シンポジウム「統計科学の新展開]」報告書

本シンポジウムは科学研究費・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」(代表：谷口正信（早稲田大学基幹理工学部)）の助成により開催された。

1.開催日時：2013 年 11 月 27 日（水）～11 月 29 日（金）
2.会場： 金沢大学サテライト・プラザ 3 階集会室

下記の目次には講演開始日時，講演者，講演者所属，頁，演題のみ掲載し、共同研究者の情報は省く。

目次

A necessary test for elliptical symmetry based on uniform distribution over Stiefel manifold

東京理科大学理工 · 教養

岩下 登志也

Karlsruher Institut für Technologie

Bernhard Klar

1. はじめに

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

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

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

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

と表される. ここに,

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

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

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

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

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

$S^{-1/2}$ は S の対称平方根の逆行列を表す。

2. 主たる理論的結果

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

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

と Iwashita and Klar (2013) の結果により次のような結果を得た.

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

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

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

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

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

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

3. 提案する検定手法と数値実験結果

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

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

として

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

とすると, Theorem 2 の結果から, $\boldsymbol{X}_j^{(k)} \sim \mathrm{EC}_p(\boldsymbol{\mu}, \Lambda)$ $(j = 1, \ldots, N; k = 1, \ldots, m)$ ならば U_k は独立に Stiefel 多様体 $\mathcal{O}(n, p)$ 上の一様分布に従うことになるので Jupp (2001) が提案した Stiefel 多様体上の一様分布に関する検定法を応用して, 検定手順を構成した. 提案する検定法に基づいて数値実験を実行した結果, 楕円母集団の下で, 提案した検定手法が有効であるが, 非楕円母集団の下で, 検出力を有しないことが判明した. この原因については, 今後の研究課題と考えている.

参考文献

Iwashita, T. and Klar, B.(2013). The joint distribution of Studentized residuals under elliptical distributions. *submitted*

Jupp, P. E.(2001). Modification of the Rayleigh and Bingham tests for uniformity of directions. *J. Multivariate Anal.*, **77**, 1–20.

Asymptotic properties of Bayesian type estimators without assuming that the Hessian matrices of log-likelihood functions converge

宮田　庸一（高崎経済大学）

ベイズ型推定量の一致性，漸近正規性などの漸近的な性質は多くの著者 (Van der Vaart 1998, Ibragimov and Hasminskii 1981, Ferguson 1996, Basawa and Rao 1980, Yoshida 2007) により研究されてきた．一般的に, 多くの研究者は Bernstein-von Mises theorem を経由してベイズ型推定量の漸近的な性質を示すために, 標本数が無限大に発散するにつれ, 対数尤度のヘシアン行列が正値定符号行列に収束することを仮定する必要があった．しかしながら White (1980, pp.727-728) で指摘されているように，前述されている内容を保障するために定常性を仮定することはしばしば強い良い条件となり，そして対数尤度のヘシアンが収束しない例も容易に作ることができる．$'$ は行列の転置を表す記号とする．ここで大きさ n の観測ベクトル $\tilde{\mathbf{Y}}$ が与えられた下での，確率パラメーター $\tilde{\boldsymbol{\Theta}} \equiv (\tilde{\Theta}_1, ..., \tilde{\Theta}_d)'$ の事後密度型関数を考える:

$$p_n(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{exp\{-nQ_n(\tilde{\mathbf{Y}}, \boldsymbol{\theta})\}b(\boldsymbol{\theta})}{\int_\Theta exp\{-nQ_n(\tilde{\mathbf{Y}}, \boldsymbol{\theta})\}b(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{1}$$

ここで $Q_n(\tilde{\mathbf{Y}}, \boldsymbol{\theta})$ を $Q_n(\boldsymbol{\theta})$ と略記し，$L_n(\boldsymbol{\theta})$ を $\boldsymbol{\theta}$ の対数尤度関数, $\pi(\boldsymbol{\theta})$ を $\tilde{\boldsymbol{\Theta}}$ の事前分布とするとき, (1) 式における $b(\boldsymbol{\theta})$ と $Q_n(\boldsymbol{\theta})$ の選び方は無数にある．例えば, $Q_n(\boldsymbol{\theta}) = -(1/n)\log L_n(\boldsymbol{\theta})$, $Q_n(\boldsymbol{\theta}) = -(1/n)\log L_n(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ である．特に後者の選び方においては，事前分布は観測ベクトル $\tilde{\mathbf{Y}}$ および説明変数（外生変数）に依存することができる．そのとき (1) の下でベイズ型推定量は以下の式で与えられる．

$$\boldsymbol{\theta}_n^B \equiv \left(\frac{\int_\Theta \theta_1 exp\{-nQ_n(\boldsymbol{\theta})\}b(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta exp\{-nQ_n(\boldsymbol{\theta})\}b(\boldsymbol{\theta})d\boldsymbol{\theta}}, ..., \frac{\int_\Theta \theta_k exp\{-nQ_n(\boldsymbol{\theta})\}b(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta exp\{-nQ_n(\boldsymbol{\theta})\}b(\boldsymbol{\theta})d\boldsymbol{\theta}} \right)'. \tag{2}$$

Hanousek (1990) は観測ベクトルが i.i.d. の場合に, ベイズ型推定量と $Q_n(\boldsymbol{\theta})$ により生成される最小コントラスト推定量は漸近的に同等であることを示した.

当日の報告では, $Q_n(\boldsymbol{\theta})$ のヘシアン行列が正値定符号行列に収束することを仮定をせずにベイズ型推定量が強一致性，漸近正規性を持つための十分条件を与えた．尚, 本報告においてはモデルが誤特定されている場合も考慮にいれているため, ベイズ型推定量と擬真値の差がほとんどいたるところ 0 に収束することを強一致性の定義として採用した．さらにいくつかの条件と観測ベクトルが α-mixing（定常である必要はない）であることのもとで，ベイズ型推定量が上記の漸近的性質を満たすことも併せて報告をした.

またロジットモデルを用いて，説明変数の振る舞いによっては対数尤度関数のヘシアンが収束しない例を与え，さらに Gourieroux and Monfort (1981) の条件を用いることでベイズ型推定量が強一致性，漸近正規性を持つための十分条件を与えた.

最後に非等分散な攪乱項を持つ AR(1) モデルにおいて, α-mixing 性を満たすための十分条件を与え，そしてあるコントラスト関数を用いたベイズ型推定量が弱い条件のもとで強一致性，漸近正規性を持つことを報告した.

尚，本報告に関して最小コントラスト推定量の漸近的性質に関する質問, 擬真値に関する質問，漸近有効性に関する質問，非等分散 AR(1) モデルにおける α-mixing に関する質問があった.

Third-order average local powers of Bartlett-type adjusted tests: Ordinary *versus* adjusted profile likelihood

柿沢 佳秀 (北大経済)

1. 設定と記号　d_X-次元の密度関数モデル $\mathcal{P} = \{f(\mathbf{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^p\}$ で未知の母数ベクトルを $\boldsymbol{\theta} = (\boldsymbol{\theta}'_{(1)}, \boldsymbol{\theta}'_{(2)})'$, $\boldsymbol{\theta}_{(1)} = (\theta_1, \ldots, \theta_{p_1})'$, $\boldsymbol{\theta}_{(2)} = (\theta_{p_1+1}, \ldots, \theta_p)'$ のように分割し, 帰無仮説 $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(1)0}$ を検定する. $\ell_{j_{i1}\ldots j_{iR_i}}(\mathbf{x},\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_{j_{i1}}} \cdots \frac{\partial}{\partial\theta_{j_{iR_i}}} \log f(\mathbf{x},\boldsymbol{\theta})$, $i = 1, \ldots, v$ の $f(\mathbf{x},\boldsymbol{\theta})$ に関する v 次キュムラントを $\nu_{j_{11}\ldots j_{1R_1},\ldots,j_{v1}\ldots j_{vR_v}}(\boldsymbol{\theta})$ と書き, $Z_j^{(N)}(\boldsymbol{\theta}) = N^{-1/2}\sum_{i=1}^N \ell_j(\mathbf{X}_i,\boldsymbol{\theta})$, $Z_{j_1j_2}^{(N)}(\boldsymbol{\theta}) = N^{-1/2}\sum_{i=1}^N \{\ell_{j_1j_2}(\mathbf{X}_i,\boldsymbol{\theta}) - \nu_{j_1j_2}(\boldsymbol{\theta})\}$ などを定義する. このとき, 情報量行列 $[\nu_{j,k}(\boldsymbol{\theta})] = -[\nu_{jk}(\boldsymbol{\theta})]$ とスコアベクトルを分割し,

$$[\nu_{j,k}(\boldsymbol{\theta})] = \begin{pmatrix} \boldsymbol{\nu}_{(1,1)}(\boldsymbol{\theta}) & \boldsymbol{\nu}_{(1,2)}(\boldsymbol{\theta}) \\ \boldsymbol{\nu}_{(2,1)}(\boldsymbol{\theta}) & \boldsymbol{\nu}_{(2,2)}(\boldsymbol{\theta}) \end{pmatrix}, \quad [Z_j^{(N)}(\boldsymbol{\theta})] = \begin{pmatrix} \mathbf{Z}_{(1)}^{(N)}(\boldsymbol{\theta}) \\ \mathbf{Z}_{(2)}^{(N)}(\boldsymbol{\theta}) \end{pmatrix}$$

と書く. ここで, パラメータ直交性, すなわち, 『任意の $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ に対し $\boldsymbol{\nu}_{(1,2)}(\boldsymbol{\theta}) = \mathbf{O}_{p_1,p_2}$』を仮定しない. $p \times p_1$ 行列 $[\mathcal{G}_{j,a}(\boldsymbol{\theta})] = \begin{pmatrix} \mathbf{I}_{p_1} \\ -\boldsymbol{\nu}_{(2,2)}^{-1}(\boldsymbol{\theta})\boldsymbol{\nu}_{(2,1)}(\boldsymbol{\theta}) \end{pmatrix}$ を定義する. 以下, 添え字については, $j_1, \ldots, j_R \in \{1, \ldots, p\}$ (j,k も同様), $a_1, \ldots, a_R \in \{1, \ldots, p_1\}$ (a,b も同様), $r_1, \ldots, r_R \in \{p_1+1, \ldots, p\}$ (r,s も同様) と約束し, それらの和 $\sum_{j_1,\ldots,j_R=1}^p$, $\sum_{a_1,\ldots,a_R=1}^{p_1}$, $\sum_{r_1,\ldots,r_R=p_1+1}^p$ に対しては Einstein 表記法を採用する.

2. 通常の尤度解析　対数尤度 $\mathcal{L}^{(N)}(\boldsymbol{\theta}) = \sum_{i=1}^N \log f(\mathbf{X}_i,\boldsymbol{\theta})$ を最大化し, (Unrestricted) MLE $\widehat{\boldsymbol{\theta}}^{(N)}$ 及び Restricted MLE $\widetilde{\boldsymbol{\theta}}^{(N)}$ での評価は, ハット/チルド記号を付ける;

$$\left.\begin{array}{ll} \mathrm{LR}^{(N)} = 2(\widehat{\mathcal{L}}^{(N)} - \widetilde{\mathcal{L}}^{(N)}), & \mathrm{grad}^{(N)} = (\widetilde{\mathbf{Z}}_{(1)}^{(N)})' N^{1/2}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)} - \boldsymbol{\theta}_{(1)0}) \\ \mathrm{R}^{(N)} = (\widetilde{\mathbf{Z}}_{(1)}^{(N)})' \widetilde{\boldsymbol{\nu}}_{(11\cdot2)}^{-1} \widetilde{\mathbf{Z}}_{(1)}^{(N)}, & \mathrm{W}^{(N)} = N(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)} - \boldsymbol{\theta}_{(1)0})' \widehat{\boldsymbol{\nu}}_{(11\cdot2)} (\widehat{\boldsymbol{\theta}}_{(1)}^{(N)} - \boldsymbol{\theta}_{(1)0}) \\ \mathrm{MR}^{(N)} = (\widetilde{\mathbf{Z}}_{(1)}^{(N)})' \widehat{\boldsymbol{\nu}}_{(11\cdot2)}^{-1} \widetilde{\mathbf{Z}}_{(1)}^{(N)}, & \mathrm{MW}^{(N)} = N(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)} - \boldsymbol{\theta}_{(1)0})' \widetilde{\boldsymbol{\nu}}_{(11\cdot2)} (\widehat{\boldsymbol{\theta}}_{(1)}^{(N)} - \boldsymbol{\theta}_{(1)0}) \end{array}\right\} \xrightarrow{d} \chi^2_{p_1}.$$

3. 修正プロファイル尤度解析　各 $\boldsymbol{\theta}_{(1)}$ に対しプロファイル対数尤度は

$$\mathcal{L}^{\mathrm{P}(N)}(\boldsymbol{\theta}_{(1)}) = \mathcal{L}^{(N)}(\check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)}))$$

で定義される. ここに, $\check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)}) = \begin{pmatrix} \boldsymbol{\theta}_{(1)} \\ \check{\boldsymbol{\theta}}_{(2)}^{(N)}(\boldsymbol{\theta}_{(1)}) \end{pmatrix}$, $\check{\boldsymbol{\theta}}_{(2)}^{(N)}(\boldsymbol{\theta}_{(1)}) = \arg\max_{\theta_{(2)}} \mathcal{L}^{(N)}(\boldsymbol{\theta})$ であり, $\widehat{\boldsymbol{\theta}}^{(N)} = \check{\boldsymbol{\theta}}^{(N)}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)})$ と $\widetilde{\boldsymbol{\theta}}^{(N)} = \check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)0})$ に注意すれば, 上記の $\mathrm{LR}^{(N)}$ は

$$\mathrm{LR}^{(N)} = 2\{\mathcal{L}^{\mathrm{P}(N)}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)}) - \mathcal{L}^{\mathrm{P}(N)}(\boldsymbol{\theta}_{(1)0})\} \xrightarrow{d} \chi^2_{p_1}$$

と書け, 局外母数を含む場合の尤度比検定は "プロファイル尤度解析" と解釈されうる. しかし, $\mathcal{L}^{\mathrm{P}(N)}(\boldsymbol{\theta}_{(1)})$ は "正真正銘な対数尤度"でなく, 『$E^{(N)}_{\theta_{(1)0},\theta^\dagger_{(2)}}[\frac{\partial}{\partial\theta_a}\mathcal{L}^{\mathrm{P}(N)}(\boldsymbol{\theta}_{(1)0})] = 0$』とならないため, Stern (1997; JRSS) は修正プロファイル対数尤度を定義している;

$$\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)}) = \mathcal{L}^{\mathrm{P}(N)}(\boldsymbol{\theta}_{(1)}) - N^{-1/2}\check{\mathbf{M}}'\check{\boldsymbol{\mathcal{G}}}\check{\boldsymbol{\nu}}_{(11\cdot2)}^{-1}\check{\mathbf{Z}}_{(1)}^{(N)} \quad [\text{ここに, } \check{Q} = Q(\check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)}))].$$

なお, $M_j(\boldsymbol{\theta}) = \frac{1}{2}\{\nu_{rr',j}(\boldsymbol{\theta}) + \nu_{r,r',j}(\boldsymbol{\theta})\}\nu^{r,r'}_{(2,2)}(\boldsymbol{\theta})$ は局外母数を含む高次漸近論の文脈でお馴染みのものであり, この量が消えないとき $\mathrm{LR}^{(N)}$ は一般的に 2 次漸近不偏でない (Hayakawa, 1975; Bio). 本報告では, 2 節の通常の統計量とその修正プロファイル版

$$
\begin{aligned}
\mathrm{LR}^{\mathrm{AP}(N)} &= 2\{\mathcal{L}^{\mathrm{AP}(N)}(\overline{\boldsymbol{\theta}}^{(N)}_{(1)}) - \mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})\}\,,\\
\mathrm{grad}^{\mathrm{AP}(N)} &= [\tfrac{\partial}{\partial\theta_a}\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})][\overline{\boldsymbol{\theta}}^{(N)}_{(1)} - \boldsymbol{\theta}_{(1)0}]_a,\\
\mathrm{R}^{\mathrm{AP}(N)} &= N^{-1}[\tfrac{\partial}{\partial\theta_{a_1}}\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})]\nu^{a_1,a_2}_{(11\cdot2)}(\check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)0}))[\tfrac{\partial}{\partial\theta_{a_2}}\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})]\,,\\
\mathrm{W}^{\mathrm{AP}(N)} &= N(\overline{\boldsymbol{\theta}}^{(N)}_{(1)} - \boldsymbol{\theta}_{(1)0})'\boldsymbol{\nu}_{(11\cdot2)}(\check{\boldsymbol{\theta}}^{(N)}(\overline{\boldsymbol{\theta}}^{(N)}_{(1)}))(\overline{\boldsymbol{\theta}}^{(N)}_{(1)} - \boldsymbol{\theta}_{(1)0})\,,\\
\mathrm{MR}^{\mathrm{AP}(N)} &= N^{-1}[\tfrac{\partial}{\partial\theta_{a_1}}\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})]\nu^{a_1,a_2}_{(11\cdot2)}(\check{\boldsymbol{\theta}}^{(N)}(\overline{\boldsymbol{\theta}}^{(N)}_{(1)}))[\tfrac{\partial}{\partial\theta_{a_2}}\mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)0})]\,,\\
\mathrm{MW}^{\mathrm{AP}(N)} &= N(\overline{\boldsymbol{\theta}}^{(N)}_{(1)} - \boldsymbol{\theta}_{(1)0})'\boldsymbol{\nu}_{(11\cdot2)}(\check{\boldsymbol{\theta}}^{(N)}(\boldsymbol{\theta}_{(1)0}))(\overline{\boldsymbol{\theta}}^{(N)}_{(1)} - \boldsymbol{\theta}_{(1)0})
\end{aligned}
$$

の局所検出力を漸近展開で比較した. ここに, $\overline{\boldsymbol{\theta}}^{(N)}_{(1)} = \arg\max_{\theta_{(1)}} \mathcal{L}^{\mathrm{AP}(N)}(\boldsymbol{\theta}_{(1)})$.

4. (平均) 局所検出力　2,3 節の統計量は, それぞれ

$$
\begin{aligned}
T^{(N)} \approx{}& \widetilde{Z}^{(N)}_a\widetilde{\nu}^{a,b}_{(11\cdot2)}\widetilde{Z}^{(N)}_b + \frac{2}{N^{1/2}}\Big(\widetilde{C}^{\mathcal{G}\,\mathcal{G}\,\mathcal{G}}_{b_1b_2b_3}\prod_{i=1}^{3}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i} + \widetilde{C}^{\mathcal{G}\,\mathcal{G}}_{b_1b_2,k_1k_2}\widetilde{Z}^{(N)}_{k_1k_2}\prod_{i=1}^{2}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\Big)\\
&+ \frac{2}{N}\Big\{\widetilde{D}^{\mathcal{G}\,\mathcal{G}\,\mathcal{G}\,\mathcal{G}}_{b_1b_2b_3b_4}\prod_{i=1}^{4}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i} + (\widetilde{D}^{\mathcal{G}\,\mathcal{G}\,\mathcal{G}}_{b_1b_2b_3,k_1k_2}\widetilde{Z}^{(N)}_{k_1k_2} + \widetilde{D}^{\mathcal{G}\,\mathcal{G}\,\mathcal{G}}_{b_1b_2b_3,k_1k_2k_3}\widetilde{Z}^{(N)}_{k_1k_2k_3})\prod_{i=1}^{3}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\\
&\quad+\widetilde{D}^{\mathcal{G}\,\mathcal{G}}_{b_1b_2,k_1k_2,k_3k_4}\widetilde{Z}^{(N)}_{k_1k_2}\widetilde{Z}^{(N)}_{k_3k_4}\prod_{i=1}^{2}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\Big\}\,,\\
T^{\mathrm{AP}(N)} \approx{}& T^{(N)} + \frac{2}{N^{1/2}}\,\widetilde{M}^{\mathcal{G}}_{b_1}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_1}\\
&+ \frac{2}{N}\Big(\frac{1}{2}\,\widetilde{M}^{\mathcal{G}}_{b_1}\widetilde{\nu}^{b_1,b_2}_{(11\cdot2)}\widetilde{M}^{\mathcal{G}}_{b_2} + {}_M\widetilde{D}^{\mathcal{G}\,\mathcal{G}}_{b_1b_2}\prod_{i=1}^{2}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i} + {}_M\widetilde{D}^{\mathcal{G}}_{b_1,k_1k_2}\widetilde{Z}^{(N)}_{k_1k_2}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_1}\Big)
\end{aligned}
$$

と確率展開され, 以下の 2 形式の調整 (Kakizawa, 2012ab,2013; JMVA,SPL) を考えた;

$$
\begin{aligned}
\#^{\mathrm{GCF}(N)} &= \#^{(N)} + \frac{2}{N}\sum_{R=2,4,6}\widetilde{\Gamma}_{b_1\cdots b_R}\prod_{i=1}^{R}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\,,\\
\#^{\mathrm{GB}(N)} &= \#^{(N)} + \frac{2}{N^{1/2}}\,\widetilde{\Gamma}^{C}_{b_1b_2b_3}\prod_{i=1}^{3}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i} + \frac{2}{N}\Big\{(\widetilde{\Gamma}^{\star}_{b_1b_2b_3b_4} + \widetilde{\Delta}_{b_1b_2b_3b_4})\prod_{i=1}^{4}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\\
&\quad+\widetilde{\Delta}_{b_1b_2b_3,k_1k_2}\widetilde{Z}^{(N)}_{k_1k_2}\prod_{i=1}^{3}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i} + \widetilde{\Gamma}^{\star}_{b_1b_2}\prod_{i=1}^{2}[\widetilde{\boldsymbol{\nu}}^{-1}_{(11\cdot2)}\widetilde{\mathbf{Z}}^{(N)}_{(1)}]_{b_i}\Big\}\,.
\end{aligned}
$$

ただし, $P^{(N)}_{\theta_{(1)0},\theta^{\dagger}_{(2)}}[\#^{\mathrm{GCF}(N)} \le x]$, $P^{(N)}_{\theta_{(1)0},\theta^{\dagger}_{(2)}}[\#^{\mathrm{GB}(N)} \le x] = \Pr[\chi^2_{p_1} \le x] + \mathrm{o}(N^{-1})$ となるようにした (こうしないとサイズが揃わないために公平な検出力比較が不可能である). 局所対立仮説 $\boldsymbol{\theta}^{\dagger(N)} \equiv (\boldsymbol{\theta}'_{(1)0}, (\boldsymbol{\theta}^{\dagger}_{(2)})')' + N^{-1/2}(\mathbf{h}'_{(1)}, \mathbf{0}'_{p_2})'$ の下で検出力を 3 次まで導出して, 以下の平均基準から $\mathrm{R}^{\mathrm{GCF}(N)}$ または $(\mathrm{R}^{\mathrm{AP}})^{\mathrm{GCF}(N)}$ の "ある意味" の 3 次最適性を示した;

$$
\operatorname*{ave}_{S_\lambda}\{\pi(\mathbf{h}_{(1)})\} = \frac{\int_{S_\lambda}\pi(\mathbf{h}_{(1)})\,d\mathbf{h}_{(1)}}{\int_{S_\lambda}d\mathbf{h}_{(1)}}\,,\quad S_\lambda = \{\mathbf{h}_{(1)} \in \mathbf{R}^{p_1} : \mathbf{h}'_{(1)}\boldsymbol{\nu}_{(11\cdot2)}\mathbf{h}_{(1)} = \lambda\}\quad(\lambda > 0)\,.
$$

氏　　名：蛭川雅之

所　　属：摂南大学経済学部

講演題目：“Two-Sample Estimation of Varying Coefficient Models via Nearest Neighbor Matching” [Artem Prokhorov (University of Sydney)との共著]

講演内容：

最初に断っておくと、講演題目は可変係数回帰モデル(varying coefficient model, “VCM”)の推定理論に関するものであるが、複数のデータセットから最近隣マッチング(nearest neighbor matching, “NNM”)を応用して新たにデータセットを構築する場合の推定理論は一般になじみが薄いと予想された。そのため、本講演では、線形回帰モデルに関する推定理論を詳細に説明し、その理論の拡張として、VCM に関する推定理論を概説するという手順を採用した。

回帰モデルに関する推定理論は、モデルに含まれる変数全てが単一のデータセットから得られることを前提として構築されている。しかし、労働経済学・公共経済学などの実証分析では、回帰モデルの推定に必要となる変数が複数のデータセットにわたって存在するケースがしばしばあり、この場合、これらのデータセットから NNM を用いて分析に必要なデータセットを作成することが慣行として行われている。この慣行はさらに、回帰モデル右辺の説明変数全体が NNM でマッチさせたデータに置き換えられるケースと、説明変数の一部のみが置き換えられるケースとの２通りに分類可能である。分析の容易さから、今回は前者のみを考察の対象とした。

本講演では、まず、NNM で構築したデータセットを用いて線形回帰モデルの係数を最小二乗法により推定する場合、最小二乗(two-sample ordinary least squares, “TSOLS”)推定量は一致性を持たないことを示した。直感的には、被説明変数と説明変数とのミスマッチに起因するモデル定式化の誤りの結果、除外変数バイアス(omitted variable bias)に類似するバイアスが発生し、TSOLS 推定量は不一致となると説明される。

TSOLS 推定量は一致性を持たない半面、その確率極限は真値の一次変換として表現される。この一次変換に対する逆写像から構築される推定量は真値に対

して一致性を持つ。この発想は、Gouriéroux, Monfort and Renault (1993)および Smith (1993)の indirect inference 推定法の一例である。そのため、ここで提案されるバイアス修正推定量を two-sample indirect inference (“TSII”)推定量と呼ぶことにした。TSII 推定量の一致性、漸近正規性および漸近分散も同時に示された。さらに、モンテカルロ実験により、TSOLS 推定量の不一致性および TSII 推定量の一致性が数値的に確認された。TSII 推定量は一致性を持つ反面、全ての変数が単一のデータセットから得られると仮定した場合の実行不可能な OLS 推定量に比べて有効性は劣後する。これは複数のデータセットを利用することに伴う情報のロスに対するコストと解釈される。

この結果を発展させ、NNM で構築したデータセットを用いて VCM をカーネル平滑化により推定する場合の理論も得られた。具体的には、線形回帰モデルの場合と同様、可変係数の局所線形(local linear, “LL”)推定量は各点で一致性を持たないことが証明された。しかし、その確率極限がやはり真値の一次変換として表現されることに基づき、バイアスを修正したノンパラメトリック推定量を幾つかのバージョンで提案し、これらを総称して bias-corrected LL (“BCLL”) 推定量と呼ぶことにした。これら BCLL 推定量に関し、一致性、漸近正規性、バイアス項の近似式、および漸近分散も明示された。

最後に、今後の研究の方向が、(i) BCLL 推定量の実行法（バンド幅選択法）およびモンテカルロ実験、(ii) 回帰モデル右辺の説明変数の一部のみが NNM でマッチさせたデータに置き換えられるケースに対する推定理論の拡張、さらには、(iii) 実際のデータを用いた実証分析に移行する予定であることに触れ、本講演を締めくくった。特に、(ii)は、サーベイ・サンプリングにおいて変数の調査漏れが発生した場合に、どのように追加調査を行えばよいかを示唆する内容とも解釈できるため、この方面への研究拡張に注力したい。

参考文献：

Gouriéroux, C., A. Monfort, and E. Renault (1993): “Indirect Inference,” *Journal of Applied Econometrics*, 8, S85 - S118.

