On the Identifiability of Markov Switching Models

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

Abstract

In the realm of interpretability and out-of-distribution generalization, the identifiability of latent variable models has emerged as a captivating field of inquiry. In this work, we delve into the identifiability of Markov Switching Models, taking an initial stride toward extending recent results to sequential latent variable models. We develop identifiability conditions for first-order Markov dependency structures, whose transition distribution is parametrised via non-linear Gaussians. Through empirical studies, we demonstrate the practicality of our approach in facilitating regime-dependent causal discovery and segmenting high-dimensional time series data.

1. Introduction

State-space models (SSMs) were a well-established approach in the early stages of probabilistic sequential modelling (Lindgren, 1978; Poritz, 1982; Hamilton, 1989). Over the recent years, recurrent neural networks have gained popularity over SSMs, thanks to the development of LSTMs (Hochreiter & Schmidhuber, 1997) and GRUs (Cho et al., 2014) that allow the modelling of longer-term dependencies. More recently, Gu et al. (2022) has achieved promising results with *structured state-spaces*. From a data-generating process perspective, these techniques can be framed as sequential latent variable models, such as the Variational Recurrent Neural Network (Chung et al., 2015) or other further extensions (Li & Mandt, 2018; Babaeizadeh et al., 2018; Saxena et al., 2021; Linderman et al., 2016; 2017). Sequential generative models have been combined with SSMs by e.g. setting them as priors (Fraccaro et al., 2017; Dong et al., 2020; Ansari et al., 2021; Smith et al., 2022). Despite these efforts in incorporating more flexible priors and developing stabler training schemes, theoretical properties such as identifiability or consistency are less studied, contrary to the early literature on SSMs.

Identifiability is an important property in a wide variety of contexts; it, in general, establishes a one-to-one correspondence between the data likelihood and the model parameters, or an equivalence class of the latter. In causal inference (Peters et al., 2017), identifiability refers to whether the underlying causal structure can be estimated from observational data. In ICA (Comon, 1994), we aim to identify the latent sources and the mapping from the latents to the observed. General nonlinear ICA is ill-defined (Hyvärinen & Pajunen, 1999). However, recent results show that identfiability can be achieved using conditional priors (Khemakhem et al., 2020). Moreover, in a framework for deep (non-temporal) latent variable models, the required access to auxiliary variables can be relaxed using a mixture prior (Kivva et al., 2022). Despite these developments, extending these results to sequential models remains an open question with limited understanding (Hälvä et al., 2021; Yao et al., 2022).

In this work, we consider the identifiability for Markov Switching Models (MSMs) (Hamilton, 1989), an extension of a Hidden Markov Model with autoregressive connections. Identifiable MSMs could serve as prior distributions to enable identifiable sequential latent variable models. We present conditions in which the conditional autorregressive distribution is identifiable for the non-linear Gaussian case. We validate our theoretical results with synthetic experiments, and showcase the approach for high-dimensional time-series segmentation. Notably, our results can achieve identifiability for first-order regime-dependent causal discovery in the absence of instantaneous effects, enabling us to study time-dependent causal structures in climate data.

2. Theoretical considerations

2.1. Identifiable Markov Switching Models

Let $\boldsymbol{x}_t \in \mathbb{R}^d$ with $t \in \{1, \dots T\}$ denote observed random variables from a discrete time series following a Markov Switching Model (MSM) (Hamilton, 1989). Such models consider discrete latent random variables $s_t \in \{1, \dots, K\}$ which condition the distribution of the observations. The conditional distribution follows an autoregressive process.

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}) = \sum_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}) p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\mathbf{s}_{1:T}), \tag{1}$$

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

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

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling.

where we consider first-order processes and the distribution is parametrised by $\boldsymbol{\theta}$. Note that the distribution of the latent states $p(\mathbf{s}_{1:T})$ is not specified. Our goal is to explore the assumptions on the model so that the transition distribution family is identifiable. Below we discuss the Gaussian family and defer the non-parametric case to Appendix B.2.

The MSM can be formulated as a finite mixture model where $c_i = p(\mathbf{s}_{1:T} = \{a_1^i, a_2^i, \dots, a_T^i\})$ represents the probability of a Markov chain with $i \in \{1, \dots, K^T\}$ and $p_{a^i_{1:T}}(\boldsymbol{x}_{1:T}) = p(\boldsymbol{x}_{1:T}|\mathbf{s}_{1:T} = a^i_{1:T})$ denotes the joint probability given some trajectory i. Let us define the family of initial and transition distributions

$$\Pi_A := \{p_a(\boldsymbol{x}_1)|a \in A\},$$
$$\mathcal{P}_A := \{p_a(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})|a \in A\}, \quad |A| = K. \quad (3)$$

Note $\mathcal{P}_A$ incorporates the first-order Markov assumption. Then, we denote the family of first-order MSMs as

$$\mathcal{H}^T(\Pi_A, \mathcal{P}_A) := \left\{ \sum_{i=1}^{K^T} c_i p_{a_1^i}(\boldsymbol{x}_1) \prod_{t=2}^{T} p_{a_t^i}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \right.$$
$$\left. a_t^i \in A, a^i_{1:T} \neq a^j_{1:T}, \forall i \neq j, \sum_{i=1}^{K^T} c_i = 1 \right\}. \quad (4)$$

We define the identification of $\mathcal{H}^T(\Pi_A, \mathcal{P}_A)$ as follows.

Definition 2.1. The family $\mathcal{H}^T(\Pi_A, \mathcal{P}_A)$ that contains first-order MSMs is said to be identifiable up to permutations, when for $p_1(\boldsymbol{x}_{1:T}) = \sum_{i=1}^N c_i p_{a_1^i}(\boldsymbol{x}_1) \prod_{t=2}^T p_{a_t^i}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p_2(\boldsymbol{x}_{1:T}) = \sum_{i=1}^M \hat{c}_i p_{\hat{a}_1^i}(\boldsymbol{x}_1) \prod_{t=2}^T p_{\hat{a}_t^i}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, $p_1(\boldsymbol{x}_{1:T}) = p_2(\boldsymbol{x}_{1:T}), \forall \boldsymbol{x}_{1:T} \in \mathbb{R}^{Td}$, if and only if $M = N$ and for each $1 \leq i \leq N$ there is some $1 \leq j \leq M$ s.t.

