情報処理学会論文誌　データベース　Vol.14　No.3　30–41 (July 2021)

オンライン活動データストリームのための非線形モデル解析

川畑 光希[1,2,a)]　松原 靖子[2,b)]　本田 崇人[3,c)]　櫻井 保志[2,d)]

受付日 2020年12月10日, 採録日 2021年4月4日

概要：Web 検索履歴等に代表される大規模時系列データは，時刻や地域，キーワードといった様々な情報とともに収集され，テンソルストリームとして扱うことができる．Web 上におけるユーザアクティビティの解析では，より高精度な将来予測を実現することが重要な課題の 1 つであるが，複雑な構造を持つテンソルストリームから将来予測に有用なパターンを発見することが問題となる．本論文では，時間，国，キーワードの 3 つ組に対する Web 検索数で構成されるテンソルストリームを効果的に解析するためのストリームアルゴリズムである CubeCast を提案する．CubeCast は与えられたテンソルストリームに含まれる潜在的な長期トレンドと季節パターンを発見し，それらを基に類似した特徴を持つ地域グループへと分解する．このとき，提案手法は次の特長を持つ．(a) 長期トレンドと季節パターンの非線形特性を単一のモデルで表現する．(b) パラメータチューニングや事前知識を必要とせず，時系列モデルやパターン変化を自動的に推定する．(c) 逐次的かつ適応的にパターン変化をとらえ，テンソルストリームを効率的に処理する．実データを用いた実験では，提案手法が将来予測に有用なパターンを効果的かつ効率的に発見できることを示し，既存の時系列予測手法と比較して，予測精度，計算時間の改善を確認した．

キーワード：時系列予測，テンソル分解，データストリーム処理

Non-linear Mining of Social Activities in Tensor Streams

Koki Kawabata[1,2,a)]　Yasuko Matsubara[2,b)]　Takato Honda[3,c)]　Yasushi Sakurai[2,d)]

Received: December 10, 2020, Accepted: April 4, 2021

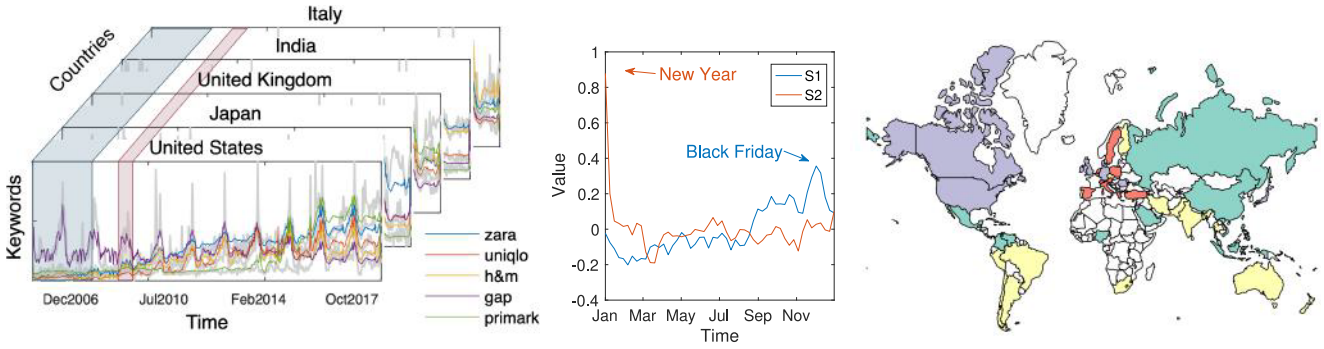
Abstract: Given a large time-evolving event series such as Google web-search logs, which are collected according to various aspects, i.e., timestamps, locations and keywords, how accurately can we forecast their future activities? How can we reveal significant patterns that allow us to long-term forecast from such complex tensor streams? In this paper, we propose a streaming method, namely, CubeCast, that is designed to capture basic trends and seasonality in tensor streams and extract temporal and multi-dimensional relationships between such dynamics. Our proposed method has the following properties: (a) it is *effective*: it finds both trends and seasonality and summarizes their dynamics into simultaneous non-linear latent space. (b) it is *automatic*: it automatically recognizes and models such structural patterns without any parameter tuning or prior information. (c) it is *scalable*: it incrementally and adaptively detects shifting points of patterns for a semi-infinite collection of tensor streams. Extensive experiments that we conducted on real datasets demonstrate that our algorithm can effectively and efficiently find meaningful patterns for generating future values, and outperforms the state-of-the-art algorithms for time series forecasting in terms of forecasting accuracy and computational time.

Keywords: time series forecasting, tensor decomposition, stream processing

1. まえがき

時系列予測は，センサネットワーク監視 [1], [2]，ユーザ行動分析 [3]，意思決定支援 [4]，等の幅広い分野で重要な役

[1] 大阪大学大学院情報科学研究科
Osaka University, Suita, Osaka 565–0871, Japan
[2] 大阪大学産業科学研究所
ISIR Osaka University, Ibaraki, Osaka 567–0047, Japan
[3] 株式会社 JDSC
JDSC Co. Ltd., Bunkyo, Tokyo 113–0033, Japan
a) koki@sanken.osaka-u.ac.jp
b) yasuko@sanken.osaka-u.ac.jp
c) takato.honda@jdsc.ai
d) yasushi@sanken.osaka-u.ac.jp



(a) Tensor stream　(b) Seasonal patterns　(c) Location-specific pattern groups

図 1　テンソルストリームと CubeCast の出力結果

Fig. 1　Modeling power of CubeCast for an online search volume tensor stream related to five apparel companies: (a) Given the original tensor (gray lines), CubeCast quickly identifies the non-linear dynamics in the latest tensor (blue), then, continuously forecasts multiple steps ahead values (red), (b) while extracting seasonal patterns common to all countries. (c) It also automatically identifies similar country groups based on their trends and seasonality, which are compressed into compact models.

割を果たす．特に，近年では IoT [*1] [5], [6] やクラウドサービスの普及による大規模データ解析の需要が高く，高精度な時系列予測には，複数の属性を持つ高次元時系列データから重要なパターンを効果的にとらえるための時系列モデルが必要となる．たとえば，マーケターは在庫管理や新製品の開発を行う際，意思決定に役立つ情報を求めて自社製品に対する顧客の反応を分析する．注目すべき点は，これまでのデータの中で何の需要が（製品の属性），どこで（地理的属性）で，どのように（時系列特性）変化しているのかを推定することである．時系列解析を通して今後必要とされる需要をあらかじめ見積もることにより，人材や資源をより効果的に利用することが可能になると考えられる．そこで本論文では，時間，国，キーワードの 3 つ組に対する Web 検索数で構成される大規模テンソルストリームを扱い，Web 検索履歴の多角的な解析と将来予測を試みる．一般に，テンソルストリームは多くのノイズを含み，高次元データの中から重要なパターンを発見することは難しい．また，多様な趣向や社会的なイベントに影響された複雑な時系列パターンは，時間とともに変化する．よって，本研究では以下の 2 つ課題に取り組む．

- 潜在トレンドの自動検出：多くのオンライン活動データは複数のトレンドを持ち，長期的な成長・衰退といった長期トレンドと周期的に現れる季節パターンに大別され，これらを抽出することで効果的に将来予測を行うことができるが，これらの時系列パターンに関する事前知識が与えられることは稀である．テンソルストリームは高次元であるため考慮すべき属性の組合せが増加するだけでなく，時系列パターンの特性はデータによって異なるため適切なモデルを設計することは容易ではない．そのため，テンソルストリームに含まれる潜在的なトレンドを自動的かつ効果的に抽出することが重要である．
- 動的変化のモデリング：本研究では，テンソルストリームのトレンドを時系列と地域の 2 方向から分析し，それらの特徴を単一の時系列パターンとして抽出することで時系列予測精度の改善を図る．また，それらの特徴の変化を監視することで適応的かつ効率的に予測モデルを更新するアルゴリズムを開発する．

本論文では，オンライン活動履歴等から得られるテンソルストリームに対する，高速かつ高精度な将来予測手法として CubeCast [7] を提案する．

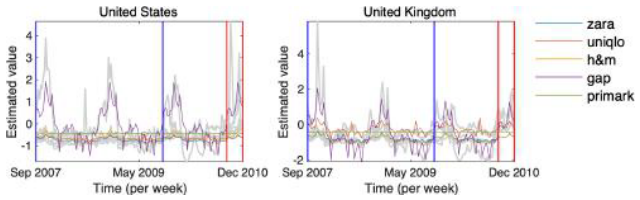
より具体的には，次の問題を扱う．テンソルストリーム $\mathcal{X}$ が与えられたとき，以下の能力を有する時系列モデル，および，将来予測アルゴリズムを開発する：

- 長期トレンド/季節パターンに基づく非線形時系列パターンの抽出とテンソルストリームの多角的解析
- ユーザの介入を必要としない特徴自動抽出
- データストリーム処理に基づくモデル学習

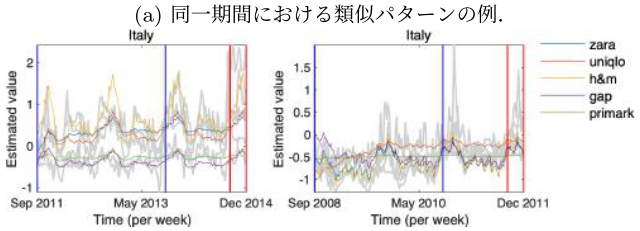
1.1　具体例

図 1 に CubeCast を用いたテンソルストリームの解析例を示す．図 1 (a) は解析に使用したテンソルデータの一部を表し，テンソルの各要素は 50 カ国において 5 つのアパレル企業の社名が検索された回数を週ごとに集計したものである．図中の灰色の線はオリジナルデータ，色の付いた線は各企業の検索数に対する提案手法のモデリング結果を示す．提案手法は，テンソルストリームの一部（青枠）を保持しながらパターン検出とモデル更新を繰り返し，1

[*1] IoT: Internet of Things



(a) 同一期間における類似パターンの例.



