シンポジウム「統計科学の現代的課題」報告書

本シンポジウムは科学研究費・基盤研究(A)「大規模複雑データの理論と方法論の総合的研究」(研究代表者:青嶋 誠(筑波大学), 課題番号:15H01678)の助成により開催した。

■日時: 2017年1月27日～1月29日

■場所: 金沢大学サテライトプラザ

■開催責任者: 星野 伸明 (金沢大学)

【 プログラム・目次 】

	国友直人＞	14:10 高橋佳苗	大阪大学医学部附属病院	診断研究における検査性能の比較	23
		14:50 石原拓磨	大阪大学大学院　医学系研究科	複数の二値評価変数をもつ臨床試験における検定法	25
		15:30 休憩			
	セッション6＜座長：塚原英敦＞	15:45 新村秀一	成蹊大学 経済学部	R.A.Fisher 以後の判別分析の新理論と遺伝子解析の新手法	27
		16:25 山本けい子	函館工業高等専門学校一般理数系	産業連関表に基づく都道府県クラスターと産業構造推移の可視化	29
		17:05 山縣 一慶	上智大学大学院理工学研究科数学領域	テキストマイニングにもとづくレビューのスコアリングを用いた映画の統計的分類	31
		17:45 竹下 佳宏	上智大学大学院理工学研究科数学領域	規模の異なる変量群をもつロジスティック回帰モデルの係数2段階推定	33
		18:25 終了			
		19:00 懇親会			
1月29日	セッション7＜座長：松井宗也＞	10:00 伊藤伸介	中央大学経済学部	国勢調査の匿名化ミクロデータの作成可能性について―地域区分に着目して―	35
		10:40 紙屋英彦	大阪経済大学経済学部	Estimation of the shape of density level sets of star-shaped distributions	37
		11:20 吉田知行	北星学園大学経済学部	正方分割表の一致率検定のための代数的方法	39
		12:20 終了			

経済不平等度を予測するためのモデルについて

広島大学大学院社会科学研究科　西埜晴久*

日本経済における経済の不平等度の研究は，2000 年代初めに橘木俊詔（1998) および大竹文雄（2005）において行われてきて，大きな関心を集めてきた．そこで，不平等度の時系列的な変動についても興味深い問題対象であると考えている．

本報告では，Nishino, H. and K. Kakamu (2011) で用いられているように対数正規分布を仮定して分位データを用いてモデル化する．なお，日本の家計調査のデータに対して対数正規分布は適合的である. そして，本報告ではそのモデルに GARCH 型の時系列モデルを組み込むことで，不平等度を予測するための時系列モデルを組み立てることを試みるものである．

不平等度を取り扱う日本のデータとして，本報告では総務省統計局の家計調査を用いる．対数正規分布は (μ, σ) の二つのパラメータをもつ分布である．一方，対数正規分布のジニ係数は

$$G = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1, \tag{1}$$

とあらわされる．(たとえば，Kleiber and Kotz (2003) を参照のこと）. つまり ジニ係数は σ だけの単調増加関数であり，σ の大きさが不平等度の大きさを表すことになる．また，2 つのパラメータを持つ分布のなかでは対数正規分布が家計調査の 5 分位のデータには AIC で当てはまりを比較する限り，もっとも適合であることが分かる．

標本サイズを n, 部分標本サイズを k としたときに，分位データは部分順序統計量として $\{X(n_1) \leq X(n_2) \leq, \ldots, \leq X(n_k)\}$ $(1 \leq n_1 < n_2 < \ldots < n_k \leq n)$ と書ける．家計調査の 5 分位データでは，$n = 10000, k = 4, (n_1, n_2, n_3, n_4) = (2000, 4000, 6000, 8000)$ となる．また，$(x_1, x_2, \ldots, x_k)$ を $\{X(n_1), X(n_2), \ldots, X(n_k)\}$ の観測値とする．また，対数変換値 $\mathbf{y} = (y_1, y_2, \ldots, y_k)' = (\log x_1, \log x_2, \ldots, \log x_k)'$ とする．なお，対数正規分布を対数変換すると正規分布になる．

部分順序統計量の同時密度は，David and Nagaraja (2003) から，

$$\begin{aligned} &f(\mathbf{y}|\boldsymbol{\theta}) = \\ &n!\frac{F(y_i)^{n_1-1}}{(n_1-1)!}f(y_1)\left\{\prod_{i=2}^{k}\frac{(F(y_i)-F(y_{i-1}))^{n_i-n_{i-1}-1}}{(n_i-n_{i-1}-1)!}f(y_i)\right\} \\ &\times\frac{(1-F(x_k))^{n-n_k}}{(n-n_k)!} \end{aligned} \tag{2}$$

と書ける．ただし，$F(\cdot)$ は累積分布関数であり，$f(\cdot)$ は確率密度関数である．

グループ•データを $(x_1, \ldots, x_k)$ とし，その対数変換を $(y_1, y_2, \ldots, y_k) = (\log x_1, \log x_2, \ldots, \log x_k)$ とする．対数正規分布を仮定しているため, $Y = \log X \sim \mathcal{N}(\mu, \sigma^2)$ となる．

*〒 739-8525 東広島市鏡山 1-2-1. Email: hnishino@hiroshima-u.ac.jp

$\mathbf{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{k,t})'$ と表示すると，GARCH(1,1) model は

$$\begin{aligned} \mathbf{y}_t &\sim f(\mathbf{y}_t|\mu_t, \sigma_t), && (3)\\ \sigma_t^2 &= \alpha_0 + \alpha_1\sigma_{t-1}^2 + \beta_1\tilde{\sigma}_{t-1}^2, \ (GARCH) && (4)\\ \sigma_t &= exp\left\{\alpha_0 + \alpha_1 \log(\sigma_{t-1}) + \beta_1 \log(\tilde{\sigma}_{t-1})\right\}, \\ &\qquad (EGARCH) && (5) \end{aligned}$$

と書ける．ただし，$\{\tilde{\sigma}_t\}$ は $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{k,t})'$ からの σ_t の推定値である．

また，推定は以下の疑似最尤法を用いる．つまり，

$$logL = \sum_{t=2}^{T} \log f(\mathbf{y}_t|\mu_t, \sigma_t) \tag{6}$$

を最大化するパラメータを推定値とする．

次に一期先予測は，例えば，EGARCH(1,1) に対しては，

$$\hat{\sigma}_t = exp\left\{\hat{\alpha}_0 + \hat{\alpha}_1 \log(\sigma_{t-1}) + \hat{\beta}_1 \log(\tilde{\sigma}_{t-1})\right\}, \tag{7}$$

とする．ただし，$\{\tilde{\sigma}_t\}$ は $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{k,t})'$ からの σ_t の推定値であり，$\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_1$ パラメータの推定値である．さらに，予測の観点からモデルを比較すると平均予測誤差の平方根 (root MSE) を

$$\sqrt{\frac{1}{T-1}\sum_{t=2}^{T}\{\hat{\sigma}_t - \tilde{\sigma}_t\}^2} \tag{8}$$

用いることにする．なお，root MSE によって家計調査のデータにモデルの当てはまりを判断すれば，共に $\alpha_1 = 0$ となる，GARCH(0,1) 次いで，EGARCH(0,1) が良いことが分かった．しかし，さらに当てはまりの良いモデルを開発することを今後検討している．

References

[1] David, H. A. and Nagaraja, H. N. (2003) *Order Statistics*, 3rd ed. Wiley, New York.

[2] Kleiber C. and Kotz, S. (2003) *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, New York.

[3] Nishino, H. and Kakamu, K. (2011) "Grouped Data Estimation and Testing of Gini Coefficients using Lognormal Distributions," *Sankhyā B* **73**, 193–210.

[4] 大竹文雄 (2005)『日本の不平等』, 日本経済新聞社.

[5] 橘木俊昭 (1998)『日本の経済格差』, 岩波新書.

組み合わせ論と無限分解可能分布のひとつの接点

南山大学　松井 宗也

研究内容

組み合わせ論では、その解析に母関数 (Generating function) がよく用いられ、諸処の問題に有効となることも多い (Wilf [5])。この母関数は、確率・統計において非負整数値の確率変数の確率母関数に対応する。本報告では、対称群の cycle index polynomials が非負整数値の無限分解可能分布（それは複合ポアソン分布と同値）の確率関数に対応することに注目する。そして無限分解可能分布に関する既知の結果をもとに、組み合わせ論の新しい定理を得る。さらに、その定理の組み合せ論的な証明や、その定理から敷衍される結果も述べる。

より具体的な内容を述べる。非負整数列 $(a_k)_{k\geq 0}$、$(b_k)_{k\geq 0}$ と $(c_k)_{k\geq 1}$ は以下の関係を満たすものとする。

$$\sum_{k=0}^{\infty} a_k u^k = \sum_{k=0}^{\infty} \frac{b_k u^k}{k!} = \exp\Big(\sum_{j=1}^{\infty} \frac{c_j u^j}{j}\Big). \tag{1}$$

このとき数列 $(a_k)_{k\geq 0}$ と $(b_k)_{k\geq 0}$ の性質を $(c_k)_{k\geq 1}$ のそれから導くという問題を考える。Bender and Canfield [1] は、数列 $(c_k)_{k\geq 1}$ と $c_0 = 1$ からなる数列が log-concave であると仮定し、そのうえで数列 (a_k) が log-concave と almost log-convex であることを示した。厳密には不等式：

$$a_{k-1}a_{k+1} \leq a_k^2 \leq \frac{k+1}{k} a_{k-1}a_{k+1},$$
$$b_{k-1}b_{k+1} \geq b_k^2 \geq \frac{k}{k+1} b_{k-1}b_{k+1}$$

が成り立つことを示した[1]。それは組み合わせ論的な証明による。研究の主結果は、対応する関係を log-convex のそれに置き換えたものへ拡張したことである。つまり、数列 $(c_k)_{k\geq 1}$ と $c_0 = 1$ の log-convexity から数列 (a_k) のそれを導いた。詳しくは以下の同値な不等式を示した。

$$a_{k-1}a_{k+1} \geq a_k^2, \qquad b_k^2 \leq \frac{k}{k+1} b_{k-1}b_{k+1}.$$

証明は、式 (1) の確率的解釈と非負整数の値を取る複合ポアソン分布の性質による。この証明方法は、最初の log-concavity の関係の別証にもなる。

次に、拡張された結果 (log-convex に関する) の組み合わせ論的な証明も与える。組み合わせ論的な証明では log-concave/convex のみならず、更に強い結果を得られ

[1] ほとんど log-convex であることは第 1 式の左側の不等式を指す。なお、第 1 式の右側の不等式と第 2 式の左側の不等式は同値である。

る。以下に述べるのは Bender and Canfield [1] の組み合わせ論による結果とその拡張の両方である。まず非負数列 (c_k) から構成される集合を以下に定義する。

$$\begin{aligned}\mathscr{X} &= \{c_1, c_2, \ldots\},\\ \mathscr{Y} &= \mathscr{X} \cup \{c_j c_k - c_{j-1}c_{k+1} : 0 < j \le k\},\\ \mathscr{Z} &= \mathscr{X} \cup \{c_{j-1}c_{k+1} - c_j c_k : 0 < j \le k\}.\end{aligned}$$

すると、式 (1) を満たす数列 (b_k) の log-concave/convex form を (c_k) のそれの多項式で以下のように表せる。

$$\begin{aligned}&(n+1)b_m b_n - m b_{m-1} b_{n+1} \in \mathbb{N}[\mathscr{Y}],\\ &m b_{m-1} b_{n+1} - (n+1) b_m b_n \in \mathbb{N}[\mathscr{Z}]/\mathbb{N}[\mathscr{X}] \quad \text{for} \quad 1 \le m \le n,\end{aligned}$$

ここで $\mathbb{N}[\mathscr{Y}]$ は、集合 $\mathscr{Y}$ の要素からなり、かつ非負整数の係数を持つ多項式である。$\mathbb{N}[\mathscr{Z}]$ と $\mathbb{N}[\mathscr{X}]$ も同様に定義される。最初の式が Bender and Canfield [1] による log-concave に関する結果であり、次の式がそれを log-convex へ拡張したものである。

さらに、式 (1) を満たす数列 (a_k) や (b_k) の convolution や binomial convolution によりできる数列の性質を与える。式 (1) の積により (a_k) の convolution や (b_k) の binomial convolution が得られる。この操作により得られる数列は、これまでに確立された convolution に対する log-convex/concave の法則に必ずしも従わない。これをみるために、離散確率変数の確率関数を数列とみなし、それにより具体例を構成した。

キーワード:Cycle index polynomials, compound Poisson, symmetric group, infinitely divisible, generating function, log-convexity.

参考文献

[1] E. A. Bender and E. R. Canfield (1996) Log-concavity and related properties of the cycle index polynomials, *J. Combin. Theory Ser. A* **74**, 57–70.

[2] B. G. Hansen (1988) On log-concave and log-convex infinitely divisible sequences and densities. *Ann. Probab.* **16**, 1832–1839.

[3] M. Matsui (2016) Log-convexity and the cycle index polynomials with relation to compound Poisson distributions, ArXiv:1609.06875.

[4] F. W. Steutel and K. van Harn (2004) *Infinite Divisibility of Probability Distributions on the Real Line.* Marcel Dekker, New York.

[5] H. S. Wilf (1994) *Generatingfunctionology.* 2nd. ed. Academic Press, San Diego.

多変量確率密度関数の二重対称性について

東京理科大学 理工学部　　生亀清貴*

Tomizawa (1985) は二元分割表において二重対称モデル，準二重対称モデル，周辺二重対称モデルを提案し, 定理「二重対称モデルが成り立つための必要十分条件は，準二重対称モデルと周辺二重対称モデルの両方が成り立つことである」を与えた. Yamamoto et al. (2012) はこの定理を多元分割表に拡張した.

本講演では多変量確率密度関数に対して二重対称性, 準二重対称性と周辺二重対称性を定義し, 二重対称性をもつ確率密度関数の分解を与える.

確率変数 $(X_1, \ldots, X_T)$ は確率密度関数 $f(x_1, \ldots, x_T)$ をもつとする. ただし

$$f(x_1, \ldots, x_T) > 0 \text{ for } (x_1, \ldots, x_T) \in D^T,$$
$$f(x_1, \ldots, x_T) = 0 \text{ for } (x_1, \ldots, x_T) \notin D^T,$$

また

$$D^T = \{(x_1, \ldots, x_T) \mid a_i < x_i < b_i;\ i = 1, \ldots, T\},$$

ただし $a_i = -\infty$ かつ $b_i = +\infty$, または a_i, b_i は共に有限とする. $(c_1, \ldots, c_T)$ を D^T 内の点とする, ただし a_i, b_i が共に有限の場合は $c_i = (a_i + b_i)/2$. また任意の i $(i = 1, \ldots, T)$ に対して, $X_i = x_i$ のとき $x_i^* = 2c_i - x_i$ とおく.

確率密度関数 $f(x_1, \ldots, x_T)$ の二重対称性を次のように定義する：$(1, \ldots, T)$ の任意の並べ替え $(\pi_1, \ldots, \pi_T)$ に対して,

$$\begin{aligned} f(x_1, \ldots, x_T) &= f(x_{\pi_1}, \ldots, x_{\pi_T}) \\ &= f(x_1^*, \ldots, x_T^*). \end{aligned}$$

また任意の k $(k = 1, \ldots, T-1)$ に対して, k 次周辺二重対称性を次のように定義する：

$$\begin{aligned} f_{X_{i_1} \ldots X_{i_k}}(x_{i_1}, \ldots, x_{i_k}) &= f_{X_{i_1} \ldots X_{i_k}}(x_{\pi_{i_1}}, \ldots, x_{\pi_{i_k}}) \\ &= f_{X_{j_1} \ldots X_{j_k}}(x_{i_1}, \ldots, x_{i_k}) \\ &= f_{X_{i_1} \ldots X_{i_k}}(x_{i_1}^*, \ldots, x_{i_k}^*) \\ (1 \le i_1 < \cdots < i_k \le T; 1 \le j_1 < \cdots < j_k \le T), & \end{aligned}$$

ただし $(\pi_{i_1}, \ldots, \pi_{i_k})$ は $(i_1, \ldots, i_k)$ の任意の並べ替え, $f_{X_{i_1} \ldots X_{i_k}}$ は $(X_{i_1}, \ldots, X_{i_k})$ の周辺確率密度関数とする.

