学術シンポジウム「統計科学の開拓」

報告書

主催：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革新的展開」
（課題番号：20H00576）

研究代表者：青嶋 誠（筑波大学）

■学術シンポジウム「統計科学の開拓」実施概要

・開催日時：2022 年 12 月 16 日（金）～12 月 18 日（日）

・会場：金沢大学サテライト・プラザ 3 階集会室（金沢市西町三番丁 16 番地）

・開催責任者：星野 伸明（金沢大学）

■プログラム

12/16(金)			
開始	終了	講演者	演題
13:00	13:05	主催者	開催挨拶
13:10	13:50	江頭健斗(筑波大学理工情報生命学術院)，矢田和善(筑波大学数理物質系)，青嶋誠(筑波大学数理物質系)	Hierarchical clustering and its asymptotic behaviors in high-dimensional settings
13:55	14:35	桃木 光輝（鹿児島大学），吉田 拓真（鹿児島大学）	Mixed effects modeling of clustered extreme values
14:40	15:20	島谷健一郎（統計数理研究所）	植物個体群へ応用できる繰り返し可能な非定常クラスター点過程
15:25	16:05	種市信裕（北海道教育大学札幌校）， 関谷祐里（北海道教育大学釧路校）	スパース性の仮定のもとでの多項分布の適合度検定統計量選択の指針とその応用
16:10	16:50	張元宗・目白大学, 篠崎　信雄　・慶應大学	Simultaneous estimation of Poisson means in two-way contingency tables under normalized squared　error loss
16:55	17:35	中島亮（東京理科大学大学院　理工学研究科） 星野瞭太（東京理科大学大学院　理工学研究科） 藤澤健吾（東京理科大学　理工学部） 田畑耕治（東京理科大学　理工学部）	正方分割表における点対称性からの隔たりを測る尺度の推定について
17:40	18:20	足立 匠 (東京理科大学大学院 理工学研究科 情報科学専攻)、安藤宗司(東京理科大学 理工学部 情報科学科)、田畑 耕治(東京理科大学 理工学部 情報科学科)	正方分割表におけるモーメントに基づく周辺同等性からの隔たりを測る尺度
12/17(土)			
開始	終了	講演者	演題
9:20	10:00	西埜晴久（広島大学大学院人間社会科学研究科）	国民生活基礎調査による日本の所得分布と所得格差
10:05	10:45	国友直人（統計数理研究所特任教授）	操作変数法の理解へ：計量生物と計量経済の邂逅
10:50	11:30	岡野遼（東京大学経済学研究科経済専攻統計学コース博士1年）	多変量分布間の回帰モデル
			昼休み
13:00	13:40	永井勇（中京大学）	GMANOVAモデルでの仮定緩和のさらなる可能性
13:45	14:25	松内 直輝（神戸大学大学院理学研究科），首藤 信通（神戸大学大学院理学研究科）	多変量正規母集団における条件付き独立性検定について
14:30	15:10	田中勝人（一橋大学名誉教授）	多重積分ブラウン運動の統計学的応用について
15:15	15:55	Bat-Erdene, A., Graduate School of Science and Engineering, Iwate University Kawasaki, S., Faculty of Science and Engineering, Iwate University Li, J., School of Computer Science, Chongqing University Altantsetseg, E., School of the Engineering and Applied Sciences, National University of Mongolia	On Asymptotic Distribution in Martingale Convergence of Supercritical Branching Processes with Poissonian Offsprings
16:00	16:40	Islam Md Ashraful (大阪大学), 鈴木譲 (大阪大学)	Forest Construction of Gaussian and Binary Variables based on WBIC
16:45	17:25	山添滉弥(千葉大学融合理工学府)・内藤貫太(千葉大学理学研究院)	多次元空間における埋め込み1次元曲線の同時信頼領域
17:30	18:10	勝又 真（神戸大学大学院理学研究科） 首藤 信通（神戸大学大学院理学研究科）	多標本問題における 2-step 単調欠測データの下での平均ベクトルの同等性検定の検出力について
12/18(日)			
開始	終了	講演者	演題
9:20	10:00	松井宗也（南山大学経営学部）	Subexponentaliy of densities of infinitely divisible distributions on the whole real line
10:05	10:45	入江薫（東京大学経済学部）	ガンマ分布の形状パラメータのベイズ推定
10:50	11:30	前園宜彦（中央大学理工学部）	カーネル型推定量を利用した推測について
			昼休み
13:00	13:40	塚原英敦（成城大学経済学部）	On a generalization of Clayton-Oakes model by R. L. Prentice
13:45	14:25	助田一晟（東京大学情報理工学系研究科M1），清智也（東京大学情報理工学系研究科）	Kendallの順位相関係数を固定した下での最小情報コピュラ
14:30	15:10	渡邊宏大（千葉大学融合理工学府）・内藤貫太（千葉大学理学研究院）	スピアマンランク行列による主成分分析のロバストネス
15:10	15:15	主催者	閉会挨拶

■報告書目次

Hierarchical clustering and its asymptotic behaviors in high-dimensional settings

Kento Egashira[a], **Kazuyoshi Yata**[b], **Makoto Aoshima**[b]

[a]**Degree Programs in Pure and Applied Sciences, Graduate School of Science and Technology, University of Tsukuba**
[b]**Institute of Mathematics, University of Tsukuba**

Hierarchical clustering is a methodology to group a set of data by building dendrogram based on a similarity or a dissimilarity between clusters so that data in a cluster are similar in the sense of pre-determined linkage function. In hierarchical clustering, one can observe a process how a cluster is combined or divided through dendrogram on graphic. Hierarchical clustering has been approved as useful tool for analysis of gene expression microarray data. In fact, applications of hierarchical clustering on gene expression microarray data are given by Eisen et al. [4], Perou et al. [8], Bhattacharjee et al. [2], among others. A characteristic of data used in Eisen et al. [4], Perou et al. [8] and Bhattacharjee et al. [2] is that the number of variables is much larger than sample size. This type of data represented by gene expression microarray data is called high-dimension, low-sample-size (HDLSS) data. Substantial work about clustering has been done on HDLSS asymptotics in recent years. Liu et al. [6] proposed a two-way split clustering called "statistical significance of clustering(SigClust)" especially for HDLSS data. Ahn et al. [1] proposed a hierarchical divisive clustering and considered its high dimensional asymptotics. Huang et al. [5] developed the SigClust by Liu et al. [6] with soft thresholding approach. Yata and Aoshima [9] gave consistency properties of sample principal component scores and applied it to clustering under high dimensional settings. Nakayama et al. [7] investigated clustering by kernel principal component analysis for HDLSS data. Borysov et al. [3] studied behaviors of hierarchical clustering under several asymptotic settings from moderate dimension through HDLSS, nevertheless it is considered that theoretical assumptions are strict for HDLSS data due to having discussions on several asymptotic settings at once. Given this background, we focused on HDLSS settings and considered asymptotic properties of hierarchical clustering with several linkage functions.

In this talk, we investigated the hierarchical clustering theoretically in the HDLSS context as dimension goes to infinity while sample size is fixed. We gave asymptotic properties of hierarchical clustering and showed the threshold to decide the asymptotic behaviors. Finally, we deliberated performances of the hierarchical clustering in numerical simulations and actual data analyses.

References

[1] Ahn, J., Lee, M.H., Yoon, Y.J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, 22, 443–464.

[2] Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 13790–13795.

[3] Borysov, P., Hannig, J., Marron, J.S. (2014). Asymptotics of hierarchical clustering for growing dimension. *Journal of Multivariate Analysis*, 124, 465–479.

[4] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14863–14868.

[5] Huang, H., Liu, Y., Yuan, M., Marron, J.S. (2015). Statistical Significance of Clustering using Soft Thresholding. *Journal of computational and graphical statistics*, 24, 975–993.

[6] Liu, Y., Hayes, D.N., Nobel, A., Marron, J.S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103, 1281–1293.

[7] Nakayama, Y., Yata, K., Aoshima, M. (2021). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis*, 185, 104779.

[8] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406, 747–752.

[9] Yata, K., Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47, 899–921.

Mixed effects modeling of clustered extreme values

鹿児島大学大学院　桃木 光輝　　鹿児島大学大学院　吉田 拓真

1　はじめに

気象データから災害リスクを定量化することは，災害対策計画や都市開発の観点から重要である．言い換えると，それは災害を引き起こすような極端な値の発生確率を予測することに相当する．しかしながら，稀な事象を対象とするこの分野では，利用可能なデータが決して多くないという問題がある．極値統計学では，数学的に保証されたパラメトリックな分布を活用することで，このような状況下でもより精度の高い予測を実現する．一方，気象データの特徴は複数の気象観測所でデータが蓄積されている点である．本研究で着目する混合効果モデルは，全ての観測所のデータを一つのモデリングに統合してデータ全体の共通情報を引き出しつつ，各観測所のデータの分布の差異を効率的・効果的に予測することができる (Sugasawa and Kubokawa 2020)．そのため，混合効果モデルは，利用可能なデータが不足しがちな極値統計学との相性が非常に良い．本研究では，混合効果モデルを応用した新たな極値統計モデルを提案し，予測の不確実性を測るための漸近理論を確立した．

極値統計学において，最も重要視されているのは，極値指数（extreme value index）と呼ばれる，分布の裾の重さに関連するパラメータの予測である．先行研究には，ベイズ階層モデルを応用した手法があるが，実データ解析において極値指数の予測結果の大きな不確実性が指摘されていた (Dyrrdal et al. 2015)．本研究では，アプリケーションの範囲をリスクの予測が難しいとされるものに限定することで，極値指数のより詳細な解析が可能な手法を実現した．

また，その他の手法には，多変量極値統計学やコピュラなどがある（Davison et al. 2012）．これらは全ての観測所のデータの同時分布をパラメトリックに解析する手法であり，主に，観測所間のリスクの依存構造に焦点を当てている．しかしながら，多変量極値統計学の場合，多くの観測所のデータを同時にモデリングすることが困難である（Huster and Wadsworth 2020）．また，災害は複雑な要因により引き起こされると考えられるが，これらの手法は位置情報以外の共変量情報をモデリングに組み込むことが困難である．本研究では，災害リスクの地理的な依存構造よりも，災害の要因を明らかにするためのモデリングの開発に専念した．

2　モデル

本研究はクラスターデータ（clustered data）

$$\left\{(Y_{ij}, \boldsymbol{X}_{ij}) \in \mathbb{R}^+ \times \mathbb{R}^p,\ i = 1, 2, \cdots, n_j,\ j = 1, 2, \cdots, J\right\}$$

を対象としており，気象データはその一例である．J がクラスターの個数，n_j が各クラスター内のデータ数であり，$(Y_{ij}, \boldsymbol{X}_{ij})$ は j 番目のクラスターにおける i 番目の観測である．ただし，Y_{ij} は目的変数であり，$\boldsymbol{X}_{ij}$ は説明変数である．

U_j を未知の分散 σ_0^2 を持つ正規分布 $N(0, \sigma_0^2)$ に従う未観測の確率変数とする．このとき，$\boldsymbol{X}_{ij} = \boldsymbol{x}$ と $U_j = u_j$ が与えられた下での Y_{ij} の条件付き分布関数 $F(y|\boldsymbol{x}, u_j) = P(Y_{ij} \leq y|\boldsymbol{X}_{ij} = \boldsymbol{x}, U_j = u_j)$ に，パレート型分布（Pareto-type distribution）

$$1 - F(y|\boldsymbol{x}, u_j) = y^{-1/\gamma(\boldsymbol{x}, u_j)} \mathcal{L}(y, \boldsymbol{x}, u_j)$$

を仮定する．ここで，$\gamma(\boldsymbol{x}, u) > 0$ が極値指数であり，$\mathcal{L}(y, \boldsymbol{x}, u)$ は任意の $s > 0$ に対して $\lim_{y\to\infty} \mathcal{L}(ys, \boldsymbol{x}, u)/\mathcal{L}(y, \boldsymbol{x}, u) \to 1$ を満足するような関数である．この分布族には t 分布やパレート分布などの裾の重い分布が多く含まれる．分布の裾における挙動を決定する極値指数の予測が本質的に重要である．本研究では，極値指数に混合効果モデル

$$\log\left\{\gamma(\boldsymbol{x}, u_j)^{-1}\right\} = \alpha_0 + \boldsymbol{\beta}_0^\top \boldsymbol{x} + u_j, \quad j = 1, 2, \cdots, J$$

を仮定する．ここで，$\alpha_0 \in \mathbb{R}$ と $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ は未知の回帰係数である．上記モデルは，$U_j \equiv 0$ の古典的な線形モデルについて研究した Wang and Tsai (2009) の拡張である．α_0 と $\boldsymbol{\beta}_0$ は全クラスター共通のパラメータである．$U_j = u_j$ は変量効果と呼ばれ，これにより説明変数 $\boldsymbol{X}_{ij} = \boldsymbol{x}$ では説明のつかない影響，すなわち，クラスター毎の分布の違いを考慮できる．また，その実現値そのものを予測することも可能である．

講演では，研究を通して得られた提案モデルの数学的性質と数値パフォーマンスについて報告した．

参考文献

[1] Davison, A.C., Padoan, S.A., and Ribatet, M. (2012) Statistical modeling of spatial extremes. *Statistical Science*. **27** 161–186.

[2] Dyrrdal, A.V., Lenkoski, A., Thorarinsdottir, T.L., and Stordal, F. (2015) Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, **26** 89–106.

[3] Sugasawa, S., and Kubokawa, T. (2020) Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*, **3** 693–720.

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

植物個体群へ応用できる繰り返し可能な非定常クラスター点過程

島谷健一郎・統計数理研究所

野外環境下の植物集団には、種子散布が親個体から近距離に限られること、あるいは適した環境が限られることから、集中分布を示すものが多い。種子散布制限に起因する集中分布を表現するモデルの基本のひとつである Thomas 過程とは、

１．　母親はランダムに分布する（定常ポアソン過程に従う）。
２．　各母親はポアソン分布に従う個数の娘を生産する。
３．　娘は母親から正規分布に従って近傍に散布される。
４．　母親は死んで、娘たちが集中分布を示す。

というアルゴリズムで点分布を生成する空間点過程である。

Thomas 過程には、大きく２つの問題がある。

１．　定常性を仮定しているが、現実には生残率は環境に依存する。すなわち、非定常な点過程モデルにする必要がある。
２．　植物は繁殖を繰り返す。だから、親集団も集中分布だったと仮定するほうが自然である。その親集団も環境に依存した非定常な分布と考えるほうがさらに自然である。

要するに、繰り返し可能な非定常点過程である必要がある。

実際のところ、非定常で集中分布を生成するアルゴリズムは容易に作ることができる。しかし、1 次モーメント、2 次モーメントなどを数学として導出しないと、データからのパラメータ推定などの統計的推定を行えない。

非定常性を加えたネイマン・スコット過程は 2000 年代に入り、数多く提唱されており、パラメータ推定法も数多く試されている。しかし、その多くが親集団と娘集団の２世代かぎりのモデルで、かつ、娘集団への非定常性しか考慮していない。より広い非定常性を加えると、2 次モーメントを初等関数で書き下せないため、何らかの代替法でパラメータ推定を行うが、計算量が膨大になりがちである。

ここでは、集中分布する非定常な親集団から集中分布する非定常な娘集団を生成する点過程モデルで、2 次モーメントまで初等関数で書き表せるものを提唱した。

娘集団の 1 次モーメント $\rho_D^{(1)}(\mathbf{x})$、2 次モーメント $\rho_D^{(2)}(\mathbf{x}, \mathbf{y})$は、母親集団の 1 次モーメント $\rho_M^{(1)}(\mathbf{x})$、2 次モーメント $\rho_M^{(2)}(\mathbf{x}, \mathbf{y})$を用いて、以下の漸化式で表される。

$$\rho_D^{(1)}(x) = \int_A \rho_M^{(1)}(z)\mathbf{E}(u)d_{nl}(x;z,s^2)dz \cdot s(x)$$

$$\rho_D^{(2)}(x,y) =$$
$$(\int_A \rho_M^{(1)}(z,w)\mathbf{E}(u)d_{nl}(x;z,s^2)d_{nl}(y;z,s^2)dz + \int_A\int_A \rho_M^{(2)}(z,w)\mathbf{E}(u(u-1))d_{nl}(x;z,s^2)d_{nl}(y;w,s^2)dzdw)$$
$$\times s(x)s(y)$$

ここで、$d_{nl}(x;\mu,\sigma^2)$ は平均 μ, 分散 σ^2 の正規分布の確率密度関数で、母親から娘への散布に対応する。第 1 項は母親が共通の姉妹に対応し、第 2 項は母親が異なる 2 つの娘個体に対応する。u は娘数という確率変数、**E( )**は期待値を表す。s**(x)** は **x** での生残率で random thinning として点分布に作用する。多くの場合、m 個の環境条件 l_i**(x)**の線形回帰式に何らかのリンク関数を施した形で表される。

環境データはサンプリング地点で計測し、ガウスカーネル平滑化などで任意の点 **x** における環境条件 l_i**(x)**として用いることが多い。母親集団の 1 次モーメント、2 次モーメントが s**(x)**を含むので、それに正規分布を乗じた漸化式の積分は当然、初等関数では表せない。

ところが、ガウスカーネルによる平滑化を回帰式の後で行うという近似を行うと、生残率は混合正規分布の形で表わされ、上の漸化式は正規分布の convolution の公式で explicit に解けてしまう。結果として、繁殖後に生残率による random thinning という非定常性が入る繰り返し可能な非定常過程は、以下のような標準形を有することが示された。

$$\rho_D^{(1)}(x)$$
$$= \sum_k C_k^0 d_{nl}(x,q_k,Q_k^2)$$

$$\rho_D^{(2)}(x,y)$$
$$= \rho_D^{(1)}(x)\rho_D^{(1)}(y)$$
$$+ d_{nl}(x;y,S_1^2)\sum_k C_k^1 d_{nl}(x,v_k,T_k^2)d_{nl}(y,v_k,T_k^2)$$
$$+ d_{nl}(x;y,S_2^2)\sum_h C_h^2 d_{nl}(x,u_h,R_h^2)d_{nl}(y,u_h,R_h^2)$$
$$+\cdots$$

ランダムな親世代から始まった k 世代目なら…で略されている項は全部で k + 1 個になる。$Q, S, T, R, \ldots$は種子散布やカーネル平滑化で用いる正規分布の分散の関数、q, v, u は環境データのサンプリング地点（の関数）、C はそれらと生残率の中のパラメータの関数である。

スパース性の仮定のもとでの多項分布の適合度検定統計量選択の指針とその応用

北海道教育大・札幌　種市信裕　　北海道教育大・釧路　関谷祐里

1. ϕ-ダイバージェンスに基づく多項分布の適合度検定統計量.

