統計科学における深化と横断的展開

日　　時：　平成 24 年 10 月 24 日（水）～ 10 月 26 日（金）
会　　場：　松江テルサ（松江勤労者総合福祉センター）４階大会議室
URL: http://www.sanbg.com/terrsa/

プログラム

10 月 24 日（水）【セッション 1～2】

セッション１　座長　内藤　貫太（島根大学）
13：30～14：00
「射影べきエントロピーを用いた異質性の検出」
野津 昭文(総合研究大学院大学), 江口 真透(統計数理研究所)
14：00～14：30
「標本歪度尖度の同時分布」
中川 重和(倉敷芸術科学大学), 橋口 博樹(埼玉大学), 仁木 直人(東京理科大学)
14：30～15：00
「計算機代数的手法の統計的因果推論への応用」
田中 研太郎(東京工業大学)

セッション２　座長　中川　重和（倉敷芸術科学大学）
15：30～16：00
「米国国債モデルと金利の期間構造」
刈屋 武昭, 山村 能郎 (明治大学大学院), 王 竹(ZW システム)
16：00～16：30
「Shrinkage Estimation of Optimal Portfolio」
白石 博(東京慈恵会医科大学)
16：30～17：00
「Analysis of CL and Estimating Function Estimators for Financial Time Series Models」
天野 友之(和歌山大学)
17：00～17：30
「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」
高橋 将宜, 伊藤 孝之(独立行政法人 統計センター)

10 月 25 日（木）【セッション 3～6】

セッション 3　座長　矢田　和善（筑波大学）

9：30～10：00

「High-Dimensional Mean Estimation via L1-Penalized Normal Likelihood」

片山 翔太（大阪大学大学院）

10：00～10：30

「LARS に基づく L1 正則化法におけるモデル選択基準の構成」

保科 架風（中央大学大学院），廣瀬 慧（大阪大学），小西 貞則(中央大学)

セッション 4　座長　小泉　和之（横浜市立大学）

11：00～11：30

「L1 型正則化法による因子分析モデルのスパース推定」

廣瀬 慧，山本 倫生(大阪大学)

11：30～12：00

「高次元小標本における幾何学的表現とその応用」

矢田 和善，青嶋 誠(筑波大学)

セッション 5　座長　廣瀬　慧（大阪大学）

13：30～14：00

「Moment convergence of Z-estimators and Z-process method for change point problems」

西山 陽一（統計数理研究所）

14：00～14：30

「金利スプレッドによる倒産確率の推定」

高橋 一（鳥取環境大学）

14：30～15：00

「Brownian Quantile を用いた逐次解析の試み」

三浦 良造

セッション 6　座長　西山　陽一（統計数理研究所）

15：30～16：00

「標本積率を用いた多変量正規性検定について」

小泉 和之（横浜市立大学），澄川 琢磨(東京理科大学大学院)

16：00～16：30

「新たな離散異分布適合度検定統計量とその海洋調査データへの適用」

柴田 里程（慶應義塾大学）

16：30～17：00

「Improved confidence intervals for quantiles」

前園 宜彦（九州大学）， Spiridon Penev（University of New South Wales）

17：00～17：30

「情報科学演習におけるグループ内変動の定量的分析」

安田 晃（島根大学）

10 月 26 日（金）【セッション 7～9】

セッション 7　座長　内藤　貫太（島根大学）

9：30～10：00

「多変量線形回帰モデルにおける AIC の漸近性質について」

伊森 晋平（広島大学大学院）

10：00～10：30

「罰則付カーネル正準相関分析における罰則最適化のための CV 規準」

永井 勇（広島大学大学院）

セッション 8　座長　安田　晃（島根大学）

11：00～11：30

「自己組織化マップと情報量規準を用いたクラスタリング手法の提案」

加藤 聡，堀内 匡(松江工業高等専門学校)

11：30～12：00

「外来待ち時間と患者満足度、および入院期間と患者満足度との関連 」

奥田 益美（松江赤十字病院，島根大学大学院）

12：00～12：30

「テンジククルマエビの成長と生存のモデルから導かれる非対称混合分布で検出された淡水流入効果」

仲 真弓，柴田 里程（慶應義塾大学大学院）

セッション 9

13：30～15：00

総合討論

世 話 人：内藤貫太（島根大学，研究分担者)，谷口正信(早稲田大学，研究代表者)

予　　算：基盤研究（A)「非対称・非線形統計理論と経済・生体科学への応用」

射影べきエントロピーによる異質性の検出

総合研究大学院大学　野津 昭文　　統計数理研究所　江口 真透

1　始めに

最尤推定はボルツマン・シャノンエントロピー (BS エントロピー) 最小化に等価であることはよく知られているが，BS エントロピー以外のエントロピーを代わりに用いることで，最尤推定とは異なる性質を持つ推定法を考えることができる．Minami and Eguchi (2002) では独立成分分析に β-エントロピーを用いており，Naito and Eguchi (2012) では U-エントロピーを密度推定に応用している．本研究では，射影べきエントロピー (Fujisawa and Eguchi, 2008) を用いる．射影べきエントロピーを用いることの利点としてロバストな推定をすることが可能となるが，本研究では新たに異質性を検出するという性質に焦点を当て，射影べきエントロピーを用いたガウシアンコピュラの推定やクラスタリングについて考察する．

2　射影べきエントロピー，クロスエントロピーと異質性の検出

データが従う分布を g とし，統計モデルを $f(x;\theta)$ とすると，射影べきクロスエントロピーは，

$$\begin{aligned} C_\gamma(g, f(\cdot;\theta)) &= -\int g(x)\kappa_\gamma(\theta) f(x;\theta)^\gamma dx, \\ \kappa_\gamma(\theta) &= \left(\int f(x;\theta)^{1+\gamma} dx\right)^{-\frac{\gamma}{1+\gamma}} \end{aligned}$$

で定義される．$H_\gamma(g) = C_\gamma(g, g)$ を射影べきエントロピーといい，射影べきエントロピーは BS エントロピーの拡張となっている (Eguchi et.al 2011)．$g(x)$ を $f(x;\theta)$ の混合分布，$f(x;\theta)$ は指数型分布族と仮定する：

$$\begin{aligned} g(x) &= \tau_1 f(x;\theta_1) + \tau_2 f(x;\theta_2), \tau_1 + \tau_2 = 1, \tau_i \geq 0, i = 1, 2, \\ f(x;\theta) &= \exp(\theta^\top t(x) - \psi(\theta)). \end{aligned}$$

このとき，射影べきクロスエントロピーは

$$C_\gamma(g, f(\cdot;\theta)) = \tau_1 C_\gamma(\theta_1, \theta) + \tau_2 C_\gamma(\theta_2, \theta)$$

となる．ここで $C_\gamma(\theta_i, \theta) = C_\gamma(f(\cdot;\theta_i), f(\cdot;\theta)), i = 1, 2$ である．$C_\gamma(\theta_i, \theta)$ は次の性質を満たす.

命題 2.1 $C_\gamma(\theta_i, \theta)$ は有界であり，

$$C_\gamma(\theta_i, \theta_i) \leq C_\gamma(\theta_i, \theta) \leq 0$$

を満たし，任意の θ に対し $\theta_t = (1-t)\theta_i + t\theta (0 \leq t \leq 1)$ とすると，$C_\gamma(\theta_i, \theta_t)$ は t に関して単調増加である.

よって $C_\gamma(g, f(\cdot;\theta))$ は有界単峰関数の重み付きの和であり，次の予想が成り立つことが直観的に理解される.

予想 2.1 θ_1 と θ_2 が十分異なっていれば，$C_\gamma(g, f(\cdot;\theta))$ は *2* つの極小値を持ち，それらの極小解は θ_1 と θ_2 にそれぞれ近い値となる.

予想 2.1 を一般的に証明することは困難であるが，より具体的に $f(x;\theta)$ の形を限定すれば証明することができ，また，シミュレーションによって極小解が存在する様子を確認することもできる．予想 2.1 において，混合分布を構成する成分を 2 つとしているが，本質的には何個でも構わない．予想 2.1 より，$\hat{g}$ を経験分布関数としたとき，ロス関数 $L_\gamma(\theta)$ を $C_\gamma(\hat{g}, f(\cdot;\theta))$ で定義すると，$L_\gamma(\theta)$ の極小値を用いることで，θ_1 や θ_2 を推定でき，母集団の異質な構造をとらえることができる．本研究ではこの性質をガウシアンコピュラの推定やクラスタリングに応用するが，紙数の制約上ガウシアンコピュラの場合のみ記述する．

3　ガウシアンコピュラの推定における異質性の検出

ガウシアンコピュラの密度関数は

$$c_{\mathrm{G}}(\boldsymbol{u};P)=\det P^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{x}_{\mathrm{G}}(\boldsymbol{u})^{\top}(P^{-1}-I_m)\boldsymbol{x}_{\mathrm{G}}(\boldsymbol{u})\right),\ \boldsymbol{u}\in[0,1]^m,$$

である (McNeil et. al., 2005)．ただし，$\boldsymbol{x}_{\mathrm{G}}(\boldsymbol{u})=(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_m))^{\top}$，$\Phi(x)$ は標準正規分布の分布関数，P は相関行列，I_m は m 次単位行列とする．データを $\boldsymbol{u}_1,\dots,\boldsymbol{u}_n$，統計モデルをガウシアンコピュラとし，射影べきクロスエントロピーの計算に適当な測度を用いた場合，γ-推定量のロス関数は定数倍を除いて，

$$L_\gamma(P) \quad = \quad -\det(P)^{-\frac{\gamma}{2(1+\gamma)}}\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\gamma}{2}\boldsymbol{x}_i^{\top}P^{-1}\boldsymbol{x}_i\right)$$

となる．ただし，$\boldsymbol{x}_i=\boldsymbol{x}_{\mathrm{G}}(\boldsymbol{u}_i), i=1,\dots,n$ とする．$L_\gamma(P)$ の極小値を γ-推定量と定義する．データが従う分布がガウシアンコピュラ $c_{\mathrm{G}}(\boldsymbol{u};P_0)$ の場合，γ-推定量の一致性と漸近正規性が成り立つ．データが従う分布が

$$c(\boldsymbol{u})=\tau_1 c_{\mathrm{G}}(\boldsymbol{u};P_1)+\tau_2 c_{\mathrm{G}}(\boldsymbol{u};P_2),\ \tau_1+\tau_2=1,\tau_i\geq 0,\ i=1,2$$

であり，$m=2$ の場合は次の命題を示すことができる．

命題 3.1 射影べきクロスエントロピー $C_\gamma(c,c_G(\cdot;P)|Q_G)$ は P の非対角成分 ρ の関数である．P_1 と P_2 を

$$P_1=\begin{pmatrix}1 & \rho_*\\ \rho_* & 1\end{pmatrix},\ P_2=\begin{pmatrix}1 & -\rho_*\\ -\rho_* & 1\end{pmatrix},$$

とし，$\gamma=1,\tau_1=\tau_2=0.5$ とする．$\rho_*>\sqrt{6-\sqrt{28}}\doteqdot 0.842$ ならば，射影べきクロスエントロピーは *2* つの極小解を持ち，それぞれ区間 $(-1,0)$ と $(0,1)$ に含まれる．

より高次元でも 2 つの相関行列を検出できることがシミュレーションの結果から分かっている．

4　まとめ

本研究では，射影べきエントロピーの特徴的な性質である予想 2.1 を用いて，母集団の異質な構造を検出する方法を提案し，それをガウシアンコピュラの推定やクラスタリングに応用した．数理的な観点からもシミュレーションの結果からも，この手法は有効に働くことが確認できた．

参考文献

[1] Shinto Eguchi, Osamu Komori, and Shogo Kato. Projective power entropy and maximum tsallis entropy distributions. *Entropy*, Vol. 13, No. 10, pp. 1746–1764, 2011.

[2] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, Vol. 99, No. 9, pp. 2053–2081, 2008.

[3] Alexander J. McNeil, Rudiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2005.

[4] Mihoko Minami and Shinto Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, Vol. 14, No. 8, pp. 1859–1886, August 2002.

[5] Kanta Naito and Shinto Eguchi. Density estimation with minimization of u-divergence. *Machine Learning*, Vol. 89, , 2012.

標本歪度尖度の同時分布

中川 重和　(倉敷芸術科学大・産業科学技術)
橋口 博樹　(埼玉大大学院・理工学研究科)
仁木 直人　(東京理科大・工)

1　はじめに

標本歪度と標本尖度は正規性検定統計量として基本的に用いられる統計量であり，その際にはそれらの分布が重要となる．標本歪度の帰無分布の性質が，そのモーメントを含めて，明らかにされていることに対し，標本尖度のそれはそれほど解明されていない [4]．本報告では，標本歪度と標本尖度の同時帰無分布を考える．なお，同時密度関数の関連研究として，[3] では標本歪度を与えたときの標本尖度の条件付き分布関数を χ^2 近似し，それを用いて同時密度関数の等高線図を描いている．それに対し，本報告は [1] と [2] の結果の自然な拡張である．

本報告では，正規標本からの標本歪度と標本尖度の同時密度関数を漸化式の形で与える (定理 1)．また，その同時モーメントも定理 2 で与える．

2　同時密度関数

大きさ n の標本 $(X_1, X_2, \ldots, X_n)$ に対し，標本歪度と標本尖度は，それぞれ $\sqrt{b_1} = m_3/m_2^{3/2}$, $b_2 = m_4/m_2^2$ である．ここで，$m_r = 1/n \sum_{i=1}^n \left(X_i - \bar{X}\right)^r$ $(r = 2, 3, 4)$，$\bar{X} = 1/n \sum_{i=1}^n X_i$.

定理 1 大きさ n の正規標本からの $(\sqrt{b_1}, b_2)$ の同時密度関数を $h_n(x, y)$ とする．このとき，

$$h_{n+1}(x,y) = \frac{\left(\frac{n}{n+1}\right)^{\frac{3}{2}}}{B\left(\frac{1}{2}, \frac{n}{2} - \frac{1}{2}\right)} \int_{-1}^{1} h_n\left(\sigma_n(x,z), \tau_n(x,y,z)\right) (1-z^2)^{\frac{n-10}{2}} dz. \tag{1}$$

ただし，

$$\sigma_n(x,z) = \left\{\sqrt{n}x - 3z + (n+2)z^3\right\} (n+1)^{-\frac{1}{2}} (1-z^2)^{-\frac{3}{2}}, \tag{2}$$

$$\tau_n(x,y,z) = \left\{ny - 4\sqrt{n}xz + 6z^2 - (n^2+3n+3)z^4\right\} (n+1)^{-1} (1-z^2)^{-2} \tag{3}$$

であり，$B(\cdot, \cdot)$ はベータ関数である．

3　同時モーメント

定理 2 大きさ n の正規標本からの $(\sqrt{b_1}, b_2)$ の同時分布の原点周りの r, s 次モーメントを $\nu_{r,s}^{(n)}$ とする．このとき，

$$\begin{aligned}
\nu_{2r,s}^{(n+1)} = {} & \frac{\left(\frac{n+1}{n}\right)^{r+s}}{B\left(\frac{1}{2}, \frac{n}{2} - \frac{1}{2}\right)} \sum_{i=0}^{2r} \binom{2r}{i} \sum_{j=0}^{s} \binom{s}{j} \sum_{k=0}^{s-j} \binom{s-j}{k} \frac{4^k \nu_{2r-i+k, s-j-k}^{(n)}}{(n+1)^{\frac{i}{2}+j+\frac{k}{2}}} \times \\
& \sum_{\ell=0}^{i} \binom{i}{\ell} 3^{i-\ell} (1-n)^{\ell} \sum_{m=0}^{j} \binom{j}{m} 6^{j-m} (n^2-n+1)^m \times \\
& B\left(j + \ell + m + \frac{i}{2} + \frac{k}{2} + \frac{1}{2}, 3r + 2s - j - \ell - m + \frac{n-1}{2} - \frac{i}{2} - \frac{k}{2}\right).
\end{aligned} \tag{4}$$

$n=3$ のときは,

$$\nu_{2r,s}^{(3)} = \prod_{i=0}^{r-1} \left(\frac{2r-2i-1}{2r-2i} \right) 2^{-r} \left(\frac{3}{2} \right)^{s}. \tag{5}$$

定理 2 の結果は，数値計算だけでなく，数式計算としても利用できる．実際，数式処理システム Mathematica に実装し，$\nu_{2r,s}^{(n)}$ $(2r+s \leq 20)$ を n に関する有理式で導出している．

定理 2 より，$\sqrt{b_1}$ の帰無分布のモーメントの漸化式が得られるが，これは [2] の結果と一致している．

系 3 定理 *2* において $s=0$ とすると，標本歪度 $\sqrt{b_1}$ の帰無分布のモーメントの漸化式

$$\begin{aligned} \nu_{2r,0}^{(n+1)} &= \frac{\left(\frac{n+1}{n}\right)^{r}}{B\left(\frac{1}{2},\, \frac{n}{2}-\frac{1}{2}\right)} \sum_{i=0}^{2r} \binom{2r}{i} \frac{\nu_{2r-i,0}^{(n)}}{(n+1)^{\frac{i}{2}}} \times \\ & \sum_{\ell=0}^{i} \binom{i}{\ell} 3^{i-\ell} (1-n)^{\ell} B\left(\ell + \frac{i}{2} + \frac{1}{2},\, 3r - \ell + \frac{n-1}{2} - \frac{i}{2}\right) \end{aligned} \tag{6}$$

を得る．

同様に，b_2 の帰無分布のモーメントの漸化式も，同時モーメントを介し，次のように得られる．

系 4 定理 *2* において $r=0$ とすると，標本尖度 b_2 の帰無分布のモーメントの漸化式は次式となる．

$$\begin{aligned} \nu_{0,s}^{(n+1)} &= \frac{\left(\frac{n+1}{n}\right)^{s}}{B\left(\frac{1}{2},\, \frac{n}{2}-\frac{1}{2}\right)} \sum_{j=0}^{s} \binom{s}{j} \sum_{k=0}^{s-j} \binom{s-j}{k} \frac{4^{k} \nu_{k,s-j-k}^{(n)}}{(n+1)^{j+\frac{k}{2}}} \times \\ & \sum_{m=0}^{j} \binom{j}{m} 6^{j-m} (n^2-n+1)^{m} B\left(j+m+\frac{k}{2}+\frac{1}{2},\, 2s-j-m+\frac{n-1}{2}-\frac{k}{2}\right). \end{aligned} \tag{7}$$

参考文献

[1] Geary, RC (1947) The frequency distribution of $\sqrt{b_1}$ for samples of all sizes drawn at random from a normal population. *Biometrika*, **34**, 68–97

[2] Mulholland, HP (1977) On the null distribution of $\sqrt{b_1}$ for samples of size at most 25, with tables, *Biometrika*, **64** (2), 401–409

[3] Shenton, LR and Bowman, KO (1977) A Bivariate Model for the Distribution of $\sqrt{b_1}$ and $b2$, *Journal of the American Statistical Association*, **72** (357), 206–211

[4] Thode, H. C. *Testing for normality (Statistics, a Series of Textbooks and Monographs)*. Marcel Dekker Inc., New York (2002).

計算機代数的手法の統計的因果推論への応用*

東京工業大学大学院社会理工学研究科 田中研太郎

今回の発表では, 各セルの確率が正である分割表における確率分布を考える. 確率変数の次元は n であるとし, $X_1, \ldots, X_n$ で表すとする. $N = \{1, 2, \ldots, n\}$ とし, その部分集合 $A = \{a_1, a_2, \ldots, a_{|A|}\} \subset N$ と確率変数集合 $X_A = (X_{a_1}, \ldots, X_{a_{|A|}})$ とを同一視したりもする. また, 確率変数の集合 (添え字の集合) を A, B, C, D, E, F, G などの大文字で表すことにし, 1 つの元だけからなる確率変数の集合 (添え字の集合) の場合には a, b, c, d, e, f, g などの小文字で表すことにする. さらに, $A \cup B$ を AB などと略して書くことにする.

このような設定のもとで, 以下のような問題を考える:

> 「ある確率分布に対していくつかの条件付き独立性が成り立つことが分かっている場合に, 他の条件付き独立性が成り立つことをそれらから導くことができるか?」

例えば, 「$a \perp\!\!\!\perp b, a \perp\!\!\!\perp b \mid c \Rightarrow a \perp\!\!\!\perp (b, c)$」などのような命題が正しいかどうか判定するといった問題を考える. つまり, 条件付き独立性の間に成り立つ関係式 (含意) を導出したい.

今回の発表では, imset を使って条件付き独立性の関係式の導出する新しい方法を提案する. 用いたアイデアはシンプルで, 元の関係式 (含意) の導出の問題を解きやすい形に変形するというものだが, それにより, Studený [1] における imset の手法では解けない問題が解けるようになった.

X_C を与えたときに X_A と X_B が条件付き独立であるとは,

$$p(X_A, X_B \mid X_C) = p(X_A \mid X_C) p(X_B \mid X_C) \tag{1}$$

が満たされることをいう. これは, $p(A, B, C)^1 p(A, C)^{-1} p(B, C)^{-1} p(C)^1 = 1$ と同値である.

条件付き独立性の関係式 (含意) の問題は一般的に以下のように表される.

$$\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^{I} \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, C. \tag{2}$$

左側のいくつかの条件付き独立性 $\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^{I}$ が成立することが分かっているときに, 右側の条件付き独立性 $A \perp\!\!\!\perp B \,|\, C$ が成立するかどうかを判定するのが考えたい問題である.

*本稿の内容は, 竹村彰通教授 (東京大学大学院情報理工学系研究科数理情報学専攻), 清智也講師 (慶応義塾大学理工学部数理科学科), Millan Studeny 教授 (Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic) との共同研究に基づいている.

提案手法について簡単に説明する. 提案手法は, 元の関係式 (含意) の導出の問題 ((2) 式) の既知の条件付き独立性の部分 (左辺) に, ある解法を実行する際に問題を変えないように他の条件付き独立性を足す, というアイデアに基づいている. そのために以下を定義する.

