科研費シンポジウム
「多変量データ解析法における理論と応用」

科学研究費・基盤研究（A）「大規模複雑データの理論と方法論の総合的研究」
研究 代表者：青嶋 誠（筑波大学）, 課題番号：15H01678）

■ **日時：2018年12月13日～12月15日**

■ **場所：広島大学理学部B707**

■ **開催責任者：栁原宏和 (広島大学)**

12月13日

13:00-13:10 Opening

セッション1, 座長：若木宏文 (広島大学)

13:10-13:50	矢田和善* (筑波大学) 青嶋誠（筑波大学） 石井晶（東京理科大学）	高次元共分散構造に関する検定の一般化について
13:50-14:30	新村秀一 (成蹊大学)	癌の高次元遺伝子解析の諸問題（2）
14:30-15:10	柳本武美 (統計数理研究所)	RCT と DNN が医療水準の向上を駆動する

企画セッション1「疫学における統計的挑戦」, 座長：伊藤ゆり (大阪医科大学)
(オーガナイザー：伊藤ゆり・福井敬祐 (大阪医科大学))

15:30-16:10	鈴木有佳 (大阪医科大学)	女性の経済・雇用に関する市区町村の指標と死亡率との関連
16:10-16:50	福井敬祐 (大阪医科大学)	マイクロシミュレーションによる大腸がん検診における死亡率減少効果
16:50-17:30	堀芽久美 (国立がん研究センター)	シミュレーションモデルによる乳がん予防・検診の効果

12月14日

セッション2, 座長：橋本真太郎 (広島大学)

9:30-10:10	清水康希* (東京理科大学) 橋口博樹（東京理大学） 岩下登志也（東京理大学）	特異ウィシャート行列の最大・最小固有値分布について
10:10-10:50	米永航志朗* (北海道大学) 鈴川晶夫 (北海道大学)	ウィッシャート行列と正規ベクトルの積の密度関数とモーメント
10:50-11:30	牧草夏実 (千葉大学)	ヒルベルト空間における正規性の検定の実際的側面
11:30-12:10	岩下登志也* (東京理科大学) Bernhard Klar (カールスルーエ工科大学) 橋口博樹（東京理科大学）	楕円対称性の必要性検定

企画セッション2 「社会疫学における統計解析の役割」, 座長：福井敬祐 (大阪医科大学)
(オーガナイザー：伊藤ゆり・福井敬祐 (大阪医科大学))

13:30-14:10	那波伸敏 (東京医科歯科大学)	地域のソーシャルキャピタル，親の心理的苦痛，児童虐待の関連：足立区子どもの健康・生活実態調査から
14:10-14:50	伊角彩 (東京医科歯科大学)	子どもの貧困と虐待における親の心理的ストレスと個人レベルのソーシャル・キャピタルの媒介効果

特別セッション, 座長：山村麻理子(広島大学)
(平成29年度ダイバーシティ研究環境実現イニシアティブ（牽引型）
「国際型ダイバーシティ研究環境実現プログラム」協力)

15:10-15:50	伊藤ゆり(大阪医科大学)	がん登録資料を用いたがん患者の生存時間解析
15:50-16:30	ソルヴァン加藤 比呂子* (Inst. of Marine Research) Sam Subbey (Inst. of Marine Research)	複雑な海洋生態システムにおける因果性推測
16:30-17:10	イリチュ美佳 (筑波大学)	データ融合に基づく要約手法

12月15日

セッション3, 座長：栁原宏和 (広島大学)

9:30-10:10	永井勇 (中京大学)	Plug-in optimization method for generalized ridge regression for MLE in GMANOVA model
10:10-10:50	内藤貫太 (千葉大学)	歪曲度のノンパラメトリック推定
10:50-11:30	地道正行* (関西学院大学) 阪智香 (関西学院大学) 宮本大輔 (東京大学) 永田修一 (関西学院大学)	探索的財務ビッグデータ解析
11:30-12:10	大瀧慈* (広島大学 & 広島大学ベンチャー企業 (株)deCult) 大谷敬子 (広島大学 & 広島大学ベンチャー企業 (株)deCult)	セミパラメトリック混合効果モデルの適用による繰り返し観測データの解析

12:10-12:20 まとめ

高次元共分散構造に関する検定の一般化について

矢田　和善　(筑波大数理物質)
青嶋　誠　　(筑波大数理物質)
石井　晶　　(東京理大理工)

1. はじめに

共分散行列に p 次の半正定値行列 $\boldsymbol{\Sigma}$ をもつ母集団を考える．母集団から $n\ (\geq 4)$ 個の p 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出する．ただし，$\boldsymbol{x}_j = (x_{1j}, ..., x_{pj})^T$ とおく．ここで，$\sigma = \mathrm{tr}(\boldsymbol{\Sigma})/p$，$\mathbf{1}_p = (1, ..., 1)^T$，$\boldsymbol{\Sigma}$ の j 番目の対角成分を σ_j とし，$\boldsymbol{\Sigma}_{\mathrm{D}} = \mathrm{diag}(\sigma_1, ..., \sigma_p)$，$\boldsymbol{\Sigma}_{\mathrm{IC}} = \sigma\{(1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T\}$ とおく．ただし，$\rho \in (0, 1)$ である．そのとき，Schott (2005, Biometrika) や Srivastava et al. (2011, JMA) は，正規分布もしくはそれに類する仮定のもとで，以下の仮説における高次元検定方式を与えた．

$$H_{0a} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{D}} \quad \text{vs.} \quad H_{1a} : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_{\mathrm{D}} \tag{1.1}$$

一方で，Srivastava and Reid (2012, JMA) は，正規分布のもと以下の仮説における高次元検定方式を与えた．

$$H_{0b} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{IC}} \quad \text{vs.} \quad H_{1b} : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_{\mathrm{IC}} \tag{1.2}$$

本講演では，まず，仮説 (1.1) と (1.2) を考えた．Yata and Aoshima (2013, JMA) で与えた「拡張クロスデータ行列法 (ECDM)」を用いることで，非正規分布においても頑健な高次元検定方式を与えた．さらに，$\boldsymbol{\Sigma}_{\mathrm{D}}$ や $\boldsymbol{\Sigma}_{\mathrm{IC}}$ を包含するような一般的な高次元共分散構造の枠組みにおける検定方式も提供した．

2. 仮説 (1.1) の検定方式

いま，$\Delta_{\mathrm{D}} = ||\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\mathrm{D}}||_F^2 = \mathrm{tr}(\boldsymbol{\Sigma}^2) - \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}}^2)$ とおく．H_{0a} のもと $\Delta_{\mathrm{D}} = 0$ となることに注意する．Yata and Aoshima (2013) は，拡張クロスデータ行列法 (ECDM) という方法論を開発した．2つの集合 $\mathcal{V}_{n(1)(k)}$, $\mathcal{V}_{n(2)(k)}$ $(k = 3, ..., 2n-1)$ を次のように定義する．

$$\mathcal{V}_{n(1)(k)} = \begin{cases} \{\lfloor \frac{k}{2} \rfloor - n_{(1)} + 1, ..., \lfloor \frac{k}{2} \rfloor\}, & \lfloor \frac{k}{2} \rfloor \geq n_{(1)} \text{のとき}, \\ \{1, ..., \lfloor \frac{k}{2} \rfloor\} \cup \{\lfloor \frac{k}{2} \rfloor + n_{(2)} + 1, ..., n\}, & \text{それ以外} \end{cases}$$

$$\mathcal{V}_{n(2)(k)} = \begin{cases} \{\lfloor \frac{k}{2} \rfloor + 1, ..., \lfloor \frac{k}{2} \rfloor + n_{(2)}\}, & \lfloor \frac{k}{2} \rfloor \leq n_{(1)} \text{のとき}, \\ \{1, ..., \lfloor \frac{k}{2} \rfloor - n_{(1)}\} \cup \{\lfloor \frac{k}{2} \rfloor + 1, ..., n\}, & \text{それ以外} \end{cases}$$

ここで，$n_{(1)} = \lceil n/2 \rceil$, $n_{(2)} = n - n_{(1)}$, $\lfloor x \rfloor$ は x 以下の最大の整数，$\lceil x \rceil$ は x 以上の最小の整数を表す．そのとき，拡張クロスデータ行列法を使えば，$\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の不偏推定量は次のように構築できる．各 $k\ (= 3, ..., 2n-1)$ で2分割した集合について，標本平均を

$$\bar{\boldsymbol{x}}_{n(1)(k)} = \frac{1}{n_{(1)}} \sum_{j \in \mathcal{V}_{n(1)(k)}} \boldsymbol{x}_j, \quad \bar{\boldsymbol{x}}_{n(2)(k)} = \frac{1}{n_{(2)}} \sum_{j \in \mathcal{V}_{n(2)(k)}} \boldsymbol{x}_j$$

とし，

$$W_n = \frac{2c_n}{n(n-1)} \sum_{i<j}^{n} \{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{n(1)(i+j)})^T (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_{n(2)(i+j)})\}^2$$

を定義する．ここで，$c_n = (n_{(1)}-1)^{-1}(n_{(2)}-1)^{-1}n_{(1)}n_{(2)}$ である．そのとき，$E(W_n) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$ となる．$\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}}^2)$ についても同様に，各 $s\ (=1,...,p)$ と $k\ (=3,...,2n-1)$ で

$$\overline{x}_{s(1)(k)} = n_{(1)}^{-1} \sum_{j \in \mathcal{V}_{n(1)(k)}} x_{sj}, \quad \overline{x}_{s(2)(k)} = n_{(2)}^{-1} \sum_{j \in \mathcal{V}_{n(2)(k)}} x_{sj}$$

とし，

$$U_n = \frac{2c_n}{n(n-1)} \sum_{i<j}^{n} \sum_{s=1}^{p} \left\{(x_{si} - \overline{x}_{s(1)(i+j)})(x_{sj} - \overline{x}_{s(2)(i+j)})\right\}^2.$$

を定義する．そのとき，$E(U_n) = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}}^2)$ となる．ここで，Δ_{D} の不偏推定量

$$\widehat{\Delta}_{\mathrm{D}} = W_n - U_n$$

を考える．そのとき，次の定理が成り立つ．

定理 1. $m = \min\{p, n\}$ とし，次を仮定する．

(C-i) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{n\Delta_{\mathrm{D}}} \to 0$ as $m \to \infty$.

そのとき，いくつかの正則条件のもと次が成り立つ．

$$\frac{\widehat{\Delta}_{\mathrm{D}}}{\Delta_{\mathrm{D}}} = 1 + o_P(1) \ \text{ as } m \to \infty.$$

定理 2. 次を仮定する．

(C-ii) $\limsup_{m\to\infty} \left\{\dfrac{n\Delta_{\mathrm{D}}}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right\} < \infty$.

そのとき，いくつかの正則条件のもと次が成り立つ．

$$\frac{\widehat{\Delta}_{\mathrm{D}} - \Delta_{\mathrm{D}}}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/n} \Rightarrow N(0,1) \ \text{ as } m \to \infty.$$

ここで，H_{0a} のもと $\mathrm{tr}(\boldsymbol{\Sigma}^2) = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}}^2)$ となることに注意し，仮説 (1.1) を次のように検定する．

$$\text{rejecting } H_{0a} \Longleftrightarrow \frac{n\widehat{\Delta}_{\mathrm{D}}}{2U_n} > z_\alpha. \tag{2.1}$$

ただし，z_α は $N(0,1)$ の上側 $100\alpha\%$ を表す．そのとき，検定方式 (2.1) による第 1 種の過誤は α に収束する．

当日は，ECDM を用いて仮説 (1.2) の高次元検定方式も与えた．さらに，$\boldsymbol{\Sigma}_{\mathrm{D}}$ や $\boldsymbol{\Sigma}_{\mathrm{IC}}$ を包含するような一般的な高次元共分散構造の枠組みにおける検定方式も提供した．

癌の高次元遺伝子解析の諸問題(2)

成蹊大学　名誉教授
新村秀一

現在、再び統計学が注目されている。そのキーワードの一つが**ビッグデータ**であり、量、多様性、速度が説明に用いられている。筆者は、既存の小標本を扱う推測統計学との関係で論じたい。小標本はｎとｐが小さいものを指すが境界は個人にとって異なる。ｎが多くてｐが小さい縦長データは、統計分析として従来の延長線上で扱える。ｎが小さくｐが大きいものを**高次元データあるいは横長データ**と呼ぶことにし、本研究のテーマである。ｎとｐがともに大きいものが**ビッグデータ**であり、ユーザーが勝手に登録したデータを利用できる Apple の個人の顔認識や、Google はペイジランクを数理計画法(MP)のマルコフ過程の応用で特許を取り、大規模なデータをインターネットでコストをかけずに集めた統計手法の集計が主役である。**横長データの代表は** Microarray である。ハーバード大学の Golub 教授らが 1999 年にサイエンスに発表した論文で 30 年以上 Microarray の遺伝子情報から 2 群の予測(判別)と新しい癌の Subclass の発見を目指してきたが芳しい成果を得ていない。すなわち、1970 年頃から行われてきたことになる。これらの研究で用いられた Microarray が無償で公開され、統計や機械学習の研究者が 2000 年以前に横長データの新しい研究テーマとして研究をしてきた。2 群の判別分析が最も適した手法であるが、めぼしい成果が出ていない。

筆者は 1971 年に大学卒業後に大阪成人病センターで、心電図の自動診断論理を判別手法で開発を行ったが医師の開発した枝分かれ論理に全くかなわなかった。これが判別分析の研究の動機になり、多くの実証研究を通して判別分析の 4 つの問題を見つけた。2015 年の科研費シンポジュームで石井さんの発表で上記の 6 種のデータが公開されていることを知り、1970 年以降研究され成果の出ていない判別分析の**問題 5** を再認識し、新しく開発した判別理論の応用問題とした。そして僅か 56 日間で簡単に解決し**癌の遺伝子解析**を完成させた。結論は**最小誤分類数(MNM)**基準による**改定 IP-OLDF(RIP)**で 6 種の Microarray の 2 群は完全に分かれ**線形分離可能(LSD)**である(**Fact3**)。さらにそれが 64 組から 239 組の遺伝子数が n_i 個以下の**部分空間(SM)**と雑音空間に分割できた（**Fact4**)。本研究では、なぜ可能であるかを報告する。

1　癌の遺伝子解析

6 種類の Microarray データを RIP で判別すると簡単に LSD すなわち MNM=0 である(**Fact3**)。そして、n 個以下の判別係数だけが 0 でなく、残りが全て 0 になる。さらに判別すると幾つかの係数が 0 になった。これを繰り返し判別し 0 になる判別係数が現れないモデルを **SM1** と呼ぶことにした。そして SM1 を省いて判別すると別の SM2 が得られた。そこで LINGO を用いた**新手法 2** を RIP で作成した。そして、多くの SM の信号部分空間と雑音空間（MNM>=1）に簡単に分離できた。

2　癌の遺伝子診断

当初、癌の遺伝子診断までが統計の役割と考えていた。しかし SM は小標本であるので簡単に統計分析ができる。そこで 8 種類の判別分析と、標準的な統計手法で分析した。しかし RIP、改定 LP-OLDF、改定 IPLP-OLDF、H-SVM とロジスティック回帰だけが全ての SM を MNM あるいは NM が 0 と正しく判別できた。それ以外の判別分析は全ての SM を正しく判別できなかった。Alon らで 56 個の SM を見つけ、RatioSV(SV 間の距離の 2 を判別スコアの範囲で割った比(%))で評価したところ、RIP と H-SVM の範囲は[3%,25.3%]と[5.8%,32.6%]である。このような簡単な判別は研究対象にならないが、**1970 年から行われ 2004 年以降の早い時期に NIH が終息宣言を出し、医学研究が行われていない状況のようだ。**6 種類の Microarray データで得られた全ての SM で求めた **RIP の判別スコア(RipDSs)**を遺伝子の代わりに変数とした新しい 56 次元のデータ(Signal

Data)を作成した。これを階層型クラスター分析で分析すると、2群が完全に2つのクラスターに分かれた。そして健常群と癌群をPCAで分析すると、**図1**のようにPrin1上の負の軸上にclass1の症例が布置し、正の軸上にclass2の症例が布置することが分かった。これはRIPに限らず改定LP-OLDFあるいはH-SVMの判別スコアで作ったSignal Dataでも、6種類のMicroarrayでも、ほぼ同じであるので、2群がLSDであることが信号の定義でないかと考えられる。しかし他の多くの研究では信号と雑音の定義が明確でない。