(b) 時系列パターンの時間発展の例.

図 2　CubeCast を用いたテンソルストリームの解析例

Fig. 2　Multi-aspect mining of CubeCast for GoogleTrends related to major apparel companies. It automatically detects (a) similar country groups based on dynamics, and (b) changes between discrete dynamics.

年後の検索数（赤枠）を予測し続ける．このとき，提案手法は以下に示すテンソルストリームの特徴を自動的かつリアルタイムに検出する．

長期トレンド検出．図 1 (a) に示すように，アメリカにおける各企業の検索数には様々な長期トレンドが見られる．GAP の検索数は，2010 年まで下降し，2014 まで上昇したのち再び下降している．他の企業の検索数は 2017 年まで次第に増加しており，特に，Primark は他の企業を追うように検索数が急激に成長している．提案手法はこのようなパターンを非線形モデルを用いて柔軟に表現することで，各企業のトレンドを効果的にとらえることができる．

季節パターン検出．図 1 (b) は提案手法が出力した 2 つの季節パターンを示す．アパレル企業のキーワード群では，ブラックフライデー（S1）や年末年始（S2）に検索数が増大する年単位の周期性を自動的に発見した．季節パターンの検出は長期トレンドの推定精度と密接に関係しており，将来予測の精度を向上させる重要な要素である．提案モデルの特徴は，長期トレンドと季節パターンの相互関係を単一の非線形ダイナミクスとして表現し，高次元テンソルに含まれる潜在的なパターンを抽出することである．

類似パターンに基づくレジーム検出．提案手法は，与えられたテンソルの一方向（たとえば，地域）等から類似した長期トレンド，季節パターンを持つ集合を検出する．本研究では，この集合をローカルグループと呼ぶ．図 1 (c) は実データから得られたローカルグループの例であり，地図中の色は 2006 年から 2008 年にかけて類似パターンを持っていた地域を示す．たとえば，**図 2** (a) に示すように，紫で示されるアメリカ，イギリス等の国々では検索数の推移が類似した傾向を持つ．さらに，本研究では各ローカルグループのメンバとグループ内の特徴を表す時系列パターン（長期トレンド，季節パターン）の構成を要約したものをレジームと定義し，データの傾向の変化に応じてレジームを変化させる．具体的な例として，図 2 (b) はイタリアにおける異なる 2 つの期間のデータであり，H&M の検索数が急激に成長したことにより，傾向が大きく変化している．提案モデルは，このような時系列パターンの変化をとらえるためにローカルグループの構成を柔軟に変化させ，適応的に将来予測を行う．

1.2　本論文の貢献

本論文では，テンソルストリームを効果的に表現する非線形モデル，および，その推定アルゴリズムを提案する．提案手法は次の特長を持つ．

- 非線形モデル：オンライン活動履歴から得られるテンソルストリームの解析に有効な非線形モデルを提案する．提案モデルは，長期トレンドと季節パターンを柔軟に表現し，これらの特徴に基づいてテンソルストリーム中の時間的，地理的な差異をとらえる．
- 特徴自動抽出：非線形モデルの構造を自動的に決定するための符号化スキームを提案し，高度なパラメータチューニングやデータに関する事前知識を必要とせずモデルパラメータを推定する．
- データストリーム処理：提案手法は，増加し続けるテンソルストリームに対し逐次的かつ適応的に非線形モデルを変化させながら重要な時系列パターンをとらえる．そのため，提案手法の計算時間はテンソルストリーム全体の長さに依存せず，高速に将来予測を行う．

2.　関連研究

時系列予測に関する研究は，データマイニング，データベース分野でさかんに取り組まれている [6], [8], [9]．自己回帰モデル（AR），線形動的システム（LDS）は代表的な手法であり，これまでに様々な拡張モデルが提案されている．SARIMA は，季節パターンを表現するための統計モデルであるが，線形モデルに基づいており，複雑な時系列パターンを表現することができない．非線形動的システム [10], [11] に基づく時系列解析では，データのドメイン知識をモデルに適用することで，効果的に時系列パターンする手法がある．しかし，ドメイン知識がつねに有効であるとは限らず，複雑なテンソルストリームの解析においては，自動的に時系列モデルを推定できることが望ましい．深層学習に基づく時系列モデルも注目を集めており，LSTM，GRU 等に代表される再帰型ニューラルネットワーク（Recurrent neural network：RNN）[12] を用いた時系列モデルが多数存在する [13], [14], [15]．深層学習は，高い表現能力を持つモデルを用い，大量のデータから時系列予測に有効な特徴量を抽出することが可能であるが，解析結果がパラメータチューニングに依存する傾向があり，モデ

ル学習には高い計算コストを要する．これらの特徴は，時間経過にともない傾向が変化しうるデータストリームの解析には適さない．

時系列データストリームのための将来予測手法として提案された RegimeCast [16] は，非線形動的システムをデータの傾向に応じて適応的に変化させることで，高速かつ効果的な将来予測を達成した．しかし，RegimeCast は季節パターンを表現するための明示的な機構がなく，オンライン活動データの解析には不十分である．

テンソルデータの解析も活発な研究分野である [17], [18], [19]．ストリーム処理を想定した手法 [20], [21], [22] では，主に次元削減や関係性の抽出による欠損値の補間，情報推薦等を目的としている．一般に，これらの手法は時系列特性を無視しており，将来の振舞いを予測する能力を持たない．テンソル分解を応用し，季節パターンを検出する手法 [23], [24] も存在するが，ストリーム処理を想定していない．多線形動的システム（Multi-Linear Dynamical System：MLDS）は，時系列特性と多線形性を同時に推定するために提案された LDS の拡張である．しかし，オンライン活動履歴では季節に依存するパターンがよく見られるため，多線形性に基づくモデルは不十分である．

本研究の目的は，非線形性を持つ時系列パターン（長期トレンド，季節パターン）を表現するストリーム処理指向の時系列予測モデルを開発することである．

3. 提案モデル

本章では，オンライン活動テンソルストリームのための時系列モデルついて述べる．

3.1 問題定義

表 1 に本研究で使用する記号の定義を示す．本研究では，時間，地域，キーワードで構成される 3 階テンソルを扱う．長さ t_c，地域数 d_l，キーワード数 d_k から成るテンソルを $\mathcal{X} \in \mathbb{R}^{t_c \times d_l \times d_k}$ と表し，テンソルの各要素 x_{tij} は単位期間中の Web 検索数の総数を示すものとする．また，各時刻において，新たなデータが観測されるたびにテンソルの長さ t_c が増加し続ける．本研究の目的は，このように刻々と増大する時系列テンソルデータ，すなわち，テンソルストリーム $\mathcal{X}$ をモデル化し，将来の振舞いを予測し続けることである．

定義 1（推定イベントテンソル） $\mathcal{X}$ をモデル化したときの推定値を $\mathcal{E} \in \mathbb{R}^{t_c \times d_l \times d_k}$ と表す．

続いて，$\mathcal{X}$ のうち，ある時刻 t_s から t_e で切り出される長さ l の部分テンソルを $\mathcal{X}_{t_s:t_e} \in \mathbb{R}^{l \times d_l \times d_k}$ のように表すとき，解析に用いる最新データの区間，および，予測対象とする区間を次のように定義する．

定義 2（カレントウィンドウ） 時刻 t_p から t_c までの区間で与えられる $\mathcal{X}^c = \mathcal{X}_{t_p:t_c}$ を時系列予測に使用する最新のテンソルデータとし，カレントウィンドウと呼ぶ．このとき，カレントウィンドウの長さを $l_c = t_c - t_p$ と定義する．

表 1　記号と定義
Table 1　Symbols and definitions.

