2021 年 12 月 3 日（金）、4 日（土）
オンライン開催

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）
シンポジウム

データサイエンス・統計学における方法論と応用の新展開

開催責任者：植木優夫（長崎大学情報データ科学部）

内容・目的

データサイエンスや統計学の重要性が社会的に認知されつつある一方で、年々拡大するデータの種類や規模、複雑さに合わせて、データを扱う技術自体も発展しています。本シンポジウムでは、データサイエンスや統計学における新しい方法論や手法とその応用、様々な領域のデータに対するモデリングや解析事例など、幅広いテーマの講演を募集します。異なる問題意識を持つ研究者が相互に交流することで、データサイエンス・統計学分野における最新の動向や問題点を理解し、問題解決や新たな研究の発展につなげることを目的とします。

【プログラム】

12 月 3 日（金）

1. 13:30-14:05 山口崇幸 (滋賀大学データサイエンス教育研究センター)
ベンチマークドーズ法とモデル平均化の機能を持つ毒性データの評価のためのソフトウェア開発の紹介

2. 14:05-14:40 吉田拓真 (鹿児島大学理工学研究科)、桃木光輝 (鹿児島大学理工学研究科)
極値統計学における予測モデリングについて

3. 14:40-15:15 本田敏雄 (一橋大学経済学研究科)、Chien-Tong Lin (Department of Statistics, National Tsing Hua University)
Forward variable selection for ultra-high dimensional quantile regression models

休憩

4. 15:55-16:30 高橋将宜 (長崎大学情報データ科学部)
多重代入法不連続デザインによる局所因果効果の推定

5. 16:30-17:05 江村剛志 (久留米大学バイオ統計センター)
高次元スコア検定を用いた生存データに基づく決定木の構築

6. 17:05-17:40 柳川　堯 (久留米大学バイオ統計センター)
Big data の統計的検定

12 月 4 日（土）

7. 10:00-10:35 廣瀬雅代 (九州大学マス・フォア・インダストリ研究所)、Malay Ghosh (University of Florida)、Tamal Ghosh (University of Florida)
A Prediction Error under Area Level Model with Arc-Sin transformation for Sample Proportions

8. 10:35-11:10 斎藤正也 (長崎県立大学シーボルト校情報システム学部情報セキュリティ学科)
感染症流行記述におけるメタ・ポピュレーションモデルの課題

9. 11:10-11:45 中杤昌弘 (名古屋大学大学院医学系研究科)
日本人を対象とした大規模 SNP データ研究の近年の成果と課題

休憩

10. 13:20-13:55 Muthu Subash Kavitha (長崎大学情報データ科学部)
Desired weight learning and controlled class-variant mechanism improve the data imbalance problems in deep network for biomedical applications

11. 13:55-14:30 田尻　涼 (佐賀大学大学院医学系研究科)、川口　淳 (佐賀大学医学部)
深層学習により生成した脳画像は解析に使用できるのか

12. 14:30-15:05 高橋雄太 (東北大学病院精神科)
精神医学データに対する機械学習的アプローチ

ベンチマークドーズ法とモデル平均化の機能を持つ毒性データの評価のためのソフトウェア開発の紹介

山口 崇幸　(西浦博氏, 茅野大志氏との共同研究)

1　はじめに

毒性データのリスク評価に用いられるベンチマークドーズ (BMD, benchmark dose) 法では, 統計モデルを観察データに適用し, 無毒性量に相当する BMD 信頼下限 BMDL (BMD lower limit) が得られる. モデルの不確かさを考慮するために複数の統計モデルの平均化が考えられており [1–3], 最も適合度の良い 3 つのモデルだけの平均化から安定した妥当性と信頼性が得られることが確認されている [4]. BMD 法のソフトウェアとしては BMDS [*1] や PROAST [*2] が知られているが, 既存の BMD 法のソフトウェアには適合度の良い少数のモデルだけの平均化が可能なものは存在しなかった. そのため, 最も適合度の良い少数のモデルだけの平均化機能を持つソフトウェア BMDMA (Benchmark dose modeling and model averaging) [*3] を開発した. BMDMA は R のパッケージとして開発されており, プログラミングで用いるための関数と日本語と英語の GUI 機能を持つ (図 1). GPLv3 のライセンスでオープンソースソフトウェアとして公開している.

2　二値用量反応データ

BMDMA が対応するデータは二値用量反応データであり, 用量, 各用量における実験の対象の数, 反応が出た対象の数からなる. それぞれの値は BMDMA のウィンドウに表示される表の dose, N, response として表示される (図 1). 以降, 観察データは D 個の用量 d_i $(i = 1, \ldots, D)$ で実験が行われたとして, 用量 d_i での実験の対象の数を N_i, 反応が出た対象の数を T_i とする.

3　モデルの推定と平均化

BMDMA は Logistic, Log-logistic, Probit, Log-probit, Gamma, Quantal linear, Weibull, Multistage2, Multistage3, Dichotomous hill の 10 の統計モデルに対応している. 各モデルには 2〜4 個のパラメータがあり, パラメータの取りうる値の範囲が定められている. いくつかのモデルに対しては, パラメータ値の制限を追加するオプションがあり, このオプションの利用の有無を設定の上でモデルの適合を行う. パラメータの推定は, 実験の対象の反応が二項分布にしたがうとして尤度を定義し, 準ニュートン法の BFGS や L-BFGS-B を用いた最尤法で行う. インデックス m の用量反応関係のモデルを p_m とし, モデルのパラメータの数を M_m $(2 \le M_m \le 4)$, パラメータを $x_1, \ldots, x_{M_m}$ とすると, 尤度は次で定義される.

$$L(x_1, \ldots, x_{M_m}) = \prod_{i=1}^{D} \binom{N_i}{T_i} p_m(d_i)^{T_i} (1 - p_m(d_i))^{N_i - T_i}. \tag{1}$$

モデル平均化は, オプションのパラメータ値の制限がないモデルから, 全モデルまたは AIC (Akaike's Information Criterion) の小さい上位 3 つのモデルに対して行う. それぞれを MA ALL, MA3 とおく. 平均化の重みは AIC から計算する. S はモデルのインデックスの集合, I_m はモデル m の AIC とすると, モデル平均化は

$$p_{\mathrm{MA}}(d) = \sum_{m \in S} w_m p_m(d), \; w_m = \frac{e^{-I_m/2}}{\sum_{m \in S} e^{-I_m/2}} \tag{2}$$

[*1] https://www.epa.gov/bmds/benchmark-dose-software-bmds-version-32-download
[*2] https://www.rivm.nl/en/proast
[*3] https://math-numerical-experiment.gitlab.io/bmdma/

によって計算される. S は MA ALL の場合はパラメータの範囲の制限がない全モデルのインデックスの集合であり, MA3 の場合はパラメータの範囲の制限がないモデルの中で AIC が小さい上位 3 つのモデルのインデックスの集合である.

4　ブートストラップ法による BMDL の計算

BMD は用量反応関係のモデルと BMR (benchmark response) から定められる. BMR は反応が出る確率の増加を表し, 過剰リスクと追加リスクがある. BMDMA のデフォルトは過剰リスクで BMR $= 0.1$ (10%) である. p を用量反応関係とする. 過剰リスクを考えるとき, BMD は BMR $= (p(d) - p(0))/(1 - p(0))$ を満たす用量 d として定義される. また, 追加リスクのとき, BMD は BMR $= p(d) - p(0)$ を満たす用量 d として定義される. BMDL は BMD の信頼限界であり, BMDMA ではパラメトリックブートストラップによって求める.

5　ソフトウェアの機能

GUI では図 1 のようなウィンドウが表示される.「ファイル」ボタンでファイルを読み込み,「解析の実行」ボタンを押して解析を開始する. 解析が終わると, グラフにデータに適合したモデルが表示され, BMDL などの値は「結果」タブに表として表示される.

データに適合したモデルは, 評価ロジックにしたがって Unusable, Questionable, Viable (Warning あり), Viable (Warning なし) のいずれかに分類される. BMDMA では MA3, MA ALL の順番で最初に Unusable でないモデルを推奨モデルとする. 評価ロジックの結果, 推奨モデルは GUI の表やレポートに表示される.

解析の結果は,「解析結果の出力・保存」ボタンを押すと, フォルダに保存することができる. 解析に用いたデータの CSV ファイル, 結果の表の CSV ファイル, 各モデルのフィッティングのグラフの PNG ファイルが保存される. 加えて, 結果の表や各モデルのグラフを含むレポートが Word 形式のファイルとして保存される. レポートは Pandoc により生成されるので, PDF などの他の形式で出力することも可能である.

謝辞

本研究は平成 31 年度・令和 2 年度食品健康影響評価技術研究「課題名: 二値反応の用量反応データを対象としたベンチマークドーズ計算ソフトウェアの開発研究 (課題番号: 1907)」の助成を受けたものです.

図 1　解析が終わった後のウィンドウの設定タブ. 右下のグラフの中に, モデル平均化を含む適合したすべてのモデルが表示されている.

参考文献

[1] M. W. Wheeler and A. J. Bailer, Risk Analysis **27**, 659–670 (2007).

[2] M. Wheeler and A. J. Bailer, Journal of Statistical Software, Articles **26**, 1–15 (2008).

[3] M. W. Wheeler and A. J. Bailer, Environmental and Ecological Statistics **16**, 37–51 (2009).

[4] K. Yoshii et al., Theoretical Biology and Medical Modelling **17**, 13 (2020).

