令和４年度科学研究費　基盤研究(A)（課題番号：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))
によるシンポジウム

多様な分野における統計科学の理論とその応用

シンポジウム報告書

令和４年１０月２７日（木） ～ １０月２９日（土）

研究分担者：田畑耕治（東京理科大学）
開催責任者：田畑耕治，安藤宗司，石井晶，中川智之（東京理科大学）

場所：
東京理科大学 野田キャンパス 7 号館 6 階 講堂とのハイブリッド（Zoom）

内容・目的：
多種多様な分野において現れる大規模複雑データの解析において，統計科学への期待と関心は高まる一方である．本シンポジウムでは，カテゴリカルデータ解析，医療統計学，高次元データ解析，ロバスト統計，ベイズ統計などの新しい理論や方法論に関する研究，実データを用いた事例研究，データサイエンス時代の統計教育法など幅広い分野からの講演を募集します．多様な分野の研究者が知識や意見を交換することで，数理データサイエンスにおける最新動向や問題点を共有し，問題解決の糸口を創出する場になることを目的とします．

共催：
東京理科大学，
東京理科大学データサイエンスセンター
東京理科大学 統計科学研究部門

科研費シンポジウム
「多様な分野における統計科学の理論とその応用」

日時：2022 年 10 月 27 日（木）〜 10 月 29 日（土）
場所：東京理科大学野田キャンパス
7 号館 6 階講堂とのハイブリッド（Zoom）

---

10 月 27 日（木）

12:30–13:00 開場
13:00–13:10 開会
13:10–15:20　セッション 1

◇「NDB サンプリングデータセットを用いた精神神経疾患の合併症の網羅的分析」
石井一夫　（公立諏訪東京理科大学, 久留米大学）
吉永泰周　（福岡歯科大学）
坂上竜資　（福岡歯科大学）
小路純央　（久留米大学）
森川渚　（久留米大学）
野原夢　（久留米大学）
野原正一郎　（久留米大学）
福本義弘　（久留米大学）

◇「連検定による重心動揺データの解析」
竹内直子　（大阪府立大学）
綿森葉子　（大阪公立大学）
久利彩子　（大阪河崎リハビリテーション大学）
有末伊織　（関西福祉科学大学）

◇「GMANOVA モデルにおける新たな経時変動の推定方法とその解釈」
永井勇　（中京大学）

15:20–15:30 休憩
15:30–17:40　セッション 2

◇「統計科学におけるエントロピー概念の誤り訂正」
得丸久文

◇「多様体学習による銀河進化の探求」
竹内努　（名古屋大学, 統計数理研究所）
Suchetha COORAY　（名古屋大学）

◇「Fréchet 距離を用いた分類問題について」
伊森晋平　（広島大学）
若木宏文　（広島大学）

10 月 28 日（金）

9:30–10:00 開場
10:00–12:10　セッション 1

◇「正方分割表における松下距離に基づく周辺同等性からの隔たり尺度」
篠田覚　(横浜市立大学)
吉本拓矢　（中外製薬株式会社）
田畑耕治　（東京理科大学）

◇「方向データのためのカーネル密度推定量のバイアス修正」
鶴田靖人　（長野県立大学）

◇「Outcome-adaptive lasso と adaptive lasso を用いた AIPW 推定量の性能評価」
本江渡　（東京理科大学大学院, ノバルティスファーマ株式会社）
安藤宗司　（東京理科大学）
土田潤　（同志社大学）
寒水孝司　（東京理科大学）

12:10–13:40　-お昼休憩-
13:40–15:50　セッション 2

◇「特異ウィシャート行列の固有値の正確分布論」
清水康希　（東京理科大学）
橋口博樹　（東京理科大学）

◇「ロジスティック回帰モデルにおける安定な非凸スパース正則化法」
塘由惟　(東京大学, 国立精神・神経医療研究センター)
小川光紀　(東京大学)
片井みゆき　(政策研究大学院大学医学, 東京女子医科大学)
大庭幸治　(東京大学)
松山裕　(東京大学)

◇「ブリッジ推定量を用いた BIC の妥当性とその周辺」
宮田庸一　（高崎経済大学）

15:50–16:00　-休憩-
16:00–18:10　セッション 3

◇「ロジスティック分布における母数推定について」
作村建紀　（法政大学）
柳本武美　（統計数理研究所）

◇「対数オッズ比の推定量と検定を改善するためのベイズ法」
小椋透　（三重大学）,
柳本武美　(統計数理研究所)

◇「共役解析の再構成と拡張の試み」
柳本武美　（統計数理研究所）

10 月 29 日（土）

9:00–9:30　開場
9:30–11:40　セッション 1

◇「大規模時空間データに対するベイズモデル」
若山 智哉　（東京大学）

◇「観察研究の効果推定値を標的集団に一般化／移送する方法の検討」
堀江悠生　（東京理科大学大学院）
篠崎智大　（東京理科大学）

◇「トーリックモデルからの直接抽出の代数的アルゴリズム」
間野修平　(統計数理研究所)
高山信毅　（神戸大学）

11:40–11:50　閉会

キャンパス情報
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝
HP:https://sites.google.com/view/sympo221027/
アクセス:https://www.tus.ac.jp/info/access/nodcamp.html
キャンパスマップ:https://www.tus.ac.jp/info/campus/noda.html
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

NDB サンプリングデータセットを用いた精神神経疾患の合併症の網羅的分析

石井 一夫[1,4], 吉永 泰周[2], 坂上 竜資[2], 小路 純央[3], 森川 渚[4], 野原 夢[4], 野原 正一郎[4], 福本 義弘[4]

[1,4]公立諏訪東京理科大学工学部, [2]福岡歯科大学口腔歯学部, [3]久留米大学高次脳疾患研究所/久留米大学医学部神経精神医学講座, [4]久留米大学医学部心臓・血管内科

日本では少子高齢化が進行し、要介護者の増加による医療・介護システムの崩壊が危惧されている。我々は、生活習慣病と精神疾患および歯科疾患が生活習慣病や介護の重症化に密接に関連していることに着目し、医療ビッグデータを用いた生活習慣病患者のメンタルケアおよびデンタルケアの研究を展開している。今回、NDB サンプリングデータセットを用いて「認知症」、「うつ病などの気分障害」、「ストレスなどの不安障害」、「睡眠障害」、「依存症」などの精神疾患と生活習慣病との関連を網羅的に調査したので報告した。このうち「認知症」と「睡眠障害」について以下にまとめた。

1. 認知症と生活習慣病との関連性調査

(1) 概要

近年、日本では少子高齢化が問題となっており、高齢化率と要介護者の急増から、医療・介護の崩壊が起こると危惧されている。要介護の要因には認知症、脳血管疾患、高齢による衰弱等があるが、中でも認知症は全体の 18.7%を占めている。認知症は主に「アルツハイマー型認知症」、「脳血管性認知症」、「レビー小体型認知症」、「前頭側頭型認知症」の 4 種に分類され、アルツハイマー型認知症が 67.6%、脳血管性認知症が 19.5%と大半を占めている。

調査にあたり利用可能なデータベースの 1 つに、厚生労働省が「高齢者の医療の確保に関する法律」に基づき、2009（平成 21）年よりレセプト情報並びに特定健診・特定保健指導情報を収集した「レセプト情報・特定健診等情報データベース(National Database of Health Insurance Claims and Specific Health Checkups of Japan: 以下、NDB)」があり、NDB は国民皆保険制度下にある日本においては国民の医療の実態を全数に近い割合で評価できることから、非常に貴重なデータであり、幅広い分野と多くの産業に活用されることが期待されている。本研究では認知症と生活習慣病の関連性調査を目的に、NDB から事前に抽出された NDB サンプリングデータセットに基づいて調査を行った。

(2) 方法

対象のデータより、ICD-10 に基づき、認知症と生活習慣病の有病者を特定し、疾患の有無と疾患の関連について効果判定を行った。疾患は、傷病名コード、診療行為コード、医薬品コードに基づいて分類した。認知症は「アルツハイマー型認知症」、「血管性認知症」を、生活習慣病は、「標準的な健診・保健指導に関するプログラム（確定版）」に基づき、

「がん」、「糖尿病」、「高血圧」、「脂質異常症(高脂血症)」、「脳血管疾患」、「虚血性心疾患」、「高尿酸血症」、「肝機能障害」、「動脈閉塞」「血圧性腎臓障害」を用いた。疾患の併存状況は、オッズ比および調整済オッズ比、カイ二条検定に基づいて評価した。

(3) 結果および考察

アルツハイマー型認知症と血管性認知症共に多くの生活習慣病に分類される疾患においてオッズ比が 1 を超えていることから、アルツハイマー型認知症と血管性認知症での傾向の多少の違いはあるが、どちらも生活習慣病との関連性が示唆された。がんは入院においてほぼ関連性が認められなかった。また、他の月についても同様の傾向が見られた。演者らは、NDB オープンデータを用いて都道府県レベルで、精神神経疾患と生活習慣病の関連を示してきた。今回、NDB サンプリングデータセットを用いて、個人レベルで、認知症と生活習慣病の合併を示すことができた。

2.睡眠障害と生活習慣病との関連性調査

(1) 概要

睡眠障害における生活習慣の乱れは、近年において大きな社会問題となっている。日本は、特に世界的にも睡眠時間が少ないとされている。また、考えられる様々な要因や背景によって、さらに質のよい睡眠が損なわれている可能性がある。睡眠障害の一つである不眠症や睡眠時無呼吸症候群は、生活習慣病の要因になる因子であると示唆されている。我々は、レセプト情報・特定健診等情報データベース(NDB)サンプリングデータセットを用いて、精神神経疾患と生活習慣病の関連を調べる網羅的な研究を開始した。今回、サンプリングデータセットの医科入院外、医科入院、DPC、調剤の４つのレセプトを用いて、睡眠障害の実態を調査し、さらには睡眠障害と生活習慣病の関連性を調査した。

(2)方法及び結果

NDB サンプリングデータセット（2011～2019 年 1 月、4 月、7 月、10 月及び、2020 年 1 月）を用いたデータ分析を行った。医科入院外、医科入院、DPC の各レセプトの傷病名コードに対し ICD-10 コードに基づき、各疾患（生活習慣病、及び精神疾患）を集計した。睡眠障害（G47）の有無について、各疾患の罹患率を特定し、オッズ比を算出して評価した。その結果、睡眠障害がある人は様々な生活習慣病（糖尿病、高血圧、高脂血症、高尿酸血症、虚血性心疾患、肝機能障害など）を併発していることや、アルコール依存症やパーキンソン病、双極性障害、統合失調症、その他の不安障害といった精神疾患を併発していることが有意に多いことが示された。

謝辞

本研究は JSPS 科研費 22K10587 の助成を受けたものである。

連検定による重心動揺データの解析

竹内直子[1,3]　綿森葉子[2]　久利彩子[3]　有末伊織[4]

1 大阪府立大学　2 大阪公立大学　3 大阪河﨑リハビリテーション大学　4 関西福祉科学大学

はじめに

連検定は、データ列のランダム性を調べる検定法であり[1)]、様々な解析ソフトで関数が用意されているため容易に活用することができる。連検定をヒト立位重心動揺データへ適用させることで立位制御の規則性を調べた。その具体的な適用方法と結果、および有用性を紹介した。立位の安定度によって加速度の向き変化に違いがあることを捉え、動揺制御の質的評価の可能性が示唆された。

背景とデータ

健康寿命の延伸に転倒予防は重要であり、自身の立位バランス能力を把握することが有効な対策になる。転倒リスクが特に高くなる動作変換時のバランス制御を質的に評価するため、力の代替として加速度に着目した。制御方法を時系列として捉えるために連検定を行った。

両脚立ちの姿勢から一歩踏み出し、片脚立位保持の後、両脚立位保持まで連続して足圧中心（Center of Pressure：COP）の計測を行った。踏み出し動作は 4 パターン（前方ステップ、前方昇段、側方ステップ、側方昇段）で、それぞれ 12 試行ずつ計測した。踏み出した後、片脚立位保持を約 10 秒間、その後両脚立位となり約 10 秒間、合計 20 秒間のデータを取得した。サンプリング周波数は 100Hz。被験者は健常成人 3 名。重心動揺計に荷重がかかり始めてから 7 秒までの各 1 秒間を解析区間 T1～T7 とし、計測最後の 1 秒間を解析区間 Tb とした。

解析

COP 加速度の向きを算出し、その角度データの偏差Δa により以下の 2 種類の 2 値化を行った。

① 「同」：$-90^\circ < \Delta a < 90^\circ$、「逆」：$-180^\circ \leqq \Delta a < -90^\circ$ または $90^\circ < \Delta a \leqq 180^\circ$

② 「右」：$-180^\circ < \Delta a < 0^\circ$、「左」：$0^\circ < \Delta a < 180^\circ$

2 種類の 2 値化で得られたデータ列それぞれに対し、解析区間ごとに、帰無仮説「ランダム」、対立仮説「混合不足」の片側検定、および、帰無仮説「ランダム」、対立仮説「混合過剰」の片側検定の連検定を行った（$p < 0.05$）。検定には統計解析ソフト R の関数 tseries::runs.test を用いた。

結果

2 値化①の場合の結果を図 1 に示す。横軸は各解析区間、縦軸は 12 試行中「混合不足」「混合過剰」「ランダム」となった頻度を表す。「混合過剰」は重心の位置が留まっている時の特徴であり、「混合不足」は不安定な状況での特徴と考えられた。2 値化②の場合の結果を図 2 に示す。「混合過剰」が現れず、「混合不足」は片脚立位・両脚立位共に見られ、「右」「左」は固まる傾向にあった。

立位の安定度によって連検定の結果が異なる傾向が見られた。さらに、従来の評価指標である軌跡長と比較したところ、2 値化①のデータに対する連検定の結果との関連がうかがえた。加速度の向き変化のランダム性を調べることにより、動揺制御の質的評価の可能性が示唆された。

