科研費（基盤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）

予測モデリングの理論と応用

開催責任者：小森　理（成蹊大学）・深谷　肇一（国立環境研究所）
日時：2021 年 10 月 22 日（金）～24 日（日）
場所：成蹊大学 6 号館 401 室（ハイブリッドで実施）

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：2021 年 9 月 24 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：2021 年 10 月 8 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

10 月 22 日（金）

特別講演
13:00-13:50 岡村　寛(中央水産研究所)
水産資源データと予測モデル

13:50-14:25 Dou Xiaoling(早稲田大学データ科学センター)
Estimation of B-spline copula

14:25-15:00 米岡　大輔(聖路加国際大学)
予測モデルのメタアナリシス

15:00-15:20 休憩

15:20-15:55 大前　勝弘(国立循環器病研究センター)
マジョリティーに対する治療効果のロバストな予測

15:55-16:30 三角　俊裕*（横浜市立大学）, 三枝　祐輔（横浜市立大学), 原　みゆひ（横浜市立大学), 藤田　豊（横浜市消防局), 古場　亮介(横浜市消防局）
横浜市における救急出場件数に対する関数動的予測モデリング

16:30-17:05 得丸　久文(カラハリプロジェクト)
概念の複雑性・信頼性予測のための受け入れ検査(デジタル言語学)

10 月 23 日（土）午前

特別講演
10:00-10:50 江口　真透(統計数理研究所)
予測のための最大エントロピーと最小ダイバージェンス

10:50-11:25 野津　昭文(静岡県立静岡がんセンター)
臨床研究における統計家の役割

11:25-12:00 岡山　大成*(千葉大学融合理工学府), 内藤　貫太(千葉大学理学研究院)
ガンマクレームによる資産過程における破産確率の漸近近似

12:00-13:00 お昼休み

特別講演
13:00-13:50 久保田　康裕(琉球大学, 株式会社シンクネイチャー)
生物多様性ビッグデータと地球環境保全

13:50-14:25 楠本　聞太郎(九州大学 農学研究院)
生物分布の時空間予測：機構論的モデル構築に向けた課題

14:25-15:00 城田　慎一郎(明治大学商学部)
Preferential Sampling に関する最近の話題について

15:00-15:20　休憩

15-20-15:55 矢島　豪太*(日本大学生物資源科学部), 中島 啓裕(日本大学生物資源科学部)
近似ベイズ計算による野生動物の集団サイズ推定 - 生成過程が複雑なデータにどう対処するか？ -

15:55-16:30 深谷　肇一(国立環境研究所)
環境 DNA による生物多様性評価のための階層モデルとベイズ決定分析

特別講演
16:30-17:20 島津　秀康(Loughborough University)
生物多様性変化の定量化とその理解

10 月 24 日（日）

特別講演

9:30-10:20 島谷　健一郎(統計数理研究所)

定量的生態学と統計モデル：食われる側からみた特定食う種の貢献度

10:20-10:55 川森　愛(統計数理研究所)

ニワトリ雛の採餌行動解析ー理論最適解から乖離する理由を探すー

10:55-11:30 永井　勇(中京大学 教養教育研究院)

ランク落ちでの無罰則推定法と不偏推定量

11:30-11:55 小森　理(成蹊大学)

A unified formulation of k-means, fuzzy　c-means and Gaussian mixture model by the Kolmogorov-Nagumo average

水産資源データと予測モデル

水産研究教育機構　水産資源研究所　漁業情報解析部
岡村　寛

海洋生物の再生産関係（親子関係）データは，不確実性が大きいことで知られ，再生産関係の推定には頑健推定と呼ばれる方法が使用されることがある（Chen and Paloheimo 1995, Shertzer and Prager 2002）．我が国の水産資源でもいくつかの種（系群）に対して頑健推定法が使用されている．一方で，再生産関係データは時系列データであり，自己相関の推定が重要である．頑健推定は外れ値のような“良くない”データに「鈍感」である必要があり，自己相関推定はデータ間の関係性に基づくので他のデータに「敏感」である必要がある．つまり，2 つの一見矛盾した問題への対応が必要となるが，その両方をうまく取り扱う頑健推定法の開発はこれまで十分になされてこなかった．そこで，頑健推定法を拡張し，外れ値に強く，かつ自己相関の正確な推定も可能となる推定法を考案した（Okamura et al. 2021）．

加入尾数 R の対数値を $r = \log(R)$と書き，親魚量を S で書くとすると，加入尾数は再生産関係 $f(S|\theta)$の周りの誤差 ε をあわせて

$$r_t = f(S_t|\theta) + \varepsilon_t$$

で与えられる（$t = 1, \ldots, T$）．このとき，残差 ε_t は，

$$\varepsilon_t \sim N\left(\rho\sqrt{\lambda_{t-1}}\varepsilon_{t-1}, \sigma^2/\lambda_t\right)$$

という形であるとする．ここで，λ_t を，

$$\lambda_t = \exp(-\phi\varepsilon_t^2)$$

と定義する．$1–\lambda_t$ は，各データが外れ値であるかどうかの確率に対応し，上の自己相関の式から，λ_t が小さくなると（外れ値である確率が大きくなると）自己相関に対する寄与が小さくなることになる．逆に，λ_t が大きい場合は，自己相関への寄与が大きくなる．$\phi \to 0$とすれば，通常の AR(1)モデルとなる．また，この確率分布は，ρ=0 とすると，Tukey の biweight 関数と似たような形になる．

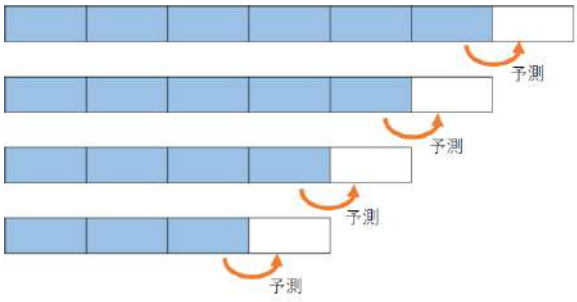


図 1. 時系列クロスバリデーションの模式図．1 年ずつデータを除いたセット（青色部分）で予測モデルを構築し，除かれたデータ（白色部分）を予測する

ϕは外れ値の程度を決めるパラメータであり，各データが再生産関係推定に大きく寄与するものか，外れ値である可能性が高いのかを決定するための鍵となるものである．そこで，ϕパラメータを加入尾数に対する時系列クロスバリデーションによって決定してやることにする（図 1）．ϕ以外のパラメータを最尤推定によって求めたとき，最もクロスバリデーションエラーが小さいϕの結果を選択する．

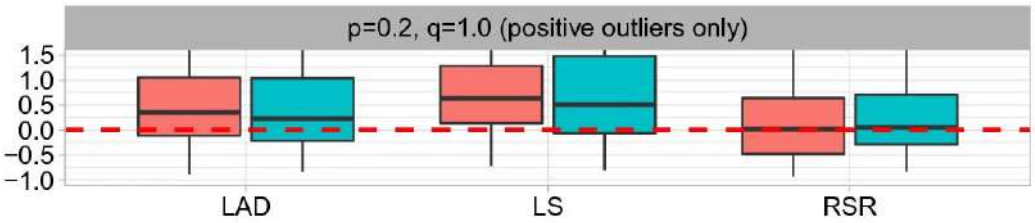


図 2. シミュレーション結果の一例．大きな外れ値があり，自己相関も高い（ρ=0.8）の場合．赤いボックスは長期加入予測，緑のボックスは短期の加入予測に対応する．最小絶対値法（LAD）と最小二乗法（LS）はバイアスした結果となるが，新しい頑健推定法（RSR）は不偏に近い推定ができている

大きな外れ値や自己相関を想定したシミュレーションによって，加入量の短期・長期両方の予測の観点で新しい方法は従来の方法より良いパフォーマンスを持つことが示され（図 2），実データに適用した結果も変わる可能性があることが示唆された（図 3）．理解志向型のモデルと応用志向型の統計手法（江崎 2020）を効果的に組み合わせることが，水産資源データの解析と理解には不可欠なことである．頑健推定と自己相関推定の問題は，再生産関係推定にとどまるものではなく様々なデータ解析に現れるものであることから，新手法は広い応用範囲を持つものと考えられる．

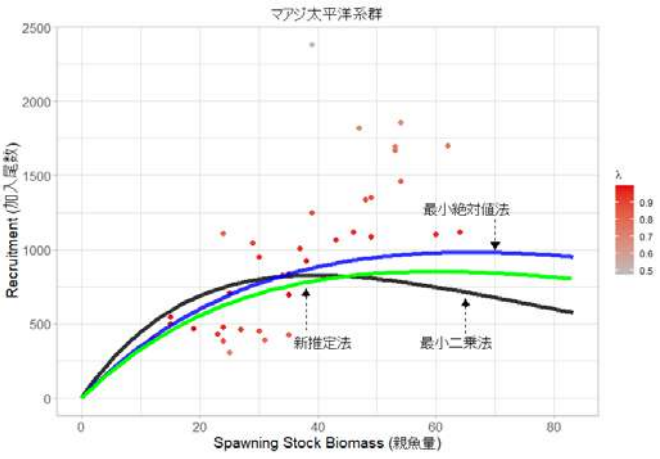


図 3. マアジ太平洋系群に新しい頑健推定モデルを適用した結果

引用文献

Chen, Y. and Paloheimo, J. E. 1995. A robust regression analysis of recruitment in fisheries. Can. J. Fish. Aquat. Sci. 52, 993–1006.

江崎　貴裕 2020. データ分析のための数理モデル入門 本質をとらえた分析のために．ソシム．

Okamura, H., Osada, Y., Nishijima, S., and Eguchi, S. 2021. Novel robust time series analysis for long-term and short-term prediction. Scientific Reports 11: 11938.

Shertzer, K. W. and Prager, M. H. 2002. Least median of squares: a suitable objective function for stock assessment models? Can. J. Fish. Aquat. Sci. 59, 1474–1481.

Estimation of B-spline copula

Xiaoling Dou*, Satoshi Kuriki†, Gwo Dong Lin‡, Donald Richards§

B-spline copula is constructed with B-spline basis functions. It is a generalization of the Bernstein copula, and its form is similar to that of the Bernstein copula (Dou, et al., 2021). For given $p + 2d + 1$ knots

$$s_{-d} = \cdots = s_0 = 0 \le s_1 \le \cdots \le s_{p-1} \le 1 = s_p = \cdots s_{p+d},$$

we can generate $d + p$ (=: n) B-spline functions on $[0, 1]$ of degree d,

$$N_i^d(s), \quad i = -d, \ldots, -1, 0, 1, \ldots, p-1.$$

For any $u, v \in [0, 1]$, the bivariate B-spline copula density is defined as

$$c(u, v; R) = \sum_{k=1}^{m} \sum_{l=1}^{n} r_{kl} \phi_{k,m}(u) \phi_{l,n}(v) \tag{1}$$

where

$$r_{kl} \ge 0, \quad \sum_{l=1}^{n} r_{kl} = q_{k,m}, \quad \sum_{k=1}^{m} r_{kl} = q_{l,n},$$

and

$$\phi_{k,m}(s) = \frac{1}{q_{k,m}} N_{k-d-1}^d(s), \quad q_{k,m} = \int_0^1 N_{k-d-1}^d(s) ds.$$

When $m = n$, it is known that the maximum correlation of the B-spline copula $c(u, v; R)$ is attained when

$$R = (r_{kl}) = \mathrm{diag}(q_{k,n})_{1 \le k \le n},$$

and we denote the B-spline copula with the maximum correlation as

$$c_n^+(u, v) = \sum_{k=1}^{n} q_{k,n} \phi_{k,n}(u) \phi_{k,n}(v), \quad u, v \in [0, 1]. \tag{2}$$

Suppose that we have a set of i.i.d. data (x_t, y_t), $t = 1, \ldots, N$. We consider methods of estimating the joint density by using the B-spline copula. As the B-spline copula has a similar form to that of the Bernstein copula, the EM algorithm methods for the Bernstein copula (Dou, et al, 2016) are applicable after some changes made on the basis functions and the constrictions on the parameter matrix R. We consider two EM algorithms for the two forms of the B-spline copula in (1) and (2), respectively. In both methods, the marginal distributions and the copulas are estimated separately. That is, for each marginal, we use the kernel method and the empirical distribution function to estimate the density function and the cumulative distribution function, respectively.

*Waseda University, † The Institute of Statistical Mathematics,‡ Academia Sinica, § Pennsylvania State University

To estimate the joint density of two random variables, the general idea is to use the general form (1) directly,

$$h(x, y; R) = c(F(x), G(y); R)\, f(x)g(y).$$

In this method, we need to determine the parameter matrix R and its size $m \times n$. The first EM algorithm is to estimate the parameter matrix R of the finite mixture distribution (1). The size of R, $m \times n$, can be determined by AIC.

When the two random variables X and Y are highly correlated, we can use the copula density of the maximum correlation case (2) and the independent copula to approximate the joint density of the data by

$$h(x, y; q, n) = (1 - q)f(x)g(y) + qc_n^+(F(x), G(y))f(x)g(y).$$

In this case, we need to find parameters q and n. For this model, except for another EM algorithm, a grid method can also accomplish this purpose.

References

[1] Dou, X., Kuriki, S., Lin, G. D., and Richards, D. (2016). EM algorithms for estimating the Bernstein copula, *CSDA*, **93**, 228–245.

[2] Dou, X., Kuriki, S., Lin, G. D., and Richards, D. (2021). Dependence properties of B-spline copulas, *Sankhya A*, **83**, 283–311.

