平成２２年度科学研究費基盤研究（A）
研究課題番号　：　1920400901
研究課題名　　：　統計科学における数理的手法の理論と応用
研究代表者　　：　谷口　正信（早稲田大学）

統計的推測方法の理論的展開とその応用
プログラム

2010 年 11 月 17 日～19 日

熊本大学大学院自然科学研究科研究棟

研究分担者　：　高田 佳和　（熊本大学）
岩佐 学　（熊本大学）

１１月 17 日（水）

１４：００～１４：０５　　開会

１４：０５～１４：４０　今田　恒久　（東海大・総合経営学部）
尤度比検定法に基づく或る種の多変量両側検定法について

１４：４０～１５：１５　高木　祥司　（奈良教育大・教育学部）
ガンマ分布の形状パラメータの推定問題について
ー漸近２次最適性の立場からー

１５：１５～１５：５０　今野　良彦　（日本女子大・理学部）
Shrinkage estimation of the mean matrix of multivariate complex normal distribution

１５：５０～１６：００　　休憩

１６：００～１６：３５　　田中　秀和　（大阪府立大・工学研究科）
高木　祥司　(奈良教育大学・教育学部)
Park, Junyong　（University of Maryland）

On conditions for a modified maximum likelihood estimator to be second order admissible

１６：３５～１７：１０　　高橋　倫也　（神戸大・海事科学研究科）

最大値の推定

１７：１０～１７：４５　　吉原　健一　（横浜国立大・工学部）
金川　秀也　（東京都市大学・知識工学部）

Parameter estimated standardized U-statistics for some dependent sequences and its application to change-point problems

１１月１８日（木）

９：３０～１０：０５　　西山　陽一　（統数研）

無限次元マルチンゲール中心極限定理の使用法

１０：０５～１０：４０　　白石　博　（東京慈恵会医科大学）
谷合　弘行　（早稲田大学）

Goodness of Fit for Randomly Censored Data

１０：４０～１１：１５　　Alexandre Petkovic　（早稲田大学・理工学術院）

Linear Regression with Deterministic Regressors and Unit Root in the Variance

１１：１５～１１：２５　　休憩

１１：２５～１２：００　　布能 英一郎　（関東学院大・経済学部）

Pooling incomplete samples の状況下における統計的推論

１２：００～１２：３５　　種市　信裕　（鹿児島大・理工学研究科）
関谷　祐里　（北海道教育大・教育学部）

多項分布の適合度検定統計量の漸近展開における離散項J_2の評価についての考察

１２：３５～１３：３５　　休憩

１３：３５～１４：１０　　吉田　知行　（北大・理学研究院）

金属考古学における統計的問題

１４：１０～１４：４５　　磯貝　英一　(新潟大・自然科学研究科)
小林　加奈
宇野　力　(秋田大・教育文化学部)

Higher order approximations by a two-stage procedure for a negative exponential distribution

１４：４５～１５：２０　　柿沢　佳秀　（北大・経済学部）

A generalized Bernstein polynomial approach to density estimation

１５：２０～１５：３０　　休憩

１５：３０～１６：０５　　百武　弘登　（九大・数理学研究院）

On comparisons of univariate normal mean and elements of multivariate normal mean

１６：０５～１６：４０　　高橋 翔　（東京理科大・理学研究科）
西山 貴弘　（東京理科大・理学部）
瀬尾 隆　（東京理科大・理学部）
今田 恒久　（東海大・総合経営学部）

ステップワイズ法による確率ベクトルの成分間の独立性の同時検定

１６：４０～１７：１５　　西山 貴弘　（東京理科大・理学部）
瀬尾 隆　（東京理科大・理学部）

平均ベクトル間の多重比較法に対する同時信頼区間とその保守性

１１月１９日（金）

９：３０～１０：０５　　小林　裕子　（筑波大・数理物質科学研究科）
矢田　和善　（筑波大・数理物質科学研究科）
青嶋　誠　（筑波大・数理物質科学研究科）

ß-Lasso 推定による頑健なモデル選択

１０：０５～１０：４０　　藤本 翔太　（大阪大学・基礎工学研究科）
狩野　裕　（大阪大学・基礎工学研究科）
Muni　S. Srivastava　（University of Toronto）

高次元データにおける幾つかの検定統計量の漸近分布について

１０：４０～１１：１５　　笹渕　祥一　（九大・芸術工学研究院）

共分散行列が未知の場合の任意の半順序対立仮説に対する平均ベクトルの均一性の検定

１１：１５～１１：２５　　休憩

１１：２５～１２：００　　矢田 和善　（筑波大・数理物質科学研究科）
青嶋　誠　（筑波大・数理物質科学研究科）

Sample Size Determination for High-Dimension, Low-Sample-Size Data

１２：００～１２：３５　　小池　健一　（筑波大・数理物質科学研究科）

台が有界な分布における台の端点の逐次推測

１２：３５～１２：４０　　閉会

尤度比検定法に基づく或る種の多変量両側検定法について

東海大学 総合経営学部 今田 恒久

1. はじめに

正規母平均に対する片側検定，両側検定の多変量正規分布への拡張として多変量片側検定，多変量両側検定の研究がある．その仮説の設定方法は一様でなく，様々な研究結果が発表されている．成分がいずれも 0 以上と仮定された多変量正規母平均の零ベクトルとの差を調べるための多変量片側検定では母集団の分散共分散行列を既知と仮定した下で Kudo (1963) が尤度比検定方式を考察し，尤度比検定統計量の確率分布を導出した上で指定した有意水準を満たす棄却限界値の決定方法を確立している．また，分散共分散行列が未知の場合は Perlman (1969) が議論しているが，未知の分散共分散行列に関係なく尤度比検定統計量の確率分布を決定することが困難であるため，分散共分散行列を動かしたときの尤度比検定統計量の確率分布の上限を導出し，指定した有意水準に対して保守的な棄却限界値を決定した．これに対して，Wang-McDermott (1998) は未知の分散共分散行列に対する十分統計量を与えた下での条件付分布を用いて指定した有意水準を満たす棄却限界値を決定した．

一方，成分はすべて 0 以上，あるいはすべて 0 以下と仮定された多変量正規母平均の零ベクトルとの差を調べるための多変量両側検定では分散共分散行列を既知と仮定した下で 2 変量正規分布の場合に Kudo-Fujisawa (1964) が尤度比検定方式を考察し，指定した有意水準を満たす棄却限界値を決定している．また，Yeh (1968) は分散共分散行列が単位行列の場合に Kudo-Fujisawa (1964) の検定方式を一般次元まで拡張した．しかし，分散共分散行列が対角行列でない場合の尤度比検定方式は変数の数が 3 以上の場合，理論と計算が複雑となり，結果は得られていない．

以上述べたように尤度比検定法に基づく多変量両側検定方式では分散共分散行列を既知と仮定した場合は一般の多変量正規分布に対して変数の数が 3 以上の場合未解決であり，分散共分散行列が未知の場合の結果は得られていない．ここでは多変量正規分布に対する多変量両側検定方式に対する未解決問題に対して分散共分散行列を既知の場合，未知の場合に分けて考える．指定した有意水準に対して保守的な棄却限界値の決定，指定した有意水準を満たす棄却限界値の近似値の求め方について議論する．

2. 多変量両側検定

p 次元確率ベクトル $\boldsymbol{X}$ は p 変量正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従うと仮定する．母平均 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ の成分について次の 2 通りのいずれかが成立すると仮定する．

$$\mu_1 \geq 0,\ \mu_2 \geq 0,\ \ldots,\ \mu_p \geq 0 \tag{2.1}$$

$$\mu_1 \leq 0,\ \mu_2 \leq 0,\ \ldots,\ \mu_p \leq 0 \tag{2.2}$$

(2.1) において少なくとも一つ $\mu_i > 0$ が成立するときは，$\boldsymbol{\mu} \geq^* \boldsymbol{0}$ と表し，(2.2) において少なくとも一つ $\mu_i < 0$ が成立するときは，$\boldsymbol{\mu} \leq^* \boldsymbol{0}$ と表す．いま，帰無仮説 H_0 とその対立仮説 H_1 を

$$H_0 : \boldsymbol{\mu} = \boldsymbol{0} \text{ vs. } H_1 : \boldsymbol{\mu} \geq^* \boldsymbol{0} \text{ または} \boldsymbol{\mu} \leq^* \boldsymbol{0} \tag{2.3}$$

と設定し，$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ からのサイズ n の標本 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ に基づき，この検定方式を考える．

3. $\boldsymbol{\Sigma}$ が既知の場合

まず，分散共分散行列 $\boldsymbol{\Sigma}$ は既知と仮定し，(2.3) に対する尤度比検定法を考える．

$$\boldsymbol{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{X}_i$$

とおく．いま，p 次元ユークリッド空間 $\boldsymbol{R}^p$ において領域 $\boldsymbol{\Theta}$ を

$$\boldsymbol{\Theta} = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_p)' \in \boldsymbol{R}^p | x_1 \geq 0, x_2 \geq 0, \ldots, x_p \geq 0 \text{ または } x_1 \leq 0, x_2 \leq 0, \ldots, x_p \leq 0\}$$

とおく．$\boldsymbol{Z}$ を $\boldsymbol{R}^p$ 内の点と考え，$\boldsymbol{\Sigma}$ による Mahalanobis 距離 $\|\cdot\|_{\boldsymbol{\Sigma}}$ に関する $\boldsymbol{Z}$ の領域 $\boldsymbol{\Theta}$ 上への直交射影を $\pi_{\boldsymbol{\Sigma}}(\boldsymbol{Z},\boldsymbol{\Theta})$ と表す．仮説 (2.3) に対する尤度比検定法では棄却限界値 c を指定し，

$$\bar{\bar{\chi}}^2 = \| \pi_{\boldsymbol{\Sigma}}(\boldsymbol{Z},\boldsymbol{\Theta}) \|^2_{\boldsymbol{\Sigma}} > c$$

であるとき，H_0 を棄却する．棄却限界値 c は指定した有意水準 α に対し，H_0 の下で

$$P(\bar{\bar{\chi}}^2 > c) = \alpha \tag{3.1}$$

を満たすように決定したい．$\bar{\bar{\chi}}^2$ の確率分布は $p \geq 3$ の場合，決定が困難である．保守的な棄却限界値の決定方法も考えられるが，他の方法としては H_0 の棄却域 $\boldsymbol{D}(c) = \{\boldsymbol{Z} \in \boldsymbol{R}^p | \bar{\bar{\chi}}^2 = \| \pi_{\boldsymbol{\Sigma}}(\boldsymbol{Z},\boldsymbol{\Theta}) \|^2_{\boldsymbol{\Sigma}} > c\}$ を構成し，$f_0(\boldsymbol{z})$ を $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$ の確率密度関数とすると，H_0 の下で

$$P(\bar{\bar{\chi}}^2 > c) = \int \cdots \int_{\boldsymbol{D}(c)} f_0(\boldsymbol{z}) d\boldsymbol{z}. \tag{3.2}$$

しかし，$p \geq 3$ の場合は $\boldsymbol{D}(c)$ の形状は複雑となり，実際に (3.2) を計算することは困難である．他方，(3.2) の近似値を得る方法としてグリッドを用いる方法がある．

4. $\boldsymbol{\Sigma}$ が未知の場合

次に分散共分散行列 $\boldsymbol{\Sigma}$ は未知と仮定し，(2.3) に対する尤度比検定法を考える．$\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i' - \boldsymbol{Z}\boldsymbol{Z}'$ とおく．仮説 (2.3) に対する尤度比検定統計量は $\boldsymbol{S}$ による Mahalanobis 距離 $\|\cdot\|_{\boldsymbol{S}}$ を用いると，

$$\bar{\bar{F}} = \frac{\| \pi_{\boldsymbol{S}}(\boldsymbol{Z},\boldsymbol{\Theta}) \|^2_{\boldsymbol{S}}}{1 + \| \boldsymbol{Z} - \pi_{\boldsymbol{S}}(\boldsymbol{Z},\boldsymbol{\Theta}) \|^2_{\boldsymbol{S}}}$$

により与えられ，指定した棄却限界値 c に対し，$\bar{\bar{F}} > c$ であるとき，H_0 を棄却する．棄却限界値 c は指定した有意水準 α に対し，H_0 の下で

$$P(\bar{\bar{F}} > c) = \alpha \tag{4.1}$$

を満たすように決定したい．$\bar{\bar{\chi}}^2$ の確率分布は決定が困難である上，(4.1) は未知である $\boldsymbol{\Sigma}$ に依存する．保守的な棄却限界値の決定方法も考えられるが，他の方法としては $\boldsymbol{\Sigma}$ に対する十分統計量 $\boldsymbol{V} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'$ の値を与えた下での条件付確率 $P(\bar{\bar{F}} > c|\boldsymbol{V})$ を考え，これが α に等しくなるように c を決定したい．$\boldsymbol{V}$ の値を与えた下での $\boldsymbol{Z}$ の条件付確率密度関数を $f_0(\boldsymbol{z}|\boldsymbol{V})$ とすると，

$$P(\bar{\bar{F}} > c|\boldsymbol{V}) = \int \cdots \int_{\boldsymbol{D}(c)} f_0(\boldsymbol{z}|\boldsymbol{V}) d\boldsymbol{z}. \tag{4.2}$$

しかし，一般に $\boldsymbol{D}(c)$ の形状は複雑であり，(4.2) を計算することは困難であるが，前節と同様にグリッドを用いて近似値を求めることは可能である．

5. 数値例による考察

多変量正規分布に対する多変量両側検定方式に対し，分散共分散行列を既知の場合，未知の場合に分けて指定した有意水準に対応する棄却限界値について議論してきた．いずれの場合も指定した有意水準を正確に満たす棄却限界値の決定は困難であるため近似法による決定方法を考えた．発表では数値実験の結果に基づき，近似法により決定した棄却限界値の保守性，精度について議論した．

参考文献

[1] Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403-418.

[2] Kudo, A. and Fujisawa, H. (1964). A bivariate normal test with two sided alternative. *Memoirs of the Faculty of Science, Kyushu University, Ser.A*, **18**, No.1, 104-108.

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

[4] Yeh, N. (1968). A multivariate normal test with two-sided alternative. *Bulletin of mathematical statistics* **13**, 85–88.

[5] Wang, Y. and McDermott, M. P. (1998). Conditional likelihood ratio test for a nonnegative normal mean vector, *Journal of the American Statistical Association* **93**, 380–386.

ガンマ分布の形状パラメータの推定問題についてー漸近２次最適性の立場からー

高木 祥司（奈良教育大学　教育学部）

1．はじめに

２母数ガンマ分布の確率密度関数は，

$$f(x;\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)\quad \text{for}\quad x>0 \tag{1}$$

と表現され，α (>0) は形状パラメータと呼ばれ，β (>0) は尺度パラメータと呼ばれる．

ここでは，ガンマ分布の形状パラメータの推定問題を考える．一般には最尤推定量（MLE）がよく用いられるが，MLE が２乗損失もとで許容的であるか，MLE のどのようなバイアス修正推定量が許容的になるかといった基本的な問題についての研究はさほど多くない．最近では，２乗損失のもとで MLE よりもよい IML 推定量が提案されている．

この研究では，漸近２次最適性理論の立場から，最適な推定量を見つけ出すことを考える．まず，パラメータ直交性のもとでの漸近２次許容性についての研究結果 (Takagi(2003)) を基盤とし，損失関数の自由な選択のために損失係数の概念 (Takagi (2006)) を取り入れることで，損失係数 k をもつ損失関数のもとで，ある推定量が漸近２次許容的および漸近２次非許容的になるための十分条件をそれぞれ与える．次に，簡単なパラメータ変換によって形 (1) をパラメータ直交性をもつガンマ分布に書き換えることで前述の十分条件を適用し，最尤推定量や IML 推定量等の基本的な推定量が，漸近２次許容的になるか漸近２次非許容的になるかを調べる．

2．パラメータ直交性のもとでの漸近２次許容性

興味ある母数 θ_1 と局外母数 θ_2 による２母数確率モデルを考える．ここで，パラメータベクトルを $\theta=(\theta_1,\theta_2)$ で表す．その確率分布はパラメータ直交性をもつとする (Cox and Reid (1987))．すなわち，フィッシャー情報量行列が次の形で表現される：$I(\theta)=\begin{pmatrix} I_{11}(\theta) & 0 \\ 0 & I_{22}(\theta)\end{pmatrix}$.

任意の損失関数を $\ell(z)$ とし，その損失係数を k とする（Takagi(2006)）．推定量の良さに関する比較基準として，興味あるパラメータ θ_1 に対する任意の推定量 $T_n=T_n(X_1,X_2,\cdots,X_n)$ のリスク関数を以下のように定義する：$R_\ell(\theta_1,\theta_2;T_n)=E\left[\ell\left(\sqrt{n}I_{11}(\theta)^{1/2}(T_n-\theta_1)\right)\right]$

考える推定量族 $\mathcal{C}$ は次の通りである：　$\mathcal{C}=\left\{\tilde{\theta}_c \mid \tilde{\theta}_c=\hat{\theta}_1+c(\hat{\theta}_1,\hat{\theta}_2)/n\right\}$．ここで，$(\hat{\theta}_1,\hat{\theta}_2)$ は最尤推定量で，$c(\cdot,\cdot)$ は任意の連続で微分可能な関数である．

いま，独立同分布 $X_1,X_2,\cdots,X_n$ での対数尤度関数を $l_n(\theta)$ で表し，$i=1,2$ に対して，

$$Z_i=n^{-1/2}\partial_i l_n(\theta),\quad Z_{ij}=n^{-1/2}\{\partial_i\partial_j l_n(\theta)-E[\partial_i\partial_j l_n(\theta)]\}$$

と定義する（$\partial/\partial\theta_i$ を ∂_i と略記）．Z_i と Z_{ij} の漸近モーメントは次の展開を仮定する．

$$E(Z_iZ_{jk})=J_{ijk}+O(n^{-1}),\quad E(Z_iZ_jZ_k)=\frac{1}{\sqrt{n}}K_{ijk}+O(n^{-3/2}).$$

推定量族 $\mathcal{C}$ に属する推定量 $\tilde{\theta}_c=\hat{\theta}_1+c(\hat{\theta}_1,\hat{\theta}_2)/n$ が漸近２次許容性をもつかどうかの判定について考える．このとき，関数 $w_{\gamma,k}(\theta_1,\theta_2)$ を次のように定義する：

$$w_{\gamma,k}(\theta_1,\theta_2)=\sqrt{I_{11}}c+\kappa_{11}+\frac{(k-3)\kappa_{31}}{6}+(\gamma-1)\frac{\partial I_{11}}{I_{11}\sqrt{I_{11}}}$$

ここで，κ_{11} と κ_{31} は最尤推定量 $\hat{\theta}_1$ のキュムラントの漸近展開項で次のように表せる：

$$\kappa_{11}(\theta)=-\frac{1}{2I_{11}\sqrt{I_{11}}}(K_{111}+J_{111})+\frac{1}{2\sqrt{I_{11}}I_{22}}J_{212},\quad \kappa_{31}(\theta)=\frac{1}{I_{11}\sqrt{I_{11}}}(-2K_{111}-3J_{111})$$

定理１： (1) 推定量 $\tilde{\theta}_c$ が損失係数 k のもとで漸近２次許容的であるための十分条件は，次の条件 [1-1] と [1-2] がともに成り立つことである．

[1-1]　すべての固定された θ_2 に対して，$\displaystyle\int_{\theta_0}^{\infty} I_{11}(\theta_1,\theta_2)^{\gamma_1}d\theta_1=\infty$ となる γ_1 が存在し，関数 $w_{\gamma_1,k}(\theta_1,\theta_2)$ が十分に大きなすべての θ_1 に対して非正である．

[1-2]　すべての固定された θ_2 に対して，$\displaystyle\int_{-\infty}^{\theta_0} I_{11}(\theta_1,\theta_2)^{\gamma_2}d\theta_1=\infty$ となる γ_2 が存在し，関数 $w_{\gamma_2,k}(\theta_1,\theta_2)$ が十分に小さなすべての θ_1 に対して非負である．

(2) 推定量 $\tilde{\theta}_c$ が損失係数 k のもとで漸近２次非許容的であるための十分条件は，次の条件 [2-1] と [2-2] の少なくとも１つが成立することである．

[2-1]　ある固定された θ_2 に対して，$\int_{\theta_0}^{\infty} I_{11}(\theta_1,\theta_2)^{\gamma_1} d\theta_1 < \infty$ となる γ_1 が存在し，関数 $w_{\gamma_1,k}(\theta_1,\theta_2)$ が十分に大きなすべての θ_1 に対して非負である．

[2-2]　ある固定された θ_2 に対して，$\int_{-\infty}^{\theta_0} I_{11}(\theta_1,\theta_2)^{\gamma_2} d\theta_1 < \infty$ となる γ_2 が存在し，関数 $w_{\gamma_2,k}(\theta_1,\theta_2)$ が十分に小さなすべての θ_1 に対して非正である．