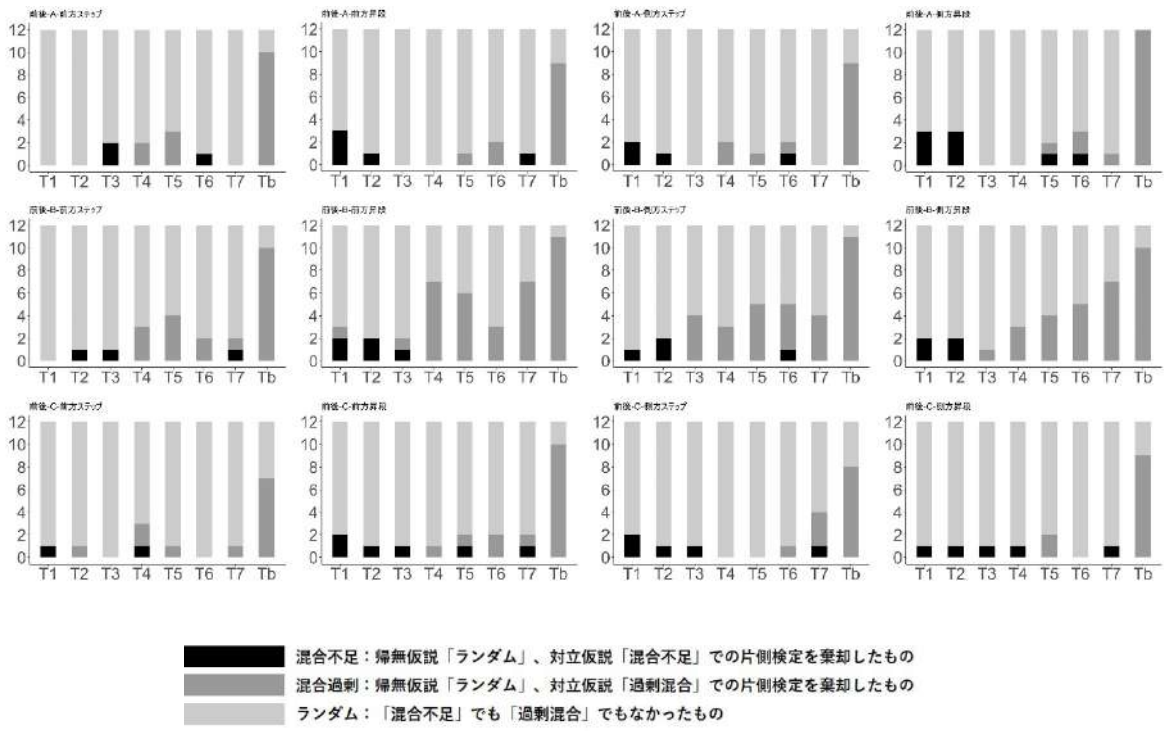

図 1　加速度の向き変化が同方向か逆方向かで 2 値化した場合の連検定の結果

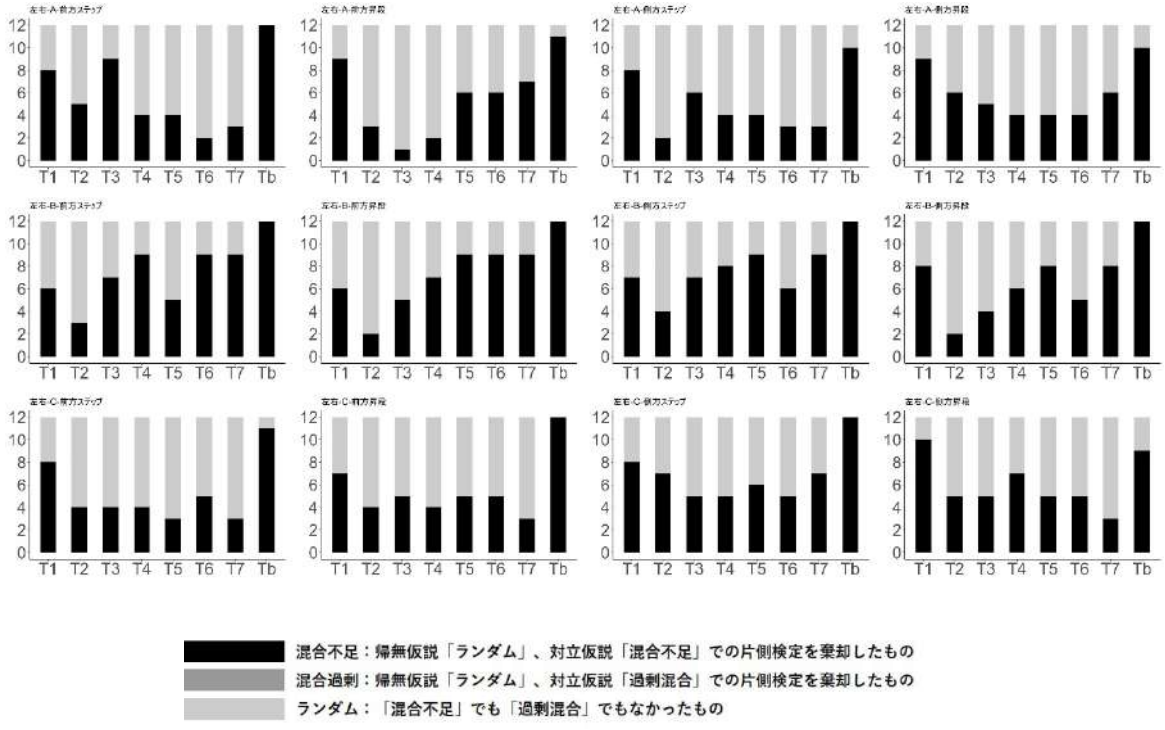

図 2　加速度の向き変化が右側か左側かで 2 値化した場合の連検定の結果

文献

1）JD Gibbons, S Chakraborti : Nonparametric statistical inference. Chapman and Hall/CRC, 2020

GMANOVA モデルにおける新たな経時変動の推定方法とその解釈 (報告書)

中京大学 教養教育研究院 永井 勇

本講演では, n 個の各個体に対して, 全ての個体で測定時点を揃えて p 回測定して得られる経時測定データの分析を考えた. このようなデータはバランス型経時測定データと呼ばれ, 各個体で測定時点が揃っていないものはアンバランス型経時測定データと呼ばれる. これらのデータの分析の目的は, データの裏に潜む経時変動を上手く捉えることある. 本講演では, バランス型経時測定データの分析について考えた.

バランス型経時測定データの分析の際には, Pothoff and Roy (1964) で提案された次の一般化多変量分散分析 (Generalized Multivariate Analysis of Variance; GMANOVA) モデルがよく使われるため, 本講演でもこのモデルを用いた;

$$\boldsymbol{Y} = \mathbf{1}_n \boldsymbol{\mu}' \boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}, \tag{1}$$

ここで, $\mathbf{1}_n$ は n 次元の全てが 1 からなるベクトル, $\mathbf{0}_r$ は r 次元の全てが 0 からなるベクトル, $\boldsymbol{Y}$ は各行が各個体で測定して得られる経時測定データからなる $n \times p$ 行列, $\boldsymbol{A}$ は各個体の特徴を表す測定時点に無関係な k 個の変数からなる $\mathrm{rank}(\boldsymbol{A}) = k$ の $n \times k$ 行列とし, $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ (各説明変数で中心化されている) を満たしているとし, $\boldsymbol{X}$ は後述のように各行が測定時点の関数からなる $p \times q$ 行列であり, これらの $\boldsymbol{Y}$, $\boldsymbol{A}$, $\boldsymbol{X}$ は既知である. また, $\boldsymbol{\mu}$ は q 次元未知ベクトル, $\boldsymbol{\Xi}$ は $k \times q$ 未知行列であり, $\boldsymbol{\mathcal{E}}$ は $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ の $n \times p$ 誤差行列とし, $\boldsymbol{\Sigma}$ は正則な $p \times p$ 未知行列とする. このモデルにおいて, $E[\boldsymbol{Y}] = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'$ の部分が, 経時測定データの分析で目的としている経時変動に対応していることを報告した. また, 測定時点を $t_1 < t_2 < \cdots < t_p$ として, $\boldsymbol{X}$ の i 列目を $(t_i^0, t_i^1, \ldots, t_i^{q-1})$ とすることは, 経時変動を測定時点の $(q-1)$ 次多項式で推定することに対応することを報告した.

このモデル (1) において, 未知の $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定としてよく使われる推定量は, 次のリスクを最小にする $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ を求めることで得られることを報告した;

$$R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma}) = \mathrm{tr}\left\{ \left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'\right)' \right\}. \tag{2}$$

実際に $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\mu}} R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma})$ と $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Xi}} R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma})$ を求めると, $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ の解となる. これらの方程式を解き $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}}$ や $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}}$ を得るために, モデル (1) において $\mathrm{rank}(\boldsymbol{X}) = q$ を仮定することが多い. この仮定は**経時変動の推定に用いる関数を制限している**と考えられることを報告した.

もし $\mathrm{rank}(\boldsymbol{X}) = q$ ならば $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ が存在するので, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}/n$, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ となる. 実際に経時変動を推定する際は $\boldsymbol{\Sigma}$ が未知のため, その不偏推定量 $\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n'/n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}\boldsymbol{Y}/(n-k-1)$ が代わりに使われる. このとき, $E[\hat{\boldsymbol{\mu}}_{\boldsymbol{S}}] = \boldsymbol{\mu}$, $E[\hat{\boldsymbol{\Xi}}_{\boldsymbol{S}}] = \boldsymbol{\Xi}$ であることが知られていることを報告した.

一方 $\mathrm{rank}(\boldsymbol{X}) < q$ の場合, $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ が存在しないため, これらの推定量は得られない. そこで本講演では, 永井 (2021, 2022) と同様, $q_1 + \cdots + q_r = q$ となる正の整数 q_i を用いて, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r)'$ ($\boldsymbol{\mu}_i$; q_i 次元ベクトル), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r)$ ($\boldsymbol{X}_i$; $\mathrm{rank}(\boldsymbol{X}_i) = q_i$ の $p \times q_i$ 行

列), $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ ($\boldsymbol{\Xi}_i$; $k \times q_i$ 行列) として, モデル (1) を次のように書き換えた;

$$\boldsymbol{Y} = \sum_{i=1}^{r} \mathbf{1}_n \boldsymbol{\mu}_i' \boldsymbol{X}_i' + \sum_{i=1}^{r} \boldsymbol{A}\boldsymbol{\Xi}_i \boldsymbol{X}_i' + \boldsymbol{\mathcal{E}}. \tag{3}$$

ここで $\boldsymbol{X}_i$ の各列が $\mathbf{0}_p$ でなければ, $r = q$ (つまり 1 列ずつ分ける) とすると, $\mathrm{rank}(\boldsymbol{X}_i) = q_i = 1$ $(i = 1, \ldots, r)$ となる. つまり, $\boldsymbol{X}$ の各列が $\mathbf{0}_p$ でない場合も含んでいることを報告した.

このように書き換えたモデルにおいても, リスク (2) と同様のリスクを最小にするような $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$ や $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ を求めて並べれば, $\mathrm{rank}(\boldsymbol{X}) < q$ でも $\boldsymbol{\mu}$ および $\boldsymbol{\Xi}$ の推定量が得られるというのが, 本講演のアイデアであった. 実際, リスク (2) の $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ と $\boldsymbol{X}$ に分割した形のものを代入すれば, モデル (3) に対応したリスク R' ができ, そのリスクを最小にする $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$ および $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ を求めればよいこととなった.

ここでリスク R' を最小にするそれぞれの推定量を $\hat{\boldsymbol{\mu}}_{i\cdot\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Xi}}_{i\cdot\boldsymbol{\Sigma}}$ $(i = 1, \ldots, r)$ として, これらを求めることを考えた. これらを求めて $(\hat{\boldsymbol{\mu}}_{1\cdot\boldsymbol{S}}, \ldots, \hat{\boldsymbol{\mu}}_{r\cdot\boldsymbol{S}})$ や $(\hat{\boldsymbol{\Xi}}_{1\cdot\boldsymbol{S}}, \ldots, \hat{\boldsymbol{\Xi}}_{r\cdot\boldsymbol{S}})$ とすると, $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定量が得られ経時変動の推定ができる. このリスク R' を $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ であることに注意して展開すると, R'=($\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$ に関連する項) + ($\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ に関連する項) と分割できることを報告した. しかし, そのままでは $\hat{\boldsymbol{\mu}}_{i\cdot\boldsymbol{\Sigma}}$ や $\hat{\boldsymbol{\Xi}}_{i\cdot\boldsymbol{\Sigma}}$ が求まりそうにないため, 永井 (2021) のアイデアを用いた. その結果, それぞれの項を最小にする $\hat{\boldsymbol{\mu}}_{1\cdot\boldsymbol{\Sigma}}, \ldots, \hat{\boldsymbol{\mu}}_{r\cdot\boldsymbol{\Sigma}}$ と $\hat{\boldsymbol{\Xi}}_{1\cdot\boldsymbol{\Sigma}}, \ldots, \hat{\boldsymbol{\Xi}}_{r\cdot\boldsymbol{\Sigma}}$ を求めると, 各 ℓ で以下となることを報告した;

$$\hat{\boldsymbol{\mu}}_{\ell\cdot\boldsymbol{\Sigma}} = (\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell)^{-1}\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\left(\frac{\boldsymbol{Y}'\mathbf{1}_n}{n} - \sum_{j<\ell}\boldsymbol{X}_j\hat{\boldsymbol{\mu}}_{j\cdot\boldsymbol{\Sigma}}\right),$$

$$\hat{\boldsymbol{\Xi}}_{\ell\cdot\boldsymbol{\Sigma}} = \left((\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y} - \sum_{j<\ell}\hat{\boldsymbol{\Xi}}_{j\cdot\boldsymbol{\Sigma}}\boldsymbol{X}_j'\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell(\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell)^{-1},$$

ここで, $\sum_{j<1}\boldsymbol{X}_j\hat{\boldsymbol{\mu}}_{j\cdot\boldsymbol{\Sigma}} = \mathbf{0}_p$, $\sum_{j<1}\hat{\boldsymbol{\Xi}}_{j\cdot\boldsymbol{\Sigma}}\boldsymbol{X}_j' = \mathbf{0}_k\mathbf{0}_p'$ である. これは, $\ell = 1$ の場合の両方が陽に求まり, それを代入することで $\ell = 2$ のときの推定量が得られることを示していることを報告した. これらにおいて $\boldsymbol{\Sigma}$ の部分を $\boldsymbol{S}$ に置き換えることで推定量が構築できる. この推定手法は永井 (2021) と同様に, 何らかの罰則などを用いていないため**最適化のための反復計算が不要**である. さらに, $\mathrm{rank}(\boldsymbol{X})$ に関する制約なしで推定が可能となるので, **経時変動の推定に用いる関数への制約がほぼ全てなくなる**ことを表していることを報告した. また, これらの推定量をモデル (3) へ入れることで, 経時変動が推定できる.

これらの推定量の解釈などについては, 当日の講演で報告した.

引用文献:

[1] Pothoff, R. F. & Roy, S. N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

[2] 永井 勇 (2021) 高次元小標本における多変量線形回帰モデルでの推定法, 2021 年度統計関連学会連合大会

[3] 永井 勇 (2022) 説明変数がランク落ちしている状況での多変量線形回帰における不偏推定量, 2022 年度統計関連学会連合大会

統計科学におけるエントロピー概念の誤り訂正

得丸久文 Tokumaru Kumon

1.　意味と恣意的に結合する「記号」から，意味が一般性・普遍性をもつ「概念」への進化

パブロフの条件反射実験が示すように，視覚・聴覚・体感などの感覚から入力される記号は，その直後に経験する餌や毒物の記憶と結びつく．この記憶が言葉の意味であり，記号と意味は個体の経験にもとづいて恣意的(経験的・偶然)に結合する．

文字が生まれ法律が公布されると，法的安定性のために，言葉に一般性が求められるようになり，概念が生まれた．概念は，数学的な群の論理で**言葉を例外なくすべての記憶と結合する**ものであり，例外をなくすために「法の番人」として法律家や裁判制度が生まれた．

2.　科学概念の起源論・学際性・有用性を検証する番人の必要性

科学概念の一般性は，起源を検証することで確かめられる．すべての科学概念は，ある特定の科学者が，それまで誰も気づかなかった不可視の現象を発見し，検証し，命名して誕生する．学際的に用いられる科学概念の起源は共通である．科学概念は，概念操作によって，不可視の現象が相互に生みだす複雑な現象を正しく把握することを可能にする．

法概念と同様に，科学概念の起源や有用性を検証する科学概念の番人が必要である．

3.　統計科学のエントロピー概念

熱力学・情報理論を問わず，エントロピー概念は混乱している．その原因は，概念の有用性を問わず, 他の分野の概念と矛盾や齟齬があっても気にしない学会の風潮にある. 個々の科学者が，概念の発生起源を調べず，間違った言語情報を鵜呑みにするため，雑音や歪が重畳し錯乱し，伝言ゲーム状態がおきている．

エントロピーを概念化するにあたってむずかしいのは，不可視の現象であることに加えて，熱力学でも定義が確立されていない. **エントロピー概念とトレードオフの関係にある情報量概念が，誤ってエントロピーと呼ばれている**．真逆な現象を同じ名前で呼ぶため混乱に拍車がかかる．

①　エントロピーの誕生と概念化（熱力学）

エントロピーは，熱力学分野でクラウジウスが発見したが，これまで起源に遡った検証が行われていなかった．八木江里氏が，クラウジウスの方程式を解き直すことで概念化し「**構成要素の配置の変化量**」と定義した．変化量だから⊿で示され正の値しかとらない．（八木他 2013）

②　エントロピー概念の学際性（情報理論）

第二次世界大戦中に生命科学や脳科学と接点をもった数学者フォン・ノイマンは，生命や知能を科学する情報理論の構築をめざした．彼は真空管式コンピュータを製造した経験から，情報理

論は熱力学的エントロピーによって決定づけられることを直感した．ところが彼は情報理論を完成させる前に，創始者の栄誉をクロード・シャノンに譲り，1948 年時点の知見を渡した．このため**シャノン理論には，ノイマンが 1950 年以降行った誤り訂正が反映されていない．**

③ **エントロピー概念の有用性（進化学，電波工学など）**

科学概念は，それが群を構成することを確かめる．群は演算に閉じているので，概念操作の結果も群となる．不可視の現象を考える上で，おおまかな全体構造を示す参照モデルが有効である．この状態で熱力学エントロピーを意識すると，まったく新しい仮説・知見を得ることができる．

4. 情報理論のエントロピー概念における矛盾の解明

「通信の数学的理論」が掲げる一般通信モデルは，通信回線上を伝搬する信号が，回線雑音によって劣化する不可視の量子力学的現象を示す参照モデルである．この二元的関係は信号対雑音比(S/N)と呼ばれ，大気を伝搬中の信号と熱雑音という二つの不可視の変数によって信号波形の乱れ(正規値からのズレ量)が生まれる．これがエントロピーであり，それは「**雑音によって生まれる信号の歪(=構成要素の配置の変化量)**」である．

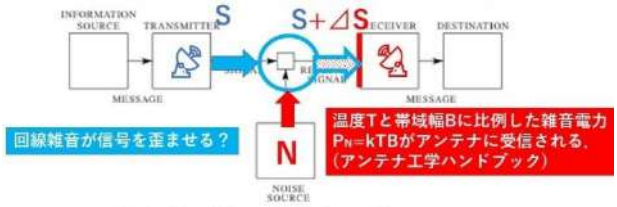

Fig. 1 — Schematic diagram of a general communication system.

