時間変化する因果関係の抽出に基づいた高速将来予測

千原 直己†,‡　松原 靖子†　藤原 廉†,‡　櫻井 保志†
† 大阪大学 産業科学研究所　‡ 大阪大学大学院 情報科学研究科
OSAKA UNIVERSITY　SANKEN OSAKA UNIVERSITY

KDD Paper　Source code

研究背景

時系列データストリーム

- 幅広い分野で日々社会の中で生成され続けている
- 観測され続ける最新のデータを用いてモデルを更新可能な手法の需要が高まっている

変数間の関係性

- e.g., 相関関係，**因果関係**，独立性
  - 因果関係：変数間の原因と結果を意味する関係性
- 時系列分析の精度向上のために長年活用され続けている

研究課題

- 既存手法の大半は因果関係が時系列データ内で変動しないことを仮定している [Runge 2018]
- 環境の移り変わりによって因果関係が変動することを考慮すると，この仮定は限定的であるといえる
  - e.g., 新商品の発売により既製品の売り上げが低下する

✨ **ModePlait: 最新のストリーミング手法**

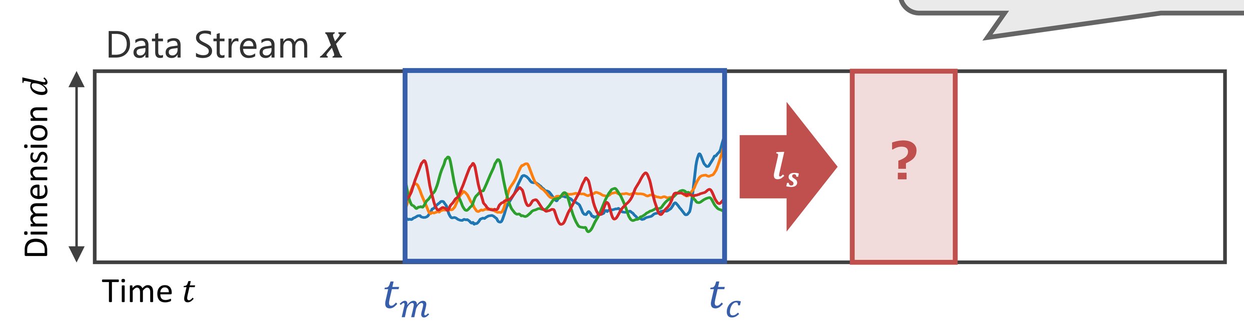
問題定義

Given: 時系列データストリーム $X = \{x(1), \dots, x(t_c), \dots\}$

Goal: 以下の重要な課題を全て達成する　(t_c: 現在時刻)

- 特徴的な時系列パターン (**レジーム**) の発見
- 時間変化する因果関係の抽出
- l_s ステップ先の値の予測

半無限長な時系列データ



提案モデル - ModePlait モデル

1. 構造方程式モデル [Pearl 2009]

$$\boxed{X_{\text{sem}}} = \boxed{B_{\text{sem}}}\, X_{\text{sem}} + \boxed{E_{\text{sem}}}$$

観測変数　因果隣接行列　外生変数

外生変数の時間変動性に着目
👉 **固有信号**

2. 固有信号の潜在的な時間ダイナミクス

潜在変数　固有値行列　モード　モードの数

$$s_{(i)}(t+1) = \Lambda_{(i)}\, s_{(i)}(t) : k_i \text{ 次元の潜在空間}$$

$$e_{(i)}(t) = g^{-1}(\Phi_{(i)}\, s_{(i)}(t)) : \text{観測空間への射影 } (\mathbb{C}^{k_i} \to \mathbb{R})$$

