シンポジウム

多様な分野のデータに対する
統計科学・機械学習的アプローチ

開催日：2020 年 9 月 28, 29 日

オンライン開催

シンポジウム「多様な分野のデータに対する統計科学・機械学習的アプローチ」

開催責任者：松井 秀俊（滋賀大学）

文部科学省科学研究費補助金　基盤研究 A（20H00576）
「大規模複雑データの理論と方法論の革新的展開」
研究代表者：青嶋 誠（筑波大学）

内容・目的

近年の計測・測定技術の発展に伴い、大量で複雑な形式を持つデータが様々な分野で取得されるようになってきました。これにより、データの形式や分析目的に応じたデータ科学的アプローチに対する需要が高まっています。本シンポジウムでは、多様な分野におけるデータに対する、統計学、機械学習の応用事例に関する講演を広く募集します。講演内容としては、新たな分析手法の提案のみならず、応用分野の側面から見た分析上の問題提起なども歓迎します。参加者の交流を通じて、知識の共有だけでなく、新たな研究の発展や問題解決に繋げる場になることを目的としています。

プログラム

9 月 28 日（月）

10:00-10:05　Opening

10:05-10:35

「New Approach to Galaxy Evolution via Manifold Learning」

竹内努（名古屋大学理学研究科）

10:35-11:05

「常微分方程式の時間に依存したパラメータの推定による新型コロナウイルス感染症の実効再生産数」

山口崇幸（滋賀大学データサイエンス教育研究センター）

11:20-11:50

「構造出力デザインのための能動学習とその材料科学問題への応用」

松井孝太（名古屋大学大学院医学系研究科）

草川隼也（名古屋工業大学大学院工学研究科）

安藤圭理（名古屋大学大学院工学研究科）

沓掛健太朗（理化学研究所革新知能統合研究センター）

宇治原徹（名古屋大学大学院工学研究科，産業技術総合研究所）

竹内一郎（名古屋工業大学大学院工学研究科，理化学研究所革新知能統合研究センター，物質・材料研究機構）

11:50-12:20

「ガウス過程分位点回帰モデリングと新生児ホルモン分泌リズムの解明」

田辺佑太（静岡大学大学院総合科学技術研究科）

荒木由布子（静岡大学大学院総合科学技術研究科）

13:30-14:00

「薬物動態パラメータの予測から目指す創薬支援」

江崎剛史（滋賀大学データサイエンス教育研究センター）

14:00-14:30

「遺伝子ネットワーク構造が予測精度に与える影響」

沖永悠一（九州大学大学院数理学府）

京極大助（兵庫県立人と自然の博物館自然・環境マネジメント研究部）
近藤聡（トヨタ自動車（株）アグリバイオ事業部農食事業室）
永野惇（龍谷大学農学部）
廣瀬慧（九州大学マス・フォア・インダストリ研究所）

14:45-15:15
「A novel metric for hyperbolic phylogenetic tree embeddings」
松本拡高（長崎大学情報データ科学部）

15:15-15:45
「高次元統計解析に基づく遺伝子発現データのノイズ削減法」
井元 佑介（京都大学高等研究院）
平岡 裕章（京都大学高等研究院）
吉脇 理雄（理化学研究所革新知能統合研究センター）
Emerson G. Escolar（理化学研究所革新知能統合研究センター）
中村 友紀（京都大学医学研究科）
山本 拓也（京都大学 iPS 細胞研究所）
斎藤 通紀（京都大学高等研究院）

16:00-16:50
特別講演：「植物科学分野のデータに対する統計科学・機械学習的アプローチと作物研究への応用」
持田恵一（理化学研究所環境資源科学研究センター，横浜市立大学木原生物学研究所，岡山大学資源植物科学研究所，理化学研究所バトンゾーン研究推進プログラム）

17:00-18:00　総合討論（zoom ブレイクアウトルーム使用）
19:00-　懇親会（オンライン）

9 月 29 日（火）

10:00-10:30
「概念語の複雑性、ならびに『きわめて大量だが信頼性の保証がない言語情報』の取り扱いについて」
得丸久文（カラハリプロジェクト）

10:30-11:00

「Dating the Nuzi Cuneiform Tablets Computationally: Analyzing Family Networks in Ancient Mesopotamia」
上田澄江（統計数理研究所元教員）
土谷隆（政策研究大学院大学）
伊藤栄明（統計数理研究所）

11:15-11:45
「Health status and repeated multiple treatments in long-term care: A panel structural VAR analysis」
菅原慎矢（東京理科大学経営学部）
石原庸博（大阪経済大学経営学部）

11:45-12:15
「Forward Variable Selection for Sparse Ultra-High Dimensional Generalized Varying Coefficient Models」
本田敏雄（一橋大学経済学研究科）
林建同（国立清華大学統計学研究所）

13:30-14:00
「クラスタリングによる正準判別の精度向上とクロスバリデーションの高速化」
三浦完太（九州大学大学院数理学府）
廣瀬慧（九州大学マス・フォア・インダストリ研究所）

14:00-14:30
「射影勾配法による高次元回帰モデリング」
川島孝行（東京工業大学情報理工学院，理化学研究所革新知能統合研究センター）

14:30-15:00
「カーネル法に基づく超高次元モデル選択」
梅津佑太（長崎大学情報データ科学部）

15:00-15:05 Closing

New Approach to Galaxy Evolution via Manifold Learning

(多様体学習による新しい銀河進化の定量化)

竹内　努 [1,2]

1. 名古屋大学理学研究科素粒子宇宙物理学専攻,

2. 統計数理研究所統計的機械学習研究センター

銀河とは，星と星間物質(ガスと星間塵)，暗黒物質からなる巨大な天体である．銀河は現在観測可能な宇宙に数千億個存在しており，我々の目に見える波長(可視光線)での宇宙の姿を形作っている．しかし宇宙誕生当時，物質分布はほぼ一様であった．銀河は平均からわずかに密度が高い領域が重力によって成長し，合体成長を経て現在の姿へと進化してきたのである(図 1)．銀河進化は周囲の銀河の密度やガス密度など，銀河の置かれた環境にも強く依存する極めて複雑な過程である．そして銀河の形成と進化は 138 億年の宇宙進化の歴史の中でも非常に重要な現象のひとつであると考えられている．

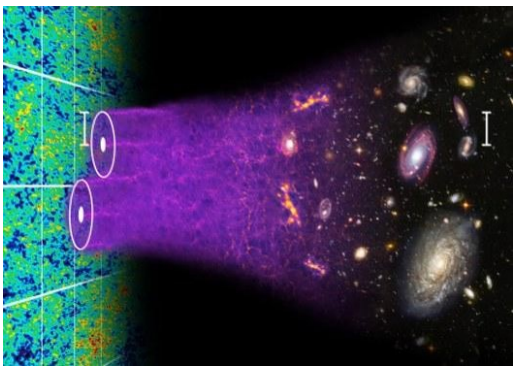


図 1: 一様な物質分布から銀河が形成され、進化する様子.

我々は 130 数億年にわたる銀河の形成・進化という複雑な物理現象を，これまでの宇宙物理学の方法とはまったく異なる，データ科学の最新手法である位相的データ解析(topological data analysis: TDA)に基づくアプローチによって新たな角度から解明することを試みている．本研究では，TDA のひとつである多様体学習(manifold learning)の方法(図 2)を用い，多波長銀河探査データに対してこの方法を適用し，銀河進化を記述する方程式の構築を行った．

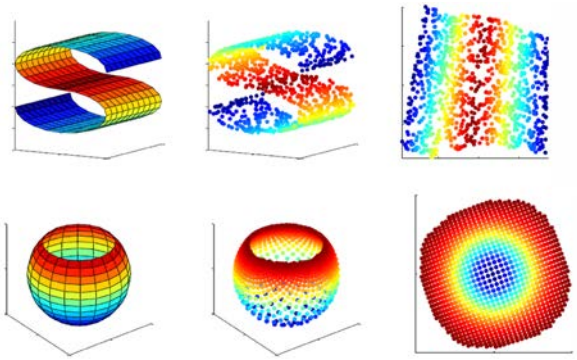


図 2: 多様体学習の典型的な例.

多波長銀河データは Reference Catalog of galaxy Spectral Energy Distributions (RCSED: Chilingarian et al. 2016)を用いた. RCSED は約 100 万個の銀河の紫外から近赤外までの測光データ (11 バンド: FUV, NUV, *u*, *g*, *r*, *i*, *z*, *Y*, *J*, *H*, *K*)および関連する物理量を含んでいる．この全 11 バンドおよび赤方偏移の揃った 30,000 個の銀河からなる volume-limited sample を構築し，12 次元特徴空間に多様体学習を用いた結果，銀河は 3 次元の光度空間に埋めこまれた 2 次元曲面で表現さ

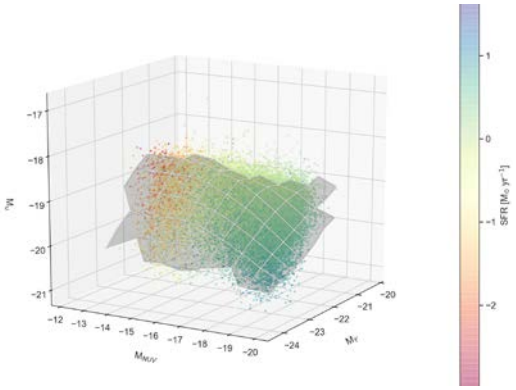


図 3: RCSED サンプルから発見された銀河多様体．強い非線型性を示す.

れることを発見した(図 3). この多様体は強い非線型性を持ち, 銀河多様体という概念が誕生した 30 年前の古典的方法論, たとえば古典的主成分解析などの方法では発見し得なかったものである.

さらに, この銀河多様体を多様体の大局的トポロジーを保持するアルゴリズムである Isomap (Tenenbaum et al. 2000)および UMAP (McInnes et al. 2018)を用いて定量化し, 2 次元多様体を記述する座標系を推定した. 主成分解析等でもよく知られているように 一般にはこれらの多様体座標に物理的意味を付与できるとは限らず, それ以上の物理的定量化が困難である例も多い. 紫外線・可視光・近赤外線に特に顕著に現れる銀河の進化は星形成率および星質量であることから, これらの主要な物理量と多様体座標の相関を検証してみたところ, 実際に星形成率および星質量が 2 つの多様体座標にほぼ対応していることが見出された.

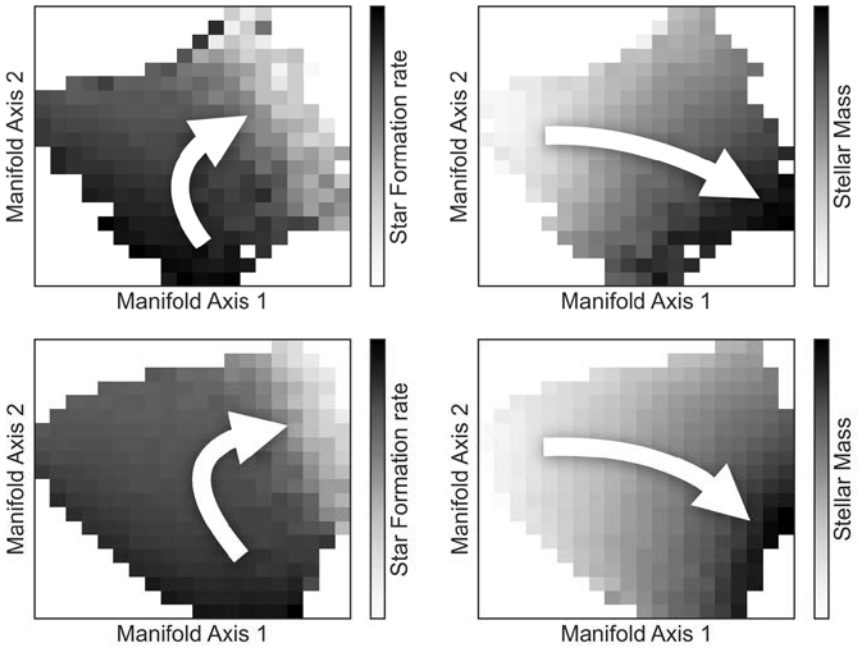


図 4: Isomap および UMAP によって求めた銀河多様体. 上段が Isomap, 下段が UMAP によって構成された銀河多様体を示す. 多様体座標上での星形成率(左)および星質量(右)の振舞いがグレースケールで示されている.