図 1 一般通信モデルを使った信号(S)・雑音(N)・エントロピー(⊿S)の相関図

現在通用している情報理論のエントロピーと，熱力学的なエントロピーとの矛盾は，なぜ生まれたのか．米国議会図書館に保管されているノイマンの手稿のなかに，1949 年 12 月にイリノイ大学で行った講演記録があり，エントロピーを情報量と結びつけて論じていた．おそらくノイマンは出版にあたって，概念を精査して誤りを自ら訂正した．そしてそれをシャノンに伝えたのだが，シャノンは，ノイマンからの誤り訂正要求を無視した．だから 1982 年の IEEE オーラルヒストリーの言葉があると考えられる．この矛盾解明は統計科学にとっても有益である．

図 2/3 情報理論エントロピーの誤り分析

Exploration of Galaxy Evolution via Manifold Learning

Tsutomu T. TAKEUCHI[1,2] and Suchetha COORAY[1,†]

1. Division of Particle and Astrophysical Science, Nagoya University, Nagoya 464-8602, Japan

2. The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

† JSPS Research Fellow (DC1)

1. Galaxy Formation and Evolution

Matter in the early Universe was almost uniform, and slightly dense regions evolve by gravity, finally into a galaxy. It was attempted to develop a theory to deal with the star formation and associated history of heavy element synthesis, under an assumption that a galaxy has formed from a single, huge gas cloud. While the research in this direction was once completed in the first half of 1980s, this was not the end of the studies of galaxy evolution. Cosmological research that has progressed in parallel has revealed that galaxies merge and grow. This indicates that the galaxy evolution is a very complicated process that strongly depends on the density of the surrounding galaxies and the gas density. In order to formulate the galaxy evolution, it is necessary to determine such a huge system of equations. Though astrophysicists have constructed the governing equations from the physical laws from the first principle before, such a method is not realistic anymore when the quantity space exceeds 10 dimensions. Galaxy surveys as of the 21st century provides hundreds (or even thousands) of physical quantities for hundreds of millions of galaxies, typical big data in both quality and quantity indeed. The feature space of the galaxy to be analyzed exceeds 100 dimensions. Therefore, the characterization of the galaxy evolution is no longer possible by the traditional method relying on phy1ical intuition.

2. Galaxy Manifold

2.1 Rise, fall, and revival of the galaxy manifold

From 1970s to the mid-1980s, classical multivariate analysis methods such as the principal component analysis (PCA) were used to combine physical quantities of galaxies in a high-dimensional space. Various (logarithmic) linear relations, so-called galactic scaling relations, have been discovered. Research to unify the scaling relations and find the fundamental relationships has led to the concept of galaxy manifolds. However, the galaxy manifold has once been almost forgotten because the classical PCA could treat only linear relations, and it remained a limited concept, though they are still useful for exploring (log)linear relations of galaxies.

Recently, we discovered a galaxy manifold that expresses the basics of galactic evolution by the Fisher EM algorithm. Because of its strongly nonlinear spatial structure, it could have never been found in previous studies based on the classical PCA. To understand the manifold, a more sophisticated method beyond a mere classification is needed. We focused on a method known as the manifold learning, one of the latest methods of data science that is completely different from conventional methodologies

2.2 Galaxy manifold constructed by manifold learning

We adopt the algorithm Isomap and UMAP (Uniform Manifold Approximation and Projection). Isomap

defines the neighboring points by using input-space distance and the distant points as a sequence of "short hops" between neighboring points. Isomap tries to find shortest paths in a graph with edges connecting neighboring data points. By construction, Isomap preserves the "surface density" of data points in the feature space. UMAP is based on differential geometry and algebraic topology. The algorithm is founded on three assumptions: 1) the data are uniformly distributed on a Riemannian manifold, 2) the Riemannian metric is locally constant (or can be approximated as such), and 3) the manifold is locally connected. From these assumptions it is possible to model the manifold with a fuzzy topological structure. Sine it defines the manifold so that the data points distribute as homogeneously as possible, it does not preserve the surface density of data points. UMAP also preserves some important structural properties, and it is more robust against noise than Isomap. Manifold learning algorithm can "unfold" a curved and/or rolled manifold in the feature space, and provide a local coordinate system on it. The resulting manifolds with local coordinates from Isomap and UMAP are presented in Fig. 1. From Figure 1, we clearly see that the galaxy manifold is two-dimensional. We also stress that two different algorithms, Isomap and UMAP yield similar two-dimensional manifolds. The difference of the two estimated manifolds is clearly seen in Fig. 1. Since Isomap preserves the density of data point cloud, we observe that the manifold has a density structure, i.e., dense and sparse regions on the manifold.

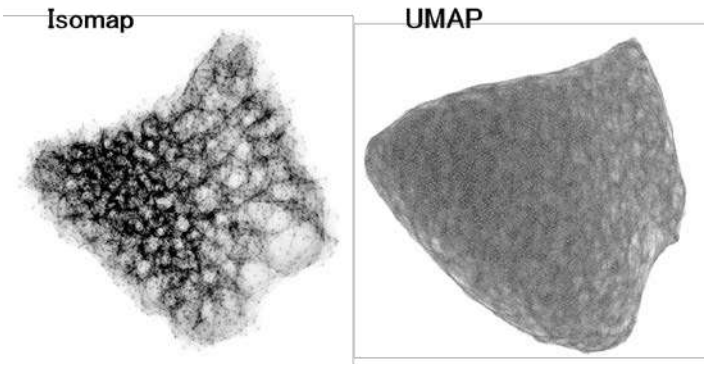

Figure 1: The "unfolded" galaxy manifold by a manifold learning algorithm Isomap and UMAP. Left and right panels show the manifolds from Isomap and UMAP, respectively. Though the global shape is slightly different from each other, they share common features on the manifold.

2.3 Result

The galaxy manifold obtained with Isomap preserve this information and reveal the speed of galaxy evolution at various stages along the manifold. e.g., galaxies passes the green valley very fast. In contrast, the galaxy manifold obtained with UMAP is imposed uniformity on the galaxy data, leading to a more robust and representative description of the observed galaxy properties e.g., galaxies evolve continuously in the feature space, without a discontinuity or "jump" on their evolutionary tracks. Thus, the galaxy manifold provides a clue to the evolutionary path of galaxies on the manifold. The SFR and stellar mass fields do not show the same evolutionary path. This supports that the galaxy merger without star formation plays a significant role in the growth of stellar mass. Next step is to fully parametrize the evolution equation of galaxies. For example, we can construct a vector field of the star formation rate on the galaxy manifold (Fig. 2).

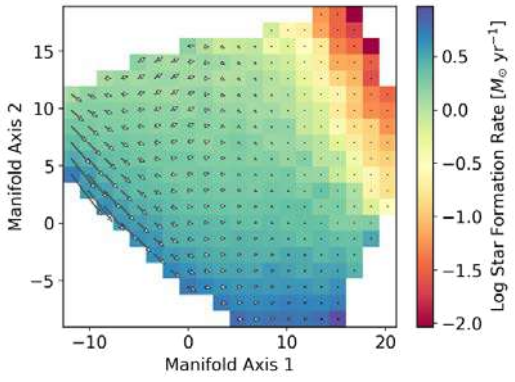

Figure 2: The vector field of star formation rate on the galaxy manifold.

Fréchet 距離を用いた分類問題ついて

伊森 晋平 [1], 若木 宏文 [1]

1: 広島大学 大学院先進理工系科学研究科

1 はじめに

本発表では，多群の混合データにおける推測問題を扱う．特に各群の母集団パラメータの推定に興味がある場合，正確に群を分類することで推定結果の向上が期待できる．Mercatanti et al. (2015) では，2 群の混合正規分布モデルにおいて，興味の対象である主要変数に補助変数を加え，変数の次元を拡張することによる，推測精度の向上について議論している．また，いくつかの仮定の下で，どのような補助変数を加えれば各個体を正しく分類する確率が上昇するかについて，理論的な考察を与えている．具体的には，補助変数間の平均差と補助変数と主要変数の相関の大きさが重要であることを示している．

一方で，この補助変数の有用性に関する結果は解析の目的と手法，データなどの設定に依存しており，異なる状況においても，同じ特徴を持つ補助変数が役立つかはわからない．Imori and Shimodaira (2019) では，不完全データ解析における有用な補助変数の選択問題として，完全データにおける損失関数を定め，それに基づく期待損失関数の推定量として，AIC (Akaike, 1974) タイプなどの変数選択規準を導出しており，それによって有用な補助変数の選択を試みている．

本発表では，各群の母集団分布の分類およびそのパラメータの推定を目的とし，正規分布間の距離として，Fréchet 距離 (Dowson and Landau, 1982) を用いた場合の分類問題に対し，有用な補助変数の条件等について考察した．

2 Fréchet 距離

2 つの p 次元多変量正規分布 $\Pi_x^{(i)} : N_p(\boldsymbol{\mu}_x^{(i)}, \boldsymbol{\Sigma}_x^{(i)})$ $(i = 1, 2)$ の分類を考える．Dowson and Landau (1982) により，2 つの正規分布間の Fréchet 距離 $F(\Pi_x^{(1)}, \Pi_x^{(2)})$ が次のように導出されている:

$$F^2(\Pi_x^{(1)}, \Pi_x^{(2)}) = \|\boldsymbol{\mu}_x^{(1)} - \boldsymbol{\mu}_x^{(2)}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_x^{(1)} + \boldsymbol{\Sigma}_x^{(2)} - 2(\boldsymbol{\Sigma}_x^{(1)}\boldsymbol{\Sigma}_x^{(2)})^{1/2}).$$

ただし，$(\boldsymbol{\Sigma}_x^{(1)}\boldsymbol{\Sigma}_x^{(2)})^{1/2}$ は非負の固有値を持つ $\boldsymbol{\Sigma}_x^{(1)}\boldsymbol{\Sigma}_x^{(2)}$ の平方根行列であり，$\mathrm{tr}((\boldsymbol{\Sigma}_x^{(1)}\boldsymbol{\Sigma}_x^{(2)})^{1/2})$ は一意に定まることに注意する．

いま，分布 $\Pi_x^{(i)}$ を母集団分布とする確率変数 $\boldsymbol{x}_i$ $(i = 1, 2)$ を次のように分割する:

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{a}_i \end{pmatrix}.$$

ただし，$\boldsymbol{y}_i \in \mathbb{R}^q$ $(q < p)$ を主要変数，変数 $\boldsymbol{a}_i \in \mathbb{R}^r$ $(r = p - q)$ を補助変数であるとする．したがって，$\boldsymbol{y}_i$ の周辺分布の推測に興味がある場合に，$\boldsymbol{a}_i$ を加えた $\boldsymbol{x}_i$ を用いて分類を行う．

さて，群間の距離が広がると，分類精度が向上し，それに伴いパラメータの推定精度も向上すると期待できる．反対に，ある変数を加えても群間の距離が広がらない場合，その変数はノイズにしかならないと考えられる．そこで本発表では，Fréchet 距離に対する冗長性をもとに有用でない補助変数を定める．すなわち，$i = 1, 2$ に対し，$\boldsymbol{\mu}_y^{(i)} = E[\boldsymbol{y}_i]$, $\boldsymbol{\Sigma}_y^{(i)} = \mathrm{Cov}[\boldsymbol{y}_i]$ と定めたとき，q 次元多変量正規分布 $\Pi_y^{(i)} : N_q(\boldsymbol{\mu}_y^{(i)}, \boldsymbol{\Sigma}_y^{(i)})$ の Fréchet 距離を $F(\Pi_y^{(i)}, \Pi_y^{(i)})$ とすると，

$$F(\Pi_y^{(1)}, \Pi_y^{(2)}) = F(\Pi_x^{(1)}, \Pi_x^{(2)})$$

であるとき，$\boldsymbol{a}_i$ を冗長な（有用でない）補助変数であるとみなす．

実際の分類では Fréchet 距離を構成するパラメータは未知であるから，各パラメータの推定量を代入することで Fréchet 距離を推定する必要がある．本発表では，観測データとして，$\boldsymbol{X}_i = (\boldsymbol{x}_{1i}, \ldots, \boldsymbol{x}_{ni})^\top \in \mathbb{R}^{n \times p}$, ただし $\boldsymbol{x}_{ti} \sim N_p(\boldsymbol{\mu}_x^{(i)}, \boldsymbol{\Sigma}_x^{(i)})$ を得ており，$\boldsymbol{x}_{ti}$ $(i = 1, 2; t = 1, \ldots, n)$ が独立である状況を想定し，$\boldsymbol{X}_1$ と $\boldsymbol{X}_2$ に基づき Fréchet 距離を推定した際の収束レートについて考察した．具体的な結果および数値シミュレーションについても当日報告した．

参考文献

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19, 716–723.

2. Dowson, D. C., & Landau, B. (1982). The Fréchet distance between multivariate normal distributions. Journal of multivariate analysis, 12, 450–455.

3. Imori, S., & Shimodaira, H. (2019). An information criterion for auxiliary variable selection in incomplete data analysis. Entropy, 21, 281.

4. Mercatanti, A., Li, F., & Mealli, F. (2015). Improving inference of Gaussian mixtures using auxiliary variables. Statistical analysis and data mining, 8, 34–48.

正方分割表における松下距離に基づく周辺同等性からの隔たり尺度

篠田　覚　　(横浜市立大学)
吉本　拓矢　(中外製薬株式会社)
田畑　耕治　(東京理科大学)

1　はじめに

行と列が同じ分類からなる正方分割表は, 医学, 教育学, 社会科学等の分野で活用されている. このような正方分割表の解析では, 多くの観測度数が主対角セルまたはその近傍に集中するため, 行と列の分類間の独立性は成り立たない. したがって, 我々は独立性ではなく対称性または周辺同等性に関心がある. たとえば, 周辺同等 (MH) モデルは行変数と列変数の各周辺分布が等しいかどうかを表す (Stuart, 1955). MH モデルの当てはまりが悪いとき, どの程度 MH モデルからの隔たりがあるかに関心があり, これまでに複数の隔たりを測る尺度が提案された (Tomizawa *et al.*, 2003; Yamamoto *et al.*, 2011; Ando *et al.*, 2021).

近年, 質的データの視覚化に関心が高まっている. しかしながら, 量的データの可視化とは異なり, 質的データの可視化の歴史は浅く, データ解析の際に活用される機会も少ない (Friendly and Meyer, 2015).

本講演では, 複数の隔たりを測る尺度に関する先行研究を整理し, 視覚的な解釈を容易にする MH モデルからの隔たりを測る尺度を提案した.

