数理解析研究所講究録
第1621巻 2009年 153-160

マルチスケール・ブートストラップ法による頻度論的およびベイズ的な信頼度計算とFDR

東京工業大学・情報理工　　下平英寿
Hidetoshi Shimodaira
Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

1 概要

データ $\mathcal{X} = (x_1, \ldots, x_n)$ から復元抽出によってブートストラップ標本 $\mathcal{X}^* = (x_1^*, \ldots, x_{n'}^*)$ を生成するが，何らかの非線形変換により $\mathcal{X}$ を $m+1$ 次元 $(m \geq 1)$ ベクトル y に変換し，確率変数 Y が近似的に次の多変量正規分布に従うと仮定する．

$$\text{データ} \quad Y|\mu \sim N(\mu, I)$$

$$\text{ブートストラップ標本} \quad Y^*|y \sim N(y, \sigma^2 I), \quad \sigma^2 = \frac{n}{n'}$$

未知の平均ベクトル μ に関する何らかの関数 $f(\mu)$ に興味があり，それを $f(y)$ で推定する場合を考える．μ に大きさ1のノイズを加えて得られたのがデータ y であったとすれば，y にわざと大きさ σ^2 のノイズを加えて得られたのが y^* である．$Y^*|\mu \sim N(\mu, (1+\sigma^2)I)$ であるから，$\sigma^2 = -1$ と形式的におけば $Y^* = \mu$ の1点分布になり，$f(y^*) = f(\mu)$ はノイズのない推定になる．これがCook and Stefanski (1994) による回帰分析の観測誤差モデルにおける simulation-extrapolation (SIMEX) アルゴリズムのアイデアであった．別のたとえをすると，真の画像 μ に大きさ1のノイズが加わった画像データ y に，さらに大きさ σ^2 のノイズを加えた y^* を考えることにより μ を復元するという，信号処理の逆フィルタに相当する．

このような考え方を仮説検定の不偏な p-値を計算するために用いたのが，Shimodaira (2002, 2004, 2008) のマルチスケール・ブートストラップ法であり，2節で説明する．適当な領域 $\mathcal{H}_0 \subset \mathbb{R}^{m+1}$ に対して $\mu \in \mathcal{H}_0$ を帰無仮説とする．たとえば階層的クラスタリングで得られたクラスタの真偽を調べる問題を考える．興味のあるクラスタが木に含まれるとき $y \in \mathcal{H}_0$，そうでないとき $y \notin \mathcal{H}_0$ と領域を定義する．

これをEfron and Tibshirani (1998)は「領域の問題」(problem of regions)と呼んでいる．$\mathcal{H}_0$のブートストラップ確率は次式で定義される．

$$\alpha_{\sigma^2}(\mathcal{H}_0|y) = P_{\sigma^2}(Y^* \in \mathcal{H}_0|y)$$

ただし，P_{σ^2}とE_{σ^2}をスケールσ^2における確率および期待値とする．応用で使うときは，B個のブートストラップ標本$y^{*1}, \ldots, y^{*B}$からindicator functionを使って

$$\alpha_{\sigma^2}(\mathcal{H}_0|y) \approx \frac{1}{B}\sum_{b=1}^{B} 1_{\mathcal{H}_0}(y^{*b})$$

と計算する．これに$\sigma^2 = -1$とするアイデアを適用すると不偏なp-値が得られる．

テクニカルなことであるが，上記の手法の導出では$\mathcal{R}_0 \cup \mathcal{R}_1 = \mathbb{R}^{m+1}$のようにパラメータ空間が2個の領域へ分割されていてその境界$\partial\mathcal{R}_0$が十分に平坦であることを仮定している．データ点yの近傍でそのように近似できない場合は，なんらかの議論が必要になる．$\mathcal{R}_0 \cup \mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{R}^{m+1}$のように3個の領域に分割される場合を議論したのがShimodaira (submitted)であり，3節で説明する．領域が2個の場合は頻度論のp-値とベイズの事後確率を一致させるような“probability matching”事前分布があるが，そのおなじ事前分布を用いていても領域が3個になるとp-値と事後確率は一般に異なる．ここで議論しているp-値は領域が2個の場合は片側検定(one-sided, $s = 1$)に相当し，領域が3個の場合は両側検定(two-sided, $s = 2$)に相当する．これらを線形につないで得られる“zero-sided” ($s = 0$)の場合が事後確率になる．