ϕ-ダイバージェンス と呼ばれる，分布間の非類似度の尺度が，Csiszár [3] と Ali and Silvey [1] によって提案された。2 つの離散分布，$p=(p_1,\ldots,p_k)'$ と $q=(q_1,\ldots,q_k)'$ の間の ϕ-ダイバージェンス 測度は，$D_\phi(p,q)=\sum_{j=1}^k q_j\phi\,(p_j/q_j)$ によって定義される。ここで，$\phi(t)$ は $t>0$ に対して定義されるいくつかの性質を満たす実凸関数である。

$X=(X_1,\ldots,X_k)'$ を多項分布 $\mathrm{Mult}_k(n,\pi)$ に従う確率変数ベクトルとする。ここで，$\sum_{j=1}^k X_j=n$, $\sum_{j=1}^k \pi_j$=1, $0<\pi_j<1, (j=1,\ldots,k)$, であり，$\pi=(\pi_1,\ldots,\pi_k)'$ は未知の確率ベクトルである。ある固定された確率ベクトル $p=(p_1,\ldots,p_k)'$ に対して，帰無仮説 $H_0:\pi=p$ を検定するために，ϕ-ダイバージェンス統計量 K_ϕ が Zografos et al. [6] によって導入された。K_ϕ は $K_\phi=2nD_\phi\,(X/n,p)$ によって定義される。ここで，$\phi(1)=\phi'(1)=0$ と $\phi''(1)=1$ である (Pardo et al. [4])。

ϕ として，凸関数 $\phi_a(t)=\{a(a+1)\}^{-1}\{t^{a+1}-t+a(1-t)\}$ $(a\neq 0,-1);=t\ln t+1-t$ $(a=0);=-\ln t-1+t$ $(a=-1)$ を選ぶと，ϕ-ダイバージェンス測度は，パワーダイバージェンス測度になる (Read and Cressie [6])。ゆえに，ϕ-ダイバージェンス統計量 K_ϕ の族は，Cressie and Read [2] によって提案されたパワーダイバージェンス統計量 R^a の族を含んでいる。パワーダイバージェンス統計量 R^a の族が，Pearson の X^2 統計量 $(a=1)$, 対数尤度比統計量 $(a=0)$ 等を含んでいる。Zografos et al. [7] は，帰無仮説 H_0 のもとで，ϕ-divergence 統計量 K_ϕ はすべて，漸近的に自由度 $\nu=k-1$ のカイ二乗分布に従うことを示した。

2. Second-order correction term.

帰無仮説 H_0 のもとで，統計量の原点まわりの s 次モーメントを $E(K_\phi^s|H_0)=E(F_\nu^s)+\frac{m_\phi(s)}{n}+o(n^{-1}),\quad (s=1,2,\ldots)$, として評価することのできる統計量 K_ϕ を考える。ここで，F_ν は，自由度 ν のカイ二乗分布に従う確率変数とする。この時，$m_\phi(s)$ は，**second-order correction term** と呼ばれる。もし，$m_\phi(s)$ の絶対値が 0 に近ければ，その統計量の分布が自由度 ν のカイ二乗分布に近いと考えることができる。ゆえに，$m_\phi(s)$ の絶対値を計算することによって，どの統計量がカイ二乗分布に近いかを調べることができる。本報告では，ϕ-ダイバージェンス統計量の中で，帰無仮説のもとでカイ二乗分布に最も近い統計量を考える。

3. Second-order correction term に関する定理

K_ϕ の帰無分布のカイ二乗分布への収束の速さに関する定理を導く。帰無仮説 H_0 が検定統計量 K_ϕ の $m_\phi(s)$ に関して，以下の定理を得る。

定理 1. $\phi(t)$ を 4 回微分可能で，$\phi^{(4)}(t)$ が $t=1$ で連続であるとする。$p_i=O(k^{-1}),\ (i=1,\ldots,k)$ と仮定すると，$m_\phi(s)=0,\ (s=1,2,\ldots)$ を満たす ϕ に対して，H_0 のもとで，k を無限大に発散させると，$4\phi'''(1)+3\phi^{(4)}(1)$ は 0 に収束する。

帰無仮説 H_0 が，$p = (1/k, \ldots, 1/k)$ の場合，symmetric な帰無仮説と呼ばれる。定理 1 は，Pardo [5, p.183] が symmetric な帰無仮説のもとにおいて，$s = 1, 2, 3$ の場合に示した結果を，symmetric な帰無仮説を含めた，より一般的な帰無仮説と $s \geq 4$ を含めたすべての s に拡張したものである。ϕ として ϕ_a を用いることによってパワーダイバージェンス統計量に適用すると，次の系を得る。

系 1. $p_i = O(k^{-1}),\ (i = 1, \ldots, k)$ を仮定すると，パワーダイバージェンス統計量 R^a の族に対して，a についての 2 次方程式 $m_{\phi_a}(s) = 0,\ (s = 1, 2, \ldots)$ の 2 つの解は，H_0 のもとで，k を無限大に発散させると，$a = 1$ と $a = 2/3$ に収束する。

系 2. $p_i = O(k^{-1}),\ (i = 1, \ldots, k)$ と仮定すると，統計量 R^1 と $R^{2/3}$ の s 次モーメントの $m_\phi(s)$ は，$m_{\phi_1}(s) = A_s k^s + O(k^{s-1}),\ \ m_{\phi_{2/3}}(s) = B_s k^s + O(k^{s-1}),\ (s = 1, 2, \ldots)$ と評価される。ここで，$A_s = (1/2)s(s-1)\left\{(S/k^2) - 1\right\},\ \ B_s = (s/27)\left\{(s-1)(S/k^2) - 3(2s-3)\right\}\ (s = 1, 2, \ldots)$ である。ただし，$S = \sum_{i=1}^{k} 1/p_i$ とする。

系 2 より，$1 \leq S/k^2 < 33/29$ のとき，$|A_1| < |B_1|$ と $|A_s| < |B_s|,\ (s \geq 2)$ が成り立ち，他方，$39/29 \leq S/k^2$ のとき，$|A_s| > |B_s|,\ (s \geq 2)$ が成り立つことがわかる。

これらの議論より，もし，より速くカイ二乗極限分布に収束する検定統計量を選びたければ，パワーダイバージェンスに基づく統計量 R^a の中では，$R^{2/3}$ や R^1 を推奨する。さらに，多項分布の適合度検定における R^1 と $R^{2/3}$ の間の比較に関しては，$1 \leq S/k^2 < 33/29$ の場合には R^1 を推奨する。

Second-order correction term を用いたこのような議論は種々の分割表の独立性検定への応用が可能である。

4. スパース性の仮定

多項分布で，サンプルサイズ n が増加すると同時にカテゴリー数 k が増加するという条件や，カテゴリー数 k の増加に伴ってサンプルサイズ n も増加する条件，分割表で，サンプルサイズが増加するときセル数が増加するという条件は，多項分布のスパース性の仮定や分割表のスパース性の仮定と呼ばれる。スパース性の仮定は，実際の問題では，カテゴリー数に対してサンプルサイズが少ない場合や，セル数に対してサンプルサイズが少ない場合として解釈される。

本報告はこのような特徴のあるデータに対する統計量選択の指針となる。

参考文献

[**1**] Ali, S. M. and Silvey, D., J. Roy. Statist. Soc. B 28 (1966) 131–142. [**2**] Cressie, N. and Read, T. R. C., J. Roy. Statist. Soc. B 46 (1984) 440–464. [**3**] Csiszár, I., Studia Sci. Math. Hungar. 2 (1967) 299–318. [**4**] Pardo, L., Pardo, M. C. and Zografos, K., J. Japan Statist. Soc. 29 (1999) 213–228. [**5**] Pardo, L., Statistical inference based on divergence measures, Chapman & Hall/CRC, 2006. [**6**] Read, T. R. C. and Cressie, N. A. C., Goodness-of-fit statistics for discrete multivariate data, Springer-Verlag, New York, 1988. [**7**] Zografos, K., Ferentions, K. and Papaioannou, T., Commun. Statist.-Theory Meth. 19 (1990) 1785–1802.

Simultaneous estimation of multiplicative Poisson means in two-way contingency tables

Yuan-Tsung Chang (Mejiro University), Shinozaki Nobuo (Keio University)

Abstract

Shrinkage estimation of Poisson means is considered when observations are given in the form of a two-way contingency table. Assuming a multiplicative Poisson model, estimators which shrink to the specified values or an order statistic in one dimension and in two dimensions are considered and are shown to dominate the maximum likelihood estimator (MLE) under normalized squared error loss.

1 Introduction

We consider two-way multiplicative model where $x_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J$, are independent random Poisson random variables with means

$$\lambda_{ij} = \lambda \alpha_i \beta_j, \quad i = 1, \ldots, I,\ \ j = 1, \ldots, J,$$

where $\alpha_i \geq 0$ and $\beta_j \geq 0$ satisfy $\sum_{i=1}^{I} \alpha_i = 1$ and $\sum_{j=1}^{J} \beta_j = 1$, respectively. We denote the one-dimensional frequencies and the total frequency by

$$x_{i+} = \sum_{j=1}^{J} x_{ij},\ i = 1, \ldots, I,\ x_{+j} = \sum_{i=1}^{I} x_{ij},\ j = 1, \ldots, J,\ x_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}.$$

As discussed in Hara and Takemura (2006) complete sufficient statistics are $\boldsymbol{x}_1 = (x_{1+}, \ldots, x_{I+})$ and $\boldsymbol{x}_2 = (x_{+1}, \ldots, x_{+J})$. The MLE of λ_{ij} is

$$\hat{\lambda}_{ij}^{ML} = \begin{cases} \dfrac{x_{i+}x_{+j}}{x_{++}} & \text{if} \quad x_{++} \neq 0 \\ 0 & \text{if} \quad x_{++} = 0. \end{cases}$$

They have given a class of improved estimators which shrink the MLE toward the origin under the normalized squared error loss. The simple one is

$$\delta_{ij}^{HT} = \frac{x_{i+}x_{+j}}{x_{++}} \left\{ 1 - \frac{d}{x_{++} + d} \right\}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The following lemma is a special case of Lemma 2.1 of Hara and Takemura (2006) and is useful to evaluate the risk of the shrinkage estimators when normalized squared error loss is concerned.

Lemma 1.1. If $g(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a real-valued function satisfying $E|g(\boldsymbol{x}_1, \boldsymbol{x}_2)| < \infty$ and $g(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$ when $x_{i+} = 0$ or $x_{+j} = 0$, then

$$E\left\{ \frac{g(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\lambda_{ij}} \right\} = E\left\{ \frac{(x_{++} + 1)}{(x_{i+} + 1)(x_{+j} + 1)} g(\boldsymbol{x}_1 + \boldsymbol{e}_i^I, \boldsymbol{x}_2 + \boldsymbol{e}_j^J) \right\},$$

where $\boldsymbol{e}_i^I$ ($\boldsymbol{e}_j^J$) is $I \times 1$ ($J \times 1$) unit vector with i-th (j-th) component 1.

Next section we consider one-dimensional shrinkage to an order statistic or a specified point.

2 One-dimensional shrinkage to an order statistic or a specified point

First, we consider one-dimensional shrinkage to an order statistic.

Let $x_{(\ell)+}$ be the ℓ-th smallest observation among $x_{1+}, \ldots, x_{I+}$. We assume that $I \geq \ell + 2$ and consider the following estimator which shrinks x_{i+} toward $x_{(\ell)+}$ when $x_{i+} \geq x_{(\ell)+}$:

$$\delta_{ij}^{(1)} = \frac{x_{+j}}{x_{++}} \left\{ x_{i+} - \varphi(W) \frac{(x_{i+} - x_{(\ell)+})^+}{W + d} \right\}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

where $W = \sum_{i=1}^{I} (x_{i+} - x_{(\ell)+})^+$, $a^+ = \max(0, a)$ and d is a positive constant. Then we have the following.

Theorem 2.1. Suppose that $\varphi(W)$ is a non-decreasing function satisfying $0 \le \varphi(W) \le 2(I-\ell-1)$ and that $d \ge \sup \varphi(W)/2$. Then $\delta_{ij}^{(1)}, i=1,\ldots,I$ improves upon the MLE $\lambda_{ij}^{ML}, i=1,\ldots,I$ under the loss function $\sum_{i=1}^{I}(\hat{\lambda}_{ij}-\lambda_{ij})^2/\lambda_{ij}$ for any $j=1,\ldots,J$.

Next we consider the estimators shrink $\hat{\lambda}_{ij}^{ML}$ to a specified non-negative values, b_i.

Let $b_i \ge 0, i=1,\ldots,I$ be given numbers and we propose the following shrinkage estimator which shrinks x_{i+} to b_i when $x_{i+} \ge b_i$:

$$\delta_{ij}^{(2)} = \frac{x_{+j}}{x_{++}}\left\{x_{i+} - \varphi(N,W)\frac{(x_{i+}-b_i)^+}{W+d(N)}\right\}, \quad i=1,\ldots,I,\ j=1,\ldots,J,$$

where $W=\sum_{i=1}^{I}(x_{i+}-b_i)^+$ and $N=\#\{i|x_{i+}\ge b_i\}$. Then we have the following.

Theorem 2.2. Suppose that $\varphi(N,W)$ is a non-decreasing function of W and satisfies $0 \le \varphi(N,W) \le 2(N-1)^+$ for any $0 \le N \le I$. Suppose that $d(N) \ge \sup_W \varphi(N,W)/2$. Then $\delta_{ij}^{(2)}, i=1,\ldots,I$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}, i=1,\ldots,I$ under the loss function $\sum_{i=1}^{I}(\hat{\lambda}_{ij}-\lambda_{ij})^2/\lambda_{ij}$ for any $j=1,\ldots,J$.
It may be noticed that the shrinkage is made only when $N \ge 2$.

Remark Theorems 2.1 and 2.2 can be generalized directly to the case of Poisson multiplicative model for a multi-way contingency tables.

Next section we also consider two-dimensional shrinkage to order statistics or to a specified point.

3 Two-dimensional shrinkage to order statistics.

Let $x_{(\ell)+}$ and $x_{+(m)}$ be the ℓ-th and m-th smallest observation among $x_{1+},\ldots,x_{I+}$ and $x_{+1},\ldots,x_{+J}$, respectively. We assume that $I \ge \ell+2$ and $J \ge m+2$ and consider the estimator which shrinks x_{i+} toward $x_{(\ell)+}$ when $x_{i+} \ge x_{(\ell)+}$ in the first dimension and shrinks x_{+j} toward $x_{+(m)}$ when $x_{+j} \ge x_{+(m)}$ in the second dimension simultaneously. To improve upon the MLE $\hat{\lambda}_{ij}^{ML}$, we propose the following estimator :

$$\delta_{ij}^{(3)} = \frac{1}{x_{++}}\left\{x_{i+} - \varphi_1(W_1)\frac{(x_{i+}-x_{(\ell)+})^+}{W_1+d_1}\right\}\left\{x_{+j} - \varphi_2(W_2)\frac{(x_{+j}-x_{+(m)})^+}{W_2+d_2}\right\}, \quad i=1,\ldots,I,\ j=1,\ldots,J,$$

where $W_1=\sum_{i=1}^{I}(x_{i+}-x_{(\ell)+})^+$ and $W_2=\sum_{j=1}^{J}(x_{+j}-x_{+(m)})^+$ and d_1 and d_2 are positive constants. Then we have the following.

Theorem 3.1. Suppose that $\varphi_1(W_1)$ and $\varphi_2(W_2)$ are non-decreasing functions satisfying $0 \le \varphi_1(W_1) \le I-\ell-1$ and $0 \le \varphi_2(W_2) \le J-m-1$, respectively. If $d_1 \ge (I-\ell-1)/(I-\ell)\sup\varphi_1(W_1)$ and $d_2 \ge (J-m-1)/(J-m)\sup\varphi_2(W_2)$. Then $\delta_{ij}^{(3)}, i=1,\ldots,I, j=1,\ldots,J$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}$ under the loss function $\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\lambda}_{ij}-\lambda_{ij})^2/\lambda_{ij}$.

Next, we consider two-dimensional shrinkage to a specified point.

Let $b_i \ge 0, i=1,\ldots,I$ and $c_j \ge 0, j=1,\ldots,J$ be given numbers. Assuming that $I,J \ge 2$, we shrink x_{i+} to b_i when $x_{i+} \ge b_i$ and x_{+j} to c_j when $x_{+j} \ge c_j$. To improve upon the MLE $\hat{\lambda}_{ij}^{ML}$, we propose the following estimator

$$\delta_{ij}^{(4)} = \frac{1}{x_{++}}\left\{x_{i+} - \varphi_1(N_1,W_1)\frac{(x_{i+}-b_i)^+}{W_1+d_1(N_1)}\right\}\left\{x_{+j} - \varphi_2(N_2,W_2)\frac{(x_{+j}-c_j)^+}{W_2+d_2(N_2)}\right\}, \quad i=1,\ldots,I,\ j=1,\ldots,J,$$

where $W_1=\sum_{i=1}^{I}(x_{i+}-b_i)^+, W_2=\sum_{j=1}^{J}(x_{+j}-c_j)^+$, $N_1=\#\{i|x_{i+}\ge b_i, i=1,\ldots,I\}$ and $N_2=\#\{j|x_{+j}\ge c_i, j=1,\ldots,J\}$. Although it may be natural to put the condition $\sum_{i=1}^{I}b_i=\sum_{j=1}^{J}c_j$, we do not need it in the following.

Theorem 3.2. Suppose that $\varphi_i(N_i,W_i)$ is a non-decreasing function of W_i and satisfies $0 \le \varphi_i(N_i,W_i) \le (N_i-1)^+$ for any $N_i \ge 0$, and that $d_i(N_i) \ge (N_i-1)^+/N_i \sup_{W_i}\varphi_i(N_i,W_i)$, for any $N_i \ge 0$, $i=1,2$. Then $\delta_{ij}^{(4)}$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}$ under the loss function $\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\lambda}_{ij}-\lambda_{ij})^2/\lambda_{ij}$.
It may be noticed that the shrinkage in the i-th dimension is made only when $N_i \ge 2$.

For the rest of the content, please refer to "Simultaneous estimation of Poisson means in two-way contingency tables under normalized squared error loss", JJSDS (2022) vol.5, issue 2 p577-628.

正方分割表における点対称性からの隔たりを測る尺度の推定について

東京理科大学大学院 理工学研究科　中島 亮
東京理科大学大学院 理工学研究科　星野 瞭太
東京理科大学 理工学部　藤澤 健吾
東京理科大学 理工学部　田畑 耕治

1. APS モデルからの隔たりを測る尺度の推定について

行変数 X と列変数 Y が順序のある同じ分類からなる $R \times R$ 正方分割表において, (i, j) セル確率を p_{ij} $(i = 1, \ldots, R;\ j = 1, \ldots, R)$ とする. 点対称 (PS) モデルは次のように定義される (Wall and Lienert, 1976):

$$p_{ij} = p_{i^*j^*} \quad (1 \leq i, j \leq R),$$

ただし, $i^* = R + 1 - i$, $j^* = R + 1 - j$ である.

アナザー点対称 (APS) モデルは次のように定義される (Kurakami *et al.*, 2017):

$$p_{ij} = p_{i^*j^*} \quad (i + j \neq R + 1).$$

モデルが与えられたデータに適合しない場合, モデルからの隔たりを測る尺度に関心がある. Iki and Tomizawa (2019) は $p_{ij} + p_{i^*j^*} > 0$ $(i = 1, \ldots, R;\ j = 1, \ldots, R)$ を仮定し, APS モデルからの隔たりを測る尺度を次のように提案した:

$$\Phi_{APS} = \frac{1}{\Delta \log 2} \sum \sum_{i+j \neq R+1} p_{ij} \log \frac{2p_{ij}}{p_{ij} + p_{i^*j^*}},$$

ただし, $\Delta = \sum \sum_{i+j \neq R+1} p_{ij}$ である.