2　提案した尺度

行と列が同じ分類からなる $r \times r$ 正方分割表において, 行変数を X, 列変数を Y とし, (i, j) セル確率を p_{ij} とする $(i = 1, \ldots, r; j = 1, \ldots, r)$. また, 任意の $i = 1, \ldots, r-1$ に対して,

$$G_{1(i)} = \sum_{s=1}^{i} \sum_{t=i+1}^{r} p_{st}, \quad G_{2(i)} = \sum_{t=1}^{i} \sum_{s=i+1}^{r} p_{st},$$

$$G_{1(i)}^{c} = \frac{G_{1(i)}}{G_{1(i)} + G_{2(i)}}, \quad G_{2(i)}^{c} = \frac{G_{2(i)}}{G_{1(i)} + G_{2(i)}},$$

$$\Delta = \sum_{i=1}^{r-1} \left(G_{1(i)} + G_{2(i)} \right), \quad G_{1(i)}^{*} = \frac{G_{1(i)}}{\Delta}, \quad G_{2(i)}^{*} = \frac{G_{2(i)}}{\Delta},$$

とする.

上記を用いて, MH モデルからの隔たりを測る尺度を次のように提案した : $\{G_{1(i)}+G_{2(i)}>0\}$ を仮定し,

$$\Phi^* = \sum_{i=1}^{r-1} \left(G^*_{1(i)} + G^*_{2(i)}\right) M_i,$$

ただし,

$$M_i = \left[\frac{2+\sqrt{2}}{2}\left\{\left(\sqrt{G^c_{1(i)}} - \sqrt{\frac{1}{2}}\right)^2 + \left(\sqrt{G^c_{2(i)}} - \sqrt{\frac{1}{2}}\right)^2\right\}\right]^{\frac{1}{2}}.$$

この提案尺度 Φ^* は $\{G^c_{1(i)}, G^c_{2(i)}\}$ と $\{1/2, 1/2\}$ の間の松下距離 M_i の重み付き和であり, M_i は距離の定義を満たす (Matusita, 1954, 1955; Read and Cressie, 1988, p.112). したがって, 視覚化において自然な解釈が可能となる尺度であると考えた.

さらに, 提案尺度に対して質的データの可視化方法を参考に視覚化を行なった. 提案尺度および視覚化の有用性については, 医学分野の実データ解析を用いて示した. また, 尺度の近似信頼区間を導出し, シミュレーションによって近似信頼区間の被覆確率を評価した.

参考文献

[1] Ando, S., Noguchi, T., Ishii, A. and Tomizawa, S. (2021). A two-dimensional index for marginal homogeneity in ordinal square contingency tables. *SUT Journal of Mathematics*, **57**, 211-224.

[2] Friendly, M. and Meyer, D. (2015). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Chapman and Hall.

[3] Matusita, K. (1954). On the estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, **5**, 59-65.

[4] Matusita, K. (1955). Decision rules based on the distance, for problems of fit, two samples, and estimation. *Annals of the Institute of Statistical Mathematics*, **26**, 631-640.

[5] Read, T.R.C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York.

[6] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412-416.

[7] Tomizawa, S., Miyamoto, N. and Ashihara, N. (2003). Measure of departure from marginal homogeneity for square contingency tables having ordered categories. *Behaviormetrika*, **30**, 173-193.

[8] Yamamoto, K., Ando, S. and Tomizawa, S. (2011). A measure of departure from average marginal homogeneity for square contingency tables with ordered categories. *Revstat*, **9**, 115-126.

方向データのためのカーネル密度推定量のバイアス修正

長野県立大学グローバルマネジメント学部助教　鶴田靖人

1　はじめに

方向データは，d 次元球面 $\mathbb{S}^d := \{\boldsymbol{x} \in \mathbb{R}^{d+1} : ||\boldsymbol{x}|| = 1\}$ 上に観測点が位置するデータである．方向データの例として，次元 $d = 2$ のときは過去の地磁気のデータなどがあり，$d \geq 3$ のときはテキストデータ（の各文書が含む単語総数を 1 に基準化したもの）がある．また，$d = 1$ のときは円周上のデータと呼ばれており，例えば風向などの角度データがある．カーネル密度推定法などのノンパラメトリック密度推定法の利点は柔軟な推定が可能である．方向データのためのカーネル密度推定量の初期の研究は，Beran (1979), Hall et al. (1987) や Bai et al. (1989) が挙げられる．方向データ $X_1, \ldots, X_n$ は独立同一分布 $f(x)$　$(x \in \mathbb{S}^d)$ に従うとする．f についてのカーネル密度推定量は，

$$\hat{f}_\kappa(x) := \frac{1}{n}\sum_{i=1}^{n} K_\kappa(x^{\mathrm{T}} X_i),$$

ただし，$K_\kappa(x^{\mathrm{T}} X_i) : \mathbb{S}^d \to \mathbb{R}$ は球面上のカーネル関数，$\kappa := \kappa(n)$ は $\lim_{n\to\infty} \kappa = \infty$ かつ $\lim_{n\to\infty} n^{-1}\kappa^{d/2} = 0$ を満たす平滑化パラメータとする．ここで，$\eta \in \mathbb{S}^d$ とおく． カーネル関数を $K_\kappa(x^{\mathrm{T}}\eta) := C_\kappa(L)^{-1} L(\kappa(1 - x^{\mathrm{T}}\eta))$ と定義する．ただし，$L(r) : [0, \infty) \to \mathbb{R}$ は微分可能とし，$C_\kappa(L) := \int_{\mathbb{S}^d} L(\kappa(1 - x^{\mathrm{T}}\eta)) w_d(x)$ は基準化定数を表す．よく用いられる誤差基準は，平均積分二乗誤差 (MISE) $\mathrm{MISE}[\hat{f}_\kappa(x)] := \int_{\mathbb{S}^d} \mathrm{E}[(\hat{f}_\kappa(x) - f(x))^2] w_d(dx)$ である（ただし，w_d は d 次元球面上のルベーグ測度を表す）．f が十分に滑らかであるという仮定の下で，MISE の収束レートは $O(n^{-4/(4+d)})$ であることが知られている (Hall et al. 1987)．つまり，次元 d が増加することで MISE の収束レートは悪化する（これを次元の呪いという）．

本研究の目的は，次元の呪いの影響を減らすためにカーネル密度推定量 $\hat{f}$ のバイアスを修正し，MISE の収束レートを改良することである．そのために，実数空間上のバイアス修正法である Jones and Foster 型修正法 (Jones and Foster 1993) と Terrell and Scott 型修正法 (Terrell and Scott 1980) を $d \geq 2$ 次元の球面上のカーネル密度推定量に適用し，バイアスを修正可能なことを示す．本研究の提案手法は，Tsuruta and Sagae (2017) が円周上 $(d = 1)$ のカーネル密度推定量のために提案したバイアス修正法を $d \geq 2$ 次元に拡張したものである．

2　提案手法

期待値 $\mathrm{E}[[\hat{f}_\kappa(x)]$ は，f をテイラー展開すると κ のべき乗に L のモーメント $\mu_l(L) := \int_0^\infty L(r) r^{(l+d-2)/2} dr$ をかけた項の線形和になる．したがって，次式のように低次のモーメントが 0 になるようなカーネル関数はバイアスを修正する．

$$\mu_0(L) \neq 0, \quad \mu_l(L) = 0, \quad l = 2, 4, 6, \ldots, p-2, \quad \mu_p(L) \neq 0, \tag{1}$$

ただし，$p \geq 2$ は偶数である．(1) を満たすカーネル関数を p 次オーダーカーネルと呼ぶ (Tsuruta and Sagae 2017)．p 次オーダーカーネルは以下のような性質を持つ．

定理 1. 緩やかな仮定の下で，p 次オーダーカーネルを採用したカーネル密度推定量のバイアスと分散はそれぞれ，$\mathrm{Bias}[\hat{f}_\kappa(x)] = O(\kappa^{-p/2})$ と $\mathrm{Var}[\hat{f}_\kappa(x)] = O(n^{-1}\kappa^{d/2})$ となる．また，平滑化パラメータとして $\kappa_* = O(n^{2/(2p+d)})$ を選択すると $\mathrm{MISE}[\hat{f}_\kappa(x)]$ の収束レートは $O(n^{-2p/(2p+d)})$ となる．

p 次オーダーカーネルの構成法として，以下のような Jones and Foster 型修正法を提案する．

p 次オーダーカーネルの関数 L を $L_{[p]}$ と表す．このとき，$p+2$ 次オーダーカーネルの関数 L を次式のように定義する．

$$L_{[p+2]}(r) := \frac{p+d}{p} L_{[p]}(r) + \frac{2}{p} r L'_{[p]}(r).$$

Jones and Foster 型修正法は低次のモーメントが 0 となるように 2 つの関数 L を足し合わせたものである．低次のモーメントが 0 となるように 2 つの異なるカーネル密度推定量をかけ合わせることでもバイアスを修正することができる．そのようなバイアス修正法として，Terell and Scot 型のバイアス修正法を提案する．ここで，$\hat{f}_{\kappa,[p]}(x)$ は p 次オーダーカーネルを採用したカーネル密度推定量を表す．

Terell and Scot 型のカーネル密度推定量を次式のように定義する．

$$\tilde{f}_\kappa(x) = \left[\hat{f}_{\kappa,[2]}(x)\right]^{1/(1-a)} \left[\hat{f}_{a\kappa,[2]}(x)\right]^{-a/(1-a)},$$

ただし， $a \in (0,1)$ とする．

Terell and Scot 型のカーネル密度推定量は以下の性質を持つ．

定理 2. 緩やかな仮定の下で，Terell and Scot 型カーネル密度推定量のバイアスと分散はそれぞれ，$\mathrm{Bias}[\tilde{f}_\kappa(x)] = O(\kappa^{-2})$ と $\mathrm{Var}[\tilde{f}_\kappa(x)] = O(n^{-1}\kappa^{d/2})$ となる．また，平滑化パラメータとして $\kappa_* = O(n^{2/(8+d)})$ を選択すると $\mathrm{MISE}[\tilde{f}_\kappa(x)]$ の収束レートは $O(n^{-8/(8+d)})$ となる．

したがって，Terell and Scot 型のカーネル密度推定量は MISE の収束レートが 4 次オーダーカーネルに対応するカーネル密度推定量であると言える．Terell and Scot 型のカーネル密度推定量はパラメータ a の最適値を求める必要がある．そこで a に依存しないバイアス修正法として，指数型のバイアス修正法を提案する．

指数型のカーネル密度推定量を次式で定義する．

$$\check{f}_\kappa(x) = \hat{f}_{\kappa,[2]}(x) \exp\left\{\frac{\hat{f}_{\kappa,[4]}(x)}{\hat{f}_{\kappa,[2]}(x)} - 1\right\}$$

なお，指数型のカーネル密度推定量は，緩やかな仮定の下でバイアス，分散と最適な MISE のオーダーはそれぞれ定理 2 と同じオーダーとなる．

参考文献

ZD Bai, C Radhakrishna Rao, and LC Zhao. Kernel estimators of density function of directional data. *Multivariate Statistics and Probability*, 27:24–39, 1989.

Rudolf Beran. Exponential models for directional data. *The Annals of Statistics*, 7(6):1162–1178, 1979.

Peter Hall, GS Watson, and Javier Cabrera. Kernel density estimation with spherical data. *Biometrika*, 74 (4):751–762, 1987.

MC Jones and PJ Foster. Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3(1):81–94, 1993.

George R Terrell and David W Scott. On improving convergence rates for nonnegative kernel density estimators. *The Annals of Statistics*, 8(5):1160–1163, 1980.

Yasuhito Tsuruta and Masahiko Sagae. Higher order kernel density estimation on the circle. *Statistics & Probability Letters*, 131:46–50, 2017.

Outcome-adaptive lasso と adaptive lasso を用いた AIPW 推定量の性能評価

本江渡 [1,2]，安藤宗司 [3]，土田潤 [4]，寒水孝司 [5]

[1] 東京理科大学大学院工学研究科情報工学専攻，[2] ノバルティスファーマ株式会社，

[3] 東京理科大学理工学部情報科学科，[4] 同志社大学文化情報学部文化情報学科，

[5] 東京理科大学工学部情報工学科

近年，医療情報データベースを用いて因果効果を推定する研究が実施できるようになった．データには多くの変数が含まれているため，交絡変数の候補となる変数の数が多くなる場合がある．交絡による因果効果の推定量のバイアスを避けるためには，興味のあるアウトカム・暴露・患者背景等の変数間の因果関係に基づいて交絡の調整に用いる変数を研究計画時に特定することが望ましいが，変数の数が多いとすべての交絡変数を研究計画時に特定することは難しい．このような場合，データ駆動型の変数選択法が一つの解決策となりうる [1, 2].

交絡変数の調整に加えてアウトカムのみに関連する変数を調整すると因果効果の推定量の分散を小さくすることができる．アウトカムに関連する変数（交絡変数とアウトカムのみに関連する変数）を選択する傾向があるデータ駆動型の変数選択法として，outcome-adaptive lasso がある [1]．Outcome-adaptive lasso は傾向スコアを推定するためのロジスティック回帰モデルのパラメータ推定のために提案されている．Outcome-adaptive lasso により算出される傾向スコアに基づく inverse probability weighted（以下，IPW）推定量は，アウトカムに関連する変数を選択する傾向がない既存のデータ駆動型の変数選択法により算出される傾向スコアに基づく IPW 推定量よりも，分散が小さいことが数値実験によって示されている．しかしながら，IPW 推定量は，傾向スコアモデルを誤特定した場合にバイアスが生じ，傾向スコアの推定値が極端な値になると推定が不安定になる．このような問題に対処するために，IPW 法にアウトカム回帰を組み合わせた二重頑健推定量を用いることができる [3].

二重頑健推定量として，傾向スコアモデルとアウトカム回帰モデルを用いる augmented inverse propensity weighted（以下，AIPW）推定量 [5, 6] がよく使用される．通常，二重頑健推定量に用いる傾向スコアモデルとアウトカム回帰モデルには同じ変数が使用され [3]，異なる変数を使用するとバイアスが生じる可能性がある [4]. outcome-adaptive lasso と adaptive lasso は，アウトカムに関連する変数に対するパラメータの推定量がオラクル性を有する．そのため，それぞれ傾向スコアモデルとアウトカム回帰モデルの構築のために用いることにより，両モデルには同じ変数が選択される傾向があると考えられる．

本研究の目的は，傾向スコアモデルとアウトカム回帰モデルのパラメータをそれぞれ outcome-adaptive lasso と adaptive lasso で推定する AIPW 推定量が傾向スコアモデルのパラメータを outcome-adaptive lasso で推定する IPW 推定量よりも性能が向上するかを評価することである．AIPW 推定量を構築する際に outcome-adaptive lasso と adaptive lasso を組み合わせることによる良さを評価するため，比較対象として通常の lasso を用いた単純な AIPW 推定量と Farrell の推定量 [7] を含めた．Farrell の推定量は，傾向スコアモデルとアウトカム回帰モデルの変数選択のためだけに lasso を適用し，lasso により選択された変数のパラメータを最尤法により推定した傾向スコアモデルとアウトカム回帰モデルに基づく AIPW 推定量である．比較対象には，adaptive lasso を用いたアウトカム回帰のみに基づく推定量も含めた．実際の臨床研究データの解析と数値実験を通して，これらの推定量の性能を比較した．

実際の臨床研究データの解析として，Vanderbilt 大学が公開しているデータ（https://hbiostat.org/data/）[8] から，右心カテーテル法の集中治療室滞在日数に対する平均因果効果を推定した．これより，outcome-adaptive lasso を用いた IPW 推定値は他の推定値より高かった．

