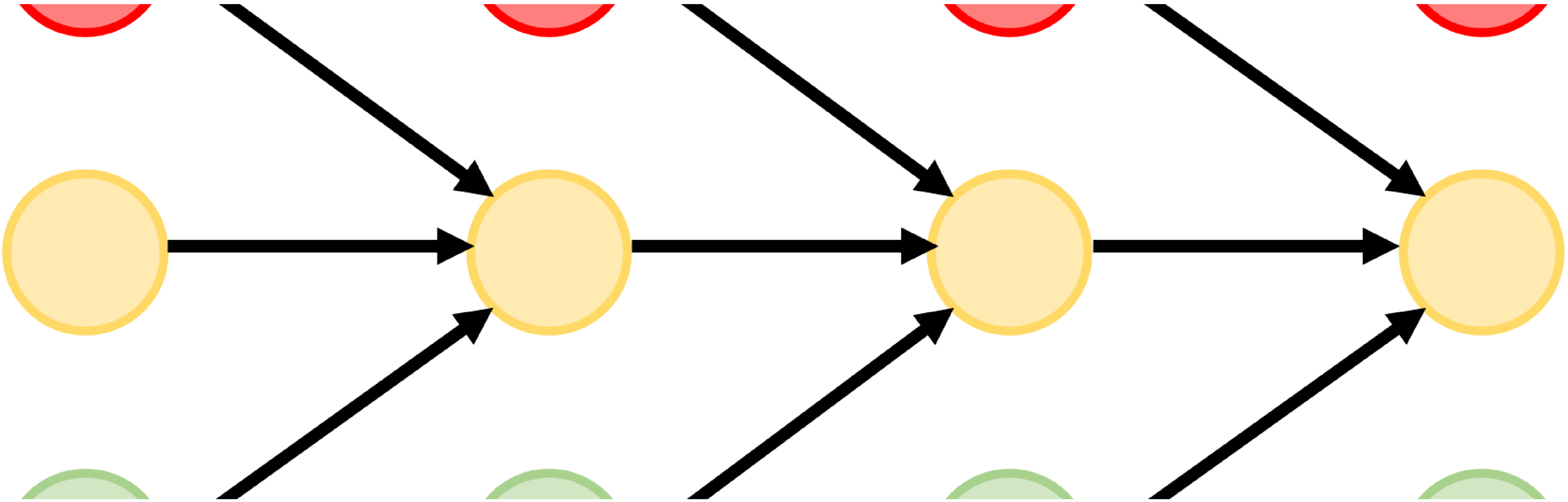


On Identification & Learning of Structured Latent Representations

Yingzhen Li

yingzhen.li@imperial.ac.uk

Starting with an “Ancient” Example…

Disentangle the representation in unsupervised fashion:

- Static information (e.g., content, style)
- Temporal information (e.g., movement)

Note: no attribute labels, learned purely in an unsupervised manner.

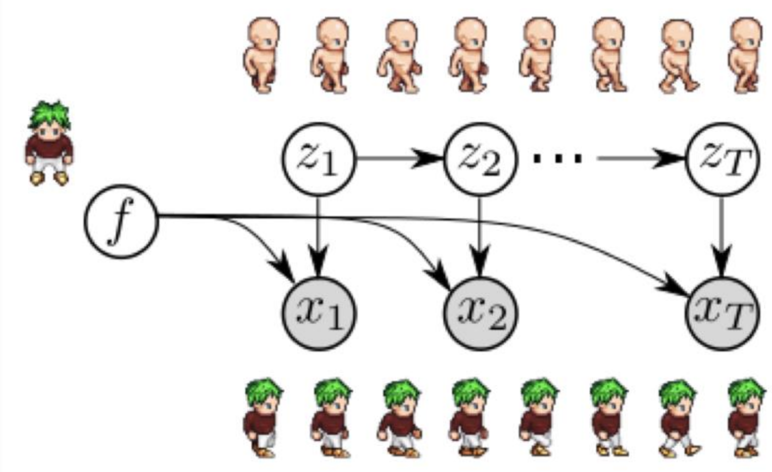
data

Generated (fix content)

Generated (fix dynamics)

Y Li and S Mandt. Disentangled Sequential Autoencoder. ICML 2018.

Disentangled Sequential Autoencoder

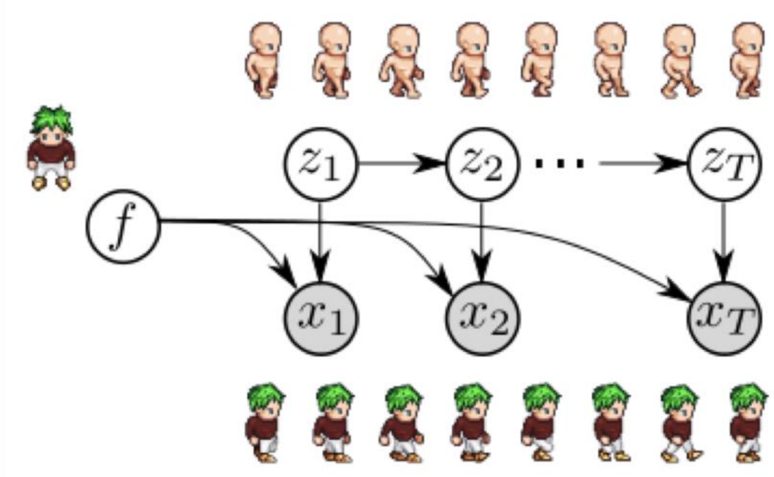


Idea:

- Build a probabilistic graphical model with f = "content" and $z_{1:T}$ = "dynamics"
- Use LSTMs to parameterise $p(z_t|z_{<t})$ and CNNs (+LSTM) to parameterise $p(x_t|f, z_t)$
- Train the model on observational data

Y Li and S Mandt. Disentangled Sequential Autoencoder. ICML 2018.