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

We note that the 2nd requirement above eliminates the permutation equivalence of e.g., $(1, 2, 3, 2)$ and $(3, 1, 2, 3)$ which would be valid in the finite mixture case with vector indexing. Yakowitz & Spragins (1968) proved the identifiability result for finite mixtures by imposing linear independence on a family of d-dimensional CDFs (see App. A for a more formal statement). We extend this result to the MSM family. First, we introduce the concept of linearly independent functions under finite mixtures as follows.

Definition 2.2. A family of functions $\{f_a(\boldsymbol{x})|a \in A\}$ is said to contain linearly independent functions under finite mixtures, if for any $A_0 \subset A$ such that $|A_0| < +\infty$, the functions in $\{f_a(\boldsymbol{x})|a \in A_0\}$ 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. Given the previous definition we have the following result (see Appendix B.1 for the proof).

Theorem 2.3. *Assume the functions in $\mathcal{P}_A^T := \Pi_A \otimes (\otimes_{t=2}^T \mathcal{P}_A)$ are linearly independent under finite mixtures, then the distribution family $\mathcal{H}^T(\Pi_A, \mathcal{P}_A)$ is identifiable as defined in Definition 2.1.*

Therefore, we need to specify linearly independent distribution families $\mathcal{P}_A^T$ to construct identifiable MSMs, which will be discussed in the next section.

2.2. Parametrisation via Gaussian distribution

We seek to define a parametric family of conditional distributions for which functions in it are linearly independent. To do so we define a Gaussian family as follows:

$$\mathcal{G}_A = \{p_a(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}(\boldsymbol{x}_{t-1}, a),$$
$$\boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a))|a \in A, \boldsymbol{x}_t, \boldsymbol{x}_{t-1} \in \mathbb{R}^d\}, \quad (5)$$

where $\boldsymbol{m}(\boldsymbol{x}_{t-1}, a)$ and $\Sigma(\boldsymbol{x}_{t-1}, a)$ define the mean and covariance matrix for the Gaussian. We further require the following *unique indexing* assumption:

$$\forall a \neq a' \in A,\ \exists \boldsymbol{x}_{t-1} \in \mathbb{R}^d,\ s.t.\ \boldsymbol{m}(\boldsymbol{x}_{t-1}, a)$$
$$\neq \boldsymbol{m}(\boldsymbol{x}_{t-1}, a') \text{ or } \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a) \neq \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a'). \quad (6)$$

In other words, for such $\boldsymbol{x}_{t-1}$ we have $p_a(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p_{a'}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ as two different Gaussian distributions. To achieve linear independence of the joint distributions and prove identifiability in the sense of Definition 2.1, we also need to introduce a family of *initial distributions* $\mathcal{I}_A$.

$$\mathcal{I}_A := \{p_a(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}(a), \boldsymbol{\Sigma}_1(a))|a \in A\}, \quad (7)$$

where we also assume unique indexing for $a \in A$, i.e.,

$$a \neq a' \in A \Leftrightarrow \boldsymbol{\mu}(a) \neq \boldsymbol{\mu}(a') \text{ or } \boldsymbol{\Sigma}_1(a) \neq \boldsymbol{\Sigma}_1(a'). \quad (8)$$

The above Gaussian distribution families paired with unique indexing assumptions satisfy conditions which favour linear independence of the joint distribution family.

Theorem 2.4. *Define the following joint distribution family under the non-linear Gaussian model*

$$\mathcal{P}_A^T = \left\{ p_{a_1, a_{2:T}}(\boldsymbol{x}_{1:T}) = p_{a_1}(\boldsymbol{x}_1) \prod_{t=2}^{T} p_{a_t}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \right.$$
$$\left. a_t \in A, \quad p_{a_1} \in \mathcal{I}_A, \quad p_{a_t} \in \mathcal{G}_A, \quad t = 2, ..., T \right\}, \quad (9)$$

with $\mathcal{G}_A$, $\mathcal{I}_A$ defined by Eqs. (5), (7) respectively. Assume:

(a1) Unique indexing for $\mathcal{G}_A$ and $\mathcal{I}_A$: Eqs. (6), (8) hold;

(a2) Continuity for the conditioning input: distributions in $\mathcal{G}_A$ are continuous w.r.t. $\boldsymbol{x}_{t-1} \in \mathbb{R}^d$;

(a3) Zero-measure intersection in certain region: there exists a non-zero measure set $\mathcal{X}_0 \subset \mathbb{R}^d$ s.t. $\{\boldsymbol{x}_{t-1} \in \mathcal{X}_0 | \boldsymbol{m}(\boldsymbol{x}_{t-1}, a) = \boldsymbol{m}(\boldsymbol{x}_{t-1}, a'), \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a')\}$ has zero measure, for any $a \neq a'$;

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

See Appendix B.2 and B.3 for proofs. The strategy can be summarised in 3 steps. (i) By using the conditional first-order Markov assumption on $p_{a^i_{1:T}}(\boldsymbol{x}_{1:T})$ we just need to show conditions for the linear independence of $\{p_{a^i_{1:2}}(\boldsymbol{x}_{1:2})\}$, and then we prove $T \geq 3$ by induction. (ii) To satisfy the conditions for linear independence of $\{p_{a^i_{1:2}}(\boldsymbol{x}_{1:2})\}$, we specify conditions on $\{p_{a^i_t}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\}$ and $\{p_{a^i_1}(\boldsymbol{x}_1)\}$ in non-parametric case. (iii) We show the Gaussian family with (a1-a3) is a special case of the above. The remaining question is how to parameterise the Gaussian moments such that assumption (a3) holds.

