Vol.2020-BIO-64 No.9
2020/12/7

線虫の細胞核の時系列データの HDP-HSMM を用いた分節化

中谷 圭志[1] 遠里 由佳子[2]

概要：近年の顕微鏡計測技術と画像処理技術の発展により、顕微鏡画像に撮影された分子や細胞の動態から特徴抽出することにより、非線形な時系列データが得られつつある。そうした時系列データを、意味のあるまとまりに区切る「分節化」は、時系列の分類や異常検知などを実現する上で重要である。本発表では、線虫の初期発生の細胞核の時系列データの分節化において、出力確率に正規分布を仮定した HDP-HSMM (Hierarchical Dirichlet Process-Hidden Semi Markov Model)の利用を提案し、その有効性を検証する。

キーワード：隠れマルコフモデル，隠れセミマルコフモデル、異常検知、ベイズ推定

1. はじめに

生命科学の分野では、生命現象の動的な知識を得ることを目的に、タイムラプス顕微鏡で撮影された分子や細胞を画像処理することで、時空間情報を数値として含む定量データが得られつつある。例えば、線虫 *Caenorhabditis elegans* (*C. elegans*)では、RNAi (RNA interference)と呼ばれる実験で 1 つの遺伝子の発現を抑制した場合に、線虫の初期発生に影響を及ぼす遺伝子群が特定され、RNAi 胚や野生胚の初期発生を微分干渉顕微鏡で撮影した大量の画像データと、画像から細胞核を物体検出して得た細胞核の時空間定量データが公開された[1]。こうした定量データを特徴抽出して得られる多変量な時系列データを解析することは、RNAi により生じた異常や遺伝子の機能を推定する上で重要である(図 1)。しかし、得られる時系列が、細胞周期などに由来する決定論的な法則に従い、非線形な性質を持つことが、その解析を難しくしている。そこで本研究では、多変量な時系列データを意味のある単位に区切る「分節化」と、分節化された時系列のモデリングをめざした。

時系列の分節化とモデリングに有効なアプローチの 1 つが、音声認識で広く利用されている隠れマルコフモデル (HMM: Hidden Markov Model、以下 HMM) である。HMM は、正規分布などの連続分布が割り当てられた潜在的な状態を仮定し、その状態遷移を時系列データから推定できる。HMM の隠れ状態の構造は、ある状態からすべての状態に遷移を許す E (Ergodic, 全遷移)型と、状態の遷移が時間に対して逆戻りできないように制限した LR(Left-to-Right)型に分類される。しかし、HMM の状態が持続する確率は、状態遷移確率のみに依存しており、継続長の増加に伴い指数的に減衰するため、状態継続長の変動を表現するには精度が不十分であるという問題があった。そこで、状態継続長分布を明示的に定義したモデルとして、隠れセミマルコフモデル(HSMM : Hidden semi Markov model[2], 以下 HSMM) が提案された。さらに、状態数の決定を自動的に行えるよう、状態遷移の事前分布に階層ディリクレ過程 (HDP: Hierarchical Dirichlet Process、以下 HDP)を導入した HDP-HMM [3]および HDP-HSMM [4]が提案された。そして、ヒトの音声の認識[5]や行動の異常検知[6]に用いられている。

そこで本研究では、*C. elegans* の初期発生の細胞核の時系列データの分節化を例に、HDP-HSMM の有効性を検証する。以下では、2 章で LR 型 HDP-HSMM について説明する。3 章で人工データおよび実データに対して HDP-HSMM を用いた実験結果を示す。4 章でまとめと今後の展望について述べる。

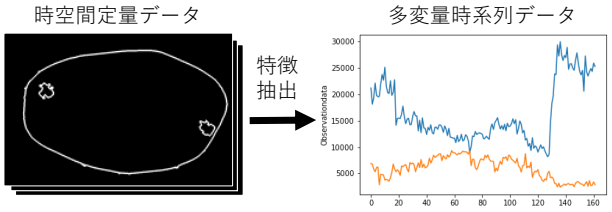

図 1 細胞核の多変量時系列データ。青：面積の時間変化、オレンジ：円形度の時間変化 (3 章にて詳述) 。

2. LR 型 HDP-HSMM

HDP-HSMM [4]は、時系列データ$y = (y_1, y_2, \ldots, y_j, \ldots, y_T)$に対して、時点$j$のデータ$y_j$の背後に、隠れた状態$x_j$を仮定し、隠れ状態を共有する上位状態(super-state) z_i を考える。そして、すべての状態において可能な遷移先が共有されており、その遷移確率が状態毎に異なる分布を構成するよう HDP を用いる。最初に、ディリクレ過程(DP: Dirichlet Process)の 1 つの構成法として、一般的に$\beta \sim \mathrm{GEM}(\gamma)$と表記される、棒折り過程 (SBP: Stick Breaking Process)を用いる。SBP は長さ 1 の棒を左から無限個に折る操作に例えられ、ベータ分布からサンプルした値で多項分布βを生成する(式 1)。次に、βを基底測度として共有した DP によって上位状態i毎に異なる遷移確率π_iを生成する(式 2)。

