科研費シンポジウム「融合する統計科学」報告書

本シンポジウムは科学研究費・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）の助成により開催した。

■日時：2018 年 11 月 30 日～12 月 2 日
■場所：金沢大学サテライトプラザ
■開催責任者：星野伸明（金沢大学）

＜ プログラムと目次 ＞

日時	演題	講演者	所属	頁
11 月 30 日				
13:00	The Stirling and Eulerian numbers in the Edo Period	Xiaoling Dou	早稲田大学	1
13:40	Ewens 抽出公式に対する正規近似とポアソン近似	佃 康司	東京大学 大学院総合文化研究科	3
14:20	Posterior sampling from some non-exchangeable priors	間野 修平	統計数理研究所	5
15:00	休憩			
15:10	統計的形状モデルを利用した大規模点群レジストレーションとその高速化	広瀬 修	金沢大学理工研究域	7
15:50	Kolmogorov-Smirnov test based on kernel estimation	Rizky Reza Fauzi	九州大学大学院数理学府	9
16:30	休憩			
16:40	Tests for high-dimensional covariance matrices and correlation matrices under the strongly spiked eigenvalue model	石井 晶	東京理科大学 理工学部 情報科学科	11
17:20	癌の高次元遺伝子解析の諸問題（1）	新村 秀一	成蹊大学	13

18:00	終了			
12 月 1 日				
9:30	A Robust-filtering Method for Small Sample Economic Time Series	国友 直人	明治大学政治経済学部	15
10:10	Applications of Distance Correlation to Time Series	松井 宗也	南山大学経営学部	17
10:50	割引因子ポアソン-ガンマ状態空間モデルによる計数時系列モデルの逐次分析	入江 薫	東京大学経済学部	19
11:30	昼休み			
13:00	Testing for changes in income inequality in Japan	西埜 晴久	広島大学社会科学研究科・経済学部	20
13:40	Particle Filtering for Non-linear State-Space Models for Wind Speeds and Directions	稗田 尚弥	東京理科大学大学院工学研究科経営工学専攻	22
14:20	The Dantzig selector for a linear model of diffusion processes	藤森 洸	早稲田大学基幹理工学部応用数理学科	24
15:00	休憩			
15:10	多様な関数を用いた経時変動の罰則付推定法	永井 勇	中京大学 国際教養学部	26
15:50	傾向性仮説と変化点モデルの様々な応用	広津 千尋	明星大学連携研究センター	28
16:30	Characterizations of indicator functions for fractional factorial designs	青木 敏	神戸大学大学院理学研究科	30
17:10	休憩			
17:20	Pooling incomplete samples による統計解析	布能 英一郎	関東学院大学経済学部	32

18:00	Pitman's Closeness Domination in Predictive Density Estimation for Two Ordered Normal Means Under α-Divergence Loss	張 元宗	目白大学	34
18:40	終了			
19:00	懇親会			

12 月 2 日				
9:30	一般欠測データの下での２標本問題における多変量正規母集団の同等性検定	野村 玲実	東京理科大学大学院理学研究科 M1	36
10:00	単調欠測データが一様共分散構造を持つ場合の平行性仮説検定と水準差の信頼区間	佐伯 悠一郎	東京理科大学大学院理学研究科 M２	38
10:40	2-step 単調欠測データのもとでの部分平均ベクトルの検定に対する修正尤度比検定統計量	川崎 玉恵	東京理科大学理学部	40
11:20	Recent cylindrical models and their application to tree data set	阿部 俊弘	南山大学理工学部	42
12:00	終了			

The Stirling and Eulerian numbers in the Edo Period

Xiaoling Dou [*], Hsien-Kuei Hwang [†]

Similar to the recurrence relation satisfied by the binomial coefficients

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k},$$

the Stirling numbers of the first kind $\begin{bmatrix} n \\ k \end{bmatrix}$ and those of the second kind $\begin{Bmatrix} n \\ k \end{Bmatrix}$ can be computed by the recurrences:

$$\begin{bmatrix} n \\ k \end{bmatrix} = \begin{bmatrix} n-1 \\ k-1 \end{bmatrix} + (n-1)\begin{bmatrix} n-1 \\ k \end{bmatrix}$$
$$\begin{Bmatrix} n \\ k \end{Bmatrix} = \begin{Bmatrix} n-1 \\ k-1 \end{Bmatrix} + k\begin{Bmatrix} n-1 \\ k \end{Bmatrix},$$

subject to proper boundary conditions. Similarly, the Eulerian numbers $\left\langle {n \atop k} \right\rangle$ and the second order version $\left\langle\!\left\langle {n \atop k} \right\rangle\!\right\rangle$ satisfy the patterns

$$\left\langle {n \atop k} \right\rangle = (n-k)\left\langle {n-1 \atop k-1} \right\rangle + (k+1)\left\langle {n-1 \atop k} \right\rangle$$
$$\left\langle\!\left\langle {n \atop k} \right\rangle\!\right\rangle = (2n-k-1)\left\langle\!\left\langle {n-1 \atop k-1} \right\rangle\!\right\rangle + (k+1)\left\langle\!\left\langle {n-1 \atop k} \right\rangle\!\right\rangle.$$

Over the last five centuries or so, these numbers emerged naturally and were studied extensively in a large number of diverse areas, ranging from finite calculus and series summations to combinatorial structures and computer algorithms, and from statistics to spline interpolations, to name a few.

While the history of the developments of these numbers in the West has been largely and factually clarified, that in the East has remained mostly obscure. In this work, we aim to provide more historical materials during the Edo Period concerning these numbers, and to specially shed further light on their evolution (including introduction and use) in the Wasan History.

The following table summarize our findings on Stirling and Eulerian numbers in the Wasan History during the Edo Period, where we also list the closely connected Bell numbers $\text{Bell}_n := \sum_{0\le k\le n} \begin{Bmatrix} n \\ k \end{Bmatrix}$:

$$\{\text{Bell}_n\}_{n\ge 1} = \{1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, \cdots\}.$$

Note that $\sum_k \begin{bmatrix} n \\ k \end{bmatrix} = n!$, and that a table of these numbers, as well as a procedure (the recurrence) of computing them, already appear in Arima's 1769 book *Shūki Sanpo*.

[*]Faculty of Science and Engineering, Waseda University, Japan
[†]Institute of Statistical Science, Academia Sinica, Taiwan

Numbers	West	East
$\begin{bmatrix} n \\ k \end{bmatrix}$	James Stirling (1730)	SEKI Takakazu (1640–1708)
$\left\{ {n \atop k} \right\}$	James Stirling (1730)	SAKA Masanobu (1782)
$\left\langle {n \atop k} \right\rangle$	Leonhard Euler (1736)	MATSUNAGA Yoshisuke (1694–1744)
$\left\langle\!\left\langle {n \atop k} \right\rangle\!\right\rangle$	Jekuthiel Ginsburg (1928)	MATSUNAGA Yoshisuke (1694–1744)
Bell numbers	G. Dobiński (1877) Eric Bell (1938)	MATSUNAGA Yoshisuke (1694–1744) ARIMA Yoriyuki (1769)

Interestingly, unlike the early developments of these numbers in the West, which are mostly *computational* and *algebraic*, those carried out by the Wasankas already are not merely computational but also were motivated by *combinatorial* problems, adding another rich dimension to the diversity and usefulness of these numbers. We will present the combinatorial connections of these numbers to certain games frequently played during this Period.

References

[1] Y. Arima (1769). *Shūki Sanpo*.

[2] E.T. Bell (1938). The iterated exponential integers. *Annals of Mathematics*. 39, 539–557.

[3] G. Dobiński (1877). Summirung der Reihe $\sum nmn! \sum \frac{n^m}{n!} \sum \frac{n^m}{n!}$ für $m = 1, 2, 3, 4, 5, \ldots$. *Grunert's Archiv*. 61, 333–336.

[4] L. Euler (1736). Euler Archive, E055.

[5] J. Ginsburg (1928). Note on Stirling's Numbers, *Amer. Math.* Monthly 35, 77–80.

[6] J. Stirling (1730). *Methodus Differentialis*.

[7] Y. Matsunaga (1694–1744). *Sanpo Zenkei*.

[8] M. Saka (1782). *Sanpō Gakkai*.

[9] T. Seki (1712). *Katsuyo Sanpo*. (M. Araki, Y. Otaka).

Ewens 抽出公式に対する正規近似とポアソン近似

佃　康司（東京大・総合文化）

Ewens 抽出公式は集団遺伝学の文脈で Warren John Ewens が 1972 年に導入した確率分布 (Ewens, 1972) であり，多くの分野で議論されてきた．各分野における Ewens 抽出公式の役割については，Crane (2016) に解説がある．特に，確率の問題として古くから考えられてきたランダム置換と関連しているほか，ノンパラメトリック統計におけるベイズ法で事前分布として用いられる Dirichlet 過程からの標本確率分割が従う確率分布として統計学や機械学習といった分野で盛んに議論されてきた．

集団遺伝学における標本抽出理論において，標本（サイズを $n(\in \mathbb{N})$ とする）から計算される統計量として標本対立遺伝子分割における要素カウント $(C_1^n, \ldots, C_n^n)$ に興味があり，Ewens 抽出公式はその確率を与える．ここで，要素カウントにおける C_j^n $(j \in \mathbb{N})$ はサイズ n の標本において j 回の観測があったクラスの数を表している．Ewens 抽出公式には，母集団（を定常分布として生成する確率過程）のスケールされた突然変異率 θ と標本サイズ $n(\in \mathbb{N})$ という二つのパラメータが存在する．これまで要素カウントや θ の完備十分統計量である種類の数（$K_n(= \sum_{j=1}^n C_j^n)$ で定義される）などの統計量に対して多くの近似が示されてきた．中でも，特によく議論されてきた近似に K_n の正規近似とポアソン近似および $(C_1^n, C_2^n, \ldots)$ のポアソン近似がある．例えば，Arratia and Tavaré (1992) を参照．

これまでの Ewens 抽出公式に対する近似の多くは，二つのパラメータのうちの片方を固定して他方を無限大に発散するという極限をとる漸近論の設定を考えて導かれてきた．一方，両方のパラメータを同時に大きくするような漸近論も自然であり，Feng (2007) の 4 節で大偏差原理が議論されているものの，これまでよく議論されてきたような他の漸近的性質についてはあまり議論されてこなかった．講演では，Tsukuda (2017, 2018) に基づいて，要素カウント $(C_1^n, C_n^2, \ldots)$ と種類の数 K_n について，n と θ がともに大きくなるような漸近論の設定のもとで講演者が導いた結果を紹介した．

参考文献

Arratia, R.; Tavaré, S. (1992). Limit theorems for combinatorial structures via discrete process approximations. *Random Structures Algorithms* **3**, no.3, 321–345.

Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31**, no.1, 1–19.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3**, 87–112; erratum, (1972) *Theoret. Population Biology* **3**, 240; erratum, (1972) *Theoret. Population Biology* **3**, 376.

Feng, S. (2007). Large deviations associated with Poisson–Dirichlet distribution and Ewens sampling formula. *Ann. Appl. Probab.* **17**, no. 5–6, 1570–1595.

Tsukuda, K. (2017). Estimating the large mutation parameter of the Ewens sampling formula. *J. Appl. Probab.* **54**, no. 1, 42–54; correction, *J. Appl. Probab.* **55**, no. 3, to appear.

Tsukuda, K. (2018). On Poisson approximations for the Ewens sampling formula when the mutation parameter grows with the sample size. *Ann. Appl. Probab.* to appear.

Posterior Sampling from some Non-Exchangeable Priors[1]

Shuhei Mano[2]

The Institute of Statistical Mathematics

January 10, 2019

Let the Dirichlet process with the measure $\theta\mu$, where μ is a probability measure, be denoted by DP. Let $F \sim \mathrm{DP}(\theta;\mu)$. It is well known that $\mathbb{P}(X_1 \in \cdot) = \mu(\cdot)$ and

$$\mathbb{P}(X_{n+1} \in \cdot | X_1, ..., X_n) = \frac{\theta}{\theta+n}\mu(\cdot) + \frac{n}{\theta+n}\Lambda_n(X_1, ..., X_n)(\cdot).$$

Here, $\Lambda_n(X_1, ..., X_n) := n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution. In Bayesian context, F is called a prior process, and the posterior distribution is called the prediction rule. The sequential sampling scheme is well known as the Blackwell-MacQueen urn scheme (1973), or the Chinese restaurant process. The prediction rule induces measures on partitions. Let the j-th firstly appear value of $(X_1, ..., X_n)$ be X_j^*, $j \in \{1, 2, ..., k\}$. Then, $\mathcal{P}_{n,k} \ni (n_1, ..., n_k)$, $n_j := \#\{i; X_i = X_j^*\}$ is a integer partition of a positive integer n. The distribution of the multiplicities of integers $(c_1, ..., c_n)$, $c_i := \#\{j; n_j = i\}$, is

$$\mathbb{P}(C_1 = c_1, ..., C_n = c_n) = \frac{n!}{(\theta)_n}\prod_{i=1}^n \left(\frac{\theta}{i}\right)^{c_i}\frac{1}{c_i!},$$

where $(\theta)_n := \theta(\theta+1)\cdots(\theta+n-1)$. This measure on partitions is called the Ewens sampling formula (1972; Antoniak 1974, Sibuya 1993). A generalization, (exchangeable) Gibbs partitions, are commonly used to characterize prior processes.

Definition 1 (Pitman 2006; M 2018)**.** Gibbs partition is the probability measure on partitions $\lambda \vdash n \in \mathbb{N} := \{1, 2, ...\}$ of the form

$$\mathbb{P}(C = c) = \frac{v_{n,l(c)}}{B_n(v,w)} n! \prod_{i=1}^n \left(\frac{w_i}{i!}\right)^{c_i}\frac{1}{c_i!}.$$

Here, $l(c) = c_1 + \cdots + c_n$ is the length, and the normalization constant is written as

$$B_n(v,w) = \sum_{k=1}^n v_{n,k}B_{n,k}(w), \qquad B_{n,k}(w) = \sum_{c\in\mathcal{P}_{n,k}} n! \prod_{i=1}^n \left(\frac{w_i}{i!}\right)^{c_i}\frac{1}{c_i!}.$$

The latter is known as the partial Bell polynomial.

Examples of Gibbs partitions are presented in Proceedings and in M (2018). The characterizations are given by Gnedin & Pitman (2005). A Gibbs partition is infinitely exchangeable iff

[1] Based on ongoing work with Jaeyong Lee at Seoul National University.

[2] E-mail: smano@ism.ac.jp

$w_i = (1-\alpha)_{i-1}$, $\alpha < 1$. Without exchangeability, the prediction rule is not available anymore (Pitman 1995, Lee et al. 2013, Proceedings) and we need another sampling scheme than the Blackwell-MacQueen urn. For a Gibbs partition, the length $l(c) = k$ is the sufficient statistic for parameters $(v_{n,k})$ and the conditional distribution is an A-hypergeometric distribution (Takayama et al. 2018):

$$\mathbb{P}(C = c|AC = b) = \frac{1}{Z_A(b;x)}\frac{x^c}{c!}, \qquad c \in \mathbb{N}_0^{n-k+1},$$

where the A-hypergeometric polynomial $Z_A(b;x)$ is defined by

$$A = \begin{pmatrix} 0 & 1 & 2 & \cdots & n-k \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}, \quad b = \begin{pmatrix} n-k \\ k \end{pmatrix}, \quad x_i = \frac{w_i}{i!},$$

and $n!Z_A(b;x) = B_{n,k}(w)$. M (2017) obtained a direct sampler for A-hypergeometric distributions, and it provides a direct sampler from a Gibbs partition without exchangeability. The sampler can be used as an alternative to the Blackwell-MacQueen urn scheme.

A typical Bayesian semiparametric setting is data $(Y_1, ..., Y_n)$ derived from a hierarchical model (MacEachern 1994; Escober & West 1995)

$$Y_i|X_i, \sigma \overset{ind}{\sim} \mathrm{N}(Y_i; X_i, \sigma), \quad \sigma \sim \pi(\sigma), \qquad X_i|P \overset{ind}{\sim} P, \quad P \sim \mathrm{DP}(\theta; \mu).$$

The marginal with the Dirichlet process, $(X_1, ..., X_n) \sim \mathbb{P}(X_1, ..., X_n)$, is sampled directly by the Blackwell-MacQueen urn scheme. Let us call the sampling from the marginal with a prior process *prior sampling.* To draw from posterior $\pi(Y, \sigma|X)$ (we will call *posterior sampling*), a Gibbs sampler is used, where we iterative draw values from conditional distributions

$$X_i|(X_{-i}, \sigma, Y), \qquad i \in [n], \qquad \sigma|(X, Y).$$

The prediction rule gives the updating rule

$$\mathbb{P}(X_i = \cdot|X_{-i}, \sigma, Y) \propto \mathrm{N}(Y_i; X_i = \cdot, \sigma)\theta\mu(\cdot) + \sum_{j=1}^{k} \mathrm{N}(Y_i; X_j{}^*, \sigma)n_j\delta_{X_j{}^*}(\cdot). \tag{1}$$

Let us consider use of a non-exchangeable Gibbs partition. Thanks to the direct sampler from a Gibbs partition introduced above, the prior sampling is straightforward. For posterior sampling, a possibility is use of an independent Metropolis–Hastings algorithm with the prior sampling. This cause poor acceptance, since the acceptance is determined based on the vector $(X_1, ..., X_n)$. Improvement of the acceptance with aid of n-exchangeability was discussed in the talk.

Remark 1. After the talk, we found that mixture modeling with non-exchangeable Gibbs partitions have been extensively discussed, where a non-exchangeable Gibbs partition with $v_{n,k} = 1$ is called a product partition model (Quintana & Iglesias 2003, JRSS B 65: 557-574). It seems that a Gibbs sampler with updating rule similar to (1) has been used. The updating rule is not consistent with infinite exchangeability.

