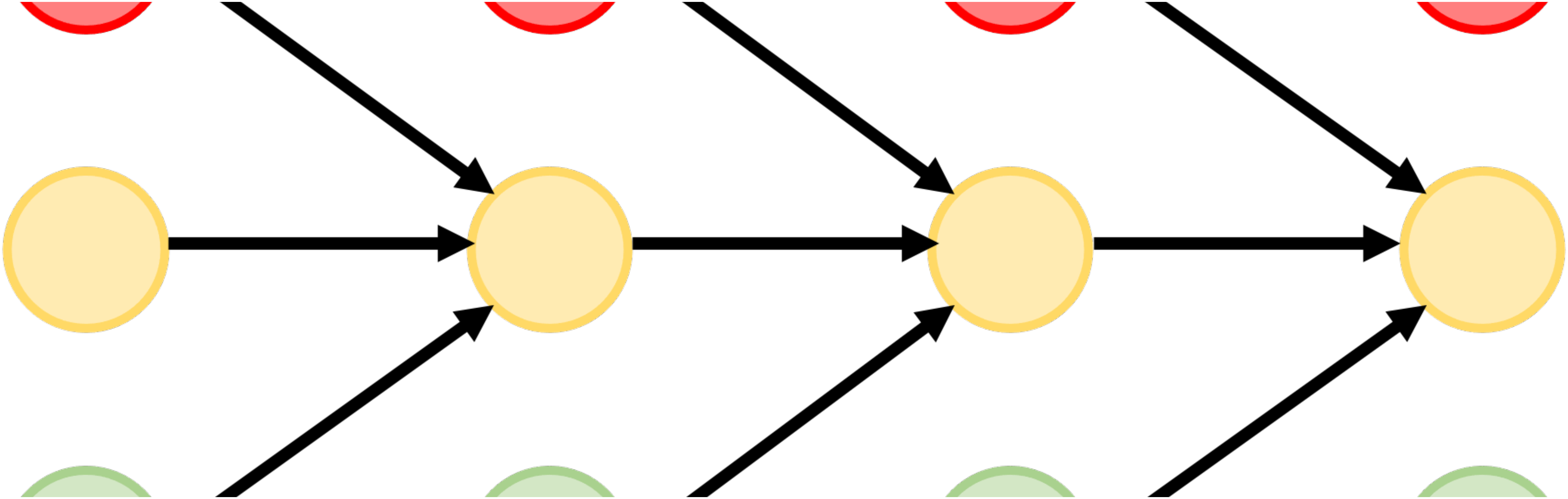


Towards Causal Deep Generative Models for Sequential Data

Yingzhen Li

yingzhen.li@imperial.ac.uk

Causality
“Theorist”

Deep Learning
“Alchemist”

Ali Rahimi. NIPS 2017 Test of Time Award presentation
Yann LeCun. My take on Ali Rahimi’s “Test of Time” award talk at NIPS 2017

Controllable Video Generation

Disentangle the representation in unsupervised fashion:

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

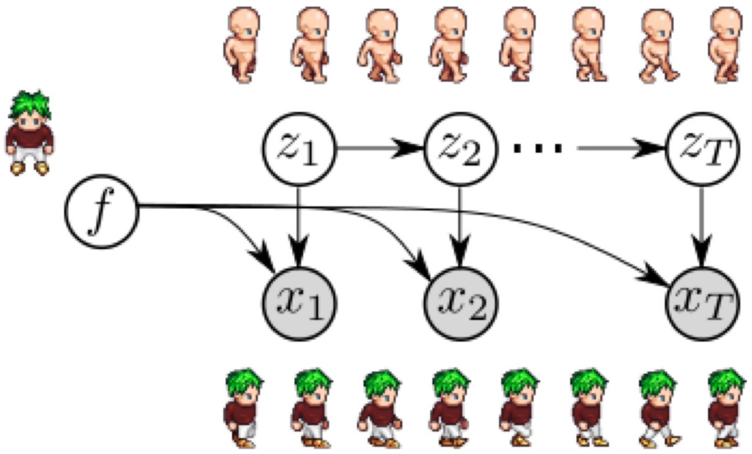
data

Generated (fix content)

Generated (fix dynamics)

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

Disentangled Sequential Autoencoder

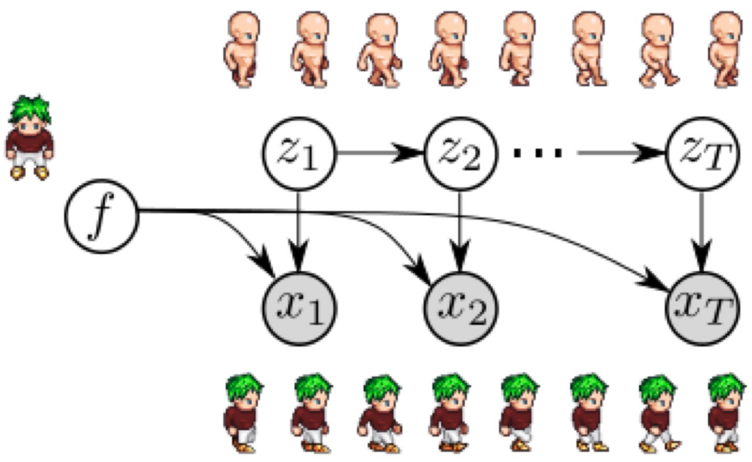


Idea:

- 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"



