The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

スパース動的モード分解におけるモード選択の統計的評価

Statistical Evaluation for Mode Selection in Sparsity-promoting Dynamic Mode Decomposition

平岡 将史 *1　河原 吉伸 1,2　鷲尾 隆 1
Masashi Hiraoka[1]　Yoshinobu Kawahara[1,2]　Takashi Washio[1]

[1]大阪大学 産業科学研究所
[1]The Institute of Scientific and Industrial Research, Osaka University

[2]理化学研究所 革新知能統合研究センター
[2]Center for Advanced Intelligence Project, RIKEN

Sparsity-promoting dynamic mode decomposition (SP-DMD) is a data-driven method for estimating a modal representation of a nonlinear dynamical system, where the modes are selected via l_1-regularization depending on the tradeoff between the quality of the representation and the number of the modes. However, the way to statistically evaluate modes selected by SP-DMD is not established. If statistical evaluation is not performed, we may not specify issues caused by different reasons such as noise and overfitting. In this paper, we propose a method to statistically evaluate modes selected by SP-DMD. We develop the method based on the combination of bootstrap and SP-DMD.

1. はじめに

近年，データ駆動型の解析アルゴリズムが注目を集めている．高次元データ収集の容易化とコンピュータの進化により，データ駆動型のアルゴリズムにより複雑なダイナミクスの時間，空間的特徴を取り出し構造推定や未来予測を行うことができるようになった．動的モード分解 (Dynamic Mode Decomposition,DMD) は，行列の特異値分解，固有値分解を用いて時系列データを空間的な情報を持つ動的モードと各動的モードの減衰 (増幅) 率，各動的モードの振幅に分解する．DMD は流体解析の分野で提案されてから様々な分野に応用されており，またその拡張としてパラメータの事前分布を導入したベイズ的動的モード分解 (Bayesian Dynamic Mode Decomposition)[1] やスパースなモデルの推定を行うスパース動的モード分解 (Sparsity-promoting Dynamic Mode Decomposition,SP-DMD)[2] がある．SP-DMD では，DMD のアルゴリズムに $L1$ 正則化項を加えることによって，DMD で求めた動的モードからデータへの寄与度が高い動的モードを選択し，ダイナミクスのモデルを推定する．

しかし SP-DMD では，ダイナミクスをよりスパースに推定するアルゴリズムは示されているが，推定したモデルの信頼性に関する考察はない．モデル推定においては推定したモデルにどの程度の信頼性があるのかを統計的に評価することは不可欠である．もしそうした評価が行われなければ，観測ノイズの影響や過学習といった問題を発見することができない．

そこで本研究では，データの定常性を仮定しサブシーケンスを用いたブートストラップ法と SP-DMD を組み合わせることで SP-DMD により推定したモデルの統計的評価を行う手法を提案する．サブシーケンスに対してブートストラップ法における繰り返しごとに SP-DMD による解析を行うのではなく，動的モードの選択によるモデル推定の部分のみを繰り返す手法によって全データから推定したモデルの不確かさを統計的に評価する．

連絡先: 平岡将史，大阪大学産業科学研究所，
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
m.hiraoka@ar.sanken.osaka-u.ac.jp

本稿の構成は以下の通りである．2 章では，DMD とその拡張である SP-DMD のアルゴリズムを示す．3 章では，3.1 節で一般的なブートストラップ法について述べる．3.2 節で SP-DMD にブートストラップ法が適用できること，ブートストラップ標本としてサブシーケンスを使用することが妥当であることを示す．3.3 節で提案手法について述べる．4 章で SP-DMD によるモード選択に提案手法によって一定の示唆が得られたことを示し，5 章でまとめと課題について述べる．

2. スパース動的モード分解

DMD では時系列データを空間的な情報を持つ動的モードと各動的モードの減衰 (増幅) 率，各動的モードの振幅に分解する．本章では DMD とその拡張である SP-DMD のアルゴリズムについて説明する [2]．DMD では，時系列データ $\Psi = \{\psi_0 \quad \psi_1 \quad \cdots \quad \psi_N\}$ から次のデータ行列を生成する．

$$\Psi_0 = \{\psi_0 \quad \psi_1 \quad \cdots \quad \psi_{N-1}\} \in \mathbb{C}^{M\times N} \tag{1}$$

$$\Psi_1 = \{\psi_1 \quad \psi_2 \quad \cdots \quad \psi_N\} \in \mathbb{C}^{M\times N} \tag{2}$$