$$\beta \sim \mathrm{GEM}(\gamma) \quad (1)$$

1 大阪電通大・院工・情報工
Dept. of Info. Eng., Grad. Sch. of Eng., O.E.C.U.
2 立命館大・情報理工
Faculty of Info. Sci. & Eng., Ritsumeikan Univ.

$$\pi_i \sim \mathrm{DP}(\alpha, \beta)\ (\mathrm{i} = 1,2,..,\infty) \quad (2)$$

そして、π_iから自己遷移を除いた$\bar{\pi}_i$を求める（式 3)。

$$\bar{\pi}_i = \frac{\pi_{ij}}{1-\pi_{ii}}\left(1-\delta_{ij}\right) \text{ where } \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

このとき HSMM の生成モデルは、下記のように記述される(式 4-10)。

$$(\theta_i, \omega_i) \sim \mathrm{H} \times \mathrm{F}\ (\mathrm{i} = 1,2,..,\infty) \quad (4)$$

$$z_s \sim \bar{\pi}_{z_{s-1}}\ (s = 1,2,\ldots,\infty) \quad (5)$$

$$D_s \sim h(\omega_{z_s}) \quad (6)$$

$$x_{t_s^1:t_s'} = z_s \quad (7)$$

$$y_{t_s^1:t_s'} = f(\theta_{x_t}) \quad (8)$$

$$t_s^1 = \sum_{s'<s} D_{s'} \quad (9)$$

$$t_s' = t_s^1 + D_s - 1 \quad (10)$$

ここで、H と F は 出力分布と状態継続長分布の基底測度であり、関数 h と f は出力分布と状態継続長分布をあらわす。上位状態系列のうちs番目の上位状態を、z_sとするとき、z_sの継続長をD_s、区間 t_s^1 と t_s' はz_sの開始と終了の時点を意味する。HDP-HSMM のパラメータの最適化には、パラメータを有限個に限定して最適化を行う、weak-limit 近似を導入したギブスサンプリングを用いることができる。

本研究では、i番目の上位状態z_iの持続時間D_iはポアソン分布fから生成し、事前分布にガンマ分布を用いる。観測された時系列データy_iは 2 次元ガウス分布から生成されたとし、事前分布として正規逆ウィシャート(NIV: Normal-Inverse-Wishart)分布を用いる(図 2)。そして、E 型の HDP-HSMM[4]を、状態遷移が LR 型になるようパラメータの更新を制限した LR 型 HDP-HSMM を構築した。

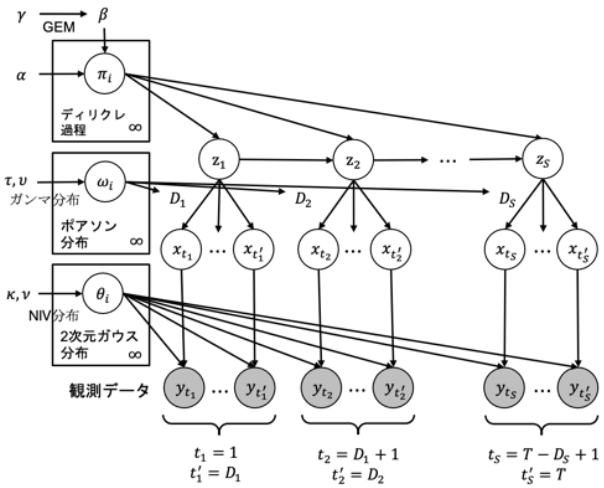

図 2　HDP- HSMM のグラフィカルモデル(式 1-10)。影付きノードは観測値を、影なしのノードは変数を、それ以外は定数を表す。

3. 実験結果

HDP-HSMM の構築では、HDP のハイパーパラメータとして、$\gamma = 6$、$\alpha = 6$を設定し、持続時間長にポアソン分布を仮定し$\tau = 60$、$\upsilon = 2$を設定した。出力分布のハイパーパラメータとしてμ_0 =(0,0)、$\sigma_0 = 2$ 次元の単位行列、$\kappa = 0.25$、$\nu = 8$を設定した。ギブスサンプリングの繰り返し回数は、1000 に設定した。Weak-limit として上位状態の最大数を 20 に制限した。