極値統計学における予測モデリングについて

吉田拓真 (鹿児島大学), 桃木光輝 (鹿児島大学)

1 はじめに

極値統計学は希な事象の予測に用いられ, 数学的には確率分布の裾の挙動を推測するための理論である. 目的変数 Y, 説明変数 $\boldsymbol{X} = (X_1, \ldots, X_p)$ に対して, $F(y|\boldsymbol{x}) = P(Y < y|\boldsymbol{X} = \boldsymbol{x})$ を $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_p)$ を与えた下での Y の条件付き分布関数とする. このとき, 極値統計学の重要な結果として, 適切な条件の下で, $\gamma(\boldsymbol{x}) > 0$ に対して,

$$\lim_{w\to\infty} P(Y > wy|\boldsymbol{X} = \boldsymbol{x}, Y > w) = \lim_{w\to\infty} \frac{1 - F(wy|\boldsymbol{x})}{1 - F(w|\boldsymbol{x})} = y^{-1/\gamma(\boldsymbol{x})} \tag{1}$$

が成立することが知られている. ここで, $\gamma(\boldsymbol{x})$ は extreme value index (EVI) と呼ばれる関数であり, ここでは, $\gamma(\boldsymbol{x}) > 0$ と仮定する. (1) から, Y がある値 w より大きい条件下での条件付き確率は $\gamma(\boldsymbol{x})$ のみに依存していることが分かるため, これを推定すれば希な事象の統計的推測が可能となる. このことから, 極値統計学では, EVI のモデリング・推定が重要であることがわかる. なお, 閾値は $\boldsymbol{x}$ に依存させてもよく, $w = w(\boldsymbol{x})$ とする.

ここからは, $\gamma(\boldsymbol{x})$ の推定を議論する. γ のノンパラメトリック推定量は Ma et al. (2020) など広く議論されているが, p が大きい場合は次元の呪いを受け, まともな推定ができない. したがって、次元の呪いを回避するための方法の開発が必要となる. Wang and Tsai (2009) は $\gamma(\boldsymbol{x}) = \exp[\boldsymbol{x}^T\boldsymbol{\theta}], \boldsymbol{\theta} \in \mathbb{R}^p$ を考えた. $\exp[\cdot]$ を用いているのは, $\gamma(\cdot) > 0$ を保証するためである. この他, Youngman (2019) は加法モデル $\gamma(\boldsymbol{x}) = \exp[\gamma_1(x_1) + \cdots + \gamma_p(x_p)], \gamma_j; \mathbb{R} \to \mathbb{R}$ を考案した. また, Li et al. (2020) は, 共変量を $(\boldsymbol{x}, t)$ としたとき, partial linear model $\gamma(\boldsymbol{x}, t) = \exp[\boldsymbol{x}^T\boldsymbol{\theta} + g(t)]$ を扱った. そして, Momoki and Yoshida (2021, preprint) は説明変数を $(\boldsymbol{x}, \boldsymbol{t})$ として, varying coefficient model $\gamma(\boldsymbol{x}) = \exp[f_1(\boldsymbol{t})x_1 + \cdots + f_p(\boldsymbol{t})x_p]$ を研究している. ここで, $\boldsymbol{t} \in \mathbb{R}^q$ はノンパラメトリック項だが, 時間変数として $q = 1$, 位置変数として $q = 2$ を想定している. 本講演では Single index model $\gamma(\boldsymbol{x}) = \exp[\alpha(\boldsymbol{x}^T\boldsymbol{\theta})], \alpha : \mathbb{R} \to \mathbb{R}$ に関して得られた結果を報告した.

2 Single index model

確率変数 $(Y, \boldsymbol{X}) \in \mathbb{R}_+ \times \mathcal{X}, \mathcal{X} \subset \mathbb{R}^p$, $\boldsymbol{X} = (X_1, \ldots, X_p), X_j \in \mathbb{R}$ に対し, $F(y|\boldsymbol{x}) = P(Y < y|\boldsymbol{X} = \boldsymbol{x})$ を $\boldsymbol{X} = \boldsymbol{x} = (x_1, \ldots, x_p)$ を与えた下での Y の条件付き生存関数とする. このとき, 我々はパレート型分布を考える:

$$P(Y > y|\boldsymbol{X} = \boldsymbol{x}) = 1 - F(y|\boldsymbol{x}) \approx y^{-1/\gamma^*(\boldsymbol{x})}. \tag{2}$$

ただし, $\gamma^*(\boldsymbol{x}) > 0, \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$ は extreme value index (EVI) である. 本講演では, EVI が single index model $\gamma^*(\boldsymbol{x}) = \gamma(\boldsymbol{x}^T\boldsymbol{\theta})$ という構造を用いた. ただし, $\gamma : \mathbb{R} \to \mathbb{R}_+$ は 1 変数関数であり, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$ は p 次元パラメータベクトルである. このとき, Y の条件付き生存関数は

$$P(Y > y|\boldsymbol{X}^T\boldsymbol{\theta} = \boldsymbol{x}^T\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{x}) \approx y^{-1/\gamma(\boldsymbol{x}^T\boldsymbol{\theta})}\{1 + O(y^{-\beta(\boldsymbol{x})/\gamma(\boldsymbol{x}^T\boldsymbol{\theta})})\}. \tag{3}$$

と書ける. ただし, $0 < \beta < \beta(\boldsymbol{x}) < \infty$ である.

推定量 $(\gamma, \boldsymbol{\beta})$ は最尤法の枠組みで推定できる. しかし, ノンパラメトリック関数 γ は関数の滑らかさを担保するために, 罰則項を付与した平滑化スプラインで推定する. すなわち,

$$\begin{aligned} U_n(\gamma, \boldsymbol{\theta}|\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left[\gamma(\boldsymbol{X}_i^T\boldsymbol{\theta})^{-1}\log\left(\frac{Y_i}{w_n(\boldsymbol{X}_i)}\right)+\log(\gamma(\boldsymbol{X}_i^T\boldsymbol{\theta}))\right] I\left(Y_i > w_n(\boldsymbol{X}_i)\right) \\ &\quad +\frac{\lambda}{2}\int_a^b\left\{\frac{d^m}{dx^m}\gamma^{(m)}(x)\right\}^2 dx \end{aligned} \tag{4}$$

を最小にする $(\gamma, \boldsymbol{\theta})$ を推定量とし, $(\hat{\gamma}, \hat{\boldsymbol{\theta}})$ と書く. 第 1 節でまとめた先行研究同様, $\gamma(\cdot) = \exp[\alpha(\cdot)]$ と書き, $(\alpha, \boldsymbol{\theta})$ を推定する話に置き換えることも可能である.

この推定量について, データ数 n が増大する下での漸近理論を構築した. 結果として, 適切な設定の下で, 以下の定理が成立する.

Theorem 1. 適切な条件を仮定する. このとき, $n \to \infty$ の下で,

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 = O(n^{-\frac{2\beta}{2\beta+1}})$$

と

$$\sup_{\boldsymbol{x}\in\mathcal{X}}|\hat{\gamma}(\boldsymbol{x}^T\hat{\boldsymbol{\theta}}) - \gamma(\boldsymbol{x}^T\boldsymbol{\theta})|^2 \leq O\left(n^{-\frac{2\beta}{2\beta+1+1/m}}\right)$$

が成り立つ.

講演内では, 提案モデルの数学的性質のみでなく, 数値実験によるモデルの正当性も報告した.

3 まとめ

本研究は多変数共変量を導入した極値統計学におけるモデリング手法の拡充が目的である. 多くのモデリングの中で, 本講演は single index model に着目し, 手法の確率, 漸近理論を構築した. 得られた結果から, 次元の呪いを回避する手法を極値統計学に持ち込むことに成功したと言える. 今後の研究としては, 共変量が膨大な場合である高次元における展開が重要である.

References

[1] Li,R., Leng,C. and You,J. (2020). Semiparametric Tail Index Regression. *Journal of Business & Economic Statistics*. In Press.

[2] Ma, Y., Wei, B. and Huang,W. (2020). A nonparametric estimator for the conditional tail index of Pareto-type distributions. *Test*. **83** 17–44.

[3] Momoki, K. and Yoshida,T. (2021). Varying coefficient models for extreme value index regression. *Preprint*.

[4] Wang, H. and Tsai, C.L. (2009). Tail index regression. *Journal of the American Statistical Association* **104** 1233–1240.

[5] Youngman. B. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *J. Amer. Statist. Assoc.* **114** 1865–1879.

Forward variable selection for ultra-high dimensional quantile regression models

一橋大学大学院経済学研究科　本田敏雄

本報告は，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 Honda and Lin(2021) に基づく．詳細は，Honda and Lin(2021) を参照されたい．

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．Lasso や SCAD などは高次元線形回帰モデルなどに対する推定法であり，高次元データ解析の必要性，重要性の高まりとともに，実用面および理論面での研究が進んできた．そして現在では，それらは高次元データ解析での標準的なツールになっている．

しかしながら次元数 p が非常に大きい場合，計算上の問題から Lasso，SCAD を適用することはできず，事前に feature screening と呼ばれる変数選択を行い，変数の絞りこみを行うことがよく行われる．

周辺モデルによる方法としては，Fan and Lv(2008)，Fan and Song(2010) などがある．被説明変数と個々の説明変数間の相関の指標による方法も多くある (He et al.(2013) など)．これらは，一度のみの適用では重要な説明変数を見落としがちであることが知られており，厳密な理論なしでのこれらの繰り返しにより，feature screening をすることも多い．

一方，繰り返しを前提とした前進型の選択法（forward type feature screening）としては，Wang(2009), Ing and Lai(2011), Cheng, Honda, and Zhang(2016), Zheng et al.(2020), Pijyan et al.(2020), Honda and Lin(2021) などがある．しかしながら分位点回帰については，十分な研究結果は得られていなかったため (Kong et al.(2019)，Tang et al.(2021+) など)，本報告の研究を行った．