数値実験では，処置とアウトカムの生成に用いる変数をそのまま傾向スコアモデルとアウトカム回帰モデルの

説明変数に含めるシナリオ（傾向スコアモデルとアウトカム回帰モデルの両方を正しく特定可能なシナリオ）と処置とアウトカムの生成に用いる変数を誤変換して傾向スコアモデルまたはアウトカム回帰モデルの説明変数に含めるシナリオ（傾向スコアモデルまたはアウトカム回帰モデル，あるいはその両方を常に誤特定するシナリオ）を設定した．outcome-adaptive lasso を用いた IPW 推定量は，傾向スコアモデルを正しく特定可能な場合でも，共変量間の相関係数が大きくなるにつれて因果効果の推定量のバイアスと分散が大きくなった．この原因として，傾向スコアの逆数による重みが極端に大きくなることが挙げられる．傾向スコアまたはアウトカム回帰を単独で用いる平均因果効果の推定量は，そのモデルを誤特定する場合，偏りと分散が大きくなった．単純に lasso を用いた AIPW 推定量は全てのシナリオにおいてバイアスが大きくなったにも関わらず，Farrell の推定量 [7] は傾向スコアモデルまたはアウトカム回帰モデルを正しく特定可能な場合，単純に lasso を用いた AIPW 推定量よりもバイアスが小さくなる傾向があった．このことから，単純に lasso を用いると縮小推定のために十分に交絡が調整できないためバイアスが生じたと考えた．outcome-adaptive lasso と adaptive lasso を用いた AIPW 推定量は，傾向スコアモデルまたはアウトカム回帰モデルを正しく特定可能な場合，他の推定量と比べてバイアスと分散が小さくなった．これは，outcome-adaptive lasso と adaptive lasso がオラクル性を有し，傾向スコアモデルとアウトカム回帰モデルの両方でアウトカムに関連する変数を選択でき，通常の lasso よりも縮小推定が起きにくいことで十分に交絡が調整されたためであると考えた．傾向スコアモデルとアウトカム回帰モデルの両方を誤特定するシナリオでは，すべての推定量のバイアスと分散が同程度に大きくなった．

outcome-adaptive lasso を用いた IPW 推定量は，実データ解析では他の推定量による推定値よりも高い値をとり，数値実験ではバイアスと分散が大きくなる傾向があった．このことから，outcome-adaptive lasso を用いた IPW 推定量は他の推定量よりもバイアスと分散が大きくなりやすいことが示唆された．この理由として傾向スコアの逆数による重みが極端に大きくなることが考えられるため，極端な重みが生じにくい stabilized weight を用いることで今回使用した IPW 推定量が改善する可能性がある．傾向スコアモデルまたはアウトカム回帰モデルを正しく特定可能な場合は，outcome-adaptive lasso と adaptive lasso を用いた AIPW 推定量のバイアスと分散は，outcome-adaptive lasso を用いた IPW 推定量よりも小さくなった．一方で，両モデルを誤特定する場合は今回検討したすべての AIPW 推定量のバイアスや分散が IPW 推定量やアウトカム回帰による推定量と同程度に大きくなった．

主要参考文献

[1] Shortreed S and Ertefaie A. Outcome-Adaptive Lasso: Variable selection for causal inference. *Biometrics* 2017; **73**: 1111–1122.

[2] Antonelli J, Cefalu M, Palmer N. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics* 2018; **72**: 1–9.

[3] Lunceford K and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**: 2937–2960.

[4] Shinozaki T and Nojima M. Misuse of regression adjustment for additional confounders following insufficient propensity score balancing. *Epidemiology* 2019; **30**: 541–548.

[5] Robins JM, Rotnitzky A, and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**: 846–866.

[6] Glynn AN and Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 2010; **18**: 36–56.

[7] Farrell M. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 2015; **189**: 1–23.

[8] Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FE Jr, Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ Jr, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *Journal of American Medical Association* 1996; **276**: 889–97.

特異ウィシャート行列の固有値の正確分布論

清水 康希 [1,2]　橋口 博樹 [1]

[1] 東京理科大学　[2] 日本学術振興会　特別研究員 (PD)

本報告では，変量の次元がサンプルサイズよりも大き場合の標本分散共分散行列，その定数倍である特異ウィシャート行列の正確な分布論について議論した．特異ウィシャート行列の密度関数は Uhlig [7] によって与えられた．Srivastava [4] は特異ウィシャート行列の固有値の同時分布を Stiefel 多様体上の積分で表現した．Shimizu and Hashiguchi [1] は，その同時分布を関数形で表示するために非斉次超幾何関数を導入し，Sugiyama [5] の積分公式を利用して最大固有値の正確な分布を求めた．最大固有値の分布関数は，特異と非特異を区別することなく統一的に表現できる．また，Shimizu and Hashiguchi [2] は，特異ウィシャート行列とそれとは独立の非特異ウィシャート行列の比の密度関数，固有値の同時密度関数，最大固有値分布などを求めた．この最大固有値分布も特異と非特異の区別なく統一的に表現できる．さらに，Shimizu and Hashiguchi [3] では，球形検定に必要な最大固有値と最小固有値の比の正確な分布を導出した．また，実際に分布の数値計算をするためにゾーナル多項式の積をゾーナル多項式の線形結合で表現するアルゴリズムを提案した．なお，Shimizu and Hashiguchi [1, 2, 3] は，実数を $\beta = 1$ として含む β-特異ウィシャート行列に関するものであり，$\beta = 2$ の複素数，$\beta = 4$ の四元数の場合も含む統一的な議論である．本報告では，Shimizu and Hashiguchi [1, 3] の成果を実数の場合（$\beta = 1$）で紹介した．

多変量正規分布 $N_m(0, \Sigma)$ に独立に従う m 次元ベクトルを $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ とし $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ とする．ただし Σ は m 次の正定値行列とする．このとき，ウィシャート行列を $W = XX^\top$ と定義し，ウィシャート分布を $W_m(n, \Sigma)$ と表す．$n \geq m$ である場合の W を非特異ウィシャート行列，$m > n$ では特異ウィシャート行列と呼ぶことにする．特異の場合の W は $H_1^\top H_1 = I_n$ となる $m \times n$ 行列 H_1 を用いることで，$W = H_1 L_1 H_1^\top$ とスペクトル分解することができる．ただし，$L_1 = \mathrm{diag}(\ell_1, \ldots, \ell_n)$ である．Uhlig [7] は，特異ウィシャート行列 W の密度関数を

$$f(W) = \frac{\pi^{(-mn+n^2)/2}}{2^{mn/2}\Gamma_n(n/2)(\mathrm{det}\Sigma)^{n/2}} \mathrm{exp}(-\mathrm{tr}\Sigma^{-1}W/2)(\mathrm{det}L_1)^{(n-m-1)/2}$$

と与えた．ただし，$\Gamma_n(\cdot)$ は多変量ガンマ関数である．m 次対称行列 A に対して，行列変数の超幾何関数を

$$ {}_pF_q(\boldsymbol{\alpha}; \boldsymbol{\beta}; X) = \sum_{k=0}^{\infty} \sum_{\kappa \in P_m^k} \frac{(\alpha_1)_\kappa \cdots (\alpha_p)_\kappa}{(\beta_1)_\kappa \cdots (\beta_q)_\kappa} \frac{\mathcal{C}_\kappa(X)}{k!} \tag{1}$$

と定義する．ただし，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$，$(a)_\kappa = \prod_{i=1}^m (a - (i-1)/2)_{\kappa_i}$ は一般化ポッホハマー記号であり，$(a)_{\kappa_i} = \prod_{j=1}^{\kappa_i}(a + j - 1)$ である．

Shimizu and Hashiguchi [1, 2] は，特異ウィシャート行列に関する固有値分布の導出で重要となる非斉次超幾何関数を次のように定義した．

定義. A を m 次の対称行列，B を n 次の対称行列とし，$m \geq n$ とする．このとき，非斉次超幾何関数を以下で定義する．

$$ {}_pF_q{}^{(m,n)}(\boldsymbol{\alpha}; \boldsymbol{\beta}; A, B) = \sum_{k=0}^{\infty} \sum_{\kappa \in P_n^k} \frac{(\alpha_1)_\kappa \cdots (\alpha_p)_\kappa}{(\beta_1)_\kappa \cdots (\beta_q)_\kappa} \frac{\mathcal{C}_\kappa(A)\mathcal{C}_\kappa(B)}{k!\mathcal{C}_\kappa(I_m)} \tag{2}$$

定理 1. $W \sim W_m(n, \Sigma)$，$m > n$ とする．このとき，W の最大固有値 ℓ_1 の分布関数は以下で表される．

$$\Pr(\ell_1 < x) = \frac{\Gamma_n((n+1)/2)(\frac{x}{2})^{nm/2}}{\Gamma_n((n+m+1)/2)|\Sigma|^{n/2}}\ {}_1F_1{}^{(m,n)}\left(\frac{m}{2}; \frac{n+m+1}{2}; -\frac{1}{2}x\Sigma^{-1}, I_n\right) \tag{3}$$

特異と非特異の場合を区別することなく，ウィシャート行列の最大固有値の正確分布は以下の定理 2 のように統一的に表現できる．

定理 2. 特異・非特異ウィシャート行列 W において，その最大固有値 ℓ_1 の分布関数は以下で表される．

$$\begin{aligned}\Pr(\ell_1 < x) &= \frac{\Gamma_t(\frac{t+1}{2})(\frac{x}{2})^{nm/2}}{\Gamma_t(\frac{m+n+1}{2})|\Sigma|^{n/2}}\ {}_1F_1\left(\frac{n}{2}; \frac{n+m+1}{2}; -\frac{1}{2}x\Sigma^{-1}\right) \\ &= \frac{\Gamma_t(\frac{t+1}{2})(\frac{x}{2})^{nm/2}}{\Gamma_t(\frac{m+n+1}{2})|\Sigma|^{n/2}}\ \exp\left(-\frac{x}{2}\mathrm{tr}\Sigma\right){}_1F_1\left(\frac{m+1}{2}; \frac{n+m+1}{2}; \frac{1}{2}x\Sigma^{-1}\right)\end{aligned}$$

ただし，$t = \min(n, m)$ である．

Sugiyama [6] は非特異の場合に，ウィシャート行列の最大最小固有値の比 $1 - \ell_m/\ell_1$ の正確な分布を求めた．$m > n$ の場合は，Shimizu and Hashiguchi [3] で次のように与えられた．

定理 3 (Shimizu and Hashiguchi [3])**.** $W \sim W_m(n, \sigma^2 I_m)$，$m > n$ とする．このとき，$x = 1-\ell_n/\ell_1$ の密度関数は以下のように表される．

$$\begin{aligned}f(x) = C &\sum_{k=0}^{\infty}\sum_{\kappa\in P_{n-1}^k}\frac{\Gamma(mn/2+k)}{(n^k k!)}\sum_{t=0}^{\infty}x^{(n-1)(n+2)/2+k+t-1}\frac{(n-1)(n+2)/2+k+t}{t!} \\ &\times\sum_{\tau\in P_{n-1}^t}\sum_{\delta\in P_{n-1}^{k+t}}\frac{\{(n-m+1)/2\}_\tau(n/2+1)_\delta}{(n+1)_\delta}\ g_{\kappa,\tau}^{\delta}C_\delta(I_{n-1})\end{aligned}$$

ただし，$g_{\kappa,\tau}^{\delta}$ はーナル多項式の積に現れる係数である．

参考文献

[1] Shimizu, K. and Hashiguchi, H. (2021). Heterogeneous hypergeometric functions with two matrix arguments and the exact distribution of the largest eigenvalue of a singular beta-Wishart matrix. *Journal of Multivariate Analysis* **183**, 104714.

[2] Shimizu, K. and Hashiguchi, H. (2022). Expressing the largest eigenvalue of a singular beta F-matrix with heterogeneous hypergeometric functions. *Random Matrices: Theory and Applications*.

[3] Shimizu, K. and Hashiguchi, H. (2022). Algorithm for the product of Jack polynomials and its application to the sphericity test, *Statistics & Probability Letters*. **187**, 109505.

[4] Srivastava, M. S. (2003). Singular Wishart and multivariate beta distributions. *Annals of Statistics*. **31**, 1537–1560.

[5] Sugiyama, T. (1967). On the distribution of the largest latent root of the covariance matrix. *The Annals of Mathematical Statistics* **38**, 1148–1151.

[6] Sugiyama, T. (1970). Joint distribution of the extreme roots of a covariance matrix. *Annals of Mathematical Statistics*. **41**, 655–657.

[7] Uhlig, H. (1994). On singular Wishart and singular multivariate beta distributions. *Annals of Statistics* **22**, 395–405.

ロジスティック回帰モデルにおける安定な非凸スパース正則化法

東京大，国立精神・神経医療研究センター　塘由惟
東京大　小川光紀
政策研究大学院大，東京女子医大　片井みゆき
東京大　大庭幸治
東京大　松山裕

はじめに

ロジスティック回帰モデルでは分離とよばれる状況で最尤推定量が存在しないことが知られており，医学研究などの実証研究においてしばしば問題となる [1, 2]．分離の状況においても推定値が得られるような推定法として Jeffreys 事前分布による罰則項を対数尤度に課す推定法が提案されており，本稿ではこの推定法を Firth の罰則付き最尤法とよぶ [3, 4, 5]．

分離の状況における推定法は検討されている一方で，モデル選択の方法はほとんど提案されていない．本研究では Lasso 法に代表されるスパース正則化法に着目する [6]．Fan と Li は，スパース推定における推定量の望ましい性質としてオラクル性（oracle property）を提案した [7]．推定量がオラクル性をもつようなスパース正則化法として，2 段階推定によるアプローチである Adaptive Lasso 法や，非凸の正則化項を用いる SCAD 法，MCP 法などが提案されている [8, 7, 9]．分離の状況においてこれらの手法の利用を検討することもできるが，調整パラメータの選択に伴って推定が不安定となる問題が生じる [10]．また，SCAD 法や MCP 法では絶対値の大きなパラメータに対して罰則を課さないことから計算が収束しないことがある．分離の状況においても安定的に推定値を計算することができ，かつ推定量がオラクル性をもつようなスパース推定法は考案されていない．

そこで本研究では，Firth の罰則付き最尤法と非凸スパース正則化法の性質に着目し，Jeffreys 事前分布に基づく罰則項と非凸スパース正則化項を組み合わせた正則化項を課す推定法を提案する．

提案法

p 次元の説明変数ベクトルを $\boldsymbol{x}_i = (x_{i,0}, \ldots, x_{i,p-1})^\top \in \mathbb{R}^p$ と表す $(i = 1, \ldots, n)$．ただし，切片に対応する成分は $x_{i,0} = 1$ である．$n \times p$ のデザイン行列を $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ と表す．$\mathbf{X}$ は列フルランクであると仮定する．p 次元の回帰係数ベクトルを $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top \in \mathbb{R}^p$ と表す．ただし，β_0 は切片を表す．標本 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ が得られた場合のロジスティック回帰モデルの対数尤度を $l(\boldsymbol{\beta}) := \sum_{i=1}^n \left\{ y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log\left(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right) \right\}$ で表す．また，対角行列 $\mathbf{W}(\boldsymbol{\beta})$ を $\mathbf{W}(\boldsymbol{\beta}) = \mathrm{diag}\left\{ w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta}) \right\}, w_i(\boldsymbol{\beta}) = \pi_i(\boldsymbol{\beta}) \left\{ 1 - \pi_i(\boldsymbol{\beta}) \right\}, \pi_i(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) / \left\{ 1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \right\}$ で定

義する．提案法では罰則付き対数尤度

$$l^{\dagger}(\boldsymbol{\beta}) := l(\boldsymbol{\beta}) - n\sum_{r=1}^{p-1} p_{\lambda_n}(|\beta_r|) + \frac{1}{2}\log\left|\mathbf{X}^{\top}\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right|$$

の最大化により推定量を定める．$p_{\lambda_n}(\theta)$ は $\lambda_n > 0$ を用いて $\theta \in [0, \infty)$ で定義される罰則関数であり，非凸スパース正則化項を構成する．

講演当日は，提案法に関する詳細な条件，提案推定量の存在性と漸近性質に関する議論を紹介したのち，数値実験を通した有限標本下での提案法の性能と実際の医学研究データへの適用例について報告した．

References

[1] Mervyn J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 43, No. 3, pp. 310–313, 1981.

[2] Adelin Albert. and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, Vol. 71, No. 1, pp. 1–10, 1984.

[3] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, Vol. 80, No. 1, pp. 27–38, 1993.

[4] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, Vol. 21, No. 16, pp. 2409–2419, 2002.

[5] Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, Vol. 108, No. 1, pp. 71–82, 2020.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.

[7] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, Vol. 96, pp. 1348–1360, 2001.

[8] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1418–1429, 2006.

