科研費（基盤 A, 15H01678）
「大規模複雑データの理論と方法論の総合的研究」
研究代表者：青嶋 誠（筑波大学）

予測モデリングとその周辺

-機械学習・統計科学・情報理論からのアプローチ-

日時：2018 年 11 月 23 日（金）～25 日（日）
場所：成蹊大学 14 号館 5 階 505 教室
https://www.seikei.ac.jp/university/aboutus/campus_uni/
開催責任者：小森　理（成蹊大学）・深谷　肇一（国立環境研究所）
後援：成蹊学園サステナビリティ教育研究センター

内容・目的：予測モデリングは理論と応用の両面で非常に重要であり，医学，生態学，生物学，工学，ビジネスなどの分野で盛んに研究されているテーマの一つです．特に近年注目されている大規模かつ複雑データを扱う予測モデリングの研究は未開拓な部分が多く，これから大いに進展する分野であることが期待されます．本シンポジウムでは様々な分野で活躍される研究者の講演を広く募集し，分野横断的な議論を通し新たな研究の方向性を探っていければと思います．

旅費の配分：講演者を中心に配分します。旅費希望の方は講演申込の際にお伝え下さい。

宿舎の斡旋：斡旋しません。

講演申込期限：2018 年 10 月 26 日（金）
　氏名・所属・講演題目を電子メールでお知らせ下さい。

予稿期限：2018 年 11 月 9 日（金）
　A4 サイズ 10 頁以内で作成し、PDF ファイルを送信して下さい。

報告書原稿：
　報告書を作成しますので、予稿とは別に報告書原稿（A4 サイズ 2 枚）も PDF ファイルで送信して下さい。

問い合せ先・講演申込先・予稿送付先・報告書原稿送付先：
〒180-8633　東京都武蔵野市吉祥寺北町 3-3-1
成蹊大学　理工学部情報科学科　小森　理
Email: komori[at]st.seikei.ac.jp　TEL: 0422-37-3764

プログラム

■11 月 23 日(金)

特別講演
10:00-10:50　　岡村寛（中央水産研究所）
水産資源学で使用される予測モデル

10:50-11:25　　中村和幸（明治大学）
データ同化における予測モデリングとその背景

11:25-12:00　　中野慎也(統計数理研究所)
宇宙科学における撮像観測とデータ同化

特別講演
13:00-13:50　　江口真透（統計数理研究所）
一般化平均を使った統計方法

13:50-14:25　　竹之内高志(はこだて未来大学)
非正規化モデルを用いた推定法

14:25-15:00　　大前勝弘（京都大学）
医学データにおける予測モデリング再考

15:00-15:20 (休憩)

特別講演
15:20-16:10　　山田誠(京都大学)
Post Selection Inference with Kernels

16:10-16:45　　松井秀俊(滋賀大学)
関数データに基づく交互作用モデルとその推定

16:45:17:20　　中山優吾（筑波大学）
カーネル主成分分析に基づく高次元データのクラスタリング

■11 月 24 日(土)

特別講演
10:00-10:50　伊勢武史(京都大学)
ビッグデータと生態学

10:50-11:25　岩山幸治(滋賀大学)
潜在変数モデルによるトランスクリプトームの予測

11:25-12:00　佐藤安弘（龍谷大学）
ゲノムワイド関連解析を応用した混植効果の推定と虫害予測

特別講演
13:00-13:50　吉田亮(統計数理研究所)
マテリアルズインフォマティクスの最前線

13:50-14:25 斎藤正也(統計数理研究所)
風疹流行モデルの構成とインフルエンザへの応用に向けての課題

14:25-15:00　城田慎一郎(UCLA)
Spatial Joint Species Distribution Modeling using Dirichlet Processes

15:00-15-20 (休憩)

15:20-15:55　植木優夫(理化学研究所)
全ゲノム配列情報を用いた疾患発症予測に向けて

15:55-16:30　新村秀一(成蹊大学)
高次元 Microarray データによる「癌の遺伝子解析」―なぜ 1970 年から研究され成功しなかったのか？

16:30-17:05　荒木由布子（静岡大学)
高次元データのための直接・間接効果を考慮した関数判別モデル

17:05-17:40　永井勇（中京大学）
大規模なデータの分析における精度行列の縮小推定法とその特徴

18:30-20:30　懇親会(板前割烹するり吉祥寺)

■11 月 25 日(日)

10:00-10:35　佐野崇(成蹊大学)
マルコフ連鎖による大学合格者歩留まり率のモデル化と予測

10:35-11:10　三枝祐輔（横浜市立大学）
救急需要予測のための時空間正規混合モデル：横浜市救急データへの応用

11:10-11:45　深谷肇一（国立環境研究所）
生態学的大規模データを統合する階層モデル：種個体数分布の広域予測

11:45-12:20　小森理(成蹊大学)
一般化平均に基づく予測モデリング

水産資源学で使用される予測モデル

岡村　寛　（中央水産研究所）

水産資源学は，漁業という営利行為から得られるデータに大きく依存する場合が多い．そのため，データには大きな不確実性が混入しがちである．また，将来の適切な漁獲量を知ることが，持続的な漁業実現の鍵となるため，統計モデルを利用した予測が早くから重視されてきた．たとえば，個体数の指標値である単位努力量あたり漁獲量（CPUE）は，そのままでは多くのバイアスが混入していると考えられ，一般化線形モデル（GLM）を利用してバイアスの除去を行うという操作がなされるが，GLM の使用は 1980 年ごろから見られる．また，機械学習手法の導入も早く，1990 年代初頭から，漁業へ加入してくる量（加入量）の予測や，漁獲量の予測に，ニューラルネットが使用されだした．2000 年代になってからは，ランダムフォレストやブースティングのようなアンサンブル法による機械学習の活用が進んでいる．本講演では，水産資源学で使用される最近の予測モデルの事例を紹介し，我々の研究チームで取り組んでいる予測モデルの使用例について紹介する．

水産資源学は大きく分けて，資源評価と資源管理に分類される（岡村・市野川 2016）．資源評価では，生残率や増加率のような個体群の基礎情報を抽出したり，個体群動態モデルを使って，個体数の状態や変化を推定したりすることが課題となっている．個体群動態モデルでは，階層モデルが頻繁に活用されるが，水産資源学では伝統的に年齢別の複雑なモデルで多様な情報を統合するモデルが使用されるため，多くのパラメータの推定が必要となり，高速な計算に対するニーズが大きかった．それ故，水産資源学では，自動微分とラプラス近似を使用した高速計算ソフトの開発・使用が進んだ．近年広く使用されつつある Template Model Builder とその事例研究について簡単に紹介する（Okamura et al. 2017, 2018a）．一方で，データが不足した中で，資源状態を予測する研究も進んでいる．ここでは，メカニスティックなモデルを構築することが難しいため，アンサンブル学習による予測モデルの適用が進んでいる．CPUE 標準化（Okamura et al. 2018b）においてアンサンブル学習の一種である gradient boosting を使った結果の紹介を行う．

資源管理においては，1990 年代にシミュレーションにより仮想現実世界を作り出し，不確実性に頑健な管理方式を選択する管理戦略評価（Management Strategy Evaluation : MSE）の考え方が構築された（Punt et al. 2016）．これは，最適戦略から頑健戦略への移行という点で大きなブレイクスルーとなった．ここでも，将来予測が重要な課題となるが，あまりデータに依存しすぎると管理に失敗するリスクが上昇する．そのため，管理方式の中の学習係数のチューニングがひとつの課題となる．今後の我が

国漁業資源の管理において，MSE を行った事例をもとに，どのような管理方式を用いるべきかについて議論する．

参考文献

岡村 寛・市野川桃子. 2016. 水産資源学における統計モデリング. 統計数理 64(1), 39-57.

Okamura, H., Yamashita, Y., and Ichinokawa, M. 2017. Ridge virtual population analysis to reduce the instability of fishing mortalities in the terminal year. ICES Journal of Marine Science 74 (9): 2427–2436.

Okamura, H., Yamashita, Y., Ichinokawa, M., and Nishijima, S. 2018. Comparison of the performance of age-structured models with few survey indices. ICES Journal of Marine Science 75(6): 2016–2024.

Okamura, H., Morita, S. H., Funamoto, T., Ichinokawa, M., and Eguchi, S. 2018b. Target-based catch-per-unit-effort standardization in multispecies fisheries. Canadian Journal of Fisheries and Aquatic Sciences 75 (3): 452–463.

Punt, A.E., Butterworth, D.S., de Moor, C. L., De Oliveira, J. A. A. and Haddon, M. 2016. Management strategy evaluation: best practices. Fish and Fisheries 17. 303–334.

データ同化における予測モデリングとその背景

中村和幸

明治大学 / JST さきがけ

1　概要

データ同化とは，計算機シミュレーションと計測データを融合し，各々単独では得られない情報を得ることを目指したものである [1]．もともと気象学・海洋学において発展してきた手法であるが，近年では，生命科学 [2]，地盤工学 [3]，材料科学 [4] など，他の分野にも拡がりつつある手法である．気象予報においては，ナビエストークス方程式などの支配方程式やモデル式に基づき，大気状態の時間発展を与えるための計算機シミュレーションがなされる．その際に精度の高い気象予報の結果を得るためには，できるだけ現在の実際の大気の状態に近い初期値を構成する必要がある．しかし，シミュレーションに必要な物理量を稠密に得ることは不可能である．また，計測値を単純にモデルに含めると，本来存在しないような物理的な不連続性を生み出すことにつながり，不適切な現象を引き起こすことがある．そこで，モデルから決まる制約と，実際に得られた過去の計測データの間のバランスを取りながら，モデルに含まれる変数を適切に推定する手法の必要性が出てくる．これを実現するのがデータ同化である．

データ同化は，直接観測できない時変状態やパラメータを推定できるようになるため，知識発見の手法として用いることもできる．また数理的には，統計的時系列解析や制御理論の分野で用いられる状態空間モデルによる状態推定の問題とみることができる．このような観点では，状態推定の困難さは，システムの非線形性，モデル化誤差，シミュレーションスキームの計算速度による制約に由来する誤差，計測誤差など様々な要因によることになる．また，それにともなって，用いるべき推定アルゴリズムや推定精度が変わることになる．この点に関する統一的なアプローチは不足しており，この問題を解決することで，これまで以上の適用分野の拡大や新たな知識発見につながると見込まれる．

講演では，データ同化の枠組みと状態空間モデルでの表現，生命科学・生態学分野における状態推定・データ同化への応用，Local Translation Error(LTE) 分析によるシステム解析とデータ同化への展開について報告した．

2　データ同化と状態空間モデル

データ同化は，時間発展する数値シミュレーションについては，各時点での全シミュレーション変数を $\boldsymbol{x}_t$ とすると，決定的な計算となるため，

$$\boldsymbol{x}_t = \tilde{\boldsymbol{f}}_t(\boldsymbol{x}_{t-1})$$

の形式で記述できる．ここで，実現象に含まれる不確かさをさらに $\boldsymbol{v}_t$ として表現すると，

$$\boldsymbol{x}_t = \boldsymbol{f}_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t)$$

となる．一方，実際の時刻 t 時点の観測を $\boldsymbol{y}_t$ とすると，これは多くの場合に計測誤差由来の不確かさをもつことから，これを $\boldsymbol{w}_t$ と書くと，

$$\boldsymbol{y}_t = \boldsymbol{h}_t(\boldsymbol{x}_t, \boldsymbol{w}_t)$$

の形式で表現できる．これらは，状態空間モデルであり，線形ガウス型の場合にはカルマンフィルタ，観測モデルのみ非線形の場合には，粒子フィルタやマルコフ連鎖モンテカルロ法などを用いることにより，フィルタや平滑化を得ることができる．

状態推定は，フィルタについては粒子フィルタなどのフィルタリング手法，平滑化についてはマルコフ連鎖モンテカルロ法や粒子スムーザなどで得ることができる．多くのデータ同化においてはフィルタを用いることが多い一方で，システムとしての不確かさが大きい，すなわち $\boldsymbol{v}_t$ が大きいような場合には，マルコフ連鎖モンテカルロ法の適用が考えられる．特に，空間相関の強いモデルの場合には，ギブスサンプリングによる平滑化推定は困難を伴う場合がある．本報告では，このような状況下におけるハミルトニアンモンテカルロ法の有効性について紹介した．

3 Local Translation Error

データ同化においては，推定アルゴリズムは不確かさのモデリングにおいて，非線形性をどの程度考慮にいれるかということが重要になる．計測の不確かさを持つ状況下でのシステムの非線形性を判断する指標として，Local Translation error[5] を用いることができると考えられる．実際に，データ同化分野においてテスト用に広く用いられる Lorenz 96 モデルに適用すると，局所時間・局所空間における非線形性の定量化ができていることが確認できたことを報告した．

参考文献

[1] 中村和幸，上野玄太，樋口知之，「データ同化 : その概念と計算アルゴリズム」，統計数理，53，pp. 211–229, 2005.

[2] K. Nakamura, R. Yoshida, M. Nagasaki, S. Miyano, and T. Higuchi, "Parameter estimation of in silico biological pathways with particle filtering towards a petascale computing," *The Proceedings of 14th Pacific Symposium on Biocomputing*, pp. 227-238, 2009.

[3] T. Shuku, A. Murakami, S. Nishimura, K. Fujisawa, and K. Nakamura, "Parameter identification for Cam-clay model in partially loading tests using the particle filter," *Soils and Foundations*, 52(2), 279-298, 2012.

[4] Ito, S., H. Nagao, T. Kasuya, and J. Inoue "Grain growth prediction based on data assimilation by implementing 4DVar on multi-phase-field model," *Science and Technology of Advanced Materials*, 18(1), 857-869, doi:10.1080/14686996.2017.1378921, 2017.

[5] N. Sviridova and K. Nakamura, "Local noise sensitivity: Insight into the noise effect on chaotic dynamics," *Chaos*, 26, 123102, DOI: 10.1063/1.4970322, 2016.

宇宙科学における撮像観測とデータ同化

中野 慎也 (統計数理研究所)

1 はじめに

地球磁気圏は，宇宙空間の中でも地球の持つ磁場の影響が及ぶ範囲を指す．地球磁気圏は，太陽風と呼ばれる太陽からのプラズマの流れの影響で非対称な形状をしており，太陽側は地上 6 万 km 程度まで，太陽と反対側には地上数百万 km 以上まで広がっている．地球磁気圏の研究では，通常，人工衛星で種々の物理量を直接観測したデータを用いる．しかし，各衛星で得られるのは点の情報に過ぎず，磁気圏の大域的な現象の全体像をつかむのが難しい．

地球磁気圏のプラズマ (電荷を持った粒子で構成されるガス) の空間分布を，遠隔から 2 次元的に捉える撮像観測は，衛星による直接観測の欠点を解決する有用な方法と考えられる．特に，2000 年から 2005 年に運用されていた人工衛星 IMAGE は，様々な手段による撮像観測を実現した衛星であり，有用なデータが取得されている．しかし撮像観測は，プラズマ密度以外の物理量について情報を得るのが難しいという欠点がある．我々はデータ同化技術を活用することにより，人工衛星 IMAGE による撮像観測データから磁気圏の大域的な現象の全体像を捉える手法の開発を進めている．

以下では，まず，Nakano et al. (2014) で提案した IMAGE 衛星の極端紫外光 (extreme ultraviolet; EUV) 撮像データをプラズマの移流モデルに同化することにより，プラズマ圏と呼ばれるエネルギーが低く濃いプラズマの分布する領域の変動を再現した成果を示す．次に，現在開発を進めている内部磁気圏統合モデルに EUV 撮像データと高速中性粒子 (energetic neutral atoms; ENA) 撮像データを同化するシステムの状況についても紹介する．

2 EUV データ同化

IMAGE 衛星では，30.4nm の波長の EUV を撮像観測している．太陽から来る紫外光のうち，30.4nm の波長のものはヘリウムイオン (He^+) に散乱されるため，これを遠隔から観測することにより，磁気圏の He^+ の分布について情報が得られる．

地球磁気圏の荷電粒子は，地球磁場の磁力線に沿って動きやすいという性質があり，そのため，通常は磁力線に沿った方向の密度変化は小さいと仮定できる．そこで，磁力線に沿った密度変化が一定の関係式を満たすと仮定すると，内部磁気圏内の He^+ 密度分布は以下の式で記述できる:

$$n(\boldsymbol{r}) = n_{\mathrm{eq}}(\boldsymbol{\rho})\left(\frac{r_{\mathrm{eq}}}{r}\right)^{\alpha}. \tag{1}$$

但し，$\boldsymbol{\rho}$ は各磁力線が赤道面に交差する位置を示す 2 次元のベクトルである．衛星で観測される EUV 画像の各ピクセル i の強度は，He^+ 密度を視線方向に積分したものに比例すると仮定でき，

$$y_i = \int_{l_i} c(\boldsymbol{r}) n(\boldsymbol{r})\, ds + \varepsilon_i. \tag{2}$$

となる．ここで，ε_i は各ピクセルの観測ノイズである．

エネルギーの低いプラズマ圏プラズマの時間発展は，以下のような方程式で記述できる:

$$\frac{\partial \bar{N}}{\partial t} - \frac{\nabla\Phi \times \boldsymbol{B}}{B^2} \cdot \frac{\partial \bar{N}}{\partial \boldsymbol{x}} = f. \tag{3}$$

ここで，$\bar{N}$ は磁力線方向に平均したプラズマ密度，Φ は電位 (電場ポテンシャル)，$\boldsymbol{B}$ は磁場を表している．内部磁気圏において，磁場は地球起源の双極子磁場がよい近似になるが，電場についてはほとんど情報がない．そこで，電位 Φ の分布を未知とし，データ同化によって推定している．

本研究では，アンサンブル変換カルマンフィルタ (ensemble transform Kalman filter; ETKF) (Bishop et al., 2001) を用いてデータ同化を行っている．ETKF は，比較的少ない計算量で大規模なシステムのデータ同化が実現できる手法として，近年，広く使われるようになっている．大規模なデータ同化の問題において，アンサンブルメンバー数 N は，状態ベクトルや観測ベクトルの次元よりはるかに小さいのが普通である．ETKF では，逆行列の計算を N の次元の空間で行うため，多くのデータ同化の問題において，非常に効率的に状態推定を実現できる．

3 内部磁気圏統合データ同化システムの開発

現在は，EUV データに加えて，高速中性粒子 (ENA) データも活用し，2 種類のデータを内部磁気圏統合モデルに同化するシステムの開発を進めている．ENA は，内部磁気圏に存在する高エネルギーの荷電粒子 (主として陽子 H^+) が，その場に分布する低エネルギーの中性粒子から電荷を受け取り，エネルギーを保持したまま中性の粒子に変化したものである．電荷を持った荷電粒子は，地球磁場による Lorentz 力を受けるため，磁力線の周りを螺旋運動し，内部磁気圏から外に出ることができないが，一旦，中性粒子つまり ENA になると宇宙空間を直進するようになる．低エネルギー中性粒子の空間分布を既知とすると，ENA を遠隔から観測することにより，内部磁気圏の高エネルギー荷電粒子の空間分布に関する情報が得られる．したがって，EUV で低エネルギー荷電粒子，ENA で高エネルギー荷電粒子の空間分布について情報が得られる．

高エネルギー粒子は，低エネルギー粒子と異なり，磁場勾配の影響を受けるため，別の方程式系で扱う必要がある．高エネルギー粒子は，磁力線方向の運動を平均化した Boltzmann 方程式で扱うことができ (Fok et al., 2001)，本研究で用いる内部磁気圏統合モデルもこの式に基づいている．高エネルギー荷電粒子は，低エネルギー荷電粒子と比べると，磁気圏のやや外側に分布しているため，荷電粒子の動きを支配する電場の推定を行う上でも，ENA データは EUV データの情報を補完する役割を果たす．

4 おわりに

最初に述べたように，地球磁気圏に関して観測から得られる情報は非常に限られている．データ同化は，物理法則の知見を直接観測できない物理量を推定に活用することができ，地球磁気圏で起こる様々な現象の全体像を捉えることに役立つと考えている．

References

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects, Mon. Wea. Rev., 129, 420–436, 2001.

