(10)「データ解析のための統計科学理論」に関する研究報告

On Multiple Comparisons of Mean Components in the Intraclass Correlation Model with Missing Data

東京理科大・理学研究科　菊地　淳
東京理科大・理学部　瀬尾　隆

本報告では，分散共分散行列が一様な構造をもつ Intraclass Correlation Model の下での繰り返し測定データにおいて，与えられたデータに欠測が生じた場合の平均成分の同等性検定について議論する．関連する研究として，Seo and Srivastava[2] 等がある．

問題を定式化するため，次のような Two-components mixed linear model を考える．

$$x_{ij} = \mu_i + \alpha_j + \epsilon_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n_i \tag{1}$$

ここで，α_j と ϵ_{ij} は互いに独立にそれぞれ平均 0，分散 σ_α^2 の正規分布と平均 0，分散 σ_ϵ^2 に従うものと仮定する．すると，$\mathrm{E}[x_{ij}] = \mu_i, i = 1, \ldots, p$ であり，$\mathrm{Var}[x_{ij}] = \sigma_\alpha^2 + \sigma_\epsilon^2 (\equiv \sigma^2)$，$\mathrm{Cov}[x_{sj}, x_{tj}] = \sigma_\alpha^2, (s \neq t)$，そして，$\mathrm{Cov}[x_{is}, x_{it}] = 0, (s \neq t)$ となる．ここで，$n_i = n, i = 1, \ldots, p$ とし，$\boldsymbol{x_j} = (x_{1j}, \ldots, x_{pj}), j = 1, \ldots, n$ とおくと，$\boldsymbol{x_j}$ は平均ベクトル $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$，分散共分散行列 Σ を持つ p 次元正規分布に従う．ただし，$\boldsymbol{\Sigma} = \sigma^2\{(1-\rho)\boldsymbol{I} + \rho \boldsymbol{e}\boldsymbol{e}'\}$ は正則であり，一様構造 (intraclass correlation form) と呼ばれる．ここで，$\boldsymbol{I}$ は $p \times p$ の単位行列，$\boldsymbol{e} = (1, \ldots, 1)'$ は $p \times 1$ の列ベクトル，$\rho \in [-(p^{-1} - 1), 1]$ である．分散共分散行列 Σ がこのような構造を持つモデルは，Intraclass Correlation Model と呼ばれる．

本報告では，欠測データの場合に対する平均成分の同等性検定について考える．次のような仮説 $\boldsymbol{H}$ を考える．

$$\boldsymbol{H} : \mu_1 = \mu_2 = \cdots = \mu_p \tag{2}$$

観測ベクトルを，$\boldsymbol{x_j} \equiv (x_{1j}, \ldots, x_{p_j j})', j = 1, 2, \ldots, n$ とおく．ただし，$p \equiv p_1 \geq \cdots \geq p_n$ である．また，$\boldsymbol{x_j}$ はそれぞれ独立に $N_{p_j}(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ に従う．ここで，Bhargava and Srivastava[1] のアイデアを利用し，次のような変換を考える．

$$\boldsymbol{z_j} \equiv \boldsymbol{C_j x_j}, \quad \boldsymbol{C_j} \equiv \boldsymbol{I}_{p_j} - \frac{\nu_j}{p_j} \boldsymbol{e_j e_j'}, \quad \nu_j \equiv 1 \pm (1-\rho)^{\frac{1}{2}} \{1 + (p_j - 1)\rho\}^{\frac{-1}{2}}. \tag{3}$$

すると，$\boldsymbol{z_j}$ はそれぞれ独立に $N_{p_j}(\boldsymbol{C_j \mu_j}, \gamma^2 \boldsymbol{I}_{p_j})$ に従う．ここに，$\gamma^2 \equiv \sigma^2(1-\rho)$ である．ここで，各ブロックで完全データを構成するように，変換されたデータを s 個のブロックに分割する．ただし，第 k ブロックは $p^{(k)} \times n^{(k)}$ 行列で，$p \geq s \geq k \geq 1$ である．さらに，

$$\overline{z}_{i\cdot}^{(k)} \equiv \frac{1}{n^{(k)}} \sum_{j=1}^{n^{(k)}} z_{ij}^{(k)}, \quad \overline{z}_{\cdot j}^{(k)} \equiv \frac{1}{p^{(k)}} \sum_{i=1}^{p^{(k)}} z_{ij}^{(k)}, \quad \overline{z}_{\cdot\cdot}^{(k)} \equiv \frac{1}{p^{(k)} n^{(k)}} \sum_{i=1}^{p^{(k)}} \sum_{j=1}^{n^{(k)}} z_{ij}^{(k)} \tag{4}$$

$$\hat{\gamma}^{(k)2} \equiv \frac{1}{f^{(k)}} \sum_{i=1}^{p^{(k)}} \sum_{j=1}^{n^{(k)}} (z_{ij}^{(k)} - \overline{z}_{i\cdot}^{(k)} - \overline{z}_{\cdot j}^{(k)} + \overline{z}_{\cdot\cdot}^{(k)})^2, \quad f^{(k)} \equiv (n^{(k)} - 1)(p^{(k)} - 1) \tag{5}$$

とおく．ここに，$z_{ij}^{(k)}$ は k 番目のブロックの i, j 成分である．よって，仮説 $\boldsymbol{H}$ の下で，それぞれのブロックで独立に，

$$\sum_{i=1}^{p^{(k)}} \left(\frac{z_{i\cdot}^{(k)} - \overline{z}_{\cdot\cdot}^{(k)}}{\gamma / \sqrt{n^{(k)}}} \right)^2 \sim \chi^2_{p^{(k)}-1}, \quad \frac{f^{(k)} \hat{\gamma}^{(k)2}}{\gamma^2} \sim \chi^2_{f^{(k)}} \quad , k = 1, \ldots, s \tag{6}$$

となる．以上により，検定統計量は次のように与えられる．

$$\frac{\sum_{k=1}^{s} \sum_{i=1}^{p} n^{(k)} (z_{i\cdot}^{(k)} - \overline{z}_{\cdot\cdot}^{(k)})^2 / p^*}{\sum_{k=1}^{s} f^{(k)} \hat{\gamma}^{(k)2} / f} \sim F_{p^*, f} \tag{7}$$

ここに，$p^* \equiv \sum_{k=1}^{s} (p^{(k)} - 1)$, $f \equiv \sum_{k=1}^{s} f^{(k)}$ であり，$F_{p^*,f}$ は自由度 p^*, f の F 分布である．

本報告では，Srivastava and Carter[3] の数値例を用いて，提案した検定の手順を示した．今後，完全データで議論している Bhargava and Srivastava[1] における平均のコントラストに対する同時信頼区間について，欠測データを仮定した考察を行う．

参考文献

[1] Bhargava, R.P. and Srivastava,M.S.(1973),"On Tukey's Confidence Intervals for the Contrasts in the Means of the Intraclass Correlation Model," *J.Royal Statist. Soc.*, B35, 147-152.

[2] Seo,T. and Srivastava, M.S.(2000), " Testing Equality of Means and Simultaneous Confidence Intervals in Repeated Measures with Missing Data," *Biometrical Journal*, 42, 981-993.

[3] Srivastava, M.S. and Carter,E.M.(1986),"The Maximum Likelihood Method for Non-Responce in Sample Survey," *Survey Methodlogy*, 12, 61-72.

疾病の地域集積性とデータ解析

国立保健医療科学院 技術評価部　高橋 邦彦

1　はじめに

疫学などの分野では, 疾病集積性の問題は重要である. 調査・対策を行う場合, どの地域を優先的に行うかを選定する必要があるからである. しかし, どんな指標であっても順に並べれば必ず, 最大値, 最小値が存在するので, 各地域の死亡率などを見る場合, 死亡率の最大の地域が本当に意味のある調査対象地域なのかどうかはわからない. そこで, その地域が他の地域と比べて有意に死亡率が高いのかどうかの検定を行う必要がある. すなわち, 対象とする空間内で, ある特定の平面領域を同定し, それが本当に有意な意味をもつのかどうかを検定することになる. この集積性の検定にはKulldorffの方法, Tangoの方法をはじめ, いくつかの検定法が提案されている. 本論では実際によく利用されているKulldorffの方法を改良する形で新たな検定法および, このような平面領域同定における検定法の評価法を提案する. さらに, この問題を通して実際問題におけるデータ解析についての考察を行う.

2　平面領域同定の検定

今, 対象とする空間が m 個の区域に分かれているとする. 確率変数 $X_1, X_2, \cdots, X_m$ は互いに独立にいずれもポアソン分布 $Po(\eta_i p_i)$ $(i = 1, 2, \cdots, m)$ に従うとし, その観測値を x_i $(i = 1, 2, \cdots, m)$ とする. ただし, $\eta_i > 0$, $0 < p_i < 1$ とする. この空間上に, ある領域 $Z(\in \mathcal{Z})$ をとったとき,

$$p_i = p \quad (i \in Z); \qquad p_i = q \quad (i \notin Z)$$

のモデルの下で, 帰無仮説 $H_0 : p = q$, 対立仮説 $H_1 : p > q$ の仮説検定問題を考える. ここで, $n(Z) := \sum_{i \in Z} x_i$, $\mu(Z) := \sum_{i \in Z} \eta_i$, $G = Z \cup Z^c$ とおくと, この検定は, 尤度比検定の考えから, 検定統計量

$$\lambda := \sup_{Z \in \mathcal{Z}} \frac{\left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{n(Z^c)}{\mu(Z^c)}\right)^{n(Z^c)}}{\left(\frac{n(G)}{\mu(G)}\right)^{n(G)}} I\left(\frac{n(Z)}{\mu(Z)} > \frac{n(Z^c)}{\mu(Z^c)}\right)$$

を用いた検定になる. ただし $I(\cdot)$ は定義関数とする. つまり, この λ を取る Z が同定したい領域であり, これが本当に有意に $p > q$ となっているのかどうかを, モンテカルロシミュレーションによって検定を行うことが出来る.

実際に疫学調査などで疾病集積性を考える場合, この同定される領域 Z は「隣接した地域」として1つの平面領域であることが望まれる. そこでKulldorffらは区域 i を中心に半径 r の円を描き, その円に含まれる領域を Z とした. この r を0から予め設定された上限まで連続的に変化させ, さらにそれを全ての i について行うことで様々な領域を取ることができる. このようにして得られる全ての領域 Z の集合として $\mathcal{Z}$ を定めた. この方法は非常に明解かつ簡便であり, 実際多くの疫学研究の場面で利用されている. しかしながら, その $\mathcal{Z}$ の定め方から円状に近い平面領域の同定には優れているが, 円状でない領域の場合にはうまく同定されないことになる. そこで本論では以下のような $\mathcal{Z}$ を定める方法 (flexible scan) を提案する.

1. ある区域 i からの距離が短い順に $k-1$ 個の区域を取り出す.
2. 取り出した $k-1$ 個の区域と i を合わせた k 個の中から長さ $l(\leq k)$ 個の組み合わせを考え, その中で i を含み, さらに全てが連結している組み合わせを Z とする.
3. l を1から k まで変化させ, 全ての Z を取り出す.

4. 以上の手順を各 i について行い, その全ての Z を要素とする集合を $\mathcal{Z}$ とする.

実際にこの方法を疾病集積性の検定に適用して, Kulldorff の方法と比較を行った. ここでは東京都と神奈川県を市区町村単位に分けた空間を対象とし, ホットスポットモデル ($p = 3q$) の下でいくつかのシミュレーションによって検討を行った. その結果, Kulldorff の方法に比べて今回の方法が様々な形状の領域に対し, その領域を的確に同定できることが分かった. また, 時間軸を入れた 3 次元空間の同定にも拡張できる.

3 領域同定における検定法の評価

一般に検定においては検出力によってその比較が行われる. しかし, 集積性の検定のような平面領域の同定の場合, その領域を正確に同定できれば良いのか, それを含んだ広い領域を同定したものも良いのか, または単に有意な領域が同定されれば良いのかなどいろいろな考え方ができ, 明確な定義はされていない. 実際, 今までの研究ではシミュレーションによって「有意な領域が同定された割合」を検出力として用いていることが多いが, これでは検定法の評価としては不十分であろう. そこで本論では, あるホットスポット領域に対して, 実際に同定された領域の長さ (区域数) と, その中に含まれるホットスポットの区域数を同時に表す 2 次元分布を考え, その評価を行う (表参照). このような評価を行うことで検定法による違いがよく観察でき, 特に本論で提案した検定法がホットスポットをきちんと含んだ形で領域をうまく同定していることが観察された.

表: 長さ 4 のあるホットスポット領域を同定する検定 ($\alpha = 0.05$) の様子

Kulldorff

Length l	Include s hot-spot regions					Total
	0	1	2	3	4	
1	1	0				1
2	0	0	351			351
3	2	0	4	0		6
4	0	0	3	0	0	3
5	2	0	2	0	0	4
6	1	0	0	0	0	1
7	0	0	0	81	0	81
8	0	0	10	18	38	66
9	0	0	2	0	26	28
10	0	0	0	29	3	32
11	0	0	1	13	1	15
12	0	0	2	4	60	66
13	0	0	0	5	62	67
14	0	0	0	10	27	37
15	0	0	0	6	37	43
Total	6	0	375	166	254	801

power=0.801

Flexible

Length l	Include s hot-spot regions					Total
	0	1	2	3	4	
1	0	0				0
2	0	0	0			0
3	0	0	0	0		0
4	0	0	0	0	138	138
5	0	0	0	3	147	150
6	1	0	0	2	200	203
7	0	1	0	4	147	152
8	0	0	2	9	107	118
9	0	0	0	10	71	81
10	1	0	2	5	28	36
11	0	0	0	0	10	10
12	0	0	0	0	2	2
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
Total	2	1	4	33	850	890

power=0.890

4 疫学調査におけるデータ解析

統計手法を現実場面への適用を考える際には, 実際の状況や現場からの要望などによって条件を課しながら理論を考慮しなくてはいけないことがある. それは統計以外の情報や知識が必要となるが, 逆に現場から起こる問題を解決するための理論を構築するという面白さもあるものと思われる. また, 今回の疾病集積性の分野では世界的に Kulldorff の方法が利用されることが多い. それは理論のわかりやすさもあるが, 手法を考案, 提案している Kulldorff 自身がパソコンのアプリケーションソフトを作成しインターネット上で無料で配布していることも一つの要因であろう. そのため, この分野では (理論を分からずとも) 集積性の検定に用いられ, 一種のゴールドスタンダード的なソフトになっている. このように新たな統計手法を考案した場合, その普及も含めた研究が必要となることもある.

制約条件つきの線形モデルにおけるLiu推定量について

東海大・理　鳥越　規央
東海大・理　道家　暎幸
東海大・教育研究所　氏家　勝巳

1. はじめに

これまではガウスマルコフモデル $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ におけるOLSEやRLSE，そしてあらゆるbiased estimatorとの比較についての研究をTrenkler([7])やAkdeniz and Kaciranlar([1])などが行ってきた．本研究ではパラメータベクトルを2つもつガウスマルコフモデル $(\boldsymbol{y}, \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \sigma^2\boldsymbol{V})$ を扱う．$\boldsymbol{X}_2$ の影響が無視できず，さらに1つのパラメータ $\boldsymbol{\beta}_1$ に制約条件がついた下で，biased estimatorの中のLiu estimatorがRLSEよりもMSE基準でよくなる条件について考察を行った．

2. 準備

$n \times 1$ 観測ベクトル $\boldsymbol{y}$，2つの $n \times p$ 説明変数行列 $\boldsymbol{X}_1$,$\boldsymbol{X}_2$，2つの $p \times 1$ パラメータベクトル $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$，$n \times 1$ 誤差ベクトル $\boldsymbol{\epsilon}$ による線形モデル

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

において $E(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$, $V(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$ を満たす $\boldsymbol{y}$ はモデル $\mathcal{M}_1 = (\boldsymbol{y}, \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \sigma^2\boldsymbol{V})$ に従うという．ここで $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^-\boldsymbol{X}_2'$, $\boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{P}_2$ によりモデル $\mathcal{M}_1$ は $\mathcal{M}_2 = (\boldsymbol{M}_2\boldsymbol{y}, \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)$ と書ける．ここで制約条件 $\boldsymbol{R}\boldsymbol{\beta}_1 = \boldsymbol{r}$ の下での $\boldsymbol{\beta}_1$ の推定量について比較を行う．今，$\boldsymbol{\beta}$ の推定量 $\tilde{\boldsymbol{\beta}}$ の評価については，MSE行列 $M(\tilde{\boldsymbol{\beta}}) = E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'$ を用いて下記のように定義する([7])．

$\boldsymbol{\beta}$ の推定量 $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2$ に対して、$\tilde{\boldsymbol{\beta}}_2$ が $\tilde{\boldsymbol{\beta}}_1$ よりも良い推定量.
$\overset{\text{def}}{\Longleftrightarrow}$ $M(\tilde{\boldsymbol{\beta}}_1) - M(\tilde{\boldsymbol{\beta}}_2)$ が非負定値行列.