Smith, A. A., Jr. (1993): “Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions,” *Journal of Applied Econometrics*, 8, S63 - S84.

Empirical Credit Risk Analysis on Euro Government Bonds
– Term Structures of Default Probabilities–

Takeaki Kariya (GSGB, Meiji University) kariya@meiji.ac.jp
Yoshiro Yamamura(GSGB, Meiji University) yyama@meiji.ac.jp
Yoko Tanokura (AMS, Meiji University) tanokura@meiji.ac.jp
Zhu Wang（ZW System）

The Financial Crisis in 2008-2009 and the European Crisis thereafter made many European states (countries) with currency integration or equivalently member states of the European Economic and Monetary Union (EEMU) confronting severe budgetary and unemployment problems, as reflected in Greek economy, where as of May 2013, there are 27 countries in the EU (European Union) among which 17 states form the EEMU. The European problem is affecting global economies through trade relations and financial markets, and naturally the world concerns about their future movements, because a collapse of the EEMU would make a significantly serious impact on the world economy.

In this paper, via interest rate (IR) differential, GB price differential, default probability (DP) and CDS, we will make a comprehensive credit risk analysis on the GB price data of the Five States; Germany, France, Italy, Spain and Greece over 2007.4-2012.3. In order to derive term structures of interest rates (TSIRs), the forward-looking GB-pricing model proposed in Kariya (1993) and applied in Kariya, Wang, Wang, Doi and Yamamura (2012) (shortened KWWDY (2012)) is applied to each monthly cross-sectional set of GB prices. And to derive term structures of default probabilities (TSDPs) relatively to DGB, the forward-looking CB (Corporate Bond) pricing model proposed in Kariya (2013) is also applied to each monthly cross-sectional set of FGB, IGB, SGB, and GrGB, which are regarded as CBs in the model, while DCBs are viewed as non-defaultable reference GBs. In derivations of TSIRs, P-differentials and TSDPs, we use each monthly cross-sectional set of GB price data at the last business day of each month where the period of analysis is 2007.4-2012.3. Our arguments are sometimes made associated with the conditions in Maastricht Treaty, which we will discuss in Section 2 and show a legitimacy of regarding DGB as a reference GB.

First after making some discussions and observations on the business cycles of the Five States and the Maastricht convergence condition of IRs, we derived TSIRs via the bond-pricing model in KWWDY (2012) and compared them. In association with the budgetary conditions and business cycles of the Five States, we found that (a) in the sub-period 2007.4-2008.6 the TSIRs of the Five States moved together at the approximately same levels, meaning that no credit differentiation was found in IR-differentials, (b) in the sub-period 2008.7-2010.3 all the IRs tended to decrease gradually as trend movements but their IR-differentials exhibited some credit differentiation, (c) in the sub-period 2010.4-2011.12 the IR-differentials tended to diverge

and (d) after that they tended to get stabilized gradually. Though F-TSIRs moved together with D-TSIR before the Euro Crisis, then they deviate. This deviation would have made the role of the German Government more important for the stability of the EEMU.

Secondly we proposed what we call CRPS measure, which measures directly credit risk via P-differential between a given GB and DGB-equivalent in terms of euros. The CRPS measure is model-free once the mean discount function of Germany is estimated. In our empirical analysis the 10-year CRPSs of Italy and Spain in the mid of the Financial Crisis 2009.3 were respectively about -10 euros and -7 euros, but in the mid of the Euro Crisis 2011.12 they jumped down to about -35 euros and -30 euros. Since the CRPS measure is additive, it can be used to measure credit risk volume (in euros) of a bond portfolio.

Thirdly the TSDPs of the Four States were derived via CB-pricing model in Kaiya (2013) and compared. The model enabled us to transform the term structures of CRPSs into TSDPs, and it turned out that the 10 year CRPSs of Italy and Spain in 2011.12 respectively corresponded to about 35% and 30% where the recovery rate was assumed to be zero. In addition we substantiated the observations on credit differentiation obtained through TSIR analysis in terms of DPs. Also it was observed that the Financial Crisis did not affect the DPs of France, did increase the DPs of Italy, Spain and Greece to the levels of 6%, 10% and 20% respectively, but after that their DPs decreased. The time series movements of the TSDPs of the Four States were associated with business cycles, Financial Crisis and Euro Crisis.

Fourthly the time series relationships between IR-differentials and DPs of each maturity were shown to be strongly linear by our regression analysis, which enables us to convert IR-differentials into DPs for each maturity and vice versa. Consequently the Maastricht convergence condition can be stated in terms of DPs. It was observed that Italy, Spain and Greece did not meet the required condition in the period of the Euro Crisis though they are the members of the EEMU, creating the instability of the EEMU system that was an economic concern in the global world. But an explicit solution that President Van Rompuy (2012) planned for a genuine integration of the EEMU will stabilized the EEMU.

Finally we made compared CDS prices of each maturity to our TSDPs by term series regression and found that CDS prices were well explained by DP levels and slopes of TSDPs. Since the CDS prices are formed in a different market by different players, this result will show that our approach and model to deriving TSDP measures are effective. In addition the regression model will enable us to use for trading CDSs.

Overall, our empirical model analysis on credit risk of the main states of the EEMU will be effective and the results therein will be useful for decision making in credit investment and risk management.

Optimal Investment in Correlated Stocks and an Index Bond for Defined Contribution Pension

Ryou YAMAGATA

Graduate Program of Science and Technology
Mathematics Division
Faculty of Science and Technology, Sophia University

1 Introduction

The purpose of this research work is to derive the optimal investment plan to hedge risks and maximize the total wealth for defined contribution pension (DCP). The target of investments are domestic and foreign stocks, domestic and foreign bonds, Real Estate Investment Trust (REIT), insurance, trust, fund, and other financial products. We consider the problem where pension members invest in financial products including stocks and index bonds. On the assumption that pension members are not day-trader and their portfolios are rebalanced once a week at most, we derive a theoretical result of the optimal investment plan with numerical examples.

2 Theory of the optimal investment problem

2.1 Basic definition for optimal investment problem

Throughout the following discussion, we assume that the market is arbitrage-free and complete. Let $t \in [0, T]$ be continuous time variable provided that $t = 0$ is the starting time of DC plan, and $t = T$ is the terminal time of the pension. Then we define the stochastic price level as

$$\begin{aligned} \frac{dP(t)}{P(t)} &= i dt + \sigma_{00} dW_0(t) \\ P(0) &= p > 0 \end{aligned}$$

where the constant i is the expected rate of inflation and $W_0(t)$ is a Brownian Motion. The volatility of the price level is σ_{00}. And we have the following properties:.

1) The price of risk-free an bond with risk-free rate R is

$$\frac{dB(t)}{B(t)} = R dt$$

2) The price process of index bond with real return r satisfies

$$\frac{dI(t)}{I(t)} = (r + i)\, dt + \sigma_{00} dW_0(t)$$

3) The price process of n stocks are given by

$$\frac{dS(t)}{S(t)} = \mu dt + \Sigma\, dW_T(t)$$

where μ is the expected of return vector on the stocks, Σ is volatility matrix of stocks, and $W_T(t)$ is Brownian motion vector with $W_k \perp\!\!\!\perp W_j$ $(k \neq j,\ k, j = 0, 1, 2, \ldots, n)$

Then we denote the market price of risk by θ, which is described as a function of r, i, μ, R and Σ.

Moreover, define the salary process of the pension plan member by

$$\begin{aligned} \frac{dY(t)}{Y(t)} &= \kappa dt + \sigma_s dW(t) \\ \kappa = \gamma + i, \quad \sigma_s &= \sigma_y + \sigma_{00}, \quad Y(0) = y > 0 \end{aligned}$$

where γ is the expected growth rate of salary, σ_y is the volatility of salary, and both are constant. The salary process $Y(t)$ can be solved in the closed form by using Ito's Lemma.

Suppose that a DC pension member contributes to an index bond and n kinds of financial products such as stocks or funds with the fixed contribute rate c (> 0). Then the portfolio weight process $\pi(t)$ is assumed to be

$$\pi(t) = \begin{pmatrix} \pi_0(t) \\ \pi_1(t) \\ \vdots \\ \pi_n(t) \end{pmatrix}$$

The first element $\pi_0(t)$ corresponds to the index bond, and $(\pi_1(t), \pi_2(t), \ldots, \pi_n(t))$ the n kinds of financial products. The member invests share of $1 - \sum_{i=0}^{n} \pi_i$ in the risk-free bond. Hence we define the wealth process $X^\pi(t)$ with an initial value of x $(0 \leq x < \infty)$ in the following manner.

$$dX^\pi(t) = X^\pi(t)\Big[R dt + \pi^\top(t)\sigma\,(\theta dt + dW(t))\Big] + cY(t)dt \quad (1)$$

where $cY(t)$ is the amount of money contributed to the index bond and the financial products at time t. In addition, we set the stochastic discount factor $H(t)$ by

$$H(t) = \exp\left\{-Rt - \frac{1}{2}||\theta||^2 t - \theta^\top W(t)\right\}$$

which adjusts for the nominal interest rate and the market price of risk. The following definition is quoted from [2]

Definition 1 *A portfolio vector $\pi(t)$ is said to be admissible if the corresponding wealth process $X^{\pi}(t)$ in (1) satisfies*

$$X^{\pi}(t) + \mathbb{E}_t\left[\int_t^T \frac{H(s)}{H(t)} cY(s)ds\right] \geq 0,\ \mathbb{P}-a.s.$$

where $\{\mathcal{F}(t)\}_t$ is the Brownian filtration and $\mathbb{E}_t$ denotes the conditional expectation given $\mathcal{F}(t)$ under $\mathbb{P}$. We call the term

$$\mathbb{E}_t\left[\int_t^T \frac{H(s)}{H(t)} cY(s)ds\right]$$

the human capital.

Assume that the utility function of the constant relative risk aversion has the form

$$u(z) = \frac{z^{1-\gamma}}{1-\gamma} \tag{2}$$

Then, as is discussed in [3], the representative pension plan member wishes to maximize the expected utility from the terminal value of the pension fund given an initial investment of $x > 0$, that is

$$\max_{\pi\in\bar{\mathcal{A}}(x)} \mathbb{E}\left[u(X^{\pi}(T)\right] \tag{3}$$

subject to

$$\begin{aligned} dX^{\pi}(t) &= X^{\pi}(t)\left[Rdt + \pi^{\top}(t)\sigma\left(\theta dt + dW(t)\right)\right] \\ &\quad + cY(t)dt \end{aligned} \tag{4}$$

$$X^{\pi}(0) = x \tag{5}$$

The problem stated in (3)-(5) is the classical terminal wealth optimization problem. Our final aim is to derive the optimal portfolio weight process $\pi(t)$ as the solution to the optimization problem.

2.2 Modeling for optimal investment problem

We consider the plan member's future contribution. For that purpose, define the present value of expected future contribution process by

$$\begin{aligned} D(t) &= \mathbb{E}_t\left[\int_t^T \frac{H(s)}{H(t)} cY(s)ds\right] \\ &= \frac{c}{\beta}\left(\exp\{\beta(T-t)\} - 1\right) Y(t) \end{aligned} \tag{6}$$

for $t \in [0, T]$ with

$$\beta = \kappa - R - \sigma_s\theta_0$$

The representation (6) yields the stochastic differential equation such as

$$dD(t) = D(t)\left[\left(R + \sigma_s\theta_0\right)dt + \sigma_Y^{\top}dW(t)\right] - cY(t)dt \tag{7}$$

Then we obtain the total wealth process

$$V(t) = X^{\pi}(t) + D(t) \tag{8}$$

It follows from (5), (7) and (8) that

$$dV(t) = dX^{\pi}(t) + dD(t)$$

We consider also that the differential discounted process of the total wealth. By Ito product rule

$$\begin{aligned} &d(H(t)V(t)) \\ &= H(t)dV(t) + V(t)dH(t) + dV(t)dH(t) \end{aligned}$$

2.3 Theoretical optimal investment solution

In view of the observation in the previous subsection, we restate the classical terminal wealth optimization problem (3)-(5) in terms of $V(t)$.

$$\max_{\pi\in\bar{\mathcal{A}}(x)} \mathbb{E}\left[u(X^{\pi}(T))\right] = \max_{\pi\in\mathcal{A}_1(x+d)} \mathbb{E}\left[u(V(T))\right] \tag{9}$$

Thus the problem (3)-(5) is replaced by (9).

Let the superscribe $(^*)$ be the corresponding optimal process hereafter, then the optimal wealth process is given by

$$V^*(t) = \frac{x+d}{H(t)} Z_1(t)$$

with

$$Z_1(t) = \exp\left\{\frac{1-\gamma}{\gamma}\theta^{\top}W(t) - \frac{1}{2}\left(\frac{1-\gamma}{\gamma}\right)^2 \|\theta\|^2 t\right\}$$

Therefore solving for $\pi^*(t)$ gives

$$\pi^*(t) = \frac{1}{\gamma}\left(\sigma^{\top}\right)^{-1}\theta + \left(\sigma^{\top}\right)^{-1}\left(\frac{1}{\gamma}\theta - \sigma_Y\right)\frac{D(t)}{V^*(t) - D(t)}$$

which is the optimal weight process of our aim.

3 Numerical work

Some numerical works using real financial data have been presented in the symposium.

References

[1] Steven E. Shreve. (2004) Stochastic calculus for finance ii continuous-time models, Springer

[2] Aihua Zhang. Cristian-Oliver Ewald. (2010) Optimal investment for a pension fund under inflation risk, Math Meth Oper Res 71:353-369

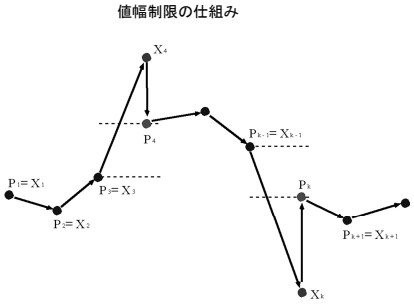
[3] Nicole El Karoui, Monique Jeanblanc-Picque. (1998) Optimization of consumption with labor income, Finance Stochastic 2, 409-440

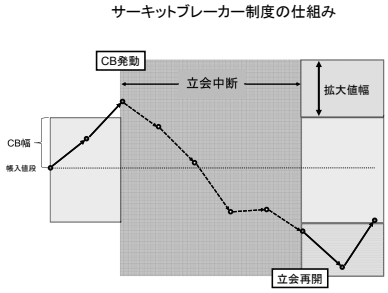
サーキットブレーカー制度を考慮した商品先物のリスク評価

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

1　先物取引の仕組み

- 先物取引
  取引価格が変動する商品について，現物の受け渡しは数ヶ月先に実行することとして，売買の約定を結ぶ取引き
- 約定値段
  先物市場で成立した売買契約による売買値段
- 約定値段決定のための制度
  - 値幅制限制度
    1 営業日に変動する最大幅を一定範囲内に制限すること
  - サーキットブレーカー制度（CB 制度）
    前日の計算区域での帳入値段を基準として，差額がＣＢ幅を超えた場合に発動
    立会を 5 分間中断し直前の CB 幅に拡大値幅を加算したＣＢ幅に変更して立会を再開





2　モデル化とパラメータ推定

2.1　問題のモデル化

データ解析をもとに，先物商品の潜在価値の差分 ΔX_t のモデルとして，1 次の自己回帰モデルを仮定

$$\Delta X_t = \phi \Delta X_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),\ t = 1, 2, \ldots.$$

ここで,

- $X_0 \equiv x_0$：初期値
- $\sigma > 0$, $|\phi| < 1$：未知パラメータ（データから推定）

CB 発動により観測できない ΔX_t が存在する
$\rightarrow$ ΔX_t の補間と (ϕ, σ^2) の推定を同時に行う
$\rightarrow$ マルコフ連鎖モンテカルロ法の利用

2.2　アルゴリズムの概略

- 約定価格の n 個の観測値 $\{P_t\}_{t=1}^n$ と，商品の潜在価値の初期値 $X_0 \equiv x_0$ を与える.

- 未知パラメータ σ^2 と ϕ について，反復計算のための初期値 $\sigma^2(0), \phi(0)$ を与える．事前分布のパラメータも与える．
- $\{\Delta X_t\}$ の中で観測不可能であったものを補間．

$$X \ \sim \ f(\ \cdot\ |x_0, \sigma^2, \phi)$$

- 補間したものも加えた $\{\Delta X_t\}$ を用いて，σ^2 と ϕ の乱数系列を発生．

$$\sigma^2 \ \sim \ \pi_1(\ \cdot\ |\Delta X, \phi)：逆ガンマ分布, \qquad \phi \ \sim \ \pi_2(\ \cdot\ |\Delta X, \sigma^2)：切断正規分布$$

- 上記 3, 4 を繰り返し，補間データと推定対象のパラメータの乱数を更新．

2.3　潜在価格の予測分布の推定および VaR の特定

マルコフ連鎖モンテカルロ法 によって乱数系列 $\{\sigma^2(i)\}_{i=N}^{N+M-1}$ と $\{\phi(i)\}_{i=N}^{N+M-1}$ が得られたら，

$$\widehat{\sigma^2} = \frac{1}{M}\sum_{i=N}^{N+M-1}\sigma^2(i), \quad \widehat{\phi} = \frac{1}{M}\sum_{i=N}^{N+M-1}\phi(i)$$

として σ^2 と ϕ を推定．

Φ を標準正規分布関数としたときのリスク評価式

- 先物商品の潜在価値 X_t が c を下回る確率

$$P\{X_t \le c\} = \Phi\left(\left\{c - \widehat{\phi}^{t-t_0}P_{t_0}\right\} \Big/ \left\{\widehat{\sigma^2}\sum_{k=0}^{t-t_0-1}\widehat{\phi}^{2k}\right\}^{1/2}\right)$$

- 下側 $100\,\alpha\,\%$ の確率の分岐となる価格（理論値）

$$\begin{aligned} c &= \widehat{\phi}^{t-t_0}\Delta X_{t_0} + \left\{\widehat{\sigma^2}\sum_{k=0}^{t-t_0-1}\widehat{\phi}^{2k}\right\}^{1/2}\Phi^{-1}(\alpha) \\ &= \mathrm{VaR}_{\alpha\times 100\%} \end{aligned}$$

3　マルコフ連鎖モンテカルロ法による VaR の推定値と検定結果

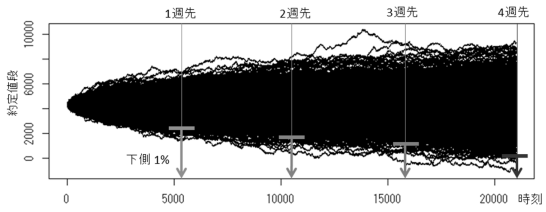
計算結果

1. パラメータの推定値

パラメータ	推定値	Geweke の事後分布収束判定
σ^2	12.54	0.428 (< 1.645)　収束
ϕ	−0.100	0.004 (< 1.645)　収束

2. VaR

推定したパラメータで1000本の path を発生



営業日	$\mathrm{VaR}_{1\%}$
1週先	2236.95
2週先	1656.27
3週先	1113.26
4週先	335.50

Statistical Arbitrage on Volatility [1]

Ngai Hang Chan
Department of Statistics
Chinese University of Hong Kong
Shatin, NT, Hong Kong
nhchan@sta.cuhk.edu.hk

Abstract

This talk discusses recent developments of statistical arbitrage and cointegration. By virtue of some of the asymptotic results about co-integration tests, a pair-trading strategy is constructed from which statistical arbitrage can be developed. In particular, a statistical arbitrage strategy is constructed for volatility trading. The talk concludes with some of the applications to financial data.

[1]Workshop in Statistical Applications and Time Series. November 28–30, 2013, Kanazawa, Japan. Joint work with Mini Pun and Philip Lee. Research supported in part by grants from HKSAR-RGC-GRF.

″ Unobserved Heterogeneity in Longitudinal Data: An Empirical Bayes Perspective″

By Roger Koenker, University of Illinois at Urbana-Champaign

Abstract

Empirical Bayes methods for Gaussian and binomial compound decision problems involving longitudinal data are considered. A new convex optimization formulation of the nonparametric (Kiefer-Wolfowitz) maximum likelihood estimator for mixture models can be used to construct nonparametric Bayes rules for compound decisions. The methods are illustrated with some simulation examples as well as an application to predicting baseball batting averages. Comparisons with nonparametric Bayesian methods based on Dirichlet process priors are also provided.