Definition 1. $E \perp\!\!\!\perp F \,|\, G$ が $A \perp\!\!\!\perp B \,|\, C$ を「ブリッジしない」とは, 以下を満たすときのことをいう.

$$(EFG) \subseteq AC \quad \text{or} \quad (EFG) \subseteq BC \tag{3}$$

このとき, 以下の定理が成り立つ.

Theorem 1. 条件付き独立性の関係式 *(*(2) 式*)* を

$$p(A_i, B_i, C_i)^1 p(A_i, C_i)^{-1} p(B_i, C_i)^{-1} p(C_i)^1 \tag{4}$$

の冪乗とかけ算だけで導出する場合を考える. いま, $A \perp\!\!\!\perp B \,|\, C$ をブリッジしない条件付き独立性の集合を $\{E_j \perp\!\!\!\perp F_j \,|\, G_j\}_{j=1}^J$ とする. このとき,

$$\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^I \cup \{E_j \perp\!\!\!\perp F_j \,|\, G_j\}_{j=1}^J \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, C \tag{5}$$

が導出可能であることと, (2) 式が導出可能であることは同値になる.

今回の発表では, imset を使って条件付き独立性の関係式を導出する新しい手法について紹介した. 実は, この手法のアイデアを発展させて, より効率的なアルゴリズムを構築できることが最近の研究で分かってきた. また, 今回は, 標本空間が有限で確率関数が正の場合のみを考えたが, より広いクラスでも同様な結果が成立するかのどうかも興味深い問題である.

参考文献

[1] Milan Studený. *Probabilistic Conditional Independence Structures.* Springer-Verlag, London, 2005.

MEIJI UNIVERSITY
GRADUATE SCHOOL OF GLOBAL BUSINESS

米国国債モデルと金利の期間構造

2012年10月24日

刈屋武昭(明治大学大学院グローバル・ビジネス研究科)
山村能郎(明治大学大学院グローバル・ビジネス研究科)
王竹(ZWシステム)

概要

- Kariya(1993), Kariya & Tsuda(1993)モデルを用いた米国債USGB価格分析モデルの推定
- 債券の属性(満期、クーポン)を考慮したクロスセクション価格モデルの推定とパフォーマンス比較
- 推定された割引率関数と金利の期間構造の導出
- 金利の期間構造とスワップレートとの比較
- 同時期におけるJGBモデルとの比較
  - 日本国債について刈屋他(2011), Kariya et al.(2012)において分析

信用リスクのない固定クーポン・プライシング・モデル

■スポットレート・アプローチ

スポットレート・プロセス$\{ru\}$を定式化し，条件付き期待値で割引率を計算—条件付きモデリング：金利変動プロセスはマルコフ

■フォワードレート・アプローチ

金利の期間構造を瞬間的フォワードレートプロセスで表現し，割引率を計算—無条件モデリング

■国債の持つ情報－金利の期間構造
- 国債市場が効率的であるとすれば、各時点での価格情報には，投資家の将来経済に対する評価としての金利の期間構造評価が潜在的に含まれている．
- 一定の相関構造を前提としてクロスセクション価格情報から投資家のみる将来経済，金利の期間構造を導出することが可能．

■現実の金利および金利変動
- マルコフ性の条件を満たさない
- フォワードレート・アプローチ
- 無条件クロスセクション価格モデル

固定クーポン国債価格モデル

■ t=0を現在時点，G 個の国債が存在.

■ 第 g 国債の CF発生時点は，以下の通り.

$$s_{g1} < s_{g2} < \cdots < s_{gM(g)} \quad (g = 1, \cdots, G)$$

■ $C_g(s)$ はCF関数で, $s = s_{gj}$でなければ、その値は 0 .

■ $D_g(s)$ を属性依存型確率的割引関数（以下は定義域）

$$0 < s \le s_{gM(g)}$$

■ **基本モデル**

$$P_g = \sum_{j=1}^{M(g)} C_g(s_{gj}) D_g(s_{gj}) \qquad (g = 1, \cdots, G)$$

realization $\boldsymbol{P_g}$ $\Longleftrightarrow$ $\{D_g(s) : 0 \le s \le s_{gM(g)}\}$

属性依存型フォワードレート期間構造の国債価格モデル

$$P_g = \sum_{m=1}^{Ma} C_g(s_{am}) D(s_{am})$$

$$D_g(s) = \bar{D}_g(s) + \Delta_g(s) \qquad \text{Mean DF + Random DF}$$

➢無条件モデリング

$$P_g = \sum_{m=1}^{M(g)} C_g(s_{gm}) \bar{D}_g(s_{gm}) + \eta_g$$

$$\eta_g = \underset{\sim}{C}_g' \underset{\sim}{\Delta}_g = \sum_{m=1}^{M(g)} C_g(s_{gm}) \Delta_g(s_{gm}),$$

$$\underset{\sim}{C}_g = (C_g(s_{g1}), \cdots, C_g(s_{gM(g)}))' : M(g) \times 1$$

$$\underset{\sim}{\Delta}_g = (\Delta_g(s_{g1}), \cdots, \Delta_g(s_{gM(g)}))' : M(g) \times 1.$$

属性依存型平均割引率関数の多項式近似

$$\bar{D}_g(s) = 1 + (\delta_{11} w_1 z_{g1} + \delta_{12} w_2 z_{g2} + \delta_{13} w_3 z_{g3}) s + \cdots + (\delta_{p1} w_1 z_{g1} + \delta_{p2} w_2 z_{g2} + \delta_{p3} w_3 z_{g3}) s^p$$

$$z_{1g} = 1, \quad z_{2g} = s_{gM(g)} \quad z_{3g} = c_g$$

$$w_1 = 1, \; w_2 = 0 \text{ or } 1, \quad w_3 = 0 \text{ or } 1$$

■ 属性を考慮した4つのモデル(w_1, w_2, w_3)
M0 モデル of (1,0,0); 基本モデル　属性効果なし
M1 モデル of (1,1,0): M0 + 満期効果,
M2 モデル of (1,0,1); M0 + クーポン効果
M3 モデル of (1,1,1); M0 +満期効果＋クーポン効果

■ Kariya and Tsuda (1994,95) は(w1,w2,w3) = (0,1,1), p=2 を考察

多項式近似した国債価格モデル

$$\sum_{m=1}^{M(g)} C_g(s_{gm}) \bar{D}_g(s_{gm})$$

$$= a_g + (\delta_{11} d_{g11} + \delta_{12} d_{g21} + \delta_{13} d_{g31}) + \cdots + (\delta_{p1} d_{g1p} + \delta_{p2} d_{g2p} + \delta_{p3} d_{g3p}),$$

$$a_g = \sum_{m=1}^{M(g)} C_g(s_{gm}) \quad d_{gij} = \sum_{m=1}^{M(g)} C_g(s_{gm}) z_{gi} s_{gm}^j \quad i\text{:属性番号(1,2,3)}, \; j\text{:多項式次数}$$

$$\underset{\sim}{x}_g = (d_{g11}, d_{g21}, d_{g31}; d_{g12}, d_{g22}, d_{g32}; \cdots\cdots; d_{g1p}, d_{g2p}, d_{g3p})' : 3p \times 1$$

$$X = (\underset{\sim}{x}_1, \underset{\sim}{x}_2, \cdots, \underset{\sim}{x}_G)' : G \times 3p$$

■回帰モデル

$$\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{\eta} \qquad \underset{\sim}{\eta} = (\eta_1, \ldots, \eta_G)'$$

$$\underset{\sim}{y} = (y_1, y_2, \cdots, y_G)' : G \times 1 \;\; with \;\; y_g = P_g - a_g$$

$$\underset{\sim}{\beta} = (\delta_{11}, \delta_{12}, \delta_{13}; \delta_{21}, \delta_{22}, \delta_{23}; \cdots\cdots; \delta_{p1}, \delta_{p2}, \delta_{p3})' : 3p \times 1$$

確率的割引率関数の分散構造の定式化:相関構造

$$Cov(D_g(s_{gj}), D_h(s_{hm})) = \sigma^2 \lambda_{gh} f_{gh\cdot jm}$$

➢満期期間による相関構造

$$\lambda_{gh} = \begin{cases} e_{gg} & (g = h) \\ \rho e_{gh} & (g \ne h) \end{cases} \quad with \quad e_{gh} = \exp(-\xi \left| s_{gM(g)} - s_{hM(h)} \right|)$$

➢CF発生時点間の割引率に係る相関構造

$$f_{gh.jm} = \exp(-\theta \left| s_{gj} - s_{hm} \right|)$$

■各銘柄の価格間の相関構造

$$Cov(\underset{\sim}{\eta}) = (Cov(\eta_g, \eta_h)) = (Cov(P_g, P_h)) = \sigma^2 (\lambda_{gh} \varphi_{gh}) \equiv \sigma^2 \Phi(\theta, \rho, \xi)$$

$$\varphi_{gh} = \sum_{j=1}^{M(g)} \sum_{m=1}^{M(m)} C_g(s_{gj}) C_h(s_{hm}) f_{gh.jm}$$

1)2つの債券の満期期間の差が大きいほど，確率的割引率の相関は小さい
2)2つのCF時点が近いほど，それに対応する確率的割引率の相関は大きい

一般化最小二乗法によるモデル推定

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\eta}$$

$$Cov(\underset{\sim}{\eta}) = (Cov(\eta_g, \eta_h)) = (Cov(P_g, P_h)) = \sigma^2(\lambda_{gh}\varphi_{gh}) \equiv \sigma^2\Phi(\theta, \rho, \xi)$$

■目的関数

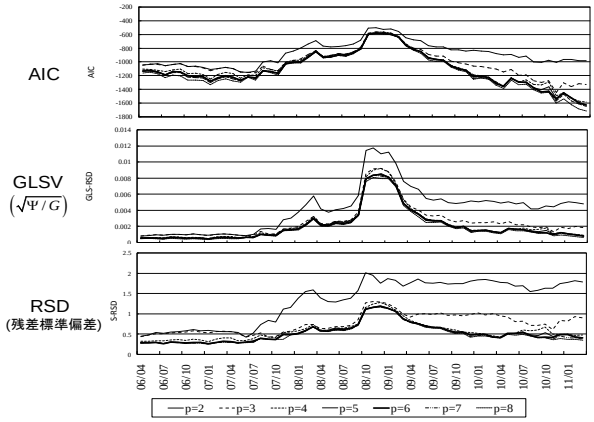
$$\psi(\underset{\sim}{\beta}, \theta, \rho) = [\underset{\sim}{y} - X\underset{\sim}{\beta}]'[\Phi(\theta, \rho, \xi)]^{-1}[\underset{\sim}{y} - X\underset{\sim}{\beta}]$$

- 満期期間の差とCF発生時点の差による効果を結合し，価格データから平均割引率関数の未知パラメータと共分散行列の未知パラメータを同時推定

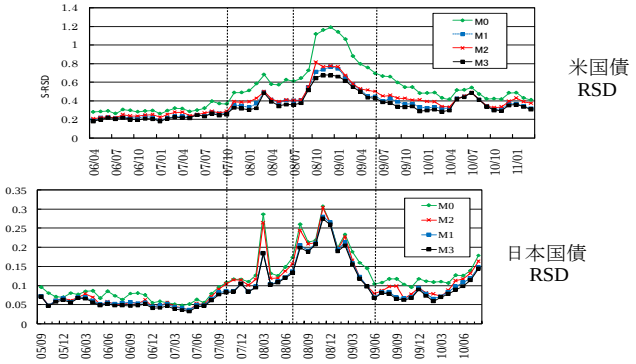
■分析期間・データ

- 金融危機を含む2006.04～2011.03まで60ヶ月
- 1年以上20年未満の満期を持つ米国財務省証券(T-Bond,T-Note)の各銘柄月末価格によるクロス・セクション分析（平均約140銘柄）

割引率関数の多項式次数pの選択：M0モデル　p=6



日米パフォーマンス比較：クーポン効果と満期効果

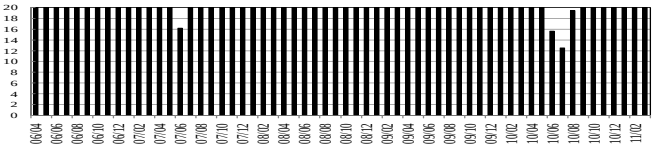


➢全体のパフォーマンスは日本国債モデル
➢属性効果については米国債

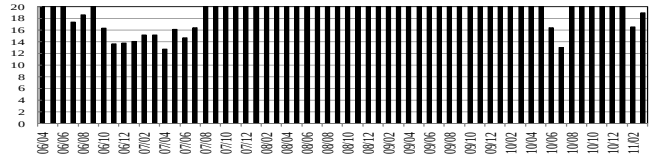
クーポン効果と満期効果の有意性

■ F ratio $=[\{\psi(0)-\psi(1)\}/\#]/[\psi(1)/df]$,
$\psi(0)$：M0の目的関数値、#：パラメータの増加数,df：M0の自由度

■満期効果：M0－M1

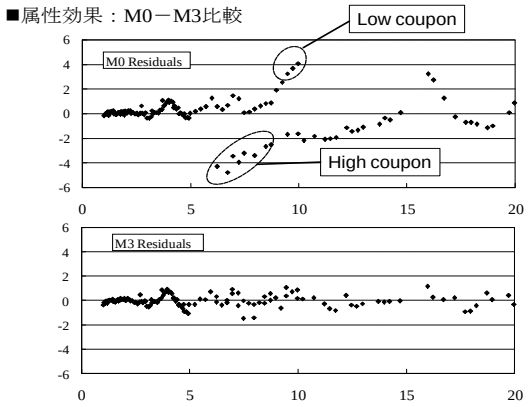


■クーポン効果：M0－M2



個別銘柄の残差：2008.11(金融危機時)

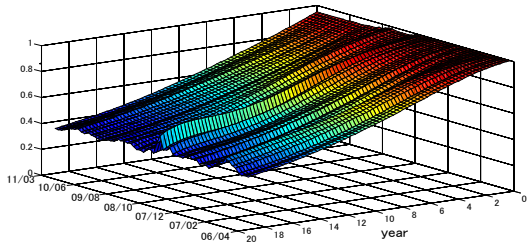
■属性効果：M0－M3比較



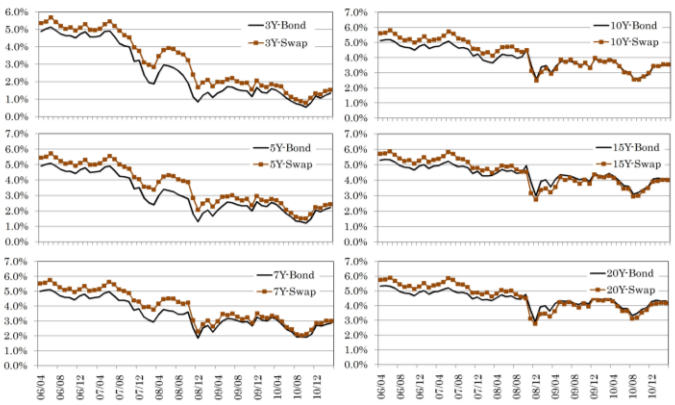
M0の割引率関数

$$R_j(s:1, z_{2j}, z_{3j}) = -\frac{1}{s}\log \bar{D}_j(s:1, z_{2j}, z_{3j})$$

$$\bar{D}_0(s) \equiv \bar{D}_0(s:1,0,0)$$



M0金利とスワップレート



要約

■米国市場においても属性依存型ﾌｫﾜｰﾄﾞﾚｰﾄ期間構造の国債価格モデルは有効

■日米の国債価格情報から導かれる属性依存型平均割引率関数の多項式構造(次数)は類似

■属性効果(クーポン効果，満期効果)は米国市場においてより顕著

(主要参考文献)
Collin-Dufresne, P. and Solnik, B. (2001) On the term structure of default premia in the swap and LIBOR markets. *Journal of Finance*, 56, 1095-1114.
Diebold, F.X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130, pp. 337–364
Heath, D., Jarrow, R.A. and Morton, A.(1992) Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica* 60, 77-105.
Kariya, T. (1993) *Quantitative Methods for Portfolio Analysis*, Springer Verlag
Kariya, T. and Tsuda, H. (1994) New bond pricing models with applications to Japanese Government Bond, *Financial Engineering and the Japanese Markets*, 1,1-20.
Kariya, T., Wang, J., Wang, A., Doi, E. and Yamamura, Y.(2012), Empirically Effective Bond Pricing Model and Analysis on Term Structures of Implied Interest Rates in Financial Crisis, *Asia-Pacific Financial Markets*, *Asia-Pacific Financial Markets*, 19, 259-292.
Nelson, C.R., Siegel, A.F. (1987). Parsimonious modeling of yield curves, *Journal of Business*, 60(4), pp.473–489
刈屋武昭他(2011)，Empirically Effective Bond Pricing Model and Analysis on Term Structures of Implied Interest Rates in Financial Crisis，統計関連学会連合大会(於九州大学)

Shrinkage Estimation of Optimal Portfolio

白石 博（東京慈恵会医科大学）

時刻 $t \in \mathbb{N}$ における m 個の資産のランダムなリターンを $\boldsymbol{X}(t) = (X_1(t), \ldots, X_m(t))'$ とし、定常性を仮定して、その期待値、分散共分散行列および自己共分散行列をそれぞれ $E\{\boldsymbol{X}(t)\} = \boldsymbol{\mu}$、$\mathrm{Var}\{\boldsymbol{X}(t)\} = \Sigma$ および $E[\{\boldsymbol{X}(t) - \boldsymbol{\mu}\}\{\boldsymbol{X}(t+k) - \boldsymbol{\mu}\}'] = R(k)$ とする。また、ポートフォリオ係数を $\boldsymbol{w} = (w_1, \ldots, w_m)$ で表す。このとき、過去、種々の基準から平均-分散最適ポートフォリオ係数が提案されたが、これらは統一的な形 $g(\theta)$ で表される。ここに、$\theta = (\boldsymbol{\mu}', vec(\Sigma)')' \in \Theta \subset \mathbb{R}^{m+m^2}$ とし、$g : \mathbb{R}^{m+m^2} \to \mathbb{R}^{m-1}$ とする。従来、最適ポートフォリオ係数 $g(\theta)$ の推定量は、パラメータ θ の推定量 $\hat{\theta}$ を plug-in した推定量 $g(\hat{\theta})$ が広く利用されている。ここに

$$\hat{\theta} = (\hat{\boldsymbol{\mu}}', vec(\hat{\Sigma})')', \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}(t), \quad \hat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} \{\boldsymbol{X}(t) - \hat{\boldsymbol{\mu}}\}\{\boldsymbol{X}(t) - \hat{\boldsymbol{\mu}}\}'$$

とする。本報告では、James and Stein (1961) らによって提案された James-Stein タイプの縮小推定量を考える。まず、$\boldsymbol{Y}(t) = \boldsymbol{X}(t) - \boldsymbol{\mu}$、$\boldsymbol{Z}(t) = vec\{\boldsymbol{Y}(t)\boldsymbol{Y}(t)' - \Sigma\}$ とし、次の仮定をする。

仮定 1　すべての $i, j, i_1, i_2, j_1, j_2 = 1, \ldots, m$ に対し

$$\sum_{k=-\infty}^{\infty} |k|^{\frac{1}{2}} |R_{ij}(k)| < \infty, \quad \sum_{k=-\infty}^{\infty} |k|^{\frac{1}{2}} |R_{i_1 j_1}(k) R_{i_2 j_2}(k) + R_{i_1 j_2}(k) R_{i_2 j_1}(k)| < \infty$$

このとき、2 つの確率過程 $\{\boldsymbol{Y}(t)\}$、$\{\boldsymbol{Z}(t)\}$ は 2 次定常過程となり、スペクトル密度行列はそれぞれ

$$f^Y(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} R(k) e^{-ik\lambda}, \quad f^Z(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} R^Z(k) e^{-ik\lambda}$$

で与えられる。また、この Cesaro sum approximations を次式で定義する。

$$f_n^Y(\lambda) = \frac{1}{2\pi}\sum_{k=-n+1}^{n-1} \left(1 - \frac{|k|}{n}\right) R(k) e^{-ik\lambda}, \quad f_n^Z(\lambda) = \frac{1}{2\pi}\sum_{k=-n+1}^{n-1} \left(1 - \frac{|k|}{n}\right) R^Z(k) e^{-ik\lambda}$$

このとき、仮定 1 の下で次の命題が得られる。

命題 1

(i) $E(\hat{\theta} - \theta) = -\frac{1}{n}\boldsymbol{a}_n$

(ii) $E\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)'\} = \frac{1}{n}B_n + \frac{1}{n^2}C_n$

(iii) $E(\|\hat{\theta} - \theta\|^r) = O\left(n^{-r/2}\right)$ for $r \geq 3$

ここに、$\boldsymbol{a}_n = (\boldsymbol{0}', vec\{2\pi f_n^Y(0)\}')'$、$B_n = \begin{pmatrix} 2\pi f_n^Y(0) & \boldsymbol{0} \\ \boldsymbol{0} & 2\pi f_n^Z(0) \end{pmatrix}$, $C_n = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 4\pi^2 f_n^{YY}(0) - 4\pi f_n^Z(0) \end{pmatrix}$
とする。また、最適ポートフォリオ係数を表す関数 $g : \theta \to \mathbb{R}^{m-1}$ に対し、次の仮定をする。

仮定 2　任意の $\boldsymbol{x} \in \Theta$ に対し、$g(\boldsymbol{x})$ は、3 回連続微分可能で、かつ任意の $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Theta$ に対し、$\|g(\boldsymbol{x}_1) - g(\boldsymbol{x}_2) \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ が成立するような定数 $L > 0$ が存在する。

また、最適ポートフォリオ係数の James-Stein タイプの推定量 $g^S(\hat{\theta})$ を次の通り定義する。

$$g^S(\hat{\theta}) = \left(1 - \frac{m-3}{n\left\|g_n(\hat{\theta})\right\|^2}\right) g(\hat{\theta})$$

ここに、$0 < \alpha \leq 1/16$ に対し

$$\|g_n(\boldsymbol{x})\| := \begin{cases} \|g(\boldsymbol{x})\| & if\ \|g(\boldsymbol{x})\| > n^{-\alpha} \\ n^{-\alpha} & otherwise \end{cases}$$

とする。

本報告では、Taniguchi and Hirukawa (2005) と同様にして、縮小推定量 $g^S(\hat{\theta})$ のが $g(\hat{\theta})$ よりも平均 2 乗誤差の意味で改良される十分条件を得る。2 つの推定量の平均 2 乗誤差の差を次式で定義する。

$$DMSE_n := E\left\{n\|g(\hat{\theta}) - g(\theta)\|^2\right\} - E\left\{n\|g^S(\hat{\theta}) - g(\theta)\|^2\right\}.$$

このとき、仮定 1 および仮定 2 の下で次の結果を得る。

定理

$$DMSE_n = \frac{1}{n}\Delta_n(m, \theta) + o\left(n^{-1}\right)$$

ここに

$$\Delta_n(m, \theta) = \frac{2(m-3)}{\|g(\theta)\|^2}\left\{tr(C_n) - \frac{2g(\theta)'C_n g(\theta)}{\|g(\theta)\|^2} - \frac{m-3}{2} + \boldsymbol{h}_n' g(\theta)\right\}$$

ただし

$$\boldsymbol{h}_n = -Dg(\theta)\boldsymbol{a}_n + \frac{1}{2}D^2 g(\theta) vec(B_n)$$

とする。

References

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379, Berkeley, Calif. Univ. California Press.