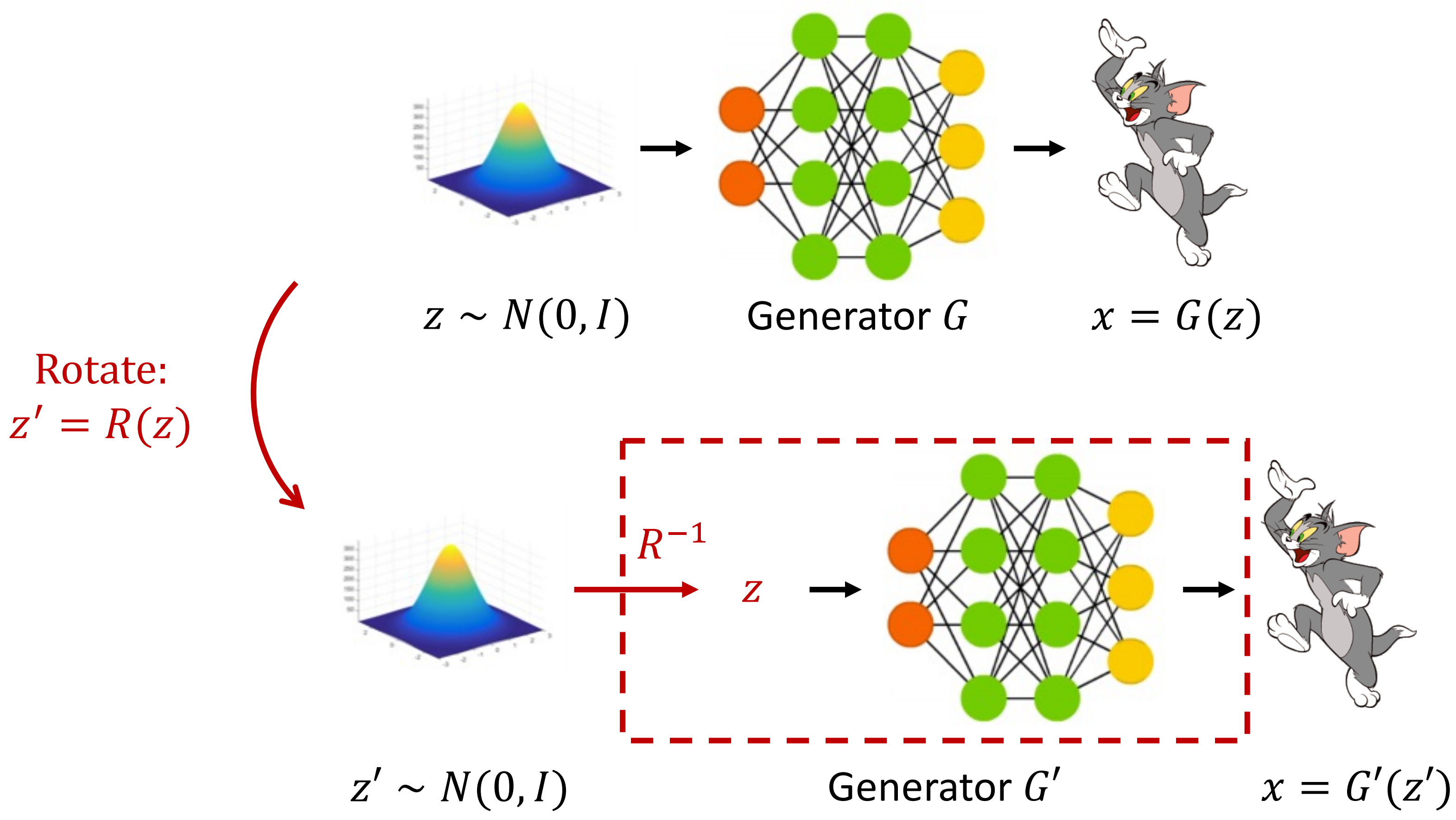
My solution back then:
Alchemy

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

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

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 Statistical/Causal Models

Workflow of causal discovery based on functional causal models:

- Write down the SCM/SEM
  - E.g. $Y = f_\theta(X) + \epsilon$
  - This defines a model $p_\theta(Y|X)$ with parameters θ
- Show identifiability
  - i.e. $p_\theta(Y|X) = p_{\theta'}(Y|X) \Leftrightarrow \theta \cong \theta'$
  - Identifiability enables causal discovery & counterfactual reasoning
- Fit the model defined by SCM to data, and do model checking
  - If pass: use the fitted model to answer causal questions

Glymour et al. Review of Causal Discovery Methods Based on Graphical Models. Sec. Statistical Genetics and Methodology, Vol. 10, 2019

Identifiability in Deep Generative Models

Workflow of causal discovery based on identifiable DGMs:

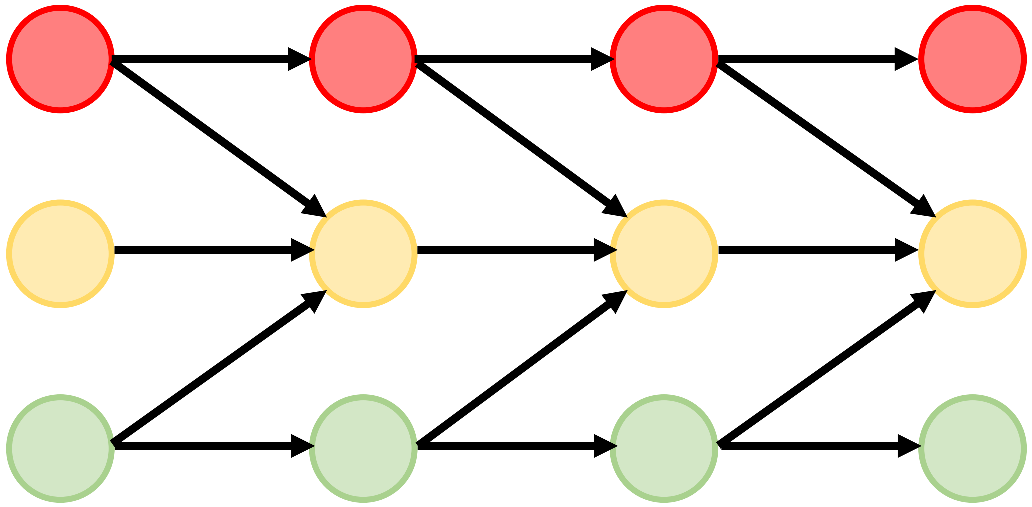
- Write down the SCM/SEM
  - E.g. $Z = g_\theta(\epsilon_1), X = f_\theta(Z) + \epsilon_2$, f_θ, g_θ can be neural networks
  - This defines a model $p_\theta(X) = \int p_\theta(X|z)p_\theta(z)dz$ with parameters θ
  - Z is unobserved
- Show identifiability
  - i.e. $p_\theta(X) = p_{\theta'}(X) \Leftrightarrow f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$
  - Identifiability enables causal discovery & counterfactual reasoning
- Fit the model defined by SCM to data, and do model checking
  - If pass: use the fitted model to answer causal 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

Causal Discovery in Time-Series

Use the information of time: “the cause happens prior to its effect”

- Granger causality, TiMINo, etc.:
  - Assume all the variables are observed
  - In most cases assume stationarity



Peters et al. Causal inference on time series using restricted structural equation models. NIPS 2013
Tank et al. Neural Granger Causality. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 08, pp. 4267-4279, 2022.

State-Dependent Causal Inference (SDCI)

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

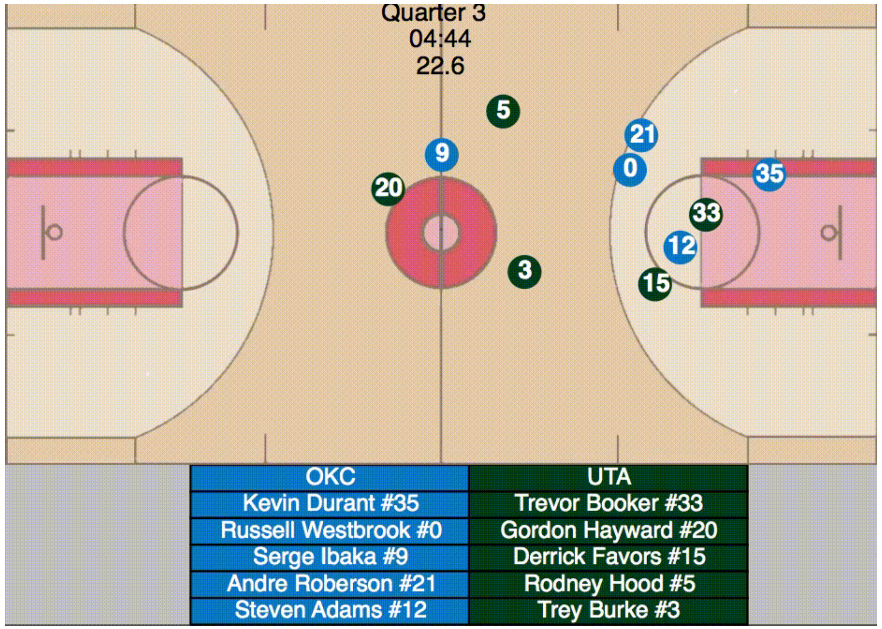
- Imagine having N agents interacting:
  - Each agent i at time step t has both its observation x_i^t and its internal discrete state s_i^t
  - Depending on the state s_i^t, x_i^t will have different functional relationship with x_j^{t+1}
- Conditional summary graph:
  - Compact summary of the causal relationship
  - When the states are all fixed to the same: reduced back to summary graph