注意:　θ_1 の範囲に応じて，積分中の ∞ を上限値に $-\infty$ を下限値に置き換えてもよい．

3．ガンマ分布の形状パラメータの推定

ガンマ分布の形状パラメータ α に対して，与えられた損失係数 k のもとで，いくつかの基本的な推定量が漸近２次許容性を持つかどうかを調べる．ただし，表現 (1) でのガンマ分布はパラメータ直交性をもたないので，パラメータ β を $\eta = \alpha\beta$ で変換すると，密度関数は

$$f(x;\alpha,\eta) = \frac{1}{\Gamma(\alpha)} \frac{1}{(\eta/\alpha)^{\alpha}} x^{\alpha-1} e^{-\alpha x/\eta} \tag{2}$$

となり，フィッシャー情報量行列は以下のようになりパラメータ直交性をもつ：

$$I(\alpha,\eta) = \begin{pmatrix} I_{11}(\alpha,\eta) & I_{12}(\alpha,\eta) \\ I_{21}(\alpha,\eta) & I_{22}(\alpha,\eta) \end{pmatrix} = \begin{pmatrix} \psi'(\alpha) - 1/\alpha & 0 \\ 0 & \alpha/\eta^2 \end{pmatrix}$$

ここで，$\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ はオイラーのディガンマ関数で，$\psi'(\alpha)$ は α による導関数である．これより，漸近２次許容性に関する２節での定理を適用できることになった．

補題１：　ガンマ分布 (2) のもとで，次の計算を得る．

(1)　$J_{111} = 0, \qquad K_{111} = \partial_1 I_{11} = \psi''(\alpha) + \dfrac{1}{\alpha^2}, \qquad J_{212} = \dfrac{1}{\eta^2}.$

(2) $\kappa_{11} = -\dfrac{\psi''(\alpha) + 1/\alpha^2}{2(\psi'(\alpha) - 1/\alpha)^{3/2}} + \dfrac{1}{2(\psi'(\alpha) - 1/\alpha)^{1/2}\alpha}, \quad \kappa_{31} = -\dfrac{2(\psi''(\alpha) + 1/\alpha^2)}{(\psi'(\alpha) - 1/\alpha)^{3/2}}.$

補題２：　フィッシャー情報量の積分に関して次が成立する：

$$\int_{\alpha_0}^{\infty} I_{11}(\alpha,\eta)^{\gamma_1} d\alpha \begin{cases} = \infty, & \gamma_1 \le 1/2 \\ < \infty, & \gamma_1 > 1/2, \end{cases} \qquad \int_{0}^{\alpha_0} I_{11}(\alpha,\eta)^{\gamma_2} d\alpha \begin{cases} = \infty, & \gamma_2 \ge 1/2 \\ < \infty, & \gamma_2 < 1/2. \end{cases}$$

定理２：　(1) 最尤推定量

$\alpha \to 0$ のとき $w_{\gamma,k} \to 1/2 - 2\,(\gamma - k/3 - 1/2)$, $\alpha \to \infty$ のとき $w_{\gamma,k} \to 1/\sqrt{2} - 2\sqrt{2}\,(\gamma - k/3 - 1/2)$

したがって，最尤推定量は，任意の損失関数のもとで，漸近２次非許容的である．

(2) 最尤尺度不変 (IML) 推定量 (Zaigraev and Podraza-Karakulska (2008))

最尤推定量は $g(\alpha) = \log\alpha - \psi(\alpha) = \log\bar{x} - \frac{1}{n}\sum_{j=1}^{n} \log x_j$ の解であるが，IML 推定量は $g(\alpha) - g(n\alpha) = \log\bar{x} - \frac{1}{n}\sum_{j=1}^{n} \log x_j$ の解である．このとき，

$\alpha \to 0$ のとき $w_{\gamma,k} \to -2\,(\gamma - k/3 - 1/2)$，$\alpha \to \infty$ のとき $w_{\gamma,k} \to -2\sqrt{2}\,(\gamma - k/3 - 1/2)$

したがって，IML 推定量は，任意の損失関数のもとで，漸近２次非許容的である．

(3) バイアス修正をした最尤推定量

$\alpha \to 0$ のとき $w_{\gamma,k} \to -2\,(\gamma - k/3)$，$\alpha \to \infty$ のとき $w_{\gamma,k} \to -2\sqrt{2}\,(\gamma - k/3)$

したがって，バイアス修正をした最尤推定量は，

(i) 損失関数 $k = 3/2$ のもとで，漸近２次許容的である．

(ii) 損失関数 $k \ne 3/2$ のもとで，漸近２次非許容的である．

(4) 最尤推定量の線形推定量 $\left(1 + \dfrac{a}{n}\right)\hat{\alpha} + \dfrac{b}{n}$

$\alpha \to 0$ のとき $w_{\gamma,k} \to \infty\ (b > 0), a + 1/2 - 2\,(\gamma - k/3 - 1/2)\ \ (b = 0), -\infty\ (b < 0)$

$\alpha \to \infty$ のとき $w_{\gamma,k} \to a/\sqrt{2} + 1/\sqrt{2} - 2\sqrt{2}\,(\gamma - k/3 - 1/2)$

したがって，線形推定量は，損失係数 k をもつ任意の損失関数のもとで，もし，

$$a < -\frac{4}{3}k - 1 \qquad \text{かつ} \qquad b > 0$$

ならば，漸近２次許容的である．

Shrinkage estimation of a mean matrix of a multivariate complex normal distribution

日本女子大学理学部 今野 良彦 (email: konno@fc.jwu.ac.jp)

The multivariate complex normal and complex Wishart distributions were first explored in Goodman [3]. These models play an important role in signal processing methods. Lillestøl [5] first investigated Stein-like shrinkage methods on simultaneous estimation of a mean vector of the complex normal model. However, shrinkage methods for these models have received less attention so far, although it is important to develop these methods beyond the maximum likelihood estimator of estimating the unknown signals in the multivariate complex normal distribution. The goal of this talk is to show how certain decision theoretical results concerning the problem of estimating a mean matrix of the real normal distribution can be extended to the complex multivariate normal case.

In this talk, we consider the problem of estimating an $m \times p$ unknown constant complex matrix $\mathbf{\Xi}$ that is observed with additive complex normal random errors in a decision theoretic set-up. Our observations are an $m \times p$ data matrix $\mathbf{Z}$ and a $p \times p$ positive definite Hermitian matrix $\mathbf{S}$, which is represented as

$$\begin{aligned} &\mathbf{Z} : m \times p \sim \mathbb{C}N_{m\times p}(\mathbf{\Xi}, \mathbf{I}_m \otimes \mathbf{\Sigma}), \\ &\mathbf{S} : p \times p \sim \mathbb{C}W_p(\mathbf{\Sigma}, n) \qquad \text{with } \mathbf{Z} \text{ and } \mathbf{S} \text{ independent}, \end{aligned} \tag{1}$$

where $n > p$, $\mathbf{\Sigma}$ is a $p \times p$ positive definite Hermitian constant matrix. Here we assume that $\mathbf{\Xi}$ and $\mathbf{\Sigma}$ are unknown. Furthermore $\mathbb{C}N_{m\times p}(\mathbf{\Xi}, \mathbf{I}_m \otimes \mathbf{\Sigma})$ and $\mathbb{C}W_p(\mathbf{\Sigma}, n)$ stand for a matrix-variate complex normal distribution with the mean matrix $\mathbf{\Xi}$ and the covariance matrix $\mathbf{I}_m \otimes \mathbf{\Sigma}$ and a complex Wishart distribution with the degree of freedom n and the parameters $\mathbf{\Sigma}$, respectively. In other words, the model (1) means that the density of $\mathbf{Z}$ with respect to the Lebesgue measure on $\mathbb{C}^{m\times p}$ is given as

$$\pi^{-mp}\mathrm{Det}(\mathbf{\Sigma})^{-m}\exp\{-\mathrm{Tr}\big((\mathbf{z}-\mathbf{\Xi})\mathbf{\Sigma}^{-1}(\mathbf{z}-\mathbf{\Xi})^*\big)\}, \qquad \mathbf{z} \in \mathbb{C}^{m\times p},$$

while the density of $\mathbf{S}$ with respect to the Lebesgue measure on $\mathbb{C}_H^{p\times p}$ is given by

$$\frac{\mathrm{Det}\,(\mathbf{s})^{n-p}\exp\big(-\mathrm{Tr}\,(\mathbf{s}\mathbf{\Sigma}^{-1})\big)}{\mathrm{Det}\,(\mathbf{\Sigma})^n\pi^{p(p-1)/2}\Pi_{k=1}^p\Gamma(n+1-k)}, \qquad \mathbf{s} \in \mathbb{C}_+^{p\times p}. \tag{2}$$

Here $\Gamma(\,\cdot\,)$ is the usual Gamma function, $\mathrm{Tr}\,(\,\cdot\,)$ and $\mathrm{Det}\,(\,\cdot\,)$ denote the trace and determinant of a square matrix, and the superscript "*" means the complex conjugate transpose of a matrix. Furthermore $\mathbb{C}^{m\times p}$, $\mathbb{C}_H^{p\times p}$, and $\mathbb{C}_+^{p\times p}$ stand for the sets of all $m \times p$ complex matrices, of all $p \times p$ Hermitian complex matrices, and of all $p \times p$ positive definite Hermitian complex matrices, respectively.

Based on $(\mathbf{Z}, \mathbf{S})$ we consider the problem of estimating the mean matrix $\mathbf{\Xi}$ with respect to a loss function

$$\mathscr{L}(\widehat{\mathbf{\Xi}}, (\mathbf{\Xi}, \mathbf{\Sigma})) = \mathrm{Tr}\,\{(\widehat{\mathbf{\Xi}}-\mathbf{\Xi})\mathbf{\Sigma}^{-1}(\widehat{\mathbf{\Xi}}-\mathbf{\Xi})^*\},$$

where an $m \times p$ random matrix $\widehat{\mathbf{\Xi}}$ is an estimator of $\mathbf{\Xi}$. The risk function corresponding to this loss function is

$$\mathscr{R}(\widehat{\mathbf{\Xi}}, (\mathbf{\Xi}, \mathbf{\Sigma})) = \mathbb{E}[\mathscr{L}(\widehat{\mathbf{\Xi}}, (\mathbf{\Xi}, \mathbf{\Sigma}))],$$

where the expectation above is taken with respect to the joint distribution of $(\mathbf{Z}, \mathbf{S})$.

This estimation problem is important since it is a prototype of estimating the regression matrix of a complex MANOVA model and of predicting multivariate responses in a linear regression complex model. We extend a large body of the results obtained by Efron and Morris [2] and Konno [4] in the multivariate real normal set-up to the complex normal set-up (1). The results in the real normal model were obtained by extensive use of the integration by parts approach, known as the Stein identity. In addition to these identities, the eigenvalue calculus is important to the development for a systematic search for shrinkage estimators. We extend these approaches to the complex normal set-up. The Stein identity for the multivariate complex normal is easily derived by using an isomorphism between real and complex variables stated in Andersen *et al.* [1] while the Stein-Haff identity was extended to the complex Wishart distribution by Svensson and Lundberg [6]. These identities and the eigenvalue calculus for the complex matrix developed in this paper are exploited to establish a systematic search for shrinkage estimators for the model (1). The detailed proof for the results is available at http://mp-w3math.jwu.ac.jp/~konno/pdf/tr10.pdf.

References

[1] H.H. Andersen, M. Højbjerre, D. Sørensen, P.S. Eriksen, Linear and Graphical Models, Springer-Verlag, New York (1995).

[2] B. Efron, C. Morris, Families of minimax estimators of the mean of a multivariate normal distribution, Ann. Statist. **4** 11–21.

[3] N.R. Goodman, Statistical analysis based on a certain multivariate complex Gaussian distribution (An introduction), Ann. Math. Statist. **34** (1963) 152-176.

[4] Y. Konno, On estimation of a matrix of normal means with unknown covariance matrix. J. Multivariate Anal. **36** (1991) 44–55.

[5] J. Lillstøl, Improved estimators for multivariate complex-normal regression with application to analysis of linear time-invariant relationships, J. Multivariate Anal. **7** (1977) 512–524.

[6] L. Svensson, M. Lundberg, Estimating complex covariance matrix, Signals, Systems and Computers, Conference Record of the Thirty-Eighth Asilomar Conference on (2004) 7-10, 2151 - 2154.

On conditions for a modified maximum likelihood estimator to be second order admissible

Hidekazu Tanaka[1], Yoshiji Takagi[2] and Junyong Park[3]

[1]Graduate School of Engineering, Osaka Prefecture University,
1-1 Gakuen-cho, Nakaku, Sakai, Osaka, 599-8531, Japan

[2]Faculty of Education, Nara University of Education,
Takabatake-cho, Nara, 630-8528, Japan

[3]Department of Mathematics and Statistics,
University of Maryland, Baltimore County,
1000 Hilltop Circle, Baltimore, MD 21250, USA

Abstract

In small sample theory of point estimation, the concept of admissibility has been studied from various point of view. On the other hand, in large sample theory, it seems that the concept of second order admissibility of estimator did not fully developed to multi-parameter case. In this article, we derive some conditions for a modified maximum likelihood estimator to be second order admissible under the normalizing quadratic loss function when the dimension of the parameter is 2.

1 Introduction

Consider a sequence of independently and identically random vectors according to some probability distribution parameterized by unknown parameters. In small sample theory, the concept of admissibility has been studied from various point of view. Especially, under the quadratic loss function it is well known that if the covariance matrix is known the sample mean is admissible in estimating the normal mean when the dimension of the parameter is 1 (Hodges and Lehmann (1950), Karlin (1958)), and 2 (Stein (1956), Brown and Hwang (1982)). Furthermore, the sample mean is inadmissible when the dimension is at least 3 (James and Stein (1961)). On the other hand, in large sample theory, Ghosh and Sinha (1981) proposed a concept of second order admissibility of point estimator to asymptotically solve the Berkson's Bioassay problem, and derived a necessary and sufficient condition for a modified maximum likelihood estimator to be second order admissible for one parameter case. Using the condition in Ghosh and Sinha (1981), Takagi (1999a) discussed the invariance property of the second order admissibility (see also Takagi (1999b)).

AMS 2000 subject classification. Primary 62F12; secondary 62C15, 62H12.

Key words and phrases. Modified maximum likelihood estimator, multi-parameter, second order admissibility.

For multi-parameter case, DasGupta and Ghosh (1983) gave a sufficient condition so that a modified maximum likelihood estimator which is asymptotically unbiased should be second order admissible under the quadratic loss function when the dimension of the parameter is 2 and the Fisher information matrix is diagonal. Also they derived a necessary condition so that such a estimator should be second order admissible when the dimension is at least 3. However, the probability distribution and the modified maximum likelihood estimator which they treated were restricted to special forms. As is shown by the results of DasGupta and Ghosh (1983), the sample mean is second order admissible in estimating the normal mean if and only if the dimension of the parameter is at most 2 when the covariance matrix is identity. Therefore, it seems that we need to consider the necessary and/or sufficient conditions when the dimension of the parameter is at most 2 or at least 3, separately. In this article, we consider the similar problem to DasGupta and Ghosh (1983) by another approach when the dimension of the parameter is 2, and give some conditions for a modified maximum likelihood estimator to be second order admissible under the normalizing quadratic loss function. As examples of these results, we give the second order admissibility and inadmissibility of (generalized) Bayes estimators of mean vector in 2 dimensional normal distribution when the covariance matrix is identity.

References

[1] Brown, L. D. and Hwang, J. T. A unified admissibility proof. *Statistical decision theory and related topics, III*, Vol. 1, 205–230, Academic Press, New York-London, 1982.

[2] DasGupta, A. and Ghosh, J. K. Some remarks on second-order admissibility in the multiparameter case. *Sankhyā* Ser. A **45** (1983), 181–190.

[3] Ghosh, J. K. and Sinha, B. K. A necessary and sufficient condition for second order admissibility with applications to Berkson's bioassay problem. *Ann. Statist.*, **9** (1981), 1334–1338.

[4] Hodges, J. L., Jr. and Lehmann, E. L. Some problems in minimax point estimation. *Ann. Math. Statistics*, **21** (1950), 182–197.

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

[6] Karlin, S. Admissibility for estimation with quadratic loss. *Ann. Math. Statist.*, **29** (1958), 406–436.

[7] Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, vol. I, 197–206. University of California Press, Berkeley and Los Angeles, 1956.

[8] Takagi, Y. Parametrization invariance with respect to second order admissibility under mean squared error. *Ann. Inst. Statist. Math.* **51** (1999), 99–110.

[9] Takagi, Y. Parametrization invariance with respect to second order admissibility under general loss function. *Sankhyā* Ser. A **61** (1999), 113–119.

最大値の推定

神戸大学・海事科学研究科　高橋 倫也

平成 22 年 11 月 17 日

概 要

極値理論に基づく仮定の下で，上に有界な台を持つ分布の上限点の推定と信頼区間の構成について述べた．

1　序と極値理論

上に有界な台を持つ確率分布について極値理論に基づく仮定の下で，分布の上限点 (endpoint，bound) の推定とその信頼区間の構成について述べる．分布の下限点や関数形が未知または計算困難な関数の最大値の推定についても同様に議論できる．

以下，次の仮定と記号を用いる：

独立で同一分布 F に従う確率変数 $X_1, X_2, \ldots, X_n$ を考える．順序統計量を $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ とし，上限点 (endpoint) を $x_F = \sup\{x : F(x) < 1\}$ とする．ここでは，n は十分大きいと仮定し，その値は未知でも良いとし，分布 F の関数形は未知の場合を考える．観測方法に応じて得られるデータに基づく上限点の推定法について述べる．

分布 F の台は上に有界，$\theta \equiv x_F = \sup\{x : F(x) < 1\} < \infty$ の場合を考える．以下 θ の推定を問題にする．ここで，F は負の Weibull 分布 $(\Psi_\alpha(x) = \exp\{-(-x)^\alpha\})$ の吸引領域に属する，$F \in \mathcal{D}(\Psi_\alpha)$，と仮定する．すなわち，

$$\theta < \infty \qquad \text{and} \qquad \lim_{x \downarrow 0} \frac{1 - F(\theta - tx)}{1 - F(\theta - x)} = t^\alpha, \quad \forall t > 0.$$

分布族 $\{\Psi_\alpha((x-\theta)/\sigma),\ \theta \in \mathbb{R},\ \sigma,\ \alpha > 0\}$ に関して，Smith (1985) より $\alpha > 2$ のとき最尤推定量は漸近有効であることが知られている．

$F \in \mathcal{D}(\Psi_\alpha)$ より，数列 $a_n > 0$ が存在して

$$\lim_{n\to\infty} P\left\{\frac{X_{n:n} - \theta}{a_n} \le x\right\} = \Psi_\alpha(x), \quad x \le 0$$

となる．ここで，上位 r 個までの極値統計量を基準化したベクトル

$$\left(\frac{X_{n:n} - \theta}{a_n}, \ \cdots, \ \frac{X_{n-r+1:n} - \theta}{a_n}\right)$$

を考える．基準化した確率ベクトルは，$\{E_i\}$ を Exp(1) に従う独立確率変数列とすると，

$$\big(-E_1^{1/\alpha},\ -(E_1+E_2)^{1/\alpha}, \ldots, -(E_1+\cdots+E_r)^{1/\alpha}\big)$$

の分布に分布収束する．

2　分布の上限点の推定

極値理論の結果を用いて，観測データに応じた分布の上限点の推定について述べた.

2.1　上位 r 個のデータ

上位 r 個のデータのみが観測されていて n が未知の場合.

α **が既知**：Weissman (1981) のピボット比

$$R_{rn} = \frac{\theta - X_{n:n}}{X_{n:n} - X_{n-r+1:n}}, \quad r \geq 2.$$

に基づく上限点 θ の信頼区間の構成法とその性質について紹介した.

α **が未知**：Weissman (1982) の 2 つのピボット比

$$W_r^{(n)} = \log \frac{\theta - X_{n-r+1:n}}{\theta - X_{n:n}} \Big/ \sum_{i=1}^{r-1} \log \frac{\theta - X_{n-r+1:n}}{\theta - X_{n-i+1:n}} \qquad (r \geq 3)$$

と

$$Q_{rm}^{(n)} = \log \frac{\theta - X_{n-m+1:n}}{\theta - X_{n:n}} \Big/ \log \frac{\theta - X_{n-r+1:n}}{\theta - X_{n-m+1:n}} \qquad (1 < m < r \leq n).$$

による上限点 θ の信頼区間の構成法とその性質について紹介した.

2.2　上位 r 個のデータが n 個

$\alpha(> 2)$ が既知の場合に情報量に基づき，上位 $r(\geq 2)$ 個までのデータを利用する相対効率とその性質について述べた.

2.3　閾値以上のデータ

論文 Einmahl and Magnus (2008) に基づき分布の上限点の推定法と，現在の最大値（世界記録）の評価法について紹介した．また，閾値以上のデータに一般パレート分布を適合して分布の上限点を推測する方法について述べた.