今回は銀河の限定された物理量の記述を行ったが, この方法の有効性が示されたことで, 今後はより広範な物理量の記述に適用できると期待できる.

常微分方程式の時間に依存したパラメータの推定による新型コロナウイルス感染症の実効再生産数

山口 崇幸

1 概要

2020 年 1 月に日本でも新型コロナウイルスの感染者が確認され, その後, 感染拡大が起こり, 毎日, 新規感染者数が発表されている. 本研究では, 新型コロナウイルスの感染のモデルとして SEIR モデルを考え, 新規感染者数からそのパラメータを推定した. モデルは B-スプラインで時間に依存したパラメータを表した常微分方程式であり, 推定には罰則付き最小二乗法を用いた. 平滑化パラメータは交差検証で定め, 定常ブートストラップで信頼区間を計算した. このモデルによる実効再生産数と緊急事態宣言における接触削減の定量化を報告する.

2 SEIR モデルと実効再生産数

モデルは図 1 のコンパートメントの SEIR モデルを考える. $S(t)$ は感受性人口 (susceptible), $E(t)$ は感染後の潜伏期間にある人口 (exposed), $I(t)$ は感染して発症した人口 (infected), $R(t)$ は診断され隔離された人口 (removed) とする. 新型コロナウイルス感染症は症状が出る前であっても感染性があるため, I だけでなく E も感染性があるとして, 次のような常微分方程式を用いる.

$$\frac{dS}{dt}(t) = -\beta(t)(E(t) + I(t))S(t) \tag{1}$$

$$\frac{dE}{dt}(t) = \beta(t)(E(t) + I(t))S(t) - \alpha E(t) \tag{2}$$

$$\frac{dI}{dt}(t) = \alpha E(t) - \gamma I(t) \tag{3}$$

$$\frac{dR}{dt}(t) = \gamma I(t) \tag{4}$$

人口は保存されているとし, S, E, I, R を総人口 N で割ることで

$$S(t) + E(t) + I(t) + R(t) = 1 \tag{5}$$

とする. $\beta(t)$ は 3 次の B-スプラインとする. このモデルでは, 実効再生産数は次で与えられる.

$$R_e(t) = \left(\frac{1}{\alpha} + \frac{1}{\gamma}\right)\beta(t)S(t). \tag{6}$$

3 推定方法

ジャッグジャパン株式会社提供による都道府県別新型コロナウイルス感染者数マップ[*1] で公開されている各自治体の発表をまとめた CSV ファイルを用いた. 1 月 15 日から 9 月 11 日までの 241 日間の新型コロナウイ

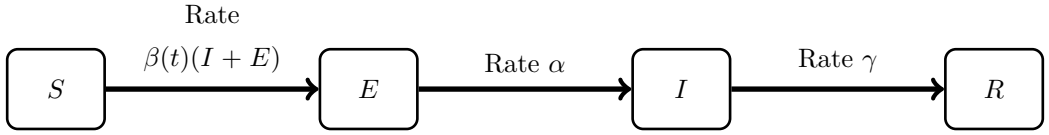


図 1 SEIR モデルのコンパートメント

[*1] https://gis.jag-japan.com/covid19jp/

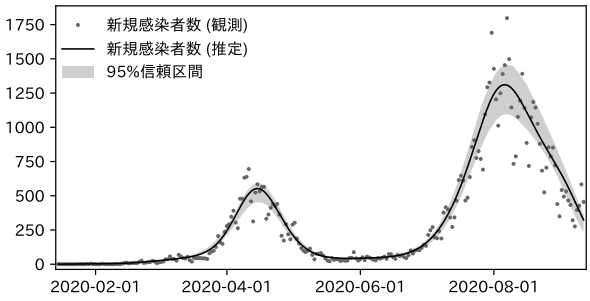


図 2　新規診断者数の推定値とモデルの適合.

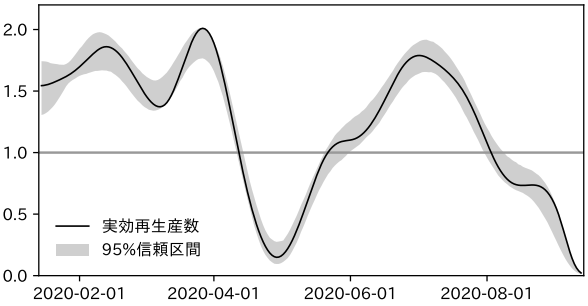


図 3　実効再生産数の推定値.

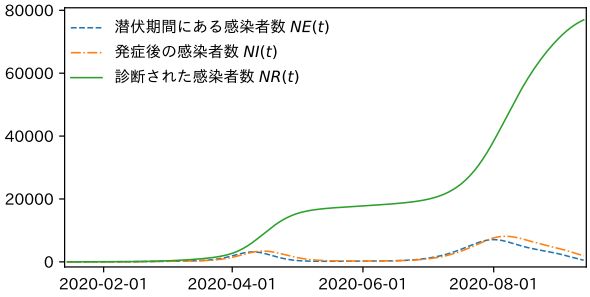


図 4　潜伏期間にある感染者数, 発症後の感染者数, 診断された感染者数の推定値.

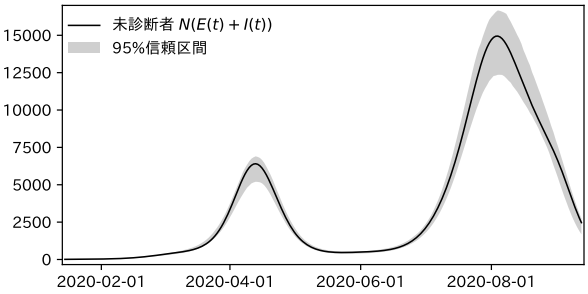


図 5　未診断者数の推定値と 95% 信頼区間.

ルスの 1 日毎の新規感染者数を使う. $M = 241$ とし, $t = 1$ から $t = M$ までのデータを使って推定を行う. 先行研究から $\alpha = 0.2$, $\gamma = 0.161$ とする. 日本の総人口は $N = 126000000$ とする.

常微分方程式 (1)-(4) は初期値を $E(0) = 6$, $I(0) = 3$, $R(0) = 0$ とし, ステップ幅 0.1 の 5 次の Runge-Kutta 法で解く. 罰則付き最小二乗法によって $\beta(t)$ を推定した. 平滑化パラメータは交差検証によって定め, 95% 信頼区間については定常ブートストラップで得た 1000 組のブートストラップ標本から, パーセンタイル法で計算した.

4　推定結果と考察

図 2 は新規感染者数の推定値と観測値である. モデルの適合を確認できる. 図 3 は実効再生産数の推定値である. 診断された感染者は診断日から約 11 日前に感染したと考えられるので, $t = M$ 付近の $\beta(t)$ の推定値は, ほとんど罰則項だけによって定められている. そのため, そのような t での推定値は, さほど信用できる値ではないことに注意が必要である. 4 月 7 日以前の $\beta(t)$ の最小値は約 0.101, 最大値は約 0.179 であり, 緊急事態宣言の期間 4 月 7 日から 5 月 25 日の最小値は約 0.013 であるので, 8 割程度の接触の削減は達成されていたと考えられる. 緊急事態宣言が解除された後, 7 月には新規感染者数の増加が急になり, 8 月上旬には新規感染者数のピークが来た. この第二波については, 緊急事態宣言など大きな対策は取られなかったが, 実効再生産数は 6 月下旬から 7 月上旬でピークとなり, その後減少し, 8 月に入ると 1 以下となっている.

図 4 は潜伏期間にある感染者数 $NE(t)$, 発症後の感染者数 $NI(t)$, 診断された感染者数 $NR(t)$ の推定値である. 図 5 は未診断の感染者 $N(E(t) + I(t))$ の推定値と 95% 信頼区間である. 8 月 31 日には未診断の感染者が 6938 人 (95% 信頼区間 $[5889, 8017]$) いたと推定される.

本研究では, 新型コロナウイルス感染症の拡大を表すために SEIR モデルを考え, 常微分方程式の B-スプラインで表される時間に依存したパラメータを推定し, 実効再生産数を得た. 加えて, 緊急事態宣言のときの接触の削減について, 定量化することができた.

構造出力デザインのための能動学習とその材料科学問題への応用

松井孝太 [1], 草川隼也 [2], 安藤圭理 [1], 沓掛健太朗 [3], 宇治原徹 [1,4], 竹内一郎 [2,3,5]

[1] 名古屋大学, [2] 名古屋工業大学, [3] 理化学研究所革新知能統合研究センター,
[4] 産業技術総合研究所, [5] 物質 · 材料研究機構

ある関数 f に対して, 与えられた所望の出力 $\boldsymbol{f}_0 \in \mathcal{Y}$ から

$$\boldsymbol{f}_0 = f(\boldsymbol{x}_0)$$

を満たすような入力 $\boldsymbol{x}_0 \in \mathcal{X}$ を見つける問題は逆問題と呼ばれる. 例えば材料科学では, シミュレータを使用した数値計算が未知の実験条件における実験結果を予測するために広く用いられている. これは, 実際に実験を行うためには時間, 資源および人的に高いコストが要求されるためである. ただし, 計算機によるシミュレーションにもしばしば大きな計算コストがかかるため, 大量の計算リソースが必要となる. 従って, 有用と思われる実験設定を事前に特定できれば, その設定に対してのみシミュレーションまたは実際の実験を実行することにより, 不要なコストを削減することができる. また, 材料科学における多くの問題は, f がブラックボックスであること, データの欠如や観測ノイズなどの様々な理由によってモデリングを行う際に不確実性が伴う. このような状況で, 不確実性を考慮せずに f の推定を行えば, 意思決定に重大な間違いが混入する可能性が高まる.

本研究では, 観測誤差を伴った構造出力を持つブラックボックス関数 f の逆問題を考察する. ここで, 構造出力とは, 要素間に構造のあるベクトル値出力を指す. このような構造として, 例えば「近い要素における関数値は類似した値を示す」というものなどが考えられる. 提案法は f に対して多出力ガウス過程モデルを仮定し,ベイズ的能動学習によってモデルによる予測と所望の出力との二乗誤差関数を最小化する. これを実現するために, 相関のあるベクトル値出力から定まる二乗誤差が従う確率分布を導出し, 本研究の問題設定における能動学習のための獲得関数を導出した. また, 提案法を材料科学における炭化ケイ素 (SiC) 結晶成長モデリングのシミュレーションデータに適用し, その有効性を検証した.