3　問題5が困難な理由

多くの研究で、癌の遺伝子解析が困難な３つの理由を取り上げているが適切でない。研究がひと段落し、あまりにも簡単に解決したので、恐らく多くの研究者が疑っているのではないかと考えた。そこで多次元遺伝子データの諸問題の事実や問題を以下のように整理することにした。

Fact3：Microarrayデータを正しくLSDと判別できるのは、RIP、改定LP-OLDFとH-SVMだけである。統計的判別関数が対応できないのは LSD を正しく判別できないためである。分散共分散行列に基づき相関比最大化基準では LSD 判別に正しく対応できない。青嶋と矢田は、多くの統計研究者と異なり高次元空間のMicroarrayをPCAで分析すると第1固有値が極端に大きくなることから、2群が高次元空間でPrin1上にある2つの球上に布置することを示した。彼らの結果との整合性を探してきた。56個のSMの和集合をRIPで判別すると、SV=-1とSV=1になる症例が多くの割合を示し、SVが-1以下あるいはSVが1以上になる症例が少ない。それを56個のSMに分割すると図2のようSV上の割合が少なく裾が長くなるようだ。

Fact4：さらに、RIPと改定LP-OLDFはMicroarrayを多数のSM(MNM=0)とそれ以外の雑音空間(MNM>=1)に分割できる。しかしH-SVMはGolubらのMicroarrayで2000個程度の係数を0にするが、SMに分割できない。この理由は意外と簡単なことが分かった。MPの中で2次計画法(QP)だけが、目的関数を最大/最小にする最適解を定義域で1個見つける。これは統計的判別関数でも同じである。このため、部分空間でFull Modelより良いモデル(部分空間すなわち癌の遺伝子)を見つけるため“Feature selection”が必要になる。Pが1万個もあれば、NP-hardになる。しかし、線形計画法(LP)で定式化された改定LP-OLDFは、簡単にn個以下の最適なMNM=0になるSMを見つけることができる。**図1**はn=3でp=2(p<n)の一般的な事例である。そして図中の三角形がこの判別モデルの実行可能領域になる。即ち三角形の全ての点が目的関数を最適化する点になる。この実行可能領域は通常の数理計画法の常識と異なり、どの点も全て最適解になる。LPは、3つの三角形の頂点を選ぶ。この頂点は、一般的にn=3の中の2個の症例の超平面の交点である。一方横長データ(n<<p)では、最大n個の交点から選ばれる。少し無理はあるが、3番目のケースを省くと横長データのn=2<p=3の状況を作れる。この場合、ケース1とケース2とb1軸で囲まれた三角形が実行可能領域になる。ケース1とケース2の交点はp=2の点であるが、ケース1とケース2がb1軸と作る交点は、b2=0であり1次元の部分空間になる。1 万次元空間の実行可能領域のイメージは難しいが、各凸体の頂点は(p-n)個以上の遺伝子の判別係数を0にしたもので、容易に(p-n)個以上の判別係数が0になる。

以上の説明を聞いてもしっくり納得できない点が残る。なぜこれほど統計的判別関数や統計手法が役に立たないかである。**報告書の図3はSMを主成分分析で分析し、Prin1を横軸に、Prin2とPrin3の散布図である。このSMの遺伝子を用いたRIPの判別スコアは各軸に直交しないで、かつ長さも短い。主成分軸に投影すればさらに短くなる。即ち完全に大きなバラツキの中に埋もれている。癌と健常は明確に分離しているが、そのばらつきは全体の中で埋もれている。この 2 群が分離しているという信号は、RIP,改定 LP-OLDF、H-SVM とロジスティック回帰でしかわからないためである。当初、SM で 2 群は完全に分かれているので、これに含まれる遺伝子を癌遺伝子と考えた。しかし、統計的判別関数や統計手法は LSD である兆候を示さない。適切な判別関数で作られる総合特性軸で初めて2群が分かれていることが分かる。しかし、2群のばらつきが全体に比べ小さく直交もしていないので、信号を検出することが困難であるが結論である。**

RCT と DNN が医療水準の向上を駆動する

柳 本 武 美: 統計数理研究所

1. 序

標題の内 RCT は randomized controlled trial (無作為化比較試験) と DNN は deep neural networks (深層神経ネット) を表す。前者は randomized clinical trial (比較臨床試験) の略として使われることも多い。本稿では DNN を分類器に限って議論する。

2. 定着した RCT と途上の DNN

今日の日本では RCT は定着している。当初は患者に最善の治療を施す必要があるのに無作為に割り付けた実験 (試験) を実施することに反感をもつ医療関係者が多くて定着しなかった。しかし、結核・リュウマチなどの分野から始まり広く受け入れられている。臨床研究の主流を占めるに至っている。一端 RCT が導入されると真面目な医療関係者が多い我が国の文化の中で、質の高い試験が実施されている。この事情は以前の標本調査の導入と軌を一にしている。

近年の目覚ましい発展に伴い DNN への期待は大きい 1), 2)。一方で懐疑的な意見をもつ研究者も多い。一時的なブームではないかとする研究者も多い 2), 3)。人の感覚・知覚・認識を模して学習をさせようとする試みは、今日で言う機械学習の当初からあった。基本的な枠組みは今日でも変わっていない。特筆すべきブレークスルーがあったと言うより、計算機器と計算技能の向上とデータの保存・転送技術の進歩・普及によっている。基本的枠組みが変わっていないことは、その推測形式がごく常識的であることを示唆している。

3. 適用範囲の違いと広さ

RCT は医療行為の効果を確認するために実施される。医療行為は一面では意図的な侵襲行為であるから、誤った印象を受けやすい。実際歴史的には無効な医療行為が歓迎されていた例が多い。一方、DNN では機械的処理で解決できそうなタスクが多いことから歓迎されている。そのタスクには、囲碁・将棋などゲームの戦術なお、多様な分野が含まれることが既に実証されている。

医療行為の効果を評価するタスクを DNN で実施することは困難である。医師の経験と印象による評価が危ういことは既に経験済みである。勿論 DNN による効果の評価更には新しい治療法を探索する試みは一部で標榜されているが、著しい成功例は見当たらない。このことは RCT と DNN の守備範囲が異なることを示している。治療の効果は医療行為の中核であるから、RCT の役割は強調するまでもない。治療法と効果は区々であって統一的な科学法則は見つけがたい。多くの各個の試験が必要とされる。

一方で、DNN が解決に大きく寄与すると期待されるタスクは多い。検査画像の分類を通じた診断支援はその実用化が視野に入っている。医療データの一つの特性は、検査時あるいは診察時の情報が、以後により精度の高い情報が得られる可能性が大きいことである。その結果多層的な推論の可能性が期待できる。

心理学では人が物事を認識･理解する段階を三つに分ける。1) 感覚 (sensation) 2) 知覚 (perception) 3) 認識 ((re)cognition) である。従来の画像認識の用語は高い認識機能を目標にしたことを示している。DNN は感覚的データから知覚的反応を導くタスクに対して性能の良い結果を示した。他の方法で情報科学の分野で探ると、医療分野での成功例が乏しい。

4. 推論形式の類似性

RCT と DNN の推論形式は見た目にも大きく異なる。しかし丁寧に見ると共通点が多い。これ

らの共通点は、常識的な科学観に基づく推論の限界を示唆する。

4.1. 統計的方法としての DNN

RCT は将に統計的方法の典型である。単に解析の段階でのみ主要な役割を果たすのではない。データ収集の骨格を解析手法に合わせてプロトコール記載する面でも、入門統計学記載されているままである。DNN も統計的方法と軌を一にする。実際分類器では以下に挙げる点で共通点が見いだされる。

1) モデルの基本は線形回帰モデルある。
2) 活性化関数は、回帰モデルで用いられてきた関数である。
3) 分類問題が多項分布 $\mathrm{Mn}(\boldsymbol{p})$ の出現確率 $\boldsymbol{p}$ の推定を通じて分類法が構成される。
4) $\boldsymbol{p}$ の推定は自然母数 $\boldsymbol{\theta} = \log \boldsymbol{p}$ を推定する一般化線形回帰モデルである 2)。
5) 母数推定の基本は cross entroy 最小化で最尤推定法である。
6) 分類器の性能は試験データで評価される。

4.2. 手法の単機能性

複雑なデータを解析する手法は大抵探索的方法である。既に分かっている知識だけではなくてデータから新しい知識を得ようとする。前節で言及した手法がそうである。RCT は将に用法・用量から対象患者・期待する効果まで細かく分かっていることを前提とする。試験の結果は、想定通りの効果が認められるか否かの判断基準が得られるだけである。入力と出力のみを扱う、単一の機能しかない。

DNN は期待するタスクを解決することを目標とする。訓練データから分類器を構築して、試験データで性能を示す。それ以外に何らかの知識を得ようとする配慮は成されない。RCT と同様に単一の機能しかない。関連分野の研究者からは、特徴量など関連知識が得られないことが重大な欠点として批判された。同じ批判は臨床薬理学に対する薬理学からの批判、ベイズ法に対する頻度論者の批判にも見られる。

解析法の性能の良さと説明の分かり易さのどちらを優先させるかは、科学観に依存する。しかし、良い性能も示せないのにその理由が分かりやすく説明できる場合は多くないと思われる。目的を一つに絞って高い性能を実現させるという括りでは DNN は RCT と同じである。

5. 統計的推論から

著しい成功に拘わらず現行の DNN は統計学的視点から見て改善したい点が多く見受けられる。既存の統計的手法とは大きく異なるから、新しい問題が多い。また、用語と概念には曖昧な点が散見される。中でも principle of MLE とか generalized loss は重要な用語であるが、その意味は曖昧である。

更に、解析法の重視とデータの収集の軽視は特に顕著である。訓練データを原料として機械学習によりタスクの解法という製品を作り出すとの視点がある。実際解析の流れを図示する、データが底辺にあって解析結果が頂上にある。データの収集法について論じることは殆どない。

アカデミアでの深層学習の研究は、データ集合の個別データを収集する費用が小さい場合が多い。そのために、データ集合を効率よく育成する視点が弱い。実際、Goodfellow ら monograph 1) でもデータについては anomality とか augmentation が紹介されているに過ぎない。

文献： 1) Goodfellow, I. *et al.* (2016). *Deep Learning* , MIT press, Cambridge. 2) 麻生英樹 (2015). In 深層学習 近代科学社，東京 3) 岡谷 貴之 (2015). MLP, 東京

女性の経済・雇用に関する市区町村の指標と死亡率との関連

鈴木有佳

大阪医科大学医学部医学科 社会・行動科学教室

【背景】

個人の社会経済状況とその健康との関連については国内外に多くの研究が存在するが、女性の社会的地位に関する地域指標と健康の関連に関する研究は限られている。米国の研究では、女性の地位に関する地域指標が、女性の死亡率ならびに男性の死亡率にも影響すると示されている。(1) 日本は、経済参加等に関して、他国と比較して男女間の差が大きいことが知られている上、国内でも、地域間の社会格差が広がっていることが指摘されており、その健康影響を明らかにする必要がある。今回、日本人女性の経済・雇用に関する市区町村の指標とそこに住む男女の死亡率との関連について検討したので、発表する。

【方法】

女性の経済・雇用に関する地域指標については、過去に米国で実施された先行研究(1)を参考に、日本の実情に合わせ、各自治体の以下の 7 指標を選定し、公的統計情報からそれぞれ算出した。経済的自立度に関する指標として、女性の業主割合（対就業状況にある女性人口）(2)、女性の大学卒業以上の割合（対 25 歳以上女性）(2)、女性の生活保護率（対 20 歳以上女性人口（10 万人あたり））(3)の 3 指標を、雇用・所得に関する指標として、女性の労働力人口の割合（対 15 歳以上女性人口）(2)、就業者中の女性割合（対全就業者）(2)、女性管理・専門職割合（対全就業者）(2)、賃金男女比（産業計）（女性/男性）(4)の 4 指標を算出した。ただし、生活保護被保護率ならびに賃金平均については、市区町村データが公開されていなかったため、都道府県データを代入した。

目的変数には、平成 22 年人口動態調査(5)ならびに平成 22 年国勢調査(2)から算出した、各市区町村の男女別全死因、循環器疾患、がん、呼吸器疾患による死亡率を用いた。対象地域は人口 100 人以上の全 1,884 市区町村、対象年齢は全年齢とした。

それぞれの自治体について、各地域指標を平均値で除した値と男女の死亡率との関連について重回帰分析を行った。尚、調整変数には各自治体の年齢中央値(2)、人口密度(2)、年間収入のジニ係数(6)、平均等価可処分所得(7)および貧困率(7)を用いた。

【結果】

重回帰分析の結果、女性の業主割合は女性総死亡率と正に相関し（β=105.6, $p<0.001$）、女性の生活保護率（β=-29.2, $p=0.01$）、女性の大学卒業以上の割合（β=-43.9, $p=0.001$）、就業者中の女性割合（β=-248.6, $p=0.01$）および賃金男女比（β=-

395.1，p=0.03）は負に相関した。循環器疾患による死亡率も同様の結果であった。さらに、がん、呼吸器疾患による死亡率と就業者中の女性割合および女性管理・専門職割合はそれぞれ正に相関し、女性の大学卒業以上の割合および女性の労働力人口の割合はそれぞれ負に相関した。

また、男性においては、女性の業主割合（β=38.3，p=0.02)、女性の大学卒業以上の割合（β=−64.9，p＜0.001)、女性管理・専門職割合（β=−55.9，p=0.03）と総死亡率に関連が認められた。循環器疾患による死亡率も同様の結果であった。さらに、がんによる死亡率は女性の生活保護率と正に相関し、女性の業主割合、女性の大学卒業以上の割合および女性の労働力人口の割合と負に相関した。呼吸器疾患による死亡率と女性の業主割合および女性管理・専門職割合は正に相関した。

【結論】

各市区町村における女性の経済的自立度ならびに雇用・所得に関する指標は、女性の死亡率に加え、その地域の男性の死亡率にも関連することが明らかになった。よって、女性の就労支援を進める等、市区町村の状況を改善することが、男女の死亡率の減少に寄与する可能性が示された。

【謝辞】

本研究は、JSPS 研究活動スタート支援 18H06403 ならびに平成 28 年度厚生労働科学研究費補助金（女性の健康の包括的支援政策研究事業：H28-女性-一般-001）の助成を受けて行われた。

本シンポジウムでの発表にあたり、ご指導を賜った谷川武先生、池田愛先生（順天堂大学)、丸山広達先生（愛媛大学)、磯博康先生（大阪大学)、本庄かおり先生（大阪医科大学）に感謝の意を表する。

References

1. Kawachi I, Kennedy BP, Gupta V, Prothrow-Stith D. Women's status and the health of women and men: a view from the States. Soc Sci Med. 1999;48:21-32.
2. 総務省統計局. 平成 22 年国勢調査. 2011.
3. 厚生労働省社会・援護局保護課. 平成 22 年度被保護者全国一斉調査　被保護人員数、年齢階級・都道府県－指定都市－中核市別.
4. 厚生労働省政策統括官付参事官付賃金福祉統計室. 平成 22 年賃金構造基本統計調査. 2011.
5. 厚生労働省. 平成 22 年人口動態調査. 2011.
6. 総務省統計局. 平成 26 年全国消費実態調査. 2016.
7. 厚生労働省. 平成 22 年国民生活基礎調査. 2011.