記号	定義
d_l, d_k	地域，キーワードの総数
t_c	現在の時刻（テンソルの長さ）
$\mathcal{X}$	テンソルストリーム $\mathcal{X} \in \mathbb{R}^{t_c \times d_l \times d_k}$
$\mathcal{X}_{t_s:t_e}$	時刻 t_s から t_e までの部分テンソル
$\mathcal{X}_{:,i}$	i 番目の時系列シーケンス $\mathcal{X}_{:,i} \in \mathbb{R}^{t_c \times d_k}$
$\mathcal{X}^c$	カレントウィンドウ $\mathcal{X}^c = \mathcal{X}_{t_p:t_c}$ $(l_c = t_c - t_p)$
$\mathcal{X}^f$	予測ウィンドウ $\mathcal{X}^f = \mathcal{X}_{t_s:t_e}$ $(t_s = t_c + l_s, t_e = t_s + l_e)$
k_z, k_v	潜在長期トレンド，潜在季節パターンの総数
$\mathbf{Z}$	長期トレンドの潜在状態 $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$, $\mathbf{z}_i \in \mathbb{R}^{k_z}$
$\mathbf{V}$	季節パターンの潜在状態 $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_t\}$, $\mathbf{v}_i \in \mathbb{R}^{k_v}$
$\mathbf{A}, \mathcal{B}$	非線形動的システム $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\mathcal{B} \in \mathbb{R}^{k \times k \times k}$ $(k = k_z + k_v)$
$\mathbf{W}$	長期トレンドの観測行列 $\mathbf{W} \in \mathbb{R}^{d_k \times k_z}$
$\mathcal{W}$	長期トレンドの観測行列集合 $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_m\}$
$\mathbf{U}$	季節パターンの観測行列 $\mathbf{U} \in \mathbb{R}^{d_k \times k_v}$
$\mathcal{U}$	季節パターンの観測行列集合 $\mathcal{U} = \{\mathbf{U}_1, \ldots, \mathbf{U}_m\}$
p	季節パターンの周期
$\mathbf{S}$	潜在季節パターン $\mathbf{S} \in \mathbb{R}^{p \times k_v}$
$\mathcal{E}$	推定イベントテンソル $\mathcal{E} \in \mathbb{R}^{t_c \times d_l \times d_k}$
m	レジーム内のローカルグループの総数
n	レジームの総数
θ	レジームパラメータ集合 $\theta = \{\mathbf{A}, \mathcal{B}, \mathcal{W}, \mathcal{U}\}$
Θ	モデルパラメータ集合 $\Theta = \{\mathbf{S}, \theta_1, \ldots, \theta_n\}$
$\mathcal{R}$	レジーム割当て $\mathcal{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_n\}$

定義 3（予測ウィンドウ） 時刻 t_c から l_s ステップ先の区間で与えられる $\mathcal{X}^f = \mathcal{X}_{t_s:t_e}$ を予測対象のデータ区間とし，予測ウィンドウと呼ぶ．ここで，$t_s = t_c + l_s$ を満たし，予測ウィンドウの長さが $l_e = t_e - t_s$ のとき，l_e ステップごとに $\mathcal{X}^f$ の値を予測する．

まとめとして，本論文で扱う問題を次のように定義する．

問題 1（l_s ステップ先予測） t_c を最新の時刻とし，長さ l_c のカレントウィンドウ $\mathcal{X}^c = \mathcal{X}_{t_p:t_c}$ $(t_c = t_p + l_c)$ が与えられる間，l_s ステップ先の長さ l_e のテンソル $\mathcal{X}^f = \mathcal{X}_{t_s:t_e}$ $(t_s = t_c + l_s,\ t_e = t_s + l_e)$ を予測し続ける．

3.2 CubeCast モデル

1.1 節で述べたように，オンライン活動データは複雑な時系列パターンで構成され，それらが時間，あるいは地域に依存して異なる特徴を持つ．本章では，これらの特徴を網羅的に表現するための新たな時系列モデルを提案する．具体的には，以下の手順でモデルを構築する．

- 非線形動的システム：多次元時系列シーケンスの中から，より少ない次元数の潜在的な時系列パターン（長期トレンド）を抽出する．
- 季節パターン：非線形動的システムに季節パターンを

とらえるための機構を加える．

- レジーム集合：上記の特徴をまとめたモデルをレジームとし，時系列パターンの特徴に対して適応的に変化させる．また，レジーム内部にローカルグループを定義し，類似パターンの検出を行う．

3.2.1 非線形動的システム

まずはじめに，提案モデルの基本となる非線形動的システムについて述べる．このモデルでは，ある単一の地域における複数キーワードの検索数の推移のような d 次元の時系列データに対して 2 つの潜在アクティビティを仮定する．

- $\mathbf{z}_t$：時刻 t における k_z 次元の潜在アクティビティ
- $\mathbf{e}_t$：時刻 t における d 次元のアクティビティ

潜在アクティビティはデータから推定される未知の情報であり，$\mathbf{e}_t$ は潜在アクティビティから実際に観測されるデータを再現したものである．潜在アクティビティの時間依存性，および，それらと実際のアクティビティとの関係性は次式で表現される．

$$\begin{aligned}\mathbf{z}_{t+1} &= \mathbf{A}\mathbf{z}_t + \mathcal{B}\mathbf{z}_t \otimes \mathbf{z}_t,\\ \mathbf{e}_t &= \mathbf{W}\mathbf{z}_t,\end{aligned} \tag{1}$$

$\otimes$ は 2 つのベクトルの外積を示し，$\mathbf{A} \in \mathbb{R}^{k_z \times k_z}$ は線形射影行列，$\mathcal{B} \in \mathbb{R}^{k_z \times k_z \times k_z}$ 非線形射影テンソルである．連続する 2 つの潜在状態は，これらのパラメータを用いて一時刻前の潜在状態から生成され，様々な事象の時間発展を数理モデルとして表現することができる．一方，$\mathbf{W} \in \mathbb{R}^{d \times k_z}$ は観測行列であり，潜在状態の線形写像によって実際のデータが表現される．非線形動的システムでは，パラメータである $\mathbf{A}$，$\mathcal{B}$，$\mathbf{W}$ と初期値 $\mathbf{z}_0$ を求めることで，d 次元の時系列データに含まれる k_z 次元の重要な時系列パターンを抽出することが可能である．

3.2.2 季節パターンの抽出

オンライン活動データの解析では，季節によって周期的に変動する時系列パターンを抽出することが重要である．通常，このような周期パターンはデータの中でつねに一定に繰り返されると仮定されるが，本研究ではこれらの傾向自身も時間とともに変化すると考える．たとえば，あるキーワードに対する関心が高まっているとき，すなわち，長期トレンドが上昇傾向を示すとき，それにともなって特定の季節に高まる関心も強くなることがある．そのため，長期トレンドと季節パターンの相互関係をモデル化するために式 (1) を拡張する．具体的には，新たに以下の 2 つの潜在アクティビティを仮定する．

- $\mathbf{v}_t \in \mathbb{R}^{k_v}$：時刻 t における潜在季節パターンの強さ．
- $\mathbf{S} \in \mathbb{R}^{p \times k_v}$：潜在季節パターン．

つまり，k_v は潜在的な季節パターンの数を示し，p は周期を示す．これらの要素を考慮した非線形動的システムを次のように定義する．

$$\begin{aligned}\begin{bmatrix}\mathbf{z}_{t+1}\\ \mathbf{v}_{t+1}\end{bmatrix} &= \mathbf{A}\begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix} + \mathcal{B}\begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix} \otimes \begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix},\\ \mathbf{e}_t &= \mathbf{W}\mathbf{z}_t + \mathbf{U}(\mathbf{v}_t \circ \mathbf{S}_{t \bmod p}),\end{aligned} \tag{2}$$

ここで，$\circ$ は 2 つのベクトルの要素積を示し，カッコは 2 つのベクトルの結合を示す．このモデルでは，ある時刻 t における 2 つの潜在状態，$\mathbf{z}$ と $\mathbf{v}$ を結合した $k = k_z + k_v$ 次元のベクトルを潜在状態とし，共通の線形・非線形射影 $\mathbf{A} \in \mathbb{R}^{k \times k}$，および $\mathcal{B} \in \mathbb{R}^{k \times k \times k}$ を用いて時系列特性をモデル化する．また，観測行列 $\mathbf{U} \in \mathbb{R}^{d \times k_v}$ を追加することで，$\mathbf{e}_t$ は 2 つの潜在アクティビティから観測され，ある時刻 t の季節パターンは行列 $\mathbf{S}$ の $(t \ mod \ p)$ 番目の列とベクトル $\mathbf{v}_t$ をかけあわせたものとなる．このモデルによって，長期トレンドの変化と相互に依存し，動的に変化する季節パターンを表現する．

3.2.3 レジーム集合の検出

これまでに述べたモデルは多次元時系列データ，すなわち，単一地域のみを対象として議論した．最後に最も重要な課題として，複数の地域を対象としたテンソルデータのための非線形モデルを提案する．本研究の目的で述べたように，テンソルを多角的に分析し，いくつかの要素に要約することで，重要なパターンを検出することが可能になる．そこで，非線形動的システムで抽出される潜在アクティビティに基づいて，地域方向，時間方向のパターン変化をとらえるためのモデルを提案する．