Taniguchi, M. and Hirukawa, J. (2005). The Stein-James estimator for short- and long-memory Gaussian processes. *Biometrika*, 92(3):737–746.

Analysis of CL and Estimating Function Estimators for Financial Time Series Models

Tomoyuki Amano (Wakayama University)

Many financial time series models have been proposed in order to represent behaviours of time series data in finance and many reserchers have investigated these models. One of the most famous financial time series models is ARCH model, which was introduced by Engle (1982). Another famous financial time series model is GARCH model, which is the extension of ARCH model and was introduced by Bollerslev (1986). Since then, a great number of theoretical and empirical studies have been conducted for them. Recently, a class of ARCH(∞) models was introduced, which includes ARCH and GARCH models as special cases. Lee and Taniguchi (2005) established the local asymptotic normality (LAN) of this model. Recently, CHARN model was proposed by Härdle and Tsybakov (1997) and Härdle et. al. (1998), which includes many financial time series models and is widely used in finance. Kato et. al. (2006) derived LAN of this model.

For financial time series models, one of the most famous and fundamental estimator is the conditional least squares estimator (CL estimator), which was proposed by Tjϕstheim (1986). CL estimator has two advantages, it can be calculated easily and it does not need the knowledge of the innovation. Hence this convenient estimator has been widely used. However Amano and Taniguchi (2008) showed the condition that CL estimator is asymptotically optimal is strict.

Another estimator for financial time series model is the estimating function estimator. The estimating function estimator was introduced by Godambe (1960, 1985) and Hansen (1982). Chandra and Taniguchi (2001) constructed the optimal estimating function estimator (G estimator) for ARCH and the random coefficient autoregressive (RCA) models based on the optimal estimating function of Godambe and showed G estimator is better than CL estimator in the sense of the sample mean squared error by simulation. Furthermore Amano (2009) applied G estimator to GARCH, RCA and Nonlinear AR models and investigated behaviours of G estimator. In Amano (2009), it is shown that G estimator is better than CL estimator in the sense of the efficiency theoretically. Amano (2009) also derived conditions that G estimator is asymptotically optimal based on LAN for these models. Since these conditions are general and natural, G estimator is a good estimator for financial time series models. Recently Kanai et. al. (2010) applied G estimator to CHARN model and showed the consistency of this estimator.

In this talk, we review results of CL and G estimators for ARCH, GARCH and CHARN models. First, we give the definitions of CL and G estimators. Next, the results of CL and G estimators for ARCH model are reviewed. Then, we review the results of CL and G estimators for GARCH model. Finally, we review the results of CL and G estimators for CHARN model.

References

[1] Amano, T. (2009). Asymptotic efficiency of estimating function estimators for nonlinear time series models, Journal of The Japan Statistical Society, 39, 209-231.

[2] Amano, T. and Taniguchi, M. (2008). Asymptotic efficiency of conditional least squares estimators for ARCH models, Statist. Probab. Lett., 78, 179-185.

[3] Bollerslev, T., (1986). Generalized autoregressive conditional heteroskedasticity, J. Econometrics. 31, 307–327.

[4] Chandra, S. A. and Taniguchi, M. (2001). Estimating functions for nonlinear time series models. Nonlinear non-Gaussian models and related filtering methods, Ann. Inst. Statist. Math., 53, 125-141.

[5] Engle, R., (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica. 50, 987–1007.

[6] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation, Ann. Math. Statist., 31, 1208-1211.

[7] Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes, Biometrika, 72, 419-428.

[8] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators, Econometrica, 50, 1029-1054.

[9] Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression, J. Econometrics, 81, 223-242.

[10] Härdle, W., Tsybakov, A. and Yang, L. (1998). Nonparametric vector autoregression, J. Statist. Plann. Inference, 68, 221-245.

[11] Kanai, H., Ogata, H. and Taniguchi, M. (2010). Estimating function approach for CHARN models, Metron, 68, 1-21.

[12] Kato, H., Taniguchi, M. and Honda, M. (2006). Statistical Analysis for Multiplicatively Modulated Nonlinear Autoregressive Model and Its Applications to electrophysiological signal analysis in humans. *IEEE Transactions on Signal Processing.* **54**, 3414-3425.

[13] Lee, S., Taniguchi, M., (2005). Asymptotic theory for ARCH-SM models: LAN and residual empirical processes, Statist. Sinica. 15, 215–234.

[14] Tjϕstheim, D. (1986). Estimation in nonlinear time series models, Stochastic Process. Appl., 21, 251-273.

経済調査における売上高の欠測値補定方法について
～多重代入法による精度の評価～

髙橋　将宜[†]、伊藤　孝之[††]

1. はじめに

日本の全事業所・企業を対象として経理項目などを調査する経済センサス-活動調査が、平成 24 年 2 月に初めて実施された。経理項目が欠測した場合に備え、多重代入法(Multiple Imputation)の研究を行った。本研究では、EDINET 情報を模擬データとして使用し、個別データの補定方法として多重代入法を評価した。様々な欠測値対処法とその限界を示し、代替法として多重代入法を導入し、フリーソフトウェア R の汎用多重代入法パッケージ Amelia を利用して、多重代入法による欠測値補定の精度評価を行った。

2. モデルとアルゴリズム

Dを$n \times p$データセットとする（nは標本サイズ、pは変数の数）。Dは、平均値ベクトルμと分散共分散行列Σで多変量正規分布しているとし、観測値 i 及び変数 j の値が欠測しているとする。$\widetilde{D}_{ij}$を観測値 i 及び変数 j のシミュレーション値とし、$D_{i,-j}$を、変数 j を除く i 行のすべての観測値とする。補定値$\widetilde{D}_{ij}$は、式(1)により算出する。

$$\widetilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \qquad (1)$$

回帰係数を算出するために必要な情報は、平均値及び分散・共分散の情報であり、これらの情報はμとΣにすべて含まれているが、Dには欠測値があるため、μとΣを完全に知ることができない。事後分布からμとΣの無作為抽出を行う手段として、Expectation Maximization with Bootstrapping (EMB)アルゴリズムを用いる。図 2.1 に示すとおり、欠測値のある不完全データをもとに、ブートストラップにより、標本サイズ n の副標本データを M 個、作成する。これら副標本に EM アルゴリズムを適用し、M 個のμとΣの点推定値を算出し、M 個の式(1)を算出して補定を行う。

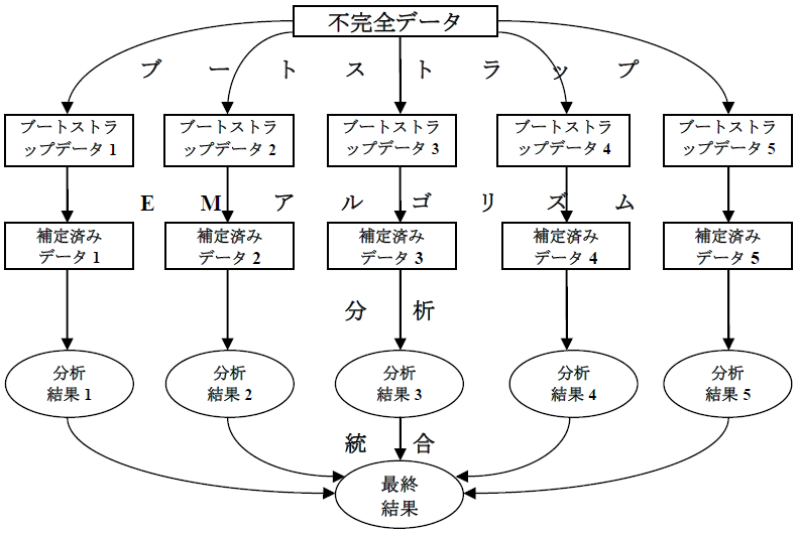
μとΣの要素は$p(p+3)/2$個あり、変数の数 p に応じて急速に増大する。様々な多重代入法プログラムで採用されている既存のアルゴリズムでは、巨大データセットの多重代入を扱えない。一方、EM アルゴリズムにブートストラップを応用する Amelia では、32,000 観測値、240 変数の実データセットの補定を行える。総パラメータ数は 29,160 個であり、4 億 2516 万 7380 個の個別要素を含む 29160×29160 共分散行列を反転できる。このアルゴリズムで扱える限界サイズは、利用可能なメモリーサイズにのみ依拠している。

---

† 独立行政法人統計センター情報技術部統計技術研究課上級研究員

†† 独立行政法人統計センター情報技術部統計技術研究課研究員

図 2.1：EMB アルゴリズムを用いた多重代入法の概念図



3. EDINET 売上高の多重代入法による補定の検証結果

本研究では、EDINET データを利用し、多重代入値と単一代入値を、それぞれ、売上高の真値と比較した。また、補定の精度評価方法として、散布図による視覚的アプローチと欠測値補定データの標準偏差を使用した。全体的に、EDINET データでは、多重代入法の方が、単一代入法と比べて、6 割強の割合で真値に近く、標準偏差の推定においても優れており、分布の再現性も高いことが分かった。

4. 多重代入法の精度評価

原則として、追加情報を収集しない限り、欠測の前提を検証することはできない。このため、長らく先行研究では、補定の精度検証法というものは見過ごされてきた。しかし、たとえ直接的な検証はできないとしても、欠測の前提及び補定モデルの間接的な検証を行うことは可能である。Amelia では、密度の比較(Comparing Densities)機能、過剰補定(Overimpute)機能、過散布初期値(Overdispersed Starting Values)機能、欠測地図(Missingness Map)機能を利用することで、実務的な補定の精度評価を行える。 本研究では、これらの補定の診断手法を駆使することにより、真の欠測メカニズムを効率よく推定できることが確認できた。

参考文献

1. Abayomi, Kobi, Andrew Gelman, and Marc Levy. (2008). “Diagnostics for Multivariate Imputations,” *Applied Statistics* vol.57, no.3: 273-291.
2. Honaker, James and Gary King. (2010). “What to do About Missing Values in Time Series Cross-Section Data,” *American Journal of Political Science* vol.54, no.2: 561–581.
3. Honaker, James, Gary King, and Matthew Blackwell. (2011). “Amelia II: A Program for Missing Data,” *Journal of Statistical Software* vol.45, no.7.
4. Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

High-Dimensional Mean Estimation via L1-Penalized Normal Likelihood

大阪大学大学院基礎工学研究科　片山 翔太[1]

2つの p 変量正規母集団 $N_p(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}^*)$ $(i=1,2), \boldsymbol{\Sigma}^* > \boldsymbol{O}$ からそれぞれ独立に n_i $(i=1,2)$ 個の独立サンプル $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ $(i=1,2; j=1,\ldots,n_i)$ が得られたとする．本研究では，高次元データ (変数の次元 p が標本サイズ n_i と同程度または大きなデータ) における，2つの母平均ベクトルの差 $\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ に関する統計的推測問題を扱う．高次元データにおける $\boldsymbol{\delta}^*$ の仮説検定問題 $H_0 : \boldsymbol{\delta}^* = \boldsymbol{0}$ v.s. $H_1 : \boldsymbol{\delta}^* \neq \boldsymbol{0}$ に関しては，Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010) などによって，高次元データに対しても検出力の高い検定方法がこれまでに様々提案されている．そこで本報告では，帰無仮説 H_0 が棄却された後の問題，すなわち，$\boldsymbol{\delta}^*$ のどの要素が 0 ではなくかつその値はどの程度かを推定する問題を考える．

近年，回帰分析モデルにおいて Lasso (Tibshirani, 1996) とよばれる，変数選択と回帰係数の推定を同時に行うことが可能な方法が提案されており，高次元データにおいて非常に良い性質を持っていることが示されている (例えば，Meinshausen and Yu, 2009)．この方法に基づくと，母共分散行列 $\boldsymbol{\Sigma}^*$ が既知ならば，$\boldsymbol{\delta}^*$ のひとつの推定量は次の目的関数を $\boldsymbol{\delta} \in \mathbb{R}^p$ に関して最小化することによって得られる:

$$Q(\boldsymbol{\delta}; \boldsymbol{\Omega}^*) = \boldsymbol{\delta}^T \boldsymbol{\Omega}^* \boldsymbol{\delta} - 2\boldsymbol{\delta}^T \boldsymbol{\Omega}^* \bar{\boldsymbol{\delta}} + 2\tau_{n,p} \|\boldsymbol{\delta}\|_1. \tag{0.1}$$

ここに，$\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{*-1}$, $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ ($\bar{\boldsymbol{X}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$ は標本平均) であり，$\|\cdot\|_1$ は ℓ_1 ノルム，$\tau_{n,p}$ はあるチューニングパラメータである．すなわち，$\tau_{n,p}$ が大きければ推定量は $\boldsymbol{0}$ に縮小され，$\tau_{n,p}$ が小さければ推定量は $\bar{\boldsymbol{\delta}}$ とほぼ同じ挙動を示す．しかしながら，ほとんどの実データ解析において $\boldsymbol{\Sigma}^*$ は未知である．その上高次元データの場合，$\boldsymbol{\Sigma}^*$ の典型的な推定量である標本共分散行列は，逆行列が定義できない，もしくは固有値の挙動が不安定になるために用いることができない．そこで本報告では，近年 Cai and Liu (2011) によって提案された，高次元データに対しても非常に良い性質を持つ $\boldsymbol{\Sigma}^*$ の推定量 $\widehat{\boldsymbol{\Sigma}}_T$ を用いる．Cai and Liu (2011) の推定量は常に逆行列が定義できるわけではないので，適当な正則化項 $a = |\lambda_{min}(\widehat{\boldsymbol{\Sigma}}_T)| + \{(\log p)/n\}^{1/2}$ を用いて $\widehat{\boldsymbol{\Omega}}_{T,a} = (\widehat{\boldsymbol{\Sigma}}_T + a\boldsymbol{I}_p)^{-1}$ を構成し，それを $Q(\boldsymbol{\delta}; \boldsymbol{\Omega}^*)$ に Plug-In して得られる次の推定量を提案する：

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta} \in \mathbb{R}^p}{\operatorname{argmin}}\, Q(\boldsymbol{\delta}; \widehat{\boldsymbol{\Omega}}_{T,a}).$$

このようにして得られる $\boldsymbol{\delta}^*$ の推定量 $\hat{\boldsymbol{\delta}}$ は，ある適当な仮定の下で，Sign Recovery を持つ．すなわち，

$$P\{\operatorname{sgn}(\hat{\boldsymbol{\delta}}) = \operatorname{sgn}(\boldsymbol{\delta}^*)\} \to 1, \quad (n_i, p) \to \infty$$

[1]日本学術振興会特別研究員 (DC1)

が成り立つ．これは，$\boldsymbol{\delta}^*$ の 0 である要素を漸近的に正確に 0 と推定できることを意味している．一方推定精度に関しては，平均二乗誤差が次のように与えられる：

$$E(\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2^2) = \frac{1}{N}\left\{\mathrm{tr}\boldsymbol{\Sigma}^*_{11.2} + \alpha(\log p)\|\boldsymbol{\Sigma}^*_{11.2}\,\mathrm{sgn}(\boldsymbol{\delta}^*)\|_2^2\right\} + o(1).$$

ここに $\boldsymbol{\Sigma}^*_{11.2}$ は，母平均ベクトルの分割 $\boldsymbol{\delta}^* = (\boldsymbol{\delta}_1^{*T}, \boldsymbol{\delta}_2^{*T})^T$, $\boldsymbol{\delta}_1^* \neq \mathbf{0}_s$ $(elementwize)$, $\boldsymbol{\delta}_2^* = \mathbf{0}_{p-s}$ に対応する母共分散行列 $\boldsymbol{\Sigma}^*$ の分割 $(\boldsymbol{\Sigma}^*_{ij})$ を用いて $\boldsymbol{\Sigma}^*_{11.2} = \boldsymbol{\Sigma}^*_{11} - \boldsymbol{\Sigma}^*_{12}\boldsymbol{\Sigma}^{*-1}_{22}\boldsymbol{\Sigma}^*_{21}$ で与えられる．また，$\alpha > 0$ は $\tau_{n,p} = \alpha\{(\log p)/N\}^{1/2}$, $N = n_1 n_2/(n_1 + n_2)$ で与えられる．

記法：本報告で用いた記法をここに要約しておく．あるベクトル $\boldsymbol{a} = (a_i)$ に対して，ℓ_q ノルム $(\sum_i |a_i|^q)^{1/q}, 1 \leq q < \infty$ を $\|\boldsymbol{a}\|_q$ で表す．また，ベクトル $\boldsymbol{a}$ に対する符号関数を $\mathrm{sgn}(\boldsymbol{a}) = (\mathrm{sgn}(a_i))$ で定義する．ここに，$\mathrm{sgn}(a_i)$ は $a_i > 0$ のとき 1, $a_i = 0$ のとき 0, $a_i < 0$ のとき -1 を返す関数である．ある行列 $\boldsymbol{A} = (a_{ij})$ に対して，その最小固有値と最大固有値をそれぞれ $\lambda_{min}(\boldsymbol{A})$, $\lambda_{max}(\boldsymbol{A})$ で表す．また，行列 $\boldsymbol{A}$ の作用素ノルム $\lambda^{1/2}_{max}(\boldsymbol{A}^T\boldsymbol{A})$ を $\|\boldsymbol{A}\|$ で表し，無限大ノルム $\max_i \sum_j |a_{ij}|$ を $\|\boldsymbol{A}\|_\infty$ で表す．明らかに，対称行列 $\boldsymbol{A}$ に対しては $\|\boldsymbol{A}\| = |\lambda_{max}(\boldsymbol{A})|$ である．

参考文献

[1] Bai, Z and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6: 311–329.

[2] Cai, T and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Society*. 106: 672–684.

[3] Chen, S. X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*. 38: 808–835.

[4] Meinshausen, N., and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*. 37: 246–270.

[5] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Jounal of Multivariate Analysis*. 99: 386–402.

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*. 58: 267–288.

LARS に基づく L_1 正則化法におけるモデル選択基準の構成

保科　架風[*]　廣瀬　慧[†]　小西　貞則[‡]

回帰モデリングにおいてモデル選択は, データの背景にある現象構造の解明や推定されたモデルの予測精度の向上において重要な問題であり, これまで AIC (Akaike, 1973) や BIC (Schwarz, 1978), Mallows の C_p (Mallows, 1973), 一般化クロスバリデーション (GCV; Craven and Wahba, 1979) などをモデル選択基準とした変数選択 (best subset selection) によるモデル選択が行われてきた. しかし, この手法はデータの次元が高くなるにつれ計算コストが莫大になることが知られており, また,「最適な変数の組み合わせの探索」という離散的なモデル選択プロセスに対し, モデル選択の不安定性やそれに起因する予測精度の悪化が指摘されている (Breiman, 1996). これに対し, lasso (Tibshirani, 1996) に代表されるスパース回帰モデリング手法では, モデルの推定やモデルに含む変数の決定などの一連のモデリングプロセスが調整パラメータによって決定されるため, 連続的なモデル選択が可能である. このスパース回帰モデリングは, 回帰係数を縮小推定することで予測精度を向上させ, また, 一部の回帰係数を厳密に 0 と推定する性質 (スパース性) によって変数選択と推定の同時性を有するという特徴が存在する. なお, lasso や elastic net (Zou and Hastie, 2005) などの L_1 タイプのノルムを正則化項として持つスパース回帰モデリング手法 (これらを "L_1 正則化法" と呼ぶ) では L_1 ノルムが 0 点で微分不可能であるため, 解析的に陽な形で推定量を表現することが困難であり, 数値的に推定値を求める必要がある. これに対し, generalized path seeking (GPS; Friedman, 2008), や coordinate descent (Friedman *et al.*, 2010) などの効果的な推定アルゴリズムが提案されている.

モデリングプロセス全体を調整パラメータによってコントロールするスパース回帰モデリングでは, 調整パラメータの最適化がモデル選択問題に対応する. これに対し,「モデルの複雑さ」を評価する「モデルの有効自由度 (Degrees of freedom)」(Ye (1998), Efron (1986, 2004)) を考慮したモデル選択基準をもとにこの最適化を行うことが考えられるが, 推定量の解析解を持たない L_1 正則化の有効自由度を解析的に陽に表すことは困難である. これに対し, Efron (2004) ではブートストラップ法によってモデルの有効自由度の不偏推定量が得られること示しているが, これは計算コストの大きさや推定したモデルの変動が問題となる. また, Zou *et al.* (2007) や Tibshirani and Taylor (2011, 2012) では lasso やその拡張された手法における有効自由度の不偏推定量を導出しているが, これらは特定の手法に関して求められた推定量である.

これに対し, 本報告では L_1 正則化法の推定アルゴリズムとして良く知られる LARS アルゴリズム (Efron *et al.*, 2004) において, その特性をうまく活かすことでモデルの有効自由度が求まることを示し, これにより種々の L_1 正則化法におけるモデルの有効自由度が得られることを報告した. また, この手法によって得られたモデルの有効自由度を基に構成されたモデル選択基準によるモデル選択の精度の検証を, 数値実験と実データへの適用によって行った.