ガウス過程分位点回帰モデリングと新生児ホルモン分泌リズムの解明

田辺佑太　　荒木由布子

静岡大学大学院総合科学技術研究科 情報学専攻

概要

本研究では, 非線形構造を有する経時データにおいて, 分位点の推移を捉えるため, 階層ベイズモデルによるガウス過程分位点回帰を提案する. 提案モデルの利点は, 階層ベイズモデリングにより, 経時測定データにおいて無視することのできない反応の個体差・同一個体のデータ相関を考慮したうえで, ガウス過程によって, 事前にパラメトリックな非線形式を特定せずに非線形構造を捉えられる点である. また, 分位点回帰によって, 平均を推定する通常の回帰では得られない有益な情報を得ることが期待できる. 提案モデルを事前にパラメトリックな非線形式を特定できない新生児ホルモン分泌量の経時測定データに適用することで, 未解明であった新生児の 1 日のホルモン分泌リズムを明らかにした.

1　階層ベイズモデルによるガウス過程分位点回帰

分位点回帰 (Koenker and Basset,1978) は, 目的変数の条件付き分位点を推定するための手法である. 分位点回帰では, 平均値ではなく中央値を対象とできるため, 目的変数の分布が非対称である場合に適応でき, 外れ値に対して頑健である. また, 任意の複数の分位点を推定することにより, 結果の比較からより多くの情報が得られることも分位点回帰の利点である. 本研究では, 特に非線形構造を有する経時データにおいて, 分位点の推移を捉える事を目的とし, 階層ベイズモデルによるガウス過程分位点回帰を提案した. 同様の目的から, 事前にパラメトリックな非線形式を指定する必要のある非線形分位点混合効果モデルが Geraci (2017) によって提案されているが, 現象を適切に表した非線形式を事前に特定することは困難なことが多い. 本研究で提案モデルの適用を行う新生児のホルモン分泌量の経時データもその一例であり, 未解明である新生児の 1 日のホルモン分泌リズムの表現に特定の非線形式を指定することはできない. 提案モデルでは, ガウス過程を用いることによって, 非線形式を事前に特定することなく, データから非線形構造を柔軟に捉えることが可能である. また, Geraci and Bottai (2007) では, 分位点混合効果モデルにおいて最尤法による推論を行う際, 点推定に MCEM アルゴリズム (Monte Carlo Expectation Maximization algorithm) による反復計算を用いる方法が取られている. この場合区間推定を行うためにはブートストラップ法を用いる必要があるため, 反復計算が必要な点推定をさらに繰り返すことになり, 計算量が膨大になる. そこで本研究では, 階層ベイズモデルを構築し, ベイズ推定によって事後分布を求めることで信用区間を得る方法を考えた. ガウス過程は, 関数の事前分布を指定し, 観測データに基づいてその事後分布を更新する手法であるため, 階層ベイズモデルに組み込むことが容易である.

2　数値実験及び実データへの適用

提案モデルは, 分位点の非線形遷移を適切に捉えるためのモデルである. この目的を達成できているかを評価するためには, 真の分位点構造と推定された分位点構造の差の大きさを評価する必要がある. つまり, 真の分位点構造が明らかなデータを発生させたうえで数値実験を行う必要があるが, このための方法が Takeuchi et al.(2006) において示されいる. Takeuchi et al.(2006) を参考に分位点が任意の真の非線形構造を持つデータを実際に生成し, 提案モデルの推定精度を評価する.

数値実験によるモデルの評価の後, 提案モデルを実データへの適用する. 分析対象のデータは, 生後 31 から 124 日（約 1〜4 か月）の 12 名の新生児を対象に, その尿中コルチゾール濃度が久留米大学医学部にて, 数日間繰り返し測定して得られたデータである.1 日あたりの測定時点数と測定間隔は個人ごとに異なり, 疎である経時測定データであるため, 階層ベイズモデルが適していると考えられる. コルチゾールの分泌リズムは非線形構造を持つことが知られているが, 新生児の分泌リズムは明らかにされていないため, 非線形式をモデルに事前に割り当てられない. そこで提案モデルを適用することを考え分析を行った.

参考文献

[1] Araki, Y., Konishi, S., Kawano, S. and Matsui, H.(2009). Functional regression modeling via regularized Gaussian basis expansions. *Ann Inst Stat Math*, **61,** 811–833.

[2] Booth, J. G. and JP. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society (Series B) Statistical Methodology* **61,** 265–285.

[3] Geraci, M. (2017). Nonlinear quantile mixed models. arXiv Preprint arXiv:1712.09981.

[4] Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing* **24,** 461–479.

[5] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8,**140–154.

[6] Hastings W. K.(1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109.

[7] Hoover, D., Rice, J., Wu, C., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85,* 809–822.

[8] Koenker, R.(2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91,** 74–89.

[9] Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94,** 1296–1310.

[10] Koenker, R. and Bassett, G.(1978) Regression quantiles. *Econometrica* **46,** 33–50.

[11] Laird, N M. and Ware, J H.(1982). Random-Effects Models for Longitudinal Data. *Biometrics***38,** 963–974.

[12] Levine R. A. and Casella G.(2000). Implementations of the monte carlo em algorithm. *In Journal of Computational and Graphical Statistics* **10** 422–439.

[13] Liu, Y. and Bottai, M.(2009). Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics* **5,** 1–22

[14] Matsui,H., Misumi,T., Yokomizoand, T. and Konishi, S.（2016）. Clustering for Functional Data via Nonlinear Mixed Effects Models, *Japanese Journal of Applied Statistics 56* 25–45.

[15] 持橋大地, 大羽成征 (2019) ガウス過程と機械学習, 講談社.

[16] McCulloch, C. E. (1997)Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *J. Am. Statist. Ass* **92,** 162–170.

[17] Rice, J, A. and Wu, C, O(2001). Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves *Biometrics* **57,** 253–259.

[18] Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric Quantile Estimation. *Journal of Machine Learning Research*, **7,** 1231–1264.

[19] Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.* **85,** 699–704.

[20] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics and Probability Letters* **54,** 437–447.

[21] Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: applications and current research areas. *The Statistician* **52** 331–350.

薬物動態パラメータの予測から目指す創薬支援

滋賀大学データサイエンス教育研究センター　江崎剛史

創薬の研究開発の初期段階において，薬の種となる化合物の毒性予測を行うことは，その薬の種の実用化に向けた可能性を適切に評価し，改良点を明確にする上で重要であると考えられる．1998 年から 2008 年の間で欧米およびアジアの市場から撤退した医薬品のうち，60% 以上が予期せぬ副作用として発現した毒性（うち，心毒性が 33%，肝毒性が 29%）を理由としていた，との報告もある[1]．一度市場に出回った医薬品が重篤な副作用を起こすことは，製薬企業で創薬研究に費やされる研究機関や費用は膨大であるため，市場撤退は製薬企業の経営に重大な影響を及ぼす要因ともなり得るだけでなく，国民の保健と福祉を脅かす要因にもなる．このため，薬の種が毒性を示す傾向を正確に予測することは，アカデミア発創薬・製薬企業において有益であり，産官学が一丸となって取り組む意義は大きい．

このような背景から，日本医療研究開発機構の支援によって「創薬支援インフォマティクスシステム構築プロジェクト」が立ち上がり，薬物の吸収・分布・代謝・排泄といった体内動態（薬物動態）と毒性（心毒性・肝毒性）予測を目的とした *in silico* 統合解析プラットフォームの構築と，そのプラットフォームを広く公開する環境整備が行われた[2]．本講演では，演者が携わった薬物動態のパラメータ予測システムの開発を中心に講演を行った．

本プロジェクトでは，実験によって得られた実測データを大量に収集し，深層学習を含む機械学習などを用いて「薬としての性質」を予測するモデルを構築した．さらに，血中の薬物濃度推移を予測する手法として，生理学的薬物動態モデルの組み込みを行った．このモデルは，体内の各組織を箱として捉え，箱に出入りする薬物の量を微分方程式で記述する．モデルの係数には，薬物動態に関する化合物のパラメータが使われる．これら化合物の薬物動態パラメータをいかに精度良く予測するかが重要となるため，大量の実験データの収集を目指して，公共データベースを利用することとした．

公共データベースは，論文や特許で発表されたデータを独自に収集して格納しているが，実験条件の詳細が登録されていない，単位の表記法が統一されていないなどの問題がある．そして稀ではあるが，間違った実験値が格納されていることもあり，データの精査が十分に行き届いていないのが現状である．また，公共データベースから収集できるデータの数は膨大であるが，企業やアカデミアなどの研究室で用いる実験条件は多岐にわたるため，ある特定の実験条件に注目した予測モデルの構築に使用できるデータは多くない．例えば化合物の水への溶解性を求める実験には，測定方法，有機溶媒の有無，設定する pH（血液中の溶解性を評価するためには pH6.2 から 7.4 が多いが，胃内では pH 1.2 が用いられることもある）といった条件の違いがある．また，ヒト肝ミクロソームを用いた代謝安定性試験では，化合物濃度，単位など，実験条件やデータの表記方法に高い多様性がある．

そこで我々は，実験条件や単位が異なるデータをまとめて予測モデルを構築することは，

予測精度の低下につながる可能性があることを明らかにし[3]，公共データの精査を行い，そのデータを格納したデータベース（DruMAP）を構築した．こうして得られた精査済みデータを用いて薬物の血中濃度推移の予測に必要な各種パラメータの予測モデルを作成した．また，さらに精度の高い予測モデルを構築するために，データ収集システムを構築することで豊富なデータの蓄積を目指した．

そして，溶解性・血液血漿タンパク結合率・膜透過性・肝ミクロソームを用いた代謝安定性などの薬物動態パラメータの予測モデルを構築した．その結果，既存の予測モデルよりも同等，あるいはより高い予測精度を持つモデルを構築することができているが，まだ改善の余地があると考えている．現在はこれらをまとめて統合システムとして公開しているが，実測値が不足している範囲のデータ収集や，それを補う新たな手法を用いた予測モデルの構築により，さらなる高い精度のシステムへと改良していきたいと考えている．

創薬研究の現場において，研究段階が進行するほど時間や費用に関わるコストは増大する．このことから，機械学習などの情報科学的な手法によって構築したモデルを用いた薬物動態・毒性予測を，研究開発の初期段階に導入することはリスク低減の観点からも有望である．本プロジェクトで構築したシステムを，創薬を指向する研究機関・製薬企業でも広く利用可能とすることで，創薬におけるインフォマティクスの役割が一新され，従来の研究開発のあり方に大きな変革をもたらすことが期待される．

【参考文献】

1. J. S. MacDonald, R. T. Robertson, Toxicity testing in the 21st century: a view from the pharmaceutical industry, *Toxicol. Sci.*, 110, 40–46 (2009)
2. 江崎剛史，渡邉怜子，川島和，夏目やよい，水口賢司，創薬支援インフォマティクスシステム構築プロジェクト：薬物動態，毒性の統合解析プラットフォーム，薬剤学, 77(4), 211-215 (2017)
3. T. Esaki, R. Watanabe, H. Kawashima, R. Ohashi, Y. Natsume-Kitatani, C. Nagao, K. Mizuguchi, Data Curation can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance, *Mol. Inf.*, 38, 1800086 (2019)

遺伝子ネットワーク構造が予測精度に与える影響

九州大学大学院数理学府　沖永　悠一
兵庫県立人と自然の博物館自然・環境マネジメント研究部　京極　大助
トヨタ自動車（株）アグリバイオ事業部農食事業室　近藤　聡
龍谷大学農学部　永野　惇
九州大学マス・フォア・インダストリ研究所　廣瀬　慧