予測モデルのメタアナリシス
Meta-analysis of prediction models

聖路加国際大学　米岡大輔

概要　本発表では、多変量メタアナリシスの手法を用いた回帰モデルの係数統合とそれにまつわる問題点を整理する。また、発展型として現在取り組んでいる予測モデルの線形予測子の統合についても時間があれば議論する。

1　はじめに

医学、疫学、教育学を始めとした学際的分野において、各研究が報告する要約統計量を統合する試みとしてメタアナリシスがある。従来のメタアナリシス研究は、例えば各研究の疾患割合などの単一指標の統合に取り組んできた。しかし近年、多変量メタアナリシスの手法が発展し複数の指標の統合が盛んに研究されるようになり、ネットワークメタアナリシスなどへの応用も見られる。この文脈においてBecker *et al.* (2007)[1]は回帰モデルの係数統合を提唱し、予測モデルのメタアナリシス的統合という分野に脚光が当たり始めた。

本発表では、回帰モデルの係数統合における様々な問題点（例えば、報告されない要約統計量の問題や、予測タスクは同じだが回帰関数が異なる場合など）を整理し、特に回帰モデル間に共変量の組の相違がある場合のバイアスとその補正について議論する。また、時間があれば発展型として現在取り組んでいる予測モデルの線形予測子の統合についても議論する。

2　提案方法: 共変量の組に相違がある場合の係数統合

手元にK個($k=1,\dots,K$)の過去研究があり、その各々がロジスティック回帰の係数ベクトル$\boldsymbol{\beta}_k$とその分散共分散行列$\boldsymbol{\Sigma}_k$を報告しているとする。各モデルには異なる組の共変量が入っているため、単純に多変量メタアナリシスで統合することは難しい状況を考える。また、手元には一つだけ実データもあり、過去研究のすべての共変量が含まれるとする。

簡単化のため$K=2$とし、一つが真モデル(1)を推定しているのに対し、もう一つが誤特定のモデル(2)を使った場合の係数統合について考える。

$$\mathrm{logit}P(Y=1|X,Z)=\alpha_0+\alpha_1X+\alpha_2Z \qquad (1)$$

$$\mathrm{logit}P(Y=1|X)=\beta_0+\beta_1X \qquad (2)$$

$\boldsymbol{\alpha}$と$\boldsymbol{\beta}$のパラメータの関係を表す$\beta_i^*=g_i(\alpha_0,\alpha_1,\alpha_2,p_{XZ})$ ($i=0,1$)は以下のスコア関数の普遍性を満たす（ただしp_{XZ}は共変量XとZの同時分布)。

$$E\left[\left\{Y-\frac{\exp(\beta_0^*+\beta_1^*X)}{1+\exp(\beta_0^*+\beta_1^*X)}\right\}\begin{pmatrix}1\\X\end{pmatrix}\right]=\mathbf{0}$$

特にXとZがともに連続変数で、$Z|X$が正規分布となるときに、このgはCramer (2003) [2]やChao *et al.* (1997) [3]の以下のような結果と一致することが示せる。

$$\beta_0^*=g_0(\alpha_0,\alpha_1,\alpha_2,p_{XZ})\approx\frac{\alpha_0+\alpha_2\gamma_0}{\sqrt{1+c^2\alpha_2^2\sigma^2}}$$

$$\beta_1^*=g_1(\alpha_0,\alpha_1,\alpha_2,p_{XZ})\approx\frac{\alpha_1+\alpha_2\gamma_1}{\sqrt{1+c^2\alpha_2^2\sigma^2}}$$

ただし、ここでは$Z=\gamma_0+\gamma_1X+\tau$、$\tau|X\sim N(0,\sigma^2)$と仮定している。共変量が全て連続の場合は、このようにgを陽に書けることがあるが、一般には難しく、上のスコア関数の普遍性の要件から陰関数gとして定義する。

このgを用いて、以下の一般化最小二乗法を解くことで、統合された係数である$\hat{\boldsymbol{\alpha}}^*=(\alpha_0^*,\alpha_1^*,\alpha_2^*)$を得る。

$$\arg\min_{\boldsymbol{\alpha}}\sum_{k=1}^{K}\left\{\hat{\theta}_k-f_k(\boldsymbol{\alpha},\hat{\boldsymbol{p}}_{\boldsymbol{XZ}})\right\}^\top\boldsymbol{\Sigma}_k^{-1}\left\{\hat{\theta}_k-f_k(\boldsymbol{\alpha},\hat{\boldsymbol{p}}_{\boldsymbol{XZ}})\right\}$$

ただし、$\hat{p}_{XZ}$は手元のデータより推定した量であり、

$$\hat{\theta}_k=\begin{cases}\hat{\boldsymbol{\alpha}}_k=(\hat{\alpha}_0,\hat{\alpha}_1,\hat{\alpha}_2)^\top & (\text{if } k=1)\\ \hat{\boldsymbol{\beta}}_k=(\hat{\beta}_0,\hat{\beta}_1)^\top & (\text{if } k=2)\end{cases}$$

$$\hat{f}_k(\boldsymbol{\alpha},\hat{\boldsymbol{p}}_{\boldsymbol{XZ}})=\begin{cases}\hat{\boldsymbol{\alpha}}_k & (\text{if } k=1)\\ (g_1(\boldsymbol{\alpha},\hat{\boldsymbol{p}}_{\boldsymbol{XZ}}),g_2(\boldsymbol{\alpha},\hat{\boldsymbol{p}}_{\boldsymbol{XZ}}))^\top & (\text{if } k=2)\end{cases}$$

3　数値実験

$K=9$とし、うち3研究は共変量X,Z、3研究はXのみ、3研究はZのみでロジスティック回帰を行った結果の係数統合を考える。比較対象として、最初のフルセットの3研究のみの多変量メタアナリシ

ス（M1)、脱落している係数を平均値補完した後の多変量メタアナリシス（M2）を用意し、バイアスとMSEを1000回のモンテカルロシミュレーションにより評価する。また、簡単化のために（脱落しているものも含めて）共変量はすべて正規分布に従うとする。

結果は以下のTableに示す通りである。M1やM2と比較して、バイアスは10-50%程度の改善、MSEは10-90%程度の改善が見られた。

バイアス

	True value of α_1				
	-3	-1	0	1	3
提案手法					
α_0	0.133	0.062	0.048	0.100	0.166
α_1	-0.377	-0.060	0.000	0.070	0.350
α_2	0.149	0.070	0.036	0.080	0.138
従来手法: フルセットのみ (M1)					
α_0	0.114	-0.005	-0.022	0.012	0.112
α_1	-0.140	0.016	-0.001	-0.003	0.131
α_2	0.066	-0.012	-0.014	0.001	0.058
従来手法: 平均値補完 (M2)					
α_0	0.687	0.375	0.204	0.378	0.687
α_1	-0.915	-0.255	-0.013	0.223	0.857
α_2	0.756	0.379	-0.015	0.326	0.688

MSE

	True value of α_1				
	-3	-1	0	1	3
提案手法					
α_0	0.058	0.024	0.013	0.030	0.074
α_1	0.251	0.015	0.002	0.016	0.239
α_2	0.049	0.019	0.008	0.021	0.048
従来手法: フルセットのみ (M1)					
α_0	0.101	0.042	0.030	0.046	0.105
α_1	0.188	0.018	0.003	0.018	0.183
α_2	0.042	0.021	0.017	0.022	0.042
従来手法: 平均値補完 (M2)					
α_0	0.480	0.149	0.049	0.151	0.481
α_1	0.888	0.070	0.001	0.055	0.780
α_2	0.576	0.148	0.008	0.111	0.477

4 結論

本研究では、メタアナリシスにおける回帰モデルの共変量の組の研究間相違の問題を、脱落変数バイスの問題と捉え直し、そのバイアスをスコア関数の普遍性の要件を考えることで解決した。更にシミュレーションと実データを用いて、既存手法よりバイアスやMSEの意味で本手法が改善されることも示した。

参考文献

[1] Becker BJ, Wu M-J. The Synthesis of Regression Slopes in Meta-Analysis. Statistical Science 2007; 22: 414-429.

[2] Cramer JS. Logit models from economics and other fields. Cambridge University Press: Cambridge, UK ; New York, 2003.

[3] Chao WH, Palta M, Young T. Effect of omitted confounders on the analysis of correlated binary data. Biometrics 1997; 53: 678-689.

マジョリティーに対する治療効果のロバストな予測

国立循環器病研究センター　大前　勝弘

ある疾患を有する対象集団に対して，2 つの治療選択 $a \in \mathcal{A} = \{0, 1\}$ のうち，どちらがより患者の状態を改善するかを予測したい．このために，被験者の治療前の状態ベクトル $x \in \mathbb{R}^p$ に応じて治療を決定する関数 $d : \mathrm{R}^p \to \{0, 1\}$ を考える．この d を治療方針と呼ぶ．

患者の状態を表す確率変数を Y とする．治療前の状態ベクトル x と治療 a を条件付けた下での Y の期待値を $Q(x, a) = \mathrm{E}[Y|X = x, A = a]$ と書くと，$d(Q(x)) = \mathrm{argmax}_a\, Q(x, a) = \mathrm{I}(Q(x, 1) > Q(x, 0))$ が自然な最適治療方針となる．ここで，I は指示関数である．つまり，条件付き期待値 $Q(x, a) = \mathrm{E}[Y|X = x, A = a]$ に関する推論を行うことで，最適治療方針を推定することができる．医学研究で得られる質の高いデータはそれほど膨大でないことと，治療選択問題における解釈性の重要さから，線形モデルを用いた治療方針の推論が有望である．例えばアウトカムが連続値で正規性を仮定できる場合には，治療方針の推定問題は，正規線形交互作用モデル

$$Y|(X = x, A = a) \sim \mathrm{N}(Q(x, a; \beta), \sigma^2)$$
$$Q(x, a; \beta) = \beta_0 + \beta_1^\top x + \beta_2 a + \beta_3^\top x a$$

のうち関心のあるパラメータ (β_2, β_3) に関する推論の問題へと帰着される．

ところが，多くの疾患と治療に関して，治療反応性は患者ごとでしばしば異なることが知られている．例えば $\{(x, a = 1, y), Q(x, 1) > Q(x, -1) > y\}$ を満たす患者のように，最適治療方針による治療選択が不合理になる場合がある．治療反応の異質性を考慮せずに対象集団全体にとって良い治療方針を模索しようとすると，このようにイレギュラーな特性をもつ学習サンプル：マイノリティーによって歪められた推論になってしまう可能性がある．そのような推論をもとに治療選択をすると，大多数の標準的な特性を持つ患者集団：マジョリティーにとって不利益を生じうる可能性もある．治療開発のコストパフォーマンスを高め，さらなる治療開発へと繋げるためには，「マジョリティーに対して有効な治療方針の推論」および「推定された治療方針が有効な集団の特定」が重要な課題となる．

本発表では，これらの問題に対応するために，治療周辺回帰モデル

$$p_\theta(y|x) = \sum_{a \in \mathcal{A}} p_\theta(y|x, a) p(a|x) \tag{1}$$

を提案した。この (1) 式に基づくパラメータ $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$ の対数尤度関数は，

$$\ell_R(\theta) = \sum_{i=1}^{n} \log \left(\sum_{a \in \mathcal{A}} p(y_i|x_i, a; \theta) p(a|x_i) \right)$$

で与えられる．従って，治療周辺回帰モデルの関心のあるパラメータに関する推定方程式は，次式のような正規線形回帰モデルの推定方程式 $\mathcal{E}_\mathrm{F}(\beta, x, a, y) = \frac{y - Q_\beta(x,a)}{\sigma^2} \frac{\partial}{\partial \beta} Q_\beta(x, a)$ の事後確率重み付き

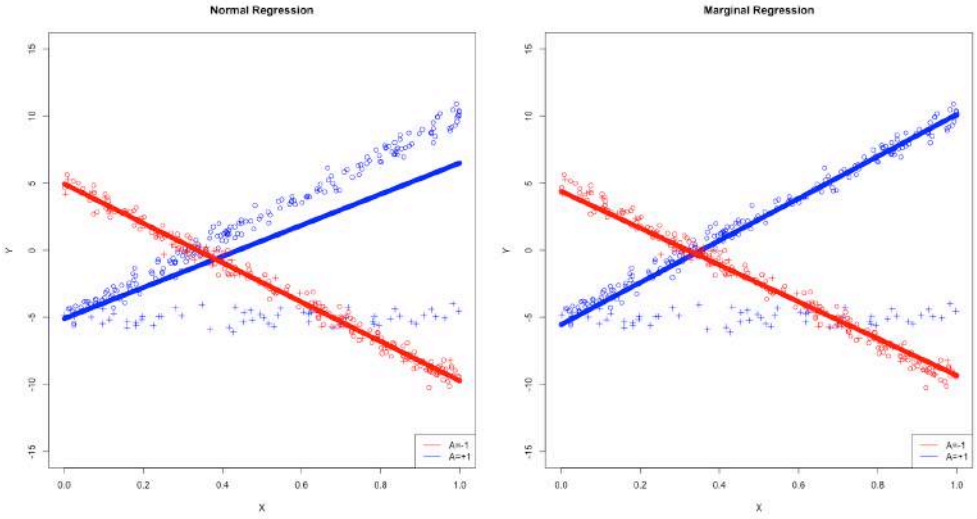


図 1　正規線形回帰モデルと治療周辺回帰モデル

平均となる.

$$\mathcal{E}_{\mathrm{M}}(\beta, x, a, y) = p_\beta(-1|x, y)\mathcal{E}_{\mathrm{F}}(\beta, x, -1, y) + p_\beta(1|x, y)\mathcal{E}_{\mathrm{F}}(\beta, x, 1, y). \tag{2}$$

ただし,

$$p_\beta(a|x, y) = \begin{cases} \dfrac{1}{1 + \exp(-\alpha_1(x)y - \alpha_0(x))} & \text{if } a = 1 \\ \dfrac{1}{1 + \exp(\alpha_1(x)y + \alpha_0(x))} & \text{if } a = -1 \end{cases}$$

であり,

$$\alpha_1(x) = \frac{Q_\beta(x, 1) - Q_\beta(x, -1)}{\sigma^2}, \alpha_0(x) = \frac{Q_\beta(x, -1)^2 - Q_\beta(x, 1)^2}{2\sigma^2} + \log\frac{p(1|x)}{p(-1|x)}.$$

正規線形回帰モデルの最尤推定量はマイノリティー集団に敏感であり，マイノリティー集団の存在によってパラメータの推定が非常に不安定になってしまう．これに対して，周辺回帰モデルの最尤推定量は，治療レジメンに不適合なマイノリティー集団の存在に対して頑健な性質を持つ（図 1 ）．これは推定方程式の以下の性質により得られる.

$$\begin{aligned} \sup_{y \le Q(x,0)} p_\beta(1|x, y)\|\mathcal{E}_{\mathrm{F}}(\beta, x, 1, y)\| &< \infty \\ \lim_{y \to -\infty} p_\beta(1|x, y)\|\mathcal{E}_{\mathrm{F}}(\beta, x, 1, y)\| &= 0 \\ \sup_{y \ge Q(x,1)} p_\beta(0|x, y)\|\mathcal{E}_{\mathrm{F}}(\beta, x, 0, y)\| &< \infty \\ \lim_{y \to \infty} p_\beta(0|x, y)\|\mathcal{E}_{\mathrm{F}}(\beta, x, 0, y)\| &= 0. \end{aligned}$$

治療周辺回帰モデルは一般化線形モデルと同様の拡張が可能であるため，様々な種類のアウトカムを改善するための治療レジメン推定問題への応用が今後期待される.