Powerful Neural Networks Can "Cheat"

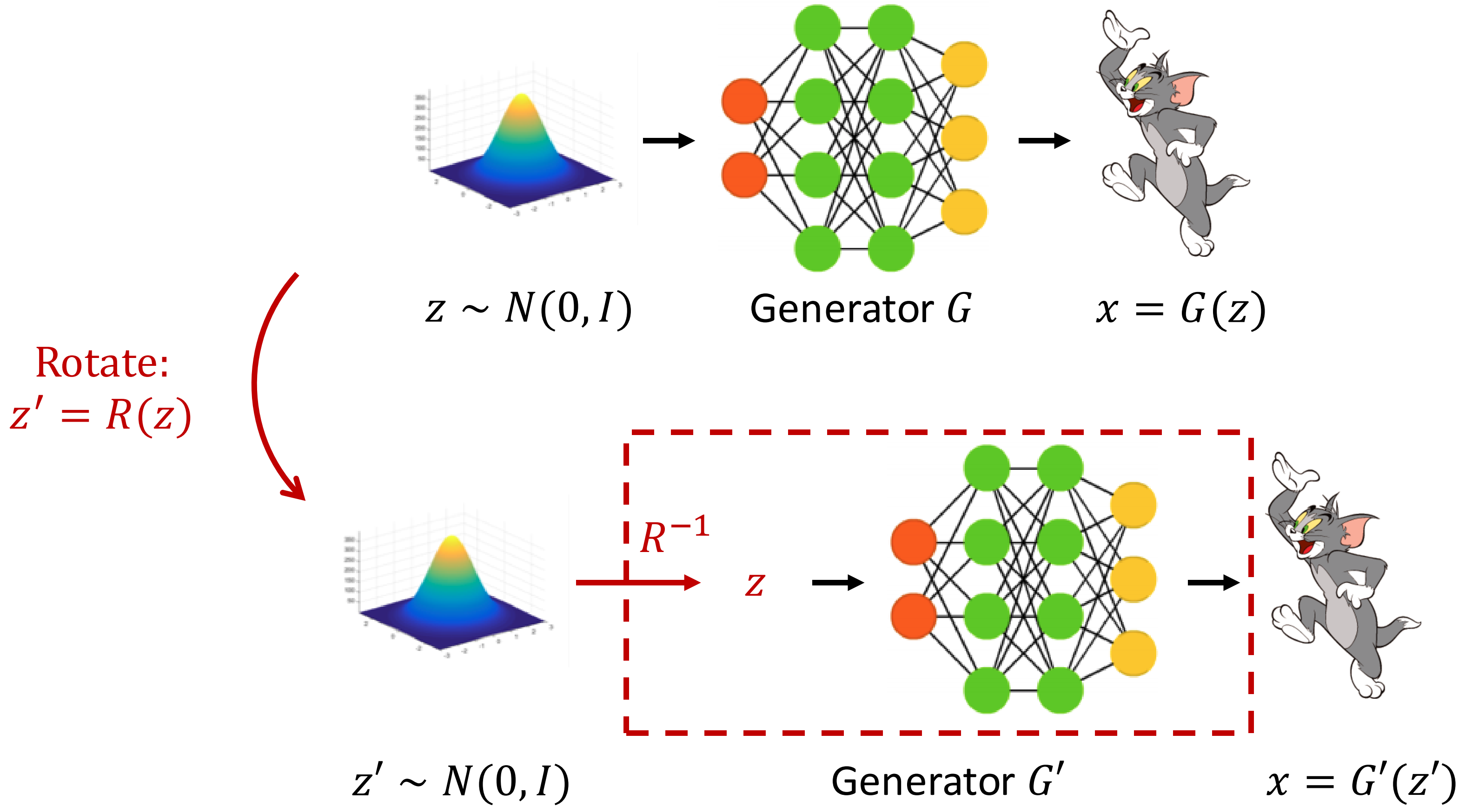


My solution back then:
Graduate student descent

Cheat in the following ways:

- The LSTM hidden cells can learn to "copy" the states
  $\Rightarrow z_t$ captures content info
- The f variable can learn the initial condition for a deterministic dynamical system
  $\Rightarrow f$ captures movement info

Powerful Neural Networks Can “Cheat”



Same results!
(non-identifiable)

Locatello et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. ICML 2019.

Identifiability in Deep Generative Models

Workflow of causal discovery based on identifiable DGMs:

- Write down the model assumptions
  - E.g. $Z = g_\theta(\epsilon_1), X = f_\theta(Z) + \epsilon_2$, f_θ, g_θ can be neural networks
  - This defines a model $p_\theta(X) = \int p_\theta(X|z)p_\theta(z)dz$ with parameters θ
  - Z is unobserved
- Show identifiability
  - i.e. $p_\theta(X) = p_{\theta'}(X) \Leftrightarrow f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$
- Fit the model to data, and do model checking
  - If pass: use the fitted model to answer representation learning questions

Khemakhem et al. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. AISTATS 2020
Kivva et al. Identifiability of deep generative models without auxiliary information. NeurIPS 2022

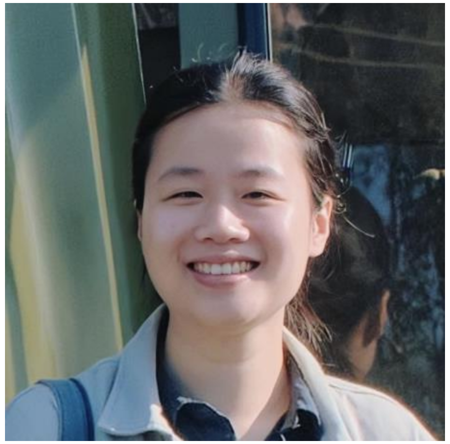
Some Important Notes

- Identifiability Proofs ≠ Learning/Estimation Guarantees
  - Assuming no model error
  - Assuming usage of consistent estimators e.g., MLE
  - Assuming abundant (e.g., infinite) amount of data
  - Assuming global optimum
- Then why should I care about this?
  - Identifiability as a fundamental concept of statistical inference
    - Pre-requisite for analyzing consistency of estimation
    - "Should I trust my discovery results when my deep generative mode fits the data?"
  - Being able to store knowledge ≠ Being able to use knowledge
    - Structured representation makes downstream use of features much easier