観測度数 n_{ij} がサンプル数 n $(n = \sum_i \sum_j n_{ij})$ の多項分布に従うと仮定し, p を $R^2 \times 1$ 多項確率ベクトルとする. すなわち,

$$p = (p_{11}, p_{12}, \ldots, p_{1R}, p_{21}, p_{22}, \ldots, p_{2R}, \ldots, p_{R1}, p_{R2}, \ldots, p_{RR})^t,$$

ここで t は転置を表す. $\hat{p}_{ij}$ を標本比率 $(\hat{p}_{ij} = n_{ij}/n)$ とし, p_{ij} を $\hat{p}_{ij}$ で置き換えたベクトルを $\hat{p}$ とする. Φ_{APS} の推定量 $\hat{\Phi}_{APS}$ の漸近バイアスを次のように与えた:

$$E(\hat{\Phi}_{APS} - \Phi_{APS}) = \frac{1}{2n} tr \left(\left[\frac{\partial^2 \Phi_{APS}}{\partial p \partial p^t} \right] (D(p) - pp^t) \right),$$

ここで $D(p)$ は p の i 番目の要素を i 番目の対角要素とする対角行列, tr は行列のトレースを表す. このとき, 次の推定量を提案した:

$$\tilde{\Phi}_{APS} = \hat{\Phi}_{APS} - \frac{1}{2n} tr \left(\left[\frac{\partial^2 \hat{\Phi}_{APS}}{\partial \hat{p} \partial \hat{p}^t} \right] (D(\hat{p}) - \hat{p}\hat{p}^t) \right),$$

ここで, $[\partial^2 \hat{\Phi}_{APS} / \partial \hat{p} \partial \hat{p}^t]$ は $[\partial^2 \Phi_{APS} / \partial p \partial p^t]$ の p_{ij} を $\hat{p}_{ij}$ で置き換えたものである.

2. RGS, CPS モデルからの隔たりを測る尺度の推定について

逆グローバル対称 (RGS) モデルは次のように定義される (Kurakami *et al.*, 2017):

$$\Delta_U = \Delta_L,$$

ただし, $\Delta_U = \sum\sum_{i+j<R+1} p_{ij}$, $\Delta_L = \sum\sum_{i+j>R+1} p_{ij}$ である.

条件付き点対称 (CPS) モデルは次のように定義される (Tomizawa, 1986):

$$p_{ij} = \tau p_{i^*j^*} \quad (i+j < R+1).$$

Iki and Tomizawa (2019) は RGS モデルと CPS モデルからの隔たりを測る尺度をそれぞれ次のように提案した:

$$\Phi_{RGS} = \frac{1}{\log 2}\left(\Delta_U^c \log\frac{\Delta_U^c}{1/2} + \Delta_L^c \log\frac{\Delta_L^c}{1/2}\right),$$
$$\Phi_{CPS} = \frac{1}{\Delta \log 2}\sum_{i+j<R+1}\sum\left(p_{ij}\log\frac{\Delta p_{ij}^*}{\Delta_U} + p_{i^*j^*}\log\frac{\Delta p_{i^*j^*}^*}{\Delta_L}\right),$$

ただし, $\Delta_U^c = \Delta_U/\Delta$, $\Delta_L^c = \Delta_L/\Delta$, $p_{ij}^* = p_{ij}/(p_{ij}+p_{i^*j^*})$ である.

APS モデルの場合と同様にして, 尺度の推定量をそれぞれ次のように提案した:

$$\tilde{\Phi}_{RGS} = \hat{\Phi}_{RGS} - \frac{1}{2n} tr\left(\left[\frac{\partial^2 \hat{\Phi}_{RGS}}{\partial\hat{p}\partial\hat{p}^t}\right](D(\hat{p}) - \hat{p}\hat{p}^t)\right),$$
$$\tilde{\Phi}_{CPS} = \hat{\Phi}_{CPS} - \frac{1}{2n} tr\left(\left[\frac{\partial^2 \hat{\Phi}_{CPS}}{\partial\hat{p}\partial\hat{p}^t}\right](D(\hat{p}) - \hat{p}\hat{p}^t)\right).$$

$\hat{\Phi}_{APS}$, $\hat{\Phi}_{RGS}$, $\hat{\Phi}_{CPS}$ の漸近バイアスを n 倍したものを Λ_{APS}, Λ_{RGS}, Λ_{CPS} とする. このとき, 次の定理及び系を得た:

定理 1. Λ_{APS} は Λ_{RGS} と Λ_{CPS} の和に等しい.

系 1. $\tilde{\Phi}_{APS}$ は $\tilde{\Phi}_{RGS}$ と $\tilde{\Phi}_{CPS}$ の和に等しい.

定理 2. $\hat{\Phi}_{APS}$, $\hat{\Phi}_{RGS}$, $\hat{\Phi}_{CPS}$ の漸近バイアスの比率は $R(R-1)/2 : 1 : (R+1)(R-2)/2$ である.

参考文献

[1] Iki, K. and Tomizawa, S. (2019). Measure of departure from point symmetry and decomposition of measure for square contingency tables. *Journal of Statistical Theory and Applications*, **19**, 526-533.

[2] Kurakami, H., Negishi, K., and Tomizawa, S. (2017). On decomposition of point-symmetry for square contingency tables with ordered categories. *Journal of Statistics: Advances in Theory and Applications*, **17**, 33-42.

[3] Tomizawa, S. (1986). Four kinds of symmetry models and their decompositions in a square contingency table with ordered categories. *Biometrical Journal*, **28**, 387-393.

[4] Wall, K. D. and Lienert, G. A. (1976). A test for point-symmetry in J-dimensional contingency-cubes. *Biometrical Journal*, **18**, 259-264.

正方分割表におけるモーメントに基づく周辺同等性からの隔たりを測る尺度

足立 匠 [1]　　安藤 宗司 [2]　　田畑 耕治 [2]

[1] 東京理科大学大学院 理工学研究科 情報科学専攻
[2] 東京理科大学 理工学部 情報科学科

順序カテゴリ $r \times r$ 正方分割表において，行変数と列変数をそれぞれ X と Y とし，(i,j) セル確率を $p_{ij} = \Pr(X = i, Y = j)$ とする $(i = 1, \ldots, r; j = 1, \ldots, r)$．正方分割表解析では，観測度数が主対角セルに集中する傾向があるため，行変数と列変数の独立性は成り立たないことが多い．そのため，独立性に代わり，行変数と列変数の対称性に関して関心がある．行変数と列変数の対称性を表す代表的なモデルとして，周辺同等 (MH) モデル (Stuart [1]) がある．

MH モデルは次のように表される．

$$p_{i\cdot} = p_{\cdot i} \quad (i = 1, \ldots, r),$$

ただし，$p_{i\cdot} = \sum_{t=1}^{r} p_{it}$，$p_{\cdot i} = \sum_{s=1}^{r} p_{si}$ である．MH モデルは様々な表現があることが知られている．行変数 X と列変数 Y の周辺累積確率を用いると，MH モデルは次のようにも表される．

$$F_i^X = F_i^Y \quad (i = 1, \ldots, r-1),$$

ただし，

$$F_i^X = \sum_{s=1}^{i} p_{s\cdot} = \Pr(X \le i), \quad F_i^Y = \sum_{t=1}^{i} p_{\cdot t} = \Pr(Y \le i).$$

周辺累積確率 F_i^X と F_i^Y の差を考えると，MH モデルは次のようにも表される．

$$G_{1(i)} = G_{2(i)} \quad (i = 1, \ldots, r-1), \tag{1}$$

ただし，

$$G_{1(i)} = \sum_{s=1}^{i} \sum_{t=i+1}^{r} p_{st} = \Pr(X \le i, Y \ge i+1), \quad G_{2(i)} = \sum_{s=i+1}^{r} \sum_{t=1}^{i} p_{st} = \Pr(X \ge i+1, Y \le i).$$

与えられたデータに対して MH モデルの当てはまりが悪い場合，(i) MH モデルよりも制約の弱いモデルを適用すること，(ii) MH モデルの当てはまりが悪い原因を分析すること，(iii) MH モデルからの隔たりの程度を測ることに関心がある．MH モデルの表現として式 (1) に注目して，これら (i) から (iii) に関する先行研究を述べる．

式 (1) に基づく周辺非同等モデルとして，Tahata and Tomizawa [2] は，m 次パラメータ周辺同等 (MH(m)) モデルを提案した．既知の m $(m = 1, \ldots, r-1)$ に対して，MH(m) モデルは次のように定義される．

$$G_{1(i)} = \prod_{k=0}^{m-1} \psi_k^{i^k} G_{2(i)} \quad (i = 1, \ldots, r-1).$$

$\psi_0 = \psi_1 = \cdots = \psi_{m-1} = 1$ のとき，MH(m) は MH モデルに一致する．MH(1) モデルは拡張周辺同等 (EMH) モデル (Tomizawa [3])，MH(2) モデルは一般化周辺同等モデル (Tomizawa [4]) に一致する．

式 (1) に基づく MH モデルよりも制約の弱いモデルとして，Tahata and Tomizawa [2] は，k 次モーメント周辺一致 (k-MME) モデルを提案した．既知の正の整数 k に対して，k-MME モデルは次のように定義される．

$$E[X^k] = E[Y^k],$$

ただし，

$$E[X^k] = \sum_{s=1}^{r} s^k p_{s\cdot}, \quad E[Y^k] = \sum_{t=1}^{r} t^k p_{\cdot t}.$$

さらに，Tahata and Tomizawa [2] は，k-MME モデルの別表現を次のように与えた．

$$\sum_{i=1}^{r-1} \left[(i+1)^k - i^k\right] G_{1(i)} = \sum_{i=1}^{r-1} \left[(i+1)^k - i^k\right] G_{2(i)}.$$

MH モデルの当てはまりが悪い原因を分析するために，Tahata and Tomizawa [2] は，「MH モデルが成り立つことと，全ての $k = 1, \ldots, r-1$ に対して k-MME モデルが成り立つことは必要十分である」という分解定理を与えた．この分解定理から，どのモーメントが一致していないことにより MH モデルの当てはまりが悪くなったのかを特定することが可能になる．

Tomizawa, Miyamoto and Ashihara [5] は，$G_{1(i)} + G_{2(i)} > 0$ $(i = 1, \ldots, r-1)$ を仮定し，式 (1) に基づく MH モデルからの隔たりの程度を測る尺度を次のように提案した．

$$\Psi = \frac{1}{\log 2} \sum_{i=1}^{r-1} \left[G^*_{1(i)} \log \frac{G^*_{1(i)}}{Q^*_i} + G^*_{2(i)} \log \frac{G^*_{2(i)}}{Q^*_i} \right],$$

ただし，

$$G^*_{1(i)} = \frac{G_{1(i)}}{\Delta}, \quad G^*_{2(i)} = \frac{G_{2(i)}}{\Delta}, \quad Q^*_i = \frac{G^*_{1(i)} + G^*_{2(i)}}{2}, \quad \Delta = \sum_{i=1}^{r-1} (G_{1(i)} + G_{2(i)}).$$

尺度 Ψ は次の性質がある．

- $0 \leq \Psi \leq 1$,
- $\Psi = 0 \Leftrightarrow G_{1(i)} = G_{2(i)}$ $(i = 1, \ldots, r-1) \Leftrightarrow$ MH モデル,
- $\Psi = 1 \Leftrightarrow$「$G_{1(i)} = 0$ かつ $G_{2(i)} > 0$」または「$G_{1(i)} > 0$ かつ $G_{2(i)} = 0$」$(i = 1, \ldots, r-1)$.

分解定理は，MH モデルの当てはまりが悪い原因を特定するには有用であるが，MH モデルからの隔たりの程度を測ることは難しい．一方，尺度 Ψ は MH モデルからの隔たりの程度を測ることには有用であるが，MH モデルの当てはまりが悪い原因を特定することは難しい．本講演では，MH モデルの当てはまりが悪い原因を特定することが可能な MH モデルからの隔たりの程度を測る尺度を提案した．提案尺度により，与えられたデータに対して MH モデルの当てはまりが悪い場合，そのデータの特徴をより詳細に分析することが可能になると期待される．

参考文献

[1] A. Stuart. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3/4):412–416, 1955.

[2] K. Tahata and S. Tomizawa. Generalized marginal homogeneity model and its relation to marginal equimoments for square contingency tables with ordered categories. *Advances in Data Analysis and Classification*, 2(3):295–311, 2008.

[3] S. Tomizawa. Diagonals-parameter symmetry model for cumulative probabilities in square contingency tables with ordered categories. *Biometrics*, 49(3):883–887, 1993.

[4] S. Tomizawa. A generalization of the marginal homogeneity model for square contingency tables with ordered categories. *Journal of Educational and Behavioral Statistics*, 20(4):349–360, 1995.

[5] S. Tomizawa, N. Miyamoto, and N. Ashihara. Measure of departure from marginal homogeneity for square contingency tables having ordered categories. *Behaviormetrika*, 30(2):173–193, 2003.

国民生活基礎調査による
日本の所得分布と所得格差

西埜晴久 *

17 Dec. 2022†

1　はじめに

国民生活基礎調査は，厚生労働省が行っている統計である．昭和 61(1986) 年から始まり，3 年ごとに大規模調査が行われている．近年では，2019 年，2016 年，2013 年，2010 年が大規模調査が行われた年である．大規模調査の年の間の各年に簡易調査が行われている．なお，2020 年調査の簡易調査は調査が中止されている．

大規模調査の年では，世帯票，健康票，所得票などがある．そのうちの所得票の個票データを用いて所得分布の分析を行う．直近の大規模調査は 2019 年であり，その際の集計客体数は世帯票・健康票で 217,179 世帯，所得票で 22,288 世帯である．

以下に公表されている所得票の数値を掲げる．

Table 1: 所得票の数値

年	調査期間	集計客体数	平均所得	中央値
1998 年（H10）	H9 年 1 月〜12 月	30506	657.7	536
2010 年（H22）	H21 年 1 月〜12 月	26115	549.6	438
2019（R1）	H30 年 1 月〜12 月	22288	552.3	437

なお，平均所得，中央値は総世帯で単位は (万円) である．

2　所得分布の推定

1998 年の個票データの総所得から平均値，中央値，ジニ係数を直接計算すると平均値 657.96, 中央値 536, ジニ係数 0.4042 となる．これらの数値は公表されている値とほぼ同じである．また，個票データを用いて，一般化ベータ分布，Singh-Maddala および第 2 種の一般化ベータ分布の所得分布に用いられる確率分布を最尤法で推定し，パラメータの推定値を得ることが出来た．

拡大乗数について：2000 年代以降の調査では，拡大乗数を用いて調整することで，平均所得などを計算している．拡大乗数とは，ある地域の国勢調査で得られる世帯数と所得票の調査世帯数との比である．実際には県別でかつ県内でも指定都市とそれ以外に分けて拡大乗数を算出している．一般的には，都市部の世帯の標本サイズが郡部の世帯の標本サイズに比べ小さいので，拡大乗数が大きくなる傾向がある．つまり，拡大乗数とは，各地域に対してウェイトを与えていることになるので，そのウェイト（拡大乗数）を考慮して対数尤度を求めた方が良い．

Table 2 には個票データ（個別）と 10 分位のデータで推定したジニ係数を掲げた．

*広島大学大学院人間社会科学研究科

†科研費シンポジウム「統計科学の開拓」金沢大学サテライト・プラザ

Table 2: ジニ係数

	H10		H22		R01	
直接	0.40420		0.40326		0.40910	
	個別	10 分位	個別	10 分位	個別	10 分位
GG	0.4096	0.4033	0.4060	0.3912	0.4120	0.4037
SM	0.4011	0.3835	0.4005	0.3970	0.4065	0.3987
GB2	0.3991	0.3824	0.4016	0.3908	0.4076	0.3954

Table 3: ξ の推定結果

	H10	H22	R01
	ξ の推定値		
上位 1%	0.2124	0.4056	0.1899
上位 0.5%	0.0418	0.1312	0.1952
集計客体数	30506	26115	22288
5000 万円以上	72	21	20

3 所得分布の上裾の推定

Moriguchi and Saez (2008)[1] は，税務統計より所得の上位 1%が占める割合と，上位 0.1%の占める割合を調べた．第二次世界大戦前の不平等度は高かったが，戦後の不平等度は小さく，米国が 1980 年代以降上位所得者の割合が上昇したのに比べ日本はそれほど上昇していないことを示した．そこでこの場合には分布の上裾がどうなっているかに関心がある．よって分布の上裾の部分を．一般化パレート分布 $GPD(\xi, \sigma, u)$ を当てはめて形状パラメータ ξ の推定を行った．形状パラメータ ξ の推定結果は Table 3 に掲げてある．

4 今後の課題

GB2 の推定は不安定なので，初期値の選び方などもう少し安定して推定する方法を考えたい．また，今回取り上げた以外の分布を推定する．

所得票のデータと世帯票のデータを接続することで各世帯の世帯数を把握することが出来る．世帯数が分かると所得を世帯人数の平方根で割った等価所得を求めることが出来る．年につれて世帯人員数（世帯の構成員数）が減少する傾向があるので，等価所得を用いると世帯人数の減少の影響を除いて所得の変化を調べることができる．

ただし，等価可処分所得については，所得票に公表されている年もある．「等価可処分所得」とは，下記によって算出される．

等価可処分所得 ＝ （当初所得 ＋ 社会保障給付 − 税金 − 社会保険料 − 掛金など） ÷ √世帯人員数

[1] 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.

操作変数法の理解へ：計量生物と計量経済の邂逅

(Toward understanding the Instrumental Variables Method in Biometrics and Econometrics)

2022 年 12 月 17 日

国友直人[1]

鍵言葉 (Key Words): 統計的因果推論, 臨床試験, 政策評価, Noncompliance, ATE, LATE, 操作変数法, 構造方程式, 識別性, TSLS, LIML, Mendelian Randomization, 遺伝子疫学

要約 ： 因果関係 (causality) は統計科学を含め諸科学にとっては基本的かつ重要な分析対象である。計量生物と計量経済の分野ではこの間、統計的因果推論 (statistical causal inference) が盛んに応用されている。本報告ではまず Rubin (1974) に始まる反実仮想 (counter-factual) モデルと Angrist, Imbens and Rubin (1996, 略して AIR) による操作変数法 (instrumental variables method) の応用の意味を説明した。次に伝統的な計量経済学 (econometrics) における同時方程式と構造方程式 (structural equation) を簡単な例を用いて説明し、構造方程式を用いた統計的因果関係の解釈を述べ、その統計的推定法を議論した。構造方程式の推定では OLS 法 (最小二乗法) は一致性を持たないので、操作変数法 (IV 法) としての Wald 法、LIML(制限情報最尤法, 分散比最小法)、TSLS(2 段階最小二乗法)、GMM(一般化積率法) などの長所と短所を説明し、さらに計量生物と計量経済などにおける統計的因果分析のさらなる課題を展望した。

また最後に遺伝子疫学における MR(Mendelian Randomization, メンデル・ランダム化解析) における操作変数法の利用に関する最近の展開についても議論した。

なお、時間の関係で十分な説明ができずに省略したところも少なくなかった。研究報告の詳細な内容を知りたい場合には未定稿ではあるが、「操作変数法の理解へ：計量生物と計量経済の邂逅」
http://www.kunitomo-lab.sakura.ne.jp/2022-11-4DP.pdf
を参照されたい。

[1]特任教授, 統計数理研究所 〒 190-8562 東京都立川市緑町 10-3

多変量分布間の回帰モデル

東京大学経済学研究科 岡野遼

東京大学総合文化研究科 今泉允聡

1. はじめに

説明変数と結果変数が共に分布の形で与えられるような回帰を考える. このような回帰は分布間回帰（distribution regression）と呼ばれ, 複雑データ解析の一分野として近年関心がもたれている. 例えば, ある年のある国における年齢別死亡率は, 横軸を年齢, 縦軸を死亡率とする密度関数として表示することで, 一次元分布に値をとるデータとみなすことができる. [1], [2] では世界 37 各国の 2013 年の年齢別死亡率分布を 1983 年のそれに回帰することで, これらの国々の年齢別死亡率分布が時間を経てどのように変化したかを分析している.