次元拡張

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

単変量データを基底ベクトルの重ね合わせで表現する

固有ダイナミクス集合：$\boxed{\mathcal{D}_{(i)} = \{\Phi_{(i)}, \Lambda_{(i)}\}}$

過去データにより位相情報が付与される

3. 特徴的な時系列パターン

$$s_{(i)}(t+1) = \Lambda_{(i)} s_{(i)}(t) \quad (1 \le i \le d)$$

$$e_{(i)}(t) = g^{-1}(\Phi_{(i)} s_{(i)}(t)) \quad (1 \le i \le d)$$

d 個の固有ダイナミクス集合

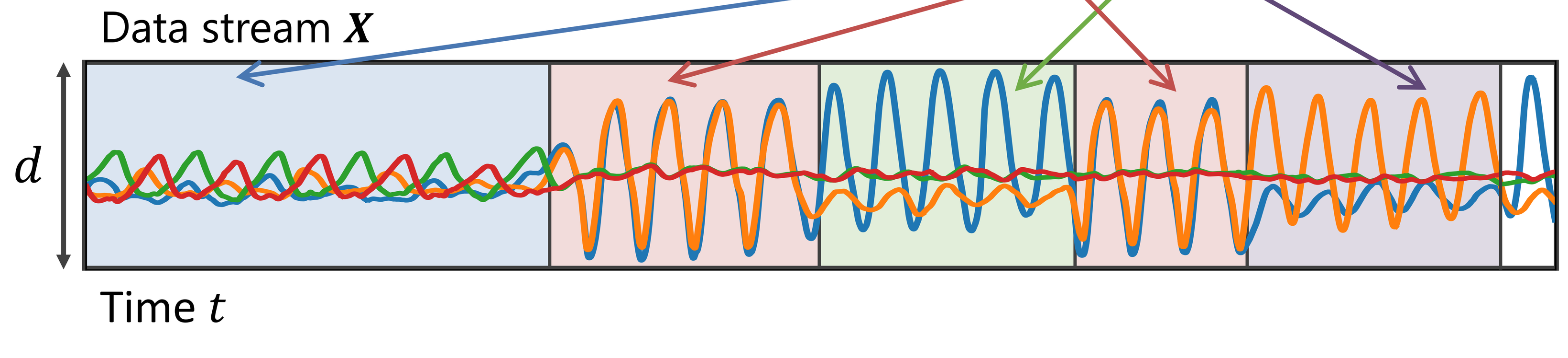
$$v(t) = W^{-1} e(t) \qquad e(t) = \{e_{(i)}(t)\}_{i=1}^{d}$$

推定値　混合行列

レジーム：$\boxed{\theta = \{W, \mathcal{D}_{(1)}, \dots, \mathcal{D}_{(d)}\}}$

4. レジームの動的遷移

- 要約のために**レジームセット** $\Theta = \{\theta_1, \theta_2, \dots, \theta_r\}$ を利用



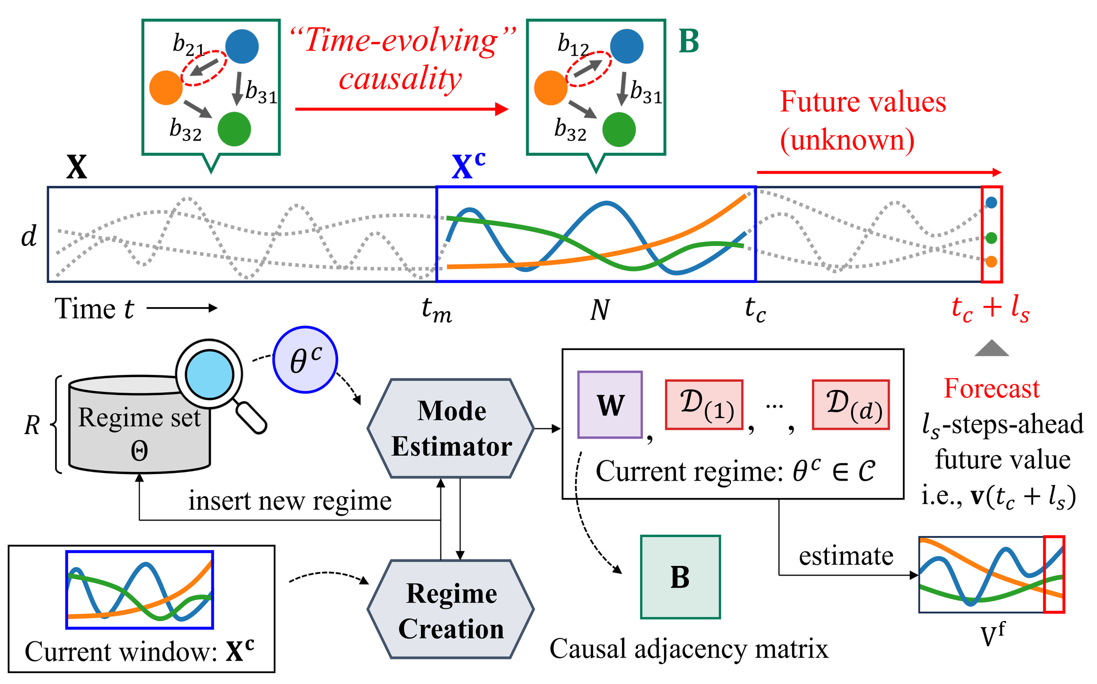
アルゴリズム - 以下の4つのステップで構成される

Step 1: ModeEstimator
適切なレジーム θ^c を探索

Step 2: RegimeCreation
未知のパターンの場合，カレントウィンドウ X^c からレジーム θ^c を推定

Step 3: ModeGenerator
因果隣接行列 B の推定と l_s ステップ先の値の予測

Step 4: RegimeUpdater
最新の観測 $x(t_c)$ を用いてレジーム θ^c の更新



$$A^{new}_{(i)} = A^{prev}_{(i)} + (g(e_{(i)}(t_c)) - A^{prev}_{(i)} g(e_{(i)}(t_c-1)))\gamma_{(i)}$$