現実の応用では，2個や3個ではなく多数の領域(互いにオーバーラップもある)$\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$があって，それらを同時に検定する状況がある．階層的クラスタリングでいえば，分類要素の部分集合の一つ一つが仮説に対応する．個々の領域の仮説検定に関しては2節の議論によって不偏なp-値$p_1, p_2, \ldots, p_K$が得られているとする．階層的クラスタリングでは$p_i < 0.05$かどうかではなく，$p_i > 0.95$かどうかに注目することがしばしばある．後者は$1 - p_i < 0.05$に注目するのと同じであるが，これは$\mathcal{H}_i^c = \mathbb{R}^{m+1} \setminus \mathcal{H}_i$を帰無仮説の領域とする仮説検定の$p$-値が$1 - p_i$であるから，もしこの検定が有意になれば「$\mu \in \mathcal{H}_i$である」と積極的に主張される．ところが帰無仮説$\mathcal{H}_1^c, \ldots, \mathcal{H}_K^c$を同時に検定して$1 - p_i < 0.05$となったクラスタを「発見」とみなすと，検定の多重性の問題が生じて，全体の誤り確率は0.05より大きくなってしまう．そこでベイズ的なfalse discovery rate (FDR)の意味での補正をp_iに適用して，「発見」されたクラスタが真実である事後確率π_iを計算する．この問題に再び

マルチスケール法の考え方を適用したのがShimodaira (in prep) であり，4 節で説明する．

2 マルチスケール・ブートストラップ法

$y=(y_1,\ldots,y_{m+1})$ を m 次元の $u=(y_1,\ldots,y_m)$ とのスカラの $v=y_{m+1}$ に分けて $y=(u,v)$ と考える．興味のある領域を次式で与える．

$$\mathcal{H}_0=\{(u,v):v\le -h(u),\, u\in\mathbb{R}^m\}$$

ここで，曲面 $\partial\mathcal{H}_0$ を表す連続関数 $h(u)$ は「十分に平坦」("nearly flat," Shimodiara 2008) と仮定する．すなわち，h の L^1 ノルムが $\|h\|_1<\infty$，フーリエ変換も $\|\mathcal{F}h\|_1<\infty$，そして h の L^∞ ノルムを $\|h\|_\infty=O(\Delta h)$ と表し，$\Delta h\to 0$ の漸近論を考える．

関数 h へ作用させる期待値オペレータ $\mathcal{E}_{\sigma^2}$ を次式で定義する．

$$(\mathcal{E}_{\sigma^2}h)(u):=E_{\sigma^2}(h(U^*)|u),\quad U^*\sim N(u,\sigma^2 I)$$

このとき，ブートストラップ確率が次式で表される．

$$\alpha_{\sigma^2}(\mathcal{H}_0|y)=\Phi\left(-\frac{v+\mathcal{E}_{\sigma^2}h(u)}{\sigma}\right)+O(\Delta h^2)$$

これはブートストラップ確率がスケール σ^2 に応じて変化する「スケール変換則」を表現する式である．一方，不偏な確率値は次式を満たす $p(\mathcal{H}_0|y)$ として定義される．

$$P_1\left(p(\mathcal{H}_0|Y)<\alpha|\mu\right)=\alpha,\quad \forall\mu\in\partial\mathcal{H}_0$$

不偏な確率値がもし存在すれば，次式が示される．

$$p(\mathcal{H}_0|y)=\Phi(-(v+\mathcal{E}_{-1}h(u)))+O(\Delta h^3)$$

したがって，ブートストラップ確率の式で形式的に $\sigma^2=-1$ とおくことにより次式が得られる．

$$p(\mathcal{H}_0|y)=\lim_{\sigma^2\to -1}\Phi\left(\sigma\Phi^{-1}(\alpha_{\sigma^2}(\mathcal{H}_0|y))\right)+O(\Delta h^2)$$

つまり，$\sigma^2=-1$ $(n'=-n)$ のブートストラップ確率から不偏な確率値が得られることになる．現実にはいくつかの $n'>0$ の値においてブートストラップ法を実行して $\sigma\Phi^{-1}(\alpha_{\sigma^2}(\mathcal{H}_0|y))$ の値を数値的に計算し，それを $\sigma^2=-1$ へ外挿する．

簡単のため $\psi(\sigma^2) = -\sigma\Phi^{-1}(\alpha_{\sigma^2}(\mathcal{H}_0|y))$ という記号を用いる．もし $\partial\mathcal{H}_0$ がなめらかな曲面ならば，h をテイラー展開したときの偶数次の係数によって $\psi(\sigma^2) = \beta_0+\beta_1\sigma^2+\beta_2\sigma^4+\cdots$ と展開できる．とくに β_0 は y から $\partial\mathcal{H}_0$ までの符号付き距離，β_1 は $\partial\mathcal{H}_0$ の曲率に相当する．$\sigma^2 = -1$ への外挿は $p(\mathcal{H}_0|y) = \Phi(-\psi(-1)) = \Phi(-\beta_0+\beta_1-\beta_2+\cdots)$ である．ところがもし $\partial\mathcal{H}_0$ がなめらかなでない場合，曲面のテイラー展開に基づく通常の漸近論を適用することができず，上記の議論も成り立たない．たとえば $\mathcal{H}_0$ が錘で y がその頂点付近にある場合，$\psi(\sigma^2) = \beta_0 + \beta_1\sigma$ は良い近似になるが，この式を $\sigma^2 < 0$ へ外挿することはできない．