Fok, M.-C., Wolf, R. A., Spiro, R. W., and Moore, T. E.: Comprehensive computational model of Earth's ring current, J. Geophys. Res., 104, 8417, 2001.

Nakano, S., Fok, M.-C., Brandt, P. C., and Higuchi, T.: Estimation of temporal evolution of the helium plasmasphere based on a sequence of IMAGE/EUV images, J. Geophys. Res., 119, 3708–3723, doi:10.1002/2013JA019 734, 2014.

一般化平均による統計方法

江口　真透
統計数理研究所

概要: 統計学のために一般化平均を積極的に援用した方法を提案する．はじめに情報幾何との関連について考察し，つぎに統計モデリングのために幾つかの応用を紹介する．ロジスティック回帰や比例ハザードモデルにおける予測関数を一般化平均を使った準線形モデルを提案する．つぎに，クラスタリングのエネルギ関数やロス関数の混合のために一般化平均がキーになることが示された．

1　一般化平均と情報幾何

1930 年に，コルモゴロフと南雲によって独立に発表された一般化平均ついて考察する．彼らは単調関数 $\phi : \mathbb{R} \to (0, \infty)$ によって正の数 x, y に対して一般化平均を

$$\mathrm{GM}_\phi(x, y) = \phi\big((1-\pi)\phi^{-1}(x) + \pi\phi^{-1}(y)\big) \tag{1}$$

と定めた．同様に x, y, z の一般化平均を考えると，定義から $\mathrm{GM}_\phi(x, \mathrm{GM}(y, z)) = \mathrm{GM}_\phi(\mathrm{GM}(x, y), z)$ がいえる．このような ‘ 平均 ’ の持つべき公理から特徴付けられる一般化平均は生成関数 ϕ の関数自由度を持つので多様な平均の考えが展開できる．典型例の算術平均，幾何平均，調和平均などが含まれる．

情報幾何は確率密度の関数空間の上に双対リーマン幾何をベースに豊かな直観を与え，確率に関連する全ての分野へ幾何的考えを構築している．その中の基本定理として，この関数空間の上でのピタゴラス定理が挙げられる (Amari-Nagaoka, 2007)．確率密度関数 $p(x)$ と $q(x)$ を混合測地線でつなぎ，一方で，$r(x)$ と $q(x)$ を指数測地線でつないだとき，この 2 つの測地線が $q(x)$ で直交するならば，またその時に限り，

$$D_0(p, r) = D_0(p, q) + D_0(q, r) \tag{2}$$

が成立する．ここで D_0 は KL ダンバージェンスとする．この性質から最尤推定と十分統計量の関係，赤池情報量規準の妥当性などが導かれる．この考察において $r(x)$ と $q(x)$ をつなぐ指数測地線とは

$$\mathrm{EG}(q, r) = \exp\big((1-\pi)\log q(x) + \pi \log r(x) - \kappa(\pi)\big) \tag{3}$$

と定められる．ここで $\kappa(\pi)$ は正規化定数とする．このように $\mathrm{EG}(q, r)$ は生成関数 $\phi = \exp$ を使って正の数の代わりに密度関数に対する一般化平均と見れる．したがって，一般の ϕ に

$$\mathrm{EG}_\phi(q, r) = \phi\big((1-\pi)\phi^{-1}q(x) + \pi\phi^{-1}r(x) - \kappa_\phi(\pi)\big) \tag{4}$$

が定まり，これを一般化指数測地線と呼ぶ (Eguchi-Komori, 2015)．同様な考えから一般化 KL ダイバージェンスを導出すると，ピタゴラス定理が示される．次の節では一般化平均を直接に統計モデリングに応用することを考察する．

2　一般化平均と統計モデリング

一般化平均 (1) は正の数の平均で在ったが，実数の場合もつぎのように定めらる．実数 x と y に対して

$$\mathrm{RGM}_\phi(x,y) = \phi^{-1}\big((1-\pi)\phi(x) + \pi\phi(y)\big) \tag{5}$$

と定める．(1) の生成関数 ϕ の代わりに ϕ^{-1} を取っていることに注意する．統計の応用としては実数の代わりに，回帰関数，予測関数，エネルギー関数，ロス関数などの実数値関数の一般化平均を考えることができる．

ロジスティック回帰において p 変数の説明変数 $X = x$ を与えたとき 2 値反応変数 y の条件付き確率関数が

$$p(y|x) = \frac{\exp\{yf(x)\}}{1+\exp\{f(x)\}} \qquad (y = 0, 1) \tag{6}$$

とする．予測関数を一般化平均によって

$$f_\tau(x,\beta,\pi) = \frac{1}{\tau}\log\big(\textstyle\sum_{k=1}^K \pi_k \exp(\tau\beta_k^\top x_k)\big) \tag{7}$$

とし，準線形予測関数と呼ぶ．ここで τ は逆温度パラメータ，$x = (x_1, \cdots, x_K)$ と $\beta = (\beta_1, \cdots, \beta_K)$ は p 変数の同じ K 分割とする．Cf. Omae et al. (2017). 逆温度パラメータ τ を極限 ∞ を取ると $f_\tau(x,\beta,\pi) = \max_{1\le k\le K} \beta_k^\top x_k$ となり，極限 $-\infty$ を取ると $f_\tau(x,\beta,\pi) = \min_{1\le k\le K} \beta_k^\top x_k$ となる．極限 0 を取ると線形予測関数に帰着される．

$Y = 0$ のときの X の条件分布 $p(x|Y=0)$ は正規分布 $N(\mu_0, \Sigma)$ に従っているが，$Y = 1$ のときの X の条件分布 $p(x|Y=1)$ は混合正規分布 $\sum_{k=1}^K \pi_k^* N(\mu_k, \Sigma)$ に従っていると仮定する．これは $Y = 0$ サンプルは均一な母集団から得られたが $Y = 1$ サンプルは非均一な異質な母集団の混合から得らた状況を考えている．このとき，$\tau = 1$ のとき

$$f_\tau(x,\beta,\pi) = \log\frac{p(x|Y=1)}{p(x|Y=0)} \tag{8}$$

が成立する．このような場合，準線形予測関数の判別の最適性が示される．

準線形予測関数の定義において説明変数 x の K 分割が必要であるが，これは教師なし学習によって構成できる．典型的にはクラスター分析によって K 分割が求まる．またスパース学習を準線形モデル (7) をロジスティック回帰 (6) に代入したモデルの対数尤度関数に (7) のパラメータ β_k の積の L_1 ペナルティを入れたものを考える方法も有力である．

比例ハザードモデル，分割クラスタリング，混合ロス関数などについても一般化平均を使うと興味深い展開ができる．線形モデルから準線形モデルへの拡張や，エネルギー関数やロス関数の混合について，幾つかの新しい知見が得られた．しかし，未だ多くの未解決問題が山積しており，これらは今後の課題と残っている．線形モデルを柔軟に結合する統計方法として着実な完成が近い将来になされることが望まれる．

参考文献

Amari, S. I., & Nagaoka, H. (2007). Methods of information geometry (Vol. 191). American Mathematical Soc.

Eguchi, S., & Komori, O. (2015). Path connectedness on a space of probability density functions. In International Conference on Networked Geometric Science of Information (pp. 615-624). Springer, Cham.

Omae, K., Komori, O., & Eguchi, S. (2017). Quasi-linear score for capturing heterogeneous structure in biomarkers. BMC bioinformatics, 18(1), 308.

非正規化モデルを用いた推定法

はこだて未来大学, 理研 AIP 竹之内高志

概要: 離散空間上の確率モデルの推定には, しばしば正規化項を得るために莫大な計算量が必要となる. そのため, 正規化項の計算を回避しつつパラメーターを推定するための手法が研究されている. 本研究では, e-混合モデルの拡張によって得られるモデルと γ-ダイバージェンスを組み合わせることで得られる推定量の性質について議論する.

1 導入

X を d 次元の離散空間 $\mathcal{X}$ 上の確率変数ベクトルとし, 確率モデル

$$\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}},\ q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\psi_{\boldsymbol{\theta}}(\boldsymbol{x})),\ Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}} \rangle$$

に着目し, パラメーター $\boldsymbol{\theta}$ を推定することを目的とする. ただし $\psi_{\boldsymbol{\theta}}$ は任意の関数, $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ は非正規化モデル, $\langle f \rangle = \sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ とし, $Z_{\boldsymbol{\theta}}$ は $\bar{q}_{\boldsymbol{\theta}}$ が確率モデルであることを要請するための正規化項とする. 高次元離散空間の場合, 正規化項 $Z_{\boldsymbol{\theta}}$ の計算にはしばしば指数オーダーの計算量が必要となるため, 確率モデルのパラメーター $\boldsymbol{\theta}$ の推定が困難である場合があり, 様々な近似法が提案されている.

データセット $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ に対応する経験分布を $\tilde{p}(\boldsymbol{x})$ とすると, 最尤推定量は経験分布 $\tilde{p}(\boldsymbol{x})$ と確率モデル $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ 間の KL ダイバージェンス最小化

$$\boldsymbol{\theta}_{MLE} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{KL}(\tilde{p}, \bar{q}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} -\log \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \quad \left(\mathrm{KL}(p,q) = \left\langle p\log\frac{p}{q} - p + q \right\rangle\right) \tag{1}$$

として定式化することができるが, 正規化項の計算に由来する計算量の問題から推定量の構成が困難であることが多い. 本稿では義球スコアと e-混合モデルの変形を用いて, 正規化項の計算を行わずに構築可能な推定量を提案する.

2 γ-ダイバージェンスを用いた推定法

2 つの正値測度 f, g に対して, γ-ダイバージェンスを以下で定義する.

$$D_\gamma(f,g) = \frac{1}{1+\gamma}\log\left\langle f^{1+\gamma}\right\rangle + \frac{\gamma}{1+\gamma}\log\left\langle g^{1+\gamma}\right\rangle - \log\left\langle fg^\gamma\right\rangle \tag{2}$$

ただし γ は正の定数であり, $D_\gamma(f,g) \geq 0$, $D_\gamma(f,g) = 0 \Leftrightarrow f \propto g$ が成り立つ. また, $\lim_{\gamma\to 0} D_\gamma(f,g) = \mathrm{KL}(f,g)$ となる. また, $\alpha(\neq 0, 1)$ を定数として, 経験分布 $\tilde{p}(\boldsymbol{x})$ と非確率モデル $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ の e-混合モデルを以下のように定義する.

$$\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} = \begin{cases} \left(\frac{n_{\boldsymbol{x}}}{n}\right)^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} & \boldsymbol{x} \in \tilde{\mathcal{X}} \\ 0 & \boldsymbol{x} \notin \tilde{\mathcal{X}} \end{cases}$$

ここで $\tilde{\mathcal{X}}$ はデータセット $\mathcal{D}$ に含まれる $\boldsymbol{x}$ からなる集合とする. 経験分布 (の累乗) との積を考えることにより, データセットに観測されていないドメインにおいては e-混合モデルの値はすべて 0 となることに注意する.

[Takenouchi and Kanamori(2017)] では以下の推定量を提案している. $\alpha \neq \alpha'$ として,

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} D_\gamma(\tilde{r}_\alpha^{\frac{1}{1+\gamma}}, \tilde{r}_{\alpha'}^{\frac{1}{1+\gamma}}) \tag{3}$$

Proposition 1. $\gamma = \frac{\alpha-1}{1-\alpha'}$ とすると, $D_\gamma(\tilde{r}_\alpha^{\frac{1}{1+\gamma}}, \tilde{r}_{\alpha'}^{\frac{1}{1+\gamma}})$ は $\boldsymbol{\theta}$ に関して凸関数となる.

Proposition 2. データを生成する分布が $p(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ を満たすと仮定する. $N(v, w)$ を平均 v, 分散 w の正規分布の密度関数, $I_{\boldsymbol{\theta}}$ をフィッシャー情報行列として, 推定量 $\hat{\boldsymbol{\theta}}$ の漸近分布は,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, I_{\boldsymbol{\theta}_p}^{-1}) \tag{4}$$

となる. したがって推定量は α, α', γ の値にかかわらず漸近有効である.

3 提案法

本稿では, e-混合モデルを以下のように拡張したモデルを考える. u を単調増加関数, ξ を u の逆関数として,

$$s_{u,\alpha}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x}) u\left(\alpha\xi(1) + (1-\alpha)\xi\left(\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})}\right)\right) \tag{5}$$

適当な変換によって一般性を失うことなく $\xi(1) = 0$ とすることができる. また, $u(z) = \exp(z)$ とすると (5) は e-混合モデルとなることに注意する. この混合モデルを用いて以下のような推定量を定義する.

$$\hat{\boldsymbol{\theta}}_{u,\alpha,\alpha'} = \operatorname*{argmin}_{\boldsymbol{\theta}} D_\gamma(s_{u,\alpha}^{1/(1+\gamma)}, s_{u,\alpha'}^{1/(1+\gamma)}) \tag{6}$$

Proposition 3. 推定量 $\hat{\boldsymbol{\theta}}_{u,\alpha,\alpha'}$ の漸近分布は

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, I_{\boldsymbol{\theta}_p}^{-1}) \tag{7}$$

となり, u, α, α' に関わらず漸近有効性を持つ.

References

[Takenouchi and Kanamori(2017)] Takashi Takenouchi and Takafumi Kanamori. Statistical inference with unnormalized discrete models and localized homogeneous divergences. *Journal of Machine Learning Research*, 18(56):1–26, 2017. URL `http://jmlr.org/papers/v18/15-596.html`.

医学データにおける予測モデリング再考

京都大学大学院医学研究科　大前　勝弘

医学分野における予測モデル，あるいはそこから算出されるリスクスコアの表現式には，その予測性能だけでなく，解釈性の高さがしばしば要請される．例えば，線形予測子のように解釈に易しいリスクスコアに結び付けられるような予測モデルのクラスの中から，より予測性能の高いモデルを選択し，比較的に小標本のデータから効率的な推定を達成することが望まれる．

しかしながら，当該分野におけるデータの特徴は，多くの場合にそれを妨げるものである．本発表では，それらの特徴の中からより顕著な性質として，(1) 異質性　(2) 小標本高次元性 ($n \ll p$) について注目し，特に 2 値の判別・予測問題において，これらの性質に適切に対応できるような予測モデリングについて再検討をする．

1　異質性と準線形予測子

異質性に対しては，共変量ベクトル x とパラメータ $\alpha = (\alpha_1, \cdots, \alpha_K)$，$\beta^\top = (\beta_1, \cdots, \beta_K)$ により定義される準線形予測子

$$f_\tau(x; \alpha, \beta) = \frac{1}{\tau} \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp\left(\tau\alpha_k + \tau\beta_k^\top x\right) \right) \tag{1}$$

による予測モデリングが提案されている [1]．これは，ある群（例えば，健常群）に対して別の群（例えば，疾患群）のデータが持つ異質性を，それぞれ等分散の正規分布と K コンポネントの混合正規分布を通じて表現した際に，これに対応する尤度比から自然に導かれる予測子である．準線形予測子は線形予測子を非線形に拡張した形で，解釈性を損なわないまま柔軟性を向上させているという点で，冒頭の課題をうまく克服できることが期待される．例えば，判別問題において汎用的な線形ロジスティック回帰モデルは，予測子部分を線形から準線形へと置き換えることにより容易に拡張される．拡張されたモデルにおける最尤法や，高次元データにおいて有用な正則化法などは線形の場合と同様にして組み合わせることが可能ではあるものの，モデルの混合表現が備える複雑性がしばしばパラメータ推定を不安定にさせる場合がある．この場合には，x の背反な分割 $\{x_1, \cdots, x_K\}$ により定義される制限準線形予測子

$$F_\tau^{Res}(x_1, \cdots, x_K, ; \alpha, \beta) = \frac{1}{\tau} \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp(\tau\alpha_k + \tau\beta_k^\top x_k) \right) \tag{2}$$

や，[2] で提案されているようなクロス L_1 罰則

$$P^{(c)}(\boldsymbol{\beta}) = \lambda^{(c)} \sum_{\ell \neq m} \sum_{j=1}^{p} |\beta_{\ell j} \beta_{m j}| \tag{3}$$

