The University of Osaka | SANKEN THE UNIVERSITY OF OSAKA
Submission ID No. 1208

Modeling Time-evolving Causality over Data Streams

Naoki Chihara, Yasuko Matsubara, Ren Fujiwara, Yasushi Sakurai
SANKEN, The University of Osaka

Source Code

Motivation - Data streams appear all around now

Motion analysis | Epidemiology | Web activity

Challenges - Causal relationsihps drift over time

- For example, the emergence of a new virus strain in a country leads to an increase in the number of infections in other countries

🧐 **How to get time-changing causal relationships?**

Problem Definition - We tackle the following challenges

Given: Multivariate data stream, i.e., $\boldsymbol{X} = \{\boldsymbol{x}(1), \dots, \boldsymbol{x}(t_c), \dots\}$

Goal: Achieve all of the following objectives

- **Find** distinct dynamical patterns / **regimes**
- **Discover** causal relationships, which changes over time / **time-evolving causality**
- **Forecast** an l_s-steps ahead future values

✨ **ModePlait: novel streaming method**

Proposed Model - ModePlait

Classical framework for causality

Key Concepts - Our model is designed based on SEM

Exogenous variables evolve over time / **inherent signals**

$$\boldsymbol{X}_{\text{sem}} = \boldsymbol{B}_{\text{sem}} \boldsymbol{X}_{\text{sem}} + \boldsymbol{E}_{\text{sem}}$$

Observed variables | Causal adjacency matrix | Exogenous variables

$x_i = \sum_{j \in Pa(i)} b_{ij} x_j + e_i$: Structural equation describing x_i

Main idea (P1): Latent temporal dynamics

😢 Each inherent signal $e_{(i)}(t)$ is only a single dimension

$\Rightarrow$ superposition of computed basis vectors (i.e., **modes**)

Latent activities | Eigenvalues matrix (Temporal dynamics)

$$\boldsymbol{s}_{(i)}(t+1) = \boldsymbol{\Lambda}_{(i)} \boldsymbol{s}_{(i)}(t) \quad : k_i\text{-dimensional space}$$

$$e_{(i)}(t) = g^{-1}(\boldsymbol{\Phi}_{(i)} \boldsymbol{s}_{(i)}(t)) \quad : \text{Projection } (\mathbb{C}^{k_i} \rightarrow \mathbb{R})$$

Augmentaion

i-th univariate inherent signal | Time-delay embedding | Modes (Spatial dynamics)

$\mathcal{D}_{(i)} = \{\boldsymbol{\Phi}_{(i)}, \boldsymbol{\Lambda}_{(i)}\}$ / **self-dynamics factor set**

$g(e_{(i)}(t)) := (e_{(i)}(t), e_{(i)}(t-1), \dots, e_{(i)}(t-h+1))$, k_i: # of modes

Main idea (P2): Dynamical patterns

Describe distinct dynamical pattern (i.e., **regime**)

$$\boldsymbol{s}_{(i)}(t+1) = \boldsymbol{\Lambda}_{(i)} \boldsymbol{s}_{(i)}(t) \quad (1 \le i \le d)$$

$$e_{(i)}(t) = g^{-1}(\boldsymbol{\Phi}_{(i)} \boldsymbol{s}_{(i)}(t)) \quad (1 \le i \le d)$$

$$\boldsymbol{v}(t) = \boldsymbol{W}^{-1} \boldsymbol{e}(t)$$

Estimated values | Mixing matrix (Relationships b/w inherent signals) | (P1) Collection of d self-dynamic factor sets

$\theta = \{\boldsymbol{W}, \mathcal{D}_{(1)}, \dots, \mathcal{D}_{(d)}\}$ / **regime**, $\Theta = \{\theta^1, \dots, \theta^R\}$ / **regime set**

$\mathcal{B} = \{\boldsymbol{B}^1, \dots, \boldsymbol{B}^R\}$ / **time-evolving causality**