$$\gamma_{(i)} = \frac{g(e_{(i)}(t_c-1))^\top P^{prev}_{(i)}}{\mu + g(e_{(i)}(t_c-1))^\top P^{prev}_{(i)} g(e_{(i)}(t_c-1))}$$

$$P^{new}_{(i)} = \frac{1}{\mu}(P^{prev}_{(i)} - P^{prev}_{(i)} g(e_{(i)}(t_c-1))\gamma_{(i)})$$

固有ダイナミクス集合 $\mathcal{D}_{(i)}$ のための更新式

モデル更新用パラメータ：$\omega = \{\{P_{(i)}\}_{i=1}^{d}, \{\epsilon_{(i)}\}_{i=1}^{d}\}$

モデルパラメータ集合：$\mathcal{F} = \{\Theta, \Omega\}$　**モデル候補**：$\mathcal{C} = \{\theta^c, \omega^c, S^c_{en}\}$

理論的分析 - 提案手法は優れた特性を有する

定理 2. ModePlait における因果探索は，ModeGenerator での因果隣接行列 B の抽出と同値である

定理 3. 各プロセスにおける ModePlait の計算時間量は少なくとも $O(N\sum_i k_i + dh^2)$ であり，高々 $O(RN\sum_i k_i + N(d^2 + h^2) + k^2)$ である