$\Psi_1 = A\Psi_0$ を満たすような行列 $A \in \mathbb{C}^{M\times M}$ を考え，Ψ_0 の特異値分解を計算する．

$$\Psi_0 = U\Sigma V^* \tag{3}$$

ただし，$\Sigma \in \mathbb{C}^{r\times r}$ は特異値を対角成分に持つ対角行列，$U \in \mathbb{C}^{M\times r}, V \in \mathbb{C}^{r\times N}$ である．

$$A \approx UF_{dmd}U^* \tag{4}$$

を満たす $F_{dmd} \in \mathbb{C}^{r\times r}$ を考えることで，次元削減した線形空間で A を考えることができる．F_{dmd} は

$$\min_A \|\Psi_1 - A\Psi_0\|_F^2 \tag{5}$$

に $\Psi_0 = U\Sigma V^*$ と $A = UF_{dmd}U^*$ を代入することによって求められる．

$$F_{dmd} = U^*\Psi_1 V\Sigma^{-1} \tag{6}$$

U の列ベクトルで張られる線形空間において，即ち $\psi_t \approx Ux_t$ となる線形空間においてデータの時間変化は

$$x_{t+1} = F_{dmd} x_t \tag{7}$$

と表される．以下の通り F_{dmd} の固有値分解を計算する．

$$F_{dmd} = \underbrace{\begin{bmatrix} y_1 & \cdots & y_r \end{bmatrix}}_{Y} \underbrace{\begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_r \end{bmatrix}}_{D_\mu} \underbrace{\begin{bmatrix} z_1^* \\ \vdots \\ z_r^* \end{bmatrix}}_{Z^*} \tag{8}$$

ただし，$\{y_1, \ldots, y_r\}$ は F_{dmd} の固有値 $\{\mu_1, \ldots, \mu_r\}$ に対応する固有ベクトル，$\{z_1, \ldots, z_r\}$ は F_{dmd}^* の固有値 $\{\overline{\mu}_1, \ldots, \overline{\mu}_r\}$ に対応する固有ベクトルである．(7)，(8) よりデータは初期値 x_0 を用いて

$$x_t = Y D_\mu^t Z^* x_0 = \sum_{i=1}^{r} y_i \mu_i^t z_i^* x_0 = \sum_{i=1}^{r} y_i \mu_i^t \alpha_i \tag{9}$$

と書くことができる．ただし，$\alpha_i = z_i^* x_0$．α はデータの初期値を反映した値であるが，各動的モードの振幅を表した値と解釈することもできる．U を左側からかけることにより，

$$\psi_t \approx Ux_t = \sum_{i=1}^{r} \phi_i \mu_i^t \alpha_i \tag{10}$$

ただし，$\phi_i = Uy_i$．この両辺のフロベニウスノルムの差の最小化問題を解いて最終的な α を決定する．即ち，上式を行列形式で表した

$$\begin{aligned}
\Psi_0 &= \Phi D_\alpha V_{and} \qquad (11)\\
\Phi &:= \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_r \end{bmatrix},\\
D_\alpha &:= \begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_r \end{bmatrix},\\
V_{and} &:= \begin{bmatrix} 1 & \mu_1 & \cdots & \mu_1^{N-1} \\ 1 & \mu_2 & \cdots & \mu_2^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mu_r & \cdots & \mu_r^{N-1} \end{bmatrix}.
\end{aligned}$$

を用いて表した下式の最小化問題を解いて最終的な α を決定する．

$$\min_{\alpha} \quad \|\Psi_0 - \Phi D_\alpha V_{and}\|_F^2 \tag{12}$$

$\Psi_0 = U\Sigma V^*$ と $\phi_i = Uy_i$ を用いれば，この最小化問題は特異値分解による次元削減を利用した形に帰着できる．

$$\min_{\alpha} \quad J(\alpha) := \|\Sigma V^* - Y D_\alpha V_{and}\|_F^2 \tag{13}$$

この最小化問題の目的関数 $J(\alpha)$ は下式と等価である．詳細な証明については [2] を参照されたい．

$$\begin{aligned}
J(\alpha) &= \alpha^* P \alpha - q^* \alpha - \alpha^* q + s \qquad (14)\\
P &:= (Y^* Y) \circ (\overline{V_{and} V_{and}^*}),\\
q &:= \overline{diag\,(V_{and}\,(\Sigma V^*)^*\, Y)},\\
s &:= trace(\Sigma^* \Sigma)
\end{aligned}$$

ただし，$\circ$ は二つの行列の要素ごとの掛け算を表し，M^* は行列 (ベクトル)M を共役転置した行列 (ベクトル)，$\overline{M}$ は行列 (ベクトル)M の複素共役な行列 (ベクトル) を表す．また，ベクトル M に対して $diag(M)$ はベクトル M の要素を対角成分に並べた対角行列を表し，行列 M に対して $diag(M)$ は行列 M の対角成分を抜き出してできるベクトルを表す．