を用いた方法などが考えられ，シミュレーションや実データ解析において有望な結果が観察されている．

2　小標本高次元と CLIP 法

小標本高次元性は，伝統的な予測モデルおよび推定方法を容易に適用しにくいという問題を生じさせている．これに対して，予測モデルを学習する前段階のステップとして事前に予測に有用そうな変数を抜粋するか，スパース性を仮定した上で予測モデルの推定アルゴリズム中に変数選択を行えるようなアイデアを取り入れるのが標準的である．

しかしながら，このようなスパース性の仮定をデータから検証することは一般には難しく，データが期待しているようなスパース性をいつも備えているかは不明瞭である．スパース性を仮定することなく，正規理論の上で高次元小標本における推定に耐えうる予測子の推定方法を考える．これにおいて合理的な一つのアイデアは，予測子に係る説明変数をいくつかのサブグループに分けておき，それぞれで学習したものをつなぎ合わせるというものである．説明変数を組 $\{x_{(1)}, \cdots, x_{(k)}\}$ に分解した場合に，それぞれで構成した線形予測子を重み付き平均の形で足し合わせたものを，Combined Linear Predictor (CLIP) 関数と呼び，

$$f_{CLIP}(x; w, \beta) = \sum_{k=1}^{K} w_k f_k(x_{(k)}; \beta_k), \tag{4}$$

$$f_k(x_{(k)}; \beta_k) = \beta_k^\top x_{(k)} \tag{5}$$

で定義する．このような CLIP 関数の中で，各標本の部分ベクトル $x_{(k)}$ のクラス条件付き分布がそれぞれ $\mathrm{N}(\mu_{k1}, \Sigma_{k1})$ および $\mathrm{N}(\mu_{k0}, \Sigma_{k0})$ に独立に従うような 2 つの正規標本 $D_1 = \{x_{(k)1i}; i = 1, 2, \cdots, n_1, k = 1, 2, \cdots, K\}$ と $D_0 = \{x_{(k)0j}; j = 1, 2, \cdots, n_0, k = 1, 2, \cdots, K\}$ を分離するもののうち，AUC(Area Under the receiver operating characteristic Curve) を最大にするパラメータ w_k と β_k は任意の $k = 1, 2, \cdots, K$ に対して

$$w_k \propto \frac{\beta_k^\top (\mu_{k1} - \mu_{k0})}{\beta_k^\top (\Sigma_{k1} + \Sigma_{k0}) \beta_k}, \tag{6}$$

$$\beta_k \propto (\Sigma_{k1} + \Sigma_{k0})^{-1} (\mu_{k1} - \mu_{k0}) \tag{7}$$

で与えられ，結果的に得られた CLIP 関数による 2 値判別は，Fisher の線形判別の拡張と見なす事ができる．このような簡単な拡張において，シミュレーションや実データ解析において有望な結果が観察されており，更なる方法論のブラッシュアップが期待される．

参考文献

[1] Omae, K., Komori, O. and Eguchi, S. (2017) Quasi-linear score for capturing heterogeneous structure in biomarkers. *BMC Bioinformatics.* 18:308

[2] Omae, K. and Eguchi, S. (2018) Quasi-linear score for capturing heterogeneous structure in biomarkers. *submitted*

Post Selection Inference with Kernels

Makoto Yamada
Kyoto University, Japan, RIKEN AIP, Japan
myamada@i.kyoto-u.ac.jp

1 Introduction

Finding a set of features in high-dimensional data is an important problem with many real-world problems such as biomarker discovery [1] and document categorization [2], to name a few. In particular, finding a set of *statistically significant* features is crucial for scientific discovery.

Recently, a novel approach called the post selection inference (PSI) has been proposed [3, 4]. PSI algorithms tend to have higher detection power than data splitting approaches. However, only *linear* approaches which are built upon LASSO or other similar linear feature selection approaches are available so far. Since real-world datasets tend to have non-linear relationship, the existing linear approaches may fail to find a set of important features; this is a critical problem in practice. Moreover, existing PSI approaches are only applicable to univariate output. Thus, the applications of existing PSI methods is limited.

In this paper, we propose a kernel based PSI method `hsicInf`, which can find *statistically significant* features from non-linear and/or structured output data such as multi-dimensional output. Specifically, we develop a PSI algorithm for independence measures, and propose the HSIC based PSI algorithm. A clear advantage of `hsicInf` over existing approaches is that it can easily handle non-linearity and structured data through kernels. Namely, it can be used for wider range of applications including multi-class classification and multi-variate regression. Through synthetic and real-world experiments, we show that the proposed approach can find a set of *statistically significant* features for both regression and classification problems.

2 HSIC based Post Selection Inference (`hsicInf`)

In this section, we propose a PSI method with kernels. More specifically, we develop a new PSI framework based on an independence measure called the Hilbert-Schmidt Independence Criterion (HSIC) [5, 6].

Problem Formulation: Let us denote an input vector by $\boldsymbol{x} = [x^{(1)}, \ldots, x^{(d)}]^\top \in \mathbb{R}^d$ and the corresponding target vector $\boldsymbol{y} \in \mathbb{R}^{d_y}$. i.i.d. samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ have been drawn from a joint probability density $p(\boldsymbol{x}, \boldsymbol{y})$. The final goal of this paper is to first screen $k < d$ features of input vector $\boldsymbol{x}$ and then test whether the selected features are of *statistically significant* association to its output $\boldsymbol{y}$.

Marginal screening and post-selection inference: In this paper, we employ an estimate of the independence measure $\widehat{I}(X_m, Y)$, which measures the discrepancy from the independence between the m-th random variable X_m and its output variable Y, where the vector of independence measures denoted by $\boldsymbol{z} = [\widehat{I}(X_1, Y), \ldots, \widehat{I}(X_d, Y)]^\top$ follows a multi-variate normal distribution with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$:

$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then, we exploit the normality of $\boldsymbol{z}$ and combine it with the post selection inference framework recently developed by [4].

In order to develop post selection inference method for the independence measure $\boldsymbol{z}$, we confirm that the problem of selecting top k features in the decreasing order of $\widehat{I}(X., Y)$ can be represented as a linear selection event in the form of $\boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b}$ in Theorem 5.2 [4].

We denote the index set of the selected k features by $\mathcal{S}$, and that of the unselected $\bar{k} = d - k$ features by $\bar{\mathcal{S}}$. The fact that k features in $\mathcal{S}$ are selected and $\bar{k}$ features in $\bar{\mathcal{S}}$ are not selected is rephrased by

$$\widehat{I}(X_m, Y) \geq \widehat{I}(X_\ell, Y), \quad \text{for all } (m, \ell) \in \mathcal{S} \times \bar{\mathcal{S}}. \tag{1}$$

Here, we have in total $k\bar{k}$ constraints written as the linear inequalities with respect to $\boldsymbol{z}$. Then, the cumulative distribution function for each selected feature can be explicitly stated as follows.

Theorem 1 *[4]. Consider a stochastic data-generating process $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If a feature-selection event is characterized by $\boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b}$ for a matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ that do not depend on $\boldsymbol{z}$, then, for any fixed vector $\boldsymbol{\eta} \in \mathbb{R}^d$,*

$$F^{[V^-(\boldsymbol{A},\boldsymbol{b}), V^+(\boldsymbol{A},\boldsymbol{b})]}_{\boldsymbol{\eta}^\top \boldsymbol{\mu}, \boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}}(\boldsymbol{\eta}^\top \boldsymbol{z}) \;\; | \;\; \boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b} \quad \sim \quad \mathrm{Unif}(0, 1),$$

where $F^{[v,w]}_{t,u}(\cdot)$ is the cumulative distribution function of the uni-variate truncated normal distribution with the mean t, variance u, and lower and upper truncation points v and w, respectively. Furthermore, using $\boldsymbol{c} := \frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}}$, the lower and upper truncation points are given as

$$V^-(\boldsymbol{A}, \boldsymbol{b}) := \max_{j:(\boldsymbol{A}\boldsymbol{c})_j < 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}, \quad V^+(\boldsymbol{A}, \boldsymbol{b}) := \min_{j:(\boldsymbol{A}\boldsymbol{c})_j > 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}. \tag{2}$$

Hilbert-Schmidt Independence Criterion: In this paper, we employ the Hilbert-Schmidt Independence Criterion (HSIC) [5, 7] with a characteristic kernel (e.g., Gaussian kernel) as an independence measure $I(X, Y)$. More specifically, we employ the block HSIC estimator [6], which is given as the average of n/B HSICs $(\widehat{\eta}_1, \ldots, \widehat{\eta}_{n/B})$ and each $\widehat{\eta}_b$ is computed from B i.i.d. samples (block size). Here, we assume that n/B is an integer.

The empirical block HSIC score asymptotically follows normal distribution when B is finite and n goes to infinity, and thus, we can use the block HSIC for PSI based on Theorem 1. Note that, to ensure Gaussian assumption, we need to have relatively large number of samples n with a finite block size B.

Post Selection Inference: We consider the following hypothesis tests:

- $H_{0,m}$: $\mathrm{HSIC}(X_m, Y) = 0 \mid \mathcal{S}$ was selected,
- $H_{1,m}$: $\mathrm{HSIC}(X_m, Y) \neq 0 \mid \mathcal{S}$ was selected.

Then, the p-value of the m-th feature is estimated by using the Theorem 1.

References

[1] E. P. Xing, M. I. Jordan, R. M. Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, 2001.

[2] G. Forman. BNS feature scaling: An improved representation over TF-IDF for SVM text classification. In *CIKM*, 2008.

[3] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[4] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[5] A. Gretton, O. Bousquet, Alex. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.

[6] Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2017.

[7] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *NIPS*, 2007.

関数データに基づく交互作用モデルとその推定

滋賀大学データサイエンス学部　　松井　秀俊

1. 概要

複数の個体に対して経時的に測定されたデータを関数化処理し，関数化データ集合に基づく解析を行う方法は関数データ解析とよばれ，生命科学をはじめとした多くの分野でその有用性が報告されている (Ramsay and Silverman, 2005; Kokoszka and Reimherr, 2017)．特に，回帰分析を関数データ解析の枠組みへ拡張した関数回帰分析は，理論，応用の両側面から多くの研究が現在も行われている．本研究では，説明変数と目的変数が共に関数データとして与えられた場合，これらの関係を表す関数回帰モデルを扱う．特に本報告では，説明変数に関する 2 次の項を導入することで，関数説明変数の任意の時点間における交互作用を考慮に入れたモデルについて説明する．スカラー目的変数のモデルに対しては Yao and Müller (2010)，関数目的変数のモデルに対しては Luo and Qi (2018) によってその推定法が提案されている．本報告では，関数目的変数に対する交互作用モデルに対して，正則化法に基づく推定法と評価法について紹介する．そして，数値実験および実際のデータ解析を通して，手法の有効性を検証する．本報告の内容の詳細については，Matsui (2017) を参照されたい．

2. 関数目的変数に対する交互作用モデル

いま，説明変数と目的変数に関して，それぞれが関数構造を背景にもつ n 個の経時測定データの組 $\{(x_i(s),\, y_i(t));\, s \in \mathcal{S} \subset \mathbb{R},\, t \in \mathcal{T} \subset \mathbb{R},\, i = 1, \ldots, n\}$ が与えられたとする．このとき，説明変数と目的変数との関係を表す関数回帰モデルを次で与える．

$$y_i(t) = \alpha(t) + \int_{\mathcal{S}} x_i(s)\beta(s,t)ds + \iint_{\mathcal{S}^2} x_i(r)x_i(s)\gamma(r,s,t)drds + \varepsilon_i(t). \tag{1}$$

ここで，$\alpha(t)$ はベースライン関数，$\beta(s,t)$，$\gamma(r,s,t)$ はそれぞれ 1 次，2 次の項に対する係数関数，$\varepsilon_i(t)$ は誤差関数とする．特に，$\gamma(r,s,t)$ は，$y_i(t)$ の各時点における，説明変数 $x_i(\cdot)$ の異なる 2 時点 r, s での交互作用の重みを表したものとみなすことができる．ベースラインおよび係数関数は，基底関数 $\boldsymbol{\phi}(s) = (\phi_1(s), \ldots, \phi_{M_x}(s))^T$ および $\boldsymbol{\psi}(t) = (\psi_1(t), \ldots, \psi_{M_y}(t))^T$ を用いて次のように表されると仮定する．

$$\alpha(t) = \sum_{l=1}^{M_y} a_l\psi_l(t) = \boldsymbol{a}^T\boldsymbol{\psi}(t), \quad \beta(s,t) = \sum_{k,l} b_{kl}\phi_k(s)\psi_l(t) = \boldsymbol{\phi}(s)^T B\boldsymbol{\psi}(t),$$
$$\gamma(r,s,t) = \sum_{h,k,l} \gamma_{hkl}\phi_h(r)\phi_k(s)\psi_l(t) = \{\boldsymbol{\phi}(s) \otimes \boldsymbol{\phi}(r)\}^T \Gamma_{(3)}^T \boldsymbol{\psi}(t).$$

ただし，$B = (b_{kl})_{kl}$，$\Gamma_{(3)}$ は 3 次元テンソル $\underline{\Gamma} = (\gamma_{hkl})_{hkl}$ を第 3 配列に関して列方向に行列化したもので，$\otimes$ はクロネッカー積を表す．さらに，説明変数 $x_i(s)$ も，基底関数展開によって $x_i(s) = \boldsymbol{w}^T\boldsymbol{\phi}(s)$ と表されるとする．ここで $\boldsymbol{w} = (w_1, \ldots, w_{M_x})^T$ は平滑化などを用いて得られる既知の値からなるベクトルとする．以上の仮定を用いると，関数回帰モデル (1) は次で表すことができる．

$$\begin{aligned} y_i(t) &= \boldsymbol{a}^T\boldsymbol{\psi}(t) + \boldsymbol{w}_i^T\Phi B\boldsymbol{\psi}(t) + (\boldsymbol{w}_i \otimes \boldsymbol{w}_i)^T(\Phi \otimes \Phi)\Gamma_{(3)}^T\boldsymbol{\psi}(t) + \varepsilon_i(t) \\ &= \boldsymbol{z}_i^T\Theta^T\boldsymbol{\psi}(t) + \varepsilon_i(t). \end{aligned}$$

ただし，$\Phi = \int \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T ds$, $\boldsymbol{z}_i = (1, \boldsymbol{w}_i^T\Phi, (\boldsymbol{w}_i \otimes \boldsymbol{w}_i)^T(\Phi \otimes \Phi))^T$ で，$\Theta = (\boldsymbol{\alpha}\ B^T\ \Gamma_{(3)})^T$ はパラメータ行列とする．

2. モデルの推定と評価

モデルに含まれるパラメータを推定するために，誤差関数 $\varepsilon_i(t)$ が，次の構造を持つと仮定する (Fan and Zhang, 2000; Shi and Choi, 2011)．

$$\varepsilon_i(t) = \tau_i(t) + e_i(t), \tag{2}$$
$$\tau_i(t) \sim GP(0, k(\cdot,\cdot)),\quad k(t,t') = \nu_1 \exp\left\{-\frac{\nu_2}{2}(t-t')^2\right\},\quad e_i(t) \overset{\text{i.i.d.}}{\sim} N(0,\nu_3).$$

ただし，$GP(0, k(\cdot,\cdot))$ 平均 0，共分散関数 $k(\cdot,\cdot)$ をもつガウス過程とし，$\nu_1 > 0, \nu_2 > 0, \nu_3 > 0$ は分散に含まれるパラメータとする．目的変数に対応するデータ $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ は，実際には各個体 i が n_i 個の時点 $t_{i1}, \ldots, t_{in_i}$ で観測されているとすると，$\boldsymbol{y}_i$ は平均 $\Psi_i\Theta\boldsymbol{z}_i$，分散共分散行列 $\Sigma_i = K_i + \nu_3 I_{n_i}$ の正規分布に従う．ただし $\Psi_i = (\boldsymbol{\psi}(t_{i1}), \ldots, \boldsymbol{\psi}(t_{in_i}))^T$，$K_i = (k(t_{ij}, t_{ij'}))_{jj'}$ とする．このことを用いて，パラメータ Θ，$\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3)^T$ を，正則化最尤法，すなわち次の正則化対数尤度関数の最大化により推定する．

$$\ell_\lambda(\Theta, \boldsymbol{\nu}) = \ell(\Theta, \boldsymbol{\nu}) - \frac{n\lambda}{2}(\mathrm{vec}\Theta)^T\Omega(\mathrm{vec}\Theta).$$

ここで，$\ell(\Theta, \boldsymbol{\nu}) = \sum_{i=1}^n \log f(\boldsymbol{y}_i|\Theta, \boldsymbol{\nu})$ は対数尤度関数，$\lambda > 0$ は正則化パラメータで，Ω は非負値定符号行列である．

いま，$\boldsymbol{\nu}$ が既知のとき，Θ は次で推定される．

$$\mathrm{vec}\hat{\Theta} = \left\{\sum_{i=1}^n X_i^T\Sigma_i^{-1}X_i + n\lambda\Omega\right\}^{-1}\left(\sum_{i=1}^n X_i^T\Sigma_i^{-1}\boldsymbol{y}_i\right).$$

ただし $X_i = \boldsymbol{z}_i^T \otimes \Psi_i$ とする．実際には $\boldsymbol{\nu}$ は未知であり，かつこれを解析的に導出することは困難である．そこで，$\boldsymbol{\nu}$ については，ここでは Newton-Raphson 法を用いる．パラメータ Θ と $\boldsymbol{\nu}$ を交互に更新することで，それぞれの推定値 $\hat{\Theta}$ と $\hat{\boldsymbol{\nu}}$ を得る．

参考文献

Fan, J. and Zhang, J. (2000), "Two-step estimation of functional linear models with applications to longitudinal data," *J. Roy. Statist. Soc. Ser. B*, 62, 303–322.

Kokoszka, P. and Reimherr, M. (2017), *Introduction to functional data analysis*, CRC Press.

Luo, R. and Qi, X. (2018), "Interaction model and model selection for function-on-function regression," *J. Comput. Graph. Statist.*, to appear.

Matsui, H. (2017), "Quadratic regression for functional response models," *arXiv preprint arXiv:1702.02009*.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis (2nd ed.)*, New York: Springer.