統計的形状モデルを利用した大規模点群レジストレーションとその高速化

広瀬 修[†]

金沢大学理工研究域[†]

1. はじめに

点群位置合わせ問題は，物体の形状を表現する２つの点集合間に対し，１つの点群をもう一方の点群に移す写像を求める問題である．点群位置合わせ問題は想定する写像の種類に応じて剛体変換と非剛体変換のものに分類され，最近では非剛体変換に基づいた点群位置合わせが，その柔軟さのため非常に活発に研究されている．

Coherent point drift (CPD) は非剛体変換に基づいた点群レジストレーション手法の代表的な手法である [Myronenko 2010]．CPD の成功の要因としてまず挙げられるのが，外れ値への耐性である．ここで，外れ値とは点群によって表現される形状とは無関係に存在する点とする．CPD は点群位置合わせ問題を混合確率分布の推定問題として定義する．その際，混合分布の構成分布の１つとして外れ値の分布を明示的に与えることが，外れ値への耐性の主要な要因である．CPD のもう１つの成功要因として挙げられるのが，非剛体換される点群に対する「変位場の滑らかさ」である．変位場の滑らかさとは，非剛体変換される点群を構成する任意の点の変位と，その他の点の変位が，その距離が近ければ近いほど相関するとした仮定である．この仮定は非常に自然な仮定であるため，CPD は多くの点群位置合わせ問題において精度の高い位置合わせ結果を与える．一方で，変位場の滑らかさの仮定が適切ではない場合，CPD は容易に位置合わせに失敗する．例えば，人間の手の形状マッチングを行う場合，人差し指と中指を構成する点は比較的近くに位置するが，その動きは逆相関する傾向があり，このような場合には変位場の滑らかさの仮定だけでは不十分であるためである．

この問題を克服する方法の１つとして挙げられるのが，教師あり学習に基づく方法である．もし人差し指と中指の動きが逆相関する傾向にあることを事前に知っていれば，その知識を位置合わせアルゴリズムに組み込むことにより，高精度の位置合わせが期待できるからである．今回，新たに開発した教師あり学習法に基づいた点群位置合わせ手法に対する位置合わせ性能の評価について報告する．開発手法は外れ値の分布を構成分布の１つとして持つ混合分布に基づいているため，外れ値への優れた耐性を有する．また，物体の形状変化モデルに，訓練データから得られる事前知識を組み込むため，変位場の滑らかさの問題を自然に解決することができる．講演では大規模な点群データに対応するための高速化についても議論を行う．

2. 実験

開発した位置合わせ手法と代表的な点群位置合わせ手法である CPD と Thin Plate Spline Robust Point Matching (TPS-RPM) [Chui 03] の性能の比較を行った．使用したデータは IMM hand データである[Stegmann 2002]．このデータは４０種類の人間の手の画像に対し，手と背景の境界部分に５６個の特徴点を人手で打点したものである．各々の特徴点は４０種類の画像で対応関係がとられるように打点されている．図１が IMM hand データ中の点群番号６に対する CPD, TPS-RPM および開発手法の位置合わせ結果を表す．図１の１段目が推定されるべき正解の手の形状を表し，２段目が位置合わせの対象となるデータで，手法の頑健性を検証するため正解データに４種類の改変を施したものである．左から順に (1)点の複製，(2)欠損，(3)外れ値の付加，(4)回転の改変を表す．図１の３段目の赤色で示された点が最適化の初期形状，すなわち，平均形状を表す．３～５段目がそれぞれ提案手法，CPD，TPS-RPM の適用結果を表す．データ (1)，(2)，(3)，(4)の全てで提案手法が正解とほぼ同一の形状の推定に成功した．CPD と TPS-RPM については，全てのデータで親指以外の指が細くなる現象が見られた．これは，変位場の滑らかさの欠点の１つを表している．ある指を形成する点が，近隣に存在する他の指を形成する点ともその変位が相関することが原因である．データ(1) に対しては全ての手法で概ね正解の形状を空いてしたが，CPD と TPS-RPM には指が細くなる現象が見られた．データ(2) に対して提案手法は欠損領域の点群の推定に成功し

た．一方で CPD と TPS-RPM は親指を含む全ての　　指が短く描画された．

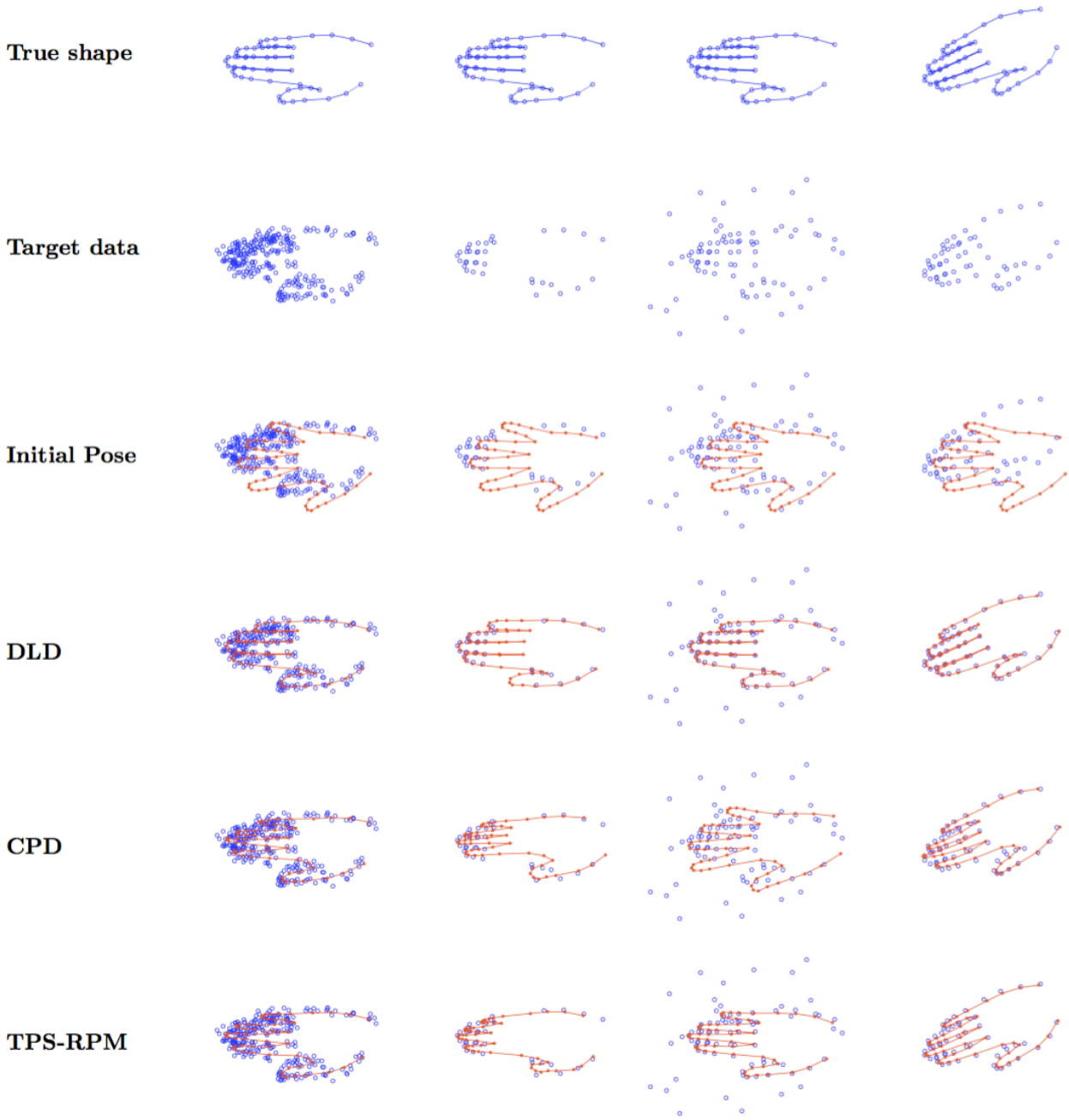

図 1．提案手法，CPD，TPS-RPM の比較．１段目の図は推定されるべき正解の形状を表す．２段目の図は位置合わせの対象となるデータ．３段目の赤の点群が最適化の初期形状を表す．手法の頑健性を検証するため正解データに４種類の改変を施した．左から順に (1)点の複製，(2)欠損，(3)外れ値の付加，(4)回転の改変を表す．３～５段目がそれぞれ提案手法，CPD，TPS-RPM の適用結果を表す．

Kolmogorov-Smirnov test based on kernel estimation

Rizky Reza Fauzi, Graduate School of Mathematics Kyushu University
Maesono Yoshihiko, Faculty of Mathematics Kyushu University

1. Boundary-free kernel distribution function estimators

Let $X_1, X_2, ..., X_n$ be independently and identically distributed random variables with an absolutely continuous distribution function F_X and a density f_X. The classical nonparametric estimator of F_X has been the empirical distribution function $F_n(x) = \sum_{i=1}^{n} I(X_i \leq x)/n$($I(\cdot)$: indicator function). Let K be a kernel function and $h > 0$ be a bandwidth. We assume that K is a symmetric (about 0) continuous nonnegative function with $\int_{-\infty}^{\infty} K(v)\mathrm{d}v = 1$, as well as $h \to 0$ and $nh \to \infty$ when $n \to \infty$. Since distribution function is actually an integral of density function, this kernel density estimator gave an idea to define a kernel distribution function estimator. Nadaraya (1964) defined it as

$$\widehat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R},$$

where $W(v) = \int_{-\infty}^{v} K(w)\mathrm{d}w$.

Under the condition that f_X has one continuous derivative f'_X, it has been proved that, as $n \to \infty$,

$$\begin{aligned} Bias[\widehat{F}_X(x)] &= \frac{h^2}{2} f'_X(x) \int_{-\infty}^{\infty} v^2 K(v)\mathrm{d}v + o(h^2), \\ Var[\widehat{F}_X(x)] &= \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} r_1 f_X(x) + o\left(\frac{h}{n}\right) \end{aligned}$$

where $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)\mathrm{d}v$. It is easy to show that r_1 is a nonnegative number.

If we deal with $\mathbb{R}^+$ or unit interval, the standard kernel distribution function estimator will suffer the so called boundary bias problem. To solve this problem, we utilise a function g which bijectively transform the support A of the random variable under consideration into $\mathbb{R}$, then doing the usual standard kernel distribution function estimation of $Y = g(X)$. Here we propose an estimator

$$\widetilde{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left[\frac{g(x) - g(X_i)}{h}\right], \quad x \in A.$$

Just as the standard one, we have a bias and a variance

$$\begin{aligned} Bias[\widetilde{F}_X(x)] &= \frac{h^2}{2} c(x) \int_{-\infty}^{\infty} v^2 K(v)\mathrm{d}v + o(h^2) \\ Var[\widetilde{F}_X(x)] &= \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n}\frac{f_X(x)}{g'(x)} r_1 + o\left(\frac{h}{n}\right), \end{aligned}$$

where $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)\mathrm{d}v$ and

$$c(x) = \frac{f'_X(x)}{[g'(x)]^2} - \frac{f_X(x)g''(x)}{[g'(x)]^3}.$$

2.Boundary-free smoothed Kolmogorov-Smirnov type test

The Kolmogorov-Smirnov (KS) test is the most popular GOF test used in practice. Unfortunately, it lacks of smoothness that can lead to smaller power at the tails, which is important in many practical applications. It is natural if one uses the naive kernel distribution function estimator in place of the empirical distribution function. Thus, instead of the standard KS statistic

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F_X(x)|$$

being used to test whether random variable X having F_X as its distribution, we can reformulate by smoothing it to

$$\widehat{D} = \sup_{-\infty<x<\infty} |\widehat{F}_X(x) - F_X(x)|,$$

where $\widehat{F}$ is the kernel distribution function estimator.

However, a new problem is raising when the support of the random variable that we are dealing with is not the entire real line, i.e. boundary problem. To overcome this problem, we propose to use our estimator in section 1 to substitute empirical distribution function in standard KS statistic. Therefore, we define the boundary-free smoothed KS type test as

$$\widetilde{D} = \sup_{-\infty<x<\infty} |\widetilde{F}_X(x) - F_X(x)|,$$

where $\widetilde{F}_X$ is our boundary-free kernel distribution function estimator. We compare a numerical study by calculating simulated power of our proposed test with $n = 50$, and then we compare it with the result of the standard KS test.

The asymptotic behaviours of our proposed test statistic are stated in the following theorems.

Theorem 1 Let X be a random variable with distribution function F_X supported on a set A. If $\widetilde{F}_X$ is the proposed boundary-free kernel distribution function estimator, then

$$\sup_{-\infty<x<\infty} |\widetilde{F}_X(x) - F_X(x)| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 2 Let X be a random variable with distribution function F_X supported on a set A. If $\widetilde{D}$ is the proposed boundary-free smoothed KS-type statistic, then

$$\lim_{n\to\infty} \Pr(\sqrt{n}\widetilde{D} \le x) = \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} \exp\left[\frac{-(2i-1)^2\pi^2}{8x^2}\right].$$

Tests for high-dimensional covariance matrices and correlation matrices under the strongly spiked eigenvalue model

Aki Ishii[a], Kazuyoshi Yata[b] and Makoto Aoshima[b]

[a] Department of Information Sciences, Tokyo University of Science
[b] Institute of Mathematics, University of Tsukuba

1 Introduction

we consider the equality test of covariance matrices when the data dimension is much larger than the sample size. Suppose we have two classes $\pi_i,\ i = 1, 2$. We define independent $d \times n_i$ data matrices, $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}],\ i = 1, 2$, for $\pi_i,\ i = 1, 2$. We assume that $\boldsymbol{x}_{ij},\ j = 1, ..., n_i$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. We assume $n_i \geq 4,\ i = 1, 2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{1(i)}, ..., \lambda_{d(i)})$ having $\lambda_{1(i)} \geq \cdots \geq \lambda_{d(i)} (\geq 0)$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)}, ..., \boldsymbol{h}_{d(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. We assume $\lambda_{2(i)} > 0$ for $i = 1, 2$, and $\lambda_{1(i)}$s are of multiplicity one in the sense that $\liminf_{d\to\infty} \lambda_{1(i)}/\lambda_{2(i)} > 1$ for $i = 1, 2$. We consider the equality test of covariance matrices as follows:

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2. \tag{1.1}$$

Aoshima and Yata [1] gave a test procedure based on the quantity of $\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)$. They also discussed sample size determination so as to have a prespecified size and power simultaneously. Li and Chen [5] considered the test problem by using the quantity of $\mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\}$. The above references discussed asymptotic properties of their test procedures when $d \to \infty$ and $n_i \to \infty$ under the following eigenvalue condition:

$$\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \ \text{ as } d \to \infty \text{ for } i = 1, 2. \tag{1.2}$$

Aoshima and Yata [2] called (1.2) the "non-strongly spiked eigenvalue (NSSE) model". On the other hand, Ishii, Yata and Aoshima [3] investigated asymptotic properties of the first principal component and considered the test problem (1.1) when $d \to \infty$ while n_is are fixed under the following eigenvalue condition:

$$\frac{\sum_{s=2}^{d} \lambda_{s(i)}^2}{\lambda_{1(i)}^2} = o(1) \text{ as } d \to \infty \text{ for } i = 1, 2. \tag{1.3}$$

Note that (1.3) implies the conditions that $\lambda_{2(i)}/\lambda_{1(i)} \to 0$ and $\lambda_{1(i)}^2/\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \to 1$ as $d \to \infty$. The condition (1.3) is generalized as

(A-ii) $\displaystyle\liminf_{d\to\infty} \left\{ \frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \right\} > 0$ for $i = 1$ and 2.

Aoshima and Yata [2] called (A-ii) the "strongly spiked eigenvalue (SSE) model" and showed that high-dimensional data often have the SSE model.

2 A new test procedure under the SSE model

Let $\Delta = ||\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2||_F^2 = \mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\}$. Under (1.3), it holds that as $d \to \infty$

$$\Delta = (\lambda_{1(1)} - \lambda_{1(2)})^2 + 2\lambda_{1(1)}\lambda_{1(2)}\{1 - (\boldsymbol{h}_{1(1)}^T\boldsymbol{h}_{1(2)})^2\} + o(\lambda_{1(1)}^2 + \lambda_{1(2)}^2). \tag{2.1}$$

We consider (2.1) as a starting point to handle the SSE model (A-ii). We give a test statistic based on (2.1) and show that it holds an asymptotic null distribution even when (1.3) is not met. Let $\overline{\boldsymbol{X}}_i = [\bar{\boldsymbol{x}}_i, ..., \bar{\boldsymbol{x}}_i]$ and

$\bar{\boldsymbol{x}}_i = n_i^{-1}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$ for $i = 1, 2$. We define the sample covariance matrix as $\boldsymbol{S}_{in_i} = (n_i - 1)^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T$ for $i = 1, 2$. We denote the dual matrix of $\boldsymbol{S}_{in_i}$ by $\boldsymbol{S}_{iD}$ and define its eigen-decomposition as follows:

$$\boldsymbol{S}_{iD} = (n_i - 1)^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i) = \sum_{s=1}^{n_i-1}\hat{\lambda}_{s(i)}\hat{\boldsymbol{u}}_{s(i)}\hat{\boldsymbol{u}}_{s(i)}^T, \tag{2.2}$$

where $\hat{\lambda}_{1(i)} \geq \cdots \geq \hat{\lambda}_{n_i-1(i)}$ (≥ 0) and $\hat{\boldsymbol{u}}_{s(i)}$ denotes a unit eigenvector corresponding to the eigenvalue $\hat{\lambda}_{s(i)}$. Note that $\boldsymbol{S}_{in_i}$ and $\boldsymbol{S}_{iD}$ share non-zero eigenvalues. If one uses the noise reduction (NR) method given by Yata and Aoshima [7], $\lambda_{1(i)}$ and $\boldsymbol{h}_{1(i)}$ are estimated by

$$\tilde{\lambda}_{1(i)} = \hat{\lambda}_{1(i)} - \frac{\mathrm{tr}(\boldsymbol{S}_{iD}) - \hat{\lambda}_{1(i)}}{n_i - 2} \text{ and } \tilde{\boldsymbol{h}}_{1(i)} = \{(n_i - 1)\tilde{\lambda}_{1(i)}\}^{-1/2}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)\hat{\boldsymbol{u}}_{1(i)}. \tag{2.3}$$

for $i = 1, 2$, where $\hat{\boldsymbol{u}}_{1(i)}$ is given in (2.2). Let $\delta_i = \mathrm{tr}(\boldsymbol{\Sigma}_i^2) - \lambda_{1(i)}^2,\ i = 1, 2$. If one applies the NR method to estimating Δ in (2.1) straightforwardly, the estimation of Δ includes huge noise for high-dimensional data. In this talk, we evaluate the amount of the noise and consider bias correction. Let $\eta = \delta_1^{1/2}/\lambda_{1(1)} + \delta_2^{1/2}/\lambda_{1(2)}$. By using the cross-data-matrix (CDM) method given by Yata and Aoshima [6] and the NR method, we estimate η by $\hat{\eta}$. We provide the following new test statistic:

$$T_{\mathrm{NR}} = \frac{(\tilde{\lambda}_{1(1)} - \tilde{\lambda}_{1(2)})^2 + 2\tilde{\lambda}_{1(1)}\tilde{\lambda}_{1(2)}\{1 - \min\{1, (\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)})^2\}\}^{1+\hat{\eta}}}{\sum_{i=1}^2 2\tilde{\lambda}_{1(i)}^2/(n_i - 1)}.$$

Proposition 2.1 (Ishii, Yata and Aoshima [4])**.** *Under (A-ii), H_0 and some regularity conditions, it holds that $T_{NR} \Rightarrow \chi_1^2$ as $\min(n_1, n_2, d) \to \infty$.*

We gave another test statistic under (A-ii) by dividing high-dimensional eigenspaces into the first eigenspace and the others. We gave asymptotic null distribution and the power of the test statistic. We demonstrated the new test procedure by using actual microarray data sets. We also considered a correlation test for high-dimensional data under the SSE model.

References

[1] Aoshima, M., Yata, K., 2011. Two-stage procedures for high-dimensional data. *Sequential Anal. (Editor's special invited paper)* 30, 356-399.