$\mathcal{W}$ を $\mathbb{R}^d$ 上の分布で有限な 2 次モーメントを持つもの全体とし, $\mathcal{W}$ に 2-Wasserstein 距離 d_W を与えてできる距離空間 $(\mathcal{W}, d_W)$ を Wassersetin 空間と呼ぶ. $\mathcal{F}$ を $\mathcal{W} \times \mathcal{W}$ 上の分布とし, $(\nu_1, \nu_2) \sim \mathcal{F}$ とする. 分布間回帰は ν_1 を説明変数, ν_2 を結果変数とするような回帰問題として定式化される.

分布間回帰のモデルは近年いくつか提案されているが, それらのほとんどは分布の次元を $d = 1$ の場合に限定している. 例えば, [1] は Wasserstein 空間の tangent bundle, [2] は最適輸送写像を用いた分布間回帰モデルをそれぞれ提案しているが, これらのモデルは $d = 1$ のときに最適輸送問題が closed form な解を持つことを利用しており, $d \geq 2$ の一般の分布に対してそのまま拡張することは困難である. 本研究では, 多変量分布であっても分布のクラスを適切に制限すれば, 最適輸送問題が closed form な解な解を持つことに注目し, 新たな分布間の回帰モデルを提案する.

2. 背景：ガウス分布間の最適輸送問題

以下では簡単のため, 分布のクラスをガウス分布族を制限した場合を考えるが, より一般にある種の楕円分布に制限しても同様に議論ができる. $\mathcal{G}$ を $\mathbb{R}^d$ 上の Gauss 分布全体とする. 非退化な共分散行列を持つ二つのガウス分布 $\mu_1 = N(m_1, \Sigma_1), \mu_2 = N(m_2, \Sigma_2) \in \mathcal{G}$ が与えられた時, それらの間の最適輸送写像 $t_{\mu_1}^{\mu_2}$ 及び Wasserstein 距離 $d_W(\mu_1, \mu_2)$ はそれぞれ以下のように陽に表されることが知られている：

$$t_{\mu_1}^{\mu_2}(x) = m_2 + \Sigma_1^{-1/2}[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}]^{1/2}\Sigma_1^{-1/2}(x - m_1),$$
$$d_W(\mu_1, \mu_2) = \sqrt{\|m_1 - m_2\|^2 + \mathrm{tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}]}.$$

また, Σ_* を非退化な共分散行列とすると, $\mu_* = N(m_*, \Sigma_*) \in \mathcal{G}$ における空間 $(\mathcal{G}, d_W)$ の接空間は, 内積空間 $T_{\mu_*} = (\mathbb{R}^d \times \mathrm{Sym}(d), G_{\mu_*})$ によって与えられることが知られている ([3]). ここで, $\mathrm{Sym}(d)$ はサイズ $d \times d$ の対称行列全体で, $z = (a, V), w = (b, W) \in \mathbb{R}^d \times \mathrm{Sym}(d)$ に対して, それらの間の内積を $G_{\mu_*}(z, w) = a^\top b + \mathrm{tr}(V\Sigma_* W)$ によって定めている. さらに, μ_* における exponential map $\exp_{\mu_*} : T_{\mu_*} \to \mathcal{G}$ は $\exp_{\mu_*}(a, V) = N(a + m_*, (V + I)\Sigma_*(V + I))$ によって与えられ, μ_* における logarithmic map $\log_{\mu_*} : \mathcal{G} \to T_{\mu_*}$ は, $\log_{\mu_*} N(m, \Sigma) = (m - m_*, \Sigma_*^{-1/2}[\Sigma_*^{1/2}\Sigma\Sigma_*^{1/2}]^{1/2}\Sigma_*^{-1/2} - I)$ によって与えられる. 一般に, 多変量分布間の最適輸送問題は解を陽に表すことができないが, 分布のクラスをガウス分布族 $\mathcal{G}$ に制限することで, このように最適輸送写像, Wasserstein 距離及び接空間の点との対応を陽に与えることが可能となる.

3. 提案するモデル

$\mathcal{F}$ を $\mathcal{G} \times \mathcal{G}$ 上の同時分布とし, $(\nu_1, \nu_2) \sim \mathcal{F}, \nu_j = N(m_j, \Sigma_j), j = 1, 2$ とする. また, 各 $j = 1, 2$ に対し, ν_j の Fréchet 平均を $\nu_{j\oplus} = N(m_{j\oplus}, \Sigma_{j\oplus})$ とする. 以下では, ν_1 を説明変数, ν_2 を結果変数とするような, $\mathcal{G}$ から $\mathcal{G}$ への回

帰を考える. [1] と同様に空間 $(\mathcal{G}, d_W)$ の幾何を用いて, ガウス分布 ν_1 及び ν_2 をそれぞれ非線形制約のない行列に変換し, 分布値データ間の回帰を行列データ間の回帰に帰着させるというアプローチを取る.

まず, 各 $j = 1, 2$ に対し, Fréchet 平均 $\nu_{j\oplus}$ において $(\mathcal{G}, d_W)$ の接空間をはり, 分布 $\nu_j = N_j(m_j, \Sigma_j)$ を $\log_{\nu_{j\oplus}} \nu_j \in S_d$ という行列に変換する. ここで, S_d は $S_d = \{(a, B) \in \mathbb{R}^{d\times(d+1)} : a \in \mathbb{R}^d, B \in \mathrm{Sym}(d)\}$ という非線形制約の無い行列の集合である. そして, 変換後の行列 $X = \log_{\nu_{1\oplus}} \nu_1$ と $Y = \log_{\nu_{2\oplus}} \nu_2$ の間に, 線形回帰モデル

$$Y = \langle X, \mathbb{B}\rangle + E, \quad \mathbb{E}[E|X] = 0$$

を仮定する. ここで, 4 次のテンソル $\mathbb{B} \in \mathbb{R}^{d\times(d+1)\times d\times(d+1)}$ が回帰パラメータであり, $\langle X, \mathbb{B}\rangle \in \mathbb{R}^{d\times(d+1)}$ は contracted tensor product と呼ばれる行列とテンソルの間の積である. さらに, 分布の次元 d が大きい時は, 回帰係数 $\mathbb{B}$ の要素数は膨大になるため, $\mathbb{B}$ は低ランクを持つことを仮定する. 最後に, 空間 $(\mathcal{G}, d_W)$ とその接空間の間の対応が一対一になるために, 確率 1 で $\langle X, \mathbb{B}\rangle \in \log_{\nu_{2\oplus}} \mathcal{G}$ となることも仮定する.

4. 推定量の構成とその性質について

実際には確率分布が直接観測されることは稀で, 我々は各分布からの離散観測のみが得られる場合が多い. そこで, 以下ではまずガウス分布のペア $(\nu_{1i}, \nu_{2i}) \sim \mathcal{F}, i = 1, ..., n$ が潜在的に生成され, その後潜在的な各分布から d 次元ベクトル $W_{jir} \sim \nu_{ji}, r = 1, ..., N$ が離散観測されるという設定を考える. 回帰パラメータ $\mathbb{B}$ が識別されるようにパラメータ空間 Θ をとり, 観測値 $W_{jir} \sim \nu_{ji}, r = 1, ..., N, i = 1, ..., n, j = 1, 2$ に基づいて $\mathbb{B}$ を推定する.

推定量の構成法は以下の通りである. まず, 離散観測 $W_{jir}, r = 1, ..., N$ を用いて, 潜在的な分布 ν_{ji} の推定量 $\hat{\nu}_{ji} = N_j(\hat{m}_{ji}, \hat{\Sigma}_{ji})$ を構成し, 推定された分布 $\hat{\nu}_{ji}, i = 1, ..., N$ を用いて, 経験 Fréchet 平均 $\hat{\nu}_{j\oplus}$ を計算する. 次に, 経験 Fréchet 平均 $\hat{\nu}_{j\oplus}$ で接空間をはり, 推定された分布 $\hat{\nu}_{ji}$ を $\hat{X}_i = \log_{\hat{\nu}_{1\oplus}} \hat{\nu}_{1i} \in S_d, \hat{Y}_i = \log_{\hat{\nu}_{2\oplus}} \hat{\nu}_{2i} \in S_d$ という行列に変換する. 最後に, 規準関数を $\hat{M}_{n,N}(\mathbb{B}) = n^{-1} \sum_{i=1}^n \|\hat{Y}_i - \langle \hat{X}_i, \mathbb{B}\rangle\|_F^2$ とおき, 最小二乗法 $\arg\min_{\mathbb{B}\in\Theta} \hat{M}_{n,N}(\mathbb{B})$ によってパラメータを推定する. ここで, $\|\cdot\|_F$ はフロベニウスノルムを表す.

こうして得られる推定量について, 一致性, 収束レート及び漸近正規性に関する以下の結果が示せる.

Theorem 1　真の Fréchet 平均 $\nu_{1\oplus}, \nu_{2\oplus}$ が既知であると仮定し, その状況で得られる推定量を $\hat{\mathbb{B}}_{n,N}$ とする. また, パラメータの真値を $\mathbb{B}_0$ とする. この時, いくつかの正則条件のもと, $\|\hat{\mathbb{B}}_{n,N} - \mathbb{B}_0\|_F = O_p(n^{-1/2} + N^{-1/4})$ が成立. さらに, N は n の数列で, $N(n) = n^q (q > 2)$ あれば, $n \to \infty$ のとき $\sqrt{n}(\mathrm{vec}(\hat{\mathbb{B}}_{n,N}) - \mathrm{vec}(\mathbb{B}_0))$ はある正規分布 $N(0, V)$ に分布収束する.

発表ではこの他, Calgary weather data を用いた実データの解析例なども紹介した.

参考文献

[1] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1-14, 2021.

[2] Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika (to appear), available at arXiv preprint arXiv:2104.09418.*

[3] Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005 – 1026, 2011.

GMANOVA モデルでの仮定緩和のさらなる可能性

(報告書)

中京大学 教養教育研究院 永井 勇

n 個の各個体に対して, 全ての個体で測定時点を揃った状態で p 回測定して得られる経時測定データと呼ばれるデータの分析を本講演では考えた. このようなデータの分析の主な目的は, データに隠れている経時変動と呼ばれる時間的な変動を上手く捉えることである. そのための分析には, Pothoff and Roy (1964) で提案された次の一般化多変量分散分析 (GMANOVA) モデルがよく使われる;

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

ここで $\mathbf{1}_n$ は全ての成分が 1 の n 次元ベクトル, $\boldsymbol{Y}$ は各行が各個体で測定して得られた経時測定データからなる $n \times p$ 行列, $\boldsymbol{A}$ は各行が各個体の特徴を表す測定時点に無関係な k 個の説明変数のデータからなる $n \times k$ 説明変数行列とし, $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ ($\mathbf{0}_r$ は全ての成分が 0 の r 次元ベクトル) を満たしているとし, $\boldsymbol{X}$ は後述のように, q は解析者が用いる関数によって決まるような, 各行が測定時点の関数からなる $p \times q$ 行列であり, これらは既知である. また, $r_{\boldsymbol{A}} = \mathrm{rank}(\boldsymbol{A})$, $r_{\boldsymbol{X}} = \mathrm{rank}(\boldsymbol{X})$ とする. さらに, $\boldsymbol{\mu}$ は q 次元の未知ベクトル, $\boldsymbol{\Xi}$ は $k \times q$ の未知行列であり, $\boldsymbol{\mathcal{E}}$ は $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ の $n \times p$ の誤差行列とし, $\boldsymbol{\Sigma}$ は未知の $p \times p$ 正定値行列とする. ここで, $\boldsymbol{A}$ が満たしている $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ は, 各説明変数 (各列) ごとにそれぞれ中心化されていることを表している.

このモデルで, p 回の測定時点を $t_1, \ldots, t_p$ ($t_1 < t_2 < \cdots < t_p$) とし, 例えば $\boldsymbol{X}$ の i 行目を $(t_i^0, t_i^1, \ldots, t_i^{q-1})$ とすると, $\boldsymbol{\mu}$ や $\boldsymbol{\Xi}$ は $t_1, \ldots, t_p$ の $(q-1)$ 次多項式の係数からなるベクトルと行列をそれぞれ表しており, これらを推定することが測定時点 $t_1, \ldots, t_p$ の $(q-1)$ 次多項式を用いて経時変動 $E[\boldsymbol{Y}]$ を推定することに対応していることを講演した. また, 過剰適合の問題もあるが, より柔軟な関数を $\boldsymbol{X}$ に用いて経時変動を推定することも可能であり, その際は, 用いる関数の重み付き和で経時変動を推定することとなり, $\boldsymbol{\mu}$ や $\boldsymbol{\Xi}$ は重みの部分に対応していることも講演で触れた. これらのことから分かるように, GMANOVA モデルであるモデル (1) における未知のベクトル $\boldsymbol{\mu}$ や行列 $\boldsymbol{\Xi}$ を推定することで経時変動が推定できる.

したがって, この GAMOVA モデルにおいて, 経時測定データからなる行列 $\boldsymbol{Y}$ や各個体の説明変数を各行に持つ説明変数行列 $\boldsymbol{A}$ などから, 未知の $\boldsymbol{\mu}$, $\boldsymbol{\Xi}$ を推定することが, 経時測定データの分析における目的である経時変動の推定において重要であることから, 本講演ではこれらを $r_{\boldsymbol{A}}$ や $r_{\boldsymbol{X}}$ の様々な状況における推定について着目した. これらの推定は, 次のリスク関数を最小にすることでよく行われている;

$$R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma}) = \mathrm{tr}\left\{(\boldsymbol{Y} - E[\boldsymbol{Y}])\right\}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - E[\boldsymbol{Y}]\right)'\},$$

ここで $E[\boldsymbol{Y}] = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'$ である.

このリスク関数 $R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma})$ を最小にするような $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ を求めることで得られるそれぞれの推定量 $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Xi}}$ は, $r_{\boldsymbol{A}} = k$ かつ $r_{\boldsymbol{X}} = q$ であれば, $\hat{\boldsymbol{\mu}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$ と $\hat{\boldsymbol{\Xi}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ の形で陽に得られることなどを講演では触れた. しかしながら, $n < k$ などの状況で起きる $r_{\boldsymbol{A}} = k$ かつ $r_{\boldsymbol{X}} = q$ を満たさない場合はこれらの推定量が陽に得られないため, リスク関数の計算を工夫するなどして最小化する必要があることから, 本講演で様々な状況での推定量について講演した.

そこで, $r_{\boldsymbol{A}} = k$ かつ $r_{\boldsymbol{X}} < q$ の仮定の下での推定量を永井 (2022a, 2022b) で提案したことを講演で報告した. また, 多変量線形回帰モデル (モデル (1) で $\boldsymbol{X} = \boldsymbol{I}_p$ としたモデル) において永井 (2021, 2022c) で提案した推定量をモデル (1) へ拡張することで, $r_{\boldsymbol{A}} < k$ かつ $r_{\boldsymbol{X}} = q$ の仮定の下での推定量を, 本講演ではサブとして提案した. さらに，本講演のメインとして，$r_{\boldsymbol{A}} < k$ かつ $r_{\boldsymbol{X}} < q$ の仮定の下での推定量を提案した. その際に用いた共通するアイデアにおいて，非常に強い仮定を置くと Koll and von Rosen (2005; Def. 4.1.3) で提案された Extended GMANOVA モデルと対応していることにも触れた．そこで提案した推定量は，一つだけ陽に得られ，他は順番に求まることを述べた．また，メインで提案した推定量の問題点にも触れ，そこを改良するアイデアとして永井 (2019) の手法が用いれると考えられることにも触れた.

GMANOVA モデル (1) において，様々な仮定とそれぞれで提案してきた推定量が以下のようにまとめられることを講演した.

表; $r_{\boldsymbol{A}}$ や $r_{\boldsymbol{X}}$ の仮定と推定手法について

仮定	$\boldsymbol{\mu}$ の推定量や $\boldsymbol{\Xi}$ の推定量	それぞれの不偏推定量
$r_{\boldsymbol{A}} = k$ & $r_{\boldsymbol{X}} = q$	$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$, $(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$	
$r_{\boldsymbol{A}} = k$ & $r_{\boldsymbol{X}} < q$	永井 (2022a)	永井 (2022b)
$r_{\boldsymbol{A}} < k$ & $r_{\boldsymbol{X}} = q$	本講演でサブとして提案	
$r_{\boldsymbol{A}} < k$ & $r_{\boldsymbol{X}} < q$	本講演のメインとして提案	アイデア段階として触れた

引用文献:

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

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

[3] 永井 勇 (2019) バランス型経時測定データにおける Extended GMANOVA モデルの解釈と新たな推定法, 多様な分野における統計科学に関する諸問題.

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

[5] 永井 勇 (2022a) GMANOVA モデルにおける新たな推定方法とその解釈, . 多様な分野における統計科学の理論とその応用.

[6] 永井 勇 (2022b) GMANOVA モデルにおける新たな推定方法と解釈, 大規模複雑データの理論と方法論〜新たな発展と関連分野への応用〜.

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

多変量正規母集団における条件付き独立性検定について

松内 直輝（神戸大学大学院理学研究科）
首藤 信通（神戸大学大学院理学研究科）

本報告においては, 多変量正規母集団における条件付き独立性検定問題について議論した. 特に, 多変量正規性を持つ p 次元確率変数ベクトル $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \boldsymbol{X}_3')'$ に対し, 母集団から得られた標本ベクトルを基にして $\boldsymbol{X}_1$ が与えられた下で, $\boldsymbol{X}_2$ と $\boldsymbol{X}_3$ の条件付き独立性を帰無仮説とする仮説検定を行うための尤度比検定を構成した. ただし, $\boldsymbol{X}_i$ は $\boldsymbol{X}$ の p_i 次元分割ベクトル, $p = p_1 + p_2 + p_3$ である.

本報告では, まず $\boldsymbol{X}_1$ が与えられた下で, $\boldsymbol{X}_2$ と $\boldsymbol{X}_3$ が条件付き独立であることと同値となる条件を求めた. p 変量正規母集団 $N_p(\boldsymbol{\mu}, \Sigma)$ の下では, 平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 Σ について, $\boldsymbol{X}$ の分割と対応する分割を

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}$$

とする. 本報告で扱う仮説検定問題の帰無仮説と

$$\Sigma_{23\cdot 1} \equiv \Sigma_{23} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{13} = O_{23}$$

が同値であることを示すことができた. ここに, $\boldsymbol{\mu}_i$ は $\boldsymbol{\mu}$ の p_i 次元分割ベクトル, Σ_{ij} は Σ の $p_i \times p_j$ 分割行列である.