Corollary 2.5. *A Gaussian parametrisation of the conditional distribution via multivariate analytic functions,* $p_a(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{m}(\boldsymbol{x}_{t-1}, a), \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a))$ *with unique indexing (Eq. 6), renders the MSM identifiable.*

This allows parametrisations via polynomials, or even neural networks with analytic activation functions (e.g. SoftPlus).

3. Estimation

Assume we have a dataset $\mathcal{D}$ with N sequences of length T generated by Eq. 1, $\boldsymbol{x}^n_{1:T} = \{\boldsymbol{x}^n_1, \ldots, \boldsymbol{x}^n_T\} \sim \mathcal{D}$. Although we do not impose any restrictions on the distribution of the states, in this section we describe our implementation in terms of a first-order stationary Markov chain as an example. Our approach uses the expectation maximisation (EM) algorithm, which is an efficient framework for maximising the likelihood of mixture models (Bishop, 2006). Below we provide the M-step of the transition distribution parametrised by neural networks and more details (including polynomial parametrisations) can be found in Appendix B.4. Since for this case the exact updates cannot be computed, we take a Generalised EM (GEM) (Dempster et al., 1977) approach where a gradient ascent step is performed

$$\boldsymbol{\theta}^{\text{new}} \leftarrow \boldsymbol{\theta}^{\text{old}} + \eta \sum_{n=1}^{N} \sum_{t=2}^{T} \sum_{k=1}^{K} \gamma^n_{t,k} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^n_t | \boldsymbol{x}^n_{t-1}, s^n_t = k), \quad (10)$$

where $\gamma^n_{t,k} = p_{\boldsymbol{\theta}}(s^n_t = k | \boldsymbol{x}^n_{1:T})$ and the update rule can be computed using back-propagation.

This approach is well-established in the literature (Bengio & Frasconi, 1994; Hälvä & Hyvarinen, 2020), and convergence is guaranteed to a local maximum of the likelihood. To ensure our method converges to the MLE solution, we will assume N is large and use stochastic gradient ascent updates.

4. Related work

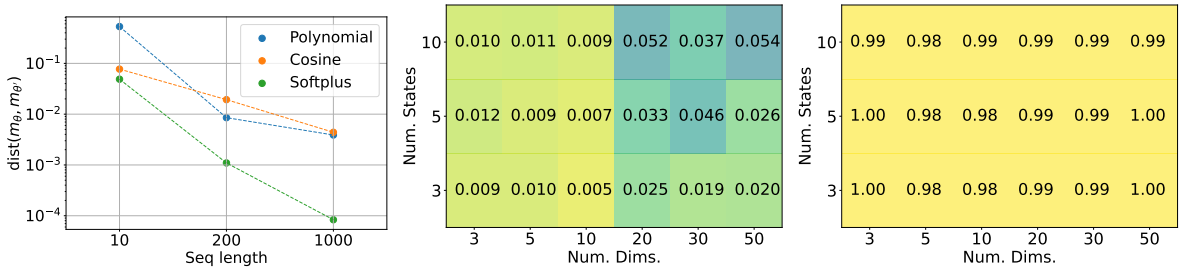
Markov Switching Models were first introduced by Poritz (1982) as switching linear auto-regressive processes. This family of state-space models re-use the forward-backward recursions (Rabiner, 1989) for tractable posterior estimation and have been studied decades ago for speech analysis (Poritz, 1982; Ephraim & Roberts, 2005) and economics (Hamilton, 1989). Frühwirth-Schnatter & Frèuhwirth-Schnatter (2006) review standard estimation approaches and applications of the MSM family. Although the majority of the proposed approaches estimate the parameters using tractable MLE solutions for their asymptotic properties, 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 the joint distribution given the states under relaxed assumptions. To the best of our knowledge, identifiability has only been explicitly studied for the first-order discrete case in An et al. (2013). Here, the central assumption is stationarity and ergodicity of the latent and observed variables, which uniquely determines the joint probability of four consecutive observations as shown by White (2001).

5. Experiments

Synthetic experiments We generate data according to our assumptions and evaluate the estimated functions in terms of the ground truth. For K components, the error is computed in terms of the averaged L2 distance using the permutation that gives lowest error. We use fixed covariances and parametrise the transition means using random cubic polynomials and networks with cosine or SoftPlus activations. When increasing dimensions, we use locally connected networks (Zheng et al., 2018) to encourage sparsity and generate stable samples. See App. C.1,C.4, and C.3 for details on the error computation, data and training.

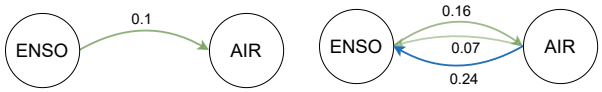
Figure 1a shows increasing the sequence length generally reduces the L2 distance error, which is expected. Considering the different function forms, we observe that the polynomials are estimated with higher error for short sequences, which could be caused by the high frequency components from the cubic terms. Figure 1b, shows low parameter errors even when increasing the number of dimensions and components. The locally connected networks are implemented by first sampling a random regime-dependent causal structure. Therefore, we evaluate the estimated causal structure in Figure 1c, which we calculate via thresholding the averaged Jacobian (see App. C.2 for details). The results show the MSM with nonlinear transitions is able to maintain high F1-scores, despite the differences in L2 distance when increasing dimensions and states (1b). Although the approach is restricted by first-order Markov assumptions, the synthetic setting shows promising directions for causal discovery with high-dimensional regime-dependent dynamics.

Regime-dependent causal discovery To motivate identifiable MSMs for regime-dependent causal discovery, we



(a) Function forms (b) L2 distance (c) Avg. F1 score

Figure 1: Synthetic experiment results. (a) L2 distance error using different transition functions with varying T. (b) L2 distance error and (c) averaged F1 score of nonlinear data (cosine activations) with increasing states and dimensions.