Shi, J. Q. and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, Boca Raton: CRC Press.

Yao, F. and Müller, H. G. (2010), "Functional quadratic regression," *Biometrika*, 97, 49–64.

カーネル主成分分析に基づく
高次元データのクラスタリング

筑波大学・数理物質科学　　中山　優吾
筑波大学・数理物質系　　矢田　和善
筑波大学・数理物質系　　青嶋　誠

1　はじめに

本講演では，高次元データのクラスタリングを考えた．

2 つの d 次元分布を Π_1, Π_2 と名付け，それぞれ平均 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ と，共分散行列 $\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ をもつと仮定する．いま，データは，p.d.f.

$$f(\boldsymbol{x}) = \varepsilon_1 f_1(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \varepsilon_2 f_2(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),\ \varepsilon_1 + \varepsilon_2 = 1\ (\varepsilon_i > 0)$$

をもつ混合分布からの標本とみなす．ここで，$f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は Π_i の p.d.f. である．この母集団から n (≥ 2) 個のデータを無作為に抽出し，データ行列を $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ とする．そのとき，$\mathrm{Var}(\boldsymbol{x}_i) = \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2 + \varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\ (= \boldsymbol{\Sigma})$ である．$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とおく．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_d]$ で $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ と分解する．さらに，$\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}] = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}$ として，$\boldsymbol{Z} = (z_{ij})$ と表記する．一方で，標本共分散行列 $\boldsymbol{S}_n = (n-1)^{-1}\sum_{j=1}^n (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)^T$ の第 j 固有値を $\hat{\lambda}_j$ とし，それに対応する固有ベクトルを $\hat{\boldsymbol{h}}_j$ とする．ただし，$\bar{\boldsymbol{x}}_n = \sum_{j=1}^n \boldsymbol{x}_j/n$ である．

2　主成分スコアにおける高次元一致性

Yata and Aoshima (2015) は，第 1 主成分スコア $s_{1j}\ (= \sqrt{\lambda_1} z_{1j})$, $j = 1, ..., n$ について，平均ベクトル間の距離 Δ に関する条件

$$\frac{\lambda_{\max}(\boldsymbol{\Sigma}_i)}{\Delta} \to 0 \ \text{ as } d \to \infty \text{ for } i = 1, 2 \tag{1}$$

のもとで

$$\operatorname*{plim}_{d\to\infty}\frac{s_{1j}}{\sqrt{\lambda_1}}=\begin{cases}\sqrt{\varepsilon_2/\varepsilon_1}, & \boldsymbol{x}_j\in\Pi_1,\\ -\sqrt{\varepsilon_1/\varepsilon_2}, & \boldsymbol{x}_j\in\Pi_2\end{cases}$$

なる一致性を示した．ただし，$\lambda_{\max}(\boldsymbol{\Sigma}_i)$ は $\boldsymbol{\Sigma}_i$ の最大固有値を表す．つまり，第 1 主成分スコアを精度よく推定できれば，その符号から高次元データを分類することができる．実際，Yata and Aoshima (2015) は規準化した標本主成分スコア

$$\hat{z}_{1j}=\{n/(n-1)\}^{1/2}\hat{\boldsymbol{h}}_1^T(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_n)/\hat{\lambda}_1^{1/2},\ j=1,...,n$$

について，適当な正則条件と

$$\frac{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta^2}\to 0\ \ \text{as } d\to\infty \text{ for } i=1,2 \tag{2}$$

のもとで

$$\operatorname*{plim}_{d\to\infty}\hat{z}_{1j}=\begin{cases}\sqrt{n_2/n_1}, & \boldsymbol{x}_j\in\Pi_1,\\ -\sqrt{n_1/n_2}, & \boldsymbol{x}_j\in\Pi_2\end{cases} \tag{3}$$

なる一致性を示した．しかしながら，条件 (2) を満たすほど Δ が十分に大きくなければ，従来の PCA を用いて高次元データを分類することは困難である．

一方で，非線形な構造をもつ場合はカーネル PCA が有効であることが知られている．そこで，高次元非線形構造まで加味したクラスタリング手法を与えるために，ガウシアンカーネル

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\gamma)\ \ (\gamma>0)$$

を用いた高次元カーネル PCA を考える．高次元におけるカーネル主成分スコアの漸近的性質を導出し，平均ベクトル間の距離 Δ だけでなく，共分散行列間の距離にもよって (3) のような主成分スコアの高次元一致性をもつことを示した．当日は，カーネル PCA と従来の PCA との理論的な比較を行い，その性能を数値実験と実データ解析を用いて検証した．

参考文献

[1] Yata, K. and Aoshima, M. (2015). Principal component analysis based clustering for high-dimension, low-sample-size data, arXiv:1503.04525.

ビッグデータと生態学

京都大学フィールド科学教育研究センター

伊勢　武史

生態学は、環境のなかで暮らす生物についての学問。研究のためにはデータの取得が不可欠だが、広大で変化にとんだ野外が主なフィールドであるため、古典的な研究ではデータの量や質に限界があることも多々あった。

時代は進み、近年では人工知能とビッグデータが台頭するようになった。このような情報科学の流れは、古き良き生態学に何をもたらすだろうか。生態学の目的が、自然環境や生物についての理解（基礎科学）や、環境や生物多様性の保全（応用科学）であるならば、そのために情報科学のツールを効果的に用いることはできないだろうか。このように考えて進めている私の近年のプロジェクトから、代表的なものを紹介した。アイデアが浮かんだらすぐに試してみたくなるゆえに雑多であるが、主な話題は以下のとおり。

ディープラーニングの応用について。ディープラーニングは近年注目を集めている人工知能の代表格であり、コンピュータ上で仮想的に脳細胞のネットワークのようなものを構築することで、画像識別などに威力を発揮している。今回は、簡便な手法で不定形な植物をテクスチャに基づいて分類する手法と、それを用いればコケ植物（図 1）・竹林・セイタカアワダチソウ・ミカン畑・カキの「ヘタスキ（黒変）」（表 2）など多くの植生・植物体を非破壊・非接触の画像データから高精度で識別できることを紹介した。

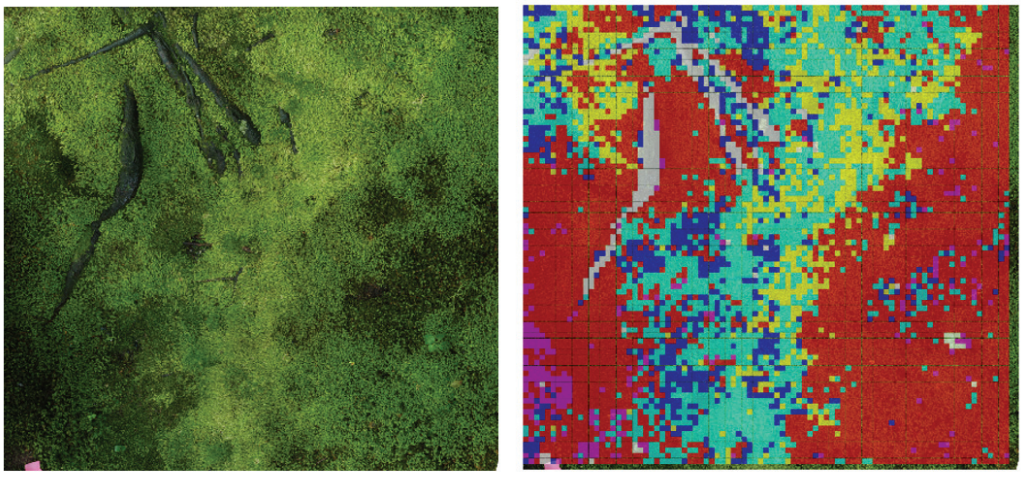

図 1 。ディープラーニングによるコケ植物の自動識別の結果。京都市無鄰菴庭園における複数のコケ植物の分布を識別することができた。

表２。ディープラーニングによるカキのヘタスキ（黒変）の非破壊検査の結果。約 2800 枚の教師画像によって構築されたモデルは、81%の精度でカキのヘタスキを検出することができた。

		深層学習の予測	
		ヘタスキなし	ヘタスキあり
実際に切った結果	ヘタスキなし	451	44
	ヘタスキあり	78	84

ディープラーニングはすばらしい性能を発揮する一方で、その内部はブラックボックスにたとえられる。人工知能が何を見ているか、人間に理解するのがむずかしいからだ。これを割り切ってとらえることで、ブラックボックス的だが精度の高い将来予測をすることができるかもしれない。その手法は、たとえば物理学的メカニズムの積み上げでボトムアップ的につくられる従来の気候予測モデルと好対照であり、相互に補完する関係にあると考えている。

ビッグデータについて。デジタルの時代になり、研究に使えるデータの量は飛躍的に増大した。たとえば前段で利用したデジカメ写真などのデジタル画像もビッグデータであり、人工衛星がリアルタイムで撮影する画像もビッグデータである。データ同化の一種である粒子フィルタを用いることで、abrupt な現象に支配される陸上生態系のシミュレーションモデルを観測結果によって最適化できることを示した。具体的には、植物が春に葉をつける展葉という現象を、過去の人工衛星観測と気象データからモデル最適化することに成功し、日本全国規模で利用可能な予測モデルを構築することができた（図２）。さらに、人工衛星画像そのものにひそむ変動を change point analysis によって検出する実験についても紹介した。ビッグデータ分析に適したスパース推定などの統計手法が導く生態学の最前線についても語り、本論をしめくくった。

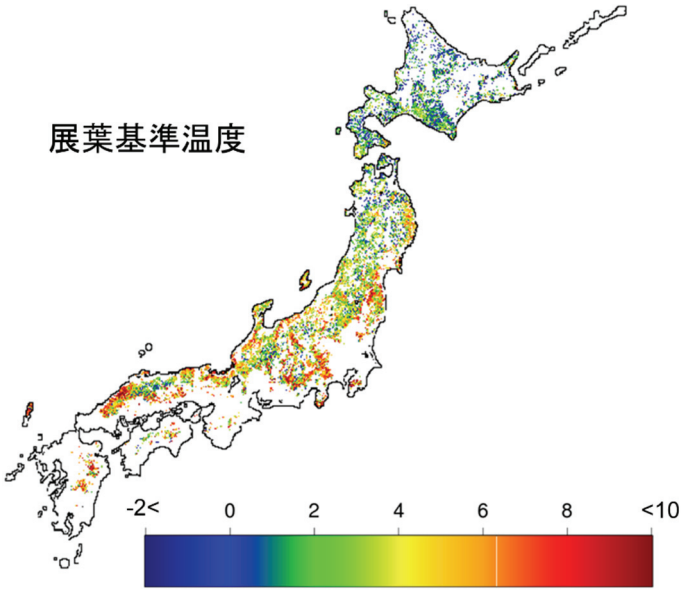

図２。粒子フィルタを用いたデータ同化により、日本全国規模での植物の季節性についてのモデル最適化に成功した。

潜在変数モデルによるトランスクリプトームの予測

岩山 幸治
滋賀大学データサイエンス教育研究センター, JST さきがけ

1 はじめに

生物が本来生息する野外環境ではその環境条件は大きく複雑に変動する．制御された実験室環境で得られてきた生物の環境応答に関する多くの結果は野外環境における観測と必ずしも一致しない [1, 2, 3]. 本研究では、未知の環境におけるトランスクリプトームの予測を行うと同時に，類似した環境応答を示す遺伝子群の抽出とその特徴づけを行うために，自然言語処理の分野で提案されたトピックモデルを元に，RNA-Seq データと環境条件の関係を記述する潜在変数モデルを提案する．

2 提案モデル

提案する生成モデルは以下の通りである．サンプル s におけるトピックの分布は，Dirichlet 分布から生成される．Dirichlet 分布のパラメータ（$\boldsymbol{\alpha}_s = (\alpha_{s1}, \ldots, \alpha sT)$）は，サンプリング時の気象条件や時刻と言った付随する情報を要素に持つベクトルを $\boldsymbol{x}_s$ とすると，$\alpha_{st} = \exp(\boldsymbol{x}_s^T \boldsymbol{\lambda}_t)$. となる．ここで、回帰係数 λ_{tk} の事前分布は平均 0，分散 σ_{tk}^2 の正規分布とする．次に，各サンプルの各遺伝子について，確率 $\boldsymbol{\theta}_s$ の多項分布でトピックを割り当てる．各トピックは平均的な発現量からの差分 $\boldsymbol{\eta}_k$ で特徴づけられ，その事前分布は平均 0，分散 l_{ti}^2 の正規分布とする．サンプル s の遺伝子 i にトピック t が割り当てられたとき、そのリードカウントは平均 $\nu_s \exp(m_i + \eta_{t,i})$，Dispersion$\phi_i$ の負の二項分布で生成される．ここで，ν_s はサンプル間の総リード数の違いを表すためのパラメータである．

事後分布を近似した分布 $q(\boldsymbol{\theta})$，$q(\boldsymbol{z})$，$q(\boldsymbol{\eta})$ について，対数周辺尤度の下界（ELBO）を最大化する変分ベイズ法で推定する．最初に $\eta_{k,i}$ の事後分布については，Doubly Stochastic Variational Inference [4] により推定する．$\eta_{k,i}$ の事後分布を平均 $\mu_{k,i}$，標準偏差 $C_{k,i}$ の正規分布とし，標準正規分布に従う乱数 $\omega_{t,i}$ を用いて，$\hat{\eta}_{t,i} = C_{t,i}\omega_{t,i} + \mu_{t,i}$ とすることで事後分布からのサンプリングを行う．そのもとでのパラメータ $C_{k,i}$ 及び $\mu_{k,i}$ に関する確率勾配を計算し，これらを更新する．トピックの分布 $q(\theta_s)$ と潜在変数の事後分布 $q(z_{si})$ は，ELBO を変分し，停留点を求めることで推定できる。Dispersion$\boldsymbol{\phi}$ 及び回帰係数 $\boldsymbol{\lambda}$ は，点推定を行う．ELBO のこれらのパラメータに関する勾配を求め，逐次更新を行う．

3 イネのトランスクリプトームの予測

実験圃場で，2 時間おき 24 時間を 1 セットとし，2013 年 6 月 13 日から 7 月 18 日まで 1 週間おきに収集されたイネのトランスクリプトームについて，7 月 4 日の 24 時間サンプリングを検証データ，その他を訓練データとしてモデルの推定を行った．圃場の近くに位置する気象庁の地上気象観測地点の気温と全天日射量について，サンプリングを行った時刻，サンプリング前の 3 時間，6 時間，12 時間，24 時間の平均値及び，田植え後日数とサンプリング時刻をサンプルの付随情報とした．

トピック数を 10 として，モデルの推定を行った．例として、時計関連遺伝子の一部の予測結果を図 1 に示す．良好に予測できている遺伝子もあるが，振

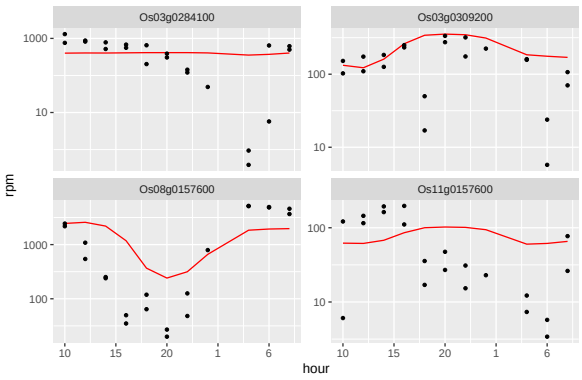

図 1: 時計遺伝子の発現量の予測結果．横軸がサンプリングの時刻，縦軸が対数 rpm 値に対応する．黒い点が実測値，赤い線で予測値を示す．

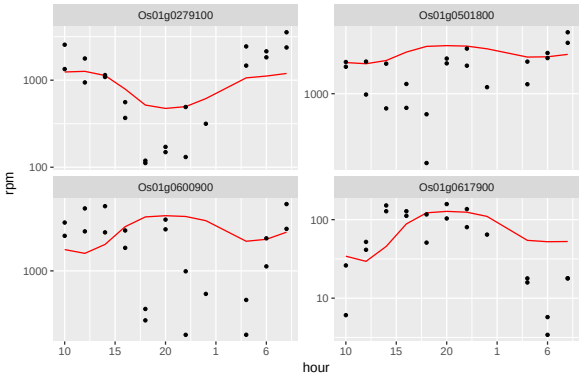

図 2: GO:0015979（photosynthesis）の遺伝子の発現量の予測結果．

幅が小さすぎるものや，発現量のピークの時刻が全く異なるものもある．次に、遺伝子オントロジーにおいて，“GO:0015979 (photosyntesis)” のアノテーションがつけられている遺伝子の一部についての発現量の予測結果を図 2 に示す．概ねよく予測できている遺伝子と，実測とは全く異なる変動を示しているものが見られる．

4　まとめ

トピックモデルをもとに，未知の環境におけるトランスクリプトームを予測するモデルを提案した．一部の遺伝子では良い予測を実現できたが，全体を少ない自由度で説明するために，得られるパターンに多様性が見られなかった．また，今回のモデルは，トピックの出現確率と環境条件の間に線形なモデルを用いたが，生物の環境応答には非線形な特性が見られる．そうした特性を取り入れるため，回帰を行わないモデルで推定されるトピックの確率と環境条件を比較し，モデルの形について検討を行っていきたい．

参考文献

[1] Cynthia Weinig, Mark C Ungerer, Lisa A Dorn, Nolan C Kane, Yuko Toyonaga, Solveig S Halldorsdottir, Trudy FC Mackay, Michael D Purugganan, and Johanna Schmitt. Novel loci control variation in reproductive timing in Arabidopsis thaliana in natural environments. *Genetics*, Vol. 162, No. 4, pp. 1875–1884, 2002.

[2] Yogesh Mishra, Hanna Johansson Jänkänpää, Anett Z Kiss, Christiane Funk, Wolfgang P Schröder, and Stefan Jansson. Arabidopsis plants grown in the field and climate chambers significantly differ in leaf morphology and photosystem components. *BMC Plant Biology*, Vol. 12, No. 6, 2012.

[3] Russell L Malmberg, Stephanie Held, Ashleigh Waits, and Rodney Mauricio. Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse. *Genetics*, Vol. 171, No. 4, pp. 2013–2027, 2005.

