修士論文 要旨

増殖バイアスを考慮した系譜データの隠れ状態推定方法の構築

数理情報学専攻 48176215　中島 蒼

指導教員　小林 徹也 教授

本修士論文の内容は，[1, 2] に基づく．

1　系列データの隠れ状態推定とその拡張

系列データの隠れ状態推定では，**隠れ Markov モデル**を土台として，効率的な手法が提案され応用されてきた [3]．そのような手法は，隠れ Markov モデルのループのないグラフィカルモデルが含意する条件付き独立性を利用している，その一例は，EM 型のモデル推定手法である **BW アルゴリズム** (Baum-Welch algorithm) である．これらの手法は，隠れ変数が分岐しながら時間発展する**樹状の隠れ Markov モデル**に一般化され，画像解析・信号処理の分野において樹状の依存関係を持つ隠れ変数の推定に使われてきた [4, 5]．樹状の隠れ Markov モデルでも，グラフィカルモデルにループが無いため，隠れ Markov モデルの効率的な手法を拡張し適用できる．樹状の依存関係をもつ隠れ変数の推定は，生物学でも先祖-子孫間での遺伝を扱うために重要であり，様々な手法が提案されてきた [6, 7]．

近年樹状の隠れ Markov モデルの更なる一般化である，**系譜** (lineage tree) からの隠れ状態推定が重要になってきている（図 1）．ここで系譜とは，あらかじめ定められた時間 T までの細胞の分裂と増殖を観察したデータである．具体的には，系譜は各細胞の親子関係と生存期間の長さ（**分裂時間** (division time)）の 2 種類の情報から成る．系譜は近年の実験技術の進歩で得られるようになった [8]．分裂時間を観測変数とし，隠れ状態として各細胞の増殖能力をモデル化すると，隠れ状態推定により各細胞の増殖能力を系譜から定量化できる．この定量化はパーシステンス (persistence) [9, 10] などの細胞集団の増殖競争による現象の予測や制御に重要である．

2　系譜データの増殖バイアスと本論文の貢献

しかし，系譜データからの隠れ状態推定は，**増殖バイアス** (survivorship bias) というデータの歪みのため，樹状の隠れ Markov モデルでの隠れ状態推定よりも困難な問題である．隠れ Markov モデルでの隠れ状態推定では，グラフィカルモデルにループが無いこと，特に，観測変数が隠れ変数に影響しないことが重要であった．しかし，系譜においては，観測変数から隠れ変数への影響が実験の打ち切り (censoring) という形で存在する（図 1）．系譜において一本の経路に着目すると，この時系列は隠れ Markov モデルになっている．この隠れ変数の個数は，観測変数である分裂時間に依存して，観察が打ち切られることで決まっている．このような形で隠れ変数は観測変数に依存しているため，隠れ Markov モデルの場合に利用した条件付き独立性は崩壊し，分裂時間が短く偏る増殖バイアスが発生する．具体的には，分裂時間が短かった経路ほどサンプルを多く含むため，系譜全体としては分裂時間が短い事象が過剰に多く含まれる．増殖バイアスは近年研究が発展している分野で [11, 12]，未だ確立した簡単な補正法はない．

本論文では，系譜データを用いた増殖バイアスの影響を受けない隠れ状態推定手法である**系譜 EM アルゴリズム** (lineage EM algorithm; LEM) を提案する．そのため，まずは系譜データにおける増殖バイアスを特徴付け，補正法を提案する．その後，この補正法と樹状の隠れ Markov モデルでの BW アルゴリズムを組み合わせ，LEM を導入する．

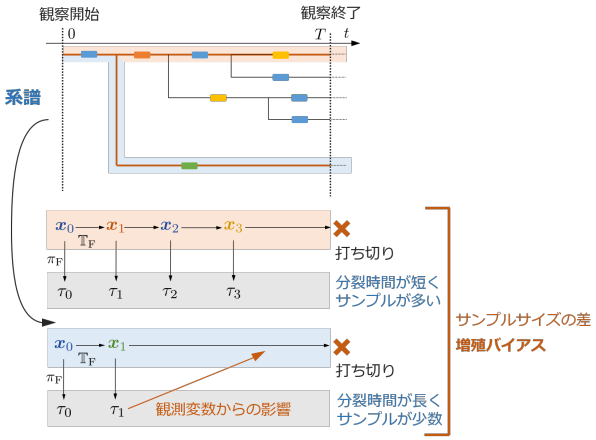

図 1. 系譜の例示と，観測変数からの打ち切りとしての隠れ状態への影響．系譜内の一本の経路（赤色・青色の四角で囲った）に着目すると，図の下部のような隠れ Markov モデルとなる．この隠れ変数の個数は観測変数の影響で決まり，分裂時間が短いほど多くサンプルに含まれる．