以下の超高次元 τ 分位点線形回帰モデルを考える．説明変数 $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ は p 次元で，第一要素は 1 とする．この p は非常に大きく，標本数 n の指数のオーダーということもありうる．

$$
\begin{aligned}
&Y = \boldsymbol{X}^T\boldsymbol{\beta}^* + \epsilon \quad \text{かつ} \quad \mathrm{E}\{\psi_\tau(\epsilon)|\boldsymbol{X}\} = 0,\\
&\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, \mathrm{E}\{\rho_\tau(Y - \boldsymbol{X}^T\boldsymbol{\beta}),\\
&\psi_\tau(t) = \tau - I(t \le 0) \quad \text{かつ} \quad \rho_\tau(t) = t\{\tau - I(t \le 0)\}.
\end{aligned}
$$

簡単のため，$j \geq 2$ で，$\mathrm{E}\{X_j\} = 0$ かつ $\mathrm{E}\{X_j^2\} = 1$ とする．

回帰係数の内非ゼロ要素をもつ添え字の集合を $\mathcal{M}$ とすると，p は非常に大きいがこの $\mathcal{M}$ の要素数 $|\mathcal{M}|$ は n に比べても十分小さいとする．そして $(Y, \boldsymbol{X})$ の n 個の独立同一標本 $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ があるとする．

三つの手法を提案しているが，ここでは我々が FR と呼んでいるもののみを述べる．まず部分モデル $S \subset [p] = \{1, \ldots, p\}$ に対する目的関数 $L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S)$ を以下に定義する．

$$L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S)$$

(1) $S = S_{k-1}$ ($S_0 = \{1\}$) として，$k = 1$ から開始とする．

$$j_k = \underset{j \in S^c}{\operatorname{argmin}} \min_{\boldsymbol{\beta}_{S \cup \{j\}}} L_n(\boldsymbol{X}_{S \cup \{j\}}^T \boldsymbol{\beta}_{S \cup \{j\}}).$$

(2) $S = S_{k-1}$ として，

$$L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S) - \min_{j \in S^c} \min_{\boldsymbol{\beta}_{S \cup \{j\}}} L_n(\boldsymbol{X}_{S \cup \{j\}}^T \boldsymbol{\beta}_{S \cup \{j\}}) > \xi_n |S| \log p_n / n$$

であれば，$S_k = S_{k-1} \cup \{j_k\}$ とおき，**(1)** に戻る．そうでない場合には，$\widehat{\mathcal{M}} = S_{k-1}$ として終了する．

(2) の ξ_n は無限大に発散さえすればよいが，最適性の議論はできないので 1 としてシミュレーション等を行った．

提案した手法に関する理論的な結果として，定理 1 で目的関数の減少についての結果を述べ，定理 2 で停止則の性質を示し，定理 3 で screening consistency を証明した．

シミュレーションにより，提案した手法と Lasso, SCAD, MCP, CQU(Wu and Yin(2015)), FR-PQU(Kong et al.(2019)) を比較した．その結果，提案した手法の有用性を示すことができた．

参考文献

Honda, T. and Lin, C. T. (2021). *Forward variable selection for ultra-high dimensional quantile regression models.* Discussion Papers 2021-02, Graduate School of Economics, Hitotsubashi University.

多重代入法不連続デザインによる局所因果効果の推定

高橋将宜（長崎大学情報データ科学部）

要旨

本研究では，潜在的結果変数の欠測部分を多重代入法で処理することで，回帰不連続デザインのように，閾値における局所的な平均因果効果を推定する方法を議論する．具体的には，Takahashi（2021）によって提案された多重代入法不連続デザイン（MIRDD: multiple imputation regression discontinuity）を紹介する．

1. はじめに

Rubin（1974）は，潜在的結果変数の枠組み（potential outcomes framework）によって因果推論を欠測データの問題として確立した．因果推論の根本問題（Holland, 1986, p.947）のため，潜在的な結果のどちらか一方しか観測されないからである．観察研究における因果推論手法の中でも，回帰不連続デザイン（RDD: regression discontinuity design）は，局所的な無作為割付けが成立しているとみなされることから，最も優れた準実験（quasi-experiment）の 1 つとされている（Lee and Lemieux, 2015, p.304; Kim and Steiner, 2016）．近年の RDD の先行研究は，潜在的結果変数の枠組みに基づいて議論されている（Lee, 2008, p.678; Imbens and Kalyanaraman, 2012, p.934; Calonico et al., 2014, p.2298）．

一方，Rubin（1987）は，欠測データの統計解析を妥当なものにする方法として多重代入法（multiple imputation）を提案した．現代の統計解析では，「多くの分野において，多重代入法は不完全データに対処するための最もよい汎用的な手法として受け入れられている」（van Buuren, 2018, p.30）と指摘されている．

このように，統計的因果推論と欠測データ解析の両方の研究分野において，Rubin（1974, 1987）は重要な役割を果たしてきた．実際，Rubin（2004, pp.167-168）は，潜在的結果変数の欠測部分を多重代入法によって処理する可能性に言及しており，統計的因果推論の手法として多重代入法を活用しようという提案もされてきている（Imbens and Rubin, 2015, pp.150-171; van Buuren, 2018, pp.241-255）．しかし，閾値の周辺における局所的な平均処置効果（LATE: local average treatment effect）の推定について，多重代入法にどのような性質があるのかは不明であった．

Takahashi（2021）は，多重代入法不連続デザイン（MIRDD）という新たな統計的因果推論手法を提案した．112 通りの異なる設定環境において，それぞれ 5,000 回の繰り返し演算を実行したモンテカルロ・シミュレーションを実行したところ，バイアス，二乗平均誤差の平方根（RMSE: root mean squared error），信頼区間のカバー率，信頼区間の長さの 4 つの観点から，従来の RDD と比較してよい結果が示された．また，追加のシミュレーションにより，Calonico et al.（2014）および Branson et al.（2019）の最新の RDD 手法と比べても遜色のない結果が示されている．さらに，Takahashi（2021）は， RDD 解析の妥当性を診断するために，MIRDD に基づくグラフ手法を提案した．Imbens and Lemieux（2008, p.622）が指摘するように，グラフによる解析は，RDD 解析における重要な要素である．

2. 閾値における LATE を推定する手法としての多重代入法

Imbens and Rubin（2015, pp.150-171）および van Buuren（2018, p.245）にならい，2 つの代入モデルを交互に使用することで，潜在的結果変数の欠測部分に対して多重代入法によるシミュレーション値を生成する． 提案手法では，2 つの代入モデルを交互に使用するので，潜在的結果変数$\tilde{Y}(1)$と$\tilde{Y}(0)$に対して，代入モデルのパラメータは異なるものが推定される．これによって，回帰の平行性の仮定が満たされていなくても，因果効果を適切に推定することができる．つまり，$Y(1)$と$Y(0)$が異なる回帰モデルであったとしても，使用できるということである．これは，局所回帰の概念に対応している．

3. モンテカルロ・シミュレーションの設定と結果

7 種類のデータ生成プロセスを用いて，さまざまなバンド幅と閾値を設定し，合計で 112 の異なる設定環境下において，5,000 回のモンテカルロ・シミュレーションを実行した．バイアス，RMSE，信頼区間のカバー率，信頼区間の長さの 4 つの指標で性能を評価した．詳細な結果は，Takahashi（2021）を参照されたい．また，Takahashi（2021）の online appendix も参照されたい．

4. ソフトウェア

本研究において提案した MIRDD を実行できるように，R パッケージ `MIRDD` を開発した．このソフトウェアと利用の手引きは，https://github.com/mtakahashi123/MIRDD からダウンロードして使用できる．R のコンソールに `MIdiagRDD(y, x, cut)` とタイプするだけで簡単に分析を実行できる．また，さまざまなオプションも引数として設定して利用可能である．

参考文献

[1] Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019), “A Nonparametric Bayesian Methodology for Regression Discontinuity Designs,” *Journal of Statistical Planning and Inference*, 202, 14-30.

[2] Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs,” *Econometrica*, 82 (6), 2295-2326.

[3] Holland, P. W. (1986), “Statistics and Causal Inference,” *Journal of the American Statistical Association*, 81 (296), 945-960.

[4] Imbens, G. W., and Lemieux, T. (2008), “Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142, 615-635.

[5] Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press.

[6] Imbens, G., and Kalyanaraman, K. (2012), “Optimal Bandwidth Choice for the Regression Discontinuity Estimator,” *The Review of Economic Studies*, 79 (3), 933-959.

[7] Lee, D. S. (2008), “Randomized Experiments from Non-Random Selection in U.S. House Elections,” *Journal of Econometrics*, 142, 675-697.

[8] Lee, D. S., and Lemieux, T. (2015), “Regression Discontinuity Designs in Social Sciences,” in *The Sage Handbook of Regression Analysis and Causal Inference*, eds. H. Best and C. Wolf, Thousand Oaks: Sage Publications.

[9] Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66 (5), 688-701.

[10] Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York, NY: John Wiley & Sons.

[11] Rubin, D. B. (2004), “Direct and Indirect Causal Effects via Potential Outcomes,” *Scandinavian Journal of Statistics*, 31 (2), 161-170.

[12] Takahashi, M. (2021), “Multiple Imputation Regression Discontinuity Designs: Alternative to Regression Discontinuity Designs to Estimate the Local Average Treatment Effect at the Cutoff,” *Communications in Statistics: Computation and Simulation*. doi: https://doi.org/10.1080/03610918.2021.1960374.