そこで Shimodiara (2008) では，通常の漸近論における h のテイラー展開のかわりに，h のフーリエ変換にもとづいた議論を試みた．$\psi(\sigma^2)$ を直接外挿するこはできないので，$\sigma^2 = \sigma_0^2$ におけるテイラー展開を k 項で打ち切って，

$$p_k(\mathcal{H}_0|y) = \Phi\left(\sum_{j=0}^{k-1}\frac{(-1-\sigma_0^2)^j}{j!}\frac{\partial^j\psi(x)}{\partial x^j}\Big|_{\sigma_0^2}\right)$$

によって $\sigma^2 = -1$ への外挿をおこなった．曲面のフーリエ変換によって得られた周波数空間において p_k のバイアスを議論することができた．k 回反復ブートストラップ法と p_k は「同等のバイアス」（h を多項式であらわした場合に，その $2k-1$ 次の成分まではバイアスが $O(\Delta h^2)$）であり，k の増加でバイアスは減少する．反復ブートストラップの計算量が $O(B^k)$ であることを考えると，マルチスケール・ブートストラップ法のほうがずっと効率がよい．

このように h がなめらかでない場合，一般に不偏検定が存在するとは限らない．Lehmann (1952) は $\mathcal{H}_0$ が錘の場合に不偏検定が存在しないことを示している．それでも p_k は k の増加とともにバイアス減少の傾向があるが，同時に棄却領域 $\mathcal{R}_0 = \{y : p(\mathcal{H}_0|y) < \alpha\}$ の境界 $\partial\mathcal{R}_0$ が激しく振動して発散してしまう（なめらかな場合に比べて h の高周波成分が大きいことが原因）から，不偏検定が存在しないことと矛盾しない．経験的には $k = 3$ 程度が両者のバランスがとれて良いようだ．

3 領域が３個の場合

Shimodaira (submitted) の議論を紹介する．十分に平坦な h_1, h_2 をつかって，

$$\mathcal{H}_0 = \{(u,v) : -d - h_2(u) \le v \le -h_1(u), \quad u \in \mathbb{R}^m\}$$

と表される場合を考える．$d \to \infty$ とすれば前節の $\mathcal{H}_0$ になる．また $\mathcal{H}_1 : v \geq -h_1(u)$, $\mathcal{H}_2 : v \leq -d - h_2(u)$ としておく．このとき，両側検定に相当する不偏な p-値は

$$p(\mathcal{H}_0|y) = 1 - |p(\mathcal{H}_1|y) - p(\mathcal{H}_2|y)|$$

と書けることが示される．$p(\mathcal{H}_1|y)$ と $p(\mathcal{H}_2|y)$ は領域が 2 個の場合の議論を $\mathcal{H}_1$ vs $\mathcal{H}_0 \cup \mathcal{H}_2$ および $\mathcal{H}_2$ vs $\mathcal{H}_0 \cup \mathcal{H}_1$ へ適用すれば良い．

領域が 2 個の場合は probability matching prior (Tibshirani 1989) を用いて事後確率を $\pi(\mathcal{H}_i|y) = p(\mathcal{H}_i|y)$, $i = 1, 2$ と仮定できる．$i = 1, 2$ に互いに矛盾しない事前分布が得られる保証はないが，Efron and Tibshirani (1998) は形式的に

$$\pi(\mathcal{H}_0|y) = 1 - (p(\mathcal{H}_1|y) + p(\mathcal{H}_2|y))$$

で $\mathcal{H}_0$ の事後確率を計算することを提案している．興味深いのは，

$$p^{(s)}(\mathcal{H}_0|y) = \pi(\mathcal{H}_0|y) + s \min(p(\mathcal{H}_1|y), p(\mathcal{H}_2|y))$$

と書くと，$s = 0$ が事後確率，$s = 1$ が片側検定，$s = 2$ が両側検定に相当する．つまり $\pi(\mathcal{H}_0|y)$ は頻度論的な "zero-sided" test の確率値と解釈することもできる．