[4] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*, pp. 1971–1979, 2014.

ゲノムワイド関連解析を応用した混植効果の推定と虫害予測

佐藤安弘[1,2*]、清水(稲継)理恵[3]、山﨑美紗子[3]、清水健太郎[3]、永野惇[2]
[1]JST さきがけ専任, [2]龍谷大学農学部, [3]チューリッヒ大学; *✉: sato.yasuhiro.36c@kyoto-u.jp

1. 背景

通常、ゲノムワイド関連研究(GWAS)では、ある生物個体がもつ表現型はその個体自身の遺伝子型によって決まると仮定される。しかし、野外の実環境下では、各個体が独立に生育しているわけではなく、様々な遺伝子型をもつ個体が集まった集団として存在する。このような条件下では、ある個体の表現型が自身の遺伝子型だけでなく、周囲の他個体にも影響されることがある。特に、固着性の生物である植物では、複数の遺伝子型が混在すると集団全体の虫害が抑制されることや、特定の遺伝子型が虫害を逃れられることが知られている。

本発表では、野外栽培したシロイヌナズナ *Arabidopsis thaliana* 野生系統を対象に、近接個体間の相互作用を考慮した GWAS を提案した。自然発生する植食性昆虫による食害を GWAS の対象表現型として扱い、近傍個体間の相互作用に関わるシロイヌナズナのゲノム領域を探索した。さらに、混植による虫害制御に向けた予測構築のために、表現型の予測値を外挿して互いに隣り合うと虫害が抑制できる系統の組み合わせを推定した。

2. 方法

自殖性の植物を想定して、GWAS の線形モデルに近接する植物個体間の相互作用を表す項を加えた。ある遺伝子座 g において 2 つの対立遺伝子が存在すると仮定して、片方の対立遺伝子状態を 1 もう一方を-1 と便宜的に表す。着目する i 番目の個体の遺伝子型を g_i と近傍 j 番目の個体の遺伝子型を g_j とすると $g_{i(j)} \in \{-1, +1\}$である。このとき、i 番目と j 番目の個体の遺伝子座 g における対立遺伝子が同じであれば、g_ig_j =1*1=1 または g_ig_j=-1*-1=1 となり、共変量に 1 が代入される。逆に遺伝子座 g における対立遺伝子が異なれば、g_ig_j =-1*1=-1 または g_ig_j=1*-1=-1 となり、共変量には-1 が代入される。これを近傍 1~K 番目の個体まで和をとり、個体数 K で割って標準化した後、共変量にかかる係数 β の効果を SNP 毎に推定することで、近傍個体間の相互作用に関わる SNP を探索した。

GWAS に用いるデータを取得するために、シロイヌナズナ 1600 個体を、2017 年 7 月と 2018 年 7 月の 2 回にわたって、スイス連邦チューリッヒ大学の野外圃場(47°23'N, 8°33'E)に移植した。200 系統は欧州、米国、日本を含む世界各地の野生系統から選ばれた。種を Jiffy-seven®に播種し、4℃で 1 週間低温処理した後、実生を 1.5 ヶ月の間、開花させないように短日条件で栽培した。栽培した植物は Jiffy-seven ごと培養土を詰めた半径 6 cm の鉢に移植して、1 区画内に 200 系統を無作為に配置しつつ合計 8 区画を野外圃場に設置した。移植後 3 週間の間、1 名の調査者が 2-3 日に 1 回の頻度で全種の食害昆虫の個体数と食痕数を記録した。

上述の近傍個体を考慮した GWAS 式 1 を基に 2 つの解析手法を試した。1 つ目は、候補遺伝子

を探索するために、通常の GWAS と同じく線形混合モデルを用いて 1 遺伝子座ごとに繰り返し Wald 検定を行った。2 つ目は、限られた SNP 数から表現型の予測値を得るために、Lasso に基づく重回帰を行った。いずれの解析でも、Minor allele frequency が 15%以上かつ隣り合う遺伝子座の連鎖が $r^2<0.8$ の計 49 万 SNP を説明変数に用いた。ノミハムシ 2 種による食痕の数を対象の表現型とし、正規性を改善するために、対数変換を施した値を応答変数とした。線形混合モデルには R の coxme パッケージに含まれる lmekin 関数を、Lasso には Python 版の glmnet 関数をそれぞれ利用した。Lasso の罰則係数 λ を決める際には、8 反復×2 年分のデータを 1 反復ごとに分割した交差確認により、検証用データ間で平均 R^2 が最大となる λ の値を採用した。単植(=周囲が全て自身と同じ系統)または混植(周囲が全て自身と違う系統)の状態を仮定し、共変量に 1 または-1 を代入することで、β の推定値から表現型の予測値を外挿した。これらの外挿を、実験に用いた 200 系統およびゲノム情報が得られた全てのペアについて行い、その中から混植条件で食痕数が減少するペアを推定した。

3. 結果と今後の課題

線形混合モデルによる GWAS の結果、近傍個体の効果について、シロイヌナズナの 1 番および 2 番染色体にそれぞれ 1 つずつ、Bonferroni の多重比較補正後 p=0.05 の水準で有意なピークを見出すことができた。また、自身の遺伝子型の効果については有意なピークは得られなかった。さらに、全体の機能を推定するため、P 値で上位 0.1%の SNP の周辺 10kbp 以内に位置する遺伝子群に対して Gene Ontology (GO)解析を適用した。Fisher の正確確率検定 p=0.05 水準で有意となった GO から最も派生的なものを探索した結果、近傍個体の効果について isoprenoid biosynthetic process および plant epidermis development の注釈が見られた。

Lasso による重回帰と交差確認の結果、196 変数の回帰式で食痕数の全変動の約半分を説明することができた。さらに、β の推定値に基づいて、単植と混植条件での予測値の差を計算したところ、実験に用いた 200 系統では混植よりも単植で虫害が増加するペアが全体的に多く、抑制されるペアは全体のわずか 12%であった。他方、ゲノム情報が既知の 2029 系統全てに対して、同様の混植-単植間で虫害予測値を比較したところ、混植よって虫害が抑制されるペアは全体の 59%と、増加するペアに比べてやや多かった。

以上の結果から、GWAS の線形モデルを拡張することで、近接する植物個体間の相互作用に関連したゲノム領域と候補遺伝子を特定することができた。Lasso に基づく食痕数の予測値からは、混植すると互いに虫害が抑制される組み合わせが複数みられた。今後はこれらのペアのうちいくつかを、実際に野外圃場で混植して予測を検証する予定である。

マテリアルズインフォマティクスの最前線

吉田 亮 [a,b]

[a] 情報・システム研究機構 統計数理研究所　[b] 物質・材料研究機構

There has been a growing interest in using machine learning (ML) to facilitate enormous savings in time and cost on the discovery and development of new materials. In this talk, I describe some key drivers of ML technologies to achieve this goal.

The first topic is focused on ML-assisted materials design. In general, the material spaces are considerably high-dimensional. For instance, the chemical space of small organic molecules is known to contain as many as 10^{60} candidates, whereas the total number of currently known molecules is at most 10^{8}. The problem entails a considerably complicated combinatorial optimization where it is impractical to fully explore the vast landscape of structure-property relationships. We developed an inverse material design algorithm by the integration of ML and quantum chemistry calculation. The objective of the design calculation is to generate promising hypothetical materials that exhibit desired properties of various kinds. The emergence of such ML algorithms to exhaustively search in such a huge space is expected to accelerate the pace of expanding the frontier in the vast universe of materials.

The second topic is on a subject of data scarcity. In recent years, various kinds of databases have begun to be developed with the aim to transform materials science into being fully data driven. However, the volume and diversity of data being accumulated remain far from enabling us to fully enjoy remarkable advances recently made in ML. A ML framework called transfer learning has the great potential to break this barrier in which various material properties, such as physical, electronic, thermodynamic, mechanical properties, are closely related to each other. For a target property with a limited supply of training data, models on physically related proxy properties are pre-trained on large amounts of data, which capture features of materials generally applicable to the target task. Re-purposing such ML-acquired knowledge on a new task provides an outstanding prediction ability as highly experienced experts are capable of rationally making inferences even on considerably less experienced tasks. We have developed a pre-trained model library which can be used to predict various properties of small molecules, polymers and inorganic solid-state materials. Along with this library, I demonstrate outstanding successful applications of transfer learning, which exploit the ML-discovered transferability underlying different properties even across different types of materials.

[1] Ikebata, H., Hongo, K., Isomura, T., Maezono, R., Yoshida, R. (2017) Bayesian molecular design with a chemical language model, *Journal of Computer-Aided Molecular Design*, 31(4):379-391.
[2] R package iqspr: https://github.com/yoshida-lab/iqspr
[3] XenonPy: Python library on representation & learning for materials data
http://xenonpy.readthedocs.io/en/latest/

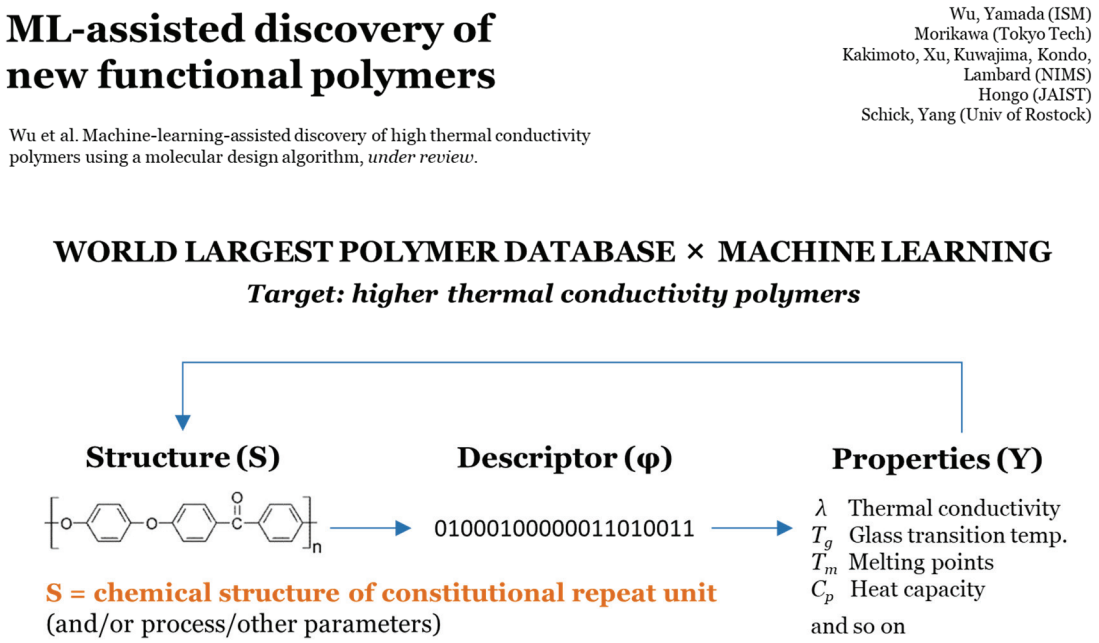

Fig 1. An illustrative example of machine-learning-accelerated materials discovery that were demonstrated in the talk at the workshop.

風疹流行モデルの構成とインフルエンザへの応用に向けての課題

[1),†]斎藤正也　[2)]西浦博　[1)]樋口知之

[1)]統計数理研究所　[2)]北海道大学医学部

1. はじめに

本発表では流行を再現する感染症数理モデルと介入策への応用の事例として風疹を取りあげ，インフルエンザへ同手法を展開する上で課題となる感染者総数の推定問題を取り上げる．

日本は風疹に対する高い集団免疫を獲得しているが，ワクチン接種政策の経緯を反映して 30～49 歳の男性を中心に陽性率が低い集団[1]が残存する．実際，2012～2013 年および今年 2018 年にアウトブレイクが発生した．妊婦が感染すると胎児に先天性風疹症候群をもたらす可能性があり，ワクチン追加接種による完全な集団免疫達成が課題となっている．

本研究では確率的メタポピュレーションモデルによって 2012～13 年の流行を再現するモデルを構成し，その応用例としてワクチンの傾斜配分の効果を評価する．ここでは風しん感染者数は悉皆調査の対象になっていることを利用した．多数の感染者が発生するインフルエンザの場合，定点医療機関による報告数から正味の感染者数を予め復元しておくことが必要となる．復元には橋本ら[6] による乗数法と呼ばれる方法が用いられてきた．これは医療機関も等確率 p_0 で定点として選ばれるとの仮定し，m 人の報告があれば総感染者は m/p_0 であろうと推定する.我々は流行期間の前後の血清陽性割合の変化が感染者割合を近似すると仮定してこれを推定する． 2009 年のパンデミック流行時にはワクチンの効果はほぼ無視できるためこの推定が可能になる．

2. 風疹の空間伝播モデル

都道府県ごとの感染動向の記述には Colizza [3]の定式化を一部変更したメタポピュレーションモデルを用いる．これは t 週目の i 県 (i=1,･･･,47)での感受性者数を $S_{i,t}$，感染者数を $I_{i,t}$，回復者数を $R_{i,t}$ としたとき，1 週間の間のこれらの数量の変化を以下の確率過程に従うとする

$$S_{i,t} = S_{i,t-1} - \Delta(S_{i,t} \to I_{i,t}) - \sum_{j\neq i}\Delta(S_{i,t} \to S_{j,t}) \quad \sum_{j\neq i}\Delta(S_{j,t} \to S_{i,t})$$
$$I_{i,t} = I_{i,t-1} + \Delta(S_{i,t} \to I_{i,t}) - \Delta(I_{i,t} \to R_{i,t}) - \sum_{j\neq i}\Delta(I_{i,t} \to I_{j,t}) \quad \sum_{j\neq i}\Delta(I_{j,t} \to I_{i,t})$$
$$R_{i,t} = R_{i,t-1} - \Delta(I_{i,t} \to R_{i,t}) - \sum_{j\neq i}\Delta(R_{i,t} \to R_{j,t}) + \sum_{j\neq i}\Delta(R_{j,t} \to R_{i,t})$$

ここで，$\Delta(\bullet \to \bullet)$ は二項分布に従う確率変数で

$$\Delta(S_{i,t} \to I_{i,t}) \sim \mathrm{Binom}\,(S_{i,t}, 1-\exp(-\beta I_{i,t}\Delta t / N_{i,t}))$$
$$\Delta(I_{i,t} \to R_{i,t}) \sim \mathrm{Binom}\,(I_{i,t}, 1-\exp(-\gamma\Delta t))$$
$$\Delta(X_{i,t} \to X_{j,t}) \sim \mathrm{Binom}\,(X_i, 1-\exp(-\rho_{ij}\Delta t)), \quad X \in \{S, I, R\}.$$

パラメータは平均感染期間を $1/\gamma = 7$ 日，i 県から j 県への移動率 $\rho_{i,j}$ を国土交通省の都道府県間流動表[4]に基づきそれぞれ設定した．風疹の感染力には季節変動があるので，それを表す β を区分的に一定と仮定してデータから推定する．i 県での t 週目の実際の報告数 $J_{i,t}$ とモデルによる予測とがポアソン分布 $J_{i,t} \sim \mathrm{Pois}\,(\bullet\,|1+\Delta(S_{i,t} \to I_{i,t}))$ で結ばれるとの仮定を表す尤度を粒子フィルタにより近似評価し，上位層で MCMC の反復を行うことで推定を実行する．

3. 推定の結果とワクチン傾斜配分の評価

時変の感染伝達係数 $\beta = \beta_t$ は $\beta_t.\gamma^{-1}.(S_{i,t}/N) \fallingdotseq \beta_t.\gamma^{-1}.(S_0/N) =: \mathbf{R}_t$ で定義される (近似) 実効再生産数と関係づけられる． $\mathbf{R}_t$ > 1 (< 1)であれば流行は拡大 (縮小) する．流行規模が小さい 2012 年の $\mathbf{R}_t$ は巨大な信頼区間を伴う不安定な推定値であっ

† saitohm@ism.ac.jp

たが，流行が大きい 2013 年の推定値は $R_t > 1$ または $R_t < 1$ を識別できる十分小さい信頼区間を伴うものであった．さらに推定値を用いてシミュレーションを行い，大都市での継続的報告(ピーク時で 100 件/週程度) と小都市での散発的な報告(数件または 0 件/週)とが同時に再現できることを確認した．詳細は論文[5]を参照されたい．

つぎに人口規模について上位 K 県 ($1 \leqq K \leqq 47$) だけに県人口に応じて所定のワクチンを配布する政策を流行に先立って実施したとき，流行による累積感染者数がどの程度減少するかを調べる．ここでは，日本の人口の 1,3,5%にあたる感受性者が接種によって免疫を獲得した状況を想定してシミュレーションを行った．その結果， 6〜7 の主要都道府県に配分するのが最も感染者数を低減することがわかった．他方，初期感染者の導入地による不確定性の方が低減幅よりも大きく，傾斜配分政策の積極的な支持は難しいとも考えられる．

4. インフルエンザへの応用と感染者数の推定

血清陽性率の変化をそのまま感染率と解釈できるのは 2009 年パンデミック流行時に限られる．しかし，あえてワクチン接種による陽性率上昇を無視して 2010/11，11/12 シーズンにも同じ方法を適用して年齢階級ごとの感染率と定点観測の報告率とを推定した．それによると，2009 年のデータに基づく報告率と不確かさの範囲で一致しており，パンデミックに基づく報告率を他シーズンに適用できるものと考えられる．結果的にワクチンによる(集団に対する正味の)陽転率上昇は限定的であると考えられる．報告率一定のもとでいくつかのシーズンの累積感染者数の推定値をレセプトまたは乗数法によるものと比較した．臨床的感染者数の 3-5 倍の不顕性感染者が存在すると解釈できる．この数値は先行研究によるものよりも過大であり，報告率一定という仮定には検証の余地があると考えられる．

謝辞　本研究は JST CREST 採択課題「大規模生物情報を活用したパンデミックの予兆、予測と流行対策策定」および JSPS 科研費 JP18K11541 の支援のもと実施されたものである．

参考文献

1. Kinoshita R, Nishiura H. Assessing herd immunity against rubella in Japan: a retrospective seroepidemiological analysis of age-dependent transmission dynamics. BMJ Open 6: 1–7, 2016.
2. Nishiura H, Kinoshita R, Miyamatsu Y, Mizumoto K. Investigating the immunizing effect of the rubella epidemic in japan, 2012-14. International Journal of Infectious Diseases 38: 16–18, 2015.
3. Colizza V, Barrat A, Barthe'lemy M, Vespignani A. The modeling of global epidemics: Stochastic dynamics and predictability. Bulletin of Mathematical Biology 68: 473–481, 2006.
4. 国土交通省. 旅客地域流動統計(平成 23 年度)・府県相互間旅客輸送人員表. 2011.
5. Saito M, Nishiura H, Higuchi T. Reconstructing the transmission dynamics of rubella in Japan, 2012-2013. PLoS ONE 13(10): e0205889. https://doi.org/10.1371/journal.pone.0205889, 2018.
6. Hashimoto S, Murakami Y, Taniguchi K, et al. Annual incidence rate of infectious diseases estimated from sentinel surveillance data. Japan. J Epidemiol 2003; 13(3): 136-41.
7. Nakamura Y, Sugawara T, Kawanohara H, Ohkusa Y, Kamei M, Oishi K. Evaluation of estimated number of influenza patients from national sentinel surveillance using the national database of electronic medical claims. Jpn J Infect Dis 2015; 68(1): 27-9.
8. Kawado, M., Hashimoto, S., Ohta, A., Oba, M.S., Taniguchi, K., Sunagawa, T., Matsui, T., Nagai, N., Murakami Y. Improvement of Influenza Incidence Estimation Using Auxiliary Information in Sentinel Surveillance in Japan. The Open Infectious Diseases Journal, 10, 29-36, 2018

Spatial Joint Species Distribution Modeling with Dirichlet Processes.