[13] van Buuren, S. (2018), *Flexible Imputation of Missing Data* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

報告書；高次元スコア検定を用いた生存データに基づく決定木の構築

江村剛志　(Email:takeshiemura@gmail.com)

久留米大学バイオ統計センター: 〒830-0011 福岡県久留米市旭町 67

分類木(または決定木)は，標本の二分割を繰り返しながら構成していく統計モデルである．生存時間データに基づく生存木は，医師が患者の属する疾患のグループを特定し，予後を予測するための極めて実践的で有用なツールである．しかしながら，生存時間データが遺伝発現量などの高次元共変量を含む場合，多重検定と擬陽性の問題により既存の手法による生存木の構築が困難になる場合がある．

分類木(または決定木)は，標本の二分割を繰り返しながら構成していく統計モデルである．二分割で出来た「木」の内部の枝と，分割の終点に出来た複数のグループ（葉）から樹木状のモデルとなる．分類木の主な目的は，興味のある個体が属するグループを特定し，その個体に関しての予測を行うことである．分類木は，医師が患者の属する疾患のグループを特定し，予後を予測するために有用なツールである．

Breiman ら[1]は分類木を非線形回帰モデルと説明している．二分割を行う基準や，分割の終点を定める基準により完成する木が異なる．２標本検定や尤度比検定が最適な分割ルールを見つけるために利用されている．分類木の概要については[2]を参照されたい．

標本が患者の生存時間の場合，生存データに基づく分類木（生存木）を構成することになる．古くから使用されている分割尺度は，ログランク検定[3,4]，情報量基準[5,6,7]，または Cox 回帰による条件付き検定[8,9]である．現状，R パッケージ *rpart* で実行可能なのは Cox モデルの下での有意性検定を用いる「条件付き推論木」である[10,11,12]).

がんの患者等から得られる多数の遺伝子発現量[13-16]を R パッケージ *rpart* に適用して生存木を構成するときには問題が生じる．１つの遺伝子発現量の効果は，Ｐ値(<0.05)またはＰ値(<0.01)の通常の基準でも遺伝子の効果が有意とみなされる（[17,18,19])が，*rpart* で行われる多重検定を行うと，有意な遺伝子が殆ど検出されなくなる．したがって，多重検定の調節を行わない検定を全ての遺伝子に適用し，有意な共変量を選択する（[14,17,18,19,20,21)．P 値閾値は，0.05，0.01，0.001 などの通常の基準で,癌の生存予測に有用となる.

本発表では上記 *rpart* の問題点が無く，高次元遺伝子の扱いに適したスコア検定に基づく生存木を構築する手法を紹介した．想定している生存時間データは [14-22]に見られる高次元共変量を含むものである．また，提案する「安定化スコア検定」検定統計量の分散を安定化するための縮小法[14]を適用するため，類似の手法であるログランク検定に基づく生存木よりも優れた予測性能を示す．本発表では，高次元遺伝子の扱いに適した「安定化スコア検定」に基づく生存木を構築する手法を紹介した．提案手法は，通常のログランク検定に比べて擬陽性の問題を軽減し，また多重検定の問題も軽減するため，より意味のある生存木を単純なチューニングで構成できる．提案した手法を肺癌患者の生存時間データに適用し，生存木がどのように構築されるかを説明した．

聴講者からいくつかの質問を頂き、本発表は終了した。

参考文献

[1] Breiman L, Friedman JH, et al. (1984). Classification and Regression Trees. New York, US, Chapman and Hall.

[2] Everitt BS, Howell DC (2005) Classification and regression trees, encyclopedia of statistics in behavioral science. Chichester, Wiley, Second Edition, pp. 287-290.

[3] Ciampi A, Bush RS, et al (1981). An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients: 1967-975. Cancer, 47(3): 621-27.

[4] Ciampi A, Thiffault J, et al. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. Computational Statistics & Data Analysis 4(3): 185-204.

[5] Radespiel-Tröger M, Rabenstein T, Schneider HT, Lausen B (2003). Comparison of tree-based methods for prognostic stratification of survival data. Artificial Intelligence in Medicine, 28(3), 323-341.

[6] LeBlanc M, Crowley J (1995) A review of tree-based prognostic models. Cancer Res Treat 75, 113-124.

[7] Bou-Hamad I, Larocque D, Ben-Ameur H (2011). A review of survival trees. Statistics Surveys, 5, 44-71.

[8] Therneau TM, Atkinson EJ (2019) rpart: Recursive Partitioning and Regression Trees. CRAN Ver 4.1-15.

[9] Hothorn T, Seibold H, Zeileis A (2020) partykit: A toolkit for Recursive Partytioning. CRAN Ver 1.2-8.

[10] Hothorn T, Hornik K, Zeileis A (2020) ctree: Conditional Inference Trees. CRAN Version 1.2-8.

[11] Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15: 651-74.

[12] Hothorn T, Everitt BS (2014). A Handbook of Statistical Analyses using R, Third Edition. CRC press.

[13] van Wieringen WN, Kun D, Hampel R, Boulesteix L (2009). Survival prediction using gene expression data: a review and comparison. Comput Stat & Data Anal 53(5): 1590-1603.

[14] Witten DM, Tibshirani R (2010) Survival analysis with high-dimensional covariates. Stat Methods Med Res 19:29-51.

[15] Emura T, Chen YH, Chen HY (2012) Survival prediction based on compound covariate under Cox proportional hazard models. PLoS ONE **7** (10). doi:10.1371/journal.pone.0047627

[16] Emura T, Matsui S, Chen HY (2019) compound.Cox: univariate feature selection and compound covariate for predicting survival. Comput Methods Programs Biomed 168: 21-37

[17] Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8: 816-824.

[18] Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 356: 11-20.

[19] Zhang Q, Wang J, et al. (2020). Weighted correlation gene network analysis reveals a new stemness index-related survival model for prognostic prediction in hepatocellular carcinoma. Aging 12(13), 13502.

Big Data の統計的検定

柳川　堯 (久留米大学バイオ統計センター)

はじめに　Big data は, 様々なデータソースから集積されたデータであり, 多様な要因の影響を受けており、多くの場合, データの信頼性や均一性が保証されていない. 他方, 統計的検定は, 多くの場合, 正規分布が仮定されていることから明らかなように高度に均質化されたデータを対象としている. また, それゆえに想定されるサンプルサイズも大きくない. 多くの統計家にとっては Big data に統計的検定など笑止千万かもしれない. しかしながら, 問題に応じて Big data からデータを切り出し層別し, 背景因子を均一化したうえで統計的検定することには意味がある.

本発表では,

- Big data から取り出された「均質化」されたデータで, しかもサンプルサイズが 500~ 1000 程度の 2 標本比較を想定し, 従来この程度のサンプルサイズの観察研究に適用されてきた統計的検定を批判的に吟味し, Bias を control したうえで効果を検定する新しい検定法を提案した.

- 簡単のため, 2 標本問題に的を絞った. 以下に, その概要を述べる.

1. 従来の検定の誤用

観察データの第一の特徴は, サンプルサイズ n が, あらかじめ所与であることである. にもかかわらず, 疫学や社会科学の広い分野に定着した統計的検定では

- 統計ソフトを利用して p 値を算出し, p 値 $< 5\%$ なら有意と判定する.

注) Neyman-Pearson 流の検定では, 予め意味ある差 δ_0 と 有意水準, および δ_0 を検出する検出力を与えておき, サンプルサイズ n_0 と棄却点 C_0 を定めて検定を行う. p 値による検定は, 従来の検定では, n_0 を n でおき替えて, 有意水準 5%で検定を行うことと対応する.

これは, 明らかに間違った適用である. 理由は, $n > n_0$ ならば, δ_0 より小さな差が n_0 を定めたときと同一の検出力で有意と判定される; いい替えれば, 「Y と X の差が意味をもつ最小の差」と定められていた δ_0 よりも小さな差, つまりゴミ, を「有意な差」と判定する可能性があるからである.

観察データの第二の特徴は, たとえ層別化しても, 比較する 2 群の対象の背景因子が均質化されていないことから生じる Bias の影響を受け, 正当な比較が行われない可能性があることである. 発表では

- Bias が無視できるほど小さい時, 適用可能な妥当な検定法を提案した. この検定は p-値 $< 10^{-12}$, ~ p-値 $< 10^{-8}$ のとき有意差ありと判定する検定である

2. Big data の検定

Big data の特徴は, 例え層別化などによって均質化されているとしても, 均質化には限度があって, Bias の存在は否定できず, 統計的検定による効果の判定が Bias の影響を受ける可能性があることである. 本節では,

- Bias の最大値が特定できることを仮定して, それを θ_{01} で表し, θ_{01} が検出される確率が, 予め与えた小さな値 γ_{01} 以下に保つ判定法を提案した.
- 提案した検定は Bias を抑える検定であるがサンプルサイズが 500~ 1000 もあれば検出力 80%以上で有意差ありとなることを示した.
- 提案した検定法のポイントは Bias の最大値を既知とした点にある.
- Big data の旨味はデータを上手に使えば Bias の最大値を見積もることができることである.
- 多重回帰モデルをデータに当てはめ, Bias の最大値を見積もる一つの方法を提案した.

A Prediction Error under Area Level Model with Arc-Sin transformation for Sample Proportions

Masayo Y. Hirose*, Malay Ghosh† and Tamal Ghosh†
*Kyushu University**, *University of Florida*†

1 Introduction

In small area estimation, it is also important to consider some situations when sample size within each area is not large enough. Such situation may often occur in several research fields.

Suppose it is interested in the analysis of binomial sample proportions. While the sampling variance in a binomial model is a function of the unknown sample mean, the arc-sin transformation is a classical transformation which achieves a known variance. This transformation is used for several real data analyses in Efron and Morris (1975) and Casas-Cordero et al. (2015). Hirose et al. (2018) also considered such transformation model in the analysis of one consciousness survey at the municipality level in Japan.