(a) (b)

Figure 2: Regime-dependent graphs generated assuming (a) linear and (b) non-linear effects. Green and blue lines indicate effects in summer and winter months respectively.

explore the task considered in 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 the settings in Saggioro et al. (2020) and train our approach using linear and nonlinear (softplus networks) transitions. The results are reported in Figure 2, where we show the regime-dependent graphs of both models. Our approach captures regimes based on seasonality, as one component is assigned to Summer months (from May to September), and the other is assigned to Winter months (from October to April). The weights of the regime-dependent graphs denote the absolute value of the Jacobian where we keep edges for values greater than 0.05. In the linear case, we only observe an effect from ENSO to AIR which occurs only during Summer. This result is consistent with both Saggioro et al. (2020) and the literature (Webster & Palmer, 1997), which suggest that ENSO has a direct effect on AIR during summer, but not in Winter. For nonlinear transitions (fig. 2b), we observe the previous result along with additional links which are not supported by evidence. Since assuming non-linear effects is more flexible, our results suggest finding additional links imply the presence of confounders that have influence on both variables, which should be expected in scenarios with few observations.

Segmentation of dancing patterns We showcase that our model can successfully segment high-dimensional and non-linear data despite not considering any continous latent variables compared to recent approaches (Dong et al., 2020; Ansari et al., 2021). We consider *salsa dancing* sequences from the CMU mocap data. Details on the data can be found in Appendix C.6. Figure 3 illustrates our iMSM using nonlinear (softplus networks) transitions in comparison to KVAE (Fraccaro et al., 2017). Our model assigns different dancing patterns into different states. For example, the first part of the sequence consists on forward and backward movements present in the majority of the training samples. This pattern is repeated at the end, which KVAE fails to classify. In general, both approaches present some limitations. For KVAE, the soft-switching mechanism creates a trade-off between modelling non-linear transitions and confidence in the component assignments. Our approach is limited by the first-order Markov assumption on the conditioned transition distribution, which prevents the model from learning high order moments (such as acceleration) that would provide richer features for higher-fidelity time series segmentations.

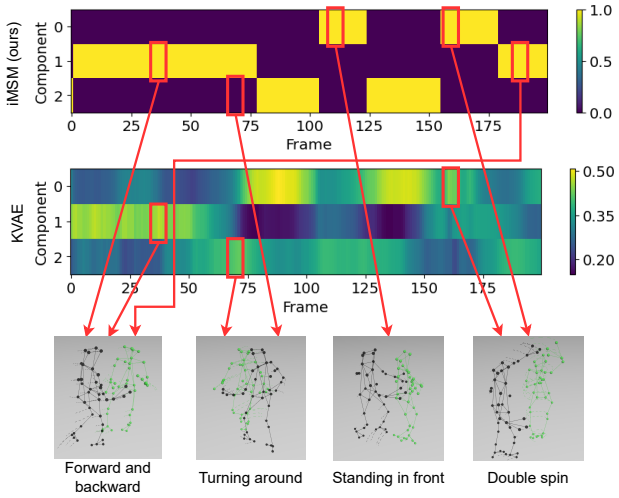


Figure 3: Posterior probability of a *salsa dancing* sequence of our approach (iMSM) and KVAE (Fraccaro et al., 2017) along with several patterns distinguished in the example.

6. Conclusions

We present identifiability analysis regarding Markov Switching Models. Key to our contribution is the parametrisation of the Gaussian mean via analytic functions, which establish identifiability of the transition functions independently of the marginal distribution of the states. We empirically verify our theoretical results with synthetic experiments, and motivate our approach for regime-dependent causal discovery and time series segmentation with real data.

Acknowledgements

Y.W. acknowledges grant support by the National Science Foundation and the Office of Naval Research.

References

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.

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

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.

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

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.

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.

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.

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

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.

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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. W. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

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.

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

White, H. *Asymptotic theory for econometricians*. Economic theory, econometrics, and mathematical economics. Emerald, Bingley, rev. ed. edition, 2001. ISBN 9780127466521.

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

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

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

where $F_a(\boldsymbol{x})$ is an d-dimensional CDF. Now consider the following finite mixture distribution family by linearly combining the CDFs in $\mathcal{F}$:

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

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 result for finite mixtures. Recall, Def. 2.2 for linear independence under finite mixtures, we state the following identifiability result

Proposition A.2. *(Yakowitz & Spragins, 1968) The finite mixture distribution family $\mathcal{H}$ is identifiable up to permutations, iff. functions in $\mathcal{F}$ are linearly independent under finite mixtures.*

B. Proofs

B.1. Proof of Theorem 2.3

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

Proposition B.1. *Consider the distribution family given by Eq. 4. Then the joint distribution in $\mathcal{H}^T(\Pi_A, \mathcal{P}_A)$ is identifiable up to permutations if and only if functions in $\mathcal{P}_A^T$ are linearly independent under finite mixtures.*

We now can prove Theorem 2.3.

Proof. From proposition B.1 we see that, $\mathcal{P}_A^T$ being linearly independent implies identifiability up to permutation for $\mathcal{H}^T(\Pi_A, \mathcal{P}_A)$ in the finite mixture sense (Definition A.1). This means for $p_1(\boldsymbol{x}_{1:T})$ and $p_2(\boldsymbol{x}_{1:T})$ defined in Definition 2.1, we have $M = N$ and for every $1 \leq i \leq N$, there exists $1 \leq j \leq M$ such that $c_i = \hat{c}_j$ and

