# Causal Discovery from Conditionally Stationary Time Series

Carles Balsells Rodas

Imperial College London

joint work with Ruibo Tu, Hedvig Kjellström and Yingzhen Li

April 21, 2023

# Table of Contents

# Motivation

# Motivation

# Motivation


Common deep learning approaches struggle performing complex tasks that are **intuitive for humans** (e.g. action recognition) [Wang and Gupta, 2018].

# Motivation


Common deep learning approaches struggle performing complex tasks that are **intuitive for humans** (e.g. action recognition) [Wang and Gupta, 2018].

To approach machines to human cognition, [Lake et al., 2017] suggest the following learning outcomes:

# Motivation


Common deep learning approaches struggle performing complex tasks that are **intuitive for humans** (e.g. action recognition) [Wang and Gupta, 2018].

To approach machines to human cognition, [Lake et al., 2017] suggest the following learning outcomes:

1. Harness **compositionality** in data.

# Motivation


Common deep learning approaches struggle performing complex tasks that are **intuitive for humans** (e.g. action recognition) [Wang and Gupta, 2018].

To approach machines to human cognition, [Lake et al., 2017] suggest the following learning outcomes:

1. Harness **compositionality** in data.
2. Build **causal models** of the world.

# Motivation


Common deep learning approaches struggle performing complex tasks that are **intuitive for humans** (e.g. action recognition) [Wang and Gupta, 2018].

To approach machines to human cognition, [Lake et al., 2017] suggest the following learning outcomes:

1. Harness **compositionality** in data.
2. Build **causal models** of the world.
3. Ground learning in intuitive **physics** and **human behaviour**.

# Example: Causal discovery in videos

- **Objective:** Causal discovery in videos (in the wild)

# Example: Causal discovery in videos

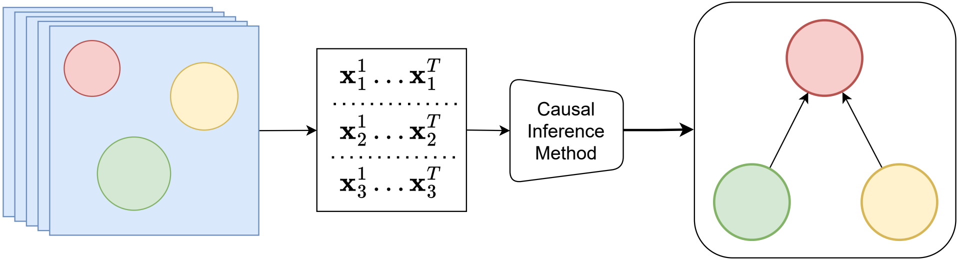
- **Objective:** Causal discovery in videos (in the wild)
- **Solution:** Interpreting the scene as a **composition of N time series**

# Example: Causal discovery in videos

- **Objective:** Causal discovery in videos (in the wild)
- **Solution:** Interpreting the scene as a **composition of N time series**
  1. Unsupervised feature extraction (region proposals, keypoints, etc)

# Example: Causal discovery in videos

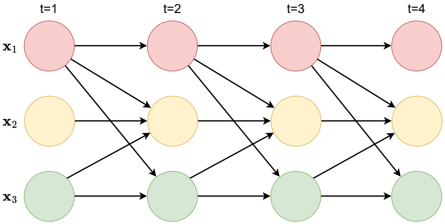
- **Objective:** Causal discovery in videos (in the wild)
- **Solution:** Interpreting the scene as a **composition of N time series**
  1. Unsupervised feature extraction (region proposals, keypoints, etc)
  2. Causal inference across samples



# Causal discovery in conditionally stationary time series

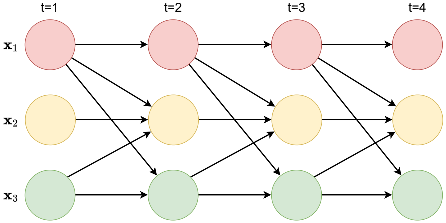
# Causal discovery in time series

- Mainly based on stationary time series
  - Non-temporal identifiability $\rightarrow$ Temporal setting
  - Granger causality (no instantaneous effects)
  - Amortised causal discovery (ACD)