確率優越性の評価基準で平均と分散の双方に順序制約がある場合の２つの正規母集団の平均の推定とその応用

目白大学・社会学部　張　元宗　　慶應大学・理工学部　篠崎 信雄

1．はじめに

制約された母数空間における推定問題が数多く考えられるが、母平均に関する代表的な線形制約条件は次のようなものが挙げられる。(1) 非負性 (2) 順序制約 (simple order) (3)simple tree order (4) 傘型順序制約である。例えば、順序制約は、年齢とともに平均値が大きくなると考えられる量（児童の身長など)、薬品の投与量とともに平均的に大きくなると考えられる反応量などの場合に考えられている。このような場合の統計的推測について、古くから様々の研究が進められてきているが、1988 年以前の研究については、Barlow et al.(1972) や Robertson et al. (1988) で詳しく解説されている。その後の発展、特に、点推定及び区間推定については、Silvapulle & Sen(2005) や van Eeden(2006) のモノグラフィによって解説されている。また、国内では張 (Chang) と篠崎 (Shinozaki) は制約条件を考慮する最尤推定量（RMLE）による改良問題について研究を推進し、Kubokawa, Tsukuma, Marchand, Perron, Strawderman らは一般ベイズ推定量の許容性およびミニマックス性などについて精力的に研究を進めている。

ここでは、母平均と分散の双方に順序制約条件がある場合の２つの正規母集団における平均の推定問題を、確率優越性 (stochastic domination) あるいは平均 2 乗誤差 (MSE) を評価基準として考える。$X_{ij}, i = 1, 2, j = 1, \ldots, n_i$ は平均 μ_i、分散 σ_i^2 の正規分布からの独立な観測値とする。ここで、μ_i と σ_i^2 は共に未知で、$\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i, s_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(n_i - 1)$ は μ_i と σ_i^2 の不偏推定量である。Oono & Shinozaki(2005) が μ_i の打ち切り型推定量

$$\hat{\mu}_1^{OS} = \min\{\bar{X}_1, \hat{\mu}^{GD}\}, \quad \hat{\mu}_2^{OS} = \max\{\bar{X}_2, \hat{\mu}^{GD}\} \tag{1.1}$$

を提案し、$\hat{\mu}_i^{OS}$ が制約条件を無視した μ_i の最尤推定量 $\bar{X}_i$ を MSE で改良するための必要十分条件を与えた。ここで、$\hat{\mu}^{GD}$ は Graybill & Deal(1959) が提案した２つの正規分布の共通母平均の推定量

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

である。しかし、分散にも順序制約条件 $\sigma_1^2 \le \sigma_2^2$ がある場合、$s_1^2 > s_2^2$ のとき、$n_1 s_2^2/(n_1 s_2^2 + n_2 s_1^2) < n_1/(n_1 + n_2)$ となり不自然である。そこで、次のような推定量

$$\hat{\mu}_1^{CS} = \begin{cases} \hat{\mu}_1^{OS}, & if\ s_1^2 \le s_2^2 \\ \min\{\bar{X}_1, \frac{n_1}{n_1+n_2}\bar{X}_1 + \frac{n_2}{n_1+n_2}\bar{X}_2\}, & if\ s_1^2 > s_2^2, \end{cases} \tag{1.2}$$

$$\hat{\mu}_2^{CS} = \begin{cases} \hat{\mu}_2^{OS}, & if\ s_1^2 \le s_2^2 \\ \max\{\bar{X}_2, \frac{n_1}{n_1+n_2}\bar{X}_1 + \frac{n_2}{n_1+n_2}\bar{X}_2\}, & if\ s_1^2 > s_2^2 \end{cases} \tag{1.3}$$

を考え、$\hat{\mu}_i^{CS}$ と $\hat{\mu}_i^{OS}$ との比較を行い、次のような結果を得た。

２．結果

大きい分散に対応する母平均 μ_2 の推定：

定理１． $\hat{\mu}_2^{CS}$ は $\hat{\mu}_2^{OS}$ より確率的に優れている。つまり、すべての $\mu_1 \le \mu_2$, $\sigma_1^2 \le \sigma_2^2$ に対して、

$$P\Big\{|\ \hat{\mu}_2^{CS} - \mu_2\ | \le d\Big\} \ge P\Big\{|\ \hat{\mu}_2^{OS} - \mu_2\ | \le d\Big\}, \quad \forall\, d > 0$$

が成立。(証明を付録に置く。)

系２． MSE を基準にしたとき、$\hat{\mu}_2^{CS}$ は $\hat{\mu}_2^{OS}$ より優れている。

小さい分散に対応する母平均 μ_1 の推定：

定理３． $\Delta = \mu_2 - \mu_1$ が十分大きいとき、$\hat{\mu}_1^{CS}$ は $\hat{\mu}_1^{OS}$ よりも大きな MSE をもつ。

次に、(μ_1, μ_2) の同時推定を考える。

定理４． $(\hat{\mu}_1^{CS}, \hat{\mu}_2^{CS})$ は $(\hat{\mu}_1^{OS}, \hat{\mu}_2^{OS})$ より次の意味で優れている、つまり、 すべての $\mu_1 \le \mu_2$, $\sigma_1^2 \le \sigma_2^2$ に対して、

$$P\Big\{\sum_{i=1}^{2}\Big(\frac{\hat{\mu}_i - \mu_i}{\sigma_i^2/n_i}\Big)^2 \le d\Big\} \ge P\Big\{\sum_{i=1}^{2}\Big(\frac{\hat{\mu}_i^{OS} - \mu_i}{\sigma_i^2/n_i}\Big)^2 \le d\Big\}, \quad \forall\, d > 0,$$

が成立。

参考文献：(1)Chang Y.-T. and Shinozaki, N. (2012) “Estimation of Ordered Means of Two Normal Distributions with Ordered Variances”, Journal of Mathematics and System Science Vol.2, No.1, pp. 1-7. (2)Chang Y,-T. Oono, Y. and Shinozaki, N. (2012), “Improved estimators for the common mean and ordered means of two normal distributions with ordered variances”, Journal of Statistical Planning and Inference, 142, 2619-2628. (3)Graybill, F.A. and Deal, R. B. (1959), “Combining unbiased estimators,” Biometrics 15, pp. 543-550. (4)Hwang, J. T. (1985), “Universal domination and stochastic domination,” Ann. Statist., Vol.13, No.1, pp. 295-314. (5)Oono, Y. and Shinozaki, N. (2005), “Estimation of two order restricted normal means with unknown and possibly unequal variances,” Journal of Statistical Planning and Inference, Vol. 131, Issue2, pp. 349-363. (6) Silvapulle, M. J., Sen, P. K. (2004). *Constrained Statistical Inference*. Wiley, New Jersey. (7) van Eeden, C., (2006). *Restricted Parameter Space Estimation Problems*. Lecture notes in Statistics 188, Springer.

謝辞：本研究は科研費 (基盤研究 (C)No.22500263) 助成を受けたものである。また、著者張元宗は目白大学の特別研究費の助成も受けています。

付録　(定理 1 の証明)

まず、$P\{|\hat{\mu}_2^{CS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2\}$ を下記のように表現できる。

$$P\left\{|\hat{\mu}_2^{CS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2\le s_2^2\right\}P\{s_1^2\le s_2^2\}+P\left\{|\hat{\mu}_2^{CS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}P\{s_1^2>s_2^2\}.$$

$\hat{\mu}_2^{CS}$ は $\hat{\mu}_2^{OS}$ より、確率的に優れているために、次の式を証明すればよい。

$$P\left\{|\hat{\mu}_2^{CS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}\ge P\left\{|\hat{\mu}_2^{OS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}. \tag{A.1}$$

$c=n_1/(n_1+n_2)$ とする。(A.1) の左辺は

$$P\left\{|c(\bar{X}_1-\mu_2)+(1-c)(\bar{X}_2-\mu_2)|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}. \tag{A.2}$$

になる。次に、変数変換 $W_i=\bar{X}_i-\mu_2,\quad i=1,2.$ を行うと、$W_1\sim N(-\Delta,\tau_1^2)$, $W_2\sim N(0,\tau_2^2)$ に従い、W_1 と W_2 は互いに独立になる。ここで、$\Delta=\mu_2-\mu_1$ であり、$\tau_i^2=\sigma_i^2/n_i$, i=1,2 である。

よって、(A.2) は

$$P\{-d\le cW_1+(1-c)W_2\le d|W_1>W_2,s_1^2>s_2^2\}. \tag{A.3}$$

になる。さらに、変数変換 $V_1=W_1-W_2,V_2=(\tau_2^2/\tau_1^2)W_1+W_2.$ を行うと、V_1 と V_2 は互いに独立で、$V_1\sim N(-\Delta,\tau_1^2+\tau_2^2),V_2\sim N(-(\tau_2^2/\tau_1^2)\Delta,\tau_2^2(\tau_1^2+\tau_2^2)/\tau_1^2)$ である。また、$W_1=\frac{\tau_1^2(V_1+V_2)}{\tau_1^2+\tau_2^2},W_2=\frac{\tau_1^2V_2-\tau_2^2V_1}{\tau_1^2+\tau_2^2}$, であるから、便宜上、$s_1^2>s_2^2$ を固定し、(A.3) は

$$\begin{aligned}
&P\left\{-d\le cW_1+(1-c)W_2\le d|W_1>W_2\right\}=P\left\{-d\le\frac{(c\tau_1^2-(1-c)\tau_2^2)V_1+\tau_1^2V_2}{\tau_1^2+\tau_2^2}\le d|V_1>0\right\}\\
&=P\left\{(c\tau_1^2-(1-c)\tau_2^2)V_1+\tau_1^2V_2\ge-d(\tau_1^2+\tau_2^2)|V_1>0\right\}-P\left\{(c\tau_1^2-(1-c)\tau_2^2)V_1+\tau_1^2V_2\ge d(\tau_1^2+\tau_2^2)|V_1>0\right\}.\\
&=E\{g(c,V_1)|V_1>0\},
\end{aligned}$$

になる。ここで、

$$g(c,v_1)=\Phi\left(\frac{((1-c)\tau_2^2-c\tau_1^2)v_1+\tau_2^2\Delta+d(\tau_1^2+\tau_2^2)}{\tau_1\tau_2\sqrt{\tau_1^2+\tau_2^2}}\right)-\Phi\left(\frac{((1-c)\tau_2^2-c\tau_1^2)v_1+\tau_2^2\Delta-d(\tau_1^2+\tau_2^2)}{\tau_1\tau_2\sqrt{\tau_1^2+\tau_2^2}}\right).$$

である。同様に、(A.1) の右辺については、次のように計算される。

$$P\left\{|\hat{\mu}_2^{OS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}=E\{g(\gamma_0,V_1)|V_1>0,s_1^2>s_2^2\},$$

ここで、$\gamma_0=n_1s_2^2/(n_1s_2^2+n_2s_1^2)$ である。

$v_1>0$ を固定した上で、$g(u,v_1)$ は $0\le u\le1$ の非減少関数であることを簡単に証明できる。また、$s_1^2>s_2^2$ に対して、$c>\gamma_0$ が成立するので、

$$P\left\{|\hat{\mu}_2-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\}\ge P\left\{|\hat{\mu}_2^{OS}-\mu_2|\le d|\bar{X}_1>\bar{X}_2,s_1^2>s_2^2\right\},$$

が成立、証明が完成した。

LMSR-method とその応用

島根大学大学院総合理工学研究科 M2 清水 祥太
島根大学大学院総合理工学研究科数理科学領域 内藤 貫太

1 はじめに

LMS-method は Cole and Green (1992) で提案された非線形回帰分析手法である (Cole (1990),Cole,Freeman and Preece(1998) も参照)．LMS-mehod は人体の成長を歪度，中央値，変動係数の 3 つのパラメータを含むべき変換によって捉える手法であった．従来の手法は，説明変数と目的変数が共に 1 次元の場合である．そのため，目的変数を多次元化した設定のもとで一般化した．一般化した LMS-method を，LMSR-method と呼ぶことにする．LMSR-method で最も重要視するパラメータは中央値である．この中央値を用いて多次元空間における 1 次元曲線を描くことによって，人体の成長過程を多次元的に捉えることができる．

2 LMSR-method

2.1 LMS-method

データ $\{(x_i, y_i) : i = 1, \cdots, n\}$ が得られたとき，x_i における y_i の歪度，中央値，変動係数を表すパラメータをそれぞれ $L(x_i), M(x_i), S(x_i)(i = 1, \cdots, n)$ とする．このとき，3 つのパラメータを用いて次の変換を考える：

$$z_i = \begin{cases} \dfrac{\left(\dfrac{y_i}{M(x_i)}\right)^{L(x_i)} - 1}{L(x_i)S(x_i)} & , L(x_i) \neq 0 \\ \dfrac{\log\left(\dfrac{y_i}{M(x_i)}\right)}{S(x_i)} & , L(x_i) = 0 \end{cases} \quad , i = 1, \cdots, n.$$

このとき

$$Z_1, \cdots, Z_n \overset{i.i.d.}{\sim} N(0, 1)$$

を仮定する．そして，未知のパラメータ $\boldsymbol{L} = (L(x_1), \cdots, L(x_n))^T$, $\boldsymbol{M} = (M(x_1), \cdots, M(x_n))^T$, $\boldsymbol{S} = (S(x_1), \cdots, S(x_n))^T$ は罰則付最尤法で推定する．$\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}$ は Natural Cubic Spline(NCS) であると仮定する．このとき，罰則付対数尤度関数 $\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S})$ は

$$\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}) = \ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}) - \frac{\alpha_L}{2}\boldsymbol{L}^T K \boldsymbol{L} - \frac{\alpha_M}{2}\boldsymbol{M}^T K \boldsymbol{M} - \frac{\alpha_S}{2}\boldsymbol{S}^T K \boldsymbol{S}$$

となる．ただし，$\alpha_L, \alpha_M, \alpha_S$ は平滑化パラメータであり等価自由度によって定められる．対数尤度関数 $\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S})$ は

$$\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}) = \sum_{i=1}^{n} \left\{ L(x_i) \log\left(\frac{y_i}{M(x_i)}\right) - \log S(x_i) \right\}$$

となる．また K は，x_i $(i=1,\cdots,n)$ にのみ依存する $n\times n$ 行列である (Green and Silverman(1994) 参照)．罰則付対数尤度関数 $\Pi(\boldsymbol{L},\boldsymbol{M},\boldsymbol{S})$ の最大化によって得られた $\boldsymbol{L},\boldsymbol{M},\boldsymbol{S}$ の推定量 $\widehat{\boldsymbol{L}}=\left(\widehat{L}(x_1),\cdots,\widehat{L}(x_n)\right)^T,\widehat{\boldsymbol{M}}=\left(\widehat{M}(x_1),\cdots,\widehat{M}(x_n)\right)^T,\widehat{\boldsymbol{S}}=\left(\widehat{S}(x_1),\cdots,\widehat{S}(x_n)\right)^T$ より

$$\left\{\left(x_i,\widehat{L}(x_i)\right):i=1,\cdots,n\right\},\left\{\left(x_i,\widehat{M}(x_i)\right):i=1,\cdots,n\right\},\left\{\left(x_i,\widehat{S}(x_i)\right):i=1,\cdots,n\right\}$$

を作り，これらを NCS で平滑化する．データ $\{(x_i,y_i):i=1,\cdots,n\}$ はこのようにして得られる 3 つの曲線 $\widehat{L}(x),\widehat{M}(x),\widehat{S}(x)$ で説明される．

2.2 LMSR-method

目的変数を多次元化した設定のもとでの LMS-method の一般化が LMSR-method である．多次元化した際に必要となる変数間の相関については 2 つのパターンを考える．

2.2.1 相関が関数のとき

データ $\{(x_i,\boldsymbol{y}_i):i=1,\cdots,n\},\boldsymbol{y}_i=(y_{i1},\cdots,y_{im})^T$ が得られたとき，x_i における y_{ik} の歪度，中央値，変動係数を表すパラメータを $L_k(x_i),M_k(x_i),S_k(x_i)$ とする．このとき，3 つのパラメータを用いて次の変換を考える：

$$z_{ik}=\begin{cases}\dfrac{\left(\dfrac{y_{ik}}{M_k(x_i)}\right)^{L_k(x_i)}-1}{L_k(x_i)S_k(x_i)} & ,L_k(x_i)\neq 0\\ \dfrac{\log\left(\dfrac{y_{ik}}{M_k(x_i)}\right)}{S_k(x_i)} & ,L_k(x_i)=0\end{cases}\qquad ,i=1,\cdots,n;\ k=1,\cdots,m.$$

そして，$\boldsymbol{Z}_i=(Z_{i1},\cdots,Z_{im})^T$ とおき

$$\boldsymbol{Z}_i\sim N_m(\boldsymbol{0},I(R(x_i))),\boldsymbol{Z}_i\perp\!\!\!\perp\boldsymbol{Z}_j\ (i\neq j\ ,i,j=1,\cdots,n)$$

を仮定する．ここで $I(R(x_i))$ はパラメータ $R(x_i)$ によって規定される相関行列である．また，$\boldsymbol{L}_k=(L_k(x_1),\cdots,L_k(x_n))^T,\boldsymbol{M}_k=(M_k(x_1),\cdots,M_k(x_n))^T,\boldsymbol{S}_k=(S_k(x_1),\cdots,S_k(x_n))^T$ $(k=1,\cdots,m)$ および $\boldsymbol{R}=(R(x_1),\cdots,R(x_n))^T$ を未知のパラメータとし，それらを罰則付最尤法で推定する．ここで，$\boldsymbol{L}_k,\boldsymbol{M}_k,\boldsymbol{S}_k(k=1,\cdots,m)$ および $\boldsymbol{R}$ は NCS であると仮定し，$\boldsymbol{L}=(\boldsymbol{L}_1^T,\cdots,\boldsymbol{L}_m^T)^T,\boldsymbol{M}=(\boldsymbol{M}_1^T,\cdots,\boldsymbol{M}_m^T)^T,\boldsymbol{S}=(\boldsymbol{S}_1^T,\cdots,\boldsymbol{S}_m^T)^T$ とおく．このとき，罰則付対数尤度関数 $\Pi(\boldsymbol{L},\boldsymbol{M},\boldsymbol{S},\boldsymbol{R})$ は

$$\Pi(\boldsymbol{L},\boldsymbol{M},\boldsymbol{S},\boldsymbol{R})=\ell(\boldsymbol{L},\boldsymbol{M},\boldsymbol{S},\boldsymbol{R})-\frac{\alpha_L}{2}\boldsymbol{L}^T\boldsymbol{K}\boldsymbol{L}-\frac{\alpha_M}{2}\boldsymbol{M}^T\boldsymbol{K}\boldsymbol{M}-\frac{\alpha_S}{2}\boldsymbol{S}^T\boldsymbol{K}\boldsymbol{S}-\frac{\alpha_R}{2}\boldsymbol{R}^TK\boldsymbol{R}$$

となる．ただし，$\alpha_L, \alpha_M, \alpha_S, \alpha_R$ は平滑化パラメータである．対数尤度関数 $\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{R})$ は

$$\begin{aligned}\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{R}) &= \sum_{i=1}^{n}\left\{\sum_{k=1}^{m}\left\{L_k(x_i)\log\left(\frac{y_{ik}}{M_k(x_i)}\right) - \log S_k(x_i)\right\}\right. \\ &\quad \left. -\frac{1}{2}\log|I(R(x_i))| - \frac{1}{2}\boldsymbol{z}_i^T I(R(x_i))^{-1}\boldsymbol{z}_i\right\}\end{aligned}$$

であり，$\boldsymbol{z}_i = (z_{i1}, \cdots, z_{im})^T$ $(i = 1, \cdots, n)$ である．$\boldsymbol{K}$ は $nm \times nm$ 行列であり，$\boldsymbol{K} = I_m \otimes K$ と定義する．ただし，I_m は $m \times m$ の単位行列であり，$\otimes$ はクロネッカー積を表している．罰則付対数尤度関数 $\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{R})$ の最大化によって得られた $\boldsymbol{L}_k, \boldsymbol{M}_k, \boldsymbol{S}_k (k = 1, \cdots, m), \boldsymbol{R}$ の推定量 $\widehat{\boldsymbol{L}}_k = (\widehat{L}_k(x_1), \cdots, \widehat{L}_k(x_n))^T$，$\widehat{\boldsymbol{M}}_k = (\widehat{M}_k(x_1), \cdots, \widehat{M}_k(x_n))^T$，$\widehat{\boldsymbol{S}}_k = (\widehat{S}_k(x_1), \cdots, \widehat{S}_k(x_n))^T (k = 1, \cdots, m)$, $\widehat{\boldsymbol{R}} = (\widehat{R}(x_1), \cdots, \widehat{R}(x_n))^T$ より

$$\left\{\left(x_i, \widehat{L}_k(x_i)\right) : i = 1, \cdots, n\right\}, \left\{\left(x_i, \widehat{M}_k(x_i)\right) : i = 1, \cdots, n\right\}, \left\{\left(x_i, \widehat{S}_k(x_i)\right) : i = 1, \cdots, n\right\}, \quad k = 1, \cdots, m,$$

$$\left\{\left(x_i, \widehat{R}(x_i)\right) : i = 1, \cdots, n\right\}$$

を作り，これら全てをNCSで平滑化する．データ $\{(x_i, \boldsymbol{y}_i) : i = 1, \cdots, n\}, \boldsymbol{y}_i = (y_{i1}, \cdots, y_{im})^T$ はこのようにして得られた $3m + 1$ 個の曲線 $\widehat{L}_k(x), \widehat{M}_k(x), \widehat{S}_k(x) (k = 1, \cdots, m), \widehat{R}(x)$ によって説明される．

2.2.2 相関が定数のとき

特に任意の $i \in \{1, \cdots, n\}$ に対して $R(x_i) = \rho$ とする．このとき，

$$\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n \overset{i.i.d.}{\sim} N_m(\boldsymbol{0}, I(\rho))$$

を仮定する．ここで $I(\rho)$ はパラメータ ρ によって規定される相関行列である．罰則付対数尤度関数 $\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho)$ は

$$\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho) = \ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho) - \frac{\alpha_L}{2}\boldsymbol{L}^T\boldsymbol{K}\boldsymbol{L} - \frac{\alpha_M}{2}\boldsymbol{M}^T\boldsymbol{K}\boldsymbol{M} - \frac{\alpha_S}{2}\boldsymbol{S}^T\boldsymbol{K}\boldsymbol{S}$$

となる．ただし，$\alpha_L, \alpha_M, \alpha_S$ は平滑化パラメータである．また，対数尤度関数 $\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho)$ は

$$\begin{aligned}\ell(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho) &= \sum_{i=1}^{n}\left\{\sum_{k=1}^{m}\left\{L_k(x_i)\log\left(\frac{y_{ik}}{M_k(x_i)}\right) - \log S_k(x_i)\right\}\right. \\ &\quad \left. -\frac{1}{2}\log|I(\rho)| - \frac{1}{2}\boldsymbol{z}_i^T I(\rho)^{-1}\boldsymbol{z}_i\right\}\end{aligned}$$

であり，$\boldsymbol{z}_i = (z_{i1}, \cdots, z_{im})^T$ $(i = 1, \cdots, n)$ である．罰則付対数尤度関数 $\Pi(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{S}, \rho)$ の最大化によって得られた推定量 $\widehat{\boldsymbol{L}}_k = (\widehat{L}_k(x_1), \cdots, \widehat{L}_k(x_n))^T$, $\widehat{\boldsymbol{M}}_k = (\widehat{M}_k(x_1), \cdots, \widehat{M}_k(x_n))^T$, $\widehat{\boldsymbol{S}}_k = (\widehat{S}_k(x_1), \cdots, \widehat{S}_k(x_n))^T (k = 1, \cdots, m)$ より

$$\left\{\left(x_i, \widehat{L}_k(x_i)\right) : i = 1, \cdots, n\right\}, \left\{\left(x_i, \widehat{M}_k(x_i)\right) : i = 1, \cdots, n\right\}, \left\{\left(x_i, \widehat{S}_k(x_i)\right) : i = 1, \cdots, n\right\},$$

$$k = 1, \cdots, m$$

を作り，これら全てをNCSで平滑化する．データ $\{(x_i, \boldsymbol{y}_i) : i = 1, \cdots, n\}, \boldsymbol{y}_i = (y_{i1}, \cdots, y_{im})^T$ はこのようにして得られた $3m$ 個の曲線 $\widehat{L}_k(x), \widehat{M}_k(x), \widehat{S}_k(x)(k = 1, \cdots, m)$ と $\widehat{\rho}$ によって説明される．

2.2.3 単調性と有界性について

人体の成長は成年期に至るまでは単調増加と見なせる．そのためNCSによって平滑化し得られた中央値を表す曲線 $\widehat{M}_k(x)(k = 1, \cdots, m)$ も単調増加となっている必要がある．そこで，推定量 $\widehat{\boldsymbol{M}}_k = (\widehat{M}_k(x_1), \cdots, \widehat{M}_k(x_n))^T (k = 1, \cdots, m)$ より

$$\left\{ \left(x_i, \widehat{M}_k(x_i) \right) : i = 1, \cdots, n \right\}, k = 1, \cdots, m$$

を作り，Berwin(2005,Sec.2.2) で提案された手法を用いて単調性の制約を加えたNCSで平滑化する．このようにして得られた曲線 $\widehat{M}_k(x)(k = 1, \cdots, m)$ は，単調増加関数となる．

また，NCSによって平滑化し得られた曲線 $\widehat{R}(x)$ は変数間の相関を表している．つまり，

$$-1 \leq \widehat{R}(x) \leq 1$$

となる必要がある．そこで，推定量 $\widehat{R} = (\widehat{R}(x_1), \cdots, \widehat{R}(x_n))^T$ より

$$\left\{ \left(x_i, \widehat{R}(x_i) \right) : i = 1, \cdots, n \right\}$$

を作り，Berwin(2005,Sec.2.3) で提案された手法を用いて有界性の制約を加えたNCSで平滑化する．このようにして得られた曲線 $\widehat{R}(x)$ は有界な関数となる．

参考文献

[1] Cole,T.J. and Green,P.J.(1992). Smoothing reference centile curves: the LMS method and penalized likelihood,*Statistics in Medicine*,**11**,1305-1319.

[2] Cole,T.J.(1990). The LMS method for constructing normalized growth standards, *European Journal of Clinical Nutrition*,**44**,45-60.

[3] Cole,T.J.,Freeman,J.V. and Preece,M.A.(1998). British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood.*Statistics in Medicine*,**17**,407-429.

[4] Green,P.J. and Silverman,B.W.(1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*,Vol.58 of *Monographs on Statistics and Applied Probability*,Chapman and Hall,London.

[5] Berwin,A.T.(2005). Shape constrained smoothing using smoothing splines,*Computational Statistics*,**20**,81-103.