参考文献

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B.N. Petrov and F. Csáki, eds.) 267-281.

[*] 中央大学大学院理工学研究科数学専攻 112-8551 東京都文京区春日 1-13-27 Email:ibu@gug.math.chuo-u.ac.jp
[†] 大阪大学大学院基礎工学研究科 560-8531 大阪府豊中市待兼山町 1 − 3 Email:hirose@sigmath.es.osaka-u.ac.jp
[‡] 中央大学理工学部 112-8551 東京都文京区春日 1-13-27 Email:konishi@math.chuo-u.ac.jp

Académian Kaidó, Budapest.

[2] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350-2383.

[3] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik.* **31**, 377-403.

[4] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

[5] Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *J. Amer. Statist. Assoc.* **99**, 619-642.

[6] Efron, B. , Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-499.

[7] Friedman, J. (2008). Fast sparse regression and classification. Technical report, Dept. of Statistics, Stanford University.

[8] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statit. Software.* **33**,

[9] Mallows, C. (1973). Some comments on C_p. *Technometrics.* **15**, 661-675.

[10] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Ser B.* **58**, 267-288.

[12] Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39**, 1335-1371.

[13] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40**, 1198-1232.

[14] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93**, 120-131.

[15] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. Ser B.* **67**, 301-320.

[16] Zou, H. Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.* **35**, 2173-2192.

L_1 型正則化法による因子分析モデルのスパース推定

廣瀬 慧, 山本 倫生
大阪大学 大学院基礎工学研究科

1 概要

因子分析モデルとは，多変量データの相関構造から，背後に存在する共通因子を見出すモデルであり，心理学，社会科学，生命科学をはじめとする諸科学の様々な分野で応用されている．観測データと共通因子を結ぶ因子負荷量は，通常，次の 2 段階推定によって求められる: (i) 因子分析モデルを最尤法によって推定する，(ii) バリマックス回転などの因子回転を用いて解釈しやすい因子を見つけ出す．“解釈のしやすさ”といっても様々な指標が考えられるが，本稿では因子負荷行列をできるだけスパースに推定することを考える．ここで，上記の 2 段階推定による因子負荷行列の推定法にはいくつか問題点がある．まず，因子分析モデルはパラメータ数が膨大となり，最尤法ではしばしば推定が不安定となる (Akaike, 1987)．特に，変数の数がサンプルサイズより大きい場合，最尤推定値を求めることはできない．また，たとえ最尤推定値を数値的に求めることができたとしても，因子回転では十分にスパースな推定を行うことができないことが多い．これらの問題に対処するために，L_1 型正則化法によって因子分析モデルを推定する．

Lasso (Tibshirani, 1996) をはじめとする L_1 型正則化法は，線形回帰モデル，グラフィカルモデル，サポートベクターマシンなどの様々なモデルに適用できるスパース推定法として広く用いられている．しかしながら，因子分析モデルにおける L_1 型正則化法はほとんど研究されていない．Ning and Georgiou (2011) や Choi et al. (2011) は，因子負荷行列に Lasso タイプの正則化推定法を適用し，先程述べた 2 段階推定より安定した推定を行うことができることを示した．しかしながら，罰則付き最尤推定法と古典的な 2 段階推定法との関係性は議論されていない．さらに，Lasso は過度に密なモデルを推定する傾向にあることが知られており，よりスパースな解を求める罰則項を用いる必要がある．

本稿では，因子分析モデルにおいて，非凸ペナルティに基づく新しい罰則付き最尤推定法を提案する．また，罰則付き最尤推定法が古典的な 2 段階推定の一般化と見なすことができることを示す．具体的には，因子負荷行列が回転の不定性を除いて一意に存在し，かつ Solution path が最尤推定値付近で連続であった場合，2 段階推定法によって得られた解は罰則付き最尤推定法によって計算することができることを示す．提案手法はチューニングパラメータの値を変えることによって，因子回転よりもスパースな解を求めることができる．さらに，全てのチューニングパラメータに対する解を効率的に求めるために，EM アルゴリズム (Rubin and Thayer, 1982) と Coordinate descent アルゴリズム (Mazumder et al., 2011) を組み合わせた新しいアルゴリズムを提案する．提案するアルゴリズムは，Lasso, SCAD, MC+ ペナルティを含む極めて広いクラスのペナルティに対して適用可能である．提案手法を遺伝子発現データに適用し，その有用性を検証する．

2　L_1 型正則化法による因子分析モデルのスパース推定

p 次元観測変数を $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ とし，その平均ベクトルと分散共分散行列をそれぞれ $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ とする．このとき，因子分析モデルは次で与えられる．

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon}$$

ただし，$\boldsymbol{\Lambda} = (\lambda_{ij})$ は $p \times m$ 因子負荷行列，$\boldsymbol{F} = (F_1, \cdots, F_m)^T$ は m 次元共通因子ベクトル，$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)^T$ は p 次元独自因子ベクトルとする．共通因子ベクトル $\boldsymbol{F}$ と独自因子ベクトル $\boldsymbol{\varepsilon}$ はそれぞれ独立に多変量正規分布に従うと仮定し，それらの平均ベクトルと共分散行列は $E(\boldsymbol{F}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \boldsymbol{I}_m$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi}$ で与えられるとする．ここで，$\boldsymbol{I}_m$ は $m \times m$ 単位行列，$\boldsymbol{\Psi}$ は $p \times p$ 対角行列である．$\boldsymbol{\Psi}$ の第 i 対角成分は独自分散と呼ばれ，ψ_i で与えられるとする．これらの仮定により，観測変数ベクトル $\boldsymbol{X}$ は，平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$ を持つ多変量正規分布に従う．

N 個の p 次元データ $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ が $N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$ から発生した時，モデルパラメータ $\boldsymbol{\Lambda}, \boldsymbol{\Psi}$ を lasso タイプの罰則付き最尤法によって推定する．ここで，罰則付き対数尤度関数は

$$\ell_\rho(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \ell(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) - N \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|)$$

で与えられる．ただし，$\ell(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ は対数尤度関数，$P(\cdot)$ は罰則項，$\rho > 0$ は正則化パラメータとする．

参考文献

Akaike, H. (1987), "Factor analysis and AIC," *Psychometrika*, 52(3), 317–332.

Choi, J., Zou, H., and Oehlert, G. (2011), "A Penalized Maximum Likelihood Approach to Sparse Factor Analysis," *Statistics and Its Interface*, 3(4), 429–436.

Kaiser, H. F. (1958), "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23(3), 187–200.

Mazumder, R., Friedman, J., and Hastie, T. (2011), "SparseNet: Coordinate Descent with Nonconvex Penalties," *Journal of the American Statistical Association*, 106, 1125–1138.

Ning, L., and Georgiou, T. T. (2011), Sparse factor analysis via likelihood and ℓ_1 regularization,, in *50th IEEE Conference on Decision and Control and European Control Conference*, pp. 5188–5192.

Rubin, D. B., and Thayer, D. T. (1982), "EM algorithms for ML factor analysis," *Psychometrika*, 47(1), 69–76.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

高次元小標本における幾何学的表現とその応用

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）

1　はじめに

マイクロアレイデータや MRI データに見られるように，情報化の進展に伴い，データの次元数 p が標本数 n よりも遥かに大きな高次元小標本データの統計解析が益々重要になってきている．この高次元小標本の特徴的な研究として，高次元小標本データ空間を幾何学的に捉えるための研究がある．Ahn et al. (2007)，Hall et al. (2005)，Yata and Aoshima (2012a) は，標本数 n を高々100 程度に固定して次元数 p を $p \to \infty$ としたときの高次元データ空間の幾何学的表現を見つけている．特に，Yata and Aoshima (2012a) は高次元小標本データの 2 つの幾何学的表現を発見し，さらに，その幾何学的表現に基づいて‘ ノイズ掃き出し法 ’とよばれる固有空間のセミパラメトリックな推定法を考案した．

本発表では，まず，Yata and Aoshima (2012a) が明らかにした幾何学的表現である球面集中現象と座標軸集中現象の境界条件を具体的に提示した．それに基づき，ノイズ掃き出し法を精密に評価した．次に，高次元データ空間における PCA について一般化したモデル設定のもので推定論を展開し，高次元空間に占めるノイズの相対量と標本数の関係に着目して，固有値・主成分スコアの推定に一致性を導いた．

2　高次元小標本における固有空間の幾何学的表現

平均に p 次の $\mathbf{0}$ ベクトル，共分散行列に p 次の非負定値対称行列 $\boldsymbol{\Sigma}$ $(\geq \boldsymbol{O})$ をもつ母集団を考える．n 個の p 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出して，データ行列 $\boldsymbol{X} : p \times n = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を定義する．ただし，$p > n$ である．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ で $\boldsymbol{\Sigma}$ を $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ と分解する．そのとき $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}$ とおき，$\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p]^T$, $\boldsymbol{z}_i = (z_{i1}, ..., z_{in})^T$ と表記する．ここで，$\boldsymbol{Z}$ の成分は，4 次モーメントが一様有界と仮定する．高次元小標本データを解析する上で鍵となるのは，データがもつ特有の幾何学的表現といえる．いま，Dual な標本共分散行列を $\boldsymbol{S}_D = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ とする．$\boldsymbol{S}_D$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ とし，$\hat{\lambda}_j$ に対する固有ベクトルを $\hat{\boldsymbol{u}}_j$ として，スペクトル分解が $\boldsymbol{S}_D = \sum_{j=1}^{n} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$ となるとする．このとき，$p \to \infty$ と適当な正則条件のもとで，$\boldsymbol{S}_D$ が

$$\frac{n}{\sum_{i=1}^{p} \lambda_i} \boldsymbol{S}_D \xrightarrow{P} \boldsymbol{I}_n \quad \text{もしくは,} \quad \frac{n}{\sum_{i=1}^{p} \lambda_i} \boldsymbol{S}_D \xrightarrow{P} \boldsymbol{D}_n$$

なるどちらかの幾何学的表現を有することを本発表で示し，その境界条件を導出した．ここで，$\boldsymbol{D}_n$ は対角成分が $O_P(1)$ となる対角行列である．

3　高次元データの固有値推定

いま，$\boldsymbol{\Sigma} = \sum_{j=1}^{m} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T + \sum_{j=m+1}^{p} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$ という分解を考える．ここで，$\boldsymbol{\Sigma}_{(1)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$，$\boldsymbol{\Sigma}_{(2)} = \sum_{j=m+1}^{p} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$ とおく．そのとき，$\boldsymbol{\Sigma}$ の固有値に対して次のモデルを仮定する．

$$\lambda_m \text{ に対して，} \lim_{p\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{(2)}^{k_m})}{\lambda_m^{k_m}} = 0 \text{ なる (有界な) ある自然数 } k_m \text{ が存在する} \quad (1)$$

(1) を満たす $\lambda_1 \geq \cdots \geq \lambda_p$ を power spiked モデルとよび，既存のモデルの拡張となっている．

Yata and Aoshima (2012a) は，高次元小標本の幾何学的表現に着目して，ノイズ掃き出し法という方法論を提案した．それは，次のような固有値の推定に基づくものである：

$$\acute{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{i=1}^{j} \hat{\lambda}_i}{n-j} \quad (j = 1, ..., n-1) \qquad (2)$$

このとき，power spiked モデルのもと以下の結果が導かれた．

定理 (Yata and Aoshima, 2012b). $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_{(2)}^2)/\lambda_j^2 = 0$ を満たす固有値 λ_j ($j \leq m$) に対して，下記の幾何学的表現に基づく条件 (C-i) のもと，$p \to \infty,\ n \to \infty$ のとき，$\acute{\lambda}_j/\lambda_j = 1 + o_P(1)$ が成り立つ．

(C-i) $$\frac{\sum_{r,s=m+1}^{p} \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{n\lambda_j^2} = o(1)$$

推定量 $\acute{\lambda}_j$ は，ノイズを除去する効果があるので，従来の推定量 $\hat{\lambda}_j$ より緩い条件で一致性をもつことを報告し，数値的に比較した．

4　高次元データの主成分スコアの推定

ノイズ掃き出し法による主成分スコアの推定を考える．データ $\boldsymbol{x}_k$ の第 j 主成分スコアは $\boldsymbol{h}_j^T \boldsymbol{x}_k = \lambda_j^{1/2} z_{jk}$ ($= s_{jk}$ とおく) である．$\boldsymbol{S}_D$ の固有ベクトルの成分を $\hat{\boldsymbol{u}}_j = (\hat{u}_{j1}, \cdots, \hat{u}_{jn})^T$ とする．推定量 (2) に基づいて，第 j 主成分スコアを $\hat{u}_{jk}(n\acute{\lambda}_j)^{1/2}$ ($= \acute{s}_{jk}$ とおく) で推定する．この推定量が，従来型の推定量 $\hat{u}_{jk}(n\hat{\lambda}_j)^{1/2}$ ($= \hat{s}_{jk}$ とおく) に比べ，高次元の枠組みで優れていることを数値的かつ理論的に示した．さらに，ノンパラメトリックな手法であるクロスデータ行列法に基づく推定量についても，Yata and Aoshima (2012b) の結果を紹介した．

Ahn, J., Marron, J.S., Muller, K.M. and Chi, Y.-Y. (2007). Biometrika, **94**, 760-766.

Hall, P., Marron, J.S. and Neeman, A. (2005). J. Roy. Statist. Soc. Ser. B, **67**, 427-444.

Yata, K. and Aoshima, M. (2012a). J. Multivariate Anal., **105**, 193-215.

Yata, K. and Aoshima, M. (2012b). PCA consistency for power spiked model in high-dimensional settings, submitted.

Moment convergence of Z-estimators and Z-process method for change point problems

Yoichi Nishiyama

The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
nisiyama@ism.ac.jp

The talk: October 25th, 2012. This report: November 2nd, 2012

1 Moment convergence of Z-estimators

For an illustration, let us consider the simplest case of i.i.d. data. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, and let us be given a parametric family of probability densities $f(\cdot;\theta)$ with respect to μ, where $\theta \in \Theta \subset \mathbb{R}^d$. Let $X_1, X_2, ...$ be an i.i.d. sequence of $\mathcal{X}$-valued random variables from this parametric model. There are two ways to define the "maximum likelihood estimator (MLE)" in statistics. One way is to define it as the maximum point of the random function

$$\theta \mapsto \mathbb{M}_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \log f(X_k;\theta),$$

while the other is to do it as the solution to the estimating equation

$$\dot{\mathbb{M}}_n(\theta) = 0,$$

where $\dot{\mathbb{M}}_n(\theta)$ is the gradient vector of $\mathbb{M}_n(\theta)$. The former is a special case of "M-estimators", and the latter is that of "Z-estimators"; see van der Vaart and Wellner (1996) for these terminologies.

It is well known that the MLE $\widehat{\theta}_n$ is asymptotically normal: it holds for any *bounded* continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ that

$$\lim_{n\to\infty} E[\psi(\sqrt{n}(\widehat{\theta}_n - \theta_0))] = E[\psi(I(\theta_0)^{-1/2}Z)],$$

where $I(\theta_0)$ is the Fisher information matrix and Z is a standard Gaussian random vector. Furthermore, it is important for some advanced theories in statistics, including asymptotic expansions and model selections, to extend this kind of results for *bounded* continuous functions ψ to that for any continuous function ψ with

polynomial growth, that is, any continuous function ψ for which there exist some constants $C = C_\psi > 0$ and $q = q_\psi > 0$ such that

$$|\psi(x)| \leq C(1 + ||x||)^q, \quad \forall x \in \mathbb{R}^d. \tag{1}$$

See the discussion in Yoshida (2011) for the importance of this problem.

We observe that, when we have an asymptotic distribution result of an estimator, namely $R_n(\widehat{\theta}_n - \theta_0) \to^d L(\theta_0)$ where R_n is a (possibly, random) diagonal matrix and the limit random vector $L(\theta_0)$ is not necessarily Gaussian, it is sufficient for the generalization to the case where ψ is a continuous function satisfying (1) to check that $||R_n(\widehat{\theta}_n - \theta_0)||$ is *asymptotically L_p-bounded* for some $p > q$, that is,

$$\limsup_{n\to\infty} E[||R_n(\widehat{\theta}_n - \theta_0)||^p] < \infty.$$

In the first part of the talk, a set of sufficient conditions for the above claim was given in a general framework of Z-estimation.

2 Z-process method for change point problems

Introduce the partial sum process

$$\mathbb{M}_n(u, \theta) = \frac{1}{n} \sum_{k=1}^{[un]} \log f(X_k; \theta), \quad \forall u \in [0, 1],$$

and consider the gradient vectors $\dot{\mathbb{M}}_n(u, \theta)$ of $\mathbb{M}_n(u, \theta)$ with respect to θ. Let $\widehat{\theta}_n$ be the MLE for the full data $X_1, ..., X_n$ as a special case of Z-estimators, that is, $\widehat{\theta}_n$ is the solution to the estimating equation $\dot{\mathbb{M}}_n(1, \theta) = 0$.

The fact that the random process

$$u \rightsquigarrow \sqrt{n}\dot{\mathbb{M}}_n(u, \theta_0) \quad \text{converges weakly to} \quad u \rightsquigarrow I(\theta_0)^{1/2} B(u)$$

in the Skorohod space $D[0, 1]$, where $u \rightsquigarrow B(u)$ is a vector of independent standard Brownian motions, is immediate from Donsker's theorem. However, it does not seem so well known that the random process

$$u \rightsquigarrow \sqrt{n}\dot{\mathbb{M}}_n(u, \widehat{\theta}_n) \quad \text{converges weakly to} \quad u \rightsquigarrow I(\theta_0)^{1/2} B^\circ(u) \tag{2}$$

in $D[0, 1]$, where $u \rightsquigarrow B^\circ(u)$ is a vector of independent standard Brownian bridges. Horváth and Parzen (1994) is apparently the first to introduce the statistic

$$\mathcal{T}_n = n \sup_{u\in[0,1]} \dot{\mathbb{M}}_n(u, \widehat{\theta}_n)^\top \widehat{I}_n^{-1} \dot{\mathbb{M}}_n(u, \widehat{\theta}_n)$$

for change point problems, where $\widehat{I}_n$ is a consistent estimator for the Fisher Information matrix $I(\theta_0)$. It is immediate from (2) and the continuous mapping theorem that

$$\mathcal{T}_n \to^d \sup_{u\in[0,1]} ||B^\circ(u)||^2.$$

In the second part of the talk, a general theory based on this idea was presented.

「金利スプレッドによる倒産確率の推定」

------ Implied 倒産確率の推定 ----------

高橋　一（鳥取環境大学）

2012 年 10 月 25 日

ある企業の Implied Default Probability をその企業発行の社債（Risky bond)と Government bond (Risk free bond)との金利差(Yield Spread) を用い推定する方法について考察する。前半部では Duffie+Singletone(1999)によるコーポレートボンド価格式に基づいた Implied probability の統計学的な推定法を Takahashi(2011)に基づき紹介する。パラメトリックモデル、ノンパラメトリックモデル共に通常は（単純な）Calibration が用いられ、前者の場合であればパラメータが OLS 推定される。ただ、問題は Calibration なるものが本当に有効なのか？また、例え有効であったとしても推定されたパラメータが如何なる性質を持つかである。Takahashi(2011)では Implied なフレームワークの中で OLE に基づく推定量が一致性や漸近正規性を持つためにはどの様なモデルを設定しなければならないかを考察し、更に Calibration の結果と比較した。

一方、Implied な方法で推定された倒産確率は多くの場合、他の方法で推定された結果と比べ大きくなっている。これは、金利スプレッドに影響を与える重要な要素としての流動性を無視している事による。本報告の後半部では Duffie+Singleton の誘導型モデルのフレームワークの中で流動性を理論的に考察する一方法を提案する。 流動性がスプレッドに与える影響について多くの実証研究が報告されてきているが、多くは上記 Duffie+Singleton モデルのような、所謂 Reduced Model のフレームワークとは独立なモデルに寄っている。本報告では、（非）流動性を債券価格決定時におけるノイズの大きさと、流動性指標を定義することにより統一的に扱うことを提案する。

流動性が阻害される要素として本報告では Bid-ask spread と No-trading period の存在を出発点とする。これらにより取引成立までに①一定の時間が必要とされると同時に、②ノイズも発生するだろう。そこで本稿では①に関しては離散時間モデル、②については価格形成時に流動的な市場より大きな（k 倍の大きさを持つ）ボラティリティーの存在を仮定する。Δtを取引間隔とすると$\Delta t = 0$ は連続時間モデル、$\Delta t > 0$ は離散時間モデルに対応している。(勿論問題はΔtの決め方である)。ポイントは非流動的な市場における基本過程の分散を$k^{\Delta t}$とし(Ω ,F,Pr)を流動的な資産を決定する確率空間、$\left(\Omega, F, \Pr^{(k^{\Delta t})}\right)$は非流動的な資産価格を決定する確率空間とする。この時、近似式 $P^{(1)}(t,T) \cong \frac{1}{C(k,\Delta t)} P^{(k)}(t,T)$ を以下の様に導出する。繰り返しになるが、重要な点は①流動性の尺度として最小取引時間 Δt の導入と流動的資産に比べ k 倍のノイズの存在、②技術的には尤度比を用いた測度変換と大数の法則である。ただし、$P^{(1)}(t,T)$は流動的

な資産の価格、$P^{(k)}(t,T)$は流動性以外は同一な資産の価格である。まず、非流動的な債券価格式に測度変換を施すことにより、

$$P^{(k)}(t,T) = E^{(1)}\left\{\exp\left[-\sum_{i=N_t}^{N_T} R(t_i)\,\Delta t\right]\frac{dQ^{(k)}}{dQ^{(1)}}\middle|F_t\right\} \cong C(k,\Delta t)\,E^{(1)}\{\exp[-\sum_{i=N_t}^{N_T} R(t_i)\,\Delta t]|F_t\}$$

$= C(k,\Delta t)P^{(1)}(t,T)$ 、但し、 $R(t) = r(t) + (1-\delta_t)\,\lambda(t)$ $N_T = \frac{T}{\Delta t}$, $N_t = \frac{t}{\Delta t}$ を得る。

