flexible parametrisations but lose transition identifiability.

6.2. Regime-Dependent Causal Discovery

We explore regime-dependent causal discovery using climate data from Saggioro et al. (2020). The data consists on monthly observations of El Niño Southern Oscillation (ENSO) and All India Rainfall (AIR) from 1871 to 2016. We follow Saggioro et al. (2020) and train identifiable MSMs with linear and non-linear (softplus networks) transitions. Figure 7 shows the regime-dependent graphs extracted from both models, where the edge weights denote the absolute value of the corresponding entries in the estimated transition function's Jacobian (we keep edges with weights ≥ 0.05). The MSMs capture regimes based on seasonality, as one component is assigned to summer months (May - September), and the other is assigned to winter months (October - April). In the linear case, the MSM discovers an effect from ENSO to AIR which occurs only during Summer. This suggests that ENSO has a direct effect on AIR during summer, but not in winter, which is consistent with Saggioro et al. (2020) and Webster & Palmer (1997). For the non-linear case (Figure 7b), additional links are discovered which are yet to be supported by evidence in climate science. But as the flexibility of non-linear MSMs allows capturing correlations as causal effects in disguise, these additional links may imply the presence of confounders influencing both variables – a likely result for cases with few observations.

6.3. Segmentation of Dancing Patterns

We consider salsa dancing sequences from CMU mocap data and Hip-Hop videos from AIST Dance DB (Tsuchida et al., 2019) to demonstrate our models' ability in segmenting high-dimensional time-series. See Appendix for details on the data (F.6), training (F.3), and additional results (F.7).

Key-point data We present a sample using key-point data in Figure 9 with both the identifiable MSM (iMSM; softplus networks) and KVAE (Fraccaro et al., 2017). The iMSM

(a) Synthetic salsa videos (pixel MSE $2.26 \cdot 10^{-4}$ per frame). (b) AIST Hip-Hop videos (pixel MSE $7.85 \cdot 10^{-3}$ per frame).

Figure 8: Reconstructions and segmentations of (a) salsa dancing and (b) AIST Hip-Hop videos, where colour boxes indicate different components. The MSE results are computed on 560 and 100 test videos for salsa and Hip-Hop, respectively.

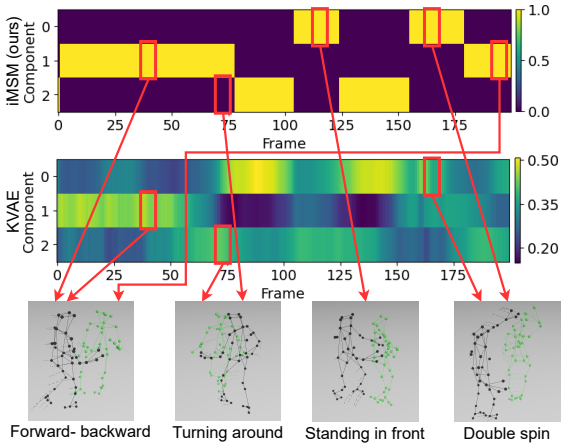


Figure 9: Posterior probability of a *salsa dancing* sequence of iMSM and KVAE (Fraccaro et al., 2017) along with several patterns distinguished in the example.

assigns different patterns to different states, e.g., the major pattern (forward-backward movements) is assigned to state 1. The iMSM also identifies this pattern at the end, which KVAE fails to recognise. Additionally, KVAE assigns a turning pattern into component 2, while iMSM treats turning as in state 1 patterns, but then jumps to component 2 consistently after observing it. The iMSM also classifies other dancing patterns into state 0. For limitations, the soft-switching mechanism restricts KVAE from confident component assignments. The iMSM's first-order Markov transitions hinders learning e.g., acceleration that would provide richer features for better segmentation.

Dancing videos We show reconstructions and segmentation results in Figure 8 to demonstrate identifiable SDS's applicability to videos. Here, we generate salsa videos by rendering 3D meshes from key-points using the renderer from Mahmood et al. (2019) (Figure 8a). Again in this experiment, one component is more prominent and used for the majority of the forward and backward patterns; and the other components are used to model spinning and other dancing patterns. Meanwhile segmentation results on AIST videos (Figure 8b) show more diverse patterns, illustrating a correlation between the discovered regimes and the moving position of the dancer and background information. The pixel MSE results also support identifiable SDSs as flexible sequential LVMs for video modelling.

7. Conclusion

We present identifiability analysis for Markov Switching Models and Switching Dynamical Systems. Key to our innovation is the establishment of MSM identifiability using Gaussian transitions with analytic functions, independently of the discrete state prior. We further extend the results to develop identifiable SDSs fully parameterised by neural networks. We verify our theory with synthetic experiments, and apply our models to regime-dependent causal discovery and high-dimensional time-series segmentation.

While our work focuses on identifiability, accurate estimation is also key to the success of causal discovery. Current estimation methods require intensive hyper-parameter tuning and are prone to state collapse in high-dimensional settings. Also variational learning for deep LVMs has no consistency guarantees unless assuming universal approximations of the q posterior (Gong et al., 2023), which disagrees with the popular use of Gaussian encoders. Future work should design efficient estimation methods, and extend the identifiability and estimation innovations to higher-order MSMs/SDSs.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

Yixin Wang was supported in part by the Office of Naval Research under grant number N00014-23-1-2590 and the National Science Foundation under Grant No. 2231174 and No. 2310831.

References

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

An, Y., Hu, Y., Hopkins, J., and Shum, M. Identifiability and inference of hidden markov models. Technical report, Technical report, 2013.

Ansari, A. F., Benidis, K., Kurle, R., Turkmen, A. C., Soh, H., Smola, A. J., Wang, B., and Januschowski, T. Deep explicit duration switching models for time series. *Advances in Neural Information Processing Systems*, 34: 29949–29961, 2021.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rk49Mg-CW.

Bengio, Y. and Frasconi, P. An input output hmm architecture. *Advances in neural information processing systems*, 7, 1994.

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

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Dong, Z., Seybold, B., Murphy, K., and Bui, H. Collapsed amortized variational inference for switching nonlinear dynamical systems. In *International Conference on Machine Learning*, pp. 2638–2647. PMLR, 2020.

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

Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *Advances in neural information processing systems*, 30, 2017.

Frühwirth-Schnatter, S. and Frèuhwirth-Schnatter, S. *Finite mixture and Markov switching models*, volume 425. Springer, 2006.

Gassiat, É., Cleynen, A., and Robin, S. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1):61–71, 2016.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=i_1rbq8yFWC.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

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

Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvarinen, A. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34: 1624–1633, 2021.

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

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hyvarinen, A. and Morioka, H. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

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

Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., , Datta, S. R., and Adams, R. P. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29, 2016.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *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.

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

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EbIDjBynYJ8.

Kruskal, J. B. 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.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Lachapelle, S., López, P. R., Sharma, Y., Everett, K., Priol, R. L., Lacoste, A., and Lacoste-Julien, S. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

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

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pp. 914–922. PMLR, 2017.

Linderman, S. W., Miller, A. C., Adams, R. P., Blei, D. M., Paninski, L., and Johnson, M. J. Recurrent switching linear dynamical systems. *arXiv preprint arXiv:1610.08466*, 2016.

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

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. BISCUIT: Causal representation learning from binary interactions. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1263–1273. PMLR, 31 Jul–04 Aug 2023a. URL https://proceedings.mlr.press/v216/lippe23a.html.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. 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.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., and Black, M. J. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL https://amass.is.tue.mpg.de.

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

Morioka, H., Hälvä, H., and Hyvarinen, A. Independent innovation analysis for nonlinear vector autoregressive process. In *International Conference on Artificial Intelligence and Statistics*, pp. 1549–1557. PMLR, 2021.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

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

Saxena, V., Ba, J., and Hafner, D. Clockwork variational autoencoders. In *Neural Information Processing Systems*, 2021.

Smith, J. T., Warrington, A., and Linderman, S. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3065601.

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

Webster, P. J. and Palmer, T. N. The past and the future of el niño. *Nature*, 390(6660):562–564, 1997.

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

Yao, W., Chen, G., and Zhang, K. Temporally disentangled representation learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=Vi-sZWNA_Ue.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RDlLMjLJXdq.

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

A. Non-parametric finite mixture models

We use the following existing result on identifying finite mixtures (Yakowitz & Spragins, 1968), which introduces the concept of linear independence to the identification of finite mixtures. Specifically, consider a distribution family that contains functions defined on $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathcal{F}_{\mathcal{A}} := \{F_a(\boldsymbol{x}) | a \in \mathcal{A}\} \tag{22}$$

where $F_a(\boldsymbol{x})$ is an d-dimensional CDF and $\mathcal{A}$ is a measurable index set such that $F_a(\boldsymbol{x})$ as a function of $(\boldsymbol{x}, a)$ is measurable on $\mathbb{R}^d \times \mathcal{A}$. In this paper, we assume this measure theoretic assumption on $\mathcal{A}$ is satisfied. Now consider the following finite mixture distribution family by linearly combining the CDFs in $\mathcal{F}$:

$$\mathcal{H}_{\mathcal{A}} := \{H(\boldsymbol{x}) = \sum_{i=1}^{N} c_i F_{a_i}(\boldsymbol{x}) | N < +\infty, a_i \in \mathcal{A}, a_i \neq a_j, \forall i \neq j, \sum_{i=1}^{N} c_i = 1\}. \tag{23}$$

Then we specify the definition of *identifiable finite mixture family* as follows:

Definition A.1. The finite mixture family $\mathcal{H}$ is said to be identifiable up to permutations, when for any two finite mixtures $H_1(x) = \sum_{i=1}^{M} c_i F_{a_i}(\boldsymbol{x})$ and $H_2(x) = \sum_{i=1}^{M} \hat{c}_i F_{\hat{a}_i}(\boldsymbol{x})$, $H_1(\boldsymbol{x}) = H_2(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, if and only if $M = N$ and for each $1 \leq i \leq N$ there is some $1 \leq j \leq M$ such that $c_i = \hat{c}_j$ and $F_{a_i}(\boldsymbol{x}) = F_{\hat{a}_j}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Then Yakowitz & Spragins (1968) proved the identifiability results for finite mixtures. To see this, we first introduce the concept of linearly independent functions under finite mixtures as follows.

Definition A.2. A family of functions $\mathcal{F} = \{f_a(\boldsymbol{x}) | a \in \mathcal{A}\}$ 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 $\{f_a(\boldsymbol{x}) | a \in \mathcal{A}_0\}$ are linearly independent.

This is a weaker requirement of linear independence on function classes as it allows linear dependency by representing one function as the linear combination of *infinitely many* other functions. With this relaxed definition of linear independence we state the identifiability result of finite mixture models as follows.

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

B. Proof of theorem 3.2

We follow the strategy described in the main text.

B.1. Identifiability via linear independence

Proposition A.3 can be directly generalised to CDFs defined on $\boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}$. Furthermore, if we have a family of PDFs[3], e.g. $\mathcal{P}_{\mathcal{A}}^T := \Pi_{\mathcal{A}} \otimes (\otimes_{t=2}^T \mathcal{P}_{\mathcal{A}})$, with linearly independent components, then their corresponding Tm-dimensional CDFs are also linearly independent (and vice versa). Therefore we have the following result as a direct extension of Proposition A.3.

Proposition B.1. *Consider the distribution family given by Eq. 7. Then the joint distribution in $\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ is identifiable up to permutations if and only if functions in $\mathcal{P}_{\mathcal{A}}^T$ are linearly independent under finite mixtures.*

The above assumption of linear independence under finite mixtures over the joint distribution implies the following identifiability result.

Theorem B.2. *Assume the functions in $\mathcal{P}_{\mathcal{A}}^T := \Pi_{\mathcal{A}} \otimes (\otimes_{t=2}^T \mathcal{P}_{\mathcal{A}})$ are linearly independent under finite mixtures, then the distribution family $\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ is identifiable as defined in Definition 3.1.*

Proof. From proposition B.1 we see that, $\mathcal{P}_{\mathcal{A}}^T$ being linearly independent implies identifiability up to permutation for $\mathcal{M}^T(\Pi_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}})$ in the finite mixture sense (Definition A.1). This means for $p_1(\boldsymbol{z}_{1:T})$ and $p_2(\boldsymbol{z}_{1:T})$ defined in Definition 3.1, we have $K = \hat{K}$ and for every $1 \leq i \leq K^T$, there exists $1 \leq j \leq \hat{K}^T$ such that $c_i = \hat{c}_j$ and

$$p_{a_1^i}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{a_t^i}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}) = p_{\hat{a}_1^j}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{\hat{a}_t^j}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}), \quad \forall \boldsymbol{z}_{1:T} \in \mathbb{R}^{Tm}.$$

[3] In this case we assume that the probability measures are dominated by the Lebesgue measure on $\mathbb{R}^{Tm}$ and the CDFs are differentiable.

This also indicates that $p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = p_{\hat{a}_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ for all $t \geq 2$, $\boldsymbol{z}_t, \boldsymbol{z}_{t-1} \in \mathbb{R}^m$, which can be proved by noticing that $p_a(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ are conditional PDFs. To see this, notice that as the joint distributions on $\boldsymbol{z}_{1:T}$ are equal, then the marginal distributions on $\boldsymbol{z}_{1:T-1}$ are also equal:

$$p_{a_1^i}(\boldsymbol{z}_1)\prod_{t=2}^{T-1} p_{a_t^i}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = p_{\hat{a}_1^j}(\boldsymbol{z}_1)\prod_{t=2}^{T-1} p_{\hat{a}_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}), \quad \forall \boldsymbol{z}_{1:T-1} \in \mathbb{R}^{(T-1)m},$$

which immediately implies $p_{a_T^i}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1}) = p_{\hat{a}_T^j}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1}), \forall \boldsymbol{z}_{T-1}, \boldsymbol{z}_T \in \mathbb{R}^m$. Similar logic applies to the other time indices $t \geq 1$, which also implies $p_{a_1^i}(\boldsymbol{z}_1) = p_{\hat{a}_1^j}(\boldsymbol{z}_1)$ for all $\boldsymbol{z}_1 \in \mathbb{R}^m$.