3. Ordinary Least Square Estimator

モデル $\mathcal{M}_3 = (\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V})$ における最小2乗法による $\boldsymbol{\beta}$ の推定量 $\hat{\boldsymbol{\beta}}$ を求める．$\boldsymbol{V}$ は正定値行列より，$\boldsymbol{V} = \boldsymbol{W}'\boldsymbol{W}$ となる正則行列 $\boldsymbol{W}$ が存在する．これを用いて $\boldsymbol{W}^{-1}\boldsymbol{y} = \boldsymbol{z}, \boldsymbol{W}^{-1}\boldsymbol{X} = \boldsymbol{U}, \boldsymbol{W}^{-1}\boldsymbol{\epsilon} = \boldsymbol{e}$ とおくとモデル $\mathcal{M}_3$ は $(\boldsymbol{z}, \boldsymbol{U}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ となる．$\boldsymbol{S} = \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$ が正則ならば $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ であり、$\boldsymbol{S}$ が正則でないならば $\boldsymbol{S}\boldsymbol{S}^-\boldsymbol{S} = \boldsymbol{S}$ なる一般逆行列 $\boldsymbol{S}^-$ を用いて $\hat{\boldsymbol{\beta}} = \boldsymbol{S}^-\boldsymbol{U}'\boldsymbol{z} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ となる．これをOrdinary Least Square Estimator (**OLSE**) という．また $\hat{\boldsymbol{\beta}}$ は $\boldsymbol{\beta}$ の不偏推定量である．これを用いるとモデル $\mathcal{M}_2$ における $\boldsymbol{\beta}_1$ のOLSE $\hat{\boldsymbol{\beta}}_1$ は次のように表せる．

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{X}_1)^-\boldsymbol{X}_1'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{y}.$$

4. Restricted Least Square Estimator とその評価

$\boldsymbol{R}$ をランク m の $m \times p$ 行列、$\boldsymbol{r}$ を $m \times 1$ ベクトルとし、$\boldsymbol{R}, \boldsymbol{r}$ とも既知とする．ここで $\boldsymbol{\beta}_1$ について $\boldsymbol{R}\boldsymbol{\beta}_1 = \boldsymbol{r}$ なる制約条件を設ける．この条件の下での $\boldsymbol{\beta}_1$ の最小2乗推定量 $\boldsymbol{b}_1$ を求めると

$$\boldsymbol{b}_1 = \hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_1'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{X}_1)^-\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}_1'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{X}_1)^-\boldsymbol{R}')^-(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_1)$$

となる．これをRestricted Least Square Estimator (**RLSE**) と呼ぶ．なお $\boldsymbol{R}\boldsymbol{\beta}_1 = \boldsymbol{r}$ の条件の下ではRLSEは $\boldsymbol{\beta}_1$ の不偏推定量となる．Trenkler([7]) の結果をもとにRLSEとOLSEの

関係について，次の (1), (2) は同値であることが言える．
(1) $\boldsymbol{b}_1$ は $\hat{\boldsymbol{\beta}}_1$ よりも良い推定量．
(2) $\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\beta}_1 - \boldsymbol{r}$ とすると $\boldsymbol{\delta}'(\boldsymbol{R}(\boldsymbol{X}_1{}'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{X}_1)^{-}\boldsymbol{R}')^{-}\boldsymbol{\delta} \leq \sigma^2$．

5. Restricted Liu Estimator

Ridge 推定量 ([3],[4]) と Stein 推定量 ([6]) の互いの利点をあわせたような推定量

$$\hat{\boldsymbol{\beta}}_d = (\boldsymbol{S}+\boldsymbol{I})^{-1}(\boldsymbol{X}_1{}'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{y} + d\hat{\boldsymbol{\beta}}_1) \qquad (0 < d < 1)$$

が，Liu ([5]) によって提案された．ただし $\boldsymbol{S} = \boldsymbol{X}_1{}'\boldsymbol{M}_2(\boldsymbol{M}_2\boldsymbol{V}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2\boldsymbol{X}_1$ である．これをモデル $\mathcal{M}_2$ での Liu Estimator という．ここで $\boldsymbol{F}_d = (\boldsymbol{S}+\boldsymbol{I})^{-1}(\boldsymbol{S}+d\boldsymbol{I})$ とおくと $\hat{\boldsymbol{\beta}}_d = \boldsymbol{F}_d\hat{\boldsymbol{\beta}}_1$ となる．また $\boldsymbol{b}_{rd} = \boldsymbol{F}_d\boldsymbol{b}_1$ を $\boldsymbol{\beta}_1$ の Restricted Liu Estimator (**RLE**) という．$E(\boldsymbol{b}_{rd}) = \boldsymbol{F}_d\boldsymbol{\beta}_1 + \boldsymbol{F}_d\boldsymbol{S}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{\delta}$ であり，$\boldsymbol{\Sigma} = \boldsymbol{S}^{-1} - \boldsymbol{S}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{S}^{-1}$ とおくと $\mathrm{cov}(\boldsymbol{b}_{rd}) = \sigma^2\boldsymbol{F}_d\boldsymbol{\Sigma}\boldsymbol{F}_d'$ となる．ここで制約条件の下，RLE と RLSE を MSE を用いて比較すると，$\boldsymbol{b}_1$ と $\boldsymbol{b}_{rd}$ の MSE 行列はそれぞれ $M(\boldsymbol{b}_1) = \sigma^2\boldsymbol{\Sigma}$, $M(\boldsymbol{b}_{rd}) = \sigma^2\boldsymbol{F}_d\boldsymbol{\Sigma}\boldsymbol{F}_d' + (\boldsymbol{F}_d - \boldsymbol{I})\boldsymbol{\beta}_1\boldsymbol{\beta}_1{}'(\boldsymbol{F}_d - \boldsymbol{I})'$ となる．$\boldsymbol{S}$ は正定値行列より，$\boldsymbol{P}'\boldsymbol{S}\boldsymbol{P} = \boldsymbol{\Delta}$ となるような直交行列 $\boldsymbol{P}$ と正値対角行列 $\boldsymbol{\Delta}$ が存在する．いま $\boldsymbol{B} = \boldsymbol{P}'\boldsymbol{\Sigma}\boldsymbol{P}$ とおくと $\boldsymbol{\Sigma}$ は非負定値行列より $\boldsymbol{B}$ も非負定値であり，$\boldsymbol{B}$ の対角成分は全て非負である．さらに $\boldsymbol{\gamma} = \boldsymbol{P}'\boldsymbol{\beta}_1$ とおくと $\boldsymbol{\beta}_1 = \boldsymbol{P}\boldsymbol{\gamma}$ となる．ここで制約条件 $\boldsymbol{R}\boldsymbol{\beta}_1 = \boldsymbol{r}$ の下，$\boldsymbol{\beta}_1$ の 2 つの推定量 $\boldsymbol{b}_1$, $\boldsymbol{b}_{rd}$ について、次の (i), (ii) は同値であることを示した．

(i) $\boldsymbol{b}_{rd}$ は $\boldsymbol{b}_1$ より良い推定量．つまり

$$M(\boldsymbol{b}_1) - M(\boldsymbol{b}_{rd}) = \boldsymbol{P}(\boldsymbol{\Delta}+\boldsymbol{I})^{-1}(1-d)^2\left(\frac{\sigma^2}{1-d}\boldsymbol{E} - \boldsymbol{\gamma}\boldsymbol{\gamma}'\right)(\boldsymbol{\Delta}+\boldsymbol{I})^{-1}\boldsymbol{P}'$$

が非負定値行列である．ここで $\boldsymbol{E} = \boldsymbol{\Delta}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{\Delta} + (1+d)\boldsymbol{B}$ である．
(ii) $\boldsymbol{E}$ は非負定値であり、$\boldsymbol{\gamma}$ を $\boldsymbol{E}$ で生成されるベクトル空間に属するベクトルとし，$\boldsymbol{E}^-$ を $\boldsymbol{E}$ の一般逆行列（$\boldsymbol{E}\boldsymbol{E}^-\boldsymbol{E} = \boldsymbol{E}$ を満たす $\boldsymbol{E}^-$）とすると、

$$\boldsymbol{\gamma}'\boldsymbol{E}^-\boldsymbol{\gamma} \leq \frac{\sigma^2}{1-d} \qquad (0 < d < 1)$$

である．

参考文献

[1] Akdeniz, F. and Kaciranlar, S. (2001) More on the new biased estimator in linear regression, *Sankhya, Series B*,**63**, 321-325.

[2] Baksalary, J. K. and Kala, R. (1983) Partial orderings between matrices one of which is of rank one. *Bulletin of the Polish Academy of Sciences, Mathematics*, **31**,5-7.

[3] Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.

[4] Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: application for non-orthogonal problems. *Technometrics*, **12**, 69-82.

[5] Liu, Ke Jian(1993) A new class of biased estimate in linear regression, *Communications in Statistics, Series A*,**22**,393-402.

[6] Stein, C.(1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**,197-206.

[7] Trenkler, G. (1987) mean square error martix comparisons among restricted least squares estimators, *Sankhya, Series A*, **49**, 96-104.

差分フィルタを用いたレーダー受信波解析

慶應大・理工・院　青木義充
慶應大・理工　　加藤　剛

1　はじめに

人工衛星に搭載されている地表面画像取得用レーダーについて考える.「地表面を覆う雲」または「衛星からの距離が非常に長い」など様々な理由により，光学カメラによる撮影では満足する地表面画像を得ることは難しい．そのため，レーダーを用いて地表面の様相を観測する必要がある．レーダーによる地表面画像の取得とは，レーダーから送信された電磁波は，地表面上で反射する際，その振幅が地表の状態 (海面，草地，コンクリートなど) 固有の反射率で変化する．この現象を利用して，受信された電磁波の振幅の変化から地表面反射率の情報を抽出することにより，地表面画像を再生する．

従来までの議論では，地表面を，多数の「固有の反射率を持った点散乱体」の集合として仮定してきた．この仮定のもとでは，受信波は各点散乱体からの反射波の重ね合わせ（離散和）として定義される．しかし，実際には地表面がそのようになっているとは考えにくく，散乱体が連続的に存在していると仮定した方が自然である．この場合，連続の濃度を持つ散乱体からの合成波として得られる．本報告では，この受信波にもとづいた画像再生手法を提案するとともに，従来より利用されている画像再生法である「パルス圧縮法」との比較を行なっている．また，本報告では議論を簡便化するために対象となる地表面が１次元であることを仮定している．

なお，本報告における各パラメタの値は，日本で開発され実際に運用されてきた衛星 $JERS-1$ に搭載された合成開口レーダーのものを使用している．

2　地表面のモデル化と受信波

今回，送信する電磁波として，地表面画像取得を目的とした合成開口レーダーで一般的に用いられているものを利用する．その特徴としては次の二点が挙げられる．第一に，送信するパルスは瞬間的（短い時間）のものではなく，ある程度の幅（送信幅）を持ったパルスであること．第二に，従来から行われている手法であるパルス圧縮を実現させるために，周波数の変調が行われていることである．

次に地表面反射率について考える．地表面上に衛星の進行方向と垂直に y 軸をとり，その原点を衛星の真下におく．点散乱体は y 軸上に存在すると仮定し，反射率は y についての非負実数値関数として定義する．従来までの議論では，点散乱体が y 軸上に離散的に存在することを仮定している．この仮定のもとでは，受信波は各点散乱体からの反射波の重ね合わせとして定義される．しかし，地表面画像を対象として考えたときには，従来の定義は妥当なものとは言えない．なぜならば，地表面上の反射点は観測幅内に連続に存在すると考えた方が自然だからである．例として，海面，草地，建造物など，対象とするものはそれぞれ固有の大きさ（広がり）を持っており，一点としては考えにくい．したがって，時刻 t で観測される受信波は，連続的な合成を考慮した積分形で考えた方がより厳密な定義となる．本報告では，このような考えをもとに，地表面反射率をある区間ごとに変化する関数（階段関数）として仮定する．この仮定のもとでは，受信波は，区間の変化する点からの波形の部分と単調に変化する部分を合わせ持った形として得られる．

3　地表面反射率の復元

地表面反射率を復元するためには,「変化点の位置」とそこでの「変化量」の2種類の情報が必要である．そこで，我々は受信波に対してフィルタ演算を行なうことで情報の抽出を実現している．

3.1　変化点の位置特定

複数の変化点が互いに遠い場合には，特に何もしないでもそれぞれの点を区別することができるが，変化点同士の距離が近い場合には，それぞれの点からの反射波が互いに重なりあい，各変化点の特定が困難になる．特に送信するパルス幅が地表面上の距離に換算した場合に決して小さくはない(地表面上の距離にして約8[km]) ので，変化点が重なりあうということは頻繁に起こるものと考えてよい．

この問題を解消する工夫として，現在実用化されており，代表的な方法がパルス圧縮技術である．この技術を利用することにより，異なる変化点間の距離がかなり近い場合（地表面上の距離にして10[m] 程度) でも，その変化点を特定することが可能になる．しかし今回の発表において，我々は新しい特定方法を提案する．その変化点の特定方法は，従来行われているパルス圧縮よりも次の二点において優れている．

1. 非常に密接した二点の変化点の特定が可能であること．(地表面上の距離にして 20[cm] 程度)
2. 地表面反射率の変化量の大小に変化点を特定する能力が左右されないという意味で頑健であること．

この主張の正当性はシミュレーションによる対比実験により確認されている．

3.2　変化量の復元

第1節に記したように，衛星レーダーによる地表面画像の取得では，地表面反射率の情報を受信波より抽出する必要がある．ところが，パルス圧縮技術を用いた従来の解析方法では，実は地表面反射率を受信波から特定することができない．実際の現場では，地表面に関する事前情報との照合をはじめとする地道な手作業によって，この欠陥を補っている．しかし，今回定式化した受信波とそのモデルを利用することにより，受信波から地表面反射率を直接特定することができる．

地表面反射率の変化を表す関数を復元するために，変化点付近での積分を考える．この式も変化点を特定するために定義した式と同様に，モデルを通して受信波の挙動を吟味し，地表面反射率の情報をうまく抽出できるように定義している．実際，各変化点の値を代入することにより，その変化点での跳躍量を高い精度で得ることができる．また，各変化点での地表面反射率の差が符号付き（位相情報付き）で得られるので，反射率の増減を表すことが可能である．この結果と先述の変化点特定の式を用いて得られている変化点の位置情報をあわせることにより，地表面反射率を表す関数を復元できる．

実用上は，変化点を特定する際に多少のずれが生じる可能性があるが，そのずれを許容した形の跳躍量特定の式も合わせて定義している．また，このときに予想される変化点のずれは地表面上の距離で20[cm] 程度に抑えられている．これらの結果についてもシミュレーションによる数値実験を行っている．

Wavelet - Vaguelette 分解による非定常雑音の処理

慶應大・理工　　加藤　剛
慶應大・理工・院　青木義充

1　目的

人工衛星のレーダーで地表面画像を取得する問題を考える．人工衛星が発した送信波は，照射された地表面の状態に応じて振幅が変化した反射波となり，人工衛星で受信される．観測時刻 t における受信波は，次の式が自然なモデル化となる．

$$H(t) = e^{2\pi i\theta_0}\int_{t-T}^{t} A(\xi(u))\, e^{2\pi i f_0(t-u)}\xi'(u)\, du + \sigma B_H(t) =: G(t) + \sigma B_H(t), \quad t > \frac{2h}{c} + T. \tag{1}$$

$B_H(t)$ は標準 fractional Brownian motion (fBm)，$\sigma > 0$ は雑音の水準，$f_0 > 0$ は送信波の周波数，θ_0 は初期位相を表す．また，$\xi(t) = \sqrt{c^2 t^2/4 - h^2}$ で ξ' は ξ の導関数，c は光速，h は衛星高度である．地表面の状態は関数 $A(\cdot)$ として表現される．したがって，$H(t)$ の観測データから地表面画像を取得する作業は，関数 $A(\cdot)$ の推定を行うことに他ならない．

受信信号の自然なモデル化である (1) は，次の 2 点において扱いの困難さを伴う．

1. 雑音が非定常であること
   IEEE の画像処理に関するシリーズ，掲載論文では，観測時刻ごとに互いに独立で，同一分布（正規分布）にしたがう確率変数を雑音のモデルとして仮定していることが多い（雑音そのものを考慮に入れていないものも少なからずある）．けれども，現実には，雑音は各時刻に依存すると考える方が自然なので，fBm を仮定した．ところが，fBm は非定常なので，フーリエ解析の枠組みでは扱いきれない．
2. データが非直接的であること
   関数 $A(\cdot)$ についての情報は，区間 $[t-T, t]$ 上の積分値としてしか得られない．一般に，推定対象に関するデータが変換された形でしか観測できない場合，そのデータを非直接的データと呼ぶ．

問題を簡単にするために，地表面の状態を区画状にモデル化する．そのとき，$A(\cdot)$ は階段関数

$$A(y) = \rho_0 I_{(-\infty,y_1]}(y) + \sum_{k=1}^{n-1}\rho_k I_{(y_k,y_{k+1}]}(y) + \rho_n I_{(y_n,\infty)}(y) \quad (\rho_k > 0)$$

で与えられる．よって，この場合，$A(\cdot)$ の推定は不連続点 $\{y_k\}$ と各不連続点間の関数値 $\{\rho_k\}$ を推定することに帰着される．さらに，$t_k = \xi^{-1}(y_k)$, $S(t) = \{t_k : t-T < t_k < t\} \cup \{t, t-T\}$ とおくとき，精度のよい近似

$$G(t) \approx e^{2\pi i\theta_0}\sum_{s,s'\in S(t)} \tilde{A}(s)\int_s^{s'} e^{2\pi i f_0(t-u)}\xi'(u)\, du, \quad \tilde{A}(\cdot) = A(\xi(\cdot)),$$

の成り立つことがわかっている．したがって，$\tilde{A}(\cdot)$ の不連続点 $\{t_k\}$ が検出できさえすれば，例えば回帰モデルによって $\{\rho_k\}(=\{\tilde{A}(t_k)\})$ を推定することが可能になる．つまり，推定の要は，

不連続点の検出である.

2 結果

不連続点の検出は, wavelet 変換の得意分野である. しかも, wavelet 変換は fBm とも相性がよい. ところが, (1) にもとづく不連続点の検出では, 推定対象の非直接性のために, wavelet 変換それだけでは有効な道具にはなりえない. wavelet 変換の応用である wavelet-vaguelette 分解を利用することによって, 目的を果たす結果を導くことができる.

vaguelette 変換を行う関数族 $\{v_{a,b}(t) : a > 0, b \in \boldsymbol{R}\}$ を

$$v_{a,b}(t) = \psi'_{a,b}(t) + 2\pi i f_0\, a\, \psi_{a,b}(t) \tag{2}$$

にとる. ここで, ψ は台が区間 $[-r, r]$ に含まれるような mother wavelet 関数, ψ' はその導関数である. 本来, vaguelette 変換に用いる関数族 $\{v_{a,b}(t)\}$ は 1 つの関数 v から $v_{a,b}(x) = a^{-1/2} v((x-b)/a)$ によって生成されるが, ここで用いる $v_{a,b}(t)$ はそうではない. この点が通常の vaguelette 変換の場合とやや異なる.

$H(t)$ の vaguelette 変換は

$$(VH)(a,b) = \int_{-\infty}^{\infty} G(t)\, v_{a,b}(t)\, dt + \int_{-\infty}^{\infty} v_{a,b}(t)\, B_H(t)\, dt = (VG)(a,b) + \sigma(VB_H)(a,b)$$

である. このとき, 次の命題を示すことができる.

命題 1 u_0 を $\tilde{A}(\cdot)$ の 1 つの不連続点とし, ちょうど T だけ離れたところに別の不連続点はないと仮定する. このとき, 次が成り立つ.

(i) 任意の $p \in (0,1)$ に対し, $|b - u_0| \le pra$ または $|b - u_0 - T| \le pra$ ならば, ある $M_p \neq 0$ が存在して, $\lim_{a \to 0} a^{-3/2} (VG)(a,b) = M_p$.

(ii) $u_0 \notin (-ar + b,\, ar + b) \cup (-ar + b - T,\, ar + b - T)$ ならば, ある $M > 0$ が存在して, $\lim_{a \to 0} a^{-5/2} |(VG)(a,b)| \le M$.