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

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

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

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

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

Lastly, if there exists $t_1 \neq t_2$ such that $a^i_{t_1} = a^i_{t_2}$ but $\hat{a}^j_{t_1} \neq \hat{a}^j_{t_2}$, then the proved fact that, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$,

$$\begin{aligned} p_{\hat{a}^j_{t_1}}(\boldsymbol{x}_{t_1} = \boldsymbol{\beta} | \boldsymbol{x}_{t_1-1} = \boldsymbol{\alpha}) &= p_{a^i_{t_1}}(\boldsymbol{x}_{t_1} = \boldsymbol{\beta} | \boldsymbol{x}_{t_1-1} = \boldsymbol{\alpha}) \\ &= p_{a^i_{t_2}}(\boldsymbol{x}_{t_2} = \boldsymbol{\beta} | \boldsymbol{x}_{t_2-1} = \boldsymbol{\alpha}) \\ &= p_{\hat{a}^j_{t_2}}(\boldsymbol{x}_{t_2} = \boldsymbol{\beta} | \boldsymbol{x}_{t_2-1} = \boldsymbol{\alpha}), \end{aligned}$$

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

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

$$\sum_{i,j} \gamma_{ij} p_{a^i_{1:t-1}}(\boldsymbol{x}_{1:t-1}) p_{a^j_t}(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = 0, \quad \forall \boldsymbol{x}_{1:t} \in \mathbb{R}^{(t-1)d} \times \mathbb{R}^d,$$

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

$$\sum_i \left(\sum_j \gamma_{ij} p_{a^j_t}(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) \right) p_{a^i_{1:t-1}}(\boldsymbol{x}_{1:t-1}) = 0, \quad \forall \boldsymbol{x}_{1:t} \in \mathbb{R}^{(t-1)d} \times \mathbb{R}^d,$$

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

□

B.2. Proof of Theorem 2.4

As described in the main text, to prove identifiability in the sense of Definition 2.1, we require 4 steps. First, we need to show linear independence of the conditional distribution family $\{p(\boldsymbol{x}_{1:T} | \mathbf{s}_{1:T})\}$, which is why the result of Theorem 2.4 refers to the linear independence of Eq. (9), which is formulated as a finite mixture. Therefore, we aim to explore the non-parametric case under some reasonable restrictions to the joint distribution family $\mathcal{P}^T_A$ to obtain linear independence under finite mixtures and render the MSM identifiable.

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

Lemma B.2. *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 independent functions 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

$$\sum_{(i,j)\in S_0} \gamma_{ij} v_j(\boldsymbol{z}, \mathbf{y}) u_i(\mathbf{y}, \boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \mathbf{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{d_y} \times \mathcal{Z}. \tag{11}$$

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. (11)) for any $\boldsymbol{\beta} \in \mathbb{R}^{d_y}$ as

$$\sum_{j\in J_0} \left(\sum_{i:(i,j)\in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) \right) v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{12}$$

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:

$$\begin{aligned} J_0 = \cup_{k\in K(\boldsymbol{\beta})} J_k(\boldsymbol{\beta}), \quad J_k(\boldsymbol{\beta}) \cap J_{k'}(\boldsymbol{\beta}) = \emptyset, \forall k \neq k' \in K(\boldsymbol{\beta}), \\ j \neq j' \in J_k(\boldsymbol{\beta}) \quad \Leftrightarrow \quad v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) = v_{j'}(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}), \quad \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned} \tag{13}$$

Then we can rewrite Eq. (12) for any $\boldsymbol{\beta} \in \mathcal{Y}_1$ as

$$\sum_{k\in K(\boldsymbol{\beta})} \left(\sum_{j\in J_k(\boldsymbol{\beta})} \sum_{i:(i,j)\in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) v_j(\boldsymbol{z}, \mathbf{y} = \boldsymbol{\beta}) \right) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}. \tag{14}$$

Recall from Eq. (13) 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. (11) holds, then for any $\boldsymbol{\beta} \in \mathcal{Y}_1$, under assumptions (b1) and (b5),

$$\sum_{j\in J_k(\boldsymbol{\beta})} \sum_{i:(i,j)\in S_0} \gamma_{ij} u_i(\mathbf{y} = \boldsymbol{\beta}, \boldsymbol{x}) = 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \quad k \in K(\boldsymbol{\beta}). \tag{15}$$

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. (13)) 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. (15) 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. □

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

Theorem B.3. *Define the following joint distribution family*
$$\left\{ p_{a_1, a_{2:T}}(\boldsymbol{x}_{1:T}) = p_{a_1}(\boldsymbol{x}_1) \prod_{t=2}^{T} p_{a_t}(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}), \quad p_{a_1} \in \Pi_A, \quad p_{a_t} \in \mathcal{P}_A, t = 2, ..., T \right\},$$
and assume Π_A and $\mathcal{P}_A$ satisfy assumptions (b1)-(b6) as follows,

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

(c2) Π_A satisfies (b3), and

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

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

Proof. We proceed to prove the statement by induction as follows.

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

(2) $T > 2$: Assume the statement holds for the joint distribution family when $T = \tau - 1$. Note that we can write $p_{a_{1:\tau}}(\boldsymbol{x}_{1:\tau})$ as
$$p_{a_{1:\tau}}(\boldsymbol{x}_{1:\tau}) = p_{a_1, a_{2:\tau-1}}(\boldsymbol{x}_{1:\tau-1}) p_{a_\tau}(\boldsymbol{x}_\tau | \boldsymbol{x}_{\tau-1}).$$
Then the statement for $T = \tau$ can be proved using Lemma B.2 by setting $u_i(\mathbf{y} = \boldsymbol{x}_{\tau-1}, \boldsymbol{x} = \boldsymbol{x}_{1:\tau-2}) = p_{a_{1:\tau-1}}(\boldsymbol{x}_{1:\tau-1}), i = a_{1:\tau-1}$, and $v_j(\boldsymbol{z} = \boldsymbol{x}_\tau, \mathbf{y} = \boldsymbol{x}_{\tau-1}) = p_{a_\tau}(\boldsymbol{x}_\tau | \boldsymbol{x}_{\tau-1}), j = a_\tau$. Note that the family spanned with $p_{a_{1:\tau-1}}(\boldsymbol{x}_{1:\tau-1}), i = a_{1:\tau-1}$ satisfies (b1) and (b2) from Π_B and $\mathcal{P}_A$ 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).