# Causal discovery in time series

- Mainly based on stationary time series
  - Non-temporal identifiability $\rightarrow$ Temporal setting
  - Granger causality (no instantaneous effects)
  - Amortised causal discovery (ACD)

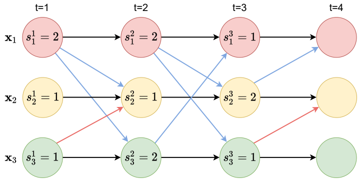


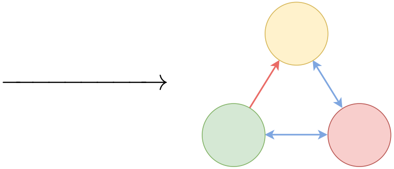

- Non-stationary time series
  - Heterogeneous data (distribution shifts but invariant full time graph)
  - SSMs with time-dependent effects (linear)
  - FCMs with time-dependent effects (GP regression)

# Causal discovery from conditionally stationary time series

**Idea:** Introduce categorical variables (states), that control the causal effects.

- Previous approaches consider the summary graph.

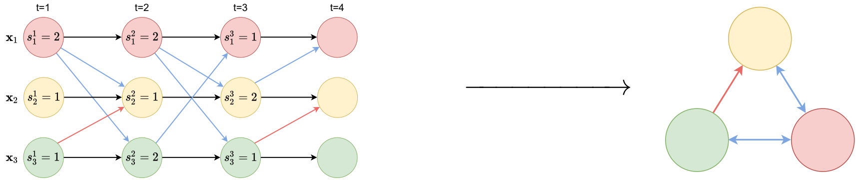


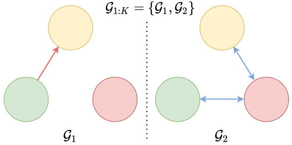
# Causal discovery from conditionally stationary time series

**Idea:** Introduce categorical variables (states), that control the causal effects.

- Previous approaches consider the summary graph.




- **Definition 1**: *Conditional summary graph*
  $\mathcal{G}_{1:K} = \{\mathcal{G}_k = \{\mathcal{V}, \mathcal{E}_k\} : 1 \leq k \leq K\}, \quad \mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$




$$\mathbf{x}_i^t \to \mathbf{x}_i^{t+1}, s_i^t = k \implies \mathbf{x}_i \to \mathbf{x}_j \in \mathcal{E}_k$$

# Causal discovery from conditionally stationary time series

**Idea:** Introduce categorical variables (states), that control the causal effects.

$$\left\{\{\mathbf{x}_i^t, s_i^t\}_{i=1}^N\right\}_{t=1}^T \sim \mathcal{D}, \quad \mathbf{x}_i^t \in \mathbb{R}^d, \quad s_i^t \in \{1,\dots,K\}$$

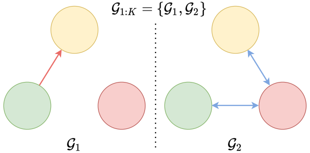
- Assumptions:
  1. Causal sufficiency: all variables are observed $\mathbf{x}_i^t \in \mathcal{V}^{1:T}$.
  2. States are observed $\{s_i^t : 1 \leq t \leq T, 1 \leq i \leq N\}$.
  3. First-order Markov setting.
  4. Additive noise model $\mathbf{x}_j^{t+1} = f_j\left(\mathbf{PA}(\mathbf{x}_j^t)\right) + \varepsilon_j^t, \quad \mathbf{PA}(\mathbf{x}_j^t) = \left(\mathbf{PA}_j^1\right)^{t-1}$.
  5. At time t, the states control the causal structure: $\mathbf{PA}(\mathbf{x}_j^t) = (\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}$.
- Model:

$$\mathbf{x}_j^t = f_j^{\mathbf{s}^{t-1}}\left((\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}\right) + \epsilon_j^t, \tag{1}$$

$$\mathbf{PA}_j^1|\mathbf{s}^{t-1} = \{\mathbf{x}_i : \mathbf{x}_j \in C_i(s_i^{t-1}), 1 \leq i \leq N\}, \tag{2}$$