地域方向の解析では，3 階テンソル $\mathcal{X}$ が与えられたとき，d_l 個の地域を m 個 $(m < d_l)$ のローカルグループに分割する．このとき，各グループが固有の観測行列 $\mathbf{W}_i$，$\mathbf{U}_i$ $(i \in \{1, \ldots, m\})$ を持つことで，すべての地域で共通する潜在ダイナミクスと，地域によって異なるダイナミクスを表現する．つまり，類似した時系列パターンを持つ地域は，類似した潜在アクティビティで構成されていることを意味する．よって，$\mathcal{X}$ の推定値 $\mathcal{E} \in \mathbb{R}^{t_c \times d_l \times d_k}$ は次式で表現される．

$$\begin{aligned}\begin{bmatrix}\mathbf{z}_{t+1}\\ \mathbf{v}_{t+1}\end{bmatrix} &= \mathbf{A}\begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix} + \mathcal{B}\begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix} \otimes \begin{bmatrix}\mathbf{z}_t\\ \mathbf{v}_t\end{bmatrix},\\ \mathbf{e}_{ti} &= \mathbf{W}_j\mathbf{z}_t + \mathbf{U}_j(\mathbf{v}_t \circ \mathbf{S}_{t \bmod p}),\end{aligned} \tag{3}$$

つまり，i 番目の地域は，自身が属する j 番目のグループの観測行列を用いて表現される．

定義 4（レジーム） $\theta = \{\mathbf{A}, \mathcal{B}, \mathcal{W}, \mathcal{U}\}$ を単一のレジームとし，$\mathcal{W}$，および $\mathcal{U}$ は m 種類のローカルグループの潜在アクティビティを表現するための観測行列集合とする．すなわち，$\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_m\}$，$\mathcal{U} = \{\mathbf{U}_1, \ldots, \mathbf{U}_m\}$ と表される．

さらに，レジーム θ で表現されるダイナミクスが時間変化する様子をとらえたい．そのため，テンソルストリームを時間方向に分割し，それぞれのダイナミクスを n 個のレ

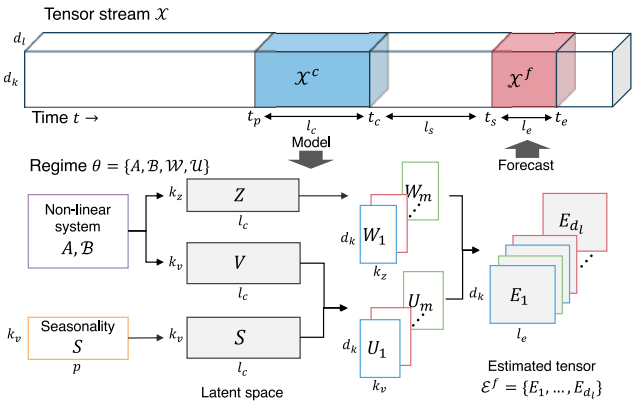


図 3　CubeCast のモデル概要図

Fig. 3　Graphical representation of CubeCast: Given a current tensor $\mathcal{X}^c$, (a) it identifies a regime θ while capturing seasonality with $\mathbf{S}$. (b) It generates latent states $\mathbf{Z}$ and $\mathbf{V}$. (c) It reports l_s-steps ahead values $\mathbf{E}_i$ at the i-th location with projection matrices $\mathbf{W}_j$ and $\mathbf{U}_j$, which capture the j-th location-specific pattern.

ジーム $\{\theta_1, \ldots, \theta_n\}$ を用いて表現し，各レジームが特有の非線形ダイナミクスとローカルグループの構成を持つものとする．したがって，本研究で求めたいモデルパラメータを以下のように定義する．

定義 5（モデルパラメータ集合）　$\Theta = \{\theta_1, \ldots, \theta_n, \mathbf{S}\}$ を n 種類のレジームと季節パターンで構成されるモデルパラメータ集合とする．

また，各レジーム内のローカルグループの割当てを次のように定義する．

定義 6（レジームメンバーシップ）　集合 $\mathcal{R}$ を n 個のレジーム Θ のレジームメンバーシップと定義し，$\mathcal{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_n\}$ と表す．各要素 $\mathbf{r}_i$ は，i 番目のレジームのローカルグループの割当てを示し，各地域の割当て $\mathbf{r}_i = \{r_1, \ldots, r_j, \ldots, r_{d_l}\}$ を d_l 個の整数で表現する $(r_j \in \{1, \ldots, m_i\})$.

提案モデルの全体図を**図 3** に示す．提案モデルは $\mathcal{X}^c$ に最も適したレジーム θ と潜在季節パターン $\mathbf{S}$ 用いて将来予測を行う．具体的には，非線形動的システムを用いて推定された初期値から潜在状態 $\mathbf{Z}$ および $\mathbf{V}$ を予測したいステップまで生成し，それぞれ観測行列を用いて予測値を生成する．このとき，i 番目の地域の推定値 $\mathcal{E}_{:,i}$ は，自信が属するローカルグループの観測行列を用いて生成される．よって，本研究の具体的な問題は，テンソルストリーム $\mathcal{X}$ に対するパラメータ集合 Θ，$\mathcal{R}$ を推定し，現在の時刻 t_c から l_s ステップ先の推定値 $\mathcal{E}^f$ を生成し続けることである．

4. アルゴリズム

本章では，オンライン活動履歴をとらえたテンソルストリームのための将来予測アルゴリズムである CubeCast について述べる．本章の目的は，大規模データストリームから効果的かつ効率的に時系列パターンを検出するために，モデル構造を自動的に決定し，複数のレジームを選択・更新することである．Algorithm 1 は CubeCast の処理の流れを示す．提案手法は，以下の手順でモデルの各要素を更新しながら将来予測を行う．

Algorithm 1 CubeCast $(\mathcal{X}^c, \Theta, \mathcal{R})$

Input: (a) Current tensor $\mathcal{X}^c$
(b) Full parameter set Θ
(c) Regime assignment set $\mathcal{R}$
Output: (a) l_s-steps-ahead future values $\mathcal{E}^f$
(b) Updated full parameter set Θ'
(c) Updated regime assignment set $\mathcal{R}'$

1: /* (I) Estimate a new regime for given data */
2: $\{\theta, \mathbf{r}\} \leftarrow$ RegimeEstimation $(\mathcal{X}^c, \mathbf{S})$;　// $\mathbf{S} \in \Theta$
3: /* (II) Update model set and detect current dynamics */
4: $\{\Theta', \mathcal{R}'\} \leftarrow$ RegimeCompression $(\mathcal{X}^c, \Theta, \mathcal{R}, \theta, \mathbf{r})$;
5: /* (III) Generate future values using a current regime */
6: $\{\theta, \mathbf{r}\} \leftarrow \underset{\theta' \in \Theta', \mathbf{r}' \in \mathcal{R}'}{\arg\min} \|\mathcal{X}^c - f(\theta', \mathbf{r}')\|$　// $f(\cdot,\cdot)$: Equation (3)
7: $\mathcal{E}^f \leftarrow f(\theta, \mathbf{r})$;　// $\mathcal{E}^f = \{e_{tij}\}_{t,i,j=t_s,1,1}^{t_e,d_l,d_k}$,
8: **return** $\{\mathcal{E}^f, \Theta', \mathcal{R}'\}$;

(1) RegimeEstimation：$\mathcal{X}^c$ から新たなレジーム θ を推定する．具体的には，非線形動的システムとローカルグループの検出，すなわち，レジーム割当て $\mathbf{r}$ と観測行列集合 $\mathcal{W}$，$\mathcal{U}$ を推定する．

(2) RegimeCompression：新たに推定したレジームと過去に推定されたレジームを比較し，MDL コストに基づいて，モデル全体をどのように更新するかを決める．また，選択されたレジームに基づいて，季節性 $\mathbf{S}$ を更新する．

(3) 最後に，現在のデータに最も適したレジーム θ を選出し，式 (3) に従って l_s ステップ先の予測値 $\mathcal{E}^f = \{e_{tij}\}_{t,i,j=t_s,1,1}^{t_e,d_l,d_k}$ を計算する．

4.1 特長自動抽出

本研究の目的は，Θ の構造を自動的に決定することである．そこで本研究では，MDL（Minimum description length）[25] に基づく符号化スキームを導入する．MDL は，モデルの複雑さを示すモデルコストと，あるモデルが与えられたときのデータの表現能力を示す符号化コストで構成され，2 つのコストの和が最小となるモデルを最適とする．すなわち，以下の最適化問題を解いて Θ を求める．

$$\Theta = \underset{\Theta'}{\arg\min} < \Theta' > + < \mathcal{X} | \Theta' >, \tag{4}$$

ここで，$< \Theta' >$ はモデルパラメータ集合 Θ' のモデル表現コスト，$< \mathcal{X} | \Theta' >$ は Θ' に対する $\mathcal{X}$ の符号化コストを与えるものとする．提案モデルのモデル表現コストは，入力データの次元数，潜在状態の次元数，および，行列のサイズで決定され，整数のユニバーサル符号長 [26] を $\log^*$ と