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

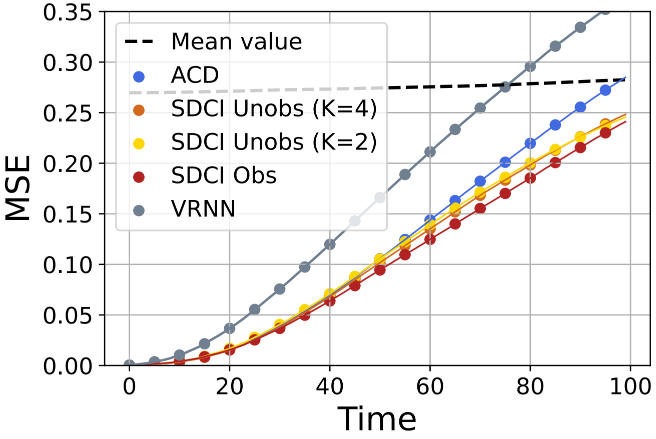
State-Dependent Causal Inference (SDCI)

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

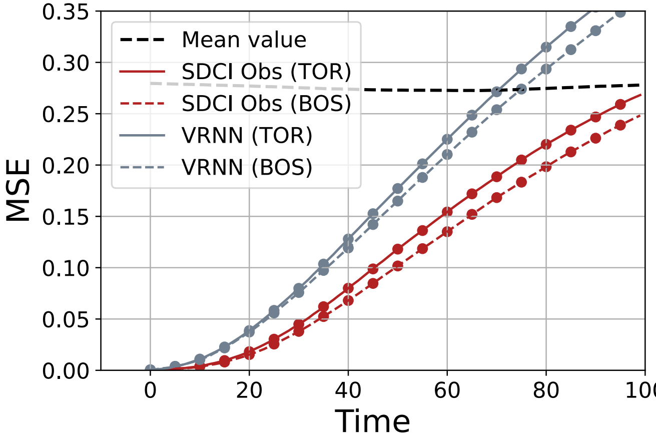
Dataset: NBA player trajectories
- multi-agent
- non-stationary



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

State-Dependent Causal Inference (SDCI)

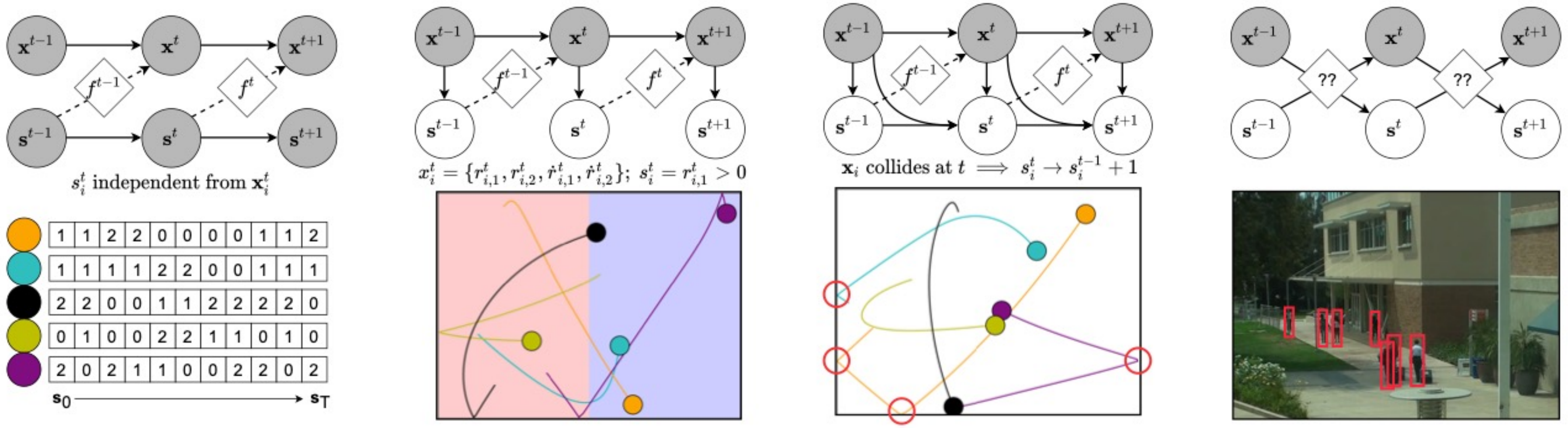
Identifiability result for SDCI (informal):

The conditional summary graph is identifiable if the states are observed.

(not realistic)

Can we do better?

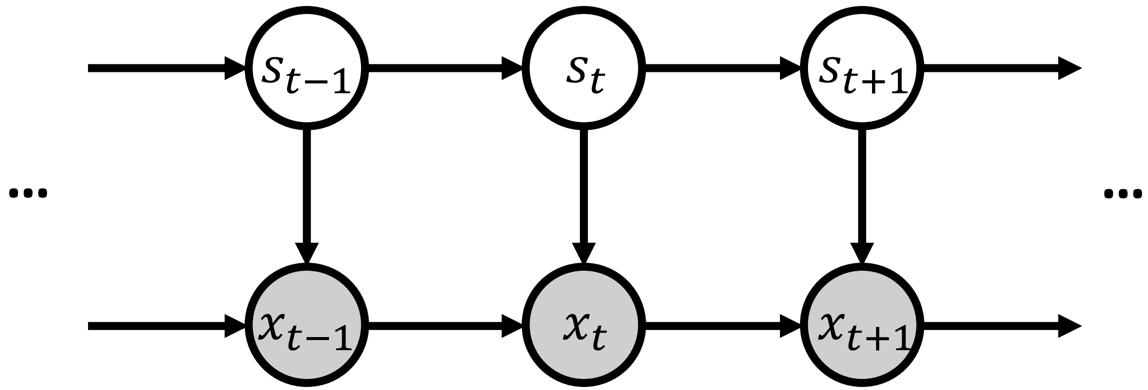
Yes, but need assumptions on how the observations and states interact



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

Identifiability in Switching Dynamic Models

Markov Switching Models (first-order):



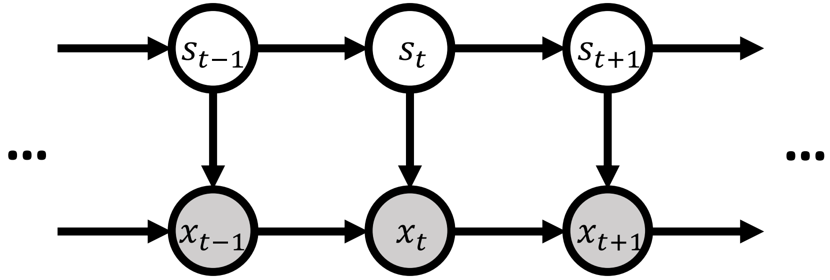
- Discrete and finite state-space: $s_t \in \{1, \dots, K\}$
- Conditional first-order Markov model: $p(x_t | x_{<t}, s_t) = p(x_t | x_{t-1}, s_t)$
  (assuming $x_0 = \emptyset$)

When does this model identifiable with observations of $x_{1:T}$ only?