$\mathbf{PA}(\mathbf{x}_j^t)$, generally not invariant in time

# Causal discovery from conditionally stationary time series




$$\mathbf{x}_i^t \to \mathbf{x}_i^{t+1}, s_i^t = k \implies \mathbf{x}_i \to \mathbf{x}_j \in \mathcal{E}_k$$

More **informative** and **compact** causal representation.

- **Theorem 1**: Under the following assumptions:
  1. Causal sufficiency: all variables are observed $\mathbf{x}_i^t \in \mathcal{V}^{1:T}$.
  2. States are observed $\{s_i^t : 1 \leq t \leq T, 1 \leq i \leq N\}$.
  3. First-order Markov setting.
  4. Additive noise model
  5. At time t, the states control the causal structure.

  the *full time graph* and *conditional summary graph* are **identifiable** from data.

# State-dependent Causal Inference

# Estimation

- Could we extend *TiMINo causality* [Peters et al., 2013] with observed states?

# Estimation

- Could we extend *TiMINo causality* [Peters et al., 2013] with observed states?

$$\mathbf{x}_j^t = f_j^{\mathbf{s}^{t-1}}\left((\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}\right) + \boldsymbol{\epsilon}_j^t \tag{3}$$

  - The direct causes of $\mathbf{x}_j^t$ depend on $\mathbf{s}^{t-1} \rightarrow K^N$ models!

# Estimation

- Could we extend *TiMINo causality* [Peters et al., 2013] with observed states?

$$\mathbf{x}_j^t = f_j^{\mathbf{s}^{t-1}}\left((\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}\right) + \boldsymbol{\epsilon}_j^t \tag{3}$$

  - The direct causes of $\mathbf{x}_j^t$ depend on $\mathbf{s}^{t-1} \rightarrow K^N$ models!

- To target real non-stationary domains we further assume.
  - Components can be shared across datapoints $\mathbf{X}_1, \mathbf{X}_2, \cdots \sim \mathcal{D}$.
  - Components can be shared across variables, i.e. $f_i^{\mathbf{k}} = f_j^{\mathbf{l}}$.

# Estimation

- Could we extend *TiMINo causality* [Peters et al., 2013] with observed states?

$$\mathbf{x}_j^t = f_j^{\mathbf{s}^{t-1}}\left((\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}\right) + \epsilon_j^t \quad (3)$$

  - The direct causes of $\mathbf{x}_j^t$ depend on $\mathbf{s}^{t-1} \to K^N$ models!
- To target real non-stationary domains we further assume.
  - Components can be shared across datapoints $\mathbf{X}_1, \mathbf{X}_2, \cdots \sim \mathcal{D}$.
  - Components can be shared across variables, i.e. $f_i^{\mathbf{k}} = f_j^{\mathbf{l}}$.
- We amortize the causal discovery task using deep learning [Löwe et al., 2020].

# Estimation

- Could we extend *TiMINo causality* [Peters et al., 2013] with observed states?

$$\mathbf{x}_j^t = f_j^{\mathbf{s}^{t-1}}\left((\mathbf{PA}_j^1|\mathbf{s}^{t-1})^{t-1}\right) + \epsilon_j^t \tag{3}$$

  - The direct causes of $\mathbf{x}_j^t$ depend on $\mathbf{s}^{t-1} \to K^N$ models!

- To target real non-stationary domains we further assume.
  - Components can be shared across datapoints $\mathbf{X}_1, \mathbf{X}_2, \cdots \sim \mathcal{D}$.
  - Components can be shared across variables, i.e. $f_i^{\mathbf{k}} = f_j^{\mathbf{l}}$.
- We amortize the causal discovery task using deep learning [Löwe et al., 2020].
- Consistency is left as future work.