また, 以下のようなパラメータの変換

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{11}^{-1}\Sigma_{12} & \Sigma_{(12)}^{-1}\Sigma_{(12)3} \\ \Sigma_{21}^{-1}\Sigma_{11}^{-1} & \Sigma_{22\cdot 1} & \\ \Sigma_{3(12)}\Sigma_{(12)}^{-1} & & \Sigma_{33\cdot(12)} \end{pmatrix} \to \begin{pmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} \\ \Psi_{31} & \Psi_{32} & \Psi_{33} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 - \Psi_{21}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_3 - \Psi_{3(12)} \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \end{pmatrix} \to \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{pmatrix}$$

を考えると $\boldsymbol{X}_1$ が与えられた下で, $\boldsymbol{X}_2$ と $\boldsymbol{X}_3$ の条件付き独立性に関する仮説検定問題は

$$H_0 : \Psi_{23} = O_{23} \quad vs. \quad H_1 : \Psi_{23} \neq O_{23}$$

と書き換えることができる.

具体的に, $\Pi : N_p(\boldsymbol{\mu}, \Sigma)$ から得られた N 個の p 次元標本ベクトル $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ が得られたとしてこの仮説検定問題における尤度比検定統計量を構成すると尤度比検定統計量は

$$\begin{aligned} -2\log\lambda &= N\left(\log(\det(S_{33\cdot 1})) - \log(\det(S_{33\cdot(12)})\right) \\ &= N\left(\log(\det(S_{33\cdot 1})) - \log(\det(S_{33\cdot 1} - S_{32\cdot 1}S_{22\cdot 1}^{-1}S_{23\cdot 1}))\right) \end{aligned}$$

となることが示された. ここに, S_{ij} は

$$\bar{\boldsymbol{y}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{y}_j, \quad S = \frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{y}_j - \bar{\boldsymbol{y}}\right)\left(\boldsymbol{y}_j - \bar{\boldsymbol{y}}\right)',$$

としたとき, S の $p_i \times p_j$ 分割行列である. また, $S_{ij\cdot 1} = S_{ij} - S_{i1}S_{11}^{-1}S_{1j}$,

$$S_{(12)} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, S_{(12)3} = \begin{pmatrix} S_{13} \\ S_{23} \end{pmatrix}, S_{33\cdot(12)} = S_{33} - S_{3(12)}S_{(12)}^{-1}S_{(12)3}$$

である.

最後に数値実験を行い, 本報告で議論している条件付き独立性の下で擬似的に発生させたデータセットから尤度比検定統計量を求め, シミュレーションで与えられる上側 $100\alpha\%$ 番目の値と, 自由度 p_2p_3 のカイ二乗分布の上側 $100\alpha\%$点の比較を行った. 数値実験の結果から p や p_1 の値よりも p_2p_3 の値が強く第 1 種過誤の精度に影響を与えることを確認した.

今後の課題としてより少ない標本でも検定の精度が保てるように分布の修正を行うことや, より緩い仮定の下での仮説検定の構成を行うことなどが考えられる.

多重積分ブラウン運動の統計学的応用について

田中 勝人（一橋大学名誉教授）

[0,1] 上で定義されたブラウン運動 $\{W(t)\}$ およびフラクショナル･ブラウン運動 $\{B_H(t)\}$ の 2 次汎関数に関連する統計量に関して，以下の (1)，(2)，(3) の報告をした．

(1) g 重積分過程

$$F_g(t) = \int_0^t \int_0^{t_1} \cdots \int_0^{t_{g-1}} W(t_g)\, dt_g\, dt_{g-1} \cdots dt_1 \qquad (g = 1, 2, ...)$$

に対して，

$$V_g = \int_0^1 F_g^2(t)\, dt$$

の分布の特性関数は，

$$\mathrm{E}(e^{i\theta V_g}) = \Big(D_g(2i\theta)\Big)^{-1/2}$$

で表されることを示した．ここで，$D_g(\lambda)$ は，関数 $\mathrm{Cov}(F_g(s), F_g(t))$ のフレッドホルム行列式 (FD) であり，FD を導出する方法を説明した．$g = 1, 2, 3$ の場合の FD は次のようになる．

$$\begin{aligned}
D_1(\lambda) &= \frac{1}{2}\,(1 + \cos\lambda^{1/4} \cosh\lambda^{1/4}),\\
D_2(\lambda) &= \frac{1}{9}\Big[2(1 + \cos\lambda^{1/6} + \cos\lambda^{1/6}\omega + \cos\lambda^{1/6}\omega^2) + \cos\lambda^{1/6} \cos\lambda^{1/6}\omega \cos\lambda^{1/6}\omega^2\Big],\\
\omega &= \frac{1 + \sqrt{3}i}{2},\\
D_3(\lambda) &= \frac{1}{16}\Big[3\cos a\, \cos b\, \cos c\, \cos d + 2\Big(\cos a\, \cos b + \cos b\, \cos c + \cos c\, \cos d\\
&\qquad + \cos d\, \cos a\Big) + \cos a\, \cos c + \cos b\, \cos d + 3\Big]\\
&\quad + \frac{\sqrt{2}}{16}\Big[\sin a\, \sin b\,(1 + \cos c\, \cos d) + \sin b\, \sin c\,(1 + \cos d\, \cos a)\\
&\qquad + \sin c\, \sin d\,(1 + \cos a\, \cos b) - \sin d\, \sin a\,(1 + \cos b\, \cos c)\Big],\\
a &= \lambda^{1/8},\ b = a\omega,\ c = a\omega^2,\ d = a\omega^3,\ \omega = \frac{1+i}{\sqrt{2}}.
\end{aligned}$$

また，g が大きくなる場合には，最小固有値（$D_g(\lambda) = 0$ の最小根）だけを使った分布でもよい近似を与えることを示した．

(2) 多重単位根統計量

$$R_g = \frac{F_g^2(1)/2}{\displaystyle\int_0^1 F^2(t)\, dt} \qquad (g = 1, 2, ...)$$

の分布についても関連する特性関数を求め，密度関数を数値積分により求めた．分布関数は，

$$P(R_g < x) \;=\; P\left(x \int_0^1 F_g^2(t)\, dt - \frac{1}{2} F_g^2(1) > 0\right)$$

$$
\begin{aligned}
&= P\left(\int_0^1 \int_0^1 K_g(s,t;x)\, dW(s)\, dW(t) > 0\right) \\
&= \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{1}{\theta} \operatorname{Im}\left[\left(D_g(2i\theta;x)\right)^{-1/2}\right] d\theta
\end{aligned}
$$

により計算することができる．ここで，$D_g(\lambda;x)$ は，

$$
K_g(s,t;x) = \frac{1}{(g!)^2}\left[x\int_{\max(s,t)}^1 \Big((u-s)(u-t)\Big)^g\, du - \frac{1}{2}\Big((1-s)(1-t)\Big)^g\right]
$$

の FD である．FD の導出には，積分方程式

$$
f(t) = \lambda \int_0^1 K_g(s,t;x) f(s)\, ds
$$

が，微分方程式

$$
f^{(2g+2)}(t) + (-1)^g \lambda x f(t) = 0
$$

および $2g+2$ 個の境界条件

$$
\begin{aligned}
f(1) &= f'(1) = \cdots = f^{(g-1)}(1) = 0, \quad f^{(g)}(1) = \frac{(-1)^{g-1}\lambda}{2g!}\int_0^1 (1-s)^g f(s)\, ds, \\
f^{(g+1)}(0) &= f^{(g+2)}(0) = \cdots = f^{(2g+1)}(0) = 0.
\end{aligned}
$$

と同値となることを用いた．ここで，$\mathrm{E}(R_g) = g+1$ という興味深い結果が得られた．

(3) 通常の単位根統計量と多重単位根統計量の間に介在するフラクショナル単位根統計量

$$
Q_H = \frac{B_H^2(1)/2}{\displaystyle\int_0^1 B_H^2(t)\, dt}
$$

の分布について考察した．ここで，$B_H(t)$ はフラクショナル・ブラウン運動 (fBm) であり，fBm は平均 0 の正規過程で，共分散関数が

$$
\mathrm{Cov}(B_H(s), B_H(t)) = \frac{1}{2}[s^{2H} + t^{2H} - |s-t|^{2H}]
$$

で与えられる．$S_H = \displaystyle\int_0^1 B_H^2(t)\, dt$ および Q_H の分布導出は未解決問題である．ここでは，S_H の近似として，

$$
T_H = \int_0^1 C_H^2(t)\, dt, \quad C_H(t) = \sqrt{2(1-H)}\, t^{2H-1}\int_0^t u^{1/2-H}\, dW(u)
$$

を提案した．T_H に関連した FD は，

$$
D_H(\lambda) = \Gamma(1-\nu) J_{-\nu}(\eta) \Big/ \left(\frac{\eta}{2}\right)^{-\nu}
$$

となる．ここで，$J_{-\nu}(\eta)$ は第 1 種ベッセル関数，また，$\eta = \sqrt{2(1-H)\lambda}/(H+1/2), \quad \nu = (2H-1/2)/(H+1/2)$ である．また，Q_H の近似として，

$$
R_H = \frac{C_H^2(1)/2}{\displaystyle\int_0^1 C_H^2(t)\, dt}
$$

を提案して，その分布特性についても考察した．

On Asymptotic Distribution in Martingale Convergence of Supercritical Branching Processes with Poissonian offsprings

Bat-Erdene A.[*], Kawasaki S.[*], Li J.[**], Altantsetseg E.[***]

[*]Faculty of Science and Engineering, Iwate University
[**] School of Computer Science, Chongqing University
[***]School of the Engineering and Applied Sciences, National University of Mongolia

Abstract

A branching process is a mathematical model of Erdos-Renyi random graphs. For a normalized process, a martingale convergence theorem holds. However, its asymptotic distribution has not been known so far. We propose a characterization of the distribution via analysis on a functional equation for the Laplace transform of the distribution. It turns out that a numerical analysis mostly coincides well with the theoretical characterization. Applications of the Branching process include a biological population, nuclear chain reactions, and the spread of computer software viruses in common. Mathematical models of these applications play a central role in figuring out the main process and predicting future extensions. Our funding works in common cases and provides an explanation for the characterization of the processes.

Problem. Let $W_n = Z_n/\lambda^n, n = 0, 1, 2, \cdots$, where Z_n is the supercritical branching process with the mean λ which is greater than 1. Then, $\{W_n\}$ is known [2] to form a martingale and converges to a random variable W_∞ on $\mathbb{R}_+ = [0, \infty)$ a.s. as $n \to \infty$.

However, not much is known about the distribution of W_∞ so far in previous studies. The paper aims to characterize the distribution by analyzing a functional equation that holds for the Laplace transform of the distribution.

Analysis and result. W_∞ is known to have a density function on $(0, \infty)$ [2, Corollary 12.1], which we denote by $w : (0, \infty) \mapsto \mathbb{R}_+$, while W_∞ has a point mass at the origin [2, Theorem 6.2]. Thus, we may write $W_\infty \sim \eta\delta(x) + w(x)$ on $x \in \mathbb{R}_+$ and $\delta(x)$ being the Kronecker delta function.
Let $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be the Laplace transform of W_∞,

$$\varphi(u) = \int_{(0,\infty)} e^{-ux} w(x)\, dx. \tag{1}$$

Mclaurin's expansion was used to approximate the function $\varphi(u)$ described in below.

$$\varphi(u) = \sum_{n\in\mathbb{N}_0} \frac{\varphi^{(n)}(0)}{n!}\, u^n \tag{2}$$

Then, upon calculating $\varphi^{(n)}(0)$, $n \in \mathbb{N}$ recursively. To the evaluate $\varphi^{(n)}(0)$, an approximation of K_n as we defined in below is necessary

$$\varphi^{(n)}(0) = -K_n\,(-\rho)^{n-1}, \qquad n \in \mathbb{N} \tag{3}$$

for appropriate functions $K_n = K_n(\lambda)$.

We write $1 + \frac{1}{\lambda} + \cdots + (\frac{1}{\lambda})^{n-2} \triangleq \rho_{n-1}$ below. We note that $\rho_n \longrightarrow \rho$. Through the recursive

relation

$$K_n = \frac{1}{\rho_{n-1}} \sum_{l=1}^{n-1} \binom{n-1}{l} \lambda^{-(n-1-l)}\, K_{n-l}(\lambda) K_l(\lambda)$$

$$= \frac{(n-1)!}{\rho_{n-1}} \sum_{l=1}^{n-1} \lambda^{-(n-1-l)}\, \frac{K_{n-l}(\lambda)}{(n-l)!} \cdot \frac{K_l(\lambda)}{l!} \cdot (n-l), \tag{4}$$

In the result stage, we propose an approximation function $K_n(\lambda)$ as follows:

Theorem 1. *For arbitrarily fixed $\lambda \in [1, \infty)$,*

$$K_n \simeq c_n \times (n-1)! \left[\, \lambda(\lambda+1) \,\right]^{-(an+b)} \cdot \Big(1 + O\big(\lambda^{-n}\big)\Big), \qquad as \quad l \to \infty \tag{5}$$

for some $c_n > 0$.

A derivative $\varphi'(u)$, instead of $\varphi(u)$ has considered. That is,

$$\varphi'(u) = \sum_{n \in \mathbb{N}_0} \frac{\varphi^{(n+1)}(0)}{n!}\, u^n = -\sum_{n \in \mathbb{N}_0} \frac{K_{n+1}(\lambda)}{n!}\, (-\rho u)^n. \tag{6}$$

Applying the expression (5) to (6), we have that the main term of $\varphi'(u)$ is given by

$$const. \times \sum_{n \in \mathbb{N}_0} \Big(-\rho \big[\, \lambda(\lambda+1) \,\big]^{-a} u \Big)^n = const. \times \frac{1}{u + \rho^{-1} \big[\, \lambda(\lambda+1) \,\big]^{a}}.$$

Now we recall that $\varphi'(u)$ is the Laplace transform of $-x\, w(x)$. Thus, by the inverse Laplace transform of $\varphi'(u)$, we found that $w(x)$ has the corresponding main component of the probability density function given by

$$const. \times x^{-1} \exp\Big(-\big[\, \lambda(\lambda+1) \,\big]^{a} x \Big), \qquad x \in (0, \infty).$$

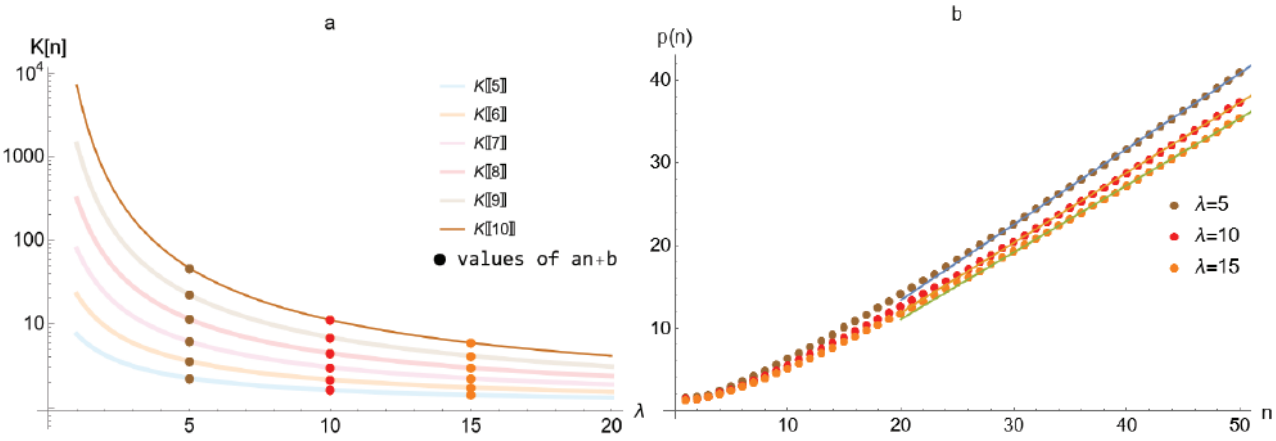

Figure 1: a. The function $K_n(\lambda)$, its relation and values of an equation $p_n = an + b$. b. Asymptotic linearity, $p_n = an + b$ at some fixed λ

References

[1] Hofstad, R. (2016). *Random Graphs and Complex Networks*, Cambridge Univ Press.

[2] Athreya, K. B., Ney, P. E. (2004). *Branching Processes*, Dover publ.

[3] Harris, T. E. (1963). *The Theory of Branching Processes*, Dover publ.

Forest Construction of Gaussian and Binary Variables based on WBIC

Ashraful Islam, Joe Suzuki
Graduate school of Engineering Science, Osaka University, Japan

Mutual information (MI) is a metric that determines the association between two random variables by measuring the amount of information that one variable holds about the other. It quantifies the dependency between them. A higher mutual information value indicates a stronger relationship between the random variables. It is related to the concept of entropy and is used in various fields such as information theory, statistics, and machine learning. Estimated joint probabilities from observed samples in each variable category combination are used to calculate MI between discrete variables. But the conventional method is inefficient for estimating the MI of a mixture of discrete and continuous random variables because the conditional probabilities for discrete variables given continuous variables cannot be determined. In this paper, we examine a new MI method that can accommodate a mixture of discrete and continuous random variables. The estimation of free energy for new MI is not easy, but we can use WBIC to calculate the corresponding free energies. After that, the Chow-Liu algorithm can be modified to take into account the new MI in order to build a forest rather than spanning trees.

This section will outline the paper's most significant findings and conclusions. Chow and Liu (1968) considered estimating mutual information between two discrete variables. When only one of the variables is Gaussian and the rest are discrete, Edwards, De Abreu, and Labouriau (2010) recalculated mutual information using the ANOVA model. If X is a Gaussian random variable, and Y and Z are discrete variables, getting maximum likelihoods of the graphical models such as $X - Z$ and $Y - X$ are simple, but obtaining maximum likelihoods for other types of graphical models as $Y - X - Z$ by maximizing observational mutual information estimates is difficult(Suzuki, 2017). The state-of-the-art avoids this problem by restricting the class to forests that separate discrete and continuous nodes.

On the other hand, the Bayesian method makes the assumption that there is infinite number of histograms, but the number of histograms that are required for accuracy is unclear to us. Each cluster also has a large sample size variance. Adjusting parameters is difficult when many clusters have no sample. The proposed approach succeeds where the other methods fail in capturing the relationship between multiple gene expressions and SNPs. Some previous works that relate to our research are as follows:

Suzuki (1993) considers MDL method and obtains total lengths for all the variables, and estimates the mutual information as:

$$J_n = I_n - \frac{(\alpha-1)(\beta-1)}{2n}\log_2 n \tag{1}$$

For large n, $J_n \leq 0$ indicates the independence of X and Y (Suzuki, 2012).