命題 2 任意の $\beta \in \boldsymbol{R}, \varepsilon > 0$ に対し, ある $M_\varepsilon > 0$ が存在して, $b \in \boldsymbol{R}$ について一様に

$$P\left\{ |(VB_H)(a,b)| \ge \varepsilon\, \sigma^\beta \right\} \le M_\varepsilon (\sigma^{-\beta} a^{H+1/2})^2.$$

尺度母数 a を $a = \sigma^q$ という形で雑音の水準 σ に依存させ, $q \in (0, \frac{1}{1+\theta-H})$, $\theta > 0$, と選ぶ. $\beta = (\frac{3}{2} + \theta)q - 1$ にとると, 命題 2 より, $\sigma(VB_H)(a,b)$ のオーダーは b について一様に $O_p(\sigma^{(3/2+\theta)q})$ 以下になる. 他方, 命題 1 から, $(VG)(a,b)$ について b または $b - T$ が不連続点近傍にあるときのオーダーは $O(\sigma^{3q/2})$, そうでないときのオーダーは $O(\sigma^{5q/2})$ 以下である. したがって, $a \to 0$ のときに

$$(VH)(a,b) \text{ のオーダー} = \begin{cases} O_p(a^{3/2}) & : \ b, b - T \text{ が不連続点の近傍} \\ O_p(a^{5/2}) \text{ または } O_p(a^{3/2+\theta}) \text{ 以下} & : \ \text{その他} \end{cases}$$

なので, 適当な敷居値を設定して b を動かせば, 理論的には $A(\cdot)$ の不連続点が見つかることになる. ただし, T だけずれたところにも余分に検出される.

Bandwidth selection for kernel smoothing in binomial regression

中国短大　　奥村英則
島根大・総合理工　　内藤貫太

1. kernel 推定量　用量反応曲線の推定を考える. K 個の用量 $x_1, \ldots, x_K$ の各 x_i で N_i 個の固体中反応した個体の数 Y_i が観測されたとする. ここですべての用量で個体の反応は独立であるとする. 反応の数 Y_i, $i = 1, \ldots, K$ はパラメータ $p_i = p(x_i)$ をもつ2項分布 $Bi(N_i, p_i)$ に従うとする. 用量反応曲線 $p(x)$ に関して知見がない場合には探索的にノンパラメトリックアプローチが使用される. 奥村・内藤 (2002) は各用量での分散不均一性を考慮した重み付き kernel 推定量

$$\hat{p}(x;h) = \frac{\sum_{i=1}^{K} Y_i w_i(x)/\hat{v}_i}{\sum_{i=1}^{K} N_i w_i(x)/\hat{v}_i}$$

を提案し, その推定量の挙動について報告した. ここで, $\hat{v}_i = \hat{p}_i(1-\hat{p}_i), \hat{p}_i = (\bar{Y}_i + \sqrt{N_i}/2)/(N_i + \sqrt{N_i})$, $w_i(x) = \phi(h^{-1}(x_i - x))/h$ であって, $\phi(x)$ はある滑らかな kernel, h はバンド幅である. 議論を簡単にするために, 用量 x_i は $x_i = (i-1)/(K-1)$, $i = 1, \ldots, K$ であるとし、N_i はすべて等しいとする. ある正則条件の下で, $\hat{p}(x;h)$ のバイアスは $O(h^2)$ より収束が遅い $O(N_1^{-1})$ の項が現れるので, その項を消去するように $\hat{p}(x;h)$ のバイアス修正推定量

$$\tilde{p}(x;h) = \frac{1}{N_1} + (1 - 2\frac{1}{N_1})\hat{p}(x;h)$$

が導出された. このとき, $\tilde{p}(x;h)$ の MSE は

$$\mathrm{MSE}[\tilde{p}(x;h)] \simeq h^4 \mu_2(\phi)^2 f(x)^2 + \frac{v(x)R(\phi)}{N_1 K h}$$

で与えられる. ここで, $\mu_2(\phi) = \int_{-1}^{1} z^2\phi(z)dz$, $R(\phi) = \int_{-1}^{1} \phi(z)^2 dz$, $f(x) = \{p^{(2)}(x) - 2(1 - 2p(x))p^{(1)}(x)^2 v(x)^{-1}\}/2$, $v(x) = p(x)(1-p(x))$ である. 本報告では, 推定量 $\tilde{p}(x;h)$ に含まれるバンド幅 h をデータに基づき自動的に決定する Plug-in 法のアルゴリズムを提案し, その性能について考察を与える.

2. Plug-in 法　ある正数 $\delta(0 < \delta << 1)$ に対して, バイアス修正推定量 $\tilde{p}(x;h)$ の 区間 $[\delta, 1-\delta]$ 上での Mean Integrated Squared Error(MISE) の最小化によって 最適バンド幅 h_{opt}

$$h_{\mathrm{opt}} = C(\phi)\left(\frac{\theta_2}{\theta_1}\right)^{\frac{1}{5}} (N_1 K)^{-1/5} \tag{1}$$

が得られる. ここで $C(\phi) = (R(\phi)/\mu_2(\phi)^2)^{1/5}$, $\theta_1 = \int_{\delta}^{1-\delta} f(x)^2 dx$, $\theta_2 = \int_{\delta}^{1-\delta} v(x)dx$ である。このとき $p(x)$ に依存する θ_1 と θ_2 の推定が必要である. θ_1 の推定量を構成するために, $p(x)$ の簡便推定量

$$\bar{p}(x;g) = \frac{1}{K}\sum_{i=1}^{K} w_i(x)\bar{Y}_i$$

を使用する. これから $f(x)$ の一致推定量

$$\bar{f}(x;g) = \frac{\bar{p}^{(2)}(x;g)}{2} - \frac{(1 - 2\bar{p}(x;g))\bar{p}^{(1)}(x;g)^2}{\bar{p}(x;g)(1-\bar{p}(x;g))}$$

が得られる． θ_1 の推定量として

$$\bar{\theta}_1(g) = \int_{\delta}^{1-\delta} \bar{f}(x;g)^2 dx$$

を使用する．ある正則条件の下で，$\bar{\theta}_1(g)$ の MSE は

$$\mathrm{MSE}[\bar{\theta}_1(g)] \simeq \left[g^2 \Delta_1(p^{(1)}, p^{(2)}, p^{(3)}, p^{(4)})) + \frac{\Delta_2(p)}{N_1 K g^5} \right]^2 + \frac{\Delta_3(p)}{N_1 K^2 g^9} \tag{2}$$

で与えれる．ここで $\Delta_1(p, p^{(1)}, p^{(2)}, p^{(3)}, p^{(4)})$, $\Delta_2(p) > 0$, $\Delta_3(p) > 0$ はそれぞれ括弧内に明示された関数に依存する定数である．ここで，$\bar{\theta}_1(g)$ の分散の主項である式 (2) の右辺第 2 項はより速く落ちることがわかる．従って，最適なバンド幅 g_{opt} は，バイアス 2 乗項に注目して

$$g_{\mathrm{opt}} = \begin{cases} \left(\dfrac{\Delta_2}{\Delta_1}\right)^{1/7} (N_1 K)^{-1/7}, & \Delta_1 < 0, \\ \left(\dfrac{5\Delta_2}{2\Delta_1}\right)^{1/7} (N_1 K)^{-1/7}, & \Delta_1 > 0 \end{cases} \tag{3}$$

で与えられる．一方，θ_2 の推定量も $\bar{p}(x;h)$ を使用して構成できるが，それに含まれるバンド幅の選択を MISE 基準で行なうと，幾分煩雑になる．θ_2 の推定量として

$$\bar{\theta}_2 = \frac{1}{K^*} \sum_{j=0}^{6} \frac{1}{N_1^j} {\sum_i}^* (\bar{Y}_i - \bar{Y}_i^2)$$

を採用する。ここで K^* は $[\delta, 1-\delta]$ に含まれる x_i の個数，${\sum_i}^*$ は，$[\delta, 1-\delta]$ に含まれる x_i に関数する和を表す．g_{opt} が導出される同じ条件の下で，$\bar{\theta}_2 - \theta_2 = O_p((N_1 K)^{-1/2})$ が成り立つ．$\bar{\theta}_2(k) = (K^*)^{-1} \sum_{j=0}^{k-1} N_1^{-j} {\sum_i}^* (\bar{Y}_i - \bar{Y}_i^2)$ とおく．より一般に $N_1 \to \infty, K \to \infty$ のとき，$K/N_1^k = O(1)$ または $K/N_1^k = o(1)$ となる k が存在すれば，$\bar{\theta}_2(k) - \theta_2 = O_p((N_1 K)^{-1/2})$ となることに注意する．ゆえに，式 (1) から h_{opt} の選択は

$$\bar{h}_{\mathrm{opt}} = C(\phi) \left(\frac{\bar{\theta}_2}{\bar{\theta}_1(g_{\mathrm{opt}})} \right)^{\frac{1}{5}} (N_1 K)^{-1/5}$$

に基づいて実行できる．また，$\bar{h}_{\mathrm{opt}}$ の相対誤差は

$$\frac{\bar{h}_{\mathrm{opt}}}{h_{\mathrm{opt}}} - 1 = O_P\left(\frac{1}{(N_1 K)^{2/7}}\right)$$

で与えられる．$\bar{h}_{\mathrm{opt}}$ を代入して得られる推定量 $\tilde{p}(x;\bar{h}_{\mathrm{opt}})$ は，一致性と漸近正規性をもつ．

実際に g_{opt} を使用するときには、Δ_1, Δ_2, Δ_3 に含まれるの未知の部分を推定する必要がある．各 Δ_i の推定には，Rule of Thumb(ROT) を適用する．すなわち，$p(x)$ の初期推定量として，パラメトリック推定量 $p(x;\hat{\beta})$ を使用する．ここで $\hat{\beta}$ はパラメータ β の推定量である．$p(x)$ の導関数 $p^{(i)}(x)$ も $p^{(i)}(x;\hat{\beta})$ で置き換える．パラメトリック推定量を式 (3) に代入して得られる g_{opt} の推定量を $\hat{g}_{\mathrm{opt}}$ とする．求める h の最適バンド幅は

$$\tilde{h}_{\mathrm{opt}} = C(\phi) \left(\frac{\bar{\theta}_2}{\bar{\theta}_1(\hat{g}_{\mathrm{opt}})} \right)^{\frac{1}{5}} (N_1 K)^{-1/5}$$

で与えられる．

3. シミュレーション 最適バンド幅 $\tilde{h}_{\mathrm{opt}}$ の挙動に関するシミュレーションの結果について報告した．また提案する選択法を実データへの適用し，その実用性を示した．

ある形に配置された二値独立試行列における連の数の分布

舟尾 暢男

大阪大学大学院基礎工学研究科
システム人間系数理科学分野

成功(S)か失敗(F)かの二値独立試行列中に特定の長さの成功連が決まった数だけ起こる分布について考えていく．連の数え方は連同士が重なり合っている場合も複数個の連が出来ているとするoverlappingな数え方をすることにする．overlappingな数え方では，例えばSSSSFSSSSSSという系列で長さ3の成功連に関していえば6個あることになり，それぞれ3回目，4回目，8回目，9回目，10回目，11回目に起こっていることになる．

本稿では特定の長さの成功連を数える二値独立試行列の並び方は直線的であるとは限定せず，円の様に確率変数の試行列の両端が繋がっているような並び方，確率変数の試行列が途中で交差するような複雑な並び方をしている場合にも分布を求めることが出来る様なアルゴリズムを提案した．例えば確率変数を8の字状に並べた場合，以下の様な連のでき方が考えられる．

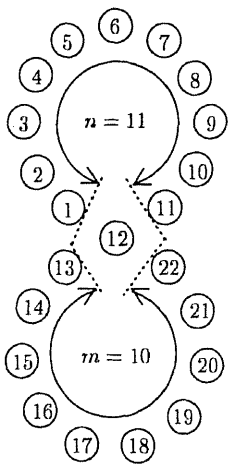

図 1: $n = 11$, $m = 10$, $k = 4$ の場合の試行列

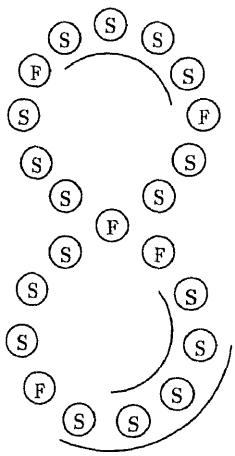

図 2: 長さ4の成功連が3個出来ている場合の例

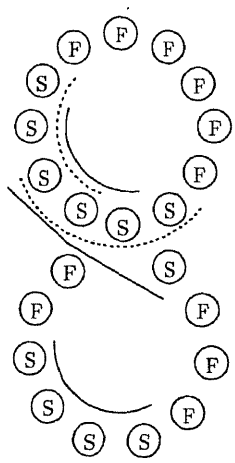

図 3: 長さ4の成功連が5個出来ている場合の例

具体的には以下のような流れで話を進めた.

(1) 特定の長さの成功連が決まった数だけ起こる分布についてのこれまでの研究の流れでは，確率変数の並べ方は直線上及び円状に限られた場合での計算アルゴリズムしか提案されてこなかったが，本稿で紹介するアルゴリズムを用いることで確率変数を複雑な形に並べた場合でも計算できることを述べた.

(2) 本稿での連の数え方を，確率変数が円状に並んだ場合，8の字状に並んだ場合の2つを例に挙げて説明した.

(3) 直線的に配置された二値独立試行列中において特定の長さの成功連が決まった数だけ起こる分布を求めることで，問題を解く為の道具である条件付き確率母関数とマーカーの使い方をみた.

(4) 確率変数の並び方が直線的でない場合においても，特定の長さの成功連の数の分布を比較的容易に求めることが出来るアルゴリズムを提案した．ここでこれまでに求められることが無かった，確率変数を8の字状に並べた場合での成功連の数の分布を求めた.

条件付き確率母関数とマーカー を用いる方法を提案したのは Ebneshahrashoob and Sobel (1990) で，ここではさまざまな条件付き確率母関数間の関係式を導き出し，それぞれの条件付き確率母関数自体を未知関数とする連立方程式を解いて確率母関数を得る方法を提案している．さらに本稿では，確率母関数間の関係式中に特定の離散パターンが出たことを表すマーカーという文字定数を入れることで，それぞれの条件付き確率母関数自体を未知関数とする連立方程式が解けて注目すべき離散パターン (例えば特定の長さの成功連) に関する分布の確率母関数が得られた時に，その文字定数に係っている式自体が即，特定の離散パターンが出たという部分の確率母関数になるのである．本稿では，この論文で提案された方法を改良することで，独立試行列が複雑な並び方をしている場合においても正確な分布を求めることを可能にする条件付き確率母関数とマーカーの使い方を提案した.

参考文献

[1] Ebneshahrashoob, M. and Sobel, M. (1990). Sooner and later waiting time problems for bernoulli trials: frequency and run quotas, *Statistics & Probability Letters*, **9**, 5-11.

A generalized Pólya urn model and related distributions

学振特別研究員 (統計数理研究所) 井上 潔司
関西大学工学部 教養数学教室　　安芸 重雄

1 はじめに

壷の中にラベル "0" の玉が α_0 個, ラベル "1" の玉が α_1 個, ..., ラベル "m" の玉が α_m 個入っているとする. 本報告では, 壷の状態を $\boldsymbol{b} = (\alpha_0, \alpha_1, ..., \alpha_m)$, 玉の総数を $|\boldsymbol{b}| = \alpha_0 + \alpha_1 + \cdots + \alpha_m$ と表す. この壷の中から random に一つの玉を取り出し, そのラベルを確認して元に戻す. この時, addition matrix $A = (a_{ij})$ $i, j = 0, 1, ..., m$ に従ってさらに玉を追加する. これを一つの試行とする. このような試行を n 回繰り返すものとする. ここで, 行列の row は取り出した玉の種類を, column は追加する玉の種類を表しており, 行列の各成分は非負整数とする. これは, もし, ラベル "i" の玉が抽出されたとき, ラベル "j" の玉を a_{ij} $(j = 0, 1, ..., m)$ 個追加することを表している. 本報告では, n 回の抽出における長さ k_i の "i"-連の数の同時分布を 4 つの異なる数え方に基づいて考察する.

今までの研究では, 数学的な便宜上のため, 各試行において, 追加する玉の数が等しい $(a_{i1} + a_{i2} + \cdots + a_{im} = a_{j1} + a_{j2} + \cdots + a_{jm}, i \neq j, i, j = 0, 1, ..., m)$ という制約が課せられた場合での考察がほとんどである. また, この制約を外しての Pólya の壷モデルの厳密な扱いは困難であることも報告されている (Johnson and Kotz (1977), Kotz *et al.* (2000) 参照). ここでは, 制約の有無に関わらず, 厳密な分布の導出が可能であることを示す. 最後に幾つかの数値例を与え, 本報告における結果が計算機を用いて実行可能であることを示す.

2 Notation

長さ k の連を数える際に 4 つの異なる数え方 (Fu and Koutras (1994) 参照) を採用する. つまり, 重複しないで数える数え方, 長さ k 以上の連を数える数え方, 重複して数える数え方, ちょうど長さ k の連を数える数え方を用い, それぞれを Type I, II, III, IV の数え方と呼ぶことにする. $X_1, X_2, ..., X_n$ は n 回の試行によって得られるラベルの列とし, この列の中に現れる長さ k_i の "i"-連の数 (Type β_i で数えたとき) を $N(n, k_i; \beta_i)$ $(i = 1, 2, ..., m)$ で表す. また, 次の関数を用意する.

$$\boldsymbol{\mu}(j; \boldsymbol{\beta}) = (\mu_1(j; \beta_1), ..., \mu_m(j; \beta_m)) \tag{2.1}$$

$$\mu_i(j; \beta_i) = \begin{cases} \left[\frac{j}{k_i}\right] & \beta_i = I, \\ I(j \geq k_i) & \beta_i = II, \\ (j - (k_i - 1))^+ & \beta_i = III, \\ I(j = k_i) & \beta_i = IV, \end{cases}$$

ここで, $(j - (k_i - 1))^+ = \max\{0, (j - (k_i - 1))\}$, $I(u)$ は u の indicator function である.

本報告では, $(N(n, k_1; \beta_1), ..., N(n, k_m; \beta_m))$ の厳密な分布を条件付確率母関数を用いて考察する.

3 Results

壷の初期状態を $\boldsymbol{b}_0 = (\alpha_{01}, \alpha_{02}, ..., \alpha_{0m})$ と仮定し, $(N(n, k_1; \beta_1), ..., N(n, k_m; \beta_m))$ の確率母関数を $\phi_n(\boldsymbol{b}_0, \boldsymbol{t}\ ; \boldsymbol{\beta})$ と表す. ここで, $\boldsymbol{t} = (t_1, ..., t_m)$ とする. 次に, $X_1 = s$ $(s = 0, 1, ..., m)$ かつ, そのときの壷の状態が $\boldsymbol{b} = (\alpha_1, \alpha_2, ..., \alpha_m)$ と仮定する. この条件の下で $X_1, X_2, ..., X_n$ の中に現れる長さ k_i の "i"-連の数についての条件付確率母関数を $\phi_{n-1}^{(i)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta})$ $i = 0, 1, ..., m$ と表す.