Lastly, if there exists $t_1 \neq t_2$ such that $a_{t_1}^i = a_{t_2}^i$ but $\hat{a}_{t_1}^j \neq \hat{a}_{t_2}^j$, then the proved fact that, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^m$,

$$\begin{aligned} p_{\hat{a}_{t_1}^j}(\boldsymbol{z}_{t_1} = \boldsymbol{\beta}|\boldsymbol{z}_{t_1-1} = \boldsymbol{\alpha}) &= p_{a_{t_1}^i}(\boldsymbol{z}_{t_1} = \boldsymbol{\beta}|\boldsymbol{z}_{t_1-1} = \boldsymbol{\alpha}) \\ &= p_{a_{t_2}^i}(\boldsymbol{z}_{t_2} = \boldsymbol{\beta}|\boldsymbol{z}_{t_2-1} = \boldsymbol{\alpha}) \\ &= p_{\hat{a}_{t_2}^j}(\boldsymbol{z}_{t_2} = \boldsymbol{\beta}|\boldsymbol{z}_{t_2-1} = \boldsymbol{\alpha}), \end{aligned}$$

implies linear dependence of $\mathcal{P}_{\mathcal{A}}$, which contradicts to the assumption that $\mathcal{P}_{\mathcal{A}}^T$ are linearly independent under finite mixtures.

We show the contradiction by assuming the case where $\mathcal{P}_{\mathcal{A}}^{t-1}$ is linearly independent for some $t > 1$, and then we consider the linear independence on $\mathcal{P}_{\mathcal{A}}^t$. We should have

$$\sum_{i,j} \gamma_{ij} p_{a_{1:t-1}^i}(\boldsymbol{z}_{1:t-1}) p_{a_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = 0, \quad \forall \boldsymbol{z}_{1:t} \in \mathbb{R}^{(t-1)m} \times \mathbb{R}^m,$$

with $\gamma_{ij} = 0, \forall i, j$. We can swap the summations to observe that from linear dependence of $\mathcal{P}_{\mathcal{A}}$, we have $\gamma_{ij} \neq 0$, for some i and j such that $\sum_j \gamma_{ij} p_{a_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = 0$.

$$\sum_i \left(\sum_j \gamma_{ij} p_{a_t^j}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) \right) p_{a_{1:t-1}^i}(\boldsymbol{z}_{1:t-1}) = 0, \quad \forall \boldsymbol{z}_{1:t} \in \mathbb{R}^{(t-1)m} \times \mathbb{R}^m,$$

which satisfies the equation with $\gamma_{ij} \neq 0$ for some i and j and thus contradicts with the linear independence of $\mathcal{P}_{\mathcal{A}}^t$.

□

B.2. Linear independence for T=2

Following the strategy as described in the main text, the next step requires us to start from linear independence results for $T = 2$, and then extend to $T > 2$. We therefore prove the following linear independence result.

Lemma B.3. *Consider two families $\mathcal{U}_I := \{u_i(\mathbf{y}, \boldsymbol{x})|i \in I\}$ and $\mathcal{V}_J := \{v_j(\boldsymbol{z}, \mathbf{y})|j \in J\}$ with $\boldsymbol{x} \in \mathcal{X}, \mathbf{y} \in \mathbb{R}^{d_y}$ and $\boldsymbol{z} \in \mathbb{R}^{d_z}$. We further assume the following assumptions:*

(b1) Positive function values: $u_i(\mathbf{y}, \boldsymbol{x}) > 0$ for all $i \in I, (\mathbf{y}, \boldsymbol{x}) \in \mathbb{R}^{d_y} \times \mathcal{X}$. Similar positive function values assumption applies to $\mathcal{V}_J$: $v_j(\boldsymbol{z}, \mathbf{y}) > 0$ for all $j \in J, (\boldsymbol{z}, \mathbf{y}) \in \mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$.

(b2) Unique indexing: for $\mathcal{U}_I$, $i \neq i' \in I \Leftrightarrow \exists\, \boldsymbol{x}, \mathbf{y}$ s.t. $u_i(\boldsymbol{x}, \mathbf{y}) \neq u_{i'}(\boldsymbol{x}, \mathbf{y})$. Similar unique indexing assumption applies to $\mathcal{V}_J$;

(b3) Linear independence under finite mixtures on specific non-zero measure subsets for $\mathcal{U}_I$: for any non-zero measure subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $\mathcal{U}_I$ contains linearly independent functions under finite mixtures on $(\mathbf{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}$.

(b4) Linear independence under finite mixtures on specific non-zero measure subsets for $\mathcal{V}_J$: there exists a non-zero measure subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$, such that for any non-zero 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 $(\boldsymbol{z}, \mathbf{y}) \in \mathcal{Z} \times \mathcal{Y}'$;

(b5) 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 non-zero measure subset $\mathcal{Z} \subset \mathbb{R}^{d_z}$ and any subset $J_0 \subset J$ such that $|J_0| < +\infty$, $\{v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) | j \in J_0\}$ 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 $\boldsymbol{z} \in \mathbb{R}^{d_z}$.

(b6) Continuity for $\mathcal{V}_J$: for any $j \in J$, $v_j(\boldsymbol{z}, \mathbf{y})$ is continuous in $\mathbf{y} \in \mathbb{R}^{d_y}$.

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

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

$$\sum_{(i,j) \in S_0} \gamma_{ij} v_j(\boldsymbol{z}, \mathbf{y}) u_i(\mathbf{y}, \boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \mathbf{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}. \tag{24}$$

Note that the choices of S_0 and γ_{ij} are independent of any $\boldsymbol{x}, \mathbf{y}, \boldsymbol{z}$ values, but might be dependent on $\mathcal{Z}$. By assumptions (b1), 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 $v_j(\boldsymbol{z}, \mathbf{y})$ out:

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

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 to assumptions (b1) - (b3).

Now define $J_0 = \{j \in A | \exists (i,j) \in S_0\}$ 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. (24)) for any $\boldsymbol{\beta} \in \mathbb{R}^{d_y}$ as

$$\sum_{j \in J_0} \left(\sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) \right) v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{25}$$

From assumption (b3) we know that the set $\mathcal{Y}_0 := \{\boldsymbol{\beta} \in \mathbb{R}^{d_y} | \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{X}\}$ can only have zero measure in $\mathbb{R}^{d_y}$. Write $\mathcal{Y} \subset \mathbb{R}^{d_y}$ the non-zero measure subset defined by assumption (b4), we have $\mathcal{Y}_1 := \mathcal{Y} \backslash \mathcal{Y}_0 \subset \mathcal{Y}$ also has non-zero measure and satisfies assumption (b4). Combined with assumption (b1), we have for each $\boldsymbol{\beta} \in \mathcal{Y}_1$, there exists $\boldsymbol{x} \in \mathcal{X}$ such that $\sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) \neq 0$ for at least two j indices in J_0. This means for each $\boldsymbol{\beta} \in \mathcal{Y}_1$, $\{v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) | j \in J_0\}$ contains linearly dependent functions on $\boldsymbol{z} \in \mathcal{Z}$. Now under assumption (b5), 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:

$$\begin{aligned} J_0 = \cup_{k \in K(\boldsymbol{\beta})} J_k(\boldsymbol{\beta}), \quad J_k(\boldsymbol{\beta}) \cap J_{k'}(\boldsymbol{\beta}) = \emptyset, \forall k \neq k' \in K(\boldsymbol{\beta}), \\ j \neq j' \in J_k(\boldsymbol{\beta}) \quad \Leftrightarrow \quad v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}), \quad \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned} \tag{26}$$

Then we can rewrite Eq. (25) for any $\boldsymbol{\beta} \in \mathcal{Y}_1$ as

$$\sum_{k \in K(\boldsymbol{\beta})} \left(\sum_{j \in J_k(\boldsymbol{\beta})} \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) \right) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{27}$$

Recall from Eq. (26) 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. (24) holds, then for any $\boldsymbol{\beta} \in \mathcal{Y}_1$, under assumptions (b1) and (b5),

$$\sum_{j \in J_k(\boldsymbol{\beta})} \sum_{i:(i,j) \in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \quad k \in K(\boldsymbol{\beta}). \tag{28}$$

Define $C(\boldsymbol{\beta}) = \min_k |J_k(\boldsymbol{\beta})|$ the minimum cardinality count for j indices in the $J_k(\boldsymbol{\beta})$ subsets. Choose $\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\beta} \in \mathcal{Y}_1} C(\boldsymbol{\beta})$:

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 $v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta})$ for all $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{\beta} \in \mathcal{Y}_1$, so that they are linearly dependent on $(\boldsymbol{z}, \mathbf{y}) \in \mathcal{Z} \times \mathcal{Y}_1$, a contradiction to assumption (b4) by setting $\mathcal{Y}' = \mathcal{Y}_1$.

2. Now assume $|J_1(\boldsymbol{\beta}^*)| = C(\boldsymbol{\beta}^*)$ w.l.o.g.. From assumption (b5), we know that for any $j \in J_1(\boldsymbol{\beta}^*)$ and $j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*)$, $v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta})$ only on zero measure subset of $\mathcal{Z}$ at most. Then as $|J_0| < +\infty$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$ has non-zero measure, there exist $\boldsymbol{z}_0 \in \mathcal{Z}$ and $\delta > 0$ such that

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

   Under assumption (b6), there exists $\epsilon(j) > 0$ such that we can construct an ϵ-ball $B_{\epsilon(j)}(\boldsymbol{\beta}^*)$ using ℓ_2-norm, such that

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

   Choosing a suitable $0 < \epsilon \leq \min_{j \in J_0} \epsilon(j)$ (note that $\min_{j \in J_0} \epsilon(j) > 0$ as $|J_0| < +\infty$) and constructing an ℓ_2-norm-based ϵ-ball $B_\epsilon(\boldsymbol{\beta}^*) \subset \mathcal{Y}_1$, we have for all $j \in J_1(\boldsymbol{\beta}^*), j' \in J_0 \backslash J_1(\boldsymbol{\beta}^*)$, $j' \notin J_1(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ due to

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

   So this means for the split $\{J_k(\boldsymbol{\beta})\}$ of any $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$, we have $J_1(\boldsymbol{\beta}) \subset J_1(\boldsymbol{\beta}^*)$ and therefore $|J_1(\boldsymbol{\beta})| \leq |J_1(\boldsymbol{\beta}^*)|$. Now by definition of $\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\beta} \in \mathcal{Y}} C(\boldsymbol{\beta})$ and $|J_1(\boldsymbol{\beta}^*)| = C(\boldsymbol{\beta}^*)$, we have $J_1(\boldsymbol{\beta}) = J_1(\boldsymbol{\beta}^*)$ for all $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$.

3. One can show that $|J_1(\boldsymbol{\beta}^*)| = 1$, otherwise by definition of the split (Eq. (26)) and the above point, there exists $j \neq j' \in J_1(\boldsymbol{\beta}^*)$ such that $v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta})$ for all $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$, a contradiction to assumption (b4) by setting $\mathcal{Y}' = B_\epsilon(\boldsymbol{\beta}^*)$. 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 $\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}^*)$ means

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

   again a contradiction to assumption (b3) by setting $\mathcal{Y} = B_\epsilon(\boldsymbol{\beta}^*)$.

The above 3 points indicate that Eq. (28) cannot hold for all $\boldsymbol{\beta} \in \mathcal{Y}_1$ (and therefore for all $\boldsymbol{\beta} \in \mathcal{Y}$) under assumptions (b3) - (b6), therefore a contradiction is reached. □

B.3. Linear independence in the non-parametric case

The previous result can be used to show conditions for the linear independence of the joint distribution family $\mathcal{P}_{\mathcal{A}}^T$ in the non-parametric case.

Theorem B.4. *Define the following joint distribution family*

$$\left\{ p_{a_1, a_{2:T}}(\boldsymbol{z}_{1:T}) = p_{a_1}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{a_t}(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}), \quad p_{a_1} \in \Pi_{\mathcal{A}}, \quad p_{a_t} \in \mathcal{P}_{\mathcal{A}}, t = 2, ..., T \right\},$$

and assume $\Pi_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{A}}$ satisfy assumptions (b1)-(b6) as follows,

(c1) $\Pi_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{A}}$ satisfy (b1) and (b2): positive function values and unique indexing,

(c2) $\Pi_{\mathcal{A}}$ satisfies (b3), and

(c3) $\mathcal{P}_{\mathcal{A}}$ satisfies (b4)-(b6).

Then the following statement holds: For any $T \geq 2$ and any subset $\mathcal{Z} \subset \mathbb{R}^m$ The joint distribution family contains linearly independent distributions for $(\boldsymbol{z}_{1:T-1}, \boldsymbol{z}_T) \in \mathbb{R}^{(T-1)m} \times \mathcal{Z}$.

Proof. We proceed to prove the statement by induction as follows. Here we set $I = J = \mathcal{A}$.

(1) $T = 2$: The result can be proved using Lemma B.3 by setting in the proof, $u_i(\mathbf{y} = \boldsymbol{z}_1, \boldsymbol{x} = \boldsymbol{z}_0) = \pi_i(\boldsymbol{z}_1), i \in \mathcal{A}$ and $v_j(\boldsymbol{z} = \boldsymbol{z}_2, \mathbf{y} = \boldsymbol{z}_1) = p_j(\boldsymbol{z}_2|\boldsymbol{z}_1), j \in \mathcal{A}$.