Edwards et al. (2010) assumed about a situation where some random variables are continuous and some are Gaussian variables. Assume, $X^{(1)}, X^{(2)}, \ldots, X^{(N)}$ are Gaussian. And, mutual information:

$$I_n(i,j) = -\frac{1}{2}\log_2\left(1-\hat{\rho}^2\right), \tag{2}$$

To use the mutual information estimator, Edwards et al. (2010) imposed the limitation that no Gaussian variables could be intermediate between the discrete ones.

Suzuki (2017) took into account the Dirichlet distribution with the proportionality $\prod_x \theta(x)^{a(x)-1}$, and derived the formula for mutual information as follows:

$$J_n = \frac{1}{n}\log_2 \frac{Q^n\left(x^n, y^n\right)}{Q^n\left(x^n\right)Q^n\left(y^n\right)} \tag{3}$$

Suzuki (2017) has found that $-log_2 Q^n(x^n, y^n)$ is nothing but the free energy of the mixture of x^n and y^n. In the field of data science, free energy refers to the amount of uncertainty or surprise present in a system or dataset. In statistical mechanics and information theory, free energy is employed to quantify

the disorder or randomness of a system. But obtaining free energy is not easy. In this circumstance, WBIC (Watanabe, 2021) is a way to figure out free energy.

The WBIC algorithm estimates the free energy of a system using Bayesian statistics, taking into account the uncertainty in the model's parameters. Again, we use the Kernel Hilbert–Schmidt independence criterion (HSIC) to verify the Mutual Information we get from the new mutual information method. In the context of statistical inference, the HSIC was developed to assess how much the embedding of the Hilbert space depends on the underlying distribution. To obtain the free energy from the mixture of discrete and Gaussian, we use stan. Stan specifies statistical models probabilistically. Stan uses Markov Chain Monte Carlo methods like the No-U-Turn sampler for continuous variable Bayesian inference.

Let, for large n, X and Y are discrete and Gaussian random variables, respectively, and their joint pdf,

$$f(x,y) = \sum_{i=1}^{2} p^x (1-p)^{1-x} . \frac{1}{\sqrt{2\pi}\sigma_i} exp\{\frac{1}{2\sigma_i}(y-\mu_i)^2\}$$

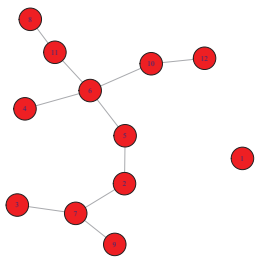

Figure 1: Forest for 'P.sojae.survey' data

And we propose the new mutual information:

$$J_n = \frac{1}{n} log \frac{Q^n(X^n, Y^n)}{Q^n(X^n).Q^n(Y^n)} = \frac{1}{n} log \frac{e^{-F_{xy}}}{e^{-Fx}.e^{-Fy}} = \frac{1}{n}(F_x + F_y - F_{xy}) \quad (4)$$

We calculate the MI from equation 4 using simulated data for a mixture of discrete and Gaussian variables.

When we use n=500, we get the following results for the mutual information: -0.008828827, HSIC: 0.000759246, and p-value: 0.2397602. These findings both confirm and imply that the variables X and Y are free to move around without influence from one another.

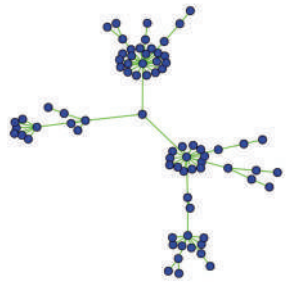

Figure 2: Forest using simulated data

We can see that the suggested mutual information estimator works for all types of variables: discrete, continuous, or a mixture of discrete and continuous variables. After calculating the mutual information from the new method, we applied it to the Chow-Liu algorithm to find the forest. Figure 1 shows the forest for P.sojae.survey data with 12 variables; we can see that the first variable is independent of the others. That is why it is separated from the other variables. Figure 2 shows a forest with 75 simulated observations with a mixture of discrete and continuous variables with 100 samples each. Therefore, depending on the previous results and the figures, the current method can construct a forest from a mixture of discrete and continuous nodes, depending on the independence of the variables. In future work, we will apply the proposed mutual information method to multiple gene expressions and SNPs (single-nucleotide polymorphism) data.

References

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, *14*(3), 462–467.

Edwards, D., De Abreu, G. C., & Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC bioinformatics*, *11*(1), 1–13.

Suzuki, J. (1993). A construction of bayesian networks from databases based on an mdl principle. In *Uncertainty in artificial intelligence* (pp. 266–273).

Suzuki, J. (2012). The bayesian chow-liu algorithm. In *The sixth european workshop on probabilistic graphical models* (pp. 315–322).

Suzuki, J. (2017). A novel chow–liu algorithm and its application to gene differential analysis. *International Journal of Approximate Reasoning*, *80*, 1–18.

Watanabe, S. (2021). Waic and wbic for mixture models. *Behaviormetrika*, *48*(1), 5–21.

多次元空間における埋め込み 1 次元曲線の同時信頼領域

千葉大・融合理工学府　山添 滉弥
千葉大・理学研究院　内藤 貫太

はじめに: 道路交通網上の交通事故や，脳神経細胞網における樹状突起スピンの分布といった，枝分かれした 1 次元曲線上で観測されるデータをネットワークデータを呼ぶ．ネットワークデータに対して，与えられたネットワーク上における密度推定の手法についてはすでに研究されている (Liu and Ruppert, 2021; McSwiggan et al., 2017)．しかしながら，ネットワークを構成する 1 次元曲線の同時信頼領域をデータから構築する方法については，いまだ研究がないようである．本研究ではこの問題を多次元空間に埋め込まれた 1 次元曲線の同時信頼領域を構築する問題として捉え，ノンパラメトリック回帰のアプローチから取り組んだ．

問題: $I=[a,b]\subset\mathbb{R}$ を閉区間とし，$\varphi_i : I\to\mathbb{R}\ (i=1,\dots,d)$ を滑らかな関数とする．このとき，$\mathbb{R}^d$ に埋め込まれた 1 次元曲線を

$$\boldsymbol{\varphi}(t)=\begin{bmatrix}\varphi_1(t)\\ \vdots\\ \varphi_d(t)\end{bmatrix}\in\mathbb{R}^d$$

で表す．確率ベクトル $\boldsymbol{Y}\in\mathbb{R}^d$ と共変量 $T\in I$ の組 $(T,\boldsymbol{Y})$ からの n 個の観測値 $(T_1,\boldsymbol{Y}_1),\dots,(T_n,\boldsymbol{Y}_n)$ に基づき 1 次元曲線 $\boldsymbol{\varphi}$ の同時信頼領域を構築したい．つまり，$M_{\boldsymbol{\varphi}}=\left\{\boldsymbol{\varphi}(t)\in\mathbb{R}^d \mid t\in I\subset\mathbb{R}\right\}$ としたとき，十分小さい $\alpha>0$ に対して，

$$\mathbb{P}\left(\mathfrak{D}\supset M_{\boldsymbol{\varphi}}\right)\geq 1-\alpha$$

となるような有界閉領域 $\mathfrak{D}\subset\mathbb{R}^d$ の構築を目指す．

設定: 1 次元曲線の推定の関連研究である Hastie and Stuetzle (1989) を参照しつつ，モデルの仮定を行う．$\mathbb{R}^d$ に埋め込まれた 1 次元曲線 $\boldsymbol{\varphi}$ に対して，任意の点 t における勾配ベクトル $\boldsymbol{\varphi}'(t)$ を $\boldsymbol{\varphi}'(t)=[\varphi_1'(t)\ \cdots\ \varphi_d'(t)]^T$ で表す．また，$\boldsymbol{n}_1(t),\dots,\boldsymbol{n}_{d-1}(t)$ を $\left\{\boldsymbol{\varphi}'(t)/\left\|\boldsymbol{\varphi}'(t)\right\|,\boldsymbol{n}_1(t),\dots,\boldsymbol{n}_{d-1}(t)\right\}$ が $\mathbb{R}^d$ の正規直交基底となるようにとる．さらに，T を I 上の密度 f_T に従う確率変数とし，$\boldsymbol{V}=[V_1\ \cdots\ V_{d-1}]^T$ を原点中心半径 r の $(d-1)$ 次元球 B_r^{d-1} 上の密度 $f_{\boldsymbol{V}}$ に従う確率ベクトルとする．

$(T_1,\boldsymbol{V}_1),\dots,(T_n,\boldsymbol{V}_n)\overset{\text{i.i.d.}}{\sim} f_T\times f_V$ と $E[\boldsymbol{V}_1]=\boldsymbol{0}$ を仮定する．確率ベクトル $\boldsymbol{Y}$ の d 次元標本 $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ として，

$$\boldsymbol{Y}_i=\boldsymbol{\varphi}(T_i)+N(T_i)\boldsymbol{V}_i$$

を考える．ただし，$\boldsymbol{Y}_i=[Y_{i1}\ \cdots\ Y_{id}]^T,\boldsymbol{V}_i=[V_{i1}\ \cdots\ V_{id-1}]^T,N(t)=[\boldsymbol{n}_1(t)\ \cdots\ \boldsymbol{n}_{d-1}(t)]$ である．

同時信頼領域の構築: 共変量 $\boldsymbol{T}=[T_1\ \cdots\ T_n]^T$ と $\boldsymbol{Y}$ の第 j-成分のデータ $\widetilde{\boldsymbol{Y}}_j=[Y_{1j}\ \cdots\ Y_{nj}]^T$ との組 $(\boldsymbol{T},\widetilde{\boldsymbol{Y}}_i)\ (i=1,\dots,d)$ に基づき，t における $\varphi_i(t)$ の局所線形推定量 $\widehat{\varphi}_i(t)(i=1,\dots,d)$ が得られる．これらを並べることにより $\boldsymbol{\varphi}(t)$ の推定量 $\widehat{\boldsymbol{\varphi}}(t)$ が構築される．さらに $\widehat{\boldsymbol{\varphi}}(t)$ を用いてその勾配ベクトルを求める：

$$\widehat{\boldsymbol{\varphi}}(t)=\begin{bmatrix}\widehat{\varphi}_1(t)\\ \vdots\\ \widehat{\varphi}_d(t)\end{bmatrix},\quad \widehat{\boldsymbol{\varphi}}'(t)=\frac{d}{dt}\widehat{\boldsymbol{\varphi}}(t).$$

局所線形推定量については Wand and Jones (1995); Fan and Gijbels (1996) などを参照されたい．$\widehat{\boldsymbol{n}}_1(t),\ldots,\widehat{\boldsymbol{n}}_{d-1}(t)$ を $\{\widehat{\boldsymbol{\varphi}}'(t)/\left\|\widehat{\boldsymbol{\varphi}}'(t)\right\|,\widehat{\boldsymbol{n}}_1(t),\ldots,\widehat{\boldsymbol{n}}_{d-1}(t)\}$ が $\mathbb{R}^d$ の正規直交基底となるようにとり，

$$
\begin{aligned}
\widehat{\mathfrak{D}}_a(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(a) - R\frac{\widehat{\boldsymbol{\varphi}}'(a+)}{\left\|\widehat{\boldsymbol{\varphi}}'(a+)\right\|} + \widehat{N}(a)\boldsymbol{v} \;\middle|\; 0 \le R \le r, \boldsymbol{v} \in B^{d-1}_{\sqrt{r^2-R^2}} \right\}, \\
\widehat{\mathfrak{D}}_J(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(t) + \widehat{N}(t)\boldsymbol{v} \;\middle|\; t \in J, \boldsymbol{v} \in B^{d-1}_{r} \right\}, \\
\widehat{\mathfrak{D}}_b(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(b) + R\frac{\widehat{\boldsymbol{\varphi}}'(b-)}{\left\|\widehat{\boldsymbol{\varphi}}'(b-)\right\|} + \widehat{N}(b)\boldsymbol{v} \;\middle|\; 0 \le R \le r, \boldsymbol{v} \in B^{d-1}_{\sqrt{r^2-R^2}} \right\}, \\
\widehat{N}(t) &= \begin{bmatrix} \widehat{\boldsymbol{n}}_1(t) & \cdots & \widehat{\boldsymbol{n}}_{d-1}(t) \end{bmatrix}
\end{aligned}
$$

とする．3 つの排反な領域を用いて同時信頼領域 $\widehat{\mathfrak{D}}(r) \subset \mathbb{R}^d$ を

$$
\widehat{\mathfrak{D}}(r) = \widehat{\mathfrak{D}}_a(r) \cup \widehat{\mathfrak{D}}_J(r) \cup \widehat{\mathfrak{D}}_b(r)
$$

として定義する (図 1 参照)．本講演では $\widehat{\boldsymbol{\varphi}}(t)$ と $\widehat{\mathfrak{D}}(r)$ の理論的性質および，信頼係数 $1-\alpha$ から $\widehat{\mathfrak{D}}(r)$ の r を決定する方法について解説を与えた．また，共変量 $\boldsymbol{T}$ がデータとして得られていない場合において，$\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ から $T_1,\ldots,T_n$ を作成する方法についても報告した．さらに，提案領域が同時信頼領域として機能していることを確認するシミュレーション実験の結果や，実データへの適用例についても報告した．

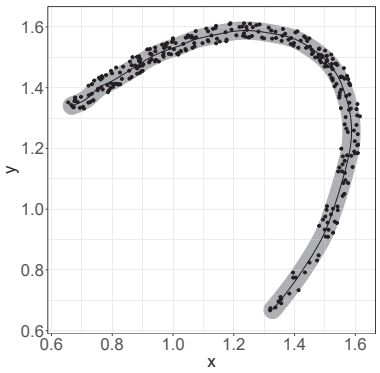

図 1　推定曲線 $\widehat{\boldsymbol{\varphi}}(t)$ から作られた同時信頼領域 $\widehat{\mathfrak{D}}(r)$ の例

参考文献

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Taylor & Francis.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.

Liu, Y. and Ruppert, D. (2021). Density estimation on a network. *Computational Statistics & Data Analysis*, 156, 107128.

McSwiggan, G., Baddeley, A., and Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44, 324–345.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.

多標本問題における 2-step 単調欠測データの下での平均ベクトルの同等性検定の検出力について

神戸大学・理・院　勝又　真
神戸大学・理・院　首藤　信通

本報告では, k 個の母集団 $\Pi^{(g)} : N_p(\boldsymbol{\mu}^{(g)}, \Sigma)$ $(g = 1, \ldots, k)$ から, 2-step 単調欠測データが得られた下で帰無仮説 $H_0 : \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(k)}$ に対する尤度比検定統計量を導出し, その第 1 種の過誤や検出力について数値的評価を与えた.

具体的には, 第 g 母集団 $\Pi^{(g)}$ $(g = 1, \ldots, k)$ から $N_1^{(g)}$ 個の p 次元標本ベクトル $\boldsymbol{x}_j^{(g,1)} = \left(\boldsymbol{x}_{1j}^{(g,1)\prime}, \boldsymbol{x}_{2j}^{(g,1)\prime}\right)'$ $(j = 1, \ldots, N_1^{(g)})$, $N_2^{(g)}$ 個の p_1 次元標本ベクトル $\boldsymbol{x}_{1j}^{(g,2)}$ $(j = 1, \ldots, N_2^{(g)})$ が観測されたと仮定し, 尤度比検定統計量が

$$
\begin{aligned}
-2\log\Lambda =& N\log\left\{1 + \frac{1}{N}\sum_{g=1}^{k}\frac{1}{N^{(g)}}\left(\boldsymbol{x}_{1\cdot}^{(g,1)} + \boldsymbol{x}_{1\cdot}^{(g,2)}\right)'\hat{\Delta}_{11}^{-1}\left(\boldsymbol{x}_{1\cdot}^{(g,1)} + \boldsymbol{x}_{1\cdot}^{(g,2)}\right)\right.\\
&\left. - \frac{1}{N^2}\left(\boldsymbol{x}_{1\cdot}^{(\cdot,1)} + \boldsymbol{x}_{1\cdot}^{(\cdot,2)}\right)'\hat{\Delta}_{11}^{-1}\left(\boldsymbol{x}_{1\cdot}^{(\cdot,1)} + \boldsymbol{x}_{1\cdot}^{(\cdot,2)}\right)\right\}\\
&+ N_1\log\left\{1 + \sum_{g=1}^{k}N_1^{(g)}\begin{pmatrix}\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(\cdot,1)}\\ \bar{\boldsymbol{x}}_{2\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{2\cdot}^{(\cdot,1)}\end{pmatrix}'\left(W^{(\cdot,1)}\right)^{-1}\begin{pmatrix}\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(\cdot,1)}\\ \bar{\boldsymbol{x}}_{2\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{2\cdot}^{(\cdot,1)}\end{pmatrix}\right\}\\
&- N_1\log\left\{1 + \sum_{g=1}^{k}N_1^{(g)}\left(\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(\cdot,1)}\right)'\left(W_{11}^{(\cdot,1)}\right)^{-1}\left(\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(\cdot,1)}\right)\right\}
\end{aligned}
$$

となることを示した. ここで, $\boldsymbol{x}_{1\cdot}^{(g,i)} = \sum_{j=1}^{N_i^{(g)}}\boldsymbol{x}_{1j}^{(g,i)}, \bar{\boldsymbol{x}}_{1\cdot}^{(g,i)} = N_i^{(g)^{-1}}\sum_{j=1}^{N_i^{(g)}}\boldsymbol{x}_{1j}^{(g,i)}$, $\bar{\boldsymbol{x}}_{1\cdot}^{(\cdot,1)} = N_1^{-1}\sum_{g=1}^{k}\boldsymbol{x}_{1\cdot}^{(g,1)}, \hat{\Delta}_{11} = N^{-1}(W_{11}^{(\cdot,1)} + W^{(\cdot,2)}), W_{11}^{(\cdot,1)} = \sum_{g=1}^{k}\sum_{j=1}^{N_1^{(g)}}$ $(\boldsymbol{x}_{1j}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,1)})(\boldsymbol{x}_{1j}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,1)})', W^{(\cdot,2)} = \sum_{g=1}^{k}\sum_{j=1}^{N_2^{(g)}}(\boldsymbol{x}_{1j}^{(g,2)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,2)})(\boldsymbol{x}_{1j}^{(g,2)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,2)})'$ $+ \sum_{g=1}^{k}\left\{N_1^{(g)}N_2^{(g)}/N^{(g)}(\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,2)})(\bar{\boldsymbol{x}}_{1\cdot}^{(g,1)} - \bar{\boldsymbol{x}}_{1\cdot}^{(g,2)})'\right\}, N^{(g)} = N_1^{(g)} + N_2^{(g)}$, $N = \sum_{g=1}^{k}N^{(g)}, N_1 = \sum_{g=1}^{k}N_1^{(g)}, N_2 = \sum_{g=1}^{k}N_2^{(g)}$ である.

また, $k = 3, p = 6, p_1 = 2, 3, 4, 5$ の場合における上記の尤度比検定統計量について, 自由度 $p(k-1) = 12$ のカイ二乗分布の上側 5% 点を棄却限界値として用いた場合の仮説検定方式と, 同様の設定の下で $N_2^{(1)} = N_2^{(2)} = N_2^{(3)} = 0$ における尤度比検定統計量について同じ棄却限界値を用いた場合の仮説検定方式の第 1 種の過誤を比較した. 表 1 はその結果の一部であり, 前者の第 1 種の過誤は ASL(miss), 後者の第 1 種の過誤は ASL(comp) で表す.