Sometime it is essential to transform back properly to the original scale to arrive at the final conclusion. However, the natural back transformation could produce a bias especially when sample size within an area is not large enough. Slud and Maiti (2006) addressed such issue with the log-transformation to derive a multiplicative bias correction terms and the estimator of these mean squared prediction errors after the back transformation is made. On the dual power transformation, Sugasawa and Kubokawa (2017) suggested a non-explicit empirical Bayes estimators and the estimators of these mean squared prediction errors.

In this study, for arc-sin transformed data, we find an explicit empirical Bayes estimators in order to correct biases. Moreover, the second-order unbiased estimators of these mean squared prediction errors are obtained explicitly, maintaining strict positivity. This study has been accepted in Statistica Sinica. (Hirose et al. 2021, in the progress of proofreading).

2 Organization of Talk

In this talk, we introduced the explicit empirical Bayes estimators of the original parameters under Fay–Herriot model (Fay and Herriot, 1979) with arc-sin transformed data. And then we showed the asymptotic unbiased estimators of these mean squared prediction errors. Finally, we tried to illustrate an example in predicting the positive rate in PCR testing for COVID-19 for each prefecture in Japan.

Acknowledgements

The first author's research was partially supported by JSPS KAKENHI grant number 18K12758 and the ISM Cooperative Research Program (2018-ISMCRP -2057).

References

[1] Casas-Cordero, C., Encina, J. and Lahiri, P. (2015), Poverty Mapping for the Chilean Comunas, In *Analysis of Poverty Data by Small Area Estimation*, (Edited by Monica Pratesi,). Wiley Series in Survey Methodology.

[2] Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**: 311-319.

[3] Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269-277

[4] Hirose, M.Y., Park, Y. and Tsuchiya, T. (2018). Analysis of a Disaster Prevention Consciousness Survey using a Small Area Explicit Model-based Approach (in Japanese), Journal of the Japan Statistical Society, **48**, 49-70

[5] Hirose, M. Y., Ghosh, M., and Ghosh, T. (2021). Arc-sin transformation for binomial sample proportions in small area estimation. Statistica Sinica, Preprint.

[6] Slud, E. V., and Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B*, **68**, 239-257.

[7] Sugasawa, S., and Kubokawa, T. (2017). Transforming response values in small area prediction. *Computational Statistics & Data Analysis*, **114**, 47-60.

感染症流行記述におけるメタ・ポピュレーションモデルの課題

斎藤　正也・竹内昌平 (長崎県立大)・山内　武紀 (昭和大学)・内田　満夫 (群馬大学)

1　はじめに

2020 年の新型コロナウイルスの流行 (COVID-19) は大都市圏で維持されており，メタ・ポピュレーションモデル (Sattenspiel and Dietz, 1995; Colizza et al., 2006) による流行記述，その予測介入策の応用が見込まれる．しかしこのモデルは多数のパラメータを含みモデルの校正が困難であるという難点もある．本研究ではメタ・ポピュレーションモデルを使っての 2020 年 10 月末までの COVID-19 の国内流行の記述可能性を検討する．そのために各都道府県の流行曲線をモデルに同化することで状態推定を行うとともに，短期間の予測誤差を計測した．これらをもとに流行動態の記述に関する制約を議論する．詳細は斎藤ら (2021) を参照のこと．

2　手法

都道府県 $i = 1, \cdots, 47$ における感受性者数 S_i，感染者数 I_i，除外人口 R_i に関する確率過程で流行を記述する．シミュレーションのタイムステップは利用できるデータ間隔に合わせて 1 日とする ($\Delta t = 1$ 日)．時点 $t-1$ から t の間の人数の変化を $\Delta X_{i,t} = X_{i,t} - X_{i,t-1}$，そのうち $X_{i,t-1}$ から $Y_{j,t}$ へ移動した分を $\Delta[XY]_{ij}$ とすると，この確率過程は

$$
\begin{array}{ll}
\Delta S_{i,t} = -\Delta[SI]_{ii} - \sum_{j\neq i}\Delta[SS]_{ij} + \sum_{j\neq i}\Delta[SS]_{ji}, & \Delta[SI]_{ii} \sim \mathrm{Binom}\left(S_{i,t}, 1-\exp\left(-\frac{\mathcal{R}_{i,t} I_{i,t}\Delta t}{N_{i,t-1} T_{\mathrm{inf}}}\right)\right), \\
\Delta I_{i,t+1} = \Delta[SI]_{ii} - \Delta[IR]_{ii} - \sum_{j\neq i}\Delta[II]_{ij} + \sum_{j\neq i}\Delta[II]_{ji}, & \Delta[IR]_{ii} \sim \mathrm{Binom}\left(I_{i,t}, \Delta t/T_{\mathrm{inf}}\right), \\
\Delta R_{i,t+1} = \Delta[IR]_{ii} - \sum_{j\neq i}\Delta[RR]_{ij} + \sum_{j\neq i}\Delta[RR]_{ji}, & \Delta[XX]_{ij} \sim \mathrm{Binom}\left(X_i, p_{ij}\right) \quad \text{for} \quad X \in \{S, I, R\}.
\end{array}
$$

上記の確率過程にはパラメータとして i 県から j 県への移動確率 p_{ij}，感染性期間 T_{inf}，実効再生産数 $\mathcal{R}_{i,t}$ が含まれている．これらをどのように設定するか，以下で述べる．

パラメータ p_{ij} は，Δt の間に，i 県の居住者の内 j 県に移動するものの割合である．ここでは「旅客地域流動調査」に基づく「府県相互間輸送人員表」のうち鉄道，航空などを含む全交通機関を集計したものに依拠した (国土交通省, 2011)．この表は i 県から j 県の年間輸送人員数 M_{ij} を与える．競合リスクモデル (Gelfand et al., 2000) に従い，最尤推定で決定する経験的な係数 Conn とともに，

$$
p_{ij} = (1-s_i)\frac{\rho_{ij}}{\sum_{j\neq i}\rho_{ij}} \quad \text{ただし} \quad \rho_{ij} = \frac{M_{ij}}{365} \times \mathrm{Conn}, \quad s_i = (1 - \sum\nolimits_{j\neq i}\rho_{ij})^{\Delta t}
$$

と定義した．

実効再生産数 $\mathcal{R}_{i,t}$ は地域内での小規模流行を仮定した予備解析 (次節で述べる) を別途行い，プラグインする．ただし，感染報告がない期間は未定義になるためランダムウォークによる進化

$$
\mathcal{R}_{i,t} = \mathcal{R}_{i,t-1} + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, 0.1^2)
$$

を導入し，確率的な外挿によりアウトブレイクとアウトブレイクの間でも実効再生産数が定義されるようにした．なお，分散 0.1^2 はデータに追随できて，かつ計算の破綻が起きない十分な値を選んだ．

実効再生産数のランダムウォークを追加したメタ・ポピュレーションモデルによる予測と実際の新規感染者数 $J^{\mathrm{obs}}_{i,t}$ との間に

$$
J^{\mathrm{obs}}_{i,t} \sim \mathrm{Poisson}(\Delta[SI]_{ii})
$$

との関係を仮定して構成される状態空間モデルでデータ同化を行う．新規感染者時系列 $J^{\mathrm{obs}}_{i,t}$ にはジャッグジャパン株式会社 (2020) の提供による令和 2 年 1 月 15 日〜11 月 11 日の毎日の都道府県別の報告 (300 時点) を用いる．

地域毎の実効再生産数の予備推定　単発の小規模流行を記述するモデルを使って，地方都市でも流行があるかぎり実効再生産を割り充てる．ここで，小規模流行とは T_{inf} 以上続く 0 人報告を区切りとする連続する感染者の報告列である．はしか・風しんのクラスタ分析に Yoshikura and Takeuchi (2016) が用いた方法を参考にした．このように定義した感染報告列 $J_{i,t}$ が負の二項分布に従うと仮定するとして実効再生産数を推定する:

$$
J_{i,t} \sim \mathrm{NegBin}\left(\mu = \frac{I_{i,t}\mathcal{R}_{i,t}\Delta t}{T_{\mathrm{inf}}}, \sigma^2 = \mu + \frac{I_{i,t}\sigma_0^2\Delta t}{T_{\mathrm{inf}}}\right).
$$

3　結果

データに見られる流行動態をモデルが再現できることを確認するためにデータ全体を同化し，フィルタ分布の平均を図 1 に示す．この図は人口規模の異なる 6 都道府県を代表として取り出したものである．大規模の東京 (ピーク時，約 300) から小規模の長崎 (同 10) からまで，規模の異なる流行が概ね再現できている．他方，細部を見ると小都市でのマイナーアウトブレイクがシミュレーションでは見逃されていることも確認できる．

感染期間を 4 日より長く取ると，データの再現性が悪化することも確認された．同じく図 1 では $T_{\mathrm{inf}} = 5, 8$ 日 とした場合のフィルタ分布の平均値も示しているが，2020 年 7 月中旬以降の東京での上昇トレンドに追随できていない．SIR モデルでは感染期間の設定によらず感染者数 $I(t)$ のスケールを変えることでほぼ等しい新規感染者系列を生み出せることが知られていることを踏まえると，この結果が感染期間を制約しているのではなく，式 (2) による再生産数の簡易推定のために適切でない再生産数が選択されていることが原因とと予想される．この式では 1 世代 (=感染期間) の間，再生産数は定数としているので，感染期間が長くない場合にはデータを平滑化する効果が入ることになる．

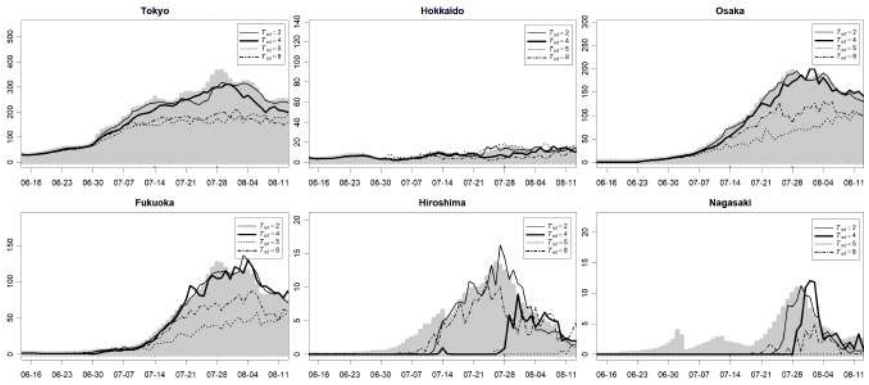


図 1: 斎藤ら (2021) より再掲．6 都道府県のデータ同化結果．棒グラフ:観察データ, 線:フィルタ分布の分位点 (2.5%,50%,97.5%).