(2) $T > 2$: Assume the statement holds for the joint distribution family when $T = \tau - 1$. Note that we can write $p_{a_{1:\tau}}(\boldsymbol{z}_{1:\tau})$ as

$$p_{a_{1:\tau}}(\boldsymbol{z}_{1:\tau}) = p_{a_1, a_{2:\tau-1}}(\boldsymbol{x}_{1:\tau-1}) p_{a_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}).$$

Then the statement for $T = \tau$ can be proved using Lemma B.3 by setting $u_i(\mathbf{y} = \boldsymbol{z}_{\tau-1}, \boldsymbol{x} = \boldsymbol{z}_{1:\tau-2}) = p_{a_{1:\tau-1}}(\boldsymbol{z}_{1:\tau-1}), i = a_{1:\tau-1}$, and $v_j(\boldsymbol{z} = \boldsymbol{z}_\tau, \mathbf{y} = \boldsymbol{z}_{\tau-1}) = p_{a_\tau}(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau-1}), j = a_\tau$. Note that the family spanned with $p_{a_{1:\tau-1}}(\boldsymbol{z}_{1:\tau-1}), i = a_{1:\tau-1}$ satisfies (b1) and (b2) from $\Pi_\mathcal{A}$ and $\mathcal{P}_\mathcal{A}$ directly, and (b3) from the induction hypothesis. □

With the result above, one can construct identifiable Markov Switching Models as long as the initial and transition distributions are consistent with assumptions (c1)-(c3).

B.4. Linear independence in the non-linear Gaussian case

As described, in the final step of the proof we explore properties of the Gaussian transition and initial distribution families (Eqs. (8) and (10) respectively). The unique indexing assumption of the Gaussian transition family (Eq. (9)) implies linear independence as shown below.

Proposition B.5. *Functions in $\mathcal{G}_A$ are linearly independent on variables $(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ if the unique indexing assumption (Eq. (9)) holds.*

Proof. Assume the statement is false, then there exists $\mathcal{A}_0 \subset \mathcal{A}$ and a set of non-zero values $\{\gamma_a | a \in \mathcal{A}_0\}$, such that

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

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 & Spragins (1968) proved that multivariate Gaussian distributions with different means and/or covariances are linearly independent. Therefore the equality above implies for any $\boldsymbol{z}_{t-1}$

$$\boldsymbol{m}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{m}(\boldsymbol{z}_{t-1}, a') \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a') \quad \forall a, a' \in A_0, a \neq a',$$

a contradiction to the unique indexing assumption. □

We now draw some connections from the previous Gaussian families to assumptions (b1-b6) in Lemma B.3.

Proposition B.6. *The conditional Gaussian distribution family $\mathcal{G}_\mathcal{A}$ (Eq. (8), under the unique indexing assumption (Eq. (9), satisfies assumptions (b1), (b2) and (b5) in Lemma B.3, if we define $\mathcal{V}_J := \mathcal{G}_\mathcal{A}, \boldsymbol{z} := \boldsymbol{z}_t$ and $\mathbf{y} := \boldsymbol{z}_{t-1}$.*

Proposition B.7. *The initial Gaussian distribution family $\mathcal{I}_\mathcal{A}$ (Eq. (10), under the unique indexing assumption (Eq. (11), satisfies assumptions (b1), (b2) and (b3) in Lemma B.3, if we define $\mathcal{U}_I := \mathcal{I}_\mathcal{A}, \mathbf{y} := \boldsymbol{z}_1$ and $\boldsymbol{x} = \mathcal{X} = \emptyset$.*

To see why $\mathcal{G}_\mathcal{A}$ satisfies (b5), notice Gaussian densities are analytic in $\boldsymbol{z}_t$. Similar ideas apply to show that $\mathcal{I}_\mathcal{A}$ satisfies (b3). With the previous results, we can rewrite the previous result for the non-linear Gaussian case.

Theorem B.8. *Define the following joint distribution family under the non-linear Gaussian model*

$$\mathcal{P}^T_\mathcal{A} = \left\{ p_{a_1, a_{2:T}}(\boldsymbol{z}_{1:T}) = p_{a_1}(\boldsymbol{z}_1) \prod_{t=2}^{T} p_{a_t}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}), a_t \in \mathcal{A}, \quad p_{a_1} \in \mathcal{I}_\mathcal{A}, \quad p_{a_t} \in \mathcal{G}_\mathcal{A}, \quad t = 2, ..., T \right\}, \tag{29}$$

with $\mathcal{G}_\mathcal{A}$, $\mathcal{I}_\mathcal{A}$ defined by Eqs. (8), (10) respectively. Assume:

(d1) Unique indexing for $\mathcal{G}_\mathcal{A}$ and $\mathcal{I}_\mathcal{A}$: Eqs. (9), (11) hold;

(d2) Continuity for the conditioning input: distributions in $\mathcal{G}_\mathcal{A}$ are continuous w.r.t. $\boldsymbol{z}_{t-1} \in \mathbb{R}^m$;

(d3) Zero-measure intersection in certain region: there exists a non-zero measure set $\mathcal{X}_0 \subset \mathbb{R}^m$ s.t. $\{\boldsymbol{z}_{t-1} \in \mathcal{X}_0 | \boldsymbol{m}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{m}(\boldsymbol{z}_{t-1}, a'), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a')\}$ has zero measure, for any $a \neq a'$;

Then, the joint distribution family contains linearly independent distributions for $(\boldsymbol{z}_{1:T-1}, \boldsymbol{z}_T) \in \mathbb{R}^{(T-1)m} \times \mathbb{R}^m$.

Proof. Note that assumptions (b1) - (b3) and (b5) are satisfied due to Propositions B.6 and B.7, and assumptions (b6) and (d2) are equivalent, and assumption (b4) holds due to assumption (d3). To show (d3) $\implies$ (b4), We first define $\mathcal{V}_J := \mathcal{G}_{\mathcal{A}}$, $\boldsymbol{z} := \boldsymbol{z}_t$, and $\mathbf{y} := \boldsymbol{z}_{t-1}$ from Prop. B.6. From (d3), $\mathcal{Y} := \mathcal{X}_0$ and note that $\mathcal{V}_J$ contains linear independent functions on $(\boldsymbol{z}, \mathbf{y}) \in \mathcal{M} \subset \mathcal{Z} \times \mathcal{Y}$ if $\mathcal{M} \neq \mathcal{Z} \times \mathcal{D}$, where $\mathcal{D}$ denotes the set where intersection of moments happen within $\mathcal{Y}$. Also by (d3), $\mathcal{D}$ has measure zero and thus, (b4) holds since $\mathcal{Y}'$ is a non-zero measure set.

Then, the statement holds by Theorem B.4. □

B.5. Concluding the proof

Below we formally state the proof for Theorem 3.2 by further assuming parametrisations of the Gaussian moments via analytic functions, i.e. assumption (a2).

Proof. (a1) and (d1) are equivalent. Following (a2), let $\boldsymbol{m}(\cdot, a) : \mathbb{R}^m \to \mathbb{R}^m$ be a multivariate analytic function, which allows a multivariate Taylor expansion. The corresponding Taylor expansion of $\boldsymbol{m}(\cdot, a)$ implies (d2). Similar logic applies to $\boldsymbol{\Sigma}(\cdot, a)$. To show (d3), we note for any $a \neq a'$ the set of intersection of moments, i.e. $\{\boldsymbol{z} \in \mathbb{R}^m | \boldsymbol{m}(\boldsymbol{z}, a) = \boldsymbol{m}(\boldsymbol{z}, a'), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a')\}$ can be separated as the intersection of the sets $\{\boldsymbol{z} \in \mathbb{R}^m | \boldsymbol{m}(\boldsymbol{x}, a) = \boldsymbol{m}(\boldsymbol{z}, a')\}$ and $\{\boldsymbol{z} \in \mathbb{R}^m | \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, a')\}$. Wlog, the set $\{\boldsymbol{z} \in \mathbb{R}^m | \boldsymbol{m}(\boldsymbol{z}, a) = \boldsymbol{m}(\boldsymbol{z}, a')\}$ is the zero set of an analytic function $\boldsymbol{f} := \boldsymbol{m}(\cdot, a) - \boldsymbol{m}(\cdot, a')$. Proposition 0 in Mityagin (2015) shows that the zero set of a real analytic function on $\mathbb{R}^m$ has zero measure unless $\boldsymbol{f}$ is identically zero. Hence, the intersection of moments has zero measure from our premise of unique indexing.

Since (d1-d3) are satisfied, by Theorem B.8 we have linear independence of the joint distribution family, which by Theorem B.2 implies identifiability of the MSM in the sense of Def. 3.1. □

C. Properties of the MSM

In this section, we present some results involving MSMs for convenience. First, we start with the result on first-order stationary Markov chains presented in section 3.1.

Corollary C.1. *Consider an identifiable MSM from Def. 3.1, where the prior distribution of the states $p(\mathbf{s}_{1:T})$ follows a first-order stationary Markov chain, i.e $p(\mathbf{s}_{1:T}) = \pi_{s_1} Q_{s_1, s_2} \dots Q_{s_{T-1}, s_T}$, where $\boldsymbol{\pi}$ denotes the initial distribution: $p(s_1 = k) = \pi_k$, and Q denotes the transition matrix: $p(s_t = k | s_{t-1} = l) = Q_{l,k}$. Then, $\boldsymbol{\pi}$ and Q are identifiable up to permutations.*

Proof. From Def. 3.1, we have $K = \hat{K}$ and for every $1 \leq i \leq K^T$ there is some $1 \leq j \leq \hat{K}^T$ such that $c_i = \hat{c}_j$. Now writing $\mathbf{s}_{1:T} = (s_1^i, ..., s_T^i) = \varphi(i)$ and $\hat{\mathbf{s}}_{1:T} = (\hat{s}_1^j, ..., \hat{s}_T^j) = \varphi(j)$, we have

$$c_i = \pi_{s_1^i} Q_{s_1^i, s_2^i} \dots Q_{s_{T-1}^i, s_T^i} = \hat{\pi}_{\hat{s}_1^j} \hat{Q}_{\hat{s}_1^j, \hat{s}_2^j} \dots \hat{Q}_{\hat{s}_{T-1}^j, \hat{s}_T^j} = \hat{c}_j.$$

Since the joint distributions are equal on $\mathbf{s}_{1:T}$, they must also be equal on $\mathbf{s}_{1:T-1}$. Therefore, we have $Q_{s_{T-1}^i, s_T^i} = \hat{Q}_{\hat{s}_{T-1}^j, \hat{s}_T^j}$. Similar logic applies to $t \geq 1$, which also implies $\pi_{s_1^i} = \hat{\pi}_{\hat{s}_1^j}$.

For $t = 1$, the above implies that for each $i \in \{1, ..., K\}$, there exists some $j \in \{1, ..., K\}$, such that $\pi_i = \hat{\pi}_j$. This indicates permutation equivalence. We denote $\sigma(\cdot)$ as such permutation function, so that for all $i \in \{1, ..., K\}$ and the corresponding j, $\pi_i = \hat{\pi}_j = \pi_{\sigma(j)}$.

For $t > 1$, the previous implication gives us that for $i, j \in \{1, \dots, K\}$, $\exists k, l \in \{1, \dots, K\}$ such that $Q_{i,j} = \hat{Q}_{k,l}$. Following the previous logic, we can define permutations that match Q and $\hat{Q}$: $i = \sigma_1(k), j = \sigma_2(l)$. We observe from the second requirement in Def. 3.1 that if $i = j$, then $k = l$ and since $\sigma_1(k) = \sigma_2(l)$, we have that the permutations $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ must be equal. Therefore, we have $Q_{i,j} = \hat{Q}_{k,l} = Q_{\sigma_1(k), \sigma_1(l)}$.