表 1: $k=3, p=6, p_1=2$ における第 1 種の過誤の比較

$N_1^{(1)}$	$N_2^{(1)}$	$N_1^{(2)}$	$N_2^{(2)}$	$N_1^{(3)}$	$N_2^{(3)}$	ASL(miss)	ASL(comp)
10	10	10	10	10	10	0.085274	0.089827
20	10	20	10	20	10	0.062790	0.063582
30	10	30	10	30	10	0.057406	0.057700
40	10	40	10	40	10	0.054916	0.055446

数値実験の結果, 完全データのみで尤度比検定を構成するよりも欠測データを利用して尤度比検定を構成する方が, 第 1 種の過誤がより正確に制御されることを確認した. また, サンプルサイズ N が小さいとき, p_1 が大きくなるほど, 第 1 種の過誤がより正確に制御されることを確認した.

次に, $\Sigma = I_p$ の場合において, 欠測パターンに対応するマハラノビス二乗距離をそれぞれ $\delta_i^2 = (\boldsymbol{\mu}_i^{(1)} - \boldsymbol{\mu}_i^{(2)})'(\boldsymbol{\mu}_i^{(1)} - \boldsymbol{\mu}_i^{(2)})$ $(i=1,2)$ とする. これらのマハラノビス二乗距離を変化させたときに $k=3, p=6, p_1=2,3,4,5$ における検出力が p_1 の影響を受けるかについて確認した. 表 2 はその結果の一部である.

表 2: $\delta_1^2 = (0.0)^2, N_1^{(1)} = N_1^{(2)} = N_1^{(3)} = N_2^{(1)} = N_2^{(2)} = N_2^{(3)} = 10$ における検出力の比較

δ_1^2	δ_2^2	$p_1=5$	$p_1=4$	$p_1=3$	$p_1=2$
$(0.0)^2$	$(0.1)^2$	0.072707	0.079148	0.083119	0.087148
$(0.0)^2$	$(0.5)^2$	0.122120	0.128863	0.134764	0.138954
$(0.0)^2$	$(1.0)^2$	0.311521	0.320074	0.328623	0.336122
$(0.0)^2$	$(1.5)^2$	0.621688	0.632784	0.642896	0.652537
$(0.0)^2$	$(2.0)^2$	0.875459	0.883849	0.891225	0.898671
$(0.0)^2$	$(2.5)^2$	0.977481	0.980229	0.982739	0.985267
$(0.0)^2$	$(3.0)^2$	0.997588	0.998116	0.998584	0.998942

数値実験の結果, サンプルサイズ N が小さいとき, p_1 が小さくなるほど検出力が高くなる傾向を確認することができた. また, 欠測データでも同様にサンプルサイズ N が増加するにつれて, 検出力が高くなる傾向を確認した.

今後の課題として, サンプルサイズ N が小さい状況下でも, ほぼ正確な仮説検定となるような分布の修正を行うことなどが挙げられる. また, 本報告で提案する仮説検定では平均ベクトルの差がどの群間にあるかを推測することができないので, これを行うには欠測データの下での平均ベクトルに関する多重比較法の構成について考える必要がある.

Subexponentiality of densities of infinitely divisible distributions on the whole real line

Muneya Matsui

Abstract

We show the equivalence of three properties for an infinitely divisible distribution: the subexponentiality of the density, the subexponentiality of the density of its Lévy measure and the tail equivalence between the density and its Lévy measure density, under monotonic-type assumptions on the Lévy measure density. The key assumption is that tail of the Lévy measure density is asymptotic to a non-increasing function or is almost decreasing. Our conditions are natural and cover a rather wide class of infinitely divisible distributions. Several significant properties for analyzing the subexponentiality of densities have been derived such as closure properties of [convolution, convolution roots and asymptotic equivalence] and the factorization property. Moreover, we illustrate that the results are applicable for developing the statistical inference of subexponential infinitely divisible distributions which are absolutely continuous.

Introduction

Let f, g be probability density functions on $\mathbb{R}$ and denote by $f * g$ the convolution of f and g:

$$f * g(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy,$$

and denote by f^{*n} the nth convolutions with itself. Throughout the paper, for functions $\alpha, \beta : \mathbb{R} \to \mathbb{R}_+$, $\alpha(x) \sim \beta(x)$ means that $\lim_{x\to\infty} \alpha(x)/\beta(x) \to 1$. We study the following characteristics for densities.

Definition 0.1. (i) *f is (right-side) long-tailed, denoted by $f \in \mathcal{L}$, if there exists $x_0 > 0$ such that $f(x) > 0$, $x \geq x_0$ and for any fixed $y > 0$ $f(x+y) \sim f(x)$.*
(ii) *f is (right-side) subexponential on $\mathbb{R}$, denoted by $\mathcal{S}$, if $f \in \mathcal{L}$ and $f^{*2}(x) \sim 2f(x)$.*
(iii) *f with dist. F is weakly (right-side) subexponential on $\mathbb{R}$, denoted by $\mathcal{S}_+$, if $f \in \mathcal{L}$ and the function $f_+(x) = \mathbf{1}_{\mathbb{R}_+}(x)f(x)/\overline{F}(0)$, $x \in \mathbb{R}$ is subexponential, i.e. $f_+ \in \mathcal{S}$. Here $\overline{F}(x) = 1 - F(x)$.*

Definition 0.2. (i) *We say that a density $f : \mathbb{R} \to \mathbb{R}_+$ is asymptotic to a non-increasing function (a.n.i. for short) if f is locally bounded and positive on $[x_0, \infty)$ for some $x_0 > 0$, and*

$$\sup_{t\geq x} f(t) \sim f(x) \quad \text{and} \quad \inf_{x_0\leq t\leq x} f(t) \sim f(x). \tag{0.1}$$

(ii) *We say that a density $f : \mathbb{R} \to \mathbb{R}_+$ is almost decreasing (al.d. for short) if there exists $x_0 > 0$ and $K > 0$ such that*

$$f(x+y) \leq Kf(x), \quad \text{for all } x > x_0,\ y > 0.$$

Notice that the al.d. property includes the a.n.i. property, and the latter is satisfied by the regularly varying functions with negative indices.

We will investigate properties of the above sort, particularly on infinitely divisible distributions μ on $\mathbb{R}$. The characteristic function (ch.f.) of μ is

$$\widehat{\mu}(z) = \exp\Big\{ \int_{-\infty}^{\infty} (e^{izy} - 1 - izy\mathbf{1}_{\{|y|\leq 1\}})\nu(dy) + iaz - \frac{1}{2}b^2z^2 \Big\}, \tag{0.2}$$

where $a \in \mathbb{R}$, $b \geq 0$ and ν is the Lévy measure satisfying $\nu(\{0\}) = 0$ and $\int_{-\infty}^{\infty}(1 \wedge x^2)\nu(dx) < \infty$. Throughout this paper, we always assume that the Lévy measure ν of μ has a density, and we denote by $\mathrm{ID}(\mathbb{R})$ the class of all infinitely divisible distributions on $\mathbb{R}$.

Main contents

We separate the cases depending on weather $\nu(\mathbb{R}) < \infty$ or $\nu(\mathbb{R}) = \infty$. Note that we use notation g also for the (non-proper) density of a Lévy measure. The followings are the results for the absolutely continuous case ($\nu(\mathbb{R}) = \infty$).

Theorem 0.3. *Let $\mu \in \mathrm{ID}(\mathbb{R})$ with $\nu(dx) = g(x)dx$ such that $\nu(\mathbb{R}) = \infty$. Let $f_0(x)$ be a density of $\mu_0 \in \mathrm{ID}(\mathbb{R})$ with $a = b = 0$ and $\nu(dx) = \mathbf{1}_{\{|x|\le 1\}} g(x)dx$. Suppose that $g_1(x) = \mathbf{1}_{\{x>1\}} g(x)/\nu((1,\infty))$ is bounded, and there exists $\gamma > 0$ such that*

$$\lim_{x\to\infty} e^{\gamma x} f_0(x) = 0. \tag{0.3}$$

For a density f of μ we consider the following properties.

$$\begin{aligned} &\text{(i)} \quad f \in \mathcal{S}_+ \quad \textit{and} \quad f \textit{ is al.d.} \\ &\text{(ii)} \quad g_1 \in \mathcal{S}_+ \\ &\text{(iii)} \quad g_1 \in \mathcal{L} \quad \& \quad \lim_{x\to\infty} f(x)/g_1(x) = \nu((1,\infty)). \end{aligned}$$

(a) If g is a.n.i., then we can choose f such that (i), (ii) *and* (iii) *are equivalent.*
(b) If g is al.d., then we can choose f such that (ii) $\Leftrightarrow$ (iii) *implies* (i).

We could remove several conditions in Theorem 0.3 by assuming the absolute integrability of the spectrally positive part $\widehat{\mu}_+(z)$.

Theorem 0.4. *Let $\mu \in \mathrm{ID}(\mathbb{R})$ with and $\nu(dx) = g(x)dx$ such that $g_1(x)$ is bounded. Suppose that $\int_{-\infty}^{\infty} |\widehat{\mu}_+(z)|dz < \infty$, which implies $\int_{-\infty}^{\infty} |\widehat{\mu}(z)|dz < \infty$, so that μ has a bounded continuous density f. Then the following relations hold between the properties* (i), (ii) *and* (iii) *of Theorem 0.3.*
(a) If g is a.n.i., then we can choose f such that (i), (ii) *and* (iii) *are equivalent.*
(b) If g is al.d., then we can choose f such that (ii) $\Leftrightarrow$ (iii) *implies* (i).

We apply our results to the consistency proof of the maximum likelihood estimation (MLE for short) for $\mu \in \mathrm{ID}(\mathbb{R})$ which is absolutely continuous. For simplicity we put $a = b = 0$ in $\widehat{\mu}(z)$ of (0.2) and assume that the spectrally positive part $\widehat{\mu}_+(z)$ is absolutely integrable.

Let $f(x;\theta)$ be the density of μ with θ a parameter vector and $g(x;\theta)$ be a density of the corresponding Lévy measure ν. Let $(X_1, \dots, X_n)$ be a random sample from $f(x;\theta_0)$ with $\theta_0 \in \Theta$ where Θ is a compact parameter space. Define the likelihood function

$$M_n(\theta) = n^{-1} \sum_{i=1}^{n} \log f(X_i;\theta).$$

MLE $\hat{\theta}_n$ maximizes the function $\theta \mapsto M_n(\theta)$. We say that a function $\alpha(x;\theta)$ is identifiable if $\alpha(\cdot\,;\theta) \neq \alpha(\cdot\,;\theta')$ every $\theta \neq \theta' \in \Theta$, i.e. $\alpha(x;\theta) \overset{a.e.}{=} \alpha(x;\theta')$ does not hold. For convenience, we only consider the symmetric or positive-half case, but we can easily generalize the result in the non-symmetric two-sided case. We use the function g_1 defined in Theorem 0.4.

Proposition 0.5. *Let $\mu \in \mathrm{ID}(\mathbb{R})$ given by* (0.2) *with $a = b = 0$ such that $\widehat{\mu}_+(z)$ is absolutely integrable. Let $g(x;\theta)$ be a symmetric or positive-half density of ν. Suppose* (i) : *$g(x;\theta)$ is identifiable, $\theta \mapsto g(x;\theta)$ is continuous in θ for every x, and $\int (\sup_{\theta\in\Theta} |\log g_1(x;\theta)|) g_1(x;\theta_0)dx < \infty$ with Θ a compact set such that $\theta_0 \in \Theta$. Suppose* (ii) : *$g_1(x;\theta)$ is bounded and a.n.i., and $g_1 \in \mathcal{S}$. Then MLE $\hat{\theta}_n$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

References

[1] Matsui, M. (2022) Subexponentialiy of densities of infinitely divisible distributions (submitted).

「ガンマ分布の形状パラメータのベイズ推定」
東京大学経済学部　　入江　薫

予稿の通り、ガンマ分布の形状パラメータのベイズ推測に関する研究について講演した。講演後には、以下の質疑および議論があった。

- 提案手法のアイディアの確認があった。Polya-inverse gamma 分布という複雑な分布を用いる先行研究とは異なり、提案手法ではベータ分布による簡易な混合表現を見つけていること、その正当化を MH 法によって行っていること、そして MH 法の受容確率が高いことを数学的に確認できることを整理した。
- 主定理の証明について補足した。ガンマ関数の逆数の累乗に対して、分子・分母に引数をずらしたガンマ関数をかけることでベータ関数を作りだすこと、残ったガンマ関数の積に Gauss' multiplication formula を適用することで単一のガンマ関数に書き直せることを指摘した。
- 提案手法が他のパラメトリックな分布に対する MCMC 法に応用可能であることに関連して、一般化された多項分布への適用可能性が示唆された。
- 事前分布として離散分布を用いるアプローチとの比較を求められ、特に不確実性の評価において離散分布によるアプローチは正確性を欠くおそれがある点を指摘した。
- 提案手法の多項ディリクレ分布への応用に関連して、高次元・大規模データの場合の計算可能性について質問があった。理論上は潜在変数のサンプリングの並列化が可能ではあるが、実際に計算時間がどれだけかかるかについては研究課題とした。
- ガンマ母集団分布という単純な設定であるので、形状パラメータの推測においても、ベイズ推定量および MSE は本当に解析的に計算できないのかという質問に対して、既知の特殊関数での記述は難しいと返答した。
- 先行研究の Polya-inverse gamma 分布が無限分解可能であることが指摘され、関連して Polya-gamma 分布などの先行研究を紹介した。

カーネル型推定量を利用した推測について

中央大学理工学部　前園宜彦

1. カーネル型推定量

$X_1, X_2, ..., X_n$ を互いに独立で同じ分布にしたがう確率変数とし，分布関数を F_X，密度関数を f_X とする．この分布関数 F_X の推定量としては経験分布関数

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad x \in \mathbf{R},$$

がある．ここで $I(A)$ は定義関数である．経験分布関数は階段関数となり連続ではない．連続になるような推定量でよく利用されるのがカーネル型推定量である．

カーネル型分布関数推定量（Nadaraya (1964)）は

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

で与えられる．ただし $k(\cdot)$ はカーネル関数とし，$\int_{-\infty}^{\infty} k(v)dv = 1$ の条件を満たす．バンド幅は $n \to \infty$ のとき $h \to 0$ かつ $nh \to \infty$ とする．このとき $K(v) = \int_{-\infty}^{v} k(u)du$ である．適当な条件の下で $n \to \infty$ のとき

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

となる．ただし $r_1 = \int_{-\infty}^{\infty} vk(v)K(v)dv$ で，多くの場合非負の値をとる．

2. 確率点推定

p-確率点は $Q(p) = F^{-1}(p)$ で定義される．経験分布関数に基づく p-確率点 $Q(p)$ の推定量は

$$\widetilde{Q}(p) = F_n^{-1}(p) = \inf\{x; F_n(x) \ge p\}$$

で与えられる．漸近平均二乗誤差は

$$AMSE(\widetilde{Q}(p)) = \frac{p(1-p)}{nf^2(Q(p))} = \frac{\{Q'(p)\}^2 p(1-p)}{n}$$

となる．Falk(1984) や Maesono & Penev(2011) では，カーネル関数 $k(\cdot)$ を利用した滑らかな推定量

$$\widehat{Q}_{p,h} = \frac{1}{h}\int_0^1 F_n^{-1}(x) K\left(\frac{x-p}{h}\right) dx$$

の漸近分布について研究している．漸近平均二乗誤差は

$$AMSE(\widehat{Q}_{p,h}) = \frac{\{Q'(p)\}^2 p(1-p)}{n} + \frac{2h}{n} Q'(p)^2 \left(-\frac{1}{2} + \frac{1}{2}\int_{-1}^{1} K^2(x)dx\right)$$

で与えられる．

3. 順位検定の連続化への応用

$X_1, X_2, \ldots, X_n$ を互いに独立で同じ母集団分布 $F(x-\theta)$ にしたがう無作為標本とする．ここで対応する確率密度関数は $f(-x) = f(x)$ を満たす原点対称な分布とする．θ は未知母数で，帰無仮説 $H_0 : \theta = 0$ に対して対立仮説 $H_1 : \theta > 0$ の検定問題を考える．この問題に対する順位検定統計量の有意確率については Lehmann & D'abrera(2006) でも指摘されているように，検定統計量の取り得る値が細かいほど有意確率が小さくなる傾向がある．

$\psi(x) = 1\ (x \geq 0),\ = 0\ (x < 0)$ とおくとき，符号検定 S とウィルコクソンの符号付き順位検定 W は

$$S = \sum_{i=1}^{n} \psi(X_i), \qquad W = \sum_{1 \leq i \leq j \leq n} \psi(X_i + X_j)$$

である．これら S, W の連続化として

$$\widetilde{S} = n - n\widetilde{F}_n(0) = n - \sum_{i=1}^{n} K\left(-\frac{X_i}{h}\right), \qquad \widetilde{W} = \frac{n(n+1)}{2} - \sum_{1 \leq i \leq j \leq n} K\left(-\frac{X_i + X_j}{2h}\right)$$

が考えられる．この統計量の分布は帰無仮説の下でも母集団分布に依存することになるが，漸近平均と漸近分散は $F(\cdot)$ に依存しない．この $\widetilde{S}$, $\widetilde{W}$ は S, W を滑らかにしたものになっている．

4. 境界バイアス縮小す推定量に基づくノンパラメトリック検定

F を仮定された分布関数とするとき，次の検定問題を考える．

$$\text{帰無仮説 } H_0 : F_X = F \quad \text{対立仮設 } H_1 : F_X \neq F$$

このような一般的な設定の下でよく利用される検定として

$$KS_n = \sup_{x \in \mathbf{R}} |F_n(x) - F(x)|, \qquad CvM_n = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \mathrm{d}F(x)$$

を検定統計量とするコルモゴロフ・スミルノフとクラーメル・フォンミーゼス検定統計量が知られている．

Rizky & Maesono(2022) は全単射関数 g を用いたバイアスを縮小するカーネル型推定量

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

を用いて

$$\widetilde{KS} = \sup_{x \in \mathbf{R}} |\widetilde{F}_X(x) - F(x)|, \qquad \widetilde{CvM} = n \int_{-\infty}^{\infty} [\widetilde{F}_X(x) - F(x)]^2 \mathrm{d}F(x)$$

を提案している．これに対して

$$|KS_n - \widetilde{KS}| \to_p 0, \qquad |CvM_n - \widetilde{CvM}| \to_p 0$$

が成り立つ．このことより，検定統計量 $\widetilde{KS}$ と $\widetilde{CvM}$ を利用した有意確率は，通常のコルモゴロフ・スミルノフ検定統計量およびクラーメル・フォンミーゼス検定統計量に対する数表を利用して検定できる．シミュレーションの結果では，検出力が高いことが示された．

On a generalization of Clayton-Oakes model by R. L. Prentice

Hideatsu Tsukahara*

January 3, 2023

When we have two failure times T_1 and T_2 and our primary interest is in the association between them, we need families of bivariate survivor functions $S(t_1, t_2)$ for statistical modeling. Postulating some desirable properties on certain conditional hazard functions, Clayton [1] derived such a family which is now called *Clayton-Oakes* model. In terms of the associated copula, it is written as

$$C(u_1, u_2) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \vee 0, \quad \theta \in [-1, \infty) \setminus \{0\}.$$

This family is an example of *Archimedean copulas*, which are closely related to the frailty model in survival analysis as shown in Oakes [3]. The *cross ratio function* of T_1 and T_2 is defined by

$$\theta^*(t_1, t_2) := \frac{S(t_1, t_2)\partial_{12} S(t_1, t_2)}{\partial_1 S(t_1, t_2)\partial_2 S(t_1, t_2)}$$

Oakes [3] shows that the cross ratio function is a function of survivor function only if and only if the associated copula is Archimedean. The Clayton-Oakes model amounts to assuming that the cross ratio function is constant.