Theorem 3.1 確率母関数 $\phi_n(\boldsymbol{b}_0, \boldsymbol{t}\ ; \boldsymbol{\beta})$, $\phi_n^{(i)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta})$ $i = 0, 1, ..., m$ は, 次の線形方程式をみたしている;

$$(3.1)\quad \phi_n(\boldsymbol{b}_0, \boldsymbol{t}\ ; \boldsymbol{\beta}) = \sum_{i=0}^{m} \frac{\alpha_{0i}}{|\boldsymbol{b}_0|} \phi_{n-1}^{(i)}(\boldsymbol{b}_0 + \boldsymbol{a}_i, \boldsymbol{t}\ ; \boldsymbol{\beta}) \qquad n \geq 1,$$

$$(3.2)\quad \phi_0(\boldsymbol{b}_0, \boldsymbol{t}\ ; \boldsymbol{\beta}) = 1,$$

$$(3.3)\quad \phi_n^{(0)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta}) = \sum_{i=0}^{m} \frac{\alpha_i}{|\boldsymbol{b}|} \phi_{n-1}^{(i)}(\boldsymbol{b} + \boldsymbol{a}_i, \boldsymbol{t}\ ; \boldsymbol{\beta}) \qquad n \geq 1,$$

$$(3.4)\quad \phi_0^{(0)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta}) = 1,$$

$$(3.5)\quad \phi_n^{(i)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta}) = \sum_{i' \neq i} \sum_{j=0}^{n-1} \frac{\alpha_i^{[j, a_{ii}]}}{|\boldsymbol{b}|^{[j, |\boldsymbol{a}_i|]}} \frac{\alpha_{i'} + j a_{ii'}}{|\boldsymbol{b}| + j|\boldsymbol{a}_i|} t_i^{\mu_i(j+1; \beta_i)} \phi_{n-j-1}^{(i')}(\boldsymbol{b} + j\boldsymbol{a}_i + \boldsymbol{a}_{i'}, \boldsymbol{t}\ ; \boldsymbol{\beta})$$
$$+ \frac{\alpha_i^{[n, a_{ii}]}}{|\boldsymbol{b}|^{[n, |\boldsymbol{a}_i|]}} t_i^{\mu_i(n+1; \beta_i)} \qquad n \geq 1,\ \ \beta_i = I, II, III, IV,\ i = 1, 2, ..., m,$$

$$(3.6)\quad \phi_0^{(i)}(\boldsymbol{b}, \boldsymbol{t}\ ; \boldsymbol{\beta}) = t_i^{\mu_i(1; \beta_i)} \qquad i = 1, 2, ..., m,$$

ここで, $a^{[x,c]} = a(a + c) \cdots (a + (x - 1)c)$, $a^{[0,c]} = 1$, $|\boldsymbol{b}_0| = \sum_{i=0}^{m} \alpha_{0i}$, $|\boldsymbol{b}| = \sum_{i=0}^{m} \alpha_i$, $\boldsymbol{a}_i = (a_{i0}, ..., a_{im})$, $|\boldsymbol{a}_i| = \sum_{j=0}^{m} a_{ij}$.

参考文献

Inoue, K. and Aki, S. (2001), Pólya urn models under general replacement schemes, *Research Report on Statistics,* **54**, Osaka University.

Inoue, K. (2003). Generalized Pólya urn models and related distributions, *Proceedings of the Symposium, Research Institute for Mathematical Science, Kyoto University*, **1308**, 29-38.

Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs : a Markov chain approach, *J. Amer. Statist. Assoc.,* **89**, 1050–1058.

Johnson, N. L. and Kotz, S. (1977), *Urn Models and Their Applications*, Wiley, New York.

Kotz, S., Mahmoud, H. and Robert, P. (2000), On generalized Pólya urn models, *Statist. Probab. Lett.,***49**, 163–173.

Stepwise smoothing 公式を利用した離散確率の計算法

関西大学　工学部　安芸 重雄

最近，確率変数の系列や配列，あるいはさまざまなグラフの頂点に対応する確率変数族などのランダムな構造体の上で，ある事象が起こるまでの待ち時間の確率分布や事象の生起数についての確率を計算することが多くなってきた（たとえば，工学的 consecutive systems の信頼性や DNA sequence における matching の問題や連に基づいた品質管理の問題など）．このような離散分布の確率を求める問題は，古くから組合せの方法を使って研究されてきた．しかし，この方法は，注目する事象が複雑になったり，確率変数の系列自体が依存性をもつようになると適用が困難になる．

この報告では，条件付期待値についての stepwise smoothing 公式を利用するという観点から，条件付確率母関数法を説明し，この方法をさらに一般化して，さまざまな場面で確率の計算が可能になることを示した．

$\mathbf{N}_0 = \{0, 1, 2, \cdots\}$ とし，X を確率空間 $(\Omega, \mathcal{F}, P)$ 上で定義された $\mathbf{N}_0$-値確率変数とする．X の分布は離散分布だから，確率母関数 $E[t^X]$ が求まれば X の分布についてはすべてわかる．さらに，$\mathcal{F}$ の sub σ-fields $\mathcal{F}_0, \mathcal{F}_1$ を考え，$\mathcal{F}_0 \subset \mathcal{F}_1$ を仮定するとき，条件付期待値についての stepwise smoothing 公式より，$E[t^X|\mathcal{F}_0] = E[E[t^X|\mathcal{F}_1]|\mathcal{F}_0]$ が成り立つ．この式により，さまざまな条件付確率母関数間の関係が記述され，X の確率母関数を求めるためのアルゴリズムを提案することができる．

現在，研究の中心になっている，分布論の問題に対する適用例を次のように整理することができる．

(1) 2つ以上の事象を同時に考える場合．

たとえば，2 種類の事象の数をそれぞれ X_1, X_2 とするとき，同時確率母関数は $E[t_1^{X_1} t_2^{X_2}]$ だから，$Y = t_1^{X_1} t_2^{X_2}$ として条件付き期待値を計算すればよい．また，2 種類の事象の待ち時間をそれぞれ W_1, W_2 とするとき，sooner event の待ち時間 $\sigma = \min\{W_1, W_2\}$ や later event の待ち時間 $\tau = \max\{W_1, W_2\}$ の確率母関数を求めるときには，マーカー付きの確率母関数を求めるのが便利である．

(2) より複雑な事象を対象にする場合．

今では多くの研究者の興味は，連の分布から一般の有限パターンやさまざまな scan statistics の標本分布へと広がっている．技術的には，sub σ-field のとり方に工夫が必要になる．

(3) 観測する確率変数列の配置を複雑にする場合．

さまざまな consecutive systems においては，確率変数の配置が問題になることが多い．確率変数が円形に並んだ場合をはじめとして，多次元格子点上や，グラフの頂点上に確率変数が配置される場合も研究の対象になっている (Aki and Hirano (2003))．

(4) 配置された確率変数の集合に依存性を入れる場合．

確率変数列に関しては，マルコフ連鎖，オーダー k の依存系列，オーダー (k,r) の依存系列，高次マルコフ連鎖などの依存性を入れた考察がなされてきた．また，グラフィカルモデルの分布論の発展にともない，Markov tree 上での連の分布の研究も行われるようになった (Aki (2001))．以上のモデルにおいては，時間的あるいは空間的な一様性を入れることにより，条件付けのための事象をうまくとれば，stepwise smoothing の考え方によって得られる条件付き確率母関数の関係式の個数を有限にすることができる．この場合は，必要ならば計算機を活用することによって，確率母関数を導くことが可能になる．

別のタイプの依存性のモデルとして，壷のモデルを考えることもできる．この場合は一般には玉の数が変化していくので，時間的な一様性は仮定できない．そのため，同じように，stepwise smoothing の考え方によって条件付き確率母関数の関係式を作っても，その個数が有限にはならない．そのため確率母関数そのものを厳密に求めることは困難になるが，確率母関数の展開を途中で打ち切った関数を厳密に求めることができる (Inoue and Aki (2002))．

(5) 事象の数の数え方に変化をつける場合．

実際の問題に対応するためには，さまざまな数え方に対処しなければならない．また，この数え方の問題は理論的にも興味ある問題を引き起こすことがある．

(6) 実数値確率変数から離散値の確率変数を構成する場合．

離散パターンに関する研究は，多くの場合，離散値をとる確率変数の族が与えられるところから出発するが，ときには背後に連続値をとる確率変数の族があり，それらをさまざまな方法で分類することによって離散値の確率変数の族が得られている場合もある．そのような場合には，さまざまな分類の仕方についての同時分布を考察することも面白い問題である．

引用文献

Aki, S. (2001). Exact reliability and lifetime of consecutive systems, *Handbook of Statistics*, **Vol. 20**, N. Balakrishnan and C. R. Rao Eds., Elsevier, 281-300.

Aki, S.and Hirano, K. (2003). Waiting time problems for a two-dimensional pattern, to appear in *Ann. Inst. Statist. Math.*

Inoue, K. and Aki, S. (2002). Generalized waiting time problems associated with patterns in Polya's urn scheme, *Ann. Inst. Statist. Math.*, **54**, 681-688.

殆ど一致の待ち時間分布

上海財経大学　　韓　清
統計数理研究所　平野勝臣

2つの独立な系列 $T = T_1 \cdots T_d$ と $R = R_1 \cdots R_n$ を考え，$1 \le i \le d, 1 \le j \le n$ に対し確率変数

$$Z_{i,j} = \begin{cases} 1 & \text{if } T_i = R_j \\ 0 & \text{otherwise} \end{cases}$$

を定義し，$Z_{i,j} = 1$ のとき，系列 T の位置 i と系列 R の位置 j で一致しているという (Fu *et al.* (1999))．このようにして，$Z_{i+t,j+t}, (t = 0, 1, \cdots)$ の $\{0,1\}$-値系列を得る．k を 1 より大きい整数とする．ウインドウサイズ k のスキャン統計量 S は

$$S = \max_{1 \le i \le d-k+1, 1 \le j \le n-k+1} \sum_{t=0}^{k-1} Z_{i+t,j+t}$$

で定義され，2つの系列間で長さ k 内で動かしたとき，一致した文字の最大個数である．2つの系列において，殆ど一致とは，長さ k のパターンで不一致は高々ひとつだけあることとする．

$X_1, X_2, \cdots$ を $\{0,1\}$-値マルコフチェインとし，初期確率を $Pr(X_1 = 1) = p_1, Pr(X_1 = 0) = p_0$，推移確率行列を $\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$，$j$-step 推移確率行列を $\mathbf{P}^j = \begin{pmatrix} p_{00}^{(j)} & p_{01}^{(j)} \\ p_{10}^{(j)} & p_{11}^{(j)} \end{pmatrix}$，$j = 0, 1, 2, \cdots$ とする．ここで $\mathbf{P}^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ とする．

本報告では，このマルコフ系列において殆ど一致がはじめて起こるまでの待ち時間 (試行数) W の確率生成母関数を与え，確率生成母関数から確率関数や確率に対する数値結果を与える．

定理．殆ど一致がはじめて起こるまでの待ち時間 W の確率生成母関数 $\phi(x)$ は

$$\phi(x) = \frac{\{p_0 x p_{01} x + p_1 x(1 - p_{00}x)\}(1 - p_{11}x)g(x)}{(p_{10}xp_{01}x)g(x) + \left(1, (p_{11}x)^{k-3}p_{10}xp_{01}x\right)\left(\prod_{j=2}^{[\frac{k-1}{2}]} A_j\right)\beta(k)h(x)}$$

で与えられる．ここに

$$\begin{aligned} g(x) &= (p_{11}x)^{k-3}p_{10}xp_{01}x\{\left(0, \frac{p_{11}x}{p_{10}xp_{01}x}\right)\left(\prod_{j=2}^{[\frac{k-1}{2}]} A_j\right) + (1,1) \\ &\quad -\frac{1}{2}(1,1)I(k \text{ is an odd }) + (1,1)\sum_{i=2}^{[\frac{k-1}{2}]}\left(\prod_{j=i}^{[\frac{k-1}{2}]} A_j\right)\}\beta(k), \end{aligned}$$

$$A_i = \begin{pmatrix} 1 & (p_{11}x)^{k-i-2}p_{10}xp_{01}x \\ -(p_{11}x)^{i-2}p_{10}xp_{01}x & 1 - (p_{11}x)^{k-4}(p_{10}xp_{01}x)^2 \end{pmatrix},$$

$$\beta(k) = \begin{cases} \begin{pmatrix} 1 \\ 1 \end{pmatrix} & \text{if } k \text{ is an odd,} \\ \begin{pmatrix} 1 \\ 1 - (p_{11}x)^{\frac{k}{2}-2}p_{10}xp_{01}x \end{pmatrix} & \text{if } k \text{ is an even} \end{cases}$$

である.

殆ど一致の待ち時間 W のある値での確率 $Pr(W=j)$ の計算は $\phi(x)$ が有理関数なので，これを展開した x^j $(j=0,1,2,\cdots)$ の係数から求めることができる．また $\phi(x)$ を微分すれば積率を得る.

p_0 と p_1 をそれぞれ p_{10} と p_{11} におけば，オーバーラップしない数え方で殆ど一致間の分布の確率生成母関数を得ることができる．さらに $X_1, X_2, \cdots, X_n$ において，殆ど一致の起こる回数 (non-overlapping) の分布の確率生成母関数も得ることができる.

連やスキャン統計量は連続システムの信頼性と密接な関係がある．(Chan *et al.* (2000), Glaz *et al.* (2001))．事実，$Pr(W \leq x)$ を使って (k-1)-within-consecutive-k-out-of-n system with dependent components の信頼度が計算できる.

また，遺伝子解析の例がある．2つの遺伝子系列の一致について調べるとき，文字パターンが完全に一致しなくても同じ機能を持つことが知られている (TATA box, DnaA box, *etc.*).

実際，$\phi(x)$ を定理から求めると，例えば k=4, 5 のときは次の様になる.

(1) $k=4$ のとき $\beta(4) = \begin{pmatrix} 1 \\ 1-p_{10}xp_{01}x \end{pmatrix}$ で,

$$\phi(x) = \frac{\{p_0xp_{01}x + p_1x(1-p_{00}x)\}p_{11}x\{2p_{10}xp_{01}x + p_{11}x - p_{11}xp_{10}xp_{01}x - (p_{10}xp_{01}x)^2\}}{1-p_{00}x-p_{10}xp_{01}x-p_{00}xp_{11}xp_{10}xp_{01}x+p_{00}xp_{11}x(p_{10}xp_{01}x)^2}.$$

(2) $k=5$ のとき $\beta(5) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ で，$\phi(x) = \{p_0xp_{01}x + p_1x(1-p_{00}x)\}g(x)/D$,

$$\begin{aligned} g(x) &= (p_{11}x)^2\{p_{11}x + 3p_{10}xp_{01}x - p_{11}xp_{10}xp_{01}x - (p_{10}xp_{01}x)^2 \\ &\quad + p_{11}x(p_{10}xp_{01})^2 - (p_{11}x)^2(p_{10}xp_{01}x)^2 - p_{11}x(p_{10}xp_{01}x)^3\}, \\ D &= 1 - p_{00}x - p_{10}xp_{01}x - p_{00}xp_{11}xp_{10}xp_{01}x - p_{00}x(p_{11}x)^2p_{10}xp_{01}x \\ &\quad - p_{11}x(p_{10}xp_{01}x)^2 + p_{00}x(p_{11}x)^2(p_{10}xp_{01}x)^2 + p_{00}x(p_{11}x)^3(p_{10}xp_{01}x)^3. \end{aligned}$$

なお，本報告は Han and Hirano (2003) に基づいている．引用など詳しいことはこの論文を参照されたい.

参考文献

Chang, G. J., Cui, L. and Hwang, F. K. (2000). *Reliabilities of Consecutive-k Systems,* Kluwwer, Dordrecht.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics,* Springer-Verlag, New York.

Fu, J. C., Lou, W. Y. W. and Chen, S. C. (1999). On the probability of pattern matching in nonaligned DNA sequences: a finite Markov chain imbedding approach, *Scan Statistics and Applications*, (eds. J. Glaz and N. Balakrishnan), 287-302, Birkhäuser.

Han, Q. and Hirano, K. (2003). Waiting time problem for an almost perfect match, *Statistics and Probability Letters* **65**, 39-49.

Statistical Problems in Automatic Recognition of Estrangelo

Kyushu University T. Sakata.
Kyushu@Institute of Design(Doctor course) S.C.Tan.

1 Introduction

We discussed about automatic recognition of Estrangelo, an old Syriac language, through clustering analysis. The motivation of our study is that such an automatic recognition engine might help to mutual understanding between peoples with different languages and cultures. Such classical languages are very important through many historically precious classical texts. Our purpose of this experiment was to reconsider the Cloksin and Fernando's result of the recognition of Estrangelo by using different features and a different classification technique. One major difference between the work of Cloksin et. al. and us is that they used word segmentation into strokes and we treated a word as a whole. It should be noted that major difficulty of recognition comes from each character has variations of shape in relatation to its position in a word(right, middle and left) and further each word has many different shapes in relation to the text in which it appeared. For example, "God" has more than 100 different shapes and "brother" has more than 50 different shapes. If we adopt the method of recognize words as a whole without segmentation like us, these variation of shapes of the same word might be a big hazard of the recognition or increases the need of a huge dictionary. Thus our aim of this research is to check the possibility of reducing the shape forms of each word by clustering them and regiter the center of the cluster as a new word in the dictionary.

2 Experiment

Our program of this experiment is described as follows.
(1) The degital images of 64×64 pixels of 50 different shapes of 20 basic words, total 1000 images, were collected.
(2) We used the original images of 4096 dimension of each shape as the feature and did not used any special extracted features.
(3) For 50 shapes of each word we applied the hierachical clustering with the rule of joining pairs having a minimal euclidian distance. The number of

cluster were varied 5 to 50 increasing at every 5 grids.
(4) The center(the mean vector of the cluster member) of each cluster was registerd as a new word. Thus our dictionary has 20× the number of the cluster shapes.
(5) The same 1000 shapes of the 20 words which were used in constructing the clusters were reinputted and compared to the cluster center and the nearest center was asigned to the shape.
Below is the list of recognition rate.

3 Conclusion

From the table we see 10 clusters for each words are enough to recognize the 20 basic words even with such varied forms. It reduces the dictionary size to $\frac{1}{5}$. However we we note that
(1)For invariance consideration, we need to use invariant features, invariant moment or polar transformation.These will be sought in the future.
(2) For classification we can try other rule, including SVM or Fuzzy classification which are tried in the future.