Finally, we can use the second requirement in Def. 3.1 to see that $\sigma(\cdot)$ and $\sigma_1(\cdot)$ must be equal. □

Proposition C.2. *The joint distribution of the Markov Switching Model with Gaussian analytic transitions and Gaussian initial distributions is closed under factored invertible affine transformations, $\boldsymbol{z}'_{1:T} = \mathcal{H}(\boldsymbol{z}_{1:T})$: $\boldsymbol{z}'_t = A\boldsymbol{z}_t + \mathbf{b}$, $1 \leq t \leq T$.*

Proof. Consider the following affine transformation $\boldsymbol{z}'_t = A\boldsymbol{z}_t + \mathbf{b}$ for $1 \leq t \leq T$, and the joint distribution of a Markov Switching Model with T timesteps

$$p(\boldsymbol{z}_{1:T}) = \sum_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}) p(\boldsymbol{z}_1|s_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t),$$

where we denote the initial distribution as $p(\boldsymbol{z}_1|s_1 = i) = \mathcal{N}(\boldsymbol{z}_1; \boldsymbol{\mu}(i), \boldsymbol{\Sigma}_1(i))$ and the transition distribution as $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t = i) = \mathcal{N}(\boldsymbol{z}_t; \mathbf{m}(\boldsymbol{z}_{t-1}, i), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, i))$. We need to show that the distribution still consists of Gaussian initial distributions and Gaussian analytic transitions. Let us consider the change of variables rule, which we apply to $p(\boldsymbol{z}_{1:T})$

$$p_{\boldsymbol{z}'_{1:T}}(\boldsymbol{z}'_{1:T}) = \frac{p_{\boldsymbol{z}_{1:T}}\left(A_{1:T}^{-1}\left(\boldsymbol{z}'_{1:T} - \mathbf{b}_{1:T}\right)\right)}{\det(A_{1:T})},$$

where we use the subscript $\boldsymbol{z}'_{1:T}$ to indicate the probability distribution in terms of $\boldsymbol{z}'_{1:T}$, but we drop it for simplicity. Note that the inverse of a block diagonal matrix can be computed as the inverse of each block, and we use similar properties for the determinant, i.e. $\det(A_{1:T}) = \det(A) \cdots \det(A)$. The distribution in terms of the transformed variable is expressed as follows:

$$\begin{aligned}
p(\boldsymbol{z}'_{1:T}) &= \sum_{s_{1:T}} p(s_{1:T}) \frac{p\left(A^{-1}\left(\boldsymbol{z}'_1 - \mathbf{b}\right)|s_1\right)}{\det(A)} \prod_{t=2}^{T} \frac{p\left(A^{-1}\left(\boldsymbol{z}'_t - \mathbf{b}\right) | \left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right)\right), s_t\right)}{\det(A)} \\
&= \sum_{i_1,\ldots,i_T} p\left(s_{1:T} = \{i_1, \ldots, i_T\}\right) \mathcal{N}\left(\boldsymbol{z}'_1; A\boldsymbol{\mu}(i_1) + \mathbf{b}, A\boldsymbol{\Sigma}_1(i_1)A^T\right) \\
&\quad \prod_{t=2}^{T} \frac{1}{\sqrt{(2\pi)^m \det\left(A\boldsymbol{\Sigma}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right)A^T\right)}} \\
&\quad \exp\Bigg(-\frac{1}{2}\Big(A^{-1}\left(\boldsymbol{z}'_t - \mathbf{b}\right) - \boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right)\Big)^T \\
&\quad \boldsymbol{\Sigma}\Big(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\Big)^{-1}\Big(A^{-1}\left(\boldsymbol{z}'_t - \mathbf{b}\right) - \boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right)\Big)\Bigg) \\
&= \sum_{i_1,\ldots,i_T} p\left(s_{1:T} = \{i_1, \ldots, i_T\}\right) \mathcal{N}\left(\boldsymbol{z}'_1; A\boldsymbol{\mu}(i_1) + \mathbf{b}, A\boldsymbol{\Sigma}_1(i_1)A^T\right) \\
&\quad \prod_{t=2}^{T} \frac{1}{\sqrt{(2\pi)^m \det\left(A\boldsymbol{\Sigma}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right)A^T\right)}} \\
&\quad \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{z}'_t - A\boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right) - \mathbf{b}\Big)^T \\
&\quad A^{-1}\boldsymbol{\Sigma}\Big(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\Big)^{-1} A^{-T}\Big(\boldsymbol{z}'_t - A\boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right) - \mathbf{b}\Big)\Bigg) \\
&= \sum_{i_1,\ldots,i_T} p\left(s_{1:T} = \{i_1, \ldots, i_T\}\right) \mathcal{N}\left(\boldsymbol{z}'_1; A\boldsymbol{\mu}(i_1) + \mathbf{b}, A\boldsymbol{\Sigma}_1(i_1)A^T\right) \\
&\quad \prod_{t=2}^{T} \mathcal{N}\Big(\boldsymbol{z}'_t; A\boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right) + \mathbf{b}, A\boldsymbol{\Sigma}\Big(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\Big)A^T\Big)
\end{aligned}$$

We observe that the resulting distribution is a Markov Switching Model with changes in the Gaussian initial and transition distributions, where the analytic transitions are transformed as follows: $bmm'(\boldsymbol{z}'_{t-1}, i_t) = A\boldsymbol{m}\left(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\right) + \mathbf{b}$, and $\boldsymbol{\Sigma}'(\boldsymbol{z}'_{t-1}, i_t) = A\boldsymbol{\Sigma}\Big(A^{-1}\left(\boldsymbol{z}'_{t-1} - \mathbf{b}\right), i_t\Big)A^T$ for any $i_t \in \{1, \dots, K\}$. □

D. Proof of SDS identifiability

D.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 D.1. 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 D.2. Let $\boldsymbol{x}_0 \in \mathbb{R}^m$ and $\delta > 0$. Let $p : B(\boldsymbol{x}_0, \delta) \to \mathbb{R}$. Define

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

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 D.3. Let $D_0 \subset D$ and $p : D \to \mathbb{R}$ be a function. We define $p|_{D_0} : D \to \mathbb{R}$ 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 D.4. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a piece-wise 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 $\boldsymbol{z} \in f^{-1}(\{\boldsymbol{x}\})$.

Lemma D.5. *If $f : \mathbb{R}^m \to \mathbb{R}^n$ is a piece-wise affine function such that $\{\boldsymbol{x} \in \mathbb{R}^n : |f^{-1}(\{\boldsymbol{x}\})| > 1\} \subseteq f(\mathbb{R}^m)$ has measure zero with respect to the Lebesgue measure on $f(\mathbb{R}^m)$, then $\dim(f(\mathbb{R}^m)) = m$ and almost every point in $f(\mathbb{R}^m)$ is generic with respect to f.*

D.2. Proof of theorem 3.5.(i)

We extend the results from Kivva et al. (2022) to using our MSM family as prior distribution for $\boldsymbol{z}_{1:T}$. The strategy requires finding some open set where the transformations $\mathcal{F}$ and $\mathcal{G}$ from two equally distributed SDSs are invertible, and then use analytic function properties to establish the identifiability result. First, we need to show that the points in the pre-image of a piece-wise factored mapping $\mathcal{F}$ can be computed using the MSM prior.

Lemma D.6. *Consider a random variable $\boldsymbol{z}_{1:T}$ which follows a Markov Switching Model distribution. Let us consider $f : \mathbb{R}^m \to \mathbb{R}^m$, a piece-wise affine mapping which generates the random variable $\boldsymbol{x}_{1:T} = \mathcal{F}(\boldsymbol{z}_{1:T})$ 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, $\boldsymbol{x}^{(0)}_{1:T} = \{\boldsymbol{x}^{(0)}, \dots, \boldsymbol{x}^{(0)}\} \in \mathbb{R}^{Tm}$ is also a generic point with respect to $\mathcal{F}$ and the number of points in the pre-image $\mathcal{F}^{-1}(\{\boldsymbol{x}^{(0)}_{1:T}\})$ can be computed as*

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

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

$$g_{i_t}(\boldsymbol{z}_t) = f(\boldsymbol{z}_t), \quad \forall \boldsymbol{z}_t \in B(\boldsymbol{z}_i, \epsilon)$$

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

Let $\delta_0 > 0$ such that

$$B(\boldsymbol{x}_{1:T}^{(0)}, \delta_0) \subseteq \bigcap_{i_1,\dots,i_T}^{n} \mathcal{G}_{i_1,\dots,i_T}(B(\{\boldsymbol{z}_{i_1}, \dots, \boldsymbol{z}_{i_T}\}, \epsilon))$$

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 where the MSM is closed under factored affine transformations.

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

Where 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 $\mathbb{R}^m$. Then, we have

$$\begin{aligned}
&\int_{\boldsymbol{x}_{1:T}\in\mathbb{R}^{Tm}} \text{Ext}\left(p|_{B(\boldsymbol{x}_{1:T}^{(0)},\delta)}\right) d\boldsymbol{x}_{1:T} \\
&= \sum_{i_1,\dots,i_T}^{s} \sum_{j_1,\dots,j_T}^{K} p\left(s_{1:T} = \{j_1, \dots, j_T\}\right) \mathcal{N}\left(\boldsymbol{x}_1; A_{i_1}\boldsymbol{\mu}(j_1) + \mathbf{b}_{i_1}, A_{i_1}\boldsymbol{\Sigma}_1(j_1)A_{i_1}^T\right) \\
&\quad \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{x}_t; A_{i_t}\boldsymbol{m}\left(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \mathbf{b}_{i_{t-1}}\right), j_t\right) + \mathbf{b}_{i_t}, A_{i_t}\boldsymbol{\Sigma}\Big(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \mathbf{b}_{i_{t-1}}\right), j_t\Big)A_{i_t}^T\right) \\
&= \sum_{i_1,\dots,i_T}^{n} \int_{\boldsymbol{x}_{1:T}\in\mathbb{R}^{Tm}} \sum_{j_1,\dots,j_T}^{K} p\left(s_{1:T} = \{j_1, \dots, j_T\}\right) \mathcal{N}\left(\boldsymbol{x}_1; A_{i_1}\boldsymbol{\mu}(j_1) + \mathbf{b}_{i_1}, A_{i_1}\boldsymbol{\Sigma}_1(j_1)A_{i_1}^T\right) \\
&\quad \prod_{t=2}^{T} \mathcal{N}\Bigg(\boldsymbol{x}_t; A_{i_t}\boldsymbol{m}\left(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \mathbf{b}_{i_{t-1}}\right), j_t\right) + \mathbf{b}_{i_t}, \\
&\quad A_{i_t}\boldsymbol{\Sigma}\Big(A_{i_{t-1}}^{-1}\left(\boldsymbol{x}_{t-1} - \mathbf{b}_{i_{t-1}}\right), j_t\Big)A_{i_t}^T\Bigg) d\boldsymbol{x}_{1:T} \\
&= \sum_{i_1,\dots,i_T}^{n} 1 = n^T = \left|\mathcal{F}^{-1}(\{\boldsymbol{x}_{1:T}^{(0)}\})\right|
\end{aligned}$$

□

We can deduce the following corollary as in Kivva et al. (2022).

Corollary D.7. *Let $\mathcal{F}, \mathcal{G} : \mathbb{R}^{Tm} \to \mathbb{R}^{Tn}$ be factored piece-wise affine mappings, with $\boldsymbol{x}_t := f(\boldsymbol{z}_t)$ and $\boldsymbol{x}'_t := g(\boldsymbol{z}'_t)$, for $1 \leq t \leq T$. Assume f and g are weakly-injective (Def. 2.1). Let $\boldsymbol{z}_{1:T}$ and $\boldsymbol{z}'_{1:T}$ be distributed according to the identifiable MSM family. Assume $\mathcal{F}(\boldsymbol{z}_{1:T})$ and $\mathcal{G}(\boldsymbol{z}'_{1:T})$ are equally 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 $\boldsymbol{x}_{1:T}^{(0)} = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_0\} \in \mathbb{R}^{Tn}$ there exists $\boldsymbol{x}_{1:T}^{(1)} \in B(\boldsymbol{x}_{1:T}^{(0)}, \delta)$ and $\delta_1 > 0$ such that both $\mathcal{F}$ and $\mathcal{G}$ are invertible on $B(\boldsymbol{x}_{1:T}^{(1)}, \delta_1) \cap \mathcal{F}(\mathbb{R}^{Tm})$.

Proof. First, we observe that since $\mathcal{F}$ is a factored mapping, if f is invertible on $B(\boldsymbol{x}_0, 2\delta) \cap f(\mathbb{R}^m)$, we can compute the inverse of $\mathcal{F}$ on $B(\boldsymbol{x}_{1:T}^{(0)}, 2\delta) \cap \mathcal{F}(\mathbb{R}^m)$ for $\boldsymbol{x}_{1:T}^{(0)} = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_0\} \in \mathbb{R}^{Tn}$ as $\mathcal{F}^{-1}(\boldsymbol{x}_{1:T}) = \{f^{-1}(\boldsymbol{x}_1), \dots, f^{-1}(\boldsymbol{x}_T)\} \in \mathbb{R}^{Tm}$, for $\boldsymbol{x}_t \in B(\boldsymbol{x}_0, 2\delta) \cap f(\mathbb{R}^m)$, $1 \leq t \leq T$. Then, $\mathcal{F}$ is invertible on $B(\boldsymbol{x}_{1:T}^{(0)}, 2\delta) \cap \mathcal{F}(\mathbb{R}^m)$.

By Lemma D.5, almost every point $\boldsymbol{x} \in B(\boldsymbol{x}^{(0)}, \delta) \cap f(\mathbb{R}^m)$ is generic with respect to f and g, as both mappings are weakly injective. As discussed previously, if $\boldsymbol{x}^{(0)} \in f(\mathbb{R}^m)$ is a generic point with respect to f, the point $\boldsymbol{x}_{1:T}^{(0)} = \{\boldsymbol{x}^{(0)}, \dots, \boldsymbol{x}^{(0)}\} \in$

$\mathcal{F}(\mathbb{R}^{Tm})$ is also generic with respect to $\mathcal{F}$, as the finite points in the preimage $f^{-1}(\{\boldsymbol{x}^{(0)}\})$ extend to finite points in the preimage $\mathcal{F}^{-1}(\{\boldsymbol{x}^{(0)}_{1:T}\})$. Then, almost every point $\boldsymbol{x}_{1:T} \in B(\boldsymbol{x}^{(0)}_{1:T}, \delta) \cap \mathcal{F}(\mathbb{R}^{Tm})$ is generic with respect to $\mathcal{F}$ and $\mathcal{G}$.

Consider now $\boldsymbol{x}^{(1)}_{1:T} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(1)}\} \in B(\boldsymbol{x}^{(0)}_{1:T}, \delta)$ such a generic point. From the invertibility of $\mathcal{F}$ on $B(\boldsymbol{x}^{(1)}_{1:T}, \delta)$, we have $|\mathcal{F}^{-1}(\{\boldsymbol{x}^{(1)}_{1:T}\})| = 1$. By Lemma D.6, we have that $|\mathcal{G}^{-1}(\{\boldsymbol{x}^{(1)}_{1:T}\})| = 1$, as $\boldsymbol{x}^{(1)}_{1:T}$ is generic with respect to $\mathcal{F}$ and $\mathcal{G}$. Then, there exists, $\delta > \delta_1 > 0$ such that on $\left(B(\boldsymbol{x}^{(1)}_{1:T}, \delta_1) \cap \mathcal{F}(\mathbb{R}^{Tm})\right) \subset \left(B(\boldsymbol{x}^{(0)}_{1:T}, 2\delta) \cap \mathcal{F}(\mathbb{R}^{Tm})\right)$ the function $\mathcal{G}$ is invertible. □

We need an additional result to prepare the proof for Theorem 3.5.(i).