Prentice [4] proposed a generalization of the Clayton-Oakes model; he wrote "A trivariate generalization with unrestricted marginal and pairwise marginal survivor functions is

$$S(t_1, t_2, t_3) = \big\{S(t_1, t_2, 0)^{-\theta} + S(t_1, 0, t_3)^{-\theta} + S(0, t_2, t_3)^{-\theta} - S(t_1, 0, 0)^{-\theta} - S(0, t_2, 0)^{-\theta} - S(0, 0, t_3)^{-\theta} + 1\big\}^{-1/\theta} \vee 0,$$

where $-1 \le \theta < \infty$." However, this is a functional equation in $S(t_1, t_2, t_3)$, and it is not clear that there exists a 3-dimensional **survivor** function $S(t_1, t_2, t_3)$ satisfying this equation. So the problem should be correctly posed in the following way.

Problem Given three bivariate sf's $S_{12}(t_1, t_2)$, $S_{13}(t_1, t_3)$, $S_{23}(t_2, t_3)$ satisfying the compatiblility conditions

$$\begin{aligned} S_{12}(t_1, 0) = S_{13}(t_1, 0) =: S_1(t_1), \\ S_{12}(0, t_2) = S_{23}(t_2, 0) =: S_2(t_2), \\ S_{13}(0, t_3) = S_{23}(0, t_3) =: S_3(t_3), \end{aligned}$$

is the function G defined by

$$G(t_1, t_2, t_3) = \big\{S_{12}(t_1, t_2)^{-\theta} + S_{13}(t_1, t_3)^{-\theta} + S_{23}(t_2, t_3)^{-\theta} - S_1(t_1)^{-\theta} - S_2(t_2)^{-\theta} - S_3(t_3)^{-\theta} + 1\big\}^{-1/\theta} \vee 0$$

*Faculty of Economics, Seijo University, 6–1–20 Seijo, Setagaya-ku, Tokyo, 157-8511, Japan, E-mail: tsukahar@seijo.ac.jp

a proper survivor function?

The problem can be stated in terms of copulas as well. Let $C_{12}(u_1, u_2)$, $C_{13}(u_1, u_3)$, $C_{23}(u_2, u_3)$ be the copula associated with $S_{12}(t_1, t_2)$, $S_{13}(t_1, t_3)$, $S_{23}(t_2, t_3)$; namely

$$\begin{aligned} S_{12}(t_1, t_2) &= C_{12}(S_1(t_1), S_2(t_2)), \\ S_{13}(t_1, t_3) &= C_{13}(S_1(t_1), S_3(t_3)), \\ S_{23}(t_2, t_3) &= C_{23}(S_2(t_2), S_3(t_3)). \end{aligned}$$

The problem is then reduced to whether the function

$$\widehat{G}(u_1, u_2, u_3) = \big\{ C_{12}(u_1, u_2)^{-\theta} + C_{13}(u_1, u_3)^{-\theta} + C_{23}(u_2, u_3)^{-\theta} - u_1^{-\theta} - u_2^{-\theta} - u_3^{-\theta} + 1 \big\}^{-1/\theta} \vee 0$$

is a (proper) copula for any given copulas C_{12}, C_{13} and C_{23}.

One can easily observe that the answer is "no"; if the answer were yes, then the Fréchet class $\mathcal{F}(F_{12}, F_{13}, F_{23})$ (the set of trivariate distributions with three given bivariate margins) would always be non-empty (see Joe [2, Section 3.4]).

In this expository note, we begin with the review of the original Clayton-Oakes model and the key concept of cross ratio, give an explicit counterexample to the above problem, and discuss under which conditions Prentice's generalization gives a valid survivor function.

References

[1] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, Vol. 65, pp. 141–151, 1978.

[2] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997.

[3] D. Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, Vol. 84, pp. 487–493, 1989.

[4] R. L. Prentice. Higher dimensional clayton–oakes models for multivariate failure time data. *Biometrika*, Vol. 103, pp. 231–236, 2016.

Kendall の順位相関係数を固定した下での最小情報コピュラ

助田 一晟 *　　清智也 †

1　最小情報コピュラ

一般に d 次元コピュラとは全ての周辺分布が $[0,1]$ 上の一様分布である d 次元の分布関数を指し, 任意の d 次元分布は Sklar の定理によりコピュラと周辺分布を用いて表現することができることから, コピュラは 2 変数間の従属性を完全に記述しているとされる. 所与の制約条件下で, 独立コピュラ ($[0,1]^2$ 上の一様密度を持つコピュラ) に Kullback-Leibler 情報量の意味で最も近いコピュラを**最小情報コピュラ**と呼ぶ. このコピュラは以下の最適化問題の解として与えられるものである（例えば Bedford et al. [1]）.

$$\text{minimize} \quad \int_0^1 \int_0^1 p(x,y) \log p(x,y) dxdy \tag{1.1}$$

$$\text{s.t.} \quad \int_0^1 p(x,y)dy = 1, \ \int_0^1 p(x,y)dx = 1 \tag{1.2}$$

$$\int_0^1 \int_0^1 h_k(x,y)p(x,y)dxdy = \mu_k \tag{1.3}$$

最小情報コピュラの離散近似版として**最小情報チェス盤コピュラ**が考えられている [3]. 最小情報コピュラと同様に, エントロピー最大化原理に則った枠組みで定式化を行うことで, 最小情報チェス盤コピュラは以下の最適化問題の解と捉えることができる.

$$\text{minimize} \quad \sum_{i=1}^n \sum_{j=1}^n \pi_{ij} \log \pi_{ij} \tag{1.4}$$

$$\text{s.t.} \quad \sum_{j=1}^n \pi_{ij} = \frac{1}{n}, \sum_{i=1}^n \pi_{ij} = \frac{1}{n} \tag{1.5}$$

$$\sum_{i=1}^n \sum_{j=1}^n h_{k,ij} \pi_{ij} = \mu_k \tag{1.6}$$

ただし, $h_1, \ldots, h_K$ は所与の関数, $h_{k,ij}$ は領域 D_{ij} の中心での $h_k(x,y)$ の値, $\theta = (\theta_1, \ldots, \theta_K)$ は未知パラメータである. この最適化問題は有限次元凸最適化問題となっており, 最適解が一意に存在する.

2　２次の最小情報コピュラへの拡張

従来の最小情報コピュラは所与の制約として 1 次の母数（ある関数の期待値の形で表されるもの）のみを扱っていた. たとえば, Spearman の順位相関係数は 1 次の母数であり, Spearman の順位相関係数を固定した状況は従来の最小情報コピュラの枠組みで扱うことができる. 一方, Spearman の順位相関係数と比較されることが多い Kendall の順位相関係数は 2 次の母数であるため, Kendall の順位相関係数を固定した状況は従来の最小情報コピュラの枠組みで扱うことができない. そこで本研究では所与の制約式を 2 次の母数である Kendall の順位相関係数としたときの最適解（最小情報チェス盤コピュラ）を考察する.

* 東京大学大学院情報理工学系研究科数理情報学専攻
† 東京大学大学院情報理工学系研究科数理情報学専攻

従来の最小情報チェス盤コピュラと同様に, Kendall の順位相関係数を $\tau(\in [-1,1])$ に固定した下での最小情報チェス盤コピュラは以下のように最適化問題の解として定式化される. ただし最後の制約はチェス盤コピュラ上の Kendall の順位相関係数 [2] を固定している.

$$(MP)\ \text{minimize}\quad \sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}\log\pi_{ij} \tag{2.1}$$

$$\text{s.t.}\quad \sum_{j=1}^{n}\pi_{ij}=\frac{1}{n},\ \sum_{i=1}^{n}\pi_{ij}=\frac{1}{n} \tag{2.2}$$

$$1-\mathrm{Tr}(\Xi\Pi\Xi\Pi^{\mathrm{T}})=\tau,\ \Pi=(\pi_{ij}),\Xi=\begin{pmatrix}1 & 0 & \dots & 0\\ 2 & 1 & \dots & 0\\ \dots & & & \\ 2 & & 2 & 1\end{pmatrix} \tag{2.3}$$

まず, 各領域の確率質量（分割表モデルにおける各マス目の値）を移動させる操作を以下のように導入する. チェス盤コピュラ全体の空間を一様コピュラ $U=\frac{1}{n^2}I$ を原点とするベクトル空間 (の部分空間) とみなし, $(n-1)^2$ 本から成る斜交基底 $(T_{ij})_{i=1,\dots,n-1,\ j=1,\dots,n-1}$ を考える. ただし, T_{ij} は i 行 j 列と $i+1$ 行 $j+1$ 列が $+1$, i 行 $j+1$ 列と $i+1$ 行 j 列が -1, それ以外が 0 となっている行列である. 任意のチェス盤コピュラ Π は以下のように一意な係数 a_{ij} を用いて表記できる：$\Pi=U+\sum_{i,j}a_{ij}T_{ij},(i=1,\dots,n-1,j=1,\dots,n-1)$

この操作に対する最適化問題 (MP) の停留条件を考えることで, 以下の主張を得た.

Lemma 1. 任意のチェス盤コピュラ Π と任意の成分移動操作 T_{ij} と $T_{i'j'}$ に対し,

$$\Pi':=\Pi+\epsilon T_{ij}-r\epsilon T_{i'j'},\ r:=\frac{\pi_{i,j}+\pi_{i+1,j}+\pi_{i+1,j}+\pi_{i+1,j+1}}{\pi_{i',j'}+\pi_{i'+1,j'}+\pi_{i'+1,j'}+\pi_{i'+1,j'+1}} \tag{2.4}$$

と定める. このとき,

$$1-\mathrm{Tr}(\Xi\Pi\Xi\Pi^{\mathrm{T}})=1-\mathrm{Tr}(\Xi\Pi'\Xi{\Pi'}^{\mathrm{T}})+O(\epsilon^2) \tag{2.5}$$

Lemma 2. チェス盤コピュラ $\Pi=(\pi_{ij})$ に対し, 任意の成分移動操作 T_{ij} を微小に行なった時の情報量の変分は

$$\log\frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}$$

Theorem 1 (2 次の最小情報チェス盤コピュラの不変量)**.** 2 次の最小情報チェス盤コピュラでは任意の (i,j) の組み $(i,j=1,\dots,n-1)$ に対し, 以下の値が一定.

$$\frac{1}{\pi_{i,j}+\pi_{i+1,j}+\pi_{i+1,j}+\pi_{i+1,j+1}}\log\frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}$$

また, 発表ではその他 Total Positivity や幾何構造等の諸性質や, 提案した最小情報コピュラモデルを実際の株価データにモーメント法を基にフィッティングさせた結果を紹介した.

参考文献

[1] Bedford, T. and Wilson, K. J., On the construction of minimum information bivariate copula families, Ann. Inst. Stat. Math., 66, 703–723, 2014.

[2] Durrleman, V., Nikeghbali, A. and Roncalli, T., Copulas Approximation and New Families, Available at SSRN, 2000.

[3] Piantadosi, J., Howlett, P., and Boland, J., Copulas with maximum entropy, Optim. Lett., 6, 99–125, 2012.

スピアマンランク行列による主成分分析のロバストネス

千葉大・融合理工学府　渡邊 宏大
千葉大・理学研究院　内藤 貫太

はじめに: ロバストな主成分分析については，これまで多くの手法が提案されている．本発表では，Marden (1999) の Rank から派生したスピアマンランク行列に基づく主成分分析のロバスト性について報告した．Han and Liu (2018) で議論されている Kendall's tau との比較についても考察した．

設定と定義: $\boldsymbol{y} \in \mathbb{R}^d$ の Spatial sign を

$$S(\boldsymbol{y}) = \begin{cases} \dfrac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} & , \quad \boldsymbol{y} \neq \boldsymbol{0} \\ \boldsymbol{0} & , \quad \boldsymbol{y} = \boldsymbol{0} \end{cases}$$

で定義する．$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ を有限な分散共分散行列 Σ を持つ d 次元確率分布 F にしたがう互いに独立な確率ベクトルとする．Spatial sign に基づくロバストな主成分分析の手法が，Marden (1999)，Han and Liu (2018) で議論されている．

母集団での Kendall's tau matrix $\boldsymbol{K} \in \mathbb{R}^{d \times d}$ は

$$\boldsymbol{K} = E_{\boldsymbol{X},\boldsymbol{Y}}\left[S(\boldsymbol{X}-\boldsymbol{Y})S(\boldsymbol{X}-\boldsymbol{Y})^T\right]$$

で定義され，その推定量は

$$\widehat{\boldsymbol{K}} = \frac{2}{n(n-1)} \sum_{j<i} S(\boldsymbol{X}_i - \boldsymbol{X}_j)S(\boldsymbol{X}_i - \boldsymbol{X}_j)^T$$

で定義される．一方，母集団でのスピアマンランク行列 $\Sigma_R \in \mathbb{R}^{d \times d}$ は

$$\Sigma_R = E_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}\left[S(\boldsymbol{X}-\boldsymbol{Y})S(\boldsymbol{X}-\boldsymbol{Z})^T\right] = E_{\boldsymbol{X}}\left[E_{\boldsymbol{Y}}\left[S(\boldsymbol{X}-\boldsymbol{Y})\right] E_{\boldsymbol{Z}}\left[S(\boldsymbol{X}-\boldsymbol{Z})\right]^T\right]$$

で定義され，その推定量は

$$\widehat{\Sigma}_R = \frac{1}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j\neq i, k\neq i, j\neq k} S(\boldsymbol{X}_i - \boldsymbol{X}_j)S(\boldsymbol{X}_i - \boldsymbol{X}_k)^T$$

で定義される．

問題: 重要な事実として，F が楕円分布のとき，Σ，$\boldsymbol{K}$ および Σ_R は同じ直交行列で対角化されることが知られている (Marden (1999)，Han and Liu (2018) を参照)．Marden (1999) においても，Σ_R の推定量 $\widetilde{\Sigma}_R$ によるロバストな主成分分析が議論されているが，その $\widetilde{\Sigma}_R$ は

$$\widetilde{\Sigma}_R = \frac{1}{n}\widehat{\boldsymbol{K}} + \left(1 - \frac{2}{n}\right)\widehat{\Sigma}_R$$

と分解される．$n \to \infty$ のとき，$\widehat{\boldsymbol{K}}$ は $\boldsymbol{K}$ に，$\widehat{\Sigma}_R$ は Σ_R に確率収束することが示される．このことから，$\widetilde{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ の比較は，漸近的には $\widehat{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ の比較となる．我々の問題は特に，$\widehat{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ はどちらがロバストな主成分分析を提供するのか？ということになる．

影響関数: ロバストネスの指標として影響関数を考える (影響関数については，Hampel et al. (1986) を参照)．F を楕円分布とする．比較の対象として，$\boldsymbol{K}$ と Σ_R の最大固有値に関する影響関数および固有ベクトルに関する影響関数を与える．

$\boldsymbol{K}$ は直交行列 $\Gamma = [\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_d]$ と対角行列 $A = \mathrm{diag}(a_1, \ldots, a_d)$，$a_1 > \cdots > a_d > 0$ によって $\boldsymbol{K} = \Gamma A \Gamma^T$ とスペクトル分解されるとする．上で述べた重要な事実から，Σ_R も同じ直交行列によって $\Gamma \Lambda \Gamma^T$，$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ とスペクトル分解できる．ただし，対応する固有値については，$\lambda_1 > \cdots > \lambda_d > 0$ となるかは定かではない．したがって，$\boldsymbol{\gamma}_1$ が属している固有値を λ_M，$M \in \{1, \ldots, d\}$，Σ_R の最大固有値と対応する固有ベクトルをそれぞれ λ_{max}，$\boldsymbol{\gamma}_L$，$L \in \{1, \ldots, d\}$ とする．このとき，$\boldsymbol{K}$ と Σ_R の最大固有値に関する影響関数 $IF_{\boldsymbol{K}}(\boldsymbol{x}, a_1, F)$，$IF_{\Sigma_R}(\boldsymbol{x}, \lambda_{max}, F)$ はそれぞれ

$$IF_{\boldsymbol{K}}(\boldsymbol{x}, a_1, F) = \boldsymbol{\gamma}_1^T A(\boldsymbol{x}) \boldsymbol{\gamma}_1 - 2a_1, \tag{1}$$

$$IF_{\Sigma_R}(\boldsymbol{x}, \lambda_{max}, F) = \boldsymbol{\gamma}_L^T B(\boldsymbol{x}) \boldsymbol{\gamma}_L - 3\lambda_{max} \tag{2}$$

で与えられる．ここで，

$$A(\boldsymbol{x}) = 2E_{\boldsymbol{X}} \left[S(\boldsymbol{x} - \boldsymbol{X}) S(\boldsymbol{x} - \boldsymbol{X})^T \right],$$
$$B(\boldsymbol{x}) = E_{\boldsymbol{X}, \boldsymbol{Y}} \left[S(\boldsymbol{X} - \boldsymbol{Y}) S(\boldsymbol{X} - \boldsymbol{x})^T + S(\boldsymbol{X} - \boldsymbol{x}) S(\boldsymbol{X} - \boldsymbol{Y})^T + S(\boldsymbol{x} - \boldsymbol{X}) S(\boldsymbol{x} - \boldsymbol{Y})^T \right]$$

である．また，$\boldsymbol{K}$ と Σ_R の固有ベクトル $\boldsymbol{\gamma}_1$ に関する影響関数 $IF_{\boldsymbol{K}}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F)$，$IF_{\Sigma_R}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F)$ はそれぞれ

$$IF_{\boldsymbol{K}}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F) = \sum_{k=2}^{d} \frac{1}{a_1 - a_k} \boldsymbol{\gamma}_k^T A(\boldsymbol{x}) \boldsymbol{\gamma}_1 \cdot \boldsymbol{\gamma}_k, \tag{3}$$

$$IF_{\Sigma_R}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F) = \sum_{k \neq M} \frac{1}{\lambda_M - \lambda_k} \boldsymbol{\gamma}_k^T B(\boldsymbol{x}) \boldsymbol{\gamma}_1 \cdot \boldsymbol{\gamma}_k \tag{4}$$

で与えられる．

ロバストネスの比較と適用例： 本講演においては，いくつかの 2 次元楕円分布のもとで，固有値，固有ベクトルの影響関数 (1)，(2)，(3)，(4) の曲面を描画することおよび固有空間の類似度の比較を通して，$\widehat{\Sigma}_R$ による主成分分析のロバスト性を報告した．外れ値を含む実データへの適用結果についても報告した．

参考文献

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley.

Han, F. and Liu, H. (2018). Eca: High-dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, 113:252–268.

Marden, J. I. (1999). Some robust estimates of principal components. *Statistics and Probability Letters*, 43:349–359.