Character recognition rates by 5 clusters for each word

brother	desire	father	follow	good	forget
82.00	82.00	82.00	88.00	84.00	100.00
god	glorify	hope	lord	light	love
86.00	92.00	84.00	74.00	94.00	74.00
obey	offer	preach	pure	rejoice	say
88.00	80.00	86.00	74.00	94.00	96.00
son	swear				
98.00	80.00		average=85.7		

Character recognition rates by 10 clusters for each word

brother	desire	father	follow	good	forget
98.00	92.00	90.00	90.00	94.00	98.00
god	glorify	hope	lord	light	love
100.00	94.00	100.00	96.00	96.00	88.00
obey	offer	preach	pure	rejoice	say
100.00	100.00	94.00	100.00	100.00	100.00
son	swear				
100.00	86.00		average=95.4		

Evaluating High Dimensional Probability Expressions Using Recursive Integration

A. J. Hayter
Georgia Institute of Technology
ajh@isye.gatech.edu

Yokohama Conference on Statistics, December 2003

Recursive integration can be used to evaluate high dimensional integral expressions as a series of lower dimensional integral calculations. In this article the evaluation of general multivariate normal integrals is first considered. Secondly, the specific case of a quadrivariate normal integral is considered.

Let $\phi_k(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the probability density function of the random variables $\boldsymbol{X} = (X_1, \ldots, X_k)$ which have a k-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. This article addresses the evaluation of probabilities defined by a set of inequalities of linear combinations of $\boldsymbol{X}$. That is, for $c_{ij} \in \Re$, $1 \leq i \leq n$, $1 \leq j \leq k+1$, attention is directed towards the evaluation of

$$P(c_{i1}X_1 + \ldots + c_{ik}X_k \leq c_{i,k+1}; 1 \leq i \leq n) = \int_S \ldots \int \phi_k(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{x} \qquad (1)$$

where the set S is defined as

$$S = \{\boldsymbol{x} : c_{i1}x_1 + \ldots + c_{ik}x_k \leq c_{i,k+1}; 1 \leq i \leq n\}.$$

It can be seen that by employing a linear transformation of $\boldsymbol{X}$ these probabilities can be expressed with different c_{ij}, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and specifically $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_k$, the k-dimensional identity matrix, may be assumed. Notice also that for the purpose of this representation, a two-sided bound on a linear combination of the random variables $\boldsymbol{X}$ can be written as two one-sided inequalities.

While the evaluation of equation (1) ostensibly requires the evaluation of a k-dimensional integral expression, it is shown how it can actually be performed by a series of recursive one-dimensional integral calculations. The number of these one-dimensional integral calculations depends upon the dimension k of the multivariate normal distribution and the number of inequalities n that determines the integration region. Whenever the number of one-dimensional integral calculations required is not exorbitant, this approach affords a practical algorithm for the numerical integration of (1). Furthermore, it enlarges the set of probabilities of the kind (1) whose evaluation is computationally feasible, in comparison to more direct numerical integration approaches to the k-dimensional integral expression.

A special case of the decomposition is addressed in Miwa et al. (2003) where it is shown that a k-dimensional orthant probability for a general multivariate normal distribution can be expressed as no more than $(k-1)!$ orthant probabilities for multivariate

normal distributions with a tri-diagonal covariance matrix. In that paper a geometrical approach was employed to decompose the orthant region. The paper also provides some examples of the computation times required to calculate the orthant probabilities using that decomposition method.

A rough estimate of the overall computational intensity of this method can then be made by using upper bounds on the number of terms in the decomposition. For example, consider the integration of a six-dimensional multivariate normal distribution over a region defined by two-sided inequalities for ten different linear combinations of the random variables. The computational intensity of this probability is no more than the equivalent of $10 + 180 + 2,880 + 40,320 = 43,390$ one-dimensional numerical integrations. This is clearly computationally feasible, and it is an immense improvement over calculating this probability by the direct numerical integration of a six-dimensional integral. If each dimension of a six-dimensional integral is specified with N grid points, then in general the direct numerical integration of the six-dimensional integral requires N^5 one-dimensional numerical integrations.

Let the random variables $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ have a quadrivariate normal distribution with mean vector $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. This article also addresses the evaluation of the two-sided orthant probability

$$P(l_i \leq X_i \leq u_i; 1 \leq i \leq 4) \tag{2}$$

where some of the l_i may be equal to $-\infty$ and some of the u_i may be equal to ∞. With a suitable linear transformation of the random variables $\boldsymbol{X}$ any probability of this kind can be expressed as a two-sided orthant probability for random variables with a mean $\boldsymbol{\mu} = \mathbf{0}$ and with $\boldsymbol{\Sigma}$ equal to a correlation matrix with all diagonal elements equal to one. It is shown how (2) can be evaluated numerically with a one-dimensional integral calculation.

References

Miwa, T., Hayter, A. J., and Kuriki, S. (2003), "The evaluation of general non-centred orthant probabilities," *Journal of the Royal Statistical Society, Series B*, 65, 223-234.

Informations contained in record data

筑波大・理工 山本泰志
筑波大・数学 赤平昌文

1 はじめに

最近, 独立に同一の連続型分布に従う確率変数から得られる記録値と記録時刻に含まれる情報量を調べることが盛んに行われている. また, 記録値の予測問題についても論じられている ([ArBN98], [HiAk02]). 本論では, 連続型分布から得られた無作為標本に基づく記録値に含まれる Fisher 情報量について, [HoN03] に基づいて述べた上で, 指数分布を含む一般のガンマ分布の場合に適用する. また, 記録値に含まれる Kullback-Leibler 情報量についても論じる.

2 定義

確率変数 X の (ルベーグ測度に関する) 確率密度関数 (p.d.f.) を $f(x,\theta)$ とする. ただし $\theta \in \Theta$ とし, Θ を $\boldsymbol{R}^1$ の開区間とする. また, $F(x,\theta)$ を $f(x,\theta)$ の累積分布関数 (c.d.f.) とする. このとき, $\theta_1, \theta_2 \in \Theta$ について, $f(\cdot,\theta_2)$ に対する $f(\cdot,\theta_1)$ の識別をするために, Kullback-Leibler (K-L) 情報量を

$$I(\theta_1 : \theta_2) = \int_{-\infty}^{\infty} f(x,\theta_1) \log \frac{f(x,\theta_1)}{f(x,\theta_2)} dx$$

によって定義する. 次に, 上位の記録時刻 (upper record time) T_m, 上位の記録値 (upper record value) R_m を

$$T_1 := 1, \quad T_m := \min\{j | j > T_{m-1}, X_j > X_{T_{m-1}}\} \quad (m = 2, \cdots, n),$$
$$R_m := X_{T_m} \quad (m = 1, \cdots, n)$$

と定義する. また, 下位の記録時刻, 記録値も同様にして定義する. なお, 本論では次の表記を用いる.

(i) $I^U_{RT}(n;\theta_1,\theta_2)$: 大きさ n の無作為標本から得られた記録数を N_n とし, 上位の記録値 $\boldsymbol{R} := (R_1, \cdots, R_{N_n})$ と上位の記録時刻 $\boldsymbol{T} := (T_1, \cdots, T_{N_n})$ の組 $(\boldsymbol{R}, \boldsymbol{T})$ のもつ K-L 情報量

(ii) $I^U_R(n;\theta_1,\theta_2)$: 大きさ n の無作為標本から得られた上位の記録値 $\boldsymbol{R}$ のもつ K-L 情報量

(iii) $I^U_M(n;\theta_1,\theta_2)$: 大きさ n の無作為標本から得られた最後 (最大) の記録値 R_m のもつ K-L 情報量

また, 下位の記録値 (lower record value) に関しても, 同様の表記 $I^L_{RT}(n;\theta_1,\theta_2)$, $I^L_R(n;\theta_1,\theta_2)$, $I^L_M(n;\theta_1,\theta_2)$ を用いる.

3 記録データの情報量

まず, $X_1, \cdots, X_n$ を互いに独立にいずれも p.d.f. $f(x,\theta)$, c.d.f. $F(x,\theta)$ をもつ分布に従う確率変数とする. このとき, 最大の記録値 R_m は順序統計量を用いれば, $R_m = \max_{1\le i\le n} X_i =: X_{(n)}$ となるから, その p.d.f. は,

$$f_{R_m}(x,\theta) = nF(x,\theta)^{n-1} f(x,\theta)$$

になる. 第 2 節より, 最大の記録値 R_m のもつ K-L 情報量は,

$$I^U_M(n;\theta_1,\theta_2) = \int_{-\infty}^{\infty} \left\{ \log \frac{f(x,\theta_1)}{f(x,\theta_2)} + (n-1) \log \frac{F(x,\theta_1)}{F(x,\theta_2)} \right\} nF^{n-1}(x,\theta_1) f(x,\theta_1) dx$$

になり, 下位の記録値については F を $1-F$ に変えればよい. また, $(\boldsymbol{R}, \boldsymbol{T})$ のもつ K-L 情報量は

$$I^U_{RT}(n;\theta_1,\theta_2) = \sum_{i=0}^{n-1} \int_{-\infty}^{\infty} \left\{ \log \frac{f(x,\theta_1)}{f(x,\theta_2)} + i \log \frac{F(x,\theta_1)}{F(x,\theta_2)} \right\} F^i(x,\theta_1) f(x,\theta_1) dx$$

になり, $I_M^U(\cdot;\theta_1,\theta_2)$ を用いれば

$$I_{RT}^U(n;\theta_1,\theta_2)=\sum_{i=1}^{n}\frac{1}{i}I_M^U(i;\theta_1,\theta_2)$$

になる. さらに $\boldsymbol{R}$ のもつ K-L 情報量は $I_R^U(n;\theta_1,\theta_2)=E_1^U+E_{\theta_1}[\log(A^U(\boldsymbol{R})/B^U(\boldsymbol{R}))]$ になる. ただし,

$$E_1^U=\sum_{i=0}^{n-1}\int_{-\infty}^{\infty}\left(\log\frac{f(x,\theta_1)}{f(x,\theta_2)}\right)F^i(x,\theta_1)f(x,\theta_1)dx,$$

$$A^U(\boldsymbol{r})=\text{coefficient of } s^{n-m} \text{ in } \prod_{i=1}^{m}\frac{1}{1-F(r_i,\theta_1)s},\quad B^U(\boldsymbol{r})=\text{coefficient of } s^{n-m} \text{ in } \prod_{i=1}^{m}\frac{1}{1-F(r_i,\theta_2)s}$$

とする. 下位の記録値については, $E_1^L, A^L(\boldsymbol{r},m)$ として F を $1-F$ とおきかえたものを用いればよい. 後はシミュレーションによる数値計算を行う.

4 例

実際の適用例として, 指数分布に関する記録データのもつ K-L 情報量について述べる. いま, $X_1,\cdots,X_n$ をたがいに独立にいずれも p.d.f. $f(x,\theta)=\theta e^{-\theta x}$ $(x>0)$; $=0$ $(x\le 0)$ をもつ指数分布 $\mathrm{Exp}(1/\theta)$ に従う確率変数とする. ただし, $\theta>0$ とする. このとき, $\lambda:=\theta_2/\theta_1$ とすれば, 最大の記録値および最小の記録値のもつ K-L 情報量は, それぞれ

$$I_M^U(n;\lambda)=-\log\lambda-(1-\lambda)\sum_{j=1}^{n}\frac{1}{j}-\frac{n-1}{n}+\lambda(n-1)\sum_{k=1}^{\infty}B(k\lambda,n+1),\quad I_M^L(n;\lambda)=-\log\lambda-(1-\lambda)$$

となる. ただし, $B(\cdot,\cdot)$ はベータ関数とする. 次に,

$$I_{RT}^U(n;\lambda)=\sum_{i=1}^{n}\frac{1}{i}I_M^U(i;\lambda),\quad I_{RT}^L(n)=\sum_{i=1}^{n}\frac{1}{i}I_M^L(i;\lambda)$$

として記録値と記録時刻の組 $(\boldsymbol{R},\boldsymbol{T})$ のもつ K-L 情報量が得られる. 最後に, 記録値 $\boldsymbol{R}$ のもつ K-L 情報量は

$$E_1^U=-(\log\lambda)\sum_{i=1}^{n}\frac{1}{i}-(1-\lambda)\sum_{i=1}^{n}\frac{1}{i}\sum_{j=1}^{i}\frac{1}{j},\quad E_1^L=-(\log\lambda)\sum_{i=1}^{n}\frac{1}{i}-(1-\lambda)\sum_{i=1}^{n}\frac{1}{i^2}$$

として, シミュレーションによりその情報量の値が得られる (下表参照).

	λ												
	0.01	0.02	0.05	0.1	0.2	0.5	1	2	5	10	20	50	100
n 1	3.62	2.93	2.05	1.40	.809	.193	0	.307	2.39	6.70	16.0	45.1	94.4
5	17.3	13.9	9.47	6.34	3.53	.777	0	1.00	6.74	17.4	39.6	107	221
10	33.5	26.7	18.0	11.9	6.43	1.33	0	1.48	9.21	23.2	51.8	139	284
50	154	121	78.1	48.8	24.3	4.11	0	2.86	15.4	37.2	81.5	216	440
100	297	230	146	88.9	42.3	6.39	0	3.52	18.2	43.4	94.6	249	508

表 $\mathrm{Exp}(1/\theta)$ から得られた最大の記録値に対する $I_M^U(n;\lambda)$ の値 (有効数字 3 桁)

参考文献

[ArBN98] Arnord, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records.* Wiley, New York.

[HiAk02] Hida, E. and Akahira, M. (2002). On the construction of prediction intervals for record values. (In Japanese), *Proc. Sympos., Res. Inst. Math. Sci., Kyoto Univ.*, **1273**, 165–177.

[HoN03] Hofmann, G. and Nagaraja, H. N. (2003). Fisher information in record data. *Metrika* **57**, 177–193.

U-統計量の凸結合のスチューデント化に基づくエッジワース展開

鹿児島大・理　大和 元
戸田 光一郎
都城高専　野町 俊文
九州大・経済　前園 宜彦

1. 序

$X_1,...,X_n$ を分布 F からの大きさ n の標本とする. 次数 k (≥ 2) の対称な kernel $g(x_1,...,x_k)$ をもつ推定可能な母数 $\theta=\theta(F)$ の推定量として U-統計量の凸結合 Y_n (Toda and Yamato (2001)) を考える. Y_n は以下のようにして与えられる: $j=1,...,k$ に対して, $w(r_1,...,r_j;k)$ を $r_1+\cdots+r_j=k$ を満たす正整数 $r_1,...,r_j$ の対称な非負関数とする. $w(r_1,...,r_j;k)$ の少なくとも 1 つは正であるとする. $j=1,...,k$ に対して, $d(k,j)=\sum^{+}_{r_1+\cdots+r_j=k} w(r_1,...,r_j;k)$ とおく. 但し, $\sum^{+}_{r_1+\cdots+r_j=k}$ は $r_1+\cdots+r_j=k$ を満たすすべての正整数 $r_1,...,r_j$ 上でとられる和を表す. $j=1,...,k$ に対して, $g_{(j)}(x_1,...,x_j)$ を

$$g_{(j)}(x_1,...,x_j)=\frac{1}{d(k,j)}\sum^{+}_{r_1+\cdots+r_j=k} w(r_1,...,r_j;k)g(\underbrace{x_1,...,x_1}_{r_1},...,\underbrace{x_j,...,x_j}_{r_j})$$

により与えられる kernel とし, $g_{(j)}(x_1,...,x_j)$ に対応する U-統計量を $U_n^{(j)}$ とする.
$w(r_1,...,r_j;k)$ の対称性より, $g_{(j)}(x_1,...,x_j)$ は対称である. また, ある j に対して $d(k,j)=0$ であるとき, 対応する $U_n^{(j)}$ は 0 とする. このとき, 統計量 Y_n は

$$Y_n=\frac{1}{D(n,k)}\sum_{j=1}^{k} d(k,j)\binom{n}{j}U_n^{(j)} \tag{1}$$

により与えられる. 但し, $D(n,k)=\sum_{j=1}^{k} d(k,j)\binom{n}{j}$ である. $j=1,...,k$ に対して, $w(r_1,...,r_j;k)$ は少なくとも 1 つが正となる非負関数であるので, $D(n,k)$ は正である. (1) の右辺の $U_n^{(\cdot)}$ の係数は非負であり, その和は 1 となるので, (1) により与えられた線形結合は凸結合である.

2. スチューデント化 Y-統計量のエッジワース展開

(1) により与えられた Y-統計量 Y_n のスチューデント化に対して, ジャックナイフ分散推定量を用いる. 即ち, $\sqrt{n}Y_n$ の分散推定量として

$$\hat{\sigma}_n^2=(n-1)\sum_{i=1}^{n}(Y_n^{(i)}-Y_n)^2$$

により与えられる $\hat{\sigma}_n^2$ を用いる. 但し, $Y_n^{(i)}$ は, X_i を除いた大きさ $n-1$ の標本に基づく (1) により与えられる Y-統計量である. kernel $g(x_1,...,x_k)$ が退化しない場合のジャックナイフ分散推定量 $\hat{\sigma}_n^2$ を用いたスチューデント化 Y-統計量 Y_n についての n^{-1} の項までのエッジワース展開を考える.
$d(k,k)>0$ とする. このとき,

$$\frac{d(k,k)}{D(n,k)}\binom{n}{k}=1-\frac{\delta_k}{n}+O\left(\frac{1}{n^2}\right),\quad \frac{d(k,k-1)}{D(n,k)}\binom{n}{k-1}=\frac{\delta_k}{n}+O\left(\frac{1}{n^2}\right)$$

を満たす定数 δ_k (≥ 0) が存在する. kernel $g(x_1,...,x_k)$ に対して, 次のようにおく.

$$\psi_c(x_1,...,x_c)=E[g(X_1,...,X_k)\mid X_1=x_1,...,X_c=x_c]\qquad (c=1,2,3)$$

$$g^{(1)}(x_1)=\psi_1(x_1)-\theta,$$

$$g^{(c)}(x_1,...,x_c)=\psi_c(x_1,...,x_c)-\sum_{i=1}^{c-1}\sum_{1\leq l_1<\cdots<l_i\leq c} g^{(i)}(x_{l_1},...,x_{l_i})-\theta\qquad (c=2,3).$$

また, kernel $g_{(k-1)}(x_1,...,x_{k-1})$ に対して, 次のようにおく.

$$\theta_{k-1} = Eg_{(k-1)}(X_1,...,X_{k-1}),$$
$$\psi_{(k-1),1}(x_1) = E[g_{(k-1)}(X_1,...,X_{k-1}) \mid X_1 = x_1],$$
$$g^{(1)}_{(k-1)}(x_1) = \psi_{(k-1),1}(x_1) - \theta_{k-1}.$$

ここで,

$$\sigma_1^2 = E[\{g^{(1)}(X)\}^2], \qquad \sigma_2^2 = (k-1)^2 E[\{g^{(2)}(X_1,X_2)\}^2],$$
$$\nu = \sigma_2^2 + \frac{2(k-1)\delta_k}{k} E[g^{(1)}(X) g^{(1)}_{(k-1)}(X)] - 2\delta_k \sigma_1^2$$

とおき, f_1, f_2 を, $g^{(1)}$, $g^{(2)}$, $g^{(3)}$ を用いて表される関数とする. さらに,

$$\tau = \frac{3E[f_1^2(X_1)]}{2\sigma_1^4} - \frac{\nu}{2\sigma_1^2}, \qquad \zeta = E[f_1(X_1) g^{(1)}(X_1)],$$

とおき, a_1, a_2, a_3 を, $g^{(1)}$, $g^{(2)}$, $g^{(3)}$, $g^{(1)}_{(k-1)}$ および f_1, f_2 を用いて表される関数とする. これらを用いることにより,

$$\hat{\sigma}_n^{-1}\sqrt{n}(Y_n - \theta) = \frac{\sqrt{n}}{\sigma_1} U_n^{**} + \frac{1}{\sqrt{n}} \Big\{ -\frac{\zeta}{\sigma_1^3} + \frac{\mu_k}{k\sigma_1} \Big\} + R_n^* + o_p^*(n^{-1})$$

と表すことができる. 但し,

$$U_n^{**} = \frac{1}{n}\sum_{i=1}^{n} \Big\{ g^{(1)}(X_i) + \frac{a_1^*(X_i)}{n} \Big\} + \frac{1}{n^2} \sum_{1\le i<j\le n} a_2(X_i,X_j) + \frac{1}{n^3} \sum_{1\le i<j<l\le n} a_3(X_i,X_j,X_l),$$
$$a_1^*(X_i) = a_1(X_i) - \frac{\mu_k}{k\sigma_1^3} f_1(X_i)$$

であり, $\mu_k = \delta_k(\theta_{k-1} - \theta)$, $E|R_n^*| = O(n^{-3/2})$. また, $o_p^*(n^{-1})$ は, 任意の定数 $c > 0$ に対して $P\left(|o_p^*(n^{-1})| \ge cn^{-1}(\log n)^{-1}\right) = o(n^{-1})$ を満たす量である.