マイクロシミュレーションモデルを用いた大腸がん検診における受診年齢上限の検討

福井敬祐[1]・加茂憲一[2]・伊藤ゆり[1]・片野田耕太[3]・中山富雄[4]
[1] 大阪医科大学　研究支援センター
[2] 札幌医科大学医療人育成センター
[3] 国立がん研究センターがん対策情報センター
[4] 国立がん研究センター社会と健康研究センター

【背景】大腸がんは、わが国で二番目に多いがんであり、早期発見の場合はほとんどが治癒可能である。1990 年代に導入された便潜血検査による大腸がん検診の死亡率減少効果は複数の無作為化比較試験で示されており、科学的根拠に基づくがん検診として推奨されている．一方で，近年，高齢のがん検診受診者の増加に伴い、検診による死亡率減少効果等の利益のみならず検診による不利益とのバランスを鑑み，検診受診年齢の上限に関する議論が活発になっている．実際に，厚生労働省の「がん検診のあり方に関する検討会」においては，検診対象年齢の議論が行われている　しかし，これまでのように利益と不利益を改めて無作為化比較試験により定量化し　具体的な年齢上限を検討することは資源や，倫理的な面から現実的でなく、欧米諸国で多用されているマイクロシミュレーション（MS）を用いるべきである．

【目的】日本のデータに基づき開発された大腸がんに関する MS モデルを用いて，便潜血検査による大腸がん検診の年齢上限の検討に活用できる資料を作成する。

【方法】2012 年時点で 30 歳である男女 100 万人の仮想的コホートを対象に，100 歳まで加齢する MS モデルを開発した．

開発は，図 1 にあげた自然史モデルにおいて，各コンパートメント間の移動を数理的にモデリングすることで行った．使用したソフトは R 言語 ver 3.3.1 である．

作成した MS モデルにおいて，2012 年時点の性・年齢階級別検診受診率を基に現実を反映したコホート（上限なしコホート）および検診年齢上限を 65, 70, 75, 80, 85 歳と設定した際の仮想コホート（上限ありコホート）をシミュレートした．　上限なしコホートと上限ありコホートにおける大腸がん死亡と有害事象発生を比較し，回避死亡および有害事象発生をそれぞれ検診の利益および不利益と定義し，比較検討を行った．

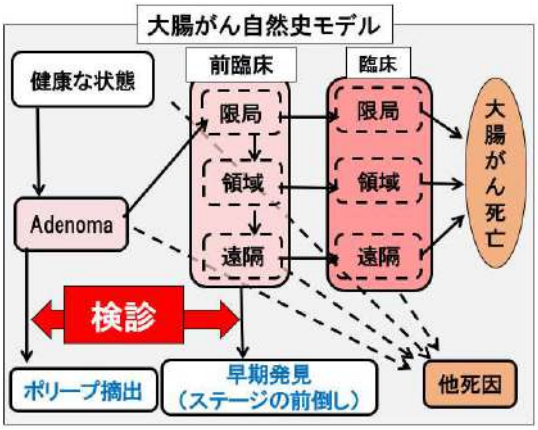

図 1 大腸がん自然史

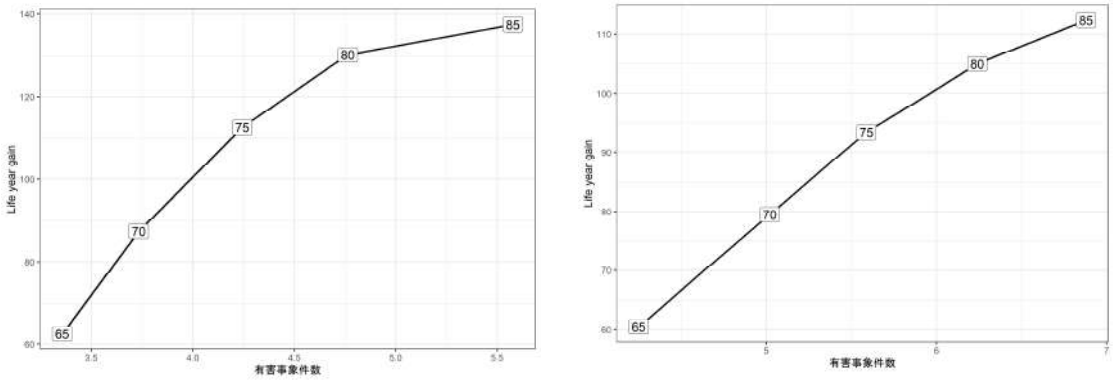

図 2　マイクロシミュレーションモデルによる上限年齢の検討結果
（左：男性, 右：女性）

【結果】不利益および利益が年齢上限の設定によりどう変化するかをみた. 男性においては, 検診年齢上限を 85 歳から 80 歳に引き下げた場合、不利益 1 件減少あたりの利益減少は 1.5 人となった。これが 80 歳から 75 歳では 2.23 人、75 歳から 70 歳では 13.2 人、70 歳から 65 歳では 16.1 人、65 歳から 60 歳では 14.3 人と算出され, 75 歳前後で大きく傾向が変化することが示唆された.

【考察】がん対策において重要な役割を担うがん検診は，対象者を適切に選択することにより有効性が高まることが期待される。特に年齢に関しては,体力の衰えが顕著となる高齢者における不利益を考慮する必要があることが MS により示された. 今後, 年齢上限の設定のために, 本研究の成果を一つの根拠資料として活用されることを期待したい.

シミュレーションモデルによる乳がん予防・検診の評価

堀 芽久美[1], 齋藤 英子[1], 大久保 亮[2], 清水 陽一[2], 小手森 綾香[2], 街 勝憲[2], 高橋 宏和[2]
1. 国立がん研究センター がん対策情報センター
2. 国立がん研究センター　社会と健康研究センター

1. がん対策とシミュレーション研究

シミュレーションの利点の一つは、現実に再現不可能な事象もモデルによって再現できることである。シミュレーション研究は質問に応じたシナリオを作成することで、今欲しい情報を生みだすことができる。

米国 NCI (**N**ational **C**ancer **I**nstitute)は大学・病院で組織されるシミュレーションモデリングネットワーク CISNET (the **C**ancer **I**ntervention and **S**urveillance Modeling **Net**work)を支援している[1]。CISNET では個人のイベント発生を確率的に表現し、イベントの発生や選択行動の確率表現をシナリオ化することで、実際には観察できない状況を再現し、その状況変化の影響を推計している。CISNET は Centers for Disease Control (CDC)や Agency for Health Care Research and Quality (AHRQ)等と連携しており、結果は最適ながん対策の策定に貢献している[2-3]。

CISNET では一つの部位に対して複数のモデリンググループがあり、それぞれが独立してモデルを構築しており、部位グループごとに NCI と外部の大学または病院にコーディネーターが選任されている。NCI におけるコーディネーターの役割には、政策決定に関わる機関との連携、がんの予防、治療、スクリーニングに関する臨床研究あるいは疫学研究に関わる研究者への情報提供体制の整備が含まれる。CISNET は研究で得た情報や結果を可能な限りオープンとし、情報を共有しており、同一分野の研究者やその結果の利用者が集まるプラットフォームを形成している。コラボレーションを望む研究者に対する参加承認手続きも明確で、国際的なコラボレーションがスムーズに行える体制が整っており、シミュレーション研究の発展こそが重要な課題とされている。NCI 外部のコーディネートセンターは、その部位グループを構成する複数のモデリンググループ間の調整を行っている。スクリーニングによる過剰診断などそれぞれの研究テーマに対して、必要なデータやアウトカムの選択など、モデリンググループ間での調整が積極的に行われる。また、CISNET では若手研究者育成は研究グループの成果として評価する仕組みで、研究計画の初期段階から育成プランが作成されている。規模の大きい計画を長期的に進めていくために必要不可欠であり、日本でも若手育成を評価する仕組みが重要だろう。

CISNET が構築した研究手法やネットワークは、日本におけるシミュレーション研究の発展に重要な手がかりである。シミュレーションですべての事象を説明するのは不可能であるが、実験研究や観察研究とともにシミュレーション研究で得られた知見を組み合わせることが、これまでより早く、効率的に、社会が求める質問に答える近道になるだろう。

2. Health behavior of breast cancer survivor simulation (HERBS) study

乳がんと診断された後の予後改善に関して日本人を対象とした研究は少なく、我が国ではいまだ乳がんサバイバーの予後改善につながる具体的なガイドラインは未整備である。

本研究は、サバイバーシップケア（運動・栄養介入、サーベイランス、患者ケア）を行った場合、乳がんサバイバーの予後がどの程度改善するかについて、現在まで介入研究では用いられていなかったマイクロシミュレーション手法を活用することで、乳がんサバイバーに効果的な介入を明らかにすることを目的とする。

本研究では以下のことを明らかにする。

1) 乳がんサバイバーにおいて、がんと診断された後の運動量がその後のアウトカム（死亡率、再発率、QOL）にどのような影響を与えているか。さらに、診断時の乳がん進行度別・年齢別に、アウトカムを改善するための最適な運動プログラムを明らかにする。
2) 乳がんサバイバーにおいて、がんと診断された後の食物繊維やイソフラボンの摂取量がその後のアウトカムにどのように影響するか。また、日本人女性の年齢・BMI別の最適な栄養摂取基準を明らかにする。
3) 乳がんサバイバーにおける診断後サーベイランスについて、利益と害の双方から検証し、最適なサーベイランスのあり方（対象年齢、サーベイランスの間隔）を明らかにする。

本研究はマイクロシミュレーション手法を活用することで、乳がんサバイバーの生存率向上・再発防止・QOL 改善のための最適な健康行動介入プログラムを具体化し、今後のガイドライン立案の基盤となることが期待される。

3. 引用文献

[1] National Cancer Institute (NCI). Cancer Intervention and Surveillance Modeling Network. Available from: https://cisnet.cancer.gov/. [Accessed December 2018].

[2] Mandelblatt JS, Stout NK, Schechter CB, van den Broek JJ, Miglioretti DL, Krapcho M, et al. Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies. Ann Intern Med. 2016;16;164(4):215-25.

[3] Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Breast Cancer Working Group of the Cancer Intervention and Surveillance Modeling Network. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 2009;151:738-47.

特異ウィシャート行列の最大・最小固有値分布について

東京理科大学大学院　　清水康希
東京理科大学・理学部　　橋口博樹
東京理科大学・理工学部　岩下登志也

1　はじめに

ウィシャート行列の固有値分布は，主成分分析の結果の安定性・信頼性を検証する際に重要である．近年 DNA マイクロアレイデータのような標本数 n より次元数 m が大きい場合のデータ解析の必要性が高まっている．このような観点から，$n < m$ の場合での固有値分布論の理論の構築が必要である．

非特異ウィシャート行列の固有値の正確分布や近似分布の研究は数多くある．しかし特異の場合は，ウィシャート行列の逆行列が存在しないことや固有値に 0 が含まれるため，近似分布の導出には，母集団固有値にある仮定をおくなどの方法が必要である．

本研究では，Srivastava(2003) が導出した特異ウィシャート行列の固有値の同時密度関数をはじめに取り上げる．そこでスティーフェル多様体上の積分を直交群上の積分に変換する定理を与えることで，最大固有値の正確分布及び，近似分布が導出できることを示す．

2　特異ウィシャート行列の最大固有値の正確分布

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．このとき，ランダム行列 $W = XX'$ は自由度 n，共分散行列 Σ の特異ウィシャート分布に従う．特異ウィシャート分布に従う W を特異ウィシャート行列という．Srivastava (2003) は特異ウィシャート行列 W のゼロでない n 個の固有値 $\ell_1, \ldots, \ell_n$ の同時密度関数を次のように求めた．

$$f(\ell_1,\ldots,\ell_n) = \frac{2^{-nm/2}\pi^{n^2/2}|\Sigma|^{-n/2}}{\Gamma_n(\frac{n}{2})\Gamma_n(\frac{m}{2})} \left(\prod_{i=1}^{n} \ell_i^{(m-n-1)/2}\right) \left(\prod_{i<j}(\ell_i-\ell_j)\right) \int_{V_{n,m}} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1}H_1L_1H_1'\right)(dH_1)\ , \quad (1)$$

ただし，$\Gamma_n(a)$ は多変量ガンマ関数，$V_{n,m} = \{H_1 : m \times n \mid H_1'H_1 = I_n\}$ となるスティーフェル多様体，(dH_1) は $V_{n,m}$ のハール測度，$L_1 = \mathrm{diag}(\ell_1, \ell_2, \ldots, \ell_n)$ を表す．もし $n = m$ ならば，$V_{m,m}$ は直交行列全体の集合 $O(m)$ である．

特異ウィシャート行列の固有値分布を導出するために，異なる型の行列を引数にもつ超幾何関数を導入する．X を m 次対称行列，Y を n 次対称行列として，$m \geq n$ とする．2 変量の超幾何関数 ${}_pF_q^{(m,n)}$ を次のように定義する．

$${}_pF_q^{(m,n)}(\boldsymbol{\alpha},\boldsymbol{\beta};X,Y) = \sum_{k=0}^{\infty}\sum_{\kappa\in P_n^k} \frac{(\alpha_1)_\kappa\cdots(\alpha_p)_\kappa}{(\beta_1)_\kappa\cdots(\beta_q)_\kappa} \frac{C_\kappa(X)C_\kappa(Y)}{k!C_\kappa(I_m)}\ ,$$

ただし，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$，$P_n^k = \{\ \kappa = (k_1, \ldots, k_n) | k_1 + \cdots + k_n = k,\ k_1 \geq k_2 \cdots \geq k_n \geq 0\}$，$C_\kappa(X)$ は分割 κ の複素ゾーナル多項式，$(a)_\kappa$ は分割 κ のポッホハマー記号である．

Theorem 1. A を m 次対称行列，$B_1 = \mathrm{diag}(b_1 \ldots, b_n)$，$m$ 次対角行列 $B = \mathrm{diag}(b_1 \ldots, b_n, 0 \ldots, 0)$，$H_1$ は $m \times n$ 行列で，m 次直交行列 H の分割 $H = (H_1, H_2)$ から構成されるものとする．この時，以下が成立する．

$$\begin{aligned}\int_{H_1\in V_{n,m}} {}_pF_q(\boldsymbol{\alpha};\boldsymbol{\beta};AH_1B_1H_1')(dH_1) &= \frac{\mathrm{Vol}(O(m))}{\mathrm{Vol}(V_{n,m})\mathrm{Vol}(O(m-n))}\int_{H\in O(m)} {}_pF_q(\boldsymbol{\alpha};\boldsymbol{\beta};AHBH')(dH)\\ &= \frac{\mathrm{Vol}(O(m))}{\mathrm{Vol}(V_{n,m})\mathrm{Vol}(O(m-n))}{}_pF_q^{(m,n)}(\boldsymbol{\alpha};\boldsymbol{\beta};A,B_1)\ .\end{aligned}$$

ただし，$\mathrm{Vol}(V_{n,m}) = \frac{2^n\pi^{nm/2}}{\Gamma_n(m/2)}$.

Corollary 1. $L_1 = \mathrm{diag}(\ell_1, \ldots, \ell_n)$，$m$ 次対角行列 $L = \mathrm{diag}(\ell_1, \ldots, \ell_n, 0, \ldots, 0)$ とする．この時，次式が成立する．

$$\int_{V_{n,m}} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1}H_1 L_1 H_1'\right)(dH_1) = \frac{\mathrm{Vol}(O(m))}{\mathrm{Vol}(V_{n,m})\mathrm{Vol}(O(m-n))}{}_0F_0{}^{(m)}\left(-\frac{1}{2}\Sigma^{-1}, L\right) .$$

Sugiyama(1967) は，次の Lemma 1 を導出し，非特異ウィシャート行列の同時密度関数に Lemma 1 を用いることで非特異フィシャート行列の最大固有値の分布関数を導いた．また Shinozaki et al.(2018) では，正規母集団の拡張である楕円母集団下で同様の手法を用いて，楕円ウィシャート行列の最大固有値の分布関数を導出した．特異ウィシャート行列の場合も同時密度関数に Lemma 1 を用いることで最大固有値の分布関数が導出できることを示す．

Lemma 1. 対角行列 $X_1 = \mathrm{diag}(1, x_1, \ldots x_n)$，$X_2 = \mathrm{diag}(x_2, \ldots x_n)$, $x_2 > \cdots > x_n > 0$ とする．

$$\begin{aligned}\int_{1>x_2>\cdots x_n>0} |X_2|^{t-(n+1)/2} C_\kappa(X_1) \prod_{i=2}^{n}(1-x_i)\prod_{i<j}(x_i - x_j)\prod_{i=1}^{n} dx_i \\ = (nt+k)(\Gamma_n(n/2)/\pi^{n^2/2})\frac{\Gamma_n(t,\kappa)\Gamma_n((n+1)/2)C_\kappa(I_n)}{\Gamma_n(t+(n+1)/2,\kappa)} ,\end{aligned}$$

ただし，$\mathrm{Re}(t) > (n-1)/2$，$\Gamma_n(\alpha, \kappa) = (\alpha)_\kappa \Gamma(\alpha)$.