Theorem D.8. *Let $\mathcal{F}$, $\mathcal{G} : \mathbb{R}^{Tm} \to \mathbb{R}^{Tn}$ be factored piece-wise affine mappings, with $\boldsymbol{x}_t := f(\boldsymbol{z}_t)$ and $\boldsymbol{x}'_t := g(\boldsymbol{z}'_t)$, for $1 \leq t \leq T$. Let $\boldsymbol{z}_{1:T}$ and $\boldsymbol{z}'_{1:T}$ be distributed according to the identifiable MSM family. Assume $\mathcal{F}(\boldsymbol{z}_{1:T})$ and $\mathcal{G}(\boldsymbol{z}'_{1:T})$ are equally distributed, and that there exists $\boldsymbol{x}^{(0)}_{1:T} \in \mathbb{R}^{Tn}$ and $\delta > 0$ such that $\mathcal{F}$ and $\mathcal{G}$ are invertible on $B(\boldsymbol{x}^{(0)}_{1:T}, \delta) \cap f(\mathbb{R}^{Tm})$. Then there exists an invertible factored affine transformation $\mathcal{H}$ such that $\mathcal{H}(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$.*

Proof. From the invertibility of $\mathcal{F}$ and $\mathcal{G}$ in $B(\boldsymbol{x}^{(0)}_{1:T}, \delta) \cap f(\mathbb{R}^{Tm})$ we can find a Tm-dimensional affine subspace $B(\boldsymbol{x}^{(1)}_{1:T}, \delta_1) \cap L$, where $\delta_1 > 0$, $B(\boldsymbol{x}^{(1)}_{1:T}, \delta_1) \subseteq B(\boldsymbol{x}^{(0)}_{1:T}, \delta)$, and $L \subseteq \mathbb{R}^{Tn}$ such that $\mathcal{H}_{\mathcal{F}}, \mathcal{H}_{\mathcal{G}} : \mathbb{R}^{Tm} \to L$ are a pair of invertible affine functions where $\mathcal{H}_{\mathcal{F}}^{-1}$ and $\mathcal{H}_{\mathcal{G}}^{-1}$ coincide with $\mathcal{F}^{-1}$ and $\mathcal{G}^{-1}$ on $B(\boldsymbol{x}_1, \delta_1) \cap L$ respectively. The fact that $\mathcal{F}$ and $\mathcal{G}$ are factored implies that $\mathcal{H}_{\mathcal{F}}, \mathcal{H}_{\mathcal{G}}$ are also factored. To see this, we observe that the inverse of a block diagonal matrix is the inverse of each block, as an example for $\mathcal{F}$, we first have that $\mathcal{H}_{\mathcal{F}}^{-1}$ must be forcibly factored since it needs to coincide with $\mathcal{F}^{-1}$.

$$\mathcal{H}_{\mathcal{F}}^{-1}(\boldsymbol{x}_{1:T}) = \begin{pmatrix} \tilde{A}_f & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \tilde{A}_f \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_T \end{pmatrix} + \begin{pmatrix} \tilde{b}_f \\ \vdots \\ \tilde{b}_f \end{pmatrix} = \begin{pmatrix} f^{-1}(\boldsymbol{x}_1) \\ \vdots \\ f^{-1}(\boldsymbol{x}_T) \end{pmatrix}$$

then we can take the inverse to obtain the factored $\mathcal{H}_{\mathcal{F}}$.

$$\begin{aligned} \mathcal{H}_{\mathcal{F}} &= \begin{pmatrix} \tilde{A}_f^{-1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \tilde{A}_f^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_T \end{pmatrix} - \begin{pmatrix} \tilde{A}_f^{-1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \tilde{A}_f^{-1} \end{pmatrix} \begin{pmatrix} \tilde{b}_f \\ \vdots \\ \tilde{b}_f \end{pmatrix} \\ &= \begin{pmatrix} A_f & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & A_f \end{pmatrix} \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_T \end{pmatrix} + \begin{pmatrix} b_f \\ \vdots \\ b_f \end{pmatrix}, \quad \text{where } A_f = \tilde{A}_f^{-1}, \text{ and } b_f = -\tilde{A}_f^{-1}\tilde{b}_f \end{aligned}$$

Since $\mathcal{F}(\boldsymbol{z}_{1:T})$ and $\mathcal{G}(\boldsymbol{z}'_{1:T})$ are equally distributed, we have that $\mathcal{H}_{\mathcal{F}}(\boldsymbol{z}_{1:T})$ and $\mathcal{H}_{\mathcal{G}}(\boldsymbol{z}'_{1:T})$ are equally distributed on $B(\boldsymbol{x}^{(1)}_{1:T}, \delta_1) \cap L$. From Prop C.2, we know that $\mathcal{H}_{\mathcal{F}}(\boldsymbol{z}_{1:T})$ and $\mathcal{H}_{\mathcal{G}}(\boldsymbol{z}'_{1:T})$ are distributed according to the identifiable MSM family, which implies $\mathcal{H}_{\mathcal{F}}(\boldsymbol{z}_{1:T}) = \mathcal{H}_{\mathcal{G}}(\boldsymbol{z}'_{1:T})$, and also $\mathcal{H}_{\mathcal{G}}^{-1}(\mathcal{H}_{\mathcal{F}}(\boldsymbol{z}_{1:T})) = \boldsymbol{z}'_{1:T}$, where $\mathcal{H} = \mathcal{H}_{\mathcal{G}}^{-1} \circ \mathcal{H}_{\mathcal{F}}$ is an affine transformation. □

From the previous result and Theorem 3.2, there exists a permutation $\sigma(\cdot)$, such that $\boldsymbol{m}_{fg}(\boldsymbol{z}', k) = \boldsymbol{m}'(\boldsymbol{z}', \sigma(k))$ for $1 \leq k \leq K$.

$$\begin{aligned} \boldsymbol{m}'(\boldsymbol{z}', \sigma(k)) = \boldsymbol{m}_{fg}(\boldsymbol{z}', k) &= A_g^{-1}\boldsymbol{m}_f\left(A_g\boldsymbol{z}' + \mathbf{b}_g, k\right) - A_g^{-1}\mathbf{b}_g \\ &= A_g^{-1}A_f\boldsymbol{m}\left(A_f^{-1}A_g\boldsymbol{z}' + A_f^{-1}\left(\mathbf{b}_g - \mathbf{b}_f\right), k\right) + A_g^{-1}\left(\mathbf{b}_f - \mathbf{b}_g\right) \\ &= A\boldsymbol{m}\left(A^{-1}(\boldsymbol{z}' - \mathbf{b}), k\right) + \mathbf{b}, \end{aligned}$$

where $A = A_g^{-1}A_f$ and $\mathbf{b} = A_g^{-1}\left(\mathbf{b}_f - \mathbf{b}_g\right)$. Similar implications apply for $\Sigma(\boldsymbol{z}, a), a \in \mathcal{A}$.

Now we have all the elements to prove Theorem 3.5.(i).

Proof. We assume there exists another model that generates the same distribution from Eq.(13), whith a prior $p' \in \mathcal{M}_{NL}^{T}$ under assumptions (a1-a2), and non-linear emmision $\mathcal{F}'$, composed by f which is weakly injective and piece-wise linear: i.e. $(\mathcal{F}\#p)(\boldsymbol{x}_{1:T}) = (\mathcal{F}'\#p')(\boldsymbol{x}_{1:T})$.

From weakly-injectiveness, 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 D.7, which implies there exists $\boldsymbol{x}^{(1)}_{1:T} \in B(\boldsymbol{x}^{(0)}_{1:T}, \delta)$ and $\delta_1 > 0$ such that both $\mathcal{F}$ and $\mathcal{F}'$ are invertible on $B(\boldsymbol{x}^{(1)}_{1:T}, \delta_1) \cap \mathcal{F}(\mathbb{R}^{Tm})$. Thus, by Theorem D.8, there exists an affine transformation $\mathcal{H}$ such that $\mathcal{H}(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$, which means that $p \in \mathcal{M}^T_{NL}$ is identifiable up to affine transformations. □

D.3. Proof of theorem 3.5.(ii)

So far we have proved the identifiability of the transition function on the latent MSM distribution up to affine transformations. By further assuming injectivity of the piece-wise mapping $\mathcal{F}$, we can prove identifiability of $\mathcal{F}$ up to affine transformations by re-using results from Kivva et al. (2022). We begin by stating the following known result.

Lemma D.9. *Let $Z \sim \sum_{j=1}^{J} \lambda_j \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$ where Z is a GMM (in reduced form). Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is a continuous piecewise affine function such that $f(Z) \sim Z$. Then f is affine.*

We can state the identification of $\mathcal{F}$.

Theorem D.10. *Let $\mathcal{F}, \mathcal{G} : \mathbb{R}^{mT} \to \mathbb{R}^{nT}$ be continuous invertible factored piecewise affine functions. Let $\boldsymbol{z}_{1:T}, \boldsymbol{z}'_{1:T}$ be random variables distributed according to MSMs. Suppose that $\mathcal{F}(\boldsymbol{z}_{1:T})$ and $\mathcal{G}(\boldsymbol{z}'_{1:T})$ are equally distributed.*

Then there exists a factored affine transformation $\mathcal{H} : \mathbb{R}^{mT} \to \mathbb{R}^{mT}$ such that $\mathcal{H}(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$ and $\mathcal{G} = \mathcal{F} \circ \mathcal{H}^{-1}$.

Proof. From Theorem D.8, there exists an invertible affine transformation $\mathcal{H}_1 : \mathbb{R}^{mT} \to \mathbb{R}^{mT}$ such that $\mathcal{H}_1(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$. Then, $\mathcal{F}(\boldsymbol{z}_{1:T}) \sim \mathcal{G}(\mathcal{H}_1(\boldsymbol{z}_{1:T}))$. From the invertibility of $\mathcal{G}$, we have $\boldsymbol{z}_{1:T} \sim (\mathcal{H}_1^{-1} \circ \mathcal{G}^{-1} \circ \mathcal{F})(\boldsymbol{z}_{1:T})$. We note that $\mathcal{H}_1, \mathcal{G}, \mathcal{F}$ are factored mappings, and structured as follows

$$\left(\mathcal{H}_1^{-1} \circ \mathcal{G}^{-1} \circ \mathcal{F}\right)(\boldsymbol{z}_{1:T}) = \begin{pmatrix} \left(h_1^{-1} \circ g^{-1} \circ f\right)(\boldsymbol{z}_1) \\ \vdots \\ \left(h_1^{-1} \circ g^{-1} \circ f\right)(\boldsymbol{z}_T) \end{pmatrix} \sim \begin{pmatrix} (\boldsymbol{z}_1) \\ \vdots \\ (\boldsymbol{z}_T) \end{pmatrix},$$

where the inverse of $\mathcal{H}_1$ is also factored, as observed from previous results. Since the transformation is equal for $1 \leq t \leq T$, we can proceed for $t = 1$ considering $\boldsymbol{z}_1$ is distributed as a GMM (in reduced form), as it corresponds to the initial distribution of the MSM. Then, by Lemma D.9, there exists an affine mapping $h_2 : \mathbb{R}^m \to \mathbb{R}^m$ such that $h_1^{-1} \circ g^{-1} \circ f = h_2$. Then,

$$\mathcal{F} = \begin{pmatrix} f \\ \vdots \\ f \end{pmatrix} = \begin{pmatrix} g \circ h_1 \circ h_2 \\ \vdots \\ g \circ h_1 \circ h_2 \end{pmatrix} = \left(\mathcal{G} \circ \mathcal{H}\right),$$

where $h = h_1 \circ h_2$. Considering the invertibility of $\mathcal{G}$ and the fact that $\mathcal{F}(\boldsymbol{z}_{1:T})$ and $\mathcal{G}(\boldsymbol{z}'_{1:T})$ are equally distributed, we also have $\mathcal{H}(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$. □

We use the previous result to prove Theorem 3.5.(ii).

Proof. We assume there exists another model that generates the same distribution from Eq.(13), whith a prior $p' \in \mathcal{M}^T_{NL}$ under assumptions (a1-a2), and non-linear emmision $\mathcal{F}'$, composed by f which is continuous, injective and piece-wise linear: i.e. $(\mathcal{F}\#p)(\boldsymbol{x}_{1:T}) = (\mathcal{F}'\#p')(\boldsymbol{x}_{1:T})$.

These are the preconditions to satisfy Theorem D.10, which implies there exists an affine transformation $\mathcal{H}$ such that $\mathcal{H}(\boldsymbol{z}_{1:T}) = \boldsymbol{z}'_{1:T}$ and $\mathcal{F} = \mathcal{F}' \circ \mathcal{H}$. In other words, the prior $p \in \mathcal{M}^T_{NL}$, and f which composes $\mathcal{F}$ are identifiable up to affine transformations. □

E. Estimation details

We provide additional details from the descriptions in the main text.