分散共分散行列に基づく判別分析の終焉

新村　秀一†

†成蹊大学経済学部 〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1

1. はじめに

最小誤分類数(**MNM**) に基づく最適線形判別関数(**OLDF**)を開発した．MNM 基準に基づく線形判別関数は混合整数計画法でしか実現できず，最初は IP-OLDF と呼ぶモデルを開発し，Fisher のアイリスデータ，スイス銀行データ，CPD データ，学生データを研究データとして用いて研究を行った．そして，1) スイス銀行紙幣データが 2 変数で線形分離可能であり MNM の単調減少性からその 2 変数を含む 16 個の判別モデルが線形分離可能である．2)判別係数と誤分類数の関係が分かり，判別係数の空間で最適凸体の内点を直接求める改定 IP-OLDF を開発した．本手法は MNM=0 のデータを理論的に判別できる唯一の判別手法であり，判別超平面上のケースの判別を正しく処理ができる．3)上記の 4 種の実データからケース数 100 倍のリサンプリング標本を作成し，MNM の近似値を高速で求める改定 IPLP-OLDF を LDF とロジスティック回帰で 100 重交差検証法を用いて比較し，改定 IPLP-OLDF が圧倒的に学習標本と検証標本で良い結果を得た．以上の結果を踏まえ 2010 年に研究成果をまとめて出版した [31].

2010年から応用研究としてMNM=0の判別に焦点を絞って研究を行うことにした．しかし MNM=0 のデータは現実に多くないが，設問の得点を説明変数として得点合計で合否判定を行う 2 群判別は，自明な MNM=0 になる線形判別関数が求まる．18 個の合否判定で LDF の誤分類確率の範囲は[2.3, 16.7]，QDF は[0.8,10.8]と大きいことが分かった．また，QDF は一方の群の変数値が一定値の場合のみ，一般化逆行列を用いた JMP [25] の QDF や正則化判別分析が 2 群に指定した全ケースを 1 群に誤判別する重大な瑕疵の原因が分かった[35] ．また「小標本のための k-重交差検証法」を開発し，上記 4 種のデータに加えて，2012 年度の筆者の統計入門の中間試験で 100 重交差検証法を行い，改定 IP-OLDF は学習標本で過学習するが，検証標本で LDF,ロジスティック回帰，ソフトマージン最大化 SVM (S-SVM) に比べて検証標本でも平均誤分類確率が一番小さいことが分かった．

2. 合否判定による 100 重交差検証法

2.1　評価に用いたデータ

MNM=0 のデータとして 2012 年の「統計入門」の中間試験のデータを用いる[35]．10 択 100 問の試験を T1 から T4 の大問 4 問に分けた．この 4 個の得点を説明変数とし，合計得点の 10%点(37 点)，50%点(63 点)，90%点(78 点)の 3 水準で合否判定を行う.10%点では大問 2 問,50%点と 90%点では大問 4 問で MNM=0 である．4 個の説明変数の判別モデルは全部で 15 個あるが，1 変数の判別は意味がないので 2 変数以上の 11 個で検討する．MNM の単調減少性から，10%点の合否判定では 4 個の判別モデルが MNM=0 で，7 個が MNM=0 でない．50%点と 90%点では 1 個だけが MNM=0 で 10 個が MNM=0 でない．これによって，MNM=0 を含む判別分析の問題が検討できる．SVM は H-SVM で MNM=0 なデータの判別を判別分析の出発点にしたことは革新的である．しかし，SVM を含めこれまで判別分析は MNM=0 のデータの実証研究を行ってこなかった．判別分析の重要な応用分野であるパターン認識では，MNM=0 のデータの判別が重要である．このため，今回 MNM=0 が自明な試験の合否判定データを用いることは適切と考える．試験の合否判定データは，多くの人が入手し容易に検証でき，合格水準を種々に変えることで 2 群のケース数の種々の割合の違いによる判別への影響を検討できるなどの利点がある．

2.2　10%水準の 100 重交差検証法

表 1 は 100 重交差検証法の結果である．1 列目は各判別手法の下にモデル番号を示す．2 列目は説明変数の数を表す.3 列は学習標本の 100 個の平均誤分類確率（MEAN1）である．4 列は検証標本の平均誤分類確率（MEAN2）で，最少のモデルを 100 重交差検証法が選ぶモデルと考える．

改定 IP-OLDF（表の「IP」）は 100 重交差検証法の学習標本で元データと同じく 4 個の判別モデルで MNM=0 であるが，検証標本では SN=2,6 だけが 0 になった.「差」は「MEAN2」と「MEAN1」との差である．一般的に，改定 IP-OLDF は学習データで過学習し，検証標本で判別の予測結果は悪くなると考えられたが，実際は異なってい

表 1.　10%水準の誤分類確率

IP	P	MEAN1	MEAN2	差	
1	4	0	0.07	0.07	
2	3	0	0	0	
3	3	0	0.03	0.03	
4	3	0.79	2.44	1.65	
5	3	2.25	4.64	2.39	
6	2	0	0	0	
7	2	1.78	3.4	1.62	
8	2	2.28	3.14	0.85	
9	2	4.88	6.63	1.75	
10	2	4.52	6.42	1.9	
11	2	4.94	7.21	2.27	
Logi	P	MEAN1	MEAN2	差	MEAN2 差
1	4	0	0.77	0.77	0.7
2	3	0	1.09	1.09	1.09
3	3	0	0.85	0.85	0.81
4	3	1.59	2.83	1.24	0.39
5	3	4.12	5.46	1.34	0.82
6	2	0	0.91	0.91	0.91
7	2	3.25	4.26	1.01	0.85
8	2	3.68	4.03	0.35	0.89
9	2	6.94	7.78	0.84	1.15
10	2	7.65	8.04	0.38	1.62
11	2	7.05	7.82	0.78	0.61
SVM	P	MEAN1	MEAN2	差	MEAN2 差
1	4	0	0.81	0.81	0.73
2	3	0	1.15	1.15	1.15
3	3	0	0.89	0.89	0.85
4	3	1.36	2.84	1.48	0.4
5	3	3.96	5.72	1.76	1.08
6	2	0	0.85	0.85	0.85
7	2	2.89	4.19	1.3	0.78
8	2	3.08	3.73	0.65	0.6
9	2	6.45	7.45	1	0.82
10	2	6.61	7.6	0.99	1.19
11	2	6.7	8.1	1.4	0.89
LDF	P	MEAN1	MEAN2	差	MEAN2 差
1	4	9.64	10.5	0.9	10.5
2	3	9.89	10.5	0.66	10.5
3	3	9.48	10.1	0.61	10.1
4	3	11.4	12	0.6	9.6
5	3	12.4	12.7	0.32	8.05
6	2	9.54	9.91	0.37	9.91
7	2	11.8	12.2	0.4	8.76
8	2	10.8	11	0.22	7.89
9	2	13.2	13.5	0.28	6.84
10	2	12.5	12.6	0.16	6.23
11	2	16.3	16.5	0.2	9.27

た．ロジスティック回帰と S-SVM は学習標本で元データと同じく 4 個の判別モデルで MNM=0 であり，検証標本で SN=1 を選んだが誤分類数は 0 でない．差は 0.77 と 0.81 で悪くない．LDF は学習標本と検証標本で平均誤分類確率が 0 のものがなく SN=3 を選んだが，平均誤分類確率は 9.48 と 10.1 と非常に悪い．MEAN2 差は，ロジスティック回帰，S-SVM，LDF の MEAN2 から改定 IP-OLDF のそれを引いたもので，範囲は[0.39, 1.62]，[0.4，1.19],[6.23,10.5]である．ロジスティック回帰と S-SVM は改定 IP-OLDF より高々1.62%悪いが，LDF は 6.23%以上悪い．

3.3　50%水準と 90%水準の合否判定

得点分布の 50%点(63 点)で合否判定を行う．ロジスティック回帰の「MEAN2 差」の範囲は[0.05, 3.01]，SVM は[-0.1，3.23]，LDF は[0.65，5.92]である．LDF は他の 3 手法が最小になるモデル 1 で改定 IP-OLDF より 5.92%も悪い．90%点の合否判定で，ロジスティック回帰の「MEAN2 差」の範囲は[-0.15,1.8]で，SVM は[0.23,1.48]であるのに対して，LDF は[7.83，22.58]と明らかに悪い．ロジスティック回帰の平均誤分類確率は，モデル選択と無関係な 9 番目のモデルで改定 IP-OLDF より少ない．

3. まとめ

本研究では，試験の合否判定データを用いて MNM=0 の判別分析の問題を検討した． MNM 基準に基づく改定 IP-OLDF は学習標本で過学習し，検証標本で汎化能力が悪いと考えられるが，実際には正規分布を仮定する LDF が学習標本と検証標本で確率分布を仮定していない S-SVM やロジスティック回帰よりも悪かった．これは Fisher の仮説を満たす現実のデータが少ないのにそれで理論構築したためと考えられる．

文　献

[25] 新村秀一,JMP による統計学とっておき勉強法，東京，講談社，2004.

[28] 新村秀一,Excel と LINGO で学ぶ数理計画法，丸善,2007.

[31] 新村秀一,最適線形判別関数．日科技連出版社,2010.

[32] 新村秀一,数理計画法による問題解決法．日科技連出版社,20

[35] 新村秀一,統計教育における判別分析の改善点．統計教育実践研究第 5 巻-統計数理研究所研究レポート 293-，36-45,2013.

カーネル密度推定法を用いた非線形判別手法の提案

山本けい子 函館工業高等専門学校　寒河江雅彦 金沢大学経済学類

1．はじめに

パターン認識における非線形判別問題は，サポートベクターマシン（SVM）[1]の出現により，大きな進展をとげている．我々は，カーネル密度推定法（以降，KDE と略す）を用いた非線形判別手法を提案する．カーネル密度推定法は，データの分布を仮定しないノンパラメトリックな手法であり，複雑な現象を確率的かつ柔軟にとらえて表現できることから様々な応用が期待される．本稿では，数字判別問題に対する 2 つの状況下での SVM との比較を通して KDE の特性を検証する．

2．カーネル密度推定法を用いた非線形判別器

多変量カーネル密度推定法(1)は，データ点にカーネル関数と呼ばれる基底関数を配置し，それらを領域内で足し合わせることによって滑らかな推定量を得るものである．

$$\hat{f}(\mathbf{x};\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i) \qquad (1)$$

ただし，$\mathbf{X}_i \quad (i=1,\cdots,n)$ はデータベクトル，K_H はバンド幅行列 H をもつカーネル関数である．

多変量データに対する KDE は，構築の困難さや推定精度の問題から直接的な適用は難しい．そこで，1 変量 KDE の積で表現するプロダクト(積型)カーネル推定法によって近似する．たとえば，2 変量プロダクト KDE は(2)式で定義される：

$$\hat{f}(\mathbf{z};h_x,h_y) = \frac{1}{n}\sum_{i,j} K_{h_x}(x-X_i)K_{h_y}(y-Y_j) \qquad (2)$$

以下に，(2)に示した KDE を利用した非線形判別の流れを示す．

1) クラス別正解確率分布の作成
学習用データに対し，クラスごとに KDE を適用し，クラス別確率分布の推定を行う．

2) 評価確率分布の作成
判別対象の評価用データに対し，KDE を適用し，評価確率分布の推定を行う．

3) 正解分布と評価分布間での類似量の算出
判別対象データに基づく評価分布と各クラスの正解分布間の類似量を平均積分二乗誤差によって算出する．

4) クラスの判別
評価分布と最も類似する正解分布のクラスへ判別する．

クラス別正解分布を作成しておくことで，実際の判別時には，判別対象データの分布推定と類似量の算出のみでクラス判別を行うことができるため，効率的かつ汎用的な手法といえる．

3．手書き数字判別問題

手書き数字判別問題に対して，カーネル密度推定法を用いた非線形判別器（数字判別器）を構築し，その判別性能を SVM と比較し，評価する．実装には，オープンソースの統計解析システム R[3]を用いた．

3.1 実験用データ

U.S. Postal Service (USPS) ZIP code datasets の手書き数字を対象とした．USPS データは 1 つの数字を 16×16 ピクセルのグレースケール値(-1 から 1 の範囲の値）によって表す正規化されたデータである．データ数を表 1 に記載する．

表 1　USPS データセット

データ	0	1	2	3	4
学習用	1194	1005	731	658	652
評価用	359	264	198	166	200
データ	5	6	7	8	9
学習用	556	664	645	542	644
評価用	160	170	147	166	177

グレースケール値は，0 から 1 の間の値に変換し確率値として使用した．

3.2 ビン化カーネル密度推定法

通常のカーネル密度推定法は，データ点に対して基底となるカーネル関数を用いる．カウント(頻度)データの場合は，カウント値に比例する重みつきのカーネル関数を用いたビン化カーネル推定法が使用される．USPS データの性質から，16×16 ピクセルにグレースケール値の高さを持つデータとみなし，ビン化カーネル密度推定法を適用した．数字判別におけるビン化カーネル推定法は(3)式で定義される．

$$\hat{f}(\mathbf{z};h_x,h_y) = \sum_{i,j} q_{ij}K_{h_x}(x-i\delta)K_{h_y}(y-j\delta) \qquad (3)$$

ただし，$q_{ij} \equiv g_{ij}/\sum_{i,j} g_{ij}$ であり，$g_{ij}, i\delta, j\delta$ は，それぞれ (i,j) ビンにおけるグレースケール値とビン中点である．

3.3 正解確率分布の作成

表 1 に示した学習用データを数字ごとに各ビンで平均し，ビン化カーネル密度推定を行う．推定

点は各ビンの中点を加えた 33×33 点，カーネル関数は，2 変量正規プロダクトカーネル，バンド幅は，各次元とも 1 に設定した．グレースケール値が標本数ではないため，理論的な議論は省く．推定した各数字の正解確率分布を図 1 に示す．

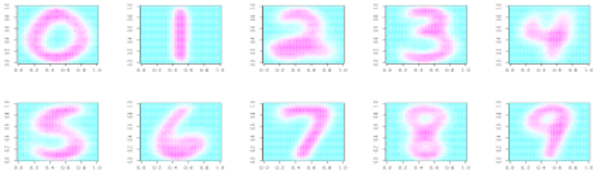


図 1　推定した正解確率分布

3.4 評価用データを用いた数字判別

3.3 で作成した"0"から"9"までの正解確率分布と評価用データで作成した評価確率分布との類似度を(4)式で与えられる平均積分二乗誤差(MISE)を用いて算出する．

$$\mathrm{MISE}[\hat{f}(\cdot;h_x,h_y)] = \int \mathrm{E}\left[\left\{\hat{f}(\mathbf{z};h_x,h_y)-f(\mathbf{z})\right\}^2\right] d\mathbf{z} \qquad (4)$$

MISE の最も小さかった(最も類似した)クラスを評価データのクラスとして分類する．

4．結果

4.1 判別性能の検証

手書き数字判別問題に対する各数字の判別性能を表 2 に示す．

表 2 に示すように，カーネル推定法を用いた数字判別器は，判別する評価用データの数字によって，判別率にばらつきがあるものの，平均して 0.82 であった．一方，SVM はすべての数字において，判別率 0.9 以上，平均して 0.94 という高性能な結果であった．

表 2 評価用データの判別率

判別率	カーネル	SVM
0	0.87	0.98
1	0.96	0.96
2	0.75	0.91
3	0.81	0.91
4	0.77	0.93
5	0.77	0.92
6	0.82	0.95
7	0.82	0.93
8	0.76	0.9
9	0.8	0.97
計	0.82	0.94

4.2 頑健性の検証

カーネル密度推定法による非線形判別器の頑健性を調べるため，ノイズを付加した手書き数字判別に関する実験を行った．

手書き数字データにノイズを加え，判別性能を評価する．16×16 ピクセルのうち，ノイズの大きさを正規乱数(平均 0，標準偏差 0.5)で与え，ノイズを付加するピクセルの割合を増やしたときの判別性能のグラフを図 2 に示す．

図 2 より，SVM の判別性能はノイズの量とともに低下するが，カーネル推定法による数字判別器の

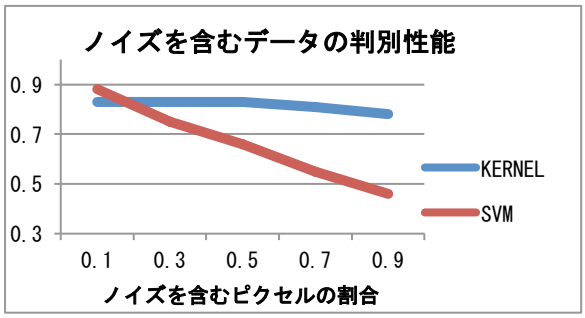


図 2　ノイズを含むデータの判別性能

性能の低下は少ないことがわかる．

4.3 クラスタリングによる判別性能の改善

学習用データに応じて数字毎に複数個の正解確率分布を作成し，判別性能の改善を試みた．学習用データを数字ごとに k-means 法によってクラスタリングし，クラスターごとに正解密度分布を推定する．評価用データでの判別の際には，各クラスターが所属していた数字へ判別する．クラスター数を変化させたときの判別性能を図 3 に示す．

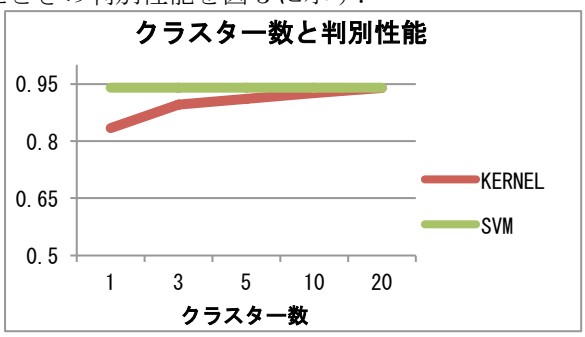


図 3　クラスター数と判別性能

図 3 より，クラスター数を増やすと，SVM に近い判別性能が得られていることがわかる．

5．考察

カーネル密度推定法を用いたパターン認識手法について，手書き数字判別の例を取り上げて検討した．カーネル密度推定法による数字判別器は，SVM の判別性能よりも劣っていたが，ノイズのあるデータに対しては，SVM よりも性能低下の影響は受けにくく，頑健性がみられた．判別する数字によって性能にばらつきが見られるため，クラスタリングによって複数個の正解密度分布を準備することで，SVM とかわらない判別性能が得られた．実用化に向け，さらなる改善が期待できる．

参考文献

[1] V.Vapnik, "The Nature of Statistical Learning Theory", Springer, (1995).
[2] M. P. Wand, M. C. Jones "Kernel Smoothing", Monographs on Statistics and Applied Probability, Chapman & Hall, (1995)
[3]R Development Core Team "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Asymptotics for M-Estimators in Time Series

Yan LIU

Department of Pure and Applied Mathematics, Waseda University, Tokyo 169-8555

November 28, 2013
at Kanazawa University

In this talk, we give conditions which guarantee the asymptotic normality of M-estimator based on the observations from autoregressive (AR) models in the time series analysis by minimizing some convex objective functions. The objective function is defined by two-variate function since the scale parameter of time series model has to be estimated as well as the coefficient parameters. In addition, we do not assume the differentiability around the true parameter of the objective functions. The results are extended to M_m estimates. Examples of the results are also provided.

First, we reviewed the history on the development of M-estimation from the point of view of relaxing the condition of the differentiability around the true value. It has been a half of a century since Hodges and Lehmann (1963) proposed a Lemma on the asymptotic normality for the estimation of location before Huber, who changed the idea into the concept of M-estimation. More precisely, Huber showed the asymptotic normality of M-estimators for convex real-valued function and investigated the robustness of the estimators in 1964. Koenker and Bassett (1978) provided the theory on the regression quantiles by employing the idea of Huber. Pollard (1991) reinvestigated the properties of convex objective function and proposed the convexity lemma. Niemiro (1992), following the line in Habermann (1989), established a Bahadur-type representation of the estimators. Further, Bose (1998) generalized the results in Niemiro (1992) to the case of M_m estimators including Oja's median, univariate robust scale estimator of Bickel and Lehmann and Hodges and Lehmann's estimator.

To evaluate the efficiency and robustness of the estimators, we focused on Martin and Yohai (1985). They summarized the concepts into four types such as Tukey's efficiency robustness, Huber's min-max robustness, Hampel's qualitative robustness and Tukey's resistance. We did not enter into the last two concepts here. Tukey's efficiency robustness is defined as the ratio of the Cramer-Rao lower bound at the parametrized joint distribution of the samples to the variance of the statistic derived from the samples. In the asymptotic case, the ratio can be written as one over the product of the asymptotic Fisher information and the asymptotic variance of the statistic. Huber (1981) formalized the min-max robustness on the asymptotic variance of the statistic at the distributions of the samples and call the statistic min-max robust if the statistic is the infimum over all statistics of the supremum over all distributions of the samples of the asymptotic variance. Since the

underlying distribution is not known generally, Huber (1964) and Huber (1973) proposed minimizing the objective function of i.i.d. samples normalized by its scale parameter.

However, the scale parameter is also unknown in time series analysis. We further generalized M-estimators to be a two-variate function, estimating the scale parameter and coefficient parameters of AR models simultaneously. The usage of two-variate function corresponds to the Whittle's estimators, which is optimal in higher order efficiency in the frequency domain in time series analysis. We suppose the objective function to be convex but may not be differentiable around the true parameter. In this case, the theory contains the L_1 norm, which means that the quantile estimators such as least absolute deviation (LAD) estimator are also considered. Thus this class of M-estimators is larger than the Whittle's estimator since the theory in the frequency domain mainly depends on L_2 norm. The estimators may be more robust beyond those estimators whose objective function defined in L_2 norm if the underlying distribution of the samples has a sufficiently large deviation. The approach for L_1 norm, which receives broad attention, has been developed in the i.i.d. case, regression case and also recently in time series case (see Koenker (2005)).

The conditions in this talk for asymptotic normality of M-estimators in time series are chiefly derived from both Hodges and Lehmann (1963)'s and Niemiro (1992)'s conditions. The former ones are given by Hodges and Lehmann and rearranged in Inagaki and Kondo (1980). The conditions are not directly assumed on the objective function but on the score function. The convergence in probability of the difference between the score function with true parameter and its $\sqrt{n}$-neighborhood is difficult to show, and we employ the latter one for proof. Although Niemiro (1992) is oriented to show that M-estimators have Bahadur-type representation, they gave a very nice structure of proof for asymptotic normality of the estimators with convex objective functions. When the objective function is convex, there exists a subgradient function of the objective funtion. The subgradient function can be regarded as the score function. The asymptotic normality of M-estimators is derived from the properties of the subgradient function. The main difficulty in time series case is that only L_2 integrability of the score function is not enough. To show the asymptotic normality, we have further to guarantee that the covariance between the components of the score function is also summable, like in Billingsley (1968). Under the conditions, we obtained the desirable results without the differentiability around the true parameter. We also derived the asymptotic variance of the estimators assuming the differentiability of the estimators. Our results agree with the existing results, which were given in the talk as examples. The extension of the results to M_m estimators was also mentioned in the talk, which can be shown by the properties of U-statistics given in Yoshihara (1976) or Denker and Keller (1983). The idea was generalized from Hoeffding (1948) and Hoeffding (1961).

The contribution of the work is that we generalized the approach of M-estimation to estimate the scale parameter and the coefficient parameters simultaneously and defined the objective function convex, so that the objective function do not have to be differentiable around the true parameters. We also gave the conditions for asymptotic normality of the estimators, which in turn can be generalized to the approach in the frequency domain.