[9] Cun Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, Vol. 38, No. 2, pp. 894–942, 2010.

[10] Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, and Georg Heinze. Separation in logistic regression - causes, consequences, and control. *American journal of epidemiology*, Vol. 187, , 08 2017.

ブリッジ推定量を用いたBICの妥当性とその周辺

高崎経済大学・経済学部　　宮田 庸一

科研費シンポジウム（東京理科大学野田キャンパス）
10月27日～10月29日

$'$ 記号は, 行列の転置を表すものとする. 観測ベクトル $(Y_i, \boldsymbol{x}_i')'$ は, 以下のスパースな線形モデルに従うとする.

$$\begin{aligned} Y_i &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + u_i, \quad (i = 1, ..., n) \\ &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + 0 \cdot x_{k_n+1,i} + \cdots + 0 \cdot x_{p_n i} + u_i, \end{aligned}$$

ただし, $\boldsymbol{x}_i = (x_{1i}, \cdots, x_{p_n,i})'$ は p_n 次元の説明変数を表すベクトルとし, Y_i は被説明変数とする. また $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, .., \beta_{k_n}^*, 0, \cdots, 0)' \in \mathbb{R}^{p_n+1}$ は真のパラメーターのベクトルとし, $\beta_j^* \neq 0$ $(j = 1, ..., k_n)$, および $\beta_j^* = 0$ $(j = k_n + 1, .., p_n)$ であるとする. u_i は期待値 $E(u_i) = 0$, 分散 $Var(u_i) = \sigma_0^2$ となる撹乱項とする. ここで, 切片以外の真のパラメーターベクトルにおける非ゼロの要素の添え字の集合を $S_{0,n} = \{j \in \{1, ..., p_n\} | \beta_{0,j} \neq 0\}$ とし, その要素の個数を $k_n := |S_{0,n}|$ で表すとことにする. また k_n と説明変数の個数 p_n は, 標本の大きさ n に依存してもよいことにする. このモデルにより生成された標本 $(Y_i, \boldsymbol{x}_i')'$ $(i = 1, ..., n)$ に対して, 以下の線形モデルを当てはめることを考える.

$$Y_i = \beta_0 + \beta_{j_1} x_{j_1 i} + \cdots + \beta_{j_k} x_{j_k i} + \epsilon_i, \qquad (i = 1, ...n),$$

ただし $j_1, ..., j_k \in \{1, ..., p_n\}$, $k = 1, ..., p_n$ とする. このとき, どの説明変数をモデルに組み込むかについては, p_n の次元がそれほど大きくない場合には, AIC, BIC などの情報量規準を用いて総当たり法を行うのが標準的なアプローチであろう. しかしその一方で, p_n の次元が高くなるにと, 評価すべきモデルの数が指数的に増加するため, 総当たり法を行うのが困難になる. このような問題に対する一つのアプローチとしては, Tibshirani (1996) による LASSO(Least absolute shrinkage and selection operator) に代表される罰則付き最小二乗推定量がある. 具体的には, $\boldsymbol{y} = (Y_1, ..., Y_n)'$, $\boldsymbol{X}_1 = (x_{ji}) : n \times p_n$ 行列, $\mathbf{1}_n = (1, ..., 1)'$, $\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{X}_1 \end{pmatrix}$ とおき, 損失関数

$$L_{0,n}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \|\boldsymbol{\beta}\|_1 \tag{1}$$

を最小にする $\boldsymbol{\beta}$ の値 $\hat{\boldsymbol{\beta}}_L$ を求める. ただし $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ とする. この推定量は LASSO 推定量と呼ばれ, $\lambda_n \to \infty$ $(n \to \infty)$ とその他のいくつかの条件のもとで漸近正規性を持つことが知られている. 一方で LASSO 推定量をモデル選択手法の観点から見たときには, いくつかの問題が生

じる．それは収束レートが $\sqrt{n}$ である漸近正規性が成り立つためのチューニングパラメーター λ_n のオーダーの仮定の下では, 真のモデルを選ぶ確率は漸近的に 1 にならないことが知られている．この問題を改善するために, 式 (1) における罰則項を, 非凸の形のものに置き換えた以下の損失関数を考え, これを最小にする推定量 $\hat{\boldsymbol{\beta}}$ を求める:

$$L_n(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|^\gamma, \tag{2}$$

ただし $0 < \gamma < 1$ とする．この推定量はブリッジ推定量と呼ばれている．当然のことながら, ブリッジ推定においても適切な λ_n の選び方が重要になる．Huang et al. (2008) では, オラクル性と呼ばれる, 推定量 $\hat{\boldsymbol{\beta}}$ が漸近正規性と, ある種の漸近有効性を持ち, なおかつ説明変数の選択に関しても一致性を持つような条件を与えている．例えば $\gamma = 1/2$ として, $k_n = k$(定数), $p_n = \log n$ とおいた場合, λ_n のオーダーに関する条件は, $\lambda_n/(n^{1/4}(\log n)^{3/4}) \to \infty$ $(n \to \infty)$, $\lambda_n = o(n^{1/2})$ となるが, そのような λ_n の選択肢は, $\lambda_n = 3n^{3/8}$ でも良いし, $\lambda_n = 10n^{3/8}$ でもよいことになる．すなわち, このような漸近理論に関する結果は λ_n の選択に関しては, 何も述べていないことになる．Huang et al. (2009), Wang et al. (2010) では, λ_n を選ぶために, ある BIC 型の規準を提案しているが, この導出に関する妥当性に関しては, 現時点では明らかになっていない．本報告では, ある疑似周辺尤度が Yamanishi (1998) により提案された Extended stochastic complexity とみなせることを紹介し, また上述した BIC が, この疑似周辺尤度に対するラプラス近似に対数を取ったものの主要項として正当化できることを報告した．

参考文献

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist. 36*(2), 587–613.

Huang, J., S. Ma, H. Xie, and C.-H. Zhang (2009). A group bridge approach for variable selection. *Biometrika 96*(2), 339–355.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Wang, M., L. Song, and X. Wang (2010). Bridge estimation for generalized linear models with a diverging number of parameters. *Statistics & probability letters 80*(21-22), 1584–1596.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theory 44*(4), 1424–1439.

ロジスティック分布における母数推定について

作村建紀 *　　柳本武美 †

1　はじめに

ロジスティック分布の密度は次で与えられる．

$$p(x \mid \mu, \tau) = \frac{\tau \exp(\tau(x-\mu))}{\{1+\exp(\tau(x-\mu))\}^2}. \qquad (1)$$

ここで，$-\infty < x < \infty$, $-\infty < \mu < \infty$, $\tau > 0$ である．この分布の詳しい解説は，[Balakrishnan, 1992] にある．この分布のパラメータ (μ, τ) の推定では，その分布のシンプルな構造にも関わらず，推定量をシンプルな形式で表せないことで知られている．標本 $\boldsymbol{x} = (x_1, \ldots, x_n)$ をこのロジスティック分布からの独立同一な無作為標本とする．また，$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ とする．$E[p; q]$ は分布 q のもとでの p についての期待値を表すとする．

本研究の目的は，パラメータ (μ, τ) のベイズ推定量を導入することと，その性能を評価することである．この目的のために，無情報事前分布として参照事前分布を仮定する [Bernardo, 1979].

2　提案推定量

本研究でのベイズ的アプローチでは，$(\tau\mu, \tau)$ の事後平均 $(\widehat{\tau\mu}, \hat{\tau})$ を扱う．正規分布の場合，この推定量をプラグインした予測分布はカルバックライブラー損失のもとで最適な予測子と一致する．このとき，密度 $p(x|\mu,\tau) = (2\pi)^{-1/2}\tau^{1/2}\exp(-\tau(x-\mu)^2/2)$ として，$(\widehat{\tau\mu}, \hat{\tau}) = (\hat{\tau}\bar{x}, (n-1)/\sum_{i=1}^n (x_i - \bar{x})^2)$ となり，また，$\hat{\mu} = \bar{x}$ を得る．この結果から，$(\hat{\mu}, 1/\hat{\tau})$ は $(\mu, 1/\tau)$ の不偏推定量になる．ロジスティック分布においてパラメータ $(\tau\mu, \tau)$ を推定対象とすることについての理論的な裏付けはないが，形式的には応用できると思われる．事前分布は，$\pi_R(\mu,\tau) \propto 1/\tau$ で与えられる [Ghosh, 2011, Example 1].

事前分布 $\pi_R(\mu,\tau)$ の下での事後分布を $\pi_R(\mu,\tau|\boldsymbol{x}) \propto p(\boldsymbol{x}|\mu,\tau)\pi_R(\mu,\tau)$ とすれば，$(\tau\mu,\tau)$ の事後平均は，

$$\begin{cases} \widehat{\tau\mu} & = E[\tau\mu; \pi_R(\mu,\tau|\boldsymbol{x})] \\ \hat{\tau} & = E[\tau; \pi_R(\mu,\tau|\boldsymbol{x})] \end{cases} \qquad (2)$$

で得られる．このとき，μ の推定量は，

$$\hat{\mu} = \frac{\widehat{\tau\mu}}{\hat{\tau}}, \qquad (3)$$

として求める．

また，参照事前分布 $\pi_R(\mu,\tau)$ のもとでの事後モード $(\hat{\mu}_{\mathrm{mode}}, \hat{\tau}_{\mathrm{mode}})$ を考えることもできる．$n = 2$ のとき，$\hat{\mu}_{\mathrm{mode}} = \bar{x}$ が成り立つ．また，$\pi_R(\mu,\tau|\boldsymbol{x}) = \tau^{n-1}\prod \frac{\exp(\tau(x_i-\mu))}{\{1+\exp(\tau(x_i-\mu))\}^2}$ であるから，尤度を改善した推定量という見方もできる．

2.1　$n = 2$ における陽な表現

$n = 2$ のときには，提案推定量は陽な表現が可能である．すなわち，$\boldsymbol{x} = (x_1, x_2)$ として，式 (2) より，

$$\widehat{\tau\mu} = \frac{2\bar{x}}{|x_1 - x_2|}, \quad \hat{\tau} = \frac{2}{|x_1 - x_2|} \qquad (4)$$

が成り立つ．さらに，式 (3), (4) より，$\hat{\mu} = \bar{x}$ が得られる．これは μ に関して不偏である．また，$1/\hat{\tau} = |x_1 - x_2|/2$ であるから，$E[1/\hat{\tau}] = 1/\tau$ が成り立つ．よって，$n = 2$ においては，$(\hat{\mu}, 1/\hat{\tau})$ は $(\mu, 1/\tau)$ の不偏推定量になる．

2.2　線形性

a, b を任意の定数，$\mathbf{1} = (1, \ldots, 1)'$ として，$\boldsymbol{y} = a\boldsymbol{x} + b\mathbf{1}$ $(a \neq 0)$ とすると，

$$\hat{\mu}(\boldsymbol{y}) = a\hat{\mu}(\boldsymbol{x}) + b, \quad \hat{\tau}(\boldsymbol{y}) = \frac{1}{a}\hat{\tau}(\boldsymbol{x}) \qquad (5)$$

が成り立つ．これは積分での変数変換によって証明できる．また，ロジスティック分布の密度はパラメータ μ について左右対称であることと，等式 (5) から，$E[\hat{\mu}(\boldsymbol{x}); p(\boldsymbol{x}|0,\tau)] = 0$ を得る．すると，$E[\hat{\mu}(\boldsymbol{x}); p(\boldsymbol{x}|\mu,\tau)] = \mu$ が成り立つ．つまり，$\hat{\mu}$ はすべての n について μ の不偏推定量である．

3　リスク比較

本節では，提案推定量の性能を調査する．比較する既存推定量として，最尤推定量 $(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}_{\mathrm{ml}})$ を考え

* 法政大学理工学部：〒184-8584 東京都小金井市梶野町 3-7-2.
† 統計数理研究所：〒190-8862 東京都立川市緑町 10-3.

表 1: $1/\check{\tau}$ のバイアス

n	$1/\hat{\tau}$	$1/\hat{\tau}_{mode}$	$1/\hat{\tau}_{ml}$
2	(0)	−0.0302	−0.3442
3	0.0034	0.0180	−0.2109
4	0.0030	0.0241	−0.1529
5	0.0018	0.0214	−0.1211
10	0.0049	0.0162	−0.0549
20	0.0017	0.0076	−0.0277

た [Johnson et al., 1995]．性能はいくつかの損失関数のもとで，その平均 $\frac{1}{m}\sum_{i=1}^{m} L(\check{\theta}_i, \theta)$ をリスクとして評価した．ここで，m はシミュレーション回数，$\check{\theta}_i$ は i 番目のサンプルから得られるパラメータ θ の推定値，θ は真値，$L(\check{\theta}_i, \theta)$ は損失関数である．すべてのケースにおいて，シミュレーションは $m = 10,000$ 回行った．(μ, τ) をパラメータの真値，$(\check{\mu}, \check{\tau})$ をその推定値とするが，パラメータ (μ, τ) の各コンポーネント μ と τ についても個別に比較した．式 (5) から，$\mu = 0, \tau = 1$ のケースのみを扱った．

3.1 $1/\tau$ についての比較

まず，$1/\check{\tau}$ に注目し，そのバイアス，二乗損失，変動係数についてリスク比較した．$n = 2$ のとき $1/\hat{\tau}$ は式 (4) で計算される．表の丸括弧の表記は理論値の結果であることを表している．提案推定量の $1/\hat{\tau}$ はほぼ不偏であることが観察された．$1/\hat{\tau}_{\mathrm{mode}}$ もほぼ不偏の傾向があるが，$1/\hat{\tau}$ のほうがより不偏の傾向が強い．ここで，$\hat{\tau}$ は式 (2) の複雑な形で計算されることに注意されたい．$1/\hat{\tau}$ は近似的不偏性を有しており，また変動係数の値も小さいことが観察された．

3.2 τ についての比較

次に，τ について調べた．τ についての提案推定量の性能は，バイアス，二乗損失，変動係数に加えて，ガンマ分布から得られる近似カルバックライブラー損失によって評価した．二乗損失と双対な近似 KL 損失では $n = 2, 3$ で良い性能を示しており，また近似 KL 損失では選択したすべての n において良好な性能を示していた．

3.3 μ についての比較

次に，μ の推定量についてのリスク比較を行った．まず，すべての n のおいて，提案推定量 $\hat{\mu}$ は不偏である．二乗損失で比較すると，$n = 3, 4, 5$ で $\hat{\mu}_{\mathrm{ml}}$ のリスクが小さく，$n = 10, 20$ で $\hat{\mu}$ のリスクが小さい結果となったが，しかしこれらのリスク値の差は小さかった．

3.4 (μ, τ) についての比較

パラメータ (μ, τ) のリスク比較をカルバックライブラー損失のもとで行った．$p = p(x|\mu, \tau)$, $\check{p} = p(x|\check{\mu}, \check{\tau})$ として，カルバックライブラー損失 $E\left[\log\left(\check{p}/p\right);\ \check{p}\right]$ においては，$n = 2$ で提案推定量が良い結果であり，それ以外は提案推定量と事後モードがほぼ同等の性能であった．また，$E[\log\left(p/\check{p}\right);\ p]$ では $n = 2, 3$ で提案推定量が優勢な結果を示していた．一方，$(\hat{\mu}_{\mathrm{ml}}, \hat{\tau}_{\mathrm{ml}})$ は値が大きい結果となった．

4 考察

ロジスティック分布のパラメータ推定において，本研究で提案したベイズ推定量の性能を見る上で，τ よりもまず $1/\tau$ に注目した．これは，推定対象を考える上でその根拠の一つである正規分布においては不偏であることから，ロジスティック分布においてもその不偏性を期待したことが理由である．実際，シミュレーションの結果から，$1/\tau$ についてはほぼ不偏であることが観測された．ベイズ推定量では，一般に不偏性を考慮していない．本研究で提案した推定量も不偏性から誘導されたものではなく，指数型分布族における最適性を有する予測子に着想を得て考えたものである．それにも関わらず，近似的不偏性が観測されることは興味深い．この非指数型分布族においての推定対象の設定と，その不偏性についての理論的根拠については，今後明らかにされることを期待したい．

参考文献

[Balakrishnan, 1992] Balakrishnan, N. (1992). *Handbook of the Logistic Distribution*. CRC Press.