表すとき，次のように定義される．

$$< t_c >=\log^*(t_c),\ < d_l >=\log^*(d_l),\ < d_k >=\log^*(d_k).$$
$$< k_z >=\log^*(k_z),\ < k_v >=\log^*(k_v),\ < p >=\log^*(p).$$
$$< \mathbf{S} >= |\mathbf{S}| \cdot (\log(p) + \log(k_v) + c_F) + \log^*(|\mathbf{S}|).$$
$$< \theta >=< k_z > + < \mathbf{A} > + < \mathcal{B} > + < \mathcal{W} > + < \mathcal{U} > .$$

ここで，$|\cdot|$ は与えられた行列内の非ゼロ要素の総数，c_F は実数の符号化に要する浮動小数点コストを示す[*2]．同様に，レジームパラメータのモデルコスト $< \theta >$ は次のように定義される．

$$< \mathbf{A} >= |\mathbf{A}| \cdot (2 \cdot \log(k) + c_F) + \log^*(|\mathbf{A}|),$$
$$< \mathcal{B} >= |\mathcal{B}| \cdot (3 \cdot \log(k) + c_F) + \log^*(|\mathcal{B}|),$$
$$< \mathcal{W} >= \Sigma_{i=1}^{m} |\mathbf{W}_i| \cdot (\log(d_k) + \log(k_z) + c_F) + \log^*(|\mathbf{W}_i|),$$
$$< \mathcal{U} >= \Sigma_{i=1}^{m} |\mathbf{U}_i| \cdot (\log(d_k) + \log(k_v) + c_F) + \log^*(|\mathbf{U}_i|).$$

続いて，Θ に対する $\mathcal{X}$ の符号化コストは，ハフマン符号化 [27] に基づき，平均 μ，分散 σ^2 の正規分布における負の対数尤度で表される．

$$< \mathcal{X}|\Theta >= \sum_{t,i,j=1}^{t_c,d_k,d_l} -\log_2 p_{\mu,\sigma}(x_{tij} - e_{tij}), \tag{5}$$

ここで，$e_{tij} \in \mathcal{E}$ は式 (3) による $x_{tij} \in \mathcal{X}$ の推定値である．まとめとして，$\mathcal{X}$ に対する Θ の総コスト $< \mathcal{X}; \Theta >$ を次のように定義する．

$$\begin{aligned} < \mathcal{X}; \Theta > &=< \Theta > + < \mathcal{X}|\Theta > \\ &=< t_c > + < d_l > + < d_k > + < p > \\ &\quad + < k_v > + < \mathbf{S} > + \sum_{i=1}^{n} < \theta_i > + < \mathcal{X}|\Theta > . \end{aligned} \tag{6}$$

4.2 RegimeEstimation

本節では，式 (6) を最小化するレジーム，およびローカルグループを求めるアルゴリズムとして RegimeEstimation を提案する．提案モデルの基本的な構成要素は (a) 非線形写像 $\mathbf{A}$，$\mathcal{B}$ (b) 観測行列集合 $\mathcal{W}$，$\mathcal{U}$，(c) 潜在季節トレンド $\mathbf{S}$ であるが，それに加えてローカルグループの数と割当てが重要な役割を持つ．これらのパラメータは互いに依存関係を持っているためすべて同時に最適化することは難しい．この問題を解決するため，RegimeEstimation は貪欲法に基づき MDL コストをより小さくするローカルグループとそのパラメータを求めることを目的とする．Algorithm 2 に RegimeEstimation の詳細を示す．全体の流れとして，(1) 与えられたテンソル $\mathcal{X}^c$ に単一のローカルグループを仮定して基本となる非線形動的システムを推定後，(2) MDL コストの減少が止まるまでローカルグループの追加と割当ての更新を繰り返すことで，自動的にモデル構造を決定する．以下ではこれら 2 つの処理について詳細に述べる．

[*2] 本論文では，$c_F = 32$ ビットとする．

Algorithm 2 RegimeEstimation $(\mathcal{X}^c, \mathbf{S})$

Input: Current tensor $\mathcal{X}^c$ and seasonality $\mathbf{S}$
Output: Regime parameter set θ and regime assignment $\mathbf{r}$

$\mathcal{W} = \phi$; $\mathcal{U} = \phi$; $\mathbf{r} = \{r_i = 1 | i = 1, \ldots, d_l\}$;
$\mathcal{W}^* = \phi$; $\mathcal{U}^* = \phi$; // candidate observation matrix set
/* Estimate a regime with a single local activity */
$\{\mathbf{A}, \mathcal{B}, \mathbf{W}, \mathbf{U}\} \leftarrow \underset{\theta' = \{\mathbf{A}', \mathcal{B}', \mathbf{W}', \mathbf{U}'\}}{\arg\min} < \mathcal{X}^c; \mathbf{S}, \theta', \mathbf{r} >$;
Push $\mathbf{W}$ into $\mathcal{W}^*$; Push $\mathbf{U}$ into $\mathcal{U}^*$;
/* Estimate local activities */
while $\mathcal{W}^*$ and $\mathcal{U}^*$ are not empty **do**
  Pop an entry $\mathbf{W}_0$ from $\mathcal{W}^*$; Pop an entry $\mathbf{U}_0$ from $\mathcal{U}^*$;
  $\theta \leftarrow \{\mathbf{A}, \mathcal{B}, \mathcal{W}_F, \mathcal{U}_F\}$; // $\mathcal{W}_F = \mathcal{W} \cup \mathcal{W}^* \cup \{\mathbf{W}_0\}$
  Initialize $\mathbf{r}^*$; Initialize $\mathbf{W}_1, \mathbf{W}_2, \mathbf{U}_1, \mathbf{U}_2$;
  $\theta^* \leftarrow \{\mathbf{A}^*, \mathcal{B}^*, \mathcal{W}_F^*, \mathcal{U}_F^*\}$; // $\mathbf{A}^* = \mathbf{A}$, $\mathcal{B}^* = \mathcal{B}$
  // $\mathcal{W}_F^* = \mathcal{W} \cup \mathcal{W}^* \cup \{\mathbf{W}_1, \mathbf{W}_2\}$, $\mathcal{U}_F^* = \mathcal{U} \cup \mathcal{U}^* \cup \{\mathbf{U}_1, \mathbf{U}_2\}$
  while $< \mathcal{X}^c; \mathbf{S}, \theta^*, \mathbf{r}^* >$ is improved **do**
    Estimate $\mathbf{r}^*$;
    Estimate $\mathbf{W}_1, \mathbf{W}_2, \mathbf{U}_1, \mathbf{U}_2$;
    Estimate $\mathbf{A}^*, \mathcal{B}^*$;
  end while
  if $< \mathcal{X}^c; \mathbf{S}, \theta^*, \mathbf{r}^* >$ is less than $< \mathcal{X}^c; \mathbf{S}, \theta, \mathbf{r} >$ **then**
    Push $\{\mathbf{W}_1, \mathbf{W}_2\}$ into $\mathcal{W}^*$; Push $\{\mathbf{U}_1, \mathbf{U}_2\}$ into $\mathcal{U}^*$;
    $\mathbf{A} \leftarrow \mathbf{A}^*$; $\mathcal{B} \leftarrow \mathcal{B}^*$; $\mathbf{r} \leftarrow \mathbf{r}^*$
  else
    Push $\mathbf{W}_0$ into $\mathcal{W}$; Push $\mathbf{U}_0$ into $\mathcal{U}$;
  end if
end while
return $\{\theta, \mathbf{r}\}$; // $\theta = \{\mathbf{A}, \mathcal{B}, \mathcal{W}, \mathcal{U}\}$

4.2.1 非線形動的システムの推定

最新のテンソル $\mathcal{X}^c$ が与えられたとき，コスト $< \mathcal{X}^c; \mathbf{S}, \theta, \mathbf{r} >$ をより小さくする $\theta = \{\mathbf{A}, \mathcal{B}, \mathcal{W}, \mathcal{U}\}$ を推定する．このとき，$\mathbf{S}$ はレジーム間で共通して使用されるため固定パラメータと考える．今，ローカルグループの数は $m = 1$ であり，レジーム割当て $r_i \in \mathbf{r}$ $(i = 1, \ldots, d_l)$ の値はすべて 1 である．また，各観測行列集合は単一の要素を持ち，$\mathcal{W} = \{\mathbf{W}\}$，$\mathcal{U} = \{\mathbf{U}\}$ と表される．

パラメータの推定では，長期トレンドの潜在状態の数を決定するため，k_z の数を 1 から順に増加させ，コストの減少が止まったときのパラメータを採用する．与えられた k_z に対し，アルゴリズムはまず非線形パラメータ $\mathcal{B}$ の値を 0 に固定し，線形パラメータ $\{\mathbf{A}, \mathbf{W}, \mathbf{U}\} \in \theta$ のみを EM アルゴリズムを用いて推定する．その後，Levenberg-Marquardt (LM) 法 [28] を用いて $\mathcal{B}$ を推定する．本研究では，$\mathcal{B}$ の対角成分 $b_{iii} \in \mathcal{B}$ $(i \in [1, k])$ のみを使用する．非線形動的システムの初期値 $\{\mathbf{z}_0, \mathbf{v}_0\}$ は各パラメータの更新時にともに推定する．