Shinichiro Shirota
Department of Biostatistics, University of California, Los Angeles

1 Introduction

Species distribution models usually attempt to explain presence-absence or abundance of a species at a site in terms of the environmental features (so-called abiotic features) present at the site. Historically, such models have considered species individually. However, it is well-established that species interact to influence presence-absence and abundance (envisioned as biotic factors). As a result, there has been substantial recent interest in joint species distribution models with various types of response, e.g., presence-absence, continuous and ordinal data. Such models incorporate dependence between species response as a surrogate for interaction.

The challenge we address here is how to accommodate such modeling in the context of a large number of species (e.g., order 10^2) across sites numbering on the order of 10^2 or 10^3 when, in practice, only a few species are found at any observed site. Again, there is some recent literature to address this; we adopt a dimension reduction approach. The novel wrinkle we add here is spatial dependence. That is, we have a collection of sites over a relatively small spatial region so it is anticipated that species distribution at a given site would be similar to that at a nearby site. Specifically, we handle dimension reduction through Dirichlet processes, enabling clustering of species, joined with spatial dependence across sites through Gaussian processes.

2 Model

Let $\mathcal{D} \subset \mathbb{R}^2$ be a bounded study region, $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be a set of plot locations where $\boldsymbol{s}_i \in \mathcal{D}$ for $i = 1, \ldots, n$, and $\boldsymbol{U}_i := \boldsymbol{U}(\boldsymbol{s}_i) \in \mathbb{R}^S$ be an $S \times 1$ latent vector of continuous variables at location $\boldsymbol{s}_i$. Under independence for the locations, the model for $\boldsymbol{U}_i$ is specified as

$$\boldsymbol{U}_i = \mathbf{B}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}_S(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{for} \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{B}$ is an $S \times p$ coefficient matrix, $\boldsymbol{x}_i$ is a $p \times 1$ covariate vector at location $\boldsymbol{s}_i$ and $\boldsymbol{\Sigma}$ is a $S \times S$ covariance matrix for species. This model has $\mathcal{O}(S^2)$ parameters, $S(S+1)/2$ parameters from $\boldsymbol{\Sigma}$ and pS parameters from $\mathbf{B}$.

Taylor-Rodríguez et al. (2017) approximate $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \mathbf{I}_S$ and replace the above model with

$$\boldsymbol{U}_i = \mathbf{B}\boldsymbol{x}_i + \boldsymbol{\Lambda}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_S(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_S), \quad \text{for} \quad i = 1, \ldots, n \tag{2}$$

where the random vectors $\boldsymbol{w}_i$ are i.i.d. with $\boldsymbol{w}_i \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ and $\boldsymbol{\Lambda}$ is an $S \times r$ matrix with $r \ll S$. Now, $\boldsymbol{\Sigma}^*$ has only $Sr + 1$ parameters, the estimation problem of $\mathcal{O}(S^2)$ parameters is reduced to that of $\mathcal{O}(S)$ parameters. We refer to this specification as the dimension reduced nonspatial model.

Although $\mathbf{\Lambda}\mathbf{\Lambda}^T$ has rank r, including the nugget variance $\sigma_\epsilon^2\mathbf{I}$ ensures that $\mathbf{\Sigma}^*$ is nonsingular. The further approximation which Taylor-Rodríguez et al. (2017) proposed is to sample the rows of $\mathbf{\Lambda}$ from a Dirichlet process mixture (DPM) using a stick-breaking representation (Sethuraman, 1994). The stick-breaking representation is attractive within a Gibbs sampling setting (see, e.g., Escobar, 1994; Escobar and West, 1995) due to a Pólya urn scheme representation which enables straightforward simulation from needed full conditional distributions.

To provide the hierarchical formulation for this model, let $\mathbf{Z} = [\boldsymbol{Z}_1 : \ldots : \boldsymbol{Z}_N]^T$ (with $\boldsymbol{Z}_j \sim H$) denote the $N \times r$ matrix whose rows make up all potential atoms. In this setup, we need a vector of grouping labels $\boldsymbol{k} = (k_1, \ldots, k_S)$ $(1 \leq k_l \leq N)$ so that the l-th row of $\mathbf{\Lambda}$ is equal to $\boldsymbol{Z}_{k_l}$. We note that $\mathbf{\Lambda}$ can be represented by $\mathbf{\Lambda} = \mathbf{Q}(\boldsymbol{k})\mathbf{Z}$ where $\mathbf{Q}(\boldsymbol{k}) = [\boldsymbol{e}_{k_1} : \ldots : \boldsymbol{e}_{k_S}]^T$ is $S \times N$ with $\boldsymbol{e}_{k_l}$ denoting the N-dimensional vector with a 1 in position k_l and 0's elsewhere. Letting $\mathbf{W} = [\boldsymbol{w}_1 : \ldots : \boldsymbol{w}_n]^T$ be the $n \times r$ spatial factor matrix, our approximate model is

$$
\begin{aligned}
\boldsymbol{U}_i|\boldsymbol{k}, \mathbf{Z}, \boldsymbol{w}_i, \mathbf{B}, \sigma_\epsilon^2 &\sim \mathcal{N}_S(\mathbf{B}\boldsymbol{x}_i + \mathbf{Q}(\boldsymbol{k})\mathbf{Z}\boldsymbol{w}_i, \sigma_\epsilon^2\mathbf{I}_S), \quad \text{for} \quad i = 1, \ldots, n, \\
\mathbf{W}^{(h)} &\sim \mathcal{N}_n(\mathbf{0}, \mathbf{C}_\phi), \quad \text{for} \quad h = 1, \ldots, r, \\
k_l|\boldsymbol{p} &\sim \sum_{j=1}^{N} p_j \delta_j(k_l), \quad \text{for} \quad l = 1, \ldots, S, \\
\boldsymbol{Z}_j|\mathbf{D}_{\boldsymbol{Z}} &\sim \mathcal{N}_r(\mathbf{0}, \mathbf{D}_{\boldsymbol{Z}}), \quad \text{for} \quad j = 1, \ldots, N, \\
Z_{1,h} &> 0, \quad \text{for} \quad h = 1, \ldots, r, \\
\boldsymbol{p} &\sim \mathcal{GD}_N(\boldsymbol{a}, \boldsymbol{b}), \\
\mathbf{D}_{\boldsymbol{Z}} &\sim \mathcal{IW}(2 + r - 1, 4\text{diag}(1/\eta_1, \ldots, 1/\eta_r)), \\
\eta_h &\sim \mathcal{IG}(1/2, 1/10^4), \quad \text{for} \quad h = 1, \ldots, r,
\end{aligned}
\tag{3}
$$

where $\mathcal{GD}_N$ is an N dimensional generalized Dirichlet distribution, $\mathbf{W}^{(h)} = (w_1^{(h)}, \ldots, w_n^{(h)})^T$ is the h-th column of $\mathbf{W}$ ($n \times 1$ vector) and is distributed as an n-variate normal vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}_\phi = [\exp(-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|)]_{i,i'=1,\ldots,n}$, i.e., a realization of a Gaussian process (GP) with exponential covariance function at the sites in $\mathcal{S}$. We refer to the above modeling specification as the dimension reduced spatial model.

References

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*, 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Taylor-Rodríguez, D., K. Kaufeld, E. M. Schliep, J. S. Clark, and A. E. Gelfand (2017). Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis 12*, 939–967.

全ゲノム配列情報を用いた疾患発症予測に向けて

理化学研究所 革新知能統合研究センター 植木優夫

近年のデータサイエンス技術の目覚ましい発展により、様々な領域でデータが大量に取得され、実際の社会問題への活用が進んでいる。医学研究における一般集団コホートやバイオバンク（UK バイオバンク（Sudlow et al. 2015）など）では、数万人以上の規模の参加者について、性別、年齢、身長、体重などの基本的な情報に加えて、職業や運動習慣、飲酒、喫煙などの生活習慣、家庭環境に至るまで様々な背景情報がデータとして取得されている。そこには、ゲノムワイド SNP アレイ（一塩基多型）や全ゲノムシーケンスデータなどの全ゲノム情報や MRI 画像なども含まれており、データの規模は年々増加の一途をたどっている。このような大量のデータには、疾患発症に関与する新規バイオマーカーや未知の遺伝子など重要な情報が含まれている可能性がある。これらの情報を用いることで、従来の疾患発症予測モデルの精度が向上すれば非常に有益である。ゲノムデータを実際の疾患発症予測に応用する試みはすでにはじまっているが、多くの複合疾患において、ゲノム情報による疾患予測精度の上乗せが十分でないことが報告されており、予測精度向上は重要な課題である。

SNP アレイは、全ゲノム配列を相関係数（r^2）によってカバーする代表的なバリアントから成るマーカーセットであり（タグ SNP；Carlson et al. 2004）、配列情報の一部しか含まれておらず、マーカー数が 60 万～100 万個程度とデータサイズが削減されている。例えば、もし仮に相関係数 90％の関係にある 2 つのバリアントがあれば、一方のみが SNP アレイに含まれる、というように冗長性が取り除かれたデザインとなっている。そのため、真に疾患発症に関与するバリアントが必ずしも SNP アレイに含まれているとは限らず、SNP アレイデータから構築された予測モデルの予測精度が不十分となるおそれがある。一方で、全ゲノムシーケンスデータには網羅的なバリアント情報が含まれるため、重要なバリアントが見落とされる可能性は小さくなり、より高い予測性能が得られる可能性がある。しかし、全ゲノムシーケンスデータの取得は現在のところ高コストであり、安価な SNP アレイが広く用いられている。

SNP アレイはゲノム情報としては不完全であるが、物理的な近傍にある SNP 同士が相関することを利用し、SNP アレイデータから全ゲノム配列情報を統計的に復元する全ゲノムインピュテーション法（Marchini & Howie 2010）が広く用いられている。様々な手法が存在するが、代表的なものとして、IMPUTE2（Howie et al. 2009）、Minimac3（Das et al. 2016）が挙げられる。その原理としては、ヒト集団を代表するハプロタイプデータを参照し、ゲノム配列を復元したいサンプルの SNP アレイデータから、共通するバリアントについて類似度の高い参照ハプロタイプを復元する、というものである。参照ハプロタイプとしては、国際 1000 人ゲノムプロジェクトのデータがよく用いられる（1000 Genomes Project Consortium 2010）。数千人規模の複数民族集団についての全ゲノムシーケンスが行われており、これを参照ハプロタイプとしている。参照ハプロタイプにはおよそ 8 千万個のバリアントが含まれるデータである。さらに最近、複数のプロジェクトから成る全ゲノムシーケンスデータを集約した巨大なデータを参照ハプロタイプに使用するプロジェクトがはじまっている（The Haplotype Reference Consortium 2016）。ところで、ヒトはディプロイド（両親からそれぞれ継承したハプロタイプ 2 つをもつ）であって、実際にはハプロタイプ自体は観察されず、2 つのハプロタイプが混ざった状態でのみ観察される。そのため、まず最初に SNP データから SHAPEIT2（Delaneau et al. 2012）などを用いて統計的にハプロタイプを構築し、得られたハプロタイプからインピュテーションを行うというプレフェージング法（ Howie et al. 2012 ）が標準的な方法となっている。プレフェージングとインピュテーションは統計的な推定であり、正確な推定が重要となる。発表では、全ゲノムシーケンスデータとインピュテーションの方法について述べ、さらにそれらを用いた予測モデリングについて俯瞰した。

参考文献

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-73.
Carlson CS et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106-20.
Das S et al. (2016) Next-generation genotype imputation service and methods. Nature Genetics 48:1284-7.
Delaneau O et al. (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9:179.
Howie B et al. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529.
Howie B et al. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955-9.
Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499-511.
Sudlow C et al. (2015) UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12: e1001779.
The Haplotype Reference Consortium (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48:1279-83.

高次元 Microarray データによる「癌の遺伝子解析」
―なぜ 1970 年から研究され成功しなかったのか？―

成蹊大学　名誉教授　新村秀一

Big データがマスコミでも取り上げられ統計にとってフォローの風が吹いているが、2000 年以前から高次元データ解析が統計や機械学習の格好のテーマとして注目され未だに芳しい成果が出ていない。これらの研究は、1970 年ごろから恐らくハーバード大学医学部の Golub 教授らが先導した Microarray データから、特定の癌を classify する研究が行われ研究に用いられた公開データを用いている。筆者も 2015 年 10 月 26 日に富山市で行われた科研費シンポジュームで米国の主要 6 研究が、Golub らが 1999 年にサイエンスに論文を発表し、2004 年に Tien が〝The New England Journal of Medicine”等に発表した 6 編の論文に用いられているデータが CSV 形式で公開されていることを知った。

1997 年から判別分析の 4 つの深刻な問題を解決するため誤分類数最小(Minimum Number of Misclassifications,**MNM**)基準に基づく IP-OLDF や Revised IP-OLDF(RIP)を開発し、IP-OLDF の説明に用いた概念図で、LDF の判別係数と NM の関係を初めて解明し**問題 1** を解決した。すなわち、RIP 以外の LDF の NM は正しくない可能性があることを示した。そして線形分離可能なデータ (LSD) の判別を統計では無視してきたが、H-SVM と RIP だけが理論的で可能であることが分かった。スイス銀行紙幣データ、日本車 44 種のデータ、試験の合否判定データで LSD 判別研究を行った。しかし、これ等の研究は注目されなかった。このことは Google Schoalar で現在 1,062 件の Citation があるが、2013 年以降が 985 件でありそれ以前の研究がほどんと注目されていないことから明らかである。

しかし 10 月 28 日から 6 種類の microarray データを RIP で判別すると MNM=0 であることが分かった (**Fact3**)。さらに Matryoshka Feature Selection Method(Method2)で簡単に多数の Small Matryoshka(SM)と呼ぶ症例数 n 個以下の遺伝子をもつ MNM=0 の小標本と MNM が 1 以上の雑音に簡単に分かれた(**Fact4**)。そこで、Springer から出版予定の”New Theory of Discriminant Analysis After R. Fisher(書籍 1)”に急遽この結果を含め 2016 年 5 月に出版した。本来統計研究としてここで終了すればよかったが、SM は小標本であるので統計で簡単に分析できる。H-SVM と RIP で判別すると、SV で簡単に癌と健常症例を判別スコアの 60%以下の狭い範囲で判別できることが **RatioSV** という統計量で分かった。医学専門家でないが、これは**癌の悪性度指標**を表すのでないかと考えた。また青嶋・矢田は多くの統計家のアプローチと異なり、高次元 PCA で群が 2 つの異なる球上に布置していることを見つけた。これを筆者の研究では 2 群が異なる狭い範囲に布置していることを PCA とクラスター分析で示した。そして”From Cancer Gene Analysis to Cancer Diagnosis”を 2017 年に Amazon から出版した。その後日本の大学、大病院、検査機器メーカー等に結果を医学的に検証してもらえないかとアプローチした。その中で唯一説明の機会を得たのは、1997 年に大学卒業後に大阪府立成人病センターで癌の疫学データの共同研究を行った鈴木先生の紹介で会った加藤先生である。開口一番「NIH がこの種の研究は意味がないという報告を出し、世界中の研究者が続々退場している。今からやっても相手にされないので無駄である」と諭された。この研究の最終目的は、医学的に役立つ癌の遺伝子の特定である。当の医学研究で意味がないと結論されたのに、今もって統計や機械学習の研究者から”Feature Selection とか Filtering…”というタイトルの論文が Research Gate に Upload されている。彼らの多くは、NIH の決定を知らないのであろう。私がこの研究

を続けているのは、NIH の判断が間違っていると考えているからである。

本来はこの問題は癌と健常の 2 群、あるいは異なった癌の 2 群の高次元遺伝子データの 2 群判別が最も適している。しかし、LSD 判別は筆者しか行っていないため、統計的な判別分析は microarray データが LSD であるという重要な情報を認識できなかったため、医学研究者が 1970 年から研究し成果が出なかっただけである。H-SVM と RIP で判別すれば簡単に MNM=0 であることが分かる（**Fact3**）。しかしこれは、彼らが SOM などのクラスター分析で 2,000 個から 12,625 個の遺伝子に絞った。

このことは、iPS 細胞研究で山中伸弥先生が 22,000 といわれる人の遺伝子から 24 個の遺伝子に絞り込んだのと同じである。この 24 個から山中 4 遺伝子を見つけたが、私が提案した癌の基本遺伝子（BGS）である。この 4 遺伝子には C-Myc という発癌性の高い遺伝子が含まれるので、苦労して L-Myc に置き換えた。この事実は、私の研究で多数の SM や BGS があることと符合する。もう少し早くこの対比に気づいておれば、iPS 細胞研究を加速するのに貢献したと思う。