[Bernardo, 1979] Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.

[Ghosh, 2011] Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. *Statistical Science*, 26(2):187 – 202.

[Johnson et al., 1995] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Volume 2*, volume 289. John wiley & sons.

対数オッズ比の推定量と検定を改善するためのベイズ法

三重大学　　小椋 透
統計数理研究所　柳本 武美

1. はじめに

独立した 2 群間における 2 値データは表 1 に示す 2×2 分割表に要約される．臨床試験においては, 2 群間における予後の違いを評価する指標の一つとして対数オッズ比がしばしば用いられる．対数オッズ比は最尤法 (MLE) で推定されることが多いが, 一つ以上のセルに 0 がある場合は $-\infty$, ∞ 又は推定不能となる．これは明らかに過小推定又は過大推定である．条件付き MLE (CMLE) は, さまざまなモデルで MLE よりも優れていることが報告 (Yanagimoto and Anraku, 1989; Yanagimoto, 1991) されているが, 対数オッズ比における CMLE は MLE と同様に一つ以上のセルに 0 がある場合は $-\infty$, ∞ 又は推定不能となる．臨床試験ではサンプルサイズが小さい場合や稀な発症確率の場合に一つ以上のセルに 0 があることは珍しくないが, その場合においてもリーズナブルな対数オッズ比の推定が必要とされる．研究者が対数オッズ比を使用する場合, 群ごとの割合や群ごとのオッズに対して関心が低いため, 群ごとの割合や群ごとのオッズを改善する必要がない場合がある．本研究はベイズ法を用いて対数オッズ比を直接推定する方法を提案した．

表 1. 2×2 分割表

観測値	Positive	Negative	Total
処理群	x	$n-x$	n
対照群	y	$m-y$	m

確率	Positive	Negative
処理群	p	$1-p$
対照群	q	$1-q$

2. 提案推定量

二項分布 $\mathrm{Bi}(n,p)$ 及び $\mathrm{Bi}(m,q)$ からの観測値をそれぞれ x 及び y とした．ここで, n と m はサンプルサイズとし, p と q は確率とした．研究者が Fisher's exact test を用いる場合は p と q の推定にも関心があると考えられる．一方, 研究者が対数オッズ比を用いる場合は p と q の推定に関心が低いと考えられる．そこで, 本研究は二つのパラメータ p と q の代わりに $\alpha = \log \frac{q}{1-q}$ と $\beta = \log \frac{p(1-q)}{q(1-p)}$ を用いてベイズ推定を行った．このとき, β は対数オッズ比になっており, β の推定は対数オッズ比の直接推定であった．ベイズ法により, Jeffreys 事前分布を用いた事後密度は

$$\pi_M(\alpha,\beta|x,y) \propto \frac{\exp(x\beta)}{\sum_{z=\max(0,t-m)}^{\min(n,t)} \binom{n}{z}\binom{m}{t-z}\exp(z\beta)} \frac{\exp((t+1)\alpha+\beta/2)}{(1+\exp(\alpha+\beta))^{n+1}(1+\exp(\alpha))^{m+1}} \times \frac{\left(1+\exp\left(\alpha(t,\beta)+\beta\right)\right)^n \left(1+\exp\left(\alpha(t,\beta)\right)\right)^m}{\exp(t\alpha(t,\beta))}$$

として表された．ただし, $t = x + y$ であり, $\alpha(t,\beta)$ は $t = n\frac{\exp(\alpha(t,\beta)+\beta)}{1+\exp(\alpha(t,\beta)+\beta)} + m\frac{\exp(\alpha(t,\beta))}{1+\exp(\alpha(t,\beta))}$ を解くことによって求められた．よって, 二つのパラメータ α と β は $(\hat{\alpha},\hat{\beta}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\alpha,\beta) \times \pi_M(\alpha,\beta|x,y)d\alpha d\beta$ として推定された．

提案推定量に対応してオッズ比の信用区間に基づいた検定は次のように構成された．帰無仮説 $\mathrm{H}_0 : \beta = 0$, 対立仮説 $\mathrm{H}_1 : \beta > 0$ の片側検定とするとき, H_1 の credibility (Ghosh et al., 2006) は

$$C(x,y) = \Pr\{\beta > 0; \pi_M(\alpha,\beta|x,y)\} = \frac{\int_0^{\infty}\int_{-\infty}^{\infty}\pi_M(\alpha,\beta|x,y)d\alpha d\beta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\pi_M(\alpha,\beta|x,y)d\alpha d\beta}$$

として表されて, 検定統計量を $T(x,y) = C(x,y)$ とした．次に, 領域 $S(x,y) = \{(w,z)|T(w,z) \geq T(x,y)\}$ を設定すると, P 値は次の式で算出された．

$$\Pr\{S(x,y)|H_0\} = \max_{\alpha}\{S(x,y); p(x,y|\alpha,0)\} = \max_{\alpha}\sum_{(w,z)\in S(x,y)}\binom{n}{w}\binom{m}{z}\frac{\exp((w+z)\alpha)}{(1+\exp(\alpha))^N}$$

3. リスク比較

提案推定量と既存の推定量の比較に次のロスが用いられた. 一つ目のロスは kullback-Leibler divergence (KLD) に基づいた e-divergence 損失で

$$L_e((\check{p},\check{q}),(p,q)) = n\check{p}\log\frac{\check{p}}{p} + n(1-\check{p})\log\frac{1-\check{p}}{1-p} + m\check{q}\log\frac{\check{q}}{q} + m(1-\check{q})\log\frac{1-\check{q}}{1-q}$$

として表された. ここで, $\check{p}$ と $\check{q}$ は推定量を表した. 二つ目のロスは L_e の双対である m-divergence 損失 (dual KLD; dKLD) で $L_m((p,q),(\check{p},\check{q})) = L_e((\check{p},\check{q}),(p,q))$ として表された. 三つ目のロスは平均二乗誤差 (mean squared error; MSE) で $L_{\mathrm{MSE}}(\check{\beta},\beta) = (\check{\beta}-\beta)^2$ として表された. ここで, $\check{\beta}$ は推定量を表した. 様々な設定でリスク比較を行い, 提案推定量は既存の推定量 (CMLE, MLE, Jeffreys 事前分布を仮定した自然パラメータの事後平均 (posterior mean of the canonical parameter; PMCP), 修正 MLE (MMLE), 修正中央値の不偏推定量 (MMUE)) と比べて, 広い範囲でリスクが低いことが確認された.

4. 実例

進行性大腸癌患者における二つの新しい化学療法 (ホモハリングトニンとカラセミド) のランダム化された第 II 相臨床試験のデータ (Parzen, 2002) を使用した. この研究の結果は 2×2 分割表として表 2 のように要約された. 提案推定量と既存の推定量で対数オッズ比の推定を行った. 検定の比較には, exact 法が用いられた. 結果は表 3 にまとめられた.

表 2. 第 II 相臨床試験のデータにおける 2×2 分割表

	主要評価			副次評価		
	Positive	Negative	Total	Positive	Negative	Total
ホモハリングトニン群	0	14	14	2	12	14
カラセミド群	0	11	11	1	10	11

表 3. 第 II 相臨床試験のデータにおける対数オッズ比の推定および検定

	推定						検定 P 値	
	提案法	PMCP	CMLE	MLE	MMLE	MMUE	提案法	Exact
主要評価	-0.2458	-0.2410	-	-	-0.2318	-0.2409	0.999	>0.999
副次評価	0.4780	0.4845	0.4911	0.5108	0.3365	0.4271	0.272	0.593

5. まとめ

臨床試験においてサンプルサイズが小さい場合や稀な発症確率の場合には, 実例のような偏りが生じることは珍しくなかった. その場合に, よく知られている MLE と CMLE による推定は $-\infty$, ∞ 又は算出不能となり, これは過小推定又は過大推定であった. 対数オッズ比を用いる場合, 群ごとの割合の改善を必要としないことが多いことから, 本研究は対数オッズ比を直接推定する方法を提案した. リスク比較において, 提案推定量のリスクは既存の推定量より安定して小さいことが確認された. 実例においても, 提案推定量の有効性が検証された.

参考文献

Gart, J. J. and Zweifel, J. R. (1967). *Biometrika*, **54**(1/2), 181–187.

Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An introduction to Bayesian analysis: theory and methods*. New York: Springer.

Haldane, J. B. (1956). *Ann Hum Genet*, **20**(4), 309–311.

Parzen, M., Lipsitz, S., Ibrahim, J. and Klar, N. (2002). *J Comput Graph Stat*, **11**(2), 420–436.

Yanagimoto, T. (1987). *Ann I Stat Math*, **39**(2), 247–261.

Yanagimoto, T. and Anraku, K. (1989). *Ann I Stat Math*, **41**(2), 269–278.

Yanagimoto, T. (1991). *Ann I Stat Math*, **43**(4), 735–746.

共役解析の再構成と拡張の試み

柳 本 武 美: 統計数理研究所

1. 序

共役解析は標本分布が指数分布に従う場合に、理解が容易な事前分布を仮定して、簡単な線形推定値が得られる推測の枠組みとして登場した。Bernardo and Smith (2000) でも基本的枠組みとして紹介されている。しかし、今日の視点で見ると仮定が無雑作である一方で、形式的で発展性に乏しい枠組みである。そこで、拡張を意識した共役解析の再構成に取り組んでいる。

本研究のもう一つの狙いは、共役分布の極限としての無情報事前分布である。無情報とは利用できる情報が限りなく小さい、と理解できる。そうすると、無情報事前分布を一般化して広い立場から研究することになる。興味深い結果を得つつあるので報告した。

2. 既存の共役解析

指数分布族に属する大きさ n の標本サイズの標本の密度を $p(\boldsymbol{x}|\theta) = \exp\{n\big(\bar{x}\theta - M(\theta)\big)\}\, a(\boldsymbol{x})$, $\theta \in \Theta$. とする。共役解析では、標本密度に対して双対な構造を持つ共役事前密度を

$$\pi_T(\theta; m, \delta) = \exp\{\delta\big(m\theta - M(\theta)\big)\}h(m, \delta) \tag{1}$$

と定義する。主要な因子は、$\bar{x}$, n が m, δ と置き換えられている。共役解析では m と $\delta > 0$ の値は主観的に与えることが基本である。

事後密度は、 $\mu^* = (n\bar{x} + \delta m)/(n + \delta)$ とおいて

$$\pi_T(\theta|\,\boldsymbol{x}) = \exp\{(n + \delta)\big(\mu^*\theta - M(\theta)\big)\}h(\mu^*, n + \delta) \tag{2}$$

と表される。この分布は事前密度と同じ形をして、μ, δ を μ^*, $n + \delta$ と入れ替えて得られる。この性質を *closed under sampling* と呼んでいる。

母数 $\mu = M'(\theta)$ は $\bar{x}$ の平均である。この母数は $\mathrm{Argmax}_\mu\, \pi(\theta|\,\boldsymbol{x}) = \mu^*$、あるいは

$$\mathrm{E}_{\mathrm{post}}[\mu] = \mu^* \tag{3}$$

で与えられる。後者の表現が成立するための十分条件は、母数空間 Θ が開集合であることである (Diaconis and Ylvisker, 1979)。この推定値は簡単な形をしている上に、 n と δ で重み付けたない文展である。解釈が容易である。

3. 共役解析への懐疑

既存の共役解析を仔細に点検すると二つの重要な欠点に気付く。

1) 事前密度の選択が安易である。
   自然母数上の一様分布が陰に仮定されている
2) 多次元母数で事前密度を仮定するためには大幅な修正が必要になる。
   正規分布 $\mathrm{N}(\mu, \sigma^2)$ の場合も扱いにくい

更には、次の二点をも指摘する。

3) 事前密度の選択では自然母数が、推定では期待値母数が用いられている。

何らかの説明が求められる

4) 事前密度の族の超母数 (μ, δ) は直交するように選びたい。

概念上全く異なる役割を果たす超母数である

既存の共役解析は、事後密度と推定値の簡素な形が重視されて、事前密度の選択はおざなりである。この現状はベイズ推論の基本的な枠組みと乖離している。ベイズ法は、事前分布を注意深く選べば、後の推論は極めて簡便である。

4. 共役解析の再構成

既存の共役解析では、事前密度 $\pi_T(\theta; m, \delta)$ に標本密度の $a(\boldsymbol{x})$ に対応する台測度 $b(\theta)$ がない。言い換えると、$b(\theta) \propto 1$ が陰に仮定されている。事前密度の選択では形式的に $\delta = 0$ とおくと、密度ではなくなることが多い。しかし、事後密度では形式的に $\delta = 0$ とおいても通常は密度になる。この事実は客観的ベイズ法の基本的な拠り所である。事前密度の仮定で暗黙の仮定を設けるのは重大な欠点である。しかも、具体的な例を調べると、この仮定は通常全く魅力的でない。

更に、標本分布が指数分布族に属するときの事前密度の表現では $M(\theta)$ に対して共役な凸関数 $N(m)$ を導入する。結局新しい事前密度は

$$\pi(\theta; m, \delta) \;=\; \exp\{\delta\big(m\theta \;-\; M(\theta) \;-\; N(m)\big)\}b(\theta)\exp(-k(m, \delta) \tag{4}$$

と表される。前者の exponent は Kullback-Leibler divergence を用いて $\mathrm{D}\big(p(\boldsymbol{x}|N'(m)), p(\boldsymbol{x}|\theta)\big)$ とも表される。即ち $\theta = N'(m)$ の時に 0 になる。この事実は値 m の別の説明を与える。

台測度 $b(\theta)$ は明示的に仮定する。この測度の選択と客観的ベイズ法の研究者が研究する無情報事前分布とは殆ど同じである。Default としての選択は Jeffreys prior であると思われる。複数次元の母数場合には、Bernardo (1979) が提唱した reference prior の定義は説得力があり第一選択であると考えられる。事後密度は (2) と同様に得られる。この定義でも性質 *closed under sampling* は成り立つ。既存の事後モード、期待値母数の事後平均に関する性質は成り立たない。それでも、推定値は μ^* を通してのみ $\bar{x}$, m　に依存する。

推定値は自然母数の事後平均が推奨される。最大の理由は標本分布が指数分布族に属する場合には plug-in 予測子 $p(\boldsymbol{y}|\hat{\theta})$ が広い範囲で最適性を満たすことである (Yanagimoto and Ohnishi, 2009)。平均値母数の事後平均ではそのような性質は満たさない。

また自然母数の事後平均は

$$\mathrm{E}_{\mathrm{post}}[\theta] \;=\; N'(\mu^*) \;+\; \partial k(\mu^*, n + \delta)/\partial m$$

と表現できる。この性質は極めて弱い条件の下に成立する。この事実は表現 (3) が逆ガウス分布、逆二項分布、 von-Mises 分布で成り立たないことと対比される。具体的には、$b(\theta)$ の選択では、規格化定数が δ のみの関数となる k とが一つの規準になる。この規準が満たされると、m と δ が直交すると共に、θ の事後平均から誘導される μ の推定値が μ^* になる。

文献 1) Bernardo, J.M. (1979). *J. Roy. Statist. Soc.. Ser. B*, **41**, 113-147. 2) Bernardo, J.M., Smith, A.F.M., (2000). *Bayesian Theory.* Wiley: Chichester. 3) Diaconis, P and Ylvisker, D. (1979). *Ann. Statist.*, **7**, 269-281. 4) Yanagimoto, T. and Ohnishi, T., (2009). *J. Statist. Plann.Inf.*, **139**, 3064-3075.