横浜市における救急出場件数に対する関数動的予測モデリング

横浜市立大学 医学部　三角 俊裕
横浜市立大学 医学部　三枝 祐輔
横浜市立大学 医学部　原 みゆひ
横浜市消防局　藤田 豊
横浜市消防局　吉塲 亮介

1. 概要

図 1 は，2008 年から 2019 年に記録された横浜市における救急車の出場要請データを示す．救急車の時間別出場件数は，午前中にピークを迎え，午後から夜間にかけて減少する特徴を持つ．また，休日・休日明け・平日に分類すると，特に休日明けの午前中に増加する傾向にある．さらに，出場件数は夏と冬に比較的多く，春と秋に少ない季節変動性の特徴がある．2008 年以降は増加傾向にあり，今後もこの傾向が継続する場合，救急車の現場到着の遅延などの問題が危惧される．そのため，限られたリソースを効率よく運用する施策が求められている．

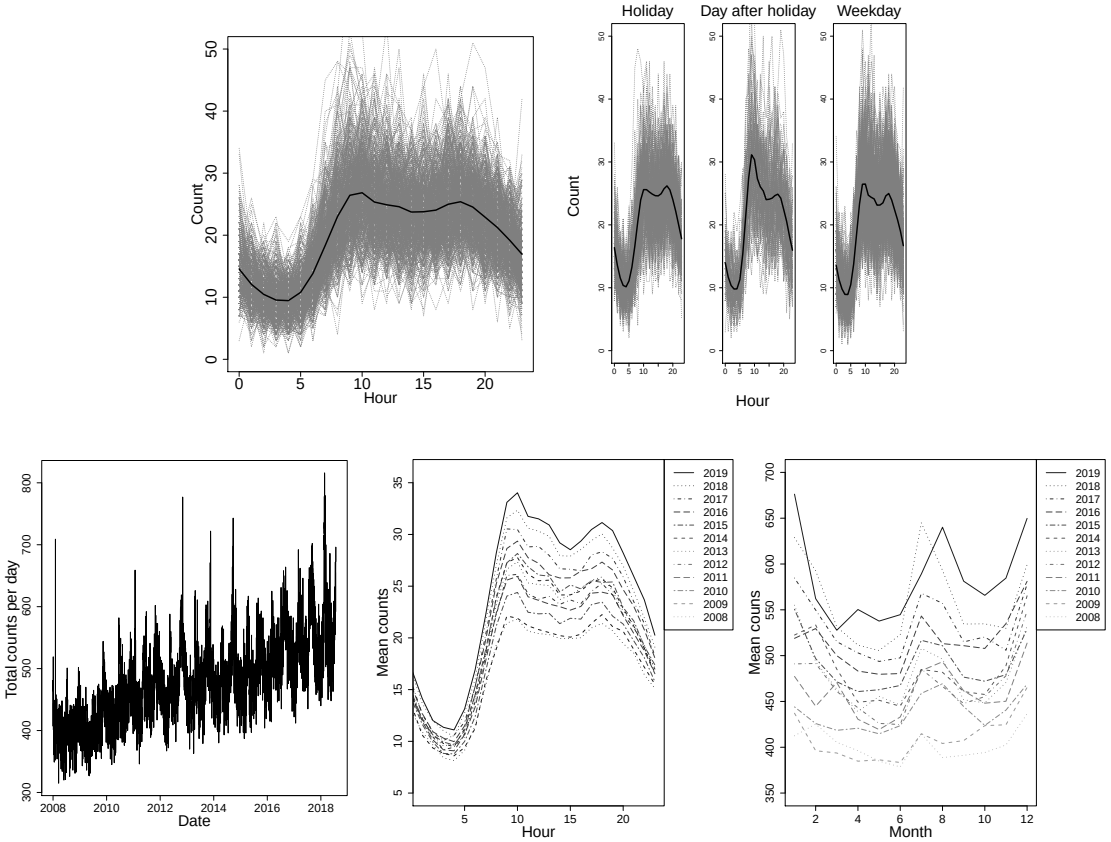


図 1: 左上：2008-2019 年からランダムに選択した 500 日分の時間別救急出場件数のスパゲッティプロットと平均曲線，右上：2008-2019 年からランダムに選択した 500 日分の時間別出場件数を休日・休日明け・平日に分類して作成したスパゲッティプロットと平均曲線，左下：1 日あたりの合計救急出場件数の時系列プロット，中央下：時間別救急出場件数の年平均曲線，右下：月別救急出場件数の年平均曲線．

本研究では，一日のある時間までに観測された救急出場件数に基づいて，以降の出場件数予測を逐次的に行う動的予測法を提案する．動的予測法の文脈では，交通量予測 (Chiou, 2012) や子供の成長予測 (Leroux, 2018) など，関数データ解析手法に基づくアプローチが注目されている．そこで，様々な特徴を持つ横浜市の救急出場データに対して，Leroux (2018) により提案された関数同時回帰モデルを拡張させた関数動的予測モデリング手法を提案した．提案手法は，横浜市の救急出場データに適用した．また，数値シミュレーションにより提案手法の有用性を検討した．

2. 関数動的予測モデル

n 日分の観測データを $\{\xi_{ij}, x_{iq}, z_{ijr}; i = 1, \ldots, n, j = 1, \ldots, J, q = 1, \ldots, Q, r = 1, \ldots, R\}$ とする．ここで，$\xi_{ij} = \xi_i(t_{ij})$ は第 i 日目の t_{ij} 時における出場件数，x_{iq} および $z_{ijr} = z_{ir}(t_{ij})$ は，それぞれ第 i 日目の q 番目の非時間依存共変量，t_{ij} 時における r 番目の時間依存共変量とする．また，出場件数の長期傾向と季節変動を考慮するために，第 i 日が属する年，月をそれぞれ $t^y_{ik}(k = 1, \ldots, K)$，$t^m_{il}(l = 1, \ldots, L)$ とし，当該出場件数を $\xi_{ikl}(t)$ と定義する．このデータに対して，以下のモデルを仮定する．

$$\xi_{ikl}(t_{ij}) = f_0(t_{ij}) + \sum_{q=1}^{Q} x_{iq} f_{1q}(t_{ij}) + \sum_{r=1}^{R} z_{ir}(t_{ij}) f_{2r}(t_{ij}) + g(t^y_{ik}, t^m_{il}) + b_i(t_{ij}) + \varepsilon_i(t_{ij}). \tag{1}$$

ここで，$f_0(\cdot)$ は母平均関数，$f_{1q}(\cdot)$ は q 番目の非時間依存共変量の係数関数，$f_{2r}(\cdot)$ は r 番目の時間依存共変量の係数関数，$g(\cdot, \cdot)$ は長期傾向と季節変動を表す関数，$b_i(\cdot)$ は日毎の変動を表す変量効果関数，$\varepsilon_i(t_{ij}) = \varepsilon_{ij}$ は観測誤差である．また，$b_i(t)$ は共分散関数 $C(s,t) = \mathrm{cov}\{(b_i(s), b_i(t)\}$ をもつガウス過程，ε_{ij} は $N(0, \sigma^2)$ にそれぞれ従い，$b_i(t)$ と $\varepsilon_i(t)$ は互いに独立であると仮定する．さらに，時間依存共変量は，季節変動の特徴をもつ気温や湿度等を想定し，$z_{irk}(t_{ij}) = \gamma_r(t_{ij}) + g_r(t^m_{ik}) + b_{ir}(t_{ij}) + \varepsilon_{ir}(t_{ij})$ のようなモデルを (1) 式と同時に仮定する．ここで，$\gamma_r(\cdot)$ は母平均関数，$g_r(\cdot)$ は季節変動を表す関数，$b_{ir}(\cdot)$ は日毎の変動を表す変量効果関数，$\varepsilon_{ir}(t_{ij}) = \varepsilon_{ijr}$ は観測誤差とする．$b_{ir}(t)$ は共分散関数 $C_r(s,t) = \mathrm{cov}\{b_{ir}(s), b_{ir}(t)\}$ をもつガウス過程，ε_{ijr} は $N(0, \sigma^2_r)$ にそれぞれ従い，$b_{ir}(t)$ と $\varepsilon_{ir}(t)$ は互いに独立であると仮定する．出場件数の長期傾向と季節変動は，$g(t^y, t^m) = \psi_{\mathrm{trend}}(t^y) + \psi_{\mathrm{seasonal}}(t^m) + \psi_{\mathrm{interaction}}(t^y, t^m)$ と表し，$\psi_{\mathrm{trend}}(\cdot)$ は長期変動項，$\psi_{\mathrm{seasonal}}(\cdot)$ は季節変動項，$\psi_{\mathrm{interaction}}(\cdot, \cdot)$ は交互作用項とする．時間別出場件数の動的予測は，Leroux et al. (2018) を参考に最良線形不偏予測量を用いて行った．

モデリングの詳細，数値シミュレーション結果および実データへの適用結果は，当日報告した．

参考文献

Leroux, A., Xiao, L., Crainiceanu, C., and Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. *Statistics in Medicine*, **37**(8):1376–1388.

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

概念の複雑性・信頼性予測のための受け入れ検査（デジタル言語学)

得丸久文(カラハリプロジェクト)

1. デジタル言語学（生命進化の延長線上にある言語的知能）

1.1 原核生物からインターネットまで一貫する生命のデジタル進化(物理層➔論理層)

デジタルは 01(ON-OFF)の電気信号にもとづく 2 進計算に限らない．むしろ生命体が 4 種の核酸で紡いだ遺伝情報を自律分散的に進化させる生命進化こそがデジタルであり，2 進計算機と離散信号である音節を使った言語と知能の進化は生命進化の派生として位置づけられる．

表 1 は真核生物，脊椎動物，哺乳類，人類の進化を概観する．すべて進化は，よりよく生きようとする生命体の意志が生みだす．第一段階は物理進化であり，小惑星衝突などによる生息環境の悪化が，それを乗り切るための低雑音環境(核，脳室，子宮)を生みだす．第二進化は論理進化であり，環境が復元されると，低雑音環境(=高 S/N 比)での高ダイナミックレンジを利用した，複雑で繊細な信号処理が可能となり，目に見えない論理進化が生まれる．

時期	誕生	物理進化	物理進化現象	論理進化現象
2BA	**真核生物**	核，共生	核の保護と細胞質共生	転写後修飾，多細胞化
530MA	**脊椎動物**	脳室	視覚・聴覚	**脊髄反射(脳室内免疫 NW)**
66MA	**哺乳類**	子宮	保育器，授乳	大きな脳，音声信号交信
130KA	**現生人類**	洞窟居住	ヒトの晩成化	ヒトの**真社会性化**，鳴き声共有
66KA	**言語的人類**	音節	無限の語彙数，母音	文法(母語の片耳聴覚)
5KA	**文明人**	文字	文明(時空を超えた共有)	概念(反射から群への進化)
50A	**知恵ある人**	Bit	ネットワーク検索	前方誤り訂正

表 1 生命のデジタル進化概要(B=10 億，M=100 万，K=千，A=年前)

1.2 離散信号音節を活用するためには低雑音環境が必要であった

ヒトの言語は，真社会性動物が共同体内部で共有する特殊な鳴き声が，相互に周波数離散的な音素性を獲得し，母音アクセントによる時間的離散性をもつ**音節**を獲得して，無意識の文法処理が始まった．その後，消えない音節である**文字**と対話する音節である **bit** が生まれた後，その特徴を活かした論理進化を確立させ，知恵ある人の遺伝子に書き込む日がまもなく訪れる．

文法処理が無意識に自然に行われるのは言語学上の最大の謎であった．筆者の仮説は① 話し手は，文法語の音節を識別子なしに挿入して修飾・接続を示す，② 聞き手は，母語を片耳で聴き，脳幹聴覚神経核の方向定位能力で文法語をベクトル解析する，というもの．(得丸 2014)

片耳聴覚への切り替えが生得的な遺伝子情報に基づいている(遺伝子に符号化されている)こと，熱帯雨林で生活するピダハンはその遺伝子が発現せず母語を両耳で聞くことから推論すると，文法のためには開墾地や都市など低雑音環境が必要である．概念は世俗から隔絶された僧院で生まれた．前方誤り訂正のためには，より低雑音の環境とより低速の反復処理が求められる．

2. 概念の複雑性を予測するために

2.1 言葉を処理する生物構造は脳室内免疫細胞ネットワークである

ヒトの言語的知能の基盤は，脊椎動物の脊髄反射回路であり，それは脳室内免疫細胞ネットワークである．五官の記憶はマイクログリア細胞に記銘されて大脳皮質に着床し，言葉の記憶とネットワーク記憶は脳脊髄液中を浮遊する B リンパ球(B 細胞)が保持する．

2.2 言葉の意味は 2 種類の記憶と 3 種類の論理によってフラクタルに複雑化する

意味の複雑化は，言葉と意味を二元的に結びつける B 細胞の内部論理が，1 対 1 の反射論理から，1 対全の群論理，全対全のネットワーク論理へとフラクタルに複雑化して実現する．言葉には 6(=2x3)通りの複雑性があることを聞き手は知っておく必要がある．

	五官記憶(グリア細胞)	意味記憶(B 細胞，ネットワーク)
1 対 1(反射論理)	**偶発的体験**	**現象の恣意的な命名**
1 対全(群論理)	**日常概念**	**科学的概念**
全対全(ネットワーク論理)	**日常概念操作**	**学際的な科学概念**

3. 概念の信頼性は予測不可能．知識再生的に受容する

3.1 脊髄反射と真社会性の制約により，知能は誤り訂正しにくい

文字や bit 化された言語情報は，通信回線上で単なる文字列・bit 列となるため，改ざんが容易である．著者没後の偽書を見破ることは難しい．

真社会性動物であるため，師が弟子に知識伝達する場合，弟子は鵜呑みにするほかない．

脊髄反射回路で言語処理するため，一度誤った知識を獲得するとそれが基準になり，正しい情報を受け入れられなくなる．

3.2 知識再生的な受容のための受け入れ時のアルゴリズムを実装する

受け入れる知識は，話し手が自分で考えたことか(A)，他者から仕入れた知識か(B)

A ➔ どのように考えたかの筋道を明らかにする ➔ 自分でも同じように考えてみる

B ➔ 誰が考えたのかを明らかにする ➔ 知識・概念の発見者を見つける ➔A に戻る

この知識再生的受容アルゴリズムが遺伝子に書き込まれたとき，ホモサピエンスが誕生する．

参考文献：得丸(2014) 母語のモノラル聴覚と文法処理－例外としてのピダハン，情処学会 2014-NL-219(23) PP 1-19

得丸(20H00576)概念の複雑性、ならびに『きわめて大量だが信頼性の保証がない言語情報』の取り扱いについて

得丸(20H00576) 概念は群である－学習における前方誤り訂正の必要性

予測のための最大エントロピーと最小ダイバージェンス

江口 真透
統計数理研究所

一般化線形モデル (GLM) は指数型分布族の中で標準パラメータから回帰関数へのリンク関数によって自然な線形モデルを導入している．更に混合効果モデル，一般化推定方程式などの工夫から広くデータ解析に用いられている．その理論的背景には最大エントロピーモデルと KL ダイバージェンスの関係の巧妙な利用がある．この発表では回帰の内容で，一般の U-エントロピーと U-ダイバージェンスを考えて最大エントロピーと最小ダイバージェンスの双対的な関係を追求する．これによって GLM の拡張を図り，GLM をより柔軟にする方法について考えたい．

1. はじめに

特徴ベクトル $\boldsymbol{X}$ から目的変数 Y の予測問題を考える．$\boldsymbol{X}=\boldsymbol{x}$ を与えたとき Y の条件付き密度 (質量) 関数の全体を $\mathcal{P}=\{p(y|\boldsymbol{x})\}$ と書くとき，回帰関数の空間 $\mathcal{R}=\{\mu(\boldsymbol{x})=\int yp(y|\boldsymbol{x})d\Lambda(y) : p\in\mathcal{P}\}$ が導かれる．GLM の枠組みでは，分散パラメータを取り込まない形であれば

$$p(y|\boldsymbol{x})=\exp\{y\theta(\boldsymbol{x})-\psi(\theta(\boldsymbol{x}))\}$$