文献

Einmahl, J. H. J. and Magnus, J. R. (2008). Records in athletics through extreme-value theory. *J. Amer. statist. Assoc.* **103**, 1382–1391.

Weissman, I. (1981). Confidence intervals for the threshold parameter. *Comm. Statist. Theory Method A* **10**, 549–557.

Weissman, I. (1982). Confidence intervals for the threshold parameter II: unknown shape parameter. *Comm. Statist. Theory Method* **11**, 2451–2474.

Parameter estimated standardized U-statistics for some dependent sequences and its application to change-point problems

K. Yoshihara (Yokohama National University)

S. Kanagawa (Tokyo City University)

1 Results for some mixing sequences

Let $\{\xi_i\}$ be a stationary sequence of random variables with common distribution function F. Let Θ be an open set in R^d. Let h be the kernel which satisfies

$$h(x,y;\theta) \;=\; h(y,x;\theta)$$

for any x, y $(\in \mathsf{R})$ and θ $(\in \Theta)$. Let θ_0 be a true value. and assume

$$\int h(x,y;\theta_0)dF(x)dF(y) \;=\; \theta_0. \tag{1}$$

Define the projection function

$$h_1(x;\theta) \;=\; \int (h(x,y;\theta)-\theta)dF(y). \tag{2}$$

and assume that

$$\int h_1^2(x;\theta_0)dF(x) \;=\; 0, \tag{3}$$

i.e., the kernel $h(x,y;\theta_0)$ generates a degenerate U-statistic. Let $T_h f\,(x) = \int h\,(x,y;\theta_0)\,f\,(y)\,F\,(dy)$ be a trace class operator for $f \in L^2\,(F\,(dx))$. If, in addition to (3),

$$\int h^2(x,y;\theta_0)dF(x)dF(y) \;<\; \infty \tag{4}$$

there exist eigenvalues $\{\lambda_i\}$ and eigenfunctions $\{\varphi_i(t)\}$ for the linear operator T_h such that

$$\sum_{i=1}^{\infty}\lambda_i^2 \;<\; \infty, \tag{5}$$

$$\int \varphi_i(x)\varphi_j(x)dF(x) = \delta_{i,j}, \tag{6}$$

$$\lim_{M\to\infty}\int\int\Big\{\sum_{i=1}^{M}\lambda_i\varphi_i(x)\varphi_i(y)-h(x,y;\theta_0)\Big\}^2 dF(x)dF(y) \;=\; 0. \tag{7}$$

Let $\{\xi_i\}$ be a stationary sequence of random variables satisfying the absolutely regular condition

$$\beta(k) \;=\; 2\sup_{n\geq 1}\Big\{\sup_{A\in\mathcal{M}_{n+k}^{\infty}}(P(A|\mathcal{M}_1^n)-P(A))\Big\}$$

where $\mathcal{M}_a^b = \sigma(\xi_a,\cdots,\xi_b)$. Let θ_0 be a true value of θ. The following theorem is shown in Yoshihara and Kanagawa (2006) when the parameter θ_0 is known.

Theorem A. *Suppose Condition A holds. Then*

$$\frac{1}{\sqrt{2\log\log n}}\max_{1\leq k\leq n}\frac{1}{\sqrt{k(n-k)}}\Big|\sum_{i=1}^{k}\sum_{j=k+1}^{n}h(\xi_i,\xi_j;\theta_0)\Big| \;\xrightarrow{D}\; \Big(\sum_{i=1}^{\infty}\lambda_i N_i^2\Big)^{\frac{1}{2}}.$$

When the parametor θ_0 is unknown, we have the following theorem.
Theorem 1. *Suppose Conditions A and B hold. Then*

$$\frac{1}{\sqrt{2\log\log n}} \max_{1\le k\le n} \frac{1}{\sqrt{k(n-k)}} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} h(\xi_i, \xi_j; \hat{\theta}_n) \right| \xrightarrow{D} \vartheta^{\frac{1}{2}},$$

where

$$\vartheta \;=\; \sum_{i=1}^{\infty} \lambda_i^2 N_i^2 + 2\sigma^2 \rho^T \sum_{i=1}^{\infty} \lambda_i N_i b_i + \rho^T B \rho. \tag{8}$$

2 Functionals of processes

Let $\{\xi_n; n \in \mathsf{Z}\}$ be a stationary stochastic process. Assume that $\{\eta_n; n \in \mathsf{Z}\}$ be a two-sided functional sequence of $\{\xi_n\}$, i.e., there is a measurable function f defined on R^{Z} such that

$$\eta_n \;=\; f(\{\xi_{n+k}; k \in \mathsf{Z}\}) = f\left(\cdots, \xi_{n-1}, \xi_n, \xi_{n+1}, \cdots\right).$$

Thus, $\{\eta_n; n \in \mathsf{Z}\}$ is necessarily a stationary stochastic process. Then we have the following result.
Theorem 2. *Let $\{\eta_n; n \in \mathsf{Z}\}$ be a 1-approximating functional with constants $\{a(k); k \ge 0\}$ of absolutely regular process with mixing coefficients $\{\beta(k); k \ge 0\}$. Let $\ell(\cdot)$ and $h(\cdot,\cdot)$ be 1-continuous functions with the same function $\phi(\epsilon)$. Suppose that*

$$\hat{\theta}_n - \theta_0 \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell(\eta_i) + o_P\left(n^{-\frac{1}{2}}\right).$$

and that for some constants $r \in (2,3)$ and $\delta \in (0, 1-(r/3))$

$$E\eta_i \;=\; 0, \quad E|\eta_i|^{r+\delta} \;<\; \infty$$

Then

$$\frac{1}{\sqrt{2\log\log n}} \max_{1\le k\le n} \frac{1}{\sqrt{k(n-k)}} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} h(\eta_i, \eta_j; \theta_n) \right| \xrightarrow{D} \vartheta^{\frac{1}{2}}.$$

References

[1] Borovkova, S., Burton, R. and Dehling H. (2001). Limit Theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Trans. A. M. S.*, **353**, 4261-4318.

[2] Gombay, and Horváth. (1998). Parameter estimated standardized U-statistics. *Asymptotic methods in Probability and Statistics*, Szyskowicz, B. (Editor), Elsevier Science B. V.

[3] Yoshihara, K. (1992). *Weakly dependent stochastic sequences and their applications, Vol I. Summation theory for weakly dependent sequences.* Sanseido, Tokyo.

[4] Yoshihara, K. (2005). *Weakly dependent stochastic sequences and their applications, Vol XV. Recent topics on limit theorems for functionals of processes with contents and index of this series.* Sanseido, Tokyo.

[5] Yoshihara, K. and Kanagawa, S. (2010). Limit theorems for maximum of standardized U-statistics defined by weakly dependent sequences. (Submitted).

無限次元マルチンゲール中心極限定理の使用法

西山陽一

統計数理研究所

〒 190-8562 東京都立川市緑町 10-3

nisiyama@ism.ac.jp

http://www.ism.ac.jp/~nisiyama/

平成 22 年 11 月 8 日

$X_0, X_1, X_2, ...$ は実数値マルコフ連鎖であるとし，その推移密度を $p(x, y)$ とする．すなわち，

$$P(X_i \in A | X_{i-1}) = \int_A p(X_{i-1}, y)dy, \quad \forall A \in \mathbf{B}(\mathbb{R}),\ i = 1, 2, ...$$

この例題を通じて，マルコフ連鎖がエルゴード的あることを仮定し，その不変分布を $\pi(dx)$ とする．確率場

$$\begin{aligned} G_n(x; p) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ 1_{(-\infty, x]}(X_i) - \int_{-\infty}^{x} p(X_{i-1}, y)dy \right\}, \\ H_n(x; p) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ 1_{(-\infty, x]}(X_i) - \int_{-\infty}^{x} p(X_{i-1}, y)dy \right\} \sqrt{p(X_{i-1}, x)} \end{aligned}$$

を導入する．後に我々は，弱収束

$$G_n(\cdot; p) \to^d B^\circ(F(Z, \cdot)) \quad \text{in } \ell^\infty([-\infty, \infty]), \tag{1}$$

$$H_n(\cdot; p) \to^d B^\circ(F(Z, \cdot))\sqrt{p(Z, \cdot)} \quad \text{in } L_2(\mathbb{R}) \tag{2}$$

が成立するための十分条件を与える．ただし $u \rightsquigarrow B^\circ(u)$ は標準ブラウン橋であり，$F(z, x) = \int_{-\infty}^{x} p(z, y)dy$ とし，また Z は B° とは独立な確率変数であって分布 $\pi(dz)$ をもつようなものである．

ひとたびこれが示されたら，次のようにして推移密度の適合度検定問題を考えることができる．各 $z \in \mathbb{R}$ に対し $x \mapsto F(z, x)$ は連続分布関数であるから，連続写像定理を用いることにより，帰無仮説 $p = p_0$ のもとで，コルモゴロフ-スミルノ

フ (KS) 型検定統計量の収束

$$\begin{aligned}\sup_{x\in\mathbb{R}}|G_n(x;p_0)| &\to^d \sup_{x\in\mathbb{R}}|B^\circ(F_0(Z,x))| \\ &=^d \sup_{u\in[0,1]}|B^\circ(u)|\end{aligned}$$

が成り立つ．ただし $F_0(z,x)=\int_{-\infty}^{x}p_0(z,y)dy$ である．また，クラメール-フォン・ミセス (CvM) 型統計量については

$$\begin{aligned}\int_{-\infty}^{\infty}|H_n(x;p_0)|^2dx &\to^d \int_{-\infty}^{\infty}|B^\circ(F_0(Z,x))|^2p_0(Z,x)dx \\ &=^d \int_0^1|B^\circ(u)|^2du\end{aligned}$$

が従う．これらの検定は漸近的に分布不変であることに注意せよ．一方，これらの検定が対立仮説

$$H_1:\quad \int_{-\infty}^{\infty}(F(z,x)-F_0(z,x))\pi(dz)\neq 0 \quad \text{for some } x \quad \text{(KS)},$$

および

$$H_1:\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|F(z,x)\sqrt{p(z,x)}-F_0(z,x)\sqrt{p_0(z,x)}|^2dx\pi(dz)>0 \quad \text{(CvM)}$$

のもとで一致性をもつことが示される．

なお，マルコフ連鎖の状態空間が多次元の場合にも類似した弱収束をしめすことはできるが，検定は漸近的分布不変にはならないことを注意しておく．

$\ell^\infty(\mathbb{R})$ 空間における弱収束 (1) を導くための付加的仮定は，ある $\mathbb{R}$ 上の π-可積分関数 $g\geq 0$ とルベーグ可積分関数 $h\geq 0$ が存在して

$$p(x,y)\leq g(x)h(y),\quad \forall x,y\in\mathbb{R}$$

が成り立つことである．

一方，$L_2(\mathbb{R})$ 空間における弱収束 (2) を導くための付加的仮定は，マルコフ連鎖が定常であることである．

参考文献

[1] Nishiyama, Y. (2010). Goodness-of-fit tests for ergodic Markov chains as clean applications of infinite-dimensional martingale central limit theorems. *Submitted for publication.*

Goodness of Fit for Randomly Censored Data

白石 博（東京慈恵会医科大学）, 谷合 弘行（早稲田大学）

例えば、ある疾病の罹患者に対する生存時間やその対数変換のような興味のある確率変数を Y_i とし、$\boldsymbol{x}_i = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')'$ を $p+q$ 次元の観測可能な共変量とする。 生物・医学統計などの多くの場面で、患者の退院や興味のある事由以外による死亡などを理由として、観測途中での観察の打ち切りにより Y_i が完全に観測出来ない場合がある。本研究は、このような打ち切りデータがランダムに存在する場合の "goodness-of-fit test" を考える。 C_i を打ち切りデータとし、その分布は $\boldsymbol{x}_i$ にのみ依存するものとする。このとき、各個体に対し、実際に観測される変量は $(\boldsymbol{x}_i, \tilde{Y}_i, \Delta_i)$ となる。ここで、

$$\tilde{Y}_i = \min(Y_i, C_i), \quad \Delta_i = \mathbb{I}(Y_i \leq C_i)$$

とし、Y_i は線形モデル

$$Y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \epsilon_i$$

に従うと仮定する。ただし、ϵ_i はある確率変数に従う誤差項とする。今、次の検定問題を考える。

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \text{ v.s. } H_n : \boldsymbol{\beta}_2 = \tfrac{1}{\sqrt{n}}\boldsymbol{h}$$

Gutenbrunner and Jurečková (1992) などは、(打ち切りデータの無い) このような線形モデルに対してランク統計量を提案した。彼らのアプローチは、Koenker and Bassett (1978) による "regression quantile statistics" の双対問題を基にしており、Hájek and Šidák (1967) によって提案された伝統的なランク統計量の自然な拡張である。本報告では、Gutenbrunner and Jurečková (1992) などによって提案されたランク統計量を、打ち切りデータが存在する場合に拡張する。

まず、ランク統計量の双対問題である "regression quantile statistics" において、Koenker and Bassett (1978) は打ち切りデータが無い場合に、"quantile coefficent" の推定量 $\hat{\boldsymbol{\beta}}_1(\tau)$ を提案しているが、打ち切りデータが存在する場合には、Portnoy (2003) などによって提案された "recursive reweighting scheme" を適用して $\hat{\boldsymbol{\beta}}_1(\tau)$ を推定する。この結果を双対問題として適用してランク統計量を提案し、その漸近分布を報告した。

格子点 $\{0 < t_1 < t_2 < \cdots < t_M < 1\}(\exists M = o(n^{1/2}))$ を設定し、$\{\tilde{Y}_1, \ldots, \tilde{Y}_n\}$ を観測列とする。また、n 個の $\tilde{Y}_i$ の内、打ち切りデータが L 個 ($L \ll n$) 存在すると仮定し、打ち切りデータの順序統計量を $C_{(l)}(l = 1, \ldots, L)$ とする。まず、観測列の t_j 変位点[1]までは打ち切りデータが無いとして、従来の " quantile regression" (QR) の手法[2]により $\hat{\boldsymbol{\beta}}_1^{(0)}(t_1), \ldots, \hat{\boldsymbol{\beta}}_1^{(0)}(t_j)$ および $\hat{\boldsymbol{\beta}}_1^{(0)}(t_{j+1})$ を構成する。次に、t_{j+1} 変位点が上回る打ち切りデータを $C_i(= C_{(1)})$ とし、$\hat{\tau}^{(0)}$ を次式で定義する。

$$\hat{\tau}^{(0)} := (1 - \hat{\alpha}_{i(j)}^{(0)})t_j + \hat{\alpha}_{i(j)}^{(0)}t_{j+1}, \quad \hat{\alpha}_{i(j)}^{(0)} := \frac{C_i - \boldsymbol{x}_{1i}'\hat{\boldsymbol{\beta}}_1^{(0)}(t_j)}{\boldsymbol{x}_{1i}'\{\hat{\boldsymbol{\beta}}_1^{(0)}(t_{j+1}) - \hat{\boldsymbol{\beta}}_1^{(0)}(t_j)\}}$$

---

[1] t_j 変位点とは、観測列 $\{\tilde{Y}_1, \ldots, \tilde{Y}_n\}$ の順序統計量を $\{\tilde{Y}_{(1)}, \ldots, \tilde{Y}_{(n)}\}$ としたときの $\tilde{Y}_{(nt_j)}$ の事である。

[2] 従来の QR による推定量とは、(1) 式の左辺において、$\hat{w}_i^{(0)}(t) = 1$ としたときの解のことである。

この $\hat{\tau}^{(0)}$ を使ってウエイト $\hat{w}_k^{(0)}(t)(k=1,\ldots,n)$ を次式で定義する。

$$\hat{w}_k^{(0)}(t) := \begin{cases} \frac{t-\hat{\tau}^{(0)}}{1-\hat{\tau}^{(0)}} & if\ \Delta_k = 0\ and\ t \geq \hat{\tau}^{(0)} \\ 0 & if\ \Delta_k = 0\ and\ t < \hat{\tau}^{(0)} \\ 1 & if\ \Delta_k = 1 \end{cases}$$

このウエイト $\hat{w}_i^{(0)}(t)$ を使って次式で定義する "weighted quantile regression" (weighted QR) の手法により、$\hat{\boldsymbol{\beta}}_1^{(1)}(t_{j+1}),\ldots,$ を $t_l(l=j+1,\cdots)$ 変位点が $C_{(2)}$ を超えるまで構成する。

$$\hat{\boldsymbol{\beta}}_1^{(1)}(t) := \arg\min_{\boldsymbol{b}} \left\{ \sum_{i=1}^{n} \hat{w}_i^{(0)}(t) \rho_t(\tilde{Y}_i - \boldsymbol{x}_{1i}'\boldsymbol{b}) \right\} \tag{1}$$

ここに、$\rho_t(u) := u(t - \mathbb{I}\{u<0\})$ とする。以下、同様に繰り返して

$$(\hat{\boldsymbol{\beta}}_1^{(0)}(t_1),\ldots,\hat{\boldsymbol{\beta}}_1^{(1)}(t_{j+1}),\ldots,\hat{\boldsymbol{\beta}}_1^{(L)}(t_M))$$

および

$$(\hat{\boldsymbol{w}}^{(-1)}(t_1),\ldots,\hat{\boldsymbol{w}}^{(0)}(t_{j+1}),\ldots,\hat{\boldsymbol{w}}^{(L-1)}(t_M))$$

を構成する。ここで、$\hat{\boldsymbol{w}}^{(l)}(t) = (\hat{w}_1^{(l)}(t),\ldots,\hat{w}_n^{(l)}(t))'$ および $\hat{\boldsymbol{w}}^{(-1)}(t) = (\Delta_1,\ldots,\Delta_n)'$ とする。このウエイトを単に $\hat{\boldsymbol{w}}(t)$ と表し、$\hat{\boldsymbol{a}}_n(\tau), \hat{b}_{ni}, \boldsymbol{S}_n$ およびランク統計量 $\boldsymbol{T}_n$ を次式で定義する。

$$\begin{aligned} \hat{\boldsymbol{a}}_n(\tau) &:= \arg\max_{\boldsymbol{a}} \left\{ \tilde{\boldsymbol{Y}}'\boldsymbol{a} | \boldsymbol{X}_1'\boldsymbol{a} = (1-\tau)\boldsymbol{X}_1'\hat{\boldsymbol{w}}(\tau),\ \boldsymbol{0} \leq \boldsymbol{a} \leq \hat{\boldsymbol{w}}(\tau) \right\} \\ \hat{b}_{ni} &:= -\int_0^1 \varphi(t) d\hat{a}_{ni}(t) \\ \boldsymbol{S}_n &:= \frac{1}{\sqrt{n}} (\boldsymbol{X}_2 - \hat{\boldsymbol{X}}_2) \boldsymbol{\Delta} \hat{\boldsymbol{b}}_n \\ \boldsymbol{T}_n &:= \frac{\boldsymbol{S}_n' \boldsymbol{Q}_n^{-1} \boldsymbol{S}_n}{A^2(\varphi)}. \end{aligned}$$

このとき、適当な正則条件の下、次を得る。

定理

(i) 帰無仮説 H_0 の下で、$\boldsymbol{T}_n$ は漸近的に自由度 q の χ^2 分布に従う

(ii) 対立仮説 H_n の下で、$\boldsymbol{T}_n$ は漸近的に自由度 q の非心 χ^2 分布に従う

References

Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.*, 20(1):305–330.

Hájek, J. and Šidák, Z. (1967). *Theory of rank tests.* Academic Press, New York.

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

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.*, 98(464):1001–1012.

Portnoy, S. and Lin, G. (2010). Asymptotics for censored regression quantiles. *J. Nonparametr. Stat.*, 22(1-2):115–130.

Linear Regression with Deterministic Regressors and Unit Root in the Variance

Alexandre Petkovic*

October 2010

Abstract

The first part of this paper derives the asymptotic distribution of the ordinary least squares estimator in a linear regression model with deterministic regressors when the variance of the innovations is a function of an integrated time series. In the second part of this paper we study the impact of heteroscedasticity on the standard t-test for the slope coefficient in a linear trend model.

Keywords: nonlinear, integrated time series, deterministic regressors, regression with heteroskedastic errors

1 Introduction

The tools for the study of a linear system of an integrated time series were introduced by Phillips (1986,1987) and Phillips and Durlauf (1986). Their results relied on weak convergence in functional spaces, the continuous mapping theorem and on weak convergence of stochastic integrals to martingales. It is with the papers of Phillips and Park (1999, 2001) that the study of the asymptotic behavior of nonlinear functions of an integrated time series started. Phillips and Park (1999, 2001) derived the asymptotic distribution of the average of a nonlinear function of integrated time series. These results where further extended by Chang and Park (2003), Jong and Wang (2005) and Shi and Phillips (2010).