Theorem 2. $W \sim W_m(n, \Sigma)$， $m > n$ とする．このとき，W の最大固有値 ℓ_1 の分布関数は次のように表される．

$$\begin{aligned}\Pr(\ell_1 < x) &= \frac{\mathrm{Vol}(O(m))}{\mathrm{Vol}(V_{n,m})\mathrm{Vol}(O(m-n))}\frac{\Gamma_n(n+1/2)(\frac{x}{2})^{nm/2}}{\Gamma_n(n+m+1/2)|\Sigma|^{n/2}}\sum_{k=0}^{\infty}\sum_{\kappa\in P_n^k}\frac{(m/2)_\kappa C_\kappa(-\frac{1}{2}x\Sigma^{-1})C_\kappa(I_n)}{(n+m+1/2)_\kappa k! C_\kappa(I_m)} \\ &= \frac{\mathrm{Vol}(O(m))}{\mathrm{Vol}(V_{n,m})\mathrm{Vol}(O(m-n))}\frac{\Gamma_n(n+1/2)(\frac{x}{2})^{nm/2}}{\Gamma_n(n+m+1/2)|\Sigma|^{n/2}}{}_1F_1{}^{(m,n)}\left(\frac{m}{2}; \frac{n+m+1}{2}; -\frac{1}{2}x\Sigma^{-1}, I_n\right) .\end{aligned}$$

3　特異ウィシャート行列の固有値分布の χ^2 近似

Corollary 1 の ${}_0F_0{}^{(m)}$ に対して， Butler and Wood(2005) で与えられるラプラス近似を適用する次の近似式が成立する．

$${}_0F_0{}^{(m)}\left(-\frac{1}{2}\Sigma^{-1}, L\right) \approx 2^n \frac{\mathrm{Vol}(O(m-n))}{\mathrm{Vol}(O(m))}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\ell_i}{\lambda_i}\right)\prod_{i<j}^{n}\left(\frac{2\pi}{c_{ij}}\right)^{1/2}\prod_{i=1}^{n}\prod_{j=n+1}^{m}\left(\frac{2\pi}{d_{ij}}\right)^{1/2} , \tag{2}$$

ただし，$c_{ij} = \frac{(\ell_i - \ell_j)(\lambda_i - \lambda_j)}{\lambda_i \lambda_j}$，$d_{ij} = \frac{\ell_i(\lambda_i - \lambda_j)}{\lambda_i \lambda_j}$ である．(1)，(2) より，$f(\ell_1 \ldots, \ell_n)$ のラプラス近似を得る．

Theorem 3. 特異ウィシャート行列 $W \sim W_m(n, \Sigma)$，$n < m$，母共分散行列 Σ の固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$，$W$ の固有値を $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_n > 0$ この時隣接固有値の比 ρ が $\rho = \max\{\frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_2}, \cdots, \frac{\lambda_n}{\lambda_{n-1}}\} \to 0$ の時，特異ウィシャート行列の第 i 固有値 ℓ_i に対して次の近似式が成立する．

$$\Pr[\ell_i < x] \sim \chi^2_{n-i+1}\left(\frac{x}{\lambda_i}\right)$$

参考文献

[1] Butler, R. W., Wood, A. T. A. (2005). Laplace approximations to hypergeometric functions of two matrix arguments, *Journal of Multivariate Analysis*, **94**(1), 1–18.

[2] Shinozaki, A., Hashiguchi, H., Iwashita, T. (2018). Distribution of the largest eigenvalue of an elliptical Wishart matrix and its simulation, *Journal of the Japanese Society of Computational Statistics*, **19**(1) 45–56.

[3] Srivastava, M. S. (2003). Singular Wishart and multivariate beta distributions, *The Annals of Statistics*, **31**(5), 1537–1560.

[4] Sugiyama, T. (1967). On the distribution of the largest latent root of the covariance matrix, *The Annals of Mathematical Statistics*, **38**(4), 1148–1151.

ウィッシャート行列と正規ベクトルの積の密度関数とモーメント

米永 航志朗　　鈴川晶夫
北海道大学経済学院　　北海道大学公共政策大学院

本報告はウィッシャート行列と正規ベクトルの積に関する密度関数とモーメント公式についてである．多変量正規分布とウィッシャート分布は多変量解析の様々な場面において現れる．例えば多変量正規分布を母集団分布として想定した場合に, 代表的な多変量解析手法である主成分分析や判別分析において, ウィッシャート分布, あるいはその関数が現れることは周知である．また, $X_1, \cdots, X_n$ を $N_k(\mu, \Sigma)$ からの独立同分布な標本としたときに, ただし, $N_k(\mu, \Sigma)$ は平均ベクトル μ, 分散共分散行列 Σ の k 次元正規分布を表しているが, 母平均 μ の不偏推定量 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ は $N_k(\mu, \Sigma/n)$ に従い, 母分散共分散行列の不偏推定量である $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$ については $(n-1)S \sim W_k(n-1, \Sigma)$, すなわち自由度 $n-1$, 分散共分散行列 Σ のウィッシャート分布となる．

以上のように, 多変量正規分布とウィッシャート分布は多変量解析において重要であるのだが, これまで統計学的な文脈において，ウィッシャート行列と正規ベクトルの積に関する結果は Bodnar&Okhrin(2011) [1], Bodnar&Mazur&Okhrin(2013) [2] そして Kotsiuba&Mazur(2016) [3] を除いてあまり議論されてこなかった．ウィッシャート行列の逆行列は逆ウィッシャート行列と呼ばれるのであるが, Bodnar&Okhrin(2011) [1] は逆ウィッシャート行列と正規ベクトルの積について密度関数の導出をした後，ファイナンスにおけるポートフォリオ理論への応用や判別分析に対する応用を議論しており, Bodnar&Mazur&Okhrin(2013) [2] ではウィッシャート行列と正規ベクトルの積の確率表現と密度関数を導出した後，その密度関数の近似について議論している．また, Kotsiuba&Mazur(2016) [3] においては逆ウィッシャート行列と正規ベクトルの積の確率表現を導出した後, それの極限分布と密度関数の近似について議論している．ここで, 確率表現という言葉は, ある確率変数と同じ分布を持つような別の確率変数によってその確率変数を表すという意味合いであると考えられる．

ところで，判別分析においては通常, 逆ウィッシャート行列と正規ベクトルとの積が大切であるのだが, ベイズ統計学の観点からみた場合には, 分散共分散行列の事後分布として逆ウィッシャート分布を得ることを考えると, ウィッシャート行列と正規ベクトルの積もまた重要である．例として 2 群判別の例を考えてみたい．すなわち, 2 つの正規母集団 $N_k(\mu_1, \Sigma)$ と $N_k(\mu_2, \Sigma)$ があり, いま新しい観測ベクトル $x = (x_1, \cdots, x_p)$ が得られたとき, それがどちらの母集団に属するかを判別する数学的基準として線形判別関数 $L(x) = (\mu_1 - \mu_2)'\Sigma^{-1}x$ を得る．そして線形判別関数の値に

よって新しい観測ベクトルがどちらの母集団に属するのかを決めることになる. 今, μ_1, μ_2 は共に既知とすれば, ウィッシャート行列と確率ベクトルの積の 1 次結合の分布が重要となるのは明白である. さらに, 新たに得られた観測ベクトルはもとの正規母集団と同じ分散共分散行列を持っているとは限らない. 例えば, 先の 2 つの正規母集団が健常者と非健常者の母集団と考えてみたい. このとき何らかの理由で非健常者のコレステロール値の分散が変わるというケースも考えられる. このような場合には判別関数の分布がどうなるのかということに関心が出てくる. 本報告ではこのようにウィッシャート行列と正規ベクトルの分散共分散行列が異なっている場合に対応した議論と平均ベクトルが既知の場合の判別関数の分布に対応した議論を行っていく.

本報告でははじめに, 異なる分散共分散行列を持つウィッシャート行列と正規ベクトルの積に関して確率表現を与えたのち, それを用いて密度関数を導出する. さらに, それらがいくつかの特殊な条件の下でどのように表されるかについて考察する. さらに分布の重要な情報を含むモーメントについて 4 次までのモーメント公式を与える. 本報告では 3 次モーメントと 4 次モーメントは確率表現に基づいてモーメント公式が計算されているが, このうち 3 次モーメントについては条件付き期待値の考え方を用いて別な表現を与えた. また, ウィッシャート行列と正規ベクトルの積に関する密度関数は積分を含むため, その形状を把握することができない. そこで, 数値積分によってグラフを描き, それとともにシミュレーションを行い, 数値積分によって描いたグラフと比較することで今回考察した理論的結果の正当性も確認する.

参考文献

[1] Bodnar, T.& Okhrin, Y.(2011) On the Product of Inverse Wishart and Normal Distributions with Applications to Discriminant Analysis and Portfolio Theory. *Scandi. J. Statist.* **38**, 311-331.

[2] Bodnar, T.& Mazur, S. & Okhrin, Y.(2013) On the exact and approximate distributions of the product of a Wishart matrix with a normal vector. *J. Multivariate Anal.* **122**, 70-81.

[3] Kotsiuba, I.& Mazur, S. (2016) On the asymptotic and approximate distributions of the product of an inverse Wishart matrix and a Gaussian vector. *Theor. Probability and Math. Statist.***93**, 103-112.

[4] Muirhead, R. J(1982). *Aspects of multivariate statistical theory.* Wiley, New York.

[5] Gupta, A. K. & Nagar, D. K(2000). *Matrix Variate Distributions.* Chapman and Hall/CRC.

[6] Mathai, A. M. & Provost, S.B(1992) *Quadratic forms in random variables.* Marcel Dekker, Inc, New York.

ヒルベルト空間における正規性の検定の実際的側面

千葉大・融合理工学府　牧草 夏実

1　はじめに

P を確率分布とするとき，データ $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ に基づく

$$\text{帰無仮説 } H_0 : P = N(m_0, \Sigma_0) \text{ v.s 対立仮説 } H_1 : P \neq N(m_0, \Sigma_0)$$

の検定を “正規性の検定” という．ここで，$m_0 = \mathbb{E}[X_1]$，$\Sigma_0 = V[X_1]$ である．ユークリッド空間では，1 変量，そして多変量でも様々な検定方法が議論されている．この正規性の検定の数理的な一般化を考え，ヒルベルト空間に値を取る確率変数に対する正規性の検定について考える．特に，Maximum Mean Discrepancy に基づく正規性の検定について，理論的側面と，実際的側面から議論を与える．

2　ヒルベルト空間における正規分布

ヒルベルト空間 $\mathcal{H}$ に値を取る確率変数 X が正規分布に従うとは，任意の $h \in \mathcal{H}$ に対して，$\mathcal{H}$ における内積 $\langle X, h\rangle_{\mathcal{H}}$ が 1 変量の正規分布の確率変数となることである．また，$\mathbb{E}\left[\|X\|_{\mathcal{H}}^2\right] < \infty$ とすると，任意の $h, h' \in \mathcal{H}$ に対し，$\langle m, h\rangle_{\mathcal{H}} = \mathbb{E}_X\left[\langle X, h\rangle_{\mathcal{H}}\right]$，$\langle \Sigma h, h'\rangle_{\mathcal{H}} = \mathrm{cov}\left(\langle X, h\rangle_{\mathcal{H}}, \langle X, h'\rangle_{\mathcal{H}}\right)$ を満たす $m \in \mathcal{H}$ と作用素 $\Sigma \in HS(\mathcal{H})$ が一意に存在し，m，Σ をそれぞれ，X の期待値，共分散作用素といい，X の従う正規分布を $N(m, \Sigma)$ と表す．ここで，$HS(\mathcal{H})$ は $\mathcal{H}$ から $\mathcal{H}$ へのヒルベルトシュミット作用素の空間である．

3　検定統計量の構築

正定値カーネル k，k に対応する再生核ヒルベルト空間を $H(k)$ とする．このとき，適当な条件の下で，$\mathcal{H}$ 上の 2 つの確率分布 P と $N(m_0, \Sigma_0)$ との違いは，MMD（Maximum Mean Discrepancy）

$$\Delta(P, N(m_0, \Sigma_0)) = ||\mu(P) - \mu(N(m_0, \Sigma_0))||_{H(k)}$$

によってそれらの違いを見ることができる ([1] 参照)．ここで，$\mu(P) = \mathbb{E}_{X\sim P}\left[k(\cdot, X)\right]$，$\mu(N(m_0, \Sigma_0)) = \mathbb{E}_{X\sim N(m_0,\Sigma_0)}\left[k(\cdot, X)\right]$ は確率分布 P，$N(m_0, \Sigma_0)$ のカーネル k による埋め込みである．$\mathcal{H}$ におけるデータ $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ を用いて，$\Delta^2 = ||\mu(P) - \mu(N(m_0, \Sigma_0))||_{H(k)}^2$ は

$$\widehat{\Delta}^2 = \left|\left|\frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i) - \mu(N(\widehat{m}, \widehat{\Sigma}))\right|\right|_{H(k)}^2$$

によって推定される．ここで，$\widehat{m}$，$\widehat{\Sigma}$ は m，Σ の一致推定量である．さらに，$\mu(N(m, \Sigma))$ が (m, Σ) に関して連続であるとき，$\widehat{\Delta}^2$ は Δ^2 の一致推定量となることが [3] によって知られている．

本発表では，この MMD に基づく正規性の検定について理論的側面と実際的側面について議論を与えた．

理論的結果として，検定統計量 $\widehat{\Delta}^2$ の漸近挙動について報告をした．特に，帰無仮説 $H_0 : P = N(m_0, \Sigma_0)$ のもとで，$n\widehat{\Delta}^2$ の漸近分布が，独立な自由度 1 の χ^2 分布の重みつき無限和であること，また，対立仮説 $H_1 : P \neq N(m_0, \Sigma_0)$ のもとでの $n\widehat{\Delta}^2$ の漸近分布が，平均が 0 の正規分布になっていること，局所対立仮説 $P = P_n = (1 - 1/\sqrt{n})N(m_0, \Sigma_0) + (1/\sqrt{n})Q$ のもとでの $n\widehat{\Delta}^2$ の漸近分布が，独立な非心 χ^2 分布の重みつき無限和であることについて報告を行った．

$n\widehat{\Delta}^2$ による正規性の検定の実際的な挙動に関し，$\mathcal{H} = \mathbb{R}^d$，$k$ をガウスカーネルとした場合で調べた結果を報告した．特に，$n\widehat{\Delta}^2$ の帰無分布の棄却点，および検出力について報告を行った．棄却点の推定として，従来のシミュレーションを用いた結果を報告するとともに，漸近帰無分布に基づく 2 つの方法との比較検討についても報告を行った．漸近帰無分布に基づく方法の一つとして，適当な定数 $c > 0$，$r \in \mathbb{N}$ の単一重み付きカイ 2 乗 $c\chi^2_r$ によって漸近帰無分布の裾を近似する方法を用いた．もう一つの方法は，あるグラム行列の固有値によって漸近帰無分布の重みを推定することで，漸近帰無分布を近似する方法であり，[2] でも議論されている．これらの方法を比較した結果，従来のシミュレーションにより導出した棄却点と，漸近帰無分布に基づく方法による棄却点は，どちらもほぼ同じ値をとっており，高次元データに対しては，実行時間を考慮すると，漸近帰無分布に基づく方法によって棄却点を導出すれば実用上十分であることがわかった．

対立仮説のもとでの挙動としては，分布 P として一様分布，指数分布，Σ_0 として，$\Sigma_0 = I_d$ の場合と，適当な共分散をもつ場合の各ケースについての検出力をシミュレーションによって求めた結果を報告した．結果としては，次元 d が大きくない場合は，どのケースでも検出力はほとんど 1 に近い値をとったが，d を大きくした場合，特に一様分布に対しては，検出力はあまり大きな値を取らない結果となった．

参考文献

[1] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schlkopf, and A. Smola (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Procesing Systems*, volume 19 of *MIT Press*, *Cambridge*.

[2] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur (2009). A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems* volume 22.

[3] J. Kellner and A. Celisse (2015). A One-Sample Test for Normality with Kernel Methods, *arXiv preprint arXiv:1507.02904v1*.

楕円対称性の必要性検定

東京理科大学理工
岩下 登志也

Karlsruher Institut für Technologie
Bernhard Klar

東京理科大学理
橋口 博樹

1. はじめに

楕円分布 (族) は, 多変量正規分布 の一般化であり, 重要な多変量分布のクラスである. それゆえ, 母集団として楕円分布を仮定することは, 多くの多変量解析にとって極めて重要であり, このような観点から, 採られた標本が楕円母集団から採られたものであるか否かを検定することは, 欠くことのできないものである.