2010 Mathematics Subject Classification: 62H05
キーワード：二重対称性, 準二重対称性, 周辺二重対称性，確率密度関数, 分解
* 〒 278-8510 千葉県野田市山崎 2641
e-mail: iki@is.noda.tus.ac.jp

一般に, 確率密度関数は

$$f(x_1,\ldots,x_T)=\alpha\Big[\prod_{i_1=1}^{T}\alpha_{i_1}(x_{i_1})\Big]\Big[\prod_{1\le i_1<i_2\le T}\prod\alpha_{i_1i_2}(x_{i_1},x_{i_2})\Big]\times\cdots$$
$$\times\Big[\prod_{1\le i_1<\cdots<i_{T-1}\le T}\prod\cdots\prod\alpha_{i_1\ldots i_{T-1}}(x_{i_1},\ldots,x_{i_{T-1}})\Big]\cdot\alpha_{1\ldots T}(x_1,\ldots,x_T),\qquad(1)$$

のように表すことが可能である, ただし $(x_1,\ldots,x_T)\in D^T$,

$$\{\alpha_i(c_i)=\alpha_{i_1i_2}(c_{i_1},x_{i_2})=\cdots=\alpha_{1\ldots T}(x_1,\ldots,x_{T-1},c_T)=1\}.$$

このとき, 確率密度関数 $f(x_1,\ldots,x_T)$ が二重対称であるための必要十分条件は, 式 (1) に次の制約を課すことである：

$$\begin{aligned}\alpha_{i_1\ldots i_m}(x_{i_1},\ldots,x_{i_m})&=\alpha_{i_1\ldots i_m}(x_{\pi_{i_1}},\ldots,x_{\pi_{i_m}})\\&=\alpha_{j_1\ldots j_m}(x_{i_1},\ldots,x_{i_m})\\&=\alpha_{i_1\ldots i_m}(x^*_{i_1},\ldots,x^*_{i_m})\\(m=1,\ldots,T;1\le i_1<\cdots<i_m\le T;1\le j_1<\cdots<j_m\le T),&\end{aligned}$$

ただし $(\pi_{i_1},\ldots,\pi_{i_m})$ は $(i_1,\ldots,i_m)$ の任意の並べ替え. 任意の k $(k=1,\ldots,T-1)$ に対して, $f(x_1,\ldots,x_T)$ の k 次準二重対称性を次のように定義する：式 (1) に対して,

$$\begin{aligned}\alpha_{i_1\ldots i_m}(x_{i_1},\ldots,x_{i_m})&=\alpha_{i_1\ldots i_m}(x_{\pi_{i_1}},\ldots,x_{\pi_{i_m}})\\&=\alpha_{j_1\ldots j_m}(x_{i_1},\ldots,x_{i_m})\\&=\alpha_{i_1\ldots i_m}(x^*_{i_1},\ldots,x^*_{i_m})\\(m=k+1,\ldots,T;1\le i_1<\cdots<i_m\le T;1\le j_1<\cdots<j_m\le T),&\end{aligned}$$

ただし $(\pi_{i_1},\ldots,\pi_{i_m})$ は $(i_1,\ldots,i_m)$ の任意の並べ替え.

このとき, 次の定理を得る.

定理： 任意の固定された k に対して $(k=1,\ldots,T-1)$, 確率密度関数 $f(x_1,\ldots,x_T)$ が二重対称性を満たすための必要十分条件は, $f(x_1,\ldots,x_T)$ が k 次準二重対称性と k 次周辺二重対称性の両方を満たすことである.

参考文献

[1] Tomizawa, S. (1985). Double symmetry model and its decomposition in a square contingency table. *Journal of the Japan Statistical Society*, **15**, 17-23.

[2] Yamamoto, K., Takahashi, F. and Tomizawa, S. (2012). Double symmetry model and its orthogonal decomposition for multi-way tables. *SUT Journal of Mathematics*, **48**, 83-102.

二標本ノンパラメトリック検定の連続化と局所漸近検出力

(Smoothed two-sample nonparametric tests and their local asymptotic powers)

Taku MORIYAMA (Graduate School of Mathematics, Kyushu University)[*1]
Yoshihiko MAESONO (Faculty of Mathematics, Kyushu University)[*2]

1. Introduction

Let $X_1, X_2, \cdots, X_m$ be independently and identically distributed random variables (*i.i.d.*) from distribution function $F(x)$ and $Y_1, Y_2, \cdots, Y_n$ be *i.i.d.* from $F(x-\theta)$ where θ is unknown location parameter. We assume that $m+n=N$ and $0 < \lim_N m/N = \lim_N \lambda_N = \lambda < 1$. We consider '2-sample problem' whose null hypothesis $H_0 : \theta = 0$ and $H_1 : \theta > 0$.

Let us define that $\psi(x) = 1$ $(x \geq 0)$, $= 0$ $(x < 0)$ and $Z_1 = X_1, \cdots, Z_m = X_m, \cdots, Z_{m+1} = Y_1, \cdots, Z_N = Y_n$. The median and Wilcoxon's rank sum test statistic are

$$M = \sum_{j=1}^{n} \psi(Y_j - Z), \quad W_2 = \sum_{1 \leq i \leq m} \sum_{1 \leq j \leq n} \psi(Y_j - X_i).$$

where Z denotes the sample median of $\{Z_1, Z_2, \cdots, Z_N\}$. For observed values $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^T$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$, we put $s = s(\boldsymbol{x}, \boldsymbol{y})$ and $w_2 = w_2(\boldsymbol{x}, \boldsymbol{y})$ are the realized value of M and W_2. If the p-values $P_0(M \geq m)$ or $P_0(W_2 \geq w_2)$ is small enough, we reject the null hypothesis H_0.

2. The smoothed nonparametric tests

We propose the following smoothed median test statistic

$$\widetilde{M} = \sum_{i=1}^{m} K^* \left(\frac{Z - X_i}{h} \right)$$

where K^* is an integral of the kernel $k^*(t)$ which satisfies

$$\int_0^\infty k^*(t)dt = 1 \quad \text{and} \quad k^*(t) = 0 \ \text{ for } \ t \leq 0$$

and h is a bandwidth which satisfies $h \to 0$ and $nh \to \infty$. Further,we propose the following Wilcoxon's rank sum test

$$\widetilde{W}_2 = \sum_{i=1}^{m} \sum_{j=1}^{n} K \left(\frac{Y_j - X_i}{h} \right)$$

where K is an integral of a symmetric kernel k. We can easily prove their asymptotic normality and that their Pitman efficiency are same as the discretes :

$$e_P[\widetilde{M}] = e_P[M^\dagger], \quad e_P[\widetilde{W}_2] = e_P[W_2]$$

where $M^\dagger$ is equivalent to M.

Keywords: nonparametric test, kernel estimator, significance probability, local asymptotic power.
[*1] e-mail: moritaku3542168@gmail.com
[*2] e-mail: maesono@math.kyushu-u.ac.jp

Table 1: Comparison of significance probabilities of M and W_2

sample size (m,n)	(10,10)	(20,20)	(30,30)	(10,20)	(20,10)	(U_m^*, U_n^*)[1]
$z_{0.9}$	2.21	2.79	1.68	7.13	7.07	3.86
$z_{0.95}$	6.37	5.66	2.17	3.44	3.41	4.21
$z_{0.975}$	3.05	2.80	3.72	16.6	14.7	4.10
$z_{0.99}$	33.7	7.54	1.56	6.05	5.29	4.11

3. Significance probability

Table 1 shows the ratio of frequency of exact p-value of W_2 getting smaller than M in the tale area Ω_α

$$\Omega_\alpha = \left\{ \boldsymbol{x} \in \boldsymbol{R}^n \;\middle|\; \frac{m(\boldsymbol{x}) - E_0(M)}{\sqrt{V_0(M)}} \geq v_{1-\alpha}, \quad \text{or} \quad \frac{w_2(\boldsymbol{x}) - E_0(W_2)}{\sqrt{V_0(W_2)}} \geq v_{1-\alpha} \right\}$$

where $v_{1-\alpha}$ is a $(1-\alpha)th$ quantile of the standard normal distribution $N(0,1)$, and $E_0(\cdot)$ and $V_0(\cdot)$ stand for an expectation and a variance under H_0, respectively. We count samples that an exact p-value of the test is smaller than the other in the tail area Ω_α, and calculate the ratio of the frequency.

Because the values in Table 1 are larger than 1, we find that W_2 tends to have smaller p-value than M. They can let one to use W_2 if one wants the small value. This comes from that the possible p-value of M is more sparse than W_2 like the sign test and Wilcoxon's signed rank test as is discussed by Maesono et al.(2016). On the other hand, we can find the ratio values of $\widetilde{M}$ and $\widetilde{W}_2$ in table 2 are close to 1.

Table 2: Relation of significance probabilities of $\widetilde{M}$ and $\widetilde{W}_2$

sample size (m,n)	(10,10)	(20,20)	(30,30)	(10,20)	(20,10)	(U_m^*, U_n^*)[1]
$z_{0.9}$	1.41	1.08	1.31	1.05	1.22	0.827
$z_{0.95}$	0.829	1.21	1.39	1.41	1.24	0.782
$z_{0.975}$	1.64	1.12	1.09	1.23	1.39	0.755
$z_{0.99}$	0.985	1.03	1.32	0.707	0.806	0.677

References

[1] Hájek, J., Sidak, Z. & Sen, P. K. (1999). Theory of rank tests, Academic press.

[2] Maesono, Y., Moriyama, T., & Lu, M. (2016). Smoothed nonparametric tests and their properties. *arXiv preprint arXiv:1610.02145*.

[1] U_m^* and U_n^* are realized values of the discrete uniform distribution $U^*(5,40)$

氏　　名：蛭川雅之

所　　属：摂南大学経済学部

講演題目：“Nonparametric Estimation and Testing on Discontinuity of Positive Supported Densities: A Kernel Truncation Approach” [Benedikt Funke (Technical University of Dortmund)との共著]

講演内容：

本講演では、未知の確率密度関数が所与の点において不連続で（もしくはジャンプが）あると疑われる場合のジャンプ幅の推定および不連続性の検定を取り扱った。密度関数の不連続性に関する問題は、統計学では 90 年代初頭に既に研究がなされていた。一方、経済学においてこの方面の研究は回帰分断デザイン(regression discontinuity design)と関連してごく最近開始された。本講演では、回帰分断デザインの共変量に用いられる経済変数の多くが非負値をとる点を踏まえ、共変量の密度関数の台が非負実数全体である場合を想定して非対称ガンマ・カーネルを利用する推定・検定手法を提案した。

まず、ジャンプ幅（＝所与の切断点における密度関数の左側・右側極限の差）の推定法は以下の手順に従う：①ガンマ・カーネルを切断点で左右に分割する；②左右各部分をそれぞれ積分値が 1 となるよう規準化し、二つの正当な切断カーネル関数を作り上げる；③対応する切断カーネル関数を用いて左側・右側極限をノンパラメトリック推定し、その差をジャンプ幅推定量とする。ただし、この推定量の分散は通常の $O(n^{-1}b^{-1/2})$を維持する一方、片側のみの平滑化を用いる手法のため、バイアスは通常の $O(b)$より遅い $O(b^{1/2})$となってしまう。このような遅いバイアス収束は検定のサイズに悪影響を及ぼすことが懸念されるため、Terrel and Scott (1980)のバイアス修正法を応用してジャンプ幅推定量のバイアスを $O(b)$ に改善した。

次に、ジャンプ幅推定量が漸近的に正規分布に従うことはリアプノフ中心極限定理を用いて証明される。この結果から、ジャンプ幅推定量に関する t 値をジャンプ幅がゼロ（＝切断点において連続）の帰無仮説の下で検定統計量として使用してよいことがわかる。そこで、二種類の漸近分散推定値に基づく検定

統計量を提案した。なお、これら漸近理論の導出に不完全ガンマ関数の級数近似が積極的に利用されている点も本研究の新規性の一つである。

また、推定・検定を実行する上で重要な問題である平滑化パラメータの選択法も提案した。具体的には、サブサンプリング法に基づく検出力最適化法を本検定に適合させて用いた。加えて、所与の切断点で密度関数が連続であるとの帰無仮説が棄却された場合に密度関数全体を推定する方法にも言及した。特に、切断点の左右で切断カーネル関数を用いて密度推定を行うと、通常のガンマ・カーネルを用いた場合と同様の統計的特性が得られることを証明した。

さらに、モンテカルロ実験では、提案されたジャンプ幅推定量のバイアスがほぼゼロとなる点、および、検出力最適化法によって計算された平滑化パラメータの値を用いた検定統計量はサイズを損なうことなく検出力を改善でき、McCrary (2008)の検定統計量に比べて高い検出力を得られる点を確認できた。Angrist and Lavy (1999)の実データを用いた応用例では、McCrary (2008)の検定では密度関数が連続であるという帰無仮説が棄却できない分断点の幾つかで、提案する検定は仮説を棄却した。これは本検定の高い検出力の裏付けと見ることができる。

最後に、今後の本研究の発展・拡張に触れる。まず、現状切断点は唯一と想定しているが、複数の切断点を特定し各点におけるジャンプ幅ゼロの同時検定法を開発すべきであろう。さらに、切断点を特定せずに密度関数のジャンプの有無を検定する方法の開発にも取り組みたいと考えている。

参考文献：

Angrist, J. D., and V. Lavy (1999): “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement,” *Quarterly Journal of Economics*, 114, 533 - 575.

McCrary, J. (2008): “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test,” *Journal of Econometrics*, 142, 698 - 714.

Terrell, G. R., and D. W. Scott (1980): “On Improving Convergence Rates for Nonnegative Kernel Density Estimators,” *Annals of Statistics*, 8, 1160 - 1163.

円周上のカーネル密度推定量とそのバンド幅選択法の漸近的性質

金沢大学人間社会環境研究科　　鶴田靖人
金沢大学経済学経営学系　　寒河江雅彦

1　はじめに

周期性を持つデータは、円周上の角度として値を取る確率変数として扱うことができるので方向データと呼ばれる．このような方向データを分析することを目的とした統計学が方向統計学である．方向統計学では周期性を持つ確率変数 $\Theta \sim f(\theta)$ を扱う．ただし，$\theta \in [-\pi, \pi)$，$f(\theta) = f(\theta + 2\pi)$ とする．

本稿は、方向データのモデリングとしてよく用いられる巻き込みコーシー (WC) 分布をカーネル関数とした WC カーネル密度推定量の漸近的性質を与える．数値実験の結果から分布が集中しているという条件の下では WC カーネル密度推定量は良い性質を持つことが分かっている．

また，新しいモーメントを定義し，p 次オーダーカーネル関数のクラスを提案する．このカーネル関数のクラスは，Hall et al.(1987) が与えた 2 次オーダーカーネル関数のクラスを p 次オーダーカーネル関数に拡張したものとなっている．p 次オーダーカーネル密度推定量は平均二乗誤差 (MISE) の収束レートが $O(n^{-2p/(p+1)})$ で漸近正規性を持つ．このカーネル関数は，Jones and Foster (1993) の加法型構成法や Terrell and Scott (1980) の乗法型構成法を用いて，2 次オーダーカーネルから 4 次以上の高次オーダーカーネル密度推定量を構成できる．

方向統計学での平滑化パラメータ (集中度パラメータ) κ の選択法として， least square cross validation (CV) 法（Hall et al. , 1987) や Direct Plug-in (PI) 法 (Mardio et al. , 2011) などがある．κ の推定についての研究は，数値実験による比較実験が主であり，選択法の理論的な性質はほとんど議論されていない．我々は CV 法と PI 法の漸近的な性質を導出した．CV 法による κ の推定量 $\hat{\kappa}$ の漸近的な収束レートは $O(n^{-1/10})$ であるが，PI 法による $\hat{\kappa}$ の漸近的な収束レートは $O(n^{-5/14})$ となる．この結果は PI 法の方が CV 法よりも優れた選択法であることを示している．

2　先行研究

対称なカーネル関数を $K_\kappa(\theta)$ とする．ただし，$\kappa > 0$ は集中度パラメータ (バンド幅の逆数に対応する平滑化パラメータ) とする．標本 $\Theta_1, \ldots, \Theta_n$ は独立同一分布 $f(\theta)$ に従うとする．カーネル密度推定量は，

$$\hat{f}_\kappa(\theta) := \frac{1}{n}\sum_{i=1}^{n} K_\kappa(\theta - \Theta_i).$$