The results obtained by Park and Phillips (1999, 2001) have been applied to various nonlinear econometric models. For example, Park and Phillips (2001) and Shi and Phillips (2010) used them to derive the distribution of the least squares estimator in a nonlinear regression model. Chang and all (2001) considered nonlinear regressions with separably additive regression functions. Park (2002) studied the possibility of modeling assets variance using a nonlinear function of an integrated time series. Studying the USD/DM exchange rate he found out that the conditional variance of the spread can be accurately modeled using the spot rate. Chung and Park (2003) considered nonstationary index models. Hu and Phillips (2004) worked on discrete choice models. Chang and Park (2005) studied the distribution of the ordinary least squares estimator of a linear regression model with integrated or stationary regressors when the error term volatility is a nonlinear function of an integrated time series. They showed that, when the volatility of the error term is a function of an integrated time series, the asymptotic distribution of the ordinary least squares is nonstandard and involves an integral with respect to the local time of a Brownian motion at the origin.

The objective of this paper is twofold. Firstly, we extend some results of Phillips and Park (1999,2001) by deriving the asymptotic distribution of a temporally weighted average of a function of an integrated time series. Secondly, we use our results to study the distribution of the ordinary least squares estimator of linear regression model when the regressors are time-deterministic and the variance of the error term is a function

*Waseda University, Center for English Language Education in Science and Engineering, email: apetkovi@aoni.waseda.jp

of an integrated time series. Using these results we also study the asymptotic distribution of the standard t-stat. In this sense our results can be seen as an extension of those derived by Chung and Park (2007). As explained bellow potential applications of our results can be found in macroeconometrics and finance.

The paper is organized as follows: Section 2 presents the model and the assumptions, Section 3 studies the asymptotic distribution of the ordinary least squares estimator, in Section 4 we consider an application of our theory and study the impact of heteroscedasticity on the level of the standard t-test, in Section 5 we propose some further applications of our theory, finally Section 5 concludes. All the proofs can be found in the appendix. Throughout this paper we will use the following notations: $\rightarrow_d$ and $\rightarrow_p$ will mean convergence in distribution and in probability, respectively. $\mathbb{N}^*$ and $\mathbb{N}$ will denote the integer and the non-negative integer, respectively.

2 The model and Assumptions

Consider the regression model given by

$$y_{n,t} = \alpha + g_{1,n}(t)\beta_1 + \cdots + g_{k,n}(t)\beta_k + \epsilon_{n,t}, \qquad t = 1, \ldots, n \tag{1}$$

where $y_{n,t}$ is the depend variable and the $g_{i,n}(t)$'s are deterministic regressors. The error term, $\epsilon_{n,t}$, is modeled as

$$\epsilon_{n,t} = \sigma(z_t)u_{n,t},$$

where $u_{n,t}$ is a martingale difference sequence with mean zero and unit variance with respect to a filtration $\mathcal{F}_{n,t}$, z_t is an integrated time series and σ a function whose properties will be specified bellow. We assume that z_t is measurable with respect to $\mathcal{F}_{n,t-1}$ implying that $(\epsilon_{n,t}, \mathcal{F}_{n,t})$ is a martingale difference sequence satisfying

$$E(\epsilon_{n,t}^2|\mathcal{F}_{n,t-1}) = \sigma^2(z_t).$$

Let $[s]$ be the larger integer smaller than s. Through this paper we will assume that each deterministic regressor $g_{i,n}(t)$ satisfies the following assumption.

Assumption 1. *Let $g_{i,n}(t)$, $t = 1, \ldots, n$, be a sequence of finite valued deterministic regressors. Then there exists a positive function $c_i(n)$, whose limit as $n \to \infty$ exists in $\bar{\mathbb{R}}$, such that*

$$sup_{r\in[0,1]}|\frac{g_{i,n}([rn])}{c_i(n)} - \breve{g}_i(r)| \to 0,$$

where $\breve{g}_i(r)$ is piecewise continuous on $[0,1]$ and satisfies $\int_0^1 |\breve{g}_i(r)|dr \neq 0$.

We assume that the process z_t is of the form

$$z_t = z_{t-1} + w_t, \tag{2}$$

where w_t follows the linear process

$$w_t = \psi(L)e_t = \sum_{k=0}^{+\infty} \psi_k e_{t-k},$$

where e_t is an iid sequence of random variables with mean zero. In this paper we set $z_0 = 0$.

Pooling incomplete samples の状況下における統計的推論

関東学院大 経済　布能 英一郎

1. Introduction 自然数 k, m を $m < k$ に選ぶ。確率変数 $\mathbf{X}$, $\mathbf{Y}$ は 互いに独立で

$$\mathbf{X} = (X_0, X_1, X_2, \cdots, X_k) \sim \text{Multinomial}(N_1; p_0, p_1, p_2, \cdots, p_k),$$
$$\mathbf{Y} = (Y_0, Y_1, \cdots, Y_m) \sim \text{Multinomial}(N_2; \frac{p_0}{\sum_{j=0}^m p_j}, \frac{p_1}{\sum_{j=0}^m p_j}, \cdots, \frac{p_m}{\sum_{j=0}^m p_j}).$$

とする。このような状況のもとで、Asano(1965) は p_i の MLE $\hat{p}_i$ が

$$\frac{x_i + y_i}{N_1\Big(1 + \frac{N_2}{\sum_{j=0}^m x_j}\Big)} \quad \text{if } i \le m, \qquad \frac{x_i}{N_1} \quad \text{if } i > m$$

で与えられることを示し、更に、$\hat{p}_i$ に関するいくつかの性質を研究した。さて、これと同様な現象は、他の分布の下でもいくつか生じている。

例 1. (Poisson) $X_1, \cdots, X_k$, $Y_1, \cdots, Y_m$ は すべて独立であって, $X_i \sim Poisson(\lambda_i)$, $i = 1, \cdots, k$, $Y_i \sim Poisson((\sum_{l=1}^k \lambda_l)(\lambda_i / \sum_{j=1}^m \lambda_j)$, $i = 1, \cdots, m$, とする。このとき、$T_x = \sum_{j=1}^k x_j$, $T_y = \sum_{j=1}^m y_j$ とおくと、λ_i の MLE $\hat{\lambda}_i$ は

$$\frac{T_x + T_y}{2} \frac{x_i + y_i}{T_x\Big(1 + \frac{T_y}{\sum_{j=1}^m x_j}\Big)} \quad \text{if } i \le m, \qquad \frac{T_x + T_y}{2} \frac{x_i}{\sum_{j=1}^m x_j} \quad \text{if } i > m.$$

例 2. (負の多項分布) 自然数 k, m は $m < k$. 確率変数 $\mathbf{X}$, $\mathbf{Y}$ は 互いに独立で

$$\mathbf{X} = (X_1, \cdots, X_m, \cdots, X_k) \sim \text{負の多項分布 } (r_1; p_1, \cdots, p_k),$$
$$\mathbf{Y} = (Y_1, \cdots, Y_m) \sim \text{負の多項分布 } (r_2; \frac{p_1}{\sum_{j=1}^m p_j}, \cdots, \frac{p_m}{\sum_{j=1}^m p_j})$$

ならば

$$\hat{p}_i = \begin{cases} \dfrac{T_x + T_y}{T_x + T_y + r_1 + r_2} \dfrac{x_i + y_i}{T_x\Big(1 + \dfrac{T_y}{\sum_{j=1}^m x_j}\Big)}, & 1 \le i \le m \\ \dfrac{T_x + T_y}{T_x + T_y + r_1 + r_2} \dfrac{x_i}{T_x}, & i > m, \end{cases}$$

である。但し $T_x = \sum_{j=1}^k x_j$, $T_y = \sum_{j=1}^m y_j$.

上記の例の場合に、Asano の方法が適用できる背景について、はっきとした見解を見出せていなかったが、今回、ある程度の見解が得られたので、このことを報告した。

2. 分布の分解 例 1. は、確率分布が Poisson × Multinomial, 例 2. は、確率分布が Negative Multinomial × Multinomial に分解されることが示せる。そして、Multinomial

の部分に対して Asano の状況を取り入れたものであることが示され、本質的に Multinomial の分布に関して Asano の問題が考えられている。

Asano の形とならない例 (1)　$\mathbf{X}$, $\mathbf{Y}$ は 互いに独立で

$$P(x_1,\cdots,x_3 \mid r_1;\mathbf{p}) \propto p_0^{r_1}p_1^{x_1}p_2^{x_2}p_3^{x_3},$$
$$P(y_1,y_2 \mid r_2;\mathbf{p}') \propto \Big(\frac{p_0}{p_0+p_1+p_2}\Big)^{r_2}\Big(\frac{p_1}{p_0+p_1+p_2}\Big)^{y_1}\Big(\frac{p_2}{p_0+p_1+p_2}\Big)^{y_2}.$$

この場合、MLE を書き下してみると、Asano の形にはならない。その原因を、次のように説明することができる。パラメーター変換 $t=p_0+p_1+p_2,\ u_i=p_i/(p_0+p_1+p_2),\ (i=0,1,2)$ を用いることで、通常のサンプリングの場合の尤度は　$L \propto t^{r_1+x_1+x_2}(1-t)^{x_3}u_0^{r_1}u_1^{x_1}u_2^{x_2}.$ となる。ところが、u_0, u_1, u_2 に関して Negative multinomial であって、multinomial ではない。また、t, $1-t$ に関しても Binomial ではない。 つまり、通常のサンプリングが、パラメーター変換後、その分布に multinomial の部分が含まれていないので、Asano の状況が生じていない。

Asano の形とならない例 (2)　$X_1,\cdots,X_l,\cdots,X_m,\cdots,X_k$, $Y_1,\cdots,Y_l$ は すべて独立で,

$$X_i \sim Poisson(\lambda_i),$$
$$Y_i \sim Poisson((\lambda_1+\cdots+\lambda_l+\cdots+\lambda_m)\frac{\lambda_i}{\lambda_1+\cdots+\lambda_l})$$

とする。 次の変数変換を用いる

$$s=\sum_{i=1}^{k}\lambda_i,\quad t=\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{k}\lambda_i},\quad u=\frac{\sum_{i=1}^{l}\lambda_i}{\sum_{i=1}^{m}\lambda_i},\quad \theta_i=\frac{\lambda_i}{\sum_{i=1}^{l}\lambda_i},\ (i\le l),$$
$$v_i=\frac{\lambda_i}{\lambda_{l+1}+\cdots+\lambda_m},\ (l<i\le m),\quad w_i=\frac{\lambda_i}{\lambda_{m+1}+\cdots+\lambda_k},\ (m<i)$$

このとき、同時分布は

$$L \propto s^{T_x+T_y}t^{(\Sigma_{i=1}^{m}x_i)+T_y}(1-t)^{\Sigma_{i=m+1}^{k}x_i}u^{\Sigma_{i=1}^{l}x_i}(1-u)^{\Sigma_{i=l+1}^{m}x_i}$$
$$\times\prod_{i=1}^{l}\theta_i^{x_i+y_i}\prod_{i=l+1}^{m}v_i^{x_i}\prod_{i=m+1}^{k}w_i^{x_i}\times\exp(-s(1+t))$$

であるから、直ちに $\hat\theta_i$, $\hat v_i$, $\hat w_i$, $\hat u$ は求められる。ところが、$\hat t$ は、

$$0=\frac{\sum_{i=1}^{m}x_i+T_y}{t}-\frac{\sum_{i=m+1}^{k}x_i+}{1-t}-\frac{T_x+T_y}{1+t}$$

という 3 次方程式の解で、Asano のような単純な形では書けない。この原因は、多項分布と Poisson 分布に完全に分離しないためである。

結論　以上の議論を踏まえて、Asano の状況が起きるのは、通常のサンプリングによる分布を分解した時に Multinomial の部分が含まれていること、および、通常のサンプリングと pooling incomplete samples の同時分布において、多項分布と他の分布とが完全に分離していることを必要とするであろうとの見地を得ることができた。

参考文献　Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. *Annals of the Institute of Statistical Mathematics* **17**, 1-13.

多項分布の適合度検定統計量の漸近展開における離散項 J_2 の評価についての考察

鹿児島大学理工学研究科 種市信裕
北海道教育大学教育学部 釧路校　関谷祐里

Yarnold (1972) derived an evaluation of discrete term J_2 of asymptotic expansion for lower probability of the distribution of Pearson's X^2 goodness-of-fit statistic for multinomial distribution under null hypothesis. Assylbekov, Ulyanov and Zubov (2008) improved the evaluation given by Yarnold. Discussion of the results are summarized as follows.

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)' \sim M_3(n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$, $\pi_j > 0$, $\sum_{j=1}^3 \pi_j = 1$. Consider simple null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{p}$, where $\boldsymbol{p} = (p_1, p_2, p_3)'$, $p_j > 0$, $\sum_{j=1}^3 p_j = 1$, and p_j, $j = 1, 2, 3$ are fixed. We also consider power-divergence family of statistics.

$$T_a(\boldsymbol{Y}) = \frac{2}{a(a+1)} \sum_{j=1}^{3} Y_j \left[\left(\frac{Y_j}{np_j} \right)^a - 1 \right],\ a \in R,$$

T_0 and T_{-1} are defined as limit.

From here, we state about an approximation of $\Pr(T_a(\boldsymbol{Y}) < c)$ under H_0. Since $Y_1 + Y_2 + Y_3 = n$, we consider random variables $\boldsymbol{X} = (X_1, X_2)'$, where

$$X_j = (Y_j - np_j)/\sqrt{n},\ j = 1, 2.$$

The components of $\boldsymbol{X}$ are concentrated on the lattice

$$L = \{\boldsymbol{x} = (x_1, x_2)' : \boldsymbol{x} = (\boldsymbol{m} - n\tilde{\boldsymbol{p}})/\sqrt{n}, \tilde{\boldsymbol{p}} = (p_1, p_2)', \boldsymbol{m} = (n_1, n_2)'\},$$

where n_j are non-negative integers.

Let

$$B^a = \{(x, y) : T_a(x, y) < c\}$$

and

$$\begin{aligned} T_a(x, y) \quad = \quad & \frac{2}{a(a+1)}(np_1 + \sqrt{n}x)\left[\left(1 + \frac{x}{\sqrt{n}p_1} \right)^a - 1 \right] \\ & + \frac{2}{a(a+1)}(np_2 + \sqrt{n}y)\left[\left(1 + \frac{y}{\sqrt{n}p_2} \right)^a - 1 \right] \\ & + \frac{2}{a(a+1)}(np_3 - \sqrt{n}(x+y))\left[\left(1 - \frac{x+y}{\sqrt{n}p_3} \right)^a - 1 \right]. \end{aligned}$$

Since the set B^a is an extended convex set, then we can evaluate $\Pr(T_a(\boldsymbol{Y}) < c)$ as follows:

$$\Pr(T_a(\boldsymbol{Y}) < c) = \Pr(\boldsymbol{X} \in B^a) = J_1 + J_2 + O(n^{-1}).$$

Siotani and Fujikoshi (1984) and Read (1984) showed that

$$J_1 = J_1(B^a) = \Pr(\chi^2_2 < c) + O(n^{-1}), \tag{1}$$

and

$$J_2 = J_2(B^a) = \frac{(N^a - nV^a)e^{-\frac{c}{2}}}{2\pi n\sqrt{\prod_{j=1}^3 p_j}} + o(1), \tag{2}$$

for arbitrary $a \in R$, where N^a is a number of points from the lattice L lying in B^a and V^a is an area of B^a. Equation (1) means that

$$\Pr(T_a < c) = \Pr(\chi^2_2 < c) + J_2(B^a) + O(n^{-1}).$$

So, the initial problem to find rate of convergence for approximation of $\Pr(T_a < c)$ is reduce to the problem of finding order of $J_2(B^a)$. Since the set B^a is an extended convex set, we can also apply Yarnold's result for $J_2(B^a)$ (Yarnold (1972), p.1557) and obtain

$$J_2(B^a) = O(n^{-\frac{1}{2}}).$$

Paper of Assylbekov, Ulyanov and Zubov (2008) showed a better estimate as follows:

$$J_2(B^a) = O(n^{-\frac{100}{146}}(\log n)^{\frac{315}{146}}), \qquad (a \in R).$$

The proof is devided into two parts.

1. Order of approximation of $J_2(B^a)$ by first summand in (2).
2. Using the results of Huxley (1993),(2003), they obtain the order of $J_2(B^a)$.

For the 1st part, we obtain the following.

Statement 1. We can write $J_2(B^a)$ in the form

$$J_2(B^a) = \frac{d}{n}(N^a - nV^a) + O(n^{-\frac{3}{4}}), \tag{3}$$

where d is a positive constant.

In this announcement, we derived the proof of Statement 1 and considered wether the methodology using for evaluating (3) can be applied for evaluating for asymptotic expansion for goodness-of-fit test statistics under local alternative or for test statistics of another model.

References

[1] Assylbekov Zh. A., Ulyanov V. V. and Zubov V. N. (2008). On approximation of goodness-of-fit statistics for discrete three dimensional data, *Tech. report of Hiroshima University.*

[2] Huxley, M. N. (1993). *Proc. London Mathematical Society.*, **(3)66**, 279–301.

[3] Huxley, M. N. (2003). *Proc. London Mathematical Society.*, **(3)87**, 591–609.

[4] Read, T. R. C. (1984). *Ann. Inst. Statist. Math.,* **36**, 59–69.

[5] Siotani, M. and Fujikoshi Y. (1984). *Hiroshima Math. J.*, **14**, 115–124.

[6] Yarnold, J. K. (1972). *Ann. Math. Statist.*, **43**, 1566–1580.

金属考古学における統計的問題

吉田 知行 (北大・理)

yoshidat@math.sci.hokudai.ac.jp

概要: 鉛同位体法では，古代青銅器に含まれる鉛の同位体比鉛同位体比を測定することによって，青銅器に含まれる鉛の産地や青銅器そのものがどこで作られたかの明らかにする．ここでは，鉛の混合の過程を推定するための数学的方法とその問題点を紹介する．

キーワード： 鉛同位体法，鉛の混合，散布図，椿井大塚山古墳，鉛インゴット

1 はじめに

鉛同位体比の散布図には，いたる所に鉛の凝集，直線状の鉛の並び，長円状の鉛の分布があるく．とくに，直線状の鉛の並びが生じる原因として，鉛の混合の可能性が指摘されてきた (馬淵，久野など)．そもそも散布図内の点の数が多すぎて，全部が鉱山の鉛から来ているとは考えられない．鉛の混合の可能性を考えるのは自然である．ところが鉛の混合があると，単純な鉛同位体法は使えない．

鉛の混合を考えるときは，同位体比より**同位体率** (百分率) の方が扱いやすい．そこでまず鉛同位体比 (Pb206/Pb204 など) から鉛同位体率 Pb204 : Pb206 : Pb207 : Pb208 を計算しておく (**作業 1**)．

講演では，椿井大塚山古墳出土鏡のデータ (馬淵 1996 科研費報告集) を例にして議論を進めた．

質量 (比)	平均	標準偏差	精度仮説
206/204	18.230	0.2109	$\pm 0.0^3 5$
207/206	0.8591	0.0078	$\pm 0.0^4 5$
208/206	2.1263	0.0138	$\pm 0.0^4 5$
204	0.013578	$0.0^2 0083$	$\pm 0.0^4 007$
206	0.247515	$0.0^2 1380$	$\pm 0.0^4 06$
207	0.212630	$0.0^2 0798$	$\pm 0.0^4 12$
208	0.526277	$0.0^2 0566$	$\pm 0.0^4 12$

表 1　椿井大塚山の統計量 ($0^2 = 00$ 等)

鉛の混合を考えるときは，同位体比 (率) の精度が問題になる．ところが鉛同位体比 (率) の持つ次の性質 (表 1 参照) のため，データの分析が困難になる．(1) 存在率のアンバランス．Pb204 だけが 1.35% 程度と少ない．(2) 同位体比の有効桁数の不一致．Pb207/Pb206 が 4 桁，Pb206/Pb204, Pb208/Pb206 は 5 桁．(3) 同位体比測定精度のアンバランス．Pb206/Pb204 の測定誤差は 0.06%，他は 0.02% と言われる．

2 鉛の混合

ここでは鉛を同位体率の 4 次元ベクトルと見なす．2 次元か 3 次元空間に射影した散布図で考える．鉛の混合に関しては次の原理が成り立つ．

基本原理： 鉛 X が鉛 $A, B, \cdots$ の混合比 $p : q : \cdots$ の混合なら，$X = pA + qB + \cdots$．

したがって，3 次元散布図内での平面上の鉛の分布，2 次元あるいは 3 次元散布図内での直線状の鉛の並びは，鉛の混合を示唆する．また鉛インゴット (ここでは単位の重さを持つ純粋の鉛) が使われていたとすれば混合比が簡単な整数比になる．

金属考古学における基本問題： 鉛同位体率のデータから鉛の混合の過程を明らかにせよ．

例． 椿井大塚山古墳の 2, 34, 5 号鏡の鉛同位体率は

$$\boxed{2} = (0.013597, 0.246250, 0.212858, 0.527295)$$
$$\boxed{34} = (0.013619, 0.247108, 0.212908, 0.526365)$$
$$\boxed{15} = (0.013607, 0.246641, 0.212876, 0.526875)$$

$\boxed{2}$ と $\boxed{34}$ を 5:4 で混ぜた鉛 $(5/9)\boxed{2}+(4/9)\boxed{34}$

$(0.\underline{013607},\ 0.\underline{246631},\ 0.\underline{212880},\ 0.\underline{526881})$

は，5 号鏡と区別できない．差は

$(0.0^4 00,\ 0.0^4 10, -0.0^4 04, -0.0^4 06)$.