On the Identifiability of Switching Dynamic Models

ICML 2024

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



[1]Imperial College London [2]University of Michigan

Motivating Switching Dynamic Models

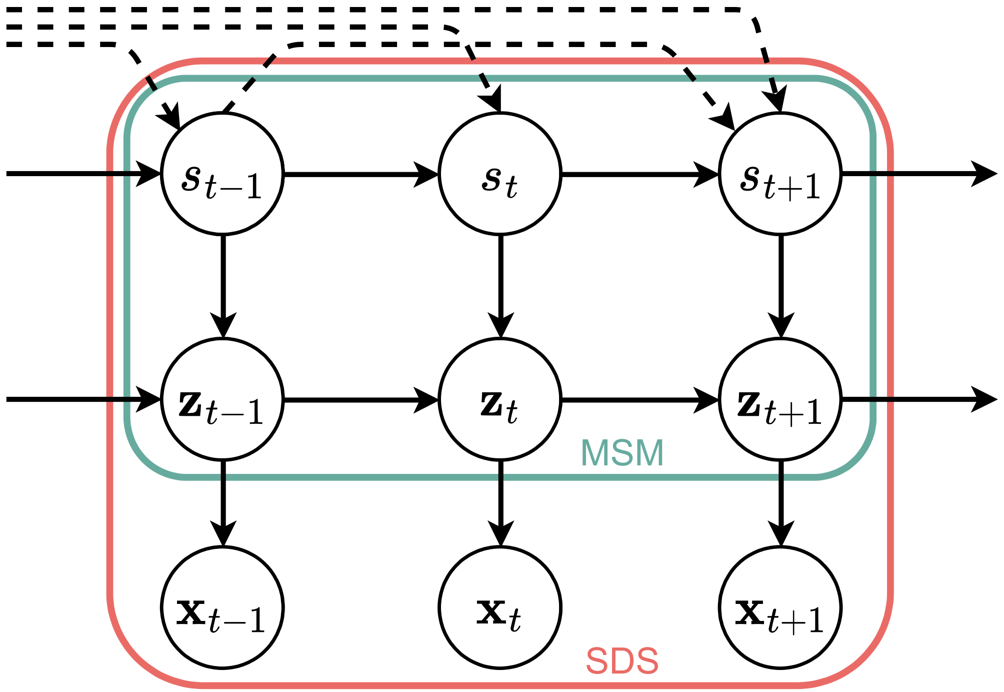
Regime-switching behaviour in time-series data:

“regimes”

- Complex behaviour due to switching between different dynamical patterns
- Within the same regime:
  - The dynamics may be stationary
  - The causal dependencies may be the same across time steps
  - The causal structure may be the same across time steps

Switching Dynamical Systems

- Observations: $\boldsymbol{x}_t \in \mathbb{R}^n$
- Continuous latent variables: $\boldsymbol{z}_t \in \mathbb{R}^m$
- Discrete latent variables: $s_t \in \{1, \ldots, K\}$



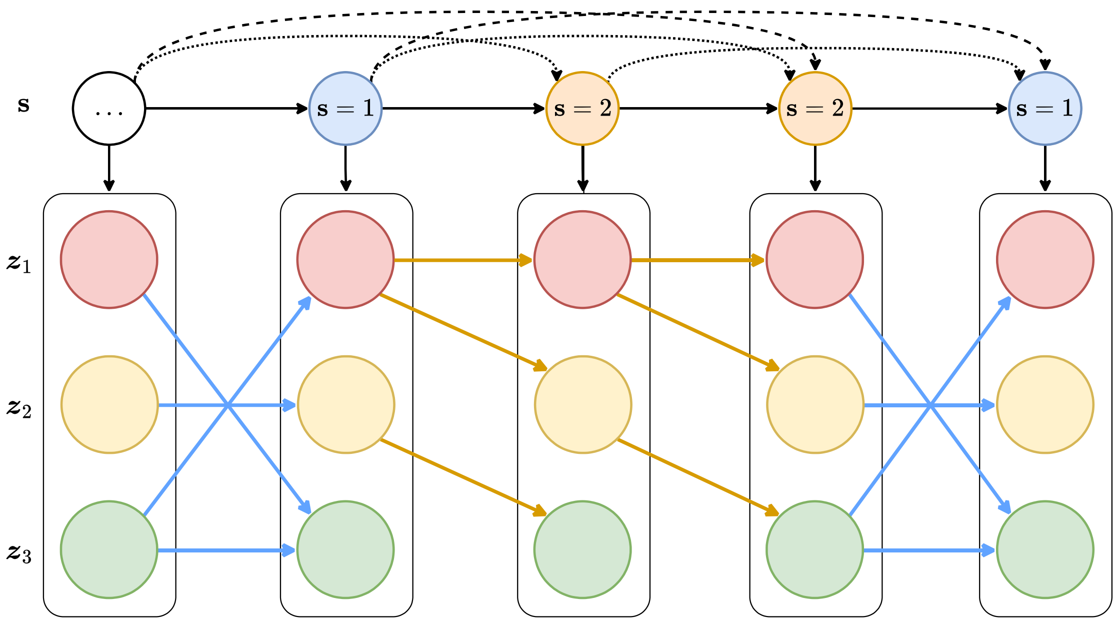
Switching autoregressive prior dynamics

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

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

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

Markov Switching Models



- Discrete latents: $\boldsymbol{s}_t \in \{1, \ldots, K\}$
- Transitions are conditionally stationary and conditional first-order Markov.

$$p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_1, \boldsymbol{s}_t) = p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \boldsymbol{s}_t)$$

Q: Under which **conditions** are the transitions **identifiable** from observations?

Identifiable Markov Switching Models

Theorem 1:

The following conditions render the **Markov Switching Model identifiable**[1] up to permutations:

1. *Unique indexing* for the states

$$\boldsymbol{i} \neq \boldsymbol{j} \iff p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{s}_t = \boldsymbol{i}) \neq p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{s}_t = \boldsymbol{j})$$

2. The transition is Gaussian, and the mean and covariance are analytic in $\boldsymbol{z}_{t-1}$

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t))$$

$\boldsymbol{m}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t)$ and $\boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t)$ can be neural networks with analytic activation functions!

[1] We refer to identifiability by means of the function form rather than parameters.

Identifiable Markov Switching Models

Proof strategy: Frame the problem as a **finite mixture problem** over paths:

$$p(\boldsymbol{z}_{1:T}) = \sum_{i=1}^{K^T} c_i p(\boldsymbol{z}_{1:T} | s_1 = k_1^i, \dots, s_T = k_T^i), \quad c_i = p(\boldsymbol{s}_{1:T} = \{k_1^i, \dots, k_T^i\}), \quad k_t^i \in \{1, \dots, K\}$$

1. Identifiability requires linear independence over the family $\left\{p(\boldsymbol{z}_{1:T}|\boldsymbol{s}_{1:T})\right\}$ [1].
2. Start with $\left\{p(\boldsymbol{z}_{1:2}|\boldsymbol{s}_{1:2})\right\}$, then prove $T > 2$ by induction.
3. Show conditions for $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)$ s.t. $\left\{p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)p(\boldsymbol{z}_{1:t-1}|\boldsymbol{s}_{1:t-1})\right\}$ are linearly independent. (non-parametric case)
4. Work out conditions for the Gaussian case:

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{m}(\boldsymbol{z}_{t-1}, s_t), \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1}, s_t)) \implies \text{analytic in } \boldsymbol{z}_{t-1}.$$

[1] Yakowitz, Sidney J., and John D. Spragins. "On the identifiability of finite mixtures." *The Annals of Mathematical Statistics* 39.1 (1968).

Identifiable Markov Switching Models

Why do we need Gaussian and analytic transition functions?