C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models

Identifiability result (informal):



The first-order Markov Switching Model is identifiable up to state permutation when:

- *Unique indexing for the states (i.e., no repeating states):*

$$i \neq j \;\Leftrightarrow\; p(x_t|x_{t-1}, s_t = i) \neq p(x_t|x_{t-1}, s_t = j)$$

- *In Gaussian case, the mean and covariance functions are analytic in x_{t-1}:*

$$p(x_t|x_{t-1}, s_t) = N(x_t; \underline{m(x_{t-1}, s_t), S(x_{t-1}, s_t)})$$

Can use neural networks with smooth activation functions!
(here identifiability means identifying the functions)

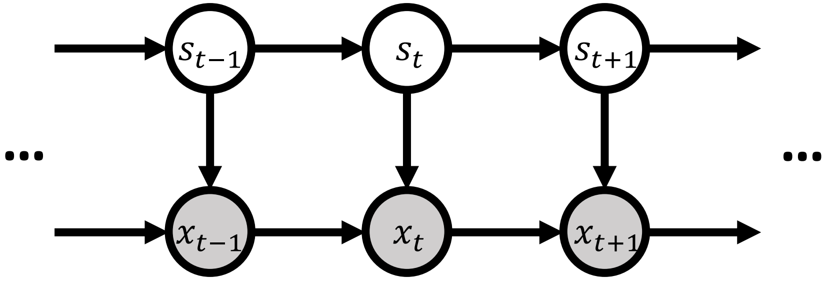
C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models

Proof sketch (informal):

Think about it as a finite mixture model over paths:

$$p(x_{1:T}) = \sum_{s_{1:T} \in \{1,\dots,K\}^T} p(x_{1:T}|s_{1:T})p(s_{1:T})$$



(1) Identifiability for finite mixture model requires linear independence of family $\{p(x_{1:T}|s_{1:T})\}$

(2) Notice the first-order Markov structure: $p(x_{1:T}|s_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{t-1}, s_t)$

⇒ Show linear independence of $p(x_{1:2}|s_{1:2})$, then prove for $T \geq 3$ case by induction

(3) Work out conditions on $p(x_t|x_{t-1}, s_t)$ to make $\{p(x_t|x_{t-1}, s_t)\, p(x_{t+1}|x_t, s_{t+1})\}$ linearly independent

⇒ Obtain certain linear independence & continuity conditions in non-parametric case

(4) In Gaussian case: work out the conditions on the mean & covariance to satisfy conditions in (3)

$$p(x_t|x_{t-1}, s_t) = N(x_t; \underline{m(x_{t-1}, s_t), S(x_{t-1}, s_t)})$$

⇒ Analytic in x_{t-1}

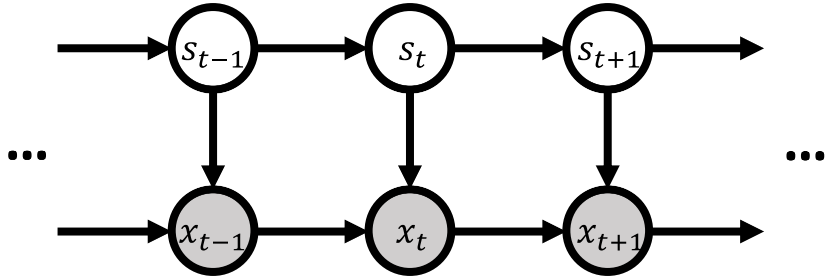
C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models

Proof sketch (informal):

Think about it as a finite mixture model over paths:

$$p(x_{1:T}) = \sum_{s_{1:T} \in \{1,\dots,K\}^T} p(x_{1:T} | s_{1:T}) p(s_{1:T})$$



- What is nice about Gaussians:

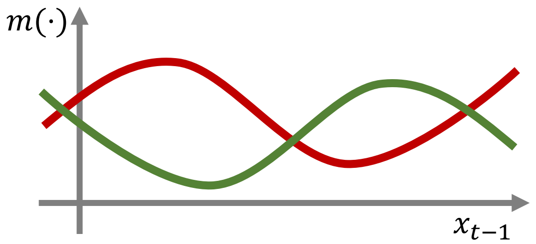
$p_{\mu_1,\Sigma_1}(x) = p_{\mu_2,\Sigma_2}(x)$ for $x \in X \subset R^d$ $\Leftrightarrow$ $\mu_1 = \mu_2, \Sigma_1 = \Sigma_2$

(non-zero measure subset)

- What is nice about analytic functions:

$f_1(x) = f_2(x)$ for $x \in X \subset R^d$ $\Leftrightarrow$ $f_1(\cdot) = f_2(\cdot)$

(non-zero measure subset)

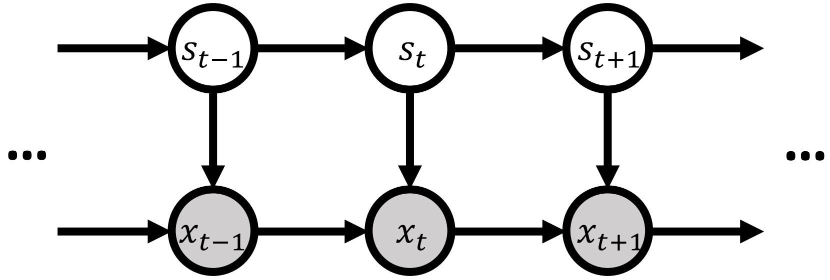


$N(x_t; m_1(x_{t-1}, s_t), S_1(x_{t-1}, s_t)) = N(x_t; m_2(x_{t-1}, s_t), S_2(x_{t-1}, s_t))$ for some (x_{t-1}, x_t) in some non-zero measure set

$\Leftrightarrow$ $m_1(\cdot, s_t) = m_2(\cdot, s_t), S_1(\cdot, s_t) = S_2(\cdot, s_t)$ (when the functions are analytic in x_{t-1})

C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models



Some simulation results:
(Estimation with stochastic EM)

Simulation settings:

- Stationary hidden state transitions (first order)
- Conditional transition ground-truth:

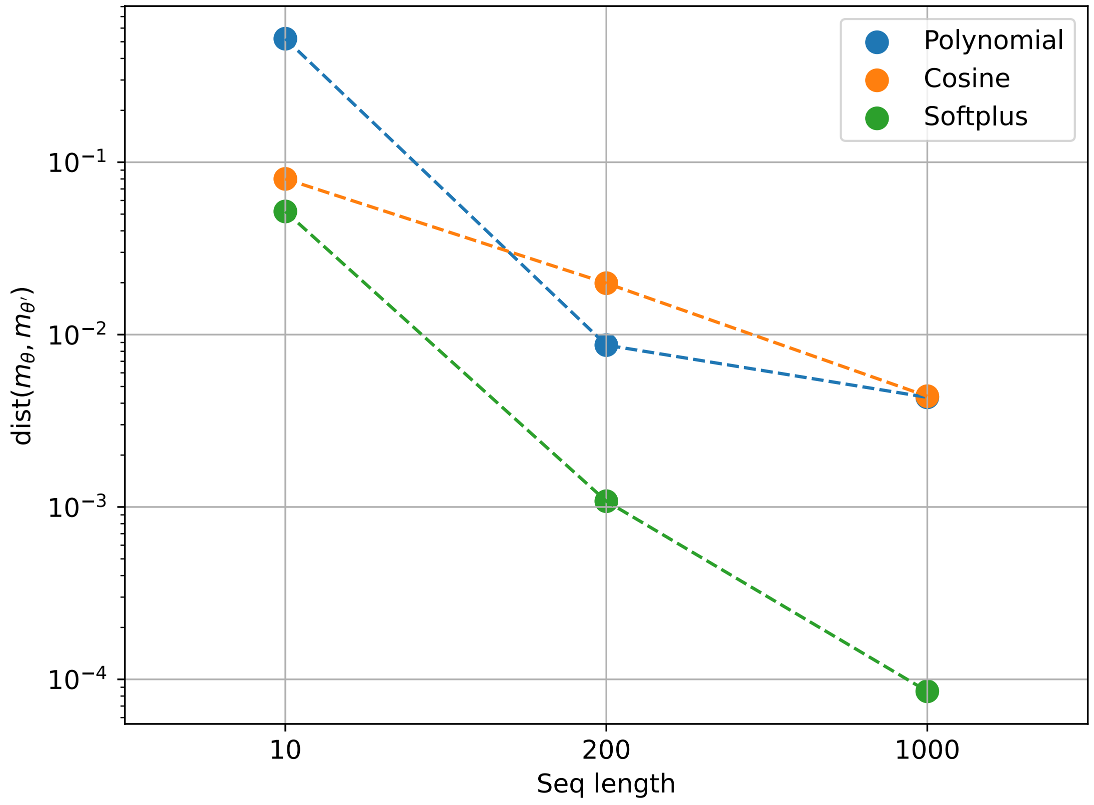
$$p(x_t|x_{t-1}, s_t) = N(x_t; m(x_{t-1}, s_t), \sigma^2 I)$$