3　系譜のモデル

本論文での系譜のモデル化について説明する．各細胞は死亡せず必ず 2 つの娘細胞に分裂する．各細胞は隠れ状態（単に**状態**と呼ぶ）$\boldsymbol{x} \in \Omega$ を持つ．状態の成す集合 Ω は有限集合の場合も連続空間の場合もあ

る．状態は分裂時に確率的に遷移する．娘細胞の状態 $\boldsymbol{x}'$ は親細胞の状態 $\boldsymbol{x}$ にのみ依存し，姉妹細胞とは独立に決まる．この確率は遷移行列 $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ で与えられる．本論文では $\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x})$ がエルゴード的 (ergodic) な場合のみを扱う．分裂時間 τ は，状態 $\boldsymbol{x}$ に依存した分布 $\pi_{\mathrm{F}}(\tau|\boldsymbol{x})$ に従う．これらの設定の下，このモデルは**状態付き齢依存分岐過程** (multi-type age-dependent branching process) [13] と呼ばれる確率過程になる．

4　増殖バイアスの特徴付けと補正法

増殖バイアスの特徴付けと補正法の提案をする．まずは議論を簡単にし増殖バイアスに集中するため，状態 $\boldsymbol{x}$ が既知であるとして分裂時間の分布 π_{F} や状態の確率遷移行列 $\mathbb{T}_{\mathrm{F}}$ を推定する問題を扱う．実際に必要な状態が未知の場合の推定手法は，次節で扱う．状態が既知であれば，π_{F} や $\mathbb{T}_{\mathrm{F}}$ は以下で定義する経験分布として推定できる：

$$\pi^{\mathcal{T}}_{\mathrm{emp}}(\tau|\boldsymbol{x}) := \frac{1}{|\mathcal{T}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{T}_{\boldsymbol{x}}}\delta(\tau-\tau_i),$$

$$\mathbb{T}^{\mathcal{T}}_{\mathrm{emp}}(\boldsymbol{x}'|\boldsymbol{x}) := \frac{\text{状態 } \boldsymbol{x} \text{ から } \boldsymbol{x}' \text{ への遷移回数}}{\text{状態 } \boldsymbol{x} \text{ からの遷移回数}}.$$

ここで，$\mathcal{T}_{\boldsymbol{x}}$ は系譜内で状態が $\boldsymbol{x}$ である葉以外の細胞の集合を表すとする．本論文では，この経験分布の収束先を調べることで，増殖バイアスの影響を特徴付けた：

定理 1 (Informal)**.** 適当な仮定の下，$T\to\infty$ の極限で，経験分布は以下のように収束する：

$$\pi^{\mathcal{T}_{\boldsymbol{x}}}_{\mathrm{emp}}(\tau|\boldsymbol{x})\to\pi_{\mathrm{B}}(\tau|\boldsymbol{x});\quad \mathbb{T}^{\mathcal{T}_{\boldsymbol{x}}}_{\mathrm{emp}}(\boldsymbol{x}'|\boldsymbol{x})\to\mathbb{T}_{\mathrm{F}}(\boldsymbol{x}'|\boldsymbol{x}).$$

ここで，$\pi_{\mathrm{B}}(\tau|\boldsymbol{x})$ は**遡及過程** (retrospective process) [2] という確率過程の分裂時間の分布であり，以下で定義される：

$$\pi_{\mathrm{B}}(\tau|\boldsymbol{x})\propto\pi_{\mathrm{F}}(\tau|\boldsymbol{x})e^{-\lambda\tau}. \tag{4.1}$$