# Generative model

- *Conditional summary graph* including **edge-types**.

$$\mathbf{W} = \left\{ w_{ijk} : 1 \leq i,j \leq N, 1 \leq k \leq K \right\}, \quad w_{ijk} \in \{0, \ldots, n_\epsilon - 1\}$$

  - **Edge-type:** Functional form of the causal effect $\{\text{"no-edge"}, f_1(\cdot), f_2(\cdot), \ldots\}$.

# Generative model

- *Conditional summary graph* including **edge-types**.

$$\mathbf{W} = \Big\{w_{ijk} : 1 \leq i, j \leq N, 1 \leq k \leq K\Big\}, \quad w_{ijk} \in \{0, \ldots, n_\epsilon - 1\}$$

  - **Edge-type:** Functional form of the causal effect $\{\text{"no-edge"}, f_1(\cdot), f_2(\cdot), \ldots\}$.
- $\mathbf{x}_i^t \in \mathbb{R}^d$ refers to some variables that we aim to predict (position, velocity, bounding box, ...).
- $s_i^t \in \{1, \ldots, K\}$ refers to state of element $i$ at time $t$.
- $z_{ij}^t \in \{0, \ldots, n_\epsilon - 1\}$: interaction from $i \rightarrow j$ at time $t$. Conditioned on $s_i^t$

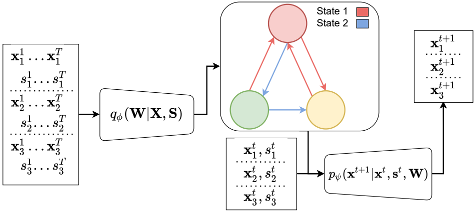
$$p(\mathbf{X}, \mathbf{W}|\mathbf{S}) = p(\mathbf{W}) \prod_{t=0}^{T-1} \prod_{j=1}^{N} p_\psi(\mathbf{x}_j^{t+1}|\mathbf{x}^t, \mathbf{s}^t, \mathbf{W}) \tag{4}$$

# State-dependent causal inference

SDCI implements a **VAE-based** approach.

1. **Encoder:** Edge-type inference.

$$q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S}) = \prod_{ijk} q_\phi(w_{ijk}|\mathbf{X},\mathbf{S})$$



# State-dependent causal inference

SDCI implements a **VAE-based** approach.

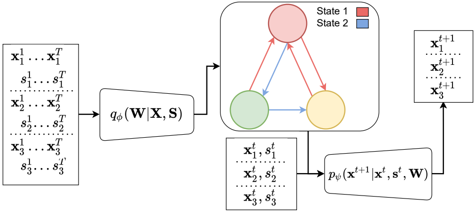
1. **Encoder:** Edge-type inference.

$$q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S}) = \prod_{ijk} q_\phi(w_{ijk}|\mathbf{X},\mathbf{S})$$

2. **Sample** and **compute** $z_{ij}^t$ at each time.

$$w_{ijk} \sim q_\phi(w_{ijk}|\mathbf{X},\mathbf{S})$$

$$z_{ij}^t = w_{ijk'}, k' = s_i^t$$



# State-dependent causal inference

SDCI implements a **VAE-based** approach.

1. **Encoder:** Edge-type inference.

$$q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S}) = \prod_{ijk} q_\phi(w_{ijk}|\mathbf{X},\mathbf{S})$$

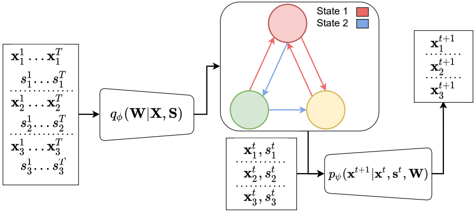
2. **Sample** and **compute** $z_{ij}^t$ at each time.

$$w_{ijk} \sim q_\phi(w_{ijk}|\mathbf{X},\mathbf{S})$$

$$z_{ij}^t = w_{ijk'}, k' = s_i^t$$

3. **Decoder:** Models the **dynamics**.