で与えられる．

Marzio et al. (2011) は sin モーメント $\eta_j(K_\kappa) := \int \sin(\theta)^j K_\kappa d\theta$ を用いて sin オーダーカーネルを提案した．p 次 sin オーダーカーネル関数 K_κ を，

$$\eta_0(K_\kappa) = 1, \quad \eta_j(K_\kappa) = 0,\ 0 < j < p, \quad \eta_p(K_\kappa) \neq 0,$$

と定義する．p 次 sin オーダーカーネル密度推定量の MISE は以下の式となる：

$$\mathrm{AMISE} = \frac{\eta_p^2(K_\kappa) R\left(f^{(p)}\right)}{(p!)^2} + \frac{1 + 2\sum_{j=1}^{\infty}\gamma_j^2(\kappa)}{2\pi n}, \tag{1}$$

ただし，$R(g) := \int \{g(\theta)\}^2 d\theta$，$\gamma_j(K_\kappa) := \mathrm{E}_K[\cos(j\theta)]$ ．式 (1) の第 1 項と第 2 項はバイアスの 2 乗と分散に対応している．

WC カーネルを以下のように定義する:

$$K_\rho(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta)}, \quad 0 < \rho < 1,$$

ただし，ρ は集中度を調節するパラメータとする．WC カーネルの特性関数は以下のように定義できる：

$$\phi_p = \rho^{|p|}.$$

WC カーネルの $\gamma_j(\rho)$は$\gamma_j(\rho) = \rho^j$ となるので，WC カーネルの AMISE は次式で与えられる：

$$\mathrm{AMISE}_{\mathrm{WC}}[\hat{f}_\rho(\cdot)] = \frac{\{1-\rho^2\}^2 R(f'')}{16} + \frac{1}{n\pi(1-\rho^2)}. \tag{2}$$

(2) について $1 - \rho^2 = h$ とおくと，次式が得られる：

$$\mathrm{AMISE}_{\mathrm{WC}}[\hat{f}_h(\cdot)] = \frac{h^2 R(f'')}{16} + \frac{1}{n\pi h}. \tag{3}$$

κ^* と同様にして (3) を最小にする最適な h^* は次式のようになる：

$$h^* = \left(\frac{8}{\pi R(f'')n}\right)^{1/3}, \quad n > 8(\pi R(f''))^{-1}. \tag{4}$$

(3), (4) より AMISE の収束レートは $O(n^{-2/3})$ となる．

WC カーネルの最適な収束レート $O(n^{-2/3})$ は同じ VM カーネルの収束レート $O(n^{-4/5})$ とは異なる． このことは，Sin 型モーメントは MISE の収束レートに対応したカーネル・モーメントとは言えないことを示唆している．また，Di Marzio et al. (2011) は， sin 型 p 次オーダーカーネルは必ずしもバイアスを修正できるわけではないと指摘した．つまり，sin 型 p 次オーダーカーネルの次数と MISE の収束レートは必ずしも対応しているわけではない．

3　高次オーダーカーネル密度推定量

本稿では $K_\kappa(\theta) := C_\kappa^{-1}(L)L(\kappa\{1-\cos(\theta)\})$ と定義する．モーメントを $\mu_l(L) := \int_0^\infty L(r)r^{(l-1)/2}dr$，ただし，$r = \kappa\{1-\cos(\theta)\}$ とする．

$K_\kappa(\theta)$ が p 次オーダーカーネル関数であるとは，正の偶数 l に対して，

$$\mu_0(L) \neq 0, \quad \mu_l(L) = 0, \quad 0 < l < p, \quad \mu_p(L) \neq 0,$$

を満たすことである．p 次オーダーカーネル密度推定量の MISE は以下の式で表せる：

$$\mathrm{AMISE} = \frac{\mu_p^2(L)R\left(\sum_{t=1}^{p/2} \frac{b_{p,2t} f^{(2t)}(\cdot)}{2t!}\right)}{\mu_0^2(L)\kappa^p} + \frac{d(L)\kappa^{1/2}}{n}, \tag{5}$$

ただし，$b_{p,2t}$，$d(L)$ は定数である．式 (5) の第 1 項と第 2 項はバイアスの 2 乗と分散に対応している．$\kappa = h^{-2}$ とおけば，(5) は実数直線上で定義されたカーネル密度推定量の MISE の形によく似ている．実際に，その収束レートは $O(n^{-2p/(p+1)})$ となる．

4　高次オーダーカーネルの構成法

$K_{\kappa,[p]}(\theta) := C_\kappa^{-1}(L_{[p]})L_{[p]}(\kappa\{1-\cos(\theta)\})$ は p 次オーダーカーネル関数を表す．

Jones and Foster (1993) に対応する加法型構成法は次の式となる：

$$L_{[p]}(r) := \frac{p+1}{p}L_{[p]}(r) + \frac{2}{p}rL'_{[p]}(r),$$

ただし，$r = \kappa\{1-cos(\theta)\}$，$L'_{[p]}(r) = dL_{[p]}(r)/dr$.

Terrell and Scott (1980) に対応する乗法型構成法は，2 次オーダーカーネル密度推定量 $\hat{f}$ の積で与えられる：

$$\hat{f}_\kappa^{[\mathrm{TS}]}(\theta) := \hat{f}_\kappa^{4/3}(\theta)\hat{f}_{\kappa/4}^{-1/3}(\theta),$$

ただし，$\mathrm{bias}_f^2[\hat{f}_\kappa^{[\mathrm{TS}]}(\theta)] = O(\kappa^{-4})$, $\mathrm{Var}_f[\hat{f}_\kappa^{[\mathrm{TS}]}(\theta)] = O(n^{-1}\kappa^{1/2})$ となる．

5　平滑化パラメータの選択法

2 次オーダーカーネル密度推定量の (5) を最小にする集中度パラメータ κ_* は以下の式で与えられる:

$$\kappa_* = \beta\psi_4^{2/5}n^{2/5}, \tag{6}$$

ただし β は定数，$\psi_r := \int_{-\pi}^{\pi} f^{(r)}f(\theta)d\theta$．本稿では κ_* の推定法として CV 法と PI 法を挙げている．これらの推定法の定義と漸近的性質について述べる．

CV 法とは，

$$\mathrm{CV}(\kappa) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\Theta_i),$$

を最小にする推定量 $\hat{\kappa}_{\mathrm{CV}}$ を求める手法である．$\hat{\kappa}_{\mathrm{CV}}$ の漸近的な性質を示す：

$$\hat{\kappa}_{\mathrm{CV}}/\kappa_* \xrightarrow{p} 1,$$

$$n^{1/10}(\hat{\kappa}_{\mathrm{CV}}/\kappa_* - 1) \xrightarrow{d} \mathrm{N}(0, \sigma_{\mathrm{CV}}^2). \tag{7}$$

(7) の収束レートは実数直線上の CV 法で推定したバンド幅推定量 $\hat{h}_{\mathrm{CV}}$ の収束レートの対応している．

PI 法とは，ψ_4 をカーネル密度推定法で推定し κ_* を推定する方法である．カーネル密度推定量 $\hat{\psi}_4(g)$ を定義する：

$$\hat{\psi}_4(g) := n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}T_g^{(r)}(\Theta_i - \Theta_j),$$

ただし，$T_g(\theta) = C_g^{-1}(S)S(g\{1-cos(\theta)\})$ は p 次オーダーカーネル関数，g は集中度パラメータ．

$\hat{\psi}_4(g)$ のバイアスを最小にする g を g_* と表す：

$$g_* := cn^{2/(p+5)},$$

ただし，c は定数．このとき，$\hat{\psi}_4(g)$ の平均二乗誤差 (MSE) は，

$$\inf_{g>0}\mathrm{MSE}[\hat{\psi}_4(g_*)] = \begin{cases} O(n^{-(2p+1)/(p+5)}) & p < 4, \\ O(n^{-1}) & p \geq 4, \end{cases}$$

となる．したがって，4 次以上の高次オーダーカーネルを用いたとき，$\hat{\psi}_4(g)$ はパラメトリックな収束レートである $O(n^{-1})$ を達成する．

また，PI 推定量は，

$$\hat{\kappa}_{\mathrm{PI}} := \beta\hat{\psi}_4(g_*)^{2/5}n^{2/5}, \tag{8}$$

で与えられる．$\hat{\kappa}_{\mathrm{PI}}$ は T_g が 2 次オーダーカーネル関数のとき，以下のような漸近正規性を持つ．

$$n^{5/14}(\hat{\kappa}_{\mathrm{PI}}/\kappa_* - 1) \xrightarrow{d} \mathrm{N}(0, \sigma_{\mathrm{PI}}^2). \tag{9}$$

(9) の収束レートは実数直線上の PI 法で推定したバンド幅推定量 $\hat{h}_{\mathrm{PI}}$ の収束レートに対応する．両者の収束レートを比較すると PI 法の方がスピードが速い．

参考文献

[1] Di Marzio, M., Panzera, A. and Taylor, C. C. (2011). *Journal of Statistical Planning and Inference* **141**, 2156-2173.

[2] Hall, P., Watson, G. S. and Cabrera, J. (1987). *Biometrika*, **74**, 751-762.

[3] Jones, M. C. and Foster, P.J. (1993). *Journal of Nonparametric Statistics* **3**, 81-94.

[4] Scott, D. W. and Terrell G. R. *Jounal of the American Statistical Association*, **82**, 1131-1146.

[5] Sheather S. J. and Jones M. C. *Jounal of the Royal Statistical Society. Series B*, **53**, 683-690.

[6] Terrell, G. R. and Scott, D. W. (1980). *The Annals of Statistics* **8**, 1160-1163.

角度データのための統計モデル

統計数理研究所　　加藤 昇吾

本報告では，加藤 (2016) のレビュー論文の内容の一部に基づき，角度データに関連した統計的話題を議論した．具体的には，(1) 角度データの解析における問題点，(2) 角度データのための統計手法の基礎，(3) 円周上のコーシー分布とその拡張，という 3 つのテーマに分けて報告を行った．

1. 角度データの解析における問題点

様々な学問分野において，角度として表される観測値が得られることがある．例えば，気象学における風向の観測はその一例である．風向は，西を $-\pi$ とし，反時計回りを正の向きとすれば，南を $-\pi/2$，東を 0，北を $\pi/2$ のように角度で表すことができる．つまり，任意の風向は $-\pi$ 以上 π 未満の角度 θ，もしくは円周上の点 $(\cos\theta, \sin\theta)$，として表現できる．他には，犯罪が起こる時刻（0 時から 24 時）を記録したデータも同様に，角度の観測と解釈することができる．その他，医学・地震学・動物行動学などの分野においても角度の観測が存在している．

角度のデータには，統計解析をする上で大きな問題がある．それは，このようなデータの解析する上では，統計学が主に対象としている実数値データのための解析手法をそのまま使うことができないという問題である．この問題は，角度には周期性があり，その位相が実数の位相と異なっていることに起因している．例えば，平均や分散などの要約統計量，実数値データのための確率分布や回帰モデルを角度のデータにそのまま応用すると，不自然な結果を得ることになってしまう．

2. 角度データのための統計手法の基礎

前節で述べた問題点を克服するため，統計学では角度データの解析法に関する研究がなされてきた．本報告においては，角度データのための統計手法の基礎として，角度データのための要約統計量および確率分布について紹介を行った．

要約統計量については，mean direction と mean resultant length を紹介した．これら 2 つの要約統計量は次のように定義される．確率変数 $\Theta_1, \Theta_2, \ldots, \Theta_n$ を $[-\pi, \pi)$ に値をとる角度の標本とする．このとき，この標本の mean direction と mean resultant length は，

$$\overline{\Theta} = \arg\left(\sum_{j=1}^{n} e^{i\Theta_j}\right), \quad \overline{R} = \frac{1}{n}\left|\sum_{j=1}^{n} e^{i\Theta_j}\right|,$$

とそれぞれ定義される．ここで，i は虚数を表す．直感的に言えば，mean direction は角度データの平均に相当し，mean resultant length は角度データのばらつきをあらわす要約統計量である．このように定義をすると，それぞれの要約統計量の値が自然に解釈をすることが可能となる．

確率分布については，フォン・ミーゼス分布と wrapped 分布を紹介した．フォン・ミーゼス分布は，確率密度関数

$$f(\theta) = \frac{1}{2\pi\mathcal{I}_0(\kappa)} \exp\left\{\kappa\cos(\theta-\mu)\right\}, \quad -\pi \le \theta < \pi; \quad -\pi \le \mu < \pi,\ \kappa \ge 0,$$

で定義される円周上の確率分布である．ここで，$\mathcal{I}_j$ は位数 j の第一種変形ベッセル関数をあらわす．フォン・ミーゼス分布は，‘円周上の正規分布’ とも呼ばれ，角度データの統計解析において中心的な役割を果たしてきた．実際，フォン・ミーゼス分布には，指数型分布族に属する，最尤推定量が陽な形で表現できる，などの利点がある．一方，正規化定数が特殊関数を含む形となる，よく知られた ‘尺度’ の変換に閉じていない，再生性が成り立たない，などの正規分布が持つ扱いやすい性質が成立しない問題点も知られている．

次に wrapped 分布の紹介を行った．X を実数に値をとる確率変数とする．また，確率変数 Θ を以下のように定義する：

$$X \equiv \Theta \pmod{2\pi}.$$

このとき，Θ の従う分布を，分布 X の wrapped（巻き込み）分布という．wrapped 分布は，実数上の分布から直接的に得ることが可能であり，分布の単峰性を保つなどの利点があるが，一般には確率密度関数が無限級数の形であらわされる等の問題点もある．

3. 円周上のコーシー分布とその拡張

本報告の最後のテーマとして，円周上のコーシー分布とその 2 変量拡張の紹介を行った．円周上のコーシー分布は，確率密度関数

$$f(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}, \quad -\pi \le \theta < \pi; \quad -\pi \le \mu < \pi,\ 0 \le \rho < 1,$$

で定義される．円周上のコーシー分布は，再生性が成り立つこと，メビウス変換に関して閉じていること，無限分解可能であること，などの扱いやすい性質を持つことが知られている（例えば，McCullagh, 1996）．一方，最尤推定に関しては，$n \le 4$ のときは最尤推定量の陽な表現が知られているものの，一般には Kent & Tyler (1988) のアルゴリズムなどを用いて数値的に求める必要がある．

最後に，円周上のコーシー分布を 2 変量分布へと拡張した Kato & Pewsey (2015) の分布を紹介した．この分布の確率密度関数は，特殊関数や無限級数などを含まない陽な形で表現することが可能である．他にも，分布のパラメータの解釈が容易である，周辺分布と条件付分布がともに円周上のコーシー分布となる，効率の良い疑似乱数の発生法が存在する，特性関数や相関係数を簡潔な形で表現できる，などの多くの扱いやすい性質を持っている．パラメータの推定に関しては，モーメント推定に関しては推定量を陽にあらわすことが可能である．最尤推定については一般に数値的に求めることが必要であるが，モーメント推定値を初期値として使用すれば，推定アルゴリズムが早く収束することが経験的に知られている．

参考文献

[1] 加藤昇吾 (2016). 「円周上のコーシー分布と関連した統計モデル」. 『日本統計学会誌』, **46**, 85–111.

[2] Kato, S. & Pewsey, A. (2015). A Möbius transformation-induced distribution on the torus. *Biometrika*, **102**, 359–370.

[3] Kent, J.T. & Tyler, D.E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, **15**, 247–254.

[4] McCullagh, P. (1996). Möbius transformation and Cauchy parameter estimation. *The Annals of Statistics*, **24**, 787–808.

The Simultaneous Multivariate Hawkes-type Point Processes and their application to Financial Markets

Naoto Kunitomo (Meiji University)

joint work with

Daisuke Kurisu, **Yusuke Amano**, and **Nao Awaya** (University of Tokyo)

In economic and financial time series we sometimes observe large jumps. Although they are relatively rare events, they often have significant influence not only on a financial market but also several different markets and macro economies. By using the simultaneous Hawkes-type models we introduce, which are multivariate point processes, it is possible to analyze the causal effects of large events in the sense of the Granger-non-causality (GNC) and the instantaneous Granger-non-causality (IGNC). We investigate the financial market of Tokyo and other markets, and apply the Granger non-causality tests. We have found several important empirical findings among financial markets and macro economies.