As described, in the final step of the proof we explore properties of the Gaussian transition and initial distribution families (Eqs. (5) and (7) respectively). The unique indexing assumption of the Gaussian transition family (Eq. (6)) implies linear independence as shown below.

Proposition B.4. *Functions in $\mathcal{G}_A$ are linearly independent on variables $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ if the unique indexing assumption (Eq. (6)) holds.*

Proof. Assume the statement is false, then there exists $A_0 \subset A$ and a set of non-zero values $\{\gamma_a | a \in A_0\}$, such that

$$\sum_{a \in A_0} \gamma_a \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}(\boldsymbol{x}_{t-1}, a), \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a)) = 0, \quad \forall \boldsymbol{x}_t, \boldsymbol{x}_{t-1} \in \mathbb{R}^d.$$

In particular, this equality holds for any $\boldsymbol{x}_{t-1} \in \mathbb{R}^d$, meaning that a weighted sum of Gaussian distributions (defined on $\boldsymbol{x}_t$) 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 $\boldsymbol{x}_{t-1}$

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

a contradiction to the unique indexing assumption. □

We now draw some connections from the previous Gaussian families to assumptions (b1-b6) in Lemma B.2.

Proposition B.5. *The conditional Gaussian distribution family $\mathcal{G}_A$ (Eq. (5), under the unique indexing assumption (Eq. (6), satisfies assumptions (b1), (b2) and (b5) in Lemma B.2, if we define $\mathcal{V}_J := \mathcal{G}_A, \boldsymbol{z} := \boldsymbol{x}_t$ and $\mathbf{y} := \boldsymbol{x}_{t-1}$.*

Proposition B.6. *The initial Gaussian distribution family $\mathcal{I}_A$ (Eq. (7), under the unique indexing assumption (Eq. (8), satisfies assumptions (b1), (b2) and (b3) in Lemma B.2, if we define $\mathcal{U}_I := \mathcal{I}_A, \mathbf{y} := \boldsymbol{x}_1$ and $\boldsymbol{x} = \mathcal{X} = \emptyset$.*

To see why $\mathcal{G}_A$ satisfies (b5), notice Gaussian densities are analytic in x_t. Similar ideas apply to show that $\mathcal{I}_A$ satisfies (b3). With the previous results, we can prove Theorem 2.4 as follows.

Proof. Note that assumptions (b1) - (b3) and (b5) are satisfied due to Propositions B.5 and B.6, and assumptions (b6) and (a2) are equivalent, and assumption (b4) holds due to assumption (a3). To show (a3) $\Longrightarrow$ (b4), We first define $\mathcal{V}_J := \mathcal{G}_A$, $\boldsymbol{z} := \boldsymbol{x}_t$, and $\mathbf{y} := \boldsymbol{x}_{t-1}$ from Prop. B.5. From (a3), $\mathcal{Y} := \mathcal{X}_0$ 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 (a3), $\mathcal{D}$ has measure zero and thus, (b4) holds since $\mathcal{Y}'$ is a non-zero measure set.

Then, the statement holds by Theorem B.3. □

B.3. Proof of Corollary 2.5

Proof. Let $\boldsymbol{m}(\cdot, a) : \mathbb{R}^d \to \mathbb{R}^d$ be a multivariate analytic function, which allows a multivariate Taylor expansion. (a1) is satisfied from our premise of unique indexing, and the corresponding Taylor expansion of $\boldsymbol{m}(\cdot, a)$ implies (a2). Similar logic applies to $\boldsymbol{\Sigma}(\cdot, a)$. To show (a3), we note for any $a \neq a'$ the set of intersection of moments, i.e. $\{\boldsymbol{x} \in \mathbb{R}^d | \boldsymbol{m}(\boldsymbol{x}, a) = \boldsymbol{m}(\boldsymbol{x}, a'), \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a')\}$ can be separated as the intersection of the sets $\{\boldsymbol{x} \in \mathbb{R}^d | \boldsymbol{m}(\boldsymbol{x}, a) = \boldsymbol{m}(\boldsymbol{x}, a')\}$ and $\{\boldsymbol{x} \in \mathbb{R}^d | \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a) = \boldsymbol{\Sigma}(\boldsymbol{x}_{t-1}, a')\}$. Wlog, the set $\{\boldsymbol{x} \in \mathbb{R}^d | \boldsymbol{m}(\boldsymbol{x}, a) = \boldsymbol{m}(\boldsymbol{x}, a')\}$ is the zero set of an analytic function $\mathbf{f} := \boldsymbol{m}(\cdot, a) - \boldsymbol{m}(\cdot, a')$. Proposition 0 in Mityagin (2015) shows that the zero set of a real analytic function on $\mathbb{R}^d$ has zero measure unless $\mathbf{f}$ is identically zero. Hence, the intersection of moments has zero measure from our premise of unique indexing.

Since (a1-a3) are satisfied, by Theorem 2.4 we have linear independence of the joint distribution family, which by Theorem 2.3 implies identifiability of the MSM in the sense of Def. 2.1. □

B.4. Estimation details

We provide additional details from the descritpions in the main text. For convenience, we denote the whole set of observations and latent variables $\{\boldsymbol{x}_{1:T}^n\}_{n=1}^N$ and $\{\mathbf{s}_{1:T}^n\}_{n=1}^N$ as $\mathbf{X}$ and $\mathbf{S}$ respectively.