E.1. Expectation Maximisation on MSMs

For convenience, the expressions below are computed from samples $\{\boldsymbol{z}_{1:T}^b\}_{b=1}^B$ for a batch of size B. Recall we formulate our method in terms of the expectation maximisation (EM) algorithm. Given some arrangement of the parameter values (θ'), the E-step computes the posterior distribution of the latent variables $p_{\theta'}(\mathbf{s}_{1:T}|\boldsymbol{z}_{1:T})$. This can then be used to compute the expected log-likelihood of the complete data (latent variables and observations),

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}') := \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{p_{\boldsymbol{\theta}'}(\mathbf{s}_{1:T}^b|\boldsymbol{z}_{1:T}^b)}\Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}^b, \mathbf{s}_{1:T}^b)\Big]. \tag{30}$$

Given a first-order stationary Markov chain, we denote the posterior probability $p_{\boldsymbol{\theta}}(s_t^b = k|\boldsymbol{z}_{1:T}^b)$ as $\gamma_{t,k}^b$, and the joint posterior of two consecutive states $p_{\boldsymbol{\theta}}(s_t^b = k, s_{t-1}^b = l|\boldsymbol{z}_{1:T}^b)$ as $\xi_{t,k,l}^b$. For this case, the result is equivalent to the HMM case and can be found in the literature, e.g. Bishop (2006). We can then compute a more explicit form of Eq. (30),

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K}\gamma_{1,k}^b \log \pi_k + \frac{1}{B}\sum_{b=1}^{B}\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{t,k,l}^b \log Q_{lk} + \\ \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K}\gamma_{1,k}^b \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_1^b|, s_1^b = k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{t=2}^{T}\sum_{k=1}^{K}\gamma_{t,k}^b \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_t^b|\boldsymbol{z}_{t-1}^b, s_t^b = k),\end{aligned} \tag{31}$$

where π and Q denote the initial and transition distribution of the Markov chain. In the M-step, the previous expression is maximised to calculate the update rules for the parameters, i.e. $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}')$. The updates for π and Q are also obtained using standard results for HMM inference (again see Bishop (2006)). Assuming Gaussian initial and transition densities, we can also use standard literature results for updating the initial mean and covariance. For the transition densities, we consider a family with fixed covariance matrices, and only the means $\boldsymbol{m}_{\boldsymbol{\theta}}(\cdot, k)$ are dependent on the previous observation. In this case, the standard results can also be used to update the covariances of the transition distributions. We drop the subscript $\boldsymbol{\theta}$ for convenience.

The updates for the mean parameters are dependent on the functions we choose. For multivariate polynomials of degree P, we can recover an exact M-step by transforming the mapping into a matrix-vector operation:

$$\boldsymbol{m}(\boldsymbol{z}_{t-1}, k) = \sum_{c=1}^{C} A_{k,c}\hat{\boldsymbol{z}}_{c,t-1}, \quad \hat{\boldsymbol{z}}_{t-1}^T = \begin{pmatrix}1 & z_{t-1,1} & \dots & z_{t-1,d} & z_{t-1,1}^2 & z_{t-1,1}z_{t-1,2} & \dots\end{pmatrix}, \tag{32}$$

where $\hat{\boldsymbol{z}}_{t-1} \in \mathbb{R}^C$ denotes the polynomial features of $\boldsymbol{z}_{t-1}$ up to degree P. The total number of features is $C = \binom{P+d}{d}$ and the exact update for A_k is

$$A_k \leftarrow \left(\sum_{b=1}^{B}\sum_{t=2}^{T}\gamma_{t,k}^b \boldsymbol{z}_t^b (\hat{\boldsymbol{z}}_{t-1}^b)^T\right)\left(\sum_{b=1}^{B}\sum_{t=2}^{T}\gamma_{t,k}^b \hat{\boldsymbol{z}}_{t-1}^b (\hat{\boldsymbol{z}}_{t-1}^b)^T\right)^{-1}. \tag{33}$$

For exact updates such as the one above, we require B to be sufficiently large to ensure consistent updates during training. In the main text, we already discussed the case where the transition means are parametrised by neural networks.

E.2. Variational Inference for SDSs

We provide more details on the ELBO objective for SDSs.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = \log \int \sum_{\mathbf{s}_{1:T}} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}) d\boldsymbol{z}_{1:T} \tag{34}$$

$$\geq \mathbb{E}_{q_{\phi,\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}|\mathbf{s}_{1:T}) p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T})}{q_{\phi}(\boldsymbol{z}_{1:T}, \mathbf{s}_{1:T}|\boldsymbol{x}_{1:T})} \right] \tag{35}$$

$$\geq \mathbb{E}_{q_{\phi}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})}{q_{\phi}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} + \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}|\boldsymbol{z}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}|\mathbf{s}_{1:T}) p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T})}{p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}|\boldsymbol{z}_{1:T})} \right] \right] \tag{36}$$

$$\geq \mathbb{E}_{q_{\phi}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})}{q_{\phi}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})} + \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) \right] \tag{37}$$

$$\approx \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T}) - \log q_{\phi}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}), \quad \boldsymbol{z}_{1:T} \sim q_{\phi} \tag{38}$$

where as as mentioned, we compute the ELBO objective using Monte Carlo integration with samples $\boldsymbol{z}_{1:T}$ from q_{ϕ}, and $p_{\boldsymbol{\theta}}(\boldsymbol{z}_{1:T})$ is computed using Eq. (20). Alternatively, Dong et al. (2020) proposes computing the gradients of the latent MSM using the following rule.

$$\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{s}_{1:T}|\boldsymbol{z}_{1:T})} \left[\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \mathbf{s}_{1:T} \right] \tag{39}$$

where the objective is similar to Eq.(17). Below we reflect on the main aspects of each method.

- Dong et al. (2020) computes the parameters of the latent MSM using a loss term similar to Eq. (17). Although we need to compute the exact posteriors explicitly, we only take the gradient with respect to $\log p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t = k)$ which is relatively efficient. Unfortunately, the approach is prone to state collapse and additional loss terms with annealing schedules need to be implemented.
- Ansari et al. (2021) does not require computing exact posteriors as the parameters of the latent MSM are optimized using the forward algorithm. The main disadvantage is that we require back-propagation to flow through the forward computations, which is more inefficient. Despite this, the objective used is less prone to state collapse and optimisation becomes simpler.

Although both approaches show good performance empirically, we observed that training becomes a difficult task and requires careful hyper-parameter tuning and multiple initialisations. Note that the methods are not (a priori) theoretically consistent with the previous identifiability results. Since exact inference is not tractable in SDSs, one cannot design a consistent estimator such as MLE. Future developments should focus on combining the presented methods with tighter variational bounds (Maddison et al., 2017) to design consistent estimators for such generative models.

F. Experiment details

F.1. Metrics

Markov Switching Models Consider K components, where as described the evaluation is performed by computing the averaged sum of the distances between the estimated function components. Since we have identifiability of the function forms up to permutations, we need to compute distances with all the permutation configurations to resolve this indeterminacy. Therefore, we can quantify the estimation error as follows

$$\text{err} := \min_{\mathbf{k}=\text{perm}(\{1,\dots,K\})} \frac{1}{K} \sum_{i=1}^{K} d(\boldsymbol{m}(\cdot, i), \hat{\boldsymbol{m}}(\cdot, k_i)), \tag{40}$$

where $d(\cdot, \cdot)$ denotes the L2 distance between functions. We compute an approximate L2 distance by evaluating the functions on points sampled from a random region of $\mathbb{R}^m$ and averaging the Euclidean distance, more specifically we sample 10^5 in the $[-1, 1]^d$ interval for each evaluation.

$$d(f, g) := \int_{\boldsymbol{x} \in [-1,1]^d} \sqrt{\left|\left| f(\boldsymbol{x}) - g(\boldsymbol{x}) \right|\right|^2} d\boldsymbol{x} \tag{41}$$

$$\approx \frac{1}{10^5} \sum_{i=1}^{10^5} \sqrt{\left|\left| f\left(\boldsymbol{x}^{(i)}\right) - g\left(\boldsymbol{x}^{(i)}\right) \right|\right|^2}, \quad \boldsymbol{x}^{(i)} \sim Uniform([-1, 1]^m) \tag{42}$$

Note that resolving the permutation indeterminacy has a cost of $\mathcal{O}(K!)$, which for $K > 5$ already poses some problems in both monitoring performance during training and testing. To alleviate this computational cost, we take a greedy approach, where for each estimated function component we pair it with the ground truth function with the lowest L2 distance. Note that this can return a suboptimal result when the functions are not estimated accurately, but the computational cost is reduced to $\mathcal{O}(K^2)$.

Switching Dynamical Systems To compute the L_2 distance for the transitions means in SDSs, we first need to resolve the linear transformation in Eq.(15). Thus, we compute the following

$$\arg\min_{h} \left\{ d\Big(f, (f' \circ h)\Big) \right\} \tag{43}$$

where f,f' compose the groundtruth $\mathcal{F}$ and estimated $\mathcal{F}'$ non-linear emissions respectively, and h denotes the affine transformation. We compute the above L_2 norm using 500 generated observations from a held out test dataset. Finally, we compute the error as in Eq. (40), and with the following L_2 norm.

$$d\bigg(\boldsymbol{m}\big(\cdot, i\big), A\hat{\boldsymbol{m}}\big(A^{-1}(\boldsymbol{z} - \mathbf{b}), \sigma(i)\big) + \mathbf{b}\big)\bigg) \tag{44}$$

which is taken from Eq.(15). Similarly, we compute the norm using samples from the ground truth latent variables, generated from the held out test dataset. To resolve the permutation $\sigma(\cdot)$, we first compute the F_1 score on the segmentation task as indicated, by counting the total true positives, false negatives and false positives. Then, $\sigma(\cdot)$ is determined from the permutation with highest F_1 score.

F.2. Averaged Jacobian and causal structure computation

Regarding regime-dependent causal discovery, our approach can be considered as a functional causal model-based method (see Glymour et al. (2019) for the complete taxonomy). In such methods, the causal structure is usually estimated by inspecting the parameters that encode the dependencies between data, rather than performing independence tests (Tank et al., 2021). In the linear case, we can threshold the transition matrix to obtain an estimate of the causal structure (Pamfil et al., 2020). The non-linear case is a bit more complex since the transition functions are not separable among variables, and the Jacobian can differ considerably for different input values. With the help of locally connected networks, Zheng et al. (2018) aim to encode the variable dependencies in the first layer, and perform similar thresholding as in the linear case. To encourage that the causal structure is captured in the first layer and prevent it from happening in the next ones, the weights in the first layer are regularised with L1 loss to encourage sparsity, and all the weights in the network are regularised with L2 loss.

In our experiments, we observe this approach requires enormous finetuning with the potential to sacrifice the flexibility of the network. Instead, we estimate the causal structure by thresholding the averaged absolute-valued Jacobian with respect to a set of samples. We denote the Jacobian of $\hat{\boldsymbol{m}}(\boldsymbol{z}, k)$ as $\boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}(\boldsymbol{z})$. To ensure that the Jacobian captures the effects of the regime of interest, we use samples from the data set and classify them with the posterior distribution. In other words, we will create K sets of variables, where each set $\mathcal{Z}_k$ with size $N_K = |\mathcal{Z}_k|$ contains variables that have been selected using the posterior, i.e. $\boldsymbol{z}^{(i)} \in \mathcal{Z}_k$ if $k = \arg\max p_{\boldsymbol{\theta}}(\mathbf{s}^{(i)}|\boldsymbol{z}_{1:T})$, where we use the index i to denote that $\boldsymbol{z}^{(i)}$ is associated with $\mathbf{s}^{(i)}$. Then, for a given regime k, the matrix that encodes the causal structure $\hat{G}_k$ is expressed as

$$\hat{G}_k := \mathbf{1}\left(\frac{1}{N_k} \sum_{i=1}^{N_k} \left| \boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}\left(\boldsymbol{z}^{(i)}\right) \right| > \tau \right), \quad \boldsymbol{z}^{(i)} \in \mathcal{Z}_k, \tag{45}$$

where $\mathbf{1}(\cdot)$ is an indicator function which equals to 1 if the argument is true and 0 otherwise. We $\tau = 0.05$ in our experiments. Finally, we evaluate the estimated K regime-dependent causal graphs can be evaluated in terms of the average F1-score over components.

F.3. Training specifications

All the experiments are implemented in Pytorch (Paszke et al., 2019) and carried out on NVIDIA RTX 2080Ti GPUs, except for the experiments with videos (synthetic and salsa), where we used NVIDIA RTX A6000 GPUs.

Markov Switching Models When training polynomials (including the linear case), we use the exact batched M-step updates with batch size 500 and train for a maximum of 100 epochs, and stop when the likelihood plateaus. When considering updates in the form of Eq. (17), e.g. neural networks, we use ADAM optimiser (Kingma & Ba, 2015) with an initial learning rate $7 \cdot 10^{-3}$ and decrease it by a factor of 0.5 on likelihood plateau up to 2 times. We vary the batch size and maximum training time depending on the number of states and dimensions. For instance, for $K = 3$ and $d = 3$, we use a batch size of 256 and train for a maximum of 25 epochs. For other configurations, we decrease the batch size and increase the maximum training time to meet GPU memory requirements. Similar to related approaches (Hälvä & Hyvarinen, 2020), we use random restarts to achieve better parameter estimates.

Switching Dynamical Systems Since training SDSs requires careful hyperparameter tuning for each setting, we provide details for each setting separately.

- For the synthetic experiments, we use batch size 100, and we train for 100 epochs. We use ADAM optimiser (Kingma & Ba, 2015) with an initial learning rate $5 \cdot 10^{-4}$, and decrease it by a factor of 0.5 every 30 epochs. To avoid state collapse, we perform an initial warm-up phase for 5 epochs, where we train with fixed discrete state parameters π and Q, which we fix to uniform distributions. We run multiple seeds and select the best model on the ELBO objective. Regarding the network architecture, we estimate the transition means using two-layer networks with cosine activations and 16 hidden dimensions, and the non-linear emission using two-layer networks with Leaky ReLU activations with 64 hidden dimensions. For the inference network, the bi-directional RNN has 2 hidden layers and 64 hidden dimensions, and the forward RNN has an additional 2 layers with 64 hidden dimensions. We use LSTMs for the RNN updates.