1

■ 概要

- 生物のオミックスデータ（遺伝子の発現データ等）から，その生物の特性の予測を考える．

- 一般的な課題として，遺伝子データが高次元になるため，それに対応した回帰モデリングを考える必要がある．
- LASSOや主成分回帰（PCR）は，高次元回帰における一般的な手法である．今回は，この２つの手法を予測において用いた．

2

■ 概要

- さらに，遺伝子ネットワーク構造についても考慮する．
- 遺伝子ネットワークの特徴として，スケールフリー性が挙げられる．スケールフリー性とは，ネットワークの次数分布が以下のようにべき乗則で表現されることである．

$$p(x) \propto x^{-\gamma} \ (\gamma > 1)$$

- 多くの頂点は他の頂点とほとんど繋がっていない一方で，一部の頂点が多くの頂点と繋がっている状態である．

3

■ 概要

- 遺伝子ネットワーク構造の違いによる予測精度の差を比較するために，所与のデータをもとに３種類の共分散行列（$\Sigma_1, \Sigma_2, \Sigma_3$）を用意した．
- Σ_1 …… 標本共分散行列
- Σ_2 …… Σ_1をもとに作成された，スケールフリー性を仮定していない共分散行列
- Σ_3 …… Σ_1をもとに作成された，スケールフリー性を仮定した共分散行列
- この度は，遺伝子ネットワーク構造がLASSOおよびPCRの予測精度に与える影響を，サンプル数を変化させながらモンテカルロシミュレーションによって調べた．

4

■ シミュレーションの手順

1. 所与のデータを用いて，共分散行列 $\Sigma_1, \Sigma_2, \Sigma_3$ を作成する．
   使用データ：ハクサンハタザオという植物の遺伝子データ
   （次元数：17205，サンプルサイズ：835個）

Σ_2, Σ_3 の作成のために，それらの逆行列 Ω_2, Ω_3 を以下の式で推定する．

$$\hat{\Omega}_j = \arg\min_{\Omega} \{\log|\Omega| - \mathrm{tr}(\Omega S) - P_j(\Omega)\} \quad (j = 2, 3)$$

ここで，

$$\begin{cases} P_2(\Omega) = \rho \sum_{i=1}^{p} \|\boldsymbol{\omega}_{-i}\|_1 \\ P_3(\Omega) = \rho \sum_{i=1}^{p} \log(\|\boldsymbol{\omega}_{-i}\|_1 + a_i) \end{cases}$$

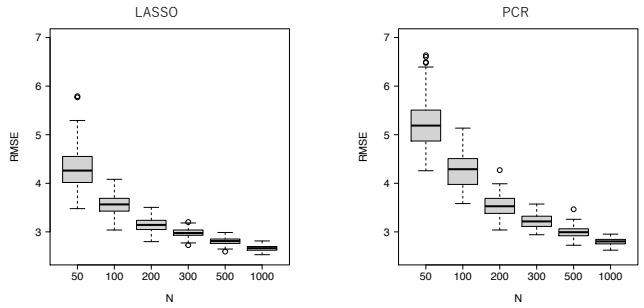
5

■ シミュレーションの手順

2. 遺伝子データが多変量正規分布に従っていると仮定して，$\Sigma_1, \Sigma_2, \Sigma_3$ からそれぞれ N 個（N はサンプルサイズ）のデータを生成する．
3. 生成した３つのデータセットそれぞれに対して，トレーニングデータとテストデータに分割してLASSOおよびPCRを行なう．予測誤差はRMSE(root mean squared error)を用いて算出する．
4. 2と3を，N の値を50, 100, 200, 300, 500, 1000と変えながらぞれぞれ100回ずつ繰り返す．（結果は箱ひげ図で描画する．）

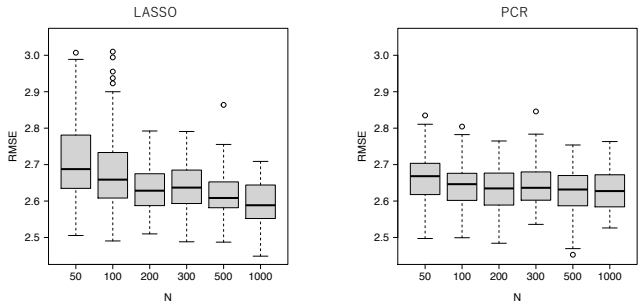
6

■ シミュレーション結果（Σ_1）



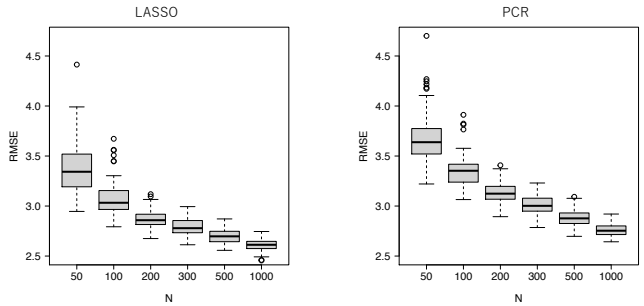
7

■ シミュレーション結果（Σ_2）



8

■ シミュレーション結果（Σ_3）



9

■ 結果

- サンプルサイズ N の増加に伴い，LASSOとPCRどちらの場合も，Σ_1, Σ_3の場合はRMSEの値が減少したのに対し，Σ_2の場合はほとんど変化がなかった．
- Σ_3において，N を十分大きくした場合のRMSEと，Σ_2における N = 50 の場合のRMSEが同じくらいの値をとっている．これは，Σ_2の相関が弱く，β の非ゼロパラメータの大部分が小さい値をとっていたため，N = 50 でも高い精度であったのだと考えられる．
- 全体的に，PCRよりもLASSOのほうが精度が高かった．これはPCRにおいて，小さな固有値に関連する予測子が予測精度に影響を与えている可能性がある．

10

■ おわりに

- 遺伝子ネットワークがランダムな場合とスケールフリー構造を持つ場合の高次元回帰の予測精度について，サンプル数を変化させながら数値シミュレーションを行なった．
- スケールフリー構造が予想される場合は，特性の予測を行なう際に，高い予測精度を得るために比較的大きなサンプルサイズが必要となることを示した．
- サンプルサイズが小さい場合の高い予測精度が得られる手法の開発は，今後の課題である．

11

A novel metric for hyperbolic phylogenetic tree embeddings

松本　拡高 [1,2], 三森　隆広 [3], and 福永　津嵩 [4]

[1] 長崎大学　情報データ科学部
[2] 理研 BDR　バイオインフォマティクス研究開発チーム
[3] 理研 AIP　医用画像解析チーム
[4] 東京大学大学院　情報理工学系研究科コンピュータ科学専攻

DNA シーケンシング技術の発達により様々な生物種のゲノム配列が決まり、遺伝子レベルや生物種レベルで様々な規模の系統樹が計算機的に再構築されるようになった。これら再構築された系統樹を用いた進化系統解析により、遺伝子間の関連解析とそれに伴う遺伝子の機能解析、進化と疾患の関連解析、感染症のダイナミクス、バクテリアの分類と同定の研究など、様々な分野で重要な知見がもたらされている。特に、容易にゲノムデータが得られる今、大規模データから遺伝子と表現型の関連解析などの研究を行うために新しい系統解析アルゴリズムが必要とされるなど、計算機的手法の需要が高まっている。

一方で、1 細胞シーケンシング技術やゲノム編集を用いた系譜記録技術などの発展にともない、1 個体内の発生・分化に伴う細胞の系統樹である細胞系譜が再構築され、進化系統解析と同様の解析が進められている。また、がんゲノム研究や免疫ゲノム研究においても、進化系統解析に基づいた研究が注目を集めている。このように、様々な研究分野において、進化系統解析の枠組みを用いた解析が有効であると考えられ、様々なデータ及び目的に適した新しい解析技術の需要が高まっている。したがって、系統解析の新しい理論やアルゴリズムの開発は、進化研究のみでなく、発生やがん研究、疫学研究などの多様な分野に貢献すると考えられる。

近年発展著しい人工知能分野において、非ユークリッド空間の 1 つである双曲空間を用いた研究が注目を集めている。階層構造が背後に存在する表現学習において、ユークリッド空間に埋め込むよりも、双曲空間に埋め込んだ方が顕著に性能が良かったという報告をきっかけに、双曲空間を用いたニューラルネットワークなど、様々な機械学習のアプローチが提案されてきている。また、単に既存の機械学習法をユークリッド空間から双曲空間へ置き換えるというだけでなく、双曲空間の特徴を活かして、階層的クラスタリングの問題を双曲空間上での座標の最適化として考えるという、新しいアプローチも提案されている。

双曲空間の表現方法としては、2 次元の場合は半径 1 の開円板であるポアンカレ円板を用いた表現がある（$D > 2$ のときは D 次元ポアンカレ球と呼ばれる）。ポアンカレ円板では、二点間の最短線は図 1(A) のように曲線を描く。このような最短線の特徴から、双曲空間上で階層構造、つまり木構造を上手く表現することが出来る。なぜなら、双曲空間上での最短線は曲線を取るが、それが木構造に近い形となり、ユークリッド空間よりもうまく木構造上の距離を表現できるからである（図 1(B)）。また、ポアンカレ円板では原点から離れるほど空間が著しく広がっており、これが木構造の枝が分岐してノードの数が増えることに対応し、木構造の広がりをうまく表現することが可能である（図 1(C)）。以上のことから、系統樹も双曲空間でうまく表現することが出来、双曲空間を用いた系統樹の可視化法が提案されている。さらに、最近では 1 細胞 RNA-seq から細胞系譜を再構築する上で双曲空間への埋め込みを活用した研究も行われている。その他にも、タンパク質相互

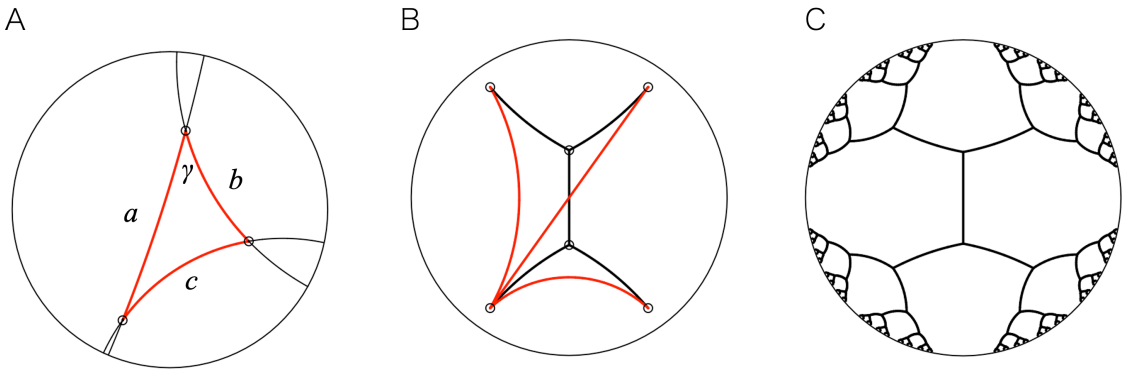


図 1　ポアンカレ円板の概要図　(A) ポアンカレ円板上の最短線と三角形　(B) 簡単な木構造を埋め込んだ場合のノード間の最短線　(C) 全ての枝長がポアンカレ距離として同じとなるような木構造の埋め込み例