[2] Aoshima, M., Yata, K., 2018. Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Stat. Sin.* 28, 43-62.

[3] Ishii, A., Yata, K., Aoshima, M., 2016. Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *J. Stat. Plan. Inference* 170, 186-199.

[4] Ishii, A., Yata, K., Aoshima, M., 2018. Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model. *J. Stat. Plan. Inference*, revised.

[5] Li, J., Chen, S.X., 2012. Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* 40, 908-940.

[6] Yata, K., Aoshima, M., 2010. Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *J. Multivariate Anal.* 101, 2060-2077.

[7] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* 105, 193-215.

癌の高次元遺伝子解析の諸問題（1）

成蹊大学　名誉教授
新村秀一

1. 癌の高次元遺伝子解析を取り巻く諸問題について

1999 年から 2004 年まで、米国の 6 研究グループが医学誌に論文を掲載し、研究に用いた microarray データをインターネット上に公開した。ハーバード大学医学部の Golub ら（1999）はサイエンスに論文を発表し、「顕微鏡などで細胞や遺伝子の生物学的な変異から癌遺伝子を見つける限界に言及し、1970 年から microarray による体系的な研究が必要であり研究しているが、大きな成果を未だ得ていない」と真摯に述べている。彼らは独自の手法を開発し、それらを Self-organizing maps(SOM)や LOO や生存時間解析等の統計手法と組み合わせている。これらのデータは、例えば癌と健常者の 2 群 100 例を、1 万個の遺伝子の発現量で判別する問題であり、最も判別に適した遺伝子をがん遺伝子と特定する研究である。しかし、幾つかの論文では SVM の利用は述べられているが明確な結果の紹介はない。そして残念なことに NIH が乳がんを除いて、この種の研究に疑問を呈したレポートを出し、医学的な研究が終わったようだ。

2. 新しい判別理論と Matryoshka Feature Selection Method(新理論 2)による癌の遺伝子解析の成功

筆者は 2015 年 10 月 25 日に富山市で開催された科研費シンポジュームで石井博士の発表で上記の 6 種のデータが公開されていることを知った。彼女から 28 日にメールで microarray を入手できる HP を知り入手した。そして 12 月 22 日までの 56 日間で癌の遺伝子解析を完成させた。

Fact3：6 種の microarray すべてが LSD すなわち最小誤分類数 MNM=0 である。

Fact4：全ての microarray は、多数の線形分離可能な Small Matryoshka(SM)とそれ以外の雑音空間に分割できる。最近まで、2 群が各 SM の部分空間で完全に分かれて識別できるので、SM に含まれる遺伝子が癌遺伝子であり、癌の Basic Gene Set(BGS)と考えていた。そしてその和集合と各 BGS は信号空間であると定義した。すなわち、

1) 分散共分散行列に基づく判別分析は LSD を正しく判別できないので、癌の遺伝子解析に全く役に立たない。
2) 改定 IP-OLDF（RIP）と改定 LP-OLDF とハードマージン最大化 SVM（H-SVM）だけが microarrays が LSD であることがわかる。しかしなぜ多くの研究者は SVM を利用しているのに、この重要な信号を見過ごしたかという問題が起きる。
3) RIP と改定 LP-OLDF だけが、microarray を多数の SM と雑音空間に分割できる。これが今回の発表のテーマである。

そして、SM は小標本であるので、簡単に分析できる。しかし、2 群が各 SM で完全に分かれているにもかかわらず、主成分分析、クラスター分析、一元配置による分散分析などで線形分離可能な事実を示さない。ロジスティック回帰は全ての SM が NM=0 であるが、分散共分散行列による LDF や QDF は NM=0 にならないものが多い。一方、RIP、改定 LP-OLDF と H-SVM は、2 つの SV で-1 以下に class1、1 以上に class2 の症例を正しく判別し、判別スコア(Discriminant Score, DS)の範囲に対する比率 RatioSV が 5%以上になるものが多い。そこで遺伝子の変わりに各 SM に含まれる遺伝子の総合特性値である RIP、改定 LP-OLDF と H-SVM の DS である RipDSs、LpDSs と HsvmDSs を変数とするデータを作成した。これらを PCA やクラスター分析で分析すると、2 群は完全に分かれる。

以上から SM で、癌と健常者の 2 群は完全に分かれているが、通常の統計手法ではそれがわからない。しかし、DS で作成したデータでは LSD の事実が簡単にわかる。すなわち RIP、改定 LP-OLDF と H-SVM の判別スコアが、高次元 microarray の信号でないかと考えるに至った。これに関する詳細は、広島のシンポジュームで報告を予定している。

3. 統計的判別分析の問題

Golub 以前にも他の遺伝子データが公開されていて、統計研究者やパターン認識などの工学者研究者は、質が高く無償で提供されたこの高次元データ(small n large p)を、格好の研究テーマとしてとらえて研究を行ってきた。多くの研究論文には”Feature Selection Methods（統計的にはモデル選択とか変数選択）”とか”Filtering”という用語を含んだものが多い。そして医学論文には見られない、次の 3 つの困難を指摘する論文もある。

1) 高次元データ(small n large p ; n<<p)の困難さ：これは、例えば100症例から1万次元の分散共分散行列を構築することが端的な事例である。2000 年以前に、国際会議で発表もあったがいつのまにかなくなった。2015年11月に六本木で行われたJMPのDiscovery SummitでJMPの創業者のSall博士が特異値分解を用いた横長データのFisherのLDFを発表した。無償で1か月借り受け、6種のmicroarrayを分析したが、誤分類数（NM）は高かった。すなわち、分散共分散に基づくLDFはLSDを正しく判別できず、癌の遺伝子解析に役に立たない点である。また、数理計画法では全く問題にならずむしろlarge n small pの方が、制約式が増えて計算時間がかかる。すなわち、この困難は2変数の相関でもって1万変数が関係づけられる統計に限定された困難である。数理計画法は変数間の関係が小さい。
2) NP-hard:1万変数の判別分析で、適切な部分モデルを選ぶことは困難である。この困難の真の意味を考えていない側面がある。即ち統計的判別関数や2次計画法で定式化されたH-SVMは、定義域で唯一の最適な判別関数が求まる。部分空間にも最適解がある場合、モデル選択で部分モデルの中から最適解を見つける必要がある。さらに問題なのはLSDでは、最小次元のBGSを含むすべてのモデルが最適解になることである。Feature Selectionで MNM=0という基準で最適化モデルを探さない限り、これ等の研究は無意味である。
3) 信号と雑音の分離：この問題は的を得ているが、「信号」の定義がはっきりしない。癌のMicroarrayデータにおける信号はMNM=0である。2017年1月の金沢における科研費シンポジュームで、私が初めて癌の遺伝子解析に成功したという説明に、青嶋氏より我々の方が先行している旨の意見があった。よく考えてみると、2015年の富山で石井氏がmicroarrayデータでPCAを行うと、第1固有値だけがスパイク上に大きな固有値になり、一般的な常識で考えられない結果になるという話を聞いた。しかし、私はmicroarrayが簡単に入手でき、これまで判別分析で4つの問題を解決してきたが、「未解決の5番目の癌の遺伝子解析」を解決していないことに気づいた。癌と健常あるいは異なった2種の癌が、遺伝子空間で完全に分かれて2つの球に布置しているというのが青嶋と矢田らの結論である。このことは、私の「2群はmicroarrayでMNM=0であり、それが多くのMNM=0であるSMとMNMが1以上の雑音空間に分割できる」という驚く結果を統計的に検証した研究であることに気づいた。

4. 癌の遺伝子解析から癌の遺伝子診断

癌の遺伝子解析は、判別分析の4つの問題と応用問題として6種のmicroarrayで全てのSMを求めることができSpringerの本で解決した問題5(Shinmura, 2016d)を指す。豊富な実証研究の成果である。

2016年になって、SMはn個以下の遺伝子で構成された小標本であるので、統計手法で簡単に分析できると考えた。しかしロジスティック回帰だけがNM=0になり、PCAやクラスター分析では２つのclassが線形分離可能な事実を示さなかった。それも仕方がないと当初は考えたが、RIP、改定LP-OLDFとH-SVMは、MNMあるいはNMが0である。そこで判別スコアの範囲に対して2つのSV間の距離2の比をRatioSVとして求めると6種の最大値は[11.67%, 38.93%]と異常に大きい。これに反して、Alonの130個のBGSは1%未満である。そして、RIPの判別スコアを遺伝子の代わりに変数として用いたデータを作成した（RipDSsデータ)。これをPCAで分析すると健常症例が第1主成分で負のある値以下に、癌症例が正のある値以上に布置することが分かった。そして各SMで求めたRipDSとPCAで求めた総合化されたRipDSを**癌の悪性度指標**と呼び、医学の素人ながら**癌の遺伝子診断**の突破口を開いたのではないかと考えた。それらの成果をまとめて、2017年にAmazonからKindle版として出版した。予約注文で600部以上がダウンロードされ幸先が良いと思ったがその後が続かない。きっとResearch Gateで出版案内したので、癌の遺伝子解析に関連した研究者がNIHの敗北宣言後も600人程度は細々といると考えられる。9月時点でRGのRead数が11万を超えたが癌の遺伝子解析関連のRead数は少ない。また7月末にラスベガスで開催されたBiocomp18で8月3日（金）に開催ホテルのLuxorで朝3時に目が覚め空港に行く7時まで,初めて約140にいたFollowingとFollowerの所属を調べた。帰国後RGがダウンし復旧後に1391人に増えた。ひょっとしてBiocomp18で遺伝子関連の専門家に注目されたかと調べてみたが、40人程度しかいないようである。また、彼らが癌の遺伝子関連のDraftを特に読んでいる事実も得られなかった。癌の遺伝子診断の成果は、専門家に検証してもらわなければ意味がないので、袋小路に入っている。

今後の課題として、青嶋・矢田らの結果に対して、筆者の研究アプローチで具体的な幾つかの事実で同じことを示していることを示す予定でいる。

[1] Aoshima M, Yata K (2017) Two-sample tests for high-dimension, strongly spiked eigenvalue models, Statistica Sinica Preprint No.ss-2016-0063R2:1-31

A Robust-filtering Method for Small Sample Economic Time Series

Naoto Kunitomo

In collaboration with Seisho Sato [1]

October 25, 2018

Key Words

Non-stationary economic time series, Errors-variables models, Measurement Error, trend and seasonality, Robust-filtering, SIML, Fourier-Inversion, Large Dimension.

Summary

We develop a new filtering method to estimate the hidden states of random variables and to handle multiple time series data, and small sample economic time series in particular. Kunitomo and Sato (2017), and Kunitomo, Sato and Kurisu (2018) have developed the separating information maximum likelihood (SIML) method for estimating the non-stationary errors-in-variables models. They have discussed the asymptotic properties and finite sample properties of the estimation of unknown parameters. We utilize their results to solve the filtering problem of hidden random variables, which gives a powerful new method of handling macro-economic time series.

Kitagawa (2010) has discussed the standard statistical filtering methods already known including the Kalman-filtering and the particle-filtering methods. Since (i) these methods depend on the underlying distributions such as the Gaussian distributions for the Kalman-filtering and (ii) the procedures essentially depend on the dimension of state variables, there may be some difficulty to extend to the high-dimension cases even when it is fixed, say 100. On the other hand , we can expect that our method has some merits when we handle small sample economic times series with non-stationarity and seasonality with many variables because our method does not depend on the specific distributions as well as the dimensions of random variables. See Kunitomo, Awaya and Kurisu (2017) for a comparison of small sample properties of the ML and SIML methods.

The most important feature of the present procedure is that it can be applicable to small sample time series data with large dimension. Also our new method has a solid mathematical and statistical foundation, which is related to Doob (1953) and Hannan (1971).

[1] We owe Akihiko Takahashi for a suggestion on our filtering method.

References

[1] Brockwell, P. and R. Davis (1990), *Time Series : Theory and Methods*, 2nd Edition, Wiley.

[2] Doob, J.L. (1953), *Stochastic Processes*, Wiley.

[3] Hannan, E.J. (1970), *Multiple Time Series*, Wiley.

[4] Kitagawa, G. (2010), *Introduction to Time Series Analysis*, CRC Press.

[5] Kunitomo, N. (2018), "Common Seasonal Factors," in preparation.

[6] Kunitomo, N. , Sato and D. Kurisu (2018), *Separating Information Maximum Likelihood Estimation for High Frequency Financial Data*, Springer.

[7] Kunitomo, N. and S. Sato (2017), "Trend, Seasonality and Economic Time Series : the Non-stationary Errors-in-variables Models," SDS-4, MIMS, Meiji University, http://www.mims.meiji.ac.jp/publications/2017-ds.

[8] Kunitomo, N., N. Awaya and D. Kurisu (2017), "Some Properties of Estimation Methods for Structural Relationships in Non-stationary Errors-in-Variables Models," SDS-3, MIMS, Meiji University.

Applications of Distance Correlation to Time Series：距離の相関係数の時系列解析への応用

南山大学　松井 宗也

研究内容

近年、「距離の相関係数」(Distance Correlation) という 2 つの確率ベクトルの依存関係を測る指標が注目されている。ここで 2 つのベクトルの次元は任意で異なっていてもよい。2005 年に Székely, G.j. により提案されて以来盛んに研究されている（詳しい定義等は [Székely et al.(2007)] や [Székely and Rizzo(2009)] 参照）。注目すべき特徴は、ケンドールやスピアマンの順位相関係数といった従来の指標では捉えることの難しい非線形な相関も検知できることである。そもそも 2 つの確率変数が独立であるとは、その同時特性関数がそれぞれの周辺特性関数の積として書けることと同値である。「距離の相関係数」の基本的なアイデアは、同時特性関数と独立な場合のそれ（各々の周辺関数の積）の距離をみて相関を測るものである。2 つの確率ベクトル $X \in \mathbb{R}^p$ と $Y \in \mathbb{R}^q$（次元は任意で異なっていてもよい）の同時特性関数を $\varphi_{X,Y}(s,t)$ とおき、またそれぞれの周辺特性関数を $\varphi_X(s)$、$\varphi_Y(s)$ とおく（ここで $p, q \in \mathbb{N}$）。また重み付き測度（重み関数（正）かけるルベーグ測度等）を $\mu(s,t)$ とおく。すると「距離の相関係数」は

$$\int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)|^2 \mu(ds, dt)$$

と重み付けした L_2 距離で定義される。この形を見れば統計量は独立ならばその時に限り 0 となることがわかる。特に Székely,G.j. らは具体的に重み関数を $\mu(ds, dt) = c_{p,q}|s|^{-\alpha-p}|t|^{-\alpha-q} ds\, dt$ とおくことで統計量を

$$\begin{aligned} T(X, Y; \mu) \;&=\; E[|X - X'|^\alpha\, |Y - Y'|^\alpha] + E[|X - X'|^\alpha]\, E[Y - Y'|^\alpha] \\ &\quad -2\, E[|X - X'|^\alpha\, |Y - Y''|^\alpha] \end{aligned}$$

と陽な形に表した。ここで (X', Y') は (X, Y) の iid コピーで、Y'' は全ての変数と独立な Y のコピーである。

本研究はこの「距離の相関係数」を 1 次元と多次元の定常な時系列へ応用するものである。主結果は論文 [Davis et al.(2018)] にまとめてある。主な目的は

１．2 つの時系列が独立かどうかを検定する、あるいは「距離の相関」で 2 つの時系列の相互依存関係を測る。

２．タイムラグをとった系列の「距離の相関」をみることで、１つの時系列が系列相関を持つかどうか検定する。

ことである。時系列解析では、多次元の系列相関を測る指標として（多次元の）自己相関関数が一般的に知られている。ここで言う相関はピアソンの積率相関係数である。非常に便利な指標であるが、非線形時系列モデルの依存関係を捉えるのが難

しいことが知られている。非線形時系列は計量ファイナンス分野において近年よく研究されていて、例えばその代表例として $GARCH(p,q)$ モデルが挙げられる。このモデルは、単に原系列の自己相関をみても依存関係を検出できず、原系列を変換したも（例えば絶対値や 2 乗を考えたもの）の自己相関をみて初めて依存関係が捉えられる。

本研究のアイデアは、そのピアソン流の相関を「距離の相関係数」に置き換えることで依存関係をより良く検出しようというものである。論文では新しく「自己距離の相関関数」という指標を自己相関関数に代わりうるものとして提案した。そして、ミキシング条件 (strong mixing) のもとで、標本「自己距離の相関関数」に関連する漸近論を導出した。まず、緩やかモーメント条件のもとで統計量の一致性を示した。その後いくつか追加的な条件を与え、依存関係がある場合と独立な場合の両方で統計量の漸近分布を理論的に導出した。独立な場合とそうでない場合は収束のオーダーが異なることも示した。検定統計量は特に独立な場合は、独立なカイ 2 乗変数の線形無限和で表され、一般に上側確率を求めるのが難しいことが知られている。これに対し、ブートスラップ法を用いた数値実験で漸近分布がうまく近似できることも示した。

2 つ目の研究である 1 系列における「自己距離の相関関数」に関しては、具体的なモデルとして $AR(p)$ モデルを考えた。モデルの適合度をみるために、パラメータ推定後に得られる残差に「自己距離の相関関数」を適応した。すると興味深いことに、iid 系列に「自己距離の相関関数」を適応した場合と比較して明らかに異なることもわかった（これは自己相関関数と同様の現象である)。その他、誤差項に裾の重い分布を仮定した場合は Székely, G.j. のオリジナルな定義では漸近論がうまく機能しないこともわかった。パラメータの推定も考えたもとでは、モーメント制約がより厳しくなるためと考えられる。我々が提案した他の重み関数によるものは、裾の厚い誤差項に関してもうまく対応することも確認できた。

実証研究も行い、アマゾンの株価収益率データや風速データなどで提案した方法の応用を試みた。

キーワード：Auto- and cross-distance correlation function, testing independence, strong mixing, ergodicity, Fourier analysis, U-statistics, AR process, residuals

参考文献

[Davis et al.(2018)] Davis, R.A., Matsui, M., Mikosch, T. and Wan, P. (2018) Applications of distance correlation to time series. *Bernoulli* **24**, 3087–3116.

[Matsui et al.(2017)] Matsui, M., Mikosch, T. and Samorodnitsky G. (2017) Distance covariance for stochastic processes. *Probab. Math. Statist.* **37**, 355–372.

[Székely et al.(2007)] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

[Székely and Rizzo(2009)] Székely, G.J. and Rizzo, M.L. (2009) Brownian distance covariance. *Ann. Appl. Stat.* **3**, 1236–1265.