Statistical Inference Associated with the Fractional Brownian Motion and Related Processes

田中 勝人（学習院大学経済学部）

1. フラクショナル **O-U** 過程

次の連続時間確率過程を考察の対象とした．

$$dY_H(t) = \alpha\, Y_H(t)\, dt + dB_H(t), \qquad (\alpha \le 0), \qquad (t \in [0, M])$$

ここで，$\{B_H(t)\}$ は，Hurst パラメータ H $(1/2 \le H < 1)$ の fBm (フラクショナル Brown 運動) であり，$\{Y_H(t)\}$ は，fO-U (フラクショナル Ornstein-Uhlenbeck) 過程と呼ばれる．ここでは，ドリフト・パラメータ α (≤ 0) の推定問題と検定問題を考察した．

2. 推定

推定量としては，LSE (最小 2 乗推定量)，MLE (最尤推定量)，MCE (最小コントラスト推定量) を取り上げて，それぞれの推定量の分布を考察した．LSE については，次の 3 つの推定量を取り上げた．

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\int_0^M Y_H(t)\, dY_H(t)}{\int_0^M Y_H^2(t)\, dt} = \alpha + \frac{\int_0^M Y_H(t)\, dB_H(t)}{\int_0^M Y_H^2(t)\, dt} \\
\hat{\alpha}_2 &= \frac{\int_0^M Y_H(t)\, \delta Y_H(t)}{\int_0^M Y_H^2(t)\, dt} = \alpha + \frac{\int_0^M Y_H(t)\, \delta B_H(t)}{\int_0^M Y_H^2(t)\, dt} \\
\hat{\alpha}_3 &= -\left(\frac{1}{H\Gamma(2H)M} \int_0^M Y_H^2(t)\, dt \right)^{-1/2H}
\end{aligned}$$

ここで，$\hat{\alpha}_1$ の定義の分子で使われている積分は，Wick 積分であり，$\hat{\alpha}_2$ の積分では，通常のリーマン和による定義が使われている．この中で，$\hat{\alpha}_1$ と $\hat{\alpha}_3$ は，観測時間 M が無限大のとき一致性をもつが，$\hat{\alpha}_2$ は一致性をもたない．また，$\hat{\alpha}_3$ の方が，$\hat{\alpha}_1$ よりも効率的であることが知られている．なお，有限の M に対して，これらの分布関数や密度関数の計算は，今のところ，未解決な問題である．

MLE $\tilde{\alpha}_{MLE}$ は，フラクショナルな場合の Girsanov の定理により，

$$\tilde{\alpha}_{MLE} = \frac{\int_0^M Q_H(t)\, dZ_H(t)}{\int_0^M Q_H^2(t)\, dv_H(t)} = \frac{U_H}{V_H},$$

で定義される．ここで，3 つの確率過程 $Z_H(t), Q_H(t), v_H(t)$ が新たに導入されている．MLE の分布を計算するために，U_H と V_H の同時積率母関数 $m(\theta_1, \theta_2)$ を，Klepstyna and Le Breton (2002) の結果を使って導出して，分布関数は次の公式に従って数値計算することができた．

$$P(\tilde{\alpha}_{MLE} < x) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{1}{\theta} \operatorname{Im}\left[m(-i\theta, i\theta x) \right] d\theta$$

ただし，

$$\begin{aligned}
m(\theta_1, \theta_2) \;=\; & e^{-M(\alpha+\theta_1)/2} \left[\left(1 + \frac{(\alpha+\theta_1)^2}{\mu^2} \right) \cosh^2 \frac{\mu M}{2} - \frac{\alpha+\theta_1}{\mu} \sinh \mu M \right. \\
& + \frac{\pi M}{4 \sin \pi H} \left\{ -\frac{(\alpha+\theta_1)^2}{\mu} I_{-H}\left(\frac{\mu M}{2} \right) I_{H-1}\left(\frac{\mu M}{2} \right) \right. \\
& \left. \left. + \mu I_{1-H}\left(\frac{\mu M}{2} \right) I_H\left(\frac{\mu M}{2} \right) \right\} \right]^{-1/2},
\end{aligned}$$

ここで，$\mu = \sqrt{\alpha^2 - 2\theta_2}$ であり，$I_\nu(z)$ は，階数 ν の変形 Bessel 関数である．

他方，MCE $\hat{\alpha}_{MCE}$ は，

$$\hat{\alpha}_{MCE} = \frac{-M/2}{\int_0^M Q_H^2(t)\, dv_H(t)}$$

で定義される．MLE と MCE については，密度関数を数値積分により計算して，グラフ表示した．また，$M \to \infty$ のときの漸近分布については，次の結果を示した．

$$\sqrt{M}\,(\tilde{\alpha} - \alpha) \ \& \ \sqrt{M}\,(\tilde{\alpha}_{MCE} - \alpha) \ \to \ \mathrm{N}(0,\, -2\alpha) \qquad (\alpha < 0)$$

なお，$\alpha = 0$ の場合は，次の結果が成り立つ．

$$M\,\hat{\alpha}_{MCE}(H, M) \ \stackrel{\mathcal{D}}{=} \ \hat{\alpha}_{MCE}(H, 1) \qquad (\alpha = 0)$$

この結果は，他の推定量についても成り立つ．したがって，$\alpha = 0$ の場合，推定量は，$M-$ consistent であるが，漸近分布は非正規である．

3. 検定

パラメータ α に関する検定問題

$$H_0 : \alpha = 0 \qquad \text{vs.} \qquad H_1 : \alpha < 0$$

を考察した．この問題は，離散的な場合の拡張された単位根近接モデル

$$y_j = \rho y_{j-1} + v_j, \quad \rho = 1 + \frac{\alpha}{T}, \quad (1-L)^{H-1/2} v_j = \varepsilon_j, \quad (j = 1, \cdots, T),$$

における単位根検定の連続時間モデル・バージョンであると解釈することができる．

検定は，MLE と MCE に基づく統計量を考えて，検出力を計算した．そのために，まず，帰無分布の分位点を計算して，次の結果を得た．

	Probability of a smaller value						
H	0.01	0.05	0.1	0.5	0.9	0.95	0.99
			$\tilde{\alpha}_{MLE}$				
0.5	−13.696	−8.039	−5.714	−0.853	0.928	1.285	2.033
0.7	−13.608	−7.964	−5.648	−0.836	0.899	1.250	1.989
0.9	−12.988	−7.415	−5.154	−0.636	0.767	1.084	1.791
			$\hat{\alpha}_{MCE}$				
0.5	−14.510	−8.856	−6.533	−1.721	−0.418	−0.302	−0.179
0.7	−14.314	−8.673	−6.359	−1.606	−0.378	−0.272	−0.161
0.9	−13.196	−7.630	−5.376	−1.027	−0.209	−0.149	−0.087

そして，検出力を計算することにより，MLE に基づく検定の方が，検出力が高くなることを見出した．

4. 今後の課題

ここで考察した fO-U 過程は，通常の O-U 過程を拡張したものであり，長期記憶性をもつような確率過程を描写するモデルとして，今後，ますます研究が進められると思われる．現時点で未解決な問題は，パラメータ α の最小 2 乗推定量の分布の導出，離散時間モデルの推定との関係，Hurst パラメータ H の推定問題などである．これらについても，今後の研究で明らかにしていきたい．

時間や変数空間上で変化する回帰係数について

佐藤 健一 （広島大学・原爆放射線医科学研究所）
冨田 哲治 （県立広島大学・経営情報学部 ）

回帰分析において、時間とともに変化する回帰係数は変化係数とよばれ、Hastie & Tibshirani (JRSS, 1993)らによって提案された。変化係数は時間軸上で変化する説明変数の効果を曲線として視覚化できるため解釈が容易である。その推定は、一般的には時間軸に沿った平滑化によって行われる。すなわち、固定された時間近傍ごとのデータを用いて回帰を繰り返し行い、その回帰係数を時間軸上に並べることで連続的な関数として推定する。

しかし、この方法では、各点ごとの信頼区間しか構成できず、変化係数曲線全体を包含する意味での同時信頼区間を構成することは困難であった。そこで、Satoh & Yanagihara(AJMMS, 2010)は、変化係数を線形な基底関数族に限定することで、成長曲線モデルにおいて曲線としての同時信頼区間を構成することに成功した。また、図 1 に示すように、従来、説明変数と時間の交互作用として示されていた複数の回帰係数をひとつの変化係数として要約でき、解析結果の解釈が容易になった。

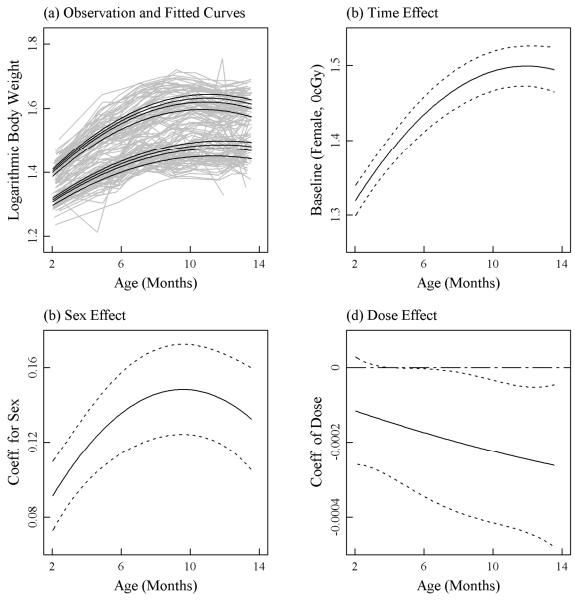


図 1. 中性子線照射後に経時測定された 124 匹のマウスの体重データ. a)個体ごとの体重データと成長曲線モデルによる推定曲線, b) メスの非照射群に対する推定曲線, c)メスに対するオスの増分を示す変化係数, d)線量効果を示す変化係

成長曲線モデルは目的変数として連続的な観測値しか扱えないため、佐藤・柳原・加茂（応用統計学, 2009）では、離散分布を目的変数とする一般化推定方程式の枠組の中で、線形構造を持つ変化係数の推測方法を提案し、線形な変化係数の利用範囲をさらに広げた（参照: 図 2)。

ここまでの研究では、変化係数として時間軸方向に変化する回帰係数を扱ったが、冨田・佐藤・柳原（応用統計学, 2010）では、空間上の位置によって変化する回帰係数を変化係数曲面として捉え、従来、平滑化が使われてきた地理的加重回帰と対比させながら、線形な変化係数の推測を空間データに対して展開した。さらに、冨田・佐藤他（長崎医学会雑誌, 2010)、Tonda, Satoh 他（Radiat. Environ. Bioph., 2012)、佐藤・冨田他（長崎医学

会雑誌，2012)においては、線形な変化係数曲面を座標平面上のハザード関数に応用することで、広島原爆被爆者の死亡危険度を従来の爆心地からの距離や被曝線量だけでなく、被爆位置によっても変わり得る解析を行った。その結果、北西方向に死亡危険度が高くなる傾向が示され、同じく北西方向で目撃情報の多かった黒い雨との関係が示唆された（参照：図 3)。

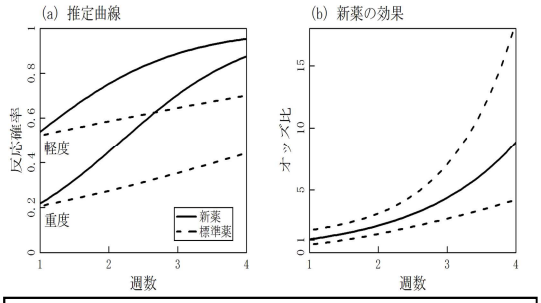


図 2. 診断群別（軽度/重度）投与群別（標準薬/新薬）の診断後 1,2,4 週で行われた検査結果（正常反応なら 1,その他 0）の経時測定データ. a)一般化推定方程式により推定された各群に対する反応曲線，b)新薬

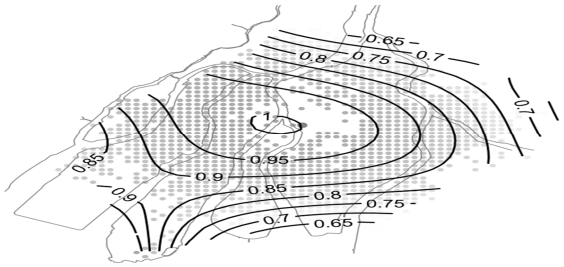


図 3. 広島市内約 6 キロ四方の地図と位置情報による原爆被爆者の死亡危険度. ハザード関数に空間上で変化する線形な変化係数（楕円放物曲面基底）を仮定し、推定されたハザード比を等高線として図示した。北西

このようにして、線形な変化係数の適用対象は連続変数から離散変数、時間から空間、そして、空間上の生存時間へと拡張されていった。この際、2 つの改良すべき点が明らかになった。1 つ目は同時信頼区間の精密化である。これに対して、冨田・佐藤（2013）は多重比較の手法を用いた改良を提案した。2 つ目は、非線形構造への対応である。これまでは、線形な変化係数の基底関数として高次の多項式を仮定し、変量選択によって適度に次数を下げることで非線形性の記述を試みていた。これに対して、佐藤・冨田（2013）では、経時測定データに対して線形性と非線形性を併せ持つ自由度の高いセミパラメトリックな変化係数を仮定し、縮小推定によって適度な硬さを持つ曲線の推定を実現している。今後の課題としては、セミパラメトリックな変化係数の同時信頼区間の評価などが挙げられる。

1. Satoh, K. and Yanagihara, H. (2010)：Estimation of varying coefficients for a growthcurve model, American Journal of Mathematical and Management Sciences, Vol 30, No 3&4, 243-256.2010. DOI:10.1080/01966324.2010.10737787
2. T. Tonda, K. Satoh, K. Otani, Y. Sato, H. Maruyama, H. Kawakami, S. Tashiro, M.Hoshi and M. Ohtaki. (2012)：Investigation on circular asymmetry of geographical distribution in cancer mortality of Hiroshima atomic bomb survivors based on risk maps：analysis of spatial survival data, Radiation and Environmental Biophysics, 51 (2), 133-141. DOI：10.1007/s00411-012-0402-4
3. 冨田哲治，佐藤健一. (2013)：線形な変化係数における信頼区間の精密化，応用統計学，42，11-21.
4. 佐藤健一，冨田哲治. (2013)：混合効果モデルを用いたセミパラメトリックな変化係数の推測について，応用統計学，42，1-10.

有限区間における変化係数の同時信頼区間の構築について

冨田哲治 [1]，佐藤健一 [2]
[1] 県立広島大学経営情報学部，[2] 広島大学原爆放射線医科学研究所

1. はじめに：経時データにおける回帰モデルにおいて，時間 t とともに変化する共変量の効果 $\beta(t)$ は変化係数とよばれている．変化係数の推定は，カーネル平滑化の要領で固定した時点周辺の近傍データに対して局所的な回帰を繰り返すことで推定され，その信頼区間も固定した各点毎に構築されるのが一般的であった．近年，Satoh and Yanagihara (2010, Amer. J. Math. Management Sci. 30:243-256) は，変化係数の関数形を線形に限定することで $t \in \mathbb{R}$ での同時信頼区間を提案した．本稿では，冨田・佐藤 (2013, 応用統計学 42:11-21) が提案した有限区間 $t \in [a, b]$ における同時信頼区間の構築法を紹介する．

2. 変化係数の信頼区間：線形な変化係数 $\beta(t) = \boldsymbol{\theta}'\boldsymbol{x}(t)$ に対して未知母数ベクトル $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ の推定量 $\hat{\boldsymbol{\theta}}$ の漸近分布が $\hat{\boldsymbol{\theta}} \sim \mathrm{N}_q(\boldsymbol{\theta}, \Omega)$ として得られているとき，変化係数 $\beta(t)$ の推定量とその漸近分布は $\hat{\beta}(t) = \hat{\boldsymbol{\theta}}'\boldsymbol{x}(t) \sim \mathrm{N}(\beta(t), \lambda^2(t))$ で与えられる．ただし，$\lambda(t) = \sqrt{\boldsymbol{x}(t)'\Omega\boldsymbol{x}(t)}$ である．ここでは，Ω の推定量を $\hat{\Omega}$ とし変化係数 $\beta(t)$ に関する $100(1-\alpha)\%$ 信頼区間，

$$\mathcal{I}_{1-\alpha}(t|u_\alpha) = \left[\hat{\beta}(t) - u_\alpha\hat{\lambda}(t),\ \hat{\beta}(t) + u_\alpha\hat{\lambda}(t)\right], \quad \hat{\lambda}(t) = \sqrt{\boldsymbol{x}(t)'\hat{\Omega}\boldsymbol{x}(t)}$$

の構築法について考える．u_α は信頼区間 $\mathcal{I}_{1-\alpha}(t|u_\alpha)$ の被覆確率 $\Pr\left(\eta(t) \in \mathcal{I}_{1-\alpha}(t|u_\alpha)\right)$ を決める閾値である．本稿では，同時信頼区間を構築するために有限区間 $t \in [a, b]$ における被覆確率が $\Pr\left(\beta(t) \in \mathcal{I}_{1-\alpha}(t|u_\alpha)\right) = 1 - \alpha$ を満たす変化係数 $\beta(t)$ の同時信頼区間を考える．そのためには，

$$\Pr(\mathcal{T} \le u_\alpha) = 1 - \alpha, \quad \mathcal{T} = \sup_{t\in[a,b]} |\mathcal{T}(t)|, \quad \mathcal{T}(t) = \frac{\hat{\beta}(t) - \beta(t)}{\lambda(t)},$$

を満たす閾値 u_α を求める必要があるが，一般にこのような u_α を求めることは困難である．そこで，ここでは，区間 $[a, b]$ を K 個の時点に離散化した時点の集合 $I_K(a, b) = \{t_k, k = 1, \cdots, K | a = t_1 < t_2 < \cdots < t_K = b\}$ における $\mathcal{T}(t)$ の値 $T_k = \mathcal{T}(t_k)$ $(k = 1, \ldots, K)$ を用いて，$\mathcal{T}$ を次式で近似する．

$$\mathcal{T} \approx T_{max} = \max_{t\in I_K(a,b)} |\mathcal{T}(t)| = \max\left\{|T_1|, |T_2|, \ldots, |T_K|\right\}.$$

K が十分大きい時，T_{max} は $\mathcal{T}$ のよい近似となる．このとき，

$$\Pr(T_{max} \le u_\alpha) = 1 - \alpha \ \Leftrightarrow\ \Pr\left(|T_1| \le u_\alpha, \ldots, |T_K| \le u_\alpha\right) = 1 - \alpha$$

が成り立つ．また，$\boldsymbol{t} = (T_1, \ldots, T_K)'$ は漸近的に多変量正規分布，$\boldsymbol{t} = (T_1, \ldots, T_K)' \rightsquigarrow \mathrm{N}_K(\boldsymbol{0}, R)$ に従う．ここで，$R = DC\Omega C'D$, $C = (\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_K))'$, $D = \mathrm{diag}(\lambda(t_1), \ldots, \lambda(t_K))^{-1}$ である．したがって，閾値 u_α は多変量正規分布から近似的に求まる．今，$\Phi_K(x|R) = \Pr(-x \le t_1 \le x, \ldots, -x \le t_K \le x|R)$ と定義すると，$\Phi_K(x|R)$ は多変量正規密度の K 次元の多重積分

$$\Phi_K(x|R) = \frac{1}{|R|^{1/2}(2\pi)^{K/2}} \int_{-x}^{x} \cdots \int_{-x}^{x} e^{-\frac{1}{2}\boldsymbol{t}'R^{-1}\boldsymbol{t}} d\boldsymbol{t},$$

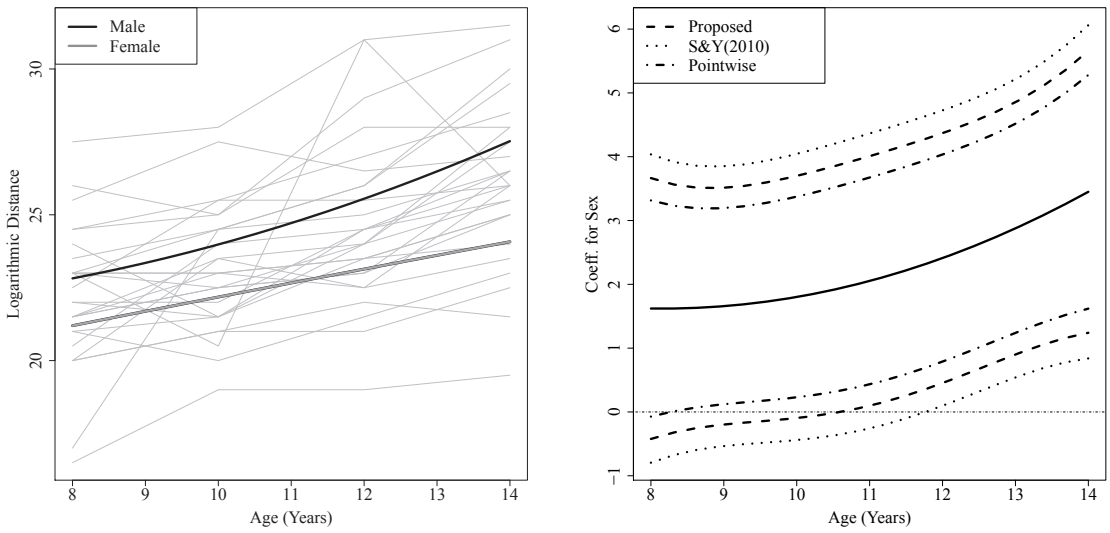


図 1　少年・少女の下顎枝の成長データにおける適合曲線と変化係数の推定値．左図：観測値の折れ線と適合曲線，右図：性差の経時変化を表す変化係数.

で表される．したがって，$\Phi_K(x_\alpha|R) = 1-\alpha$ を満たす x_α を用いて $u_\alpha = x_\alpha$ とすることで，$t \in [a,b]$ における被覆確率が $\Pr\left(\beta(t) \in \mathcal{I}_{1-\alpha}(t|x_\alpha)\right) \approx 1-\alpha$ を満たす同時信頼区間 $\mathcal{I}_{1-\alpha}(t|x_\alpha)$ が構築される．多変量正規分布における多重積分の計算法については，Genz and Bretz (2009) が詳しく，Genz (1992, 1993) および Genz and Bretz (2002) のアルゴリズムに基づく方法が統計ソフト R の mvtnorm パッケージ (Hothorn, Bretz and Genz 2001) で提供されており，1000 次元までの計算に対応している．

3. 適用例：Satoh and Yanagihara (2010) の解析例で用いられた経時データを用いて，変化係数の信頼区間：(1) 各点毎の信頼区間，(2) $t \in \mathbb{R}$ における同時信頼区間 (Satoh and Yanagihara 2010)，(3) $t \in [a,b]$ における同時信頼区間 (冨田・佐藤 2013)，の比較を行い提案法の有効性について検証する．データは Potthoff and Roy (1964) に掲載された少年 16 人および少女 11 人の下顎枝の長さ (mm) の成長データ (8 歳から 14 歳まで 2 年間隔で計 4 回測定) である．年齢 t における下顎枝の長さを $y(t)$ とし，a を性別 (女性なら $a=0$，男性なら $a=1$) を表す変数とする．このとき，期待値 $\mathrm{E}[y(t)]$ が変化係数 $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t))'$ を用いて，$\mathrm{E}[y(t)] = \boldsymbol{a}'\boldsymbol{\beta}(t) = \beta_1(t) + \beta_2(t)a, \quad \beta_j(t) = \boldsymbol{\theta}_j'\boldsymbol{x}(t)$, と記述されているとする．ただし，$\boldsymbol{a} = (1,a)'$ である．このとき，$\beta_1(t)$ は女性の下顎枝の経時変化を表す変化係数，$\beta_2(t)$ は性差の経時変化を表す変化係数である．変化係数の関数形に放物線を仮定し基底を $\boldsymbol{x}(t) = (1,t,t^2)'$ とし，未知母数の推定には変量効果モデル (例えば，Laird and Ware 1982) を利用した．測定年齢が 8 歳から 14 歳なので，$t \in [8,14]$ における被覆確率が 0.95 である同時信頼区間を考える．提案法により同時信頼区間の閾値 $x_{0.05} = 2.365$ と求まる．一方，Satoh and Yanagihara (2010) が提案した $t \in \mathbb{R}$ での同時信頼区間の閾値は $\sqrt{c_{3,0.05}} = 2.795$ である．また，時点 t を固定した時の信頼区間の閾値は $q_{0.05/2} = 1.960$ である．図 1 に，$\beta_1(t)$ の推定値を実線，提案法を $\mathcal{I}_{0.95}(t|2.365)$ を破線，Satoh and Yanagihara (2010) の $\mathcal{I}_{0.95}(t|2.795)$ を点線，各点毎の信頼区間 $\mathcal{I}_{0.95}(t|1.960)$ を示す．図 1 より，提案法による同時信頼区間は Satoh and Yanagihara (2010) よりも狭い領域であり，より精度の高い信頼区間が得られ，提案法の有効性が確認された．また，各点毎の信頼区間は 2 つの同時信頼区間に比べて狭くなっていることに注意されたい．

経時テキストデータに対する多次元尺度法の応用

大分大学　和泉志津恵、広島大学　佐藤健一

１。はじめに

近年、注目されているビッグデータの一つに経時的に観測されたテキストデータがある。テキストデータに含まれるキーワードの時系列的な出現パターンを抽出し、その分類を行うことで、テキストデータの特徴的な変化として要約できる可能性がある。

一方、横軸に測定時間、縦軸にキーワード出現の有無をとった散布図を考えると、出現パターンの抽出方法として二値離散データに対する平滑化が考えられる。平滑化の手法としては大きく分けて、１）局所線形回帰を繰り返し行うカーネル平滑化、２）スプライン基底などを用いて大域的に推定を行うパラメトリック回帰、がある。ここでは、後者に大別される混合効果モデルを用いた推定方法（例えば、 Brumback et al. (JASA, 1999)、 佐藤・冨田(応用統計学, 2013)）を適用し、その推定結果を利用した分類方法と視覚化を提案する。