と仮定される．cf. Hastie et al. (2009). 標準パラメータ集合が実軸 $\mathbb{R}$ まで拡張できると仮定すると標準リンクモデル $\theta(\boldsymbol{x},\boldsymbol{\alpha})=\boldsymbol{\alpha}^\top\boldsymbol{x}$ が無理のない形で回帰モデルが定式化できる．これより，標準パラメータから平均値パラメータの変換から回帰関数は $\mu(\boldsymbol{x},\boldsymbol{\alpha})=\eta(\boldsymbol{\alpha}^\top\boldsymbol{x})$ と表される．ここで $\eta(\theta)=(\partial/\partial\theta)\psi(\theta)$ とする．このように Y がベルヌイ分布，ポアソン分布，ガンマ分布などに従っているときでも，$\theta(\boldsymbol{x})$ のモデル化によって自然な回帰モデルが提案されている．GLM の枠組みにおいて (1) はエントロピー最大分布となっており，パラメータ推定は尤度法，あるいはクルバック・ライブラー（KL）ダイバージェンス最小化が実行されている．この関係を一般化して，そのクラスの中でデータをより柔軟に学習する方法を求めたい．ここで $\mathcal{P}$ 上のエントロピーは

$$H_0(p)=\int p(y|x)\log p(y|x)dydx$$

で与えれ，対数尤度関数はデータ $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)_{i=1}^n\}$ に対して

$$\ell(\boldsymbol{\alpha},\mathcal{D})=\sum_{i=1}^n\{y_i\boldsymbol{\alpha}^\top\boldsymbol{x}_i-\psi(\theta(\boldsymbol{\alpha}^\top\boldsymbol{x}_i))\}$$

と書けて y_i に関して線形であることが特徴となっている．

2. 一般化エントロピーとダイバージェンス

エントロピーとダイバージェンスの一般化のため生成関数の集合を $\mathcal{U}=\{U:\mathbb{R}\to[0,\infty) : U'(t)>0, U''(t)>0\}$ 用意する, cf. Eguchi (2006). スカラー倍の不定性を除くために $U'(1)=1$ を課す．$U(t)$ を単調増加な凸関数とする．この $\mathcal{U}$ の U を使って，空間 $\mathcal{P}$ に交差エントロピーを

$$H_U(p,q)=-\int\{p(y|\boldsymbol{x})\xi(q(y|\boldsymbol{x}))-\int U(\xi(q(y|\boldsymbol{x}))\}p(\boldsymbol{x})d\Lambda(\boldsymbol{x},y)$$

と定めて，U-交差エントロピーと呼ぶ．ここで $\xi(s)$ は $U'(t)$ の逆関数とする．このとき，

$$H_U(p,q)\geq H_U(p)$$

が成立し，等号は $p=q$ のときに限ることが知られている．ここで $H_U(p)=H_U(p,p)$ とし，$H_U(p)$ を U-エントロピーと呼ぶ．このことから U-ダイバージェンス

$$D_U(p,q)=H_U(p,q)-H_U(p)$$

が定まる．与えられたデータ $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)_{i=1}^n\}$ から統計モデル $\mathcal{M}=\{p(y|x,\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\}$ が建てられたとき，U-損失関数を

$$L_U(\boldsymbol{\theta},\mathcal{D})=-\sum_{i=1}^n\{\xi(p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}))-\int U(\xi(p(y|\boldsymbol{x}_i,\boldsymbol{\theta})))d\Lambda(y)\}$$

と定めて未知パラメータ $\boldsymbol{\theta}$ の U-推定量 $\hat{\boldsymbol{\theta}}_U=\mathrm{argmax}_{\theta\in\Theta}L_U(\boldsymbol{\theta},\mathcal{D})$ を導入する．U-損失関数 $L_U(\boldsymbol{\theta},\mathcal{D})$ は交差エントロピー $H_U(p,q)$ の $p(y|\boldsymbol{x})$ はデータ $\mathcal{D}$ の経験分布関数，$q(y,x)$ はモデル $p(y|\boldsymbol{x},\boldsymbol{\theta})$ を代入したものとなっている．

$\mathcal{U}$ の中から $U_0=\exp$ を取れば $H_{U_0}(p)$ はボルツマン・シャノンのエントロピー，$D_{U_0}(p,q)$ は KL ダイバージェンス，$L_U(\boldsymbol{\theta},\mathcal{D})$ は負の対数尤度関数に帰着され，GLM の基礎をあたえる．一方で，$\mathcal{U}$ の中から

$$U_\beta(t)=\tfrac{1}{\beta+1}(1+\beta t)^{\frac{\beta+1}{\beta}}$$

を採用すると，$\xi(s)=\frac{1}{\beta}(s^\beta-1)$ となるので，べきクロスエントロピーは

$$H_\gamma(p,q)=-\int\{\tfrac{1}{\beta}p(y|\boldsymbol{x})[q(y|\boldsymbol{x})^\beta-1]-\tfrac{1}{\beta+1}q(y|\boldsymbol{x})^{\beta+1}\}p(\boldsymbol{x})d\Lambda(\boldsymbol{x},y)$$

が導かれる．

このように生成関数集合 $\mathcal{U}$ に対してモデルと推定の組を $\mathcal{U}\times\mathcal{U}$ から (U_0,U_1) を取り，U_0-エントロピー最大モデル $\mathcal{M}_{U_0}$ と U_1-推定量 $\hat{\boldsymbol{\alpha}}_{U_1}$ を取ると対角集合　$\Delta:=\{(U,U):U\in\mathcal{U}\}$　に共通な構造が現れる．すなわち，U_0-エントロピー最大モデルは

$$\mathcal{M}_{U_0}=\{p_0(y|\boldsymbol{x},\boldsymbol{\alpha}):=U_0'\big(y\boldsymbol{\alpha}^\top\boldsymbol{x}-\psi_0(\boldsymbol{\alpha}^\top\boldsymbol{x})\big):\boldsymbol{\alpha}\in\mathbb{R}^d\}$$

で与えられ，U_1-ダイバージェンス推定の損失関数はモデル $\mathcal{M}=\{p(y|\boldsymbol{x},\boldsymbol{\alpha}):\boldsymbol{\alpha}\in\mathbb{R}^d\}$ に対して

$$L_{U_1}(\boldsymbol{\alpha},\mathcal{D})=-\sum_{i=1}^n\{U_1'^{-1}(p(y_i|\boldsymbol{x}_i,\boldsymbol{\alpha}))-\int U_1(p(y|\boldsymbol{x}_i,\boldsymbol{\alpha}))dy\}$$

で与えられる．従って，$(U_0,U_1)=(U,U)$ と取れば (1) は

$$L_U(\boldsymbol{\alpha},\mathcal{D})=-\sum_{i=1}^n\{y_i\boldsymbol{\alpha}^\top\boldsymbol{x}_i-\kappa(\boldsymbol{\alpha}^\top\boldsymbol{x}_i)\}$$

と書ける．ここで $\kappa(\boldsymbol{\alpha}^\top\boldsymbol{x})=\psi_0(\boldsymbol{\alpha}^\top\boldsymbol{x})-\int U(y\boldsymbol{\alpha}^\top\boldsymbol{x}-\psi_0(\boldsymbol{\alpha}^\top\boldsymbol{x}))dy$. このように $L_U(\boldsymbol{\alpha},\mathcal{D})$ は負の対数尤度関数と等価な関係が成立する．対角集合 Δ の中で最適なものを探す方法についてはいくつか考えられるが決定的なものは未だ得られてない．応用として生物種の生息分布の推定のためのマックスエントを β-マックスエントに拡張されたものが考えられている, cf. Komori & Eguchi (2019)．今後，様々な応用が成される中で有効なものが提案されることが期待される．

参考文献

Eguchi, S. (2006). Information geometry and statistical pattern recognition. Sugaku Expositions, 19(2), 197-216.

Komori, O., & Eguchi, S. (2019). Statistical methods for imbalanced data in ecological and biological studies. Springer Japan.

臨床研究における統計家の役割

静岡県立静岡がんセンター
臨床研究支援センター
野津昭文

臨床研究にはさまざまな種類(表 1)がある。診療記録を見返して、治療の効果や安全性について集計を行う研究は、後ろ向きの研究と言われ、臨床研究では重要な役割を担っている。しかし、後ろ向き研究には、一般には多くのバイアスが含まれており、結果の信頼性は高いとは言えない。そこで、収集すべき項目や評価方法をあらかじめ決めてから、データを収集する研究があり、前向きの研究と言われる。後ろ向き研究とは異なり、前向き研究では、結果として得られたデータの質を担保するために、多くの工夫がなされている。例えば、医療従事者だけでなく、データの管理を行うデータマネージャー、研究の進捗を管理する研究事務局、データ収集の補助を行う臨床研究コーディネーター、統計解析を行う統計家が関わることがある。これらの職種がそれぞれの領域で専門的な知識を持って関わることで、臨床研究によって質の高いデータを得ることができる。統計家は、研究の計画、実行、解析、解釈のすべてに少なからず関わることになる。

予測モデルの作成についても、上述のように、どのようにデータを取得するかという問題がある。後ろ向きの研究から得られたデータを使う場合もあれば、前向きの研究から得られたデータを使う場合もある。予測モデルの作成は、ともすれば、統計モデルの話に限定して語られることがあるが、どのように取得されたデータを、どのように使用するかという観点から、すなわち研究デザインの観点から議論することも重要である。統計家は、研究目的を果たすために、必要であれば研究組織についても考える場合もある。つまり、ここでも、研究の計画、実行、解析、解釈のすべてに少なからず関わることになる。

本発表では、病院で勤務している統計家が、臨床研究の試験デザインや臨床研究を進める際の組織、さらにそこでの統計家の役割について、当院の例を通して話をした。臨床研究は段階を踏んで進められること、各段階ですべきこと・行わないことについても解説した。さらに、病院の医療従事者と予測モデルを作成する研究を行った際に感じたことや、当院以外の事例も参照しつつ、研究デザインと予測モデルの関連についても話をした。

表 1：臨床研究の種類

	観察研究	介入研究
後向き研究	症例報告、症例集積報告、 コホート研究、症例対照研究	
前向き研究	コホート研究、症例対照研究	パイロット研究、忍容性試験、 第I, II, III, IV相試験

参考文献

井村裕夫(監修): NIH 臨床研究の基本と実際, 丸善, 2016
木原雅子/木原正博(訳): 医学的研究のデザイン, メディカル・サイエンス・インターナショナル, 2014
大橋靖雄 荒川義弘: 臨床試験の進め方, 南江堂, 2006

ガンマクレームによる資産過程における破産確率の漸近近似

千葉大・融合理工学府　岡山 大成
千葉大・理学研究院　内藤 貫太

資産過程のクラメル・ルンドバーグモデルでは，クレーム額を IID 確率変数とし，時刻 t までのクレーム回数を強度 λ のポアソン過程 $N = (N_t)_{t\geq 0}$ を用いて表現する．ここで，i 番目のクレーム額を表す確率変数 Y_i の分布を F_Y とし，$\mathbb{E}[Y_i] = \mu$ とする．初期資産を $x \geq 0$，安全付加率を θ とすると，保険会社の資産過程 $X = (X_t)_{t\geq 0}$ は

$$X_t = x + (1+\theta)\lambda\mu t - \sum_{i=1}^{N_t} Y_i$$

と表される．初期資産 x の X_t についての破産時刻を $\tau(x) = \inf\{t > 0 : X_t < 0\}$ とし，有限時間破産確率を $\psi(x,T) = \mathbb{P}(\tau(x) \leq T)$ とする．このモデルが純益条件を満たすとき，破産確率 $\psi(x) = \lim_{T\to\infty}\psi(x,T)$ のラプラス変換は

$$\mathcal{L}\psi(s) = \frac{1}{s} - \frac{\mu\theta}{s\mu(1+\theta) - 1 + m_Y(-s)}$$

となることが知られている (清水 (2018))．ここで $m_Y(s)$ は Y のモーメント母関数である．

資産過程 X_t において，自然数 n に対して，

$$\theta \to \frac{\theta}{\sqrt{n}},\quad \lambda \to n\lambda,\quad Y_i \to \frac{Y_i}{\sqrt{n}}$$

という置き換えによって得られる資産過程

$$\tilde{X}_t = x + \left(1 + \frac{\theta}{\sqrt{n}}\right)(n\lambda)\frac{\mu}{\sqrt{n}}t - \frac{1}{\sqrt{n}}\sum_{i=1}^{\tilde{N}_t} Y_i$$

をスケーリングによる資産過程と呼ぶ (Cohen and Young(2020))．ここで，$\tilde{N}_t$ は強度 $n\lambda$ のポアソン過程であり，n は例えば観測期間に対応している．極限定理として，

$$\tilde{X}_t \xrightarrow{d} x + W_t \quad (n\to\infty)$$

が成り立つことが知られている．ここで，W_t は $(\theta\lambda\mu, \lambda\mathbb{E}[Y^2])$-ブラウン運動である (Schmidli(2017,Proposition 5.19), 清水 (2018,6.3.2 節))．

n に依存している $\tilde{X}_t$ の破産確率を $\psi_n(x)$ とする．$n = 1$ のとき，$\tilde{X}_t$ は元の資産過程 X_t そのものである．したがって，X_t の破産確率 $\psi(x)$ は，$n = 1$ のときの $\tilde{X}_t$ の破産確率 $\psi_1(x)$ である．いま，極限定理が成り立っており，この極限資産過程 (ブラウン運動) の破産確率を $\psi_D(x)$ とすると，形式的に $\psi_n(x) \to \psi_D(x)$ $(n\to\infty)$ である．よって，$\psi(x) = \psi_1(x)$ が $\psi_D(x)$ で近似できているのかが問題となる．$\psi_n(x) \to \psi_D(x)$ $(n\to\infty)$ であり，$\psi(x)$=$\psi_1(x)$ であるから，$\psi(x)$ の $\psi_D(x)$ による近似は $\psi_n(x)$ を陽に求めることで考察できる．

本発表では特に，形状パラメータが α，尺度パラメータが β のガンマ分布にクレーム額が従う場合を考える．そのとき，$n \to \infty$ において $\psi_n(x)$ は以下の漸近展開をもつ：

$$\psi_n(x) = \psi_D(x)\left\{1 - \frac{1}{\sqrt{n}}A_1(x|\theta,\alpha,\beta) + \frac{1}{n}A_2(x|\theta,\alpha,\beta)\right\} + o\left(\frac{1}{n}\right)$$

ここで，

$$\begin{aligned}
\psi_D(x) &= e^{-\gamma x}, \quad \gamma = \frac{2\theta\beta}{\alpha+1}, \\
A_1(x|\theta,\alpha,\beta) &= \left\{\frac{2\theta(\alpha+2)}{3(\alpha+1)} + \frac{\theta(\alpha-1)}{3(\alpha+1)}\mathbf{1}(x=0)\right\}(1-\gamma x), \\
A_2(x|\theta,\alpha,\beta) &= \frac{\theta^2(\alpha+2)}{9(\alpha+1)^2}\left\{2(\alpha+2)\gamma^2x^2 - 3(3\alpha+5)\gamma x + 6(\alpha+1) + \frac{3(\alpha^2-1)}{\alpha+2}\mathbf{1}(x=0)\right\}
\end{aligned}$$

である．この展開は Cohen and Young(2020) で議論された展開の一つの一般化となっている．

また，$\psi_n(x)$ について，任意の $x \geq 0$ において

$$\ell_n(x) < \psi_n(x) < u_n(x)$$

が成り立つような $\ell_n(x)$ と $u_n(x)$ を与えることができる (Cohen and Young(2020))．$\psi_n(x)$ の漸近展開式において，$o(n^{-1/2})$ 項を落とした

$$\psi_{n,1}(x) = \psi_D(x)\left\{1 - \frac{1}{\sqrt{n}}A_1(x|\theta,\alpha,\beta)\right\}$$

に対して，任意の $x \geq 0$ において

$$\ell_n(x) < \psi_{n,1}(x) < u_n(x)$$

が成り立つ．これによって，

$$|\psi_n(x) - \psi_{n,1}(x)| < \mathbf{1}(\psi_n(x) \geq \psi_{n,1}(x))|\psi_n(x) - \ell_n(x)| + \mathbf{1}(\psi_n(x) < \psi_{n,1}(x))|\psi_n(x) - u_n(x)|$$

が成り立ち，$\psi_{n,1}(x)$ は Cohen and Young(2020) で与えられた限界よりも $\psi_n(x)$ の良い近似を与えている．

本講演において，$A_i(x|\theta,\alpha,\beta)(i=1,2)$ の導出について解説を与えるとともに，この展開式の精度と応用例について報告を行った．

参考文献

[1] 清水泰隆 (2018). 保険数理と統計的方法，共立出版.