判別分析の研究が、LSD という簡単な判別もできないことに誰も気づかなかったのは、判別研究の集団的な間違いだと考える。これは推測であるが、2000 年以降の医学研究では一切判別関数が用いられず、クラスター分析を重用しているので classify という言葉が用いられている。一部 SVM を取り上げられているが成果は示されていない。この事実を判別分析の研究者は、真摯に反省すべきであろう。一方医学研究では、高橋博士が 24 個総てを細胞に入れて万能細胞の塊を作った。当初山中先生は、彼が工学部出身で生物学的常識に無知なため間違ったと考えられたようだ。細胞培養は時間がかかる実験のため、1 個の遺伝子を培養し当たりを付けるのがこの分野に共通する間違いだと考えている。これまでの医学研究で細胞や遺伝子が癌によって顕微鏡で変異を観察する。そしてそれらの中から癌化を起こすものを特定し 1 個の癌遺伝子を見つける。このような oncogene が 50 個から 100 個見つかっているが、これ等をすべて用いて判別しても恐らく MNM=0 にならないであろう。一方、癌と健常の平均値を t 検定や Welche 検定し、正と検定されるものを癌遺伝子の候補とする研究も多い。これらの研究は全て 1 変数だけのアプローチで、多変量的な考察にかけている。SM の多くの t 検定では、正で棄却されるもの、負で棄却されるものの他、棄却されない 0 の遺伝子が含まれる。すなわち、MNM=0 になるためには、これ等の 3 種の遺伝子の適切な組み合わせが必要である。2017 年に京大医学部の和田名誉教授が「丸山ワクチン患者家族の会」で講演されたことを知って、不躾であったが大量の論文等などを送った。暫くして血液から代表的な oncogenes を使った癌と健常のリスク判定を行っている GeneScience を紹介していただいた。癌患者に各種治療などを受けている方、健常に隠れ癌患者が多く含まれていて 2 割ほどの誤分類確率である。恐らく管理された症例に限定しても、oncogenes だけで MNM=0 にならないと考えている。

私の主張は、RG の週ごとに集計結果で、10 月 21 日に 3,992 件の Read 数で日本の Top であるとの報告が来た。Citation が 1,337 件で Google と食い違いがあるが、2015 年に退官後の研究費の自己負担を削減し世界に情報発信しようと思い立ち RG の利用を始めた。結果としては、それまで見向きもされなかった研究が「癌の遺伝子解析」に成功し注目された結果である。しかし、米国の 6 研究グループに私の遺伝子診断の検証を呼び掛けているが返事がない。また RG に専門医の登録が少ないようで彼らが研究を見ている例も少ない。そこで、NIH の決断が間違っているということを示し、改めて専門医に協力を依頼するため、”High-dimensional Microarrays Data Analysis -Cancer Gene Diagnosis and Malignancy Indexes by Microarray-“を出版し、特に米国の癌の専門医できれば在米の日本人医師に共同研究を呼び掛ける予定である。

高次元データのための直接・間接効果を考慮した関数判別モデル

静岡大学大学院　総合科学技術研究科　情報学専攻
荒木由布子

1　はじめに

時間や空間の経過に伴い変動する観測値・測定値を関数として捉え, その集合から有効に情報を抽出する関数データ解析は (Ramsay and Silverman, 2005)，近年の複雑な構造を有する高次元データ，例えば，脳の 3 次元 Structural MRI, 交通量データ，密に観測された小児の成長に伴う脳波データなどから，情報を効率的に抽出し，解析できる手法として注目されつつある（J. M. Chiou et al. 2012，Chen et al. 2018, Reiss and Ogden, 2010, Araki et al. 2009, 2013.）．本研究では，時間や空間の変動に伴い観測される変数を関数データとして捉え, その変数のどのタイミングや地点が目的変数と説明変数の媒介変数として機能しているかを発見するための直接的・間接的な影響を考慮した (Pearl, 黒木 2009) 判別モデルを構築する．このような問題は関数データ構造方程式モデルで考慮することができるが（Lindquist, M.A., 2012, Zhao et al. 2018), ここで考慮する関数データは実際には離散点で観測されさらに高次元データである場合を想定しているため, 関数化に工夫が必要となる．また，汎用性の高い安定した推定量を得るため，新たに開発したモデルの評価規準を導出して最適なモデルを選択する．

2　関数構造方程式モデル

ここでは, 関数媒介モデル (Lindquist 2012, Zhao et al. 2018) と関数回帰モデル（Araki et al. 2009, 2013) をもとに構築した関数構造方程式モデルによる直接・間接効果を考慮した判別について説明する．

群を表すグループ変数 G と媒介変数としての説明変数 X, 割り付け変数としての説明変数 Z に関して大きさ n のデータ $\{(g_\alpha, x_\alpha(t), z_\alpha); t \in [0,1], \alpha = 1, 2, \cdots, n\}$

が観測されたとする． $x_\alpha(t)$ は Karhunen-Loeve 展開により表される関数データ $x_\alpha(t) = \sum_{m=1}^{M} \lambda_{\alpha m}\phi_m(t)$ であり，ここで $\lambda_m, \phi_m(t), m = 1, \cdots, M$ は共分散関数 $Cov(X(s), X(t))$ の m 番目に大きい固有値とそれに対応する固有関数である．また $g_\alpha \in \{1, 2\}$ はグループ変数で, $g_\alpha = l$ は $x_\alpha(t)$ が第 l 群 $(l = 1, 2)$ に属することを表す. このとき, 関数構造方程式モデルは以下で定義される.

$$x_\alpha(t; z_\alpha) = \beta_{01}(t) + \beta_2(t)z_\alpha + \epsilon_\alpha(t), \tag{1}$$

$$\log\left\{\frac{Pr(g_\alpha = 1|x_\alpha, z_\alpha)}{Pr(g_\alpha = 0|x_\alpha, z_\alpha)}\right\} = \beta_{02}(t) + \beta_3 z_\alpha + \int_0^1 \beta_4(t)x_\alpha(t; z_\alpha)dt,\ t \in [0, 1]. \tag{2}$$

モデル式 (1) のパラメータ推定には最小二乗推定を, (2) 式には正則化最尤法を用い, 一般化ベイズ型モデル選択規準によりモデルの評価を行う．提案したモデルは, fMRI データや MRI 画像データなど, 高次元データを観測に含む分析に有用であると考えられる.

参考文献

[1] Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009). Functional logistic discrimination via regularized basis expansions. *Communication in Statistics, Theory and Methods*, 38, 2944-57.

[2] Araki, Y., Kawaguchi, A. and Yamashita, N. (2013). Regularized logistic discrimination with basis expansions for the early detection of Alzheimer's disease based on three-dimensional MRI data. *Advances in Data Analysis and Classification*, 7(1), 109-119.

[3] Chen, Y, Goldsmith, J and Ogden, T. (2018). Functional data analysis of Dynamic PET data, *J Am Stat Assoc.*, DOI: 10.1080/016214529.2018.1497495.

[4] Chiou, J. M. (2012). Dynamic functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics*, 6(4), 1588-1614.

[5] Lindquist, M. A., (2012). Functional causal mediation analysis with an application to brain connectivity. *J Am Stat Assoc.*, 107 (500), 1297-1309.

[6] Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer.

[7] Reiss, P. and Ogden, T. (2010). Functional generalized linear models with images as predictors, *Biometrics* 66, 61-69.

[8] Pearl, J (著）, 黒木 学 (翻訳). (2009). 統計的因果推論 – モデル · 推論 · 推測 –, 共立出版.

大規模なデータの分析における精度行列の縮小推定法とその特徴

中京大学 国際教養学部 永井 勇

情報化が進み各個体に対するデータが多く得られるようになり, それに伴いデータの構造が複雑化し大規模になっている. このように各個体に対するデータが大規模になると, 遺伝子データのように, 集めた個体数 n よりも各個体に対するデータ (つまり変数) p のほうが多い状況が頻繁に起きる. このような状況においては, 従来の統計学に基づいた予測などが困難なため, 新たな予測モデリングなどが必要となる. そこで本講演では, 後述する精度行列に着目し, そこでの縮小推定法を提案した. さらに, 縮小に用いたパラメータをいくつかのロス関数に基づいて最適化した値が陽に求まることを示し, その特徴についても触れた.

本講演ではデータを $\boldsymbol{Y}$ ($n \times p$ 既知行列) とし, $\boldsymbol{Y}$ の各行の分布に関する仮定や n と p の大小関係などの仮定は置かず, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Y})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ だけ仮定した. ここで, $\boldsymbol{\Sigma}$ は未知の $p \times p$ 正定値行列であり, $\mathrm{vec}(\cdot)$ は vec 作用素, $\otimes$ はクロネッカー積である (これらの定義は例えば Lütkepohl (1996) を参照). またこの仮定は多変量データの分析では広く置かれる仮定であり, これは各個体の真の分散共分散行列が未知の $\boldsymbol{\Sigma}$ で, 個体間は無相関であることを表している.

従来の予測モデリングに基づいて予測などを行う際 (例えば, 判別分析を行う際や多変量回帰モデルなどで変数選択を行う際など) には, $\boldsymbol{\Sigma}$ の推定量 $\boldsymbol{S}$ の逆行列 (精度行列) が必要な場合が多い. 分析法やモデルにより $\boldsymbol{S}$ の求め方が異なる場合があるが, $\boldsymbol{Y}$ だけを用いた場合 $\boldsymbol{S} = \boldsymbol{Y}'(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n'/n)\boldsymbol{Y}/(n-1)$ がよく用いられる (Schott (2017) など参照). ここで, $\mathbf{1}_n$ は全ての成分が 1 の n 次元ベクトルである. また, $\boldsymbol{S}$ の固有値 $d_1, \ldots, d_p$ は全て非負である. この推定量 $\boldsymbol{S}$ において, $\mathrm{rank}(\boldsymbol{S}) < p$ の場合, 精度行列 $\boldsymbol{S}^{-1}$ が存在しない. また, $\mathrm{rank}(\boldsymbol{S}) = p$ であっても $d_1, \ldots, d_p$ の内一つでも 0 に近い値の場合, $\boldsymbol{S}^{-1}$ が不安定になってしまう. そこで, 以下のような手法を用いて $\boldsymbol{S}^{-1}$ の代わりの行列を使うことが提案されている;

- $\boldsymbol{S}$ の Moore-Penrose 型一般逆行列 (Schott, 2017, Sec. 5) を用いる手法
- $\boldsymbol{S}$ の非対角成分を全て 0 にした行列の逆行列を用いる手法 (Srivastava, Katayama & Kano (2013) など)
- リッジ型の罰則を付けて精度行列の縮小推定をした行列を用いる手法 (Wang, Pan, Tong, & Zhu (2015))

本講演では, これらの手法の中でリッジ型の罰則を付ける手法に着目した.

リッジ型の罰則を付ける手法は, $\boldsymbol{S}^{-1}$ の代わりに, 二つの正のパラメータ λ と α を導入した $\lambda(\boldsymbol{S} + \alpha\boldsymbol{I}_p)^{-1}$ を用いる手法である. この手法は, $\boldsymbol{S}$ の固有値 d_i に対し α (> 0) を加えることで, ある j で $d_j = 0$ であっても $d_j + \alpha > 0$ となり, $(\boldsymbol{S} + \alpha\boldsymbol{I}_p)^{-1}$ が存在するようにしている手法である. また, いくつかの d_i が小さな値であっても, $\alpha > 0$ より $(\boldsymbol{S} + \alpha\boldsymbol{I}_p)^{-1}$ は $\boldsymbol{S}^{-1}$ より安定する. この推定量は, $\boldsymbol{S}$ に $\alpha\boldsymbol{I}_p$ を加えた行列の逆行列を用いることで縮小推定となり, λ で全体のスケールを調整している. しかし, この手法には以下の問題がある;

① $\boldsymbol{S}$ の p 個の固有値 $d_1, \ldots, d_p$ に対して, 全て一様に同じパラメータ α で調整している
② λ と α の同時最適化のための反復計算が必要となる

特に①の問題は, 実際のデータでよく起こるように, $\boldsymbol{S}$ の固有値の中に十分大きな値と非常に小さな値が混在している場合に大きな問題となる. 具体的には, 全ての固有値に対して一様に α を加えているため, 非常に小さな固有値に対しては $\boldsymbol{S}^{-1}$ の不安定さを回避できるほど十分な調整ができていない一方で, 十分大きな固有値にも同時に余分な調整がされてるという問題である. そこで, 本講演では上述した二点の問題を回避する手法を提案した.

そこで本講演では, Hoerl and Kennard (1970), Yanagihara, Nagai and Satoh (2009) などで用いられている一般化リッジ回帰による推定法を, このリッジ型の罰則付推定に適応することを考えた. 一般化リッジ回帰による推定法を適応すると次の形となった;

$$\hat{\boldsymbol{S}}^{-1}(\lambda,\boldsymbol{\theta}) \overset{\text{def.}}{=} \lambda(\boldsymbol{S}+\boldsymbol{Q}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{Q}')^{-1},$$

ここで $\boldsymbol{Q}$ は $\boldsymbol{Q}'\boldsymbol{S}\boldsymbol{Q}=\mathrm{diag}(d_1,\ldots,d_p)$ ($d_1,\ldots,d_p$ を対角に並べた対角行列) となる直交行列, $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)$ ($\theta_i\geq 0$, $i=1,\ldots,p$) であり, λ はリッジ型の罰則付推定と同じく正のパラメータである. 次に, この推定量における $(p+1)$ 個のパラメータの最適化を考えた.

本講演ではこれらのパラメータの最適化のために, 例えばリッジ型の罰則を付けた推定量 (Wang, Pan, Tong & Zhu, 2015) で用いられている次のロス関数を考えた;

$$L^{[2]}(\lambda,\boldsymbol{\theta}) \overset{\text{def.}}{=} \mathrm{tr}\left\{\left(\hat{\boldsymbol{S}}^{-1}(\lambda,\boldsymbol{\theta})\boldsymbol{\Sigma}-\boldsymbol{I}_p\right)^2\right\}.$$

他にも Stein's ロス (James & Stein, 1961) なども考えた. 本講演では, これらのロス関数を最小にする $\boldsymbol{\theta}$ が陽に求まることを示した. 例えば $L^{[2]}(\lambda,\boldsymbol{\theta})$ を最小にする $\boldsymbol{\theta}$ の第 i 成分は,

$$\hat{\theta}_i^{[2]}(\lambda,\boldsymbol{\Sigma})=\begin{cases} 0 & (\text{if } \dot{\varphi}_i^{[2]}(\lambda)>0) \\ \infty & (\text{if } \dot{\varphi}_i^{[2]}(\lambda)\leq -\lambda(\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q})_{ii}^2) \\ \dfrac{-d_i\times\dot{\varphi}_i^{[2]}(\lambda)}{\dot{\varphi}_i^{[2]}(\lambda)+\lambda(\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q})_{ii}^2} & (\text{上記以外}) \end{cases},$$

となる. つまり, λ と $\boldsymbol{\Sigma}$ を用いて陽な形で得られる. ここで $(\boldsymbol{A})_{ij}$ が行列 $\boldsymbol{A}$ の第 (i,j) 成分を表し, $\dot{\varphi}_i^{[2]}(\lambda)=d_i(\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q})_{ii}-\lambda(\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q})_{ii}^2-\sum_{j>i}^{p}\dfrac{\lambda d_i(\boldsymbol{Q}'\boldsymbol{\Sigma}\boldsymbol{Q})_{ij}^2}{d_j+\theta_j}$, ただし $\sum_{j>p}^{p}(\cdot)=0$ である.

この最適化したパラメータは, d_i が 0 や小さすぎる場合は $\hat{\theta}_i^{[2]}(\lambda,\boldsymbol{\Sigma})=\infty$ となり, d_i がある程度大きい場合は $\hat{\theta}_i^{[2]}(\lambda,\boldsymbol{\Sigma})=0$ となり, その他の場合は d_i の大きさに応じた値となっている. つまり, 各固有値 d_i の大きさに対応して各パラメータ $\hat{\theta}_i^{[2]}(\lambda,\boldsymbol{\Sigma})$ の大きさが柔軟に変化している. この最適なパラメータは, $\dot{\varphi}_i^{[2]}(\lambda)$ の形より, $\hat{\theta}_p^{[2]}(\lambda,\boldsymbol{\Sigma})$ が陽に得られ, $\hat{\theta}_i^{[2]}(\lambda,\boldsymbol{\Sigma})$ を用いて $\hat{\theta}_j^{[2]}(\lambda,\boldsymbol{\Sigma})$ $(j<i)$ が求まる形で, p 番目から順番に最適なパラメータが得られる.

本講演では λ の最適化などのために, CV (交差検証) 法を用いた. 他のロス関数や各ロス関数に基づいた最適なパラメータや数値実験による比較などについては講演で報告した.

引用文献:

[1] Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

[2] James, W. & Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, **1**, 361–379.

[3] Lütkepohl, H. (1996). *Handbook of Matrices*, John Wiley & Sons.

[4] Schott, J. R. (2017). *Matrix Analysis for Statistics* (Third Edition).

[5] Srivastava, M. S., Katayama, S. & Kano, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.*, **114**, 349–358.

[6] Yanagihara, H., Nagai, I. & Satoh, K. (2009). A bias-corrected C_p criterion for optimizing ridge parameters in multivariate generalized ridge regression. *Jpn. J. Appl. Stat.*, **38**, 151–172 (in Japanese).

[7] Wang, C., Pan, G., Tong, T., & Zhu, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica*, **25**, 993–1008.