Algorithm 3 RegimeCompression $(\mathcal{X}^c, \Theta, \mathcal{R}, \theta, \mathbf{r})$

Input: (a) Current tensor $\mathcal{X}^c$
(b) Full parameter set Θ and regime assignment set $\mathcal{R}$
(c) Candidate regime θ and regime assignment $\mathbf{r}$
Output: Updated model set Θ^* and regime assignment set $\mathcal{R}^*$

/* Search an optimal regime within Θ */
$\{\theta^*, \mathbf{r}^*\} \leftarrow \underset{\theta' \in \Theta, \mathbf{r}' \in \mathcal{R}}{\arg\min} < \mathcal{X}^c; \mathbf{S}, \theta', \mathbf{r}' >$;
if $< \mathcal{X}^c; \mathbf{S}, \theta, \mathbf{r} >$ is less than $< \mathcal{X}^c; \mathbf{S}, \theta^*, \mathbf{r}^* >$ **then**
  $\Theta^* \leftarrow \Theta \cup \theta$; $\mathcal{R}^* \leftarrow \mathcal{R} \cup \mathbf{r}$;
  $\theta^* \leftarrow \theta$; $\mathbf{r}^* \leftarrow \mathbf{r}$; // Replace an optimal regime with a new regime
else
  $\Theta^* \leftarrow \Theta$; $\mathcal{R}^* \leftarrow \mathcal{R}$;
end if
while $< \mathcal{X}^c; \mathbf{S}, \theta^*, \mathbf{r}^* >$ is improved **do**
  Estimate θ^*; // $\theta^* \in \Theta^*$
  Estimate $\mathbf{S}$; // $\mathbf{S} \in \Theta^*$
end while
return $\{\Theta^*, \mathcal{R}^*\}$;

4.2.2　ローカルグループの推定

続いて，ローカルグループの推定アルゴリズムについて説明する．ローカルグループの割当ては，分割する次元数に応じて組合せが増大し，計算コストが高くなるという問題が生じる．そこで，スタックを用いた効率的なアルゴリズムを使用する．今，$\mathcal{W}^*$ と $\mathcal{U}^*$ を各観測行列の候補を格納するスタックとし，初期状態として，4.2.1 項で求めた行列 $\{\mathbf{W}, \mathbf{U}\}$ が格納される．スタック $\mathcal{W}^*$ と $\mathcal{U}^*$ が空となるまで $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{U}_1, \mathbf{U}_2\}$，$\{\mathbf{A}^*, \mathcal{B}^*\} \in \theta^*$，ローカルグループの割当て $\mathbf{r}^*$ を交互に更新する．r_i^* の更新では，パラメータが与えられたときのコスト $< \mathcal{X}^c_{:,i} | \mathbf{A}, \mathcal{B}, \mathbf{W}_j, \mathbf{U}_j >$ を小さくする $j \in \{1,2\}$ を割り当てる．最後に，分割を試したモデル θ^*，$\mathbf{r}^*$ と分割前のモデル θ，$\mathbf{r}$ とのコストを比較し，前者がコストを小さくする場合，$\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{U}_1, \mathbf{U}_2\}$ を次の分割候補としてスタックに格納する．そうでない場合は，$\mathbf{W}_0, \mathbf{U}_0$ を最適解としてローカルグループの数を増やし $(m = m + 1)$，$\mathbf{r}$ のインデックスを更新してそのグループの分割を停止する．

4.3　RegimeCompression

ここでは，時間経過によるレジームの変化をとらえるアルゴリズムである RegimeCompression について述べる．Algorithm 3 に示すように，RegimeCompression は，MDL コストに基づいてモデル追加の必要性を判断する．具体的には，テンソル $\mathcal{X}^c$ が与えられたとき，過去に検出したレジーム集合 $\{\Theta, \mathcal{R}\}$ の中から選択したレジーム $\theta^*, \mathbf{r}^*$ と，RegimeEstimation で推定したレジーム $\theta, \mathbf{r}$ のコストを比較する．$\theta, \mathbf{r}$ を用いたときのコストが小さければそれらをモデル集合へと追加し，そうでなければ推定したモデルを破棄して過去のモデル $\theta^*, \mathbf{r}^*$ を用いて $\mathcal{X}^c$ を表現する．最後に，ローカルグループの割当て $\mathbf{r}^*$ を固定してレジームパラメータ θ^* と潜在季節パターン $\mathbf{S}$ を交互に更新し，モデル全体を $\mathcal{X}^c$ に最適化させる．

本研究では，提案手法を用いてストリーム処理を行う前に，潜在季節の数 k_v とパラメータ $\mathbf{S}$ を初期化する．与えられたテンソルを，周期に沿って変形し，行列 $\mathbf{X} \in \mathbb{R}^{p \times d}$ を得る．その行列へ独立成分分析（ICA）を適用し，得られた k_v 個の独立成分を $\mathbf{S}$ とする．その後，$\mathbf{S}$ に基づき，RegimeEstimation を適用して θ を求める．この処理を $k_v = 1, 2, 3, \ldots,$ と繰り返し実行し，コスト $< \mathcal{X}; \mathbf{S}, \theta, \mathbf{r} >$ が最小となるような k_v，$\mathbf{S}$ を初期値とする．

定理1　CubeCast の計算量は $O(nd_l d_k)$ である．

証明1　ある時刻 t における式 (2) の計算量は，潜在状態の射影 $O(k^2)$，推定イベントの推定 $O(d_k k_z + d_k k_v)$ の合計となる．RegimeEstimation は，長さ l_c のテンソルが与えられたとき，d_l 個の地域に対してローカルグループの 2 分割を繰り返し，最適なグループ数 m とモデルパラメータの推定を行う．よって，繰り返し回数を $\#iter$ とすると，この最適化に必要な計算コストは $O(\#iter \cdot l_c d_l (k^2 + d_k k_z + d_k k_v))$ である．RegimeCompression は過去のレジームから現在のデータに対して最適なものを選択するために $O(n l_c d_l (k^2 + d_k k_z + d_k k_v))$ の計算コストを要する．ここで，繰り返し回数 $\#iter$，ウィンドウサイズ l_c，および潜在状態の数 k, k_z, k_v は非常に小さい定数であるため，CubeCast 全体の計算量は $O(nd_l d_k)$ である．

5.　評価実験

本研究では，提案手法の性能を評価するため，次の問いに関する実験を行った．

(1) 時系列モデリングに対する提案手法の有効性
(2) 提案手法の時系列予測精度
(3) 提案手法の計算コスト

実験には，Intel Xeon W-2123 3.6 GHz quad core CPU，128 GB のメモリを搭載した Linux マシンを使用した．データセットは，GoogleTrends [*3] から収集した，指定したキーワードに対する週ごとの Google 検索数の推移を示す時系列データである．**表 2** に示す各キーワードセットに対し，2004 年 1 月から 2018 年 3 月までの期間，GDP 上位 50 カ国の時系列データを収集し，3 階のテンソルデータとした．また，得られたデータは各次元ごとに標準化（z-normalization）して使用した．以下に示すように，比較手法は代表的な時系列モデルから選出した．

- RegimeCast [16]：非線形動的システムに基づく将来予測アルゴリズム．提案論文に従い，潜在状態の数 $k = 4$，モデルの階層数 $h = 2$，モデル推定の閾値 $\epsilon = 0.5 \cdot \|\mathcal{X}_c\|$ とした．