3.1 人工データによる評価

最初に、隠れ状態が既知の人工データを用いて、モデルの定量的な評価を行った。2 次元の時系列を作るため、人工データは、上限を 10,30,50,70,90、下限を 20,40,60,80,100 とする 5 種類の一様分布からランダムに 10 組を選択し、それぞれ、標準偏差 0.05 の正規乱数をノイズとして加え、長さ 50 の時系列を生成し、正解の状態数が 9 の 2 次元の時系列データを作成した (図 3、計 500 時点)。

実験では、E 型と LR 型の HDP-HMM と、E 型と LR 型の HDP-HSMM の 4 モデルを比較した(表 1)。最も文節化が最も正確なのは、E 型の HDP-HSMM となった。これは、人工データが状態継続長を考慮しないためと考えられる。LR 型 HDP-HSMM は、LR 型 HDP-HMM と比べて、状態を維持することで、正解率は改善するが、初期値に依存する性質がみられた(図 4)。

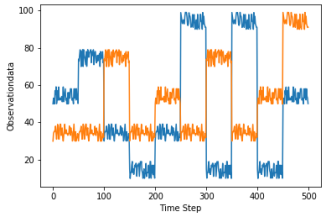

図 3　人工データ

表 1　人工データに対する状態推定の精度比較

モデル	遷移構造	状態数	正解率(%)
HDP-HMM	E 型	8	80
	LR 型	7	0.4
HDP-HSMM	E 型	9	100
	LR 型	7	30

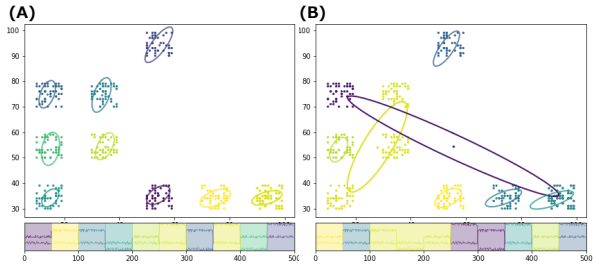

図 4　人工データの文節化の結果。(A)E 型 HDP-HSMM、(B)LR 型 HDP-HSMM。(A-B) 上のパネルは各上位状態に割り当てられた 2 次元正規分布とデータの散布図を、下のパネルは入力された時系列とその文節化の結果を表しており、上下のパネルの上位状態の種類は色で対応をとっている。

3.2 実データによる評価

線虫初期発生を微分干渉顕微鏡で撮影したタイムラプスの 2 次元画像データ[7]から、AB 細胞と呼ばれる細胞の核の輪郭を、ImageJ の ROI manager を使って手作業で抽出し、時空間定量データを作成した(5 秒毎、計 164 時点)。得られた時空間定量データから 2 次元の時系列データを作成するため、円形度や面積を特徴量として求めて得た時系列 y_t(図 1)と、その階差時系列$\Delta y_t = y_t - y_{t-1}$を作成した(図 5)。なお、ある輪郭が面積 S 周囲長 L を持つとき、円形度は $4\pi S/L^2$ で表され、真円に近いほど最大値 1 に近い値に、輪郭が複雑なほど最小値 0 に近い値になる。

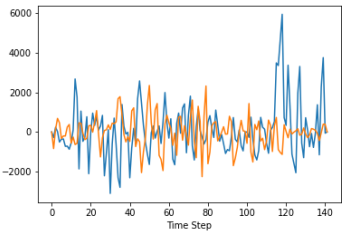

図 5　図 1 の時系列から作成した階差時系列。青：面積の階差時系列、オレンジ：円形度の階差時系列。

元となる時系列と階差時系列の 3 とおりの組み合わせに対して、相関係数により多重共線性の調査を行った(図 6)。相関係数の値は、面積と面積の階差時系列間で-0.1、円形度と円形度の階差時系列間で-0.2、面積と円形度の時系列間で-0.7 となった。面積を特徴量として得た時系列と、その階差時系列の組み合わせが、相関係数が最も低いことから、最適な組み合わせであると判断した。