We divide the continuous observation period $[0, T]$ to the discrete observation periods $I_i^n = (t_{i-1}^n, t_i^n]$ $(i = 1, \cdots, n)$. The initial time is $t_0^n = 0$ and we interpret I_i^n as the $i-$th day, but it may be possible to have more finer observations. Let the observable price process be $P_j(t)$ $(j = 1, \cdots, d\ ; t_{i-1}^n < t \leq t_i^n, i = 1, \cdots, n)$ and in $s \in I_i$ we consider the (negative) log-returns of prices $Y_j^n(s)$ $(t_{i-1}^n < s \leq t_i^n)$ be $Y_j^n(s) = -\log[P_j(s)/P_j(t_{i-1}^n)]$ $(j = 1, \cdots, d; i = 1, \cdots, n)$. Let the first stopping time when $Y_j^n(s)$ exceeds the threshold u_j in $s \in I_i$ be $\tau_j^n(i, 1)$. Define $X_j^n(s) = Y_j^n(s)$ for $s \in t_{i-1}^n \leq s \leq \tau_j^n(i, 1)$ and $X_j^n(s) = X_j^n(\tau_j^n(i, 1))$ for $s \in [\tau_j^n(i, 1), t_i^n]$. We define the simple counting processes $N_j^{n*}(s, u_k)$ by the number of stopping times that $X_j^n(s)$ exceed u_j $(j = 1, \cdots, d)$ for a particular j but not for other $k \neq j$ by the time s. (For the simplicity, we can assume that the jumps of the counting process $N_j^{n*}(s, u_k)$ can occur once at each t_i^n of the end of each intervals $(t_{i-1}^n, t_i^n]$.) Then we notice that as $n \longrightarrow \infty$ the interval length goes to zero, that is, $\max_{i=1,\cdots,n} |t_i^n - t_{i-1}^n| \longrightarrow 0$ and the simple counting process $N_j^{n*}(s, u_k)$ converges to $N_j^*(s, u_k)$. The resulting counting process can be interpreted as the limiting process in the high frequency asymptotics.

The simple point processes $N_j^*(t)$ $(j = 1, \cdots, d)$ satisfies the standard conditions that as $\Delta t \to 0$ $P(N_j^{n*}(t+\Delta t, u_j) - N_j^{n*}(t, u) = 1|\mathcal{F}_t^n) = \lambda_j^n(t, u_j)\Delta t + o_p(\Delta t)$, $P(N_j^{n*}(t+\Delta t, u_j) - N_j^{n*}(t, u) > 1|\mathcal{F}_t^n) = o_p(\Delta t)$, $P(N_k^{n*}(t+\Delta t, u_j) - N_j^{n*}(t, u_j) \geq 1|\mathcal{F}_t^n) = o_p(\Delta t)$ for $k \neq j$, where $\mathcal{F}_t^n$ is the $\sigma-$field generated by the information at t. Also we define the simple proint processes $N_{jk}^{n*}(s, u_{jk})$ by the number of stopping times that $X_j^n(s)$ exceed u_j $(j = 1, \cdots, d)$ for a particular j and $X_k^n(s)$ exceed u_k $(k = 1, \cdots, d; k \neq j)$ for a particular k and other $X_l^n(s)$ $(l \neq j, k)$ do not exceed u_l by the time s in I_i^n. Then we introduce the point processes $N_{jk}^{n*}(t, u_{jk})$ with co-jumps of N_j and N_k by $P(N_j^{n*}(t + \Delta t, u_j) - N_j^{n*}(t, u_j) = N_k^{n*}(t + \Delta t, u_k) - N_k^{n*}(t, u_k) = 1|\mathcal{F}_t) = \lambda_{jk}^{n*}(t, u_{jk})\Delta t + o_p(\Delta t)$, $P(N_j^{n*}(t + \Delta t, u_j) - N_j^{n*}(t, u_j) > 1|\mathcal{F}_t) = o_p(\Delta t)$, $P(N_k^{n*}(t + \Delta t, u_k) - N_j^{n*}(t, u_k) > |\mathcal{F}_t) = o_p(\Delta t)$ for $k \neq j$.

When we have co-jumps of two point processes, we define the point process

$$N_j^n(s,u_j) = N_j^{n*}(s,u_j) + N_{j,k}^{n*}(s,u_{jk}) \ (j \neq k; j,k = 1,\cdots,d) \tag{1}$$

and the corresponding conditional intensity functions are given by

$$\lambda_j^n(t,u_j) = \lambda_j^{n*}(t,u_j) + \lambda_{j,k}^{n*}(t,u_{jk}) \ . \tag{2}$$

Then the resulting processes correspond to as the marginal point processes of the vector point process. We use the self-exciting form of intensity function as

$$\lambda_j^{n*}(t,u) = \lambda_{j0}^* + \int_{-\infty}^t \sum_{i=1}^p c_{ji}(X^n(s-))g_{ji}(t-s)dN_{J_i}^{n*}(s,u) \ , \tag{3}$$

where the index set is defined as $J_i = i \ (i = 1,\cdots,d)$, $J_i = 1, 1+(i-d) \ (i = d+1,\cdots,2d-1),\cdots, \ J_p = 1,\cdots,d$. Sequentially define $N_i^{n*}(s,u_i) = N_i^{n*}(s,u) \ (i = 1,\cdots,d)$, $N_{d+1}^{n*}(s,u) = N_{1,2}^{n*}(s,u),\cdots,$ and $N_p^{n*}(s,u) = N_{1,\cdots,d}^{n*}(s,u)$ with the intensities $\lambda_{j,k}^{n*}(t,u_{jk})$. In this formulation we use the damping function as $g_{ji}(t-s) = e^{-\gamma_{ji}(t-s)}$ and the impact function as $C(X) = (c_{ji}(x))$. Since we are interested in large jumps, it is important to use the probability function in the tails. We use the generalized Pareto (tail) probability function for $x > u_j \ (j = 1,\cdots,d)$ as

$$\begin{aligned} P(X_j^n(s) > x | X_j^n(s) > u_j, \mathcal{F}_s) &= \frac{\left[1+\frac{\xi_j}{\sigma_j(s)}x\right]^{-1/\xi_j}}{\left[1+\frac{\xi_j}{\sigma_j(s)}u_j\right]^{-1/\xi_j}} \\ &= \left[1+\frac{\xi_j}{\sigma_j^*(s)}(x-u_j)\right]^{-1/\xi_j} \ , \end{aligned} \tag{4}$$

we set $\sigma_j^*(s) = \xi_j u_j + \sigma_j(s)$.

References

Florens, J-P. and D. Fougere (1996) "Noncausality in Continuous Time," *Econometrica*, 64, 1195-1212.

Grothe, O., V. Korniichuk and H. Mannera (2014)"Modeling multivariate extreme events using self-exciting point processes," *Journal of Econometrics*, 182, 269-289.

Hawkes, A. G. (1971) "Point Spectra of Some Mutually Exciting Point Processes," *Journal of the Royal Statistical Society. Series B*, 33-3, 438-443.

Kunitomo,N., A. Ehara and D. Kurisu (2016) "A Causality Analysis of Financial Markets by Multivariate Hawkes-type Models,"(in Japanese), CIRJE, Discussion Paper J-278, CIRJE, University of Tokyo.

ウェーブレット分散と参照形式を利用した商品先物の暴落予想

上智大学大学院理工学研究科数学領域・院　本山　要
上智大学理工学部情報理工学科　加藤　剛

1　研究目的と背景

本研究は, ウェーブレット解析の中でも特にウェーブレット分散と呼ばれる概念を活用し, 東京商品取引所 (TOCOM) における原油先物価格について, 暴落の前兆現象をとらえることを目的としたものである. 最終的には, 暴落特有の特徴をもった参照図を作成し, 新たなデータとの参照形式で暴落の注意喚起や警報に使用することを目標としている.

金融市場や金融商品売買に関する報道が絶えず行われている. 2008 年のリーマンショック, その後の中国の不動産バブル, ギリシャ危機,Google の時価総額が世界首位獲得, 次期アメリカ大統領が決定し金融市場に影響を及ぼす等々. 世界のあらゆる出来事が金融市場に影響を与え複雑化している. また, 日本国内に注目すれば,2001 年から確定拠出年金制度の導入,NISA(少額投資非課税制度), あるいは仮想通貨ビットコインの出現になどより, 個人が資産運用や投資に関わる機会が増えている現実がある. 先物取引の適切なリスク指標をつくることは, 機関投資家だけでなく個人投資家にも市場参加を助長し, 市場の活性化につながる. しかし,2009 年 5 月に相場の急激な変動がおきた際の緊急処置として新しくサーキットブレーカー制度が採用されたこともあり, 理論的裏付けがあるリスク指標に関する研究は少ない. なお, 石油を選択した理由は, 価格の下落が頻発しているためである.

2　参照形式による暴落予想

参照形式とは, ある特徴を有する状況の情報を参照図として保存しておき, 将来, 参照図と似ている情報をもった状況が現れた場合, 同じような状況が起こる可能性があると予測する方法である. 本研究は, この方法を東京商品取引所 (TOCOM) が扱う原油先物価格の暴落予想に応用したものである. なお, 値動きの状況を表す情報として様々なパラメータが存在するが, ウェーブレット分散の有効性を確かめるため, ウェーブレット分散の時間的推移のみを値動きの状況を表す情報としている.

3　データ解析

3.1　参照図の作成

本研究では, 暴落を原油におけるサーキットブレーカー発動条件 2000 円の 75% である 1500 円を上回る下落がある場合の値とし, 時間の間隔とあわせて次のように定義する.

定義 1 暴落時刻

次の不等式を満たす X_t が存在するとき, 時刻 t を暴落時刻とする.

$$X_{t-(2^{16}+2^{14})} - X_t > 1500$$

幅を 2^{16}, すらしを 2^{14}, 着目スケールを 2^{10} とした. 前処理した 2014 年 9 月 1 日 0 時 0 分 0 秒から 2015 年 12 月 30 日 23 時 59 分 59 秒までの原油の約定値段のデータ数は 20, 768, 150 であったので, 1263 個のウェーブレット分散が得られた. 得られたウェーブレット分散の個数と同じ 1263 回暴落時刻を調べた. その結果, 1263 回中 111 回が暴落時刻に該当することがわかった.

ウェーブレット分散から 4 点のパターンを構成するイメージを図 1 に示す. 図 1 で, 垂線が下りている時刻を暴落が起きたか否かを調べた時刻 T とする. 時刻 T に対して, 1〜4 の範囲でそれぞれ計算された 4 つのウェーブ

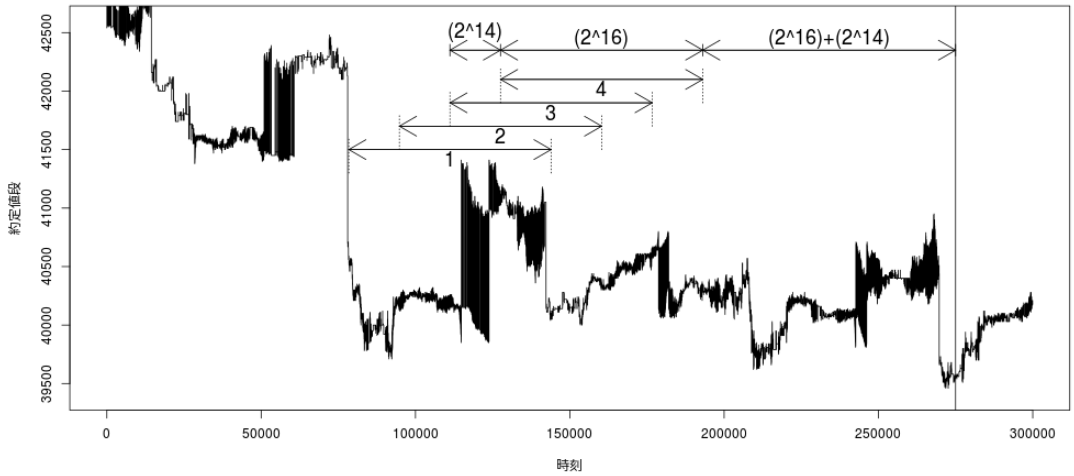

図 1　暴落があるか調べた時刻とそれに対応するウェーブレット分散の対応図

レット分散の値のパターンが対応する. これを, ずらし 2^{14} でずらしていくことにより, 1260 個のパターンが得られる.

新たなデータから計算したウェーブレット分散にもとづくパターンと類似したものが参照図に存在するとき, 暴落に対する注意喚起や警報を行う. パターンの類似度は距離で測る.

3.2　参照図の作成

以上の議論を踏まえて, 参照図の作成方法を説明する.

1. 使用するデータの期間を決定
2. パターンを構成するウェーブレット分散の個数を決定 (探索的に 4 個から 10 個までとした)
3. 1 つの参照図を構成するパターン数を決定 (4 個から 6 個までとした)
4. パターンを作成する方法を決定

(a) quantile based　　(b) cluster based

3.3　結論

次の 2 点で作成した参照図を評価した.

- **信頼性**　参照図がもつ, いずれかのパターンに近いパターンが見つかった時に, 実際に暴落が起こる割合
- **性能**　参照図がもつパターンが, 全体の暴落の回数のうち, 近いパターンとして検出できた割合

結論として, 信頼性と性能の両方の値が十分に高いといえる参照図はみつからなかった. しかし, 信頼性は高く性能が低い参照図, 逆に, 信頼性は低いが性能が高い参照図, さらに細かく見れば, 信頼性が高いパターンを発見できた. したがって, これらを組み合わせることで, より実用的な参照図を作成できる可能性がある.

主な参考文献

[1] Nason, G. P., "Wavelet Methods in Statistics with R", *Springer*, (2008).
[2] 東京商品取引所ホームページ, http://www.tocom.or.jp/jp/

リスク計測モデルのバックテスト法について

塚原 英敦（成城大学 経済学部）

1　はじめに — 問題の背景

リスク計測モデリングとは，ファイナンス・データに対する統計モデルを設定し，パラメータ推定を行い，リスク尺度の推定値を得るという一連の作業を指す．そして，(1) 現在用いている統計モデル・推定法のパフォーマンスをモニターする．(2) 統計モデル・推定法の相対的な比較を行うというのが，一般に言われているバックテストの 2 つの目的である（McNeil et al. [7]）．前者の側面を重視して述べれば，バックテストはモデルと推定法の妥当性をチェックする過程で用いられる手法であり，金融リスク管理を適切に行うためには不可欠である．バリュー・アット・リスク（VaR）の場合，よく用いられているバックテスト法は，いわゆる VaR 超過数に基づくものであり，直観的にも容易に理解できる方法である（Campbell [2]）．

Consultative Document of Basel Committee on Banking Supervision (October 2013) では，2008・9 年の金融危機における VaR の欠陥を指摘した上で，規制資本量決定のためのリスク尺度として，VaR に代わって期待ショートフォールを採用することを提案している．しかし，同じ文書において，バックテスト法としては VaR に基づく方法の使用を求めている．その理由としては，多くの人々が VaR は期待ショートフォールや他のリスク尺度よりもバックテストするのが簡単であると主張していることが挙げられる．その根拠として，期待ショートフォールに対する既存のバックテストは，帰無分布についてのパラメトリックな仮定に基づいていること，そして検定統計量の帰無分布に対する漸近的な近似が必要であることが指摘されている．この議論によれば，リスク尺度のバックテスト可能性（backtestability）とは，モデルによらない（model-free）やり方で，小標本に基づいてバックテストができることと解釈されるが，VaR の場合はこれが満たされる非常に稀な例であり，他のリスク尺度についてこの条件を課すのは無理な注文であると言わざるを得ない．

2　顕在化可能性

近年，バックテスト可能性の数学的な表現として，顕在化可能性（elicitability）という概念が注目を集めている（Gneiting [6]）．大雑把に言えば，統計的汎関数 $T(F)$ が分布族 $\mathscr{F}$ に対して顕在化可能（elicitable）であるとは，ある評点関数 S が存在して，すべての $F \in \mathscr{F}$ に対して $t = T(F)$ が $t \mapsto \mathrm{E}^F[S(t, L)]$ を最小化するただ 1 つの値であることである．ここで，L は分布 F に従う確率変数である．VaR が顕在化可能であることは容易にわかるが，期待ショートフォールやそれを含むクラスである歪みリスク尺度（distortion risk measures）はこの顕在化可能性の条件を満たさないことが最近示された（Ziegel [8]）．さらに，整合性かつ法則不変性をもつリスク尺度で顕在化可能なものは，expectile しかないことも証明された（Bellini and Bignozzi [1]）．ただし，expectile は直観的な理解に欠けるという短所もあり，実用には程遠いであろう．