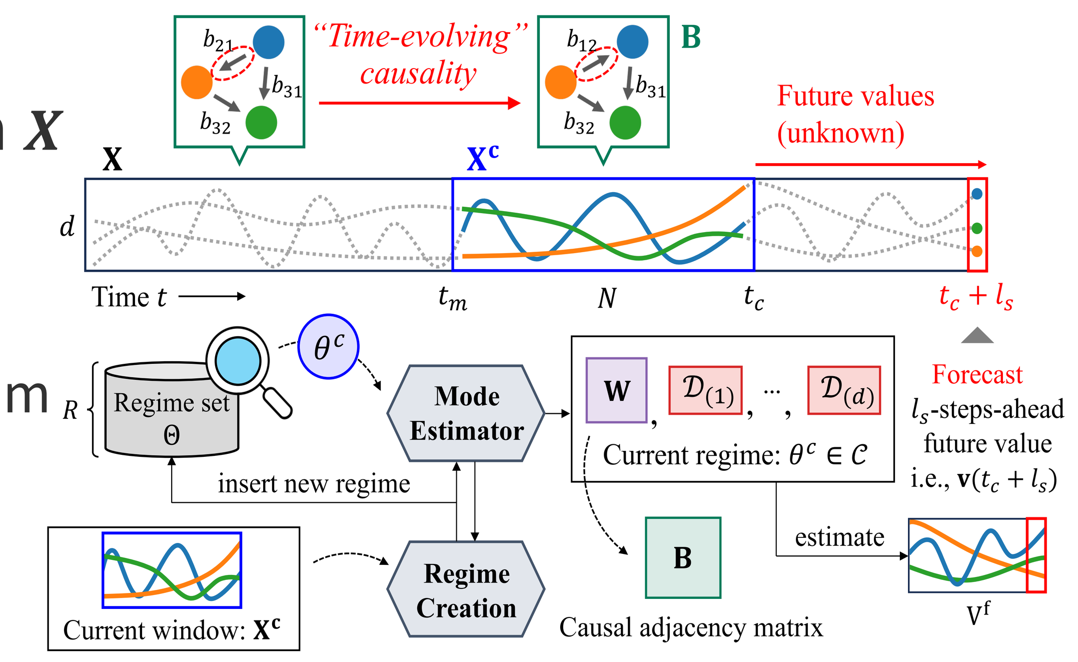
Optimization algorithm

Given:

- Multivariate data Stream $\boldsymbol{X}$

Estimate:

- Full parameter set
  $\mathcal{F} = \{\Theta, \Omega\}$, Ω: update param
- Model candidate
  $C = \{\theta^c, \omega^c, S^c_{en}\}$
- Time-evolving causality
  $\mathcal{B} = \{\boldsymbol{B}^1, \dots, \boldsymbol{B}^R\}$, R: # of regimes
- l_s-steps ahead future value $\boldsymbol{v}(t_c + l_s)$, t_c: current time point



Theoretical analysis - See Lemma 2 & 3 for details

- It theoretically discovers causal relationships
- It is practical for semi-infinite data streams

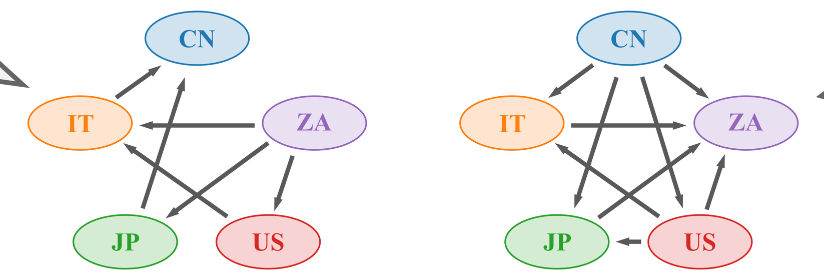
Details in paper

Experiments - Answer the essential questions

Q1. Effectiveness - Epidemiological data stream

Base of arrows is cause, head is effect

longest and toughest lockdowns in Shanghai

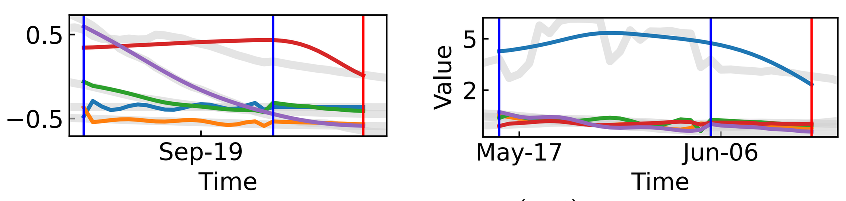


(a-i) January 8, 2021 (a-ii) May 19, 2022

(a) Causal relationships at different time points

Health officials report a new lineage of the coronavirus in South Africa

Accurate forecast based on the current distinct dynamical patterns



(c-i) September 27, 2021 (c-ii) June 5, 2022

(c) Snapshots of 10 days-ahead future value forecasting

Q2. Accuracy - Causal discovery and Forecasting

Table 3: Causal discovering results with multiple temporal sequences to encompass various types of real-world scenarios.