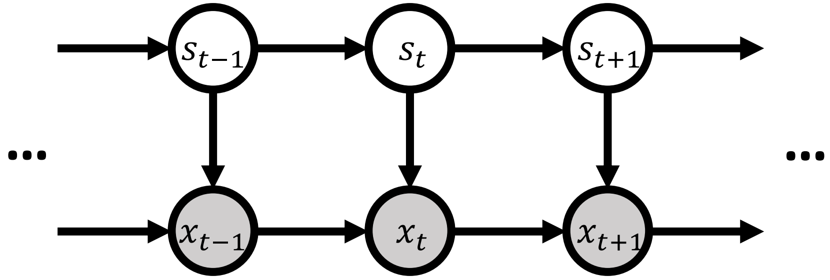
- Three types of ground-truth m function:
  1. Polynomial (cubic function)
  2. Randomly initialised neural network with cosine activations
  3. Randomly initialised neural network with softplus activations

C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models

Some simulation results:
(Estimation with stochastic EM)





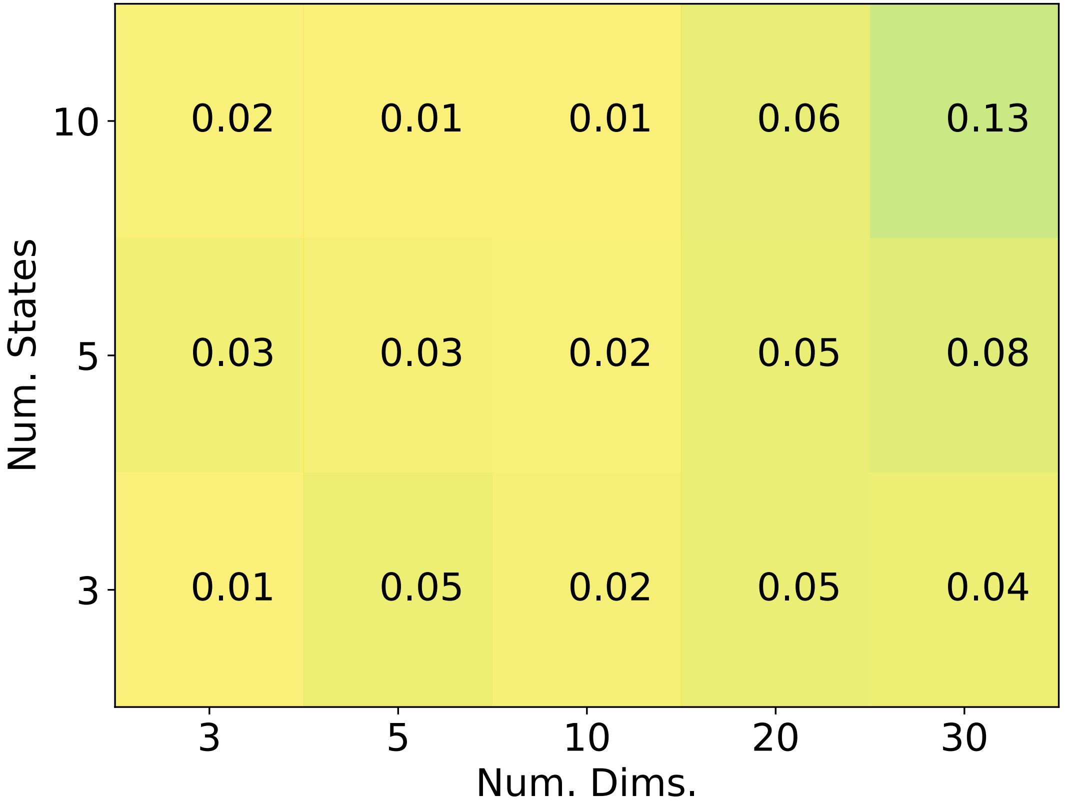
Error metric:

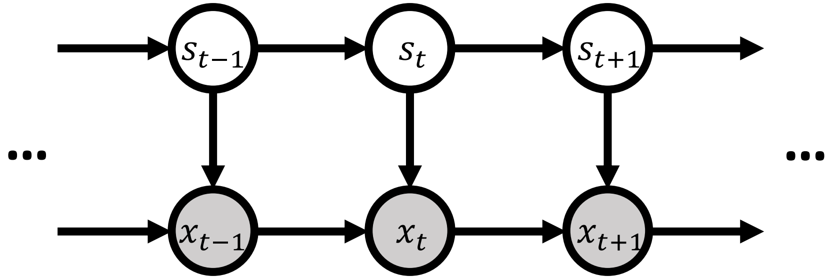
- ℓ_2 distance between ground-truth and estimated functions (after state-matching & average over states)

C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Identifiability in Switching Dynamic Models

Some simulation results:
(Estimation with stochastic EM)



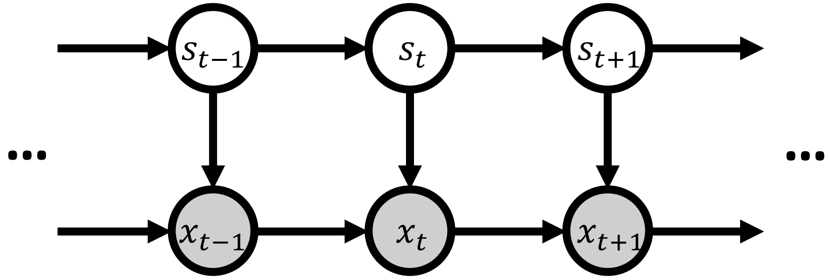


Scalability of the estimation method:

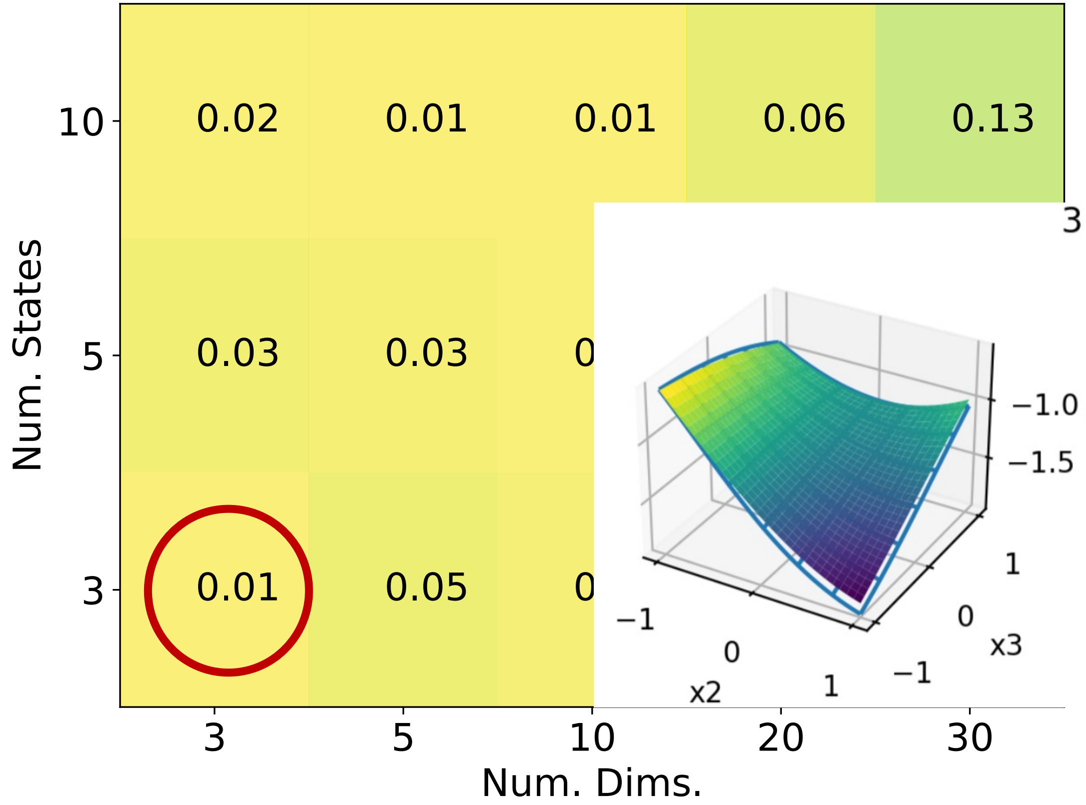
- Locally connected network assumption: on avg. 3 variables interact