顕在化可能性はバックテストの目的 (2) のためにはある程度有効である一方，モデルのモニタリングという目的 (1) との関連性は明らかではない．例えば，平均は顕在化可能であるが，平方誤差に基づいて，モデルに依存しない形でバックテストすることは，漸近近似を用いるかパラメトリックな分布を仮定しない限り不可能である．

3 Davis の較正性

観測される損失と補助変数の系列 $(L_n, \boldsymbol{X}_n)_{n\in\mathbb{N}}$ に対して，モデル $\mathscr{P} = \{P^z : z \in \mathscr{Z}\}$ は $(\mathbb{R}^{d+1})^\infty$ 上の確率分布族であり，それに対して条件付分布列 $F_1^z(l) := P^z(L_1 \le l)$, $F_n^z(l) := P^z(L_n \le l \,|\, \mathscr{F}_{n-1})$, $n = 2, 3, \ldots$ が定義される．このとき，統計的汎関数 T に対して，較正関数 $\phi : \mathbb{R}^2 \to \mathbb{R}$ と，狭義単調増加な可予測正規化列 $b = (b_n)$ で，$\lim_{n\to\infty} b_n = \infty$, P^z-a.s., $\forall z \in \mathscr{Z}$ を満たすものが存在し，

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} \phi(L_k, T(F_k^z)) = 0, \quad P^z\text{-a.s., } \forall z \in \mathscr{Z}$$

が成り立つとき，T は $\mathscr{P}$ で (ϕ, b) **較正可能**と呼ばれる（Davis [3]）．例えば，VaR_α^k に対しては，$\phi(l, t) = \mathbf{1}_{[t,\infty)}(l) - \alpha$, $b_n = n$ ととればよいことが大数の法則から明らかである．

実用上は，観測値 $(L_1, \boldsymbol{X}_1), \ldots, (L_{k-1}, \boldsymbol{X}_{k-1})$ が得られ，$T(F_k^z)$ の予測値 τ_k が何らかのアルゴリズムで計算されたとき，この予測値の “妥当性”を $J_n(L_1, \ldots, L_n, \tau_1, \ldots, \tau_n) = b_n^{-1} \sum_{k=1}^n \phi(L_k, \tau_k)$ が十分 0 に近いかどうかで判断できる．これは，Dawid [4, 5] による逐次予測（prequential）原理「予測評価の妥当性基準は，損失の実現値と予測の実際値のみに依存すべきである（予測値がどう計算されたかは無関係）」を満たしている．

Davis は非常に広い条件の下で VaR が較正可能なのに対して，期待ショートフォールのような平均型の汎関数については，較正可能のためにはかなり厳しい条件が必要であることを示した．しかし，較正可能性もまた漸近的な概念であり，バックテスト可能性を表現する十分に満足できる答であるとは言えないであろう．

以上の議論から，バックテスト可能性という概念の数学化は非常に困難な，あるいは不可能な問題のように見受けられる．

参考文献

[1] F. Bellini and V. Bignozzi. On elicitable risk measures. *Quantitative Finance*, 15:725–733, 2015.

[2] S. D. Campbell. A review of backtesting and backtesting procedures. *Journal of Risk*, 9:1–17, 2007.

[3] M. H. A. Davis. Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33:67–93, 2016.

[4] A. P. Dawid. Present position and potential developments: some personal views. statistical theory. the prequential approach (with discussion). *Journal of the Royal Statistical Society (A)*, 147:278–292, 1984.

[5] A. P. Dawid. Prequential analysis. In C. B. Read S. Kotz and D. L. Banks, editors, *Encyclopedia of Statistical Sciences, Update Volume 1*, pages 464–470. Wiley-Interscience, 1997.

[6] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

[7] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, Princeton, New Jersey, second edition, 2015.

[8] J. F. Ziegel. Coherence and elicitability. *Mathematical Finance*, 26:901–918, 2016.

交互作用を含む観察データにおける傾向スコア解析手法による結果の差異

加葉田 大志朗[1,2], 山本 紘司[1], 新谷 歩[2]

1) 大阪大学 大学院医学系研究科 臨床統計疫学寄附講座
2) 大阪市立大学　大学院医学研究科　医療統計学講座

1．はじめに

観察研究により得られたデータの解析では、交絡等のバイアスへの対処が必須となる。特に近年医学分野においては、交絡の除去を目的として、傾向スコアを利用した解析手法が普及している。傾向スコアは治療に割り当てられる確率を背景情報などから計算したものである。現在提案されている主流な傾向スコア解析としては、類似した傾向スコアで対照群を選ぶマッチング法、傾向スコアの逆数を用いて重み付けするIPTW(Inverse Probability of Treatment Weighted)法、IPTW法と多変量回帰モデルを組み合わせたDR(Doubly Robustness)法がある。特にマッチング法は解釈がしやすく、臨床で利用されることも多い。しかしデータによっては各手法も使い分けが必要となるが、その詳細は知られていない。

2. ケーススタディ

本研究では、舌癌患者における予防的リンパ郭清の効果を検証する目的で得られた実際の観察研究データの解析より着想を得た。この解析では、傾向スコアによるマッチング法と、3因子交互作用項を含む多変量回帰モデルを用いて解析を行った。その結果傾向スコアでは治療が有効であることが示されたが、多変量回帰モデルにおいては治療効果と年齢・腫瘍サイズの間に交互作用が確認された。つまり治療効果は術前の年齢と腫瘍サイズにより変化し、中年層(40～60歳)で腫瘍サイズが5mm以上の症例において治療は有効であることが示唆された。一方、高齢あるいは若年層で5mm以上の症例においては、予後を悪化させる可能性があることが示唆された。

この解析では感度解析として、IPTW法とDR法を用いた解析も行ったが、どちらもマッチング法のように治療効果があるという結果は得られなかった。

3. 検証内容

上記の観察研究データ解析におけるマッチング法は、傾向スコアが最も近い症例同士をマッチングするSimple nearest neighbor matchingを用いた。当該手法による解析では基本的にATT(Average Treatment effect for Treated)、主に治療が行われた症例における治療効果を推定することが知られている[1)]。今回のマッチング法を用いた解析では、中間年齢層で腫瘍サイズが大きい、つまり臨床的に最も治療をうける確率が

高い症例における平均的な治療効果が反映されたと考えることができる。

IPTW法とDR法はマッチング法とは異なり解析に用いる症例数が減少しないため、最近では臨床論文でも用いられることが増えてきた。しかし基本的にIPTW法とDR法はATE(Average Treatment Effect)すなわち、集団全体の平均的な効果を検証している。そのため当事例のように治療効果が対象症例の背景によって変化する状況においては、集団全体の平均的治療効果を推定すると、治療を受けた症例における治療効果が十分に検出されないことが考えられる。またIPTW法とDR法においても重みを調整することで、ATTと同様の解析結果を得られることが提唱されているが[2)]、実際にそれらを用いた臨床論文は目にすることが少ない。

傾向スコアを用いた手法比較論文においてはATEに着目した報告が行われることが多く、当事例のように、治療効果が交互作用を持つデータにおける解析手法については言及されることが少ない。しかし医学領域において治療効果が患者背景により変化することは多くのケースで考えられ、その影響を考慮して解析手法を選択することは重要であると考える。

本研究では当事例のように治療効果との交互作用を含むデータにおいて、各種傾向スコア解析手法を適用した際の結果の違いを検証し、目的に応じた適切な解析手法を提案することを目的とし、当事例に類似した擬似データを発生させ、モンテカルロシミュレーションにより各種傾向スコア解析手法を適用した結果の比較を行った内容を報告した。

1) Stuart, E. A., Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science : A Review Journal of the Institute of Mathematical Statistics 25(1): 1–21, 2010.

2) Hirano, K., and Imbens, G. W., Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services and Outcomes Research Methodology 2(3–4): 259–278, 2001.

診断研究における検査性能の比較

高橋　佳苗　（大阪大学医学部附属病院）
山本　紘司　（大阪大学大学院医学系研究科）

１．はじめに

医療において、スクリーニング検査は疾患の早期発見・早期治療のために重要な役割を果たしている。スクリーニング検査の性能をあらわす指標としては、感度（sensitivity：SE）、特異度（specificity：SP）、陽性適中率（positive predictive value：PPV）および陰性的中率（negative predictive value：NPV）がしばしば用いられている。SE は罹患している対象者をスクリーニング検査によって正しく罹患していると判断する確率であり、SP は罹患していない対象者を正しく罹患していないと判断する確率である。一方、PPV は対象者のスクリーニング検査の結果が陽性であるとき、当該対象者が実際に疾患に罹患している確率であり、NPV はスクリーニング検査の結果が陰性であった対象者が実際に疾患に罹患していない確率である。SE、SP、PPV、NPV は臨床的に有用であり、患者の治療方針の決定に影響を及ぼす。

2 つのスクリーニング検査法の性能を比べる目的で上記の指標を比較する際、SE および SP については McNemar 検定が一般的に用いられている。しかし、PPV および NPV については主流となる解析手法は確立しておらず、いくつかの手法が提案されているのみである(Leisenring et.al, 2010, Moskowitz and Pepe, 2006, Kosinski, 2013)。さらに、提案されているいずれの手法も、中心極限定理が用いられており、対象者の少ない小規模臨床試験への適用に問題がないかは未だ検討されていない。

さらに、スクリーニング検査の性能については、本来 1 つの指標のみで判断するよりも、複数の指標から総合的に判断されるべきである。2 つのスクリーニング検査間で複数の指標を比較する際は、すべての指標の優越性を確認するよりも、1 つの指標の優越性と他の全ての指標での非劣性を組み合わせたデザインの方が実臨床の場に即していると考えられる。

そこで本講演では、①PPV および NPV の比較に対する解析手法を小規模臨床試験へ適用した場合の性能評価、②いずれか 1 つの指標の優越性が認められれば他の指標は非劣性を確認するのみでよいとする、優越性と非劣性を組み合わせた多重エンドポイントの解析手法(Bloch et al, 2007, Muscha and Turan, 2012) の性能評価について報告した。

２．方法および結果

①PPV および NPV の解析手法を小規模臨床試験へ適用した場合の性能評価

多項分布をもとに中心極限定理とデルタ法を利用したもの（恒等変換、対数変換、ロジット変換)、回帰分析の枠組みを利用したもの（一般化スコア統計量、重み付き一般化スコア

統計量）の解析手法がすでに提案されている。Kosinski (2013)で報告されている方法を用いて、表 1 の設定でシミュレーション用データを生成し、症例数を 10 例から 50 例とした場合の各手法の Type 1 error rate 及び検出力を求めた。

2 つの診断検査間の PPV、NPV に差がないとしたとき(図 1)、名目有意水準に最も近い検出力となったのは重み付き一般化スコア統計量を用いた方法であった。重み付き一般化スコア統計量を用いた方法は、2 つの診断検査間の PPV、NPV に差があるとしたとき(図 2)の検出力も高く、小規模臨床試験においては最も性能の良い方法と考えられる。

<u>②優越性と非劣性を組み合わせた多重エンドポイントの解析手法の性能評価</u>

表 2 に示すシナリオについて、①と同様に Kosinski (2013)の方法を用いてシミュレーション用データを作成した。有病割合は 0.4、非劣性マージンは 0.1、名目有意水準は 0.05、症例数は 50 例から 500 例とした。SE および SP の解析には McNemar 検定を、PPV および NPV の解析には重み付き一般化スコア統計量を用いた方法を適用した。

Bloch et al. (2007)の方法と Muscha and Turan (2012)の方法間で結果に大きな違いは見られなかった。また、優越性と非劣性の両方を評価するデザインの検出力が優越性の評価のみと比較して大きく変わらない場合もあることから、いずれか 1 つの指標の優越性が認められれば他の指標は非劣性を確認するのみでよいとする、優越性と非劣性を組み合わせた多重エンドポイントの枠組みでデザイン・解析することの有用性が示唆された。

参考文献

- Bloch, D. A. et al. A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Statistics in medicine* 2013, **26**:1193-1207.
- Kosinski, A. S. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in medicine* 2013, **32**:964-977.
- Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 2000; **56**:345-351.
- Mascha, E. J. and Turan, A. Joint hypothesis testing and gatekeeping procedures. *Anesthesia and Analgesia* 2012, **114**:1304-1317.
- Moskowitz, C. S. and Pepe M. S. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials* 2006; **3**:272-279.

複数の二値評価変数をもつ臨床試験における検定法

石原　拓磨　(大阪大学大学院・医学系研究科)
山本　紘司　(大阪大学大学院・医学系研究科)

臨床試験の検証的段階では，薬剤や治療処置による治療効果を評価するために複数の評価変数を主要評価変数として用いる場合がある．本発表では特に複数の評価変数のうち，少なくとも 1 つの変数について優越性の根拠が得られ，その他すべての変数に対しては非劣性の根拠が得られるときにのみ治療効果があると判断するような試験を考える．ここでは簡単のため 2 群比較を想定し，2 つの群 $(i = 1, 2)$ の p 個の評価変数に対して，次のような複合帰無仮説 H_0 を考える．

$$H_0 : \left\{ \max_{1 \le j \le p} \mu_j \le 0 \right\} \cup \left\{ \min_{1 \le j \le p} (\mu_j + \epsilon_j) \le 0 \right\} \tag{1}$$

ここで μ_j は，第 i 群における j 番目の評価変数の処置効果を $\mu_{ij} (i = 1, 2; j = 1, ..., p)$ としたときに $\mu_j = \mu_{1j} - \mu_{2j}$ で表される処置効果の群間差であり，$\epsilon_j (j = 1, ..., p)$ は $\epsilon_j \ge 0$ の非劣性マージンである．

このとき Bloch et al.（2001），Perlman and Wu（2004），Nakazuru et al. (2014) は上記の仮説（1）に対する検定手順を提案した．

これらの研究では，処置効果の差は連続変数として議論されている．そこで本発表では，特に評価変数が二値変数である場合において，第 i 群の j 番目の評価変数に対する標本比率を $\hat{\pi}_{ij} (i = 1, 2; j = 1, ..., p)$，第 j 番目の評価変数の標本比率の群間差を $\Delta_j = \hat{\pi}_{1j} - \hat{\pi}_{2j}$ としたとき，次の場合に帰無仮説（1）を棄却する検定手順を次のように提案した（有意水準 α）:

$$\chi^2 \equiv \boldsymbol{\Delta}^t \hat{\boldsymbol{\Sigma}^t}^{-1} \boldsymbol{\Delta} > \chi^2_\alpha \quad \text{and}$$
$$\frac{\Delta_j + \epsilon_j}{\sqrt{\frac{\hat{\pi}_{1j}(1-\hat{\pi}_{1j})}{n_1} + \frac{\hat{\pi}_{2j}(1-\hat{\pi}_{2j})}{n_2}}} > Z_\alpha \quad \text{for} \quad j = 1, ..., p.$$

ただし，$\boldsymbol{\Delta}^t = (\Delta_1, ..., \Delta_p)^t$ であり，χ^2_α 及び Z_α はそれぞれ χ^2 分布及び標準正規分布の上側 α パーセント点である．

提案した検定手順の性能についてはモンテカルロシミュレーションを用いて検出力と Type I error rate を評価した．シミュレーションや実際のデータ解析結果を報告した．

参考文献

Nakazuru, Y., Sozu, T., Hamada, C. and Yoshimura, I. (2014). A new procedure of one-sided test in clinical trials with multiple endpoints. *Japanese Journal of Biometrics,* **35**, 17-35.

Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics,* **40**, 549-567.

Perlman, M.D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics,* **60**, 276-280.

Bloch, D.A., Lai, T.L. and Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics,* **57**, 1039-1047.

Bloch, D.A., Lai, T.L., Su, Z. and Tubert-Bitter, P. (2007). A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Statistics in Medicine,* **26**, 1193-1207.