$$p_\psi(\mathbf{x}_i^{t+1}|\mathbf{x}^t,\mathbf{s}^t,\mathbf{W})$$



# State-dependent causal inference

- **Objective**: ELBO estimation

$$\log p(\mathbf{X}|\mathbf{S}) \geq -KL\left(q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S})||p(\mathbf{W})\right) + \mathbb{E}_{q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S})}\left[\log p_\psi(\mathbf{X}|\mathbf{W},\mathbf{S})\right]. \quad (5)$$

# State-dependent causal inference

- **Objective**: ELBO estimation

$$\log p(\mathbf{X}|\mathbf{S}) \geq -KL\left(q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S})||p(\mathbf{W})\right) + \mathbb{E}_{q_\phi(\mathbf{W}|\mathbf{X},\mathbf{S})}\left[\log p_\psi(\mathbf{X}|\mathbf{W},\mathbf{S})\right]. \quad (5)$$

- We can marginalise the states when these are **hidden** $\implies \{\{\mathbf{x}_i^t\}_{i=1}^N\}_{t=1}^T \sim \mathcal{D}$.

$$p(\mathbf{X},\mathbf{W},\mathbf{S}) = p(\mathbf{W}) \prod_{t=0}^{T-1} p_\psi(\mathbf{x}^{t+1}|\mathbf{x}^t,\mathbf{s}^t,\mathbf{W})p(\mathbf{s}^{t+1}|\mathbf{x}^{t+1}). \quad (6)$$

$$q_\phi(\mathbf{W},\mathbf{S}|\mathbf{X}) = q_\phi(\mathbf{W}|\mathbf{X})q_\phi(\mathbf{S}|\mathbf{X})$$

We now lack **identifiability guarantees** $\implies$ more assumptions, restrictions.

# Experiments

# Experiments – Baseline comparison

- Linear data

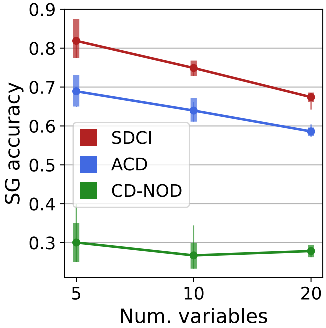
$$\mathbf{x}_j^{t+1} = \alpha \mathbf{x}_j^t + \sum_{i \neq j}^{N} \beta_k \mathbf{x}_i^t + \boldsymbol{\epsilon}_j^t, \quad k = \left(\tilde{\mathcal{E}}_{s_i^t}\right)_{ij}$$

| Method | SG Accuracy 2-edge | 3-edge const | 3-edge free |
|---|---|---|---|
| TdCM (T=100) | 65.17 ± 2.65 | 63.67 ± 1.61 | 63.50 ± 1.62 |
| CD-NOD (T=100) | 39.33 ± 2.59 | 35.25 ± 2.51 | 28.58 ± 2.66 |
| SAEM (T=100) | 47.75 ± 3.67 | 39.04 ± 2.38 | 51.44 ± 3.81 |
| TdCM (T=1000) | 68.25 ± 2.29 | 61.17 ± 2.28 | 62.00 ± 2.14 |
| CD-NOD (T=1000) | 50.08 ± 2.59 | 42.08 ± 2.17 | 41.58 ± 2.02 |
| SAEM (T=1000) | 47.38 ± 4.10 | 25.93 ± 2.82 | 28.49 ± 3.28 |
| ACD (T=50) | 60.45 ± 1.60 | 87.00 ± 2.56 | 49.25 ± 3.05 |
| SDCI (T=50) | **97.08** ± 1.05 | **90.17** ± 2.22 | **64.00** ± 2.93 |
| | **CSG Accuracy** 2-edge | 3-edge const | 3-edge free |
| SDCI (T=50) | 98.08 ± 0.64 | 76.04 ± 2.05 | 65.45 ± 1.99 |