C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

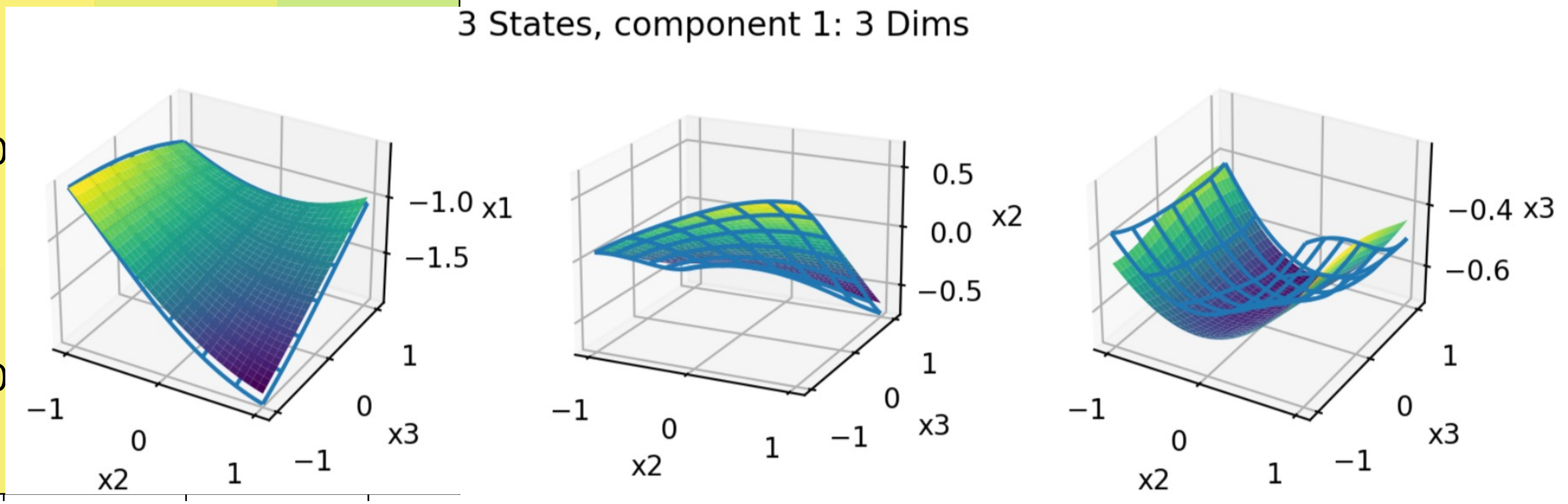
Identifiability in Switching Dynamic Models



Some simulation results:
(Estimation with stochastic EM)



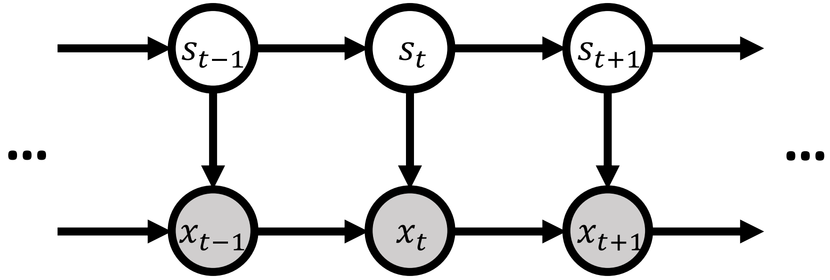
Scalability of the estimation method:

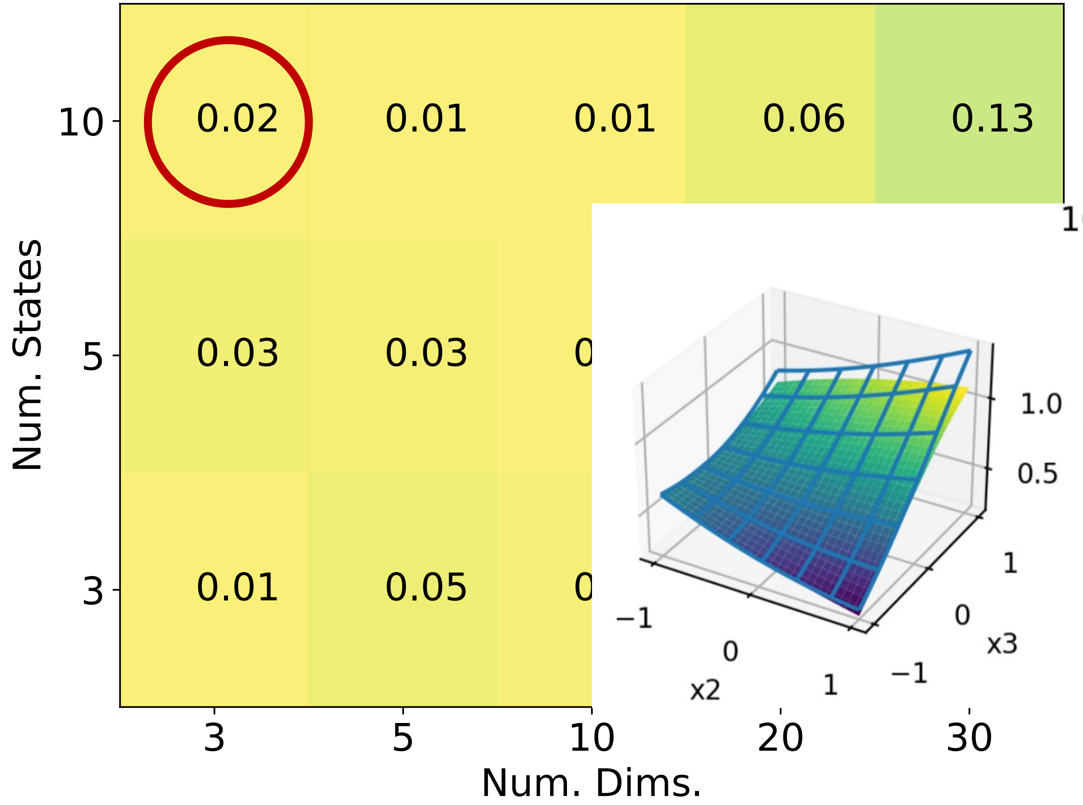


C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

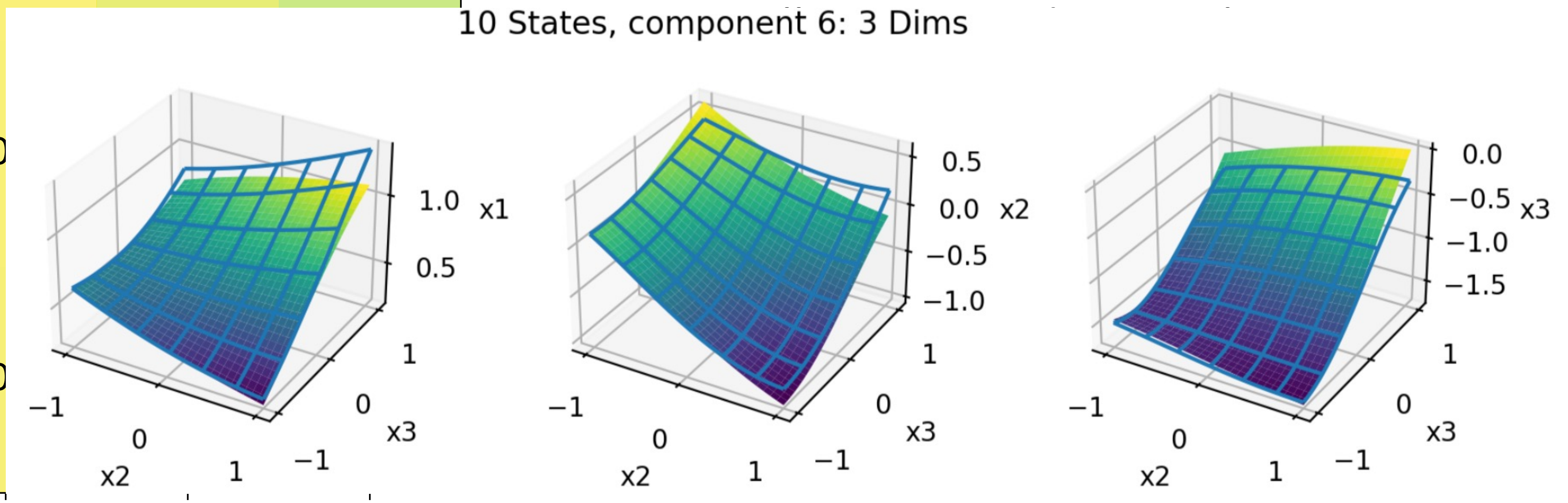
Identifiability in Switching Dynamic Models

Some simulation results:
(Estimation with stochastic EM)



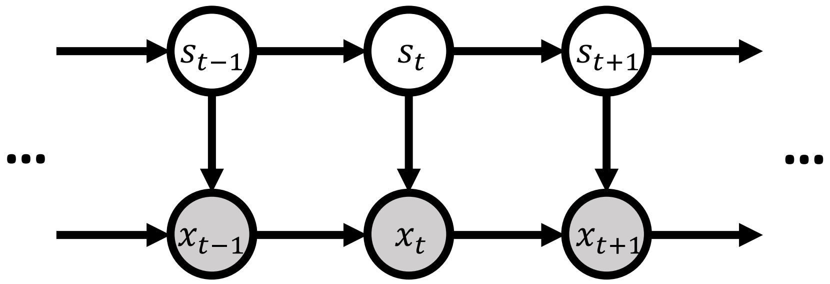


Scalability of the estimation method:



C Balsells Rodas, Y Wang and **Y Li.** On the identifiability of Markov Switching Models. In preparation

Some Discussions



On the proof strategy and indications:

- Cannot use the proof strategy of HMM identifiability results
  - Simply because the dynamic is not fully controlled by latent state transitions
- The proof makes NO assumption on $p(s_{1:T})$ and can identify the joint $p(s_{1:T})$
  - Works for ANY dynamic model for the states $s_{1:T}$
  - The marginal $p(x_{1:T})$ can thus be non-stationary and higher-order Markov
  - Direct extension to global regime settings by making $s_1 = s_2 = \cdots = s_T$
- Easily extendable to include observed "control signals" $u_{1:T}$:

$$p(x_{1:T}, s_{1:T} | u_{1:T}) = p(x_{1:T} | s_{1:T}) p(s_{1:T} | u_{1:T})$$

Gassiat et al. Inference in finite state space non parametric hidden Markov models and applications. Stat Comput 26, 61–71, 2016
Allman et al. Identifiability of parameters in latent structure models with many observed variables. Ann. Stat. 37, 3099–3132, 2009

Some Discussions

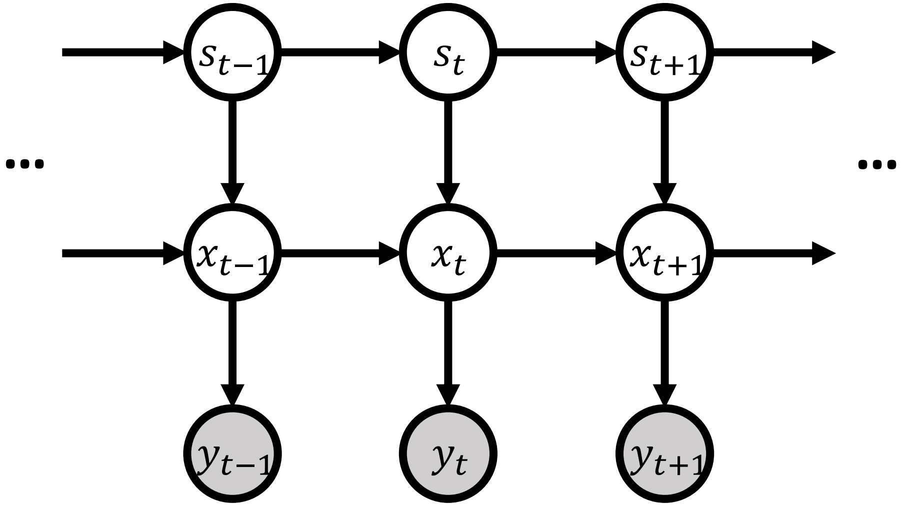
Future extensions:

- Go for higher-order Markov conditional transitions (with time lag $M > 1$):

$$p(x_t | x_{<t}, s_t) = p(x_t | x_{t-M:t-1}, s_t)$$

  - Better assumptions for e.g., neuron activity data, energy & climate time-series

- Lift the continuous states $x_{1:T}$ to latent space:
  - More realistic for video & other high-dimensional data
  - Potential application in model-based RL



- Beyond time series?

Fraccaro et al. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. NeurIPS 2017
Hafner et al. Mastering Atari with Discrete World Models. ICLR 2021