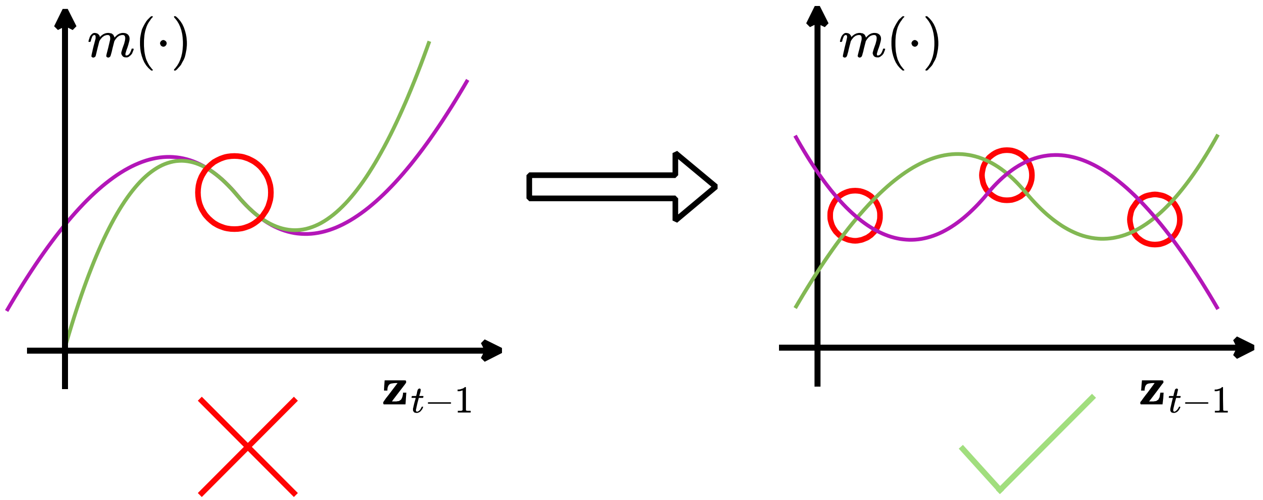
Think about linear independence of $\left\{p(\boldsymbol{z}_{1:t-2}, \boldsymbol{z}_{t-1}|\boldsymbol{s}_{1:t-1})p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, s_t)\right\}$

- Gaussians

$$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2 \iff p(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), p(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \text{ are linearly independent.}$$

- Analytic functions:

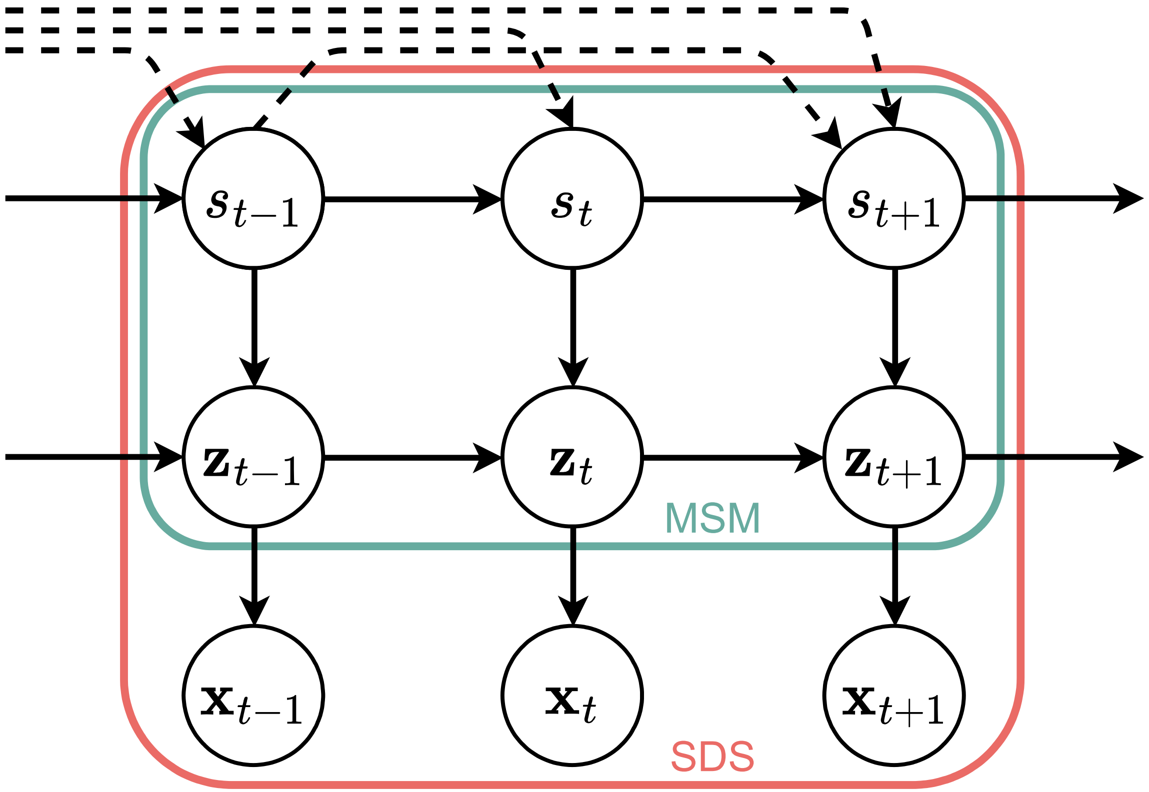
$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d, \mu(\mathcal{X}) \neq 0 \iff \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})$$



Identifiable Switching Dynamical Systems

- Frame the problem as an extension of iMSM and Kivva et al. (2022).

$$\boldsymbol{x}_t = f(\boldsymbol{z}_t) + \boldsymbol{\epsilon}_t, \quad \boldsymbol{z}_{1:T} \sim p(\boldsymbol{z}_{1:T}), \quad \boldsymbol{\epsilon}_t \sim p_{\boldsymbol{\epsilon}}, \quad \boldsymbol{x}_t \in \mathbb{R}^n, \boldsymbol{z}_t \in \mathbb{R}^m \quad n \geq m$$