- データ全体の長さに依存しない計算時間

Details in paper

評価実験 - 以下の全ての疑問に対して回答する

Q1. 有効性 - 疫病データストリームに対する出力例

主要５カ国における COVID-19の感染者数で構成される

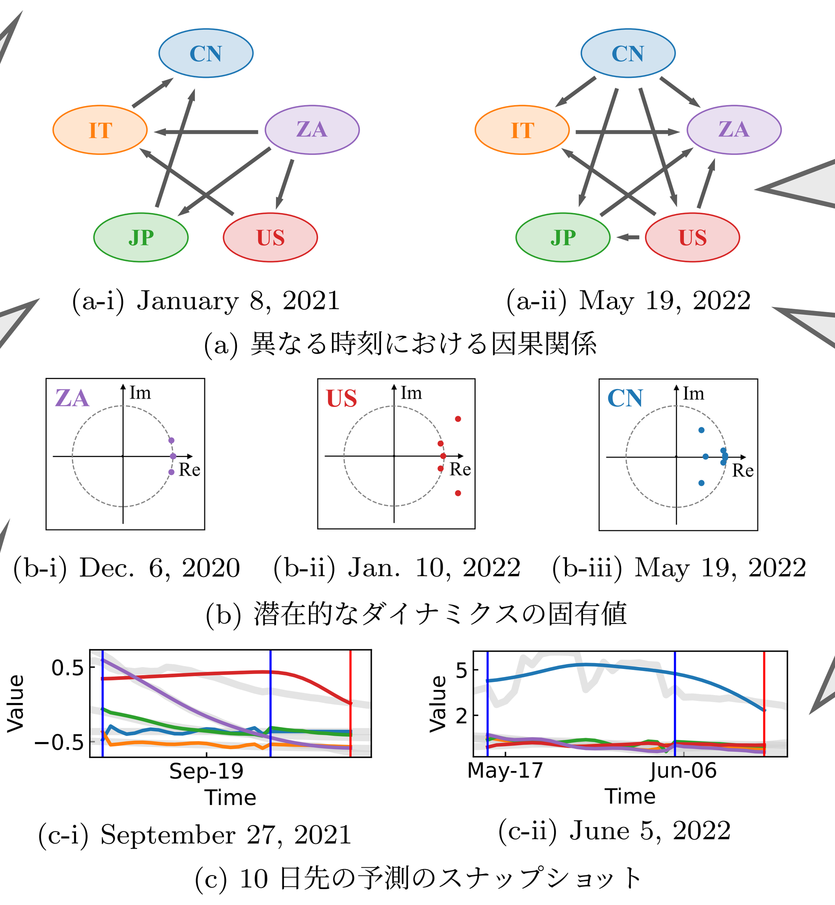
南アフリカ保健省が最新の変異株 501.V2 の発見を発表

固有信号が従う動的システムの動的特性を示す

各矢印の基部は原因を先端は結果を意味している

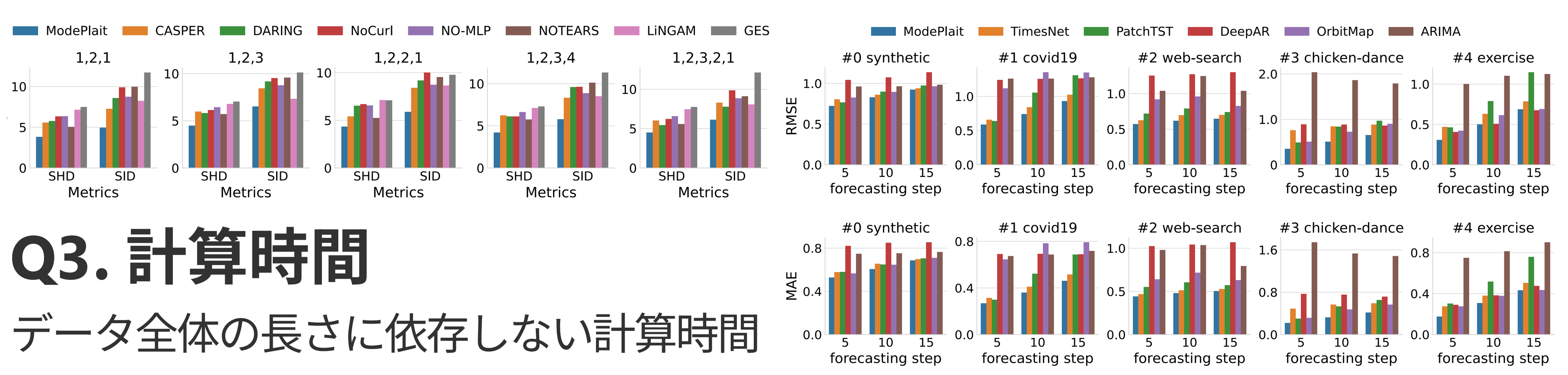
上海における過去最大のロックダウン

時系列パターンの遷移を考慮した高精度な将来予測を達成



Q2. 正確性

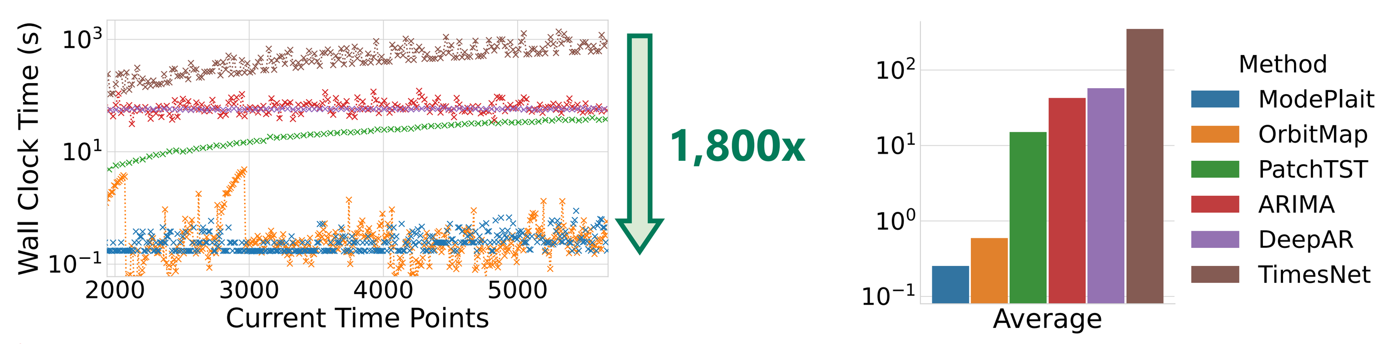
因果探索 - SHD, SID　**時系列予測 - RMSE, MAE**



Q3. 計算時間

データ全体の長さに依存しない計算時間

各時刻における計算時間



平均計算時間

🔥 **ModePlait の優れた性能を実証**

まとめ - ModePlait は以下の特性を全て満たす

Effective: 特徴的な時系列パターン (レジーム) の遷移に基づいて時間変化する因果関係を抽出する

Accurate: 理論的に因果関係を抽出し，正確に将来を予測することで最新の比較手法を上回る精度を達成した

Scalable: データの長さに依存しない高速な処理が可能である