最終的に最小化問題の答え α_{dmd} は (14) 式を最小化する α で，

$$\alpha_{dmd} \quad = \quad P^{-1} q \tag{15}$$

となる．

DMD では動的モードは r 個計算されるが，全ての動的モードが同等にデータに寄与しているわけではない．そこでデータへの寄与が小さな動的モードを取り除き，スパースなモデル推定を行うために提案された拡張が SP-DMD である．DMD のアルゴリズムに $L1$ 正則化項を加えることにより，ダイナミクスのモデル選択をより簡単に行うことができる．具体的には動的モードの係数を決定する (13) 式に $L1$ 正則化項を加えた (16) 式を解いて α を決定する．

$$\min_{\alpha} \quad J(\alpha) + \gamma \sum_{i=1}^{r} |\alpha_i| \tag{16}$$

3. 提案手法

SP-DMD によって，ダイナミクスのモデル推定を容易に行うことが可能になった．しかしモデル推定に不可欠である，推定したモデルの統計的な評価を行う手法は考察されていない．そこで本研究では，サブシーケンスを用いたブートストラップ法により各動的モードの係数 α_i を調べ，SP-DMD によるモデル推定の統計的評価を行う手法を提案する．

3.1 ブートストラップ法

ブートストラップ法は，リサンプリング法と呼ばれる統計学の手法の一種である．複雑な理論を知らずとも，コンピュータによる反復計算によって統計量の偏りや分散，あるいは分布の推定を行うことができる．そのアルゴリズムは，観測データ

$$D = \{d_1, d_2, \cdots, d_N\} \tag{17}$$

に対して $\{1, 2, \cdots, n\}$ から等確率でランダムに整数を選ぶ．これを N 回繰り返して得られた整数列を $\{i_1, i_2, \cdots, i_N\}$ とする．ある整数が複数回選ばれてもよい．得られた整数列の要素を添え字とする要素を観測データから取り出してできるデータ集合 $\{d_{i_1}, d_{i_2}, \cdots, d_{i_N}\}$ を，

$$D^* = \{d_{i_1}, d_{i_2}, \cdots, d_{i_N}\} \tag{18}$$

とする．B 回繰り返して得られた B 個のデータを $\left\{D^{*1}, D^{*2}, \cdots, D^{*B}\right\}$ とする．B 個のデータからそれぞれ統計量を計算する．

ブートストラップ法の基本的なアイデアは，こうして得られた統計量のばらつきが，母集団から幾度もデータをサンプリングして統計量を計算した場合のばらつきと近いものになっているというものである．母集団の分布を仮定できない場合や複雑な統計量の分布の推定が必要な場合に，理論ではなくコンピュータによる反復計算によってそれらを近似できることがブートストラップ法の利点である．