２．経時的テキストデータの出現パターンの抽出と分類

まず、n 時点において集められたテキストデータから形態素解析により m 個の高出現頻度の単語をキーワードとして抽出する。そして、抽出した m 個のキーワードの各時点における出現の有無から n 行 m 列の 2 値行列を生成する。次に、出現の有無を目的変数として、測定時点を用いた直線といくつかの節点を持つ折れ線を基底とするセミパラメトリックなロジスティック回帰を行う。このとき、折れ線にリッジ型の罰則を与えることは、Brumback et al. (1999) の提案手法から、折れ線の回帰係数にランダム効果を仮定した一般化線形混合モデルを考えることに等しく、罰則パラメータと回帰係数は同時に推定される。

次に、キーワードごとに推定された理論曲線の分類を考える。理論曲線を要約するもっとも自然な要約量は回帰係数ベクトルである。しかしながら、回帰係数ベクトルの取り得る範囲は広く、また理論曲線が類似していても回帰係数ベクトルとしては大きな差を持つことがある。そこで、本稿では理論曲線の値が(0, 1)区間に限定されることに着目し、節点の位置と測定時点の両端における理論値を理論曲線の要約量として用いることを考える。このとき、要約量の次元は回帰係数ベクトルの長さに等しく、出現パターンが類似していれば要約量も類似する。

さらに、要約量に対して k-means 法を適用することで出現パターンを分類する。結果として得られる代表的な要約量は経時的にプロットすることで代表的な出現パターンとして視覚化できる。また、出現パターンの形状は失われるが、すべてのキーワードの要約量とその代表点に対して多次元尺度法を行うことで分類結果は端的な 2 次元散布図として表現できる。

３．実データへの応用

提案した経時的に観測されたテキストデータの出現傾向の抽出と分類の方法および視覚化について、実データに適用した例を紹介する。ここでは、広島市長が世界に向けて発表した平

和宣言の内、1950 年を除く 1947 年-2012 年の 65 時点に観測された日本語テキストデータを用いる。このテキストデータ中のキーワードの出現の有無が時間とともにどのように変化するのか、そして、どのキーワードが共通の出現傾向を持つのかに関心がある。経時テキストデータに RMeCab（石田, 2008）を用いて形態素解析を行い、高出現頻度の上位 52 単語を抽出した。この結果、得られた観測値からなる 52 行 65 列の 2 値行列に対して、統計解析ソフトウェア R（R Core Team, 2012）を用いて、提案方法によるキーワードに関する要約を行った。キーワードごとの出現傾向を 5 つのクラスタに分類すると、クラスターに含まれるキーワードとその数は、出現確率が増加・減少する C1｛強い、軍縮、実験、国家、連帯、生存｝の 6 個、0.8 前後でほぼ一定の確率を保つ C2｛世界、平和、人類、広島、戦争、市民、迎える｝の 7 個、確率が増加し続ける C3｛都市、実現、会議、政府、求める、市長、国連、地球、条約、開催、心、御霊、援護、声、未来、思い、世紀｝の 17 個、0.5 前後でほぼ一定の確率を保つ C4｛人々、人間、体験、禁止、努力、道、新た、誓う、決意、記念、確立、破壊、恒久｝の 13 個、1950-70 年代の初期の頃に増加し一定の確率を保つ C5｛核兵器、被爆、核、廃絶、原爆、ヒロシマ、国際、訴える、犠牲｝の 9 個となった。多次元尺度法によって出現傾向の要約量とクラスタ平均を 2 次元散布図に示すことができる。この図では、提案方法によるキーワードの出現確率の経時的な出現傾向の分類結果とキーワードとの関係が表されており、従来の時間順序に依存しない 2 値の同時生起に基づく多次元尺度法の結果と異なる点が多い。

ここでは平和宣言を要約する 5 つのクラスターを実学的な視点から解釈し、特徴づけを試みる。分類 C1 に含まれるキーワードは、国際関係を表すと考える。東西冷戦の終結とともない、旧ソ連（主にカザフスタン）を含む大国が 1996 年を最後にそれぞれ地下核実験を停止した。核実験に対する懸念や軍縮に対する期待が表れたキーワードの出現確率は年とともに増加していき、1990 年頃から減少し始める。次に、分類 C2 に含まれるキーワードは、広島の普遍的な平和観を表すと考える。観察期間の間に市長が 8 回交代したものの、キーワードの出現確率は 0.8 位の高い値に保たれている。次に、分類 C3 に含まれるキーワードは、平和の担い手の変化を表すと考える。1982 年に開催された第 2 回国連軍縮特別総会を契機に国から都市というより小さな市民社会の単位において平和の実現をめざすことを表すキーワードの出現確率は年々増加していく。次に、分類 C4 に含まれるキーワードの出現確率は観察期間を通して 0.5 程度に保たれてことから、分類 C2 のように高頻度ではないものの、やはり、ある程度の普遍的な要素と考える。最後に、分類 C5 に含まれるキーワードは、1945 年からの空白の 10 年を表すと考える。このように提案方法を用いることで平和宣言を端的に 5 つの要素として要約でき、社会背景を考慮した解釈も可能となった。

４．まとめ

経時的に観測されたテキストデータにおいて、キーワードの出現の有無に着目し、キーワードの経時的な出現傾向を混合効果モデルのもとで推定曲線として要約し、これを分類・視覚化する方法を提案した。実データへ適用して得られた結果の実学的な解釈は、本方法の妥当性を検討する上で重要な要素となり得た。

連絡先: 和泉 志津恵 (Email: shizue@oita-u.ac.jp)

人口動態統計に基づく人間の寿命限界の推定

華山宣胤

尚美学園大学芸術情報学部情報表現学科

E-mail: nob-hanayama@jcom.home.ne.jp

1. Introduction

In the field of biology theories of aging are roughly divided into two major groups. One is consisting of damage theories and the other is consisting of program theories. According to damage theories we age because our systems break down over time. Meanwhile, in the program theories, it is considered that we age because there is an inbuilt mechanism that tells us to die. If the damage theories are true, we can survive any longer by avoiding damaging our organism. If the program theories are true, on the other hand, we cannot survive longer than the upper limit of longevity with any effort.

Considering the above our aim in this study is to see if the damage theories are true, which means that there exist a upper limit of human longevity, or the programing theories are true, which means that there does not exist such a limit, by applying the extreme value theory to an analysis of data for Japanese oldest old survivors obtained using the method of extinct cohort (Wilmoth, Andreev, Jdanov, and Glei, 2007).

2. Generalized Pareto model

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent lifetime of individuals with common distribution function F and Denote an arbitrary term in the X_i sequence by X. Then, according to the extreme value theory, if it can be considered that $u > 100$ is a large enough constant, the distribution function of $X-u$ conditional on $X > u$ is approximately the generalized Pareto distribution, which is written as

$$F(y;\gamma,a) = \begin{cases} 1-(1+\gamma y/a)^{-1/\gamma} & \gamma \neq 0 \\ 1-\exp(-y/a) & \gamma = 0 \end{cases}. \tag{2}$$

An important property of this distribution is that the upper limit of distribution varies depending on parameter values. That is, if $\gamma < 0$ then the upper limit of the distribution is finite, that is $0 < Y < -a/\gamma \equiv \omega$, and if $\gamma \geq 0$, then the upper limit of the distribution is infinite, that is, $0 < Y < \infty$. Considering those properties, this study is laid out as follows (1) to test the hypothesis $\gamma = 0$, (2) to estimate the upper limit of life for each cohort.

Let M_{ij} indicate the numbers of survivors who are $100+i$ years old ($i = 1, 2, \cdots, I$) in the year $1976+j$ ($j = 1, 2, \cdots, 26$), where I is supposed to be sufficiently large so that $M_{Ij} = 0$ for all j, and the numbers of deaths in a year are obtained as difference of M_{ij}, that is, $m_{ij} = M_{ij} - M_{i+1,j+1}$. Now assume that individuals die in the year $1976+j$ at the age of $100+i$ with probabilities given by $p_j = F(i+1,\gamma_j,a_j) - F(i,\gamma_j,a_j)$. Then the contribution of m_{ij} to the likelihood is $\left(M_{ij}/m_{ij}\right) q_{ij}^{m_{ij}} (1-q_{ij})^{M_{ij}-m_{ij}}$, where $q_{ij} = p_j / (1-F(i,\gamma_j,a_j))$, hence the likelihood is proportional to

$$\prod_{i=1,j=1}^{I,J} \binom{M_{ij}}{m_{ij}} q_{ij}^{m_{ij}} (1-q_{ij})^{M_{ij}-m_{ij}}. \tag{2}$$

Thus models assumed on p_{ij} thus fall into the class of binomial regression models, and the parameters in the generalized Prato distribution are estimated by maximizing the likelihood function (2).

3. Numbers of survivors obtained using the method of extinct cohort

The numbers of deaths by year when they were born are indicated in Vital Statistics in Japan. So, the population size for a cohort at a certain age is estimated by summing all future deaths for the cohort with the method of extinct cohort, which is one whose members are assumed to have all died by the end of the observation period.

The method of extinct cohort is explained as follows. Let $D_{ij}^{(U)}$ indicate the number of deaths who are at the age of i in the year j and were born in the year $j-i-1$ and $D_{ij}^{(L)}$ indicate the number of deaths who are at the age of i in the year j and were born in the year $j-i$. Then, assuming that there is no migration or error, the number of survivors who have reached the age of i, say S_{ij}, is

$$S_{ij} = D_{i,j}^{(L)} + \sum_{k=1}^{\infty} \left(D_{i+k,j+k}^{(U)} + D_{i+k+1,j+k}^{(L)} \right). \tag{3}$$

4. Result

Table 1 indicates the ML estimates of the parameter γ and ω obtained by applying the method of extinct cohort to the data for male and female deaths. As shown in the table, the estimates of γ are negative for all male cohorts with sufficiently small p-values except the cohort 1968 for male and the cohorts 1944, 1954 and 1959 for female. Besides, the upper limit of life, that is ω, is estimated between 115 and 262, while those are estimated between 117 and 165 for females with sufficiently small p-values.

Table 1. Estimates for male (left) and female (right)

Year	Gamma	P-value (one side)	Omega	P-value (one side)
1948	-0.16	0.08	117.06	0.00
1953	-0.13	0.13	115.77	0.00
1958	-0.22	0.01	108.41	0.00
1963	0.12	0.88		
1968	-0.08	0.09	127.38	0.00
1973	-0.01	0.43	262.00	0.00
1978	-0.07	0.04	129.43	0.00
1983	-0.09	0.00	123.44	0.00
1988	-0.13	0.00	116.62	0.00
1993	-0.10	0.00	122.40	0.00

Year	Gamma	P-value (one side)	Omega	P-value (one side)
1944	0.18	0.85		
1949	-0.07	0.16	128.71	0.00
1954	0.07	0.88		
1959	0.03	0.70		
1964	-0.09	0.01	121.78	0.00
1969	-0.12	0.00	117.83	0.00
1974	-0.03	0.11	165.00	0.00
1979	-0.05	0.00	143.40	0.00
1984	-0.12	0.00	120.00	0.00
1989	-0.12	0.00	120.83	0.00

References

Arssen, K., de Haan, L. (1994). On the maximal life span of humans. Mathematical Population Studies, Vol. 4, 259-281.

Kannisto, V. (1999). Trends in the mortality of the oldest-old, Statistics, Registries, and Science: Experiences from Finland (ed. Alho, J.), Statistics Finland, 177-194.

Keiding, N. (1990). Statistical inference in the Lexis diagram. Philosophical Transactions of the Royal Society of London Series A, Vol. 332, 487-509.

Wilmoth, J.R., Andreev, K., Jdanov, D. and Glei, D.A. (2007). Methods Protocol for the Human Mortality Database. Vachonhttp://www.mortality.org/Public/Docs/MethodsProtocol.pdf.

Acknowledgment. This work was supported by Grant-in-Aid for Scientific Research 24500346.

大規模経済系データにおける様々な多重代入法アルゴリズムの検証

高橋　将宜[†]　伊藤　孝之[††]

はじめに[1]

データが欠測している場合、利用可能なデータサイズが縮小し、効率性が低下する。さらに、観測値と欠測値との間に体系的な差異が存在する場合、統計分析の結果に偏りが発生するおそれがある。したがって、実際の統計分析においては、何らかの形で欠測値に対処することがほとんど常に必須なことであり、欠測データの対処法として多重代入法(Multiple Imputation)[2]が提唱されてきた(Rubin, 1987)。

多重代入法と一口に言っても、ソフトウェアに実装されているアルゴリズムには様々な方法があり、マルコフ連鎖モンテカルロ法(MCMC: Markov chain Monte Carlo)、完全条件付指定(FCS: Fully Conditional Specification)、EMB(Expectation-Maximization with Bootstrapping)の 3 つを有力なものとして挙げられる。現時点において、3 つのアルゴリズム間の優劣は判然としていない。また、多重代入法の擬似データ数(M)をいくつに設定すればよいかについて、明確な答えは見つかっていない。そこで、本稿では、大規模データセットとしての経済センサス‐活動調査の速報データを用いて、3 つのアルゴリズムの比較検証を行った。さらに、経済センサス‐活動調査の速報データに基づくシミュレーションデータを用い、多重代入法の擬似データ数(M)をいくつに設定すればよいかについて検証を行った。

1. 多重代入法の理論

多重代入法では、観測データを条件として、欠測データの事後分布を構築し、この事後分布からの無作為抽出を行うことで、補定にまつわる不確実性を反映させた M 個($M > 1$)の補定済データセットを生成することにより、欠測値を M 個のシミュレーション値に置き換える。これら M 個の補定済データセットを別々に使用して統計分析を行い、しかるべき手法により結果を統合し、点推定値を算出する。

2. 多重代入法の M 数

シミュレーションでは、一般的に、数百以上の副標本($M > 100$)を生成する必要があり、コンピュータの能力が許す限り多くの繰返しを行うべきだと考えられるが、元来、Rubin (1987)によると、多重代入法の M は非常に小さい数字で十分だとされている。一方、近年では、M 数に関して Rubin (1987)への反論が展開されているものの、十分な結論を得るにいたっていない(Hershberger and Fisher, 2003; Carpenter and Kenward, 2007; Bodner, 2008) 。

† 独立行政法人統計センター統計情報・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製表部管理企画課経済センサス業務推進室統計専門職

[1] 本研究の分析結果は、総務省・経済産業省『平成 24 年経済センサス‐活動調査』の速報結果の調査票情報を基に著者が独自集計したものである。また、本稿の内容は、筆者の個人的見解を示すものであり、機関の見解を示すものではない。

[2] 「多重代入法」とは、Multiple Imputation の訳である。総務省統計局及び統計センターでは、Imputation の訳語として「補定」を用いているが、Multiple Imputation の訳語としては「多重代入法」の呼び名が一般的に流通している(高橋, 伊藤, 2013, p.20)。よって、本稿においても、「多重代入法」の用語を用いる。

3. データセットと分析

3.1 経済センサス - 活動調査の速報データを用いた分析

2012 年 2 月に実施された経済センサス - 活動調査の速報データ（産業大分類 I の単独事業所（個人経営以外））のデータ（観測数 277,263）を用い、R の多重代入法パッケージ Amelia (EMB)、MICE (FCS)、Norm (MCMC)の比較検証を行った。欠測の発生メカニズムは、MAR に基づき、売上高（自然対数）データの 20%（55,500 個）を人工的に欠測させた。また、資本金（自然対数）データの 5%（13,600 個）を無作為に人工的に欠測させ（MCAR）、事業従事者数（自然対数）には欠測を発生させていない（欠測率 0%）。シードを 100 個生成し、3 つの多重代入法プログラムにより分析を行った。分析結果の詳細は、当日報告する。

3.2 シミュレーションデータを用いた多重代入データセット数 M の検証

自然対数変換した経済センサス - 活動調査の速報データの情報（平均値、分散・共分散など）を基に、多変量正規分布によって観測数 1000、3 変量のシミュレーションデータセットを 3 つ生成した。シミュレーションデータの基になったデータは、3.1 項と同じく、産業大分類 I の単独事業所（個人経営以外）を用いた。Amelia、MICE、Norm において、シミュレーションデータセット（欠測率 = 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%）に多重代入（M = 2, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500）を施し、多重代入済データセットを用いて、$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log(\text{事業従事者数}_i)$における$\hat{\beta}$とその標準誤差の推定を行った。分析結果の詳細は、当日報告する。

4. 結語

本研究では、Amelia と MICE の補定値の精度には大きな差がないことが分かった。一方、Amelia の計算処理速度は極めて速く、Norm は 27 万×3 変量のデータセットを分析することができなかった。多重代入法の擬似データ数 M については、概ね 5～10 では少なすぎ、20～50 程度が適切だと考えられる。欠測率に応じて、20%未満ならば M = 20、20%～30%ならば M = 30、30%～40%ならば M = 40、40%～50%ならば M = 50 といった具合に設定することが適切だと思われる。また、欠測率に関わらず、M = 100 を超えて得られるものは非常に少ない。欠測率が 50%を超え始めると、たとえ M 数を数百まで拡大したとしても、補定値の精度を保証できなくなると考えられる。

参考文献

[1] Bodner, Todd E. (2008). “What Improves with Increased Missing Data Imputations?,” *Structural Equation Modeling* vol.15, pp.651-675.

[2] Carpenter, James R. and Michael G. Kenward. (2007). *Missing Data in Clinical Trials—A Practical Guide*. Birmingham: UK National Health Service, National Co-ordinating Centre for Research on Methodology.

[3] Hershberger, Scott L. and Dennis G. Fisher. (2003). “A Note on Determining the Number of Imputations for Missing Data,” *Structural Equation Modeling* vol.10, no.4, pp.648-650.

[4] Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[5] 高橋将宜, 伊藤孝之. (2013). 「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～」, 『統計研究彙報』第 70 号 no.2, 総務省統計研修所, pp.19-86.

国勢調査ミクロデータを用いた匿名化技法の有効性の検証*

明海大学経済学部　伊藤 伸介**
(独)統計センター　星野 なおみ

1. はじめに

わが国においては、平成 25 年末に平成 12 年国勢調査の匿名データの作成・提供が開始された。その一方で、将来的には、小地域分析用の匿名データ等、別のタイプの国勢調査の匿名データの要望が出てくる可能性があり、その予備的な研究としてミクロデータに対する匿名化技法の適用可能性を検証することは有用であると考えられる。そこで、本稿では、各種匿名化技法を用いて作成された国勢調査の匿名化ミクロデータを対象に、秘匿性と有用性に関する実証研究を行うことによって、匿名化技法の有効性の検証を行った。

2. 本研究における匿名化ミクロデータの作成方法

本研究では、国勢調査を例に匿名化ミクロデータを試行的に作成した。本研究で使用するデータは、平成 17 年国勢調査(以下「国調」と略称)の個票データをもとに特定の地域(以下「地域 A」と呼称)のレコードから作成したテストデータ(約 100,000 レコード)である。このテストデータには、個人単位で抽出した一般世帯の世帯主のレコードのみが含まれている。

匿名化ミクロデータの作成手順は以下のとおりである。(1)国調のテストデータに対して、労働力状態、従業上の地位と年齢についてはリコーディングを、世帯人員と年齢に関してはトップコーディングをそれぞれ適用した。(2)リコーディングとトップコーディングを施したデータに対して様々な標本抽出率によるサンプリングを行った(本稿では、サンプリング率が 10%の場合の結果に基づいて議論する)。(3)(1)と(2)の匿名化技法が適用されたデータに対して、スワッピングを適用した。具体的なスワッピングの手順としては、(1)世帯主との続き柄等の 11 のキー変数を用いて標本一意を計測した上で、スワッピングの対象となるレコードを選出し、(2)スワッピングの対象レコードの中で優先度の高いレコードをスコアに基づいて探索し、(3)対象レコードに対してターゲット・スワッピングとランダム・スワッピングを適用している。なお、スワッピング率については、1%、2%、3%、5%、10%、20%、30%が設定されている。

3. 国勢調査ミクロデータを用いた秘匿性の評価―マッチングの試み―

本研究では、秘匿性の評価方法の 1 つとして、外部情報とミクロデータのマッチングに焦点を当てる。具体的には、国調以外の政府統計のミクロデータを外部情報とみなした上で、匿名化ミクロデータと外部情報とのマッチングを試みた。本研究においては、スワッピングが施された10%抽出の国調の匿名化ミクロデータに対して、平成 20 年住宅・土地統計調査 (以下「住調」と略称)の地域 A に該当するレコードを含む個票データ(約 10,000 レコード)とのマッチングの実験を行った。住調については、国調のテストデータと調査区が重複するように対象レコードが選定される。

本実験では、国調と住調の両方の共通の調査事項から選ばれた県市区町村番号、世帯人員、性別といった変数を含む様々なキー変数の組み合わせに基づいて、国調と住調のマッチングを行った。本実験においては、国調と住調においてキー変数の区分を合わせた上で、マッチングを行う。また、本実験では、国調の匿名化ミクロデータにおいて上記のキー変数で一意になったレコードを対象に、住調とのマッチングを行っている。さらに、マッチングにおける国調と住調の調査年次の違いについては、国調のレコードにおいて年齢を加算することによって、年次の調整がなされている。

国調の匿名化ミクロデータと住調の個票データのマッチングに関する実験結果から、主として、以下の点が明らかになった。第 1 に、キー変数の数が多くなるにつれて、国調の匿名化ミクロデータにおける標本一意の数が増大していることがわかるが、この傾向は、ターゲット・スワッピングだけでなく、ランダム・スワッピングの場合にも当てはまることが確認された。また、ターゲット・スワッピングとランダム・スワッピングのいずれにおいても、スワッピング率が高くなるにつれて、国調の匿名化ミクロデータにおける標本一意の数が小さくなっている。その理由

---

* 本稿の内容は個人的な見解を示すものであり、統計センターの見解を表すものではないことに留意されたい。

** (独)統計センター非常勤研究員

としては、スワッピングによって、標本一意に該当するレコードが度数 2 以上のセルに該当するグループに移動したことが考えられる。

第 2 に、スワッピング率が上昇するにつれて、マッチングされたレコードの中でスワッピングされたレコードの比率が高くなっているものの、スワッピング率が 30%の場合でもその比率は最大で約 60%となっている。このことは、本実験においてスワッピング率を上げても、国調の匿名化ミクロデータのレコードの中で住調の個票データとマッチングされたレコードの一部に対してのみ、スワッピングが施されていることを意味する。

第 3 に、本実験では国調の標本一意に対する「真のマッチング」[1]の比率が算出されているが、標本一意に占める真のマッチングの比率を確認すると、ターゲット・スワッピングおよびランダム・スワッピングのいずれについてもその比率は約 1%～2%あって、非常に低くなっている。このことは、国調と住調においてマッチングされたレコードの組については、同じ調査区番号を有していてもそれらが同一の世帯のレコードに該当する可能性が低いことを示している。

4. スワッピングの有効性の評価に関する研究

本節では、主として「特殊な一意(special uniques)」のレコードを対象に適用されたスワッピングの有効性を検証するために、スワッピングが適用された匿名化ミクロデータ(以下「スワッピング済データ」)を対象に、その有用性と秘匿性の評価を行った。本研究では、Shlomo *et al.*(2010)に基づいて，有用性と秘匿性の評価指標を作成した。有用性の評価指標(DU)は、絶対距離の平均値(average absolute distance)で与えられる。具体的には、原データとスワッピング済データの両方で集計値を作成した上でセルごとの度数の差の絶対値に関する平均値が算出される。一方、秘匿性の評価指標(DR)については、原データにおけるクロス表の中で度数 1 であるセル数に対して、スワッピング済データにおけるクロス表の中で度数 1 でありかつスワッピングされていないセル数の比率が算出されている。

これらの有用性と秘匿性の評価指標に基づいて、本研究は、R-U マップ(R-U Confidentiality Map)を作成し、ターゲット・スワッピングとランダム・スワッピングについて、有用性と秘匿性の相対比較を試みた。なお、R-U マップで使用する有用性と秘匿性の評価指標に関しては、キー変数の中のあらゆる 2 変数の組み合わせについて計算された評価指標の平均値がそれぞれ用いられている。

R-U マップの結果から主として以下の点が明らかになった。第 1 に、スワッピング率が上昇するほど、情報量損失値としての DU の数値が大きくなり、DR の数値が小さくなる傾向にあることが確認できる。第 2 に、ランダム・スワッピングと比較して、ターゲット・スワッピングにおける DU の数値が大きく、DR の数値が小さくなることが確認できる。このことは、ターゲット・スワッピングについては、ランダム・スワッピングと比較した場合、有用性は低いものの秘匿性は相対的に高いことを示唆している。

5. むすびにかえて

本稿では、国勢調査のミクロデータを用いて、スワッピング等の各種匿名化技法の有効性の検証を行った。外部情報とのマッチングの試みとして、国調の匿名化ミクロデータと住調の個票データとのマッチングを行ったが、マッチングの精度は高くないことから、外部情報とのマッチングの可能性という観点から見た場合、露見のリスクは低いとみなすことができる。また、本研究では、R-U マップをもとに、スワッピング済データにおける有用性と秘匿性の検証も行った。有用性あるいは秘匿性に関する閾値を設定することができれば、適切なスワッピング率およびスワッピングの方法を選択することが可能になることがわかった。

参考文献

Shlomo, N., Tudor, C., Groom, P. (2010) “Data Swapping for Protecting Census Tables”, Domingo-Ferrer, J. and Magkos, E.(eds) *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings*, Springer, pp.41-51.