命題 1 $d(k,k) > 0$ とし, $1 \le i_1 \le \cdots \le i_k \le k$ に対して $E|g(X_{i_1},...,X_{i_k})|^2 < \infty$ とする. $E|g(X_1,X_2,X_3,...,X_k)|^9 < \infty$ とし, $\varepsilon > 0$ に対して $E|g(X_1,X_1,X_2,...,X_k)|^{4+\varepsilon} < \infty$ とする. このとき,

$$\sup_{-\infty<x<\infty} \left| P\Big(\hat{\sigma}_n^{-1}\sqrt{n}(Y_n - \theta) \le x\Big) - P\Big(\frac{\sqrt{n}}{\sigma_1} U_n^{**} + \frac{1}{\sqrt{n}} \Big\{ -\frac{\zeta}{\sigma_1^3} + \frac{\mu_k}{k\sigma_1} \Big\} \le x\Big) \right| = o(n^{-1})$$

が成り立つ.

U_n^{**} に対して Maesono (1996) の結果を用い, テイラー展開を用いて展開への近似を行うことによりスチューデント化 Y-統計量に対するエッジワース展開が得られる.

命題 2

$$\sup_{-\infty<x<\infty} \left| P\Big(\hat{\sigma}_n^{-1}\sqrt{n}(Y_n - \theta) \le x\Big) - H_n^*(x) \right| = o(n^{-1}).$$

但し,

$$H_n^*(x) = \Phi(x) + \phi(x)\frac{1}{6\sqrt{n}}\Big(v_1x^2 + v_2 - \frac{6\mu_k}{k\sigma_1}\Big) + \phi(x)\frac{1}{72n}\Big\{ v_3x^5 + \Big(v_4 + \frac{v_1\mu_k}{k\sigma_1}\Big)x^3$$
$$+ \Big[v_5 + 12\frac{\mu_k}{k\sigma_1}(v_2 - 2v_1) - 72\frac{\mu_k}{k\sigma_1^5}\Big(\frac{1}{2}e_1 + e_2\Big)\Big]x - 36\Big(\frac{\mu_k}{k\sigma_1}\Big)^2 \Big\}.$$

(定数や関数の定義, および命題 2 の条件等詳細については Yamato et al. (2003) を参照されたい.)

参考文献

[1] Maesono (1996), *J. Japan Statist. Soc.*, **26**, No. 2, 189–207.
[2] Toda, K. and Yamato, H. (2001), *J. Japan Statist. Soc.*, **31**, No. 2, 225–237.
[3] Yamato, H., Toda, K., Nomachi, T. and Maesono, Y. (2003), (to appear).

U-統計量の凸結合のスチューデント化に基づく エッジワース展開（応用例）

鹿児島大学・理　大和　元
戸田光一郎
都城高専　野町俊文
九州大学・経済　前園宜彦

1　序

$\theta(F)$ を分布 F の推定可能な母数とし、$g(x_1,\ldots,x_k)$ を次数 $k\geq 2$ の対称なカーネルとし、$X_1,\cdots,X_n$ を分布 F からの任意標本 とする。$\theta(F)$ の推定量として、U-統計量の凸結合 Y_n は Toda and Yamato [1] により、提案されている：

$$Y_n=\frac{1}{D(n,k)}\sum_{j=1}^{k}d(k,j)\left(\begin{array}{c}n\\ j\end{array}\right)U_n^{(j)} \tag{1}$$

ただし、$D(n,k)=\sum_{j=1}^{k}d(k,j)\left(\begin{array}{c}n\\ j\end{array}\right)$ であり、$g_{(j)}$ は、カーネル g の凸結合によって与えられる関数である。$U_n^{(j)}$ は、カーネル $g_{(j)}$ に対応する U-統計量を表す。$d(k,j)=0$ ならば、対応する $U_n^{(j)}=0$ とする。

2　Y-統計量のステューデント化について

推定可能な母数 θ のまわりでジャックナイフ分散推定量を用いたステューデント化 $Y-$ 統計量の Edgeworth 展開は次のようになる。

[命題 1]

$$\sup_{-\infty<x<\infty}\left|P(\hat{\sigma}_n^{-1}\sqrt{n}(Y_n-\theta)\leq x)-H_n^{*}(x)\right|=o(n^{-1}) \tag{2}$$

が成り立つ。ただし、

$$\begin{aligned}H_n^{*}(x) &= \Phi(x)+\phi(x)\frac{1}{6\sqrt{n}}\left(v_1x^2+v_2-\frac{6\mu_k}{k\sigma_1}\right)+\phi(x)\frac{1}{72n}\left\{v_3x^5+(v_4+\frac{v_1\mu_k}{k\sigma_1})x^3\right.\\ &\left.+\left[v_5+12\frac{\mu_k}{k\sigma_1}(v_2-2v_1)-72\frac{\mu_k}{k\sigma_1^5}(\frac{1}{2}e_1+e_2)\right]x-36(\frac{\mu_k}{k\sigma_1})^2\right\}\end{aligned} \tag{3}$$

条件と記号、および次の述べる応用例の詳細については、 Yamato et al. [2] を参照されたい。

3　応用例

[例 1]　3 次の中心積率 ($\theta=\int(x-\mu)^3dF(x)$) を考える。$\mu$ は F の平均である。カーネルは、

$$g(x_1,x_2,x_3)=\frac{1}{3}(x_1^3+x_2^3+x_3^3)-\frac{1}{2}(x_1^2x_2+x_1^2x_3+x_1x_2^2+x_2^2x_3+x_1x_3^2+x_2x_3^2)+2x_1x_2x_3$$

となり、対応する $Y-$ 統計量は、$Y_n=d(3,3)n^2\sum_{j=1}^{n}(X_j-\bar{X})^3/(6D(n,3))$ である。ただし、$\bar{X}=\sum_{j=1}^{n}X_j/n$ である。分布 F は密度関数を持ち、F に従う確率変数 X について、$E|X|^{27}<\infty$ であると仮定する。(i) 分布 F が原点 O に関して対称である場合は、θ のまわりのエッジワース展開と期待値のまわりのエッジワース展開は $o(n^{-1})$ のオーダーで一致する。(ii) 分布 F が原点 O に関して対

称でない場合は、θ のまわりのエッジワース展開と期待値のまわりのエッジワース展開との差は、μ_k において異なる。ただし、V-統計量, S-統計量, LB-統計量に対する μ_k は次の通りである。

$$\mu_k = \begin{cases} -\frac{k(k-1)}{2} m_3' & (V, S - \text{統計量}) \\ -k(k-1)m_3' & (LB - \text{統計量}) \end{cases}$$

[例 2] $k \geq 3$ にたいして、カーネル

$$g(x_1, x_2, \ldots, x_k) = x_1 x_2 \cdots x_k$$

を考える。このカーネルは、推定可能な母数 $\theta(F) = \mu^k$ を与える。分布 F は、密度関数を持ち、$\mu(>0)$ に関して対称であると仮定する。また F に従う確率変数 X について、$E|X|^9 < \infty (k = 3, 4)$ または $E|X|^{2k+\epsilon} < \infty (k \geq 5)$ であると仮定する。この場合、

$$e_1 = 0,\ e_2 = (k-1)\mu^{3k-4} m_2^2,\ \sigma_1^2 = \mu^{2k-2} m_2,\ v_1 = v_2 = \frac{3(k-1)\sqrt{m_2}}{\mu},\ v_3 = -\frac{9(k-1)^2 m_2}{\mu^2},$$

$$v_4 = \frac{6\{\mu^2 m_4 - 6\mu^2 m_2^2 + (k-1)(k-5)m_2^3\}}{\mu^2 m_2^2},\ v_5 = -\frac{9\{2\mu^2 m_4 - 4\mu^2 m_2^2 + (k-1)(2k-3)m_2^3\}}{\mu^2 m_2^2},$$

$$\mu_k = \begin{cases} \frac{k(k-1)}{2}(m_2^2 + \mu^2 - \mu^3)\mu^{k-3} & (V, S - \text{統計量}) \\ -k(k-1)(m_2^2 + \mu^2 - \mu^3)\mu^{k-3} & (LB - \text{統計量}) \end{cases}$$

が得られる。μ_k は θ のまわりと期待値のまわりのエッジワース展開の差を与える。

[例 3] カーネル

$$g(x_1, x_2, x_3) = \frac{1}{3}\{I(x_1 > x_2 + x_3) + I(x_2 > x_1 + x_3) + I(x_3 > x_1 + x_2)\}$$

を考える。ただし、$I(A)$ は事象 A の定義関数である。このカーネルは NBU(New Better than Used) の性質を持つ度合いを計る尺度である推定可能な母数 $\theta(F) = E[1 - F(X_1 + X_2)]$ を与える。分布 F が (i) 一様分布 $U(0, 1)$ である場合と、(ii) パラメータが 1 である指数分布 $e(1)$ である場合について、θ のまわりのエッジワース展開を与えた。V-統計量, S-統計量, LB-統計量に対する μ_k の値は次の通りである。

(i) $U(0, 1)$ の場合 : $\mu_k = \frac{1}{4}$ $(V, S - \text{統計量})$ および $\mu_k = \frac{1}{2}$ $(LB - \text{統計量})$

(ii) $e(1)$ の場合 : $\mu_k = -\frac{5}{12}$ $(V, S - \text{統計量})$ および $\mu_k = -\frac{5}{6}$ $(LB - \text{統計量})$

次に次数が 2 のカーネルを考える。

[例 4] probability weighted moment : $\theta = \beta_1 = \frac{1}{2}E[\max(X_1, X_2)] = E[XF(X)]$ に対応するカーネル

$$g(x_1, x_2) = \frac{1}{2}\max(x_1, x_2) = \frac{1}{2}[x_1 I(x_1 \geq x_2) + x_2 I(x_1 < x_2)]$$

を考える。分布 F が一様分布 $U(0, 1)$ であるとすると、$\theta = 1/3$, $g_{(1)}(x_1) = x_1/2$, $\theta_1 = 1/4$ を得る。V-統計量, S-統計量, LB-統計量に対する μ_k の値は次の通りである。

$$\mu_k = -\frac{1}{12}\ (V, S - \text{統計量})\ \text{および}\ \mu_k = -\frac{1}{6}\ (LB - \text{統計量})$$

参考文献

[1] Toda, K. and Yamato, H.: *J. Japan Statist. Soc.* **31**, No.2, 225–237 (2001)

[2] Yamato, H., Toda, K., Nomachi, T. and Maesono, Y. (投稿中)

ロジットモデルによる倒産確率の推定

慶応義塾大学・理工　川戸健司

1.　序

倒産確率の推定は、企業の安全性を見るうえで非常に重要な指標であるため、正確なモデルを作ることは社会的にとても有用であると考えられ、今まで多くの研究者がこれを研究してきた。今回扱う「財務諸表データを用いた倒産確率の推定」という方法は、その中でも最も古くから研究されている分野である。

この推定方法の概略は，次のようである。

（ｉ）財務諸表の数値から、倒産・非倒産を判別できる指標を見つける。
（ii）その指標を変量としたモデルを作る。
（iii）得られたモデルに調査したい企業のデータを入力することにより、その企業の倒産確率を推定する。

ところが、これまでに世の中に出ている財務諸表データを用いた倒産確率の推定に関する結果を調べてみると、データ解析の基礎知識をきちんともっているならば、まずあり得ないことをして結論を出しているものが相当数に上ることがわかった。一番の大きな問題は、単純な線形モデルを使って倒産確率のモデル式を作っていることである（Excel の統計パッケージの影響か)。線形モデルでは、当然のことながら被説明変量（予測値）の値域が［0、1］区間に収まる保証はない。線形モデルを使って結論を出している場合、モデル式にもとづく確率の算出結果が 0 未満または 1 を超えた場合、その値を強引に 0 または 1 とおいている。これでは、意味のある結果にはならない。

世の現実を見ると、一般に，データ解析のイロハを無視した解析がまかり通っていることに気がつく。このシンポジウムのテーマは「データ解析のための統計科学理論」であるので、特に新味のない統計理論であっても、まっとうな統計理論をデータにもとづいてきちんと使ったならばどのような結果が得られるかを、ロジットモデルによる倒産確率の推定を例にして報告する。

2.　実データを使った倒産確率の推定

本報告では，日経 NEEDS にある 1 部上場企業を対象に解析を行った結果を紹介する．

まず、説明変量選択倒産企業すべてについて倒産要因を調べると、ほぼ次の 7 種類に整理できることがわかった。() 内は、その要因が強く反映される財務諸表の項目である。

a.　売上の減少（売上、経常利益）

b. 関係会社の倒産による損失（貸倒れ損失）
c. 投資などの失敗による損失（資産処分損，長・短期借入金）
d. 事件や天災による損失（特別損失）
e. 制度の変化による損失（なし）
f. 債務超過（負債、総資産）
g. 資金調達の困難（流動資産、流動負債）

a～gに出てきた財務指標の各項目を、次の指標に置き換えてモデルに取り込む。

指標	モデルに取り込む際の変換式
売上の増減	（直前期売上－２期前売上）／資本金
経常利益の増減	（直前期経常利益－２期前経常利益）／資本金
売上	直前期売上／資本金
経常利益	直前期経常利益／資本金
負債比率	負債合計／総資産
特別損失	直前期特別損失／資本金
資産処分損	直前期資産処分損／資本金
流動比率	流動資産／流動負債

資本金や総資産で割っているのは、企業規模の大小によらない指標とするためである。

実データによって箱形図を描いてこれらの指標の分布を観察した結果、経常利益、負債比率、特別損失、流動比率の４つが、説明変量として適切であると判断できた。まず、これら４指標すべてを用いたロジットモデルを組み、その後説明変量の有意性を見てなるべく簡潔なモデルにした結果、倒産確率を推定するモデル式として、最終的に次のものが得られた。

$$\mathrm{logit}(\mu) = -13.766 - 3.3166 \cdot \text{経常利益} + 16.928 \cdot \text{負債比率}$$

3. 結果の検証

試みに、使用したデータをこのモデル式に当てはめ、倒産確率を算出した。確率 0.5 を境界として、倒産企業の中で倒産確率 ＞ 0.5 となったもの、非倒産企業の中で倒産確率 ＜ 0.5 となったものの企業数を調べたところ、次のようになった。

倒産企業　28/32（87.5%）　非倒産企業　60/64（93.75%）

ここで、倒産企業の中で倒産確率が 0.5 未満（倒産の可能性は 5 割以未満）となってしまった企業の倒産要因について調べてみると、財務諸表上に表れない原因で倒産していたか、もしくは財務諸表そのものに虚偽があった企業であった。

リスク細分型保険の純保険料推定のためのロバスト回帰分析

三重大学　工　竹内一郎
東京工業大学　数理・計算科学　金森敬文

1 はじめに

損害保険各社は, 1998年の規制緩和以来, 顧客のリスクファクターを細分化して保険料を設定する「リスク細分型保険」の販売を始めている. リスク細分型保険における純保険料推定は, リスクファクターを説明変数 $\boldsymbol{X}$, 支払われる保険金額を被説明変数 Y とし, 条件付平均 $E[Y|\boldsymbol{X}]$ を推定する回帰分析として次のように定式化される:

$$Y = \mu(\boldsymbol{X}) + \epsilon(\boldsymbol{X}), \quad E[\boldsymbol{X}] = 0. \tag{1}$$

この回帰分析の特徴は, 誤差項 $\epsilon(\boldsymbol{X})$ が **非対称** で **不均一**($\boldsymbol{X}$ に依存) となることである. 非対称/不均一な誤差項のもとで条件付平均を推定したい場合には, 従来のロバスト回帰, 変数変換などを利用するとバイアスが生じる. 本報告では, 分位点回帰 [1] を利用し, 誤差項が上記のような性質を持つ場合に適用可能なロバスト回帰推定量を提案する [2, 3]. 不均一な誤差項のモデルとして, (1) 均一分散モデル, (2) 位置尺度モデル, (3) 一般化位置尺度モデルを考察し, それぞれのモデルにおけるロバスト推定量を構築する. 提案する推定量を自動車保険データ解析へ適用し, その有効性を示す.

2 従来のロバスト回帰, 変数変換法に関して

非対称/不均一な誤差項のもとでの回帰分析は多くの応用問題に現れる. これらの目的が条件付平均 $E[Y|\boldsymbol{X}]$ の推定でなければ, Y を対数変換して $E[\log Y|\boldsymbol{X}]$ を求めたり, ロバスト回帰によって中央値などを求めたりすることができ, はずれ値の影響を受けない回帰分析が可能となる. 保険料推定問題の特徴は, 誤差項が非対称/不均一で, かつ, 条件付平均 $E[Y|\boldsymbol{X}]$ の推定が必要となることである.

純保険料推定では, 無事故や軽事故に対する保険金に比べて大事故に対する保険金が非常に大きな値 (はずれ値) をとるため, 最小二乗推定量などのロバストでない推定量を用いるとばらつきが大きくなる. ロバスト推定量を用いることによりはずれ値の影響を軽減することができる. しかし, 誤差項の分布が非対称なときにはこれらは条件付平均 $E[Y|\boldsymbol{X}]$ に対して大きなバイアスを生じてしまう.