一方でブートストラップ法はあらゆる問題に適用できるわけではない [3]．多くの統計的推測の問題は，未知の分布関数 F

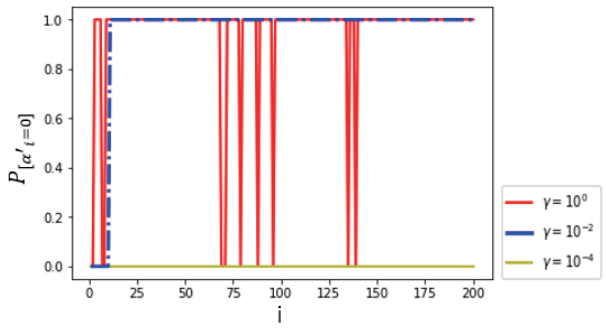

図 1: 全データに対して SP-DMD による解析を行った結果

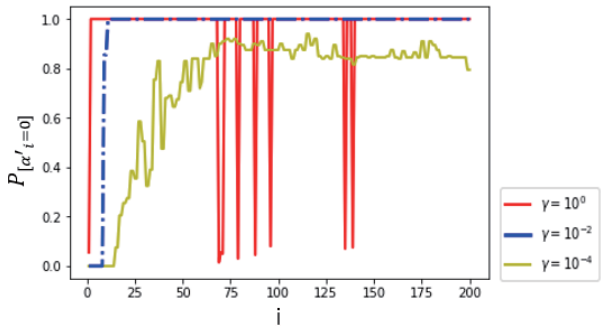

図 2: $s = 100$ の場合の提案手法の実行結果

に従う互いに独立な確率変数 $Y_1, \cdots, Y_n$ に基づいて，F に依存するパラメータ $\theta = \theta(F)$ を推測する問題として定式化できる．ここで，θ が F のみ，即ち未知の母集団のみに従うパラメータであることが重要な点である．例えば，測定誤差を伴う回帰係数の偏りなどはこれを満たさない．ブートストラップ法がうまく機能するための 1 つの十分条件は，パラメータ $\theta = \theta(F)$ が，F の連続的な変化に対して連続的に変動することである．すなわち F に近い分布関数 $F' = (1-\epsilon)F + \epsilon G$ を考えたとき，パラメータ $\theta(F')$ も $\theta(F)$ に近く，$\theta(F') = \theta(F) + o(\epsilon)$ が成り立つことである．ここで，$o(\epsilon)$ は，ϵ が 0 に近づくとき 0 に近づく量を表している．しかし，この条件は分散の推定の場合などには必ずしも必要でない．

いま $\theta = \theta(F)$ の推定量を $\hat{\theta} = \hat{\theta}(Y_1, \cdots, Y_n)$ とする．また F_n を，下式で定義される経験分布関数とする

$$F_n = \frac{1}{n}\sum_{j=1}^{n} \delta(Y_j \leq y) \tag{19}$$

実際のデータ解析においては，$\hat{\theta}$ は $\hat{\theta} = \theta(F_n)$ が使われることが多い．

さて，例えば推定量 $\hat{\theta}$ を用いて θ を推定する際には，推定量 $\hat{\theta}$ の良さを評価しなければならない．また θ に関する仮説を検証する必要もしばしば起こる．このような問題は適当な関数 $T(\hat{\theta}, F) = t(\hat{\theta}, \theta(F))$ に対して期待値

$$H(F) = E_F[T(\hat{\theta}(Y_1, \cdots, Y_n), F)] \tag{20}$$

を計算する問題に帰着できる．ただし，E_F は分布関数 F の下での期待値を表し，また関数 $T(\cdot, \cdot)$ は問題に応じて決められるものである．

ブートストラップ法は，与えられた問題が (20) によって定式化できる場合，この $H(F)$ を推定するための一般的な手法である．

また，通常経験分布関数 F_n が真の分布関数 F の良い近似となっているためにブートストラップ法が多くの局面で機能することも補足として付け加える．これは即ち，ある程度大きな標本数が必要であることを意味する．

3.2 本研究におけるブートストラップ法

3.1 節でブートストラップ法が適用可能な問題について議論したが，本節では SP-DMD に対してブートストラップ法が適用可能か議論する．

動的モードの振幅 α がダイナミクスのみに依存することは明らかである．また多くの一般的な場合，ダイナミクスの連続的な変化に対して α の変化も連続的であることが予想される．また，本研究は定性的な試みに留まっているために定式化できないが，α の推定量の良さを評価するものであるから (20) 式の形に帰着することは明らかと言ってもよいであろう．

また，ブートストラップ標本は通常時系列データの連続性を保存しない．本研究では時系列データの連続性を保存する為，ブートストラップ標本としてデータのサブシーケンスを用いる．サブシーケンスとは時系列データの一部分を抜き出したものであり，本研究ではサブシーケンスの長さはハイパーパラメータである．

ブートストラップ標本としてサブシーケンスを用いることの妥当性について補足する．時系列全データ $\{\psi_0 \quad \psi_1 \quad \cdots \quad \psi_N\}$ の元であるダイナミクスが定常状態であると仮定すれば，一定以上の長さのサブシーケンスの元であるダイナミクスはすべて一意な組み合わせのモードによって構成されている．各サブシーケンスの初期値の違いを調整することでサブシーケンス一つ一つをダイナミクスの特徴を反映したブートストラップ標本として用いることができる．

3.3 提案手法

本研究における提案手法ではまず，時系列データ

$$\Psi = \{\psi_0, \psi_1, \cdots, \psi_N\} \tag{21}$$

に対して SP-DMD を適用して解析を行う．x をサブシーケンスの長さ，t' を $0 \leq t' \leq N - x$ を満たすランダムな整数とする．x, t' を用いて作成したサブシーケンス

$$\Psi_x = \{\psi_{t'} \quad \psi_{t'+1} \quad \cdots \quad \psi_{t'+x}\} \tag{22}$$

を用いて，SP-DMD の動的モードの選択部分だけを実行し，各動的モードの係数 α を求める．即ち

$$\min_{\alpha} \quad \|\Psi_x - \Phi D_\alpha V'_{and}\|_F^2 + \gamma \sum_{i=1}^{r} |\alpha_i| \tag{23}$$