次に、簡単化のため t=0, T=t とし$P^{(1)}(0,t) \cong \frac{1}{C(k,\Delta t)}P^{(k)}(0,t)$ が導出される。ここで、簡単な例として、正規ランダムウォークモデルを仮定すると

$$P^{(k)}(0,t) = \exp\{-\sum_{i=0}^{N_t}(1-\delta)\,\lambda(t_i)\,\Delta t\}\,E^{(1)}\{\exp\{-\sum_{i=0}^{N_t} r(t_i)\,\Delta t\}\frac{dQ^{(k)}}{dQ^{(1)}}\}。$$

大数の法則を用い$\frac{dQ^{(k)}}{dQ^{(1)}}$を近似すると、

$$E^{(k)}\left\{e^{-\sum_{j=1}^{N} r(t_{j-1})\Delta t}\right\} = \left(\frac{1}{k}\right)^{\frac{\Delta t}{2}N} E^{(1)}\left\{e^{-\sum_{j=1}^{N} r(t_{j-1})\,\Delta t}\exp\left\{-\frac{1}{2\Delta t}\left(\frac{1}{k^{\Delta t}}-1\right)\sum_{j=1}^{N_t}\Delta t\varepsilon^2{}_j\right\}\right\}、$$

以上より流動的な市場における確率測度$Q^{(1)}$の基で十分大きなN_tに対し、

$$\exp\left\{-\frac{N_t}{2}\left(\frac{1}{k^{\Delta t}}-1\right)\frac{1}{N_t}\sum_{j=1}^{N_t}\varepsilon^2{}_j\right\} \cong \exp\left\{-\frac{N_t}{2}\left(\frac{1}{k^{\Delta t}}-1\right)\right\}。$$

これより、

$$P^{(k)}(0,t) = E^{(k)}\{\exp\{-\sum_{i=0}^{N_t} R(t_i)\,\Delta t\}\}$$

$$= \exp\{-\sum_{i=0}^{N_t}(1-\delta)\,\lambda(t_i)\,\Delta t\}\left(\frac{1}{k}\right)^{\frac{\Delta t}{2}N} E^{(1)}\left\{e^{-\sum_{j=1}^{N} r(t_{j-1})\,\Delta t}\exp\left\{-\frac{1}{2\Delta t}\left(\frac{1}{k^{\Delta t}}-1\right)\sum_{j=1}^{N_t}\Delta t\varepsilon^2{}_j\right\}\right\}$$

$$\cong \left(\frac{1}{k}\right)^{\frac{\Delta t}{2}N}\exp\{\frac{N_t}{2}(\frac{1}{k^{\Delta t}}-1)\}\,E^{(1)}\left\{\exp\{-\sum_{i=0}^{N_t}(1-\delta)\,\lambda(t_i)\,\Delta t\}\,e^{-\sum_{j=1}^{N} r(t_{j-1})\,\Delta t}\right\}$$

$$= \left(\frac{1}{k}\right)^{\frac{\Delta t}{2}N}\exp\{\frac{N_t}{2}(\frac{1}{k^{\Delta t}}-1)\}P^{(1)}(0,t)$$

即ち、近似公式として、

$$P^{(1)}(0,t) \cong \left(\frac{1}{k}\right)^{-\frac{\Delta t}{2}N}\exp\{-\frac{N_t}{2}(\frac{1}{k^{\Delta t}}-1)\}P^{(k)}(0,t)$$

が得られる。

最後にコメントとして、流動的な債券市場は極限として与えられる事を挙げておく；

$$\lim_{\Delta t\to 0}\left(\frac{1}{k}\right)^{-\frac{\Delta t}{2}N}\exp\{-\frac{N_t}{2}(\frac{1}{k^{\Delta t}}-1)\} = 1$$

Brownian quantiles を用いた逐次解析の試み

一橋大学名誉教授　三浦良造

数理ファイナンスの分野に於いて、Brownian quantiles と呼ばれる量がある。これは確率過程の連続な軌跡を観察データとするときの”順序統計量“に該当する。参考文献（[1]-[7],[11],[12],[14]）に見られるように、すでに確率分布論はある程度できており、数理ファイナンス分野のエキゾチックオプションの定義と価格理論に使われている（[8],[9],[13]）。この量は、確率過程に対する統計量なので、それ自身が確率過程であり逐次的な扱いも可能である。従って、ここでは数理統計学の逐次解析に用いることを試みる。今回の発表では、その試みと展望を、幾つかの推測問題の形に乗せて紹介する。それぞれまだ準備的段階ではあるが、アイデアと推論の枠組みについて述べる。すべて、逐次解析における Truncated case である([10])。

ドリフト付きのブラウン運動　$X_t=\mu \cdot t+\sigma W_t$　を考える。ここでσ は既知としておく。t は（0、T）の範囲にあるとする。T<∞とする。W_t　はウィーナー過程である。W_0　=0.

μ =0 を帰無仮説とする統計的検定問題を考える。

:(1)　Kを[O、∞]内の任意の数とする。[0,1]区間に属する任意のα に対して、軌跡があるレベル以下に滞在する時間の長さがα T　であるとき、そのレベルをα -quantile（m(α)と書く）と呼ぶが、このm(α)がK以下である確率がすでに求められている。それを利用して、m(α)の値域に棄却レベルK,あるいはK(α)を定める。さて、逐次解析としては、時間区間　[0,T]内で、つまり期末に至る途中で軌跡が K より上に居る滞在時間がα T を超えるときに停止して帰無仮説を棄却する。

この方式は、[0,1]区間内のすべてのα について定義できる。次に考えるべきことは、どのα を使うのが良いかである。一つの考え方は、上記の停止時刻が最も早いもの、つまり停止時刻の期待値が最も小さいものを選ぶことである。Brownian quantiles の確率分布論がすでに出来ているので、これは原理的に可能である。もう一つ考えるべき側面は、検出力であるが、上記の K の決まり方が一意であれば、選択の余裕はない。

:(2) 次に変化点検出問題について考える。μ の値がゼロから正の値に変化する時刻　t_0 を検出する問題である。t_0 を期間[0,T]内の任意の時刻とする。時刻　t_0 以前と以後の軌跡がそのレベルの高さに於いて顕著に異なっていれば、μ の値が変化したものを判断する。判断の基礎となる統計量は、二つの期間　$[0, t_0]$と$[t_0, T]$,あるいは[0,T]と$[t_0, T]$に「おけるm(α)を用いるのが一つの手段である。さらに、統計量としてm(α)を用いずに、ノンパラメトリックな二標本比較のように、Wilcoxonの順位和統計量を用いることも可能であるが、未だ確率分布論が不明である。

:（注意）。現実における観測は離散時刻に行われるので、上記の連続時間表現の量は実際の統計的推測では、empirical に扱われなければならない。そのためには、離散時刻の確率過程の連続時間への収束が議論される必要がある。その基本的一例として、Ngo Hoang Long (2010)　がある。

本研究集会に於いて、発表者の新しい試みを紹介する機会が得られ、さらに出席者からコメントをいただけたことは、今後の研究の励みになる。ここに感謝の意を表する次第である。

参考文献（主に Brownian quantiles と Rank に関するもの）

:[1]. Akahori, J. (1995). “Some formulae for a new type of path-dependent option.” Ann. Appl. Probab. 5. 383-388.

:[2]. A.N. Borodin and P. Salminen. (2002). Handbook of Brownian Motion - Facts and Formulae, 2nd. edition　p.256. (the first edition was published in1996)　Birkhauser.

:[3]. Dassios, A. (1995). “The distribution of the quantile of a Brownian motion with drift and the pricing of related path-dependent options. Ann. Appl. Probab. 5. 389-398.

:[4]. Dassios,A.(2005). “On the quantiles of Brownian motion and their hitting times.” Bernoulli 11(1), 29–36.

:[5]. Embrecht, P., Rogers, L.C.G. and Yor, M. (1995). “A proof of Dassios’s representation of the α-quantile of　Brownian motion with drift.” Annals of Applied Probability. 5,757-767.

:[6]. Fujita, T.(1997). ”On the price of α-percentile options.” Working Paper Series #24, Faculty of Commerce　Hitotsubashi University.

:[7]. Fujita, T. (2000). “A note on the joint distribution of α, ß percentiles and its applications to the option pricing.” Asia-Pacific Financial Markets. Vol.7(4), 339-344.

:[8]. Fujita, T. and Miura, R.(2002). “Edokko Options: A New Framework of Barrior Options.” Asia-Pacific Financial Markets. Vol.9(2),December, 141-151.

: [9]. Fujita, T. and Ishizaka, M.(2002). “An application of New Barrier Options (Edokko Options) for Pricing Bonds with Credit Risk”. Hitotsubashi Journal of Commerce and Management, Vol.37(1) pp.17-23.

:[10]. 三浦良造 (1972) .“逐次解析の統計的構造“　数理解析研究所高級録　第150巻　1972年　49-70.

:[11]. Miura, R. (1992). Miura,R.(1992). "A Note on Look-Back Options Based on Order Statistics", Hitotsubashi Journal of Commerce and Management. Vol. 27, No.1, November 1992.pp. 15-28.

:[12]. Miura,R. & Fujita,T. (2006). "The Distribution of Continuous Time Rank Processes” (with Takahiko Fujita), Mathematical Economics, Vol. 9, 2006

:[13]. Miura,R. (2007). “Rank Process, Stochastic corridor and Applications to Finance.” 529-542. Chapter 26.　Advances in Statistical Modeling and Inference: Essays in Honor of Kjell Doksum. World Scientific.2007.

:[14]. Ogawa,S.& NGO,H.L. (2011). “On the discrete approximation of occupation time of diffusion processes.” Electronic Journal of Statistics. Vol.5. 1374-1393.

:[15]. Yor, M.(1995).”The distribution of Brownian quantiles.” Journal of Applied Probability. 2, 405-416.

標本積率を用いた多変量正規性検定について

横浜市立大学 国際総合科学部　　小泉　和之
東京理科大学大学院 理学研究科　澄川　琢磨

本報告では, 得られた連続値データが正規分布に従っているかを調べるために標本積率をもとにした検定理論を考えた. その際に用いる歪度, 尖度であるが多変量データに対する歪度, 尖度は 1 変量とは異なり, 様々な定義がされているが, ここでは, Srivastava (1984) で用いられている多変量歪度, 尖度を用いる. 具体的には Koizumi et al. (2009) によって提案されている Srivastava (1984) の多変量歪度, 尖度をもとにした総括的な正規性検定統計量 (MJB_1 と呼ぶ) の改良を Wilson-Hilferty 変換を用いることにより行った. また, 次元 p が大きいもと (p は N を超えない) で近似精度の改良を行うため, F 分布を用いた近似の適用も行った. 最後に, これらの近似精度を数値的に比較するためにモンテカルロ・シミュレーションを行った.

MJB_1 は Srivastava (1984) による多変量標本歪度, 尖度を用いた統計量であるのでまず, その定義を紹介する. $\boldsymbol{x}$ は, 母平均 $\boldsymbol{\mu}$, 分散共分散行列 Σ である多次元連続分布に従う p 次元確率ベクトルとする. このとき, $\Gamma = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p)$ を $\Sigma = \Gamma D_\lambda \Gamma'$ となるような直交行列とする. ただし, $D_\lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $\lambda_1, \lambda_2, \ldots, \lambda_p$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0)$ は Σ の固有値である. Srivastava (1984) は主成分得点を用いて多変量歪度, 尖度を

$$\beta_1^2 = \frac{1}{p}\sum_{i=1}^{p}\left\{\frac{\mathrm{E}[(y_i-\theta_i)^3]}{\lambda_i^{3/2}}\right\}^2,\ \beta_2 = \frac{1}{p}\sum_{i=1}^{p}\frac{\mathrm{E}[(y_i-\theta_i)^4]}{\lambda_i^2}$$

として与えている. ここに, $y_i = \boldsymbol{\gamma}_i'\boldsymbol{x},\ \theta_i = \boldsymbol{\gamma}_i'\boldsymbol{\mu}$ $(i = 1, 2, \ldots, p)$ である. ただし, 多変量正規性のもとでは $\beta_1 = 0$, $\beta_2 = 3$ である.

また, その分布からの大きさ N の独立な観測ベクトルを $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ とする. これら N 個の標本から $\boldsymbol{\mu}$, Σ の最尤推定量として標本平均, 標本共分散行列をそれぞれ

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_j,\ S = \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})'$$

とする. 次に, S の固有値を $\omega_1, \omega_2, \ldots, \omega_p$ $(\omega_1 > \omega_2 > \cdots > \omega_p > 0)$ とすると, ある直交行列 $H = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_p)$ が存在して $H'SH = D_\omega = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_p)$ となる. このとき, Srivastava (1984) は多変量標本歪度, 尖度を

$$b_1^2 = \frac{1}{N^2 p}\sum_{i=1}^{p}\left\{\sum_{j=1}^{N}\frac{(y_{ij}-\overline{y}_i)^3}{\omega_i^{3/2}}\right\}^2,\ b_2 = \frac{1}{Np}\sum_{i=1}^{p}\sum_{j=1}^{N}\frac{(y_{ij}-\overline{y}_i)^4}{\omega_i^2} \tag{1}$$

として与えている. ただし, $y_{ij} = \boldsymbol{h}_i'\boldsymbol{x}_j,\ \overline{y}_i = \sum_{j=1}^{N} y_{ij}/N$ であり, $\boldsymbol{h}_i$ は S の第 i 直交ベクトルである.

この多変量標本歪度, 尖度を用いた総括的な多変量正規性検定統計量として, Koizumi et al. (2009) は次の統計量を提案している.

Theorem 1 (Koizumi et al. (2009)) $b_1^2,\ b_2$ をそれぞれ (1) 式で定義される多変量標本歪度, 尖度とする. このとき

$$MJB_1 = Np\left\{\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right\}$$

は十分大きな N に対して, 漸近的に自由度 $p+1$ の χ^2 分布に従う.

これは 1 変量における JB 統計量の自然な拡張となっている．しかし, この統計量の極限分布である χ^2 分布の自由度が次元数 p に依存しているため, p が大きいときに多変量標本歪度 b_1^2 の影響が強くなることが懸念される．そこで, 極限分布の自由度が次元数 p に依存しない統計量を提案するために Wilson-Hilferty 変換を用い, さらに, 近似精度の改良のため, Seo and Ariga (2011) で導出された b_2 に対する正規化変換統計量を用いることによって次の検定統計量を提案した．

Theorem 2 b_1^2, b_2 をそれぞれ (1) 式で定義される多変量標本歪度, 尖度とする．このとき

$$MJB_2 = \frac{9p}{2}\left\{\left(\frac{Nb_1^2}{6}\right)^{\frac{1}{3}} - 1 + \frac{2}{9p}\right\}^2 + \frac{Np}{24}\left\{-e^{-b_2+3} + 1 + \frac{6}{N}\left(1+\frac{2}{p}\right)\right\}^2 \tag{2}$$

とする．すると, $\forall x \in \mathbb{R}$ に対して

$$\lim_{p\to\infty}\lim_{N\to\infty} \Pr(MJB_2 \le x) = G_2(x)$$

が成り立つ．ここに, $G_2(x)$ は自由度 2 の χ^2 分布の累積分布関数である．

これにより, 検定統計量の極限分布の自由度が次元 p に依存しない形になっていることが確認できる．しかし, N が小さいもとあるいは次元 p と N が近い状況下では近似精度が良くないことが心配されるので, その近似精度を改良するために, MJB_2 の期待値の漸近展開を与えた．摂動法を用いることにより, MJB_2 の期待値について次の定理を得ることができた．

Theorem 3 MJB_2 を (2) 式で与えられる統計量とする．このとき

$$E[MJB_2] = 2 + \frac{15}{N} - \frac{84}{Np} + o(N^{-1}, p^{-1})$$

となる．

ところで,

$$n = \frac{4Np}{15p - 84} + 2$$

とする．このとき, $T \sim 2F_{2,n}$ なる確率変数を考えると,

$$E(T) = \frac{2n}{n-2} = 2 + \frac{15}{N} - \frac{84}{Np} \approx E(MJB_2).$$

となることより, 極限分布である χ_2^2 分布の上側パーセント点の代わりに $2F_{2,n}$ 分布の上側パーセント点を用いた近似も提案した．

参考文献

[1] Koizumi, K., Okamoto, N. and Seo, T. (2009). On Jarque-Bera tests for assessing multivariate normality. *Journal of Statistics: Advances in Theory and Applications*, **1**, pp. 207–220.

[2] Seo, T. and Ariga, M. (2011). On the distribution of sample measure of multivariate kurtosis. *Journal of Combinatorics, Information & System Sciences*, **55**, pp. 179–200.

[3] Srivastava, M. S. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statistics & Probability Letters*, **2**, pp. 263–267.

新たな離散異分布適合度検定統計量と その海洋調査データへの適用

柴田里程（慶應義塾大学理工学部)

1　離散分布の適合度検定統計量

与えられたデータに対する確率モデルの有効性を主張するには最終的には適合度検定に頼るしかない．しかし最近では，このような古典的な検定を省略し，頭のなかだけで考えたいくつかのモデルを AIC などのモデル選択基準で比較しその中での最良なモデルを最終結論とすればよい，という安易な風潮も見受けられる．モデル選択はあくまでも，どのモデルも適合度に問題はない場合に，すこしでも簡素で理解しやすいモデルを選択したいという欲求から生まれた方法である．つまり，相対的な比較でしかなく，それ自体で有効性を主張する力はもっていない．

すでに適合度検定統計量については十分研究し尽くされているようにも思えるが，それは連続分布の場合であって，離散分布の場合はピアソン χ^2 がほとんど唯一の統計量といってよく，特に，独立であっても同分布でない場合の研究ははほとんど手がつけられていない現状にある．そこで，Victoria University of Wellington の E.Khmaladze との共同研究の過程で発見した，ピアソン χ^2 の一般化を紹介するとともに，これを応用することで独立異分布の場合にも使える適合度検定統計量が容易に導かれることを示すとともに，その有効性を検証した結果を報告した．

2　離散異分布

異分布であることが避けられない一つの実例として，報告者がここ数年オーストラリアの CSIRO との共同研究として取り組んできた NPF(Northern Prawn Fishery) 海洋調査データをとりあげた．この調査では，エビのトロール漁の影響を調べるため，実験用に設定した領域でトロール漁を行う前と後の数回浚渫を行い，捕捉された海洋生物の種ごとにその個体数と重量を計測している．ここでの問題は，浚渫が船からかごを降ろしそれを引っ張りながら海底を底ざらいする形であるため，かごがいっぱいになりそうになるとそこで引き上げてしまい浚渫面積が一定しないところにある．つまり，n 回の浚渫の結果得られたある種の個体数を $N_1, N_2, ..., N_n$，対応する浚渫面積を $\alpha_1, \alpha_2, ..., \alpha_n$ とすればもっとも簡単なポアッソン分布モデルでも $N_i \sim Po(\alpha_i\lambda), i = 1, 2, ..., n$ のように異分布となり，種によっては集落を形成することが多いことも考慮したトーマス分布なら $Tho(\alpha_i\lambda, \phi), i = 1, 2, ..., n$ のような異分布を考える必要がある．このような異分布性は，連続分布なら適当な変換を導入することにより解消できることも多いが，離散分布，特に個数の分布に対してはこのような変換はかえって分布を複雑にするだけでほとんどメリットがないことが多い．

3　離散異分布の扱い

離散異分布を正面から扱う困難を避けるため，よく行われる便宜的な扱いとしては

1. N_i/α_i のように標準化し正規分布 $N(\lambda, \lambda/\alpha_i)$ あるいは $N(\lambda, \sigma^2)$ で近似する．
2. 0 カウントが多いことを反映するため負の二項分布 $NB_N(\alpha_i, p)$ を用いる．
3. $\log E(N_i) = \log\alpha_i + \log\theta$ のような GLIM(Generalized Linear Model) を用いる．

などがあるが，このような解析は現象の本質に迫るサイエンスというよりは単なる技法の適用であり，判明するのはごく表面的な事実だけである．下手をすれば虚像を見ているだけかもしれない．このような場合に解析者はサイエンティストとしてどう責任をとるつもりなのであろうか．

4　ピアソン χ^2 の一般化

まず，古典的な独立同分布な観測 $X_1, X_2, ..., X_n$ の場合を考えてみる．$p_i = P(X_k = x_i), i = 1, 2, ..., m$ としたとき，ピアソンの χ^2 はいうまでもなく $\chi^2 = ||\boldsymbol{Y}_n||^2 \simeq \chi^2(m-1)$ で与えられる．ここで $\simeq$ は漸近分布の意味で用いており，

$$Y_{in} = \frac{\#\{X_k = x_i,\ k = 1, 2, ..., n\} - np_i}{\sqrt{np_i}},\ i = 1, 2, ..., m$$

漸近 χ^2 分布は $\boldsymbol{Y}_n = (Y_{1n}, Y_{2n}, ..., Y_{mn})^T \simeq N\left(\boldsymbol{0},\ I - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^T\right)$ であることにもとづいている．ただし $\sqrt{\boldsymbol{p}}^T = (\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_m})$.