SIR モデルには集団免疫の形成以外には感染性を変化させる過程が含まれていないため予測性能は概して低くく，その定量化が困難である．そこで，上昇トレンドが継続する区間を抽出して 2 週間の予測誤差を計量した．小規模流行ほど偶発性が高いため，予測誤差が大きくなる．特に，誤差が 100%以内であるのは，北海道，東京，埼玉，大阪，京都のみである．

4　おわりに

本報告では，確率的メタ・ポピュレーションモデルとよばれる地域毎の確率的 SIR モデルを人の流入に基づいて連結したモデルを用いて，COVID-19 の日本での令和 2 年 10 月までの流行状況を分析した．インフルエンザのような感染者数がどの地域でも連続した流行が見られる感染症と異なり，モデルに含まれる力学パラメータを十分に制約するだけの情報が得られないという問題があった．ややアドホックではあるが，可能な限り観察データから情報を抽出するために，今回の解析では地域毎の実効再生産数の事前推定値を使う方法を取った．完全にベイズ的な方法と比較すると不確かさの評価にバイアスが入るなどの問題もあるが，確率的メタ・ポピュレーションモデルのようなフレキシビリティの高い力学モデルを限られた計算資源で活用する場合には現実的な方法と考えられる．

参考文献

Sattenspiel, L., Dietz, A. (1995). Structured epidemic model incorporating geographic mobility among regions, Mathematical Biosciences, 128: 71-91. PMID: 7606146

Colizza, V., Barrat, A., Barthe'lemy, M., Vespignani, A. (2006). The modeling of global epidemics: Stochastic dynamics and predictability, Bulletin of Mathematical Biology, **68**, 473-481.

Yoshikura, H., Takeuchi, F. (2016). Measles and Rubella: Scale Free Distribution of Local Infection Clusters, Jpn. J. Infect. Dis., **69**, 293-299.

Nishiura, H., Yan, P. Sleeman C.K. and Mode, C.J. (2012). Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks, *J Theor Biol*, **294**, 294:48-55. doi: 10.1016/j.jtbi.2011.10.039

国土交通省 (2011). 旅客地域流動調査–府県相互間旅客輸送人員表（全機関）. `https://www.e-stat.go.jp/stat-search/file-download?statInfId=000027669878&fileKind=0`

Saito, M.M., Nishiura, H., Higuchi, T. (2018). Reconstructing the transmission dynamics of rubella in Japan, 2012-2013, *PLOS ONE*, **13**, 10:e0205889. doi: 10.1371/journal.pone.0205889

Gelfand A.E., Ghosh S.K., Christiansen C., Soumerai S.B., McLaughlin T.J. (2000). Proportional hazards models: a latent competing risk approach, *Appl Stat*, **49**:385-397.

ジャッグジャパン株式会社 (2020). 都道府県別新型コロナウイルス感染者数マップ. `https://gis.jag-japan.com/covid19jp/`

Ali, S.T., Wang, L., Lau, E.H.Y., Xu, X.K., Du, Z., Wu, Y., Leung, G.M., Cowling, B.J. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions, *Science*, **369**, 1106-1109

斎藤, 竹内, 山内, 内田 (2021). COVID-19 流行動態の再構成によるメタ・ポピュレーションモデルの記述性能評価, 統計数理, 掲載予定

日本人を対象とした大規模 SNP データ研究の近年の成果と課題

名古屋大学大学院医学系研究科
総合保健学専攻　ヘルスケア情報科学分野
実社会情報健康医療学
中枋　昌弘

DNA の存在が明らかになって以降、疾患・体質に寄与する(関連する)遺伝的要因の探索が試みられ続けてきた。近年の測定技術の進歩の結果、ヒト DNA から全ゲノム中の一塩基多型(Single nucleotide polymorphism, SNP)を容易かつ安価に測定できるようになり、疾患・体質関連 SNP の探索研究に目覚ましい発展をもたらした。このような研究はゲノムワイド関連研究(Genome-wide association study, GWAS)と呼ばれ、最初期の GWAS から 10 年以上経過した現在も行われ続けている。

GWAS が行われるようになって数年後、GWAS のメタ解析(GWAS メタ解析)も実施されるようになってきた。GWAS メタ解析は、研究に参加する各グループが自身の手持ちのデータで GWAS を実施し、その解析結果を中央解析チームに提供、中央解析チームがメタ解析を実施して単独チームでは成し得なかった大規模な GWAS を実現するというアプローチである。このアプローチの利点は、研究グループ間で個々人の SNP データをやり取りする必要がない点である。そのため倫理面での障壁が低いことから多数の研究グループが参画できる。結果、大規模なサンプルサイズで GWAS を実施でき、低エフェクトサイズな SNP の検出が可能になる。一方で、各グループで解析を行う必要があるため、複雑な解析を行うことは容易ではないという欠点も挙げられる。これらの欠点を踏まえた上で GWAS メタ解析を遂行する必要がある。

このように SNP データの蓄積が進み、GWAS メタ解析も実施されるようになった結果、GWAS の規模は年々増大している。最初期の GWAS は数百例規模だったのに対し、近年の研究は日本人のみを対象とした GWAS であっても数万〜数十万例、複数人種を対象とした人種横断型 GWAS は数百万例の規模までに及んでいる。結果、低エフェクトサイズな SNP の検出に加え、従来実施が容易ではなかった SNPxSNP 交互作用や SNPx 環境因子交互作用の検出も可能になりつつある。更に GWAS の研究規模の拡大に加えて、新たな遺伝統計学的手法も開発されてきた結果、疾患や体質関連 SNP の探索以外にも、メンデルランダム化による因果推論、遺伝子レベルに SNP を集約する解析(gene-based analysis)等、従来の SNP 研究の枠組みを超えたゲノムデータ解析が可能となり、ゲノム研究の更なる進展が期待される。

複数人種を対象とする人種横断型 GWAS は、検出力の増加するため低エフェクトサイズ SNP の探索や人種間の比較が可能なことから、近年大規模**日本**プロ**人**ェクトによる実施が

をんである。その一方で、日本人な**対**の**象**定人種を対象とした GWAS を**とし**そかにすることもできない。な**た**ならば、参**大規模** 1〜**デ**のように、日本人(**ー**は**タ**ア**人**ア人)**象研**的な SNP の**究**定やメ**の近**ズムの探索を行うことができると可能**年**があるからである。

以上のように、人種横断的 GWAS、人種**象研**的 GWAS の**成**方が**果**用である。**課**後も、**成**アプローチを**題**行して行うことで、遺伝要因と形質間の新たな関連が次々と発見されることだ**し**う。

参**大規模**：

1. Spracklen CN, Horikoshi M, Kim YJ, Lin K, Bragg F, Moon S, et al. Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature. 2020;582: 240-245.
2. Lin Y, Nakatochi M, Hosono Y, Ito H, Kamatani Y, Inoko A, et al. Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. Nature communications. 2020;11: 3175.
3. Koyanagi YN, Nakatochi M, Ito H, Kasugai Y, Narita A, Kawaguchi T, et al. Genotype-stratified GWAS meta-analysis reveals novel loci associated with alcohol consumption. medRxiv 2021.06.02.21258094

Desire weight learning with class-variant loss for Imbalanced segmentation

Muthu Subash Kavitha[a], Novanto Yudistira[b], Takio Kurita[c]
[a]School of Information and Data Sciences, Nagasaki University, [b]Intelligent System laboratory, Faculty of Computer Science, Brawijaya University, Indonesia, [c]Graduate School of Advanced Science and Engineering, Hiroshima University

Introduction

Class imbalanced datasets occur in many real-world applications where the class distributions of data are highly imbalanced. When dealing with an imbalanced data, the minority class is typically of the most interest. It is more challenging for a model to learn the characteristics of examples of minority class, and to differentiate this class from the majority class. The imbalanced learning problem has drawn a significant amount of interest from academia, industry, and government funding agencies. Most specifically, the data imbalance problems in the medical community often leads wrong prediction results[1]. Thus, analyzing imbalanced multi-class data is a significant challenge in the medical diagnosis. Conventional cost-insensitive learning algorithm could perform well on binary classification problem, and they might be biased and ignore the importance of the class of interest in multi-class prediction tasks. Hence in this study, the data imbalance problem is addressed on the medical image thyroid cancer dataset.

Re-weighting and class variant mechanism

As an end-to-end network training, we utilize a Unet-like convolution neural network (CNN) model. The incomplete image features of patch-based training are not appropriate for the multi-class imbalanced data set segmentation, we feed whole image as input into the segmentation network. Furthermore, in multi-class joint object segmentation, the data sets that do not consist of all occurrences of classes, poorly influences the activation in the following layers and sometimes it leads to zero gradient updates in the back-propagation process. Hence, additional weighting mechanism is needed to increase the instances of underrepresented classes[2]. To improve the co-occurrence of classes and control the false positive regions, we proposed a re-weighting and class variant mechanism. For reweighting, we considered the number of pixels of the foreground relative to the total number of pixels of all the foreground classes.