[*3] https://trends.google.com/trends/

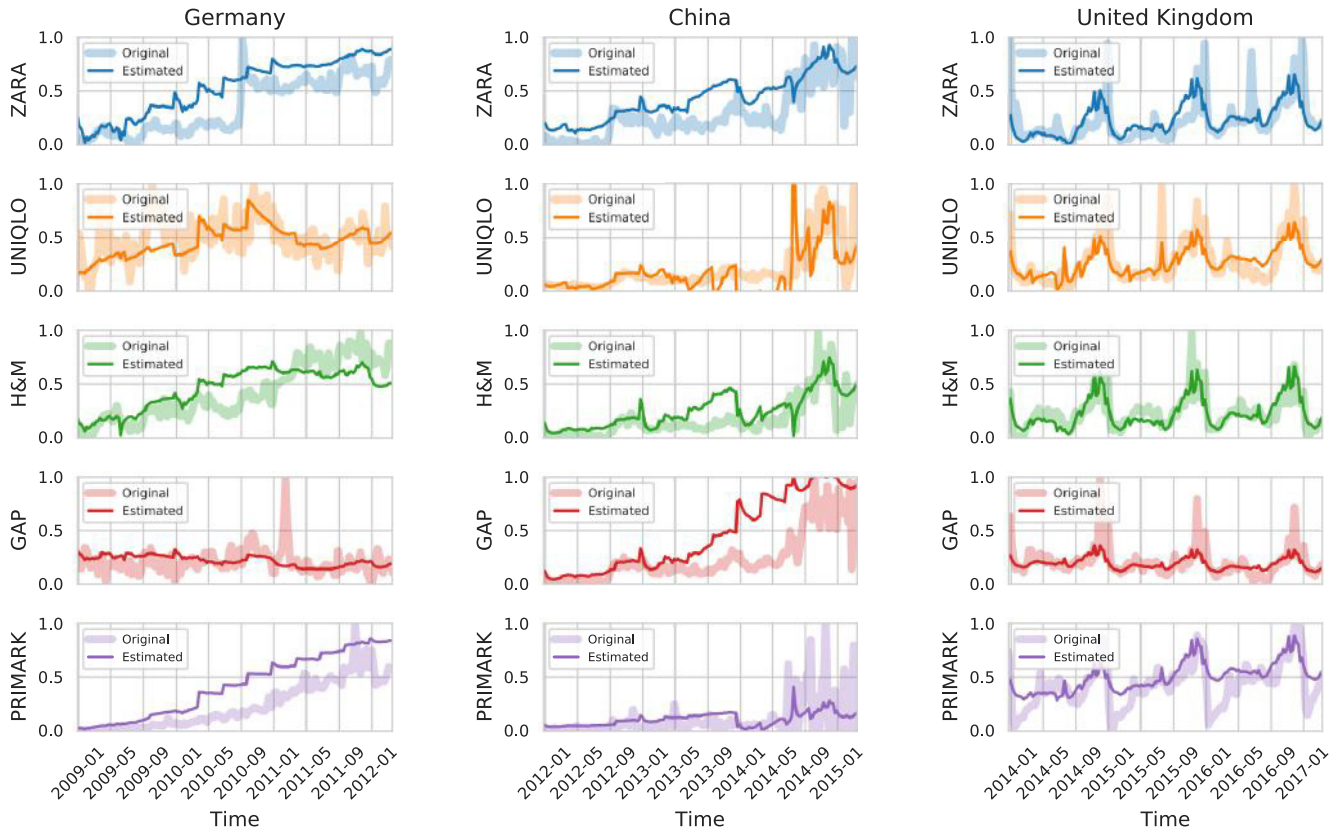


図 4　提案手法を用いた時系列モデリングの例

Fig. 4　Fitting results of CubeCast for five apparel companies on GoogleTrends. CubeCast incrementally and automatically identifies sudden changes in dynamical patterns including the latent trends, seasonality and structure of groups of similar countries.

表 2　データセットの概要

Table 2　Dataset description.

ID	Dataset	Query
#1	Apparel	zara, uniqlo, h&m, gap, primark
#2	Chatapps	facebook, LINE, slack, snapchat, twitter, telegram, viber, whatsapp
#3	Hobby	soccer, baseball, basketball, running, yoga, crafts
#4	LinuxOS	debian, ubuntu, centos, redhat, fedora, opensuse, steamos, raspbian, kubuntu
#5	PythonLib	numpy, scipy, sklearn, matplotlib, plotly, tensorflow
#6	Shoes	booties, flats, heels, loafers, pumps, sandals, sneakers

- SARIMA [29]：季節変動パターンを考慮した自己回帰モデル．AIC 基準を用いて $\{1, 2, 4, 8\}$ の中から最適なパラメータを選択した．
- MLDS [30]：多線形性を考慮した状態空間モデル．本研究では，国，キーワード方向のランクをそれぞれ $\{2, 4\}$，$\{4, 8\}$ と変化させた．
- LSTM/GRU [12]：各モデルにおいて，50 ユニットの RNN を 2 層使用した．ユニット数は，各層のユニット数を $10, 20, \ldots, 100$ と変化させたとき，6 つのテンソルストリームの開始 2 年間のうち，10%の検証データの平均損失が最良となるものを選択した．ネットワークの学習では，中間層で 0.5%のドロップアウトを適用し，Adam [31] を用いて最適化した．

5.1　時系列モデリングに対する有効性

本節では，オンライン活動テンソルストリームに対する CubeCast の表現能力について検証する．図 1，図 2 に示したように，提案手法は大規模テンソルストリームから，時系列パターンの時間変化，地域間の差異を同時にとらえながら効果的に将来予測を行うことが可能である．**図 4** は，同様のデータセットに対する提案手法のモデル推定結果を示している．実験では，カレントウィンドウの長さを 104 ステップ（2 年）とし，13 ステップ（四半期）ごとに Algorithm 1 を実行した．また，季節パターンの周期を 52 ステップ（1 年）とし，$\mathbf{S}$ は 2004 年から 2006 年のデータを用いて初期化した．ここでは，3 つの時期におけるそれぞれの国が特徴的な時系列パターンを持つ中，提案手法は

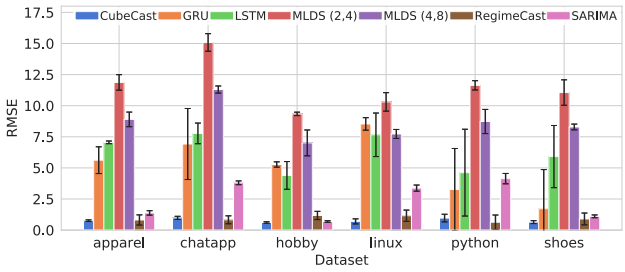


図 5　各予測ウィンドウの平均予測精度の比較

Fig. 5　Average forecasting accuracy of CubeCast: our method is consistently superior to its competitors for all datasets (lower is better).

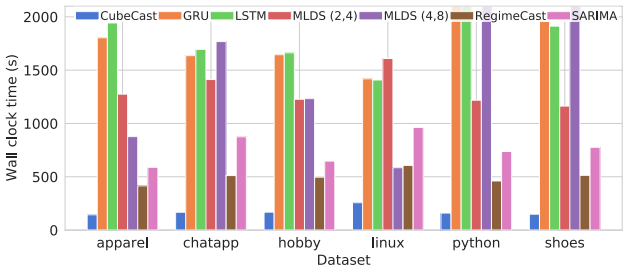


図 6　各ウィンドウでの平均計算時間の比較

Fig. 6　Average wall clock time on GoogleTrends: CubeCast can quickly provide a forecast while detecting regime shifts and important patterns (lower is better).

すべての傾向をうまくとらえることに成功している．特徴的な傾向として，2009 年から H&M がドイツに参入したことによる Web 検索数の大幅な上昇が見られる．2014 年には，中国で各キーワードの検索数が急上昇しており，他国では季節パターンの成長が見られた．提案手法を用いることで，こうした様々な傾向の変化に適応し，それらの特徴をレジームとして抽出しながら将来予測を行うことに成功した．

5.2　予測精度の比較

次に，テンソルストリームの将来予測に対する提案手法の性能を評価する．各比較手法について，前節と同様の実験設定で将来予測を行った．**図 5** は，各予測ウィンドウに対する平均自乗誤差（RMSE）の比較である．線形モデルある SARIMA，MLDS はテンソルストリームに含まれる複雑なトレンドをとらえることができず，長期予測には適さない．RegimeCast は，長期トレンドを表現する非線形動的システム（式 (1)）のみを用いた手法であり，季節パターンを表現することができない．提案モデルは式 (1) を拡張し，テンソルストリームに含まれる季節パターンをとらえることにより，高い予測精度を示した．一方，LSTM，GRU は提案モデルと比べてより表現能力の高い非線形構造を持つ時系列モデルであるが，各カレントウィンドウに対して，汎化性能を持ったモデルを推定することが困難であった．結果として，長期予測に対して有効な時系列パターンをとらえることができない．提案手法は，最新のデータに対してモデルの複雑さを自動的に変化させ，効果的な長期予測に成功した．

5.3　計算時間の比較

最後に，提案手法の計算時間について述べる．**図 6** は，各データセットにおいてリアルタイム予測を行ったとき，各ウィンドウで要した計算時間の平均であり，提案手法は比較手法に対して高速に動作することを示す．

図 7 は，提案手法に与えるテンソルの各次元（時系列の長さ，国の数）を変化させたときの計算時間の変化を示す．貪欲法に基づく手法でテンソルをローカルグループへ分割することにより，提案アルゴリズムの計算コストは入力データの国の数に対して線形である．以上のように，提案手法はテンソルストリームのリアルタイム解析に適した性能を有する．

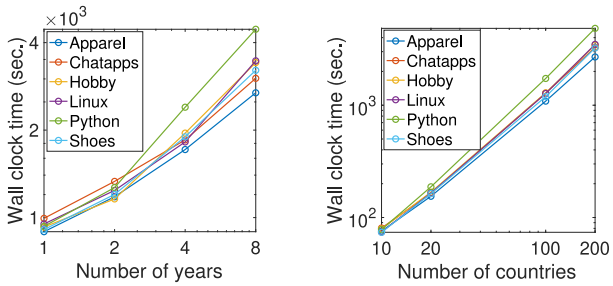


図 7　テンソルサイズに対する提案手法の計算時間

Fig. 7　Average wall clock time vs. tensor stream size, i.e., duration (t_c) and number of countries (d_t). CubeCast scales linearly with respect to the time and target mode for division into several groups.