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ を p-次元 (列) 確率ベクトルとする. $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ により対応する $p \times N$ 観測行列 とすると, 標本平均ベクトル 及び標本共分散行列 それぞれ

$$\bar{\boldsymbol{X}} = N^{-1} X \boldsymbol{j}_N, \tag{1}$$

$$S = n^{-1} X Q_N X', \qquad n = N - 1 \geq p, \tag{2}$$

と表される. ここに,

$$\boldsymbol{j}_N = (1, \ldots, 1)' \in \mathbb{R}^N, \qquad Q_N = I_N - N^{-1} \boldsymbol{j}_N \boldsymbol{j}_N'. \tag{3}$$

本報告では, スケール化された残差の同時分布, 即ち $p \times N$ の確率行列

$$W = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N] = S^{-1/2} [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N] Q_N = S^{-1/2} X Q_N, \tag{4}$$

に基づく新たな統計量を考案し, 楕円対称性の必要条件検定の新たな手法の提案をした. ここに,

$$\boldsymbol{W}_i = S^{-1/2} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}), \qquad i = 1, \ldots, N, \tag{5}$$

$S^{-1/2}$ は S の対称平方根の逆行列を表す。

2. 主たる理論的結果

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ を $\boldsymbol{X} \sim \mathrm{EC}_p(\boldsymbol{0}, \Lambda)$ のランダムコピー, $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ を $p \times N$ の観測行列 (observation matrix) とする. 次のような左球形分布 $\mathrm{LS}_{p \times N}(\phi)$ のサブクラス $\mathfrak{F}_{p \times N}$

$$\mathfrak{F}_{p \times N} = \{X(p \times N) \sim \mathrm{LS}_{p \times N}(\phi_X);\ \text{the distribution of } X\boldsymbol{a} \text{ depends on } \boldsymbol{a} \in \mathbb{R}^N \text{ only through } \boldsymbol{a}'\boldsymbol{a}\} \tag{6}$$

と Iwashita and Klar (2014) の結果により次のような結果を得た.

Theorem 1. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ は独立に $\mathrm{EC}_p(\mathbf{0}, \Lambda)$ に従うとし $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ とおく. S を *(2)* により定義される標本共分散行列として $p \times N$ 確率行列 $Y = S^{-1/2}X$ とする. このとき

$$Y \sim \mathrm{SS}_{p\times N}(\phi_Y). \tag{7}$$

さらに, (3) の定数行列 Q_N は $\mathrm{rank}(Q_N) = n$ の直交射影行列であるから, $KK' = Q_N$, $K'K = I_n$ を満足する $N \times n$ (実) 行列 K が存在する. これを利用して Theorem 1 を発展されると, 次の結果を得る.

Theorem 2. *Let* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ は独立に $\mathrm{EC}_p(\boldsymbol{\mu}, \Lambda)$ 従う確率ベクトルとし, $X = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N]$ とおく. さらに, S を *(2)* により定義される標本共分散行列とする. このとき, $n \times p$ 確率行列

$$U = K'Q_NX'(nS)^{-1/2} = K'X'(nS)^{-1/2} \tag{8}$$

は *Stiefel* 多様体 $\mathcal{O}(n, p)$ 上の一様分布に従う. ここに, Q_N は *(3)* で定義された $N \times N$ 行列である.

3. 提案する検定手法と数値実験結果

$\{\boldsymbol{X}_i^{(k)}\}_{i=1}^N$ $(k = 1, 2, \ldots, m)$ を p-次元確率ベクトル $\boldsymbol{X}$ のランダムコピー,

$$X_{(k)} = \left[\boldsymbol{X}_1^{(k)}, \ldots, \boldsymbol{X}_N^{(k)}\right], \qquad S_{(k)} = n^{-1}X_{(k)}Q_NX'_{(k)},$$

として

$$U_k = K'_{(k)}X'_{(k)}(nS_{(k)})^{-1/2}, \qquad n = N - 1 \geq p, \tag{9}$$

とすると, Theorem 2 の結果から, $\boldsymbol{X}_j^{(k)} \sim \mathrm{EC}_p(\boldsymbol{\mu}, \Lambda)$ $(j = 1, \ldots, N; k = 1, \ldots, m)$ ならば U_k は独立に Stiefel 多様体 $\mathcal{O}(n, p)$ 上の一様分布に従うことになるので Pycke (2010) が提案した円周上のの一様分布に関する検定法を応用して, 検定手順を構成した. 提案する検定法に基づいて数値実験を実行した結果, 楕円母集団の下で, 提案した検定手法が有効であるが, 歪楕円母集団の下で, 検出力を有しないことが判明した. また，非正規性の指摘がされている Iris Setosa deta についても，変数選択をすることで帰無仮説を棄却できたが，すべてで帰無仮説を棄却することはできなかった．検定手順を含め, 今後の研究課題と考えている.

謝辞

本研究は JSPS 科研費 18K11198(岩下), 18K03428(橋口，岩下) の助成をうけたものです。

参考文献

Iwashita, T. and Klar, B.(2013). The joint distribution of Studentized residuals under elliptical distributions. *J. Multivariate Anal.*,**128**, 203–209.

J.-R. Pycke. (2010). Some tests for uniformity of circular distributions powerfull against multimodal alternatives, *Can. J. Statist.*, **38**, 80–96.

Community-level social capital, parental psychological distress, and child physical abuse: A multilevel mediation analysis

Nobutoshi Nawa, Aya Isumi, Takeo Fujiwara

Department of Global Health Promotion, Tokyo Medical and Dental University, Tokyo, Japan.

Abstract

Clarifying modifiable risk factors and mediators is crucial for developing a strategy to prevent child maltreatment. The purpose of this study was to investigate the association between community-level social capital and physical abuse towards children, and the mediating effect of parental psychological distress by multilevel mediation analyses. We analyzed data from a population-based study of first-grade elementary school children (6–7 years old) in Adachi Ward, Tokyo, Japan, which was conducted in 2015. The caregivers of first-grade students from all elementary schools in Adachi Ward (N = 5,355) were asked to respond to a questionnaire assessing parents' self-reported physical abuse (hitting and beating) and neighborhood social capital. Among them, 4,291 parents returned valid responses (response rate: 80.1%). We performed multilevel analyses to determine the relationships between community-level parental social capital and physical abuse, and further multilevel mediation analyses were performed to

determine whether parental psychological distress mediated the association. Low community-level social capital was positively associated with physical abuse (both beating and hitting) after adjustment for other individual covariates. Multilevel mediation analyses revealed that community-level parental psychological distress did not mediate the association. Fostering community-level social capital might be important for developing a strategy to prevent child maltreatment, which may have a direct impact on abusive behavior towards children.

ACKNOWLEDGEMENTS

We are particularly grateful to the staff members, central office of Adachi City Hall for conducting the survey and everyone who participated in the surveys. In particular, we would also like to thank Mayor Yayoi Kondo, Mr. Syuichiro Akiu, Mr. Hideaki Otaka, and Ms. Yuko Baba of Adachi City Hall, all of whom contributed significantly to completion of this study. This study was supported by a Health Labour Sciences Research Grant, Comprehensive Research on Lifestyle Disease from the Japanese Ministry of Health, Labour and Welfare (H27-Jyunkankito-ippan-002), and Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS KAKENHI Grant Number 16H03276, 16K21669 and 16J11423), St. Luke's Life Science Institute Grants, and the Japan Health Foundation Grants.

子どもの貧困と虐待における
親の心理的ストレスと個人レベルのソーシャル・キャピタルの媒介効果

Mediating effects of parental psychological distress and individual-level social capital on the association between child poverty and maltreatment in Japan

伊角　彩

東京医科歯科大学　国際健康推進医学分野

1. はじめに

貧困は子ども虐待のリスクと考えられているが、親の心理的ストレスや個人レベルのソーシャル・キャピタル（SC）が貧困と子ども虐待との関連を媒介しているかは検討されていない。そこで本研究では、足立区の小学 1 年生を対象とした悉皆調査を用いて、親の心理的ストレスや SC の媒介効果を明らかにすることを目的とした。

2. 方法

平成 27 年 7 月および 11 月に、足立区立小学校に在籍する小学 1 年生の保護者 5,355 名（有効回答率 80.1%）を対象に実施された質問紙調査「子どもの健康・生活実態調査」のデータを用いて、ロジスティック回帰分析およびポワソン回帰分析を行った。本研究では、①世帯年収 300 万円未満、②生活必需品の非所有、③支払い困難経験のいずれかひとつでも該当する場合、生活困難にある状態と定義した。虐待傾向については、身体的虐待、ネグレクト、心理的虐待を含む 9 種類の虐待の頻度を尋ねた。分析には、生活困難と虐待傾向に欠損のない 3,944 名のデータを用いた。

3. 結果

生活困難と虐待はどの虐待においても関連があることが確認された（表 1）。さらに媒介分析により、親の心理的ストレスは生活困難と身体的虐待、また生活困難と心理的虐待との関連の 60%以上を説明することが明らかになった（表 2）。一方、SC は生活困難といずれかの虐待の 10%しか説明していなかった（表 2）。さらに構造方程式モデリングによって、親の心理的ストレスと SC が同時に媒介していることがわかった（図 1）。

4. 考察

本研究より、親の心理的ストレスが軽減するようにサポートすることが、貧困が子ども虐待に与える影響を緩和させるのに効果的かもしれないことが示唆された。

表 1 ポワソン回帰分析およびロジスティック回帰分析の結果

		Prevalence of	Crude		Model 1		Model 2		Model 3		Model 4	
Outcome: Any Maltreatment		maltreatment (%)	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI
Poverty	No	37.1	Ref		Ref		Ref		Ref		Ref	
	Yes	47.8	**1.29**	**1.15–1.44**	**1.23**	**1.10–1.38**	1.08	0.96–1.22	**1.21**	**1.08–1.35**	1.07	0.95–1.21
Age		N/A	**0.98**	**0.97–0.99**	**0.99**	**0.98–1.00**	**0.99**	**0.98–1.00**	**0.99**	**0.98–1.00**	**0.99**	**0.98–1.00**
Education		N/A	**0.91**	**0.85–0.97**	0.94	0.88–1.01	0.96	0.90–1.03	0.95	0.89–1.02	0.97	0.90–1.04
K6 total score		N/A	**1.05**	**1.04–1.06**			**1.15**	**1.04–1.06**			**1.05**	**1.04–1.06**
SC mean score		N/A	**1.14**	**1.07–1.21**					**1.11**	**1.04–1.18**	1.05	0.98–1.12
Outcome: Physical Abuse			OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Poverty	No	10.8	Ref		Ref		Ref		Ref		Ref	
	Yes	16.8	**1.67**	**1.36–2.05**	**1.43**	**1.16–1.78**	1.10	0.88–1.38	**1.35**	**1.09–1.68**	1.07	0.85–1.35
Age		N/A	**0.95**	**0.93–0.97**	**0.96**	**0.94–0.98**	**0.96**	**0.94–0.98**	**0.96**	**0.94–0.98**	**0.96**	**0.94–0.98**
Education		N/A	**0.74**	**0.64–0.84**	**0.80**	**0.70–0.91**	**0.83**	**0.72–0.95**	**0.82**	**0.72–0.94**	**0.84**	**0.73–0.96**
K6 total score		N/A	**1.10**	**1.09–1.13**			**1.10**	**1.08–1.12**			**1.19**	**1.07–1.12**
SC mean score		N/A	**1.39**	**1.24–1.57**					**1.30**	**1.15–1.47**	**1.18**	**1.04–1.33**
Outcome: Neglect			OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Poverty	No	12.2	Ref		Ref		Ref		Ref		Ref	
	Yes	18.8	**1.67**	**1.38–2.04**	**1.69**	**1.38–2.08**	**1.44**	**1.27–1.79**	**1.62**	**1.32–1.99**	**1.41**	**1.14–1.75**
Age		N/A	0.99	0.97–1.01	1.00	0.98–1.01	1.00	0.98–1.02	1.00	0.98–1.01	1.00	0.98–1.02
Education		N/A	0.97	0.86–1.09	1.04	0.92–1.18	1.07	0.95–1.22	1.07	0.94–1.21	1.09	0.96–1.24
K6 total score		N/A	**1.17**	**1.05–1.09**			**1.07**	**1.05–1.09**			**1.06**	**1.04–1.08**
SC mean score		N/A	**1.30**	**1.16–1.45**					**1.25**	**1.12–1.41**	**1.18**	**1.05–1.32**
Outcome: Psychological Abuse			IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI
Poverty	No	28.8	**Ref**		Ref		Ref		Ref		Ref	
	Yes	38.5	**1.34**	**1.18–1.51**	**1.27**	**1.12–1.44**	1.09	0.95–1.24	**1.24**	**1.09–1.41**	1.08	0.94–1.23
Age		N/A	**0.98**	**0.97–0.99**	**0.98**	**0.97–0.99**	**0.98**	**0.97–0.99**	**0.98**	**0.97–0.99**	**0.98**	**0.97–0.99**
Education		N/A	**0.90**	**0.84–0.97**	0.95	0.88–1.02	0.97	0.90–1.05	0.96	0.89–1.04	0.97	0.90–1.05
K6 total score		N/A	**1.06**	**1.05–1.07**			**1.06**	**1.04–1.07**			**1.05**	**1.04–1.07**
SC mean score		N/A	**1.15**	**1.07–1.24**					**1.12**	**1.04–1.20**	1.05	0.98–1.13

IRR: Incidence Rate Ratios**, OR**: Odds Ratios, **Bold**: p<0.05

表 2 媒介分析結果

Outcome: Any Maltreatment	CDE (IRR)	95% CI	NIE (IRR)	95% CI	TE (IRR)	95% CI	Percent mediated
Psychological distress	1.08	0.99 —1.18	**1.12**	**1.09 —1.15**	**1.21**	**1.11 —1.32**	**59.1**
SC	**1.21**	**1.11 —1.32**	**1.02**	**1.01 —1.03**	**1.23**	**1.13 —1.34**	**9.6**
Outcome: Physical Abuse	CDE (OR)	95% CI	NIE (OR)	95% CI	TE (OR)	95% CI	Percent mediated
Psychological distress	1.10	0.89 —1.39	**1.25**	**1.18 —1.32**	**1.38**	**1.11 —1.73**	**69.9**
SC	**1.35**	**1.10 —1.68**	**1.05**	**1.03 —1.09**	**1.43**	**1.16 —1.77**	**14.8**
Outcome: Neglect	CDE (OR)	95% CI	NIE (OR)	95% CI	TE (OR)	95% CI	Percent mediated
Psychological distress	**1.44**	**1.17 —1.79**	**1.16**	**1.11 —1.22**	**1.67**	**1.35 —2.05**	**28.7**
SC	**1.62**	**1.31 —1.98**	**1.05**	**1.02 —1.08**	**1.69**	**1.37 —2.07**	**8.6**
Outcome: Psychological Abuse	CDE (IRR)	95% CI	NIE (IRR)	95% CI	TE (IRR)	95% CI	Percent mediated
Psychological distress	1.09	0.98 —1.21	**1.14**	**1.11 —1.17**	**1.24**	**1.11 —1.37**	**60.2**
SC	**1.24**	**1.11 —1.38**	**1.02**	**1.01 —1.04**	**1.27**	**1.14 —1.41**	**9.2**

CDE: Controlled Direct Effects, **NIE**: Natural Indirect Effects, **TE**: Total Effects
IRR: Incidence Rate Ratios**, OR**: Odds Ratios, **Bold**: p<0.05

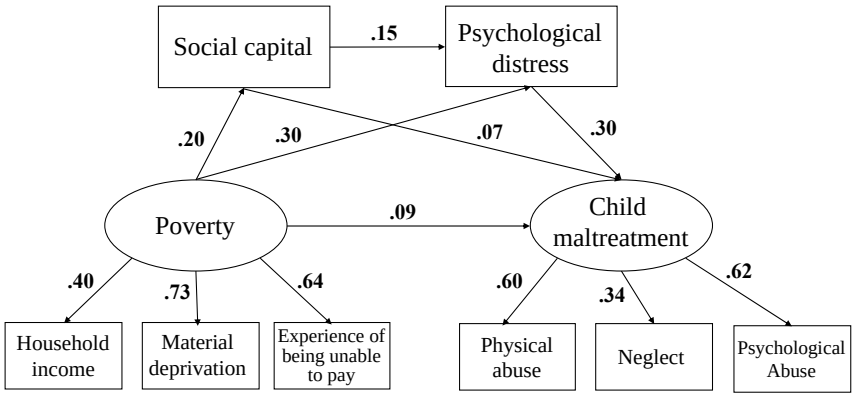

図 1 構造方程式モデリングによる結果

がん登録資料を用いたがん患者の生存時間解析

大阪医科大学　研究支援センター　医療統計室　伊藤　ゆり

はじめに

がん登録はがんの発生状況とその予後を把握し、がん対策に活用するために、わが国では 1950 年代から地域がん登録として宮城県から開始した。地域において発生したがんを全て登録する悉皆調査である。これにより、がんの罹患率（人口 10 万人あたりの年間がん発生率）が把握される。登録されたがん患者情報は人口動態統計および住民基本台帳（住民票）と照合され予後が把握される。これにより、がん患者の生存率が把握可能となる。本報告では、がん登録資料を活用したがん患者の生存時間解析の各手法について、適用事例を交えて紹介する。