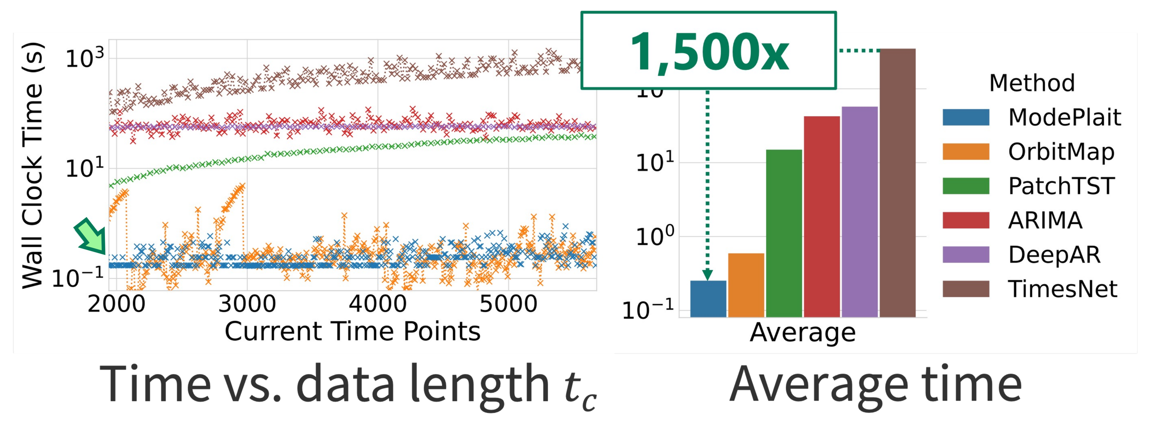
Models	ModePlait		CASPER		DARING		NoCurl		NO-MLP		NOTEARS		LiNGAM		GES	
Metrics	SHD	SID	SHD	SID	SHD	SID	SHD	SID	SHD	SID	SHD	SID	SHD	SID	SHD	SID
1, 2, 1	**3.82**	**4.94**	5.58	7.25	5.75	8.58	6.31	9.90	6.36	8.74	5.03	9.95	7.13	8.23	7.49	11.7
1, 2, 3	**4.48**	**6.51**	5.97	8.44	5.81	9.17	6.13	9.51	6.44	8.77	5.69	9.56	6.79	7.33	7.03	10.1
1, 2, 2, 1	**4.32**	**5.88**	5.41	8.41	6.54	9.17	6.69	10.0	6.55	8.72	5.23	9.54	7.12	8.65	7.08	9.77
1, 2, 3, 4	**4.21**	**5.76**	6.22	8.33	6.12	9.58	6.10	9.61	6.62	8.87	5.73	10.1	7.10	8.50	7.29	11.3
1, 2, 3, 2, 1	**4.50**	**6.11**	6.02	8.28	5.45	7.77	6.20	9.83	6.56	8.83	5.57	9.11	7.46	8.05	7.74	12.1

Table 4: Multivariate forecasting results for both synthetic and real-world datasets. We used forecasting steps $l_s \in \{5, 10, 15\}$.

Models		ModePlait		TimesNet		PatchTST		DeepAR		OrbitMap		ARIMA	
Metrics		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
#0 synthetic	5	**0.722**	**0.528**	0.805	0.578	0.768	0.581	1.043	0.821	0.826	0.567	0.962	0.748
	10	**0.829**	**0.607**	0.862	0.655	0.898	0.649	1.073	0.849	0.896	0.646	0.966	0.752
	15	**0.923**	**0.686**	0.940	0.699	0.973	0.706	1.137	0.854	0.966	0.710	0.982	0.765
#1 covid19	5	**0.588**	**0.268**	0.659	0.314	0.640	0.299	1.241	0.691	1.117	0.646	1.259	0.675
	10	**0.740**	**0.361**	0.841	0.410	1.053	0.523	1.255	0.693	1.353	0.784	1.260	0.687
	15	**0.932**	**0.461**	1.026	0.516	1.309	0.686	1.265	0.690	1.351	0.792	1.277	0.718
#2 web-search	5	**0.573**	**0.442**	0.626	0.469	0.719	0.551	1.255	1.024	0.919	0.640	1.038	0.981
	10	**0.620**	**0.481**	0.697	0.514	0.789	0.604	1.273	1.044	0.960	0.717	1.247	1.037
	15	**0.646**	**0.505**	0.701	0.527	0.742	0.571	1.300	1.069	0.828	0.631	1.038	0.795
#3 chicken-dance	5	**0.353**	**0.221**	0.759	0.490	0.492	0.303	0.890	0.767	0.508	0.316	2.037	1.742
	10	**0.511**	**0.325**	0.843	0.564	0.838	0.535	0.886	0.753	0.730	0.476	1.863	1.530
	15	**0.653**	**0.419**	0.883	0.592	0.972	0.654	0.862	0.718	0.903	0.565	1.792	1.481
#4 exercise	5	**0.309**	**0.177**	0.471	0.275	0.465	0.304	0.408	0.290	0.424	0.275	1.003	0.748
	10	**0.501**	**0.309**	0.630	0.381	0.789	0.518	0.509	0.382	0.616	0.377	1.104	0.814
	15	0.687	**0.433**	0.786	0.505	1.147	0.758	**0.676**	0.475	0.691	0.434	1.126	0.901

Q3. Scalability

- It requires only constant computational time w.r.t. a data stream length



Time vs. data length t_c | Average time

🔥 **ModePlait outperforms its competitors**

Conclusion - ModePlait has following properties:

Effective: it discovers time-evolving causality

Accurate: its performance is confirmed by experiments

Scalable: it does not depend on stream length