6.　むすび

本論文では，大規模テンソルストリームのための非線形モデルである CubeCast を提案した．提案モデルは，長期トレンドと季節パターンの時間変化を同一空間上に表現する非線形動的システムを用いて時系列予測に有用な時系列パターンをとらえることができる．また，検出された時系列パターンに基づく地域のグループ化により，柔軟な動的システムを得ることができる．実データを用いて提案手法の性能を評価し，既存の時系列モデルと比較して予測精度，計算時間ともに改善することに成功した．

謝辞　本研究の一部は JSPS 科研費，JP17H04681，JP18H03245，JP19J11125，JP20H00585，JST さきがけ JPMJPR1659，JST 未来社会創造事業 JPMJMI19B3，総務省 SCOPE 192107004，ERCA 環境研究総合推進費 JPMEERF20201R02 の助成を受けたものです．

参考文献

[1] Matsubara, Y. and Sakurai, Y.: Dynamic Modeling and Forecasting of Time-Evolving Data Streams, *KDD*, pp.458–468 (2019).

[2] Hooi, B., Liu, S., Smailagic, A. and Faloutsos, C.: BeatLex: Summarizing and Forecasting Time Series with Patterns, *ECML PKDD*, pp.3–19, Springer (2017).

[3] Matsubara, Y., Sakurai, Y., Faloutsos, C., Iwata, T. and Yoshikawa, M.: Fast mining and forecasting of complex time-stamped events, *KDD*, pp.271–279 (2012).

[4] Kiciman, E. and Richardson, M.: Towards Decision Support and Goal Achievement: Identifying Action-Outcome Relationships From Social Media, *KDD*, pp.547–556 (2015).

[5] Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M.: Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions, *Future Gener. Comput. Syst.*, Vol.29, No.7, pp.1645–1660 (2013).

[6] Morales, G.D.F., Bifet, A., Khan, L., Gama, J. and Fan, W.: IoT Big Data Stream Mining, *KDD, Tutorial*, pp.2119–2120 (2016).

[7] Kawabata, K., Matsubara, Y., Honda, T. and Sakurai, Y.: Non-Linear Mining of Social Activities in Tensor Streams, *KDD*, pp.2093–2102, Association for Computing Machinery (2020).

[8] Sakurai, Y., Matsubara, Y. and Faloutsos, C.: Mining and Forecasting of Big Time-series Data, *SIGMOD, Tutorial*, pp.919–922 (2015).

[9] Sakurai, Y., Matsubara, Y. and Faloutsos, C.: Mining Big Time-series Data on the Web, *WWW, Tutorial*, pp.1029–1032 (2016).

[10] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: Non-Linear Mining of Competing Local Activities, *WWW* (2016).

[11] Thinh, M.D., Yasuko, M. and Yasushi, S.: Real-time Forecasting of Non-linear Competing Online Activities, 情報処理学会論文誌データベース（TOD），Vol.13, No.2 (2020).

[12] Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y.: Recurrent Neural Networks for Multivariate Time Series with Missing Values, *Sci. Rep.*, Vol.8, No.1, p.6085 (2018).

[13] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R. , Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, Vol.29, No.6, pp.82–97 (2012).

[14] Qin, Y., Song, D., Cheng, H., Cheng, W., Jiang, G. and Cottrell, G.W.: A Dual-stage Attention-based Recurrent Neural Network for Time Series Prediction, *IJCAI*, pp.2627–2633, AAAI Press (2017).

[15] Yao, S., Hu, S., Zhao, Y., Zhang, A. and Abdelzaher, T.: DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing, *WWW*, pp.351–360 (2017).

[16] Matsubara, Y. and Sakurai, Y.: Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences, *KDD*, pp.1045–1054 (2016).

[17] Kolda, T.G. and Bader, B.W.: Tensor Decompositions and Applications, *SIAM Review*, Vol.51, No.3, pp.455–500 (2009).

[18] Ye, J., Sun, L., Du, B., Fu, Y., Tong, X. and Xiong, H.: Co-Prediction of Multiple Transportation Demands Based on Deep Spatio-Temporal Neural Network, *SIGKDD*, pp.305–313 (2019).

[19] Cai, Y., Tong, H., Fan, W., Ji, P. and He, Q.: Facets: Fast Comprehensive Mining of Coevolving High-order Time Series, *KDD*, pp.79–88 (2015).

[20] Sun, J., Tao, D. and Faloutsos, C.: Beyond Streams and Graphs: Dynamic Tensor Analysis, *KDD*, pp.374–383 (2006).

[21] Zhou, S., Vinh, N.X., Bailey, J., Jia, Y. and Davidson, I.: Accelerating Online CP Decompositions for Higher Order Tensors, *KDD*, pp.1375–1384 (2016).

[22] Song, Q., Huang, X., Ge, H., Caverlee, J. and Hu, X.: Multi-Aspect Streaming Tensor Completion, *KDD*, pp.435–443 (2017).

[23] Takahashi, T., Hooi, B. and Faloutsos, C.: AutoCyclone: Automatic Mining of Cyclic Online Activities with Robust Tensor Factorization, *WWW*, pp.213–221 (2017).

[24] Song, H.A., Hooi, B., Jereminov, M., Pandey, A., Pileggi, L.T. and Faloutsos, C.: PowerCast: Mining and Forecasting Power Grid Sequences, *ECML/PKDD* (2017).

[25] Grünwald, P.D., Myung, I.J. and Pitt, M.A.: *Advances in minimum description length: Theory and applications*, MIT press (2005).

[26] Rissanen, J.: A universal prior for integers and estimation by minimum description length, *The Annals of Statistics*, pp.416–431 (1983).

[27] Rissanen, J.: Modeling by shortest data description, *Automatia*, Vol.14, pp.465–471 (1978).

[28] Moré, J.J.: The Levenberg-Marquardt algorithm: Implementation and theory, *Numerical Analysis*, pp.105–116 (1978).

[29] Durbin, J. and Koopman, S.J.: *Time Series Analysis by State Space Methods*, Oxford University Press, 2 edition (2012).

[30] Rogers, M., Li, L. and Russell, S.J.: Multilinear Dynamical Systems for Tensor Time Series, *NIPS*, pp.2634–2642 (2013).

[31] Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol.abs/1412.6980 (2015).

川畑 光希

2016 年熊本大学工学部情報電気電子工学科卒業．2018 年同大学大学院博士前期課程修了．2021 年大阪大学大学院情報科学研究科情報システム工学専攻博士後期課程修了．博士（情報科学）．2019 年大阪大学情報科学研究科日本学術振興会特別研究員（DC2）．2021 年 4 月より大阪大学産業科学研究所助教．DEIM Forum 2016 最優秀論文賞，WebDB Forum 2018 最優秀論文賞，学生奨励賞，企業賞，2019 年度コンピュータサイエンス領域奨励賞（データベースシステム），等受賞．データマイニング，データストリーム処理の研究に従事．日本データベース学会会員．

松原 靖子

2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学大学院博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．2012 年 NTT コミュニケーション科学基礎研究所 RA．2013 年日本学術振興会特別研究員（PD）．2014 年熊本大学大学院自然科学研究科助教．この間，カーネギーメロン大学客員研究員．2016 年国立研究開発法人科学技術振興機構さきがけ研究者．2019 年 5 月より大阪大学産業科学研究所准教授．2016 年度日本データベース学会上林奨励賞，情報処理学会山下記念研究賞．2018 年度 IPSJ/ACM Award for Early Career Contributions to Global Research，ACM Recognition of Service Award，2020 年度マイクロソフト情報学研究賞，電気通信普及財団第 36 回テレコムシステム技術賞等受賞．2018～2019 年度日本データベース学会理事．大規模時系列データマイニングに関する研究に従事．ACM，電子情報通信学会，日本データベース学会各会員．

本田 崇人

2015 年熊本大学工学部情報電気電子工学科卒業．2017 年同大学院博士前期課程修了．2020 年同大学院博士後期課程修了．博士（工学）．2017 年より日本学術振興会特別研究員（DC1）．2020 年より大阪大学産業科学研究所特任助教．現在，株式会社 JDSC データサイエンティスト．WebDB Forum 2018 最優秀論文賞，DEIM Forum 2020 優秀論文賞，2020 年度情報処理学会コンピュータサイエンス領域奨励賞（データベースシステム）等受賞．大規模時系列テンソルマイニングの研究に従事．日本データベース学会会員．

櫻井 保志

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013 年熊本大学大学院自然科学研究科教授．2019 年より大阪大学産業科学研究所産業科学 AI センターセンター長・教授．本会平成 18 年度長尾真記念特別賞，平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010 年）等受賞．データマイニング，データストリーム処理，センサーデータ処理，Web 情報解析技術の研究に従事．ACM，IEEE，電子情報通信学会，日本データベース学会各会員．

（担当編集委員　田中　剛）