割引因子ポアソン-ガンマ状態空間モデルによる計数時系列データの逐次分析

入江　薫

科研費シンポジウム「融合する統計科学」において、二日目にあたる十二月一日の午前のセッションにて、「割引因子ポアソン-ガンマ状態空間モデルによる計数時系列データの逐次分析」という演題で研究成果を発表した。

発表は予稿の内容に沿って行い、公表済みの論文（Chen et al., 2018）および進行中のプロジェクト（Irie et al., 2018）の二つの研究について発表した。シンポジウムの主旨を踏まえ、提案するモデルや計算手法の詳細は省き、主に使用しているデータについて詳しく紹介した。具体的にはニュースウェブサイトへのアクセス数のデータであるが、(1)これは非負の整数値を取る多変量の時系列データでありモデリングに工夫を要すること、(2)また、高頻度で（毎 30 秒ごとに）次々と観測されるために事後・予測分析を短時間で終える必要があり、この点でもモデリングおよび計算手法に工夫を要すること、これらの統計学上の問題意識に加えて、(3)応用上でも同種のデータ分析の需要が高まっており、主にマーケティングの問題として産業においても関心が高まっていることを述べた。提案モデル・手法の解説においては、実際のデータに適用した結果を主に事後・予測分布を図示することで説明した。

当日の質疑応答では、主にデータの構造について確認の質問を多く受けた。たとえば、ウェブサイトへのアクセス数をどう定義するかという点については一定の恣意性があり、ここでは既定の時間内に同じ IP アドレスからの連続したアクセスがあった場合に、そのドメイン間のフローにカウントすると説明した。そのため、ドメイン A からドメイン A へのフロー（同一ドメイン内のフロー）も存在することや、対象のウェブサイト外への（からの）アクセスも定義できることを指摘した。

また、技術的な点については大きく省略したものの、モデルの一部には Two-way ANOVA というよく知られた統計モデルと同様の構造があり、数理統計学を専門とする先生方から数々の指摘があった。特に、ANOVA のパラメータの解釈は識別のための仮定に大きく依存することや、個別効果や独自効果の平均に制約を入れる場合には外れ値に敏感になる問題が議論された。これらの点については共同研究者とともに認識していたが、応用の研究として普段発表する際にはあまり質問されないため、今回の議論は非常に有意義であった。

発表内容のうち半分は未発表のものであるため、今回得られたコメントをもとに論文の執筆を急ぎ、早くに公表することを目指したい。

Testing for changes in income inequality in Japan

Haruhisa Nishino
Graduate School of Social Sciences, Hiroshima University.*

A lot of attention was paid to Japanese income inequality around 2000. Tachibanaki (2005) and Ohtake (2008) are famous works in this field. These works were originally published in Japanese. The English translations were published later. They agree that the Japanese income inequality has risen constantly from the 1970s to 2000s. In recent years Piketty (2014) suggested that the income inequalities in developed countries including Japan became wider using tax statistics. This paper investigates how much the change in Japanese income inequality after 2000 is.

Ohtake and Saito (1998) argue that inequality in the 1980s and 1990s can be explained mainly by population ageing. The baby boomer generation in Japan, who was born in 1947-49, became older in the 1980s and 1990s when the proportion of older workers was larger. Consequently, the income inequality was getting wider, because some of older workers are at senior positions and richer; other workers are not rich.

Moriguchi and Saez (2008) and World Inequality Database indicate that top income shares are increasing in Japan, the U.S., and other developed countries after the 1980s. Pickety (2014) also uses top income share statistics.

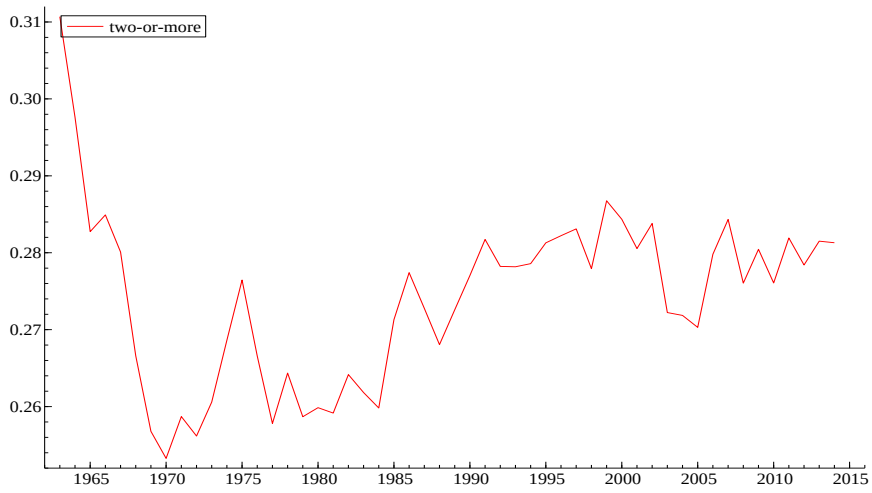

Figure 1: Gini coefficients: Two-or-more-persons households (1963–2014)

*1-2-1 Kagamiyama, HigashiHiroshima, Hiroshima 739-8525 Japan, Email: hnishino@hiroshima-u.ac.jp

As an example, Figure 1 gives Gini coefficients based on Family Income and Expenditure Survey (Two-or-more-persons households) in Japan: (1963–2014). It indicates that inequality in Japan does not seem to be increasing since 2000.

This paper proposes two test statistics for testing change in income inequality by using grouped income data in Family Income and Expenditure Survey in Japan. One is a parametric test, which assumes lognormal distribution and follows Nishino and Kakamu (2011). The other is a nonparametric test proposed by Beach and Davidson (1983). The nonparametric test does not need to assume a parametric distribution, which is useful. The nonparametric test is available only for decile data while the parametric one for quintile data also. That is, more various datasets are available for the parametric test than for nonparametric one. In addition, the nonparametric test needs calculating local variances for each group, which is usually unavailable. In this paper, local variances are extrapolated by assuming a parametric model.

These tests lead to the following interpretation that the inequalities at 1974, 1986, 1998 become wider statistically significantly for two-or-more persons households and for workers' households in common. It indicates that inequality is increasing largely in the 1980s and 1990s. It cannot be judged that income inequality is increasing since the 2000s.

References

[1] Beach, C. M. and Davidson, R. (1983) "Distribution-Free Statistical Inference with Lorenz Curves and Income Shares", *Review of Economic Studies*, **50**, 723–735.

[2] Moriguchi, C. and Saez, E. (2008) "The Evolution of Income Concentration in Japan, 1886-2005: Evidence from Income Tax Statistics" *The Review of Economics and Statistics*, **90**, 713–734.

[3] Nishino, H. and Kakamu, K. (2011) "Grouped data estimation and testing of Gini coefficients using lognormal distributions," *Sankhya B*, **73**, 193–210.

[4] Ohtake, F. (2008) "Inequality in Japan", *Asian Economic Policy Review*, **3**(1), 87–109.

[5] Ohtake, F and Saito, M. (1998) "Population aging and consumption inequality in Japan," *The Review of Income and Wealth*, **44**, 361–381.

[6] Piketty, T. (2014) *Capital in the Twenty-First Century*, translated by A. Goldhammer, Cambridge: Harvard University Press.

[7] Tachibanaki, T. (2005) *Confronting Income Inequality in Japan: A Comparative Analysis of Causes, Consequences, and Reform*, Cambridge: The MIT Press.

[8] World Inequality Database: *http://wid.world/* (accessed 25 September 2018).

Particle Filtering for Non-linear State-Space Models for Wind Speeds and Directions

Naoya Hieda[a], Takayuki Shiohama[b,*]

[a] *Graduate School of Engineering, Tokyo University of Science*
6-3-1 Niijuku, Katsushika, Tokyo, 125-8585 JAPAN
[b] *Department of Information and Computer Technology*
Tokyo University of Science
6-3-1 Niijuku, Katsushika, Tokyo, 125-8585 JAPAN

Abstract

The state space form is a useful framework for estimating unobserved state variables from some given observations. The applications can be found in diverse areas of natural science and engineering such as ecology, epidemiology, meteorology and economics and finance. The wind speeds and directions have complex time series probability structures involving highly non-Gaussian and nonlinear transition. In this study, we consider a simulation-based inference using the sequential Monte Carlo methods for computing the posterior distributions for the state variables given all available observations. We propose an alternative approach that allows us to extend the methods of importance sampling distributions incorporating with the class of circular Markov transition densities. The resulting methods are compared with various resampling schemes with real data applications.

Keywords: circular data, EM-algorithm, state-space model, particle filter

1. Introduction

Circular (or directional) data refer to data recorded as points for which directions are measured, typically in the fields of biology, geography, medicine, and astronomy. For such data, which are usually expressed in terms of compass angles or pairs of sine and cosine variables, the beginning and end of the scale in the domain coincide. Owing to this periodicity, analyzing circular data is challenging because traditional statistics are not meaningful, and may even be misleading when the particular definition of the domain is ignored. Recent developments in circular data analysis using the statistical computing software, R, are summarized in Pewsey et al. (2013). Although most circular data are in the form of time series, little research has been carried out in the field of circular time series analysis compared with the number of circular time series modeling approaches.

In general, three main approaches are used to model circular time series. The first method is used to obtain circular-valued random variables by wrapping; one example is the wrapped autoregressive process of Breckling (2012). The second approach is based on a link function

*Corresponding author
Email addresses: 4417621@ed.tus.ac.jp (Naoya Hieda), shiohama@rs.tus.ac.jp (Takayuki Shiohama)

that maps a line onto a circular domain, called a linked autoregressive moving average process. This model was proposed by Fisher & Lee (1994). The last approach specifies the density of the conditional distribution, including the Markov process of Wehrly & Johnson (1980), Möbius transformation of Kato (2010), and hidden Markov models of Holzmann et al. (2006). Abe et al. (2017) studied thecircular Markov process of Wehrly & Johnson (1980) and obtained theoretical circular autocorrelation structures under simple model assumptions. According to their results, circular autocorrelations are determined by the mean resultant length of the underlying circular density of the process. Abe et al. (2018) considered the circular Markov processes whose concentration parameter could be time-varying.

Many data in directional time series applications display nonlinear features such as heteroskedasticity and a nonlinear relationship between wind direction and speed. These features become more and more relevant as the length of the observed time series increases and as the series itself is subject to changes in the dynamic structure. In this paper, we address the circular process of Wehrly & Johnson (1980), which allows time-varying concentration parameters. The proposed model can incorporate the time-varying autocorrelations of the observed circular time series. For this purpose, we introduce a simple nonparametric regression model to the model parameter with time-varying observed exogenous variables, which cause a reasonable fit of the observed time series. The proposed models are then used to illustrate how wind direction and speed are related to the time-varying parameters. For further detail on the time series analysis of wind direction, see, for example, Breckling (2012), Ailliot et al. (2006), and Fuentes et al. (2005).

In an applications in meteorology, bivariate data consists of wind speeds and directions are often modeled by using projected normal distributions (Mardia & Jupp (2009)). However, time series modeling for such a bivariate dataset is not sufficiently studied in the literature. The dataset of wind speeds and directions are called cylindrical data because circular wind direction data are observed along with linear wind speeds ones. Lagona et al. (2015) proposed a hidden Markov model for analyzing cylindrical time series, and the proposed model can adequately explain circular–linear correlation, and temporal autocorrelation of the observed data. The state space modeling using circular random variable is considered in Mazumder & Bhattacharya (2017) and Kurz et al. (2016). In this study, we extend existing circular state-space models to cope with cylindrical time series.

Sequential Monte Carlo (SMC) methods are the set of simulation-based methods which provide a convenient and attractive approach to computing posterior distributions. Over the last few years, there has been a proliferation of scientific papers on SMC methods and their applications. Several closely related algorithms, under the names of bootstrap filters, condensation, particle filters, Monte Carlo filters, interacting particle approximations and survival of the fittest, have appeared in several research fields. In general, the parameter estimation in Gaussian and linear state-space model can be done by usual maximum likelihood estimation (MLE). However, the EM algorithm is used for estimating model parameters in the state space model, and it turns out be more robust than the direct solution of the MLE. The recent development of EM algorithms in SMC methods are explained in Kantas et al. (2015).

The Dantzig selector for a linear model of diffusion processes

Kou Fujimori
Waseda University

Let us consider the following model given by the linear stochastic differential equation:

$$X_t = X_0 + \int_0^t \Theta^\top \phi(X_s) ds + \sigma W_t, \tag{1}$$

where $\{W_t\}_{t\geq 0} := \{(W_t^1, \dots, W_t^p)\}_{t\geq 0}$ is a p-dimensional standard Brownian motion, Θ is a $p \times p$ sparse deterministic matrix, $\sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_p)$ is a $p \times p$ diagonal matrix and $\phi(x) = (\phi_1(x_1), \dots, \phi_p(x_p))^\top$ for $x = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$ is a smooth $\mathbb{R}^p$-valued function. We will propose some estimators for the true values (Θ^0, σ^0) of (Θ, σ) based on the observation of $\{X_t\}_{t\geq 0}$ at $n+1$ equidistant time points $0 =: t_0^n < t_1^n < \dots < t_n^n$, under the high-dimensional and sparse setting, *i.e.*, $p \gg n$ and the number of nonzero components of the true value Θ^0 is relatively small.

To deal with high-dimensional and sparse parameters, various kinds of estimators for regression models have been discussed. One of the most famous estimation methods is the l_1-penalized method called Lasso proposed originally by Tibshirani (1996), which has been studied for regression models with high-dimensional and sparse parameters in various models including the ones of stochastic processes.

On the other hand, a relatively new estimation procedure called the Dantzig selector was proposed for linear regression models by Candés and Tao (2007) as follows.

$$\hat{\beta}_D := \arg\min_{\beta \in \mathcal{C}} \|\beta\|_1, \quad \mathcal{C} := \left\{ \beta \in \mathbb{R}^p : \sup_{1\leq j\leq p} |Z^{j\top}(Y - Z\beta)| \leq \lambda \right\},$$

where $\lambda \geq 0$ is a tuning parameter. When $\lambda = 0$, the Dantzig selector coincides with the classical estimators such as the LSE in general cases and the MLE in Gaussian noise cases. For $\lambda > 0$, the Dantzig selector searches for the sparsest β within the given distance of the classical estimators. The Dantzig selector has been studied well especially for *i.i.d.* models. For example, Bickel et al. (2009) showed that the Dantzig selector has some properties similar to Lasso estimator for linear regression models in the sense of the consistency. In addition, as well as Lasso, the Dantzig selector has variable selection consistency for some regression models. Fan et al. (2016) showed the varable selection consistency of

the Dantzig selector for general single index models by using the irrepresentable conditions which are obtained from the KKT condition of the optimization problem. The Dantzig selector also has a good potential to be applied for other models including the models of stochastic processes. For instance, Antoniadis et al. (2010) applied this method to estimate regression parameter for Cox's proportional hazards model and proved the obtained estimator has the consistency. Fujimori (2017) studied the variable selection consistency of the Dantzig selector for the proportional hazards model and construct asymptotically normal estimators for the regression parameter and the cumulative baseline hazard function. Moreover, it is well-known that the Dantzig selector for linear models has computational advantages since it can be solved by a linear programming, while Lasso demands a convex program.

This talk dealt with the Dantzig selector for the linear models of stochastic processes (1) to estimate the drift matrix Θ^0 and prove the consistency in the sense of l_q norm for every $q \in [1, \infty]$ and the variable selection consistency under some appropriate conditions. Moreover, using the variable selection consistency, construction of a new estimator which has an asymptotic normality was discussed. We can prove the consistency of the Dantzig selector by the standard way which is similar to Bickel et al. (2009). However, since dealing with the KKT conditions of the Dantzig selector for our model is more difficult due to the complicated structure of the model than those of *i.i.d.* models, it may be hard to obtain the same results as Fan et al. (2016) concerning the variable selection consistency. Therefore, the another type of variable selection consistency by using a thresholding method was proposed in this talk. This talk is based on the paper Fujimori (2018)

References

Antoniadis, A., Fryzlewicz, P. and Letué, F. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37**, no.4, 531-552. (2010).

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, no. 4, 1705-1732. (2009).

Candés, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, no.6, 2313-2351. (2007).

Fan, Y., Gai, Y. and Zhu, L. Asymtotics of Dantzig selector for a general single-index model. *J. Syst. Sci. Complex.* **29**, no.4, 1123-1144. (2016).

Fujimori, K. Cox's proportional hazards model with a high-dimensional and sparse regression parameter. arXiv:1710.10416[math.ST]. (2017).

Fujimori, K. The Dantzig selector for a linear model of diffusion processes. *To appear in Stat. Inference Stoch. Process.* (2018).

Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, no.1, 267-288. (1996).

多様な関数を用いた経時変動の罰則付推定法

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, 多くの分野で収集され分析がされている. この経時測定データの分析では, 経時変動と呼ばれるデータに潜む変動を上手く捉えることが一つの目的である. 本講演では, i 番目の個体における第 j 時点 t_j での測定値を $y_i(t_j)$ と表し, $t_1, \ldots, t_p$ は全ての個体で共通とした.

このとき, 未知の q 個の係数 $m_1, \ldots, m_q$ を用いて $m_1 + m_2 t_j + \cdots + m_q t_j^{q-1}$ とすると, 経時変動を測定時点の $(q-1)$ 次の多項式で推定できる. ここで $\boldsymbol{m} = (m_1, \ldots, m_q)'$ とすると, この項は $(1, t_j, \ldots, t_j^{q-1})\boldsymbol{m}$ となる. この項を平均的な経時変動を表す項とした. 一方で, 各個体の性別などからなる k 個の説明変数の各変数に対しても同様に $(q-1)$ 次の多項式で経時変動を推定することを考えると, 未知の係数 ξ_{ij} を用いて $\left(\sum_{i=1}^{q} \xi_{1i} t_j^{i-1}, \ldots, \sum_{i=1}^{q} \xi_{ki} t_j^{i-1}\right)$ の形でモデル化できた. ここで, (i,j) 成分が ξ_{ij} からなる $k \times q$ 未知行列を $\boldsymbol{\Xi}$ とすると, この項は $(1, t_j, \ldots, t_j^{q-1})\boldsymbol{\Xi}'$ となり, これが i 番目の個体の説明変数 $\boldsymbol{a}_i$ (k 次元ベクトル) に関する経時変動を表現する項とした. この二つの項の表現より, $\boldsymbol{X}$ の j 行目を $(1, t_j, \ldots, t_j^{q-1})$ とし, i 番目の個体の経時測定データ $\boldsymbol{y}_i = (y_i(t_1), \ldots, y_i(t_p))'$ の経時変動全体を測定時点の $(q-1)$ 次多項式で捉えられるとすると, 誤差ベクトル $\boldsymbol{\varepsilon}_i$ を用いて, 各個体に対して $\boldsymbol{y}_i' = \boldsymbol{X}\boldsymbol{m} + \boldsymbol{X}\boldsymbol{\Xi}'\boldsymbol{a}_i + \boldsymbol{\varepsilon}_i$ とモデル化できた. ここで, $\boldsymbol{\varepsilon}_i$ は $E[\boldsymbol{\varepsilon}_i] = \boldsymbol{0}_p$ (p 次元ゼロベクトル), $\mathrm{Cov}[\boldsymbol{\varepsilon}_i] = \boldsymbol{\Sigma}$ ($p \times p$ 未知正定値行列), $\boldsymbol{\varepsilon}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_j$ $(j \neq i)$ とした. さらにここで, n を個体数, $\boldsymbol{1}_n$ を全成分が 1 の n 次元ベクトル, i 行目が i 番目の経時測定データ $\boldsymbol{y}_i'$ からなる $n \times p$ 既知行列 $\boldsymbol{Y}$, i 行目が i 番目の個体の説明変数 $\boldsymbol{a}_i'$ からなる $n \times k$ 既知行列 $\boldsymbol{A}$ (ただし $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$), i 行目が $\boldsymbol{\varepsilon}_i'$ からなる $n \times p$ 誤差行列 $\boldsymbol{\mathcal{E}}$ を用いると, 前述の各個体に対するモデルは次の形でまとめることができた;

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

このモデルで q 次元未知ベクトル $\boldsymbol{m}$ と $k \times q$ 未知行列 $\boldsymbol{\Xi}$ を推定することで, $\boldsymbol{Y}$ の経時変動全体を測定時点 $t_1, \ldots, t_p$ の $(q-1)$ 次多項式で捉えることができる. また, このモデルは Pothoff and Roy (1964) により提案された GMANOVA (一般化多変量分散分析) モデルと同じ形である. さらに, $\boldsymbol{X}$ の (i,j) 成分を t_j に関する柔軟な関数に置き換えることで, それらの関数の重み付和により経時変動全体を捉えることも考えられる. このとき, 従来の推定法では $\boldsymbol{Y}$ へ過剰に適合してしまうが, これを回避する推定法を Nagai (2011) で提案した.