誤差項の分布が非対称/不均一な場合, 被説明変数 Y に Box-Cox 変換等の種々の変数変換を適用することが有効である. 変換後のデータが対称/均一に分布するなどの望ましい統計的性質を持てば, 標準的なデータ解析技術を利用してよい推定量を構築できる. しかし, 条件付平均 $E[Y|\boldsymbol{X}]$ を推定する問題では, ロバスト統計を用いた場合と同様に変換バイアス (Transformation Bias) [4] と呼ばれるバイアスを生じてしまう.

3 分位点回帰を用いたロバスト回帰

準備として条件付分位点を定義する. 確率密度関数 $p(\boldsymbol{X}, Y)$ の $\boldsymbol{X} = \boldsymbol{x}$ における条件付 $q \in (0,1)$ 分位点 $f_q(\boldsymbol{x})$ は, $F_{Y|\boldsymbol{X}}(f_q(\boldsymbol{x})|\boldsymbol{x}) = q$ となるような $f_q(\boldsymbol{x})$ として定義される. ここで, $F_{Y|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ は条件付分布 $P(Y|\boldsymbol{X} = \boldsymbol{x})$ の累積分布関数である. 条件付分位点を推定するさまざまな推定量が提案されている [1]. これらは中央値の一致推定量である L_1 回帰推定量の拡張であり, Y のはずれ値に対してロバストである.

[均一分散モデル]: まず, (1) において, 誤差項 $\epsilon(\boldsymbol{X})$ が $\boldsymbol{X}$ に依存しない場合を考える. この場合, 条件付 q 分位点 $f_q(\boldsymbol{x})$ と条件付平均 $\mu(\boldsymbol{x})$ の関係は, すべての $\boldsymbol{x}$ において, 次のように表される:

$$\mu(\boldsymbol{x}) = f_q(\boldsymbol{x}) - F_\epsilon^{-1}(q).$$

ここで, F_ϵ は ϵ の累積分布関数である. これを利用すると, なんらかの q 分位点回帰推定量 $\hat{f}_q$ を用いて条件付平均関数 $\mu(\boldsymbol{x})$ の「部分的にロバストな」推定量を次のように構築できる:

$$\hat{\mu}(\boldsymbol{x})_{\mathrm{Homo}} = \hat{f}_q(\boldsymbol{x}) + \hat{c}_q, \tag{2}$$

ここで, $\hat{c}_q$ は $F_\epsilon^{-1}(q)$ の最小二乗推定値である. $\hat{\mu}_{\mathrm{Homo}}$ のうち, 関数の「形状」を特徴づける大部分のパラメータは分位点回帰によってロバストに推定され, 関数の「位置」を特徴づけるひとつのパラメータ $\hat{c}_q$ のみがロバストでない算術平均 (最小二乗法) により推定される.

[位置尺度モデル]: 次に条件付分布の位置だけでなく尺度も $\boldsymbol{X}$ に依存するような回帰モデルを考える. これは (1) における誤差項が $\epsilon(\boldsymbol{X}) = \sigma(\boldsymbol{X}) \cdot \epsilon$ と表現される場合である. このとき, 簡単な計算により, すべての $\boldsymbol{x}$ に対し, 条件付 q 分位点 $f_q(\boldsymbol{x})$ と条件付平均 $\mu(\boldsymbol{x})$ の間には

$$f_q(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) F_\epsilon^{-1}(q) \tag{3}$$

という関係が成り立つ. 異なる $q_1, q_2 \in (0,1)$ に対して (3) を用い, これを $\mu(\boldsymbol{x})$, $\sigma(\boldsymbol{x})$ に関する二次の連立方程式とみなすと, 条件付平均 $\mu(\boldsymbol{x})$ は二つの条件付分位点 $f_{q_1}(\boldsymbol{x})$ と $f_{q_2}(\boldsymbol{x})$ の線形和 (ただし, 係数の和は 1) として表すことができる. これを利用すると, 次のような推定量を考えることができる:

$$\hat{\mu}_{\mathrm{LSM}}(\boldsymbol{x}) = \hat{\beta}\hat{f}_{q_1}(\boldsymbol{x}) + (1-\hat{\beta})\hat{f}_{q_2}(\boldsymbol{x}). \tag{4}$$

ここで, $\hat{\beta}$ は最小二乗推定量で推定される. (2) と同様に, $\hat{\mu}_{\mathrm{LSM}}$ の大部分のパラメータが Y のはずれ値に対してロバストな分位点回帰により推定される.

[一般化位置尺度モデル]: 位置尺度モデルにおいて $N \geq 3$ 個以上の分位点回帰を利用することにより, 次のような推定量を構築することができる:

$$\hat{\mu}_{\mathrm{GLSM}}(\boldsymbol{x}) = \textstyle\sum_{n=1}^{N} \hat{\beta}_n \hat{f}_{q_n}(\boldsymbol{x}), \ \ \sum_{n=1}^{N} \beta_n = 1.$$

ここで, $\hat{f}_{q_1}, \ldots, \hat{f}_{q_n}$ は分位点回帰により, $\hat{\beta}_1, \ldots, \hat{\beta}_n$ は最小二乗推定量により推定される.

4 自動車保険データ解析

上述の推定量を北米の自動車保険データ解析に適用した. データ数 100000 個それぞれにおける 5 個リスクファクターから支払い保険金の期待値を推定する問題である. ブートストラップを用いた分析により, このデータが位置尺度モデルにより近似できることがわかった. また, 推定量 (4) は, 条件付平均の推定量として, 従来のものよりも (平均二乗誤差の意味で) 良いことが示された.

参考文献

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

[2] I. Takeuchi, Y. Bengio, and T. Kanamori. Robust regression with asymmetric heavy-tail noise distributions. *Neural Computation*, 14(10):2469–2496, 2002.

[3] T.Kanamori and I.Takeuchi. Robust estimation of conditional mean by the linear combination of quantile regressions. Technical Report B-394, Tokyo Institute of Technology, Dept. of Math. and Comp. Science, 2003.

[4] J. Neyman and E. L. Scott. Correction for bias introduced by a transformation of variables. *Ann. Math. Stat.*, 31:643–655, 1960.

Unbiased Estimation of Functionals under Random Censorship

北海道大学·経済　鈴川 晶夫

1　はじめに

打切りデータに基づく推定問題について考える. n個の個体の生存時間を表す正値確率変数$X_1,\ldots,X_n$が互いに独立で, 未知の分布F (生存関数$\bar{F}=1-F$)に従うとする. 個体iに対して, Y_iを打切りまでの時間を表す正値確率変数とする. ランダム右側打切りデータにおいては, X_i, Y_i, $i=1,\ldots n$ は直接観測されず,

$$(Z_i,\delta_i)=(\min(X_i,Y_i),\ I(X_i\le Y_i)),\quad i=1,\ldots,n$$

が観測される. ただし, $I(A)$を集合Aの定義関数とする.

打切り時間$Y_1,\ldots,Y_n$は互いに独立に未知の分布G(生存関数$\bar{G}=1-G$)に従い, $X_1,\ldots,X_n, Y_1,\ldots,Y_n$は互いに独立であることを仮定する.

本報告では, 既知のF-可測関数φに対して, $\int\varphi\,dF$ の推定問題を考える.

2　Kaplan-Meier 積分とその表現

F_nをFに対するKaplan-Meier (1958) 推定量(KM推定量)とする. $\int\varphi\ dF$の自然な推定量である$\int\varphi\,dF_n$はKaplan-Meier積分(KM積分)とよばれる.

Hを$Z_i=\min(X_i,Y_i)$の分布関数とし, $\tau=\inf\{z\,;\,H(z)=1\}$とおく. Stute and Wang (1993)は, $\int\varphi\,dF_n$が$\int_0^\tau\varphi\,dF$の強一致推定量であることを示した. また, $\int\varphi\,dF_n$の分布収束は, Gill (1983), Schick et al. (1988), Yang (1994), Stute (1995) によって調べられた.

打ち切り時間分布のKM推定量を用いてKM積分を表現することができる. すなわち, G_nを打ち切り時間分布Gに対するKM推定量とするとき, KM積分$\int\varphi\,dF_n$を

$$\int\varphi\,dF_n=n^{-1}\sum_{i=1}^{n}\frac{\delta_i\varphi(Z_i)}{\bar{G}_n(Z_i-)}$$

と表現することができる. GのKM推定量であるG_nは, ほとんどの統計ソフトで計算できることから, この表現を用いることによって, KM積分$\int\varphi\,dF_n$を容易に計算できる

KM積分は$\int_0^\tau\varphi\,dF$の推定量としてバイアスをもち(Mauro 1985; Zhou 1988; Stute 1994), φが有界でない場合にはそのバイアスは深刻であることが指摘されている(Stute 1994). KM積分に対する上の表現式は, Gが既知である場合における$\int_0^\tau\varphi\,dF$の不偏推定量を示唆する. Gが既知のとき,

$$S_{0n}(G)=n^{-1}\sum_{i=1}^{n}\frac{\delta_i\varphi(Z_i)}{\bar{G}(Z_i-)}$$

とおくと, $S_{0n}(G)$は$\int_0^\tau\varphi\,dF$の不偏推定量である. KM積分$\int\varphi\,dF_n$は, $S_{0n}(G_n)$ (G_nの差込) に他ならない. この差込によって, バイアスが生じる. しかし, 不偏推定量$S_{0n}(G)$は, 漸近的にKM積分$S_{0n}(G_n)$よりも大きな分散をもつ.

3　不偏推定量

打ち切り時間分布Gが既知の場合において, $\theta=\int_0^\tau\varphi\,dF$の不偏推定について考える. 次の推定量を考える.

$$\hat{\theta}_n(\varphi_1,\varphi_0)=n^{-1}\sum_{i=1}^{n}\{\delta_i\,\varphi_1(Z_i)+(1-\delta_i)\,\varphi_0(Z_i)\}$$

ただし, 関数φ_1とφ_0はFに依存しない. しかし, これらは既知のGに依存してもよい.

これらの関数を, $\varphi_1(z)=\varphi(z)/\bar{G}_n(z-)$, $\varphi_0(z)\equiv 0$ のように選んだ場合には, $\hat{\theta}_n(\varphi_1,\varphi_0)$ はKM積分 $\int\varphi\, dF_n$ に一致する.

推定量$\hat{\theta}_n(\varphi_1,\varphi_0)$ がFによらず$\theta=\int_0^\tau \varphi\, dF$の不偏推定量であるための必要十分条件は, 関数$\varphi_1$と$\varphi_0$が次の条件を満たすことである.

$$\bar{G}(z-)\varphi_1(z)+\int_0^{z-}\varphi_0(y)\,dG(y)=\varphi(z)\quad \text{for any } 0<z\le\tau$$

$$\int_0^{\tau_G}\varphi_0(y)\,dG(y)=0\quad \text{if}\quad \tau_G=\tau$$

これらの条件を満たす関数φ_1とφ_0を求めることにより, 不偏推定量

$$U_n(\gamma;G)=n^{-1}\sum_{i=1}^{n}\left\{\delta_i\frac{\varphi(Z_i)}{\bar{G}(Z_i-)}+(1-\delta_i)\gamma(Z_i)-\int_0^{Z_i-}\frac{\gamma(y)}{\bar{G}(y)}\,dG(y)\right\}$$

が導出される. ただし, $[0,\tau]$上の関数γは次の条件を満たす.

$$\int_0^{z-}\frac{|\gamma(y)|}{\bar{G}(y)}\,dG(y)<\infty\quad \text{for any } 0<z\le\tau$$

$$\gamma(\tau_G)\{G(\tau_G)-G(\tau_G-)\}=0\quad \text{if}\quad \tau_G=\tau$$

不偏推定量$U_n(\gamma;G)$の分散は次で与えられる.

$$n\times \mathrm{Var}[U_n(\gamma;G)]=\int_0^\tau\frac{\{\varphi(x)\}^2}{\bar{G}(x)}\,dF(x)-\left\{\int_0^\tau\varphi(x)\,dF(x)\right\}^2$$
$$+\int_0^\tau\bar{F}(x)\{\gamma(x)\}^2\,dG(x)-2\int_0^\tau\frac{\gamma(x)}{\bar{G}(x)}\left\{\int_x^\tau\varphi(t)\,dF(t)\right\}dG(x)$$

また, この分散を最小化するという意味での最適なγは,

$$\gamma_{opt}(x)=\{\bar{H}(x)\}^{-1}\int_x^\tau\varphi(t)\,dF(t)$$

により与えられる. 最適な不偏推定量$U_n(\gamma_{opt};G)$は, KM積分と漸近的に同等である.

参考文献

[1] Gill, R. D. (1983). Large sample behavior of the product limit estimator on the whole line. *Ann. Statist.* **11**, 49-58.

[2] Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457-481.

[3] Mauro, D. (1985). A combinatoric approach to the Kaplan-Meier estimator. *Ann. Statist.* **13**, 142-149.

[4] Schick, A., Susarla, V. and Koul, H. (1988). Efficient estimation of functionals with censored data. *Statist. and Decisions* **6**, 349-360.

[5] Stute, W. (1994). The bias of Kaplan-Meier integrals. *Scand. J. Statist.* **21**, 475-484.

[6] Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23**, 422-439.

[7] Stute, W. and Wang, J. L. (1993). The strong law under random censorship. *Ann. Statist.* **21**, 1591-1607.

[8] Yang, S. (1994). A central limit theorem for functionals of the Kaplan-Meier estimator. *Statist. Probab. Letters* **21**, 337-345.

[9] Zhou, M. (1988). Two-sided bias bound of the Kaplan-Meier estimator. *Probab. Theory Rel. Fields* **79**, 165-173.

多項母集団の一様性検定統計量における近似について

帯広畜産大学・畜産　種市信裕
北海道教育大学釧路校・教育　関谷祐里

1　多項母集団の一様性検定

表 1 の $s \times r$ 分割表において、各列の和が固定されている r 個の s 項分布からなる積多項分布モデルを考える。すなわち、

$$(X_{1j},\ldots,X_{sj})' \sim \mathrm{Mult}_s(n_j;p_{1j},\ldots,p_{sj}), \quad (j=1,\ldots,r)$$

とする。ここで、$\sum_{i=1}^{s} X_{ij} = n_j$, $(j=1,\ldots,r)$, $0 < p_{ij} < 1$, $(i=1,\ldots,s; j=1,\ldots,r)$, $\sum_{i=1}^{s} p_{ij} = 1$, $(j=1,\ldots,r)$ である。このとき、一様性の帰無仮説

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{ir} \equiv q_i, \qquad (i=1,\ldots,s)$$

を検定するためのパワーダイバージェンス統計量は、

$$R^a = 2\sum_{j=1}^{r} n_j I^a(\boldsymbol{p}_j^*, \boldsymbol{q}^*)$$

によって与えられている（ Read and Cressie [2, pp. 23–24] ）。ここで、

$$I^a(\boldsymbol{p}_j^*, \boldsymbol{q}^*) = \begin{cases} \dfrac{1}{a(a+1)}\displaystyle\sum_{i=1}^{s} p_{ij}^* \left\{ \left(\frac{p_{ij}^*}{q_i^*} \right)^a - 1 \right\} & (a \neq 0, -1) \\ \displaystyle\sum_{i=1}^{s} p_{ij}^* \log\left(\frac{p_{ij}^*}{q_i^*} \right) & (a=0) \\ \displaystyle\sum_{i=1}^{s} q_i^* \log\left(\frac{q_i^*}{p_{ij}^*} \right) & (a=-1), \end{cases}$$

$p_{ij}^* = X_{ij}/n_j$, $q_i^* = X_{i\cdot}/n$, $X_{i\cdot} = \sum_{j=1}^{r} X_{ij}$, $n = \sum_{j=1}^{r} n_j$ である。パワーダイバージェンス統計量 R^a の族は、対数尤度比統計量（R^0）やカイ 2 乗統計量（R^1）を含んでいる。また、$R^{\frac{2}{3}}$ は、多項分布の適合度検定統計量に対して Cressie and Read [1] と Read and Cressie [2] により推奨された統計量に対応する。もし、

$$n_j/n \to \nu_j \ (0 < \nu_j < 1) \text{ for each } j, \text{ as } n \to \infty$$

表 1: $r \times s$ 分割表

母集団 category	1	$\cdots$	r	計
A_1	X_{11}	$\cdots$	X_{1r}	$X_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_s	X_{s1}	$\cdots$	X_{sr}	$X_{s\cdot}$
計	n_1	$\cdots$	n_r	n

を仮定するならば、R^a は H_0 のもとで a の値によらず漸近的に自由度 $(r-1)(s-1)$ のカイ二乗分布に従うことが知られている。この漸近分布を用いて、通常検定のために $\Pr\{R^a \leq x|H_0\} \approx A_0(x)$ ただし、

$$A_0(x) = \Pr\left\{\chi^2_{(r-1)(s-1)} \leq x\right\}$$

という近似を用いる。ここで χ^2_f は自由度 f のカイ二乗分布に従う確率変数とする。

2 カイ二乗近似の改良

本報告においては、$A_0(x)$ で与えられるカイ二乗近似の改良のために以下の2種類の検定統計量 R^a の分布の近似を提案する。

(1) 連続分布を仮定した多変量エッジワース展開による近似。つまり、$\Pr\{R^a \leq x|H_0\} \approx A_1(x)$ ただし、

$$A_1(x) = \Pr\left\{\chi^2_{(r-1)(s-1)} \leq x\right\} + \frac{1}{24n}\sum_{j=0}^{3} w_j \Pr\left\{\chi^2_{(r-1)(s-1)+2j} \leq x\right\}$$

という形での近似。

(2) H_0 のもとでの $R^a = W^a + O_p(n^{-3/2})$ なる展開式 W^a を考え、$(W^a - \gamma_a)/\sqrt{\delta_a}$ の期待値と分散がそれぞれ $\chi^2_{(r-1)(s-1)}$ の期待値と分散である $(r-1)(s-1)$, $2(r-1)(s-1)$ と $o(n^{-1})$ まで一致するように γ_a と δ_a を求め、これらを用いて $\Pr\{R^a \leq x|H_0\} \approx A_2(x)$ ただし、

$$A_2(x) = \Pr\left\{\chi^2_{(r-1)(s-1)} \leq \frac{x-\gamma_a}{\sqrt{\delta_a}}\right\}$$

とする近似。

3 数値実験による近似の比較

本報告においては、積多項モデルから直接計算をおこなうことによって得られる検定統計量 R^a の正確な確率を用いた数値実験により、カイ二乗近似 $A_0(x)$、連続分布を仮定したエッジワース展開による近似 $A_1(x)$ およびモーメント修正による近似 $A_2(x)$ の性能の比較をおこなう。

参考文献

[1] Cressie, N. and Read, T. R. C. : Multinomial goodness-of-fit tests, *J. R. Statist. Soc.*, B, **46** (1984), 440-464.