3　平面状配置の探索

椿井大塚山の鉛を例に説明する．

作業 2-1: 鉛同位体率の 3 次元散布図内の 4 点以上からなる平面状配置をさがす．

作業 2-2: 各平面の生成点をさがす．生成点とは，その平面に属する鉛の源になった 3 つの鉛である．

その分析と確率論的考察から次のことが分かる．

結論 2-1：32 の鉛のうち 27 の鉛がひとつの平面上にある (残差百万分の 20 未満).
結論 2-2 (精度仮説)：鉛同位体比の測定精度は極めて高い．データの最後の桁まで信用できる (表 1).
結論 2-3 (3 つの親仮説)：椿井大塚山古墳出土鏡の鉛は中国の 3 系統の鉛の混合で得られた．

平面状分布を探し出すのに回帰分析の繰り返しを用いた．一般の鉛同位体率の散布図ではこれほど多くががひとつの平面上にあることは考えられない．これは鉛の混合の結果と考えるのが自然であろう．

残差の分析から，鉛同位体率の絶対誤差 (質量依存誤差を除く) は，従来いわれてきた誤差よりも桁違いに小さい．質量依存誤差は，分散，定数項を除く回帰係数，直線関係，平面関係，ユークリッド距離，マハラノビス距離に影響しない．

椿井大塚山の場合，3 つある生成点の有力候補は，中国北方の山東省香奇鉱山 (X)，江南地方の湖南省瑪瑙山鉱山 (Z)，それに中国四川省三星堆遺跡の青銅器の鉛 (Y) である (鉱山は不明)．2 号鏡 (方格規矩四神鏡) と 37 号鏡 (画文帯神獣鏡) にはそれぞれ 11 パーセントと 31 パーセントの三星堆の鉛が含まれている．鉛同位体比の測定結果から，三星堆や殷周の古い鉛が日本列島に入ってきたとの説はこれまでにもあった．

4　直線状配置の探索

鉛の混合の具体的過程を推測したい．

作業 3-1：散布図内の直線状配置の探索.
作業 3-2:直線状配置での混合比.
作業 3-3:整数混合比の探索.

鉛同位体率の散布図で，AXB が直線状の並びなら，X は A と B の混合で得られた可能性がある．

3 つの鉛 AXB の直線度を，直線 AB と X の「距離」として定義する．もし AXB が直線なら，直線度はゼロである．距離としては，ユークリッド距離，マハラノビス距離などの他，$\sin(\angle A+\angle B)$ が考えられる．また，グラスマン多様体 $G(4,2)$ 内の点が集積している所は直線状の並びを示す．このような代数幾何的アプローチも考えられる．

混合比 $p:q$ が整数混合比になっているかは，p/q (または q/p) の連分数展開か，ファレイ数列を使う．

椿井大塚山の場合，2–15–34–36，7–3–25–14–24，5–31–20，2–12–6，1–5–12 等が直線状並びである．

椿井大塚山の鉛の製造に関する仮説:

・直線 37–2–1 はライン D とよばれる (場淵)．この順番に新しくなる．三星堆の青銅器のスクラップの繰り返しによる (新井説).
・残りの鏡 (すべて三角縁神獣鏡) の鉛はライン D 上の鉛と江南の鉛の混合の繰り返しで得られた.
・$2+36 \to 15,35;\ 2+6 \to 12;\ 1+12 \to 5;\ 7+24 \to 3,25,14;\ 5+20 \to 31$.
・2 や 34 などで鉛インゴットが使われた.

5　問題点と文献

この方法の根本的問題点がいくつかある．

・平面状分布をさがすときに使った回帰分析 (強い多重共線性)．さらに親を捜すときの多重比較.
・精度仮説．残差分析により間接的に証明されたが，質量分析器を使えば直接の証明が可能である.

・馬淵久夫『考古学と自然科学』55 (1996)
・吉田知行『季刊邪馬台国』103 (2009)
・吉田知行『情報考古学会誌』(準備中)

Higher order approximations by a two-stage procedure for a negative exponential distribution

磯貝 英一　(新潟大学 自然科学研究科)
小林 加奈
宇野 力　　(秋田大学 教育文化学部)

$X_1, X_2, X_3, \ldots$ は互いに独立で次の同一の指数分布に従う確率変数列とする.

$$f_{\mu,\sigma}(x) = \sigma^{-1} \exp\{-(x-\mu)/\sigma\}\, I(x > \mu)$$

但し, $I(\cdot)$ は定義関数で, 位置母数 $\mu \in (-\infty, \infty)$ と尺度母数 $\sigma \in (0, \infty)$ はともに未知である. 過去の経験等から $\sigma > \sigma_L > 0$ となる尺度母数の下界 σ_L が既知であると仮定する. 任意に与えられた $d \in (0, \infty)$ と $\alpha \in (0, 1)$ に対して, 大きさ n の無作為標本 $X_1, \ldots, X_n$ に基づいて, 区間幅が d, 信頼度が $1-\alpha$ の位置母数 μ に対する信頼区間 I_n を構成したい. すなわち, すべての μ, σ, d, α に対して, $P\{\mu \in I_n\} \geq 1-\alpha$ となるようにしたい.

$$X_{n(1)} = \min\{X_1, \ldots, X_n\}, \quad U_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_{n(1)}) \qquad (n \geq 2)$$

とするとき, 区間幅が d の位置母数 μ に対する信頼区間を $I_n = [\, X_{n(1)} - d,\, X_{n(1)}\,]$ で与える. $n(X_{n(1)} - \mu)/\sigma$ は上で述べた指数分布 $f_{0,1}$ に従うので,

$$P\{\mu \in I_n\} = P\{n(X_{n(1)} - \mu)/\sigma \leq (dn/\sigma)\} = 1 - \exp\{-(dn/\sigma)\}$$

となる. よって, 標本の大きさ n が

$$n \geq \frac{a\sigma}{d} \equiv C \quad \text{ここで, } a = \log(1/\alpha)$$

を満たすとき, すべての μ, σ, d, α に対して, $P\{\mu \in I_n\} \geq 1-\alpha$ となる. C を最適標本数とよぶ. C には未知母数 σ が含まれているので, 逐次手法を用いる.

次のような二段階法を定義する.

$$m = m(d) = \max\left\{m_0,\, \left[\frac{a\sigma_L}{d}\right]^* + 1\right\}$$

但し, $m_0\,(\geq 2)$ はあらかじめ与えた整数, $[x]^*$ は x より小さい最大整数である. 初期標本 $X_1, \ldots, X_m$ によって

$$N = N(d) = \max\left\{m,\, \left[\frac{b_m U_m}{d}\right]^* + 1\right\}$$

を求める. ここで, b_m は自由度 2, $2(m-1)$ のエフ分布 $F_{2,2(m-1)}$ の上側 $100\alpha\,\%$ 点である. 初期標本と第 2 段階の標本 $X_{m+1}, \ldots, X_N$ を合わせて, 大きさ N の標本によって μ の信頼区間を $I_N = [\, X_{N(1)} - d\,,\, X_{N(1)}\,]$ で与える. このとき, すべての μ, σ, d, α に対して, $P\{\mu \in I_N\} \geq 1-\alpha$ (一致性) が成り立つ.

N の定義式において $T = b_m U_m/d$, $S = [T]^* + 1 - T$ とし, $0 \le x \le 1$ に対して, $r_1(x) = P(S \le x) - x$ とする. また,

$$r_d^* = -\int_0^1 r_1(x)dx \quad \text{および} \quad \eta_d^* = r_d^* C^{1/2}$$

とする. このとき, 次の結果を得た.

定理 1. $d \to 0$ のとき, 次が成り立つ.

$$\text{(i)} \quad E(N-C) = \eta_0 + \frac{1}{2} + O(C^{-1/2}) \qquad \text{ここで, } \eta_0 = \frac{a\sigma}{2\sigma_L}$$

$$\text{(ii)} \quad E(N-C) = \eta_0 + \frac{1}{2} + \eta_d^* C^{-1/2} + O(C^{-1})$$

であり,

$$|\eta_d^*| \le \frac{7}{12}\sqrt{\frac{2\sigma_L}{\pi\sigma}} + o(1) \quad \text{ここで,} \quad \frac{7}{12}\sqrt{\frac{2\sigma_L}{\pi\sigma}} < \frac{7}{12}\sqrt{\frac{2}{\pi}} \fallingdotseq 0.46543\,.$$

定理 2. $d \to 0$ のとき, 次が成り立つ.

$$\text{(i)} \quad P\{\mu \in I_N\} = 1 - \alpha + \frac{1}{2}A_1 C^{-1} + O(C^{-3/2})$$

$$\text{(ii)} \quad P\{\mu \in I_N\} = 1 - \alpha + \frac{1}{2}A_1 C^{-1} + A_2 B_d C^{-3/2} + o(C^{-3/2})$$

ここで,

$$A_i = \frac{\alpha a}{(i-1)!}\left(a^2\frac{\sigma}{\sigma_L}\right)^{(i-1)/2} \qquad (i = 1, 2)$$

$$B_d = a^{-1}\left(\frac{\sigma_L}{\sigma}\right)^{1/2}\eta_d^* - \left(\frac{\sigma_L}{\sigma}\right)^{1/2} C^{-1/2} E\{(T-C)S\}$$

であり,

$$|B_d| \le \frac{7}{12}\sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_L}{a\sigma} + \frac{1}{\sqrt{3}} + o(1)\,.$$

参考文献

Aoshima, M., Aoki, M. (2000). Two-stage procedure having exact consistency and second-order properties for the s best selection. *Sequential Anal.* **19**, 115–131.

Aoshima, M., Takada, Y. (2000). Second-order properties of a two-stage procedure for comparing several treatments with a control. *J. Japan Statist. Soc.* **30**, 27–41.

Mukhopadhyay, N., Duggan, W. (1999). On a two-stage procedure having second-order properties with applications. *Ann. Inst. Statist. Math.* **51**, 621–636.

A generalized Bernstein polynomial approach to density estimation

柿沢 佳秀 (北大経済)

確率分布 (密度) 関数, 回帰関数, スペクトル分布 (密度) 関数... の推定は異なる問題であるが互いに共通点があり, カーネル平滑化によるノンパラメトリックな関数推定法は Rosenblatt (1956) と Parzen (1962) 以後めざましい発展を遂げ, カーネル平滑化法の他にスプライン法, ウェーブレット法, 局所多項式法... も議論された. また, 分布の台が $[0,1]$ である場合の確率密度推定法として Vitale (1975) から始まった Bernstein 法も, 近年の文献では種々の関数推定問題に適用されている (関数近似理論では古典的な Weierstrass の近似定理の別証明として古くから知られたものである).

Weierstrass の近似定理: 有界閉区間 $[0,\Delta]$ で連続な関数 $G(x)$ は m 次の Bernstein 多項式

$$B(x;m,G) \equiv \sum_{j=0}^{m} G\Big(\frac{j\Delta}{m}\Big) b_{j,m}\Big(\frac{x}{\Delta}\Big)$$

で一様に近似できる (一般性を失うことなく閉区間 $[0,1]$ を考えるのが普通であろうが, スペクトル推定の場合は周期性と対称性から $[0,\pi]$ で考える慣例があり $\Delta=\pi$ を念頭においている).

$B(x;m,G)$ 及び $B'(x;m,G) = \frac{m}{\Delta}\sum_{j=0}^{m-1}\Big\{G\Big(\frac{(j+1)\Delta}{m}\Big) - G\Big(\frac{j\Delta}{m}\Big)\Big\} b_{j,m-1}\Big(\frac{x}{\Delta}\Big)$ に含まれる未知分布 G を経験分布に置き換えて得られる確率分布 (密度) 推定量は, 非負性が保証され, かつ, 境界バイアスが生じないという特徴をもつ (バイアス, 分散, MSE の漸近公式が Vitale (1975) に議論されたが, 強一致性, 漸近正規性及び MISE の漸近公式は Ghosal (2001), Babu et al. (2002), Kakizawa (2004) による). また, Kakizawa (2006) は $B'(x;m,G)$ を定常時系列のスペクトル推定に適用し, バイアス, 分散, MISE の漸近公式及び漸近正規性を示している.

関数近似理論の枠組みで $B(x;m,G)$ 及び $B'(x;m,G)$ を一般化する試みがなされており, 本報告では Cao (1997) による一般化 (式変形は Rao (2005) に与えられている)

$$\begin{aligned} B(x;m,s_m,G) &\equiv \frac{1}{s_m}\sum_{j=0}^{m}\sum_{r=0}^{s_m-1} G\Big(\frac{(j+r)\Delta}{m+s_m-1}\Big) b_{j,m}\Big(\frac{x}{\Delta}\Big), \quad s_m \in \mathbf{N} \\ &= \frac{1}{s_m}\sum_{r=0}^{s_m-1} G\Big(\frac{r\Delta}{m+s_m-1}\Big) + m\sum_{j=0}^{m-1} c_j^{(m,s_m,G)} \int_0^{x/\Delta} b_{j,m-1}(x)\,dx \quad (1) \end{aligned}$$

を検討した. ここに $c_j^{(m,s_m,G)} \equiv \frac{1}{s_m}\Big\{G\Big(\frac{(j+s_m)\Delta}{m+s_m-1}\Big) - G\Big(\frac{j\Delta}{m+s_m-1}\Big)\Big\}$.

Rao (2005) は確率分布 (密度) 推定において $B(x;m,s_m,G)$ 及び

$$B'(x;m,s_m,G) = \frac{m}{s_m\Delta}\sum_{j=0}^{m-1}\Big\{G\Big(\frac{(j+s_m)\Delta}{m+s_m-1}\Big) - G\Big(\frac{j\Delta}{m+s_m-1}\Big)\Big\} b_{j,m-1}\Big(\frac{x}{\Delta}\Big) \quad (2)$$

を適用している. なお, 密度関数 $F' = f \in \mathcal{C}[0,\Delta]$ の推定量を提案するときには, (2) 式の右辺を近似多項式として採用しておけば $s_m \in \mathbf{N}$ の制限が不要であったことに注意する (Rao (2005) は $s_m \in \mathbf{N}$ を仮定した).

本報告の 1 つの目的は $G(x) = \int_0^x g(t)\,dt$ とした (2) 式の右辺を正線形作用素

$$\frac{m}{s_m\Delta}\sum_{j=0}^{m-1} b_{j,m-1}\Big(\frac{x}{\Delta}\Big)\int_{\frac{j\Delta}{m+s_m-1}}^{\frac{(j+s_m)\Delta}{m+s_m-1}} g(\lambda)\,d\lambda$$

とみて, そのリスケール版として再定義される $\{\mathcal{L}_m(g,x)\}$ (命題 1 の (3)) を連続関数 g に対する近似多項式列として提案することであった:

命題 1. 任意の $g \in \mathcal{C}[0,\Delta]$ に対して

$$\lim_{m\to\infty} \mathcal{L}_m(g,x) = g(x) \text{ uniformly in } x \in [0,\Delta]$$

となるための必要十分条件は $\lim_{m\to\infty} s_m/m = 0$ である. ここに

$$\mathcal{L}_m(g,x) \equiv \frac{m+s_m-1}{s_m\Delta}\sum_{j=0}^{m-1} b_{j,m-1}\Big(\frac{x}{\Delta}\Big)\int_{\frac{j\Delta}{m+s_m-1}}^{\frac{(j+s_m)\Delta}{m+s_m-1}} g(\lambda)\,d\lambda\,, \quad s_m \in (0,\infty) \tag{3}$$

(Cao (1997) は,「任意の $G \in \mathcal{C}[0,\Delta]$ に対して $B(x;m,s_m,G) \to G(x)$ の一様収束性の必要十分条件が $\lim_{m\to\infty} s_m/m = 0$ である」を示しているが, そこでは $s_m \in \mathbf{N}$ を仮定しており, 従って命題 1 は Cao (1997) の単なる微分として得られるものではない).

命題 2 は, (3) を動機として自然に提案される推定量の漸近バイアス項の振る舞いを確率密度 $f \in \mathcal{C}^2[0,1]$ 及びスペクトル密度 $f \in \mathcal{S}^2 \equiv \{$ 原点対称で周期 2π をもつ $\mathcal{C}^2$ 級の関数 $\}$ について示している (なお, $\mathcal{C}^2$ 級の仮定を緩めた場合のオーダーも導いた).

命題 2. $\lim_{m\to\infty} s_m/m = 0$ を仮定する. もし $g \in \mathcal{C}^2[0,\Delta]$ ならば

$$\mathcal{L}_m(g,x) - g(x) - \frac{s_m(\Delta-2x)}{2m}\,g'(x) - \frac{x(\Delta-x)}{2m}\,g''(x) = \mathrm{o}(m^{-1}) + \mathrm{o}(s_m/m)$$

uniformly in $x \in [0,\Delta]$. なお, 境界 $x = 0$ ($x = \Delta$ も同様) では, 2 項分布の確率関数 $b_{j,m-1}(0)$ が $j = 0$ の 1 点に退化するから,

$$\mathcal{L}_m(g,0) - g(0) = \begin{cases} \dfrac{s_m\Delta}{2m}\,g'(0) + \mathrm{O}(s_m^2/m^2)\,, & g'(0) \neq 0 \\ \dfrac{s_m^2\Delta^2}{6m^2}\,g''(0) + \mathrm{o}(s_m^2/m^2)\,, & g'(0) = 0\,. \end{cases}$$

本報告の主要な目的として, 確率密度関数 (独立同一分布設定)/スペクトル密度関数 (定常時系列設定) の正線形作用素 (3) による密度推定量を提案し, $s_m \equiv 1$ に相当する確率密度/スペクトル密度推定量 (Vitale (1975)/Kakizawa (2006)) の諸結果 (漸近分散公式, AMISE の性質, 漸近正規性) を拡張した. また, 確率密度推定量について多項式次数と第 2 パラメータの LSCV 選択及びその数値例も報告した.