- For the synthetic videos, we vary some of the above configurations. We use batch size 64 and train for 200 epochs with the same optimiser and learning rate, but we now decrease it by a factor of 0.8 every 80 epochs. Instead of running an initial warm-up phase, we devise a three-stage training. First, we pre-train the encoder (emission) and decoder networks, where for 10 epochs the objective ignores the terms from the MSM prior. The second phase is inspired by Ansari et al. (2021), where we use softmax with temperature on the logits of π and Q. To illustrate, we use softmax with temperature as follows

$$Q_{k,:} = p(s_t|s_{t-1} = k) = Softmax(\mathbf{o}_k/\tau), \quad t \in \{2, \dots, T\} \tag{46}$$

  where $\mathbf{o}_k$ are the logits of $p(s_t|s_{t-1} = k)$ and τ is the temperature. We start with $\tau = 10$, and decay it exponentially every 50 iterations by a factor of 0.99 after the pre-training stage. The third stage begins when $\tau = 1$, where we train the model as usual. Again, we run multiple seeds and select the best model on the ELBO. The network architecture is similar, except that we use additional CNNs for inference and generation. For inference, we use a 5-layer CNN with 64 channels, kernel size 3, and padding 1, Leaky ReLU activations, and we alternate between using stride 2 and 1. We then run a 2-layer MLP with Leaky ReLU activations and 64 hidden dimensions and forward the embedding to the same RNN inference network we described before. For generation, we use a similar network, starting with a 2-layer MLP with Leaky ReLU activations and 64 hidden dimensions, and use transposed convolutions instead of convolutions (with the same configuration as before).

- For the salsa dancing videos, we use batch size 8 and train for a maximum of 400 epochs. We use the same optimiser and an initial learning rate of 10^{-4} and stop on ELBO plateau. We use a similar three-stage training as before, where we pre-train the encoder-decoder networks for 10 epochs. For the second stage, we start with $\tau = 100$ and decay it by a factor of 0.975 every 50 iterations after the pre-training phase. As always, we run multiple seeds and select the best model on ELBO. For the network architectures, we use the same as in the previous synthetic video experiment, but we use 7-layer CNNs, we increase all the network sizes to 128, and we use a latent MSM of 256 dimensions with $K = 3$ components. The transitions of the continous latent variables are parametrised with 2-layer MLPs with SofPlus activations and 256 hidden dimensions.

F.4. Synthetic experiments

For data generation, we sample $N = 10000$ sequences of length $T = 200$ in terms of a stationary first-order Markov chain with K states. The transition matrix Q is set to maintain the same state with probability 90% and switch to the next state with probability 10%, and the initial distribution π is the stationary distribution of Q. The initial distributions are Gaussian components with means sampled from $\mathcal{N}(0, 0.7^2\mathbf{I})$ and the covariance matrix is $0.1^2\mathbf{I}$. The covariance matrices

Table 2: Mean and 95-CI of the L_2 distances reported in Figure 4a (computed across 5 datasets), where we experiment with increasing sequence lengths and different network types.

Network type	Sequence length (T)		
	10	100	1000
Cubic	4.466 ± 0.814	0.795 ± 0.448	0.077 ± 0.115
Cosine	0.276 ± 0.238	0.041 ± 0.032	0.037 ± 0.020
Softplus	2.249 ± 1.045	0.078 ± 0.137	0.005 ± 0.002

Table 3: Mean and 95-CI of the L_2 distances reported in Figure 4b (computed across 5 datasets), where we experiment with increasing dimensions and states.

States (K)	Observation dimensions (m)					
	3	5	10	20	30	50
3	0.003 ± 0.005	0.004 ± 0.004	0.004 ± 0.002	0.015 ± 0.005	0.012 ± 0.006	0.064 ± 0.010
5	0.004 ± 0.002	0.008 ± 0.003	0.009 ± 0.005	0.023 ± 0.007	0.032 ± 0.005	0.052 ± 0.029
10	0.011 ± 0.008	0.014 ± 0.007	0.013 ± 0.003	0.031 ± 0.017	0.042 ± 0.005	0.054 ± 0.023

of the transition distributions are fixed to $0.05^2\mathbf{I}$, and the mean transitions $\boldsymbol{m}(\boldsymbol{z}, k), k = 1, \ldots, K$ are parametrised using polynomials of degree 3 with random weights, random networks with cosine activations, or random networks with softplus activations. For the locally connected networks (Zheng et al., 2018), we use cosine activation networks, and the sparsity is set to allow 3 interactions per element on average. All neural networks consist of two-layer MLPs with 16 hidden units.

When experimenting with SDSs, we use $K = 3$, $T = 200$, and $N = 5000$ of a 2D latent MSM with cosine network transitions and fixed covariances . We then further generate observations using two-layer Leaky ReLU networs with 8 hidden units. For synthetic videos, we render a ball on 32×32 coloured images from the 2D coordinates of the latent variables. When rendering images, the MSM trajectories are scaled to ensure the ball is always contained in the image canvas.

In Table 2 and Table 3 we show the mean and 95-CI from the experiments reported in Figure 4a and Figure 4b respectively; where we use datasets generated using 5 different seeds. In Figure 10, we show visualisations of some function responses for the experiment considering increasing variables and states (figs. 4b and 4c). For $K = 3$ states and $m = 3$ dimensions, we achieve $3 \cdot 10^{-3}$ L2 distance error on average and the responses in Figure 10a show low discrepancies with respect to the ground truth. Similar observations can be made with $K = 10$ states in Figure 10b, where the L_2 distance error is 11^{-2} on average. Furthermore, Table 4 shows details of the results reported in Figure 6.

F.5. ENSO-AIR experiment

The data consists of monthly observations of El Niño Southern Oscillation (ENSO) and All India Rainfall (AIR), starting from 1871 to 2016. Following the setting in Saggioro et al. (2020), we more specifically use the indicators *Niño 3.4 SST Index*[4] and *All-India Rainfall precipitation*[5] respectively. In total, we have $N = 1$ samples with $T = 1752$ time steps and

[4]Extracted from `https://psl.noaa.gov/gcos_wgsp/Timeseries/Nino34/`.

[5]Extracted from `https://climexp.knmi.nl/getindices.cgi?STATION=All-India_Rainfall&TYPE=p&WMO=IITMData/ALLIN`.

Table 4: Mean and 95-CI of the L_2 distance and state F_1 scores reported in Figure 6 (computed across 5 datasets), where we experiment with increasing observed dimensions.

Metric	Obs. dimensions					
	2	5	10	50	100	Image
L_2 dist.	0.083 ± 0.083	0.036 ± 0.043	0.033 ± 0.063	0.082 ± 0.136	0.054 ± 0.083	0.384 ± 0.257
State F_1	0.991 ± 0.008	0.999 ± 0.001	0.998 ± 0.003	0.994 ± 0.015	0.999 ± 0.001	0.978 ± 0.019

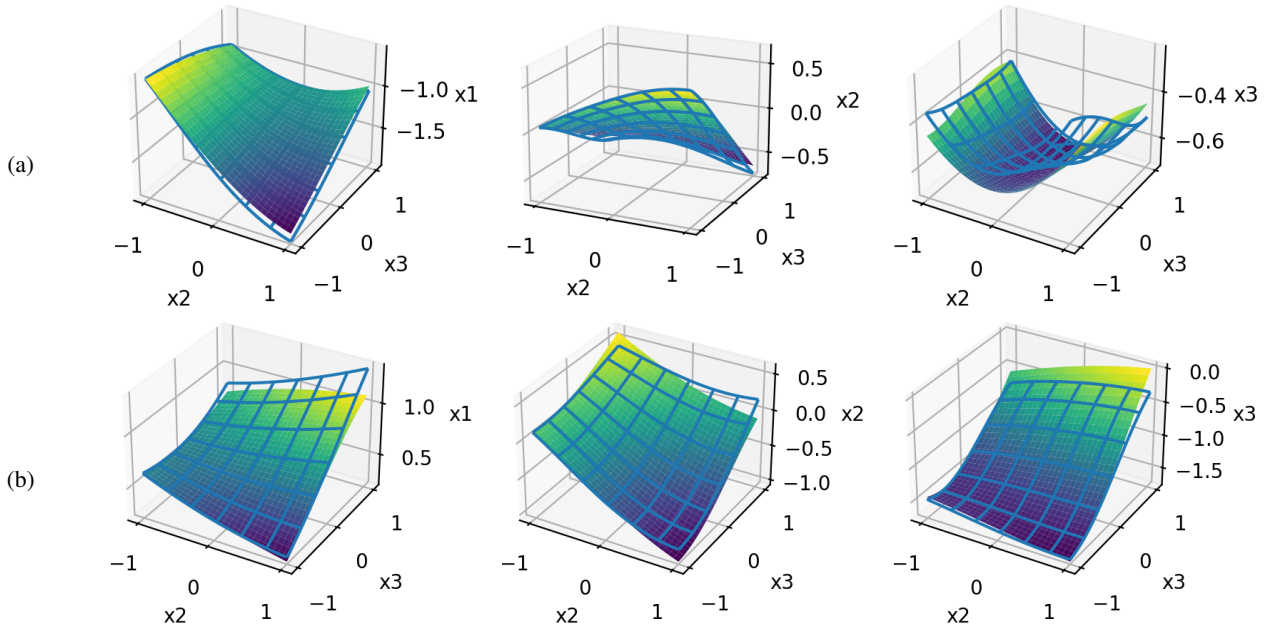


Figure 10: Function responses using 3 dimensions where we vary x_2 and x_3. Column i shows the response with respect to the i-th dimension. The blue grid shows the ground truth function for (a) 3 states showing component 1, and (b) 10 states showing component 6.

consider $K = 2$ components.

In the main text, we claim that our approach captures regimes based on seasonality. To visualise this, We group the posterior distribution by month (fig. 11), where similar groupings arise from both models, and observe that one component is assigned to Summer months (from May to September), and the other is assigned to Winter months (from October to April). To better illustrate the seasonal dependence present in this data. We show the function responses assuming linear and non-linear (softplus networks) transitions in Figures 12a and 12b respectively. In the linear case, we observe that the function responses on the ENSO variable are invariant across regimes. However, the response on the AIR variable varies across regimes, as we observe that the slope with respect to the ENSO input is zero in Winter, and increases slightly in Summer. This visualisation is consistent with the results reported in the regime-dependent graph (fig. 7). In the non-linear case, we now observe that the responses of the ENSO variable are slightly different, but the slope differences in the responses of the AIR variable with respect to the ENSO input are harder to visualise. The noticeable difference is that the self-dependency of the AIR variable changes non-linearly across regimes, contrary to the linear case where the slope with respect to AIR input was constant.

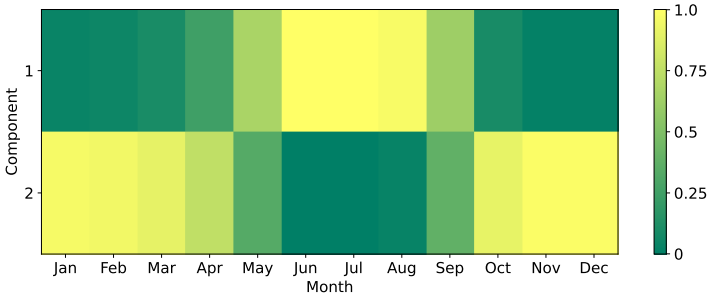
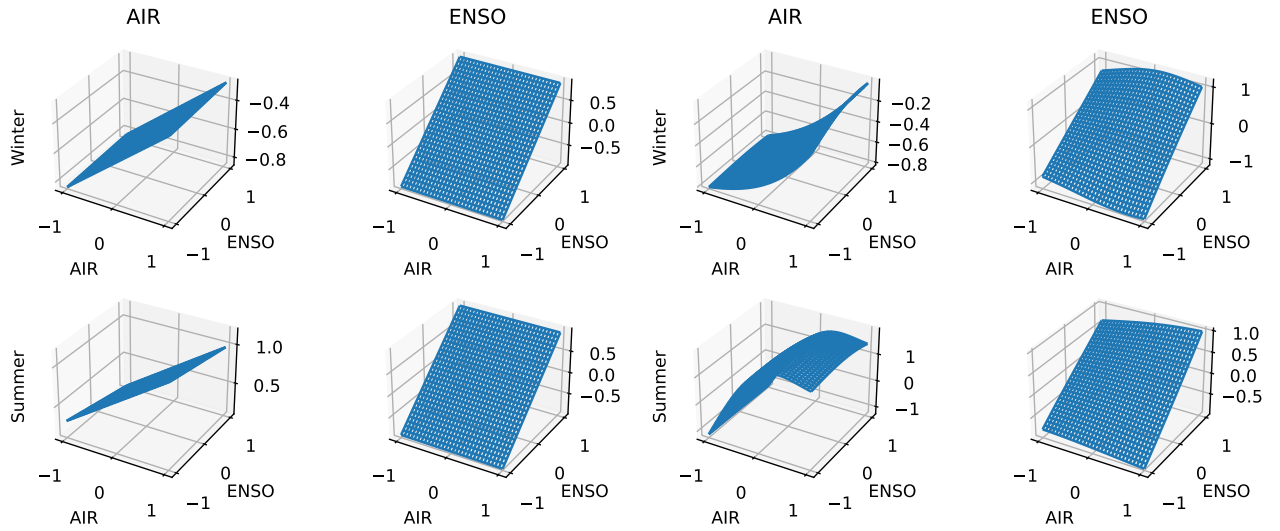


Figure 11: Posterior distribution grouped by month.