R. A. Fisher 以後の判別分析の新理論と遺伝子解析の新手法

成蹊大学　経済学部　特任教授　新村秀一

Fisher は「Fisher の仮説」を考えることで、計算機の助けを借りないで判別理論を確立した。しかし、Fisher の仮説を検定する良い統計量がないこと、誤分類確率や判別係数の標準誤差が定式化できなかったこと、などの重要な事実に多くの研究者や利用者は注意を払わなかった。Fisher も指摘する通り、「Fisher の仮説を満たさないデータに Fisher の LDF を適用してはいけない」が、それが分からないため多くの分野に適用され、一見成果を上げてきた。しかし大学卒業後の 1971 年に「心電図の自動解析システムの診断論理を LDF と QDF で 4 年間アプローチして、医師の開発した枝分かれ論理に歯が立たなかった」。この経験から、医学診断や、各種の格付けや、試験の合否判定を得点合計で判別すると、線形分離可能（MNM=0）であっても誤分類確率が最高 20%になることが実証研究で分かった。即ち既存の判別理論には、世界中の誰も研究していない 4 つの深刻な瑕疵があることを判別分析の多くの実証研究で見つけ、数理計画法による 5 個の最適判別関数を開発し解決してきた。また、判別分析は推測統計手法でないので、「小標本のための 100 重交差検証法（新手法 1）」を開発し、 誤分類確率と判別係数の 95%信頼区間を求めることができた。また、検証標本の平均誤分類確率（M2）が最小モデルを Best モデルとして選ぶと、Vapnik が提起した汎化能力に優れたモデルを選ぶことになる。各手法の全ての変数の組み合わせモデルで Best モデルを選び、8 種の異なった LDF（改定 IP-OLDF、改定 LP-OLDF、改定 IPLP-OLDF、H-SVM、SVM4、SVM1、ロジスティック回帰、Fisher の LDF）の Best モデルの中で最小の M2 を比較して、多くの場合で改定 IP－OLDF の Best モデルの M2 が最小であった。Vapnik の定義した汎化能力は、単に固定された p 変数の LDF での話であることが理解されていない。また、新手法 1 は、LOO 法よりはるかに優れている。以上で 4 種の深刻な問題を全て解決したが、2015 年の富山での科研費シンポジュームで「世界的に著名な Microarray データが公開されているのを知り、それらを判別すると、世界中の統計や医学研究者が高次元空間の分析と称し、特に Feature Selection の研究を 15 年以上行ってきたが(問題 5)が、改訂 IP-OLDF で解決した（Matroska Feature Selection Method、新手法 2)。そして Microarray データは、MNM=0 になる小さな部分空間の排他的な和集合になっていることが分かった」。即ち問題 5 の認識と同時に、僅か 41 日間で新手法 2 を開発した。これによって、新手法 1 と新手法 2 で Fisher 以後の新しい判別理論を確立した（New Theory of Discriminant Analysis After R. Fisher, Springer(2017)）。

参加者と以下の点について真摯に議論したい.

1) Fisher は Fisher の仮説に基づいて判別分析理論を確立したが，限界を知っていた．田辺の指摘[48]の他，検証にアヤメのデータを用いている点，仮説を満たさない場合に QDF が提案されていることが重要である．
2) Fisher 以後，RDA や LASSO 等の分散共分散に基づく理論が開発され，これらが Fisher の後継者と考えられているが間違いである．Fisher 理論を適用してはいけない医学分野で，Cox は Cox 回帰やロジスティック回帰を提案したが，彼こそが Fisher の正当な後継者である．そして Vapnik は MP で判別分析にアプローチした。統計や OR の分野を避け、多くの実証研究をパターン認識の分野で汎化能力や LSD 判別やカーネル SVM を広めた第三世代の後継者である．
3) 新村は，既存の判別理論には 4 つの問題があることを実証研究で見つけ，それらを解決した．その応用研究として，統計や医学の多くの研究者が従来の統計手法で 10 年から 15 年以上に渡り高次元空間の分析を試みたが成果は得られなかった．しかし 3 種類の OLDF は高次元空間を少数の遺伝子空間に自然に縮約し，Matroska 構造をもつこと，Datasets は少数の排他的な SM の和集合であることを示した．すなわち，これらの各 SM は統計的に小標本であり，通常の統計手法で分析可能である．
4) LASSO 等のアプローチは無意味である．通常のスイス銀行や日本車データで LSD であることが分からず、Feature Selection をできないのに高次元でできると考えるのは論理的ではない．

産業連関表に基づく
都道府県クラスターと産業構造推移の可視化

山本けい子（函館高専）・寒河江雅彦（金沢大学）

概要：産業連関表データの新たな活用法として，地域間の比較や経済構造の推移に着目した分析を行う．平成 17 年版都道府県単位の産業連関表（取引基本表の内生部門）データに基づいて，都道府県間の経済構造のクラスター分析を行った．また，平成 17 年版と平成 23 年版の同一都道府県における産業連関表データの主成分分析を通して，産業構造の推移の可視化を試みた．これらの結果から，産業連関表を用いた地域間の比較や経年変化の抽出方法について議論する．

1　はじめに

産業連関表は，国内経済において一定期間（通常 1 年間）に行われた財・サービスの産業間取引を一つの行列に示した統計表で，全国版，地域版，都道府県版など，5 年ごとに作成されており，主に，地域内の生産波及効果の計算や経済構造の分析に用いられている．我々は，恣意的な選択なく作成過程も統一的な産業連関表データを，地域間で比較することにより産業構造の類似性を抽出できるのではないかと考え，地域クラスタリングを試みた．本稿では地域として都道府県単位を設定し，地域間の類似度として，行列ノルムを用いてクラスター分析した結果について議論する．また，同一地域における産業構造の推移を抽出するため，平成 17 年版と平成 23 年版でのデータをそれぞれ主成分分析し，主成分の比較や可視化を通して手法の有効性を検証する．

2　産業連関表

産業連関表は，経済構造を総体的に明らかにし，経済波及効果分析 や経済指標の基準改定を行うための基礎資料を提供することを目的として，総務省を中心に各省庁で 5 年ごとに作成される統計表である．ある産業に新たな需要が発生した場合にどのように生産が波及するか計算することができることから，さまざまな経済分析に用いられている．一定期間において，財・サービスが各産業部門間でどのように生産され，販売されたかについて，行列の形で一覧表にとりまとめられており，取引基本表，投入係数表，逆行列係数表などから成り立っている．

本稿では，都道府県単位で作成されている産業連関表のうち平成 17 年版の大分類から取引基本表の中間投入・中間需要を表す行列（内生部門データ）および平成 23 年版の大分類から内生部門データを用いた分析を行う．

3　分析方法

3.1　データ

平成 17 年版産業連関表を各都道府県の Web サイトからダウンロードし，取引基本表の大分類の内生部門データを抽出した．ただし，都道府県によって，部門数が若干異なっており，部門を分割・結合して，全国データに記載の 34 部門に統一した．また，平成 23 年版については，大分類が 37 部門へと拡大されているため，分析に使用した都道府県については，平成 17 年版と同じ 34 部門となるようデータを結合した．これらの操作により，都道府県ごとに 34 部門 × 34 部門の行列データを作成した．

3.2　方法

3.2.1　都道府県クラスターの可視化

3.1 で都道府県ごとに作成した 34 × 34 の行列データに基づいて，都道府県間の類似性を算出し，都道府県の経済構造のクラスター分析を行う．クラスター分析は，各データを 1 つのクラスター（まとまり）としてスタートし，クラスター間の距離（類似度）に基づいて，2 つのクラスターを逐次的に結合していく階層的クラスタリングを採用し，樹形図によって可視化した．距離の指標には行列の要素ごとのノルム

$$d\left(\mathbf{X}^{(A)}, \mathbf{X}^{(B)}\right) = \sqrt{\sum_{i=1}^{34}\sum_{j=1}^{34}\left(\mathbf{X}_{ij}^{(A)} - \mathbf{X}_{ij}^{(B)}\right)^2}$$

を用いた．ただし，都道府県 A の行列データを $\mathbf{X}^{(A)}$，都道府県 B の行列データを $\mathbf{X}^{(B)}$ とする．ノルムの他に距離の指標として，主成分や非負値行列因子分解などの利用も考えられる．なお，実装には，R 言語を使用した．

3.2.2　経済構造推移の可視化

同一都道府県における平成 17 年版データと平成 23 年版データを比較することにより，経済構造の推移を可視化する．各年の特徴を的確に捉えながら推移の概観を

可視化するため，34 × 34 の行列データを主成分分析によって次元縮約し，得られた主成分を可視化することによって推移の読み取りを試みる．主成分は，データのもつ情報を多く保持するものから第 1 主成分，第 2 主成分，・・・として得られることから，産業構造の特徴を第 1 主成分と第 2 主成分として抽出し，これらを各年で図示することによって推移を可視化した．

4　結果

4.1　都道府県クラスターの可視化

取引基本表のデータは，各都道府県内で取引される金額であることから，都道府県によって金額の単位が大きく異なる．この金額の違いに着目する場合は，34×34 の行列データそのものを使用したクラスター分析を行う．一方，金額よりもその産業の取引構造に関心がある場合は，34 × 34 の行列データを標準化したものを使用してクラスター分析を行う．標準化しないデータと標準化したデータを用いた分析結果を図 1 に示す．

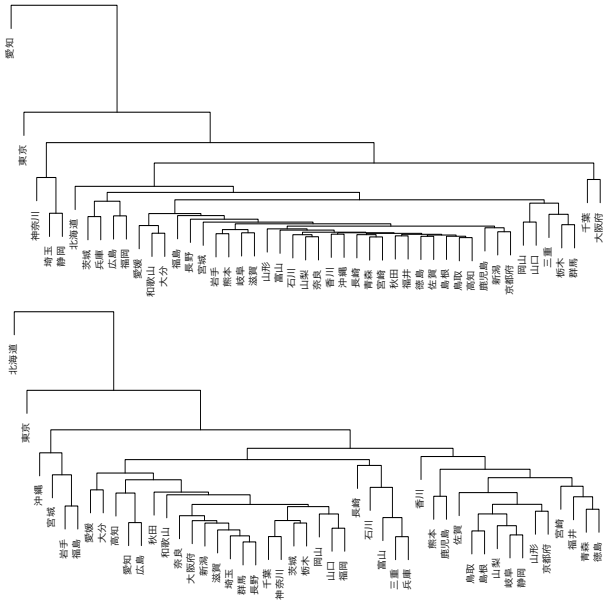

図 1　都道府県のクラスター分析 (上/下：標準化なし/あり)

図 1 の標準化しないデータを用いた都道府県クラスターの樹形図では，愛知が特出しており，次いで東京，神奈川／埼玉・静岡のクラスターとその他の都道府県でクラスターが形成されていることがわかる．また，標準化したデータでは，北海道と東京がそれぞれクラスターを形成していることがわかる．

4.2　経済構造推移の可視化

平成 17 年と平成 23 年のデータをそれぞれ主成分分析分析し，産業（需要側）の 34 部門についての第 1 主成分と第 2 主成分を図示して比較する．図 2 に愛知県と東京都の例を示す．軸ラベルのカッコ内は寄与率で，元のデータの情報をどの程度表しているかの目安となる数値である．

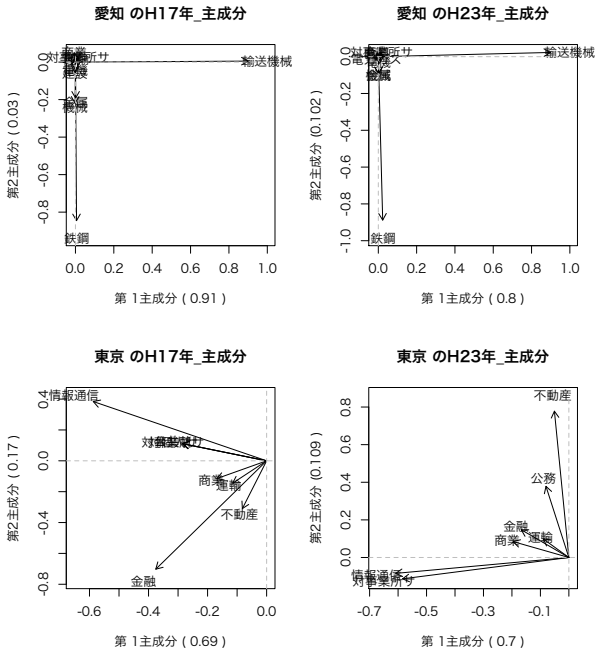

図 2　愛知県と東京都の産業構造の推移

図 2 より，愛知県では，第 1 主成分の寄与率が高く，「輸送機器」と「鉄鋼」の 2 大産業を有し，平成 17 年と平成 23 年では，産業構造にほとんど変化がみられていない．一方，東京都の場合は，第 1 主成分と第 2 主成分を合わせて 80% の寄与率であり，平成 17 年では「情報通信」と「金融」，「サービス」系に特徴的な傾向がみられたが，平成 23 年では，「サービス」系が「情報通信」と同程度まで伸びており，さらに「不動産」も一つの大きな特徴として現れていることがわかる．

5　考察

本稿では，産業連関表の新たな活用方法として，地域クラスタリングと産業構造の推移の可視化法について提案・検証した．通常，経済データの分析には専門的な知識が必須となるが，ここで扱った手法は，クラスター分析や主成分分析と可視化を組み合わせたデータマイニング手法であるため，経済に関する事前知識は不要である．ただし，結果の解釈については，専門家の協力が想定される．また，産業が 34 部門という多く複雑な状況を次元縮小法によって特徴抽出し可視化することは，容易な解釈と直感的なイメージが得られるという利点もある．

今回使用した産業連関表（取引基本表の内生部門）は各都道府県内のみで取引される金額のため，今後は，都道府県内外との移出・移入も含めた経済構造の分析へ展開が期待できる．

テキストマイニングにもとづくレビューのスコアリングを用いた映画の統計的分類

上智大学大学院理工学研究科数学領域・院　山縣　一慶

1　研究目的と背景

今日の日本は, 武士道に由来する武道や茶道などの伝統的様式のみならず, 映画や漫画, アニメなど現代文化が世界的に注目されており, それらは「クール・ジャパン」として総称される. しかし, 統計手法を用いた現代文化の分析は少なく, 分析の需要は今後高まっていくと考えられる. 本研究では, 数ある” クールジャパン” の 1 つである映画に注目する. 中でも, 定義の明確でない映画のジャンル分けに焦点を当て, テキストマイニングを用いた新たな映画のジャンル分けを行う.

テキストマイニングは, 単語や文節で区切り, 各々の出現件数や傾向を解析し, 有用な情報を取り出すことが目的である. 日本語で書かれたテキストを分析する際, 語の最小単位である形態素に分解する必要があり, 一連の工程を形態素解析という. 本研究では形態素解析ソフトの 1 つである TTM [5] を使用した. TTM は同義語を登録（以下, 辞書作り）することで, 同義語の出現件数を分析することも可能である. 下の表は, 辞書作りの一例である. 辞書 1 では, 我々が”SF” という語から連想するであろう語を登録した. 対して辞書 2 は, 同義語辞書（シソーラス）の Weblio [7] を用いて”SF” の同義語を登録した. また, 映画「スター・ウォーズ　フォースの覚醒」に関するレビュー 493 件に対して作成した辞書 2 つと TTM を用い, レビュー中で”SF” の語が使われている件数をレビュー点数ごとにまとめ, 比較した. 対象レビューは Yhoo!映画 [6] から Ruby 言語で作成したプログラムを用いて抽出している.

辞書番号	単語	同義語
1	SF	宇宙, ロケット, 船, 星, 彗星, 惑星, 月, NASA
2	SF	エスエフ, サイエンス, フィクション, 空想科学小説, 近未来小説

辞書 1 を用いた抽出件数		
レビュー点数	抽出語	出現件数
3	SF	23
4	SF	24
5	SF	18

辞書 2 を用いた抽出件数		
レビュー点数	抽出語	出現件数
3	SF	1
4	SF	4
5	SF	5

辞書内容の違いが語の出現件数に影響を及ぼすことがわかる.