[1]本研究における真のマッチングとは、国調の匿名化ミクロデータと住調の個票データの間で 1 対 1 にマッチングされたレコードにおいて、調査区番号も同一であることが確認されることである。マッチングされたレコードが真のマッチングに該当するレコードかどうかを確認するために、本実験では、統計センターの内部資料に基づいて、国調と住調でマッチングされたレコードの組を対象に、それらのレコードの調査区番号が同一か否かについて検証作業を行った。

ピットマン確率分割に付随する極限分布への一考察

鹿児島大学　大和　元

1　始めに

Pitman's sampling formula に従う確率分割について、分割の要素の個数 K_n の漸近分布に関連した事項を考えるが、K_n の分布は次で与えられる。

$$P(K_n = k) = \frac{\theta^{[k:\alpha]}}{\theta^{[n]}} c(n,k,\alpha)\alpha^{-k}$$

ここで、$c(n,k,\alpha) = (-1)^{n-k}C(n,k,\alpha)$ であり、 $C(n,k,\alpha)$ は C-number と言われ次式の右辺の係数である：$(st)^{(n)} = \sum_{k=1}^{n} C(n,k,s)t^{(k)}$、ただし、$t^{(k)} = t(t-1)\cdots(t-k+1)$。

補題 1.1 (Pitman (1996), Yamato and Sibuya (2000))

$$\mu'_r = \lim_{n\to\infty} E\Big[\Big(\frac{K_n}{n^\alpha}\Big)^r\Big] = \Big(1+\frac{\theta}{\alpha}\Big)^{[r]} \frac{\Gamma(\theta+1)}{\Gamma(\theta+1+r\alpha)} = \Big(\frac{\theta}{\alpha}\Big)^{[r]} \frac{\Gamma(\theta)}{\Gamma(\theta+r\alpha)}.$$

この μ'_r は、密度関数が

$$\frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} x^{\frac{\theta}{\alpha}} g_\alpha(x) \quad (0<\alpha<1,\ \theta>-\alpha) \qquad (\star)$$

で与えられる分布の r 次の積率である。ここで、g_α はパラメータ α の Mittag-Leffler distribution $(ML(\alpha))$ の密度関数である。 密度関数 g_α $(0<\alpha<1)$ は次式を満たす一意な関数である。

$$\int_0^\infty x^p g_\alpha(x)dx = \frac{\Gamma(p+1)}{\Gamma(p\alpha+1)} \quad (\forall\, p>-1)$$

(Pitman (1996) p.5 and Pitman (2002), p.20-21).
密度関数 $(\star)$ を持つ分布を Pitman-Mittag-Leffler distribution と言う事にし、$PML(\theta,\alpha)$ で表す。 $PML(0,\alpha) = ML(\alpha)$ である。法則収束を $\xrightarrow{d}$ で表すと、補題 1.1 から次が得られる。

補題 1.2 (Pitman (1996), Pitman (2002))

$$\frac{K_n}{n^\alpha} \xrightarrow{d} PML(\theta,\alpha) \quad \text{as} \quad n\to\infty$$

分布 $PML(\theta,\alpha)$ を、従って本質的には $ML(\alpha)$ を、より分かり易い形で表すことが本研究の目的である。以下に示す様に、パラメータを特別な場合に限れば表す事が出来た。

2 PML の分布

命題 2.1 $|X|$ $(X \sim N(0,2))$ は Mittag-Leffler 分布 $ML(1/2)$ に従い、

$$E(|X|^p) = \frac{\Gamma(p+1)}{\Gamma(\frac{1}{2}p+1)} \quad (p > -1).$$

命題 2.2 $r = 1, 2, \cdots$ について、$X_1, \cdots, X_r$ は独立で、$N(0,2)$ に従うとする。

$$Y_r = |X_1| \cdot |X_2|^{\frac{1}{2}} \cdot |X_3|^{\frac{1}{2^2}} \cdots |X_r|^{\frac{1}{2^{r-1}}}$$

は Mittag-Leffler 分布 $ML(1/2^r)$ に従う。 或は、 $r = 1, 2, \cdots$ について、$Z_1, \cdots, Z_r$ は独立で、$N(0,1)$ に従うとする。

$$Y_r = 2^{1-1/2^r}|Z_1| \cdot |Z_2|^{\frac{1}{2}} \cdot |Z_3|^{\frac{1}{2^2}} \cdots |Z_r|^{\frac{1}{2^{r-1}}}$$

は Mittag-Leffler 分布 $ML(1/2^r)$ に従う。

$Y_2 = |X_1| \cdot |X_2|^{\frac{1}{2}}$ の密度関数は

$$f_{|X_1|\cdot|X_2|^{1/2}}(x) = \frac{2}{\pi}\int_0^\infty \exp\Big\{-\frac{1}{4}\Big(\frac{x^2}{t^2} + t^4\Big)\Big\}dt \quad (x \geq 0)$$

自由度 ν のカイ 2 乗分布の密度関数と分布関数を $h_\nu(x)$、$H_\nu(x)$ とおく。

命題 2.3 パラメータ $\alpha = 1/2$、$\theta > -1/2$ について、Pitman-Mittag-Leffler 分布 $PML(\theta, \alpha)$ の分布関数と密度関数は、

$$F_{\theta,1/2}(x) = H_{2\theta+1}\Big(\frac{x^2}{2}\Big), \quad f_{\theta,1/2}(x) = xh_{2\theta+1}\Big(\frac{x^2}{2}\Big), \quad x > 0.$$

$\theta > -\alpha$, $\alpha = 1/4$ に対して、PML の密度関数は、

$$f_{\theta,\alpha}(x) = \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)}x^{\frac{\theta}{\alpha}}\frac{2}{\pi}\int_0^\infty \exp\Big\{-\frac{1}{4}\Big(\frac{x^2}{t^2} + t^4\Big)\Big\}dt$$

$\alpha \neq 1/2^r (r = 1, 2, \cdots)$ の場合の PML の分布を調べることが、今後の問題である。

参考文献

Pitman, J. (1996), Notes on the two parameter generalization of Ewens' random partition structure (unpublished manuscript).

Pitman, J. (2002), Poisson-Kingman Partitions, Technical Report 625, Dept. Statistics, U.C. Berkeley.

Yamato H. and Sibuya, M. (2000), Moments of some statistics of Pitman sampling formula, *Bull. Inform. Cybern.*, **32**, 1–10

確率分割の標本と予測量: 生態学への応用

2013-11-27/29 金沢大学 報告書

慶應義塾大学　渋谷　政昭

要約　　ピットマン確率分割の条件付き確率分割 $S \in \mathcal{P}_{n,k}|(S \in \mathcal{P}_{\nu,\kappa}, n < \nu)$ からランダムにサブサンプルを取り出す逆過程を調べ, $S \in \mathcal{P}_{n,k}$ による (ν, κ) の予測を行う.

まえがき　　生態学調査の基本データは観測・捕獲した種 C_i の個体数 $c_i,\ 1 \le i \le k$, である. $n := \sum_{i=1}^{k} c_i,\ k$, はそれぞれ, 観測・捕獲した**個体総数**, **種の数**である. ここでは $\{c_1, \ldots, c_k\}$ がピットマン確率分割 EPSF $(n; \theta, \alpha)$ であることを前提とする. その主要統計量である種の数 $K_n := \sum_{j=1}^{n} S_j, (S_j$ は $(c_i = j)$ となる C_i の数) の p.m.f., $\mathbb{P}\{K_n = k\} =: f_n(k)$ が満たす前進方程式と陽な表現 は

$$f_{n+1}(k) = \frac{n - k\alpha}{\theta + n} f_n(k) + \frac{\theta + (k-1)\alpha}{\theta + n} f_n(k-1),\ 1 \le k \le n, \tag{1}$$

$$f_n(k) = \frac{1}{(\theta|-1)_n} S_{n,k}(\theta|-1)_n(\theta|-\alpha)_k,\ 1 \le k \le n, \tag{2}$$

ただし $S_{n,k} := S_{n,k}(-1, -\alpha, 0)$ で $S_{n,k}(a, b, c)$ は一般スターリング数である. この分布を EPSF-K $(n; \theta, \alpha)$ で表す. これから EPSF (θ, α) の条件付き分布

$$w(s; n, k) := \mathbb{P}\{S = s|(S \in \mathcal{P}_{n,k})\} = \frac{1}{S_{n,k}(-1, -\alpha, 0)} \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{(1-\alpha|-1)_{j-1}}{j!} \right)^{s_j}. \tag{3}$$

が定まる. これが θ に依らないことに注意. ピットマン確率分割を拡張した**順列置換不変ギッブス確率分割** の概念が, 上の条件により特徴付けられるので. 次節の議論は順列置換不変ギッブス確率分割一般について成り立つ.

種の数 K_n の逆過程　　$f_n(k)$ は三角配列 $\{(n, k);\ 1 \le k \le n < \infty\}$ の上のマルコフ過程, あるいは酔歩と考えられる. その逆過程として, 条件付きピットマン確率分割 (3) から 1 個の個体をランダムに等確率で択んで除くことを考える. K_ν から $K_{\nu-1}$ への推移確率は

$$\mathbb{P}\{K_{\nu-1} = \kappa - 1|K_\nu = \kappa\} = S_{\nu-1,\kappa-1}/S_{\nu,\kappa}$$
$$\mathbb{P}\{K_{\nu-1} = \kappa|K_\nu = \kappa\} = (\nu - 1 - \kappa\alpha) S_{\nu-1,\kappa}/S_{\nu,\kappa} = 1 - S_{\nu-1,\kappa-1}/S_{\nu,\kappa}.$$

この酔歩において個体数 n における種の数 K_ν の分布 (初期条件に依存する) を一般に $g_n(k) := \mathbb{P}\{K_n = k\}$ で表す. 上の推移確率から $g_n(k)$ は次の後退方程式を満たす.

$$g_n(k) = (n - k\alpha) \frac{S_{n,k}}{S_{n+1,k}} g_{n+1}(k) + \frac{S_{n,k}}{S_{n+1,k+1}} g_{n+1}(k+1), \quad 1 \le k \le n. \tag{4}$$

この式により $S|(S \in \mathcal{P}_{\nu,\kappa})$ からの確率標本の p.m.f., $g_n(k),\ 1 \le k \le n$ が漸次定まる. 詳しくは酔歩 $g_k(n) = g_k(n; \nu, \kappa, \alpha)$ は $g_\kappa(\nu; \nu, \kappa, \alpha) = 1$ から出発して $g_1(1; \nu, \kappa, \alpha) = 1$ に到る, 平行四辺形 $\diamond[\nu, \kappa] := \{(n, k); \max(1, n - \nu + \kappa) \le k \le \min(\kappa, n)\}$ の上で (4) を満たすが, 境界では

$$g_k(n; \nu, \kappa, \alpha) = 0, k > \min(\kappa, n), \quad \& \quad S_{n,1}/S_{n+1,1} = 1/(n - \alpha).$$

この逆過程で, 大きさ n のサブサンプルの種の期待数が次式で定まる:

$$E(K_n|S\in\mathcal{P}_{\nu,\kappa})=\frac{1}{S_{\nu,\kappa}}\sum_{k=1}^{\min(\kappa,\nu-n)}S_{\nu-k,\kappa-1}(1-\alpha|-1)_{k-1}\left(\binom{\nu}{k}-\binom{\nu-n}{k}\right),\quad 1<n<\nu. \tag{5}$$

(4) より三角配列上の (ν,κ) から $(\mu,\lambda)\in\diamondsuit[\nu,\kappa]$ に到る下降酔歩が定まり, その逆の上昇酔歩が定まるがその表現は複雑となる.

予測　上の結果に基づき, 次の予測を考える. 観測・捕獲データ $t\in\mathcal{P}_{n_0,k_0}$ を $S|(S\in\mathcal{P}_{\nu,\kappa})$ からの確率標本と前提し, t から κ を推定する. α は t による適当な推定値を用いるとし, ν は $\nu=n_0+1,n_0+2,\ldots$ と動かすこともできるが, 任意に固定する. $(n_0,k_0)\in\diamond[\nu,\kappa]$ の条件から可能な値は $k_0\le\kappa\le k_0+\nu-n_0$ に限られる. 有限の標本空間から有限のパラメータ空間の推定量を定める・

予想

$$\hat{\kappa}=\mathrm{argmax}_\lambda g_{n_0}(k_0;\nu,\lambda)$$

を予測量とすると $\hat{\kappa}$ は k_0 の非減少関数である.

さらにその性質を調べることになるが, α の MLE 推定量が適切ではないことが問題である.

応用例　Osa Peninsula, Costa Rica, の 2 地点における甲虫捕獲調査の次のような結果が示されている. (Colwell, et al. 2012).　(a) $(n,k)=(976,140)$, (b) $(n,k)=(237,112)$,
大きなデータ (a) から 237 個を抽出したときの種の数のモードは 69 である.
小さなデータ (b) から $n=976$ の $g_n(k)$ を予測すると最尤推定値は 234 である.
(b) の方が種の数が遥かに豊富である.

参考文献

[1] Colwell, R.K. et al.(2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages, *Journal of Plant Ecology The Annals of Applied Probability* , **15-1**, 3–21.

[2] Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles, *Mathematical Sciences*, **138**, 5674–5685. (original —Russian version: Zapiski Nauchnnykh Seminarov ROMI, **325**, 2005, 83–102.)

[3] Lijoi, A., Prünster, I. and Walker, S.G. (2008) Baysian nonparametric estimators derived from conditional Gibbs structures, *The Annals of Applied Probability*, **18-4**, 1519–1547.

[4] Pitman, J. (2006) *Combinatorial Stochastic Processes*, Lecture Notes in Mathematics, **1875**, Springer, New York, NY.

[5] Sibuya, M. (2013) Prediction in Ewens-Pitman Sampling Formula and random samples from number-partitions, *Annals of the Institute of Statistical Mathematics*, (in print).

Semiparametric Efficiency for the Quantile-Regression-Based L-Estimation

Takayuki Shiohama (Tokyo Univerisity of Science)
Hiroyuki Taniai (Waseda Univerisity)

Abstract

We study a semiparametrically efficient estimation for a quantile-regression-based L-estiamtors. The L-estimators are based on linear combinations of the linear quantile regression estimator, introduced by Koenker and Bassett (1978), and our proposed L-estimators are based on the one-step estimator of the type discussed by Hallin and Werker (2003) and Hallin et al. (2008). In constructing the estimator, we use the residual signs and ranks based optimal score function with a general reference density. The asymptotic properties of the proposed estimators are discussed. Their finite sample properties are illustrated through a set of Monte Carlo simulations and an empirical applications.

1. Introduction

Quantile regression introduced by Koenker and Bassett (1978) is similar to the standard linear regression, expect that we replace conditional expectations $E(Y|X)$ by a conditional quantile $F^{-1}_{Y|X}(\tau)$. Quantile regression provides us a more complete summary of the conditional distribution. A throughout review of QR can be found in Koenker (2005).

Linear combinations of order statistics, called *L-estimators*, have long been of interest as estimates for the classical location model for robust alternatives. The problem of constructing asymptotically efficient estimators of a finite dimensional Euclidian parameter θ in a semi-parametric model with an infinite dimensional nuisance parameter F has attracted considerable attention throughout last two decades.

2. Quantile-Regression-based L-Estimation

Let $F_{Y|X}(y)$ and $F^{-1}_{Y|X}(\tau) = \inf\{y : F_{Y|X}(y) \geq \tau\}$ denote the conditional distribution function and the τ-quantile of Y given X, respectively. The model of our concern is represented as observations $\{y_t\}_{t=1}^n$ from statistical experiments $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y}^n), \mathcal{P}^n_Y)$ satisfying

$$Y_t = \boldsymbol{X}_t^\top \boldsymbol{\beta}(\tau) + \varepsilon(\tau)_t, F_\varepsilon^{-1}(\tau) = 0,$$

with some covariate vector $\boldsymbol{X}_t = [1, X_{2,t}, \ldots, X_{d,t}]^\top$, for $t = 1, \ldots, n$, and an unknown regression parameter to be estimated $\boldsymbol{\beta} \in \mathbb{R}^d$. Equivalently, we have by taking the τ-quantile on the both hand side of (2.1), conditionally on$\boldsymbol{X}_t$, we obtain the value of $\boldsymbol{\beta}$ in the relation $F^{-1}_{Y|X}(\tau) = \boldsymbol{X}_t^\top \boldsymbol{\beta}(\tau)$. This is known as basic conditional quantile regression model.

In this paper, we consider the asymptotic properties of L-estimation of the linear models via quantile regression and its one-step estomator. Our discussion focuses on the semiparametrically efficient estimation which is based on the quantile regression process $\widehat{\beta}^{QR}_n(\cdot)$ dedined, following Koenker and Bassett (1978), as

$$\widehat{\beta}^{(QR)}_n(\tau) = \underset{\boldsymbol{b}\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^n \rho_\tau(Y_i - X_i^\top \beta), \quad \tau \in \mathcal{T},$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is called check function.

Let $J : (0, 1) \to \mathbb{R}$ be the weight function. For the parameter of interest is

$$\phi_J \circ \boldsymbol{\psi}_\tau(P_Y) := \int_0^1 J(\tau)\boldsymbol{X}^\top \boldsymbol{\beta}(\tau)d\tau = \boldsymbol{X}^\top \Big(\int_0^1 J(\tau)\boldsymbol{\beta}(\tau)d\tau \Big),$$

and its QR based estimator is, for $\tau_i = \frac{i}{n+1}, i = 1, \ldots, n$,

$$\phi_J \circ \boldsymbol{\psi}_\tau(\mathbb{P}_{n,Y}) := \frac{1}{n} \sum_{i=1}^{n} J\left(\tau_i\right) \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_n^{(QR)}(\tau_i)$$

The procedure that we will apply here to achieve semiparametric efficiency is based on the invariance principle, as introduced by Hallin and Werker (2003). By following Hallin and Werker (2003), a semiparametrically efficient procedure can be obtained by projecting $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{b}_n;f}$ on some σ-field to which the generating group for $\{P^{(n)}_{\boldsymbol{b}_n,f} | f \in \mathcal{F}^\alpha\}$ becomes *maximal invariant*. For the quantile-restricted regression model, such a σ-field is studied by Hallin et al. (2008) and found to be generated by signs and ranks of the residuals.

Here, let us denote the sign of a residual as $S_{\boldsymbol{b}_n,i}$, the rank of a residual as $R^{(n)}_{\boldsymbol{b}_n,i}$, and the σ-field generated by them as

$$\mathcal{B}^{(n)}_{\boldsymbol{b}_n} := \sigma(S_{\boldsymbol{b}_n,1}, \ldots, S_{\boldsymbol{b}_n,n}; R^{(n)}_{\boldsymbol{b}_n,1}, \ldots, R^{(n)}_{\boldsymbol{b}_n,n}).$$

A "good" inference should be based on

$$\underset{\sim}{\Delta}^{(n)}_{\beta,f} := E^{(n)}_{\beta,f}[\Delta^{(n)}_{\beta,f} | \mathcal{B}^{(n)}_\beta] = \frac{1}{n^{1/2}} \sum_{i=1}^{n} E^{(n)}_{\beta,f}[\varphi_f(F^{-1}(U_{\beta,i}))|\mathcal{B}^{(n)}_\beta]\boldsymbol{X}_i$$

where $U_{\beta,i} := F(\xi_{\beta,i}) = F(Y_i - \boldsymbol{\beta}^T \boldsymbol{X}_i)$ is the i.i.d. uniform on $[0, 1]$ under $P_{\beta,f}$.

Then we propose a one-step estimator for the quantile regression coefficient processes and discuss its semiparametric inference. From Theorem18.6 of Kosorok (2008), we need to show that the composite function ϕ is Hadamard differentiable, which yields our proposed one-step L-estimators are semiparametrically efficient.

References

Hallin, M., Vermandele, C., and Werker, B. J. M. (2008). Semiparametrically efficient inference based on signs and ranks for median-restricted models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(2):389–412.

Hallin, M. and Werker, B. J. M. (2003). Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli*, 9(1):137–165.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett, Jr., G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*, Springer Series in Statistics. Springer, New York.

On a tangent space for the coefficient functions of Quantile Regression.

谷合 弘行 (Hiroyuki TANIAI, 早稲田大学)

シンポジウム「統計科学の新展開」 2013.11/27-29, 金沢大学サテライト・プラザ

本発表の概要: 一般の分位数回帰モデルに従う確率変数は、一様確率変数を引数とするランダム係数 (random coefficient) 表現を持つ。本発表では、これを一種の混合分布モデル (mixture model) とみなすことによって無限次元パラメータである係数関数の接空間を導くことを試みた。また、さらにこの結果を用いた漸近有効な推定量の構成についても議論した。以下がその概要である：

We observe $\boldsymbol{X}_t = (Y_t, \boldsymbol{Z}_t)$, $1 \le t \le n$ where Y_t is a random variable and $\boldsymbol{Z}_t$ is a d-dimensional random vector, satisfying $F_Y^{-1}(\tau|\boldsymbol{Z}_t = \boldsymbol{z}_t) = \boldsymbol{z}_t^\top \boldsymbol{\beta}(\tau)$, for all $\tau \in (0,1)$. Here, we denote the conditional quantile by $F_Y^{-1}(\tau|S) := \inf\{y : P^Y\{Y \le y|S\} \ge \tau\}$ with some set S given. This is the QR model (cf. Koenker (2005) and/or Koenker and Xiao (2006)) so has an interpretation as the following Random Coefficient Regression model:

$$Y_t = \boldsymbol{Z}_t^\top \boldsymbol{\beta}(U), \quad U \overset{\mathrm{d}}{\sim} \text{Uniform[0,1]}, \tag{1}$$

where $\boldsymbol{\beta}(\cdot) : (0,1) \to \mathbb{R}^d$ is the vector of non-decreasing functions. In particular, if we assume the 1st entry of $\boldsymbol{Z}_t$ is 1 and set $\boldsymbol{\beta}(u) = (\theta_1 + b(u), \theta_2, \ldots, \theta_d)^\top$, then $\boldsymbol{\beta}(u)$ is studied by the location model with restriction to the innovation's u-quantile to be 0. But the same $\boldsymbol{\beta}(u)$ can be studied in a scale model with u-quantile being 1, by setting $\boldsymbol{\beta}(u) = (\theta_1 c(u), \ldots, \theta_d c(u))^\top$. So, in order to have the efficiency gain, different perspective yields different regularity conditions.

In this talk, we focus on the modelling (1), where Y is generated by deterministic function $\boldsymbol{\beta}$, uniform r.v. U, and "random nuisance parameter" $\boldsymbol{Z}$ (cf. Pfanzagl (1982)). The distribution of $\boldsymbol{X}$ is now be seen from

$$P_{\boldsymbol{\beta},\Gamma}(A \times B) = \int_B P^Y(A; \boldsymbol{\beta}|\boldsymbol{z})\Gamma(d\boldsymbol{z}), \quad A \subset \mathbb{R},\ B \subset \mathbb{R}^d$$

where $\Gamma := P^{\boldsymbol{Z}}$ plays a role of mixing distribution. The crucial difference is that the parameter of our interest, $\boldsymbol{\beta}(\cdot)$, is infinite-dimensional.