作用ネットワークを双曲空間に埋め込むと、座標を元にタンパク質の機能的な意義をなどを捉えることができることや、生物の機能として嗅覚の認識構造が双曲空間によって理解できるというような報告もあり、双曲空間を用いた生命科学研究が広がりを見せている。以上のように、階層構造を持つ生物データの解析において双曲空間への埋め込みは有効で、系統樹解析においても双曲空間を用いることが有効なアプローチであると期待される。しかしながら、双曲空間を用いた系統解析の理論やその応用に関する研究はほとんどなく、発展途上であるといえる。本研究は、双曲空間を用いた系統解析のための新しい距離を提案するとともに、それに基づく応用解析を提案した。

高次元統計解析に基づく遺伝子発現データのノイズ削減法

井元 佑介 [1]，平岡 裕章 [1,2,3]，吉脇 理雄 [3]，Emerson G. Escolar [3]，
中村 友紀 [1,4]，山本 拓也 [1,5]，斎藤 通紀 [1,2,5]

[1] 京都大学高等研究院ヒト生物学高等研究拠点，[2] 京都大学高等研究院高等研究センター，
[3] 理化学研究所革新知能統合研究センター，[4] 京都大学医学研究科，[5] 京都大学 iPS 細胞研究所

E-mail: imoto.yusuke.4e@kyoto-u.ac.jp

本講演では，高次元統計解析，特に青嶋–矢田理論 [1, 4] を一部で応用して，シングルセル遺伝子発現データの観測ノイズを削減する手法を提案する．本研究成果の参考資料として，特許発明情報（TLO 京都・発明情報：https://www.tlo-kyoto.co.jp/patent/post-490.html）もご参照いただきたい．

近年，ゲノム解析技術が急速に発展しており，シングルセル遺伝子発現解析（single-cell RNA sequencing, scRNA-seq）という技術によって，単一細胞内の全遺伝子の RNA 発現情報（シングルセル遺伝子発現データ，scRNA-seq データ；図 1 参照）が収集できるようになった．遺伝子発現は細胞の変化（分化）を支配する重要な要因であるため，シングルセル遺伝子発現データのデータ解析によって細胞分化の構造を理解しようという試みが世界中で行われている．しかしながら，遺伝子は数万種以上存在するため，シングルセル遺伝子発現データは超高次元データとなる．さらに，シングルセル遺伝子発現データはサンプル数（細胞数）は高々数千程度，観測ノイズが発生，時間連続なデータが採取できないといった特徴がある．これらの特徴は，従来のデータ解析手法が想定していたデータの性質の範疇を超えているため，従来のデータ解析手法ではシングルセル遺伝子発現データが持つ情報を正しく活用できない．特に，次に述べる観測ノイズ付きの高次元データで発生する “**次元の呪い**” は深刻な問題である．

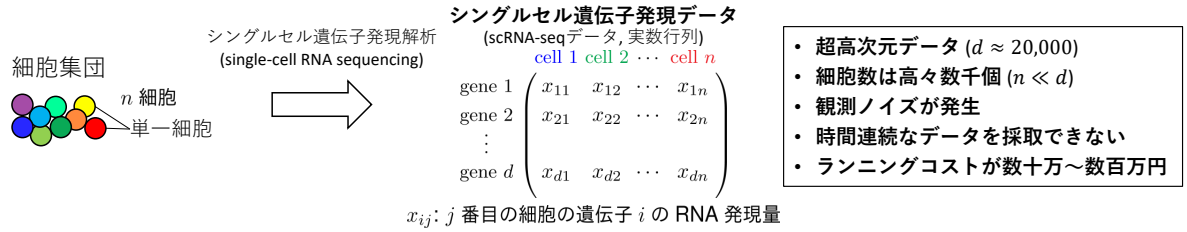


図 1　シングルセル遺伝子発現データ．

次元の呪いとは，観測ノイズを含む高次元データの距離や統計量を計算するときに，ノイズが蓄積することによって，真のデータの情報を正しく抽出できない問題である [2, 3]；図 2 参照．シングルセル遺伝子発現解析では，RNA 分子のコピーと増幅を繰り返し実施し，ライブラリ DNA と呼ばれる分子に変化させた後に次世代シーケンサーでランダムサンプリングを行うことで遺伝子発現情報を抽出する．このときに，ドロップアウト（コピーの失敗），増幅が一定でない，全ての RNA をサンプリングできないといった技術的問題が原因で観測ノイズが発生してしまう．さらに，次元数に相当する遺伝子数は哺乳類では 2 万種以上存在する．したがって，シングルセル遺伝子データは次元の呪いによる悪影響を避けられない．この問題は近年，本研究分野で認知されつつあり，次元削減などの前処理による対策が行われている．しかしながら，これらの前処理ではノイズ以外の情報も失ってしまうため，単一細胞が持つ詳細な情報を十分に活用することができない．

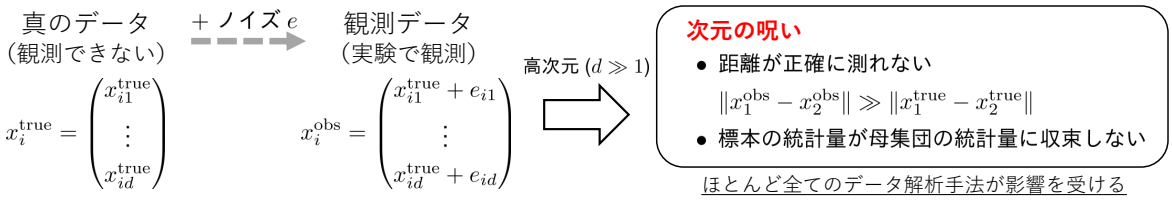


図 2　次元の呪いの概略図.

そこで本研究では，次元を削減することなく，シングルセル遺伝子発現データの観測ノイズのみを削減する手法 “**RECODE**”（resolution of the curse of dimensionality）を提案する．本手法はシングルセル遺伝子発現データの観測ノイズを統計的に分析し，そのノイズ分布の特性に基づいてノイズ部を分離し，さらに，固有値修正法 [4] に基づいて非ノイズ部を修正する手法である；図 3 参照．本手法の主な特徴は，データフォーマットを修正しないので，全てのデータ解析手法の前処理として利用可能であること，パラメータが不用であること，実験プラットフォームやバージョンには依存しないことが挙げられる．本講演では，提案手法の実験データへの適用事例も紹介する．

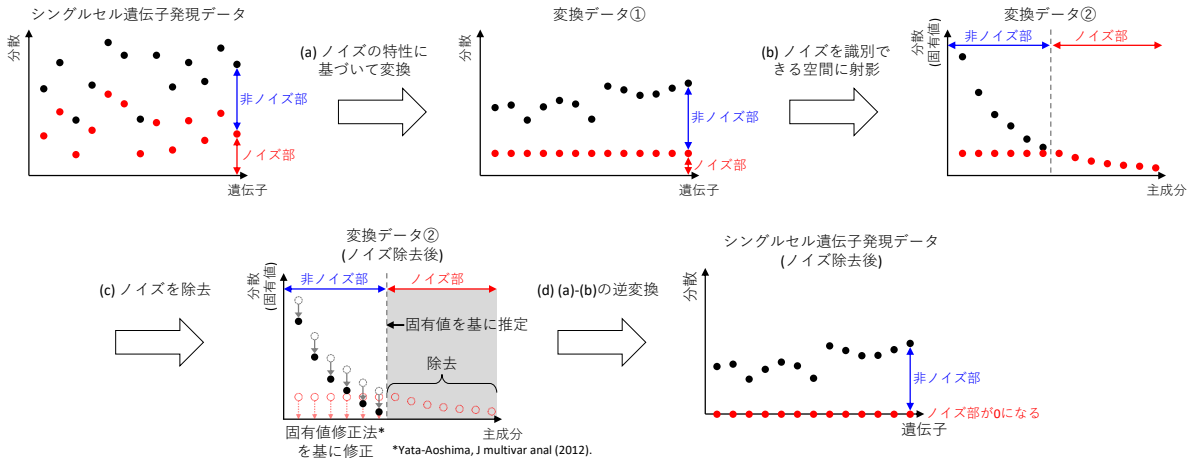


図 3　ノイズ削減法（RECODE）の概略図.

参考文献

[1] 青嶋誠，矢田和善．高次元の統計学（統計学 One Point 11 巻）．共立出版, 2019.

[2] P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *J. Royal Stat. Soc. Series B*, 67(3):427–444, 2005.

[3] K. Yata and M. Aoshima. PCA consistency for non-Gaussian data in high dimension, low sample size context. *Comm. Stat. Theory Methods*, 38(16–17):2634–2652, 2009.

[4] K. Yata and M. Aoshima. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivar. Anal.*, 105(1):193–215, 2012.

植物科学分野のデータに対する統計科学・機械学習的アプローチと作物研究への応用

持田恵一
理化学研究所環境資源科学研究センター
横浜市立大学木原生物学研究所
岡山大学資源植物科学研究所
理化学研究所バトンゾーン研究推進プログラム

近年の計測・測定技術の発展に伴い、植物あるいは作物の研究分野でも大量で複雑な形式を持つデータが高速かつ大量に取得されるようになり、それらのデータを活用した研究が幅広く進められている。生物分野での計測・測定技術の発展は、実験室における植物研究だけでなく、作物の生産性の向上を目的とした研究でも新しいアプローチを可能にしている。これらの背景をうけて、本講演では、私たちの最近の研究から、植物に由来するデータに関して、バイオインフォマティクスはもとより統計科学や機械学習的な解析を交えた研究を紹介した。

遺伝子発現データにもとづいて機械学習による遺伝子発現ネットワークの推定を有用遺伝子探索に利用した例として、イネ科作物の病害の 1 つである紋枯病(もんがれびょう)の抵抗性に重要な遺伝子を同定した研究について紹介した。紋枯病は、イネやトウモロコシなど穀物の重要病害であり、土壌に生息する植物病原糸状菌 *Rhizoctonia solani* によって引き起こされるが、紋枯病に対する新たな防除策の開発が求められている。モデル草本植物のミナトカモジグサ集団から見出された、紋枯れ病抵抗性が異なる系統の感染後時系列トランスクリプトームデータについて、機械学習アルゴリズムの 1 つであるランダムフォレスト法に基づく遺伝子制御ネットワークの推定を行い、抵抗性が異なる系統間でネットワークを比較した。それにより、BdWRKY38 遺伝子と BdWRKY44 遺伝子は、抵抗性系統のネットワーク特異的にネットワークハブを構成していることを見出した。これらの二つの遺伝子の発現を抑制した系統を作出し、紋枯病抵抗性の変化を観察したところ、BdWRKY38 遺伝子でより顕著に抵抗性の喪失が観察されたまた、ミナトカモジグサの紋枯病罹病性系統で、BdWRKY38 遺伝子を恒常的に発現すると紋枯病の感染に対して顕著な抵抗性を示したことから、これららのハブ遺伝子は、紋枯病抵抗性に関わる遺伝子発現を制御する因子であることが示唆された。

また、農業上重要な作物の形質(穀物の収量や果実の質など)に関する時系列データ解析の応用研究を紹介した。農業上重要な作物の形質はしばしば成長の後期に顕在化するものであり、それまでの成長履歴における植物と環境の相互作用の積み重ねの結果を反映していることから、作物の成長にそって、植物と環境の相互作用を理解するとともに、その相互作用に関わる遺伝要因を明らかにすることは、気候変動などの環境変化に対して適応性が高