Recall we formulate our method in terms of the expectation maximization (EM) algorithm. Given some arrangement of the parameter values (θ'), the E-step computes the posterior distribution of the latent variables $p_{\theta'}(\mathbf{S}|\mathbf{X})$. This can then be used to compute the expected log-likelihood of the complete data (latent variables and observations),

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}') := \mathbb{E}_{p_{\boldsymbol{\theta}'}(\mathbf{S}|\mathbf{X})}\left[\sum_{n=1}^N \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}^n, \mathbf{s}_{1:T}^n)\right]. \tag{16}$$

Given a first-order stationary Markov chain, we denote the posterior probability $p_{\boldsymbol{\theta}}(s_t^n = k|\mathbf{X})$ as $\gamma_{t,k}^n$, and the joint posterior of two consecutive states $p_{\boldsymbol{\theta}}(s_t^n = k, s_{t-1}^n = l|\mathbf{X})$ as $\xi_{t,k,l}^n$. 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. (16),

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{n=1}^N\sum_{k=1}^K \gamma_{1,k}^n \log \pi_k + \sum_{n=1}^N\sum_{t=2}^T\sum_{k=1}^K\sum_{l=1}^K \xi_{t,k,l}^n \log Q_{lk} + \\ \sum_{n=1}^N\sum_{k=1}^K \gamma_{1,k}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_1^n|, s_1^n = k) + \sum_{n=1}^N\sum_{t=2}^T\sum_{k=1}^K \gamma_{t,k}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^n|\boldsymbol{x}_{t-1}^n, s_t^n = k),\end{aligned} \tag{17}$$

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 result 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:

$$\boldsymbol{m}(\boldsymbol{x}_{t-1}, k) = \sum_{c=1}^C A_{k,c}\hat{\boldsymbol{x}}_{c,t-1}, \quad \hat{\boldsymbol{x}}_{t-1}^T = \begin{pmatrix}1 & x_{t-1,1} & \dots & x_{t-1,d} & x_{t-1,1}^2 & x_{t-1,1}x_{t-1,2} & \dots\end{pmatrix}, \tag{18}$$

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

$$A_k \leftarrow \left(\sum_{n=1}^N\sum_{t=2}^T \gamma_{t,k}^n \boldsymbol{x}_t \hat{\boldsymbol{x}}_{t-1}^T\right)\left(\sum_{n=1}^N\sum_{t=2}^T \gamma_{t,k}^n \hat{\boldsymbol{x}}_{t-1}\hat{\boldsymbol{x}}_{t-1}^T\right)^{-1}. \tag{19}$$

In the main text we already discussed the case where the transition means are parametrised by neural networks.

C. Experiment details

C.1. L2 distance computation

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

$$\text{err} := \min_{\mathbf{k}=\text{perm}(\{1,\dots,K\})} \frac{1}{K}\sum_{i=1}^K d(\boldsymbol{m}(\cdot, i), \hat{\boldsymbol{m}}(\cdot, k_i)), \tag{20}$$

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 $\mathbb{R}^d$ and averaging the euclidean distance, more specifically we sample 10^5 in the $[-1,1]^d$ interval for each evaluation.

$$d(\boldsymbol{f},\boldsymbol{g}) := \int_{\boldsymbol{x}\in[-1,1]^d} \sqrt{\left|\left|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x})\right|\right|^2} d\boldsymbol{x} \approx \frac{1}{10^5}\sum_{i=1}^{10^5} \sqrt{\left|\left|\boldsymbol{f}\left(\boldsymbol{x}^{(i)}\right)-\boldsymbol{g}\left(\boldsymbol{x}^{(i)}\right)\right|\right|^2}, \quad \boldsymbol{x}^{(i)} \sim Uniform([-1,1]^d) \tag{21}$$

Note that the cost of resolving the permutation inderterminacy 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 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 $\mathcal{O}(K^2)$.

C.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 inputs 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 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 $\hat{\boldsymbol{m}}(\boldsymbol{x},k)$ as $\boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}(\boldsymbol{x})$. To ensure that the Jacobian captures the effects for 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 $\mathcal{X}_k$ with size $N_K = |\mathcal{X}_k|$ contains variables that have been selected using the posterior, i.e. $\boldsymbol{x}^{(i)} \in \mathcal{X}_k$ if $k = \arg\max p_{\boldsymbol{\theta}}(\mathbf{s}^{(i)}|\mathbf{X})$, where we use the index i to denote that $\boldsymbol{x}^{(i)}$ 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

$$\hat{G}_k := \mathbf{1}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\left|\boldsymbol{J}_{\hat{\boldsymbol{m}}(\cdot,k)}\left(\boldsymbol{x}^{(i)}\right)\right| > \tau\right), \quad \boldsymbol{x}^{(i)} \in \mathcal{X}_k, \tag{22}$$

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.

C.3. Training specifications

All the experiments are implemented in Pytorch (Paszke et al., 2019) and carried out on NVIDIA RTX 2080Ti GPUs. When training polynomials (including the linear case), we use 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. (10), e.g. neural networks, we use ADAM optimizer (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.