On comparisons of univariate normal mean and elements of multivariate normal mean

九州大学数理学研究院　百武弘登

一変量正規分布 $N(\mu_0, \sigma^2)$ の平均 μ と多変量正規分布 $N_k(\boldsymbol{\mu}, \Sigma)$ の平均ベクトル $\boldsymbol{\mu}$ の各成分との比較について考察した。ただし、$\boldsymbol{\mu} = (\mu_1, ..., \mu_k)'$、また共分散行列は一様共分散構造 $\Sigma = \sigma^2\{(1-\rho)I_k + \rho \mathbf{1}_k \mathbf{1}_k'\}$ を仮定する。$N(\mu_0, \sigma^2)$, $N_k(\boldsymbol{\mu}, \Sigma)$ からの標本をそれぞれ、$(x_{01}, ..., x_{0m})$, $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ とする。また、基本統計量として、$\bar{x}_0 = \sum_{j=1}^m x_{0j}/m$, $\bar{\boldsymbol{x}} = \sum_{j'=1}^n \boldsymbol{x}_{j'}/n$, $v_0 = \sum_{j=1}^m (x_{0j} - \bar{x}_0)^2$, $v_1 = \sum_{j'=1}^n (\boldsymbol{x}_{j'} - \bar{\boldsymbol{x}})'(\boldsymbol{x}_{j'} - \bar{\boldsymbol{x}})$, $v_2 = \sum_{j'=1}^n (\boldsymbol{x}_{j'} - \bar{\boldsymbol{x}})'\mathbf{1}k\mathbf{1}_k'(\boldsymbol{x}_{j'} - \bar{\boldsymbol{x}})$ を用いる。そこで、仮説

$$H_0 : \ \boldsymbol{\mu} = \mu_0 \mathbf{1}_k, \quad H_1 : \ \boldsymbol{\mu} \neq \mu_0 \mathbf{1}_k$$

の検定と、$\mu_i - \mu_0$, $(i = 1, ..., k)$ の同時信頼区間の構成を近似的に与えた。後者は、$\rho = 0$ であれば、コントロールとの多重比較となり、Dunnett の方法を直接用いればよい。このような問題は歯科矯正において、咬み合わせの尺度があり、矯正が必要な人の矯正前と矯正後の尺度を矯正の必要のない人の尺度と比較するときなどに適用される。

上記以外に用いる統計量として、相関係数 ρ の推定量が重要である。その最尤推定量 $\hat{\rho}$ は$-1/(k-1) < \rho < 1$ となるような方程式

$$c_{13}\rho^3 + c_{12}\rho^2 + c_{11}\rho + c_{10} = 0$$

の解である。ただし、この方程式の係数は次の通りである。

$$\begin{aligned}
c_{13} &= nk(k-1)^2 v_0, \\
c_{12} &= -(k-1)\{nk(k-2)v_0 - (k-1)mv_1 + mv_2\}, \\
c_{11} &= -(k-1)\{nkv_0 - (2m+nk)v_1\}, \\
c_{10} &= (m+nk)(v_1 - v_2).
\end{aligned}$$

検定に関しては、尤度比検定統計量 λ_1 を導出し、$\lambda_1^{2/(m+nk)}$ の分布がベータ分布で近似できることを示した。m, n はそれぞれ $N(\mu_0, \sigma^2)$, $N_k(\boldsymbol{\mu}, \Sigma)$ からの標本数である。よく知られていることであるが、$-2\log\lambda_1$ が漸近的に χ_k^2 に従うことも利用できるが、標本数がかなり大きくないと近似がよくない。実際、この場合でもシミュレーションでは、標本数が 50 あたりから、状況によっては良い近似であろうという程度であった。

また、別のアプローチとして2次形式統計量 $t_1 = n(\bar{\boldsymbol{x}} - \bar{x}_0\mathbf{1}_k - \mu^*)'S^{-1}(\bar{\boldsymbol{x}} - \bar{x}_0\mathbf{1}_k - \mu^*)$ を用いた検定についても述べた。ただし、$\boldsymbol{\mu}^* = \boldsymbol{\mu} - \mu_0\mathbf{1}_k$ で、S は Σ の適当な推定量である。このような統計量は多変量正規分布の平均の推測によく用いられるものである。しかし、t_1 の分布を導出することは困難であるので、$m = n$ として、次のような分布関数の漸近展開を求めた。

$$F(x) = G_k(x) + \frac{1}{4r(n-1)} \sum_{l=0}^{2} \omega_l G_{k+2l}(x) + O(n^{-2})$$

ただし、$G_q(x)$ は χ_q^2 の分布関数であり、$\omega_0 = r(k-3) - 1,\ \omega_1 = -2\{r(k-1)+1\},\ \omega_2 = r(k+1) + 3$ である。これを用いることによって、仮説検定を近似的に与えることができる。さらに、シェッフェ流に $\boldsymbol{\mu}^* = \boldsymbol{\mu} - \mu_0 \mathbf{1}_k$ の同時信頼区間も与えることができる。

また、同時信頼区間ついては、Dunnett の方法をもとに次のような近似を与えた。

$$\mu_i - \mu_0 \in \bar{x}_i - \bar{x}_0 \pm \hat{d}_\alpha \sqrt{(\frac{1}{m} + \frac{1}{n}) s^2}, \quad i = 1, ..., k$$

ただし、$\nu s^2 = v_0 + \frac{1}{1-\hat{\rho}}(v_1 - \frac{\hat{\rho} v_2}{1 + (k-1)\hat{\rho}})$ で、$\hat{d}_\alpha$ は

$$\int_{-\hat{d}_\alpha}^{\hat{d}_\alpha} \cdots \int_{-\hat{d}_\alpha}^{\hat{d}_\alpha} \frac{\Gamma((k+\nu)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2} |\Lambda(\hat{\rho})|^{1/2}} (1 + \boldsymbol{t}' \Lambda(\hat{\rho})^{-1} \boldsymbol{t}/\nu)^{(k+\nu)/2} d\boldsymbol{t} = 1 - \alpha$$

の解である。ただし、$\Lambda(\rho) = (1/(m+n))\{m(1-\rho) I_k + (m\rho + n) \mathbf{1}_k \mathbf{1}_k'\}$, $\nu = (m-1) + k(n-1)$ である。ここでの被積分関数は、自由度 ν の多変量t分布の密度関数である。$\Lambda(\rho)$ は

$$\boldsymbol{z} = \frac{(\bar{\boldsymbol{x}} - \bar{x}_0 \mathbf{1}_k) - \boldsymbol{\mu}^*}{\sigma \sqrt{1/m + 1/n}}$$

の共分散行列である。

以上で与えた推測法はすべて近似であるため、シミュレーションにより、どの程度近似が良いかを検証した。その結果、尤度比検定におけるベータ分布による近似は標本数が小さいときにおいても、やや保守的ではあるが概ね良好であることがわかった。また、同時信頼区間の近似も標本数が小さい場合でも良い近似となっていた。さらに、ρ の推定の観点からも予想できたが、両方のシミュレーションとも、k が大きいほうが良い近似となっていることも検証された。

これらの結果は次の論文にまとめている。

T. Furukatsu, D. Shimamoto and H. Hyakutake, On comparison of univariate normal mean and elements of multivariate normal mean, Advances Appl. Stat. 2010 (to appear).

ステップワイズ法による確率ベクトルの成分間の独立性の同時検定

東京理科大学大学院 理学研究科　髙橋　翔
東京理科大学 理学部　西山 貴弘
東京理科大学 理学部　瀬尾　隆
東海大学 総合経営学部　今田 恒久

多変量正規母集団のもとで, 確率ベクトルの成分間の独立性の検定について考える.

$\boldsymbol{x} = (x_1, \boldsymbol{x}'_{(2)})'$ を平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 Σ の p 変量正規母集団からの確率ベクトルとする. ここで, $\boldsymbol{x}_{(2)} = (x_2, \ldots, x_p)'$ とする. 本報告では, x_1 と $\boldsymbol{x}_{(2)}$ のどの成分間に相関があるかの検定を考え, 独立性の同時検定を行うための閉検定手順 (Marcus, Peritz and Gabriel (1976)) に基づくステップダウン式多重比較法を提案した. さらに, ステップアップ式多重比較法 (Dunnett and Tmahane (1992)) による独立性の同時検定手法を提案し, それぞれの検定手法の検出力をモンテカルロ・シミュレーションにより数値的に比較を行った.

数の集合 $\{2, \ldots, p\}$ の基数 q の部分集合全体の成す族を M_q とする. $m = \{\ell_1, \ldots, \ell_q\} \in M_q$ ($\ell_1 < \cdots < \ell_q$) に対し, $q+1$ 次元の確率ベクトル $(x_1, x_{\ell_1}, \ldots, x_{\ell_q})'$ の分散共分散行列を $\Sigma^{(q,m)}$ とおき

$$\Sigma^{(q,m)} = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12}^{(q,m)'} \\ \boldsymbol{\sigma}_{12}^{(q,m)} & \Sigma_{22}^{(q,m)} \end{bmatrix}$$

と分割する. このとき, x_1 と $\boldsymbol{x}_{(2)}^{(q,m)} = (x_{\ell_1}, \ldots, x_{\ell_q})'$ の独立性の仮説

$$H_0^{(q,m)} : \boldsymbol{\sigma}_{12}^{(q,m)} = \boldsymbol{0} \quad \text{vs.} \quad H_1^{(q,m)} : \boldsymbol{\sigma}_{12}^{(q,m)} \neq \boldsymbol{0}$$

を考える. $N_p(\boldsymbol{\mu}, \Sigma)$ からの N 個のランダム標本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ に対し, $\overline{\boldsymbol{x}} = N^{-1}\sum_{i=1}^N \boldsymbol{x}_i$, $A = \sum_{i=1}^N (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$ とおき, A の部分行列 $A^{(q,m)}$ の分割を次のように与える.

$$A^{(q,m)} = \begin{bmatrix} a_{11} & \boldsymbol{a}_{12}^{(q,m)'} \\ \boldsymbol{a}_{12}^{(q,m)} & A_{22}^{(q,m)} \end{bmatrix}.$$

このとき, 仮説 $H_0^{(q,m)}$ に対する尤度比は

$$\Lambda^{(q,m)} = \frac{(\det A^{(q,m)})^{\frac{N}{2}}}{a_{11}^{\frac{N}{2}} (\det A_{22}^{(q,m)})^{\frac{N}{2}}}$$

で与えられ, $H_0^{(q,m)}$ のもとで修正尤度比検定統計量 $-2\tau \log \Lambda^{(q,m)}$ が漸近的に自由度 q の χ^2 分布に従うことを理論的に示した. ただし, $\tau = 1 - (q+4)/2N$ である. さらに, χ^2 分布への近似精度をモンテカルロ・シミュレーションにより数値的に確認した. また, F_q を $H_0^{(q,m)}$ 全体が成す集合とし, $F = \cup_{q=1}^{p-1} F_q$ とおくと, 族 F は閉じている. このとき, F に対する同時検定を行うための閉検定手順に基づくステップダウン式多重比較法の構築を考える. 本報告では, 修正尤度比統計量 $-2\tau \log \Lambda^{(q,m)}$ を用いて, 以下の手順で構成される手法を提案した.

Step 1. $H_0^{(p-1,m)}$ を検定する.

Case 1. $-2\tau \log \Lambda^{(p-1,m)} > \chi^2_{p-1}(\alpha)$ ならば, $H_0^{(p-1,m)}$ を棄却し, Step 2 へ進む.

Case 2. $-2\tau \log \Lambda^{(p-1,m)} \leq \chi^2_{p-1}(\alpha)$ ならば, F に含まれる全ての仮説を保留し, 検定を終了する.

Step 2. F_{p-2} に含まれる全ての仮説 $H_0^{(p-2,m)}$ を検定する.

Case 1. $-2\tau \log \Lambda^{(p-2,m)} > \chi^2_{p-2}(\alpha)$ ならば, $H_0^{(p-2,m)}$ を棄却する.

Case 2. $-2\tau \log \Lambda^{(p-2,m)} \leq \chi^2_{p-2}(\alpha)$ ならば, $H_0^{(p-2,m)}$ および $H_0^{(p-2,m)}$ から誘導される全ての仮説を保留する.

$\cup_{q=1}^{p-3} F_q$ に含まれる全ての仮説が保留されるならば, 検定を終了し, そうでなければ Step 3 へ進む.

Step 3. Step 2 で保留にならなかった F_{p-3} に含まれる仮説を検定する.

同様にして最大 Step $p-1$ まで検定を続ける.

次に, 以下の独立性の仮説の検定を行うために, ステップアップ式多重比較法による同時検定手法を提案した.

$$H_{1i} : \sigma_{1i} = 0 \quad \text{vs.} \quad K_{1i} : \sigma_{1i} \neq 0, \quad i = 2, 3, \ldots, p.$$

H_{1i} に対する修正尤度比検定統計量は先ほどと同様に $-2\eta \log \Lambda_{1i},\ \eta = 1 - 5/2N$ と導出でき, 漸近的に自由度 1 の χ^2 分布に従う. ただし, 尤度比 Λ_{1i} は A の部分行列 A_{1i} を用いて次のように与えられる.

$$\Lambda_{1i} = \frac{|A_{1i}|^{\frac{N}{2}}}{a_{11}^{\frac{N}{2}} a_{ii}^{\frac{N}{2}}}, \quad A_{1i} = \begin{bmatrix} a_{11} & a_{1i} \\ a_{i1} & a_{ii} \end{bmatrix}.$$

$L_{1i} \equiv -2\eta \log \Lambda_{1i}$ とし, 標本に基づき計算した $L_{12}, L_{13}, \ldots, L_{1p}$ を大きさの順に並べたものを

$$L_{12}^{(2)} \leq L_{13}^{(3)} \leq \cdots \leq L_{1p}^{(p)}$$

とする. また, $L_{12}^{(2)}, L_{13}^{(3)}, \ldots, L_{1p}^{(p)}$ に対応する仮説をそれぞれ $H_{12}^{(2)}, H_{13}^{(3)}, \ldots, H_{1p}^{(p)}$ と表す.
$H_{12}, H_{13}, \ldots, H_{1p}$ がすべて正しいとき, 各 $m = 2, 3, \ldots, p$ に対して

$$\Pr\left\{(L_{12}, L_{13}, \ldots, L_{1m}) \leq (c_2, c_3, \ldots, c_m)\right\} = 1 - \alpha, \quad c_2 \leq c_3 \leq \cdots \leq c_p$$

を満たすように棄却限界値 $c_2, c_3, \ldots, c_p$ を決定する. ただし

$$(L_{12}, L_{13}, \ldots, L_{1m}) \leq (c_2, c_3, \ldots, c_m)$$

は $L_{12}^{(2)} \leq c_2, L_{13}^{(3)} \leq c_3, \ldots, L_{1m}^{(m)} \leq c_m$ を表す. このとき, ステップアップ式多重比較法による独立性の同時検定手法を以下の手順で構築した.

Step 1. $H_{12}^{(2)}$ を検定する.

Case 1. $L_{12}^{(2)} > c_2$ ならば, $H_{12}^{(2)}, H_{13}^{(3)}, \ldots, H_{1p}^{(p)}$ 全てを棄却し, 検定を終了する.

Case 2. $L_{12}^{(2)} \leq c_2$ ならば, $H_{12}^{(2)}$ を保留し, Step 2 へ進む.

Step 2. $H_{13}^{(3)}$ を検定する.

Case 1. $L_{13}^{(3)} > c_3$ ならば, $H_{13}^{(3)}, H_{14}^{(4)}, \ldots, H_{1p}^{(p)}$ 全てを棄却し, 検定を終了する.

Case 2. $L_{13}^{(3)} \leq c_3$ ならば, $H_{13}^{(3)}$ を保留し, Step 3 へ進む.

Step 3. $H_{1(4)}$ を検定する.

同様にして最大 Step $p-1$ まで検定を続ける.

本報告では, 分散共分散行列に構造を仮定し, 提案した閉検定手順に基づくステップダウン式多重比較法とステップアップ式多重比較法の検出力を, いくつかのパラメータについてモンテカルロ・シミュレーションにより数値的に比較を行った.

参考文献

[1] Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, **87**, 162–170.

[2] Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

平均ベクトル間の多重比較法に対する同時信頼区間とその保守性

東京理科大・理　西山　貴弘
東京理科大・理　瀬尾　隆

多変量正規母集団における平均ベクトルに関する多重比較法についての同時信頼区間を考える．多重比較法には代表的な比較として対比較と対照比較があるが，平均ベクトル間のすべてのペアの差（対比較）に関する保守的な近似同時信頼区間を構成する手法の一つとして，多変量　k y-K　法　F j k　が提案されている．これに関連して，対照比較の場合における保守的な手法が　によって提案されている．本報告ではこれらの手法について紹介し，特に対照比較の場合について，母集団数が　個の場合の保守性に関する不等式を与え，モンテカルロ・シミュレーションによって保守性の程度を数値的に与えた．

一般に，k 個の多変量正規母集団において $\boldsymbol{\mu}_i$ を i 番目の母集団からの平均ベクトルとし，$\boldsymbol{\mu}_i$ を並べた $p \times k$ 未知行列を $\boldsymbol{M}$　$[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k]$ とする．さらに，$\widehat{\boldsymbol{M}}$　$[\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k]$ を $\boldsymbol{M}$ の推定量とし，　$\boldsymbol{X}$ $\sim N_{kp}\ \boldsymbol{0}, \boldsymbol{V} \otimes \boldsymbol{\Sigma}$ とする．ここで，$\boldsymbol{X}$　$[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]$　$\widehat{\boldsymbol{M}} - \boldsymbol{M}$，$\boldsymbol{V}$　$[v_{ij}]$ は $k \times k$ 既知行列，$\boldsymbol{\Sigma}$ は $p \times p$ 未知行列とする．また，$\boldsymbol{S}$ は $\nu\boldsymbol{S}$ が $\widehat{\boldsymbol{M}}$ と独立で，$\nu\boldsymbol{S} \sim \mathrm{W}_p\ \boldsymbol{\Sigma}, \nu$ となるような $\boldsymbol{\Sigma}$ の不偏推定量とする．

このとき，一般に平均ベクトル間の多重比較法における同時信頼区間は次のような形で表現される．

$$\boldsymbol{a}'\boldsymbol{M}\boldsymbol{b} \in \left[\ \boldsymbol{a}'\widehat{\boldsymbol{M}}\boldsymbol{b} \pm t\ \boldsymbol{b}'\boldsymbol{V}\boldsymbol{b}\ ^{1/2}\ \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}\ ^{1/2}\right], \quad \forall \boldsymbol{a} \in \mathbb{R}^p - \{\boldsymbol{0}\},\ \forall \boldsymbol{b} \in \mathbb{B}.$$

ここで $\mathbb{B}$ は k 次元空間における部分集合であり t^2 は次のような $T^2_{\max}$ 型統計量

$$T^2_{\max}\quad \underset{\boldsymbol{b}\in\mathbb{B}}{\mathrm{x}} \left\{ \frac{\boldsymbol{X}\boldsymbol{b}\ '\boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{V}\boldsymbol{b}} \right\}$$

の上側　$\alpha\%$ 点である　このとき　の被覆確率（　y）は正確に　$-\alpha$ になる

本報告では対比較と対照比較について考えている．特に k 番目の母集団をコントロールとする対照比較について考えると，$\boldsymbol{e}_i$ を i 番目が　である k 次元単位ベクトルとし，

$$\mathbb{B}\quad \mathbb{D} \equiv \{\boldsymbol{d} \in \mathbb{R}^k\quad \boldsymbol{d}\quad \boldsymbol{e}_i - \boldsymbol{e}_k,\ i\quad, \ldots, k-\ \}$$

とおくと，対照比較に対する同時信頼区間は次のように表される．

$$\boldsymbol{a}'\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_k\ \in \left[\ \boldsymbol{a}'\ \widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_k\ \pm t_{\mathrm{c}}\ d_{ik}\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}\ ^{1/2}\right], \forall \boldsymbol{a} \in \mathbb{R}^p - \{\boldsymbol{0}\},\ i\quad, \ldots, k-\ .$$

ここで，d_{ik}　$v_{ii} -\ v_{ik}$　v_{kk} であり，t^2_{c} は次のような $T^2_{\max\cdot\mathrm{c}}$ 統計量

$$T^2_{\max\cdot\mathrm{c}}\quad \underset{\boldsymbol{d}\in\mathbb{D}}{\mathrm{x}} \left\{ \frac{\boldsymbol{X}\boldsymbol{d}\ '\boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{V}\boldsymbol{d}} \right\}$$

$$\underset{i=1,\ldots,k-1}{\mathrm{x}} \left\{\ \boldsymbol{x}_i - \boldsymbol{x}_k\ '\ d_{ik}\boldsymbol{S}^{\ -1}\ \boldsymbol{x}_i - \boldsymbol{x}_k\ \right\}$$

の上側　$\alpha\%$ 点である．

実際にこの同時信頼区間を構成するためには，$T^2_{\max\cdot\mathrm{c}}$ 統計量の上側　$\alpha\%$ 点の値が必要になるが，一般に正確な値を求めることは非常に困難である．そのため，保守的な近似同時信頼区間を構成する手法が　（　）によって提案されており，この手法による近似同時信頼区間

$$\boldsymbol{a}'\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_k\ \in \left[\ \boldsymbol{a}'\ \widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_k\ \pm t_{\mathrm{c}\cdot V_1}\quad d_{ik}\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}\ ^{1/2}\right], \forall \boldsymbol{a} \in \mathbb{R}^p - \{\boldsymbol{0}\},\ i\quad, \ldots, k-$$

が保守的である，すなわち

$$\left\{\boldsymbol{a}'\ \boldsymbol{\mu}_i-\boldsymbol{\mu}_k\ \in\left[\boldsymbol{a}'\ \widehat{\boldsymbol{\mu}}_i-\widehat{\boldsymbol{\mu}}_k\ \pm t_{\mathrm{c}\cdot V_1}\left(d_{ik}\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}\right)^{1/2}\right]\right.$$
$$\left.\forall\boldsymbol{a}\in\mathbb{R}^p-\{\boldsymbol{0}\},\ i\quad,\ldots,k-\ \right\}\geq\ -\alpha$$

であると予想されている．ただし，$t^2_{\mathrm{c}\cdot V_1}$ は $\boldsymbol{V}\quad\boldsymbol{V}_1$ のときの $T^2_{\max\cdot\mathrm{c}}$ 統計量の上側 $\alpha\%$ 点であり，$\boldsymbol{V}_1$ は全ての $i,\ j\quad\leq i<j\leq k-\quad$ に対して $d_{ij}\quad d_{ik}\quad d_{jk}$ を満たす行列である．

この予想は，　　によって k　　の場合が，N　y　　によって k　　の場合が証明されている．また，この手法の保守性の程度が k　　，それぞれの場合について N　y　　8，N　y　　で議論されている．しかし，$k\geq$　では未解決な予想として残されており，本報告ではこれらの証明のアイデアを拡張することにより，k　　の場合の保守性，および保守性の程度に関する以下の定理を与えた．

定理． k　　のとき 任意の正定値行列 $\boldsymbol{V}$ について次の不等式が成り立つ

$$-\alpha\quad Q\ t^*_{\mathrm{c}},\boldsymbol{V}_1,\mathbb{D}\ \leq Q\ t^*_{\mathrm{c}},\boldsymbol{V},\mathbb{D}\ <Q\ t^*_{\mathrm{c}},\boldsymbol{V}_2,\mathbb{D}\ .$$