い作物を育種することや、環境変化に伴う収穫物の変化を予測することで収益を安定化する精密農業に有用であると考えられる。主要穀物の大麦について、その出穂期を農業形質の事例として、大麦の成長過程において、どのような遺伝子型と環境との相互作用が、出穂期の早晩に影響するか明らかにすることを目的として、ゲノム・エピゲノム・トランスクリプトーム・植物ホルモンの精密分析といった網羅的な生物データの取得とともに、統計科学・機械学習的アプローチを取り入れた解析を進めている。多様な大麦 274 系統について過去 20 年以上にわたり圃場で調査された出穂期データが蓄積されている。1996 年から 2016 年にわたって計測された出穂期データと、大麦においてこれまでに同定されている出穂関連遺伝子の DNA の多型データを用いて、それらの遺伝的多様性と出穂性との関連性を調べた。大麦の系統と生育シーズンを説明変数とし、播種日から出穂日までの日数を目的変数とした回帰モデルは、自由度調整済み決定係数が 0.89 であり、出穂日の変動性の約 90% を説明するといえた。一方、出穂関連遺伝子の DNA の多型データと生育シーズンを説明変数とした場合は、自由度調整済み決定係数は 0.61 であり、約 30%の説明不足を解消する遺伝要因が依然として潜在していることが示唆された。さらに、様々なセンサーやイメージング技術と、深層学習等を活用したコンピュータビジョンの技術の進展は、植物や作物の形質データを収集するフェノーム解析において応用範囲を拡大している。フェノーム解析技術の進展は、作物の成長状態と生育環境の関係性を調べることを可能にしつつある一方、高次元データの次元削減の手法の重要性が増している。

最後に、市民参加型のデータ収集の事例として株)ユーグレナが進める「みんなのミドリムシプロジェクト」を紹介した。日本全国各地からミドリムシを収集し、新しい特徴を持つものを探し出す市民参加型プロジェクトを 2019 年から開始した。一般の方からの協力を得て、全国のさまざまな地域にいるミドリムシを収集し、理研の解析基盤により解析することで、新たな機能の発見につなげることを目指している。そのため、日本各地のサンプルの収集場所やミドリムシの有無などを統合する地図システムを作成し、参加者の方々と研究室をつなぐインタフェースとして活用している。将来的に、多くのサンプルの情報が掲載されれば、サンプルの地域性を見出したり、他の生態地理的なデータと統合したりすることによる新しい発見を期待している

植物の表現型が、植物と環境のどのような相互作用により表現型が決まっているかを理解するために、生物学、コンピュータ科学、数学、工学をはじめとする分野を横断した取り組みが求められる。また、農作業の省力化や農作物の生産性の向上へのニーズの高まりは、作物の栽培自動化や変動環境における成長予測のための要素技術として IT・IoT 分野の融合が進みつつある。これらの背景を受けて、植物科学分野の多様なデータに関する、統計学・機械学習的アプローチの重要性を今後ますます高まると考えられる。

概念の複雑性、ならびに『きわめて大量だが信頼性の保証がない言語情報』の取り扱いについて

得丸久文(カラハリプロジェクト)

１．はじめに：学際的な科学概念と「信頼性の保証がない言語情報」と取り組む必要性

「多様な分野のデータに対する統計科学・機械学習的アプローチ」について考えるにあたり，概念の複雑性について我々がほとんどなんの理論も解析手法も持ち合わせていないこと、インターネットと検索エンジンが提示してくれる「きわめて大量だが信頼性の保証がない学際的な言語情報」をどう処理すればよいかについてもまだ手法が構築されていないこと、この二つを避けて通ると、何を学習するにしても、意味を正しく受容する保証がなく、学習する意味がない。

２．概念の意味の複雑性

① 先行研究：概念の意味を正しく受容することを考察した先行研究としては、ヴィゴツキー「思考と言語」、ピアジェ「知能の心理学」がある。

② 群論理の適用：　ピアジェは群性体という言葉を用い、合成性、可逆性、連合性、同一性、同義性の５つの条件を満たすことを概念に要求する。これは概念操作の結果を意味あるものとするためには、数学的群論の条件を満たせばよいということである。

③ 概念の複雑さの次数管理：　概念の複雑さの次数管理：ヴィゴツキーは科学的概念と生活概念を区分し、科学的概念を真の概念とする。「コトバの暗記それとのとの結合それ自体は概念の形成をもたらさない．（略）被験者のまえに，概念を形成することなしには解くことのできないような課題が発生しなければならない．」「この概念あるいはコトバの意味の発達過程は，有意的注意，論理的記憶，抽象，比較，区別のような一連の機能の発達を要求する．」「コトバは，はじめは，もっとも初歩的なタイプの一般化である．子どもは，自分自身の発達につれて初歩的な一般化からだんだんとより高次なタイプの一般化へと移行し，そうして真の概念の形成でもってこの過程をおえる．」この一般化の過程は、それ自体がひとつの知的営為であり、それを経由することで複雑さの次数がひとつ上がるフラクタルな構造を示す。

表１　記号から概念への段階的な複雑化

次数	呼称	適用論理	結果・意味
0	記号	1 対 1（反射）	反射的行動
0	言葉記号	1 対 1（反射）	五官記憶の想起
1	生活(具象)概念	1 対全（群）	五官記憶の総合化
2	一次論理概念	1 対全（群）	生活概念の操作(類・関係性)
3	分野科学概念	1 対 1	学際統合前の観察
4	科学的概念	1 対全（群）	思考記憶の総合化

ピアジェとヴィゴツキーの概念に関する考察は、すべての科学や学際研究において順守すべき一般性をもつので、辞書に複雑次数を示し、徹底させるとよい、と筆者は考える。

3．　デジタル三段階進化：

① 脳室内免疫細胞ネットワークで言語処理：デジタル言語学は、言葉は脳内で音韻波形を模した抗原と、それと特異的に結合する抗体によってネットワークを構築すると仮説する。この考え方は、アレキサンドリアのヘロフィロスとエラシストラトス以来の伝統的考えであるが、19 世紀以降の大脳皮質による言語処理という考えとは一線を画す。

② 物理層（＝脳外）における信号の三段階進化：言語的人類は、6 万 6000 年前に喉頭降下がおきて、論理成分である音素とモーラを有する音節を獲得した。文字列は、正書法の知識を持つ脳にとって「消えない音節」である。電子情報は、文字列をコード変換表にもとづいて 0/1 のビット値に変換したもので、インターネット検索エンジンに対して対話する音節として機能する。

③ 論理層(＝脳内)三段階進化：人類の知能は、物理層における信号進化の成果を脳の使い方に反映させることで進化した。

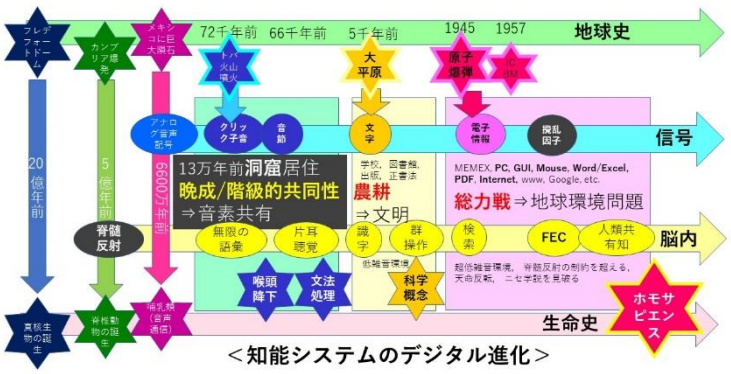
(i)　無意識の文法処理ができるのは、母語を片耳で聴き取ることによって、脳幹聴覚神経核の方向定位能力を転用して、文法的音節の音韻波形をベクトル処理している。

(ii)　文明が技術や知識の連続的発展を可能にした時、僧院や大学といった低雑音環境で、経済活動や家族から解放されて、ひたすら学習に励む階級を生みだした。これが群としての概念操作を可能にした。

(iii)　インターネットと検索エンジンのおかげで、キーワード検索によって必要な論文や書籍をみつけることが容易になった。我々は、きわめて大量だが信頼性の保証がない学際的な言語情報を容易に入手できる。これを活用するためには、脊髄反射の制約と、言語の持つ階級的共同性の制約を乗り越える必要がある。

表 2　知能の三段階デジタル進化

	引き金	誕生	獲得したもの	場所	時期
1	喉頭降下	**音節(音素とモーラ)**	無限の語彙、 無意識の文法処理	南アフリカ	66　千年前
2	農耕と王朝支配	**文字(消えない音節)** **僧院・大学という低雑音環境**	文明(知識の連続的発展) **概念(群として作用する)**	大平原	5 千年前
3	総力戦とソ連の ICBM	**電子化(対話する音節)、きわめて大量だが信頼性の保証がない言語情報**	PC, インターネット, www, 検索エンジン、前方誤り訂正(FEC)	アメリカ	20　世紀末



＜知能システムのデジタル進化＞

Dating the Nuzi Cuneiform Tablets Computationally: Analyzing Family Networks in Ancient Mesopotamia

Sumie Ueda [*], Takashi Tsuchiya [†], Yoshiaki Itoh [‡]

September 30, 2020

Making an electronic version of the name index Nuzi Personal Names, we estimate the published year of each cuneiform tablet of the Nuzi town in ancient Mesopotamia to understand the archaeological studies computationally. All these tablets enable us to reconstruct the social and economic life of Nuzi in the middle of the 2nd millennium B. C. E. The tablets, are on land transaction, marriage, loan, slavery contracts etc. We reconstruct family trees and social networks of Nuzi by using the kinship data in Nuzi Personal Names. We formulate the least squares problem with linear inequality constraints, the farther of a person is at least 22.5 years older than the person, contractors were living at the time of the contract, etc. We estimate the published years of the cuneiform tablets, together with the birth year of each person listed in them, consistently with the trees and networks. The number of tablets seems to increase by logistic growth. It may show the dynamics of concentration of lands or other properties into few powerful families in a period of about seventy years and most of them are in about forty years. We compare our study with a well known archaeological study by M. P. Maidman for the Nuzi documents. The archaeological study seems to support our present computational studies.

[*]Institute of Statistical Mathematics, Tachikawa Tokyo
[†]National Graduate Institute for Policy Studies, Minato-ku Tokyo
[‡]Institute of Statistical Mathematics, Tachikawa Tokyo

References

[1] I. J. Gelb, P. M. Purves, and A. A. MacRae, Nuzi personal names (The University of Chicago Press, Chicago Illinois, (1943). (http://oi.uchicago.edu/pdf/oip57.pdf)

[2] Y. Itoh, M. Ishiguro, S. Ueda, and K. Makino, Estimating population from the kinship data in an ancient society (CDROM), Report of Research Grant 09204245, Ministry of Education, Science and Culture of Japan, 1998) (in Japanese).

[3] M. P. Maidman, The Tehiptilla family of Nuzi -a genealogical reconstruction, Journal of Cuneiform Studies 28 (3),(1976) 127-155 .

[4] M. P. Maidman, Nuzi texts and their uses as historical evidence (Society of Biblical Literature, Atlanta, 2010).

[5] M. P. Maidman, Nuzi, the Club of the Great Powers, and the Chronology of the Fourteenth Century, KASKAL, 8(2011) 77-139.

[6] K. Makino, Social change reflected in the source of false adoption contracts at Nuzi. Shigaku 60(1), (1991) 91-119 (in Japanese with English title).