ピアソンの χ^2 は余りにも有名であるのであまり注意も払われないが，$\boldsymbol{Y}_n$ の漸近分布が $\boldsymbol{p}$ に依存するにも関わらず，その２乗ノルムが $\boldsymbol{p}$ に依存しない一定の χ^2 分布に従うのは，漸近分散共分散が射影行列になっているという極めてラッキーな状況をうまく利用しているからである．したがって，$\boldsymbol{Y}_n$ の一部の要素にもとづいた統計量などはもはやこのようなきれいな性質は持たない．しかし，すこし発想を変えれば漸近分散共分散を射影行列であることは保ちながら自由に変化させることができる．

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n - \langle \boldsymbol{Y}_n, \boldsymbol{r} \rangle \frac{\boldsymbol{r} + \sqrt{\boldsymbol{p}}}{1 + \left\langle \sqrt{\boldsymbol{p}}, \boldsymbol{r} \right\rangle} \ \simeq\ N(0,\ I - \boldsymbol{r}\boldsymbol{r}^T).$$

あきらかに，常に $||\boldsymbol{Z}_n||^2 \simeq \chi^2(m-1)$ であり $\boldsymbol{r} = \sqrt{\boldsymbol{p}}$ にとれば $\boldsymbol{Z}_n = \boldsymbol{Y}_n$ で，ピアソンの χ^2 に戻る．

5　異分布の場合への応用

前節の結果は，パラメータ $\boldsymbol{p}$ によらない漸近分布をもつように変換できるだけでなく，$\boldsymbol{r}$ を自由に選択できることから，検定の幅を広げるのにも役立つと思われるが，異分布の場合にはこのアイディアが本質的な役割を果たす．$p_{ki} = P(X_k = x_i), i = 1, 2, ..., m$ とし，

$$\eta_{ki} = \frac{I_{(X_k = x_i)} - p_{ki}}{\sqrt{p_{ki}}},\ i = 1, 2, ..., m,\ k = 1, 2, ..., n$$

を標本ごとに定義すれば

$$\boldsymbol{Z}_k = \boldsymbol{\eta}_k - \langle \boldsymbol{\eta}_k, \boldsymbol{r}_k \rangle \frac{\boldsymbol{r}_k + \sqrt{\boldsymbol{p}_k}}{1 + \left\langle \sqrt{\boldsymbol{p}}_k, \boldsymbol{r}_k \right\rangle},\ k = 1, 2, ..., n$$

の算術平均の漸近分散共分散は $\boldsymbol{r}_k$ を選ぶことでかなり自由に定められ，

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{Z}_k \ \simeq\ N\left(\boldsymbol{0},\ I - \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{r}_k \boldsymbol{r}_k^T\right)$$

が成立する．特に $\boldsymbol{r}_k = \boldsymbol{r}$ のように同じベクトルにとれば，独立同分布の場合とまったく同じように

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{Z}_k \ \simeq\ N\left(\boldsymbol{0},\ I - \boldsymbol{r}\boldsymbol{r}^T\right).$$

となり，$\|\boldsymbol{W}_n\|^2$ は自由度 $m-1$ の漸近 χ^2 分布を持つ．パラメータ推定を含む場合も，MDE(Minimum Distance Estimate) を用いる限り同じような変換で射影行列となる漸近分散共分散を自由に選ぶことができることも報告したが，ここではその詳細を省略する．

Improved Confidence Interval for Quantile

前 園 宜 彦　　九州大学数理学研究院
Penev, Spiridon　The University of New South Wales

1. 序

$X_1, X_2, \ldots, X_n$ を母集団分布 $F(x)$(密度関数 $f(x)$ を持つ) からの無作為標本とする．このとき p-確率点 $Q(p) = inf\{x : F(x) \geq p\}$ の推定問題に対して，密度関数のカーネル推定に基づく，カーネル型確率点推定量

$$\hat{Q}_{p,h_n} = \frac{1}{h_n}\int_0^1 F_n^{-1}(x)K(\frac{x-p}{h_n})dx \tag{1}$$

について報告した．ここで $F_n^{-1}(x)$ は経験分布関数の逆関数，$K(\cdot)$ は適当な条件を満たすカーネル関数で，$h_n \to 0, (n \to \infty)$ はバンド幅である．Maesono & Penev (2009) は標準化カーネル型確率点推定量のエッジワース展開を残差項 $o(n^{-1/2})$ まで求めた．本報告ではジャックナイフ分散推定量の一致性を示し，スチューデント化カーネル型確率点推定量のエッジワース展開を残差項 $o(n^{-1/2})$ まで議論する．スチューデント化には次のジャックナイフ型分散推定量を使う．

$$\hat{\sigma}_n^2 = (n-1)\sum_{i=1}^{n}\{\hat{Q}_{p,h_n}^{(i)} - \hat{Q}_{p,h_n}\}^2$$

ここで $\hat{Q}_{p,h_n}^{(i)}$ は X_i を除いた $n-1$ 個のサンプルに基づく対応する推定量であるが，簡単のために同じバンド幅を利用する．このときスチューデント化カーネル型確率点推定量

$$\hat{\sigma}_n^{-1}\sqrt{n}\{\hat{Q}_{p,h_n} - Q(p)\} \tag{2}$$

の分布のエッジワース展開を求める．なお本報告ではバンド幅について，次の仮定を置く．

$$h_n = o(n^{-1/4}) \quad \text{かつ} \quad \lim_{n\to\infty}(n^{1/4}h_n)^{-k}n^{-\beta} = 0$$

ただし $\beta > 0$ で k は正の自然数である．典型的なバンド幅は $h_n = n^{-1/4}(\log n)^{-1}$ である．

2. スチューデント化 $\hat{Q}_{p,h_n}$ のエッジワース展開

まず分散推定量の漸近表現を求め，それを利用してスチューデント化統計量の漸近表現を求めることができる．

Lemma 適当な正則条件の下で Hoeffding 分解された次の漸近表現を求めることができる．

(i)

$$\begin{aligned}\hat{\sigma}_n^2 &= \sigma_n^2 + n^{-1}h_n^{-2}(e_{5n}+e_{6n}) + n^{-1}B_{1n} + 2n^{-1}h_n^{-1}\hat{B}_{1n} + 2n^{-2}h_n^{-2}B_{2n} \\ &\quad + o_\ell(n^{-1/2})\end{aligned}$$

(ii)

$$\begin{aligned} & \hat{\sigma}_n^{-1}\sqrt{n}\{\hat{Q}_{p,h_n} - Q(p)\} \qquad (3) \\ = \;& \nu_n + d_{1n}A_{1n} + d_{2n}\Lambda_{2n} + d_{3n}\Lambda_{3n} + d_{3n}n\hat{\Lambda}_{1n} + o_\ell(n^{-1/2}) \end{aligned}$$

ここで

$$P(|o_\ell(n^{-1/2})| \geq n^{-1/2}(\log n)^{-1}) = o(n^{-1/2})$$

である.

この表現を使うとエッジワース展開を次のように求めることができる.

Theorem カーネル関数 $K(\cdot)$ と分布関数に関する適当な条件の下で

$$\begin{aligned} & P\Big(\sqrt{n}\hat{\sigma}_n^{-1}\{\hat{Q}_{p,h_n} - Q(p)\} \leq x\Big) \\ = \;& \Phi(x) - \phi(x)\Big\{\frac{\delta}{\sigma\sqrt{n}} + \frac{-2x^2-1}{6n^{1/2}\sigma_n^3}e_{1n} + \frac{-x^2-1}{2n^{1/2}\sigma_n^3 h_n}e_{2n} \\ & + \frac{1}{nh_n^2}\Big[\frac{x}{4\sigma_n^2}(2e_{5n} - e_{6n}) + \frac{-x^3+3x}{2\sigma_n^4}e_{3n} \\ & - \frac{x^3-3x}{3\sigma_n^4}e_{4n} + \frac{x^5+2x^3-41x}{8\sigma_n^6}e_{2n}^2\Big]\Big\} + o(n^{-1/2}) \end{aligned}$$

が成り立つ.

Remark 実際に利用する多くのカーネルは対称な関数が使われている (Sheather & Marron (1990, JASA) など). もし $K(-x) = K(x)$ ならば $\int_{-1}^{1} K'(x)dx = 0$ となり漸近展開は

$$\begin{aligned} & P\Big(\sqrt{n}\hat{\sigma}_n^{-1}\{\hat{Q}_{p,h_n} - Q(p)\} \leq x\Big) \\ = \;& \Phi(x) - \phi(x)\Big\{\frac{\delta}{\sigma\sqrt{n}} + \frac{-2x^2-1}{6n^{1/2}\sigma_n^3}e_{1n} + \frac{-x^2-1}{2n^{1/2}\sigma_n^3 h_n}e_{2n} \\ & + \frac{1}{nh_n^2}\Big[\frac{x}{2\sigma_n^2}e_{5n} - \frac{x^3-3x}{3\sigma_n^4}e_{4n}\Big]\Big\} + o(n^{-1/2}) \end{aligned}$$

になる.

さらに高次のオーダーのカーネルでは, $\int_{-1}^{1} K''(x)dx = 0$ の仮定を満たすものがある. もしこの仮定を満たせば漸近展開は

$$\Phi(x) - \phi(x)\Big\{\frac{\delta}{\sigma\sqrt{n}} + \frac{-2x^2-1}{6n^{1/2}\sigma_n^3}e_{1n} + \frac{-x^2-1}{2n^{1/2}\sigma_n^3 h_n}e_{2n}\Big\} + o(n^{-1/2})$$

と簡略化される.

情報科学演習期間におけるグループ内変動の定量的分析

Quantitative analysis for variance within group in medical informatics practice

安田晃Akira Yasuda,　津本周作Shusaku Tsumoto

はじめに

学生がいくつかのグループに分かれ，それぞれグループが独立して一定期間自主学習を行う際，初期の学習は混沌とし，学習が進むにつれ獲得する知識，情報量，それらを整理するための論理的な手段などを得るようになることを我々は多次元尺度構成法，グラフィカル対数線形モデリングなどを用い報告してきた．しかし，その際，知識，手段など個人が得る情報量にはグループ内で様々な分散があることは概観できるがグループを構成している学生間で表れる手段，情報量の差異を直接見ることはできない．

そこで本研究では 14 グループ内の変動を，グループを構成する学生個人と質問項目の固有ベクトルとして定量的に求め，学生の変動を面積として図示することを提案する．

方法

看護学科 1 年生および編入生に対する 2009 年度後期情報科学演習終了後，表 1 に示した 19 項目からなる学習態度評価シート（評価シート）に自己の学習態度を自記させた．

この評価シートでは個人と質問項目の行列の要素に個人が判断した学習態度が Yes, No のいずれかが入る．我々は個人の得点と，その個体が反応したカテゴリ得点は等質な値をとるという等質性の仮定から，各演習日ごとの等質性分析を行い，19 項目と対象となった学生の数量化を行った．

グループ内での学生がばらばらに議論するならば，評価シートの回答パターンも異なると考える．そこで学グループに注目し，得られた学生の数量化された固有ベクトルをグループごとに 2 次元プロットし，3 人以上の学生が出席していた場合多角形を作成する．その多角形の面積が大きければ思考の多様性が，小さければ均一な思考性を有していると考える．

表 1　学習態度評価シート

1)多目的な発想や統合的な連想ができたか
2)課題に含まれる重要なテーマに気づいたか
3)既に学んだ知識の整理ができたか
4)課題から様々な疑問点や学習項目を抽出できたか
5)抽出した学習項目を重要度に応じて順位付けできたか
6)グループ全員に共通な学習項目を設定できたか
7)自分独自の学習項目を設定できたか
8)基本的な事項を学ぼうとしたか
9)発展的・応用的な事項を学ぼうとしたか
10)学習計画ごとに自らの到達目標を設定できたか
11)学習計画の時間配分は適切であったか
12)問題解決をするための具体的方法を見い出せたか
13)自己学習に十分な時間と努力を注いだか
14)自分が設定した到達目標を達成できたか
15)自分の考えを簡明かつ論理的に説明できたか
16)他者の考えを理解しようと努めたか
17)自分の考えと異なる意見に対しても柔軟な態度がとれたか
18)討論や発表の時間配分に留意したか
19)グループの一員として問題解決への建設的貢献をしたか

結果

各演習日ごとの評価シートの内的整合性を KR-20 でみたところ 0.781 から 0.863 であり，妥当であると判断した．

図 1 に演習 2 回目の 14 グループそれぞれに得られた学生に対する第 2 固有値に対する固有ベクトルまでを用い，2 次元プロットした．表 2 にはそれらの面積をすべての演習日について示した．

図 1 を見ればグループ 11，1 の面積が大きく，グループ内での思考過程に分散が見られていると思われる．これらのグループは I 軸，II 軸に射影した場合，I 軸 II 軸に大きな範囲を得る．他のグループの面積はグループ 1，11 と比較し大きな面積ではない．

考察

グループ内の変動を確認するとき，得られた評価シートより個人間の点相関係数行列，あるいは何らかの距離行列を求めるが，今回の手法はグループ内の変動を 14 グループ間で定量的に，ビジュアルに見ることができる．

ただ，ここではプロットした 2 次元における I 軸，II 軸の意味を言及していない．個人を頂点とした平面の面積を求めにすぎず，結果を視覚化しただけである．得られる固有値は min（質問のカテゴリ数－質問数，回答者数－1）であり，解は 19 次元である．このよう

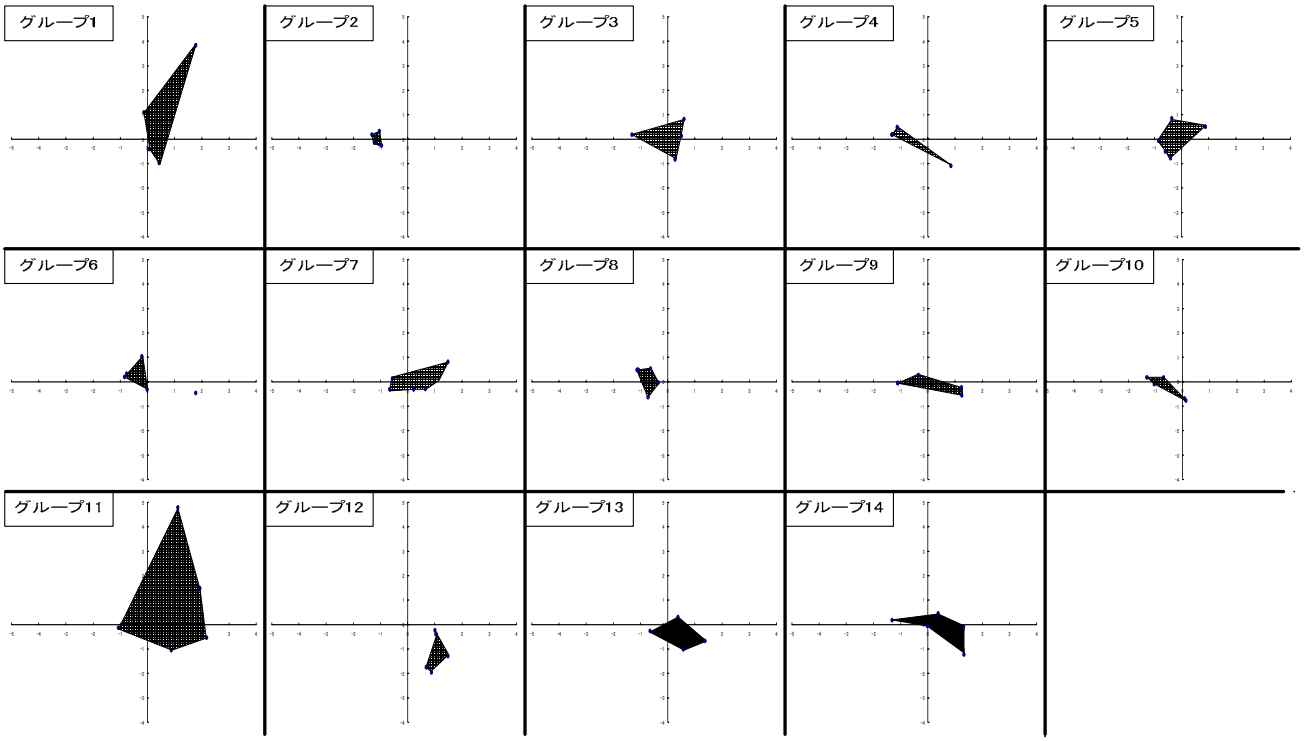


図1　演習2回目の各グループの多角形表示

表2　演習期間における各グループの多角形面積

	演習1日	2日	3日	4日	5日	6日	7日	8日	9日	10日	11日
グループ1	7.53	13.42	2.22	7.51	2.64	4.06	8.68	18.08	9.81	16.55	3.87
2	1.33	0.64	2.41	1.09	0.43	0.91	0.78	0.00	0.00	0.00	0.00
3	3.78	6.41	0.72	6.27	3.94	0.22	6.55	1.75	7.78	4.28	27.94
4	1.32	2.05	1.41	6.29	0.00	0.00	0.00	6.19	25.44	11.43	0.00
5	3.39	6.08	4.45	8.10	5.28	2.52	33.98	0.03	11.40	0.28	5.16
6	0.22	2.33	8.26	0.18	0.00	1.88	0.00	0.38	0.00	0.00	0.00
7	15.83	6.00	25.70	2.74	5.09	1.60	13.55	5.53	5.31	5.55	9.22
8	19.38	2.28	1.40	0.65	0.09	0.00	0.55	0.00	2.43	2.01	0.00
9	5.25	3.82	6.61	3.75	12.50	4.67	12.86	7.81	2.68	–	10.45
10	7.12	1.64	1.13	4.18	1.37	0.26	8.10	6.02	4.33	0.00	1.41
11	12.58	42.97	5.63	26.33	21.67	32.04	0.00	7.12	4.51	0.00	11.13
12	1.92	2.54	15.22	9.04	13.97	7.63	10.02	12.52	10.08	4.65	10.02
13	26.18	5.98	28.23	15.26	3.49	8.56	16.75	3.52	12.75	17.21	13.01
14	7.53	6.36	7.50	9.55	11.14	18.31	4.14	7.28	1.13	1.24	0.55

セル内の0.00は2人の出席，あるいは3人以上の出席であるが同一の自記により多角形が得られなかったもの．欠損値は全員欠席

な状況下での2次元プロットでは解に対して寄与が著しく少ないことが考えられる．しかし，各演習期間のプロット，それによって得られた多角形においてはグループを構成している個人を関数としてすべてではないが何らかの分散を形成しているものと考える．今後は3次元のようなより高次の空間からのアプローチも考えている．

演習日が独立していると考えられないので，演習日間の比較はできないと考えるが，学習行動把握のための手法のひとつとして提案したい．

多変量線形回帰モデルにおける AIC の漸近性質について

広島大学大学院理学研究科　伊森 晋平

正規線形回帰モデルに対する変数選択問題は, 実データ解析においてしばしば生じる重要な問題の一つである. モデル選択規準はこれまでに多く提案されているが, 中でも KL 情報量を基にしたモデル選択規準 (情報量規準) として, Akaike (1973,1974) により提案された赤池情報量規準 (AIC) は特に有名である. これまでに Shibata (1983), Nishii (1984) をはじめとして, AIC などの情報量規準の漸近最適性に関する議論は様々な角度から行われてきており, その特徴づけが行われている. しかしながら, これらの性質は目的変数の次元が単変量のときに示されているだけであり, 目的変数の観測時点数が多変量や高次元の場合については, Yanagihara, et al. (2012) による Nishii(1984) の結果の拡張しか見受けられない. さらに, この結果は真のモデルが候補のモデルの集合全体に含まれているという仮定の下で導出されているため, 真のモデルが候補のモデルの集合全体に含まれない状況における AIC の漸近性質に対する問題は残されたままである. 特に高次元における漸近性質は, 従来用いられる大標本理論とは異なり, 高次元特有の漸近論を用いる必要があるため, 単変量での結果が素直に拡張されるとは限らない. そこで本研究では多変量解析でしばしば用いられる多変量線形回帰モデルにおける変数選択問題に対し, 情報量規準の漸近性質について報告した.

本研究では以下のモデルを扱う. $\boldsymbol{Y}$ を $n \times p$ 目的変数行列, $\boldsymbol{X}$ を $n \times k$ 説明変数行列とする. n はサンプル数, p は各目的変数の次元数, そして k は説明変数の個数を表す. k_j 個の要素から構成される添え字集合 $j \subset \{1, \ldots, k\}$, 及び j に対応する $n \times k_j$ 説明変数行列 $\boldsymbol{X}_j$ を定義する. 例えば, $j = \{1, 2, 4\}$ ならば, $\boldsymbol{X}_j$ は $\boldsymbol{X}$ の $1, 2, 4$ 番目の列ベクトルにより構成された説明変数行列を表している. 候補のモデル j に対して以下を仮定する.

$$\boldsymbol{Y} = \boldsymbol{X}_j \boldsymbol{\Theta}_j + \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_j^{1/2}, \ \boldsymbol{\mathcal{E}} \sim N_{n \times p}(\boldsymbol{O}_{n \times p}, \boldsymbol{I}_{np}),$$

$\boldsymbol{\Theta}_j$ は $k_j \times p$ 未知回帰係数, $\boldsymbol{\Sigma}_j$ は $p \times p$ 未知共分散行列である. 一方で, 真のモデル j_0 を以下で定義する.

$$\boldsymbol{Y} = \boldsymbol{\Gamma}_* + \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_*^{1/2}, \quad \boldsymbol{\mathcal{E}} \sim N_{n \times p}(\boldsymbol{\mathcal{O}}_{n \times p}, \boldsymbol{I}_{np}).$$

ただし, $\boldsymbol{\Sigma}_*$ は真の共分散行列である. 真のモデルにおける $\boldsymbol{Y}$ の対数尤度は,

$$\ell(\boldsymbol{\Sigma}_*, \boldsymbol{\Gamma}_* | \boldsymbol{Y}) = -\frac{np}{2} \log 2\pi - \frac{n}{2} \log |\boldsymbol{\Sigma}_*| - \frac{n}{2} \mathrm{tr}\{(\boldsymbol{Y} - \boldsymbol{\Gamma}_*) \boldsymbol{\Sigma}_*^{-1} (\boldsymbol{Y} - \boldsymbol{\Gamma}_*)'\},$$

である.