[2] Cohen, A. and Young, V. R. (2020). Rate of convergence of the probability of ruin in the Cramér-Lundberg model to its diffusion approximation, *Insurance: Mathematics and Economics*, 93, 333-340.

[3] Schmidli, H. (2017). *Risk Theory*. Springer Actuarial Lecture Notes, Springer.

生物分布の時空間予測：機構論的モデル構築に向けた課題

楠本聞太郎（九州大学農学研究院）

将来的な生物多様性損失を防ぐための保全戦略の構築と実践が急務となっている。生物多様性は地理的空間上に不均等に分布しており、我々は限られた保全リソースの最適配分を考えなければならない。生物の空間分布情報は、生物多様性のホットスポットを特定し、様々な社会経済的制約の下で、最大限の保全効果を得るための空間プランニングにとって必須の情報である。保全効果を長期的に持続させるためには、現在の生物の分布だけでなく、気候変動による生息適地の変化も見越した保全エフォート配分が重要になる。これを実現するためには、それぞれの種が気候変動にどのように応答し、その結果として生物多様性の分布がどのように変化するかを予測できなければならない。本講演では、種分布モデリングによる時空間予測のトレンドをレビューし、現状の情報ギャップを指摘する。そして、情報ギャップを緩和するための解析的な方法と、データ収集の取り組みを紹介した。

現在、最も広く普及している分布モデリング手法は、生物の観測（在）データとそれが得られた場所の環境要因との相関関係に基づくモデリングである。相関ベースの方法は、特別な生物学的メカニズムを考慮しないにも関わらず、多くの場合で、対象生物の現在の空間分布（在・不在）をよく判別する。この方法は、最小限の情報（在データと環境データ）でモデリングできることから、多数の生物種を扱う研究にも適用しやすいのが強みである。一方で、生物学的メカニズムを考慮しないことへの批判もある。相関ベースモデルに対して、機構論的モデルは、生物の分布レンジを規定するメカニズムとそれらを制御する環境条件の関係を明らかにし、地理的空間に投影する。実際の地理分布は、複数のメカニズムが複合的に働いた結果であり、全てのメカニズムを統合したモデルを構築することは困難である。よって、純粋な機構論的モデルは、部分的にしか分布の制限要因を反映せず、分布レンジを過大評価する傾向がある。このように、相関ベースモデル、機構論的モデルいずれにも制約や問題点がある。2 つのアプローチの予測結果の違いは、構築したモデルを別の地理的空間や、将来環境に投影するときに顕著になるようである。このようなモデル間の不一致性について、どちらのモデルが正しいかを判断するのは難しい。機構論的モデルには生物学的なリアリズムがあるが、メカニズムに対する仮定が間違っていたり、重要な生物学的プロセスを見落としていたりすると、相関ベースよりもむしろ悪い予測結果になることもある。モデル間の不一致性は、モデリングに用いた環境要因では捕捉できないメカニズムが働いていることを示唆しているため、異なるアプローチで構築されたモデルを比較しながら、見落とされている生物学的

メカニズムを考察する必要がある。

双方の強みを活かすために、相関ベースモデルに生物学的なリアリズムを反映させたハイブリッド型アプローチも提案されている。ハイブリッド型アプローチの最も単純な例は、メカニズムを直接的に制御する環境変数を選択する方法である。そのほか、分布を制限するメカニズムの情報を使って、相関ベースモデルの学習を行う領域を操作する方法や、局所スケールで得られた個体群レベルの高解像度データを使って相関ベースモデルを構築し、それを広い時空間スケールに投影する方法が提案されている。

生物学的メカニズムに関するデータを得るには、多くの実験・観測情報が必要であり、ほとんどの分類群で、そのような情報は利用可能でない。このことが、機構論的モデルやハイブリッドモデルの多種系への適用への障壁になっている。これに対する解析的な解決策として、階層モデルやインバースモデルによって、異なるタイプのデータをフュージョンする方法が着目されている。階層モデルの例としては、多種系で利用可能な種特性や系統情報を用いて、メカニズムに関連するニッチパラメータを補完（推定）する方法や、環境応答の種間差を潜在変数として多種の分布予測を行う連結種分布モデル（joint-SDM）がある。インバースモデルは、「生物学的メカニズムに関するデータ→機構論的モデル→地理的空間への投影」という流れとは逆に、「地理分布→機構論的モデルのパラメータ推定」という流れで解析を行う。これは、少数の種についての詳細な観測・実験データと、多種系で利用可能な在データの両方の強みを活かせる点で有望である。

最後に、情報不足に対する究極の解決策は、新しいデータの取得である。本講演では、日本産樹木種の種子に関する生理学的情報を収集し、それに基づいて機構論的（ハイブリッド型）分布モデルの構築を目指すプロジェクトを紹介した。樹木のような固着生物の場合、ある場所での生存個体の在情報が分布モデリングに用いられる。しかし、その場所の環境が、種子の貯蔵・発芽から、個体の成長・生存、繁殖フェノロジーまで、統合的に適した環境なのか、将来的にも好適環境が保たれるかどうかはわからない。本研究では、既存の文献情報、データベースに加え、圃場・実験室での発芽実験を通じた、種子の温度依存的な特性に関する情報収集と、それらと分布情報との関係の検証を行う予定である。

Some recent topics for Preferential Sampling

Shinichiro Shirota

October 23, 2021

The basic issue of preferential sampling (PS Diggle et al., 2010) is bias in the sampling of spatial locations where point referenced response data are collected and the potential impact on inference regarding the response surface over the region. The canonical illustrative example addresses the objective of inferring about environmental exposures. If environmental monitors are only placed in locations where environmental levels tend to be high, then interpolation based upon observations from these stations will necessarily produce only high predictions. The obvious remedy lies in suitable spatial design of the locations. For example, a random or space-filling design (Saltzman and Nychka, 1998) for locations over the region of interest is expected to preclude such bias.

However, sampling may not be designed in this fashion. Environmental researchers may install monitoring stations where they expect to find high exposure levels; ecologists may tend to sample where they expect to find individuals. Recognizing the possibility of such bias, can prediction be revised to adjust for it? This is the intention of PS modeling. Specifically, there are typically two questions: (i) is there evidence of a PS effect? and (ii) can we improve prediction in the presence of PS? That is, we do not seek to remove PS, we do not propose to revise the data collection. Rather, we seek to acknowledge its presence and attempt to mitigate its impact.

We introduce the Bayesian modeling approach for PS based on the point processes (especially, log Gaussian Cox processes, Møller et al., 1998) for observed locations, then briefly review some recent topics on PS. Our main contribution here is to expand the issue of sampling bias to the case of bivariate response. If responses at location $\mathbf{s}$, say $(Y_1(\mathbf{s}), Y_2(\mathbf{s}))$ are dependent then prediction of say $Y_1(\mathbf{s}_0)$ at an unobserved location $\mathbf{s}_0$ can benefit from using all of the observed Y_2's in addition to all of the observed Y_1's.

Then, the question here becomes the following. Suppose Y_1 and Y_2 are strongly dependent, that is, for a given region D, realizations of the response surfaces, $Y_1(\mathbf{s}) : \mathbf{s} \in D$ and $Y_2(\mathbf{s}) : \mathbf{s} \in D$ are highly correlated. Then, if there is sampling bias in the subsets, $\mathcal{S}_1$ and $\mathcal{S}_2$ in D where Y_1 and Y_2 are sampled, respectively, can we demonstrate the presence of a

bivariate preferential sampling effect? Can we improve co-kriging through a model which captures bivariate preferential sampling?

We consider two illustrative examples here. One focuses on realizations over a spatial region of ozone and temperature in the summer season (May-September). It is well established that high temperatures are a driver of ozone formation, suggesting expected strong positive dependence. The other considers, for a site, mean diameter at breast height (MDBH) over the site and trees per hectare (TPH) associated with the site. Here, the "constant yield law" (Weiner and Freckleton, 2010) argues that the total yield/biomass on a plot at equilibrium, is roughly constant regardless of the number of individuals, i.e., size of individuals will decrease as density increases, suggesting strong negative dependence.

In the bivariate setting we have $\mathcal{Y}_1$ with associated $\mathcal{S}_1$ and $\mathcal{Y}_2$ with associated $\mathcal{S}_2$. This raises two possibilities: (i) $\mathcal{S}_1 = \mathcal{S}_2$ and (ii) $\mathcal{S}_1 \neq \mathcal{S}_2$. Above, the ozone and temperature fall under (i) since they are collected at the same set of monitoring stations. We imagine a single point pattern but bivariate geostatistical response. The MDBH and TPH fall under (ii) because the sites for the measurements can be different due to different data collection protocols.

References

Diggle, P., R. Menezes, and T. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C 59*, 191–232.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics 25*, 451–482.

Saltzman, N. and D. Nychka (1998). *DI, a design interface for constructing and analyzing spatial designs.* Springer Science and Media.

Weiner, J. and R. P. Freckleton (2010). Constatn final yield. *Annual Reviews of Ecology, Evolution, and Systematics 41*, 173–192.

近似ベイズ計算による野生動物の集団サイズ推定
生成過程が複雑なデータにどう対処するか？

矢島 豪太・中島啓裕
（日本大学生物資源科学部）

はじめに

野生動物の多くの種は，複数個体から成る集団を形成して生活している．その集団サイズ（単位集団内の個体数）は，動物の社会生態を明らかにするうえで基礎となる情報である．また，個体数密度を推定したり，個体群の動態を予測したりするうえでも，集団サイズとその時間変化についての情報は不可欠である．しかし，観測時に集団内のすべての個体を数え上げることは困難であるにもかかわらず，従来の研究の多くは，こうした「偽陰性」を考慮せず，観測値をそのまま真の集団サイズと見なして解析・推論を行ってきた．

近年，生態学の分野で自動撮影カメラが広く利用されるようになってきた（Burton et al. 2015）．自動撮影カメラは，赤外線センサーに動物が反応すると自動で写真を撮影するシステムのことである．自動撮影カメラは一度設置してしまえば，昼夜問わず記録を行えること，また動物に警戒されずに記録を行えることから，野生動物のデータ取得における大きなブレークスルーとなった．

しかし，自動撮影カメラによる観測データから集団サイズを推定することは，従来の直接観察よりもいっそう困難である．第一に，撮影可能な面積が限られているため，個体の検出率が群れの広がりの影響を受け，個体の検出・非検出が独立に起こらない．第二に，陸生の社会性動物の多くは，互いに重複のある固定的な行動圏を持つことが多く，自動撮影カメラのような受動型の検出機器には，複数集団が複数回撮影されるのが普通である．このため，集団サイズの推定には，その地点に生息する集団数と集団サイズを同時に推定する必要が生じる．すなわち，カメラのデータを利用して，正確な集団サイズの推定を行うためには，集団内の個体が観測されるプロセスを明示的に表現した新たな統計モデルが必要である．さらに，複雑なモデルのパラメータ推定を効率よく行うためのアルゴリズムの検討が不可欠である．

そこで，本研究の目的は，(1) 自動撮影カメラで取得したデータに適用可能な，集団サイズ推定のための統計モデルを構築すること，(2) 考案したモデルを房総半島（千葉県南部）のイノシシ個体群に実際に適用することの 2 つである．モデルのパラメータ推定は，近似ベイズ計算（Approximate Bayesian Computation，以下 ABC）を利用した．

統計モデルと推定アルゴリズム

統計モデルの構築は，真の集団数とサイズが決定される生態学的なプロセス（生態プロセス）と，観測値を得るまでの観察プロセスを明示的に分離した階層ベイズモデルを利用した．

まず初めに生態プロセスを考える．調査エリア内に，空間的に独立なI個の自動撮影カメラが設置されているとする．あるカメラ$i(i = 1, \dots, I)$によって撮影される可能性のある（カメラと行動圏が重複する）集団数K_iは，パラメータμをもつポアソン分布に従うとする．また，各集団の真の集団サイズ$N_{i,k}$ $(k = 1, \dots, K_i)$も，パラメータλをもつポアソン分布に従うとする．

$$K_i \sim Poisson(\mu)$$

$$N_{i,k} \sim Poisson(\lambda)$$

次に，観測プロセスを考える．i番目のカメラの前を利用するk番目の集団がカメラに観測される回数$C_{i,k}$もやはりパラメータυをもつポアソン分布に従うと仮定する．従って，i番目のカメラでは，$L_i(=\sum_{k=1}^{K_i} C_{i,k})$ 回の観測が得られることになる．また，観測集団サイズ$y_{i,j}(j = 1, \ldots, L_i)$ は，個体の検出が独立ではないことを想定して，真の集団サイズ$N_{i,k}$とα，βをパラメータにもち，二項指標変数$z_{i,j}$を導入したベータ二項分布と見なす．

$$y_{i,j} \sim Binomial(N_{i,k}, p_{i,j})$$
$$p_{i,j} = p'_{i,j} \times z_{i,j} + 1.0 \times (1 - z_{i,j})$$
$$p'_{i,j} \sim Beta(\alpha, \beta)$$
$$z_{i,j} \sim Bernoulli(\Omega)$$

$z_{i,j}$は，Ωをパラメータとするベルヌーイ分布に従うとする．この変数を導入したのは，野生動物の集団は，集合性が低い状態（$z_{i,j} = 1$）と高い状態（$z_{i,j} = 0$）に二分できることが多いと考えられたためである（例えば，採食時は個体間距離が離れやすいのに対し，移動時は1列になり集合性が高くなる）．ここでは，$z_{i,j} = 0$のときに検出率は 1 とした．すなわち集合性が十分に高い場合は，カメラに全ての個体が撮影されると仮定した．

このモデルを既存の確率的プログラミング言語（Jags や Stan などの MCMC を利用するためのソフトウェア）で実装するためには工夫が必要になる．なぜなら，MCMC のステップ毎に集団数K_iの値が変化し，それに伴い真の集団サイズ$N_{i,k}$の長さも変化するためである．そこで，より効率的な推定手法として，近似ベイズ計算（Approximate Bayesian Computation, 以下 ABC）を利用した．ABC は，適切な要約統計量を選択し，データ生成可能なモデルであれば，尤度関数を評価せずにパラメータの事後分布を推定できる（小林 2011）．ここでは推定効率を考え，逐次モンテカルロ法を応用した ABC-SMC（Toni et al. 2009）を利用することにし，観測集団サイズ$y_{i,j}$と観測回数L_iの平均と分散を要約統計量として使用した．

当日の発表では，モンテカルロ・シミュレーションによるモデルの性能を検証した結果を報告した．また，千葉県房総半島で実際に撮影されたイノシシの集団に適用した結果についても報告した．

参考文献

[1]. 小林弦矢.「ABC サンプラーの最近の発展について」. 日本統計学会誌, **41**, pp. 155–180, 2011.