Competing risk の扱い : Net survival, excess hazard model

がん患者の生存時間解析におけるイベント発生は当該がん死亡であることが理想とされるが、その正確な把握は困難である。がん登録資料を用いた生存時間解析においては、全死因の死亡を用いることとなるが、当該がん死亡以外の死亡の影響を極力除去したいと考える。そこで、対象集団の期待死亡を生命表に基づいて算出し、それよりも過剰な死亡（excess mortality, excess hazard）を当該がん死のイベントととらえる考え方を適用している（図 1）。それにより推定される生存率を net survival と呼ぶ。また多変量解析では、Poisson regression model を適用した excess hazard model を使用する。

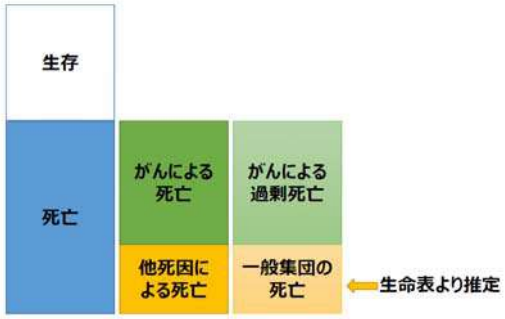

図 1. がん過剰死亡の考え方

長期生存率の計測：Period analysis

がん患者の長期生存率を計測する際、通常の方法（Cohort 法）では、同一期間に診断された患者コホートを追跡し、累積生存率を算出する。その場合、10 年、20 年生存率と長期予後を知りたい場合、それだけ古くさかのぼったがん患者の情報に基づくこととなり、得られた生存率は outdate なものとなる。簡易生命表により平均余命を推定する方法と同じ方法を用いて、長期生存率を推計するのが period analysis である。最新の医療の状況を反映した長期生存率が提供可能となるとして、わが国でも適用が始まった。

新しい予後指標：Conditional survival

通常、がんの診療の場面で参照される生存率は 5 年生存率であり、患者は診断から 5 年間診断時に提示されたがん生存率を参照し続けることとなる。5 年が治癒の目安と為れてきたためである。しかし、がんの長期生存が可能となった今、患者の社会復帰において、診断からの経過年数に応じたその後の予後に関する指標が求められてい

る。上述の period analysis を適用した 10 年生存率を用いて、診断から 1 年以上生存した患者のその後の 5 年生存率（1 年以上生存者の 6 年生存率に該当）、2 年以上・・・と、X 年以上生存した者の X+5 年生存率を Conditional 5-year survival として計測し、患者や医療現場で活用してもらう試みをはじめた（図 2）。

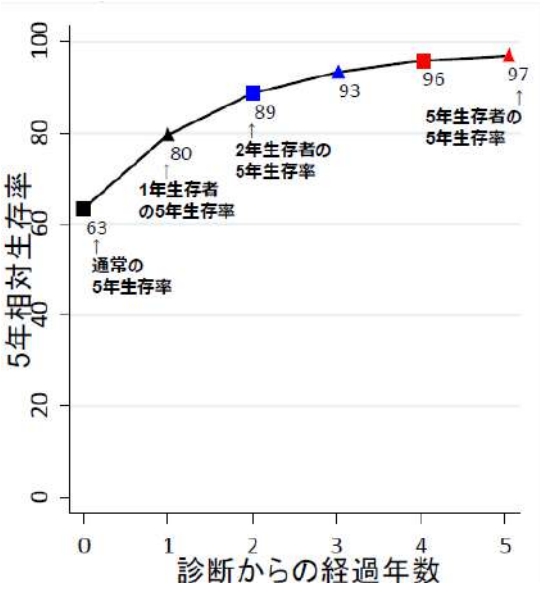

図 2. Conditional 5-year survival

治癒モデル：Cure fraction model

過剰死亡という考え方を用いると、過剰死亡が起こらなくなったときがすなわち一般集団と死亡リスクが同じになった時点と考えられ、統計的な治癒（Population Cure）として推定可能になる。非治癒患者の生存関数にワイブル関数や経験的な分布を当てはめて、治癒割合および非治癒患者の median survival time を推定する。この二つの推定値のトレンドを用いて、がん医療の評価を行った。例えば卵巣がんでは、非治癒患者の生存時間が延長するだけでなく、標準的化学療法がガイドラインにより適用されてから治癒割合が大きく向上した。

予後予測モデル：樹木構造接近法（CART）

生存時間解析における樹木構造接近法は Segal や LeBlanc and Crowley らにより紹介された。これをがん登録資料に適用するために、Competing risk を考慮した生存解析に拡張した（杉本. 2013）。がん専門病院における院内がん登録資料と診療科データベースをリンケージし、非小細胞肺がん患者の予後予測モデルを構築した。診断から 10 年間予後追跡された 1,612 名の症例データを用いて、相対生存に基づく樹木構造接近法を適用したところ、12 の予後の異なるグループに分類された（図 3）。

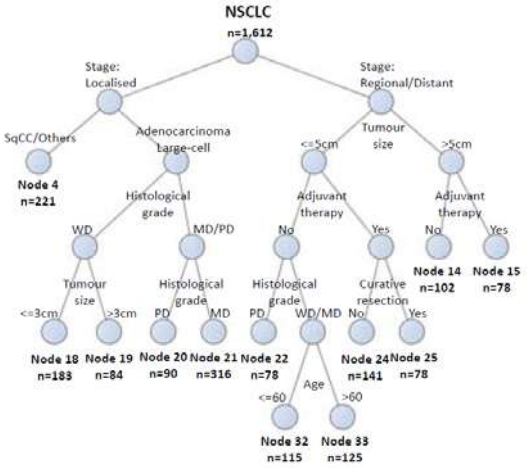

図 3. 非小細胞肺がんの回帰樹木

結びにかえて

がん登録資料を活用した生存解析の結果は、がん患者、医療現場で役立てるために強く求められているが、わが国におけるその適用事例の報告はまだ少ない。また、統計的な側面での発展性も見込まれるため、多くの研究者の参入が期待される。臨床・統計・疫学の研究者が協働してさまざまな課題解決に取り組みながら社会のニーズに合致した統計資料を発信していきたい。

複雑な海洋生態システムにおける因果性推測

ソルヴァン加藤比呂子, Sam Subbey

Marine mammals research group, Institute of Marine Research, Bergen, Norway

近年の地球温暖化に伴う海洋生態システムと海洋生物種の資源量の変化，生物種と環境要因との因果関係を予測することは，学術的・商業的に重要である．バーレンツ海の漁業管理に関する海洋調査では、海洋生物種や環境要因を含めた高次元の時系列データが観測される．それらのデータを元に、International Council for the Exploration of the Sea (ICES)のワーキンググループが統合的な海洋生態システム査定をおこなっている．毎年のレポートにはICESが推奨する多変量解析に基づいた分析結果がまとめられている．それらの手法は時間軸上のダイナミクスを考慮しないために、海洋エコシステムに関する生物種と環境要因の因果関係は探ることはできない．そこで、Ozaki（2012)が提案した，Granger(1969)と Geweke(1982)のペアワイズ因果性推測(Partial pairwise causality)と赤池(1968)の相対パワー寄与率（Total causality）を統合した因果性推測を変量間相互関係を推測する方法として適用した．

提案手法

観測された k 次元時系列データ $\mathbf{x}_t = (x_1(t), x_2(t), \cdots, x_k(t))'$, $t = 1, \cdots, N$ とする（ここで，$(\cdot)'$ は転置の記号）．これらのデータは，以下に示すような多変量自己回帰（Multivariate auto-regressive, MAR）過程の実現値と仮定する：

$$\mathbf{x}_t = \sum_{m=1}^{M} \mathbf{A}_m \mathbf{x}_{t-m} + \boldsymbol{\varepsilon}_t,$$

ここで M は自己回帰の次数，$\mathbf{A}_m$ は自己回帰モデルの係数，そして $\boldsymbol{\varepsilon}_t$ は平均ゼロベクトル，分散共分散行列 Σ に従う多変量正規分布に従うとする．自己回帰係数は最小二乗法や Yule-Walker 法などで推定され、予測誤差系列により共分散行列が計算できる．自己回帰係数とフーリエ変換から，周波数 f に対する周波数応答関数 $\mathbf{F}_f$ が求まり，以下のようなパワースペクトルが求まる：

$$\mathbf{P}_f = \mathbf{F}_f \Sigma \mathbf{F}_f^* = \begin{pmatrix} p_{11f} & p_{12f} \cdots p_{1kf} \\ p_{21f} & p_{22f} \cdots p_{2kf} \\ \vdots & \\ p_{k1f} & p_{k2f} \cdots p_{kkf} \end{pmatrix}, \quad 0 \le f \le 0.5\Delta$$

ここで $\mathbf{F}_f^*$ は $\mathbf{F}_f$ の共役複素数，Δ は観測値のサンプリング間隔とする．$\mathbf{P}_f$ の非対角成分はクロスパワースペクトルである．もし，Σ の非対角成分が非常に小さい，すなわち各変量の予測誤差は独立であることが仮定できると，i 番目の変量 x_i のパワースペクトルは他の変量 x_j からの周波数応答関数 F_{ijf} と予測誤差の分散 σ_{jj}^2 の影響を含む項の和

$$p_{iif} = \left|F_{i1f}\right|^2 \sigma_{11}^2 + \cdots + \left|F_{iif}\right|^2 \sigma_{ii}^2 + \cdots + \left|F_{ikf}\right|^2 \sigma_{kk}^2$$

の形で示すことができ，j 番目の変量 x_j からのノイズの影響 $r_{ij\,f}$ を以下のように示すことができる：

$$r_{ij\,f} = \frac{\left|F_{ij\,f}\right|^2 \sigma_{jj}^2}{\left|p_{ii\,f}\right|} \in [0,1]$$

予測誤差 σ^2 とパワースペクトル $p(f)$ との関係を用いると、このノイズ寄与率は以下のように、全ての変数が関係しているモデルと、全ての変数から 1 つだけ変数を外したモデルの予測誤差の差の形として拡張することができる：

$$\log \sigma_{i\hat{}j}^2 - \log \sigma_{ii}^2 = \int_{-1/2}^{1/2} \log p_{ii}^{(j)}(f)df - \int_{-1/2}^{1/2} \log p_{ii}(f)df$$

$$= \int_{-1/2}^{1/2} \log \frac{p_{ii}^{(j)}(f)}{p_{ii}(f)} df = \int_{-1/2}^{1/2} \log \frac{p_{ii}(f) - \left|\alpha_{ij}(f)\right|\sigma_{jj}^2}{p_{ii}(f)} df$$

$$= \int_{-1/2}^{1/2} \log(1 - \frac{\left|\alpha_{ij}(f)\right|\sigma_{jj}^2}{p_{ii}(f)})df$$

我々はさらに赤池情報量規準に基づく枠組みで，この因果関係の有意性に関する規準を加えた．

シミュレーション実験

3 変量と 5 変量の時系列データを生成し、Granger と Geweke の因果性分析との比較、サンプル数の違いによる Sensitivity、因果性の強度に関して、提案手法のパフォーマンスを検証した．

実データ分析

バーレンツ海の生態システムにおける食物連鎖で重要な 4 種類の海洋生物（シシャモ，タラ，オキアミ，ニシン）のバイオマス時系列データを用い，シシャモの年齢毎に生物種間のフィードバック関係を考察した．2−3 年齢のシシャモを含むフィードバックシステムは，1 年齢もしくは 4 年齢のシシャモを含むフィードバックシステムよりもより多くの生物種間の相互関係を示した．それらはバーレンツ海の食物連鎖に関するこれまでの先行研究を裏付け，シシャモがバーレンツ海域生態システムの食物連鎖に関連する生物種間の重要な駆動源になることを明らかにした．本稿が提案する手法は，海洋学研究で対象となる複雑な生態系において，生物種間，環境因数間の因果関係を推測する一手段として有用であると考えられる．

参考文献

Akaike, H. (1968). On the use of a linear model for the identification of feedback systems, *Annals of the Institute of Stattistical Mathematics*, **20**, 425-439.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series, *Journal of the American Statistical Association*., **77**, 304-313.

Granger, C.W.J. (1969). Investigating causal relations by economietric models and cross-spectral methods, *Econometrika*, **37**, 424-438.

Ozaki, T. (2012). Time series modeling of neuroscience data, Chapman & Hall/CRC Boca Raton.

Summarization Method Based On Data Fusion

Mika Sato-Ilic

Faculty of Engineering, Information and Systems, University of Tsukuba,
Tsukuba, Japan
mika@risk.tsukuba.ac.jp

1. Introduction

The recent growth of data requires a proportional increase in speed of analysis, as with the problem of how to predict a future value when multiple data sets are observed in real time. At this time, how to summarize the multiple data sets with data fusion is important. Therefore, this presentation discusses a data fusion method as a summarization method based on our recent researches [1], [2]. The research focused on how to obtain comparable estimates of predicting values of dependent variables in regression analysis when we observe multiple different data sets simultaneously. These data sets have various forms. For example, if we observe high-dimension and low-sample size (HDLSS) data and the data is classified into several subclasses by using an adaptable classification of the high dimension, then the obtained subclasses are one of the multiple data sets.

The main role of ordinary linear regression analysis is to obtain the estimate of predicted values as projected values in a linear subspace spanned by vectors of independent variables. However, if the data set has been observed simultaneously from multiple different data sources, then we must create different linear subspaces to estimate the different predicted values corresponding to the different datasets. So, we cannot compare the different predicted values, since the linear subspaces are different.

In order to solve this problem, we have proposed a method to obtain comparable predicted values obtained from different datasets by utilizing a fuzzy clustering result and an orthogonal projector which projects two different vectors corresponded with the two different dependent variables to the same intersection of the two different linear subspaces. [1], [2] From this, since the different predicted values from different data sources can be obtained in the common space, we can compare the different predicted values.

2. Comparable Predicted Values in Regression Analysis

Two $\mathrm{n} \times \mathrm{p}$ matrices A_1 and A_2 which consist of independent p different variables are shown as: $A_t = \left(\boldsymbol{a}_1^{(t)}, \cdots, \boldsymbol{a}_p^{(t)}\right)$, $\boldsymbol{a}_r^{(t)} = \left(a_{1r}^{(t)}, \cdots, a_{nr}^{(t)}\right)'$, $r = 1, \cdots, p$, $t = 1,2$. $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are vectors of dependent variables, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of regression coefficients, and $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are error vectors whose elements follow the same normal distribution. We utilize a result of fuzzy clustering for 3-way data [3] as a common scale of A_1 and A_2, since in this method, the same K clusters are

obtained through A_1 and A_2 when $\mathrm{T} = 2$. Then, we obtain the following result of the fuzzy clustering: $U_t = (\boldsymbol{u}_1^{(t)}, \cdots, \boldsymbol{u}_K^{(t)})$, $\boldsymbol{u}_k^{(t)} = (u_{1k}^{(t)}, \cdots, u_{nk}^{(t)})'$, $k = 1, \cdots, K$, $t = 1,2$. We define the following criterion: $c_l^{(t)} \equiv \sum_{k=1}^{K} abs(cor(\boldsymbol{u}_k^{(t)}, \tilde{\boldsymbol{a}}_l))$, $l = 1, \cdots, 2p$, $t = 1,2$, where $\mathrm{cor}(\mathbf{a}, \mathbf{b})$ means the correlation between $\mathbf{a}$ and $\mathbf{b}$, $\mathrm{abs}(*)$ means an absolute value of $*$, and $\{\boldsymbol{a}_1^{(1)}, \cdots, \boldsymbol{a}_p^{(1)}, \boldsymbol{a}_1^{(2)}, \cdots, \boldsymbol{a}_p^{(2)}\} \equiv \{\tilde{\boldsymbol{a}}_1, \cdots, \tilde{\boldsymbol{a}}_p, \tilde{\boldsymbol{a}}_{p+1}, \cdots, \tilde{\boldsymbol{a}}_{2p}\}$. Then $c_l^{(t)}$ shows a degree of contribution of $\tilde{\boldsymbol{a}}_l$ to t-th data $A_t, t = 1,2$. Therefore, we can recreate the independent vectors for each t as follows: $\tilde{A}_1 = (\tilde{\boldsymbol{a}}_1^{(1)}, \cdots, \tilde{\boldsymbol{a}}_s^{(1)})$, $\{\tilde{\boldsymbol{a}}_j^{(1)} | c_j^{(1)} > \varepsilon\}$, $j = 1, \cdots, s$, $\tilde{A}_2 = (\tilde{\boldsymbol{a}}_1^{(2)}, \cdots, \tilde{\boldsymbol{a}}_w^{(2)})$, $\{\tilde{\boldsymbol{a}}_j^{(2)} | c_j^{(2)} > \varepsilon\}$, $j = 1, \cdots, w$, $s + w = 2p$, where ε is a given positive threshold value which is selected as at least one same vector is included for both $\tilde{A}_1$ and $\tilde{A}_2$. Then we can calculate orthogonal projectors $\tilde{P}_1$and $\tilde{P}_2$ using $\tilde{A}_1$ and $\tilde{A}_2$ as: $\tilde{P}_t \equiv \tilde{A}_t(\tilde{A}_t{}'\tilde{A}_t)^{-1}\tilde{A}_t{}'$, $t = 1,2$, which project a vector to two different linear subspaces, $\tilde{V}_1$ and $\tilde{V}_2$. In order to obtain a common subspace of $\tilde{V}_1$ and $\tilde{V}_2$, we utilize an orthogonal projector, $\tilde{P}_{1\cap 2}$, which projects a vector to the intersection of subspaces $\tilde{V}_1$ and $\tilde{V}_2$, denoted as $\tilde{V}_{1\cap 2}$. Then $\tilde{P}_{1\cap 2}$ is obtained as follows by using $\tilde{P}_1$ and $\tilde{P}_2$ [4], [5], [6].

$$\tilde{P}_{1\cap 2} = 2\tilde{P}_1(\tilde{P}_1 + \tilde{P}_2)^{-}\tilde{P}_2 = 2\tilde{P}_2(\tilde{P}_1 + \tilde{P}_2)^{-}\tilde{P}_1,$$

where $(*)^{-}$shows the Moore-Penrose generalized inverse matrix of $*$. We obtain the following equations:

$$\tilde{\boldsymbol{b}}_t = \tilde{P}_{1\cap 2}\boldsymbol{b}_t, \quad t = 1,2.$$

Since the estimates $\tilde{\boldsymbol{b}}_1$ and $\tilde{\boldsymbol{b}}_2$ are obtained in the same common linear subspace, $\tilde{V}_{1\cap 2}$, we can compare $\tilde{\boldsymbol{b}}_1$ and $\tilde{\boldsymbol{b}}_2$ mathematically.