このとき, $\mathcal{J}_n$ を候補のモデルの全体とすると, モデル $j \in \mathcal{J}_n$ に対する KL 情報量を基にし

た損失 $\boldsymbol{L}_n(j)$, 期待損失 $\boldsymbol{R}_n(j)$ は以下のように定義される.

$$\begin{aligned}
\boldsymbol{L}_n(j) &= -2\mathrm{E}[\ell(\boldsymbol{\Sigma}_j, \boldsymbol{\Theta}_j|\boldsymbol{Y})]|_{\boldsymbol{\Sigma}_j=\hat{\boldsymbol{\Sigma}}_j, \boldsymbol{\Theta}_j=\hat{\boldsymbol{\Theta}}_j} + 2\mathrm{E}[\ell(\boldsymbol{\Sigma}_*, \boldsymbol{\Gamma}_*|\boldsymbol{Z})] \\
&= n\log|\boldsymbol{\Sigma}_*^{-1}\hat{\boldsymbol{\Sigma}}_j| + n\mathrm{tr}(\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}_j^{-1}) + \mathrm{tr}\{(\boldsymbol{\Gamma}_* - \boldsymbol{X}_j\hat{\boldsymbol{\Theta}}_j)\hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{\Gamma}_* - \boldsymbol{X}_j\hat{\boldsymbol{\Theta}}_j)'\} - np \\
&= n\log|\boldsymbol{T}_j| + \mathrm{tr}\{\boldsymbol{T}_j^{-1}(n\boldsymbol{I}_p + n\boldsymbol{\Psi}_j + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_j}\boldsymbol{\mathcal{E}})\} - np, \\
\boldsymbol{R}_n(j) &= n\mathrm{E}[\log|\boldsymbol{T}_j|] + \mathrm{E}[\mathrm{tr}\{\boldsymbol{T}_j^{-1}((n+k_j)\boldsymbol{I}_p + n\boldsymbol{\Psi}_j\}\}] - np.
\end{aligned}$$

ただし,

$$\hat{\boldsymbol{\Theta}}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{Y}, \quad \hat{\boldsymbol{\Sigma}}_j = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{P}_{\boldsymbol{X}_j}^{\perp}\boldsymbol{Y}, \quad \boldsymbol{T}_j = \boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2}$$

であり, $\boldsymbol{P}_{\boldsymbol{X}_j}^{\perp} = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j}$, $\boldsymbol{P}_{\boldsymbol{X}_j} = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$, さらに, $\boldsymbol{\Psi}_j$ は非心行列を意味している. すなわち, $\boldsymbol{\Psi}_j$ は以下のように定義される.

$$\boldsymbol{\Psi}_j = \frac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2}.$$

モデル j が真のモデルを含む場合には, $\boldsymbol{\Psi}_j = \boldsymbol{\mathcal{O}}_{p\times p}$ である.

また, モデル j におけるリスクの推定量として AIC を定義する.

$$\mathrm{AIC}(j) = n\log|\boldsymbol{T}_j| + 2k_jp + p(p+1).$$

このとき, 高次元の枠組みで AIC は漸近最適性を持つ. すなわち,

$$\lim_{p/n\to c_0}\frac{\mathrm{E}[\boldsymbol{L}_n(\hat{j}_a)]}{\boldsymbol{R}_n(j_*)} = 1.$$

ただし, $c_0 \in (0,1)$, $\hat{j}_a = \underset{j\in\mathcal{J}_n}{\mathrm{argmin}}\mathrm{AIC}(j)$, $j_* = \underset{j\in\mathcal{J}_n}{\mathrm{argmin}}\boldsymbol{R}_n(j)$ である.

参考文献

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

[3] Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *Ann. Statist.*, **12**, 758–765.

[4] Shibata, R. (1983). Asymptotic mean efficiency of regression variables. *Ann. Inst. Statist. Math.*, **35**, 415–423.

[5] Yanagihara, H., Wakaki, H. & Fujikoshi, Y. (2012). A Consistency Property of the AIC for Multivariate Linear Models When the Dimension and the Sample Size are Large. *TR No 12-08, Statistical Research Group, Hiroshima University.*

罰則付カーネル正準相関分析における罰則最適化のための CV 規準

広島大学大学院 理学研究科 永井 勇 (*E-mail:* inagai@hiroshima-u.ac.jp)

Hotelling (1936) により提案された正準相関分析 (Canonical Correlation Analysis; CCA) は, $E[\boldsymbol{y}] = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{yy}}$, $E[\boldsymbol{x}] = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{xx}}$, $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{yx}}$ の確率変数 $\boldsymbol{y}$ と $\boldsymbol{x}$ の線形結合の相関を最大化することで, $\boldsymbol{y}$ と $\boldsymbol{x}$ の線形関係を分析する手法である (詳細は Srivastava (2002) などを参照). つまり, $\mathrm{Cor}(\boldsymbol{a}'\boldsymbol{y}, \boldsymbol{b}'\boldsymbol{x})$ を最大にする適当な大きさのベクトル $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ を求めて, 分析を行う手法である. ここで, $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$, $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$, $\boldsymbol{\Sigma}_{\boldsymbol{yx}}$ は全て未知であり, $\det(\boldsymbol{\Sigma}_{\boldsymbol{yy}}) \neq 0$, $\det(\boldsymbol{\Sigma}_{\boldsymbol{xx}}) \neq 0$ とする. CCA の目的は次で定式化される:

$$\arg\max_{\boldsymbol{a},\boldsymbol{b}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yx}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b}}}.$$

ここで, スケールの自由度に対する制約として $\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} = 1$ と $\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} = 1$ を加えると, 以下の形で書きかえることができる:

$$\arg\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yx}}\boldsymbol{b} \ \text{ under } \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} = 1,\ \boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} = 1. \tag{1}$$

CCA は, Doeswijk *et al.* (2011) から分かるように, 現在でも多くの応用領域で用いられている. また, (1) の解 $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ を求めるためには, 次のラグランジュの未定乗数法がよく用いられる:

$$\mathcal{L} = \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yw}}\boldsymbol{b} - \frac{\theta_a}{2}(\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} - 1) - \frac{\theta_b}{2}(\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} - 1),$$

ここで, θ_a, θ_b は非負の未定乗数である.

$\partial\mathcal{L}/(\partial\boldsymbol{a})|_{\boldsymbol{a}=\boldsymbol{\alpha}} = \boldsymbol{0}$, $\partial\mathcal{L}/(\partial\boldsymbol{b})|_{\boldsymbol{b}=\boldsymbol{\beta}} = \boldsymbol{0}$ を解くと, $\boldsymbol{\beta}$ は $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{yx}})'\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}}$ の最大固有値に対応する固有ベクトルとして得られ, $\boldsymbol{\alpha}$ は最大固有値と $\boldsymbol{\beta}$ から得られる. したがって, それぞれの推定量を $\boldsymbol{S}_{\boldsymbol{xx}}$, $\boldsymbol{S}_{\boldsymbol{yx}}$, $\boldsymbol{S}_{\boldsymbol{yy}}$ とすると, $\boldsymbol{S}_{\boldsymbol{xx}}^{-1}(\boldsymbol{S}_{\boldsymbol{yx}})'\boldsymbol{S}_{\boldsymbol{yy}}^{-1}\boldsymbol{S}_{\boldsymbol{yx}}$ の固有値問題を解くことで $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ の推定量 $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ が得られる.

CCA は, $\boldsymbol{y}$ と $\boldsymbol{x}$ の線形結合の線形関係しか見ることができない. つまり, $\boldsymbol{y}$ と $\boldsymbol{x}$ 間に非線形関係がある場合に, CCA では捉える事が出来ない. そこで 赤穂 (2000) などにより, カーネル正準相関分析 (Kernel CCA; KCCA) が提案されている. KCCA では, $\boldsymbol{y}$ や $\boldsymbol{x}$ を関数 $\boldsymbol{\phi}(\cdot)$ や $\boldsymbol{\varphi}(\cdot)$ を用いて, $\boldsymbol{\phi}(\boldsymbol{y})$ や $\boldsymbol{\varphi}(\boldsymbol{x})$ と変換し, $\boldsymbol{z} = \boldsymbol{\phi}(\boldsymbol{y})$ と $\boldsymbol{w} = \boldsymbol{\varphi}(\boldsymbol{x})$ として CCA を行うことで, 非線形関係を捉える形での CCA が可能となる. しかしながら, $\boldsymbol{\phi}(\cdot)$ や $\boldsymbol{\varphi}(\cdot)$ に柔軟な関数を用いた場合, 過剰に相関を最大化するという問題がある. そこで赤穂 (2000) などでは罰則を用いてこの問題を回避する手法 (Penalized KCCA; PKCCA) が提案されている. しかし, 従来の PKCCA においては, 罰則パラメータを解析者が恣意的に決定している.

そこで本発表では, PKCCA において罰則パラメータの最適化のための交差検証 (Cross Validation; CV) 法を提案する. 計算時間の短縮および議論を簡単にするため, $\boldsymbol{y}$ と $\boldsymbol{w} = \boldsymbol{\varphi}(\boldsymbol{x})$ に対する PKCCA を考える. CCA と同様に考えると,

$$\arg\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yw}}\boldsymbol{b} \ \text{ under } \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} = 1,\ \boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{ww}}\boldsymbol{b} = 1,$$

を求めることとなる. ここで, $\boldsymbol{\Sigma}_{\boldsymbol{yw}} = \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{w})$, $\boldsymbol{\Sigma}_{\boldsymbol{ww}} = \mathrm{Var}(\boldsymbol{w})$ である. PKCCA は, このラグランジュの未定乗数法に対して, $\boldsymbol{w}$ に関する制約を加える手法であり, 罰則付ラグランジュの未定乗数法は次の形で提案されている:

$$\mathcal{L}_\lambda = \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yw}}\boldsymbol{b} - \frac{\theta_a}{2}(\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} - 1) - \frac{\theta_b}{2}\{\boldsymbol{b}'(\boldsymbol{\Sigma}_{\boldsymbol{ww}} + \lambda\boldsymbol{P})\boldsymbol{b} - 1\}, \tag{2}$$

ここで λ は非負罰則パラメータ, $\boldsymbol{P}$ は適当な大きさの既知非負定値罰則行列である. この罰則付ラグランジュの未定乗数法は,

$$\arg\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yw}}\boldsymbol{b} \ \text{ under } \boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{a} = 1,\ \boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{ww}}\boldsymbol{b} + \lambda\boldsymbol{b}'\boldsymbol{P}\boldsymbol{b} = 1,$$

に対するラグランジュの未定乗数法となっている. 二つ目の条件の第二項 $\lambda\boldsymbol{b}'\boldsymbol{P}\boldsymbol{b}$ により, $\boldsymbol{b}$ に対する罰則を付けていると考えることができる.

CCA と同様に $\partial\mathcal{L}_\lambda/(\partial\boldsymbol{a})|_{\boldsymbol{a}=\boldsymbol{\alpha}} = \boldsymbol{0}$, $\partial\mathcal{L}_\lambda/(\partial\boldsymbol{b})|_{\boldsymbol{b}=\boldsymbol{\beta}} = \boldsymbol{0}$ を解くと, $(\boldsymbol{\Sigma}_{\boldsymbol{ww}} + \lambda\boldsymbol{P})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{yw}})'\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}$ $\boldsymbol{\Sigma}_{\boldsymbol{yw}}$ の固有値問題などとなり, PKCCA の目的となる $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ が得られる. よって, それぞれの推定量 $\hat{\boldsymbol{\alpha}}_\lambda$, $\hat{\boldsymbol{\beta}}_\lambda$ は $(\boldsymbol{S}_{\boldsymbol{ww}} + \lambda\boldsymbol{P})^{-1}(\boldsymbol{S}_{\boldsymbol{yw}})'\boldsymbol{S}_{\boldsymbol{yy}}^{-1}\boldsymbol{S}_{\boldsymbol{yw}}$ の固有値問題から得られる. 本発表では, (2) における λ の最適化のための CV 法を提案する.

λ の最適化を行うためには, 新たに得られたデータ $\boldsymbol{y}^*$, $\boldsymbol{w}^* = \boldsymbol{\varphi}(\boldsymbol{x}^*)$ に対して, 今あるデータで (2) を解いて得られる $\hat{\boldsymbol{\alpha}}_\lambda$, $\hat{\boldsymbol{\beta}}_\lambda$ を用いて, $\hat{\boldsymbol{\alpha}}_\lambda\boldsymbol{S}_{\boldsymbol{y}^*\boldsymbol{w}^*}\hat{\boldsymbol{\beta}}_\lambda$ を最大化することが目的となる. そこで本研究では λ の最適化のために, n 個のサンプル $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1,\ldots,n}$, $\boldsymbol{w}_i = \boldsymbol{\varphi}(\boldsymbol{x}_i)$, $(i = 1, \ldots, n)$ を用いたときの CV 法を提案する. $\boldsymbol{z}_i = (\boldsymbol{w}_i', \boldsymbol{y}_i')'$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ として CV 法を考えるため, $\boldsymbol{Z}$ から i 番目と j 番目を除いたデータ $\boldsymbol{Z}_{[-i,-j]}$ を用いて, 以下の手順で λ の最適化を行う:

1. $\boldsymbol{S}_{\boldsymbol{ww}}^{[-i,-j]}$,$\boldsymbol{S}_{\boldsymbol{wy}}^{[-i,-j]}$, $\boldsymbol{S}_{\boldsymbol{yy}}^{[-i,-j]}$ を $\boldsymbol{Z}_{[-i,-j]}$ から求める.
2. $\left(\boldsymbol{S}_{\boldsymbol{ww}}^{[-i,-j]} + \lambda\boldsymbol{P}\right)^{-1}\left(\boldsymbol{S}_{\boldsymbol{yw}}^{[-i,-j]}\right)'\left(\boldsymbol{S}_{\boldsymbol{yy}}^{[-i,-j]}\right)^{-1}\boldsymbol{S}_{\boldsymbol{yw}}^{[-i,-j]}$ の固有値問題を解いて, 整理することで $\hat{\boldsymbol{\alpha}}_\lambda^{[-i,-j]}$, $\hat{\boldsymbol{\beta}}_\lambda^{[-i,-j]}$ を得る.
3. $C_\lambda^{[-i,-j]} = \hat{\boldsymbol{\alpha}}_\lambda^{[-i,-j]'}(\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{w}_i - \boldsymbol{w}_j)'\hat{\boldsymbol{\beta}}_\lambda^{[-i,-j]}$ を計算する.
4. $\hat{\lambda} = \max_\lambda \sum_{i\neq j}^n C_\lambda^{[-i,-j]}$ として罰則パラメータの推定量 $\hat{\lambda}$ を得る.

この CV 法において, i 番目と j 番目を除いて得られる $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ の推定量に対して, i 番目と j 番目の共分散に対する評価を $C_\lambda^{[-i,-j]}$ で与えている. 数値実験に関しては, 当日の発表で報告する.

引用文献:

[1] 赤穂 昭太郎 (2000) カーネル正準相関分析. 2000 年情報論的学習理論ワークショップ.

[2] Doeswijk, T. G., Hageman, J. A., Westerhuis, J. A., Tikunov, Y., Bovy. A. & van Eeuwijk, F. A. (2011) Canonical correlation analysis of multiple sensory directed metabolomics data blocks reveals corresponding parts between data blocks. *Chemometr. Intell. Lab.*, **107**, 371–376.

[3] Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

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

自己組織化マップと情報量規準を用いたクラスタリング手法の提案

A Study on Clustering Method by using Self-Organizing Map and Information Criterion

加藤　聡 [1]　　**堀内　匡** [2]

Satoru Kato　　Tadashi Horiuchi

1. はじめに

自己組織化マップ (SOM) を用いたクラスタリング [1] は，学習後のコードベクトル同士の距離の変化に着目してクラスタ境界を検出する手法であり，k-means 法などと比較して，初期状態の違いによる結果のばらつきが少ないことや，任意形状のクラスタ抽出が可能であることが特長である．しかしながら，隣接セル間のコードベクトル間の距離に基づく「距離ベース」のクラスタ抽出法であるため，個々のクラスタのサイズやデータ密度が大きく異なる場合に，クラスタ境界の判定に伴うしきい値の設定が困難になるという問題がある．

一方，個々のデータ同士のユークリッド距離の変動に注目せずに，データ群の局所的な「まとまりの良さ」を評価してクラスタを見出す手法を考えることもできる．これは，データ集合の分布の状態に注目してクラスタとしてのもっともらしさを評価することから,「分布ベース」のアプローチと考えることができ，Pelleg ら [2] は，k-means 法にベイズ型情報量規準 (BIC)[3] を用いた再帰的なクラスタ分割を導入した手法である x-means 法を提案している．

情報量規準を用いた分布ベースのアプローチは，SOM によるクラスタリング手法にも比較的容易に導入できると考えられる．そこで本稿では，SOM を用いたクラスタリング手法において, 情報量規準に基づいたクラスタ抽出法を提案し，クラスタリング実験によって提案手法の有効性について述べる．

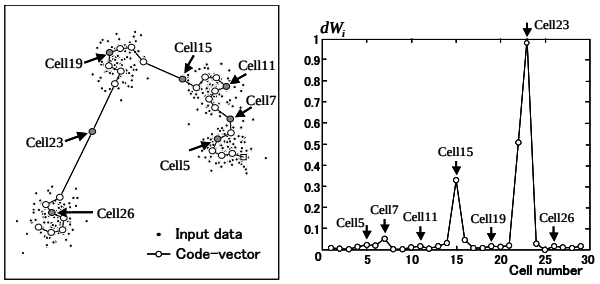
2. 提案手法

2.1 SOM を用いたクラスタ候補の抽出

Kohonen によって提案された SOM[4] の学習後に得られるマップでは，競合層上で隣接するセル間のコードベクトルが，データ空間上においても隣接しているという「位相保持写像」がなされており，さらに，入力データ空間でのデータの疎密が，学習後のコードベクトルの分布に反映されるという特徴がある（図 1(a) 参照）．ここで，セルが 1 次元的に並んだ 1 次元 SOM を学習に用いて，学習後にセル番号を横軸，隣接セル同士のコードベクトル間距離を縦軸とするような「データ密度ヒストグラム」を作成する．ヒストグラムにおける上向きのピークを機械的に検索していくと，クラスタをなすと思われるデータ集合を抽出することができる．

2.2 情報量基準を用いたクラスタ候補の併合

データ密度ヒストグラムでは，クラスタ境界部分以外にも複数のピークが現れる．したがって，前述の方法で



(a) 学習後のコードベクトルの分布　(b) データ密度ヒストグラム

図 1: 学習終了後の SOM のコードベクトル分布とデータ密度ヒストグラムの例（4 クラスタデータ）

抽出されたデータ集合は，本来のクラスタの部分集合のようなものとなっている．提案手法では，クラスタ候補の併合の際に用いる評価基準として赤池情報量規準（AIC）を採用し，以下に示す手順でクラスタリングを行う．

A. クラスタリング対象データを 1 次元 SOM に学習させ，学習結果からデータ密度ヒストグラムを作成する．

B. データ密度ヒストグラムから初期のクラスタ候補集合を生成し，個々のクラスタ候補に対して，競合層のセルの並びに準じた通し番号を付ける．

C. 通し番号が連続するクラスタ候補の任意のペアに注目し，AIC に基づいたクラスタ候補の選択的な併合を行う．

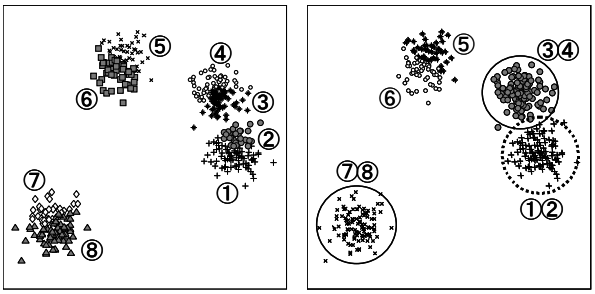
手順 C は，具体的には以下の手順によって行われる．

C1. 2 つのクラスタ候補を仮に併合したとき，仮併合後のクラスタに対して単一分布モデルあるいは二分布モデルを当てはめた場合の情報量規準の値 $\mathrm{AIC}_{\mathrm{single}}$ と $\mathrm{AIC}_{\mathrm{twin}}$ を算出し，当てはめた分布モデルの違いによる情報量規準の値の変化量 ΔAIC を以下によって求める．

$$\Delta\mathrm{AIC} = \mathrm{AIC}_{\mathrm{single}} - \mathrm{AIC}_{\mathrm{twin}} \qquad (1)$$

C2. 番号が隣接するクラスタ候補のすべての組み合わせについて手順 C1 を行った上で，ΔAIC が最も小さいクラスタ候補のペアを併合して 1 つの新たなクラスタ候補とする．その後，クラスタ候補全体の通し番号を再度付け直す．

[1]松江工業高等専門学校 情報工学科
[2]松江工業高等専門学校 電子制御工学科



(a) 初期クラスタ候補群　(b) 3 回目の併合操作後

図 2: 情報量規準を用いたクラスタ候補の併合過程

C3. クラスタ数が指定の数に達するまで，C1〜C2 の過程を繰り返す.

提案手法の実行例として，図 1(b) に示したデータ密度ヒストグラムから得られる初期のクラスタ候補群は図 2(a) のようになる．その後，手順 C を 3 回繰返した後のクラスタ候補の併合後の様子を図 2(b) に示す.

2.3 情報量規準を用いたクラスタ数の推定

クラスタ併合の判定に用いられる $\Delta\mathrm{AIC}(= \mathrm{AIC}_{\mathrm{single}} - \mathrm{AIC}_{\mathrm{twin}})$ は，単一分布モデルおよび二分布モデルそれぞれにおける情報量規準の差分であり，2 つのクラスタ候補を単一の分布とみなした方が良い場合に $\mathrm{AIC}_{\mathrm{single}} < \mathrm{AIC}_{\mathrm{twin}}$ となり，逆に 2 つの分布とみなした方が良い場合には $\mathrm{AIC}_{\mathrm{single}} > \mathrm{AIC}_{\mathrm{twin}}$ となる．したがって，ΔAIC の符号に着目すれば，クラスタ候補の併合処理を適切な段階で中断することができ，x-means 法の場合と同様に，クラスタ数の自動的な推定が可能になると考えられる.

3. クラスタ抽出実験

実験では，図 1(a) に示した 4 クラスタからなる人工データと，Iris データセット [5] をクラスタリング対象データとして使用した．表 1 は，人工データに対してクラスタ併合を進めたときの，クラスタ併合のレベルとその併合レベルにおける各併合候補の ΔAIC の最小値 $\Delta\mathrm{AIC}_{\mathrm{min}}$ を示したものである．このデータセットのクラスタ数は 4 であり，表 1 では，クラスタ数を 5 個から 4 個に併合するまでは $\Delta\mathrm{AIC}_{\mathrm{min}}$ の値が負となり，4 個以下に併合する場合には $\Delta\mathrm{AIC}_{\mathrm{min}}$ の値が正であることが分かる.