[2]. Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E. and Boutin, S. 2015. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology* **52**: 675–685.

[3]. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

環境DNAによる生物多様性評価のための階層モデルとベイズ決定分析

深谷 肇一, 今藤 夏子, 松崎 慎一郎, 角谷 拓

国立環境研究所

fukaya.keiichi@nies.go.jp

環境 DNA（eDNA）メタバーコーディングは，非侵襲的かつ費用対効果の高い生物多様性の計測方法として広く応用されている．しかし，その多段階のワークフローの中では様々な要因で偽陰性が生じるため，eDNA メタバーコーディングによる種の検出は完全ではない．eDNA メタバーコーディングの多段階のワークフローにおける不完全な検出はまた，研究デザインの問題とも関連する．具体的には，利用可能な資源を各段階にどう配分すれば調査効率を最適化できるかという問題である．

本研究では，対象地域内の複数のサイトで標本を収集する eDNA メタバーコーディング研究を対象に，多種サイト占有モデルの拡張を提案する (Fukaya *et al.* 2021)．二値の検出データを扱う従来の多種サイト占有モデル (Dorazio & Royle 2005) とは異なり，このモデルはハイスループットシーケンサーの出力である配列リード数の変動を記述するものである．提案モデルは eDNA メタバーコーディングの階層的なワークフローと種間の異質性を明示的に説明する部分モデルから構成された階層モデルであり（Fig. 1），ワークフローの異なる段階における種の検出可能性の変動要因を分析可能である．また，ベイズ決定分析の枠組みを導入することで，限られた予算の下で種検出の有効性を最適化する研究デザインをモデルとそのパラメータの事後分布に基づいて特定できる．

具体的に，関心のある調査領域から J 個のサイトが選択され，eDNA メタバーコーディングを用いて合計 I 種の出現が調査される状況を想定する．サイト j $(= 1, \ldots, J)$ で得られた反復標本 k $(= 1, \ldots, K_j)$ における各種の配列リード計数を $\boldsymbol{y}_{jk} = (y_{1jk}, \ldots, y_{Ijk})$，シーケンス深度を $N_{jk} = \sum_i y_{ijk}$ と表す．モデルは以下のように表される．

$$\boldsymbol{y}_{jk} \sim \text{Multinomial}(N_{jk}, \boldsymbol{\pi}_{jk}) \tag{1}$$

$$\boldsymbol{\pi}_{jk} = \frac{u_{ijk} r_{ijk}}{\sum_{m=1}^{I} u_{mjk} r_{mjk}} \tag{2}$$

$$r_{ijk} \sim \text{Gamma}(\phi_i, 1) \tag{3}$$

$$u_{ijk} \sim \text{Bernoulli}(z_{ij} \theta_i) \tag{4}$$

$$z_{ijk} \sim \text{Bernoulli}(\psi_i) \tag{5}$$

$$(\log \phi_i, \text{logit } \theta_i, \text{logit } \psi_i) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6}$$

u_{ijk} と r_{ijk} はそれぞれ，サイト j の反復 k における種 i の DNA 配列の有無と，（配列が反復に含まれる場合は）その相対頻度を表す潜在変数であり，z_{ij} はサイト j における種 i のサイト占有を表す潜在変数である．

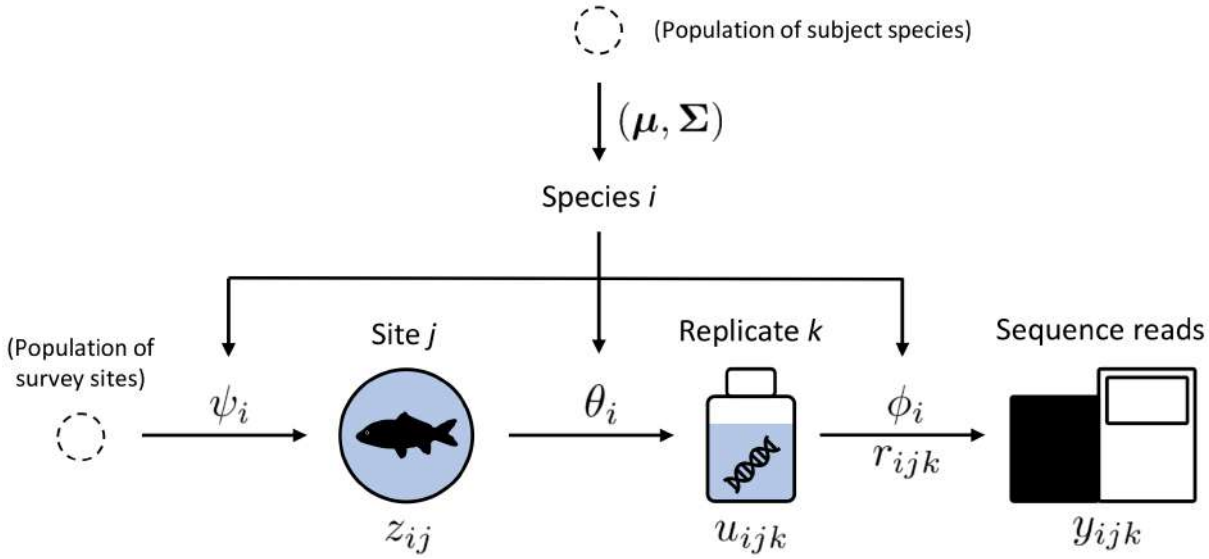


Fig. 1. モデルの概要図．潜在変数 z（種のサイト占有を表す），潜在変数 u（反復標本における種の DNA 配列の有無を表す），データ y（配列リード数）の間に，種レベルのパラメータ（サイト占有確率 ψ，配列捕捉確率 θ，配列相対優占度 ϕ）で特徴付けられる階層的な依存構造を仮定する．配列リード数は，反復ごとの各種の配列の有無（u）と優占度（r）を条件とした多項分布に従うと仮定する．種レベルのパラメータは，平均ベクトル $\boldsymbol{\mu}$ と共分散行列 $\boldsymbol{\Sigma}$ を持つ群集レベルの正規事前分布に従うと仮定する．Fukaya *et al.* (2021)，CC ライセンス 表示 4.0 国際.

ϕ_i，θ_i，ψ_i はそれぞれ，種の配列の相対優占度，捕捉確率，および種のサイト占有確率と解釈でき，いずれもワークフローの異なる段階における検出可能性の変動を特徴付ける種固有のパラメータである（Fig. 1）．

このモデルを霞ヶ浦流域の淡水魚群集に適用したところ，種の検出可能性には顕著な不均一性があり，特定の種が偏って検出される恐れがあることが明らかとなった．サイト占有確率が低い種は捕捉確率が低く，配列の優占度も低いため，検出されにくい傾向にあった．期待される配列リード数は種によって最大 23.5 倍も異なると予測された．この結果に基づいて異なる研究デザインの間で種検出の有効性（期待される検出種数）を予測した結果，各反復に数万リードが確保されている限り，サイト内反復を複数確保することが種検出の有効性を高める上で好ましいことが示唆された．

本枠組みによって，eDNA メタバーコーディングがより誤差に対して頑健なものとなり，生物群集をより効率的にモニタリングできるようになることが期待される．

References

Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.

Fukaya, K., Kondo, N.I., Matsuzaki, S.S. & Kadoya, T. (2021) Multispecies site occupancy modeling and study design for spatially replicated environmental DNA metabarcoding. *Methods in Ecology and Evolution*. (to appear)

Quantifying biodiversity change and its implication

Hideyasu SHIMADZU
Department of Mathematical Sciences, Loughborough University, UK
H.Shimadzu@lboro.ac.uk

Introduction

Quantifying biodiversity and its change has been a central interest in ecological sciences, and a large number of metrics have been developed. The modern foundation of measuring biodiversity commonly consists of three ways:

α-diversity: the number of species found within a location;

β-diversity: the number of species commonly (not) found between locations; and

γ-diversity: the number of species in a whole region.

Ideally, biodiversity metrics representing these three aspects should recognise the dimensionality of change and the ecological processes that lead to shifts in community composition. However, it is not always the case and, in fact, even widely used measures such as the number of species can fail to capture some changes (Shimadzu, 2018). This fact underlines an urgent need for a better understanding of the nature of those metrics in terms of depicting biodiversity changes.

Temporal changes in ecological communities

Let $\lambda_i(t) > 0$ be the expected abundance of the i-th species at time t. The expected total-abundance of the community is then the sum of each species' expected abundance as $\lambda(t) = \sum_{i=1}^{s} \lambda_i(t)$. Consider now an equation for the expected abundance of the i-th species as

$$\lambda_i(t + dt) = (1 + \alpha_i(t)dt)\lambda_i(t), \tag{1}$$

where $\alpha_i(t)$ is the instantaneous change rate at time t that drives the increase or the decrease of species abundance. The change rate is a legitimate indicator to quantify, literally, the turnover of the i-th species, based on its abundance, for the time increment dt. However, as the change rate is in most cases unknown, it needs to be determined by the trajectory of the expected abundance, $\lambda_i(t) > 0$, itself as

$$\alpha_i(t)dt = \frac{\lambda_i(t + dt) - \lambda_i(t)}{\lambda_i(t)} = \frac{d\lambda_i(t)}{\lambda_i(t)} = d\log\left(\lambda_i(t)\right).$$

Although the continuous form of the expected abundances, $\lambda_i(t)$, $i = 1, 2, \ldots, s$, is unknown, we can estimate it using observations at discrete time points. We therefore rewrite Equation (1) in a discrete form representing a community dynamics model over a relatively short time interval, between times t and $t + h$ ($h > 0$), within which the difference equation is still valid. By integrating the instantaneous change rate, $\alpha_i(t)$, over the time interval, the community dynamics can be described as

$$\lambda(t + h) \doteq \sum_{i=1}^{s}\left(1 + \int_{t}^{t+h} \alpha_i(v)dv\right)\lambda_i(t) = \sum_{i=1}^{s}\left(1 + \log\left(\frac{\lambda_i(t + h)}{\lambda_i(t)}\right)\right)\lambda_i(t).$$

Note that $\lambda(t) = \sum_{i=1}^{s} \lambda_i(t)$. The change rate and the instantaneous change rate may therefore hold the following relationship for the short time period h,

$$\frac{\lambda(t+h) - \lambda(t)}{\lambda(t)} \doteq \sum_{i=1}^{s} \log\left(\frac{\lambda_i(t+h)}{\lambda_i(t)}\right) p_i(t), \tag{2}$$

where $p_i(t) = \lambda_i(t)/\lambda(t)$ is the relative abundance of the i-th species at time t. The collection of the relative abundances of the community — the relative abundance distribution $\boldsymbol{p}(t) = \{p_1(t), p_2(t), \ldots, p_s(t)\}$ — satisfies the conditions that $p_i(t) > 0$ and $\sum_{i=1}^{s} p_i(t) = 1$. From Equation (2), Shimadzu et al. (2015) define the turnover measure, D, between times t and $u, (u > t)$ as

$$\begin{aligned} D(t:u) = \sum_{i=1}^{s} d_i(t:u) &= \sum_{i=1}^{s} \log\left(\frac{\lambda_i(u)}{\lambda_i(t)}\right) p_i(t) \\ &= -\sum_{i=1}^{s} \log\left(\frac{p_i(t)}{p_i(u)}\right) p_i(t) + \log\left(\frac{\lambda(u)}{\lambda(t)}\right) \\ &= D_1(\boldsymbol{p}(t):\boldsymbol{p}(u)) + D_2(\lambda(t):\lambda(u)). \end{aligned} \tag{3}$$

If the expected abundance, $\lambda_i(t)$, is modelled via GLMs with the log link function as

$$\log(\lambda_i(t)) = \sum_{j=1}^{m} \beta_{ij} x_j(t),$$

where $\{x_j(t)\}$ are environmental variables, and $\{\beta_{ij}\}$ are the parameters to be estimated, Equation (3) reveals that it becomes possible to identify drivers that influence the turnover measure, D, in an additive form as

$$D(t:u) = \sum_{j=1}^{m} \sum_{i=1}^{s} \beta_{ij}(x_j(u) - x_j(t))p_i(t).$$

Since the contribution of the species and the environmental factors are all additive, putting

$$\begin{aligned} d_i(t:u) &= \sum_{j=1}^{m} \beta_{ij}(x_j(u) - x_j(t))p_i(t) \text{ and} \\ d_j(t:u) &= \sum_{i=1}^{s} \beta_{ij}(x_j(u) - x_j(t))p_i(t), \end{aligned}$$

the contribution of the i-species and of the j-th environment variable to the turnover measure, D, is respectively identified.

References

Shimadzu, H. (2018). On species richness and rarefaction: size- and coverage-based techniques quantify different characteristics of richness change in biodiversity, *Journal of Mathematical Biology* **77**: 1363–1381. http://dx.doi.org/10.1007/s00285-018-1255-5.

Shimadzu, H., Dornelas, M. and Magurran, A. E. (2015). Measuring temporal turnover in ecological communities, *Methods in Ecology and Evolution* **6**(12): 1384–1394. http://dx.doi.org/10.1111/2041-210x.12438.