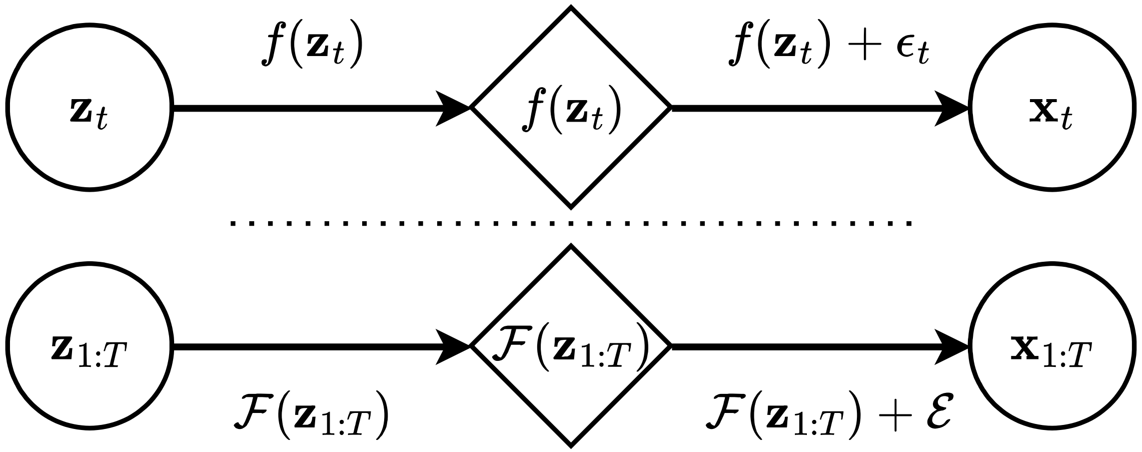
Kivva, Bohdan, et al. "Identifiability of deep generative models without auxiliary information." *NeurIPS* 2022.

Identifiable Switching Dynamical Systems

Theorem 2: Assume an identifiable MSM.

- If f, is piece-wise linear and weakly-injective,
  - MSM is identifiable up to affine transformations.
- If f, is continuous, piece-wise linear, and injective,
  - MSM and f are identifiable up to affine transformations.

Proof sketch:



Kivva et al (2022) Khemakhem et al. (2020)

$$\mathcal{F}(\mathbf{z}_{1:T}) = (f(\mathbf{z}_1), \ldots, f(\mathbf{z}_T))^T$$

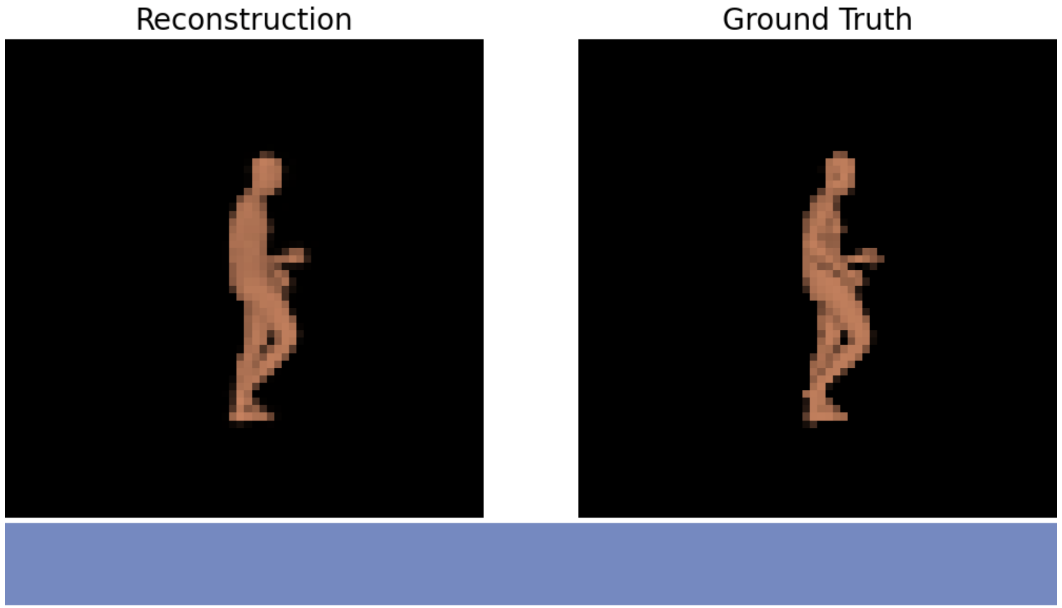
Kivva, Bohdan, et al. "Identifiability of deep generative models without auxiliary information." *NeurIPS* 2022.

Khemakhem, Ilyes, et al. "Variational autoencoders and nonlinear ica: A unifying framework." AISTATS 2020.

Experiments

- High-dimensional video sequences of synthetic meshes from CMU mocap data.

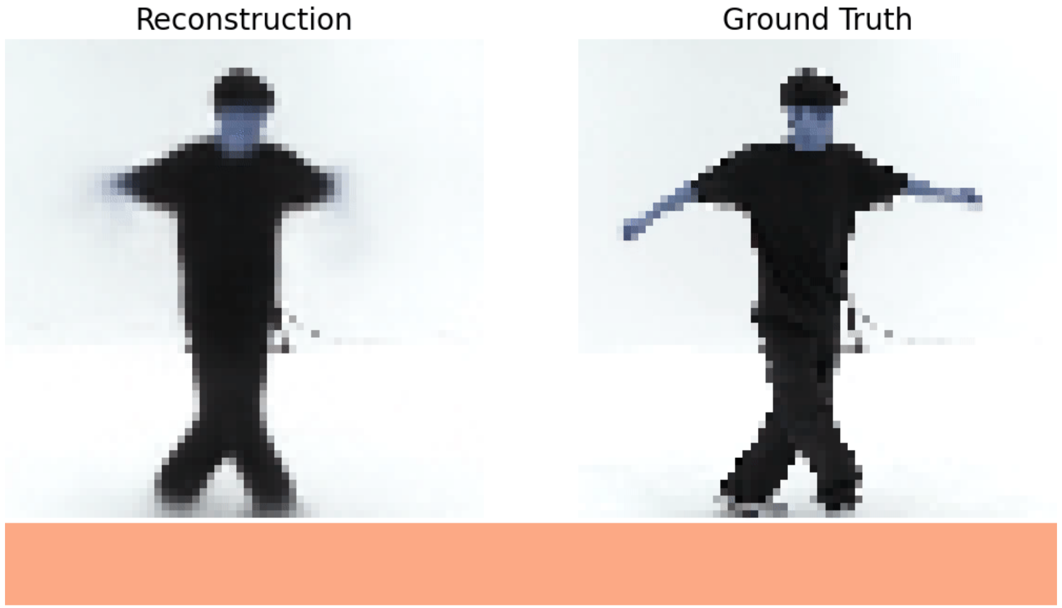
Check paper!



Experiments

- High-dimensional video sequences from the AIST Dance DB – Hip-hop.

Check paper!