マルコフ連鎖による大学合格者歩留まり率のモデル化と予測

佐野 崇、山本 真基
成蹊大学 理工学部 情報科学科

2018 年 11 月 9 日

概要

大学合格者数に対して実際に入学する人数の割合を歩留まり率という。歩留まり率を推定し、できるかぎり定員に近い人数を入学させることは、多くの大学において重要な課題である。本研究では、学生の志望校の変遷にマルコフ性を仮定することで、マルコフ連鎖の定常状態として、各大学への入学者数が決まるというモデルを提案する。このモデルを仮定すると、遷移行列を推定できれば、各大学への入学者の割合を一意に求めることが可能になる。我々は、遷移行列の要素を、大学の偏差値や定員などの公開データで推定する方法を提案し、定常状態が実際の入学者の分布を定性的に再現することを示した。さらに、この結果を用いて、望ましい入学者数のために設定すべき合格者数の考察を行った。

あらかじめ決められた定員にできるだけ近い入学者をとることは、多くの大学において重要な課題である。そのためには、合格者のうち何割が実際に入学するかを推定し、入学者数が望ましい人数になるよう合格者数を調整する必要がある。合格者のうち入学する人数の割合を、製造業の言葉を借用して、歩留まり率と言う。あらかじめ歩留まり率がわかっており、かつ歩留まり率が合格者数に依存しないのならば、(合格者数)×(歩留まり率)=(定員) を満たすように合格者数を設定すれば良い。

歩留まり率には多くの要素が影響するため、予測は一般に困難である。例えば、他の大学の合格率は大きく影響するはずである。特に、互いに併願の多い 2 つの大学に対しては、一方の合格率が下がれば、もう一方の歩留まり率は上がると予測できる。このように、大学間の相関を把握することは、歩留まり率の推定に役立つと考えられる。

本研究では、学生の入学する大学の変遷がマルコフ過程であると仮定したモデルを提案する。すなわち、ある 1 人の学生が入学する大学は、短い期間には 1 つに決まっているが、次の期間に移る時にある確率で別の大学に遷移する。この遷移がマルコフ的であると仮定する。すると、大学間の相関は、マルコフ過程の遷移行列として表現できる。直感的には、上位の大学から下位の大学への遷移確率は大きくなり、逆の遷移確率は小さくなると予想できる。遷移行列の規約性・非周期性を仮定すると、マルコフ過程の定常分布は一意に求めることができ、これが各大学への入学者の割合に対応する。また、大学の難易度は遷移行列の対角成分に対応している。

遷移行列の要素を直接求めることは難しいが、本稿では予備校の発表する大学の偏差値や、定員等を用いて遷移行列を構築する方法を提案した。このように作られた遷移行列によって、数値実験の結果、実際の入学者の分布をある程度再現することが確かめられた。

本来行いたいことは、入学者数の分布予測だけではなく、それに基づいた歩留まり率の推定と適切な合格者

数の設定である。本モデルを用いることで、次のような合格者の設定法が可能になると考えられる。まず、各年度において遷移行列を推定する。そうすることで、実際に設定した合格者数と、推定された入学者数の相関を得ることができる。この相関が各年度で大きく変化しないと仮定すると、新しい年度の合格者数と入学者数の関係が推定できる。この関係から、望ましい入学者数が得られるよう合格者数を決めるという戦略が取れる。

本稿で扱ったモデルは非常に単純化されており、改善すべき点は多くある。遷移行列の構築には、予備校などから併願に関するデータを提供してもらえば、さらに現実に即したものになるだろう。また、受験はするが大学に入らない集団や、推薦入試などで一般受験を受けずに合格する学生の数を取り除くことも必要かもしれない。

今回はモデルのパラメータは 1 つに固定していたが、本来は真の分布に近づけるよう最適化する必要がある。これまでの研究から、固有ベクトルが参照ベクトルに近づくよう、行列のパラメータを変更する最適化は非常に不安定であることが判明しつつある。目的関数は凸ではなく、様々な局所解があると考えられる。この点について、何らかの改善したアルゴリズムを開発することは今後の目標である。

救急需要予測のための時空間正規混合モデル：横浜市救急データへの応用

横浜市立大学 医学部　三枝 祐輔
横浜市立大学 医学部　三角 俊裕
横浜市立大学 医学部　窪田 和巳
横浜市立大学 医学部　山中 竹春
横浜市消防局　藤田 豊
横浜市消防局　金子 由佳

1. はじめに

近年，救急出場件数は全国的に増加傾向にあり[1]，横浜市においても同様に増加傾向である．今後も増加傾向が持続した場合，救急車の現場到着遅延などの問題が危惧されている．そのため，今後の横浜市における救急需要を予測することは重要な課題である．横浜市消防局は，横浜市立大学と協定を結び，救急需要予測に取り組み，平成 30 年 5 月に予測で使用した救急出場データのオープンデータ化および予測結果の公開を行った[2]．本研究では，「横浜市の救急需要予測に関する研究にかかる協定」にもとづき，救急出場データの提供を受け，救急出場件数の予測を行った．

データは横浜市消防局が保有する 15 年間分，約 250 万件の救急搬送記録を用いた．データには，出場ごとの覚知時刻，出場要請があった場所の緯度経度情報が含まれるため，任意の地域における 1 時間ごとの出場件数を求めることができる．さらに，住居区分（市内在住，市外在住，国外在住），初診時傷病程度（軽症，中等症，重症以上，不取扱い），事故種別（急病，一般負傷，交通事故，転院搬送），行政区（横浜市 18 区）などの付随情報も含まれる．加えて，人口動態データ（横浜市の XX 歳以上の人口，日中の流入人口，外国人延べ宿泊者数など），気象データ（平均気温，前日との気温差，日照時間など），暦（月，連休後の平日など），その他の救急関連情報（#7119 普及率，救急車適正利用広報費など）などの出場件数に影響を与え得るデータも利用可能である．

本研究では将来の救急出場件数の予測を目的として，時空間統計モデルを用いた解析を行った．

2. 時空間統計モデル

救急出場件数を時系列データとして扱った予測に関する研究は，これまでに数多く行われている．たとえば，Channouf et al. (2007) は自己回帰移動平均モデルを導入し，Matteson et al. (2011) は因子モデルを導入した．一方，救急出場の時間と出場先座標（時空間データ）を用いた予測に関する研究は数少なく，たとえば Setzler et al. (2009) はニューラルネットワークを用いて需要予測を行った．特に Zhou et al. (2015) は出場件数にポアソン点過程を仮定した時空間正規混合モデルを用いた予測を行った．以下，Zhou et al. (2015) の方法を導入した．

任意のピリオド t における総出場件数を n_t とするとき，n_t は強度 δ_t のポアソン分布に従うと仮定する $(t = 1, \ldots, T)$．このときピリオド t における i 番目の出場先の座標 s_{ti} は，任意の 2 次元座標 s に対して，密度 $f_t(s)$ をもつ分布に従うとする，ただし，$f_t(s)$ は K 個のコンポーネントからなる正規混合分布であり，任意の t に対して次のように表された：

$$f_t(s; \{p_{tj}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = \sum_{j=1}^{K} p_{tj} \phi(s; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

ここに，ϕ は平均ベクトル $\boldsymbol{\mu}_j$，分散共分散行列 $\boldsymbol{\Sigma}_j$ の 2 次元正規密度関数，p_{tj} は時間に依存する混合比である．

[1] http://www.fdma.go.jp/html/intro/form/kinkyugyoumu_h22_houkoku.html
[2] http://www.city.yokohama.lg.jp/shobo/qq/prediction/

出場件数に周期性があると仮定する．周期長を B $(B < T)$ としたとき，任意のピリオド t が周期上のピリオド b に対応する：$b \bmod B = t \bmod B$. このとき次式を得た．

$$f_t(s; \{p_{tj}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = f_b(s; \{p_{bj}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}).$$

次に混合比 p_{bj} の時系列相関を表現する準備として，次の多項ロジット変換を考える：

$$\pi_{br} = \log \frac{p_{br}}{1 - \sum_{j=1}^{K-1} p_{bj}} \quad (b = 1, \ldots, B; r = 1, \ldots, K-1).$$

混合比の多項ロジット変換 π_{br} が次の条件付き自己回帰 (CAR) 事前分布に従うと仮定する：

$$\pi_{br} | \pi_{-br} \sim \text{Normal}\left(c_r + \rho_r[(\pi_{b-1,r} - c_r) + (\pi_{b+1,r} - c_r)], v_r^2\right) \quad (b = 1, \ldots, B; r = 1, \ldots, K-1),$$

ただし $c_r \sim \text{Normal}(0, 10^4)$, $\rho_r \sim \text{Uniform}(0, 0.25)$, $v_r^2 \sim \text{Uniform}(0, 10^4)$ とする．

推定の準備として，他のパラメータが次の事前分布に従うと仮定する：

$$\mu_j \sim \text{Normal}(\xi, \kappa^{-1}), \quad \Sigma_j^{-1} | \beta \sim \text{Wishart}(2\alpha, (2\beta)^{-1}) \quad (j = 1, \ldots, K),$$

ただし $\beta \sim \text{Wishart}(2g, (2h)^{-1})$, $\alpha = 3$, $g = 1$, $\xi = (\xi_1, \xi_2)^\top$, $\kappa = diag(1/R_1^2, 1/R_2^2)$, $h = diag(10/R_1^2, 10/R_2^2)$ とし，推定を行うときは ξ_1 と ξ_2 は全観測度数の 2 次元空間座標の中央値，R_1 と R_2 は全観測度数の 2 次元空間座標の範囲長を与える．以上の仮定の下で酔歩連鎖メトロポリスヘイスティング法を用いて，各パラメータのベイズ推定量を得た．

本研究では Zhou et al. (2015) の方法を横浜市のデータに応用した．データ解析の結果の詳細は当日報告した．

参考文献

Channoouf, N., L'Ecuyer, P., Ingolfsson, A. and Avramidis, A. (2007). The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health Care Management Science* **10**, 25–45.

Matteson, D. S., McLean, M. W., Woodard, D. B. and Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics* **5**, 1379–1406.

Setzler, H., Saydam, C. and Park, S. (2009). EMS call volume predictions: A comparative study. *Computers and Operations Research* **36**, 1843–1851.

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G. and Micheas, A. C. (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association* **110**, 6–15.

生態学的大規模データを統合する階層モデル
―種個体数分布の広域予測―

国立環境研究所　深谷肇一
琉球大学　楠本聞太郎　　琉球大学　塩野貴之
琉球大学　藤沼潤一　　琉球大学　久保田康裕

生態学において，広域での生物多様性評価は基礎と応用の両面から重要な問題である．生態群集を構成する各種の個体数（種個体数分布）に基づく評価はとりわけ重要であるが，群集レベルでの個体の計数調査（範囲内の個体を数え上げるための局所的なプロットを設置して行う）は一般的に多くの労力と資金を要するため，広域に適用することは困難である．

一方，より簡便な生態群集の調査法として，プロットに現れる種ごとの個体数ではなく，出現のみを記録する方法がある．種の出現を調べることは一般的に個体を数え上げることよりも簡単であり，計数調査と比較して多数のプロットを広範囲に設置することができる．また，生態群集を構成する個々の種を見れば，博物館標本や出現記録，専門家の査定などに基づいて，その地理的分布範囲が大まかに分かっている場合も少なくない．

生態群集の個体計数データとは対照的に，局所的な群集調査から得られる種の出現データと種の地理的分布情報はすでに多数の蓄積がある場合があり，これらは非常に大きな空間範囲における種個体数分布の評価に利用できる可能性がある．本講演では，これら 2 種類の「生態学的大規模データ」を統合して，広域の種個体数分布を予測するための階層モデルを提案した．また，このモデルを蓄積されている植生調査データと地理分布データに適用し，日本に分布する木本種の種個体数分布を推定した結果を報告した．

モデル

推測の対象となる I 種が生息する，ある地理的領域が J 個の地理メッシュに分割されており，メッシュ j（$j = 1, \ldots, J$）には，各種の出現が調査される局所プロットが $K_j \geq 0$ 個反復して設置されているとする．メッシュ j 内のプロット k における種 i の検出（1）・不検出（0）を y_{ijk} と表す．メッシュ j 内のプロット k の面積を a_{jk} と表す．また，各メッシュにおいて，局所プロットによる種の出現情報とは独立に，メッシュレベルでの種の分布情報（種の分布地図や地域の生息種目録など）が得られている場合があるものとする．

局所的に得られる二値データ（種の検出・不検出）から各種の個体密度を推定するために，各メッシュにおいて個体の空間配置が種ごとに独立な均一ポアソン点過程に従うこと，

およびプロットが重なって設置されることはないことを仮定する．メッシュ j における種 i の分布を z_{ij}（分布していれば $z = 1$，そうでなければ $z = 0$），種が分布してる場合の条件付き密度を d_{ij} とおくと，以上の仮定の下で種の検出過程は以下のようにモデル化される．

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk}) \tag{1}$$

$$p_{ijk} = 1 - \exp(-z_{ij} d_{ij} a_{jk}) \tag{2}$$

$$z_{ij} \sim \text{Bernoulli}(\psi_{ij}) \tag{3}$$

ここで z_{ij} は，局所プロットデータおよび地理分布データによって部分的に観測される潜在状態変数であり，$\psi_{ij} = \Pr(z_{ij} = 1)$ は種の分布確率である．ψ_{ij} および d_{ij} はそれぞれ，変量効果を含むロジット線形モデルと対数線形モデルによってモデル化する．

木本群集データへの適用

上記のモデルを蓄積されている植生調査データと地理分布データに適用し，日本に分布する木本種の種個体数分布を推定した．局所プロットによる出現データとして，学術研究調査および環境省の自然環境保全基礎調査の中で行われた自然林の植生調査データを利用した．このデータには日本に生息する木本種のほぼ全てを網羅する 1,248 種が出現しており，今回の研究ではその全てをモデル化の対象とした．また，対象の地域（日本）全体を 4,684 個の 10 km 平方メッシュに分割し，その全てをモデル化の対象とした．メッシュレベルでの種分布情報として，植物標本，種の出現記録，日本植物分布図譜による分布地図，および各都道府県による種目録を利用した．

日本全体での木本の総個体数は約 204 億と推定された．これは，全球における木本個体数の推定値 (Crowther *et al.* 2015, *Nature* **525**:201–205) のおよそ 0.671% である．国連食糧農業機関（FAO）の 2015 年の統計によると，全球の森林面積に対する日本の自然林面積の割合は 0.367% であることから，本研究の結果は独立に行われた全球規模の推測と大きく食い違うものではないようである．また，モデルの予測が局所的な個体数計数データと整合的であるかどうかを調べるために，重要生態系監視地域モニタリング推進事業（モニタリングサイト 1000），森林生態系多様性基礎調査，および学術研究において収集された個体計数データ（毎木データ）とモデル予測値の関係を調べた．その結果，各データセットにおいて，対数個体数の実測値と予測値の間には緩やかな正の相関関係が認められ，やや過少予測の傾向は見られるものの，概ね 1:1 の線の周りに分布することが分かった．過少予測の傾向は，環境の異質性や種間相互作用によって生じる集中分布がモデルで説明されていないことが原因であると考えられる．

一般化平均に基づく予測モデリング

成蹊大学　小森　理

線形モデルから非線形モデルへの拡張にはさまざまな方法がある．リンク関数を用いさまざまな応答変数を扱えるようにした一般化線形モデル [7], 非線形性を変量ごとに限定した一般化加法モデル [3], 木構造を仮定した tree-based model[1]．そしてニューラルネットワークとそれを拡張した深層学習が近年注目を浴びている．しかしモデルの簡潔性と結果の解釈可能性が非線形モデルの難点として上げられる．そこで教師なし学習であるクラスターを用いて，そのクラスター内では単純な線形モデルを仮定し（クラスター内でのデータの等質性の仮定），クラスター間の異質性は柔軟にモデル化することを考える．それが一般化平均を用いた準線形モデルである．一般化平均は Kolmogorov-Nagumo 平均とも呼ばれる [2, 8]．予測や判別問題におけるクラスター構造の重要性については，例えば乳癌などのデータは質的に均一ではなく，いくつかの subtype が存在し，その subtype により予後の善し悪しが大きく影響されることも知られている [10]．またデータの異質性を考慮した解析手法も近年注目を浴びている [6, 9]．

単調増加関数 ϕ, クラスター数を K とすると，変量 $x \in \mathbb{R}^p$ に対する一般化平均に基づく準線形モデルは

$$\phi^{-1} \quad \sum_{k=1}^{K} \pi_k \phi(\eta_k(x))$$

と書ける．$\eta_k(x)$ はクラスター k における線形モデルであり，$0 < \pi_k < 1$ はクラスターの混合比とする．ロジスティック回帰モデルにおいて，予測子に上記の準線形モデル (但し $K = 2$ とし $\eta_2(x)$ に定数を仮定する) を考えると非対称ロジスティックモデルが導出される [4]．この非対称ロジスティック回帰モデルの推定方程式に現れる重み関数によってサンプルのサイズの偏りが軽減されることがこのモデルの特徴である．またこの非対称ロジスティックモデルに二重頑健性を持たせたモデルも考えることができ [5]，データのサンプリングバイアスを考慮した推定が可能となる． またポアソン点過程の強度のモデリングにも上記の準線形モデリングを適用することができ，生息分布予測に効果的であることが分かってきた．当日の発表では一般化平均に基づく準線形モデリングを予測の問題に適用した具体例をいくつか紹介し，その有用性を議論する．

参考文献

[1] **Clark, L. A. and Pergibon, D.** Tree-based models. , *Statistical Models in S*, AT & T Bell Laboratories California 1992.

[2] **Eguchi, S. and Komori, O.** Path Connectedness on a Space of Probability Density Functions. , *Geometric Science of Information: Second International Coference, GSI 2015*, Cham: Springer International Publishing, 2015. p. 615.

[3] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

[4] Komori, O., Eguchi, S., Ikeda, S., Okamura, H., Ichinokawa, M. and Nakayama, S. (2016). An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution* **7**, 249–260.

[5] ——— , , ——— , , Saigusa, Y., Okamura, H. and Ichinokawa, M. (2017). Robust bias correction model for estimation of global trend in marine populations. *Ecosphere* **8**, 1–9.

[6] Li, Y., Wu, F.-X. and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics* **325**, 325–340.

[7] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, New York: Chapman & Hall.

[8] Naudts, J. (2011). *Generalised Thermostatistics*, London: Springer.

[9] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S. L., Chou, K. and Pearson, M. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* **1**, 1–10.

[10] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E. and Børresen-Dale, A. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–10874.