References

[1] Sato-Ilic, M., Knowledge-based Comparable Predicted Values in Regression Analysis, In C. H. Dagli (ed) Procedia Computer Science, Elsevier, 114, 216-223 (2017)

[2] Sato-Ilic, M., Cluster-Scaled Regression Analysis for High-Dimension and Low-Sample Size Data, Advances in Smart Systems Research, 7, 1, 1-10 (2018)

[3] Sato-Ilic, M., Individual Compositional Cluster Analysis, In C. H. Dagli (ed) Procedia Computer Science, Elsevier, 95, 254-263 (2016)

[4] Anderson, W. N., and R. J. Duffin., Series and Parallel Addition of Matrices, Journal of Mathematical Analysis and Applications, 26, 576-594 (1969)

[5] Ben-Israel, A., and T. N. E. Greville, Generalized Inverses, Theory and Applications, 2nd ed., Springer (2003)

[6] Krafft, O., An Arithmetic-Harmonic-Mean Inequality for Nonnegative Matrices, Linear Algebra and Its Applications, 268, 243-246 (1998)

Plug-in optimization method for generalized ridge regression for MLE in GMANOVA model

中京大学 国際教養学部 永井 勇

各個体に対して経時的に測定することで得られるデータは経時測定データと呼ばれ, 様々な分野で収集・分析がされている. この経時測定データの分析においては, データに潜む経時的な変動 (経時変動) を上手く捉えることが一つの目的である. 本研究では, 全ての個体で測定時点が揃っている, つまり, n 個全ての個体に対して時点 $t_1, \ldots, t_p$ で測定した経時測定データを考えた. このとき, Potthoff and Roy (1964) により提案された次の GMANOVA (一般化多変量分散分析) モデルを用いることで経時変動が推定できることが知られている;

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

ここで, $\boldsymbol{Y}$ は各行が各個体の経時測定データからなる $n \times p$ 行列, $\boldsymbol{1}_n$ は全ての成分が 1 の n 次元ベクトル, $\boldsymbol{\mu}$ は q 次元未知ベクトル, $\boldsymbol{X}$ は $p \times q$ 既知行列 (詳しくは後述), $\boldsymbol{A}$ は各行が各個体の性別などからなる $n \times k$ 既知 (説明変数) 行列, $\boldsymbol{\Xi}$ は $p \times q$ 未知行列, $\boldsymbol{\mathcal{E}}$ は $n \times p$ 誤差行列である. 本研究では, $\mathrm{rank}(\boldsymbol{X}) = q$, $\boldsymbol{A}$ は $\mathrm{rank}(\boldsymbol{A}) = k$ で中心化されている (つまり $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$, $\boldsymbol{0}_k$ は k 次元ゼロベクトル) とし, $\boldsymbol{\mathcal{E}}$ の各行が独立に $N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ に従う確率変数ベクトル, $\boldsymbol{\Sigma}$ は $p \times p$ 未知正定値行列, $n-k-p+q-2 > 0$ とした. ここで, $(\boldsymbol{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}')$ の部分が $\boldsymbol{Y}$ の経時変動を表しており, 例えば $\boldsymbol{X}$ の j 行目を $(1, t_j, \ldots, t_j^{q-1})$ とすることは, 経時変動を測定時点 $t_1, \ldots, t_p$ の $(q-1)$ 次の多項式で推定することに対応している.

このモデルにおいて, $\boldsymbol{\mu}$ や $\boldsymbol{\Xi}$ の最尤推定量はそれぞれ以下となる;

$$\begin{cases} \hat{\boldsymbol{\mu}} = (\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{Y}'\boldsymbol{1}_n/n, \\ \hat{\boldsymbol{\Xi}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X})^{-1}, \end{cases}$$

ここで $\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \boldsymbol{1}_n\boldsymbol{1}_n'/n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}\boldsymbol{Y}/(n-k-1)$ である. このとき, $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$ と $\boldsymbol{\mathcal{E}}$ の各行への正規分布などの仮定より, $(n-k-1)\boldsymbol{S} \sim W_p(n-k-1, \boldsymbol{\Sigma})$ (自由度が $(n-k-1)$ で尺度行列が $\boldsymbol{\Sigma}$ の p 次元ウィッシャート分布, Schott (2017) Sec. 11.7 など参照) である.

この推定量において, $\boldsymbol{A}'\boldsymbol{A}$ の固有値 $d_1, \ldots, d_k$ の中に小さな値が現れると $\hat{\boldsymbol{\Xi}}$ が不安定になってしまう. そこで, Yanagihara, Nagai and Satoh (2009), Nagai (2011) などで用いられている多変量一般化リッジ回帰による推定法を用いた次の推定量を考えた;

$$\hat{\boldsymbol{\Xi}}(\boldsymbol{\theta}) = \boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{X}\big(\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}\big)^{-1},$$

ここで, $\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{A}'\boldsymbol{A} + \boldsymbol{Q}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{Q}'$, $\boldsymbol{Q}$ は $\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{Q} = \mathrm{diag}(d_1, \ldots, d_k)$ となる直交行列, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ $(\theta_i \geq 0)$ はリッジパラメータと呼ばれるパラメータである.

次にこの $\boldsymbol{\theta}$ の最適化を考えた. そのために, $\hat{\boldsymbol{\Xi}}(\boldsymbol{\theta})$ を用いた予測値 $\hat{\boldsymbol{Y}}(\boldsymbol{\theta}) = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}'\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}(\boldsymbol{\theta})\boldsymbol{X}'$ を評価する関数として, 次で定義される予測平均二乗誤差 (PMSE) を用いた;

$$\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})] = E_{\boldsymbol{U}}\left[E_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\Big(\boldsymbol{U} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})\Big)\boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{U} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})\Big)'\right\}\right]\right],$$

ここで $\boldsymbol{U}$ は $\boldsymbol{Y}$ と独立で同一の分布に従う確率変数, $E_{\boldsymbol{B}}[\cdot]$ は確率変数 $\boldsymbol{B}$ に関する期待値を表している. この $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})]$ は, 今のデータ $\boldsymbol{Y}$ から構築した予測値 $\hat{\boldsymbol{Y}}(\boldsymbol{\theta})$ と新たに $\boldsymbol{Y}$ と同じモデルから得られたデータ $\boldsymbol{U}$ との平均残差を評価している関数と見ることもできる.

本研究では, この PMSE$[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})]$ を最小にする最適なリッジパラメータ $\boldsymbol{\theta}$ を求めることを考えた. このとき, $\boldsymbol{U}$ が $\boldsymbol{Y}$ と独立に同一分布に従う確率変数より PMSE$[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})] = E_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}))'\}] + 2q\{1 + \mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M}_{\boldsymbol{0}_k})\}$ なので, 第一項の期待値を求めて $\boldsymbol{\theta}$ に関する項を最小にすることで最適な $\boldsymbol{\theta}$ が得られる. しかしながら, そのまま最適な $\boldsymbol{\theta}$ を求めても未知の行列である $\boldsymbol{\Xi}$ と $\boldsymbol{\Sigma}$ が残るため, 以下の二種類の手法が考えられる;

① 既知の行列からなる $E_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}))'\}]$ の推定値を作り, その最小化を行う (Yanagihara, Nagai & Satoh (2009), Yanagihara & Satoh (2010) など参照)

② PMSE$[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})]$ を最小にする $\boldsymbol{\theta}$ を求め, 未知の行列にそれぞれの推定量を代入する (Nagai, Yanagihara & Satoh (2012) など参照)

本研究では $\hat{\boldsymbol{\Xi}}(\boldsymbol{\theta})$ において②の手法を提案した. このとき, $\hat{\boldsymbol{Y}}(\boldsymbol{\theta})$ に $\boldsymbol{S}^{-1}$ が必要な点や $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$ などに注意して, 分割行列の逆行列などの行列自体の性質 (Lütkepohl (1996), Koll & von Rosen (2005) など参照), Wishart 分布や二次形式の独立性の性質など (Muirhead (1984), Siotani, Hayakawa & Fujikoshi (1985), Fujikoshi, Ulyanov & Shimizu (2010), Schott (2017) など参照) を用いて, Satoh, Kobayashi and Fujikoshi (1997) と同様に期待値を計算した.

その結果, リッジパラメータ $\boldsymbol{\theta}$ に関する項に着目すると, 適当な関数 $f_i(\theta_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ を用い,

$$\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})] = [\boldsymbol{\theta} \text{ に依存しない項}] + \sum_{i=1}^{k} f_i(\theta_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma}),$$

という形で展開できた. よって, PMSE$[\hat{\boldsymbol{Y}}(\boldsymbol{\theta})]$ を最小にする最適な $\boldsymbol{\theta}$ (各成分は非負) は, 各成分ごとに非負の範囲で $f_i(\theta_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ を最小にする θ_i を求めればよい. つまり $\tilde{\theta}_i(\boldsymbol{\Xi}, \boldsymbol{\Sigma}) = \arg\min_{\theta_i \geq 0} f_i(\theta_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ を各 i で求めて, それを並べれば最適な $\boldsymbol{\theta}$ が得られる. このとき, $\tilde{\theta}_i(\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ に未知の行列 $\boldsymbol{\Xi}$ と $\boldsymbol{\Sigma}$ が残ってしまうため, 推定量 $\hat{\boldsymbol{\Xi}}$ と $\boldsymbol{S}$ をそれぞれに代入した.

$f_i(\theta_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ や $\tilde{\theta}_i(\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ の具体的な形などや数値実験による比較などは当日報告した.

引用文献:

[1] Fujikoshi, Y., Ulyanov, V. V. & Shimizu, R. (2010). *Multivariate Statistics*, John Willey & Sons.

[2] Kollo, T. & von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*, Springer.

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

[4] Muirhead, R. J. (1984). *Aspects of Multivariate Statistical Theory*, John Willey & Sons.

[5] Nagai, I. (2011). Modified C_p criterion for optimizing ridge and smooth parameters in the MGR estimator for the nonparametric GMANOVA model. *Open J. Stat.*, **1**, 1–14.

[6] Nagai, I., Yanagihara, H. & Satoh, K. (2012). Optimization of Ridge Parameters in Multivariate Generalized Ridge Regression by Plug-in Methods. *Hiroshima Math. J.*, **42**, 301–324.

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

[8] Satoh, K., Kobayashi, M. & Fujikoshi, Y. (1997). Variable selection for the growth curve model. *J. Multivariate Anal.*, **60**, 277–292.

[9] Schott, J. R. (2017). *Matrix Analysis for Statistics* (Third ed.), John willer & Sons.

[10] Siotani, M., Hayakawa, T. & Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press.

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

[12] Yanagihara, H. & Satoh, K. (2010). An unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.

歪曲度のノンパラメトリック推定

千葉大学大学院理学研究院　内藤貫太

はじめに: $\mathbb{R}^d$ におけるサイズ N の対応のある 2 つの点群 $\boldsymbol{X} = [X_1 \cdots X_N]^T$ と $\boldsymbol{Y} = [Y_1 \cdots Y_N]^T$ を考える．ここで，X_i と Y_i は d 次元確率ベクトルである．特にその組 (X_i, Y_i) は i 番目のオブジェクトに関する (X, Y) の観測値である．

ここで議論する問題は，この $\mathbb{R}^d$ の 2 つの変数 X と Y の類似性，“調和度”，をどのように捉え，そしてデータである点群 $\boldsymbol{X}$ と $\boldsymbol{Y}$ に基づきどのように推定するかということである．

具体例としては例えば $d = 2$ として，ヒトの (右大腿長,右上腕長)$(= X)$ と同一人の (左大腿長,左上腕長)$(= Y)$ が挙げられる．この場合，腕と足の長さの左右での調和をどのように捉えるのかを考えることになる．

本講演では，この “調和度” として，擬等角写像論で現れる歪曲度を用いることを考える．そして歪曲度のノンパラメトリック核型推定について議論する．

歪曲度: この 2 つの変数を対応づける変換 f の存在を仮定する．変数 X と Y の調和度はこの変換 f の性質を通して議論しよう．特に，擬等角写像論において等角写像からの乖離を測る尺度である歪曲度と呼ばれる指標をもって調和度とすることを提案する．

$x = [x_1 \cdots x_d]^T \in \mathbb{R}^d$ に対しそのノルムを $|x| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^d x_i^2}$ とする．$d \times d$ 行列 A の作用素ノルムは

$$||A|| = \sup\{|Ay| : y \in \mathbb{R}^d, |y| = 1\}.$$

で定義される．A の作用素ノルムは $A^T A$ の最大固有値の平方根に他ならない．$\mathbb{R}^d$ から $\mathbb{R}^d$ の関数

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_d(x) \end{bmatrix}, \ x \in \mathbb{R}^d,$$

を考える．ここで，g_i は $\mathbb{R}^d$ から $\mathbb{R}$ への滑らかな関数である．g の x におけるヤコビ行列は

$$Dg(x) = \left(\frac{\partial g_i(x)}{\partial x_j} \right)_{1 \le i,j \le d}$$

で定義され，この行列式の絶対値はヤコビアンと呼ばれ $J_g(x) = |\det(Dg(x))|$ で表す．

いま，Δ を $\mathbb{R}^d$ の領域とし，f を Δ から $\mathbb{R}^d$ への同相写像とする．また，

$$\kappa(x) = \kappa(x|f) = \frac{||Df(x)||^d}{J_f(x)}, \tag{1}$$

とする．f が K-擬等角とは， Δ 上で $\kappa(x|f) \leq K$ となるような $K > 0$ が存在することである（[1]）. 変換 f の歪曲度 $\mathfrak{d} = \mathfrak{d}(f)$ は

$$\mathfrak{d} = \mathfrak{d}(f) = \max_{x \in \Delta} \kappa(x|f). \tag{2}$$

と定義される．Δ 上で常に $\kappa(x|f) \geq 1$ であり，$\kappa(x|f) \equiv 1$，すなわち 1-擬等角は等角写像であることと同値である．歪曲度で調和を測るためには，$\kappa(x)$ の推定が必要である．この $\kappa(x)$ が各 x においての f の等角写像からの乖離を測っていることになる．

素朴な考えでは，適当な平滑化で得られる f のノンパラメトリック推定量 $\hat{f}$ を用いて $\widehat{\mathfrak{d}} = \mathfrak{d}(\hat{f})$ とすることで，歪曲度のノンパラメトリック推定量が構成される．しかしながらこれはあまり筋がよろしくなく，特に $\hat{f}$ の微分を要するので，挙動の不安定性が知られている．

したがって，$\mathfrak{d}$ に現れる $\kappa(x)$，これを構成している f のヤコビ行列 $Df(x)$ それ自体をノンパラメトリックに推定することを考える．

ノンパラメトリック推定: 多変量回帰モデル

$$Y_i = f(X_i) + \boldsymbol{\varepsilon}_i \ , \quad i = 1, ..., N, \tag{3}$$

を想定する．ここで，$\boldsymbol{\varepsilon}_i (i = 1, ..., N)$ は i.i.d. で $E[\boldsymbol{\varepsilon}_i] = 0$，$V[\boldsymbol{\varepsilon}_i] = \Sigma > O$ を満たす．

このモデルにおいて，局所 2 次回帰（[2]）を用いることにより，$Df(x)$ のノンパラメトリック核型推定量が構成される．$\kappa(x|f)$ はその定義から，

$$\kappa(x) = \prod_{i=1}^{d} \sqrt{\frac{\lambda_1(x)}{\lambda_i(x)}}, \tag{4}$$

と表現される．ここで，$\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_d(x) > 0$ は $Df(x)Df(x)^T$ の固有値である.