定量的生態学と統計モデル：

食われる側からみた特定食う種の貢献度

島谷健一郎 (統計数理研究所)

特定の天敵による死亡は、食われる側の死亡の中でどのくらいの割合を占めるか。Hoso and Shimatani (2020) では、野外における標識・再捕獲データと実験室で得た捕食成功データを階層ベイズモデルの中で統合することで、あるカタツムリ種が、天敵の１つであるヘビに、成熟して繁殖に至るまでどのくらい捕食されるかを定量的に推定した。さらに、ヘビに対してはトカゲのようにシッポを自切することで捕食を回避する場合もあり、それが成熟数をどのくらい増やすかにより、進化的適応度を推定した。

このような定量的評価は、科学における営みの基本のはずである。にもかかわらず、現実にはこうした成果はなかなか伝統ある生態学の学術誌には掲載されない。その原因として、以下のような理由が考えられる。

１．　生態学が自然科学として未成熟で、多くの仮説や法則性は定性的記述にとどまる。定量的生態学の必要性は認識されていても、どう実践するか、科学方法論（科学哲学）が確立されていない。
２．　統計モデルを用いて未知数量を推定したときの科学方法論（科学哲学）が確立していない。そのため、同様な問題は、統計モデルを用いた推定を要する科学全般に見られる。
３．　両者を融合させた、生態学において統計モデルを使って定量的仮説を検証する科学方法論（科学哲学）が確立されていない以前に、その必要性が認識されていない。

国内ではあまり議論されていないが、生態学の科学哲学は古くから議論されている。ただ、その多くは数理生態学に軸足を置いており、生態学における数理モデルの役割や、物理学と同じような科学方法論が適用可能か、などを論じている。2000 年代以降、先端の統計モデルは生態学で普通に適用されているが、そんな統計学を含めて論じている生態学の科学哲学論文は見当たらない。

統計学の科学哲学も古くから研究されている。ただその大半は、今日に至るもなお古典検定と古典ベイズ（事前共役分布を用いる場合など事後確率が計算可能な場合）に関するもので、今日の階層ベイズモデルや状態空間モデルに関する科学哲学は見当たらない。それどころか、最近のマルコフ連鎖モンテカルロ法

（MCMC）を用いるベイズモデルと古典ベイズを混同したような論文が公表されている。

逆に、統計学の業界から、階層ベイズモデルを用いる科学哲学も提起されている。ただ、そこでベースとされるのは、ポパー（反証可能な仮説が科学的仮説、などを主張）の時代、すなわちフィッシャーらにより統計学が確立されていく時代以前の科学哲学にとどまりがちで、その後 50 年以上に及ぶ科学哲学の進展を踏まえた内容とは言い難い。

このような状況では、先端統計モデルを用いる生態学の科学哲学に関する議論が喚起されなくて当然である。このような現状についてまとめたのが、島谷（2021）や Shimatani (2021) である。

本講演では、最初に階層ベイズモデルを用いて定量的評価を行った研究事例を紹介し、それから、統計モデルを用いる生態学の科学哲学の未成熟さについて検討し、現在進行中の試みと展望について議論した。

参考文献

Hoso Masaki, Shimatani K. Ichiro. (2020) Life-History Modeling Reveals the Ecological and Evolutionary Significance of Autotomy. The American Naturalist 196: 690-703.

島谷健一郎 (2021) 推定目的の統計モデルを用いる帰納推論における数学・演繹の位置付けと役割. Linkage.

Shimatani, I.K. (2021) Philosophy of statistical sciences: The roles of mathematics and statistical model in estimation and other inductive inferences. 科学基礎論学会誌

ニワトリ雛の採餌行動解析―理論最適解から乖離する理由を探す―

統計数理研究所　川森愛

[背景]

言語を持たない動物が何を考えて行動を決定するのか，その生々しい思考過程を知ることは容易ではない．しかし，思考の産物である行動を説明することによって，ある程度窺い知ることはできるはずである． 本発表ではニワトリ雛の採餌行動データを対象とし，複数の認知モデルを比較検討することで認知メカニズムの探究を試みる．

採餌における基本の行動原理は効率の最大化である．特に，餌場がパッチ状に分布する時の餌場滞在時間を考えるモデルとして，最適餌パッチ利用モデル(Marginal Value Theorem, 以下 MVT と表記)が提唱されている[1]．餌場にある餌は有限であるため，探索とともに発見に時間がかかるようになる．この状態を収益逓減と呼ぶ．一方，餌場間の移動中は収益 0 の大損状態なので，頻繁に餌場を変えるのも得策ではない．そのため，一つの餌場にいつまで滞在するかの最適点が問題となるのである．

今，収益逓減のルールとして餌と餌の時間間隔が指数関数的に増加すると考える．具体的に， n 個目と n-1 個目の餌の間隔が ar^{n-2} の指数関数で増加するとする．この時，MVT から予想される最適滞在時間 S*を求める式は大変複雑であり，このような計算を実際の動物が実装できているかは疑問である．しかしながら，行動データを見る限り理論を全く無視しているわけでもない．何らかの簡易的プロセスを用いて近似計算を行い，だいたいの最適性を実現していると考える方が現実的である．

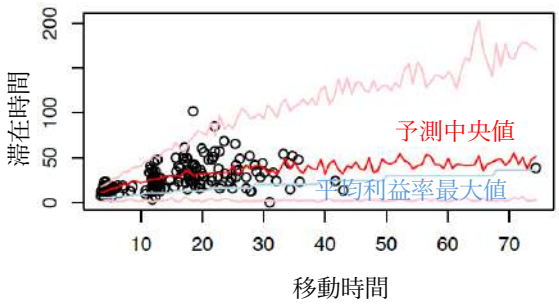
[データ]

行動実験では，両端に餌場がある I 字型装置内を用い，ニワトリ雛に自由探索させた．餌場に到達すると自動的に粟が一粒ずつ供給されるが，供給までの待ち時間は指数関数的に増加する．装置の足場はトレッドミルとなっており，移動距離を人為的に 3 種類の長さに操作した．このように MVT が想定する餌環境を再現した実験で，ニワトリ雛が餌場に入ってから離脱するまでの滞在時間とその餌場に入るまでの移動時間の間の関係を調べた．ニワトリ雛の装置内座標（1 次元データ）は超音波センサーにより自動的に，ニワトリ雛の行動とは無関係におよそ一定間隔で記録された．従ってデータとしては，滞在開始時刻として餌供給開始時刻(f)，餌場内での最終観測時刻(l)および餌場外での最初期観測時刻(m)の 3 つの時刻がある．また，離脱から次の滞在開始までの時間を移動時間(t)として算出した．これらのデータを用いてヒヨコの離脱時刻を説明する確率的意思決定モデルを作成し，行動データとのフィッティングを調べた．

[行動データへのフィッティング比較]

作成したモデルに対しパラメータを最尤推定した上で，移動時間 1 点につき 100 回のシ

ミュレーションを行うことにより，滞在時間の予測値を求めた．最も成功したモデルの予測結果を一例として図に示す．赤線が予測中央値，上下のピンク線が予測 95%範囲，青線は最適滞在時間の近似値，黒丸は実際の行動データを示している．行動データは移動時間とともに分散が増大していく特徴がみられるが，このモデルではその様子をうまく捉えられている．



（図）モデルに基づいた滞在時間予測

複数の認知アルゴリズムを仮定したモデルを作成し検証した結果，ニワトリ雛が理論通りの最適滞在時間そのものを計算できている可能性は低いと判断できた．また，採餌効率を随時モニタリングするアルゴリズムにより滞在時間を決めると想定したモデルは全般的に，行動の分散を説明できる点で良いモデルであった．さらに，そのようなモデルの中でもどの変数を使用するかが重要であり，瞬間利益率のように認知的負荷が小さい指標を使用するモデルがニワトリ雛の行動にはよく当てはまった．

参考文献

[1] E. L. Charnov (1976) “Optimal foraging: the marginal value theorem”, Theoretical Population Biology, 9:129–136

ランク落ちでの無罰則推定法と不偏推定量

永井 勇[†]

[†] 中京大学 教養教育研究院

2021 年 10 月 22 日-24 日
予測モデリングの理論と応用

概要

- 多変量線形回帰モデルでよく置かれる仮定

仮定
説明変数からなる行列が列フルランク

- この仮定を満たさないデータの分析のために様々な手法が提案されている
  - 例 Lasso 型罰則付推定

問題点 最適化のための反復計算が必要・不偏性がない

⇒ 反復計算が不要な手法での新しい推定法の提案

- 推定量が**陽に求まる**ように一工夫した結果
  - 提案する推定量の問題点など

⇒ 改良した推定量の構築

- 反復計算が不要で, 列フルランク性がない場合に, 不偏性を持つ推定量の構築

目次

ここからの話

モデルとリスク関数

- 多変量線形回帰モデル; $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}}$
  $\boldsymbol{Y}$; $n \times p$ 目的変数行列
  $\mathbf{1}_n$; 1 が n 個並んだ定数ベクトル
  $\boldsymbol{\mu}$; p 次元未知ベクトル
  $\boldsymbol{A}$; $n \times k$ 説明変数行列
  $\boldsymbol{\Xi}$; $k \times p$ 未知行列
  $\boldsymbol{\mathcal{E}}$; $n \times p$ 誤差行列
  - $p = 1$; 重回帰モデル

仮定 $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$, $\mathbf{0}_k$ は k 次元の 0 ベクトル, $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$, $\mathrm{rank}(\boldsymbol{\Sigma}) = p$ ($\boldsymbol{\Sigma}$ は未知)

リスク $R(\boldsymbol{\mu}, \boldsymbol{\Xi}) \overset{\text{def.}}{=} \mathrm{tr}\{(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{A}\boldsymbol{\Xi})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{A}\boldsymbol{\Xi})'\}$

推定 最小二乗推定量; $\arg\min_{\boldsymbol{\mu},\boldsymbol{\Xi}} \mathrm{R}(\boldsymbol{\mu}, \boldsymbol{\Xi})$

従来の推定での問題点と本研究の目的

推定量 $\hat{\boldsymbol{\mu}} = \boldsymbol{Y}'\mathbf{1}_n/n$, $\boldsymbol{A}'\boldsymbol{A}\hat{\boldsymbol{\Xi}} = \boldsymbol{A}'\boldsymbol{Y}$ の解

- $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Xi}}) = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Xi}} R(\boldsymbol{\mu}, \boldsymbol{\Xi})$

→ $\mathrm{rank}(\boldsymbol{A}) = k$ ならば $\hat{\boldsymbol{\Xi}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}$

よく置かれる仮定
$\mathrm{rank}(\boldsymbol{A}) = k$

つまり 説明変数からなる行列が列フルランク

問題点 $\mathrm{rank}(\boldsymbol{A}) < k$ の際に, どう $\boldsymbol{\Xi}$ を推定する？

例 $n < k$ のときの推定方法

本研究の大まかな目的
$\mathrm{rank}(\boldsymbol{A}) < k$ のときの $\boldsymbol{\Xi}$ の推定法の構築
- 反復計算が**不要**な推定法

$\mathrm{rank}(\boldsymbol{A}) < k$ の状況での従来の推定法

- $\mathrm{rank}(\boldsymbol{A}) < k$ での $\boldsymbol{\Xi}$ の従来の推定法
  - $(\boldsymbol{A}'\boldsymbol{A})^{-1}$ が存在しない際の $\boldsymbol{\Xi}$ の推定法

(i) 罰則付推定法
  例 Lasso 型 (Katayama & Imori, 2014), Ridge 型 (Yanagihara & Satoh, 2010), elastic net 型 (Lasso + Ridge) など

(ii) 変数選択 (Oda & Yanagihara, 2021) してから推定
  - $\boldsymbol{A}$ の列数 (k) を減らす

(iii) 主成分分析を用いてから推定
  - $\boldsymbol{A}$ を主成分分析後に推定
  名前 主成分回帰 (Fujiwara, Minamidani, Nagai & Wakaki, 2013)

(iv) 一般逆行列を用いて推定
  別名 疑似逆行列 (e.g., Srivastava, 2002, A.6)

本研究 **もっとシンプルな**推定法の提案

ここからの話

モデルとリスク関数の書き換え

元 $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}} \to \boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)$ と分割

- $\boldsymbol{A}$ の分割に対応して $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1', \ldots, \boldsymbol{\Xi}_r')'$ と分割
- $\boldsymbol{A}_i$; $n \times k_i$ 行列, $1 \le r \le k$, $k_1 + \cdots + k_r = k$
  $\boldsymbol{\Xi}_i$; $k_i \times p$ 行列 (未知)

仮定 $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ $(i = 1, \ldots, r)$

$$\to \boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i + \boldsymbol{\mathcal{E}}$$

Note $\boldsymbol{\Xi}$ を推定 $\Leftrightarrow$ $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ を推定

リスク $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r) \overset{\text{def.}}{=}$

$$\mathrm{tr}\left\{\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i\right)'\right\}$$

Note $R(\boldsymbol{\mu}, \boldsymbol{\Xi}) = R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$

書き換えたリスク関数の問題点

- $\hat{\boldsymbol{\Xi}}_i \overset{\text{def.}}{=} \arg\min_{\boldsymbol{\Xi}_i} R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ $(i = 1, \ldots, r)$

問題点 $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ を直接展開した場合の問題点

直接計算した際の問題点
$\hat{\boldsymbol{\Xi}}_i$ のために $\hat{\boldsymbol{\Xi}}_j$ $(i \neq j)$ が**全部必要**

- 初期値設定が必要
- $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ が収束するまで繰り返し計算が必要

→ $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ の展開を一工夫すると,
$\boldsymbol{\Xi}_i$ に関連する項 $= \boldsymbol{\Xi}_j$ $(i < j)$ **だけ**が含まれる形

工夫して計算した結果
$\hat{\boldsymbol{\Xi}}_i$ のために $\hat{\boldsymbol{\Xi}}_j$ $(i < j)$ **だけ**が必要

提案する推定量とその利点

- $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ の展開を一工夫して最小化の結果

$\hat{\boldsymbol{\Xi}}_i$ $(i = 1, \ldots, r)$ を求めた結果

$$\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right),$$

ただし, $\sum_{j>r}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j = \mathbf{0}_n\mathbf{0}_p'$.

結局 r 番目から順に**陽に**求まる
- $\boldsymbol{\Xi}_1$ から順に求まるようにもできる

利点 $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ $(i = 1, \ldots, r)$ は $\mathrm{rank}(\boldsymbol{A}) = k$ より**緩い**
$\mathrm{rank}(\boldsymbol{A}) = k \Rightarrow \mathrm{rank}(\boldsymbol{A}_i) = k_i$
$\mathrm{rank}(\boldsymbol{A}_i) = k_i \not\Rightarrow \mathrm{rank}(\boldsymbol{A}) = k$

提案する推定量の問題点など

- $\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right)$ の問題点

(I) r の選択
- r を大きくする → 仮定を満たしやすい
  $\leftrightarrow$ $\boldsymbol{A}$ の構造をあまり維持しない
  r を小さくする → 仮定を満たしにくい
  $\leftrightarrow$ $\boldsymbol{A}$ の構造をなるべく維持する

仮定 $\mathrm{rank}(\boldsymbol{A}_i) = k_i$

(II) $\boldsymbol{A}_i$ $(i = 1, \ldots, r)$ の選択
- $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ の条件下での選択

(III) $\hat{\boldsymbol{\Xi}}_i$ は不偏推定量**ではない**
後述 $(\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'$ の項を工夫すると, 不偏推定量にもなる

ここからの話

ここからの話

一般逆行列を用いた場合と同等?!