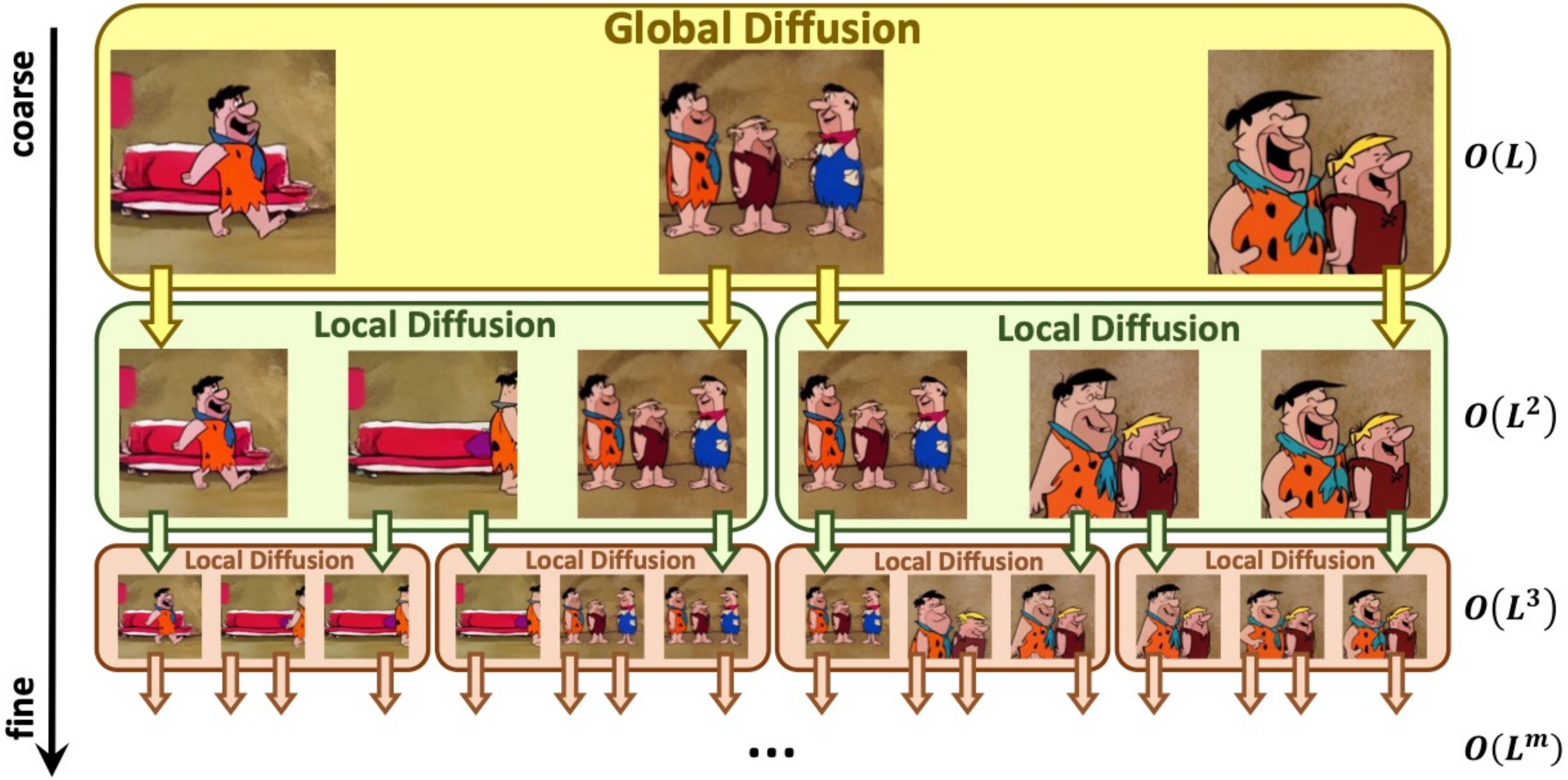
Identifiability in Deep Generative Models

Workflow of causal discovery based on identifiable DGMs:

- Write down the SCM/SEM
  - 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. $p_\theta(X) = p_{\theta'}(X) \Leftrightarrow f_\theta \cong f_{\theta'}, g_\theta \cong g_{\theta'}$
  - Identifiability enables causal discovery & counterfactual reasoning
- Fit the model defined by SCM to data, and do model checking
  - If pass: use the fitted model to answer causal 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

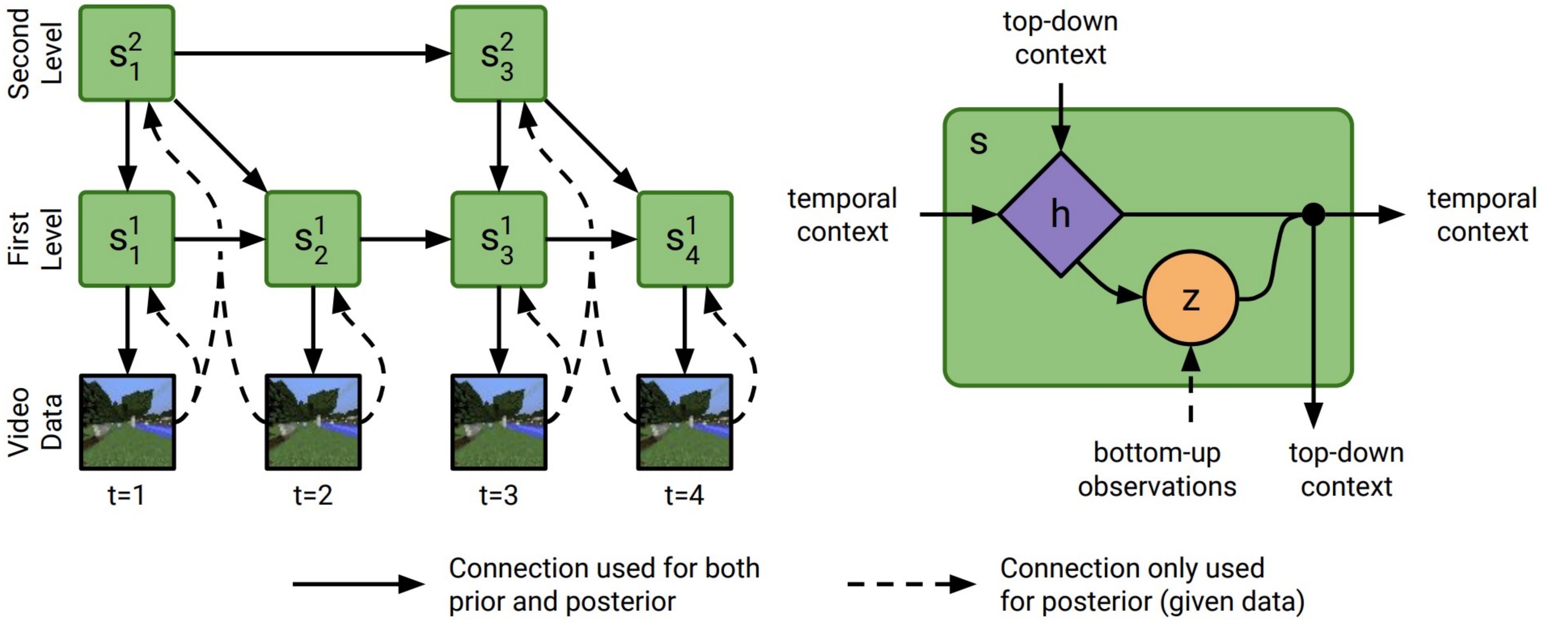
SOTA Video Generation Models are "Non-Causal"



- "Non-causal": future observations to help on generating past observations

Yin et al. NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation. arXiv:2303.12346

SOTA Video Generation Models are "Non-Causal"



- "Non-causal": Identifiability in hierarchical DGMs very difficult

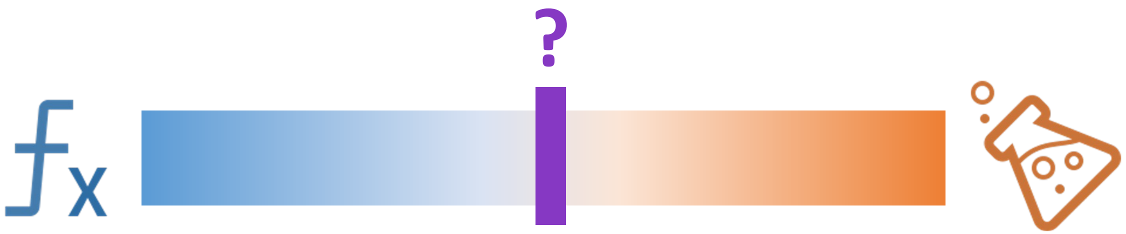
Hafner et al. Clockwork Variational Autoencoders. NeurIPS 2021

End-to-End Causal DGMs: Ever Possible?

My personal opinions:

- Leave low-level representation learning to perception models
  - Deep Learning methods provide impressive results now
  - Can leverage multi-modality data (which usually don't share the same SCM)
- Identifiable DGMs on perception representations
  - Much easier than handling "raw pixels" directly
  - Take benefits from multi-modality perception models

"Scientific Alchemy": figure out the theoretical limits, leave the rest to perception



THANK YOU!

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

Thanks to my awesome collaborators:


Stephan Mandt


Carles Balsells-Rodas


Ruibo Tu


Hedvig Kjellström


Yixin Wang