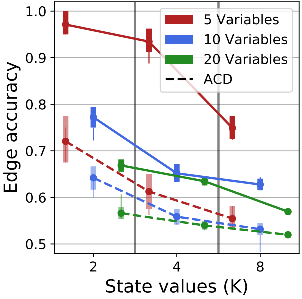
- Spring data

$$\mathbf{f}_{ij} = -\delta_k(\mathbf{r}_i - \mathbf{r}_j), \quad \{\delta_0 = 0, \delta_1 = 1\}$$

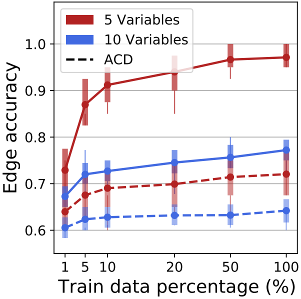
$$k = z_{ij}^t, \quad \ddot{\mathbf{r}}_i = \sum_{j=1}^{N} \mathbf{f}_{ij} \quad \mathbf{x}_i = \{\mathbf{r}_i, \dot{\mathbf{r}}_i\}$$



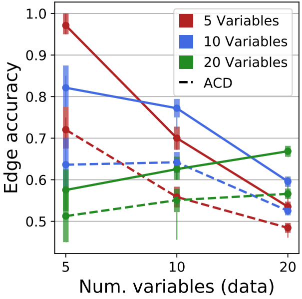
# Experiments – Spring




(a) Increasing variables and states.




(b) Data efficiency.



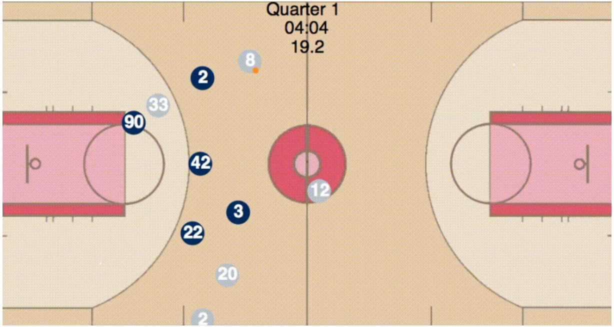

(c) Generalisation.

# Experiments – NBA

- We test SDCI on **realistic scenarios**, like NBA player trajectories.

# Experiments – NBA

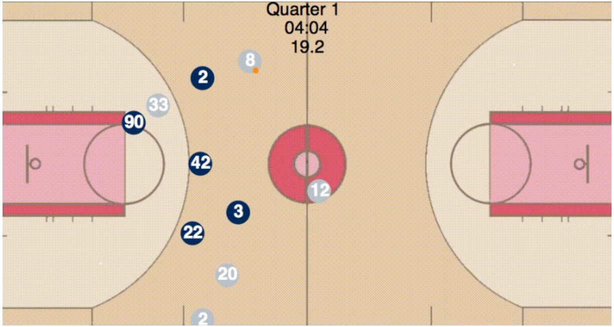
- We test SDCI on **realistic scenarios**, like NBA player trajectories.

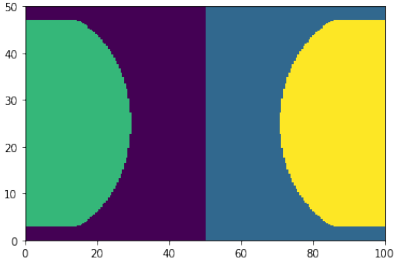



- We sample long trajectories $T \approx 200$, with 2D position and velocity ($\mathbf{x}_i^t \in \mathbb{R}^4$).
- Total training size: $\sim$150k samples.

# Experiments – NBA

- We test SDCI on **realistic scenarios**, like NBA player trajectories.



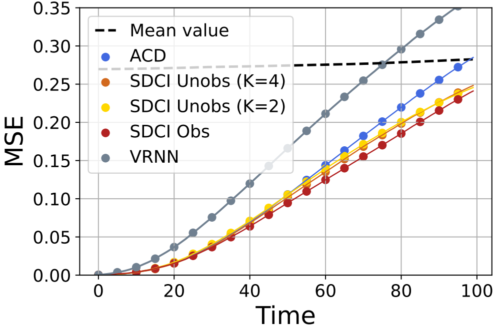




- We sample long trajectories $T \approx 200$, with 2D position and velocity ($\mathbf{x}_i^t \in \mathbb{R}^4$).
- Total training size: $\sim$150k samples.