The larger proportions of the cancer datasets either consist of only one occurrence of the remnant tissue or no occurrence of classes. The lymph node occurrences were very limited. The weights of lymph node (w_l) and remnant tissue (w_r) is defined as,

$$w_l = \frac{f_l}{f_l+f_r} \ , \ w_r = \frac{f_r}{f_l+f_r} \qquad (1)$$

Where f_l and f_r are number of pixels of the foreground classes of the lymph node and remnant tissue, respectively. The total number of pixels of all the ground truth classes can be represented as $N_f = f_l + f_r$.

Suppose if the total number of pixels of either f_l or f_r is zero, then the weights w_l or w_r becomes zero. Consider if the weights of the entire image pixels are,

$$w_l + w_r = 1 \qquad (2)$$

then the weights can be calculated as,

$$w_l = 1 - w_r \;\; ; \;\; w_r = 1 - w_l \qquad (3)$$

The class weighting of each individual class is calculated by using the summation of the weights of the entire image by multiplying the square root of all its inverse class frequencies. By substituting Eqs (1) and (3) we get the following reweighting term,

$$RW = w_l \sum_{i=1}^{f_r} (S_r - P_r)^2 + w_r \sum_{i=1}^{f_l} (S_l - P_l)^2 \qquad (4)$$

The parameters *S* and *P* represents ground truth and predicted segmentation, respectively. Furthermore, to reduce the number of false positives the reweight term is unified with our newly introduced class-variant dice loss. It not only forces the inner joint intersection but also minimizes the outer joint which maximizes the number of true positives while minimizing the number of false positives. Hence, it is straight forward in creating flexibility and balancing of both false positives and false negatives. The class-variant dice loss function is defined as,

$$L_{CD} = 1 - \frac{2\sum_{i=1}^{I}\sum_{j=1}^{J} S_{ij}P_{ij} + m}{\sum_{i=1}^{I}\sum_{j=1}^{J} S_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} P_{ij} + m} + \frac{2\sum_{i=1}^{I}\sum_{j=1}^{J} (S_{ij} - P_{ij})^2 + m}{\sum_{i=1}^{I}\sum_{j=1}^{J} S_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} P_{ij} + m} \qquad (5)$$

where s_{ij} and p_{ij} represent ground truth and prediction at each pixel with *i* row and *j* column, respectively. To prevent division by zero, the smoothing parameter *m* is added to both the denominator and numerator. Eqn (5) is twofold objectives. On the one hand, the right term of the equation aims at extracting the patterns of activation, associated with high confidence related to true location. On the other hand, the left term of the equation explains the common activation between the predicted and ground truth masks, which resembles the dice loss. The multi-class L_{CD} is calculated as the accumulation of all classes using the following equation.

$$L_{CD} = \sum_{k}^{K} L_{CD_k} \qquad (6)$$

where *k* indicates output class channel for each class. The parameter *K* represents the total number of classes. The unified reweighting and class variant objective function is used to train the network to learn the uncertain pixel regions of each class. It computes the summation of class weight multiplied with L_{CD}, which is defined as

$$L_{RWCD} = \sum_{k}^{K} RW_k L_{CD_k}$$

Additionally, we set the network to train and control the uncertainty over batches of images using the stratified sampling method. Likewise, the proposed mechanism makes the model efficient by maximizing appropriate weights regarding lower instance classes and, at the same time, it can control the model parameters from getting stuck in local minima. Hence it can efficiently handle the imbalanced segmentation over the whole image.

References

1. H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009
2. H. Phan, M. Krawczyk-Becker, T. Gerkmann & A. Mertins. DNN and CNN with weighted and multi-task loss functions for audio event detection. *ArXiv* abs/1708.03211 (2017)

深層学習により生成した脳画像は解析に使用できるのか

田尻涼[1]、　川口淳[2]
[1]佐賀大学大学院医学系研究科、[2]佐賀大学医学部

1. 初めに

近年の脳画像解析において，精度の向上のために，複数のモダリティより得られた画像を統合解析するマルチモダル脳画像解析が行われる[1]．しかし，マルチモダル脳画像解析では撮影モダリティが増えるにつれ，被験者の負担，医療コストなどの問題より，すべての画像が揃った complete dataset (CD)の取得は難しくなる．多くの場合，一部のモダリティ画像に欠損が生じた incomplete dataset (ID)が得られ，全ての画像が揃っていない被験者を削除した dataset (listwise deletion dataset LDD)での解析が行われる．リストワイズ削除は直感的で操作が簡単であるが，欠損の割合が大きい場合，サンプルサイズの減少を招き，バイアスを混入させる恐れがあり，信頼性に問題がある．この問題の解決のために，本研究では敵対的生成ネットワーク(generative adversarial networks: GAN)[2]を用いて欠損画像を生成，補完を行い，疑似完全データ(pseudo-complete dataset: PCD)を作成，これを解析に用いることを提案する．本研究の目的はこの PCD と LDD で解析した場合のどちらの方が疾患の判別を精度よく解析できるか検討を行うことである．

2. 方法

MRI 画像から PET 画像を生成する GAN として Pix2Pix[3]と CycleGAN[4]を使用し，14 例の画像で MRI 画像を PET 画像へ変換するように学習を行った．マルチモダル脳画像解析として Alzheimer's disease (AD)の判別のため，100 例の MRI 画像と PET 画像の CD を使用し，multi-block sparse multivariate analysis (MSMA)法[5]を行った．加えて，CD よりランダムに 50 例或いは 80 例の PET 画像のみを欠損させた ID に対し，LDD と Pix2Pix 或いは CycleGAN にて MRI 画像より生成された PET 画像で欠損を補完した PCD を作成した．この LDD と PCD に MSMA 法を実行した．50 例の検証用のデータにおいて LDD と PCD の MSMA 法の結果を用いて AD を予測した．この判別性能を LDD と PCD それぞれで receiver operating characteristic 曲線の曲線下面積 (area under the curve: AUC)を計算し，LDD と PCD の比較に DeLong's test を用いた．データは Alzheimer's Disease Neuroimaging Initiative より PET 画像と MRI 画像の両方を持つ患者データを用いた．

3. 結果および考察

初めに Pix2Pix と CycleGAN によって生成された画像を図 1 に示す．CD を MSMA 法で解析し，5 スコア抽出し，このうち PET の寄与率が 79%と最も高かった第 2 スコアについて着目した．このスコアの AUC は 0.77 であった．欠損が 50 例の場合，LDD で求めた第 2 スコアの AUC は 0.775 であった．Pix2Pix 或いは CycleGAN での AUC はそれぞれ 0.735, 0.747 (p=0.38, 0.53)，80 例の場合，LDD が 0.615, PCD がそれぞれ 0.661,0.528 (p=0.08, 0.04)であった．これ

より，LDD のサンプルサイズが 20 例以下のように極めて小さくなるとき，Pix2Pix を用いた PCD の 100 例で解析を行うのは有用である可能性がある．一方，同条件で CycleGAN を用いた PCD は LDD より劣る結果となった．この差異は Pix2Pix と CycleGAN の画像生成の精度の差（Pix2Pix の peak signal to noise ratio は 26.7, CycleGAN は 24.7)であると考えられた．

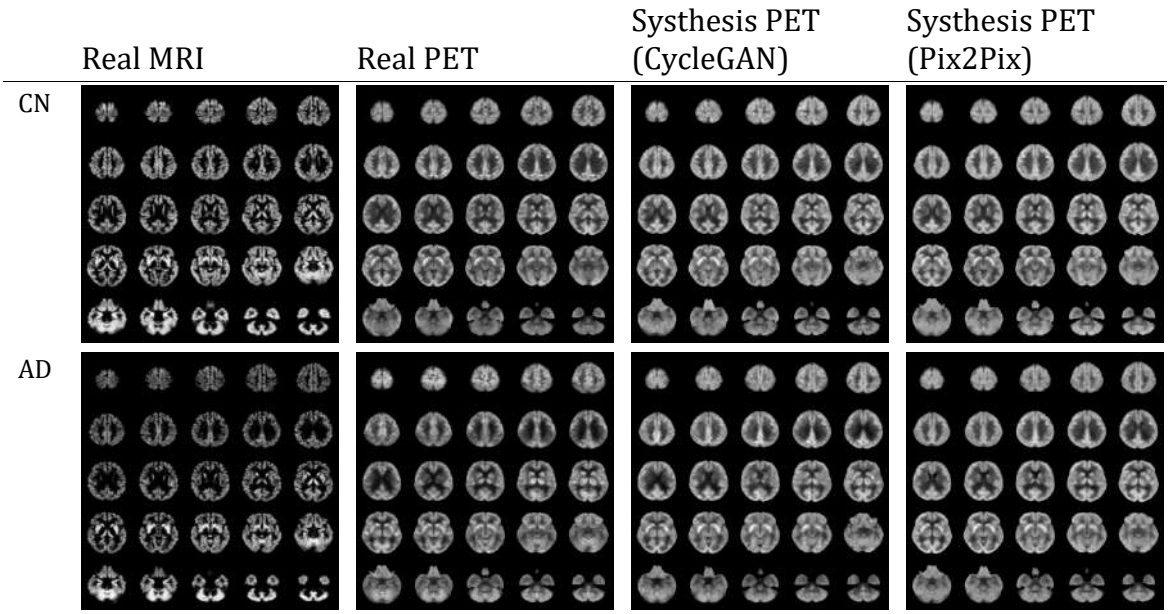


図 1: GAN による画像生成の結果．左の列より本物の MRI, PET 画像，本物の MRI 画像より CycleGAN，Pix2Pix を用いて生成された PET 画像を表示した．上段は健常者，下段は AD 患者のものである．

4. 結論

PET 画像の欠損を LDD の 20 例とするよりも, Pix2Pix による生成画像を用い，PCD の 100 例で解析した方が AD の判別性能が高い傾向が見られた．

参考文献

1. Calhoun, Vince D, and Jing Sui. 2016. “Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link (S) in Complex Mental Illness.” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 1 (3): 230–44.