GMANOVA モデルを用いた場合, $\boldsymbol{Y}$ の経時変動全体を推定する際に用いる関数は, 平均的な経時変動の項 $(= \boldsymbol{1}_n \boldsymbol{m}' \boldsymbol{X}')$ も 各説明変数に関する経時変動の項 $(= \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}')$ も同じ関数系 (例えば, 同じ次数の多項式) を用いた形に限られる. そこで本講演では, 平均や説明変数により異なる関数系を用いた推定を考えた. つまり例えば, 平均的な経時変動は $(q_0 - 1)$ 次多項式で, いくつかの説明変数 $\boldsymbol{A}_1$ に関する経時変動は $(q_1 - 1)$ 次多項式で, 残りの説明変数 $\boldsymbol{A}_2$ に関する経時変動は $(q_2 - 1)$ 次多項式で推定することを考えた. このとき, 上記の GMANOVA モデルと同様に考えると, 次のような Extended GMANOVA モデル (Kollo & von Rosen (2005) Definition 4.1.3 を少し変形したモデル) による推定を考えることになった;

$$\boldsymbol{Y} = \boldsymbol{1}_n \boldsymbol{\mu}' \boldsymbol{X}_0' + \boldsymbol{A}_1 \boldsymbol{\Xi}_1 \boldsymbol{X}_1' + \boldsymbol{A}_2 \boldsymbol{\Xi}_2 \boldsymbol{X}_2' + \boldsymbol{\mathcal{E}},$$

ここで $\boldsymbol{\mu}$ は q_0 次元未知ベクトル, $\boldsymbol{X}_i$ は j 行目が $(1, t_j, \ldots, t_j^{q_i - 1})$ などからなる $p \times q_i$ 既知行列 $(i = 0, 1, 2;\ \mathrm{rank}(\boldsymbol{X}_i) = q_i \leq p)$, $\boldsymbol{A}_i$ は $n \times k_i$ 既知行列 $(i = 1, 2;\ \mathrm{rank}(\boldsymbol{A}_i) = k_i \leq n$, ただし $\boldsymbol{A}_i'\boldsymbol{1}_n = \boldsymbol{0}_{k_i})$, $\boldsymbol{\Xi}_i$ は $k_i \times q_i$ 未知行列 $(i = 1, 2)$, $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ は $n \times p$ 誤差行列であ

る. また, 元の Kollo and von Rosen (2005) での定義や推定などでは $\mathbf{1}_n$, $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ や $\boldsymbol{X}_0$, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ に対応する部分などに階層構造を仮定しているが, 本講演ではこの構造を仮定しなかった. この $\boldsymbol{\mu}$, $\boldsymbol{\Xi}_1$, $\boldsymbol{\Xi}_2$ を推定することで, 各項に対して $(q_i - 1)$ 次多項式などを用いて経時変動全体が推定できる. 例えば, 平均的な経時変動は 3 次の, 性別に関する経時変動は 2 次の, 他の変数に関する経時変動は 4 次の多項式を用いた形で経時変動全体の推定ができる.

しかしながら, 経時変動全体が複雑な場合, 測定時点の多項式を複数組み合わせて推定しても経時変動全体を捉えることができない場合がある. そこで, 多項式だけを用いるのではなく, Nagai (2011) と同様に柔軟な関数を用いて経時変動全体を推定することが考えられる. そこで本講演では, 上記のモデルにおいて, $\boldsymbol{X}_0$ や $\boldsymbol{X}_1$ を測定時点の $(q_0 - 1)$ 次および $(q_1 - 1)$ 次の多項式を用いる項, $\boldsymbol{X}_2$ を柔軟な関数を用いる項として考えた. これにより, 平均的な経時変動やいくつかの説明変数 $\boldsymbol{A}_1$ に対しては測定時点の多項式を用い, 別の説明変数 $\boldsymbol{A}_2$ に対しては柔軟な関数を用いて経時変動が推定できる. しかし, 多項式の部分を単に柔軟な関数に置き換えて従来の推定法で未知の $\boldsymbol{\mu}$ や $\boldsymbol{\Xi}_1$, $\boldsymbol{\Xi}_2$ を推定すると, 目的である経時変動全体ではなく経時測定データ $\boldsymbol{Y}$ へ過剰に適合してしまうという問題が起きる. 本講演では, Nagai (2011) と同様に多変量一般化リッジ回帰による推定法 (Yanagihara, Nagai and Satoh, 2009) を用いて, この問題を回避することを考えた.

しかし, 従来の推定法では $\boldsymbol{\Xi}_i$ の推定量を得るために $\boldsymbol{\Xi}_j$ $(i \neq j)$ が必要となるため, 柔軟な関数を用いて推定する項 $(= \boldsymbol{A}_2\boldsymbol{\Xi}_2\boldsymbol{X}_2')$ だけに着目し, $\boldsymbol{\Xi}_2$ の推定量の部分にのみ罰則を付けることが難しい. そこでまず本講演では, vec 作用素などの性質 (例えば Lütkepohl (1996) 参照) を用いて, $\boldsymbol{\Xi}_1$ の推定量を用いずに $\boldsymbol{\Xi}_2$ の推定量を得る手法を提案した. その結果, Nagai (2011) の元となった Hoerl and Kennard (1970) により提案された手法を用いることで, 柔軟な関数を用いる部分の推定量に罰則を付けることが可能であることを示した.

次に, 過剰適合を回避するために導入した罰則パラメータの最適化について考えた. そのために, Mallows (1973) や Yanagihara and Satoh (2010) などと同様の手法で, C_p 型情報量規準を構築し, さらに $n - k_1 - k_2 - p - 2 > 0$ と $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} N_p(\mathbf{0}_p, \boldsymbol{\Sigma})$ を仮定した下での C_p 型情報量規準のバイアスを補正した規準も構築した. そして最後に, これらの情報量規準を工夫して展開することで情報量規準を最小にする罰則パラメータが陽に得られることを示した. 数値実験による比較などについては当日報告した.

引用文献:

[1] Hoerl, A. E. & Kennard R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

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

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

[4] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

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

[7] 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).

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

傾向性仮説と変化点モデルの様々な応用

広津千尋（明星大学連携研究センター）

1. 序論

用量反応解析においては，通常，厳密な反応曲線を想定することが難しいために単調性，凸性，Ｓ字性のような形状制約がよく設定される．それらのうち正規分布モデルに対する単調仮説については，isotonic regression がよく知られている．しかしながら，それは Bartholomew によってやや直観的に導入され，とくにこのような制約のある母数空間に対する最適性は自明ではない．さらに制約付き最小二乗法は，計算及び分布論が複雑で，凸性，Ｓ字性問題，正規分布以外の確率モデル，あるいは 2 元配置交互作用問題への拡張には困難が伴う．一方，著者等のアプローチは Hirotsu (1982)で導かれた一般制約仮説に対する検定の完全類を基にしており，その意味での最適性を持っている．それによると単調性，凸性，Ｓ字性仮説それぞれに対し，単純, 2 重, 3 重累積和に基づく単調増大な統計量が示唆される．本稿ではそのうち規準化最大対比，および規準化二乗和を用いる方法について論ずる．

一方，これらの形状制約は変化点モデルと密接な関係がある．すなわち，単調性，凸性，Ｓ字性仮説はそれぞれ，段差変化点，スロープ変化点，変曲点モデルと対応する(Hirotsu and Marumo, 2002)．例えば，段差変化点モデルを表す対比は単調対比の典型であり，逆にすべての単調対比は段差変化点対比の一意正係数線形結合で表される．例えばＰＭＤＡでは副作用自発報告が収集され，その経年変化が解析されている.その場合，単調増加傾向をいち早く検出すると同時に，増加傾向の生じた時点を推測することは応用上も大変有意義である．同様のことが凸性，Ｓ字性問題についても示される．最大対比統計量は，これら変化点モデルに対する efficient score 検定を与える．すなわち，従来統計学の二つの異なる流れの中で研究されてきた制約仮説と変化点問題を統合的に扱う事が出来る．

以上のように，累積和に基づく方法は様々な確率モデル，様々な実質科学上の問題を理論的に一貫した方法で扱い，効率の良い計算アルゴリズムを与えることが出来る．正規分布，2 項分布，Poisson 分布についてはこれまでいろいろな機会に発表しているので，本論では最近の研究である独立な2 × 2表の系列への応用について述べる．重要な例としてケース・コントロール研究がある．

2. 問題の概要

二つの 2 項分布母集団を K層上で比較する問題を考える．オッズ比の一様性が仮定出来る場合は例えば Gart (1970)により共通オッズ比に関する推論を行えば良い．一様性検定については最尤法や，Breslow & Day 検定が良く知られている他，Zelen (1971)の正確検定も利用出来る．正確推論に関しては Agresti (1992)のサーベイが参考になる．本論はとくにK層に自然な順序がある場合に，単調性，および凸性を対立仮説とする検定について論ずる．これらは上記の論文では論じられていないが，単調性に対する最大対比検定は Hirotsu 他(2001)の特殊ケースに当たり，その特殊ケースはまた太田他(2003)で詳しく論じられている．

本論ではまずこの検定について，基礎統計量の規準化に用いられている漸近分散を正確分散に置き換える改良を行い，Fleiss 他(2003)のロジット線形モデルによる交互作用解析と比較する．その結果は良く整合する．次にオッズ比に関する凸性仮説検定の定式化を行い，新たに最大対比正確検定を提案する．これら二つの問題について累積χ^2乗検定も与える．

3. 定式化

第K層のデータをy_{ijk}，その確率を$p_{ijk}(p_{i\cdot k}=1)$で表す，ただし，$i=1,2$ は二つの 2 項分布母集団，$j=1,2$は成否の反応，そして$k=1,\cdots,K$は層を表す．この時，y_{ijk}を並べたベクトルを$\boldsymbol{y}$として，確率分布は$g(\boldsymbol{y})=\prod_{k=1}^{K}\prod_{i=1}^{2}\left\{y_{i\cdot k}!\prod_{j=1}^{2}\left(p_{ijk}^{y_{ijk}}/y_{ijk}!\right)\right\}$で与えられる．ここで 対数線形模型：$\log p_{ijk}=\gamma_k+\varphi_{ik}+\tau_{jk}+\omega_{ijk}$を想定した上，周辺和$y_{i\cdot k}$および$y_{\cdot jk}$を与えた相似検定を考えると，条件付き確率分布は$g(\boldsymbol{y},\boldsymbol{\omega})=\prod_{k=1}^{K}C^{-1}(\omega_k)b(y_{11k})\exp(\omega_k y_{11k})$という簡単な形になる．ただし，$\boldsymbol{y}$は改めて$y_{11k}$を並べたベクトルとし，$\omega_k$が推論対象である対数オッズ比である．この先，$\omega_k$に関する単調性，および凸性検定は 2 項分布や，Poisson 分布等の 1 母数指数分布族に対する方法(Hirotsu，2017)にならい，カーネルおよび漸化式遂行の不等式の変更を適切に行うことにより実行出来る．

参考文献

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science* 7, 131-177.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical methods for rates and proportions.* Wiley Series in Probability and Statistics, New York.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2 x 2 tables with fixed marginals. *Biometrika* 57, 471-475.

Hirotsu C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* 69, 567-577.

Hirotsu, C. (2017). *Advanced analysis of variance*. Wiley Series in Probability and Statistics, New York.

Hirotsu, C. and Marumo, K. (2002). Change point analysis as a method for isotonic inference. *Scand. J. Statist.* 29, 125–138.

Hirotsu, C., Aoki, S., Inada, T. & Kitao, Y. (2001). An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis, *Biometrics* 57, 0769-778.

太田絵里，青木敏，広津千尋(2003). $2\times2\times K$分割表における単調仮説の検定．応用統計学 32，107－126.

Zelen, M. (1971). The analysis of several 2 x 2 contingency tables. *Biometrika* 58, 129-137.

Characterizations of indicator functions of fractional factorial designs

神戸大学大学院理学研究科　青木 敏

1. はじめに

実験計画法は計算代数統計における主要な研究対象の一つであり，中でも，Fontana, Pistone and Rogantin (2000, *JSPI*) により導入された指示関数 (indicator function) による一部実施計画の特徴付けは，古典的な接近法のひとつである．指示関数の構造は，2 水準の一部実施計画に関しては完全に解明されている．例えば，レギュラーな一部実施計画 D_1 の定義関係 $x_1x_2x_4 = x_1x_3x_5(= x_2x_3x_4x_5) = 1$ は，その指示関数 f_1 の各項に対応する．また，一部実施計画 D_2 の指示関数 f_2 の形は複雑であるが，定数項はこの計画の一部実施度 (3/8) を，1 次の項は因子 x_4 以外の水準数が同数であることを，2 次の項は因子 4 が他の 3 因子と直交することを，それぞれ表している．

D_1

x_1	x_2	x_3	x_4	x_5
1	1	1	1	1
1	1	−1	1	−1
1	−1	1	−1	1
1	−1	−1	−1	−1
−1	1	1	−1	−1
−1	1	−1	−1	1
−1	−1	1	1	−1
−1	−1	−1	1	1

D_2

x_1	x_2	x_3	x_4
1	1	1	1
1	1	−1	−1
1	−1	1	−1
−1	1	−1	−1
−1	−1	1	−1
−1	−1	−1	1

D_3

x_1	x_2	x_3
0	0	0
0	1	2
0	2	1
1	0	2
1	1	1
1	2	0
2	0	1
2	1	0
2	2	2

$$f_1(x_1,x_2,x_3,x_4,x_5) = \frac{1}{4} + \frac{1}{4}(x_1x_2x_4 + x_1x_3x_5 + x_2x_3x_4x_5)$$

$$f_2(x_1,x_2,x_3,x_4) = \frac{3}{8} - \frac{1}{8}x_4 + \frac{1}{8}(x_1x_2 + x_1x_3 - x_2x_3) + \frac{1}{8}(x_1x_3x_4 + x_2x_3x_4) + \frac{3}{8}x_2x_3x_4$$

このように，2 水準計画については，分解能や aberration の概念は，指示関数の係数と直接的な関係がある．一方で多水準計画に関しては，水準を有理数とする標準的な設定においては，多水準，あるいは水準数が因子ごとに不均一な場合などの一般的な計画に対する指示関数の構造は解明されていない．例えば計画 D_3 は，定義関係が $x_1+x_2+x_3 = 0 \pmod 3$ のレギュラーな一部実施計画であるが，その性質を指示関数 f_3 から読み取るのは容易ではない．

$$f_3(x_1,x_2,x_3) = 1 - \frac{3}{2}(x_1+x_2+x_3) + \frac{1}{2}(x_1^2+x_2^2+x_3^2) + \frac{1}{4}(x_1x_2 + x_1x_3 + x_2x_3)$$
$$+\frac{27}{2}x_1x_2x_3 + \frac{3}{4}(x_1^2x_2 + x_1^2x_3 + x_1x_2^2 + x_2^2x_3 + x_1x_3^2 + x_2x_3^2) - \frac{3}{4}(x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2)$$
$$-\frac{33}{4}(x_1^2x_2x_3 + x_1x_2^2x_3 + x_1x_2x_3^2) + \frac{9}{2}(x_1^2x_2^2x_3 + x_1^2x_2x_3^2 + x_1x_2^2x_3^2) - \frac{9}{4}x_1^2x_2^2x_3^2$$

2. 指示関数の性質

水準を有理数とする一般の一部実施計画について，その性質を指示関数の係数に結びつける方法を説明する．$x_1, \ldots, x_n$ を n 個の因子とする．$j = 1, \ldots, n$ について，因子 x_j の水準の集合を $A_j \subset \mathbb{Q}$ とおく．$\mathbb{Q}$ は有理数の集合である．A_j の要素数を $r_j = \#A_j$ とし，$r_j \geq 2$ と仮定する（$j = 1, \ldots, n$）．$D = A_1 \times \cdots \times A_n$ を，因子 $x_1, \ldots, x_n$ の完全実施計画とよび，その部分集合 $F \subset D$ を一部実施計画という．計画 F の指示関数は，$x_1, \ldots, x_n$ の多項式として一意的な表現

$$f(x_1, \ldots, x_n) = \sum_{\boldsymbol{a} \in L} \theta_{\boldsymbol{a}} \boldsymbol{x}^{\boldsymbol{a}}, \tag{1}$$

をもつ．ただし，$\boldsymbol{x}^{\boldsymbol{a}} = \prod_{j=1}^{n} x_j^{a_j}$,

$$L = \{\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{Z}_{\geq 0}^n \ : \ 0 \leq a_j \leq r_j - 1, \ j = 1, \ldots, n\}$$

であり，$\mathbb{Z}_{\geq 0}$ は非負整数の集合である．F の指示関数は，F 上で 1，$D \setminus F$ で 0 をとる観測値ベクトル $\boldsymbol{y}$ に対する D 上の補間多項式に他ならない．

多項式 $f \in \mathbb{Q}[x_1, \ldots, x_n]$ が，ある一部実施計画 $F \subset D$ の指示関数であることは，f と f^2 が D 上で交絡することと同値である．従って，式 (1) で表される f が，ある計画の指示関数であるとき，

$$\begin{aligned} \sum_{\boldsymbol{a} \in L} \theta_{\boldsymbol{a}} \boldsymbol{x}^{\boldsymbol{a}} &= \left(\sum_{\boldsymbol{a} \in L} \theta_{\boldsymbol{a}} \boldsymbol{x}^{\boldsymbol{a}} \right)^2 \mod I(D) \\ &= \sum_{\boldsymbol{a}_1 \in L} \sum_{\boldsymbol{a}_2 \in L} \theta_{\boldsymbol{a}_1} \theta_{\boldsymbol{a}_2} \boldsymbol{x}^{\boldsymbol{a}_1 + \boldsymbol{a}_2} \mod I(D) \end{aligned}$$