ここで $Q\ t^*_{\mathrm{c}},\boldsymbol{V},\mathbb{D}$　　$\{\ \boldsymbol{X}\boldsymbol{d}\ '\ \nu\boldsymbol{S}^{\ -1}\ \boldsymbol{X}\boldsymbol{d}\ \leq t^*_{\mathrm{c}}\ \boldsymbol{d}'\boldsymbol{V}\boldsymbol{d}\ \mathrm{f}\quad\mathrm{y}\ \boldsymbol{d}\in\mathbb{D}\}$ であり $t^*_c\quad t^2_{\mathrm{c}\cdot V_1}/\nu$，また $\boldsymbol{V}_1$ はすべての $i,j\quad\leq i<j\leq\quad$ に対して $d_{ij}\quad d_{i5}\quad d_{j5}$ を満たす行列　$\boldsymbol{V}_2$ はすべての $i,j\quad\leq i<j\leq\quad$ に対して $\sqrt{d_{ij}}\quad|\sqrt{d_{i5}}-\sqrt{d_{j5}}|$ を満たす行列である．

さらに本報告では，いくつかのパラメータの場合について，$T^2_{\max\cdot\mathrm{c}}$ 統計量の上側　$\alpha\%$ 点および被覆確率をモンテカルロ・シミュレーションにより求めた．また，得られた定理に基づき，k の場合の対照比較に対する同時信頼法の保守性の程度を数値的に与えた．

参考文献

[ ]N　y　　O　　fi　　f　　w　　*SUT Journal of Mathematics* **43**　–

[ ]　　fi　　f　　f　　*Hiroshima Mathematical Journal* **23**　8 –

[ ]　　F jk　Y　　A　　k y　j　　f　　*Journal of the American Statistical Association* **89**　–

[ ]　　N　y　　8　O　　fi　　f　　*Journal of Statistical Planning and Inference* **138**　8–

β-Lasso推定による頑健なモデル選択

小林　裕子 (筑波大学大学院・数理物質科学研究科)
矢田　和善 (筑波大学大学院・数理物質科学研究科)
青嶋　誠　(筑波大学大学院・数理物質科学研究科)

1. はじめに

本講演では，異常値に対して頑健なモデルの推定と，新しいモデル選択の基準を提案した．データを生み出す真の分布とモデルとの近さを，Basu et al. (1998) が提案した β-ダイバージェンスで測った．真の分布は異常値と潜在分布の混合分布であると考え，β-ダイバージェンスに基づいて，異常値に対して頑健に潜在分布を推定した．潜在分布には多次元混合正規分布モデルを仮定し，高次元のパラメータを効率よく推定するために，β-尤度最大化と Lasso を融合させた，β-Lasso 推定を新しく提案した．構築したモデルを予測の意味で評価するために，β-Lasso 推定の偏りを補正した新しいモデル選択基準を提案した．提案したモデル選択基準の性能が数値的に確認された．

2. 異常値に対して頑健な推定量

d 次元データ $\boldsymbol{x}$ を生成する真の分布の密度関数を $g(\boldsymbol{x})$ とする．未知の $g(\boldsymbol{x})$ を近似するモデルをクラス数 k の混合正規分布 $f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)$ とするとき，$f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)$ の $g(\boldsymbol{x})$ に対する近さを β-ダイバージェンス

$$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)) = \frac{1}{\beta(1+\beta)} \int_{\boldsymbol{R}^d} g(\boldsymbol{x})^{1+\beta} d\boldsymbol{x} \\ \int_{\boldsymbol{R}^d} \frac{f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)^\beta}{\beta} g(\boldsymbol{x}) d\boldsymbol{x} + \int_{\boldsymbol{R}^d} \frac{f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)^{1+\beta}}{1+\beta} d\boldsymbol{x} \quad (\beta > 0)$$

で測る．$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_k))$ が小さいほど，モデル $f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)$ は真の $g(\boldsymbol{x})$ に近いと考えられる．いま，$c_\beta(\boldsymbol{\theta}_k) = \int_{\boldsymbol{R}^d} (\beta+1)\ {}^1 f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)^{\beta+1} d\boldsymbol{x}$ とおき，β-尤度を

$$\ell_\beta(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)) = \frac{1}{n} \sum_{\alpha=1}^{n} \frac{f(\boldsymbol{x}_\alpha, k|\boldsymbol{\theta}_k)^\beta \quad 1}{\beta} \quad c_\beta(\boldsymbol{\theta}_k) \quad (\beta > 0)$$

で定義する．そのとき，$\hat{\boldsymbol{\theta}}_{k,\beta} = \underset{\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k}{\operatorname{argmax}} \ell_\beta(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k))$ を用いて構築したモデル $f(\boldsymbol{x}, k|\hat{\boldsymbol{\theta}}_{k,\beta})$ を，データ数 n が大きいとき漸近的に最適なモデルと考える．

真の分布は，異常値 $\boldsymbol{x}_\star$ が開領域 $\boldsymbol{R}_o^d$ に確率 $\tau \in (0,1)$ で混入すると仮定する．このとき，$\int_{\boldsymbol{R}_o^d} \psi(\boldsymbol{x}) d\boldsymbol{x} = 1, \quad \sup_{\boldsymbol{x} \in \boldsymbol{R}_o^d} \psi(\boldsymbol{x}) < \infty$ をみたすような $\psi(\boldsymbol{x})$ を $\boldsymbol{x}_\star$ の密度関数とし，$f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ をクラス数 $k_\star$ とパラメータ $\boldsymbol{\theta}_\star$ をもつ混合正規分布によって表現される潜在分布の密度関数とする．$\int_{\boldsymbol{R}_o^d} f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star) d\boldsymbol{x} = \delta\ (> 0)$ とするとき，真

の分布の密度関数は，十分小さいτとδによって$g(\boldsymbol{x}) = (1 \quad \tau)f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star) + \tau\psi(\boldsymbol{x})$で表されるとする．このとき，混合比率$\xi_j,\ j = 1, ..., k$に関するLasso型罰則付き$\beta$-尤度として

$$\ell_{\beta,\lambda}(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)) = \ell_\beta(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)) \quad \lambda p(\boldsymbol{\theta}_k), \quad p(\boldsymbol{\theta}_k) = \sum_{j=1}^{k\ \ 1} \xi_j$$

を定義し，$\hat{\boldsymbol{\theta}}_{k,\beta,\lambda} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_k}{\operatorname{argmax}}\, \ell_{\beta,\lambda}(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k))$なる推定量を考えた．ここで，$f(\boldsymbol{x}, k|\boldsymbol{\theta}_k)$について適当な正則条件を課すと，次の結果が導かれた．

定理． $n \to \infty$，$\lambda \to 0$のとき，確率1で$\ell_{\beta,\lambda}(\boldsymbol{\theta}_k; f(\boldsymbol{x}, k|\boldsymbol{\theta}_k))$の局所最大解$\hat{\boldsymbol{\theta}}_{k,\beta,\lambda}$が存在し，次が成り立つ．

$$\begin{aligned} \| \hat{\boldsymbol{\theta}}_{k,\beta,\lambda} \quad \hat{\boldsymbol{\theta}}_{k,\beta} \| &= O_p(\lambda), \\ \| \hat{\boldsymbol{\theta}}_{k,\beta,\lambda} \quad \boldsymbol{\theta}_{k,\beta} \| &= O_p(n^{\ 1/2}) + O_p(\lambda) \end{aligned}$$

3．β-ダイバージェンス最小化に基づくモデル選択

β-ダイバージェンスを用いて$g(\boldsymbol{x})$とモデル$f(\boldsymbol{x}, k|\boldsymbol{\theta}_{k,\beta})$の近さを測るとき，次の結果が導かれた．

定理． クラス数kが$k \quad k_\star$のとき，$\tau \to 0,\ \delta \to 0$のもとで次が成り立つ．

$$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{k,\beta})) = O(\tau^2) + O(\tau^{1+\beta}) + O(\delta^{1+\beta})$$

[β-Lasso モデル選択基準]

$\tau \to 0$，$\delta \to 0$のとき，次で定義される$\mathrm{IC}_{\beta,\lambda}(k)$が小さいモデルを，予測の意味で潜在分布に近いモデルとして選択する．

$$\mathrm{IC}_{\beta,\lambda}(k) = \quad 2n\ell_\beta(\hat{\boldsymbol{\theta}}_{k,\beta,\lambda}; f(\boldsymbol{x}, k|\hat{\boldsymbol{\theta}}_{k,\beta,\lambda})) + 2b_\beta(k, \hat{\boldsymbol{\theta}}_{k,\beta,\lambda})$$

いま，バイアス項$b_\beta(k, \hat{\boldsymbol{\theta}}_{k,\beta,\lambda})$を$b_\beta(k, \hat{\boldsymbol{\theta}}_{k,\beta,\lambda}) = p_k$とおく．$t_d = d(d+3)/2 + 1$とおく．そのとき，次の定理が導かれた．

定理． $\tau^2/t_d = o(n^{\ 1}),\ \tau^{1+\beta}/t_d = o(n^{\ 1}),\ \delta^{1+\beta}/t_d = o(n^{\ 1})$のもとで次が成り立つ．

$$k_\star = \underset{k}{\operatorname{argmin}} \lim_{n\to\infty} \frac{1}{n} E_G[\mathrm{IC}_{\beta,\lambda}(k)]$$

βの値は，この定理の収束条件を満たす範囲で選択する．そのとき，潜在分布に近いモデルを$\mathrm{IC}_{\beta,\lambda}(k)$によって平均的に選択することができた．

高次元データにおける幾つかの検定統計量の漸近分布について

藤本 翔太 [1], 狩野 裕 [1], Muni S. Srivastava[2]
[1] 大阪大学基礎工学研究科
[2]Department of Statistics, University of Toronto

正規分布に基づく多変量 1 標本問題を考える．すなわち，$\boldsymbol{X}_p^{(1)}, \cdots, \boldsymbol{X}_p^{(n)}$ が独立に同一の多変量正規分布 $N_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ に従うとする．ここで，$\boldsymbol{\Sigma}_p$ は任意の p に対して正定値行列であることを仮定する．いま，次の検定問題を考える：$H_0 : \boldsymbol{\mu}_p = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\mu}_p \neq \boldsymbol{0}$. この検定問題に対する伝統的な方法は，Hotelling の T^2 統計量を用いる方法であるが，標本サイズ n が変数の次元 p よりも小さな場合は，標本共分散行列が特異になり，定義されない．Dempster (1958, 1960) は T^2 統計量における標本共分散行列 $\boldsymbol{S}_{n,p}$ をそのトレース $\mathrm{tr}\boldsymbol{S}_{n,p}$ で置き換えた統計量を提案した：$T_D^2 = n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p}/\mathrm{tr}\boldsymbol{S}_{n,p}$. ここに，$\bar{\boldsymbol{X}}_{n,p}$ は標本平均である．Bai and Saranadasa (1996), Fujikoshi (2004), Srivastava (2007) は T_D^2 を用いた検定を行うために，(n, p) が共に大きくなるという高次元漸近理論の枠組みでの漸近分布を導出している．Bai and Saranadasa (1996) は T_D^2 の漸近正規性を条件

$$\max_{1\leq i\leq p} \lambda_i = o\left(\sqrt{\mathrm{tr}\boldsymbol{\Sigma}_p^2}\right) \tag{1}$$

の下で示している．ここに，λ_i $(i = 1, 2, \cdots, p)$ は $\boldsymbol{\Sigma}_p$ の固有値である．また，Bai and Saranadasa (1996) では別の統計量 $T_{BS}^2 = n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p} - \mathrm{tr}\boldsymbol{S}_{n,p}$. も提案しており，その漸近正規性も同じ条件の下で示している．近年， Fujikoshi (2004) や Srivastava (2007) によって T_D^2 と T_{BS}^2 の漸近正規性が

$$0 < \lim_{p\to\infty} \frac{\mathrm{tr}\boldsymbol{\Sigma}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4). \tag{2}$$

の下でも導出されている．なお，T_D^2 と T_{BS}^2 はデータ行列の直交行列による変換 $\boldsymbol{X}_p^{(i)} \to c\Gamma\boldsymbol{X}_p^{(i)}$ に関して不変な統計量である．ここに，$c > 0$ であり，$\Gamma^T\Gamma = I_p$ である．また，I_p は $p \times p$ の単位行列を表す．しかしながら，これらの統計量はデータの単位変換 $\boldsymbol{X}_p^{(i)} \to D\boldsymbol{X}_p^{(i)}$ に関して不変ではない．ここに，$D = \mathrm{diag}(d_1, \cdots d_p), d_i > 0$ である．そこで，Srivastava and Du (2008) はデータの単位変換に関して不変な統計量 $T_S^2 = n\bar{\boldsymbol{X}}_{n,p}^T(\mathrm{diag}\boldsymbol{S}_{n,p})^{-1}\bar{\boldsymbol{X}}_{n,p}$ を提案している．ここに，$\mathrm{diag}\boldsymbol{A}$ は行列 $\boldsymbol{A}$ の対角要素だけを残し，非対角要素を全て 0 とした行列である．さらに，高次元漸近理論の枠組みにおける漸近正規性を次の条件の下で示している：

$$0 < \lim_{p\to\infty} \frac{\mathrm{tr}\mathcal{R}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4), \tag{3}$$

ここに，$\mathcal{R}_p$ は $\boldsymbol{\Sigma}_p$ の相関行列である．

条件 (1) または (2) が満たされるならば，$p \to \infty$ のとき $r := (\mathrm{tr}\boldsymbol{\Sigma}_p)^2/\mathrm{tr}\boldsymbol{\Sigma}_p^2 \to \infty$ である．一方で，Cauchy-schwarz の不等式から $r \leq p$ であり，等号成立はある定数 $c > 0$ に対して $\boldsymbol{\Sigma}_p = cI_p$ であるときかつそのときに限られることがわかる．したがって，これらの条件は $\boldsymbol{\Sigma}_p$ が単位行列の定数倍に近い場合でしか成立ちそうにない．同じことが条件 (3) に対してもいえる．

具体的に例を見てみると，条件 (1), (2), (3) は $\boldsymbol{\Sigma}_p$ が例えば次のような場合は成立つ：

(a) Sphericity Model : $\boldsymbol{\Sigma}_p = aI_p$. ここに，$a > 0$.
(b) Autoregressive Model : $\boldsymbol{\Sigma}_p = a(\rho^{|i-j|})$. ここに，$a > 0$, $0 < \rho < 1$.

これらの条件が Autoregressive Model の場合に満たされることは，Szegö の定理 (e.g., Grenander and Szegö 1958) を使って示される．一方，条件 (1), (2), (3) は次の場合に満たされない．

(c) Compound Symmetry Structure : $\mathbf{\Sigma}_p = a(1-\rho)I_p + a\rho\mathbf{1}_p\mathbf{1}_p^T$. ここに，$a > 0,\ 0 < \rho < 1$ であり，$\mathbf{1}_p$ は要素すべてが 1 の p 次元列ベクトルを表す.

条件が満たされないことは，Compound Symmetry Structure の固有値が $\lambda_1 = a(1-\rho+p\rho),\ \lambda_2 = \cdots = \lambda_p = a(1-\rho)$ であることから容易に確認される. これは次のモデルの特別な場合である：

(d) Spiked Model (Johnstone, 2001) : $\lambda_i = c_i p^{\alpha_i}\ \ (i = 1, 2, \cdots, \ell),\ \ \lambda_j = c_j\ \ (j = \ell+1, \cdots, p)$. ここに，$\lambda_1 \geq \cdots \geq \lambda_p$ は $\mathbf{\Sigma}_p$ (または $\boldsymbol{\mathcal{R}}_p$) の固有値であり，$c_1 \geq \cdots \geq c_p > 0,\ \alpha_i \geq 0,\ m < p$ はすべて p に依存しないものする.

Spiked Model において，条件 (1) は $\alpha_1 < 1/2$ のとき満たされ，条件 (2) は $\alpha_1 \leq 1/4$ のとき満たされる. しかし，どちらの条件も $\alpha_1 \geq 1/2$ のときは満たされない. 以上の考察から，先行研究の条件は，$\mathbf{\Sigma}_p$(または $\boldsymbol{\mathcal{R}}_p$) の幾つかの固有値が大きく突出した場合に満たされないことがわかる.

次で提案する $\mathbf{\Sigma}_p$(または $\boldsymbol{\mathcal{R}}_p$) に関する新しい条件は，上記のすべての例を含んでいる：$\delta_1 \geq \delta_2 \geq \delta_k$ $(k = 3, 4, \cdots)$ であるようなある定数 $\delta_i > 0$ が存在して，

$$0 < \lim_{p\to\infty} \mathrm{tr}\left(\frac{\mathbf{\Sigma}_p}{p^{\delta_i}}\right)^i < \infty\ \ (i = 1, 2, \cdots). \tag{4}$$

本研究の目的は，検定統計量 $T_D^2,\ T_{BS}^2, T_S^2$ の高次元漸近理論の枠組みにおける漸近分布を条件 (4) の下に拡張し，より一般的な条件の下でも帰無仮説 H_0 を検定できるようにすることである. 本報告では，漸近分布は共分散行列に依存して変化することを示し，特に，ある共分散行列に対しては対応する漸近分布のパーセント点が容易にもとまらない場合があることを示す. そのような場合の対処法も簡単に紹介する. 漸近分布の導出に関する理論的な詳細，共分散行列と漸近分布の関係を示す例，そして簡単な数値実験の結果は当日報告する.

参考文献

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

[2] Dempster, A. P. (1958), A high dimensional two sample significance test, *The Annals of Mathematical Statistics*, **29**, 995–1010.

[3] Dempster, A. P. (1960), A significance test for the separation of two highly multivariate small samples, *Biometrics*, **16**, 41–56.

[4] Fujikoshi, Y., Himeno, T. and Wakaki, H. (2004), Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size, *Journal of the Japan Statistical Society*, **34**, 19–26.

[5] Grenander, U. and Szegö, G (1958), *Toeplitz forms and their applications*, 2nd edition, Chelsea Publishing Company, New York.

[6] Johnstone, I.M. (2007), On the distribution of the largest eigenvalue in principal components analysis, *The Annals of Statistics*, **29**, 295–327.

[7] Srivastava, M.S. (2007), Multivariate theory for analyzing high dimensional data, *Journal of the Japan Statistical Society*, **37**, 53–86.

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

共分散行列が未知の場合の任意の半順序対立仮説に対する平均ベクトルの均一性の検定

笹渕 祥一（九州大学　大学院　芸術工学研究院）

$A=\{a_1,a_2,\ldots,a_k\}$ をある有限集合とし、$\preceq$ を A 上のある半順序とする。

定義1 (Sasabuchi, Inutsuka and Kulatunga (1983))　$p\times k$ 実行列 $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ に対して、「$a_i\preceq a_j$ ならば $\boldsymbol{\theta}_i\le\boldsymbol{\theta}_j$」が成り立つとき、「$(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ は半順序 $\preceq$ に関して isotonic である」という。ここで、$\boldsymbol{\theta}_i\le\boldsymbol{\theta}_j$ は $\boldsymbol{\theta}_j-\boldsymbol{\theta}_i$ の成分がすべて非負であることを意味する。

定義2 (Sasabuchi et al. (1983))　$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k$ を与えられた k 個の p 次元実ベクトルとし、$\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_k$ を与えられた k 個の $p\times p$ 正定値行列とする。$(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ が半順序 $\preceq$ に関して isotonic であるという制約の下で

$$\sum_{i=1}^{k}(\boldsymbol{x}_i-\boldsymbol{\theta}_i)'\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{x}_i-\boldsymbol{\theta}_i)$$

を最小にする $p\times k$ 実行列 $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ を、「$\boldsymbol{\Lambda}_1^{-1},\ldots,\boldsymbol{\Lambda}_k^{-1}$ を重みとする $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k$ の Multivariate Isotonic Regression（多変量単調回帰）」と呼び、M.I.R. と書く。特に、$p=1$ のとき、M.I.R. は通常の Isotonic Regression（単調回帰）である。

$\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{iN_i}$ を、p 次元正規分布 $N_p(\boldsymbol{\mu}_i,\ \lambda_i\boldsymbol{\Sigma})$ からの大きさ N_i の無作為標本とする $(i=1,\ldots,k)$。ただし、λ_i は既知の正の数、$\boldsymbol{\Sigma}$ は未知とし、$N_1+\cdots+N_k>p+k$ とする。次の検定問題を考える。

帰無仮説 $\mathrm{H}_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_k$

対立仮説 $\mathrm{H}_1:(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k)$ は半順序 $\preceq$ に関して isotonic

（ただし、$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_k$ ではない。）

Sasabuchi,Tanaka and Tsukamoto (2003), Sasabuchi (2007) は、次の（i）,（ii）の設定の下で、この問題を考察した。

（i）$\lambda_1=\lambda_2=\cdots=\lambda_k$.

（ii）$\preceq$ は simple order（単純順序）、すなわち、$a_1\preceq a_2\preceq\cdots\preceq a_k$.