To determine the tangent spaces for $\boldsymbol{\beta}$ and for Γ, we define the subfamilies

$$\mathfrak{P}_\Gamma := \{P_{\boldsymbol{\beta},\Gamma} : \boldsymbol{\beta} \in \mathcal{M}^d\}, \qquad \mathfrak{Q}_{\boldsymbol{\beta}} := \{P_{\boldsymbol{\beta},\Gamma} : \Gamma \in \mathscr{G}\},$$

with $\mathcal{M}$ being the set of non-decreasing functions on $(0,1)$ and $\mathscr{G}$ being the set of distributions which $\boldsymbol{Z}$ might have. A tangent cone for $\boldsymbol{\beta}(\cdot)$ (at P) is denoted by $T_P(\mathfrak{P}_\Gamma)$, which is the set of (score) functions g satisfying $P[g^2] < \infty$ and

$$\frac{dP_\delta}{dP} = 1 + \delta(g + r_\delta), \qquad P[r_\delta^2] = o(\delta^0)$$

for the path $P_\delta \to P$ (in $\mathfrak{P}_\Gamma$) as $\delta \to 0$. Now, denote the true value of parameters $(\boldsymbol{\beta}_0(\cdot), \Gamma_0)$ and think of a contamination $\boldsymbol{\beta}(\cdot, \delta)$ such that $\boldsymbol{\beta}_\delta(u) := \boldsymbol{\beta}(u, \delta) \to \boldsymbol{\beta}_0(u)$ as $\delta \to 0$. For given $u \in (0,1)$, this $\boldsymbol{\beta}(u, \delta)$ may differ at $\delta \neq 0$ from one parametrization to another. Also, as for the partial derivative $\partial_\delta \boldsymbol{\beta}(u', \delta') := \frac{\partial}{\partial \delta}\boldsymbol{\beta}(u, \delta)|_{(u,\delta)=(u',\delta')}$, we may think of the function $\partial_\delta \boldsymbol{\beta}(\cdot, 0)$ as $\delta^{-1}(\boldsymbol{\beta}_\delta - \boldsymbol{\beta}_0)(\cdot) + o(\delta^0)$ so that each $\partial_\delta \beta_i(\cdot, 0) \in L_2(P_{\boldsymbol{\beta}_0,\Gamma_0})$ if $\boldsymbol{\beta}_\delta$ is such that $\sqrt{n}$-consistent. In fact, each $\partial_\delta \beta_i(\cdot, 0)$ can be assumed square-integrable under $P^Y(\cdot; \boldsymbol{\beta}_0|\boldsymbol{z})$ as well.

Proposition 1 *The tangent space for $\boldsymbol{\beta}(\cdot)$ is found to be*

$$T_{P_{\boldsymbol{\beta}_0,\Gamma}}(\mathfrak{P}_\Gamma) = \{A_\Gamma \boldsymbol{h} : h_i \in L_2(P^Y(\cdot; \boldsymbol{\beta}_0|\boldsymbol{z})),\ i = 1, \ldots, d\},\ \textit{where}$$
$$(A_\Gamma \boldsymbol{h})(y) := -\frac{d}{du}\Big(\frac{\boldsymbol{z}^\top \boldsymbol{h}(u)}{\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u)}\Big)\Big|_{u=(\boldsymbol{z}^\top \boldsymbol{\beta}_0)^{-1}(y)}. \quad (\boldsymbol{\beta}_0'(u) = \frac{d}{du}\boldsymbol{\beta}_0(u))$$

On the other hand, those for Γ will be $T_{P_{\boldsymbol{\beta},\Gamma_0}}(\mathfrak{Q}_{\boldsymbol{\beta}}) = \{(y, \boldsymbol{z}) \mapsto k(\boldsymbol{z}) : {}^\exists k \in T_{\Gamma_0}(\mathscr{G})\}$.

This $A_\Gamma : L_2(P^Y(\cdot; \boldsymbol{\beta}_0|\boldsymbol{z}))^d \to L_2(P_{\boldsymbol{\beta}_0,\Gamma})$ is the score operator for $\boldsymbol{\beta}(\cdot)$. So the score operator for the full model seems to be $A(\boldsymbol{a}, \boldsymbol{b}) \equiv A_\Gamma \boldsymbol{a} + \boldsymbol{b}^\top \nabla k(\boldsymbol{z})$, for $(\boldsymbol{a}, \boldsymbol{b}) \in H := L_2(P^Y(\cdot; \boldsymbol{\beta}_0|\boldsymbol{z}))^d \times \mathbb{R}^d$. Then its adjoint $A^* : L_2(P_{\boldsymbol{\beta}_0,\Gamma_0}) \to H$ can be formally given by

$$A^* f = \Big(P_{\boldsymbol{\beta}_0,\Gamma_0}\Big[\frac{\boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u_0)}\frac{\partial}{\partial y}\big(\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u_0) \cdot f(\boldsymbol{x})\big)\Big]\Big/ dP^Y(y; \boldsymbol{\beta}_0|\boldsymbol{z}),\ P_{\boldsymbol{\beta}_0,\Gamma_0}[\nabla k(\boldsymbol{z}) f(\boldsymbol{x})]\Big)$$

with $u_0 = (\boldsymbol{z}^\top \boldsymbol{\beta}_0)^{-1}(y)$. Using these operators, the efficient influence function for each $\beta_i(u)$, $i = 1, \ldots, d$, can be obtained as $A(A^*A)^{-1}\tilde{\boldsymbol{\chi}}_{i\cdot}(y)$ as in eq.(25.30) of van der Vaart (1998). In our case, $\tilde{\boldsymbol{\chi}}_{i\cdot}(y) = (\Gamma[\partial_\delta \boldsymbol{\beta}(u,0)\partial_\delta \boldsymbol{\beta}(u,0)^\top / \boldsymbol{z}^\top \boldsymbol{\beta}_0'(u)]^{-1}\Gamma[\partial_\delta \beta_i'(u,0)\partial_\delta \boldsymbol{\beta}(u,0)],\ \mathbf{0}_d)$. More specifically, we have the following result:

Proposition 2 *The efficient function for each $\boldsymbol{\beta}_i(u)$ is found to be $\tilde{\psi}_i = A_\Gamma \boldsymbol{k}_{i\cdot}$ where $\boldsymbol{k}_{i\cdot}(u) = (k_{i1}(u), \ldots, k_{id}(u))^\top$ is the solution of the following ODE*

$$k_{ij}'' - \frac{\boldsymbol{z}^\top \boldsymbol{\beta}_0''(u)}{\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u)}k_{ij}' + \Big\{\frac{(\boldsymbol{z}^\top \boldsymbol{\beta}_0''(u))^2 - \boldsymbol{z}^\top \boldsymbol{\beta}_0'''(u)}{\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u)}\Big\}k_{ij}$$
$$= \boldsymbol{\beta}_0'^\top(u)\boldsymbol{\beta}_0'(u)\partial_{u\delta}\beta_j(u,0)\Gamma\Big[\frac{\partial_\delta \boldsymbol{\beta}(u,0)\partial_\delta \boldsymbol{\beta}(u,0)^\top}{\boldsymbol{z}^\top \boldsymbol{\beta}_0'(u)}\Big]^{-1}\partial_\delta \beta_j(u,0)$$

with conditions $k_{ij}(0) = k_{ij}(1) = 0$.

As for the construction of the estimator which realizes this efficient influence function, we made some conjectures and left it as a future work.

参考文献

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *J. Amer. Statist. Assoc.*, 101(475):980–990.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*, volume 13 of *Lecture Notes in Statistics*. Springer-Verlag, New York. With the assistance of W. Wefelmeyer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

順序カテゴリ正方分割表における2変量t分布型対称モデルについて

東京理科大学理工学部　生亀(いき)清貴[1]
東京理科大学理工学部　富澤貞男

1. はじめに

行と列が順序のある同じ分類からなる $R \times R$ 正方分割表において, (i, j) セル確率を p_{ij} とする $(i = 1, \ldots, R;\ j = 1, \ldots, R)$. 対称モデル (Bowker, 1948) は次のように定義される:

$$p_{ij} = p_{ji} \quad (i < j).$$

このモデルはセル確率の対称構造を示している. 線形対角パラメータ対称モデル (Agresti, 1983) は次のように定義される:

$$p_{ij} = \theta^{j-i} p_{ji} \quad (i < j).$$

このモデルは対称的なセル確率の比が主対角線からの距離 $j - i$ に依存して指数的に変化するという構造を示している. 特に $\theta = 1$ とおいたこのモデルは対称モデルである.

2変量正規分布に従う確率変数 U と V を考える, ただし $\mathrm{E}(U) = \mu_1$, $\mathrm{E}(V) = \mu_2$, $\mathrm{Var}(U) = \mathrm{Var}(V) = \sigma^2$, $\mathrm{Corr}(U, V) = \rho$. このとき結合確率密度関数 $f(u, v)$ は, 次の関係をみたす:

$$\frac{f(u, v)}{f(v, u)} = \delta^{v-u},$$

ただし,

$$\delta = \exp\left[\frac{\mu_2 - \mu_1}{\sigma^2(1 - \rho)}\right].$$

したがって, 本来データが連続量で周辺分散が等しい潜在的な2変量正規分布に従うと想定される場合において, いくつかの切断点を設けて正方分割表データを構成するとき, 線形対角パラメータ対称モデルはよく適合すると考えられる. 詳細については Yamamoto et al. (2007) も参考にされたい.

本講演では順序カテゴリ正方分割表の解析において, 潜在分布として周辺分散の等しい2変量t分布が想定される場合に適切であると考えられるモデルを提案した.

[1]〒 278-8510 千葉県野田市山崎 2641
e-mail: iki@is.noda.tus.ac.jp

2. 2 変量 t 分布型対称モデル

自由度 m の 2 変量 t 分布に従う確率変数 U と V を考える, ただし $\mathrm{E}(U) = \mu_1$, $\mathrm{E}(V) = \mu_2$, $\mathrm{Var}(U) = \mathrm{Var}(V) = m\sigma^2/(m-2)$ $(m > 2)$, $\mathrm{Corr}(U, V) = \rho$. このとき結合確率密度関数 $h(u, v)$ は,

$$h(u,v) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}}\left(1 + \frac{Q(u,v)}{m}\right)^{-\frac{m+2}{2}},$$

ただし,

$$Q(u,v) = \frac{1}{\sigma^2(1-\rho^2)}\Big[(u-\mu_1)^2 - 2\rho(u-\mu_1)(v-\mu_2) + (v-\mu_2)^2\Big].$$

また $h(u, v)$ は次の関係をみたす:

$$\Big(h(u,v)\Big)^{-\frac{2}{m+2}} - \Big(h(v,u)\Big)^{-\frac{2}{m+2}} = \alpha_m(v-u) \quad (u < v),$$

ただし,

$$\alpha_m = \frac{2(\mu_1-\mu_2)}{m\sigma^2(1-\rho)}\Big(2\pi\sigma^2\sqrt{1-\rho^2}\Big)^{\frac{2}{m+2}}.$$

$R\times R$順序カテゴリ正方分割表に対して次のモデルを提案した. 固定した m $(m > 2)$ に対して,

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \beta_m(j-i) \quad (i < j),$$

ただし β_m は未知パラメータとする. このモデルはデータが連続量で周辺分散が等しい潜在的な自由度 m の t 分布に従うと想定されるとき, よく適合すると考えられる. このモデルを t 分布型対称モデルと呼ぶことにする（TS(m) によって記す）. 特に $\beta_m = 0$ とおいたこのモデルは対称モデルである. このモデルは p_{ij} の $-2/(m+2)$ 乗と p_{ji} の $-2/(m+2)$ 乗の差が主対角線からの距離 $j-i$ に比例することを示している.

参考文献

Agresti, A. (1983). A simple diagonals-parameter symmetry and quasi-symmetry model. *Statistics and Probability Letters*, **1**, 313-316.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572-574.

Yamamoto, H., Iwashita, T. and Tomizawa, S. (2007). Decomposition of symmetry into ordinal quasi-symmetry and marginal equimoment for multi-way tables. Austrian Journal of Statistics 36, 291-306.

順序カテゴリ正方分割表における対称性のモデルと分解および併合した表に基づく対称性に関する尺度

東京理科大学大学院・理工学研究科　島田 文香
東京理科大学大学院・理工学研究科　田中 弥生
大阪大学・医学部　山本 紘司
東京理科大学・理工学部　富澤 貞男

第 1 部　Sum-symmetry モデルの提案とその分解

第一部では, モデルを 2 つ提案した。まず, 新たな対称構造をみるモデルとして次のモデルを提案した:

$$\sum_{(i,j)\in R(x)} p_{ij} \quad = \quad \sum_{(i,j)\in R(x)} p_{ji} \quad (3 \leq x \leq 2r-1),$$

ここに,

$$R(x) = \{(i,j) \mid i+j = x, i < j\},$$

である. ここではこのモデルを Sum-symmetry(SS) モデルと呼ぶ. 次に下記のモデルを導入した:

$$\sum_{(i,j)\in R(x)} p_{ij} \quad = \quad \Delta \sum_{(i,j)\in R(x)} p_{ji} \quad (3 \leq x \leq 2r-1).$$

このモデルを条件付き Sum-symmetry(CSS) モデルと呼ぶ. また, $\Delta = 1$ とおいた CSS モデルは SS モデルである. このとき次の分解定理を得る:

定理 1: SS モデルが成り立つための必要十分条件は, CSS モデルと GS モデルの両方が成り立つことである.

$r \times r$ 正方分割表において, (i,j) セル観測度数を n_{ij} とし, m_{ij} を対応するセル期待度数とする. $\{n_{ij}\}$ は多項分布に従うと仮定する. このとき, モデル M の適合度を検定するための尤度比カイ二乗統計量は

$$G^2(M) = 2\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

で与えられる. $\hat{m}_{ij}$ はモデル M の下での m_{ij} の最尤推定値である. このとき, 次の定理を得る:

定理 2.　検定統計量 $G^2(SS)$ は, 検定統計量 $G^2(CSS)$ と $G^2(GS)$ の和に同等である.

第 2 部　正方分割表における併合した表を用いた対称性に関する尺度

1. はじめに

行および列に対するカットポイント s および t $(1 \leqq s < t \leqq r-1)$ を用いて, カテゴリを併合して $r \times r$ 分割表を 3×3 分割表にする. 全部で $\binom{r-1}{2}$ $(= (r-1)(r-2)/2)$ 枚の併合した分割表を作成することができ, それぞれの併合した分割表を T_{st} とする. このとき, T_{st} 表における (i,j) セル確率を $G_{ij}^{(s,t)}$ $(i = 1,2,3; j = 1,2,3)$ とする. 対称モデルが成り立たないとき, その隔たりの程度を測ることに関心があり, 第 2 部では併合した分割表を用いて対称性からの隔たりを測る尺度を提案した.

2. 提案する尺度

$\{p_{ij} + p_{ji} \neq 0\}$ を仮定する. ここに $\delta_{st} = \sum\limits_{i \neq j}\sum G_{ij}^{(s,t)}$ $(1 \leqq s < t \leqq r-1)$ として,

$$G_{ij}^{*(s,t)} = \frac{G_{ij}^{(s,t)}}{\delta_{st}}, \quad G_{ij}^{c(s,t)} = \frac{G_{ij}^{(s,t)}}{G_{ij}^{(s,t)} + G_{ji}^{(s,t)}} \quad (i = 1,2,3; j = 1,2,3),$$

とする. このとき, 対称モデルからの隔たりを測る尺度を次のように提案した:

$$\Psi^{(\lambda)} = \frac{1}{\binom{r-1}{2}} \sum_{1 \leqq s < t \leqq r-1} \sum \Psi_{st}^{(\lambda)} \quad (\lambda > -1),$$

ただし

$$\Psi_{st}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1} \sum_{i<j} \sum \left(G_{ij}^{*(s,t)} + G_{ji}^{*(s,t)} \right) I_{ij}^{(\lambda)} \left(\left\{ G_{ij}^{c(s,t)}, G_{ji}^{c(s,t)} \right\}; \left\{ \frac{1}{2}, \frac{1}{2} \right\} \right),$$

ここに

$$I_{ij}^{(\lambda)}(\cdot;\cdot) = \frac{1}{\lambda(\lambda+1)} \left[G_{ij}^{c(s,t)} \left\{ \left(\frac{G_{ij}^{c(s,t)}}{1/2} \right)^{\lambda} - 1 \right\} + G_{ji}^{c(s,t)} \left\{ \left(\frac{G_{ji}^{c(s,t)}}{1/2} \right)^{\lambda} - 1 \right\} \right].$$

$\lambda = 0$ のときは $\lambda \to 0$ の極限で定義した.

部分尺度 $\Psi_{st}^{(\lambda)}$ $(1 \leqq s < t \leqq r-1)$ は T_{st} 表における対称性からの隔たりを測る尺度である. ここに $I_{ij}^{(\lambda)}(\cdot;\cdot)$ は, $\{G_{ij}^{c(s,t)}, G_{ji}^{c(s,t)}\}$ と $\{1/2, 1/2\}$ の間の power-divergence であり, 特に $\lambda = 0$ のときは Kullback-Leibler 情報量である. 提案尺度は以下の性質をもつ.

- $0 \leq \Psi^{(\lambda)} \leq 1$
- $\Psi^{(\lambda)} = 0 \Leftrightarrow$ 対称モデルが成り立つ
- $\Psi^{(\lambda)} = 1 \Leftrightarrow$ 対称モデルからの隔たりが最大である

正方分割表における累積確率を用いた非対称性のモデルの分解

東京理科大学大学院 理工学研究科　安藤宗司
大阪大学　医学部　山本紘司
東京理科大学 理工学部　富澤貞男

1. はじめに

行変数と列変数のカテゴリに順序のある同じ分類からなる正方分割表データの解析においては，分類間の相互関連性が強いため，独立性を考えることに意味がないことが多い．そのため分類間の独立性に代わって対称性に関するモデルが用いられる．対称性に関するモデルとして，対称モデル（Bowker, 1948），周辺同等モデル（Stuart, 1955），条件付対称モデル（McCullagh, 1978），対角パラメータ対称モデル（Goodman, 1979），線形対角パラメータ対称モデル（Agresti, 1983），2比パラメータ対称モデル（Tomizawa, 1987），拡張周辺同等モデル（Tomizawa, 1993）等がある．また，累積確率に基づくモデルとして，累積対角パラメータ対称モデル（Tomizawa, 1993），累積線形対角パラメータ対称モデル（Miyamoto, Ohtsuka and Tomizawa, 2004），累積2比パラメータ対称モデル（Tomizawa, Miyamoto, Yamamoto and Sugiyama, 2007），累積準対称モデル（Miyamoto et al., 2004），累積拡張準対称モデル（Tomizawa et al., 2007）等がある．本報告では，累積確率に基づく新たなモデルを提案し，有用な結果と解釈が得られることを適用例と共に示した．また，累積線形対角パラメータ対称モデル，累積2比パラメータ対称モデルの当てはまりの悪い場合に，より詳細に解析するために有用であるモデルの分解定理を与えた．

2. モデル提案

正方 $R \times R$ 分割表において，(i, j) セル確率を p_{ij} とする（$i = 1, \ldots, R; j = 1, \ldots, R$）．累積確率 $\{G_{ij}\}$，$i \neq j$，を次のように定義する．

$$G_{ij} = \sum_{s=1}^{i} \sum_{t=j}^{R} p_{st} \quad (i < j), \quad G_{ij} = \sum_{s=i}^{R} \sum_{t=1}^{j} p_{st} \quad (i > j).$$

累積2パラメータ周辺対称-1モデルを次のように提案した：

$$G_i^+ = \Delta_1 \sum_{k=i+1}^{R} \Omega^{k-i} G_{ki} \quad (i = 1, \ldots, R-1),$$

ただし，

$$G_i^+ = \sum_{k=i+1}^{R} G_{ik},$$

である．また，累積2パラメータ周辺対称-2モデルを次のように導入した：

$$G_i^- = \Delta_2 \sum_{k=1}^{i-1} \Lambda^{i-k} G_{ki} \quad (i = 2, \ldots, R),$$

ただし，

$$G_i^- = \sum_{k=1}^{i-1} G_{ik},$$

である．特に $\Delta_1 = 1$（$\Delta_2 = 1$）のとき，累積1パラメータ周辺対称-1モデル（累積1パラメータ周辺対称-2モデル）である．

3. 非対称性モデルの分解

次のように累積2比パラメータ対称モデルの分解を与えた．

定理1　$t = 1, 2$ に対して，累積2比パラメータ対称モデルが成り立つための必要十分条件は，累積拡張準対称モデル，累積2パラメータ周辺対称-t モデルとD-t モデルのすべてが成り立つことである．

さらに，定理1より累積線形対角パラメータ対称モデルの分解も与えた．
また，累積2比パラメータ対称モデルの別の分解として次の定理を得た．

定理2　累積2比パラメータ対称モデルが成り立つための必要十分条件は，累積拡張準対称モデルと拡張周辺同等モデルが成り立つことである．

さらに，定理2より累積線形対角パラメータ対称モデルの分解と対称モデルの分解も与えた．

正方分割表における
拡張パリンドロミック対称モデルと対称性の分解

三枝　祐輔　(東京理科大学大学院・理工学研究科)
田畑　耕治　(東京理科大学・理工学部)
富澤　貞男　(東京理科大学・理工学部)

1. 提案するモデル

本講演では，McCullagh [3] によって導入されたパリンドロミック対称 (PS) モデルの拡張を与えた．行変数 X_1 と列変数 X_2 が同じ分類からなる $r \times r$ 正方分割表を考える．(i,j) セルの観測値の出現確率を p_{ij} $(i=1,\ldots,r; j=1,\ldots,r)$ とおく．このとき，対称 (S) モデルは次のように定義される (Bowker [2]; Bishop, Fienberg and Holland [1])：

$$p_{ij} = \psi_{ij} \quad (i=1,\ldots,r; j=1,\ldots,r), \quad \text{ただし } \psi_{ij} = \psi_{ji}.$$

累積確率を次のように導入する：

$$G_{ij} = \sum_{s=1}^{i}\sum_{t=j}^{r} p_{st} = \mathrm{P}(X_1 \le i, X_2 \ge j) \quad (i<j),$$

$$G_{ji} = \sum_{s=j}^{r}\sum_{t=1}^{i} p_{st} = \mathrm{P}(X_1 \ge j, X_2 \le i) \quad (i<j).$$

このとき，Sモデルは次のようにも表すことができる：

$$G_{ij} = \Psi_{ij} \quad (i \ne j), \quad p_{ii} = \Psi_{ii}, \quad \text{ただし } \Psi_{ij} = \Psi_{ji}.$$

m 比一般化パリンドロミック対称 (PS(m)) モデルを次のように提案した (Saigusa, Tahata and Tomizawa, [4])：固定した m $(m=1,\ldots,r-1)$ に対して，

$$G_{ij} = \begin{cases} \Delta_i^{(m)} \dfrac{\alpha_i}{\alpha_{j-1}} \Psi_{ij} & (i<j), \\ \Psi_{ij} & (i>j), \end{cases} \qquad p_{ii} = \Psi_{ii}, \quad \text{ただし } \Psi_{ij} = \Psi_{ji},$$

ここに，

$$\Delta_i^{(m)} = \prod_{k=0}^{m-1} \Delta_k^{i^k}.$$

PS(1) モデル，PS($r-1$) モデルはそれぞれ PS モデル，一般化パリンドロミック対称 (GPS) モデル (McCullagh [3]) である．

2. 対称モデルの分解

さらに，本講演では，PS(m) モデルを用いたSモデルの分解を与えた．

行変数と列変数の原点まわりの k 次モーメント一致 (MO(k)) 構造は次のように表される：固定した k (≥ 1) に対して，

$$E(X_1^k) = E(X_2^k).$$

特に $k=1$ のとき，ME (平均一致) と記す.

累積部分対称 (CSS) モデルは次のように定義される (Tomizawa, Miyamoto and Ouchi [5])：

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

平均値まわりの l 次モーメント一致 (MA(l)) 構造は次のように表される：固定した l (≥ 2) に対して，

$$\mu_l^{X_1} = \mu_l^{X_2}, \quad \text{ただし } \mu_l^{X_t} = E((X_t - E(X_t))^l) \quad (t = 1, 2).$$

特に $l=2$ のとき，VE (分散一致) と記す．また，歪度一致 (SE) 構造は次のようになる：

$$\frac{\mu_3^{X_1}}{(\mu_2^{X_1})^{3/2}} = \frac{\mu_3^{X_2}}{(\mu_2^{X_2})^{3/2}}.$$

尖度一致 (KE) 構造は次のようになる：

$$\frac{\mu_4^{X_1}}{(\mu_2^{X_1})^{2}} = \frac{\mu_4^{X_2}}{(\mu_2^{X_2})^{2}}.$$

このとき，次の定理および系を得た.

定理 1. 固定した m ($m = 1, \ldots, r-1$) に対して, S モデルが成り立つための必要十分条件は，PS(m) モデル，MO(k) モデル ($k = 1, \ldots, m$) および CSS モデルのすべてが成り立つことである.

系 1. (i)，(ii)，(iii) は同値である.

(i) S モデルが成り立つ.

(ii) 固定した m ($m = 2, \ldots, r-1$) に対して，PS(m) モデル，ME モデル，MA(l) モデル ($l = 2, \ldots, m$) および CSS モデルのすべてが成り立つ.

(iii) $r \geq 5$ のとき，PS(4) モデル，ME モデル，VE モデル，SE モデル，KE モデルおよび CSS モデルのすべてが成り立つ.

参考文献

[1] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge.

[2] Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association* **43**, 572-574.

[3] McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* **65**, 413-418.

[4] Saigusa, Y., Tahata, K. and Tomizawa, S. (2014). An extended asymmetry model for square contingency tables with ordered categories. *Model Assisted Statistics and Applications*, to appear.

[5] Tomizawa, S., Miyamoto, N. and Ouchi, M. (2006). Decompositions of symmetry model into marginal homogeneity and distance subsymmetry in square contingency tables with ordered categories. *Revstat: Statistical Journal* **4**, 153-161.