Summary

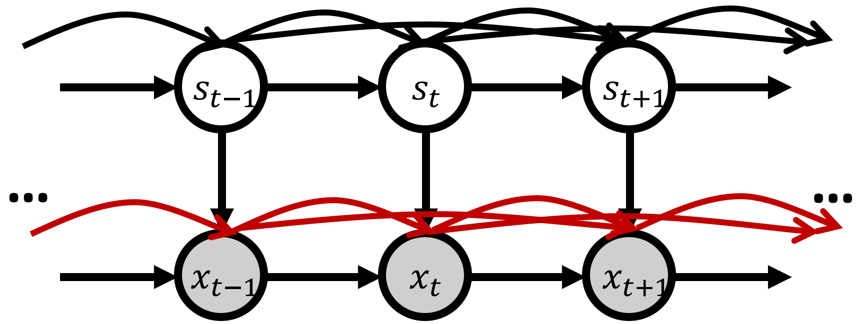
- We establish identifiability for Switching Dynamical Systems.
- Assumptions directly linked to deep generative models.
- Our proof is a major result beyond classical HMM theory.

Check paper!

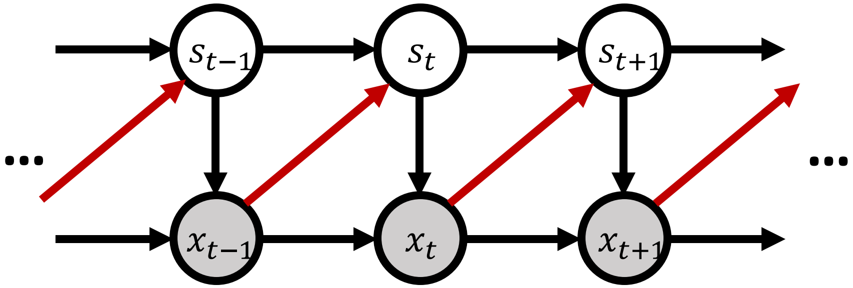
Allman, Elizabeth S, et al. "Identifiability of parameters in latent structure models with many observed variables." (2009)
Gassiat, Élisabeth, et al. "Inference in finite state space nonparametric hidden Markov models and applications." (2016)

Work In-Progress & Future Extensions

- Go for more structured distributions:



higher-order Markov transitions
(e.g., neuroscience & climate data)



observation-dependent state transitions
(critical in e.g., model-based RL)

C Balsells Rodas, Y Wang, P Mediano and **Y Li**. Identifying Nonstationary Causal Structures with High-Order Markov Switching Models. UAI 2024 workshop
C Balsells Rodas, R Tu, **Y Li and** H Kjellstrom. Causal Discovery from Conditionally Stationary Time Series. UAI 2022 Causal Representation Learning workshop

Some Important Notes (Again)

- Identifiability Proofs $\neq$ Learning/Estimation Guarantees
  - Assuming no model error
  - Assuming usage of consistent estimators e.g., MLE
  - Assuming abundant (e.g., infinite) amount of data
  - Assuming global optimum
- Can we say something closer to the practices?

Some Questions That We Can Ask

$Z = g_\theta(\epsilon_1), X = f_\theta(Z) + \epsilon_2, f_\theta, g_\theta$ can be neural networks

- In VAE (i.e., DGMs learned with amortized VI) context:

Estimation consistency

If θ' maximizes ELBO
with data $p_d(x) = p_\theta(x)$,

Then would $f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$?

- No model error
- ~~Maximum likelihood estimation~~
- Infinite amount of data
- Global optimum

Symmetry in learned representations

If θ, θ' both maximize ELBO,

Then would $f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$?

- ~~No model error~~
- ~~Maximum likelihood estimation~~
- Infinite amount of data
- Global optimum

Symmetry in Learned Representations?

$$Z = g_\theta(\epsilon_1), X = f_\theta(Z) + \epsilon_2, \epsilon_2 \sim N(0, \sigma^2 I)$$

- A typical strategy for proving identifiability for DGMs:

  Given: $p_\theta(x) = p_{\theta'}(x)$ and assume invertibility/injectiveness of f_θ & $f_{\theta'}$

  - From noisy observations to noiseless observations
    - $p_\theta(x) = p_{\theta'}(x) \quad \Rightarrow \quad f_\theta\big(g_\theta(\epsilon_1)\big)$ and $f_{\theta'}\Big(g_{\theta'}(\epsilon_1)\Big)$ are equally distributed
  - Removing the "volume" term (with further assumptions)
    - $\log p_\theta\left(z = f_\theta^{-1}(x)\right) + \cancel{\log\left|\frac{d f_\theta^{-1}(x)}{dx}\right|} = \log p_{\theta'}\left(z = f_{\theta'}^{-1}(x)\right) + \cancel{\log\left|\frac{d f_{\theta'}^{-1}(x)}{dx}\right|}$
  - Analyze the equivalence class $f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$

  e.g., $\Phi(f_\theta^{-1}(x)) = A\Phi(f_{\theta'}^{-1}(x)) + b$ if g_θ induces an ExpFam distribution

Symmetry in Learned Representations?

- Let's look at the ELBO:

$$ELBO(x,\theta,\phi) \coloneqq \underbrace{E_{q_\phi(z|x)}[\log p_\theta(x|z)]} + \underbrace{H[q_\phi(z|x)]} + \underbrace{E_{q_\phi(z|x)}[\log p_\theta(z)]}$$

$= E_{q_\phi(z|x)}\left[\frac{1}{\sigma^2}\|x - f_\theta(z)\|_2^2\right] + C(\sigma)$
(reconstruction error)

Entropy
(volume)

Stats. for inferred
latent representations
e.g., using ExpFam:
$= E_{q_\phi(z|x)}\left[\langle \lambda_0, \Phi(z)\rangle\right] - Z(\lambda)$

- Differences?
  - Encoder: $f_\theta^{-1}(x)$ vs $q_\phi(z|x)$
  - Additional reconstruction error term

Symmetry in Learned Representations?

$$Z = g_\theta(\epsilon_1), X = f_\theta(Z) + \epsilon_2, \epsilon_2 \sim N(0, \sigma^2 I)$$

- Possible strategy for analyzing symmetry in learned representations:

Given: $ELBO(x, \theta, \phi) = ELBO(x, \theta', \phi')$ are optimal solutions,

- Assume invertibility/injectiveness of f_θ & $f_{\theta'}$,

- Assume one of the following scenarios:
    - Near deterministic regime: $\sigma \to 0$
    - Optimally learned output noise variance: optimizing and share σ
    - Auxiliary info available for prior only: use $p(z|u)$ instead of $p(z)$
    - (each scenario needs different assumptions on $q_\phi(z|x)$)

$$E_{q_\phi(z|x)}[\log p_\theta(z)] = E_{q_{\phi'}(z|x)}[\log p_{\theta'}(z)]$$

Symmetry in Learned Representations?