式中の λ は**集団増殖率** (population growth rate) であり，容易に観測できる量である．

上の定理の帰結として，状態遷移 $\mathbb{T}_{\mathrm{F}}$ の推定はバイアスを受けないが，分裂時間 π_{F} の推定はバイアスを受け，間違って π_{B} が推定されると分かる．しかしこのバイアスは，(4.1) を用いて π_{B} を π_{F} に変換することで，容易に補正できる．

5　系譜 EM アルゴリズム (LEM)

LEM の概要について説明する．前節の結果より，何かしらの手法で隠れ状態推定を行い π_{B} と $\mathbb{T}_{\mathrm{F}}$ が推定できれば，増殖バイアスを補正してモデルを正しく推定できると分かった．本論文では π_{B} と $\mathbb{T}_{\mathrm{F}}$ の推定を，樹状の隠れ Markov モデルに拡張された BW アルゴリズムで行う．この推定と補正の組を LEM と名付ける．LEM では π_{B} と $\mathbb{T}_{\mathrm{F}}$ を推定するために，E-ステップと M-ステップを交互に反復する．E-ステップにおいては，現在のモデル $(\pi_{\mathrm{B}},\mathbb{T}_{\mathrm{F}})$ と観測データ $\mathcal{D}$ から状態の事後確率 $\mathbb{P}(\boldsymbol{x}_i|\mathcal{D},\pi_{\mathrm{B}},\mathbb{T}_{\mathrm{F}})$ を計算する．この事後確率は**信念伝搬法** (Blief Propagation) という手法で効率的に計算される [3]．M-ステップにおいては，現在のモデル $(\pi_{\mathrm{B}},\mathbb{T}_{\mathrm{F}})$ を，E-ステップで推定した経験分布にフィットするように最尤推定で更新する．

LEM の正当性は人工データに対する推定で確認された．また，LEM を大腸菌の系譜データに適用し，増殖に関連し数世代に渡って遺伝する情報を発見した．

6　まとめ

本論文では，増殖バイアスの影響を受けない系譜データからの隠れ状態推定手法である LEM を提案した．本論文で用いた増殖バイアスの補正法は，BW アルゴリズム以外の手法にも適用でき拡張性が高い．今後の課題は，LEM の適用範囲を広げるため，一般化されたモデルを扱うことである．また，推定された隠れ状態と観測可能な量を結びつける手法を提案するのも課題である．LEM は，本論文で示した有用性に加え，これらの拡張によって応用範囲が広がることで，系譜データを解析する基礎的な手法になっていくだろう．

参考文献

[1] Nakashima S, Sughiyama Y, Kobayashi TJ (2018) Deciphering latent growth-states from cellular lineage trees. *bioRxiv*.
[2] Sughiyama Y, Nakashima S, Kobayashi TJ (2019) Fitness response relation of a multitype age-structured population dynamics. *Phys. Rev. E* 99(1):012413.
[3] Christopher B (2006) *Pattern Recognition and Machine Learning*. (Springer-Verlag New York).
[4] Chou KC, Willsky AS, Benveniste A (1994) Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control* 39(3):464–478.
[5] Laferte JM, Perez P, Heitz F (2000) Discrete markov image modeling and inference on the quadtree. *IEEE Transactions on Image Processing* 9(3):390–404.
[6] Hormoz S, Desprat N, Shraiman BI (2015) Inferring epigenetic dynamics from kin correlations. *Proceedings of the National Academy of Sciences* 112(18):E2281–E2289.
[7] Kuchen EE, Becker N, Claudino N, Hofer T (2018) Long-range memory of growth and cycle progression correlates cell cycles in lineage trees. *bioRxiv*.
[8] Hashimoto M, et al. (2016) Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113(12):3251–3256.
[9] Bigger JW (1944) Treatment of staphyloeoeeal infections with penicillin by intermittent sterilisation. *Lancet* pp. 497–500.
[10] Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. *Science* 305(5690):1622–1625.
[11] Hoffmann M, Olivier A (2016) Nonparametric estimation of the division rate of an age dependent branching process. *Stochastic Processes and their Applications* 126(5):1433 – 1471.
[12] Marguet A (2017) A law of large numbers for branching Markov processes by the ergodicity of ancestral lineages. *ArXiv e-prints*.
[13] Harris TE (1963) *The Theory of Branching Processes*. (Springer-Verlag Berlin Heidelberg).