この問題を、（i）,（ii）の条件がない場合に拡張して考察する。

$\overline{\boldsymbol{X}}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij},\ i=1,\ldots,k,\ \ \overline{\boldsymbol{X}}=\left(\sum_{i=1}^{k}\frac{N_i}{\lambda_i}\right)^{-1}\sum_{i=1}^{k}\frac{N_i}{\lambda_i}\overline{\boldsymbol{X}}_i,$

$S=\sum_{i=1}^{k}\frac{1}{\lambda_i}\sum_{j=1}^{N_i}(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_i)'$

とおく。

次の検定統計量 $\bar{T}^2$ を考える。

$$\bar{T}^2 = \sum_{i=1}^{k} \frac{N_i}{\lambda_i} (\widehat{\boldsymbol{\mu}}_i - \overline{\boldsymbol{X}})' S^{-1} (\widehat{\boldsymbol{\mu}}_i - \overline{\boldsymbol{X}})$$

ここで、$(\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k)$ は、$\frac{N_1}{\lambda_1} S^{-1}, \ldots, \frac{N_k}{\lambda_k} S^{-1}$ を重みとする $\overline{\boldsymbol{X}}_1, \ldots, \overline{\boldsymbol{X}}_k$ の M.I.R.。

また、次の p 個の統計量 T_m を導入する $(m = 1, 2, \ldots, p)$。

$$T_m = \sum_{i=1}^{k} \frac{N_i}{\lambda_i} (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}})' S^{-1} (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}}) - \frac{1}{s_{mm}} \sum_{i=1}^{k} \frac{N_i}{\lambda_i} (\overline{X}_{im} - \widehat{\mu}_{im})^2$$

ここで、$\overline{X}_{im}$ は $\overline{\boldsymbol{X}}_i$ の第 m 成分、s_{mm} は S の (m, m) 成分であり、$(\widehat{\mu}_{1m}, \ldots, \widehat{\mu}_{km})$ は $\frac{N_1}{\lambda_1}, \ldots, \frac{N_k}{\lambda_k}$ を重みとする $\overline{X}_{1m}, \ldots, \overline{X}_{km}$ の Isotonic Regression $(m = 1, 2, \ldots, p)$。

さらに、次の検定統計量を考える。

$$T_{MIN} = \min_{1 \le m \le p} T_m$$

Sasabuchi et al. (2003), Sasabuchi (2007) は、(i) , (ii) の設定の下で、概ね以下の（１）～（４）の結果を示している。

（１）帰無仮説 H_0 の下での $\bar{T}^2, T_{MIN}$ の分布は、$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ に依存しない。

（２）$m = 1, 2, \ldots, p$ に対し、帰無仮説 H_0 の下での T_m の分布は、$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ にも $\boldsymbol{\Sigma}$ にも依存しない。従って、T_m による検定は相似検定となる。

（３）同じ棄却点を用いて検定を行うと、$\bar{T}^2, T_m, T_{MIN}$ による検定は、すべて、検定の有意水準（サイズ）が等しくなる。

（４）T_m, T_{MIN} による検定は、$\bar{T}^2$ による検定に比べて、統計量の計算がはるかに簡単であるだけでなく、一様に真に検出力が高い。

今回、(i) , (ii) の条件がなくても、（１）～（４）がすべて成り立つという結果を得た。Sasabuchi et al. (2003), Sasabuchi (2007) において得られた諸結果が、今回の報告にあるような統計量の修正を行うことにより、simple order のみでなく、tree order、umbrella order 等、任意の順序制約対立仮説に対する平均ベクトルの均一性の検定に関して 適用できることになる。

今回の講演では、以上の内容を、研究の時間的流れに沿って、次の順序に再構成して報告を行った。

１．問題設定：simple order（単純順序） の場合

２．検定統計量 $\bar{T}^2$

３．統計量 T^* に関する考察（T^* は $\bar{T}^2$ の上側確率を求めるために用いた補助的な統計量）

４．$\bar{T}^2$-test より検出力が高い検定

５．一般化：任意の半順序による制約下の対立仮説へ

台が有界な分布における台の端点の逐次推測

小池　健一　　筑波大学・数理物質科学研究科

1. はじめに

非正則な確率分布の典型例として，一様分布に対する逐次推定問題については，多くの研究がある．例えば，尺度母数をもつ一様分布，すなわち　,θ　$\theta \in \mathbb{R}$ 上の一様分布 U　,θ については，　y　　C　　C　k　　j　　等がある．また，位置母数をもつ一様分布 U θ　　/　,θ　　/　については，W　　，Ak　　k　　などがある．一方，一般の非正則分布に関する逐次推測問題を扱った文献は多くはない．　k　　y y　　8　は，　w　f　　y　　において，逐次点推定問題を考えている　k　　y y　8　　k　　y y　C　　も参照

最近，K　k　　で，有界な台をもつ位置尺度母数分布族について，その位置母数の逐次区間推定方式および逐次点推定方式方式が得られ，その漸近有効性が示されている．これは，非正則な場合を広く扱ったものであり，特に具体的に分布の形状を特定しなくても良いという利点がある．本講演では，有界な台をもつ確率分布族について，その台の端点の逐次区間推定方式および逐次点推定方式方式について考える．このような推定方式は，切断分布族に対する推測に応用できる．例えば，網による魚の捕獲に対する網目選択性の問題　　8　の　　節，F y　　などを参照 に適用可能である．

2. 逐次区間推定方式

$X_1, X_2, \ldots$ を，互いに独立にいずれも ルベーグ測度に関する 確率密度関数 $f_0\ x$　$\theta \in \mathbb{R}^1$ をもつ確率分布に従う確率変数列とする．ただし，$f_0\ x$ は台 θ_1, θ_2 をもつ，すなわち

A
$$f_0\ x \begin{cases} > & \theta_1 < x < \theta_2\ , \\ & \text{それ以外} \end{cases}$$

であって，台の内部で θ_1, θ_2 を除いて θ_1, θ_2 の近傍で C^2 級で次を満たすとする．

A　$\underset{x \to \theta_1 + 0}{\ }\ x　\theta_1　^{\gamma_1} f_0\ x　　g_1\ \theta_2　\theta_1\ ,　\underset{x \to \theta_2　0}{\ }\ \theta_2　x　^{\gamma_2} f_0\ x　　g_2\ \theta_2　\theta_1\ .$

ただし，$\gamma_i >$　　i　　，　，$g_1\ \theta_2　\theta_1$ と $g_2\ \theta_2　\theta_1$ は，$\theta_2　\theta_1$ の狭義単調減少，連続正値関数とする．

$X_{(1:n)}$　　$1 \le i \le n\ X_i$　$X_{(n:n)}$　　x$1 \le i \le n\ X_i$ とおき，θ_1 を信頼区間 $[X_{(1:n)}　d, X_{(1:n)}]$ で区間推定することを考える．$\theta_2　\theta_1$ が既知のとき

$$P\{X_{(1:n)}　d　\theta_1　X_{(1:n)}\}　P\left\{　n^{1/(\gamma_1+1)}\ X_{(1:n)}　\theta_1　n^{1/(\gamma_1+1)} d\right\}$$
$$\approx \int_0^{n^{1/(\gamma_1+1)} d} f_U\ u\ du$$
$$\text{x}　\frac{g_1\ \theta_2　\theta_1}{\gamma_1} n d^{\gamma_1+1}　,　n \in \mathbb{N}$$

となる．ただし，“≈” は $n^{1/(\gamma_1+1)}\ X_{(1:n)}$　θ_1 の漸近分布による近似を表す．従って，　<　< に対して，n　n　　$\frac{(\gamma_1+1)\log\alpha}{g_1(\theta_2　\theta_1) d^{\gamma_1+1}}$ であれば

$$\text{x}　\frac{g_1\ \theta_2　\theta_1}{\gamma_1} n d^{\gamma_1+1}$$

となる．よって，n は，$\theta_2 \quad \theta_1$ が既知のとき，漸近的に最適な標本数となる．ところが，$\theta_2 \quad \theta_1$ は未知であるので，これをレンジ $R_n \quad X_{(n:n)} \quad X_{(1:n)}$ で置き換えた停止則

$$\tau_1 \quad \mathrm{f} \quad n \quad n_0 \,\middle|\, n \quad \frac{\gamma_1}{g_1 \ R_n \ d^{\gamma_1+1}}$$

を考える．ただし，n_0 は初期標本数とする．このとき次が成り立つ．

定理 1. A と A の下で次が成り立つ．

$$\lim_{d\to 0+} P\{X_{(1:\tau_1)} \quad d \quad \theta_1 \quad X_{(1:\tau_1)}\} \qquad \text{漸近一致性}$$

$$\tau_1/n \xrightarrow{\text{a.s.}} \quad d \to$$

$$E\ \tau_1\ /n \ \to \quad d \to \qquad \text{漸近有効性}$$

また，二段階法を用いた θ_1 の区間推定方式を考えることもでき，その漸近一致性や漸近有効性を示すことができる．純逐次法，二段階法の大きな違いは初期標本数にある．前者は d に依存しないように取れるが，後者はそうではない．すなわち，区間幅 d が小さいようなときには，ある程度大きな初期標本数を必要とする．

また，θ_1 と同様に，θ_2 の逐次区間推定方式について考えることもできる．

3. 逐次点推定方式

ここでは標本抽出に対する費用も考慮した上で，未知母数の点推定方式について考える．まず，θ_1 の逐次点推定方式を考慮の対象とする．ここでは，θ_1 を $X_{(1:n)}$ で点推定するものとする．$U \quad n^{1/(\gamma_1+1)} \ X_{(1:n)} \quad \theta_1$ の漸近密度は

$$f_U\ u \quad g_1\ \theta_2 \quad \theta_1\ u^{\gamma_1} \ \mathrm{x} \qquad \frac{g_1\ \theta_2 \quad \theta_1}{\gamma_1} u^{\gamma_1+1}$$

より，U^2 の漸近期待値は

$$E\ U^2 \ \approx \int_0^\infty g_1 u^{\gamma_1+2} \ \mathrm{x} \qquad \frac{g_1}{\gamma_1} u^{\gamma_1+1} \quad du \quad \left(\frac{\gamma_1}{g_1}\right)^{2/(\gamma_1+1)} \Gamma\left(\frac{\gamma_1}{\gamma_1}\right)$$

となる．ただし，Γ はガンマ関数とする．このとき，K k の L と同様に，

$$E\ U^2 \ \to C \quad n \to \infty$$

となることが示せる．ここでは，さらに次を仮定する．

B $\qquad E\ U^2 \ \to h_1\ \theta_2 \quad \theta_1 \qquad n \to \infty\ .$

となり，$h_1\ \theta_2 \quad \theta_1$ は $\theta_2 \quad \theta_1$ の正値増加連続関数となる．

θ_1 を $X_{(1:n)}$ で推定したときのリスクを

$$r_n^{(1)} \qquad E\ X_{(1:n)} \quad \theta_1\ ^2 \quad dn$$

とする．ただし，$d >$ は観測ごとのコストとする．このとき，$U \quad n^{1/(\gamma_1+1)} \ X_{(1:n)} \quad \theta_1$ より $r_n^{(1)} \approx h_1\ \theta_2 \quad \theta_1\ n^{\ 2/(\gamma_1+1)} \quad dn$ と近似される．これを n を変数とする関数とみなすと，$n \quad n \qquad \left\{\frac{2h_1(\theta_2 \quad \theta_1)}{(\gamma_1+1)d}\right\}^{(\gamma_1+1)/(\gamma_1+3)}$ で最小値 $r_{n^{**}}^{(1)} \qquad h_1\ \theta_2 \quad \theta_1 \left\{\frac{d(\gamma_1+1)}{2h_1(\theta_2 \quad \theta_1)}\right\}^{2/(\gamma_1+3)} \left(\frac{\gamma_1+3}{\gamma_1+1}\right)$ をとる．ところが，$\theta_2 \quad \theta_1$ は未知であるので，これをレンジ R_n で置き換えた停止則

$$\tau_2 \qquad \mathrm{f}\left\{ n \quad m_d^{(1)} \,\middle|\, n \qquad \frac{h_1\ R_n}{\gamma_1 \qquad d}^{(\gamma_1+1)/(\gamma_1+3)} \right\}$$

を考える．ただし，$m_d^{(1)}$ は初期標本数で $d^{\ l} \quad m_d^{(1)} \quad o\ d^{\ (\gamma_1+1)/(\gamma_1+3)} \qquad < l < \ \gamma_1 \qquad / \ \gamma_1$ を満たすとする．このとき次が成り立つ．

定理 2. A と B の下で次が成り立つ．

$$\tau_2/n \ \xrightarrow{\text{a.s.}} \quad d \to \qquad E\ \tau_2\ /n \quad \to \quad d \to \qquad r_{\tau_2}^{(1)}/r_{n^{**}}^{(1)} \ \to \quad d \to$$

この場合にも，漸近有効な二段階推定方式を構築できる．θ_2 に関しても同様である．

さらに，数値例として，定理 の $\tau_1, [X_{(1:\tau_1)} \quad d, X_{(1:\tau_1)}]$ の被覆確率について考えたが，狙った精度に非常に近い値を得た．

Sample Size Determination for High-Dimension, Low-Sample-Size Data

矢田 和善　（筑波大学大学院数理物質科学研究科）
青嶋 誠　　（筑波大学大学院数理物質科学研究科）

1. はじめに

高次元小標本 (HDLSS) における統計的推測について，新しい方法論を提案した．HDLSS データとは，数千から数万の次元数に対して，数十から百程度の標本数からなるデータのことをいう．最近，Aoshima and Yata (2010) は，HDLSS における平均ベクトルについて，重要な 8 つの推測問題を提示し，目標精度に到達するための一連の理論と方法論を与えた．本講演では，Aoshima and Yata (2010) で扱った 8 つの推測問題のうち，平均ベクトルの推測を扱った 2 つの推測問題，要求されるバンド幅をもつ信頼領域問題，要求される有意水準と検出力をもつ 2 標本問題を紹介し，構築法を与えた．さらに，Aoshima and Yata (2010) では扱わなかった問題として，要求される半径をもつ信頼領域，要求される幅をもつノルムの信頼区間について，それらの構築法も与えた．各種推測には目標精度を設定し，必要な標本数を 2 段階の標本抽出で決定することで，HDLSS における漸近的な精度を保証することが理論的かつ数値的に確認された．

2. 要求されるバンド幅をもつ信頼領域

母数が未知の k 個の p 次元母集団 $\ _i,\ i = 1, ..., k$ における各母平均ベクトル $\ _i$ の 1 次結合 $\ = \sum_{i=1}^{k} b_i\ _i$ を推定する．いま，平均ベクトルの 1 次結合 $\ = \sum_{i=1}^{k} b_i\ _i$ を推定する．ただし，各母集団 $\ _i$ と母共分散行列 $\boldsymbol{\Sigma}_i$ には適当な正則条件を課す．各母集団から抽出される大きさ n_i の標本に基づいて $\boldsymbol{T}_{\mathbf{n}} = \sum_{i=1}^{k} b_i \overline{\boldsymbol{X}}_{in_i}$ を定義する．ただし，$\boldsymbol{n} = (n_1, ..., n_k)$, $\overline{\boldsymbol{X}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}/n_i$ とする．本講演では，損失関数 $||\boldsymbol{T}_{\mathbf{n}}\ \ ||^2$ について，与えられる $\delta = o(p^{1/2}) > 0$ によりバンド幅を要求した

$$R\ _{\mathbf{n}} = \{\ \in R^p : \max\{\ \delta + \boldsymbol{\Sigma}_{\mathbf{n}},\ 0\}\ \ ||\boldsymbol{T}_{\mathbf{n}}\ \ ||^2\ \ \delta + \boldsymbol{\Sigma}_{\mathbf{n}}\}$$

なる領域を考えた．ここで $\boldsymbol{\Sigma}_{\mathbf{n}} = \sum_{i=1}^{k} b_i^2 \mathrm{tr}(\boldsymbol{\Sigma}_i)/n_i$ である．そのとき，与えられる $\alpha \in (0, 1)$ に対して，

$$P_{\boldsymbol{\theta}}\ \ \in R\ _{\mathbf{n}}$$

なる信頼領域を求める．各母集団の標本共分散行列 $\boldsymbol{S}_{in_i} = (n_i\ \ 1)^{\ 1} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij}$ $\overline{\boldsymbol{X}}_{in_i})(\boldsymbol{X}_{ij}\ \ \overline{\boldsymbol{X}}_{in_i})^T$ に基づいて，$\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}} = \sum_{i=1}^{k} b_i^2 \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i$ とおく．そのとき，$||\boldsymbol{T}_{\mathbf{n}}$ $||^2$ について次の結果が導かれた．

定理. 適当な正則条件と，$p \to \infty$, $n_i \to \infty$, $i = 1, ..., k$ のもとで，次が成り立つ.

$$\frac{||\boldsymbol{T}_{\mathbf{n}} \quad ||^2 \quad \widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}}{\sqrt{2\sum_{i,j} b_i^2 b_j^2 \mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)/(n_i n_j)}} \Rightarrow N(0,1)$$

上記の定理に基づき，各母集団の標本数は

$$n_i \quad \frac{z_{\alpha/2}\sqrt{2}}{\delta} |b_i| \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \sum_{j=1}^{k} |b_j| \mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \quad (= C_i, \text{ say})$$

を満たす最小の整数とする．そのとき次の結果が導かれた.

定理. 適当な正則条件と $p \to \infty$ のもとで，次が成り立つ.

$$\mathrm{f}\, P_{\boldsymbol{\theta}} \quad \in R\frown_{\mathbf{n}}$$

２段階推定法

標本数 n_i は未知母数を含むため，2 段階推定法を考えた．各 $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ に，事前情報から得られる既知の下限 $\sigma_{i\star}$ $(\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} > \sigma_{i\star} > 0)$ を仮定し，$\sigma_{i\star}/\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \in (0,1)$, $p \to \infty$ を仮定する．いま，$\tau_\star = \min_{1 \le i \le k} |b_i| \sqrt{\sigma_{i\star}} \sum_{j=1}^{k} |b_j| \sqrt{\sigma_{j\star}}$ とおいて，初期標本数を $m = \max\{4, [z_{\alpha/2}\sqrt{2}\tau_\star/\delta] + 1\}$ と定義する．ここで，$[x]$ は x を越えない最大の整数を表す．各母集団から m 個の初期標本ベクトルを抽出し，$\boldsymbol{S}_{im(1)} = (m_1 \quad 1)\ ^1 \sum_{j=1}^{m_1} (\boldsymbol{X}_{ij} \quad \overline{\boldsymbol{X}}_{im_1})(\boldsymbol{X}_{ij} \quad \overline{\boldsymbol{X}}_{im_1})^T$, $\boldsymbol{S}_{im(2)} = (m_2 \quad 1)\ ^1 \sum_{j=m_1+1}^{m} (\boldsymbol{X}_{ij} \quad \overline{\boldsymbol{X}}_{im_2})(\boldsymbol{X}_{ij} \quad \overline{\boldsymbol{X}}_{im_2})^T$ を計算する．ここで，$m_1 = [m/2] + 1$, $m_2 = m \quad m_1$ とし，$\overline{\boldsymbol{X}}_{im_1} = \sum_{j=1}^{m_1} \boldsymbol{X}_{ij}/m_1$, $\overline{\boldsymbol{X}}_{im_2} = \sum_{j=m_1+1}^{m} \boldsymbol{X}_{ij}/m_2$ とする．そのとき，各母集団の標本数を

$$N_i = \max\left\{ m, \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta} |b_i| \mathrm{tr}(\boldsymbol{S}_{im(1)} \boldsymbol{S}_{im(2)})^{1/4} \sum_{j=1}^{k} |b_j| \mathrm{tr}(\boldsymbol{S}_{jm(1)} \boldsymbol{S}_{jm(2)})^{1/4} \right] + 1 \right\}$$

で定義する．各母集団から追加の $N_i \quad m$ 個の標本ベクトルを抽出し，初期標本と追加標本を合併して $\boldsymbol{T}_{\mathbf{N}} = \sum_{i=1}^{k} b_i \overline{\boldsymbol{X}}_{iN_i}$ と $\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = \sum_{i=1}^{k} b_i^2 \mathrm{tr}(\boldsymbol{S}_{iN_i})/N_i$ を定義する．ここで，$\boldsymbol{N} = (N_1, ..., N_k)$ である．このとき，次の結果が導かれた.

定理. 適当な正則条件と $p \to \infty$ のもとで，次が成り立つ.

(i) $\liminf P_{\boldsymbol{\theta}}(\quad \in R\frown_{\mathbf{N}}) \quad 1 \quad \alpha$; (ii) $\limsup |E_{\boldsymbol{\theta}}(N_i \quad C_i)| < 1$;

(iii) $Var_{\boldsymbol{\theta}}(N_i) = o(p^{1/2}/\delta)$

本講演では他に，要求される半径をもつ信頼領域，要求される幅をもつノルムの信頼区間，要求される有意水準と検出力をもつ 2 標本問題について，高次元小標本のもと目標精度を満足するよう，必要な標本数を決定する方法論についても報告した.