によって各動的モードの係数を決定する．ただし，V'_{and} は V_{and} を左から x 列取り出したものである．(23) 式の第一項

$$\|\Psi_x - \Phi D_\alpha V'_{and}\|_F^2 \tag{24}$$

は左から U^* をかけることで通常の DMD と同様に次元削減を利用した形に帰着できる．

$$J'(\alpha) := \|U^*\Psi_x - Y D_\alpha V'_{and}\|_F^2 \tag{25}$$

とすると，全データの場合と同様に $J'(\alpha)$ は

$$\begin{aligned} J'(\alpha) &= \alpha^* P' \alpha - q'^* \alpha - \alpha^* q' + s' \qquad (26) \\ P &:= (Y^* Y) \circ (\overline{V_{and} V_{and}^*}), \\ q &:= \overline{diag\,(V_{and}\,(U^* \Psi_x)^*\, Y)}, \\ s &:= trace(\overline{U^* \Psi_x (U^* \Psi_x)^*}) \end{aligned}$$

と表すことができ，最小化問題の答え α' は (25) 式を最小化する α で，

$$\alpha' = P'^{-1} q' \qquad (27)$$

となる．(22) 式から (26) 式を B 回繰り返し (t' は繰り返しごとにランダムに選ぶ)，各回において各動的モードの係数 α'_i が 0 であるかどうか調べる．

α'_i が 0 である回数が多ければ多いほどその動的モードが含まれるモデルの不確かさは大きい．例えば，B 回の繰り返し全てにおいて $\alpha'_1 = 0$ であれば，データの元となったダイナミクスが ϕ_1 を含んだモデルであるという結論には大きな不確かさが伴う．

4. 実験結果

実験データは [4] により生成された人工データに $[(-5) \times 10^{-5}, 5 \times 10^{-5}]$ の一様分布乱数をノイズとして加えたデータを用いた．

実験結果の評価指標として，各動的モードの係数 α'_i が 0 であるかどうか B 回調べたうち α'_i が 0 であった回数を $B_{[\alpha'_i=0]}$ とした時，$P_{[\alpha'_i=0]} = B_{[\alpha'_i=0]}/B$ で定義される $P_{[\alpha'_i=0]}$ を調べる．$P_{[\alpha'_i=0]}$ が高いほど，サブシーケンスの中に i 番目の動的モードが含まれる確率が低いことを示す．

図 1 は全データに対して SP-DMD を実行した結果である．横軸は各動的モードの添え字であり，DMD によって r 個求めたうちのどの動的モードであるかを表している．縦軸は $P_{[\alpha'_i=0]}$ である．$L1$ 正則化項の係数の大きさによって異なる折れ線が引かれている．全データを解析した結果は $P_{[\alpha'_i=0]} = 0$ もしくは $P_{[\alpha'_i=0]} = 1$ であり，$L1$ 正則化項の係数の大きさによって異なるモデルが選択されている ($P_{[\alpha'_i=0]} = 0$ を満たす動的モードによってダイナミクスが構築されていると推定している) ことがわかる．図 2 は $x = 100$ の場合の提案手法の実行結果である．図 1 と図 2 を比較すると，同じ $L1$ 正則化項の係数であっても，全データに SP-DMD を用いて求めたある動的モードがあらゆるサブシーケンスの中に存在するわけではないこと，また存在する確率がわかる．

5. むすび

提案手法により，SP-DMD のモデル選択の評価を単一データから行った．SP-DMD のモデル選択について，各動的モードがサブシーケンス中に存在する確率という形で一定の指標を与えた．サブシーケンス中に存在する確率が低い動的モードを用いたモデルの推定には大きな不確かさが伴うと言える．一方で，各動的モードがサブシーケンス中に存在するかどうかという定性的な評価に留まった為，各動的モードの振幅の信頼区間を求めるなど定量的な評価を行うことが今後の課題である．

参考文献

[1] T. Takeishi, Y. Kawahara, Y. Tabei, T. Yairi. *Bayesian dynamic mode decomposition.* Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence (IJCAI'17), pp.2814-2821, 2017.

[2] Mihailo R. Jonanović, Peter J. Schmid, Joseph W. Nichols. *Sparsity-promoting dynamic mode decomposition.* Phisics of Fludis 26, 024103.2014.

[3] 汪金芳， 田栗正章， 手塚集， 樺島祥介， 上田修功. 計算統計 I:確率計算の新しい手法．岩波書店．2003.

[4] J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, Joshua L.Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems.* Society for Industrial and Applied Mathematics.2016. pp151, ALGORITHM9.6.getSparseData.m.