[2] Read, T. R. C. and Cressie, N. A. C. : *Goodness-of-fit statistics for discrete multivariate data*, (1988), Springer.

Asymptotic Expansion for Distributions of Test Statistics for Profile Analysis in Elliptical Populations

東京理科大学・理学研究科　三浦　徳仁
東京理科大学・理学部　瀬尾　隆

1. はじめに

多変量分布の平均ベクトルに対するプロフィール分析で用いられる検定統計量の帰無分布について考察する．二標本問題において，プロフィール分析とは，(i) 各プロフィールが平行かどうか，(ii) 各プロフィールが平行であるとき，さらに一致しているかどうか，(iii) 各プロフィールが一致しているとき，さらに水平であるかどうか，等の検定を行う分析（Rencher(1995) 参照）である．本報告ではプロフィール分析の前に行う，一標本問題における平均ベクトルの成分に対する同等性検定（$H_0 : C\boldsymbol{\mu} = \mathbf{0}$）における検定統計量

$$T_1^2 \equiv N(C\overline{\boldsymbol{X}})'(CSC')^{-1}(C\overline{\boldsymbol{X}}),$$

および二標本問題における平行プロフィールの検定（$H_0 : C\boldsymbol{\mu}^{(1)} = C\boldsymbol{\mu}^{(2)}$）における検定統計量

$$T_{21}^2 \equiv (\overline{\boldsymbol{X}}_{(1)} - \overline{\boldsymbol{X}}_{(2)})'C'\left[\left(\frac{1}{N_1} + \frac{1}{N_2}\right)CS_pC'\right]^{-1}C(\overline{\boldsymbol{X}}_{(1)} - \overline{\boldsymbol{X}}_{(2)})$$

の帰無分布について考察する．

母集団分布として正規分布を仮定した場合，二つの検定統計量 T_1^2, T_{21}^2 は F 分布によりパーセント点が表現されるが，楕円母集団のもとでは一般にそうではない．本報告では楕円母集団のもとでの T_1^2, T_{21}^2 の漸近展開を考えることにより，非正規性の影響を調べる．ここに $\overline{\boldsymbol{X}}$ は標本平均ベクトル，S は標本分散共分散行列，$\overline{\boldsymbol{X}}_{(j)}$ は第 j 母集団からの標本平均ベクトル $(j = 1, 2)$，S_p はプールされた標本分散共分散行列である．C は，ランクが $p-1$ で $C(1\,1\cdots1)' = \mathbf{0}$，$CC' = I_{p-1}$ を満たす $(p-1)\times p$ 行列である．

2. 楕円分布

$p \times 1$ の確率ベクトル $\boldsymbol{X}$ が次の形の密度関数をもつとき，$\boldsymbol{X}$ は p 変量楕円分布に従うという．(Muirhead(1982) 参照)．

$$c_p|\Sigma|^{-\frac{1}{2}}h((\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \quad c_p\text{は正定数}, h\text{はある非負関数}.$$

また特性関数は

$$\phi(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{\mu})\psi(\boldsymbol{t}'\Lambda\boldsymbol{t}), \quad \phi\text{はある関数}.$$

このとき，$\boldsymbol{X}$ の平均ベクトルは $\boldsymbol{\mu}$，分散共分散行列は $-2\phi'(0)\Lambda$ で与えられる．

3. 楕円母集団のもとでの漸近展開

T_1^2, T_{21}^2 の構造から，一般性を失うことなく，母集団の分散共分散行列は単位行列としてよく，$\overline{\boldsymbol{X}}, W = (N-1)/NS + (\overline{\boldsymbol{X}} - \boldsymbol{\mu})(\overline{\boldsymbol{X}} - \boldsymbol{\mu})', \overline{\boldsymbol{X}}_{(j)}, W_{(j)} = (N_j - 1)/N_j S_{(j)} + (\overline{\boldsymbol{X}}_{(j)} - \boldsymbol{\mu}^{(j)})(\overline{\boldsymbol{X}}_{(j)} - \boldsymbol{\mu}^{(j)})'$ に対し，

$$\begin{cases} \overline{\boldsymbol{X}} = \boldsymbol{\mu} + \dfrac{1}{\sqrt{N}}\boldsymbol{z} \\ W = I_p + \dfrac{1}{\sqrt{N}}Z \end{cases} \qquad \begin{cases} \overline{\boldsymbol{X}}_{(j)} = \boldsymbol{\mu}^{(j)} + \dfrac{1}{\sqrt{N_j}}\boldsymbol{z}_{(j)} \\ W_{(j)} = I_p + \dfrac{1}{\sqrt{N_j}}Z_{(j)} \end{cases}$$

$(j = 1, 2)$ とおき，摂動展開することによって，帰無仮説のもとでの T_1^2, T_{21}^2 の分布の漸近展開を求め，次の定理を得る．

【定理 1】 T_1^2 の分布は

$$\Pr(T_1^2 \le t^2) = \Pr(\chi_{p-1}^2 \le t^2) + \frac{1}{4N}\sum_{j=0}^{2} b_j \Pr(\chi_{p+2j-1}^2 \le t^2) + o(N^{-1})$$

と漸近展開され，上側 100α %点 $t_1^2(\alpha)$ は

$$t_1^2(\alpha) = \chi_{p-1}^2(\alpha) + \frac{2}{N}\chi_{p-1}^2(\alpha)(b_3 + b_4\chi_{p-1}^2(\alpha)) + o(N^{-1})$$

で与えられる．

【定理 2】 T_{21}^2 の分布は

$$\Pr(T_{21}^2 \le t^2) = \Pr(\chi_{p-1}^2 \le t^2) + \frac{1}{8N}\sum_{j=0}^{2} d_j \Pr(\chi_{p+2j-1}^2 \le t^2) + o(N^{-1})$$

と漸近展開され，上側 100α %点 $t_{21}^2(\alpha)$ は

$$t_{21}^2(\alpha) = \chi_{p-1}^2(\alpha) - \frac{1}{4N(p-1)}\chi_{p-1}^2(\alpha)\left\{d_0 - \frac{d_2}{p+1}\chi^2(\alpha)\right\} + o(N^{-1})$$

で与えられる．

これらの定理における係数等，詳細は省略する．

最後に，上で求めた検定統計量の漸近展開の結果とシミュレーションの結果を比較し，近似の精度を調べる．

参考文献

[1] Muirhead, R. J., *Aspects of Multivariate Statistical Theory.* John Wiley and Sons, 1982.

[2] Rencher, A. C., *Methods of Multivariate Analysis.* John Wiley and Sons, 1995.

繰り返し測定データにおける多重比較について

九州大学・数理学研究院　百武　弘登

1. はじめに

k個の母集団があり、第i母集団からの観測値を$\boldsymbol{y}_{ir}=(y_{ir,1},\cdots,y_{ir,p})'$ $(i=1,\cdots,k,\ r=1,\cdots,n_i)$とする。ただし、$y_{ir,j}$は第$i$母集団の$r$番目の個体の時点$t_j$での観測値である。このとき、$y_{ir,j}$に対して、

$$y_{ir,j}=f(t_j;\boldsymbol{\beta}_i)+\varepsilon_{ir,j}, \tag{1}$$

と仮定する。ただし、f は非線形の既知関数、$\varepsilon_{ir,j}$ は誤差、$\boldsymbol{\beta}_i=(\beta_{i1},\cdots,\ \beta_{iq})'$ は未知パラメータで、$q\leq p$とする。また、$\boldsymbol{f}_i=(f(t_1;\boldsymbol{\beta}_i),\cdots,f(t_p;\boldsymbol{\beta}_i))'$ とおくと、$\boldsymbol{y}_{ir}=\boldsymbol{f}_i+\boldsymbol{\varepsilon}_{ir}$ と表せる。ただし、$\boldsymbol{\varepsilon}_{ir}=(\varepsilon_{ir,1},\cdots,\varepsilon_{ir,p})'$ であり、$E[\boldsymbol{\varepsilon}_{ir}]=\mathbf{o}$, $\mathrm{Var}[\boldsymbol{\varepsilon}_{ir}]=\boldsymbol{\Sigma}$、そして、$\boldsymbol{\varepsilon}_{ir}$は独立と仮定する。このとき、パラメータ$\boldsymbol{\beta}_i$ の非線形関数 $g_i=g(\boldsymbol{\beta}_i)$ に関して比較を行うための同時信頼区間の構成をする。たとえば、g_i としては薬物動態モデル(Davidian and Giltinan (1995) 参照) $f(t;\boldsymbol{\beta}_i)=\beta_{i1}te^{-\beta_{i2}t}$ のtに関する最大値$\beta_{i1}/(e\beta_{i2})$ を考えることがある。

2. 対比較

$\boldsymbol{\beta}_i$ のOLSEを$\hat{\boldsymbol{\beta}}_i$ とし、$\boldsymbol{V}_i=\sum_r(\boldsymbol{y}_{ir}-\hat{\boldsymbol{f}}_i)(\boldsymbol{y}_{ir}-\hat{\boldsymbol{f}}_i)'$ とする。ただし、$\hat{\boldsymbol{f}}_i=(f(t_1;\hat{\boldsymbol{\beta}}_i),\cdots,f(t_p;\hat{\boldsymbol{\beta}}_i))'$ である。Seber and Wild (1989) のように、テーラー展開により、

$$\hat{\boldsymbol{f}}_i=\boldsymbol{f}(\hat{\boldsymbol{\beta}}_i)\ \approx\ \boldsymbol{f}(\boldsymbol{\beta}_i)+\boldsymbol{F}_{\cdot}^{(i)}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)$$

$$\hat{g}_i=g(\hat{\boldsymbol{\beta}}_i)\approx g(\boldsymbol{\beta}_i)+\boldsymbol{g}_{i\cdot}'(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)$$

と近似できる。ただし、$\boldsymbol{F}_{\cdot}^{(i)}=(\partial\boldsymbol{f}_i/\partial\boldsymbol{\beta}_i')$, $\boldsymbol{g}_{i\cdot}'=(\partial g_i/\partial\beta_{i1},\ \cdots\ ,\partial g_i/\partial\beta_{iq})$ である。ここで、$\boldsymbol{\varepsilon}_{ir}$ が正規分布 $N_p(\mathbf{0},\Sigma)$ に従うとすれば、$g(\boldsymbol{\beta}_i)+\boldsymbol{g}_{i\cdot}'(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)$ は近似的に$N\Big(0,\ \boldsymbol{g}_{i\cdot}'(\boldsymbol{F}_{\cdot}^{(i)'}\boldsymbol{F}_{\cdot}^{(i)})^{-1}\boldsymbol{F}_{\cdot}^{(i)'}\boldsymbol{\Sigma}\boldsymbol{F}_{\cdot}^{(i)}(\boldsymbol{F}_{\cdot}^{(i)'}\boldsymbol{F}_{\cdot}^{(i)})^{-1}\boldsymbol{g}_{i\cdot}/n_i\Big)$ に従うし、$\boldsymbol{V}=\boldsymbol{V}_1+\cdots+\boldsymbol{V}_k$ は近似的にウィシャート分布$W_p(\boldsymbol{\Sigma},\ \nu)$ に従う。ただし、$\nu=n_1+\cdots+n_k-k$ である。つまり、$S=V/\nu$ はΣ の推定量として考えられる。これらの近似より、ペアごとの比較に対する同時信頼区間は

$$g_i-g_{i'}\in\hat{g}_i-\hat{g}_{i'}\pm q\sqrt{\boldsymbol{a}_i'\boldsymbol{S}\boldsymbol{a}_i/n_i+\boldsymbol{a}_{i'}'\boldsymbol{S}\boldsymbol{a}_{i'}/n_{i'}}\ \ for\ all\ i\neq i' \tag{2}$$

により近似できる。ただし、$\boldsymbol{a}_i=\boldsymbol{F}_{\cdot}^{(i)}(\boldsymbol{F}_{\cdot}^{(i)'}\boldsymbol{F}_{\cdot}^{(i)})^{-1}\boldsymbol{g}_{i\cdot}$ であり、$\sqrt{2}q$ はスチューデント化範囲分布の両側α点である。

3. コントロールとの比較

$\boldsymbol{y}_{0r} = (y_{0r,1}, \cdots, y_{0r,p})'$ $(r = 1, \cdots, n_0)$ をコントロール母集団からの観測値とする。このとき、$g_i (i = 1, \cdots, k)$ と g_0 の比較に対して、$g_i - g_0$ の同時信頼区間は、前節の近似を用いると

$$g_i - g_0 \in \hat{g}_i - \hat{g}_0 \pm d\sqrt{\boldsymbol{a}_i' S \boldsymbol{a}_i / n + \boldsymbol{a}_0' S \boldsymbol{a}_0 / n} \ for\ i = 1, \cdots, k \tag{3}$$

で与えられる。ただし、dはダネットの同時信頼区間に対するα点である。

分布の自由度をνとしたとき、Hyakutake (2003) の (2) に対するシミュレーションにより、被覆確率が設定した値より大きく傾向がある。そこで (3) に対する自由度を

$$\frac{1}{k}\sum_{i=1}^{k} \frac{(\boldsymbol{b}_i'\boldsymbol{b}_i + \boldsymbol{b}_0'\boldsymbol{b}_0)^2}{(\boldsymbol{b}_i'\boldsymbol{b}_i)^2 + (\boldsymbol{b}_0'\boldsymbol{b}_0)^2 + 2(\boldsymbol{b}_i'\boldsymbol{b}_0)^2}\nu, \tag{4}$$

により近似する。

近似の精度をシミュレーションにより検証をおこなった。モデルとしては薬物動態モデル

$$f(t; \boldsymbol{\beta}_i) = \beta_{i1} t e^{-\beta_{i2} t} \tag{5}$$

を用い、(5) の t に関する最大値 $g_i = \beta_{i1} e^{-1}/\beta_{i2}$ の同時信頼区間を5000組構成した。パラメータの設定は Table 1 で与えている。シミュレーションの結果は Table 2 である。これより、(4) の自由度の近似が良好であることがわかる。ロジスティックモデルにおける変曲点の比較についてもシミュレーションを行ったが、似たような結果となった。

Table 1. パラメータ

Population	0	1	2	3
β_{i1}	0.8	0.9	1.0	0.8
β_{i2}	0.6	0.5	0.4	0.4

Table 2. シミュレーション結果

	$k=2$		$k=3$	
n	Σ_1	Σ_2	Σ_1	Σ_2
5	.9508	.9512	.9604	.9508
	(.9556)	(.9550)	(.9640)	(.9542)
8	.9520	.9500	.9502	.9506
	(.9540)	(.9508)	(.9520)	(.9530)
12	.9514	.9516	.9544	.9506
	(.9534)	(.9522)	(.9550)	(.9524)

参考文献

Davidian and Giltinan: Nonlinear Models for Repeated Measurement Data, Chapman & Hall/CRC (1995).

Hyakutake: Biom. J., 45, 772-780 (2003).

Seber and Wild: Nonlinear Regression, Wiley.(1989)

ある種の楕円母集団での分布と共分散行列の検定について

富士大学・経　早川　毅

0 序　最近 Bayes Inference, Prediction Inference 等の分野で, 例えば Ng (2002, Bayes Inference), Khan (2002, T-variate), Ng (2000, Prediction Inference) etc, 問題設定において母集団分布が次の様な条件をみたしている場合が取り扱われている。

$X = [x_1, x_2, \cdots, x_n]_{m\times n}$ $(m \le n)$ の同時密度関数が

$$(1) \qquad |\Lambda|^{-n/2} g(\Lambda^{-1} X X') = |\Lambda|^{-n/2} g\left(\sum_{i=1}^{n} x_i' A x_i\right)$$

と表現されるとする. 本報告では,

(1) 条件 (C) のもとでの行列変数の分布

(2) 正規母集団のもとでの共分散行列の一様性に関する尤度比規準の分布を母集団数 k が十分に大きく, 各母集団からの標本数がそれ程は大きくはない場合についての漸近展開

を与える.

1. 二次形式の分布

$X_{m\times n}$ が (C) を持つとする. $A = A' > 0$, $Z = XAX'$.
$g(\cdot)$ は適当な範囲で展開が可能とする.

このとき, Z の同時密度関数について2つの表現を与える. また Löwner の意味での分布関数, Z の最大固有根の分布関数, 及び Z の密度, 分布関数等を与える. 例えば, Z の密度関数の中級数表示は次で与えられる.

$$f(Z) = \frac{\pi^{\frac{1}{2}mn}}{\Gamma_m\left(\frac{n}{2}\right)|A|^{m/2}|\Lambda|^{n/2}} |Z|^{\frac{1}{2}(n-m-1)} \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!} \sum_{\kappa} \frac{C_\kappa(A^{-1}) C_\kappa(\Lambda^{-1}Z)}{C_\kappa(I_n)},$$

$$P\{Z<\Omega\} = \frac{\pi^{\frac{1}{2}mn}\Gamma_m\left(\frac{m+1}{2}\right)}{\Gamma_m\left(\frac{n+m+1}{2}\right)|\Lambda|^{n/2}|A|^{m/2}}\,|\Omega|^{n/2}$$
$$\sum_{k=0}^{\infty}\frac{g^{(k)}(0)}{k!}\sum_{\kappa}\frac{\left(\frac{n}{2}\right)_\kappa}{\left(\frac{n+m+1}{2}\right)_\kappa}\frac{C_\kappa(A^{-1})\,C_\kappa(\Lambda^{-1}\Omega)}{C_\kappa(I_n)}$$

2 Non-central case

$X_{m\times n}$ の密度関数が $|\Lambda|^{-n/2}g(tr\Lambda^{-1}(X-M)(X-M)')$ の場合に $Z=XX'$ の分布についても同様の考察を行い、類似の結果を得る.

3 Dirichlet 分布

$X=[X_1,X_2,\cdots,X_k]_{m\times n}$, $X_i: m\times n_i$, $i=1,2,\cdots k$ の同時密度関数が

$$\prod_{i=1}^{k}|A_i|^{-n_i/2}g(tr\Lambda_i^{-1}X_iX_i')$$

とする。 $\Lambda_1=\Lambda_2=\cdots=\Lambda_k=\Lambda$ のもとで、 $A_i=X_iX_i'$, $A=\sum_{i=1}^{k}A_i$ とするとき、 $U_i=A^{-1/2}A_iA^{-1/2}$, $i=1,2,\cdots,k$ は Dirichlet 分布とする.

4 共分散行列の等値性.

k個の正規母集団の共分散行列の等値に関する尤度比規準が Dirichlet 変数で表示できることより Elliptical 分布の場合に適用させる. このとき、母集団の数kは十分に大きいが、母集団標本数は適当な大きさとした時の分布の漸近展開を与える.

参考文献

[1] Khan, S., Jour. Mult. Analysis. 83, 124−140 (2002)

[2] Ng, V. M., Commun. Statist.−Theory. Meth., 29, 477−483 (200)

[3] Ng, V. M., Jour. Mult. Analysis, 83, 409−414 (2002)