2. Goodfellow, Ian J. *et.al.* 2014. “Generative Adversarial Networks.” http://arxiv.org/abs/1406. 2661.

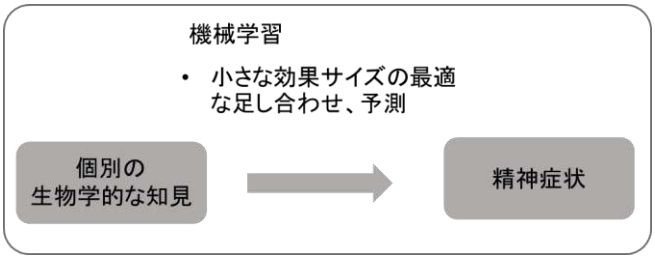
3. Isola, Phillip, *et.al.* 2017. “Image-to-Image Translation with Conditional Adversarial Networks.” http://arxiv. org/abs/1611. 07004.

4. Zhu, Jun-Yan, *et.al.* 2017. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.” http://arxiv. org/abs/1703. 10593.

5. Kawaguchi, Atsushi. 2018. “Supervised Sparse Components Analysis with Application to Brain Imaging Data.” *Neuroimaging-Structure, Function and Mind*. IntechOpen.

精神医学データに対する機械学習的アプローチ

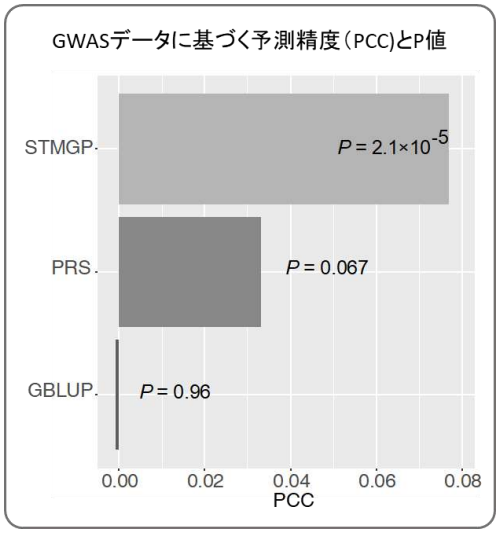
東北大学病院精神科　高橋雄太



精神科医療においては、現在面接のみに診断や治療方針が依存しており、診断の観察者間での一致率の低さが問題となっている。精神疾患を評価するための客観的な指標として、バイオマーカーの開発が期待されているが、精神疾患のバイオマーカー研究では、個別のマーカーの効果サイズが小さく、ノイズと真の効果の区別がしづらいことや、マーカー間に複雑な関係性があるといった課題が知られている。今回は、我々が行っている、統計的・機械学習的アプローチで、生物学的研究の知見と実際に観察される精神症状のギャップを埋めるための研究を紹介する。

【ゲノムワイド遺伝子多型データからのうつ症状の予測】

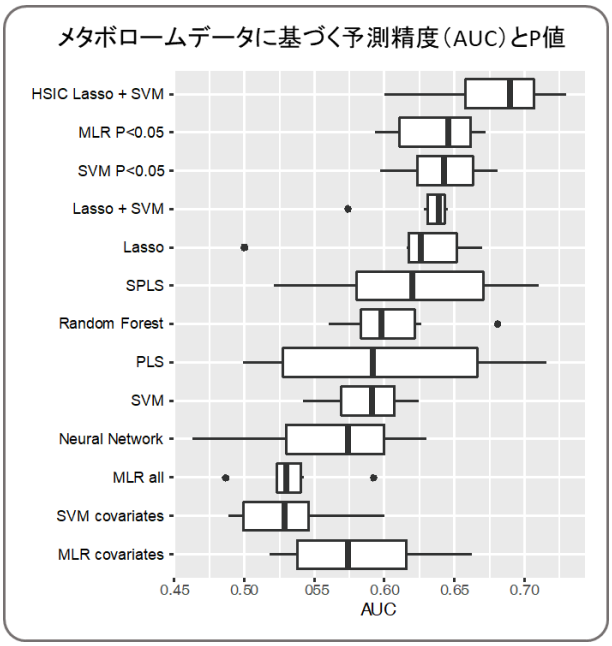
まずはゲノムワイド関連解析（GWAS）データに基づいたうつ症状の予測について紹介する。近年、精神疾患と関連する遺伝子多型が多数報告されているが、個別の遺伝子多型の効果サイズは小さいこともあり、既存の方法では過学習に陥ってしまい、テストデータではあまり有意な予測精度は示されていない。我々は、過学習を抑えることによって、予測精度を上げるという戦略のもと、共同研究者が開発した STMGP 法を用いて、うつ症状を遺伝子多型データから予測する研究を行った。STMGP 法は、機械学習（罰則付き回帰）の手法を用いて、予測精度の向上を目指した予測モデルである。結果は、右図のように、STMGP 法は、頻用されているポリジェニックリスクスコア法や線形混合モデル法（GBLUP 法）よりも、過学習を抑えることによって、高い予測精度を示し、効果的に遺伝子多型の効果を組み合わせることに成功した(Takahashi Y, *et al. Transl Psychiatry* 2020)。その一方、予測精度は臨床応用できる高さではなく、将来的には、遺伝子間・遺伝子環境間の相互作用をモデルに含めることや、他の次元の情報を加えることでさらに予測精度を上げることが期待される。



【血漿中代謝物データからのうつ症状の予測】

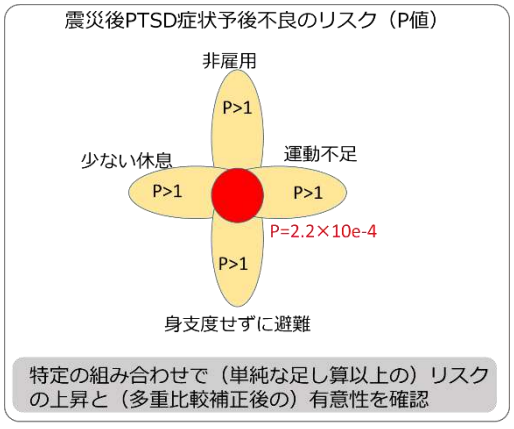
次に血漿中の代謝物データからうつ症状を予測する研究について紹介する。最近の研究より、血漿中の代謝物のうち、うつ症状と関連するものが複数報告されている。しかし、個別の代謝物の効果サイズは小さく、ノイズと区別がしづらく、さらに、代謝物間や代謝物と表現型の間に非線形の関係性が指摘されている。我々は非線形性も考慮して変数を選択する手法を用いることで、既存の手法よりも高い予測精度を示せないかと仮説を立てた。そこで、HSIC 統計量を用いて非線形性も考慮して変数を選択する手法

（HSIC Lasso 法）で代謝物を選択した後、サポートベクトルマシン（SVM）で予測する方法を用いた。この手法では、変数選択をしない SVM や Lasso 法よりも高い予測精度を示すことに成功した(Takahashi Y, *et al. Transl Psychiatry* 2020)。今後は、この予測に使われた代謝物がうつ症状に関連する明らかにし、さらに予測精度を上げるために、遺伝子多型などの情報をモデルに入れることを検討している。



【リスク間相互作用を考慮した疫学的リスク因子の組み合わせの検討】

東日本大震災の体験により心的外傷後ストレス障害（PTSD）が生じた方の中には、症状が遷延している方も多いことが分かっている。症状が遷延するリスクが同定されれば、災害後に効率的に社会的な資源を導入することが可能となる。しかし、症状遷延のためのリスク因子は複数提唱されているが、どれも効果サイズは小さく、実際には多くの人が複数のリスク因子を抱えているのに、リスク因子間の相互作用についてはこれまで十分に調べられてこなかった。今回我々は、PTSD の症状を遷延させるような、リスク因子の組み合わせを網羅的に検討した。網羅的に組み合わせの検討を行う場合、組み合わせのパターンが膨大な数になるため、古典的な統計手法では、多重比較補正が厳しすぎたり、計算が不可能となるといったりした問題が生じる。そこで、我々は、パターンマイニングを用いた機械学習手法（MP-LAMP 法）を用いて、family-wise error rate を適切に調整しつつ、並列計算で解析を行った。解析の結果、個別のリスク因子では、PTSD の症状遷延と有意に関連するものはなかったが、組み合わせを検討すると、56 通り（15 因子）が有意に PTSD 症状遷延と関連があることが分かった。さらに、これらの組み合わせを個別に検討すると、個別のリスク因子と比較して、組み合わせたリスク群では効果サイズにも大きな変化が見られていた。 (Takahashi Y, *et al. Sci Rep* 2020)



Reference

1. **Takahashi Y**, *et al*. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Transl Psychiatry* 2020; **10**(1)**:** 157.
2. **Takahashi Y**, *et al*. Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. *Transl Psychiatry* 2020; **10**(1)**:** 294.
3. **Takahashi Y**, *et al*. Machine learning to reveal hidden risk combinations for the trajectory of posttraumatic stress disorder symptoms. *Sci Rep* 2020; **10**(1)**:** 21726.