- We hand-craft a state function to study SDCI-observed.

# Experiments – NBA

- SDCI-observed achieves comparable performance to other sequential generative baselines in **forecasting**.
- SDCI-unobserved learns regimes where dynamical changes occur.



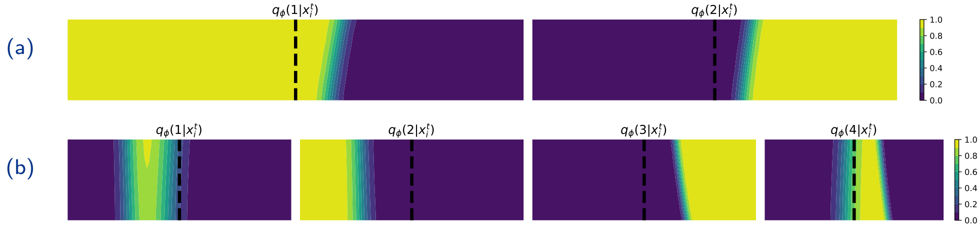




Figure: Learned regimes from SDCI on the NBA dataset using (a) $K = 2$ and (b) $K = 4$.

# Conclusions

# Conclusions

- Our **goal** is to learn representations from sequential data in real non-stationary domains.

# Conclusions

- Our **goal** is to learn representations from sequential data in real non-stationary domains.
- We developed State-dependent causal inference (SDCI) for causal discovery in conditional time series data.

# Conclusions

- Our **goal** is to learn representations from sequential data in real non-stationary domains.
- We developed State-dependent causal inference (SDCI) for causal discovery in conditional time series data.
- The hidden state setting allows us to model non-stationary behaviours present in realistic scenarios.

# Conclusions

- Our **goal** is to learn representations from sequential data in real non-stationary domains.
- We developed State-dependent causal inference (SDCI) for causal discovery in conditional time series data.
- The hidden state setting allows us to model non-stationary behaviours present in realistic scenarios.
- We showcase models as SDCI could be leveraged for data interpretability.

# References I


Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017).
Building machines that learn and think like people.
*Behavioral and brain sciences*, 40.

Löwe, S., Madras, D., Zemel, R. S., and Welling, M. (2020).
Amortized causal discovery: Learning to infer causal graphs from time-series data.
*ArXiv*, abs/2006.10833.

Peters, J., Janzing, D., and Schölkopf, B. (2013).
Causal inference on time series using restricted structural equation models.
*Advances in Neural Information Processing Systems*, 26.

# References II


Wang, X. and Gupta, A. (2018).
Videos as space-time region graphs.
In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417.

# Appendix

# Causal discovery methods – Overview

- Constrain-based methods

  C.I. tests
  $X_1 \perp\!\!\!\perp X_2 | X_3$

  E.g. PC algorithm, FCI, ts-PC, ...

# Causal discovery methods – Overview

- Constrain-based methods

  C.I. tests
  $X_1 \perp\!\!\!\perp X_2 | X_3$

  E.g. PC algorithm, FCI, ts-PC, ...

- Score-based methods

  Greedy Equivalence Search (GES)
  For time series: Learning *Dynamic Bayesian Networks*

  E.g. DYNOTEARS, ...

# Causal discovery methods – Overview

- Constrain-based methods

  C.I. tests
  $X_1 \perp\!\!\!\perp X_2 | X_3$

  E.g. PC algorithm, FCI, ts-PC, ...

- Score-based methods

  Greedy Equivalence Search (GES)
  For time series: Learning *Dynamic Bayesian Networks*

  E.g. DYNOTEARS, ...

- Functional model-based methods

  Functional models represent cause and effect
  $X_1 = f_1(X_2, X_3, \epsilon_1)$

  E.g. VAR, (neural) Granger causality, ...