- 提案した推定量;

$$\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right)$$

注意点 $E[\hat{\boldsymbol{\Xi}}_i] \neq \boldsymbol{\Xi}_i$ $(i = 1, \ldots, r)$

→ $E[\boldsymbol{Y}] = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi}$ の不偏推定量**ではない**

$\Leftrightarrow$ 一般逆行列を用いた場合; $E[\boldsymbol{Y}]$ の不偏推定量

不偏推定量ではない $\hat{\boldsymbol{\Xi}}_i$ $(i = 1, \ldots, r)$ で,
不偏推定量で予測するのとほぼ**同等の予測精度**

⇒ $\boldsymbol{\Xi}_i$ の不偏推定量を用いれば...
より予測精度を上げられるのでは？

改良結果

- 求め方を変えると，
  - 改めて $\tilde{\Xi}_i = \arg\min_{\Xi} R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ $(i = 1, \ldots, r)$

$\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ を求めた結果

$$\tilde{\Xi}_i = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\tilde{\Xi}_j\right),$$

ただし，$\sum_{j>r} \boldsymbol{A}_j\tilde{\Xi}_j = \boldsymbol{0}_n\boldsymbol{0}_p'$，ここで $\boldsymbol{Q}_1 = \boldsymbol{A}_1$，

$\boldsymbol{Q}_i = \left(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{\boldsymbol{Q}_j}\right)\boldsymbol{A}_i$ $(i > 1)$ とし，$\boldsymbol{P}_M = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'$ であり，$\mathrm{rank}(\boldsymbol{Q}_i) = k_i$ を仮定．

$\tilde{\Xi}_i$ の性質; 不偏推定量

一工夫の実際

- $R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ の展開結果 ($\Xi_1, \ldots, \Xi_r$ に関する項);

$$-2\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_i') + \sum_{i=1}^{r}\sum_{j=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

- $\mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j') = \mathrm{tr}(\boldsymbol{A}_j\Xi_j\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i')$ より
  この項=対称行列の成分の総和

Step 1 対角成分の和と非対角成分の和で分けると，

$$\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i') + \sum_{i=1}^{r}\sum_{j\neq i}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

Step 2 この項 $= 2\sum_{i=1}^{r}\sum_{j>i}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$

Step 3 $\hat{\Xi}_j$ $(j > \ell)$ を既知として $\hat{\Xi}_\ell$ を求めた

一工夫の途中までで止めてみる

- Step 1 (対角成分の和と非対角成分の和で分けた形) で止めると，
  $R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ を展開結果 ($\Xi_1, \ldots, \Xi_r$ に関する項);

$$-2\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_i') + \sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i')$$
$$+\sum_{i=1}^{r}\sum_{j\neq i} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

→ Ξ_ℓ に関する項だけピックアップする
この項は $i = \ell \neq j$ と $i \neq \ell = j$ と $i \neq \ell \neq j$ で分割

$$-2\mathrm{tr}(\boldsymbol{A}_\ell'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_\ell') + \mathrm{tr}(\boldsymbol{A}_\ell\Xi_\ell\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$
$$+2\sum_{i\neq\ell} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$

→ これを最小にする Ξ_ℓ を求める

$\tilde{\Xi}_\ell$ とその変形 (1/3)

- Ξ_ℓ について，最小化する関数;

$$-2\mathrm{tr}(\boldsymbol{A}_\ell'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_\ell') + \mathrm{tr}(\boldsymbol{A}_\ell\Xi_\ell\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$
$$+2\sum_{i\neq\ell} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$

→ $\boldsymbol{A}_\ell'\boldsymbol{A}_\ell\tilde{\Xi}_\ell = \boldsymbol{A}_\ell'\left(\boldsymbol{Y} - \sum_{i\neq\ell} \boldsymbol{A}_i\tilde{\Xi}_i\right)$ $(\ell = 1, \ldots, r)$ の解

→ $\tilde{\Xi}_1 = (\boldsymbol{A}_1'\boldsymbol{A}_1)^{-1}\boldsymbol{A}_1'\left(\boldsymbol{Y} - \sum_{i>1} \boldsymbol{A}_i\tilde{\Xi}_i\right)$，

$\boldsymbol{A}_2'\boldsymbol{A}_2\tilde{\Xi}_2 = \boldsymbol{A}_2'\left(\boldsymbol{Y} - \boldsymbol{A}_1\tilde{\Xi}_1 - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right)$ の解．

$\tilde{\Xi}_\ell$ の変形 (2/3) と仮定

- $\tilde{\Xi}_1$ を $\tilde{\Xi}_2$ の式に代入して変形すると

$$\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2\tilde{\Xi}_2 = \boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\left(\boldsymbol{Y} - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right)$$

の解．ここで，$\boldsymbol{P}_M = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'$．

仮定 $\mathrm{rank}((\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2) = k_2$ を仮定すると，

$$\tilde{\Xi}_2 = \{\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2\}^{-1}\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\left(\boldsymbol{Y} - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right).$$

⇒ $\mathrm{rank}((\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2})\boldsymbol{A}_3) = k_3$ も仮定すると，

$$\tilde{\Xi}_3 = \{\boldsymbol{A}_3'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2})\boldsymbol{A}_3\}^{-1} \times$$
$$\boldsymbol{A}_3'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}_1})\boldsymbol{A}_2})\left(\boldsymbol{Y} - \sum_{i>3} \boldsymbol{A}_i\tilde{\Xi}_i\right).$$

$\tilde{\Xi}_\ell$ の変形 (3/3) の結果

- 繰り返していくと，

$\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ を求めた結果

$$\tilde{\Xi}_i = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\tilde{\Xi}_j\right),$$

ただし，$\sum_{j>r} \boldsymbol{A}_j\tilde{\Xi}_j = \boldsymbol{0}_n\boldsymbol{0}_p'$，ここで $\boldsymbol{Q}_1 = \boldsymbol{A}_1$，

$\boldsymbol{Q}_i = \left(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{\boldsymbol{Q}_j}\right)\boldsymbol{A}_i$ $(i > 1)$ とし，$\boldsymbol{P}_M = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'$ であり，$\mathrm{rank}(\boldsymbol{Q}_i) = k_i$ を仮定．

Note $(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{\boldsymbol{Q}_j})$ は射影行列

注意点 $\mathrm{rank}(\boldsymbol{Q}_r) = k_r$ のために $n \geq k$ が必要

変形結果の性質

- $\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ の利点

$\tilde{\Xi}_i$ の性質; 不偏性

$E[\tilde{\Xi}_i] = \Xi_i$ $(i = 1, \ldots, r)$

Note $E[\tilde{\Xi}_i] \neq \Xi_i$ $(i = 1, \ldots, r)$

- $E[(\tilde{\Xi}_1', \ldots, \tilde{\Xi}_r')'] = (\Xi_1', \ldots, \Xi_r')' = \Xi$

⇒ $\mathrm{rank}(\boldsymbol{A}) < k$ でも
Ξ の不偏推定量が得られる

Note $E[\boldsymbol{Y}] = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\Xi$ の不偏推定量でもある

Note 一般逆行列を用いた場合は，
$E[\boldsymbol{Y}]$ の不偏推定量だが，Ξ の不偏推定量でない

$E[\tilde{\Xi}_i] = \Xi_i$ の証明の概略

目的 $E[\tilde{\Xi}_i] = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'E\left[\boldsymbol{Y} - \sum_{j>i} \boldsymbol{A}_j\tilde{\Xi}_j\right] = \Xi_i$
$(i = 1, \ldots, r)$ を示す

1. $(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\boldsymbol{1}_n = \boldsymbol{0}_{k_i}$ $(i = 1, \ldots, r)$ となる
2. $(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\boldsymbol{A}_j = \boldsymbol{0}_{k_i}\boldsymbol{0}_{k_j}'$ $(j < i \leq r)$ となる
3. $(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\boldsymbol{A}_i = \boldsymbol{I}_{k_i}$ $(i = 1, \ldots, r)$ となる
4. $E\left[\boldsymbol{Y} - \sum_{j>i} \boldsymbol{A}_j\tilde{\Xi}_j\right] = \boldsymbol{1}_n\boldsymbol{\mu}' + \sum_{j<i} \boldsymbol{A}_j\Xi_j + \boldsymbol{A}_i\Xi_i$
となる

結果 $E[\tilde{\Xi}_i] = \boldsymbol{0}_{k_i}\boldsymbol{0}_p' + \sum_{j<i} \boldsymbol{0}_{k_i}\boldsymbol{0}_{k_j}'\Xi_j + \Xi_i = \Xi_i$ $(i = 1, \ldots, r)$.

ここからの話

まとめ

- 多変量線形回帰モデル $\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\Xi + \boldsymbol{\mathcal{E}}$ での
従来の推定の問題点

従来 $\mathrm{rank}(\boldsymbol{A}) = k$ を仮定している

→ $\mathrm{rank}(\boldsymbol{A}) < k$ のときの推定法を構築

従来 罰則付推定，$\boldsymbol{A}$ から変数選択して推定，
$\boldsymbol{A}$ を主成分分析して回帰，一般逆行列を使う

⇒ 目的；もっと**シンプルな**推定法の構築

- $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)$，$\Xi = (\Xi_1', \ldots, \Xi_r')'$ として，
Ξ の推定 $\Leftrightarrow$ $\Xi_1, \ldots, \Xi_r$ の推定

結果 仮定が**かなり**緩く，反復計算が**不要**な推定法
  - 分割数 (r) を大きくすれば，仮定をかなり緩和可能
- 仮定を追加して，Ξ の不偏推定量の構築

今後の課題

1. r ($\boldsymbol{A}$ の分割数) や $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ の中身の選択
2. $\mathrm{rank}(\boldsymbol{A}) < k$ での Ξ の不偏推定量の数値実験
3. $n \geq \sum_{i=1}^{j} k_i$ で $n < \sum_{i=1}^{j+1} k_i$ のような n で
Ξ の一部の不偏推定量の構築
   注意点 不偏推定量 $(\tilde{\Xi}_i)$ のために $n \geq k = \sum_{i=1}^{r} k_i$ が必要
   - ある程度の標本数で，途中までの不偏性を持つ推定量が
できないか？
4. 罰則付推定量の構築
   - $\mathrm{rank}(\boldsymbol{A}) < k$ で予測平均二乗誤差 (PMSE) を
より小さくする推定量が作れないか？
5. 他のモデル (GMANOVA モデル) などへの拡張

参考文献

1. Fujiwara, M., Minamidani, T., Nagai, I. and Wakaki, H. (2013). Principal components regression by using generalized principal components analysis. *J. Jpn. Stat. Soc.*, **43**, 57–78.
2. Katayama, S. & Imori, S. (2014) Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *J. Multivariate Anal.*, **132**, 138–150.
3. Oda, R. & Yanagihara, H. (2021) A consistent likelihood-based variable selection method in normal multivariate linear regression. *Smart Systems and IoT: Innovations in Computing*, (in press).
4. Srivastava, M. S. (2002) *Methods of Multivariate Statistics*, Wiley-Interscience.
5. Yanagihara, H. and Satoh, K. (2010). A unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.

終

A unified formulation of k-means, fuzzy c-means and Gaussian mixture model by the Kolmogorov-Nagumo average

Osamu Komori
Seikei University
Kichijoji Kitamachi, Musashino, Tokyo 180-8633, Japan
komori@st.seikei.ac.jp
Shinto Eguchi
The Institute of Statistical Mathematics
Midori-cho, Tachikawa Tokyo 190-8562, Japan
eguchi@ism.ac.jp

October 26, 2021

Abstract

Clustering is a major unsupervised learning algorithm and is widely applied in data mining and statistical data analyses. Typical examples include k-means, fuzzy c-means and Gaussian mixture models, which are categorized into hard, soft and model-based clusterings, respectively. We propose a new clustering, called Pareto clustering, based on the Kolmogorov-Nagumo average, which is defined by a survival function of the Pareto distribution. The proposed algorithm incorporates all the aforementioned clusterings plus maximum-entropy clustering. We introduce a probabilistic framework for the proposed method, in which the underlying distribution to give consistency is discussed. We build the minorize-maximization algorithm to estimate the parameters in Pareto clustering. We compare the performance with existing methods in simulation studies and in benchmark dataset analyses to demonstrate its highly practical utilities.

keywords: k-means; fuzzy c-means; Gaussian mixture model; maximum-entropy clustering; Kolmogorov-Nagumo average; generalized energy function; Pareto distribution

1 Introduction

In data analysis or data mining, there are two fundamental types of methodologies: clustering and classification (Rokach & Maimon, 2005). Clustering, which is categorized as an exploratory paradigm, detects the underlying structure behind the data and grasps the rough image before proceeding to more intensive and comprehensive data analysis (Tukey, 1980; Dubes & A.K.Jain, 1980). On the other hand, classification predicts unknown class labels of test data based on models constructed by training data with known class labels. The former is called supervised learning, while the latter is called unsupervised learning in pattern recognition (Jain, 2010).

Clustering algorithms fall roughly into three categories: hierarchical, partitioning, and mixture model-based algorithms (Ghosh & Dubey, 2013). In hierarchical clustering, each observation is considered as one cluster in the initial setting. Then clusters are merged recursively based on a similarity matrix defined beforehand. The resultant clusters are expressed as a dendrogram. The partition algorithm starts with a fixed number of clusters and searches for the cluster centers to minimize an objective function such as the squared distances between the centers and observations. It finds the centers simultaneously. The model-based algorithm assumes a mixture of probability distributions, which generates the observations, and assigns the distributions to one of the mixture components. A Gaussian-mixture distribution-based approach is widely used in this context.

In this paper, we propose a new clustering, called Pareto clustering in the framework of quasi-linear modeling (Komori *et al.*, 2016; Omae *et al.*, 2017; Komori *et al.*, 2017). It combines the cluster components by the Kolmogorov-Nagumo average (Naudts, 2011) in a flexible way. We consider a generalized energy function as an objective function to estimate cluster parameters, which is an extension of the energy function proposed by Rose *et al.* (1990). The objective function consists of a survival function of the Pareto distribution, which is widely used in extreme value theory (Beirlant *et al.*, 2004). We investigate the consistency of the parameters, resulting in the underlying probability distribution of the generalized energy function. We find that k-means (Cox, 1957; MacQueen, 1967) and fuzzy c-means (Bezdek *et al.*, 1984) have the underlying probability distributions with singular points at the cluster centers. This fact shows a clear difference from the model-based clustering such as a Gaussian-mixture modeling. Moreover, we show that the quasi-linear modeling based on the the Kolmogorov-Nagumo average connects k-means, fuzzy c-means and a Gaussian-mixture modeling using the hyper parameters of the generalized energy function. See (Hathaway & Bezdek, 1995; Yu, 2005) for the discussion of the relation between k-means and fuzzy c-means.