$$E_{q_\phi(z|x)}[\log p_\theta(z)] = E_{q_{\phi'}(z|x)}[\log p_{\theta'}(z)]$$

- Why looking into this term?
  - We use q to extract latent structure in practice!
  - Example: exponential family prior and "conjugate" approximate posterior:

$$p_\theta(z) = \exp[\langle\lambda(\theta), \Phi(z)\rangle] - Z\big(\lambda(\theta)\big)$$
$$q_\phi(z|x) = \exp[\langle\lambda(\phi, x), \Phi(z)\rangle] - Z\big(\lambda(\phi, x)\big)$$

$$\Rightarrow \; E_{q_\phi(z|x)}[\Phi(z)] = AE_{q_{\phi'}(z|x)}[\Phi(z)] + b$$

$$\Rightarrow \; E_{q_{\phi^*}(z|x)}[\Phi(z)] = E_{p_\theta^*(z)}[\Phi(z)]$$

Take Away

- Identifiability: a fundamental question of structural representation learning
  - i.e. should you expect the learned representations to match the ”structures” in data
  - The field has quite substantial advances since ~2020
  - Our work provides strong theoretical results for (deep) switching dynamical systems
- We need to bring identifiability theory closer to practice
  - Option 1: Getting closer to maximum likelihood
  - Option 2: Accept bias/errors in current tech, and analyze them
- Challenge 😈: Theory for DDPMs & auto-regressive LLMs?
  - No learning for probabilistic representations
  - Studying symmetries within neural networks?

THANK YOU!

Questions? Ask now, or email:
yingzhen.li@imperial.ac.uk

PS: hiring post-docs in "AI for Chemistry"

https://arxiv.org/abs/2305.15925

Thanks to my awesome collaborators:

Carles Balsells-Rodas

Yixin Wang

Pedro Mediano

Ruibo Tu

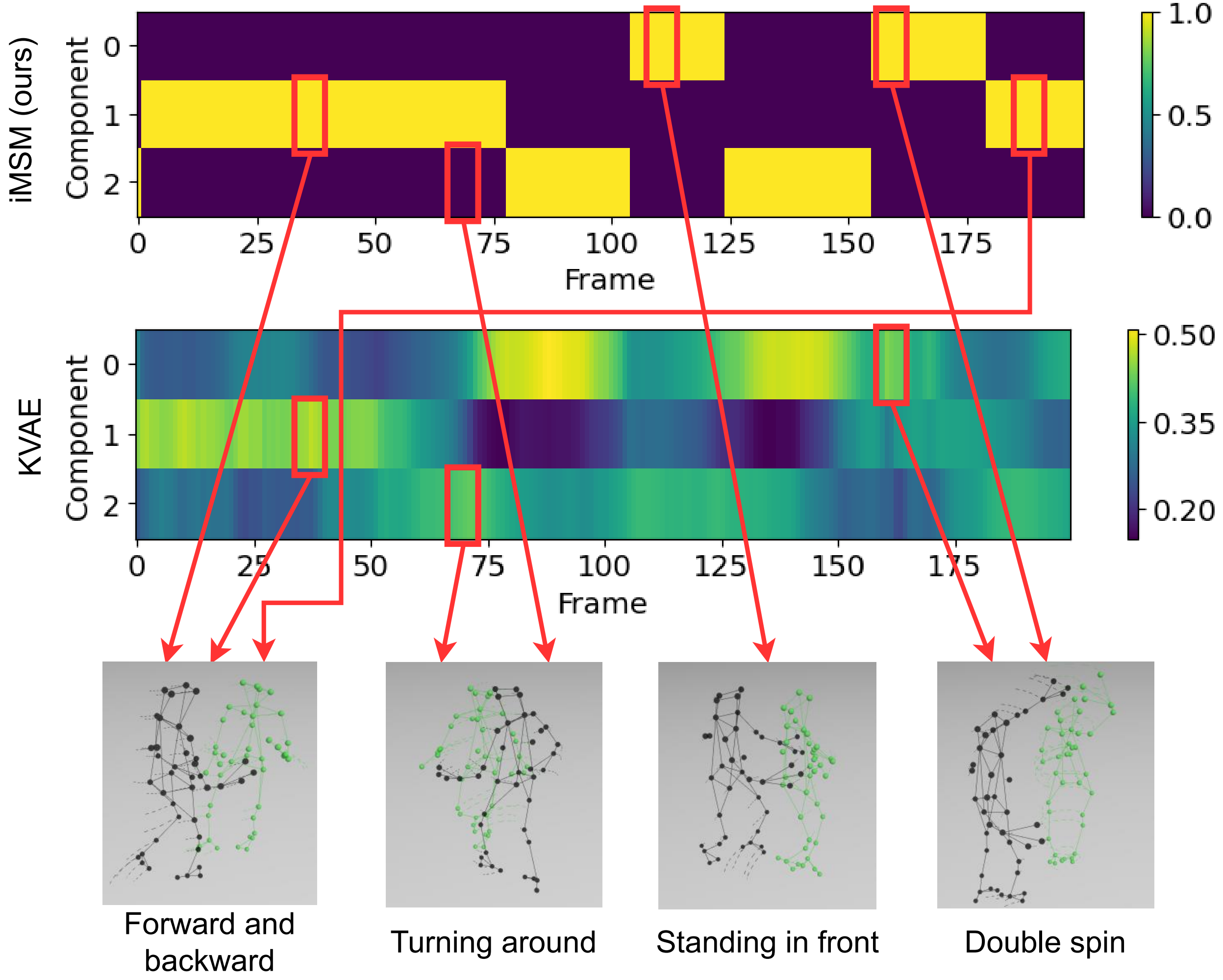
Hedvig Kjellström

Stephan Mandt

Additional Slides

Identifiability in Markov Switching Models

- Experiment: discovering dancing patterns
  - Data: CMU mocap
  - Estimation: Generalised EM
  - DL Baseline: KalmanVAE