従来の辞書作りでは, 同義語辞書（シソーラス）を用いることが多い. 短時間で同義語を見つけることができるが, 分類対象の内容に合った同義語を抽出できないという欠点を持つ. 同義語抽出の先行研究としては, 藤村 [2] が挙げられる. 藤村の研究では, ノート PC に関するレビューをもとに, 肯定的な意味を持つ語と否定的な意味を持つ語を分別するためのスコアリングを提案している. また, Anne [1] は ”personalization” と”customization” という曖昧な定義を持つ 2 語が分類可能か否か研究している. Anne は 両語について書かれたテキスト 883 件と 1544 件を対象に, テキスト内でそれぞれの語と同時に出現する語（以下, 共起語）の出現件数を調べた. そして, テキスト内である語と同時に出現する共起語の出現件数の相関から 2 語が分類可能であることを述べている. 本研究では, 映画に関するテキストデータから共起語の出現件数を測定し, クラスター分析を用いて, 互いに関連性の強い語で構成された語群を得た. これらの語群を辞書登録し, 映画分類に用いる辞書 (共起語辞書) を作成した. 映画に関するレビューから共起語辞書で語の抽出を行い, それぞれの特徴をスコアリングすることで新たな映画分類を

得るのが本研究の目的である.

2　共起語の出現件数を用いたスコアリング手法

本研究では, IMDB [3] の視聴ランキングに掲載され, レビュー数が充実している映画 32 本を分類対象とした. 対象映画のレビューは,Yhoo!映画から, Ruby 言語を用いて作成したプログラムで抽出した. また, 映画レビューが, 映画内容についての文章とレビュー著者の感想について書かれた文章で構成されていると考え, それぞれについて言及される際に使われる単語群をジャンル属性, 感情属性とした. 提案するスコアリングは, 国土交通省 [4] で提案された評価アンケートに対する重みづけの手法に, 対象データの語数による重みを加えたものである.

以下, 共起語辞書を用いたときのスコアリングについて説明する. 映画番号 i について書かれた, 評価点数 j のレビューにおける, ジャンル属性に含まれる単語の総出現件数を e_{ij} とし, 感情属性に含まれる単語の総出現件数を c_{ij} とする. また, ジャンル属性, 感情属性に含まれる特徴的な語の番号を k とする. この時, 映画番号 i について書かれた評価点数 j のレビューにおける, ジャンル属性に属する属性番号 k の語の出現件数を c_{ijk}, 感情属性に属する属性番号 k の語の出現件数を e_{ijk} とする. 以上を用いて, スコアリングの定義を行った. 映画番号 i について書かれた評価点数 j のレビューにおけるジャンル属性と感情属性の重みを, それぞれ,

$$wc_{ij} = 100 * \frac{c_{ij}}{e_{ij} + c_{ij}} \qquad we_{ij} = 100 * \frac{e_{ij}}{e_{ij} + c_{ij}}$$

とする. また, 映画番号 i における点数 j のレビュー総単語数を s_{ij} とする. このとき, 映画番号 i における, ジャンル属性に所属する k 番目の特性が持つスコア C_{ik} と, 感情属性に所属する k 番目の特性が持つスコア E_{ij} を

$$C_{ik} = \sum_{j=1}^{5} \left(wc_{ij} \cdot \frac{c_{ijk}}{\sum_{k=1}^{22} c_{ijk}} \cdot \frac{s_{ij}}{\sum_{j=1}^{5} s_{ij}} \right) \qquad E_{ik} = \sum_{j=1}^{5} \left(we_{ij} \cdot \frac{e_{ijk}}{\sum_{k=1}^{8} c_{ijk}} \cdot \frac{s_{ij}}{\sum_{j=1}^{5} s_{ij}} \right)$$

とする. これらのスコアリングを用いて 32 本の映画全てに対し各属性番号ごとのスコアをつけた.

3　スコアリング結果の一例

下の表はジャンル属性のスコアリングの例である. 比較することでタイトル特有の強い特性の存在が確認できる.

	ジャンル属性											
属性番号	1	2	3	4	5	6	7	8	9	10	11	12
特性	アクション	冒険	犯罪	ドキュメンタリ	ホラー	ミュージカル	ミステリー	ロマンス	SF	短編	スポーツ	西部劇
シックス・センス	3.703532	0.142865	0.226495	3.062464	3.111151	0	1.509333	0.307835	0.335978	0	0.225144	0
シンドラーのリスト	4.39665	0.841639	0.691965	4.738463	0.285452	0.521444	0.498867	0.877908	0.783925	0	0.408401	0.079285
スター・ウォーズ	3.245434	0.722894	0.324432	3.091519	0.379477	0.225578	0.10671	0.403934	0.634732	0	0.599412	0.042307
ダークナイト	3.3562	0.434343	1.080887	1.996075	2.615922	0.132659	0.168272	0.868955	0.713987	0	0.419804	0.017806
ローマの休日	3.673402	0.704474	0.256732	4.485298	0.492521	0.931809	0.074177	1.429511	0.045543	0	1.221595	0.091086

参考文献

[1] Anne, S. and Johanna, B., "Applying text-mining to personalization and customization", Expert Systems with Applications, 39, 10049 - 10058 (2012).

[2] 藤村 滋, 豊田 正史, 喜連川 優, Web からの評判および評価表現抽出に関する-考察", 社会法人情報処理学会研究報告, (2004).

[3] IMDB, http://www.imdb.com/chart/top/

[4] 国土交通省 公共事業評価システム研究会, " 公共事業評価の基本的考え方 (公共事業評価システム研究会報告) について", 国土交通省, (2002).

[5] TinyTextMiner β version , http://mtmr.jp/ttm/

[6] Yahoo!映画, http://movies.yahoo.co.jp/

[7] Weblio 類語辞典, http://thesaurus.weblio.jp/

規模の異なる変量群をもつロジスティック回帰モデルの係数 2 段階推定

上智大学大学院理工学研究科数学領域・院　竹下　佳宏

1　研究の背景

21 世紀に DNA の全塩基配列情報を把握することが可能となり，ゲノムワイド関連解析（以下，GWAS）が行われるようになった [1]．GWAS では一塩基多型（以下，SNP）と呼ばれる数十万単位ある特定の塩基配列情報に着目して，生物の性質をひもとく．遺伝情報の解析手法では，古典的な統計解析が親しまれているが，機械学習の技術が応用されるなど，最近では新しい技術も活用されている [2]．本研究では，遺伝子情報の解析に用いられるロジスティック回帰モデルの新たな係数推定方法を提案する．

2　従来の研究手法

ある i 番目の個体に対し，病気である $(y_i = 1)$ かそうではない $(y_i = 0)$ かという 2 値の応答変数を，遺伝子の情報や，その個体の環境情報を考慮して説明するモデルとして，ロジスティック回帰モデルがある．一般には病気の有無以外の 2 値の応答変数にもこのモデルは用いられるが，本研究では病気の有無についてのみ言及する．

m_e 個の環境要因 $\{E_{i,k}\}_{k=1}^{m_e}$ と m_s 個の遺伝要因である SNP 情報 $\{SNP_{i,j}\}_{j=1}^{m_s}$ のうち j 番目の SNP のみを用いて，個体 i が病気にかかる確率 $p_{i,j}$ を次のロジスティック回帰モデルで表現する [2].

$$\log\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \alpha_{0,j} + \sum_{k=1}^{m_e} \alpha_{k,j} E_{i,k} + \beta_{snp,j} SNP_{i,j} \qquad (1)$$

SNP 情報は数十万単位あるため，一度に SNP をすべて解析するのではなく，1 つの SNP 情報に対してその都度モデルを構築し、その中から病気を十分に説明できるようなモデルを考えることで，病気に関連を持つ SNP を見つける．実際はモデル式 (1) の回帰係数に推定値を代入して $p_{i,j}$ について解き，その値があるしきい値を上回れば病気 $(y = 1)$，下回れば健康 $(y = 0)$ と予測する．

従来の研究手法においては，以下の点で問題があった．

- SNP ごとにすべての回帰係数を推定し直すが，本来は環境要因は SNP から独立しているため，環境要因の回帰係数が j 番目の SNP に依存していることが不自然である
- SNP の数だけ環境要因の回帰係数を推定するのは計算効率が悪い
- 1 つのモデルにつき 1 つの SNP しか考慮しないため，複数の SNP の交互作用が説明できない

これらの問題を解決する回帰係数の推定を本研究で行った．

3　回帰係数の 2 段階推定

3.1　推定方法の概要

ロジスティック回帰モデルは，応答変数が 0 か 1 かの 2 値の他に，応答変数が生起または反応の確率である場合にも利用できる [3]．現実には生起の確率そのものが応答変数として観測されることはほとんどないため，標本全体に対する生起数の割合で応答変数の値を推定する．

この考え方を利用して，2 段階で回帰係数の推定を行う．

1. 標本の個体 $i \in \{1, 2, \ldots, n\}$ について，i と同じ環境下で罹患している標本数を求め，標本全体に対する割合で罹患確率 $\widehat{p}_{e,i}$ を推定
2. 次のような環境要因だけのロジスティック回帰モデルを考え，係数 $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{m_e})$ を推定
$$\log\left(\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right) = \alpha_0 + \sum_{k=1}^{m_e} \alpha_k E_{i,k} \qquad (2)$$
3. 回帰係数の推定値 $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_k)_{0 \le k \le m_e}$ を用いて，個体 i の罹患確率の推定値 $\widetilde{p}_{e,i}$ を次の式で計算
$$\widetilde{p}_{e,i} = \frac{\exp(\widehat{\boldsymbol{\alpha}} \boldsymbol{E}_i')}{1 + \exp(\widehat{\boldsymbol{\alpha}} \boldsymbol{E}_i')}$$
ここで，$\boldsymbol{E}_i = (1, E_{i,1}, \ldots, E_{i,m_e})$
4. $\widehat{p}_{e,i}$ と $\widetilde{p}_{e,i}$ の差，すなわち，環境要因だけでは説明できない罹患確率の部分を SNP 情報を用

いて説明．回帰係数の問題としては，次のモデル式で回帰係数 $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{m_s})$ を推定

$$\log\left(\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right) - \log\left(\frac{\widetilde{p}_{e,i}}{1-\widetilde{p}_{e,i}}\right) = \beta_0 + \sum_{j=1}^{m_s} \beta_j SNP_{i,j} \tag{3}$$

2 段階に分けることにより，環境要因の回帰係数は SNP の影響を受けずに推定できる．そのため, (2) の回帰係数は (1) 式とは異なり, 添字 j が消去される．

3.2 環境要因の回帰係数の推定

3.2.1 応答変数の補正

現実のデータでは，$\widehat{p}_{e,i} = 0, 1$ もあり得る．その場合，(2) と (3) が定義できない．そこで，十分小さい $\delta > 0$ を用いて，$\widehat{p}_{e,i}$ を次のように補正する．

$$\widehat{p}_{e,i}(\delta) = \begin{cases} \delta & : \ \widehat{p}_{e,i} = 0 \\ \widehat{p}_{e,i} & : \ 0 < \widehat{p}_{e,i} < 1 \\ 1-\delta & : \ \widehat{p}_{e,i} = 1 \end{cases} \tag{4}$$

$\widehat{p}_{e,i}$ を $\widehat{p}_{e,i}(\delta)$ で置き換えて，3.2.2 節以降で説明する回帰係数の推定を行う．

3.2.2 最尤法による推定

環境要因の回帰係数 $\boldsymbol{\alpha}$ は，$\{\widehat{p}_{e,i}(\delta)\}_{i=1}^n$ を応答変数の値として用いた最尤法で推定する．実際には最尤解は陽な形で得ることができないので，ニュートン法によって数値的に求める．

$\{\widehat{p}_{e,i}(\delta)\}$ にもとづく最尤解 $\widehat{\boldsymbol{\alpha}}_\delta$ は，補正量 $\delta > 0$ の影響を受けており，本来の $\{\widehat{p}_{e,i}\}_{i=1}^n$ にもとづく解 $\widehat{\boldsymbol{\alpha}}$ とはずれがある．けれども，次の命題から，$\widehat{\boldsymbol{\alpha}}_\delta$ が $\widehat{\boldsymbol{\alpha}}$ の近似解となっていることがわかる．

命題 $\widehat{\boldsymbol{\alpha}}$ を補正がない場合 ($\delta = 0$) の $\{\widehat{p}_{e,i}\}_{i=1}^n$ による最尤推定量とするとき，

$$\lim_{\delta \to +0} \mathbb{E}\left\{(\widehat{\boldsymbol{\alpha}}_\delta - \widehat{\boldsymbol{\alpha}})^2\right\} = 0$$

よく知られているように，最尤推定量 $\widehat{\boldsymbol{\alpha}}$ は漸近的に正規分布に従う．命題の結果から，十分小さな δ について，$\widehat{\boldsymbol{\alpha}}_\delta$ も近似的に正規分布に従う．よって，$\widehat{\boldsymbol{\alpha}}_\delta$ について正規性にもとづく有意性の検定ができる．

3.3 遺伝要因の回帰係数の推定

3.1 節の (3) 式にもとづいて，SNP の回帰係数 $\boldsymbol{\beta}$ を推定する．回帰モデルとしては，重回帰モデルにしたがって最小自乗解を得る．以下，環境要因で説明し残した部分を次のように定める．

$$\boldsymbol{r} = (r_1, \ldots, r_n)'$$
$$r_i = \log\left\{\frac{\widehat{p}_{e,i}(\delta)}{1-\widehat{p}_{e,i}(\delta)}\right\} - \log\left\{\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right\}$$

また，SNP 情報を $n \times (m_s + 1)$ 行列 $\boldsymbol{S}$ として表すとき，最小自乗解は次のように表される．

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{r} \tag{5}$$

ただし，行列 $\boldsymbol{S}$ の 1 列目は切片項に対応するため，成分が 1 だけの列とする．

ところが，一般に SNP の情報はサンプル数 n よりも SNP 数 m_s の方が多いため，$\boldsymbol{S}$ にはランク落ちが生じる．そこで，(5) ではなく，$\boldsymbol{\beta}$ は $\boldsymbol{S}$ のムーア・ペンローズ型一般逆行列 $\boldsymbol{S}^+$ を用いた次の形で推定する．

$$\widehat{\boldsymbol{\beta}}_+ = \boldsymbol{S}^+\boldsymbol{r}$$

推定量 $\hat{\boldsymbol{\beta}}_+$ は次の性質を持つ [4].

1. $\widehat{\boldsymbol{\beta}}_+ = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{r} - \boldsymbol{S}\boldsymbol{\beta}\|$
2. $\hat{\boldsymbol{\beta}}_+$ は不偏推定量
3. 任意の不偏推定量 $\tilde{\boldsymbol{\beta}}$ に対して，

$$V\left(\widehat{\boldsymbol{\beta}}_+\right) \leq V\left(\widetilde{\boldsymbol{\beta}}\right)$$

ここで，$\boldsymbol{V}(\cdot)$ は分散共分散行列で，不等号は 2 次形式に関する大小関係を表す．

4 数値実験

上記の推定方法を実際のデータに適用し，従来の推定方法との比較を行った．数値実験の結果もシンポジウム当日に発表した．

参考文献

[1] 鎌谷直之．遺伝統計学入門, 岩波書店, 2007.

[2] 上辻茂男．”ゲノムワイド関連研究に学ぶ遺伝統計学”, 計算機統計学，25 巻，第 1 号, 17-39 (2012).

[3] 中村永友．多次元データ解析法 (R で学ぶデータサイエンス 2), 共立出版, 2010.

[4] 柳井晴夫，竹内啓．射影行列・一般逆行列・特異値分解, 東京大学出版, 1983.