が成り立つ．ただし，$I(D)$ は，完全実施計画 D 上で 0 となる $\mathbb{Q}[x_1, \ldots, x_n]$ の多項式の集合で，D の計画イデアルとよばれる．$I(D)$ の被約グレブナー基底 G に関する $\sum_{\boldsymbol{a}_1 \in L} \sum_{\boldsymbol{a}_2 \in L} \theta_{\boldsymbol{a}_1} \theta_{\boldsymbol{a}_2} \boldsymbol{x}^{\boldsymbol{a}_1 + \boldsymbol{a}_2}$ の標準表示を $r = \sum_{\boldsymbol{a} \in L} \mu_{\boldsymbol{a}} \boldsymbol{x}^{\boldsymbol{a}}$ と書く．多項式 f が，ある計画の指示関数であることは，代数方程式系 $\theta_{\boldsymbol{a}} = \mu_{\boldsymbol{a}}, \ \boldsymbol{a} \in L$ を満足することと同値である．この代数方程式系に，サイズ，直交性などの条件に対応する制約を付け加える．このようにして構築された代数方程式系の解集合は，与えられた性質をもつすべての一部実施計画に対応する．それを同値類に分類することで，与えられた性質をもつ一部実施計画の分類を得ることができる．

3. $2^3 \times 3, 2^4 \times 3$ 計画の一部実施計画の分類

計算代数ソフトウェア Macaulay2 により，$2^3 \times 3$ 計画の一部実施計画のすべての直交計画，および，$2^4 \times 3$ 計画の一部実施計画のすべての強度 3 の直交計画を列挙し，同値類に分類した．いずれも，3 つの同値類に分類されることが分かった．詳細は，Aoki (2018), arXiv:1810.08417 を参照されたい．

Pooling incomplete samples による統計解析

関東学院大経済　布能 英一郎 (Eiichiro Funo)

1. Introduction

Asano(1965) は、On estimating multinomial probabilities by pooling incomplete samples という paper を発表した (Ann. Inst. Stat. Math., **17**, 1-17)。Pooling incomplete samples という名前は、これに由来する。後に、Johnson, Kotz, Balakrishnan(1997) は、分布論を網羅した書籍 *Discrete Multivariate Distributions* (Wiley) にて、第 35 章 Multinomial disributions の中で、Incomplete and modified multinomial disributions という節を設け (第 10 節, 71-72 ページ)、ここで Asano の成果を解説している。それによると

$\boldsymbol{X} = (X_1, X_2, \cdots, X_k)$ は、multinomial$(N_x; p_1, \cdots, p_k)$ に従い、$\boldsymbol{Y} = (Y_1, \cdots, Y_m\ (m < k)$ は、$\boldsymbol{X}$ とは独立で、multinomial$\Big(N_y; \frac{p_1}{\sum_{i=1}^m p_i}, \cdots, \frac{p_m}{\sum_{i=1}^m p_i}\Big)$ に従うものとする。このとき、p_i の最尤推定量 $\hat{p}_i$ は

$$\hat{p}_i = \frac{x_i + y_i}{N_x\Big(1 + \frac{N_y}{\sum_{j=1}^m x_j}\Big)} \quad (i \le m,) \qquad \hat{p}_i = \frac{x_i}{N_x} \quad (i > m)$$

である。

更に、次のことも知られている。

(i) $\hat{p}_i$ は p_i の minimum variance unbiased estimator である (Asano, 1965)

(ii) $\hat{p}_i$ は、自乗損失下で許容的 (布能, 2000)

(iii) Pooling incomplete samples による多項分布からの 2 標本問題において、Total information = Within information + Between information (以後、これを Kullback 情報量の直和分解と略記) が成り立つ。(布能, 2011)

注 Kullback(1959, 1968) は、2 標本問題において、仮説 H_1 を「2 つの母集団は同じ」、仮説 H_2 を「2 つの母集団は異なる」に設定したとき、Between Information を H_2 の下における「(最良) 推定量とパラメータの間の Kullback 距離」、Within Information を「H_1 の下での推定量と、仮説 H_2 の下での推定量の間の Kullback 距離」、Total Information を「H_2 の下での推定量と、H_1 の下でのパラメータ間の Kullback 距離」と定めた。そして、(通常の) 多項分布において、Kullback 情報量の直和分解 が成り立つことを示した。

2. その後の発展と課題

Pooling incomplete samples によるサンプリングは、多項分布の場合に限らず、一般の分布の下で考えることが出来る。「カテゴリーが減少する場合、減少後のカテゴリーには、元の生起確率を比例配分したものを割り当てる」ものであるから、離散分布にはこの考え方をそのまま当てはめられる。ところが、多項分布の場合に成り立っていることが、他の分布のとき、あるいは、多項分布の場合でも、Asano が設定した Pooling incomplete samples の枠組から外れると、成り立たないことが生じる。

2.1 unbiasedness Asano が示した方法は、Pooling incomplete samples を伴う他の分布に適用できる場合もあれば、うまくゆかない場合もある。

2.2 許容性の証明 Pooling incomplete samples を伴う多項分布の場合、Stepwise Bayes 法は、の許容性の証明法として「相性が良い」と言える。それは、Brown(1981) が「sample space が有限、母数空間がコンパクトのとき、任意の許容的推定量に対して、sequence of priors(あるいは、1 つの prior) が存在し、これに対する Stepwise Bayes 解 (あるいは、1 つの prior に対する Bayes 解) となっている」という定理を示しているからである。他方、Stepwise Bayes 法は、母数空間がコンパクトでないときには、必ずしもうまくゆくとは限らない。

2.3 2 標本問題における Kullback 情報量の直和分解 多項分布以外の Pooling incomplete samples では、成り立たない場合の方が多いと言わざるを得ない。

たとえば、Pooling incomplete samples を伴う負の多項分布の場合、「負の多項分布で自然に考えられる Pooling incomplete samples 」と思える確率モデルで計算したところ、Kullback 情報量の直和分解が成り立たかった。他方、負の多項分布 = 負の二項分布 × 多項分布 なので、この「多項分布」の部分で Pooling incomplete samples によるモデルを設定したところ、Kullback 情報量の直和分解が成立した。

更に、パラメータを分離して Kullback 情報量の直和分解が成立するような確率モデルの構築も行った。

今後の課題 カテゴリーが減少した際、「Kullback 情報量の直和分解が成立」するような Pooling incomplete samples 風の確率モデルを探すことも必要だが、Kullback 情報量の直和分解が成立しない原因を突き止めることも重要なことである。願わくば、Total information と Within information + Between information の差が、何らかの意味を持つ量であることを突き止められたらと思う。

文献 文中で十分な検索情報を与えたもの以外に、[1] 布能英一郎 (2000). Missing, Pooling を伴う離散サンプリングの統計的推測. 京都大学数理解析研究所講究録 No 1161, 141-158. [2] 布能英一郎 (2011). On estimating multivariate discrete probabilities by pooled incomplete samples and related topics. 京都大学数理解析研究所講究録 No 1758. [3] Kullback, S.(1968). *Information Theory and Statistics,* Revised edition. Dover.

Pitman's Closeness Domination in Predictive Density Estimation for Two Ordered Normal Means Under α-Divergence Loss

Yuan-Tsung Chang (Mejiro University) Nobuo Shinozaki (Keio University)
William, E. Strawderman (Rutgers University)

1 Introduction

We consider Pitman closeness domination in predictive density estimation problems when the underlying loss metric is α-divergence $\{D(\alpha)\}$, a loss introduced by Csiszàr (1967).The underlying distributions considered are normal, including the distribution of the observables, the distribution of the variable whose density is to be predicted, and the estimated predictive density which will be taken to be of the plug-in type. Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be two independent random normal variables, where $\mu_1 \leq \mu_2$. Under the above restriction we wish to predict a normal population with mean equal to the larger mean, μ_2, and variance equal to σ^2, $\tilde{Y} \sim N(\mu_2, \sigma^2)$. We consider different versions of this problem, depending on the $\sigma_i^2, i = 1, 2$ are unknown and/or satisfy the additional order restriction, $\sigma_1^2 \leq \sigma_2^2$.

The class of $\alpha-$divergence losses is given by

$$D_\alpha\{\hat{p}(\tilde{y}|y), p(\tilde{y}|\psi)\} = \int f_\alpha\Big(\frac{\hat{p}(\tilde{y}|y)}{p(\tilde{y}|\psi)}\Big)p(\tilde{y}|\psi)d\tilde{y}, \tag{1}$$

where, for $-1 \leq \alpha \leq 1$

$$f_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2}(1 - z^{(1+\alpha)/2}), & |\alpha| < 1 \\ z \log z, & \alpha = 1 \\ -\log z, & \alpha = -1. \end{cases} \tag{2}$$

Here KL loss corresponds to $\alpha = -1$, and $\alpha = 1$ is sometimes referred to as reverse KL loss.

Chang and Strawderman (2014, JMVA, Theorem 2.1) have derived the general form of D_α loss for the case of normal models and shown that it is a concave monotone function of quadratic loss and is also a function of the variances (observed, predicand, and plug-in).

An alternative criterion to evaluate the goodness of estimators was introduced by Pitman (1937) as follows:

Let T_1 and T_2 be two estimators of θ. Then T_1 is closer to θ than T_2 if Pitman nearness (PN) of T_1 compared to T_2

$$PN_\theta(T_1, T_2) = P\{|T_1 - \theta| < |T_2 - \theta|\} > 1/2.$$

For the case, when the estimators are equal with positive probability, Nayak (1990),Gupta and Singh (1992) defined the modified Pitman nearness (MPN) of T_1 compared to T_2. Setting $MPN_\theta(T_1, T_2) = P\{|T_1 - \theta| < |T_2 - \theta||T_1 \neq T_2\} = P\{|T_1 - \theta| < |T_2 - \theta|, T_1 \neq T_2\}/P\{T_1 \neq T_2\}$ T_1 is closer to θ than T_2 if $MPN_\theta(T_1, T_2) > 1/2$.

Here is a brief review of some of the relevant literature for the problem of estimating the mean. Let $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$, $s_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(n_i - 1)$ be the unbiased estimators of μ_i and σ_i^2, respectively, based on samples of size n_i from two normal populations, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ respectively. When the variances are unknown, the unbiased estimator

$$\hat{\mu}^{GD} = \frac{n_1 s_2^2}{n_1 s_2^2 + n_2 s_1^2}\bar{X}_1 + \frac{n_2 s_1^2}{n_1 s_2^2 + n_2 s_1^2}\bar{X}_2$$

was proposed by Graybill and Deal (1959)and they gave a necessary and sufficient condition on n_1 and n_2 for $\hat{\mu}^{GD}$ to have a smaller variance than both $\bar{X}_1$ and $\bar{X}_2$.

When estimating the ordered means, Oono and Shinozaki (2005, JSPI) proposed truncated estimators of $\mu_i, i = 1, 2$,

$$\hat{\mu}_1^{OS} = \min\{\bar{X}_1, \hat{\mu}^{GD}\}, \quad \hat{\mu}_2^{OS} = \max\{\bar{X}_2, \hat{\mu}^{GD}\}, \tag{3}$$

and showed that $\hat{\mu}_i^{OS}$ dominates the $\bar{X}_i$ in terms of MSE if and only if MSE of $\hat{\mu}^{GD}$ is not larger than that of $\bar{X}_i$ to estimate μ_i when $\mu_1 = \mu_2$.

When there are order restrictions given on both means and variances, $\mu_1 \leq \mu_2, \sigma_1^2 \leq \sigma_2^2$, Chang, Oono and Shinozaki (2012, JSPI)have proposed

$$\hat{\mu}_i^{CS} = \begin{cases} \hat{\mu}_i^{OS}, & if\ s_1^2 \leq s_2^2 \\ \min\left\{\bar{X}_1, \frac{n_1}{n_1+n_2}\bar{X}_1 + \frac{n_2}{n_1+n_2}\bar{X}_2\right\}, & if\ s_1^2 > s_2^2 \end{cases} \tag{4}$$

They show that $\hat{\mu}_2^{CS}$ stochastically dominates $\hat{\mu}_2^{OS}$, but $\hat{\mu}_1^{CS}$ cannot dominate $\hat{\mu}_1^{OS}$ even in term of MSE when $\mu_2 - \mu_1$ is sufficient large. $\hat{\mu}_2^{CS}$ is also Pitman closer to μ_2 than $\hat{\mu}_2^{OS}$.

2 Pitman closeness in predicting density function under the $D(\alpha)$ Loss Metric

In this section we will establish Pitman closeness results under the $\{D(\alpha)\}$ loss metric for certain predictive density estimation problems involving two normal populations when the means are ordered. Here we only discuss the unknown variance cases.

Definition 1. Given two predictive density estimates $\hat{f}_1(\tilde{y}|x)$ and $\hat{f}_2(\tilde{y}|x)$ of a density $f(\tilde{y}|\psi)$ based on data x from a distributions $X \sim g(X|\psi)$, $\psi \in \Omega$, $\hat{f}_2(\tilde{y}|x)$ is closer to $f(\tilde{y}|\psi)$ than $\hat{f}_1(\tilde{y}|x)$ with respect to the $D(\alpha)$ metric under the modified Pitman closeness criterion, if $\forall \psi \in \Omega$,

$$P_\psi\{D_\alpha(\hat{f}_2(\tilde{y}|x), f(\tilde{y}|\psi)) < D_\alpha(\hat{f}_1(\tilde{y}|x), f(\tilde{y}|\psi))|\hat{f}_2(\tilde{y}|x) \neq \hat{f}_1(\tilde{y}|x)\} \geq 1/2,$$

with strict inequality for some $\psi \in \Omega$.

2.1 Case when variances are unknown and unrestricted

It is assumed that $\mu_1 \leq \mu_2$ and that no restriction is given on unknown σ_i^2. It is shown that plug-in predictive density with $\hat{\mu}_i^{OS}$, (3), is Pitman closer to the true predictive density than plug-in predictive density with $\bar{X}_i$ under D_α loss. We wish to predict the density of a future observation from a normal population with mean μ_i and unknown variance σ^2, i.e. we wish to predict the density

$$f(\tilde{y}) \sim N(\mu_i, \sigma^2).$$

Let

$$\bar{X}_i \sim N(\mu_i, \sigma_i^2/n_i), \quad S_i^2 \sim \sigma_i^2\chi_{n_i-1}^2,\ i = 1, 2$$

are independent, where $\mu_1 \leq \mu_2$ and it is desired to predict the density of a future independent variable $Y_i \sim N(\mu_i, a\sigma^2), i = 1, 2$, where $a > 0$ is known. We have the following main result.

Theorem 1. The plug-in predictive density estimate $\hat{f}_i^{OS}(\tilde{y}) \sim N(\hat{\mu}_i^{OS}, a\widehat{\sigma^2})$ is Pitman closer to $f(\tilde{y}|\mu_i, a\sigma^2)$ than $\hat{f}_i^{\bar{X}_i}(\tilde{y}) \sim N(\bar{X}_i, a\widehat{\sigma^2})$ for all $\mu_1 \leq \mu_2$ and $\sigma_i^2, i = 1, 2$ under the $D(\alpha)$ metric for all $-1 \leq \alpha \leq 1$ and every estimator $\widehat{\sigma^2}$ if and only if $\hat{\mu}^{GD}$ is Pitman closer to μ than $\bar{X}_i$ for all σ_1^2 and σ_2^2 when $\mu_1 = \mu_2 = \mu$.

2.2 Case when variances are ordered

We give Pitman closeness domination results in predictive density estimation when unknown variances σ_1^2 and σ_2^2 satisfy the order restriction $\sigma_1^2 \leq \sigma_2^2$.

First we estimate the density of a future observation from a normal population with mean μ_2 and variance $\sigma^2 = a\sigma_2^2$, where a is known, i.e. we estimate the density

$$f(\tilde{y}) \sim N(\mu_2, a\sigma_2^2).$$

Theorem 2. The plug-in predictive density estimate $\hat{f}^{CS}(\tilde{y}) \sim N(\hat{\mu}_2^{CS}, a\widehat{\sigma_2^2})$ is Pitman closer to $f(\tilde{y}|\mu_2, a\sigma_2^2)$ than $\hat{f}^{OS}(\tilde{y}) \sim N(\hat{\mu}_2^{OS}, a\widehat{\sigma_2^2})$ under the $D(\alpha)$ metric for all $-1 \leq \alpha \leq 1$ and for any estimator $\widehat{\sigma_2^2}$.

Next we consider estimating the predictive density with smaller variance σ_1^2, $N(\mu_1, a\sigma_1^2)$.

Theorem 3. The plug-in predictive density estimate $\hat{f}^{CS}(\tilde{y}) \sim N(\hat{\mu}_1^{CS}, a\widehat{\sigma}_1^2)$ can not be Pitman closer to $f(\tilde{y}|\mu_1, a\sigma_1^2)$ than $\hat{f}^{OS}(\tilde{y}) \sim N(\hat{\mu}_1^{OS}, a\widehat{\sigma_1^2})$ when $\mu_2 - \mu_1$ is sufficient large, under the $\{D(\alpha)\}$ metric for all $-1 \leq \alpha \leq 1$ and for any estimator $\widehat{\sigma_1^2}$.

一般欠測データの下での2標本問題における多変量正規母集団の同等性検定

東京理科大学・理　野村 玲実
東京理科大学・理　八木 文香
東京理科大学・理　瀬尾　隆

1　はじめに

一般欠測データの下で, 2つの多変量正規母集団の同等性検定問題について考える．すなわち，$\boldsymbol{\mu}^{(\ell)}$ と $\boldsymbol{\Sigma}^{(\ell)}, \ell = 1, 2$ をそれぞれ第 ℓ 母集団の平均ベクトルと分散共分散行列とするとき, この検定問題は $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)}$ である平均ベクトルと分散共分散行列の同時検定を考えることと同じであり，完全データの下では，尤度比検定統計量とその修正尤度比検定統計量が与えられている (Muirhead (1982), Srivastava (2002) など参照). また 1 変量の場合については，Zhang et al. (2012) で議論され, 正確な尤度比検定が与えられている. 一方, 欠測データの下での同時検定については, 単調型欠測の場合の議論があり, Hao and Krishnamoorthy (2001) や Hosoya and Seo (2015, 2016) などで尤度比検定統計量の導出とその帰無分布に対する近似上側パーセント点などが与えられている. ただし, Hao and Krishnamoorthy (2001) や Hosoya and Seo (2015) では 2-step 単調欠測データの下での 1 標本問題における同時検定を議論しており, Hosoya and Seo (2016) では 2-step 単調欠測データの下での多標本問題について議論している. このように単調欠測データの下での同時検定についての議論はあるが, 単調でない一般欠測データの下での同時検定の議論はない. そこで本報告では一般欠測データの下での 2 標本問題に対する同時検定問題, すなわち, 2 つの多変量正規母集団の同等性検定問題を考え, その尤度比検定統計量を導出し, 尤度比検定を与える.