また，表 2 は，提案手法において SOM の初期状態を変化させて，上記のクラスタ数推定を 100 回行ったときの，推定されたクラスタ数とその頻度である．人工データに対しては，100 回中 99 回の試行において，正しいクラスタ数（=4）が推定されている．なお，同様のクラスタ数推定を Iris データに対して行った場合，100 回中 81 回の試行において Iris のカテゴリ数に等しいクラスタ数（=3）が推定された．ただし，クラスタリングの観点から，Iris データを 3 クラスタとみなすことの妥当性については議論の余地がある [6].

以上のことから，ΔAIC の符号に注目し，その値が負となるクラスタ候補のペアが存在しなくなった段階でクラスタ併合を止めることで，クラスタ数の推定と各クラスタの抽出を同時に実現できることが分かった.

表 1: ΔAIC の最小値の変化

Number of clusters						
8→7	7→6	6→5	5→4	4→3	3→2	2→1
$\Delta\mathrm{AIC}_{\mathrm{min}}$						
-113	-110	-102	-97.6	36.7	317	747

表 2: 推定クラスタ数の頻度

	推定されたクラスタ数の頻度			
	2 クラスタ	3 クラスタ	4 クラスタ	5 クラスタ
人工データ	0	1	**99**	0
Iris データ	3	**81**	2	14

4. まとめ

本稿では，SOM を用いたクラスタリング手法に対して情報量規準を適用することを検討し，SOM によって得られた初期クラスタ候補を，情報量規準 AIC に基づいて選択的に併合して行くクラスタリング手法を提案した．人工データおよび UCI の Iris データセットを用いたクラスタリング実験から，情報量規準を用いることによって，SOM を用いたクラスタリング手法において，しきい値設定を行うことなくクラスタ抽出ができること，および，クラスタ数の自動推定が可能であることが確認できた.

今後は，さまざまなデータセットに対して，さらに検証を重ねることが課題である.

参考文献

[1] 寺島幹彦, 白谷文行, 山本公明, 自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法, 電子情報通信学会論文誌, Vol.J79-D-II, No.7, pp.1280–1290, 1996.

[2] D. Pelleg, and A. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proc. of the 17th International Conference on Machine Learning, pp.727–734, 2000.

[3] 小西貞則, 北川源四郎, 情報量規準, 朝倉書店, 2004.

[4] T. Kohonen: Self-Organizing Maps, Springer-Verlag, 1995.

[5] UCI Machine Learning Repository, http://www.ics.uci.edu/˜mlearn/MLRepository.html

[6] D.-W. Kim, K.H. Lee and D. Lee, Fuzzy Clustering Validation Index based on Inter-cluster Proximity, Pattern Recognition Letters, Vol.24, pp.2561–2574, 2003.

外来待ち時間と患者満足度、および入院期間と患者満足度との関連

奥田益美[1) 2)]、安田晃[3)] 津本周作[3)]

1) 松江赤十字病院、2) 島根大学医学系研究科、3)島根大学医学部医学科医療情報学講座

1. はじめに

「患者満足」は「顧客満足」から発展し、消費者である患者の満足に注目した概念であり、米国では 1950 年代から研究が報告されている。患者満足度調査は、日本でも患者確保対策や患者中心の医療への関心の高まりなどから報告が多数みられるようになってきた。しかし医療は生命に直接関与し、「患者にとって一時的に不快な状況」が「患者が健康を取り戻す」、あるいは「その人にとって最良の状態に達することを目指す」構造となっているため、一般のサービスとは異なる満足の捉え方が必要である。

Donabedian はケアの質評価の枠組みを「構造」「過程」「成果」の 3 つで構成した。「構造」は専門職者の人数や経験年数、医療設備、病院環境等、「過程」は診察治療や看護の技術、接遇等、「成果」は 5 年生存率などの有効性、および合併症や転落転倒などの安全性に代表される「健康の質」と、患者の視点での「患者満足」に大別される。サービスとしての医療の特徴は、情報の非対称性、無形性、サービスの供給と需要の同時性など他のサービスと同様の特徴のほか、継続性や状況の不確実性が生命と直結しているため、消費者の不安と恐怖が大きいこと、価値性が高いため競争原理が働きにくいこと、生命に直結するからこそ受け手側は供給者が不快になるような反応はしづらいことなども特徴である。

患者満足度調査は、量的な質問紙調査と面接や投書内容の分析などに代表される質的調査に大別される。質問紙調査の分析では、全体的な患者満足度や「再度受診する場合この病院を選びますか」「家族や友人に紹介しますか」などの行動意志尺度を従属変数とした、重回帰分析、共分散構造分析、ロジスティック回帰分析などが多く行われているが、外来患者と入院患者の質問項目の構造の定量的な分析による比較や、外来待ち時間と患者満足度との関連、入院患者における入院期間と満足度との関連を調査した研究はほとんど見られなかったので、今回明らかにしたいと考えた。

2. 調査方法

平成 15 年度より松江赤十字病院で行われている無記名自記式質問紙調査結果のうち、平成 23 年 3 月 1 週間のデータを使用した。同病院は病床数 645 床、1 日平均外来患者数 760 名、1 日平均入院患者数 580 名、平均在院日数 14.9 日、病床利用率 90.5%、職員数 1100 名の地方の大規模総合病院である。外来では通路等に質問紙と回収箱を設置し、病棟では新入院患者には入院受付で、入院中患者には病室で質問紙を配布し、ナースステーションに回収箱を設置回収した。

＜**調査内容**＞属性として性別、年代、外来患者では予約時間から診察等までの待ち時間：30 分未満、30～60 分、60～90 分、90 分以上、入院患者では入院期間:7 日未満、7 日～13 日、14 日～1 ヵ月、1 ヵ月以上、満足度は、外来患者では接遇項目に加え「プライバシー配慮」「診察治療」「職員説明」「案内表示」「全体として満足(以下全体満足)」の 13 項目、入院患者では「食事」「病院環境」の 2 項目を追加した 15 項目を分析に使用した。満足度評価は「非常に満足」「満足」「普通」「やや不満」「非常に不満」の 5 件法であった。

＜**分析方法**＞満足者は「非常に満足」「満足」を選択した患者とし、記述統計、比率の差の検定を行った。質問項目間の類似構造は「非常に満足」を「満足度高」、「満足」を「満足度中」、「普通」以下を「満足度低」とし、対応分析、ウォード法によるクラスタ分析で検討した。「全体満足」への影響要因はロジスティック回帰分析により検討した。統計学的な有意差は $p<0.05$ で有意とした。

3. 結果

＜**配布数と回収数**＞外来では配布 650 枚、回収 301 枚、有効回答 229 枚（有効回答率 35.2%）、入院では配布 995 枚、回収 284 枚、有効回答 198 枚（有効回答率 19.9%）であった。

＜**属性**＞外来患者では男性 59 名、女性 84 名、性別不明 86 名、年代は 40 歳未満および不明 51 名、40 歳代 28 名、50 歳代 35 名、60 歳代 54 名、70 歳以上 61 名、待ち時間別では、30 分未満 92 名、30～60 分 82 名、60～90 分 37 名、90 分以上 18 名だった。入院患者では男性 100 名、女性 82 名、不明 16 名、年代は 40 歳未満および不明 36 名、40 歳代 11 名、50 歳代 23 名、60 歳代 43 名、70 歳以上 85 名、入院期間別では 7 日未満 43 名、7 日～13 日 50 名、14 日～1 カ月 46 名、1 カ月以上 59 名だった。

＜**満足者比率の推移**＞外来では待ち時間が長いと満足者比率の低下傾向が認められたのに対し、入院では入院 7 日～13 日で一度上昇後緩やかに低下または横ばい傾向で、外来に比較し分散が小さい傾向が認められた（図 1・2）。外来で特に低評価だったのは「困っていると声がけ」「案内表示」、入院では「食事」「案内表示」だった。待ち時間または入院期間別「全体満足」と評価での対応分析の結果から、外来は待ち時間 60 分未満と 60 分以上の 2 群、入院は 7 日未満、7 日～1 ヵ月、1 ヵ月以上の 3 群で分析した。

＜**質問項目間の類似構造**＞外来のクラスタ分析では、待ち時間 60 分未満では「案内表示」1 項目、診察治療過程に関する項目、個別の接遇項目、不特定多数の接遇項目の 4 クラスタが形成された。待ち時間 60 分以上では接遇項目と診察治療過程項目に大別された（図 3・4）。入院のクラスタ分析では、入院 7 日未満では「案内表示」「食事」とその他の項目で 2 クラスタを形成し、以後は前述 2 項目に加え「病院環境」とその他の項目で 2 クラスタを形成した（図 5・6・7）。

＜**「全体満足」への影響要因**＞外来では「プライバシー配慮」「案内表示」「診察治療」「困っていると声がけ」が選択され、的中判別率は 78.9%であった（$p<0.01$）。入院では「プライバシー配慮」「傾聴」「診察治療」が選択され、的中判別率は 92.9%であった（$p<0.01$）。

＜**外来と入院の満足者比率の比較**＞全項目で外来患者は有意に入院患者より満足者比率が低かった（$p<0.01$）。

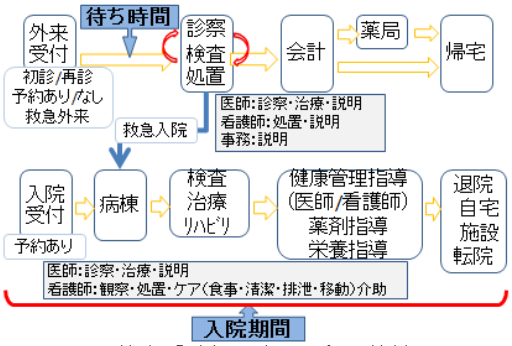


図8 外来受診と入院の過程の比較

4. 考察

外来では待ち時間により質問構造に変化が見られた。待ち時間 60 分未満では診察治療過程、案内表示、不特定多数の相手に対する接遇、個人に対する接遇に分類されたのに対し、待ち時間 60 分以上では診察治療過程と接遇に分類された。入院では環境とその他に分類され、入院期間が長くなってもその構造に大きな変化はなかった。外来では診察治療とその過程が重要視され、過程では説明や病院内移動の必要性からサービスと認識されやすく接遇と待ち時間が重要となるのに対し、入院は生活そのものであり、快適に過ごすための環境が重視されるという外来と入院の特徴（図 8）が結果に現れたと考えられた。

5. 結論

地方の大規模総合病院で外来および入院患者に自記式質問紙調査法による患者満足度調査を行い、待ち時間別、入院期間別の質問項目の定量的な分析による外来と入院の比較を行った。病院には「健康の質」を高めるために受診または入院するが、「患者満足」を高めるためには、外来では診察治療とその過程におけるサービス、入院では生活の快適性が重要と示唆された。医療サービスは医療側に力が偏在し「他者指向性」が不足しているため、この視点でのサービス改善が満足度を高めると思われる。

今回の質問紙は項目が偏り医療サービスの特徴を捉えていないため、質評価の分類や枠組み、医療のプロセスに沿ってバランスのとれた項目を用意し配布回収方法も検討する必要がある。さらには医療者と患者で非対称性である「患者満足」をどう比較するかが今後の課題である。

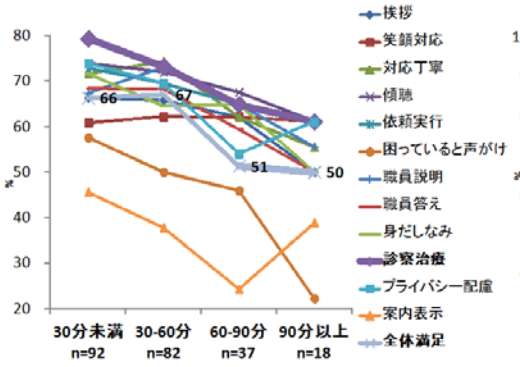


図1 外来患者 待ち時間別満足者比率推移 n=229

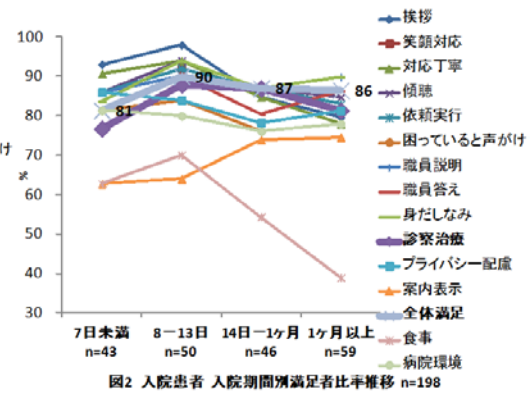


図2 入院患者 入院期間別満足者比率推移 n=198

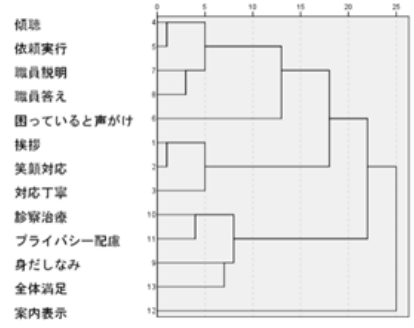


図3 待ち時間60分未満患者 クラスタ分析 n=174

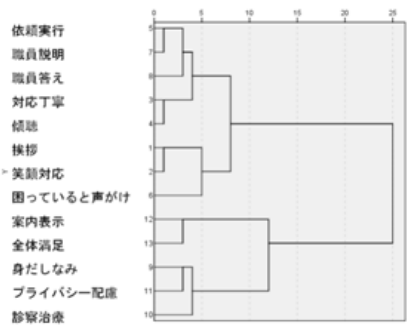


図4 待ち時間60分以上患者 クラスタ分析 n=55

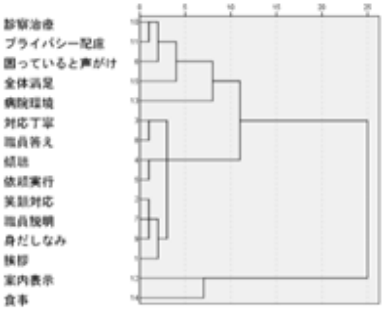


図5 入院7日未満患者 クラスタ分析 n=43

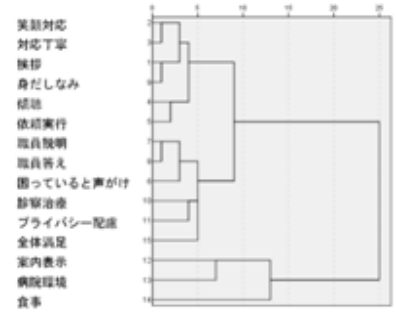


図6 入院7日－1ヵ月患者 クラスタ分析 n=96

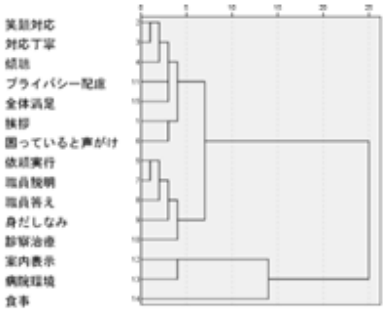


図7 入院1ヵ月以上患者 クラスタ分析 n=59

テンジククルマエビの成長と生存のモデルから導かれる非対称混合分布で検出された淡水流入効果

慶應義塾大学　仲 真弓　　慶應義塾大学　柴田 里程

1　はじめに

2002 年から 2005 年にかけ，オーストラリア・クイーンズランド州の Fitzroy 川，Calliope 川，Boyne 川の河口において淡水の流入が及ぼす生態系への影響を調べるため大規模な調査が行われた．すでに，調査データを利用してテンジククルマエビやバラマンディなど代表的な生物についてさまざまな側面からの解析が行われているが (Halliday *et al*, 2007)，あまりはっきりした結果が得られているとはいえない．そこで，Calliope 川で観測されたデータに対し，テンジククルマエビの成長と生存率のモデルを合わせて得られる非対称混合分布を用いて，淡水流入の効果が塩分濃度の変化を介した効果以外に存在するかどうか検証し，本研究集会で結果を報告した．

2　モデルの構築

テンジククルマエビは，海で生まれ，幼少期から河口で数か月滞在し，再び海に戻って産卵するという，約１年のライフサイクルを持つ．今回の調査では，新月または満月の時期に２週間，あるいは４週間おきに河口において捕獲されたテンジククルマエビの甲殻の長さ（単位は 1mm，小数点以下は切り捨て) が記録されている．そこで，ある時点で観測された集団の甲殻の長さの分布は，ひとつ前の時点で観測された集団の分布が成長と生存によって変化した分布と，新たに河口へ移動してきたいくつかの集団の分布の混合分布であると考え，モデルを構築した．

2.1　テンジククルマエビの成長のモデルと生存率のモデル

成長に関しては，Staples & Heales (1991) が水槽での生育実験の結果から求めた，水温と塩分濃度を説明変数としたテンジククルマエビの脱皮の間隔と脱皮による甲殻の長さの増分のモデルを用いた．ただし，甲殻の長さの増分に関するモデルについては，塩分濃度に関する項の係数をそのまま用いると，論文内の他の箇所の結果と矛盾することが明らかになったため，その項の係数を修正し，

$$\begin{aligned} t_m - t_{m-1} &= 13.919 - 0.411T + 0.027(T-25)^2 - 0.014S + 0.001(S-30)^2 + 0.201x_{m-1} \\ x_m - x_{m-1} &= 0.039 + 0.012T - 0.002(T-25)^2 - 0.00126S - 0.0004(S-30)^2 + 0.023x_{m-1} \end{aligned}$$

を，それぞれ脱皮の間隔と長さの増分のモデルとして本研究で用いた．ただし，t_m を m 回目の脱皮を行った日，x_m を m 回目の脱皮を行った後の甲殻の長さ，T を水温, S を塩分濃度 (‰) とする．

つぎに生存率のモデルとして，Wang & Haywood(1999) がオーストラリア北方の Carpentaria 湾で行なった調査データから求めた，テンジククルマエビの甲殻の長さに依存した生存率のモデルを利用した．そのモデルは甲殻の長さが x のときの瞬間の死亡率が $\alpha e^{\beta x}$ で与えられるという指数ハザードモデルでしかないが，実際にデータにはよくあてはまるモデルであると報告されている．このモデルから，成長率 g(mm/週間) が一定のとき，時刻 t_0（単位は１週間）における長さを x_{t_0} とすると，時刻 $t_0+\delta > t_0$ までの生存率が

$$q(x_{t_0}, \delta, g) = \exp\left(-\frac{\alpha}{\beta g}\left(e^{\beta(x_{t_0}+g\delta)} - e^{\beta x_{t_0}}\right)\right)$$

で与えられる．以降 α と β の値に関しては，論文内で与えられた推定値 $\hat{\alpha} = 1.594, \hat{\beta} = -0.2919$ を用いた．

2.2 二つの種類の集団の分布の混合分布

はじめに成長のモデルと生存率のモデルを用いて，河口に滞在していた集団の分布を考えた．成長のモデルから，成長率が甲殻の長さによらず一定と近似できることが導かれるため，各観測日間の水温と塩分濃度から求めた成長率 g を用いれば，ある時点で甲殻の長さが密度関数 $f(x)$ をもつ分布に従っていて，その集団が δ 週間後まで河口に滞在していたときに従う分布のモデルは，

$$H(x) = c_1 \int^x f(y - g\delta)q(y, \delta, g)dy \tag{1}$$

と書ける．

つぎにもう一つの集団のモデルとして，新たに海から移動してきた集団について考えた．ここで，生まれた直後の甲殻の長さは正規分布 $\mathrm{N}(\mu_0, \sigma^2)$ に従うと仮定すると，生まれてから δ 週間後の長さを x，海での成長率を g_0，μ_0 が十分小さいと仮定し $e^{-\beta\mu_0} \approx 1$ と近似とすれば，この集団の甲殻の長さの分布のモデルは $\delta' = \delta + \frac{\mu_0}{g_0}$ を用いて，

$$G(x, \delta', \sigma) = c_2 \int^x \phi\left(\frac{y - g_0\delta'}{\sigma}\right) q(y, \delta', g_0)dy \tag{2}$$

と書ける．水温 30 ℃，塩分濃度 35 ‰のときの成長率は 1.02 と計算できるので，以後 $g_0 = 1$ を仮定した．

このように，河口に滞在していた集団の分布 (1) と新たに海から移動してきた集団の分布 (2) が得られたので，これらの混合分布を考えた．新たに海から移動してきた分布に関しては，観測されたデータから，一つまたは二つの異なる δ' を持つ集団が移動してきたと考え，ある時点で観測された甲殻の長さの分布のモデルとして，

$$pH(x) + p_1G(x, \delta_1', \sigma_1) + p_2G(x, \delta_2', \sigma_2)$$

を考えた．ただし，河口に滞在していた割合にあたる p には河口から海へ戻る割合も含まれている．

3 結果

今回十分データのある 19 ケースに対して Cramer-von Mises 距離を最小にするようパラメータ $p, p_1, \delta_1', \delta_2', \sigma_1, \sigma_2$ を推定し，得られたパラメータを用いて Cramer-von Mises 統計量による適合度検定を行ったところ，全てのケースをモデルで説明することができた．このことから，河口における淡水の流入の影響が，塩分濃度の変化を通じてテンジククルマエビの成長に影響をあたえていることが明らかになった．さらに，推定したパラメータの値からも淡水の流入による影響を見ることができ，河口に滞在していた集団の分布の割合を決定する p の値は，淡水の流入時期とその量に大きく関係していることも判明した．

参考文献

Halliday, I. *et al.* (2007). Environmental flows for sub-tropical estuaries: understanding the freshwater needs of estuaries for sustainable fisheries production and assessing the impacts of water regulation, Final Report FRDC Project No. 2001/022 Coastal Zone Project FH3/AF

Wang, Y.G. and Haywood, MDE. (1999). Size-dependent natural mortality of juvenile banana prawns Penaeus merguiensis in the Gulf of Carpentaria, Australia, *Marine and freshwater research*, **50**, 313-317.

Staples, D.J. and Heales, D.S. (1991). Temperature and salinity optima for growth and survival of juvenile banana prawns *Penaeus merguiensis*, *Journal of Experimental Marine Biology and Ecology*, **154**, 251-274.