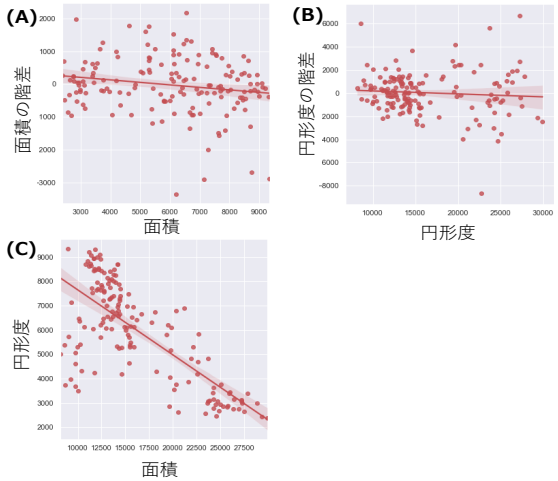

図 6　時系列データ間の散布図の比較。(A)面積の時系列と面積の階差時系列、(B)円形度の時系列と円形度の階差時系列、(C)面積と円形度の時系列

そこで、面積の時系列と階差時系列を二次元時系列データとして HDP-HSMM に用いた。野生胚の細胞分裂の上位状態の遷移は、時間に対して逆戻りできないことを想定して LR 型と、E 型の比較を行った(図 7)。

LR 型の HDP-HSMM では 3 種類の上位状態が認識され、これは細胞周期で、間期 1 種類と細胞分裂期 2 種類に相当することが示唆された(図 8)。E 型の HDP-HSMM では、間期をより細分化した形に分節化していることを確認した。

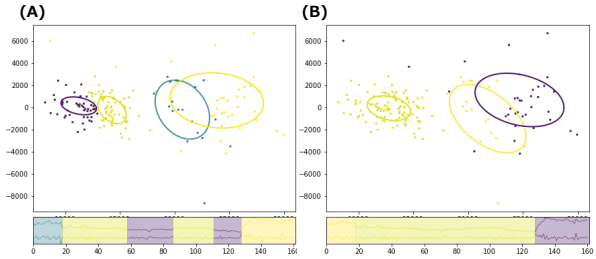

図 7　実データの文節化の結果。(A)E 型 HDP-HSMM、(B)LR 型 HDP-HSMM。(A-B) 上のパネルは各上位状態に割り当てられた 2 次元正規分布とデータの散布図を、下のパネルは入力された時系列とその文節化の結果を表しており、上下のパネルの上位状態の種類は色で対応をとっている。

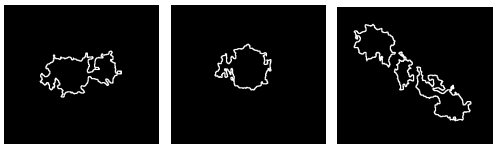

図 8　図 7B の HDP-HSMM の左から順に 3 種類の上位状態に代表的な細胞核の形状。

4. おわりに

LR 型 HDP-HSMM を実装し、人工データや実データから得られた時系列を用いて分節化を行った。

今後の課題として、状態継続長を考慮した人工データの作成と評価や、時系列データの欠損値の影響の評価がある。特徴量として面積と円形度に着目したが、楕円フーリエ係数の利用も検討したい。HSMM の拡張として、リカレント HSMM の導入も考えられる。

謝辞　本研究の成果の一部は科研費 (16K00414, 19K12226) の補助による。

参考文献

[1] Kyoda, K. et al.. WDDD: Worm Developmental Dynamics Database. Nucleic Acids Res. 2013, vol. 41, p. D732-D737.
[2] Murphy, K. P.. Hidden Semi-Markov Models (HSMMs). Technical Report. 2002.
[3] Fox, E. B. et al.. An HDP-HMM for Systems with State Persistence. In Proc. Int. Conf. Mach. Learn. 2008.
[4] Johnson, M. J. and Willsky, A. S.. Bayesian Nonparametric Hidden Semi-Markov Models, J. Mach. Learn. Res. 2013, vol. 14, p. 673-701.
[5] Taniguchi, T. et al.. Bayesian Double Articulation Analyzer for Direct Language Acquisition from Continuous Speech Signals, IEEE Trans. Cogn. Dev. Syst. 2016, vol. 8, no. 3, p. 171-185.
[6] Fuse, T. and Kamiya K.. Statistical Anomaly Detection in Human Dynamics Monitoring Using a Hierarchical Dirichlet Process Hidden Markov Model, IEEE Trans. Intel. Transp. Syst. 2017, vol. 18, no. 11, p. 3083-3092.
[7] Gönczy, P. et al.. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III, Nature. 2000, vol. 408, p. 331-336.