したがって，$Df(x)$ のノンパラメトリック推定量から $Df(x)Df(x)^T$ の推定量が構成されるから，その行列の固有値を用いることで $\kappa(x|f)$ のノンパラメトリック推定量が構成される．

研究集会においては，このようにして得られる歪曲度のノンパラメトリック推定量の漸近挙動と数値的挙動について報告を行った．

参考文献

[1] Väisälä J. *Lectures on n-Dimensional Quasiconformal Mappings. Lecture Note in Mathematics* 229: Springer,1971.

[2] Ruppert D and Wand MP. Multivariate locally weighted least squares regression. *Ann. Statist.*, 1994; 22: 1346-1370.

探索的財務ビッグデータ解析

地道 正行 *, 宮本 大輔 **, 阪 智香 *, 永田 修一 *
* 関西学院大学 商学部
** 東京大学 大学院情報理工学系研究科

2018 年 12 月 15 日
科学研究費シンポジウム「多変量解析法における理論と応用」
広島大学理学部 B707

概要

本研究では, Bureau van Dijk (BvD) 社 [1] から提供されるデータベース Osiris から抽出された世界 157 カ国の全上場企業 (一般事業会社, 上場廃止企業含む) の主要財務情報 (売上高, 営業利益, 総資産など 84 項目, 33 年分) の財務データ (粗データ) を扱った.

このデータが収められたファイルを, コンピュータ (ソフトウェア) で利用できる形式に変換する工程を「前処理」 (preprocessing) と呼び, その整形されたファイルを R に読み込み, 実際にデータ解析が行えるオブジェクト形式に変換する工程を, データ解析環境 R の統合開発環境である RStudio 上で行うため, Grolemund and Wickham (2016) にならって「データラングリング」(data wrangling) と呼んでいる. 前処理には, Unix コマンドとインタプリタを用いて, データ解析環境 R に読み込める形式に変換し, R を用いてデータの不備などの修正を行いデータを分析・解析できるファイル形式 (CSV, RDS ファイル) として出力することによって行った. また, データラングリングは, 近年注目されているクラスター・コンピューティング・システム Spark と SparkR パッケージを用いて分析できるオブジェクト形式に変換した. さらに, Tukey (1977) によって提唱された探索的データ解析 (Exploratory Data Analysis: EDA) に基づき, 可視化 (data visualization) によって得られた知見を利用して, 統計モデリング (statistical modeling) を行った.

次に, このデータに対して探索的データ解析を行った. 具体的には, 2015 年に時点を固定し, データ可視化 (data visualization) を行った結果として, 売上高, 従業員数, 総資産のそれぞの (1 変量) 分布は極度に右に歪んだものであることがわかり, 各ペアの 2 変量分布も同様の結果となった. この結果に対して, 対数をとることによって歪みを修正したところ, 若干左に歪んでいることが分かった. この知見を利用して, Azzalini (1985), Azzalini and Capitanio (2014) によって提案された非対称分布 (非対称正規分, 非対称ティー分布) を売上高の対数に当てはめたところ, 非対称ティー分布の当てはまりが良いことがわかった. この結果から, 売上高を従業員数と総資産によって説明するためにコブ・ダグラス型生産関数 (Cobb and Douglas (1928)) を応用し, 両辺の対数をとった, いわゆる両対数モデル (double-log model) を利用して統計モデリング (statistical modeling) を行ったところ, 誤差項に非対称ティー分布を仮定したものが回帰診断の結果として最も良いことが分かった. さらに, 赤池情報量規準 (Akaike Information Criterion: AIC) を用いてモデル選択 (model selection) を行い, 交差確認法によって評価 (model evaluation) を行った結果も, 上記の結果が肯定されるものとなった.

本研究のデータ (ファイル) を処理する工程は Unix のシェルスクリプトと R スクリプトで一元管理し, 探索的データ解析の過程も意味のある結果を文書化する工程は, Sweave を利用して LaTeX ファイルに R コードを埋め込む形式で動的に生成した. さらに, これらの全工程を `Makefile` にスクリプトを記述し, Unix の `make` コマンドにより

[1] `https://www.bvdinfo.com/en-gb/`

自動実行することによって, 再現可能研究 (reproducible research) を行うことを試みた.

参考文献

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: pp. 267–281.

[2] Azzalini, A. (1985) A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, Vol. 12, No. 2, pp. 171–178.

[3] Azzalini, A. with the collaboration of A. Capitanio (2014) *The Skew-Normal and Related Families*, Cambridge University Press, Institute of Mathematical Statistics Monographs.

[4] Cobb, C. W. and P. H. Douglas (1928) A theory of production, *American Economic Review*, Vol. 18, pp. 139–165.

[5] Efron, B. and T. Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press.

[6] Gandrud, C. (2015) *Reproducible Research with R and RStudio*, Second Edition, CRC Press.

[7] James, G., D. Witten, T. Hastie, and R. Tibshirani (2013) *An Introduction to Statistical Learning with Applications in R*, Springer.

[8] 地道正行 (2017-a)『R による対数非対称正規線形モデルによる財務データの統計モデリング』, 商学論究, 第 64 巻, 第 5 号, pp. 159–185, 2017 年 3 月, 関西学院大学商学研究会.

[9] 地道正行 (2017-b)『R を利用した対数非対称分布族にもとづく財務データの統計モデリング』, 経済学論究, 第 71 巻, 第 2 号, pp. 141–174, 2017 年 9 月, 関西学院大学経済学部研究会.

[10] 地道正行 (2018-a)『探索的財務ビッグデータ解析 –前処理, データラングリング, 再現可能性–』, 商学論究, 第 66 巻, 第 1 号, pp. 1–32, 2018 年 9 月, 関西学院大学商学研究会.

[11] 地道正行 (2018-b)『探索的財務ビッグデータ解析 –データ可視化, 統計モデリング, モデル選択, モデル評価, 動的文書生成, 再現可能研究–』, 商学論究, 第 66 巻, 第 2 号, pp. 1–41, 2018 年 12 月, 関西学院大学商学研究会.

[12] Jimichi, M., Miyamoto, D., Saka, C. and Nagata, S. (2018) *Visualization and Statistical Modeling of Financial Big Data: Log-Linear Modeling with Skew Error*, *Japanese Journal of Statistics and Data Science*, Vol. 1, No. 2, pp. 347–371, `https://doi.org/10.1007/s42081-018-0019-1`

[13] 地道正行, 豊原法彦 (2018) 『景気先行指数の動的文書生成にもとづく再現可能研究』, 豊原法彦編著『関西経済の構造分析』, 第 5 章, pp. 77-111, 中央経済社.

[14] Konishi, S. and G. Kitagawa (2008) *Information Criteria and Statistical Modeling*, Springer.

[15] Leisch, F. (2002) *Sweave: Dynamic generation of statistical reports using literate data analysis*, In Wolfgang Härdle and Bernd Rönz, editors, Compstat 2002 - Proceedings in Computational Statistics, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

[16] Ryza, S., U. Laserson, S. Owen, and J. Wills (2016) *Advanced Analytics with Spark*, O'Reilly.
(玉川 竜司訳 (2016)『Spark による実践データ解析』, オライリー・ジャパン.)

[17] Saka, C. and M. Jimichi (2017) Evidence of inequality from accounting data visualisation, *Taiwan Accounting Review*, Vol. 13, No. 2, pp. 193–234.

[18] 下田倫大, 師岡一成, 今井雄太, 石川 有, 田中裕一, 小宮篤史, 加嵜長門 (2016)『詳解 Apache Spark』, 技術評論社.

[19] Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Co.

[20] Unwin, A. (2015). *Graphical Data Analysis with R*, Chapman and Hall/CRC.

[21] Xie, Y. (2015) *Dynamic Documents with R and knitr, Second Edition*, CRC Press.

謝辞

本研究の一部は以下の研究費より助成を得ていることに感謝の意を述べたい:

- 科学研究費 基盤研究 C:「グラフィカル・データ・アナリシスによる格差研究と社会環境会計による解決方法の提案」(2016 年 ~2018 年), 課題番号: 16K04022, 研究代表者: 阪 智香
- 平成 30 年度 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 課題: 「財務ビッグデータの可視化と統計モデリング」, 課題番号: jh181001-NWJ, 研究代表者: 地道 正行
- 関西学院大学 図書館 図書費 B, 個人研究費

また, BvD 社の増田 歩氏にはデータの抽出に関して多大なるご協力いただいた. ここに感謝の意を述べる.

セミパラメトリック混合効果モデルの適用による繰り返し観測データの解析

大瀧　慈、大谷敬子（広島大学、(株) deCult

１．はじめに

必ずしも直線的でない経時的トレンドが想定される個体差を伴う繰り返し観測データに対して、汎用性に富み扱い易く解析結果の解釈が容易と思われるセミパラメトリック混合効果モデルおよびその適用による解析法について、実臨床データを対象にした R の関数による分析を例示しながら紹介する。

[脳梗塞患者の回復状況に関する ADC データ]

脳梗塞の患者さんの「病変」と「正常部」の ADC という細胞密度を表す（MRI により計測される）数値の比（rADC)に関心を持っています。(表１） これまでの経験的知見として、rADC は、脳梗塞急性期（発症 1-2 週）は 1.0 未満となり、亜急性期から慢性期（発症 3-6 週）にかけて 1.0 以上に上昇する傾向を把握しています。(図１）これは、細胞が虚血により膨化して、あたかも密度が上昇したような状態となり、その後細胞が死んで脱落することにより、密度が低下することに起因します。ここで、脳梗塞の原因別（心臓が原因の心原性、血管が原因の動脈硬化性など）で、この密度の変化の速度に差があるか否か、検討してみたい。（これは、かつて広島大学大学病院脳神経外科からの実際に受けた検討依頼状です。）

表１．脳梗塞の患者さんの ADC データ（出典:広島大学大学病院脳神経外科）

sid	pid	age	sex	day	type	tADC	nADC	rADC
1	1	59	2	0	1	440	693	0.635
2	1	59	2	0	1	497	897	0.554
3	1	59	2	6	1	420	820	0.512
4	1	59	2	0	1	580	827	0.701
5	1	59	2	6	1	570	810	0.704
6	1	59	2	0	1	850	1050	0.810
7	1	59	2	6	1	1000	1200	0.833
14	3	65	1	0	2	674	876	0.769
15	3	65	1	7	2	730	710	1.028
16	4	65	1	0	2	270	602	0.449
17	4	65	1	7	2	540	720	0.750
中略								
138	64	74	2	1	2	540	870	0.621
139	64	74	2	17	2	940	1100	0.855
140	67	81	1	0	1	308	770	0.400
141	67	81	1	0	1	359	613	0.586
142	67	81	1	7	1	612	644	0.950

Legend

sid	sample identification code
pid	patient identification code
sex	1: male, 2: female
age	age at the 1st obervation
day	time(day) from onset
type	stroke type
	1: embolization
	2: lacunar
	3: atheromatous
tADC	ADC at lesion site
nADC	ADC at normal site
rADC	ratio of tADC to nADC

２．モデル

ADC データについて、患者 (i) の病型 (type_i) を 2 個のダミー変数（$z_{1i}:\text{type}_i=2$ の場合 1、その他の場合 0、と $z_{2i}:\text{type}_i=3$ の場合１、その他の場合 0）で表現し、性別 (sex_i) と発症時年齢 $\left(\text{age}_i\right)$ を、それぞれ、$\text{s}_i=1.5-\text{sex}_i$ （男性で 0.5、女性で-0.5）および、$a_i=\left(\text{age}_i-70\right)/10$ と変換し、その第 j 番目の計測値の測定日 (d_{ij}) を時間変数とし、rADC の常用対数値 (y_{ij}) を目的変数とし、下記のような繰り返し測定値に関するパラレルプロファイルモデルを想定した。

$$y_{ij}=(1,z_{1i},z_{2i})\left\{\Theta_0+\sum_{\text{l}=1}^{3}\left(d_{ij}-c_{\text{l}}\right)\cdot H\left(d_{ij}-c_{\text{l}}\right)\Theta_{\text{l}}\right\}\left(1,s_i,a_i,s_i\cdot a_i\right)^{'}\ +\ \eta_i\ +\ \varepsilon_{ij},\quad (1)$$

ただし、$\Theta_{\text{l}}=\begin{pmatrix}\beta_{\text{l}}^{(0)} & \gamma_{\text{l}}^{(0)} & \delta_{\text{l}}^{(0)} & \lambda_{\text{l}}^{(0)}\\ \beta_{\text{l}}^{(1)} & \gamma_{\text{l}}^{(1)} & \delta_{\text{l}}^{(1)} & \lambda_{\text{l}}^{(1)}\\ \beta_{\text{l}}^{(2)} & \gamma_{\text{l}}^{(2)} & \delta_{\text{l}}^{(2)} & \lambda_{\text{l}}^{(2)}\end{pmatrix},\ \text{l}=0,1,2,$ は、（固定効果）未知母数行列であり、$\left(c_0,c_1,c_2\right)=\left(0,7,14\right)$ である。 また、$H(t)=1\ \left(t\ge 0\right),=0\ \left(t<0\right)$ は Heaviside 関数であり、$\varepsilon_{ij}\sim N(0,\sigma^2),\ j=1,...,m_i,\ \eta_i\sim N(0,\tau^2),\ i=1,...,n,$ 互いに独立な測定誤差および個人差を表すランダム変動項である。

モデル（１）は、病型（群）を規定する長さｇの 2 値ダミー変数と、性別や年齢などの共変数が、それぞれ、$\mathbf{z}_i=(1,\text{z}_{i1},...,\text{z}_{g-1})'$ および $\mathbf{x}_i=(1,x_1^{(i)},...,x_u^{(i)})'$ で表され、目的変数のトレンドが折れ線で表現できることを想定すると、下記のように一般化できる。

$$\mathbf{y}_i=\begin{pmatrix}(\mathbf{h}_{i1}^{'}\otimes I_u)\mathbf{x}_{i1}^{'}\otimes\mathbf{z}_i\\ \vdots\\ (\mathbf{h}_{im_i}^{'}\otimes I_u)\mathbf{x}_{im_i}^{'}\otimes\mathbf{z}_i\end{pmatrix}vec\left(\Theta_0,\Theta_1,...,\Theta_k\right)+\eta_i\mathbf{1}_{m_i}+\boldsymbol{\varepsilon}_i,\qquad (2)$$

ただし、$\mathbf{h}_{ij}=(1,(d_{ij}-c_1)H(d_{ij}-c_1),...,(d_{ij}-c_k)H(d_{ij}-c_k))',\ j=1,...,m_i,\ i=1,...,n.$

モデル(2)は、目的変数の経時的トレンドが折れ線で記述されているため、モデル中で使用されている母数と時間変数や共変数との関連性の把握が容易である。また、各モデル間に自然な階層関係を導入することができるので、変数選択による最適モデル探索を含んだデータ解析が R などの既存のソフトウェアパッケージを利用して容易に実施できる。なお、本モデルは、通常の線形重回帰モデルの自然な拡張となっている。

３．データ解析と ADC データの解析結果

既述の ADC データに対して、モデル(1)を用いて R の関数 lmer による解析を行った。その方法と結果については、研究会の当日に発表する。