F.6. Salsa experiment

Mocap sequences The data we consider for this experiment consists on 28 *salsa dancing* sequences from the CMU mocap data. Each trial consists of a sequence with varying length, where the observations represent 3D positions of 41 joints of both participants. Following related approaches (Dong et al., 2020), we use information of one of the participants, which should be sufficient for capturing dynamics, with a total of 41×3 observations per frame. Then, we subsample the data by a factor of 4, normalise the data, and clip each sequence to $T = 200$.



(a) (b)

Figure 12: Function responses of the ENSO-AIR experiment assuming (a) linear and (b) non-linear softplus networks. Each row shows the function responses for Winter and Summer respectively.

Video sequences To train SDSs, we generate 64×64 video sequences of length $T = 200$. The dancing sequences are originally obtained from the same CMU mocap data, but they are processed into human meshes using Mahmood et al. (2019) and available on the AMASS dataset. The total processed samples from CMU salsa dancing sequences are 14. To generate videos, we subsample the sequences by a factor of 8, and augment the data by rendering human meshes with rotated perspectives and offsetting the subsampled trajectories. To do so, we adapt the available code from Mahmood et al. (2019), and generate 10080 train samples and 560 test samples. In figure 14 we show examples of reconstructed salsa videos from the test dataset using our identifiable SDS, where we observe that the method achieves high-fidelity reconstructions.

Additionally, in Figure 13 we provide forecasting results on the test data for 50 future frames from the last observation. As a reference, at each frame we compare the iSDS predictions with a black image (indicated as baseline). As we observe, the prediction error increases rapidly. More specifically, the averaged pixel MSE for the first 20 predicted frames is 0.0148, which is high compared to the reconstruction error ($2.26 \cdot 10^{-4}$). Nonetheless, this result is expected for the following reasons. First, the discrete transitions are independent of the observations, which can trigger dynamical changes that are not aligned with the groundtruth transitions. Second, the errors are accumulated over time, which combined with the previous point can rapidly cause disparities in the predictions. Finally, our formulation does not train the model based on prediction explicitly. Although the predicted frames are not aligned with the ground truth, in Figure 15 we show that our iSDS is able to produce reliable future sequences, despite some exceptions (see 6th forecast). In general, we observe that our proposed model can be used to generate reliable future dancing sequences. We note that despite the restrictions assumed to achieve identifiability guarantees (e.g. removed feedback from observations in comparison to Dong et al. (2020)), our iSDS serves as a generative model for high-dimensional sequences.

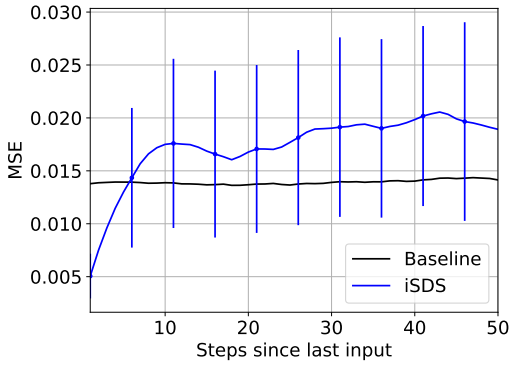


Figure 13: Forecasting averaged pixel MSE and standard deviation (vertical lines) of our iSDS using test data sequences.

F.7. Real dancing videos

To further motivate the applications of our identifiable SDS in challenging realistic domains, we consider real dancing sequences from the AIST Dance DB (Tsuchida et al., 2019). As in the previous semi-synthetic video experiment, we focus

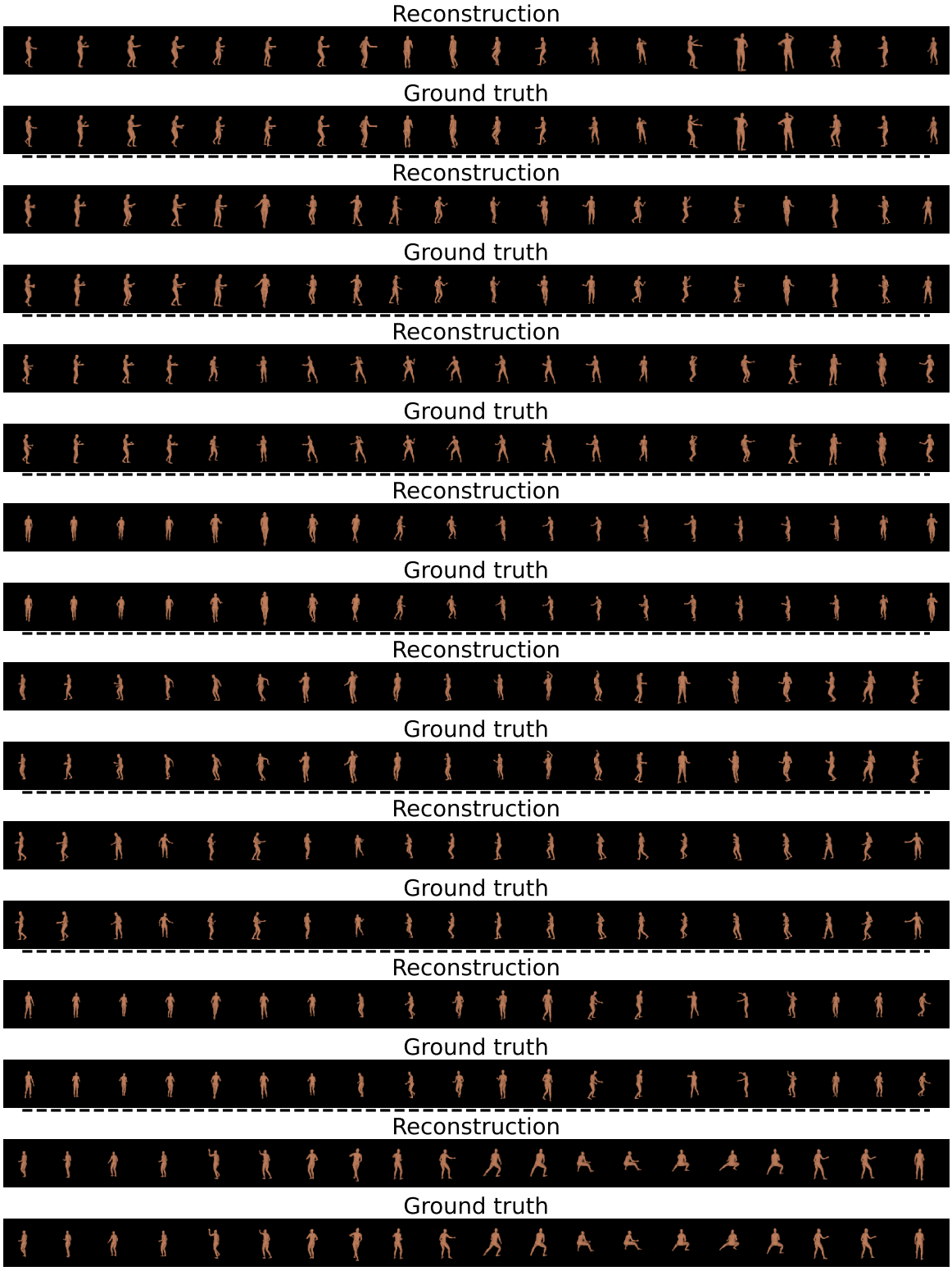


Figure 14: Reconstruction and ground truth of salsa dancing videos from the test dataset.

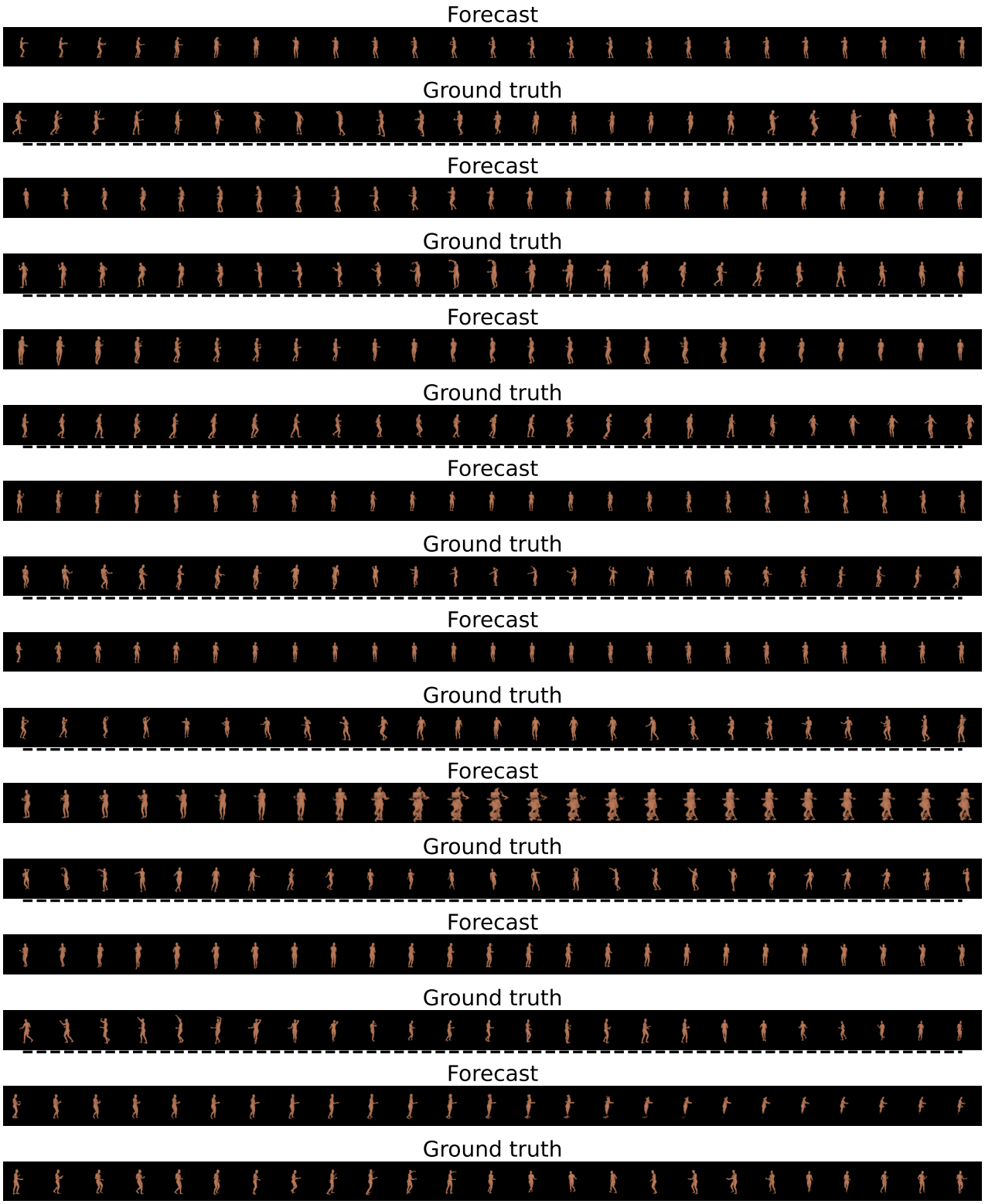


Figure 15: Forecasts and corresponding ground truth of future $T = 50$ frames from salsa test samples.

on segmenting dancing patterns from high-dimensional input. The data contains a total of 12670 sequences of varying lengths, which include 10 different dancing genres (with 1267 sequences each), different actors, and camera orientations. We focus on segmenting sequences corresponding to the *Middle Hip Hop* genre, where we leave 100 sequences for testing. We process each sequence as follows: (i) we subsample the video by 4, (ii) we crop each frame to center the dancer position, (iii) we resize each frame to 64×64, and (iv) we crop the length of the video to $T = 200$ frames.

For training, we adopt the same architecture and hyper-parameters as in the salsa dancing video experiment (See Appendix F.3; except we set a batch size of 16 this time). Here we also include a pre-training stage for the encoder-decoder networks, but in this case we include all the available dancing sequences (all genres, except the test samples). In the second stage where the transitions are learned, we use only the *Middle Hip Hop* sequences. We note that training the iSDS in this dataset is particularly challenging, as we find that the issue of state collapsed reported by related works (Dong et al., 2020; Ansari et al., 2021) is more prominent in this scenario. To mitigate this problem, we train our model using a combination of the KL annealing schedule proposed in Dong et al. (2020) and the temperature coefficient proposed in Ansari et al. (2021), which we already included previously. We start with a KL annealing term of 10^4 and decay it by a factor of 0.95 every 50 iterations, and with $\tau = 10^3$, where we decay it by a factor of 0.975 every 100 iterations. We run this second phase for a maximum of 1000 epochs.

We show reconstruction and segmentation results in Figure 16, where we observe our iSDS learns different components from the dancing sequences. In general we observe that the sequences are segmented according to different dancing moves, except for some cases where only a prominent mode is present. We also note that different prominent modes are present depending on background information. For example, the *blue* mode is prominent for white background, and the *teal* mode is prominent for combined black and white backgrounds. Such findings indicate the possibility of having data artifacts, in which the model performs segmentation based on the background information as they could be correlated with the dancing dynamics. Furthermore, our iSDS reconstructs the video sequences with high fidelity, except for fine-grained details such as the hands. Quantitatively, our approach reconstructs the sequences with an averaged pixel MSE of $7.85 \cdot 10^{-3}$. We consider this quantity is reasonable as it is an order of magnitude higher in comparison to the previous salsa videos, where the sequences were rendered on black backgrounds.

Figure 16: Reconstruction, ground truth, and segmentation of dancing videos from the AIST Dance DB test set.