[7] NUOPT, version 9, Mathematical Systems Inc. (2007).

[8] T. Tsuchiya, Interior-point Algorithms, Information Geometry and Optimization Modeling, Proceedings of the Institute of Statistical Mathematics Vol. 61, No. 1, 3-16 (2013) (in Japanese with English Summary)

[9] S. Ueda, Statistical Mathematics Approach to Human Sciences, (PhD Thesis, Graduate University for Advanced Studies, 2010) (in Japanese).

[10] S. Ueda, K. Makino, and Y. Itoh, Reconstructing family trees in ancient population from the Nuzi personal names, Proceedings of the Institute of Statistical Mathematics 53, 285-295 (2005) (in Japanese with English summary).

[11] Ueda, S., Makino, K., Itoh, Y., & Tsuchiya, T. (2015). Logistic growth for the Nuzi cuneiform tablets: Analyzing family networks in ancient Mesopotamia. Physica A: Statistical Mechanics and its Applications, 421, 223-232.

[12] Yamashita, H., Yabe, H. & Tanabe, T. A globally and superlinearly convergent primal-dual interior point trust region method for large scale constrained optimization. Math. Program. 102, 111-151 (2005).

Health status and repeated multiple treatments in long-term care: A panel structural VAR analysis

Shinya Sugawara* Tsunehiro Ishihara†

This study analyzes the dynamic relationship between health status and expenditures on repeated multiple treatments, which are typical in long-term care. To facilitate causal inferences where complex dynamic interdependencies exist between many variables, we adopt a structural vector autoregression model for panel data of individuals. The model is estimated using a Bayesian shrinkage method which can simultaneously employ estimation and model selecttion for the lag length. Then, we employ a counterfactual analysis using impulse response functions. We analyze monthly claims data in the context of long-term care in Japan, where social insurance covers many formal services for elderly care at home. Our empirical analysis reveals several patterns of dependency between service utilization and their effects. In particular, we found that day care and outpatient rehabilitation share similar utilization patterns and also result in similar levels of improvement in health status, which implies that appropriate targeting can improve the effectiveness of service provision.

Keywords: Panel structural vector autoregression model; long-term care expenditure; long-term care insurance in Japan; high-dimensional data; Bayesian shrinkage estimation; causal inference;

*Corresponding. Tokyo University of Science. Email: shinya_sugawara@rs.tus.ac.jp
†Osaka University of Economics

Forward variable selection for sparse ultra-high dimensional generalized varying coefficient models

一橋大学大学院経済学研究科　本田敏雄

本研究は，台湾の国立清華大学統計学研究所の林建同氏との共同研究である．以下の通り報告した．

近年データ収集技術の進歩により，様々な分野において非常に多くの高次元データが利用可能になり，それらのデータ解析の必要性が高まってきた．ここでは，説明変数の数を p，観測値の数を n とすると，$n \ll p$ の場合を考える．この種のデータに対しては，Lasso，SCAD などの正則化推定量が使われることが多いが，p があまりにも大きな場合 (超高次元) には，Lasso，SCAD を実行することができない．

その場合，スクリーニングとして最初に明らかに被説明変数とは無関係と思われる変数 1 を除くのが一般的である．スクリーニング法としては，

1. 周辺モデルによる方法
2. モデルは用いず，被説明変数とそれぞれの説明変数の関係の指標を用いる方法
3. ここで扱う，モデルを用いる前進型変数選択法

などがある．本報告では以下の一般化変動変数モデルを扱い，係数関数はスプライン関数で近似する．$\boldsymbol{g}^*(z)$ の非ゼロ成分は少数である．その添え字集合を $\mathcal{M}$ とする．

$$f(y|\boldsymbol{x}, z) = \exp\{y\boldsymbol{x}^T\boldsymbol{g}^*(z) - b(\boldsymbol{x}^T\boldsymbol{g}^*(z)) + c(y)\}$$
$$\boldsymbol{g}^*(z) = (g_1^*(z), \ldots, g_p^*(z))^T$$

ここで記号を定義する．$S \subset \{1, \ldots, p\}$ として，この部分モデルの対数尤度関数は以下の通りである．

$$\ell_n(\boldsymbol{X}_S^T\boldsymbol{g}_S(Z)) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \boldsymbol{X}_{iS}^T\boldsymbol{g}_S(Z_i)), \quad \ell(y, \theta) = y\theta - b(\theta).$$

L 次元の B スプライン基底を用いて，$\boldsymbol{X}_S^T\boldsymbol{g}_S(Z)$ を $\boldsymbol{W}_S^T\boldsymbol{\beta}_S$ とする．$\widehat{\boldsymbol{\beta}}_S$ が最尤推定量である．

アルゴリズムは以下の通り．

(1) $S = S_{k-1}$ とおく．

$$j_k = \underset{j \in S^c}{\operatorname{argmax}} \max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}})$$

(2) j_k を加えて $\ell_n(\boldsymbol{W}_{S_{k-1}}^T \widehat{\boldsymbol{\beta}}_{S_{k-1}})$ が十分に改善されるかチェック．具体的には $S = S_{k-1}$ として，

$$\max_{j \in S^c} \max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}}) - \ell_n(\boldsymbol{W}_S^T \widehat{\boldsymbol{\beta}}_S) > L\xi_n |A| \log p_n / n$$

ならば，$S_k = S_{k-1} \cup \{j_k\}$ として **(1)** へ．そうでなければ $\widehat{\mathcal{M}} = S_{k-1}$ で終了．

我々の研究では以下の結果を得た．

- スクリーニング一致性を証明したが，やや制約的な仮定が必要であった．
- $n = 200, 400$ かつ $p = 1000$ などで，正規回帰モデル，ポアソン回帰モデル，ロジスティック回帰モデルにおいて，Lasso，SCAD，NIS とシミュレーションで比較した．この三つの中では SCAD が最善であった．提案した方法は SCAD と同等あるいはそれ以上のパフォーマンスを示した．$p = 1000$ は比較のためである．
- 情報量基準を使った停止則では早めに停止する傾向があり，何らかの工夫をする必要がある．また greedy algorithm ではあるが，不必要な変数を多く選ぶことはなかった．
- 簡略化した sequentially conditional screening 法もほぼ同じパフォーマンスをしめした．多発性骨髄腫データ $(p = 44760)$ ではこれを用いた．

参考文献

Honda, T. & Lin, C. T. (2020). Forward variable selection for sparse ultra-high dimensional generalized varying coefficient models. Forthcoming in JJSD. A previous version is available as Discussion Papers #2020-01, Graduate School of Economics, Hitotsubashi University.

クラスタリングによる正準判別の精度向上と クロスバリデーションの高速化

三浦 完太 (九州大学大学院数理学府)
廣瀬慧（九州大学マス・フォア・インダストリ研究所）

1　クラスタリングによる正準判別の精度向上

1.1　背景

正準判別は多群からなる高次元データを低次元に射影し、視覚的に見やすくする手法であり、圧縮先でのユークリッド距離から教師ありの判別にも用いることができる。しかし課題として、既存の正準判別では一度の射影で判別を行うため、手書き文字の判別のような群の数が多いデータの中で似たような群（例えば「王」と「玉」など）に対して誤判別率が高くなり、結果的に全体の判別精度が落ちてしまう現状があった。そこで、ある程度似た群を集めた仮想的な副群をクラスタリングにより作成し、その副群の中でさらに判別を行うことで数値的精度が向上することが経験的にわかってきた。

本稿では、これを実現する新たなアルゴリズムを提案する。また、その手法におけるクロスバリデーションの高速化を試みる。数値実験と実データ解析を通じて、提案手法の有用性を検証する。

1.2　提案手法の概要

提案する手法は副群への判別と、副群の中での判別の 2 段階の正準判別による手法となる。いま、n 個の p 次元データ $\boldsymbol{x}$ がその所属の群を表すラベル y とセットになって $\{(y_i, \boldsymbol{x}_i)\ (i = 1, 2, ..., n)\}$ として与えられたとする。この時ラベルは群の個数を J として $(y_i \in \{1, 2, ..., J\})$、観測されたベクトルは $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ と表せる。また、各群の集合を $G = \{G_1, G_2, ..., G_J\}$ とし、特に上付き符号を所属する群で表した j 群の i 番目のデータを $\boldsymbol{x}_i^{(j)} (\in G_j)$ と表す。

$$G_j \ = \ \{\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, ..., \boldsymbol{x}_{n_j}^{(j)}\}, \quad j = 1, 2, ..., J$$

ここで、クラスタリングにより m 個の副群（metagroups）からなる集合 $M_m = \{C_1, C_2, ..., C_m\}$ を与えたと仮定する。この副群は G の部分集合 $C_i \subset G (i = 1, 2, ..., m)$ であり、$C_i \cap C_j = \phi (i \neq j), \cup_{i=1}^{m} C_i = \{G_1, ...G_J\}$ を満たす。この副群に基づき、M_m の上で $C_1, ..., C_m$ への判別が第一段階の判別である。$\boldsymbol{x}$ が C_i に判別されたとする。この時第 2 段階の判別は、C_i が単一の群からなる集合すなわち $C_i = \{G_j\}$ であれば $\boldsymbol{x}$ は G_j に判別されたとし、もし C_i が 1 つ以上の群からなる集合 $C_i = \{G_{i_1}, ..., G_{i_j}\}$ ならば C_i の上で $G_j \in C_i$ への判別を行う。

1.3　副群の作成方法

副群を作成するためのクラスタリングのアルゴリズムを、誤判別率の最小化を目的として、一個抜きクロスバリデーションにより最適化されるものとして提案する。方法は段階的であり、群を一つずつ複合していくものである。

まず最初の副群の集合を最初の群の集合 $M^{(0)} = \{C_1^{(0)}, ..., C_J^{(0)}\}$ $(C_j^{(0)} = \{G_j\} : j = 1, ..., J)$ として決定する。次の副群の集合は、前の副群 $M_J^{(0)}$ の 2 つの元の和集合を一つの副群としてこれを新たな副群の集合 $M^{(1)} = \{C_1^{(1)}, ..., C_j^{(0)} \cup C_k^{(0)}, ..., C_J^{(1)}\}$ とする。対象となる $C_j^{(0)}$ と $C_k^{(0)}$ は、$M^{(0)}$ の 2 つの元全ての組み合わせのうち、アルゴリズム全体での誤判別率が最小となる対とする。これを誤判別率が更新されなくなる ($M^{(T)}$ での最小値 $> M^{(T+1)}$ での最小値) まで続けて、$M^{(T)}$ をクラスタリングを終えた最終的な副群の集合とする。これによって常に誤判別率が最小となるクラスターが作成されるため、正準判別の精度を向上させることができる。

2　クロスバリデーションの高速化

提案手法により精度の向上が示されたが、この手法には計算量の面で課題が残されている。正準判別には $p \times p$ の行列の固有値計算で $O(p^3)$ の計算量が含まれており、これにクロスバリデーションの実行と副群の作成過程で計算量が膨大になってしまう。そこで、このボトルネックとなる固有値計算の部分をクロスバリデーションの外に出すことによって計算量を軽くすることを検討する。

本稿では正準判別の問題を、リッジ回帰の問題に置き換えて、リッジ回帰におけるクロスバリデーションの高速化を通じて正準判別の高速化を行う。

3　数値実験

提案手法による多群判別の精度を誤判別率により評価した結果を以下に示す。各データセットの誤判別率は一個抜きクロスバリデーションにより求める。似たような群をもつデータで実験を行うことを前提とし、擬似的に発生させたデータと実データに対して実験を行う。

参考文献

[1] Cawley, G. C., & Talbot, N. L. (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition, 36(11), 2585-2592.