国勢調査の匿名化ミクロデータの作成可能性について―地域区分に着目して―*

中央大学経済学部　伊藤 伸介**
(独)統計センター　星野 なおみ
総務省統計局　　阿久津 文香

1. はじめに

わが国では、平成 25 年 12 月と平成 26 年 3 月にそれぞれ、平成 12 年と 17 年の国勢調査の匿名データの提供が開始されている。一方、わが国の国勢調査の提供済匿名データでは、都道府県及び人口 50 万以上の市区が最小の地域区分となっているが、より詳細な地域区分を用いた分析が可能なミクロデータに対するニーズが存在すると思われることから、国勢調査の匿名データに関しては、より詳細な地域区分の提供可能性が検討されてよいと考える。そこで、本稿では、秘匿性と有用性の両面から、詳細な地域区分の作成可能性を検討する。

2. 「地域の人口規模の閾値」に基づく秘匿性の検証

匿名化ミクロデータ(個票データに匿名化処理が施されたデータ)の作成・提供においては、秘匿性に関する閾値を設定した上で、その閾値を超えない形で、様々な匿名化技法が適用されることが考えられる。例えば、Hawala(2001)は、アメリカ人口センサスの Public Use Microdata Sample の作成において用いられる 10 万人という地域区分の閾値に関して、その秘匿性に関する事後検証を行うために、母集団一意(population unique)の比率を用いて地域の人口規模と秘匿性の指標との関連性を明らかにしている。本研究では、平成 22 年国勢調査の A 県の調査票情報(個票データ)を用いて、地域の閾値を変更した場合のキー変数におけるリコーディング(区分統合)の可能性を探った。

最初に、秘匿性に関する第 1 の研究として、伊藤・星野(2014)に基づき、リコーディングを行ったデータに対して、①住宅の建て方、②住居の種類、③性別、④配偶者の有無、⑤国籍、⑥労働力状態、⑦従業上の地位、⑧年齢、⑨産業と⑩職業の 10 個のキー変数をもとに母集団一意(population unique)の比率を計測した。具体的には、10 変数の中で住宅の建て方等の 7 変数については、提供済匿名データ(現在提供されている国勢調査の匿名データ)の区分が用いられているが、残りの 3 変数である年齢、産業と職業については、様々なリコーディングおよびトップコーディングを施した上で、全 81 パターンにおける母集団一意の比率の計測を行った。なお、リコーディングにおいては、いわゆる「0.5%基準(単変量において母集団の 0.5%を下回る区分を統合すること)」を考慮した区分も設定されている。

本研究では、地域の人口規模と属性の分類区分との関連性を実証的に明らかにするために、A 県を対象にし、地域の閾値については、(1)人口 20 万人以上地域(県庁所在市に該当する地域も含む)、(2)人口 10 万人以上地域、(3)人口 5 万人以上地域、(4)人口 3 万人以上地域、(5)人口 2 万人以上地域、(6)人口 1 万人以上地域、(7)人口 5000 人以上地域、および(8)人口 1000 人以上地域の 8 パターンを設定し、それに該当する 20 地域を選定した。

人口 20 万人以上地域を対象にした場合の母集団一意の比率の算出結果によれば、年齢、産業と職業が原区分である場合の母集団一意の比率は約 18.43%であるが、年齢のみを 5 歳区分に統合した場合、比率が 8.83%に減少することがわかった。その一方で、産業や職業についてのみ匿名データと同様の区分でリコーディングを行っても、母集団一意の比率はそれぞれ 17.47%と 17.64%であり、大きな変化は見られないことが確認された。

一方、本研究では、地域の人口規模と母集団一意の比率との間にトレードオフの関係がみられることが実証的に確認された。このことは、年齢、職業、産業と分類区分が設定された場合に、母集団一意の比率に関する閾値を適切に定めることができれば、秘匿の観点から、提供可能な地域の人口規模の閾値を導出することが可能なことを意味している。

つぎに、秘匿性に関する第 2 の研究として、年齢、産業、職業と地域の人口規模に秘匿処理を施した場合の母集団一意の比率に対する影響に関する重回帰分析を行った。具体的には、母集団一意の比率を被説明変数とし、説明変数に関しては、年齢のリコーディングおよびトップコーディングに関する 9 パターン、産業のリコーディングに関する 3 パターン、職業のリコーディン

---

* 本稿は、伊藤・星野・阿久津(2016b)に基づいている。なお、本稿の内容は個人的な見解を示すものであり、統計センターの見解を表すものではないことに留意されたい。

** (独)統計センター非常勤研究員

グに関する 3 パターンのそれぞれに関するダミー変数(パターンに該当する場合には 1、そうでない場合には 0)、および対数変換された地域の人口規模をモデルに設定した。

本分析結果によれば、地域の統合が、年齢、産業、職業と比較しても母集団一意の比率に大きな影響を及ぼすことが確認された。このことから、秘匿性の観点から見た場合、匿名データの作成において、地域区分の設定が相対的に重要な要素を占めることが定量的にも明らかになっている。

3. 情報量損失に基づいた有用性の定量的な評価

つぎに、本研究では、様々な地域を対象に有用性の検証を行った。ミクロデータにおける有用性の定量的な評価方法については、クラーメルの V といった関連性の指標の算出や原データからの絶対距離の平均値(average absolute distance)の計測等を行うことが考えられる(伊藤・星野(2014))。また、情報量損失に関する指標の 1 つであるエントロピーを用いて、秘匿の観点から許容可能な分類区分の組み合わせに関する情報量損失の計測を行うことも可能である(伊藤ほか(2016a))。一方、本研究では、原区分からリコーディングを行った場合の距離を計測することによって、情報量損失の計測を行う。そのために、原区分と統合区分におけるクロス表の差に関する指標を作成し、原区分と統合区分におけるクロス表の差の検証を行った。具体的には、区分統合を行った場合、リコーディング後の度数をリコーディングの対象となった区分数で除することによって、度数の按分を行う。つぎに、リコーディング前のクロス表とリコーディング後に按分済みの度数が入力された表を用いて、情報量損失を算出した。

最初に、分析対象である 20 地域を対象に、81 の変数のパターンにおいて算出される情報量損失値に関する基本統計量を算出した。人口規模が小さい地域については、情報量損失値の平均値が相対的に小さくなっているだけでなく、標準偏差も小さいことが確認できる。このことから、地域の人口規模が小さい場合には、区分統合しても情報量損失の変化が小さいことが推察される。

つぎに、情報量損失と地域規模との関連性を確認したところ、地域の人口規模が大きくなるほど、情報量損失が大きくなる傾向にあることが実証的に明らかになった。その一方で、地域の人口規模が相対的に小さい場合、情報量損失における大きな違いが見られないことを確認することができる。

4. むすびにかえて

本稿では、国勢調査の匿名化ミクロデータを用いて、地域の人口規模に基づく閾値を設定した場合の母集団一意の比率と情報量損失の検証を行った。本研究の結果を踏まえると、個人単位で抽出した匿名化ミクロデータにおいては、地域の人口規模と母集団一意の比率との間にトレードオフの関係があることを実証的に明らかにすることができた。さらに、情報量損失に関する分析結果からは、地域の人口規模が大きいほど、情報量損失が大きくなること、秘匿処理が情報量損失に及ぼす影響は、属性によって異なることがわかった。

一方、本分析結果によれば、年齢、産業、職業を含むキー変数を用いて算出された母集団一意の比率をもとに、秘匿性の閾値を適切に設定することができれば、提供済匿名データにおいて定められている地域の人口規模 50 万以上という区分より細かな地域区分の設定についても、議論することが可能であると思われる。

こうした議論に基づいて、秘匿性と有用性の両面からの実証研究を踏まえた上で、複数ファイルにおける地域詳細化データや年齢等の属性に関する詳細化データの作成可能性を模索する必要があるように思われる。

参考文献

Hawala, S.(2001) “Enhancing the "100,000 rule" On the Variation of the Per Cent of Uniques in A Microdata Sample and the Geographic Area Size Identified on the File”, Proceedings of the Annual Meeting of the American Statistical Association.

伊藤伸介・星野なおみ(2014)「国勢調査ミクロデータを用いたスワッピングの有効性の検証」『統計学』107 号，2014 年 9 月 30 日，1～16 頁

伊藤伸介・星野なおみ・阿久津文香(2016a)「国勢調査における匿名化ミクロデータの有用性と秘匿性の定量的な評価」『製表技術参考資料』No.32, 1～33 頁

伊藤伸介・星野なおみ・阿久津文香(2016b)「国勢調査ミクロデータに対する匿名化措置の可能性に関する研究」『製表技術参考資料』No.34，1～59 頁

Estimation of the shape of density level sets of star-shaped distributions

Hidehiko Kamiya*
Faculty of Economics, Osaka University of Economics

December 2016

Elliptically contoured distributions generalize the multivariate normal distributions in such a way that the density generators need not be exponential. In this way, the class of elliptically contoured distributions includes, for example, distributions whose tails are heavier than those of the multivariate normal distributions. However, as the name suggests, elliptically contoured distributions remain to be restricted in that the similar density contours need to be elliptical. Hence, skewed distributions are not included in this class.

Kamiya, Takemura and Kuriki [5] proposed star-shaped distributions in which the density level sets are allowed to be arbitrary similar star-shaped sets (see also [6], [4]). Essentially the same idea can be found in v-spherical distributions by Fernández, Osiewalski and Steel [2] and center-similar distributions by Yang and Kotz [7]. Asymmetry is allowed in star-shaped distributions. Hence, besides distributions which are symmetric with respect to the center such as elliptically contoured distributions and l_q-spherical distributions, the class of star-shaped distributions includes asymmetric distributions such as multivariate skewed exponential power distributions.

Kamiya, Takemura and Kuriki [5] studied distributional properties of star-shaped distributions, e.g., the independence of the "length" and the "direction," and the robustness of the distribution of the "direction." However, they did not investigate inferential problems about star-shaped distributions. From the perspective of [5], the most important problem in the inference for star-shaped distributions is the estimation of the shape of their density contours

In this paper, we propose a nonparametric estimator of the shape of the density contours. The point is that the density of the usual direction under a star-shaped distribution is in one-to-one correspondence with a function which determines the shape of the density contours. Thus, by nonparametrically estimating the density of the direction, we can obtain a nonparametric estimator of the shape. We prove its strong consistency with respect to the Hausdorff distance.

Our main result is the following:

*This work was partially supported by JSPS KAKENHI Grant Number 25400201.

Theorem 0.1. *Let $x_1, \dots, x_n \in \mathcal{X} = \mathbb{R}^p \setminus \{\mathbf{0}\}$, $p \geq 2$, be an i.i.d. sample from a star-shaped distribution $h(r(x))dx$. Let $\hat{f}_n(u) = (C(\eta)/(n\eta^{p-1})) \sum_{i=1}^n L((1 - u^T u_i)/\eta^2)$ be a kernel estimator (Hall, Watson and Cabrera [3], Bai, Rao and Zhao [1]) of the density $f(u)$ of $u = x/\|x\| \in \mathbb{S}^{p-1}$, $x \sim h(r(x))dx$, based on $u_i = x_i/\|x_i\|$, $i = 1, \dots, n$.*

Assume the equivariant function $r : \mathcal{X} \to \mathbb{R}_{>0}$ under the action of the positive real numbers is continuous and normalized so that $\int_{\mathbb{S}^{p-1}} r(u)^{-p} du = 1$, and that $L : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is bounded and satisfies $\int_0^\infty L(v) v^{(p-3)/2} dv > 0$ and $\int_0^\infty \sup_{w:\,|\sqrt{w}-\sqrt{v}|<1} L(w) \cdot v^{(p-3)/2} dv < \infty$. Moreover, suppose $\eta = \eta_n > 0$ is taken in such a way that $\eta_n \to 0$ and $n\eta_n^{p-1}/\log n \to \infty$ as $n \to \infty$.

Then, $\hat{Z}_n = \{\hat{f}_n(u)^{1/p} u : u \in \mathbb{S}^{p-1}\}$ is a strongly consistent estimator of the shape $Z = \{x \in \mathcal{X} : r(x) = 1\}$ of the density contours of the star-shaped distribution in the sense that the Hausdorff distance $\delta_H(\hat{Z}_n, Z)$ between $\hat{Z}_n$ and Z satisfies

$$\delta_H(\hat{Z}_n, Z) \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

In addition, $\hat{\mathcal{Z}}_n = \bigcup_{0 \leq c \leq 1} c\hat{Z}_n$ is a strongly consistent estimator of $\mathcal{Z} = \bigcup_{0 \leq c \leq 1} cZ$: $\delta_H(\hat{\mathcal{Z}}_n, \mathcal{Z}) \to 0$ a.s.

References

[1] Z. D. Bai, C. R. Rao and L. C. Zhao, Kernel estimators of density function of directional data, *Journal of Multivariate Analysis* **27** (1988), 24–39.

[2] C. Fernández, J. Osiewalski and M. F. J. Steel, Modeling and inference with v-spherical distributions, *Journal of the American Statistical Association* **90** (1995), 1331–1340.

[3] P. Hall, G. S. Watson and J. Cabrera, Kernel density estimation with spherical data, *Biometrika* **74** (1987), 751–762.

[4] H. Kamiya and A. Takemura, Hierarchical orbital decompositions and extended decomposable distributions, *Journal of Multivariate Analysis* **99** (2008), 339–357.

[5] H. Kamiya, A. Takemura and S. Kuriki, Star-shaped distributions and their generalizations, *Journal of Statistical Planning and Inference* **138** (2008), 3429–3447.

[6] A. Takemura and S. Kuriki, Theory of cross sectionally contoured distributions and its applications, Discussion Paper Series 96-F-15, July 1996, Faculty of Economics, University of Tokyo.

[7] Z. Yang and S. Kotz, Center-similar distributions with applications in multivariate analysis, *Statistics & Probability Letters* **64** (2003), 335–345.

正方分割表の一致率検定のための代数的方法

吉田 知行 (北星学園大学 経済学部)
t-yoshida@hokusei.ac.jp

概要: 2 元以上の正方分割表の一致率の検定について，正確な p-値を求める方法を紹介する．この方法は，フィッシャーの並べ替え検定，フィッシャーの正確確率法の流れをくむ方法で，代数的には，対称群のデータセットへの作用と指標理論の応用分野である．また，有限群上のランダムウォーク (RW) とも関係している．
キーワード： 正確確率法，並べ替え検定，有限群上のランダムウォーク，分割表の列挙問題，有限群の表現.

一致率検定の正確な p-値を求める方法とその背景にある代数理論を紹介した．主結果は次である．正方分割表 $X=(x_{ij})$ の周辺和を $\boldsymbol{a}=(x_{i+})$, $\boldsymbol{b}=(x_{+j})$ とする．周辺分布が $\boldsymbol{a},\boldsymbol{b}$ であるような正方分割表の集合を $\mathsf{tab}(\boldsymbol{a},\boldsymbol{b})$ とする．$P(k)$ を，一致数が k 以上になる確率とする:

$$P(k)=\mathrm{Prob}(\boldsymbol{x}\in\mathsf{tab}(\boldsymbol{a},\boldsymbol{b})\mid \mathrm{Tr}(\boldsymbol{x})\geq k).$$

数列 $\{q(k)\},\{p(k)\}$ $(k=0,1,\cdots,n)$ を次で定義する：

$$\prod_{\lambda}{}_2F_0(-a_i,-b_i;z)=\sum_{k\geq 0}\binom{n}{k}k!q(k)z^k,$$

$$p(k)=\sum_{j=k}^{n}(-1)^{j-k}\binom{j}{k}q(j),\quad \sum_{k=0}^{n}p(k)z^k=\sum_{k=0}^{n}q(k)(z-1)^k$$

0 ここで ${}_2F_0$ は超幾何多項式である．このとき $P(k)$ は次で与えられる．

$$\sum_{k=0}^{n}P(k)z^k=1+z\sum_{k=1}^{n}q(k)(z-1)^{k-1}$$

この結果は，対称群の直積 $S_n\times S_n$ のデータセット $\mathsf{DS}(\boldsymbol{a},\boldsymbol{b}$ への可移な作用と，対称群上のランダムウォークから来ている:

$$\boxed{S_n \text{ 上の RW}}\longrightarrow\boxed{S_n\times S_n/S_n^{\Delta}\text{ 上の RW}}\longrightarrow\boxed{\mathsf{DS}(\boldsymbol{a},\boldsymbol{b})\text{ 上の RW}}\longrightarrow\boxed{\mathsf{tab}(\boldsymbol{a},\boldsymbol{b})\text{ 上の RW}}$$

本来なら，分割表をランダムサンプリングして，その中から $\mathrm{Tr}(\boldsymbol{x}) \geq k$ となるものの割合を求める．しかし一致数の場合は超幾何多項式で表せたのである．一致数の p-値は，比較言語学，とくにポリアによる二項検定法やオズワルトのシフトに現れる．後者は安本美典による日本語の起源探求に使われた．

参 考 文 献-

[吉田 07] 吉田知行，分割表の一致率検定への有限群論と組合せ論の応用，代数学シンポジウム，於神戸大学．ネットに pdf ファイルあり．

[吉田 09] 吉田知行，Finite Gelfand pairs and Markov chain Monte-Carlo method,「表現論と組合せ論」RIMS 講究録第 **1689** (201),164-170,