2　最尤推定量と反復法

第 ℓ 母集団 ($\ell = 1, 2$) に対して互いに独立で欠測値を含む p 次元観測ベクトルがそれぞれ $n^{(\ell)}$ 個あるとする. このとき $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)}$ である仮説検定問題を考える．尤度比検定統計量を与えるため, 最尤推定量の導出について考える．同じ欠測パターンをもつ観測ベクトルを 1 つのグループとし, そのグループの総数を $K^{(\ell)}$ とする. また, 第 k 番目のグループにおける観測ベクトルの数を $n_k^{(\ell)} (k = 1, 2, \ldots, K^{(\ell)})$ とする．ここに $\sum_{k=1}^{K^{(\ell)}} n_k^{(\ell)} = n^{(\ell)}, \ell = 1, 2$ である．このとき, k 番目のグループに属する観測ベクトル $\boldsymbol{x}_{kj}^{(\ell)} (k = 1, 2, \ldots, K^{(\ell)}, j = 1, 2, \ldots, n_k^{(\ell)}, \ell = 1, 2)$ の欠測部分を除くような変換行列 $\mathbf{B}_k^{(\ell)}$ を $\boldsymbol{x}_{kj}^{(\ell)}$ にかけ，データの変換を行う．すなわち $\boldsymbol{z}_{kj}^{(\ell)} = \mathbf{B}_k^{(\ell)} \boldsymbol{x}_{kj}^{(\ell)}$ と変換する (Srivastava and Carter (1986) 参照). この $\boldsymbol{z}_{kj}^{(\ell)}$ に対して, 正規性を仮定し, 対数尤度関数を微分することによって, $\boldsymbol{\mu}^{(\ell)}$ の最尤推定量は

$$\widehat{\boldsymbol{\mu}}^{(\ell)} = \left[\sum_{k=1}^{K^{(\ell)}} n_k^{(\ell)} \mathbf{B}_k^{(\ell)'} \widehat{\boldsymbol{\Lambda}}_k^{(\ell)-1} \mathbf{B}_k^{(\ell)} \right]^{-1} \left[\sum_{k=1}^{K^{(\ell)}} n_k^{(\ell)} \mathbf{B}_k^{(\ell)'} \widehat{\boldsymbol{\Lambda}}_k^{(\ell)-1} \bar{\boldsymbol{z}}_k^{(\ell)} \right], \ell = 1, 2$$

となり, $\boldsymbol{\Sigma}^{(\ell)}$ の最尤推定量 $\widehat{\boldsymbol{\Sigma}}^{(\ell)}$ は

$$\sum_{k=1}^{K^{(\ell)}} n_k^{(\ell)} \mathbf{B}_k^{(\ell)'} \widehat{\boldsymbol{\Lambda}}_k^{(\ell)-1} \mathbf{B}_k^{(\ell)} = \sum_{k=1}^{K^{(\ell)}} \mathbf{B}_k^{(\ell)'} \widehat{\boldsymbol{\Lambda}}_k^{(\ell)-1} \widehat{\mathbf{U}}_k^{(\ell)} \widehat{\mathbf{U}}_k^{(\ell)'} \widehat{\boldsymbol{\Lambda}}_k^{(\ell)-1} \mathbf{B}_k^{(\ell)}$$

を満たす解となる. ただし

$$\begin{aligned}\widehat{\mathbf{\Lambda}}_k^{(\ell)} &= \mathbf{B}_k^{(\ell)}\widehat{\mathbf{\Sigma}}^{(\ell)}\mathbf{B}_k^{(\ell)'}, \quad \widehat{\mathbf{U}}_k^{(\ell)} = \mathbf{Z}_k^{(\ell)} - \mathbf{B}_k^{(\ell)}\widehat{\boldsymbol{\mu}}^{(\ell)}\boldsymbol{e}'_{n_k^{(\ell)}}, \quad \bar{\boldsymbol{z}}_k^{(\ell)} = \frac{1}{n_k^{(\ell)}}\sum_{\alpha=1}^{n_k^{(\ell)}}\boldsymbol{z}_{k\alpha}^{(\ell)},\\ \mathbf{Z}_k^{(\ell)} &= (\boldsymbol{z}_{k1}^{(\ell)}, \boldsymbol{z}_{k2}^{(\ell)}, \ldots, \boldsymbol{z}_{kn_k^{(\ell)}}^{(\ell)}), \quad \boldsymbol{e}_{n_k^{(\ell)}} = (1, 1, \ldots, 1)'.\end{aligned}$$

同様に帰無仮説 ($H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}, \mathbf{\Sigma}^{(1)} = \mathbf{\Sigma}^{(2)}$) の下での $\boldsymbol{\mu}$ や $\mathbf{\Sigma}$ の最尤推定量 $\widetilde{\boldsymbol{\mu}}, \widetilde{\mathbf{\Sigma}}$ も求めることができる. $\widehat{\mathbf{\Sigma}}^{(\ell)}$ や $\widetilde{\mathbf{\Sigma}}$ の正確な値は求めることが困難なため, Srivastava and Carter (1986) のアイデアによる反復法を用いて数値的に求める. $\widehat{\mathbf{\Sigma}}^{(\ell)}$ を例として以下に手順を示す. まず, $\mathbf{\Sigma}_i^{(\ell)}, \boldsymbol{\mu}_i^{(\ell)}$ を第 i 近似値とし, $\mathbf{\Delta}_i^{(\ell)}$ を分散共分散行列の第 i 更新量とする. 初期値 $\mathbf{\Sigma}_0^{(\ell)}$ は完全データ部分から計算した分散共分散行列の最尤推定量とする. このとき, 反復式は $\mathbf{\Sigma}_{i+1}^{(\ell)} = \mathbf{\Sigma}_i^{(\ell)} + \mathbf{\Delta}_{i+1}^{(\ell)}$ と書くことができる. ここに $\mathrm{vec}(\mathbf{\Delta}_{i+1}^{(\ell)}) = (\mathbf{Q}_i^{(\ell)})^{-1}\mathrm{vec}(\mathbf{E}_i^{(\ell)})$ で

$$\mathbf{Q}_i^{(\ell)} = \sum_{k=1}^{K^{(\ell)}}(n_k^{(\ell)}\mathbf{D}_{ik}^{(\ell)} \otimes \mathbf{D}_{ik}^{(\ell)} - \mathbf{D}_{ik}^{(\ell)} \otimes \mathbf{F}_{ik}^{(\ell)} - \mathbf{F}_{ik}^{(\ell)} \otimes \mathbf{D}_{ik}^{(\ell)}), \quad \mathbf{E}_i^{(\ell)} = \sum_{k=1}^{K^{(\ell)}}(n_k^{(\ell)}\mathbf{D}_{ik}^{(\ell)} - \mathbf{F}_{ik}^{(\ell)})$$

であり, $\mathbf{D}_{ik}^{(\ell)}, \mathbf{F}_{ik}^{(\ell)}$ は

$$\begin{aligned}\mathbf{D}_{ik}^{(\ell)} &= \mathbf{B}_k^{(\ell)'}(\mathbf{B}_k^{(\ell)}\mathbf{\Sigma}_i^{(\ell)}\mathbf{B}_k^{(\ell)'})^{-1}\mathbf{B}_k^{(\ell)},\\ \mathbf{F}_{ik}^{(\ell)} &= \mathbf{B}_k^{(\ell)'}(\mathbf{B}_k^{(\ell)}\mathbf{\Sigma}_i^{(\ell)}\mathbf{B}_k^{(\ell)'})^{-1}\mathbf{U}_{ik}^{(\ell)}\mathbf{U}_{ik}^{(\ell)'}(\mathbf{B}_k^{(\ell)}\mathbf{\Sigma}_i^{(\ell)}\mathbf{B}_k^{(\ell)'})^{-1}\mathbf{B}_k^{(\ell)}\end{aligned}$$

である. ただし, $\mathbf{U}_{ik}^{(\ell)} = \mathbf{Z}_k^{(\ell)} - \mathbf{B}_k^{(\ell)}\boldsymbol{\mu}_i^{(\ell)}\boldsymbol{e}'_{n_k^{(\ell)}}$ である. よって尤度比は

$$\lambda = \frac{\prod_{k=1}^{K^{(1)}}|\widehat{\mathbf{\Lambda}}_k^{(1)}|^{\frac{1}{2}n_k^{(1)}}\prod_{k=1}^{K^{(2)}}|\widehat{\mathbf{\Lambda}}_k^{(2)}|^{\frac{1}{2}n_k^{(2)}}}{\prod_{k=1}^{K^{(1)}}|\widetilde{\mathbf{\Lambda}}_k^{(1)}|^{\frac{1}{2}n_k^{(1)}}\prod_{k=1}^{K^{(2)}}|\widetilde{\mathbf{\Lambda}}_k^{(2)}|^{\frac{1}{2}n_k^{(2)}}}$$

で与えられる. ただし, $\widetilde{\mathbf{\Lambda}}_k^{(\ell)} = \mathbf{B}_k^{(\ell)}\widetilde{\mathbf{\Sigma}}\mathbf{B}_k^{(\ell)'}$ である. 尤度比検定統計量 $-2\log\lambda$ の極限分布は自由度 $p(p+3)/2$ のカイ二乗分布であり, その近似精度をモンテカルロ・シミュレーションにより評価した.

参考文献

[1] Hao, J. and Krishnamoorthy, K. (2001). Inferences on a normal covariance matrix and generalized variance with monotone missing data. *Journal of Multivariate Analysis*, **78**, 62-82.

[2] Hosoya, M. and Seo, T. (2015). Simulaneous testing of the mean vector and the covariance matrix with two-step monotone missing data. *SUT Journal of Mathematics*, **51**, 83-98.

[3] Hosoya, M. and Seo, T. (2016). On the likelihood ratio test for the equality of multivariate normal populations with two-step monotone missing data. *Journal of Statistical Theory and Practice*, **10**, 673-692.

[4] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Willey, New York.

[5] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. Willey, New York.

[6] Srivastava, M. S. and Carter, E. M. (1986). The maximum likelihood method for non-response in sample surveys. *Survey Methodology*, **12**, 61-72.

[7] Zhang, L., Xu, X. and Chen, G. (2012). The exact likelihood ration test for equality of two normal populations. *The American Statistician,* **66**, 180-184.

単調欠測データが一様共分散構造を持つ場合の平行性仮説検定と水準差の信頼区間

東京理科大学・理　佐伯 悠一郎
東京理科大学・理　八木 文香
東京理科大学・理　瀬尾 隆
防衛大学校・数学　百武 弘登

1　はじめに

本研究ではプロフィール分析における平行性仮説検定とその水準差について考える．平行性仮説検定問題は，単調欠測データの下で $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \cdots, \boldsymbol{\mu}^{(g)}$ を g 個の母集団の平均ベクトルとすると，次のように表される．

$$H:\ \boldsymbol{\mu}^{(\ell)} - \boldsymbol{\mu}^{(g)} = \gamma^{(\ell)}\mathbf{1}_p,\ \ell = 1, 2, \ldots, g-1 \text{ vs. } A:\ \text{not } H,$$

ただし，$\gamma^{(\ell)}$ は水準差と呼ばれる未知パラメータであり，$\mathbf{1}_p = (1, 1, \cdots, 1)'$ とする．平行性仮説検定は Srivastava (1987) 等で完全データに対して尤度比検定統計量と漸近分布を用いた近似パーセント点が与えている．また欠測データの下での議論については Onozawa, Takahashi and Seo (2013) で 2-step 単調欠測データに対し，2 群におけるホテリングの T^2 型検定統計量を用いた検定と，多群に対する尤度比検定統計量を与えている．どちらも検定統計量の漸近分布と近似上側パーセント点を提案している．一方，Sawatsuhashi and Hyakutake (2016) では観測ベクトルにランダム効果モデルと一様共分散構造を仮定した場合の 2 群における，一般の単調欠測データの下でのプロフィール分析が議論され，検定統計量とその正確な分布が得られている．一様共分散構造とは繰り返し測定データに対して仮定される分散共分散行列の構造の一つである．本報告では Sawatsuhashi and Hyakutake (2016) を基に，一様構造を持つ一般の単調欠測データに対して，多群の場合の平行性仮説検定と，2 群かつ 2-step 単調欠測データに対し水準差の信頼区間について考える．

2　平行性仮説検定

k-step 単調欠測データとして，$\boldsymbol{x}_{ij}^{(\ell)} \sim N_{p_i}(\boldsymbol{\mu}_i^{(\ell)}, \boldsymbol{\Sigma}_i)$ $(i = 1, 2, \ldots, k\ ,\ j = 1, 2, \ldots, n_i^{(\ell)})$ を第 ℓ 群の第 i ステップ内 j 番目にある観測ベクトルとする．ただし，$p = p_1 > p_2 > \cdots > p_k$ である．また $\boldsymbol{\mu}_i^{(\ell)}$ は $\boldsymbol{\mu}^{(\ell)}$ の第 1 から第 p_i 成分で構成される $(p_i \times 1)$ ベクトルであり，$\boldsymbol{\Sigma}_i$ は $\boldsymbol{\Sigma}$ の p_i 次主座小行列とする．本報告では分散共分散行列の構造をとし，尤度比検定統計量の正確な分布を導出した．

$$\boldsymbol{\Sigma} = \sigma^2\big\{(1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p'\big\},\ \text{ただし，} -1/(p-1) < \rho < 1.$$

観測ベクトルを $\boldsymbol{C}_i\mathbf{1}_{p_i-1} = \mathbf{0}$ ，$\boldsymbol{C}_i\boldsymbol{C}_i' = \boldsymbol{I}_{p_i}$ を満たす $(p_i - 1) \times p_i$ の対比行列 $\boldsymbol{C}_i$ を用いて変換する．具体的な行列 $\boldsymbol{C}_i$ については Sawatsuhashi and Hyakutake (2016) などを参照されたい．変換後の観測ベクトルは $\boldsymbol{y}_{ij}^{(\ell)} = \boldsymbol{C}_i\boldsymbol{x}_{ij}^{(\ell)} \sim N_{p_i-1}(\boldsymbol{\eta}_i^{(\ell)}, \tau_0\boldsymbol{I}_{p_i})$ となる．ここで，$\boldsymbol{\eta}_i^{(\ell)} = \boldsymbol{C}_i\boldsymbol{\mu}_i^{(\ell)}$ ，$\tau_0 = \sigma^2(1-\rho)$ である．さらに $p_{k-i+1} = \sum_{h=1}^{i} q_h$ となる q_m $(m = 1, 2, \ldots, k)$ に対し，$\boldsymbol{y}_{ij}^{(\ell)}$ を $(q_1 - 1)$ 次元，q_2 次元 $, \ldots, q_k$ 次元ごとに区切り $\boldsymbol{z}_{ms}^{(\ell)}$ とする．m は縦に分割後のブロック番号であり，s はサンプル番号を意味する．ただし，$M_m^{(\ell)} = \sum_{i=1}^{k-m+1} n_i^{(\ell)}$ とする．$\boldsymbol{\mu}^{(\ell)}$ についても同様に分割し $\boldsymbol{\xi}_m^{(\ell)}$ とすると $\boldsymbol{z}_{1s}^{(\ell)} \sim N_{q_1-1}(\boldsymbol{\xi}_1^{(\ell)}, \tau_0\boldsymbol{I}_{q_1-1})$ ，$\boldsymbol{z}_{ms}^{(\ell)} \sim N_{q_m}(\boldsymbol{\xi}_m^{(\ell)}, \tau_0\boldsymbol{I}_{q_m})$, $m = 2, 3, \ldots, k$ となる．このことから，一般の k-step の場合の尤度関数を与えることができるので $H : \boldsymbol{\xi}^{(1)} = \boldsymbol{\xi}^{(2)} = \cdots = \boldsymbol{\xi}^{(g)}$ vs. A : not H の下での尤度比は $M_m = \sum_{\ell=1}^{g} M_m^{(\ell)}$, $\overline{\boldsymbol{z}}_m^{(\ell)} = (1/M_m^{(\ell)})\sum_{s=1}^{M_m^{(\ell)}} \boldsymbol{z}_{ms}^{(\ell)}$ ，$\overline{\boldsymbol{z}}_m = (1/M_m)\sum_{\ell=1}^{g}\sum_{s=1}^{M_m^{(\ell)}} \boldsymbol{z}_{ms}^{(\ell)}$ とすると

$$\lambda_k = \Big(1 + \sum_{m=1}^{k}\sum_{\ell=1}^{g} M_m^{(\ell)}\big(\overline{\boldsymbol{z}}_m - \overline{\boldsymbol{z}}_m^{(\ell)}\big)'\big(\overline{\boldsymbol{z}}_m - \overline{\boldsymbol{z}}_m^{(\ell)}\big) \Big/ \sum_{m=1}^{k}\sum_{\ell=1}^{g}\sum_{s=1}^{M_m^{(\ell)}}\big(\boldsymbol{z}_{ms}^{(\ell)} - \overline{\boldsymbol{z}}_m^{(\ell)}\big)'\big(\boldsymbol{z}_{ms}^{(\ell)} - \overline{\boldsymbol{z}}_m^{(\ell)}\big)\Big)^{-(1/2)\nu_k}$$

となる．また，$\nu_k = M_1(q_1-1) + M_2 q_2 + \cdots + M_k q_k$ である．さらにこの尤度比を用いた検定統計量 T_k の分布について

$$T_k = \frac{\nu_k - g(p-1)}{(g-1)(p-1)}\big(\lambda_k^{-\frac{2}{\nu_k}} - 1\big) \sim F_{(g-1)(p-1),\ \nu_k - g(p-1)}$$