[2] Takada, T., Mita, A., Maeno, A., Sakai, T., Shitara, H., Kikkawa, Y., Moriwaki. K.,Yonekawa, H. and Shiroishi, T. (2008). Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits. Genome Research, 18, 500-508.

射影勾配法による高次元回帰モデリング

須藤　隼 [1]　川島　孝行 [1,2]　金森　敬文 [1,2]

[1] 東京工業大学
[2] 理化学研究所革新知能統合研究センター

1　はじめに

高次元データにおいて回帰を行うための手法として，その有効性から LASSO (Tibshirani, 1996) が広く用いられている．具体的に，LASSO は以下の最適化問題 (以下，LASSO-主問題と呼ぶ) で表される．

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 \leq z, \end{aligned} \tag{1}$$

ただし，$\mathbf{Y}\in\mathbb{R}^n$ は応答変数ベクトル，$\mathbf{X}\in\mathbb{R}^{n\times p}$ は説明変数行列，$\boldsymbol{\beta}\in\mathbb{R}^p$ は回帰係数ベクトル，n はサンプルサイズ，p は説明変数の次元を表す．一般には，LASSO-主問題 (1) を次のラグランジュ緩和問題 へと変形し，既存の最適化手法 (例えば，座標降下法 (Friedman et al., 2007) など) を用いて解かれることが多い．

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1,$$

ただし，λ はラグランジュ乗数を表す．

別な方向性として，直接，LASSO-主問題 (1) を解くことも考えられる．最適化手法として，次の射影勾配法 (Bertsekas, 1999) を用いる．

$$\beta^{(t+1)}=\mathrm{Proj}_{(\|\boldsymbol{\beta}\|_1\leq z)}\left(\beta^{(t)}+\eta_t\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(t)})\right), \tag{2}$$

ただし，$\beta^{(t)}$ と η_t は，それぞれ t 回目の反復における回帰係数ベクトルとステップサイズ，また，$\mathrm{Proj}_{(S)}(\mathbf{b})=\underset{\boldsymbol{\theta}\in S(\subset\mathbb{R}^p)}{\mathrm{argmin}}\|\boldsymbol{\theta}-\mathbf{b}\|_2$ である．射影勾配法 (2) で LASSO-主問題 (1) を解く際に Proj の計算が必要となる．既存研究 (Duchi et al., 2008) では，単体への射影 (Held et al., 1974) を経由することで Proj の計算を可能としている．

2 LASSO 以外への拡張

統計的な理論の観点で，LASSO よりも優れた正則化法が提案されている．Adaptive LASSO (Zou, 2006) では，$\|\boldsymbol{\beta}\|_1$ の部分を $\sum_{j=1}^{p} a_j|\beta_j|$ $(a_j \geq 0)$ とすることで，LASSO による bias を減らすことに成功した．

本研究では，adaptive LASSO の場合でも，Duchi et al. (2008) で提案されているアルゴリズムと同じような Proj の計算を導出した．また，SCAD (Fan and Li, 2001) や MCP (Zhang, 2010) といった非凸な正則化法の場合でも，上記の Proj の計算が適用できる形へと変形を行うことで，射影勾配法に基づく推定を実現した．

References

Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l1-ball for learning in high dimensions. In Cohen, W. W., McCallum, A., and Roweis, S. T., editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 272–279. ACM.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Held, M., Wolfe, P., and Crowder, H. P. (1974). Validation of subgradient optimization. *Mathematical Programming*, 6(1):62–88.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

カーネル法に基づく超高次元モデル選択

長崎大学 情報データ科学部　梅津 佑太

1　はじめに

$(y, \boldsymbol{x}) \sim \mathrm{P}_{Y,\boldsymbol{X}}$ に対する加法回帰モデル $\mathbb{E}[y \mid \boldsymbol{x}] = \sum_{j=1}^{d} f_j^*(x_j)$ を考える．真の回帰構造はスパース，つまり，$S^* = \{f_{j*} \neq 0\}$ に対して $s = |S^*| \ll d$ であるとする．説明変数の次元 d がサンプルサイズ n よりも圧倒的に大きいような超高次元の設定を考える．$(y, \boldsymbol{x})$ と同じ分布に従う独立な標本 $(y_i, \boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ に基づき，高い確率で S^* を含むようなモデル $\hat{S}$ を選択することを考える (SSP: sure screening property)．

$\omega = X^\top \boldsymbol{y}/n$ なるスコアに基づき，$\hat{S} = \{j \mid |\omega_j| \geq \lambda_n\}$ でモデル選択を行う手法がFan and Lv (2008) で提案された (SIS: sure independence screening)．ここで，λ_n は適当なオーダーで減衰する実数列である．SIS は非常にシンプルではあるものの，線形回帰モデルにおいては，$\log d = \mathrm{o}(n^\kappa)$ $(\exists \kappa > 0)$ なる超高次元で SSP が成り立つ．SIS を拡張することで，非線形モデルに対するスクリーニングが数多く提案されてきた (例えば，Fan and Lv, 2014; Liu et al., 2015; Desboulets, 2018)．これまで提案されてきた手法は，SIS で用いられるスコア ω_j がピアソン相関や周辺回帰係数の推定量であることに着目して得られたものである．

本稿ではまず，SIS で選択されるモデルそのものに着目することで，新たなスクリーニング法 (KRIS) が得られることについて説明する．KRIS は，適当な条件のもと，SSP が成り立つ．ただし，KRIS に対する SSP は λ_n に関する適当なオーダーに対して成立するという主張であり，実際にデータに適用する際には恣意性が残ってしまう．そこで，KRIS で用いるスコアの漸近分布を利用することで，漸近的に expected false positive rate をコントロールするように λ_n を決定する手法を提案する．

2　KRIS: Kernel Regression based Independence Screening

以下，$\boldsymbol{y}$ は中心化されている，つまり，$\sum_{i=1}^{n} y_i/n = 0$ とし，スクリーニング手法として以下を考える．

定義 1 (KRIS)**.** カーネル関数を k_j とする再生核ヒルベルト空間 (RKHS) $\mathcal{H}_j$ に属する関数 $\{\hat{f}_j\}$ を

$$\hat{f}_j = \underset{f_j \in \mathcal{H}_j}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f_j(x_{ij}))^2 + \lambda_n \|f_j\|_{\mathcal{H}_j}, \qquad j = 1, \ldots, d \tag{1}$$

とする．このとき，$S^{\mathrm{KRIS}} = \{j \mid \hat{f}_j \neq 0\}$ とする．ただし，$\|\cdot\|_{\mathcal{H}_j}$ は $\mathcal{H}_j$ の RKHS ノルムである．

表現定理と KKT 条件より，最適化問題 (1) は実際に解く必要はない．実際，スケーリングされたカーネル行列を $K_j = (k_j(x_{pj}, x_{qj})/n)_{p,q}$ とすれば，$S^{\mathrm{KRIS}} = \{j \mid \|\boldsymbol{y}\|_{K_j}/\sqrt{n} \geq \lambda_n\}$ となることが容易に分かる．ただし，適当な大きさのベクトル $\boldsymbol{v}$ と行列 A に対して，$\|\boldsymbol{y}_j\|_A = \boldsymbol{v}^\top A \boldsymbol{v}$ である．また，カーネル関数として線形カーネル $k_j(x, y) = xy$ を考えれば，$\|\boldsymbol{y}\|_{K_j}/\sqrt{n} = |\omega_j|$ となるから，S^{KRIS} は SIS の自然な拡張である．適当な条件のもと，KRIS は SSP を持つことを示すことができる．その際，カーネル関数 k_j の固有値 μ_{jl} の減衰レートに応じて，次のようにとれば十分である:

$$\mu_{jl} \sim l^{-\kappa}\ (\kappa > 1) \Rightarrow \lambda_n = \mathrm{O}(n^{(1-\kappa)/(2+3\kappa)}), \quad \mu_{jl} \sim e^{-\kappa l}\ (\kappa > 0) \Rightarrow \lambda_n = \mathrm{O}(n^{-1/3}(\log n)^2).$$

3　しきい値の選択

KRIS は SSP を持つとはいえ, 実際にはしきい値 λ_n の選択に恣意性が残る. そこで, KRIS のスコア $\|\boldsymbol{y}\|_{K_j}^2/n = \boldsymbol{y}^\top K_j \boldsymbol{y}/n$ の漸近分布を用いることで FDR をコントロールするようにしきい値を決定するための方法を考える. 具体的には, grobal null hypothesis $\mathrm{H}_0 : f_j^* \equiv 0,\ j = 1, \dots, d$ のもとで以下が成り立つ.

定理 1. k_j をマーサーカーネルとし, その固有値を $\mu_{j1} \geq \mu_{j2} \geq \cdots \geq 0$ とする. このとき,

$$\frac{1}{n}\boldsymbol{y}^\top K_j \boldsymbol{y} \xrightarrow{\mathrm{d}} \frac{1}{\sigma}\sum_{l=1}^{\infty} \mu_{jl} z_l^2, \qquad j = 1, \dots, d \tag{2}$$

が成り立つ. ただし, $\sigma^2 = \mathbb{E}[y^2], z_l \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$ である.

Gretton et al. (2008) と同様に, 右辺の分布を形状パラメータとスケールパラメータがそれぞれ

$$\alpha_j = \frac{\sigma^4 \mathrm{tr}(K_j)^2}{v_j}, \qquad \beta_j = \frac{v_j/n}{\sigma^4 \mathrm{tr}(K_j)}$$

であるようなガンマ分布で近似する. ただし,

$$v_j = (\mathbb{E}[y^4] - 3\sigma^4)\sum_{i=1}^{n} k_j(x_{ij}, x_{ij})^2 + 2\sigma^4 \mathrm{tr}(K_j^2)$$

である. 帰無仮説のもとで $\sigma^2 = \mathbb{E}[y^2]$ や $\mathbb{E}[y^4]$ は帰無仮説のもとで明らかに不偏推定可能であり, これを用いて近似した (2) の累積分布関数を F_j とする. このとき, 任意の $r \in \{1, \dots, d\}$ に対して $\lambda_n = F_j^{-1}(1 - r/d)^{1/2}$ ととれば, expected positive false positive rate を

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{|S^{\mathrm{KRIS}} \cap S^{*c}|}{|S^{*c}|}\right] = \limsup_{n\to\infty} \frac{1}{d-s}\sum_{j \in S^*} \mathrm{P}\left(\frac{1}{n}\boldsymbol{y}^\top K_j \boldsymbol{y} \geq \lambda_n^2\right) \leq \frac{r}{d}$$

とコントロールできる.

参考文献

Desboulets, L. D. D. (2018) "A review on variable selection in regression analysis," *Econometrics*, Vol. 6, p. 45.

Fan, J. and Lv, J. (2008) "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B*, Vol. 70, pp. 849–911.

——— (2014) "Sure independence screening," *Wiley StatsRef: Statistics Reference Online*, pp. 1–8.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008) "A kernel statistical test of independence," in *Advances in neural information processing systems*, pp. 585–592.

Liu, J., Zhong, W., and Li, R. (2015) "A selective overview of feature screening for ultrahigh-dimensional data," *Science China Mathematics*, Vol. 58, pp. 1–22.