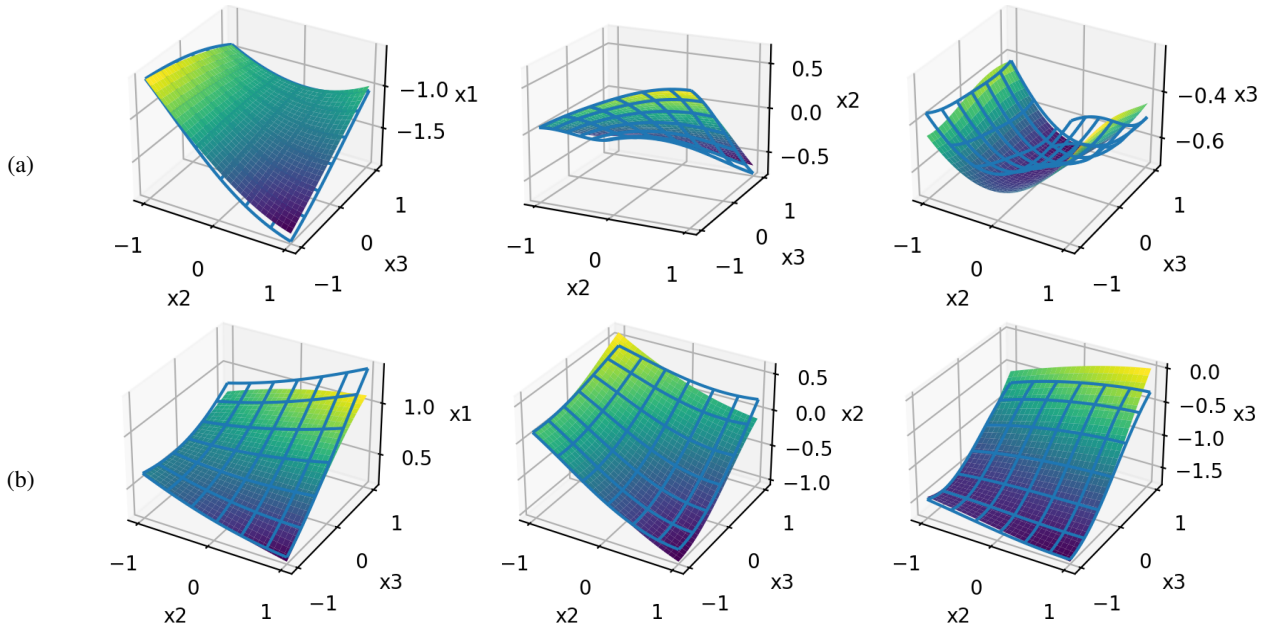


Figure 4: Function responses using 3 dimensions where we vary x_2 and x_3. Column i shows the response with respect to the i-th dimension. The blue grid shows the ground truth function for (a) 3 states showing component 1, and (b) 10 states showing component 6.

C.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 of the transition distributions are fixed to $0.05^2\mathbf{I}$, and the mean transitions $m(\boldsymbol{x}, k), k = 1, \dots, K$ are parametrised using polynomials of degree $P = 3$ 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 on two-layer MLPs with 16 hidden units.

In Figure 4, we show visualisations of some function responses for the experiment considering increasing variables and states (figs. 1b and 1c). Recall that for $K = 3$ states and $d = 3$ dimensions we achieve $9 \cdot 10^{-3}$ L2 distance error and the responses in Figure 4a show low discrepancies with respect to the ground truth. Similar observations can be made with $K = 10$ states in Figure 4b, where the L2 distance error is $\cdot 10^{-2}$.

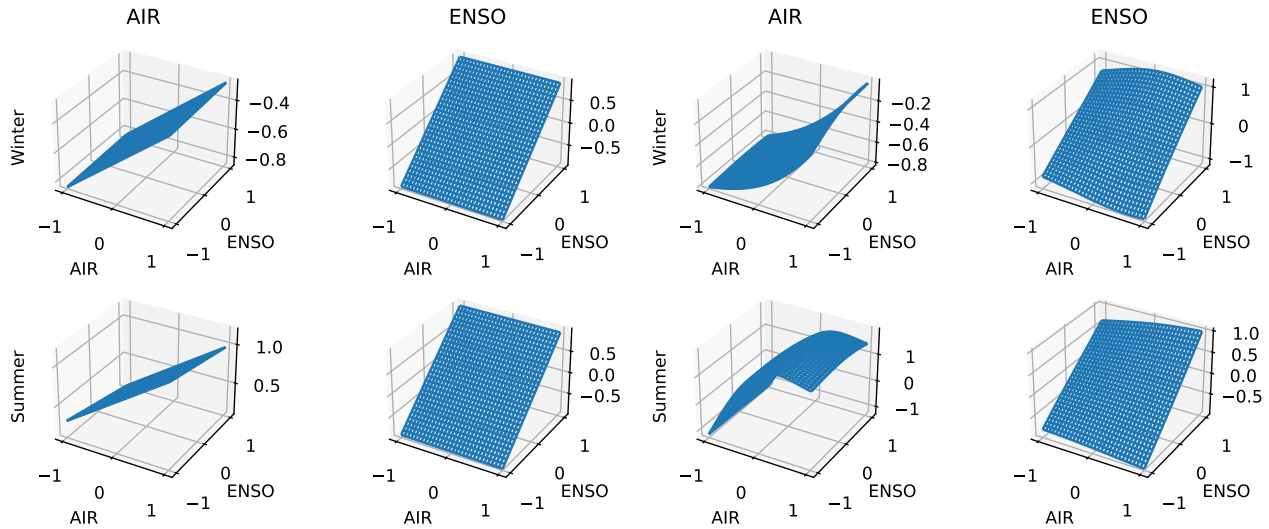
C.5. ENSO-AIR experiment

The data consists on 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 we use the indicators *Niño 3.4 SST Index*[2] and *All-India Rainfall precipitation*[3] respectively. In total, we have $N = 1$ samples with $T = 1752$ time steps, and 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. 5), 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

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

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



(a) (b)

Figure 6: 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.

illustrate the seasonal dependence present in this data. We show the function responses assuming linear and non-linear (softplus networks) transitions in Figures 6a and 6b 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 accross 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. 2). In the nonlinear 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.

C.6. Salsa experiment

The data we consider for this experiment consists on *salsa dancing* sequences from the CMU mocap data[4]. Each trial consists of a sequence with varying length, where the observations consists of the 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 $T = 200$.

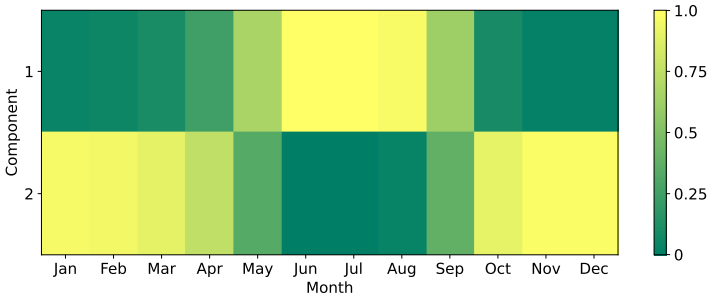


Figure 5: Posterior distribution grouped by month.

[4]Extracted from `http://mocap.cs.cmu.edu/`