であることが示される．補足として，いくつかのパラメータ設定でのモンテカルロ・シミュレーションを行ったところ，数値的にも F 分布に一致することを確認している．

3　水準差の信頼区間

観測ベクトル $\boldsymbol{x}_{ij}^{(\ell)}$ $(i = 1,2\ ,\ j = 1,2,\ldots,n_i^{(\ell)}\ ,\ \ell = 1,2)$ は $N_{p_i}(\boldsymbol{\mu}_i^{(\ell)}, \boldsymbol{\Sigma}_i)$ に従うとし，平行性 $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} = \gamma \mathbf{1}_p$ が認められているとする．欠測データにおけるパラメータ ρ, σ^2 の MLE(最尤推定量) は容易に求められないので，これらを既知とみなし最後に推定量を代入する．$n_i^* = (n_i^{(1)} n_i^{(2)})/(n_i^{(1)} + n_i^{(2)})$, $\overline{\boldsymbol{x}}_i^{(\ell)} = (1/n_i^{(\ell)}) \sum_{j=1}^{n_i^{(\ell)}} \boldsymbol{x}_{ij}^{(\ell)}$, $\boldsymbol{u}_i = \overline{\boldsymbol{x}}_i^{(1)} - \overline{\boldsymbol{x}}_i^{(2)}$ とおくと ρ が既知の場合の γ の MLE は $\widehat{\gamma} = \sum_{i=1}^2 n_i^* \tau_{1i}^{-1} \mathbf{1}_{p_i}' \boldsymbol{u}_i / \sum_{i=1}^2 n_i^* \tau_{1i}^{-1} p_i$ である．ただし，$\tau_{1i} = 1 + (p_i - 1)\rho$ である．ここで $\boldsymbol{V}_i^{(\ell)} = \sum_{j=1}^{n_i^{(\ell)}} \big(\boldsymbol{x}_{ij}^{(\ell)} - \overline{\boldsymbol{x}}_i^{(\ell)}\big)\big(\boldsymbol{x}_{ij}^{(\ell)} - \overline{\boldsymbol{x}}_i^{(\ell)}\big)'$, $\boldsymbol{V_i} = \sum_{\ell=1}^2 \boldsymbol{V}_i^{(\ell)}$, $\boldsymbol{S}_i = (1/n_i - 2)\boldsymbol{V}_i$, $n_i = \sum_{\ell=1}^2 n_i^{(\ell)}$ と定義し，Siotani, Hayakawa and Fujikoshi (1985) で与えられている完全データの下での MLE を基に，パラメータ ρ と σ^2 の推定量として $\widetilde{\rho} = \sum_{i=1}^2 (\widehat{\tau}_{2i} - \mathrm{tr}(\boldsymbol{S}_i))/\sum_{i=1}^2 \{p_i(p_i-1)\}\widetilde{\sigma}^2$, $\widetilde{\sigma}^2 = \sum_{i=1}^2 \mathrm{tr}(\boldsymbol{V}_i)/\sum_{i=1}^2 (n_i - 2)p_i$ を提案する．ただし $\widehat{\tau}_{2i} = \mathbf{1}_{p_i}' \boldsymbol{S}_i \mathbf{1}_{p_i}$ とする．$\widetilde{\rho}$, $\widetilde{\sigma}^2$ は共に MLE ではないが，どちらも 2-step 単調欠測データの全ての観測値を使用した統計量である．Jensen の不等式とこれらの推定量を用いて γ の $100\alpha\%$ 近似信頼区間

$$\widetilde{\gamma} - \left(\sum_{i=1}^2 \frac{n_i^* p_i^2}{\widehat{\tau}_{2i} t_{n_i-2}^2(\frac{\alpha}{2})}\right)^{-\frac{1}{2}} < \ \gamma\ < \widetilde{\gamma} + \left(\sum_{i=1}^2 \frac{n_i^* p_i^2}{\widehat{\tau}_{2i} t_{n_i-2}^2(\frac{\alpha}{2})}\right)^{-\frac{1}{2}}$$

が得られた．ただし，$\widetilde{\gamma} = \sum_{i=1}^2 \widetilde{\tau}_{1i}^{-1} n_i^* \mathbf{1}_{p_i}' \boldsymbol{u}_i / \sum_{i=1}^2 \widetilde{\tau}_{1i}^{-1} n_i^* p_i$, $\widetilde{\tau}_{1i} = 1 + (p_i - 1)\widetilde{\rho}$ であり，$t_{n_i-2}(\alpha/2)$ は自由度 $n_i - 2$ の t 分布の上側 $100(\alpha/2)\%$ 点である．

4　シミュレーション結果

前節で導出した近似信頼区間を C1 とし，完全データ部分のみを使用した信頼区間を C2 とすると，$100\alpha\%$ 信頼区間 C2 は $\widehat{\gamma}_1 - t_{n_1-2}(\alpha/2)((n_1^* p_1^2)/\widehat{\tau}_{21})^{-\frac{1}{2}} < \ \gamma\ < \widehat{\gamma}_1 + t_{n_1-2}(\alpha/2)(n_1^* p_1^2/\widehat{\tau}_{21})^{-\frac{1}{2}}$ である．有意水準 5%，繰り返し回数を 1,000,000 回とし，被覆確率を比較した．結果として C1 の方が被覆確率が高く，かつその確率は 0.95 を大きく上回らない，及び狭い区間であることが確認できた．

参考文献

[1] Onozawa, M., Takahashi, S. and Seo, T. (2013). Tests for profile analysis based on two-step monotone missing data. *Discussiones Mathematicae Probability and Statistics*, **33**, 171-190.

[2] Sawatsuhashi, K. and Hyakutake, H. (2016). Profile analysis for random effects model in two sample problem with monotone missing. *Far East Journal of Theoretical Statistics*, **52**, 235-251.

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

[4] Srivastava, M. S. (1987). Profile analysis of several groups. *Communications in Statistics -Theory and Methods*, **16**, 909-926.

2-step 単調欠測データのもとでの 部分平均ベクトルの検定に対する修正尤度比検定統計量

東京理科大・理　川崎　玉恵
東京理科大・理　瀬尾　隆

1　はじめに

1 標本，2 標本問題における部分平均ベクトルの仮説検定問題を 2-step 単調欠測データのもとで議論した．部分平均ベクトルの仮説検定問題について，欠測値を含んでいない完全データの場合は，尤度比を用いた検定統計量が 1 標本問題では Rao (1949) や Giri (1964) などで議論されており，これらは Rao の U 統計量と呼ばれている．欠測値を含むデータに対しては，Kawasaki and Seo (2016) が 1 標本問題における 2-step 単調欠測データのもとでの最尤推定量と尤度比検定統計量を導出している．

本報告では，2-step 単調欠測データのもとでの 1 標本，2 標本問題における部分平均ベクトルの仮説検定問題について，1 標本問題では Kawasaki and Seo (2016) で与えた尤度比検定統計量を，2 標本問題では最尤推定量を導出し，尤度比検定統計量を与えたのち，検定統計量の帰無分布を大標本漸近枠組みの下で漸近展開することで，それぞれの修正尤度比検定統計量を与えた．さらにこれらの精度について，モンテカルロ・シミュレーションにより数値的評価を行った．

2　1 標本問題における部分平均ベクトルの仮説検定問題

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_1}$ が $N_p(\boldsymbol{\mu}, \Sigma)$ に，$\boldsymbol{x}_{N_1+1}, \boldsymbol{x}_{N_1+2}, \ldots, \boldsymbol{x}_N$ が $N_{p_1+p_2}(\boldsymbol{\mu}_{(12)}, \Sigma_{(12)(12)})$ に従う互いに独立な確率ベクトルとする．ただし，$\boldsymbol{\mu}$ と Σ はそれぞれ次のよう分割される．

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{10} \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{(12)} \\ \boldsymbol{\mu}_3 \end{pmatrix}, \quad \Sigma = \left(\begin{array}{cc|c} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \hline \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{array}\right) = \begin{pmatrix} \Sigma_{(12)(12)} & \Sigma_{(12)3} \\ \Sigma_{3(12)} & \Sigma_{33} \end{pmatrix} \tag{1}$$

また，$\boldsymbol{x}_j$ をそれぞれ $p_1 \times 1, p_2 \times 1, p_3 \times 1$ の確率ベクトルとし，$\boldsymbol{x}_j = (\boldsymbol{x}'_{1j}, \boldsymbol{x}'_{2j}, \boldsymbol{x}'_{3j})' = (\boldsymbol{x}'_{(12)j}, \boldsymbol{x}'_{3j})'$ と表す．ただし，$\boldsymbol{x}_{ij}, i = 1, 2, 3, j = 1, 2, \ldots, N_1$ は $p_i \times 1$，$p = p_1 + p_2 + p_3$ であり，$N_2 = N - N_1$ である．このとき以下の仮説検定問題を考えた．

$$H_0 : \boldsymbol{\mu}_{(23)} = \boldsymbol{\mu}_{(23)0} \text{ given } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_{10} \text{ vs. } H_1 : \boldsymbol{\mu}_{(23)} \neq \boldsymbol{\mu}_{(23)0} \text{ given } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_{10}$$

この問題における尤度比検定統計量 $-2\log\lambda$ は Kawasaki and Seo (2016) で与えられている．ただし

$$\lambda = \left(\frac{|\widetilde{\Phi}_{(12)(12)}|}{|\widehat{\Psi}_{11}| \cdot |\widehat{\Psi}_{22}|}\right)^{-N/2} \left(\frac{|\widetilde{\Phi}_{33}|}{|\widehat{\Psi}_{33}|}\right)^{-N_1/2}$$

であり，記号の説明については Kawasaki and Seo (2016) を参照されたい．この尤度比検定統計量における帰無分布の漸近展開を以下の大標本漸近枠組み

$$\delta_i = \frac{N_i}{N} \to \text{ positive constants as } N_1, N_2 \to \infty, \ i = 1, 2$$

の下で行い，修正尤度比検定統計量を導出した．ただし，摂動展開は以下を用いる．

$$\boldsymbol{z}_F = \begin{pmatrix} \boldsymbol{z}_{F1} \\ \boldsymbol{z}_{F2} \\ \boldsymbol{z}_{F3} \end{pmatrix} = \sqrt{N_1} \begin{pmatrix} \overline{x}_{F1} \\ \overline{x}_{F2} \\ \overline{x}_{F3} \end{pmatrix},\ V_F = \begin{pmatrix} V_{F11} & V_{F12} & V_{F13} \\ V_{F21} & V_{F22} & V_{F23} \\ V_{F31} & V_{F32} & V_{F33} \end{pmatrix} = \sqrt{N_1 - 1}(S_F - I_p),$$

$$\boldsymbol{z}_L = \begin{pmatrix} \boldsymbol{z}_{L1} \\ \boldsymbol{z}_{L2} \end{pmatrix} = \sqrt{N_2} \begin{pmatrix} \overline{x}_{L1} \\ \overline{x}_{L2} \end{pmatrix},\ V_L = \begin{pmatrix} V_{F11} & V_{F12} \\ V_{F21} & V_{F22} \end{pmatrix} = \sqrt{N_2 - 1}(S_L - I_{p_{(12)}})$$

また，$\overline{\boldsymbol{x}}_F = (\overline{\boldsymbol{x}}'_{F1}, \overline{\boldsymbol{x}}'_{F2}, \overline{\boldsymbol{x}}'_{F3})'$ と S_F はそれぞれ $N_1 \times p$ 部分の，$\overline{\boldsymbol{x}}_L = (\overline{\boldsymbol{x}}'_{L1}, \overline{\boldsymbol{x}}'_{L2})'$ と S_L はそれぞれ $N_2 \times (p_1 + p_2)$ 部分の標本平均ベクトルと不偏標本分散共分散行列とする．

以上から，尤度比検定統計量 $-2\log\lambda$ の特性関数は

$$\mathrm{E}[e^{it(-2\log\lambda)}] = (1-2it)^{-(p_2+p_3)/2} \left\{ 1 + \frac{(p_2+p_3)c}{2N}\{(1-2it)^{-1} - 1\} \right\} + o(N^{-1})$$

と展開することができた．ただし

$$c = \frac{1}{p_2+p_3} \left\{ \frac{1}{2}p_2(2p_1+p_2+2) + \frac{1}{\delta_1}p_3(p+3) - \frac{1}{\delta_1}p_3(p_3+2) \right\}.$$

よって，$-2\log\lambda$ の帰無分布に対する漸近展開は

$$\Pr(-2\log\lambda \le x) = G_{p_2+p_3}(x) + \frac{(p_2+p_3)c}{2N}\{G_{p_2+p_3+2}(x) - G_{p_2+p_3}(x)\} + o(N^{-1})$$

のように与えられた．ただし，$G_f(x)$ は自由度 f の χ^2 分布の分布関数とする．さらに，バートレット修正係数を $\rho = 1 - c/N$ と提案した．

3　2標本問題における部分平均ベクトルの仮説検定問題

2-step 単調欠測データのもとでの 2 標本問題における部分平均ベクトルの仮説検定問題

$$H_0 : \boldsymbol{\mu}^{(1)}_{(23)} = \boldsymbol{\mu}^{(2)}_{(23)} \text{ given } \boldsymbol{\mu}^{(1)}_1 = \boldsymbol{\mu}^{(2)}_1 \text{ vs. } H_1 : \boldsymbol{\mu}^{(1)}_{(23)} \neq \boldsymbol{\mu}^{(2)}_{(23)} \text{ given } \boldsymbol{\mu}^{(1)}_1 = \boldsymbol{\mu}^{(2)}_1 \qquad (2)$$

を考えた．ただし，$\boldsymbol{\mu}^{(i)}_1, \boldsymbol{\mu}^{(i)}_{(23)}$ の分割は 1 標本問題のときの (1) 式と同様であり，$i = 1, 2$ は母集団を表す添え字とする．

仮説検定問題 (2) における尤度比検定統計量を導出し，1 標本問題と同様に大標本漸近枠組みのもとで漸近展開を行い，修正尤度比検定統計量の導出をおこなった．また，100 万回のモンテカルロ・シミュレーションにより本報告で与えた修正尤度比検定統計量についての数値的評価を行った．

参考文献

[1] Giri, N. C. (1964). On the likelihood ratio test of a normal multivariate testing problem. *Ann. Math. Stat.*, **35**, 181–189.

[2] Kawasaki, T. and Seo, T. (2016). A test for subvector of mean vector with two-step monotone missing data. *SUT J. Math.*, **52**, 21–39.

[3] Rao, C. R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhyā*, **9**, 343–364.

Recent cylindrical models and their application to tree data set

Toshihiro Abe[1] & Ichiro Ken Shimatani[2]
[1] Nanzan University, Nagoya, Japan
[2] The Institute of Statistical Mathematics, Tokyo, Japan

1 Introduction

If the objective variable is circular, we need multivariate versions of circular probability distributions, and in the two-dimensional case, there are two types. The first is when two objective variables are circular; then, the probability distribution is defined on the torus. The second type is when one is circular and the other is linear; then, the probability distribution should be defined on $S^1 \times \mathbb{R}$, or if linear variables take only non-negative values, the domain is $S^1 \times [0, \infty)$. In both cases, the probability distribution is defined on a "cylinder" and such distributions are called "cylindrical distributions." The latter type of data are commonly seen, for example, in velocity data, which consist of a speed and a direction. Often, these show specific correlations. In such cases, correlations are present between the linear and circular variables in the form that the variances over the circular variables become high when speeds are slow. In this talk, by focusing on cylindrical distributions on $S^1 \times [0, \infty)$, we introduce those distributions that are mathematically tractable, namely the probability density function (pdf) can be expressed by elementary functions or, at most, well-known special functions, and provide examples of their applications.

2 Cylindrical distributions

Let (Θ, L) be a pair of circular and linear random variables, respectively. When we observe cylindrical data, if the linear components increase, the degree of concentration around a certain direction also tends to increase. This section introduces models that satisfy these conditions.

Johnson and Wehrly (1978) proposed a cylindrical distribution whose joint pdf is given by

$$f_{JW}(\theta, l; \mu, \nu, \kappa) = \frac{\sqrt{\nu^2 - \kappa^2}}{2\pi} \exp\{-\nu l + \kappa l \cos(\theta - \mu)\}, \quad (\theta, l) \in [-\pi, \pi) \times [0, \infty), \tag{1}$$

where $0 \leq \kappa < \nu$ and $-\pi \leq \mu < \pi$. The distribution is obtained by maximizing the (Shannon) entropy with $E(L)$, $E(L\cos\Theta)$ and $E(L\sin\Theta)$ constant (see, Mardia, 1975, p. 352). The functional form of the density indicates that as the length part $(= l)$ increases, the concentration around the location $(= \mu)$ also increases. Clearly, the mode of the density always occurs at $l = 0$.

As a more flexible symmetric cylindrical distribution, we show a special case of the density in Abe and Ley (2017), whose joint pdf is given by

$$f_{WM}(\theta, l; \mu, \kappa, \alpha, \beta) = \frac{\alpha\beta^\alpha}{2\pi\cosh(\kappa)} l^{\alpha-1} \exp\left[-(\beta l)^\alpha \{1 - \tanh(\kappa)\cos(\theta - \mu)\}\right], \tag{2}$$

where $(\theta, l) \in [-\pi, \pi) \times [0, \infty)$, $\alpha > 0$ and $\beta > 0$ are the linear scale and shape parameters, and $-\pi \leq \mu < \pi$ and $\kappa \geq 0$ are the circular location and concentration parameters, respectively. We

call this cylindrical distribution the Weibull–von Mises distribution. The density (2) is basically proportional to the product of the Weibull and von Mises (replacing κ with κl) distributions,

$$f_{WM}(\theta, l) \propto l^{\alpha-1} \exp\left[-(\beta l)^{\alpha}\right] \times \exp\left\{(\beta l)^{\alpha} \tanh(\kappa) \cos(\theta - \mu)\right\},$$

divided by the normalizing constant.

As in Abe and Ley (2017), another cylindrical distribution is obtained by replacing the Weibull distribution with the Gamma distribution. The pdf is given by

$$f_{GM}(\theta, l; \mu, \kappa, \alpha, \beta) \quad = \quad C^{-1} l^{\alpha-1} \exp\left[-\beta l \left\{1 - \tanh(\kappa) \cos(\theta - \mu)\right\}\right], \tag{3}$$

where $(\theta, l) \in [-\pi, \pi) \times [0, \infty)$, $\alpha, \beta, \gamma, \kappa > 0$, and $-\pi \leq \mu < \pi$. The normalizing constant C is given by

$$\int_{-\pi}^{\pi} \int_{0}^{\infty} l^{\alpha-1} \exp[-\beta l\{1 - \tanh(\kappa) \cos(\theta - \mu)\}] dl d\theta \quad = \quad \frac{2\pi\Gamma(\alpha)(\cosh(\kappa))^{\alpha} P_{\alpha-1}(\cosh(\kappa))}{\beta^{\alpha}}.$$

We call this distribution the Gamma–von Mises distribution.

2.1 Parameter estimation

The most preferable orientation of asymmetry for focal tree i is written as

$$\mu_i(a) = \arg\left(a\overrightarrow{\boldsymbol{S}} + \sum_{j=1}^{n_i} CI^{ji}\overrightarrow{\boldsymbol{x}_j\boldsymbol{x}_i}\right), \tag{4}$$

where n_i is the number of competitors, the summation covers all the competitors, and $\overrightarrow{\boldsymbol{S}} = (0, -1)$. The parameter a quantifies the relative importance of solar radiation and competitors with a given size and distance. By substituting $\mu = \mu_i(a)$ into the symmetric cylindrical distributions and letting $(\theta_1, l_1), \ldots, (\theta_n, l_n)$ be a sample of n independent and identically distributed couples of circular and linear observations drawn from a cylindrical distribution. Then, the likelihood function can be expressed, in the cases of the basic three symmetric distributions, namely the Johnson–Wehrly, Weibull–von Mises, and Gamma–von Mises distributions, respectively as

$$L_{JW}(\mu, \kappa, \nu) \quad = \quad \prod_{i=1}^{n} f_{JW}(l_i, \theta_i; \mu, \kappa, \nu), \tag{5}$$

$$L_{WM}(\mu, \kappa, \alpha, \beta) \quad = \quad \prod_{i=1}^{n} f_{WM}(l_i, \theta_i; \mu, \kappa, \alpha, \beta), \tag{6}$$

$$L_{GM}(\mu, \kappa, \alpha, \beta) \quad = \quad \prod_{i=1}^{n} f_{GM}(l_i, \theta_i; \mu, \kappa, \alpha, \beta). \tag{7}$$

The likelihood functions of the other distributions can be written in the same way. In general, it is difficult to give closed-form expressions for the maximum likelihood estimates (MLEs); hence, numerical methods should be used to find the solutions.

References

[1] Abe, T. & Ley, C. (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, **4**, 91–104.