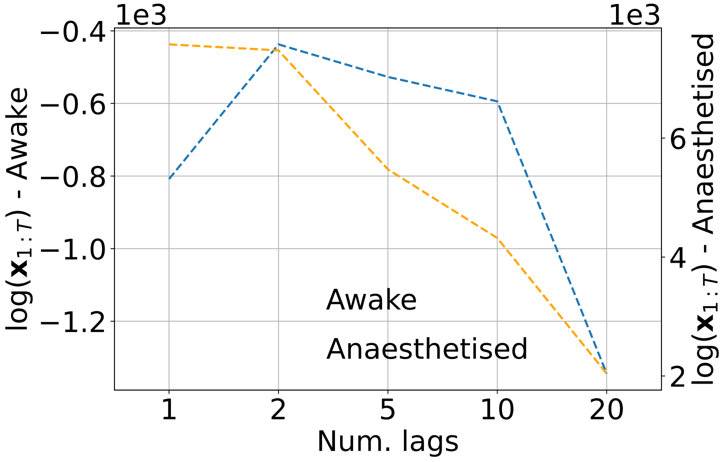


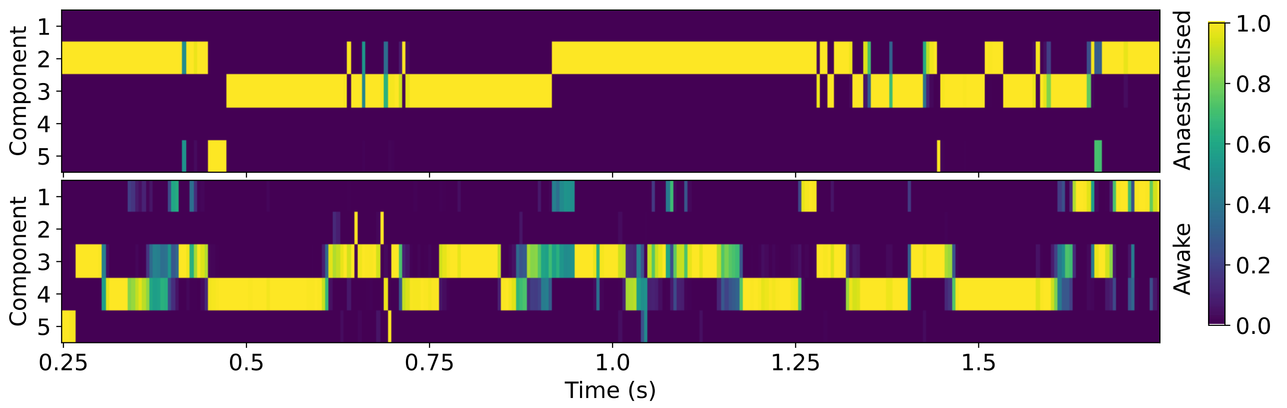
Fraccaro et al. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. NeurIPS 2017
C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Switching Dynamic Models. ICML 2024

Higher-Order Switching Dynamics

Neuro Activity Data Analysis:

- Recorded from Monkeys in (a) normal awake, and (b) induced anaesthetized status
- Idea: understand neuro activity by segmenting recorded signal into "regimes"

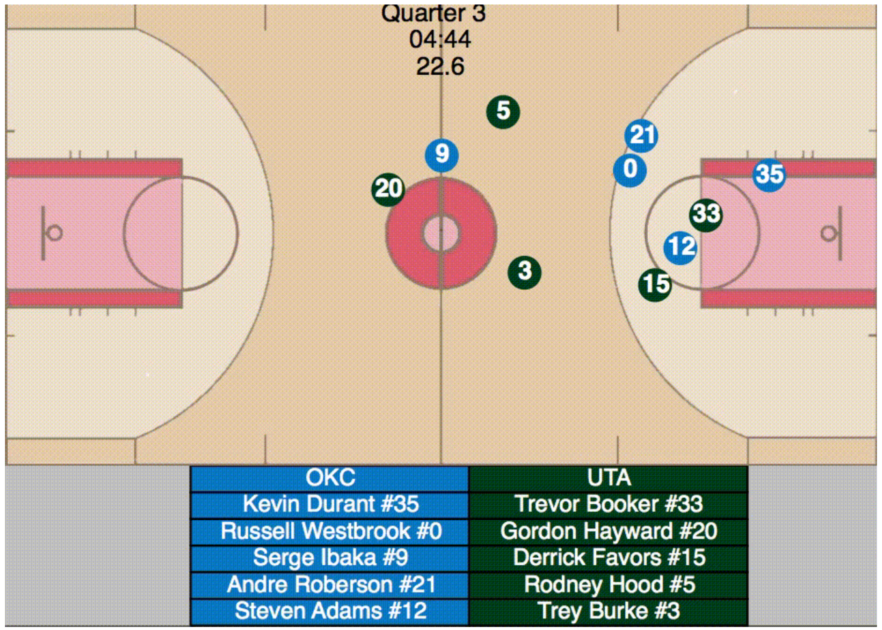




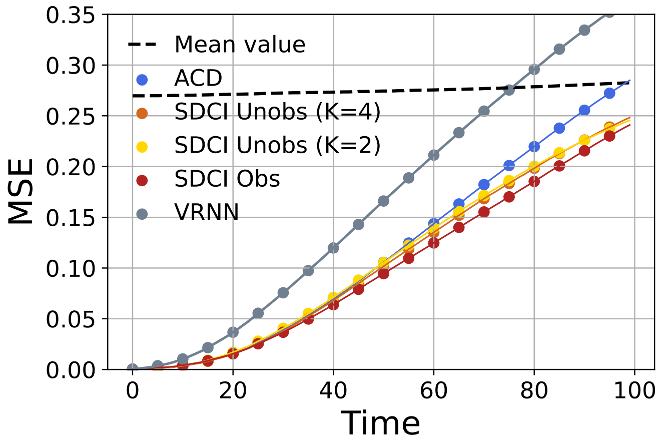
State-Dependent Causal Inference (SDCI)

Causal discovery & sequence modelling for non-stationary time series:

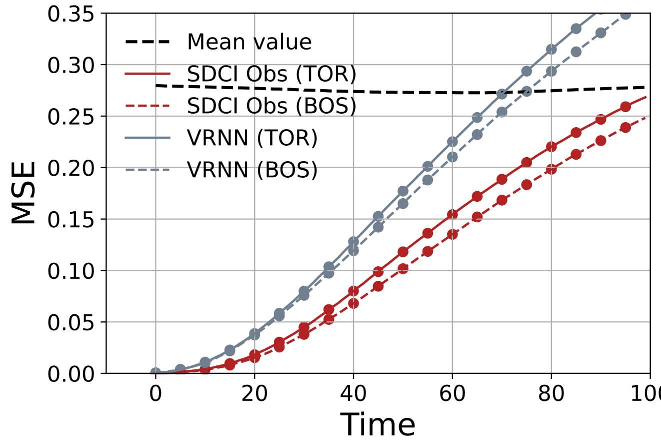
Dataset: NBA player trajectories
- multi-agent
- non-stationary



Forecasting error:



Train on full data



Train on Boston Celtics only

Learned hidden state visualisation:

C Balsells Rodas, R Tu, **Y Li and** H Kjellstrom. Causal Discovery from Conditionally Stationary Time Series. UAI 2022 Causal Representation Learning Workshop