大規模時空間データに対するベイズモデル

東京大学経済学研究科　若山智哉

1　はじめに

携帯端末の普及に伴い膨大な人口データが取得され、その活用が求められている。少なくとも東京都では、過去のある時間、ある場所にどれだけの人数が滞在しているかが分かるようになっている。このようなデータは、実に様々な場面で役に立つ。例えば、交通計画による混雑や交通渋滞の緩和、タクシーや宅配便の効率化、区画ごとの消費促進、災害時の避難誘導あるいは犠牲者数推定などに活用することができる。本報告ではこうしたデータから知見を抽出し、将来の意思決定につながるモデルの開発を紹介した。

2　提案モデル

都市の人口データの分析では、都市の構造を考えてモデルを構築する必要があった。例えば住宅街では時間ごとの滞在人口の変動は小さいが、オフィス街では日中は大きく増加するが、夜中や朝は住宅街の半分以下の人数になることもある。これらの傾向は近隣の地区にも波及し、結果として全体の人口分布図に表れる。つまり、空間依存的なデータである。

2.1　観測誤差モデル

$y_{ts}(\tau)$ を $t \in \{1,\dots,T\}$ 日に地点 $s \in \{1,\dots,N\}$ で観測されたデータとする。$\tau \in \{\tau_1,\dots,\tau_K\}$ は観測点であり、毎日全ての地点で K 回観測されるとする。

このデータに対して、次のような観測誤差モデルを考えた。

$$y_{ts}(\tau) = z_{ts}(\tau) + \varepsilon_{ts}, \qquad \varepsilon_{ts} \sim N(0, e_s^2), \qquad t = 1,\dots,T, \qquad s = 1\dots,N,$$

ここで、ε_{ts} は t や s に対して独立な誤差項であり、e_s^2 は地点ごとの未知分散である。また、z_{ts} は次のガウス過程に従うと仮定した。

$$z_{ts} \sim \mathcal{GP}(f_t, \eta_s^2 R(\phi_s)),$$

ここで、f_t は平均関数、$R(\phi) = \rho_{\phi_s}(d)$ は共分散関数である。つまり、$\widetilde{R}_s$ を (i,j) 成分が $\eta_s^2 \rho_{\phi_s}(|\tau_i - \tau_j|)$ な $K \times K$ 行列とすると、

$$(z_{ts}(\tau_1),\dots,z_{ts}(\tau_K)) \sim N\left((f_{ts}(\tau_1),\dots,f_{ts}(\tau_K)), \widetilde{R}_s\right),$$

のように表せる。

K 次元のベクトルを確率過程のパスの有限部分集合とすることには実用上の利点がある。もし通常の多変量解析のように $K \times K$ 行列をそのまま推定する場合は、$K \times (K-1)/2$ 個のパラメータが各地点で必要になってしまう。しかし、確率過程のパスの有限部分集合と見ることで、共分散カーネルのパラメータ（今回のケースでは η_s と ϕ_s）のみを推定すればよく、計算上有益である。

2.2 因子モデルの導入

さらに、次の因子モデルを導入した。

$$\begin{aligned}\boldsymbol{z}_t &= (B \otimes I_K)\boldsymbol{x}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(\mathbf{0}, (\widetilde{R}_s)) \\ \boldsymbol{x}_t &= G\boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \Lambda),\end{aligned}$$

ここで、$\boldsymbol{z}_t := (\boldsymbol{z}_{t1}, \boldsymbol{z}_{t2}, \ldots, \boldsymbol{z}_{tN}) := (z_{t1}(\tau_1), \ldots, z_{t1}(\tau_K), z_{t2}(\tau_1), \ldots, z_{tN}(\tau_K))$ は NK 次元ベクトル、$\boldsymbol{\nu}_t$ はその誤差項、$\boldsymbol{x}_t := (\boldsymbol{x}_{t1}, \boldsymbol{x}_{t2}, \ldots, \boldsymbol{x}_{tM}) := (x_{t1}(\tau_1), \ldots, x_{t1}(\tau_K), x_{t2}(\tau_1), \ldots, x_{tM}(\tau_K))$ は MK 次元ベクトル、B は次の $N \times M$ 行列である。$(M < N)$

$$B = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ b_{21} & 1 & 0 & \ldots & 0 \\ b_{31} & b_{32} & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ b_{M1} & b_{M2} & b_{M3} & \ldots & 1 \\ b_{M+1,1} & b_{M+1,2} & b_{M+1,3} & \ldots & b_{M+1,M} \\ \vdots & \vdots & & \ddots & \vdots \\ b_{N,1} & b_{N,2} & b_{N,3} & \ldots & b_{N,M} \end{pmatrix}.$$

これは、N 地点のトレンドを、より少数の M 地点のトレンドで説明する因子モデルになっている。また、因子負荷行列 B について、対角成分に 1 が並ぶようにデザインし識別可能性を保証した。

ここで重要なのは、因子をどのように決定するかという問題であった。一変量問題でも因子自体や因子の数を決めるのは難しく、領域知識に基づいて重要そうなものを揃えるという方法が一般的である (Prado et al., 2021)。この選択の問題点は、必要そうな因子を主観的に選択する必要があり、肝心な因子が選択されているかどうかが不明な点である。そこで我々は縮小事前分布を用いた因子選択の方法も提案した。また、因子負荷行列に空間依存構造を導入した。

3 数値実験とデータ解析

数値実験を通して、時空間データに対する提案手法の有用性を確かめるために。また、実際に人流データに対して提案手法に休日効果を加えらものを実装し、何が読み取れるかについて議論した。

参考文献

Prado, R., M. A. Ferreira, and M. West (2021). Time series: Modeling, computation, and inference.

観察研究の効果推定値を標的集団に一般化／移送する方法の検討

東京理科大学 大学院工学研究科 情報工学専攻　堀江 悠生
東京理科大学 工学部 情報工学科　篠崎 智大

あるアウトカムに対する治療効果の検証には，内的妥当性が期待されるランダム化比較試験が求められることが多い．しかし，ランダム化を伴うような介入試験の参加者は厳しい適格基準によって選択され，さらに様々な理由で研究参加に至るため，現実に治療実施を想定する集団（標的集団）と様々な点で異なることが多い．特に，アウトカムのリスク因子に関して研究参加集団と標的集団が異なっていると，いずれかの効果指標でリスク因子による治療効果の修飾が生じる．その結果，試験集団全体に対する治療効果の推定値が，標的集団での治療効果と大きく異なることがある．このような状況は，外的妥当性が欠如している一例である．近年，効果の一般化可能性（generalizability）あるいは移送可能性（transportability）の達成を目指す手法が潜在アウトカムの枠組みで議論され，外的妥当性のための一般化／移送を意図する際にも，内的妥当性を高めることを目的とする交絡調整で用いられるセミパラメトリック推定量を適用できることと，そのための識別仮定が整理されつつある（Dahabreh et al., 2019a, 2019b, 2020; Lesko et al., 2017）．

上記の取り組みは，ランダム化比較試験の結果を，異なる標的集団に一般化／移送する枠組みで考えられてきた．しかし，現実には観察研究で交絡調整を通して得た治療効果の推定値を，異なる標的集団に反映したい状況も多く存在する．JASTIS 研究（社会と新型タバコに関するインターネット調査研究）ではインターネット調査集団の結果を日本人全体集団の結果に一般化することを試みている（Tabuchi et al., 2021）．また，Web アンケートから生活習慣に関する有病率を推定し、サンプリング確率の逆確率を用いて Web アンケートの対象集団からより広範な標的集団に結果の一般化を試みた研究もある（Ferri-García et al., 2021）．このとき，ランダム化比較試験での一般化／移送の方法を，観察研究での一般化／移送の方法としてそのまま適用することができれば簡単であるが，観察研究での一般化／移送を難しくする技術的な問題が少なくとも 3 つ存在する．1 つ目は観察研究内での交絡調整の必要性，2 つ目は交絡調整に必要な変数と一般化／移送に必要な変数が異なり得る可能性，3 つ目は研究開始後に割り付けられるランダム化比較試験と異なり，治療が研究参加に影響する可能性である．このような技術的な障壁はこれまでの研究では全く指摘されておらず，方法論を統一的に整理した研究もなされていない．

本研究では，これまでランダム化比較試験の一般化／移送で検討された推定量の中でも逆確率重み付け推定量（inverse probability weighting ; 以下，IPW 推定量）に着目し，上記 3 つの技術的な問題を考慮した観察研究の治療効果の推定値を標的集団に一般化／移送する IPW 推定量を提案する．反事実記法を用いた代数によって上記問題点の状況を整理し，一般化／移送された効果の識別可能条件を得て IPW 推定量を導出した．本発表では、移送について提案する IPW 推定量の詳細や数値実験による性能評価の結果を報告した．

参考文献

[1] Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*. 2019a; **75**: 685–694.

[2] Dahabreh IJ, Robertson SE, Hernán MA. On the relation between g-formula and Inverse probability weighting estimators for generalizing trial results. *Epidemiology*. 2019b; **30**: 807-812.

[3] Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*. 2020; **39**: 1999–2014.

[4] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2017; **28**: 553–561.

[5] Tabuchi T, Kiyohara K, Hoshino T, Bekki K, Inaba Y, Kunugita N. Awareness and use of electronic cigarettes and heat-not-burn tobacco products in Japan. *Addiction*. 2016; **111**: 706–713.

[6] Ferri-García R, Rueda M, Cabrera-León A. Self-perceived health, life satisfaction and related factors among healthcare professionals and the general population: analysis of an online survey, with propensity score adjustment. *Mathematics*. 2021; **9**: 791; Available at: https://doi.org/10.3390/math9070791.

トーリックモデルからの直接抽出の代数的アルゴリズム

間野　修平　(統計数理研)[*1]
高山　信毅　(神戸大理)[*2]

状態空間を $[m] := \{1, 2, \ldots, m\}$ とし，行空間に $(1, \ldots, 1)$ を含む行列 $B = (a_{ij}) \in \mathbf{Z}^{d\times m}$ と $h \in \mathbf{R}^m_{>0}$ が定める確率分布の集合

$$\mathcal{M}_{B,h} := \{p \in \mathrm{int}(\Delta_{m-1}) : \log p \in \log h + \mathrm{rowspan}(B)\}$$

をトーリックモデル，もしくは離散指数型分布族，対数アフィンモデル，という．

母数 $\psi \in \mathbf{R}^{d'}_{>0}$ と $\phi \in \mathbf{R}^d_{>0}$ があり，ϕ を局外母数とする．

$$B = \begin{pmatrix} A \\ \tilde{A} \end{pmatrix} \in \mathbf{Z}^{(d+d')\times m}, \quad A = (a_{ij}) \in \mathbf{Z}^{d\times m}, \quad \tilde{A} = (\tilde{a}_{ij}) \in \mathbf{Z}^{d'\times m},$$

$(1, \ldots, 1) \in \mathrm{rowspan}(A)$ に対し，$\mathcal{M}_{B,h}$ を

$$p_j = \frac{h_j}{Z(\phi, \psi)} \prod_{i\in[d]} \phi_i^{a_{ij}} \prod_{k\in[d']} \psi_k^{\tilde{a}_{kj}}, \quad j \in [m]$$

とパラメトライズする．多項抽出，もしくはポアソン抽出された標本における状態 j をとる観察の数を u_j $j \in [m]$ で表すとき，A, b が定める集合

$$\mathcal{F}_b(A) := \{u : Au = b, u \in \mathbf{N}^m\}, \quad \mathbf{N} := \{0, 1, 2, \ldots\}$$

を ϕ に対する最小十分統計量 $b \in \mathbf{N}A := \sum_{j\in[m]} \mathbf{N}a_j$ に付随する b ファイバーという．ここで a_j は A の j 列ベクトルである．b による条件付き確率分布は

$$\mathbf{P}(U = u|AU = b) = \frac{1}{Z_A(b; y)} \frac{y^u}{u!}, \quad u \in \mathcal{F}_b(A), \quad y^u := \prod_{j\in[m]} y_j^{u_j}, \quad u! := \prod_{j\in[m]} u_j! \quad (1)$$

である．母数を $y_j := h_j \prod_{k\in[d']} \psi_k^{\tilde{a}_{kj}} \in \mathbf{R}_{>0}$ とした．正規化定数

$$Z_A(b; y) := \sum_{u\in\mathcal{F}_b(A)} \frac{y^u}{u!}$$

は GKZ 超幾何多項式，もしくは A 超幾何多項式と呼ばれる．整数格子 $\mathbf{N}A$ の凸包は多面体錘を成し，$b \notin \mathbf{N}A$ のとき $Z_A(b; y) = 0$ と規約する．(1) の確率函数が表す分布を A 超幾何分布と呼ぶ．

[*1] 〒190-8562 東京都立川市緑町 10-3　統計数理研究所 数理・推論研究系
e-mail: smano@ism.ac.jp
[*2] 〒657-8501 兵庫県神戸市灘区六甲台町 1-1　神戸大学 大学院理学研究科
e-mail: takayama@math.kobe-u.ac.jp

$\mathbf{N}A$ に埋め込まれた有界な整数格子で半順序

$$v \in \mathbf{N}A \ \text{ and } \ v - a_j \in \mathbf{N}A \quad \Rightarrow \quad v - a_j \prec v$$

を備え，最大元 b と最小元 0 を持つものをマルコフ束 $\mathcal{L}_A(b)$ と呼ぶ．以下のアルゴリズムは状態空間を $\mathcal{L}_A(b)$ とするマルコフ連鎖である．状態は元に 1 対 1 に対応し，推移は近傍，つまり $u \prec v$ を満たす元の組において v から u に起こる．

アルゴリズム **1** (Gauss–Manin ベクトル $q(b;y)$, 行列 $T(b;y)$ は講演で与えた)．

入力: 行列 $A \in \mathbf{Z}^{d\times m}$, 十分統計量 $b \in \mathbf{N}_0 A$, パラメタ $y \in \mathbf{R}^d_{>0}$.
出力: A, b が定める A 超幾何分布に従うランダムベクトル u.

Step 1: Gauss–Manin ベクトル $q(b;y)$ を初期化．
Step 2: $t = 1, n = \deg(b)$ とする．
Step 3: 規格化しない推移確率 $\tilde{e}(b, b-a_j; y) = \{T(b;y)q(b;y)\}_j, \forall j \in [m]$ を計算．
Step 4: $[0,1]$ を比 $\tilde{e}(b, b-a_1; y) : \tilde{e}(b, b-a_2; y) : \cdots : \tilde{e}(b, b-a_m; y)$ に分割．$b - a_j \notin \mathcal{L}_A(b)$ のとき j 番目は 0.
Step 5: $[0,1]$ の一様乱数をとり，区間 $\tilde{e}(b, b-a_j; y)$ に入れば $j_t = j$ とする．
Step 6: $\deg(b - a_{j_t}) = 0$ であれば Step 7, そうでなければ $q_k(b - a_{j_t}; y) = y_{j_t}^{-1}\{\tilde{P}_{j_t}(b;y)q(b;y)\}_k,\ k \in 0 \cup [r-1]$ を計算，$t \leftarrow t+1, b \leftarrow b - a_{j_t}$ とし，Step 3.
Step 7: $(u_1, \ldots, u_m)$, $u_j := |\{t : j_t = j, t \in [n]\}|$ を出力．

行列 $\tilde{P}_j(b;y)$ はいかなる行列 A に対してもある微分作用素環のイデアルの Gröbner 基底に対する標準形を計算することで得られる．$\{0, a_1, \ldots, a_m\}$ の凸包を単位単体 $\{0, e_1, \ldots, e_d\}$ で測った体積を $r = \mathrm{vol}(A)$ で表した．

定理 1. アルゴリズム 1 による A 超幾何分布に従うランダムベクトルの直接抽出の時間計算複雑性は $O(\max\{m, r\}rn)$ である．ここで $n = \deg(Z_A(b;y))$, $r = \mathrm{vol}(A)$. ただし有理数演算のコストを $O(1)$ とした．

本講演ではアルゴリズム 1 とその導出について説明し，二元分割表の独立性のないモデルについて計算機実験の結果を報告した．結果をまとめる．メトロポリス連鎖の有効サンプルサイズと比較して，分割表 1 個の生成にかかる時間は 800 倍ほどであった．計算に要した時間は，現れる推移確率の計算を尽くした後は，分割表の個数にほぼ比例した．モジュラ算法に基づく並列化の有効性を検証した．推移確率の計算はコア数を増やすだけでは短縮できないが，それを終了した後の計算時間はコア数にほぼ反比例した．

参考文献

[1] Mano, S., Takayama, N., Algebraic algorithm for direct sampling from toric models, arXiv: 2110.14992