4 FDR計算

多数の領域 $\mathcal{H}_i$, $i = 1, \ldots, K$ について不偏な p-値を p_i, $i = 1, \ldots, K$ とする．ここでは 3 節の考え方は用いずに，2 節の方法が適用できるとしておく．$p_i > \alpha$ (たとえば $\alpha = 0.95$) なら $\mu \in \mathcal{H}_i^c$ が棄却されて $\mu \in \mathcal{H}_i$ であるという「発見」をしたと判断される．z-値を $z_i = \Phi^{-1}(p_i)$, $i = 1, \ldots, K$ で定義すると，棄却定数 $c = \Phi^{-1}(\alpha)$ を用いて $z_i > c$ なら「発見」とされる．z_i の値は

$$Z_i|\xi_i \sim N(\xi_i, 1), \quad i = 1, \ldots, K \quad \text{(一般に独立でない)}$$

の確率変数の実現値と見なすことができ，帰無仮説 $\mu \in \mathcal{H}_i^c$ は $\xi_i \leq 0$ に相当することが示せる．ここで，ξ_i, $i = 1, \ldots, K$ が未知の密度関数 $g(\xi)$ に従う確率変数の実現値であったと考えよう．$g(\xi)$ は point mass を含まないと仮定しておく．このとき，

$$\pi_i = 1 - P(\xi_i \leq 0|Z_i > z_i)$$

を計算したい．$P(\xi_i \leq 0|Z_i > c)$ は「発見」が実際には誤りである事後確率であり，FDRの一種である．これまでのFDRに関する文献では，帰無仮説 $\xi_i = 0$ vs 対立仮説 $\xi \neq 0$ の両側検定を扱うものがほとんどであるのに対して，ここで計算するのは帰無仮説 $\xi_i \leq 0$ vs 対立仮説 $\xi > 0$ の片側検定に相当するFDRである．

もし $g(\xi)$ が既知ならばFDR計算は（少なくとも数値的に）可能である．そこで Z_i が正規混合分布

$$Z \sim \int \phi(z-\xi)g(\xi)\,d\xi$$

に従う確率変数であることを利用して $g(\xi)$ を推定するアプローチが考えられる．たとえば $g(\xi)$ が2成分正規混合分布などと仮定してパラメトリックに $g(\xi)$ を推定する方法を実装してみたところ，十分に実用可能であった．しかしここではマルチスケール法を利用した別のアプローチ (Shimodaira in prep) を紹介したい（図1参照）．

まず形式的に確率変数 $Z_i^*, Z_i^{**}, i = 1, \ldots, K$ を導入する．

$$Z_i^*|z_i \sim N(z_i, \sigma^2), \quad Z_i^{**}|z_i^* \sim N(z_i^*, 1)$$

スケール $\sigma^2 > 0$ は2節で用いたものと無関係である．すると，$Z_i^*|\xi_i \sim N(\xi_i, 1+\sigma^2)$ だから，$\sigma^2 = -1$ と形式的におけば (Z_i^*, Z_i^{**}) の同時分布から (ξ_i, Z_i) の同時分布がわかるというアイデアである．ここで (X_1, X_2) が2変量正規分布（平均ゼロ，分散1，相関係数 ρ）に従うときの分布関数を次式で定義しておく．

$$\Phi_\rho(x_1, x_2) = P(X_1 \leq x_1 \wedge X_2 \leq x_2)$$

すると，

$$P(Z_i^* \leq 0 \wedge Z_i^{**} > c|z_i) = \Phi_{-\frac{\sigma}{\sqrt{1+\sigma^2}}}\left(-\frac{z_i}{\sigma}, -\frac{c-z_i}{\sqrt{1+\sigma^2}}\right)$$

$$P(Z_i^{**} > c|z_i) = \Phi\left(-\frac{c-z_i}{\sqrt{1+\sigma^2}}\right)$$

は容易に数値計算できる．これを z_i の重み関数として $z_1, \ldots, z_K$ の平均を計算すれば，$P(Z_i^* \leq 0 \wedge Z_i^{**} > c)$ と $P(Z_i^{**} > c)$ の推定値が得られるから，

$$\widehat{\mathrm{FDR}}_{\sigma^2} = \frac{\sum_{i=1}^K P(Z_i^* \leq 0 \wedge Z_i^{**} > c|z_i)}{\sum_{i=1}^K P(Z_i^{**} > c|z_i)}$$

は $\mathrm{FDR}_{\sigma^2} := P(Z_i^* \leq 0|Z_i^{**} > c)$ の推定値になる．これをいくつかの $\sigma^2 > 0$ の値で計算して $\sigma^2 = -1$ へ外挿したものがFDRであることを示せる．すなわち

$$P(\xi_i \leq 0|Z_i > c) = \lim_{\sigma^2 \to -1} \mathrm{FDR}_{\sigma^2}$$

である．このアプローチでは $g(\xi)$ を直接推定することなくFDRが計算される．

最後にブートストラップ確率を直接用いるアプローチについて述べておく．本節で議論してきたFDRの計算法は領域の問題に限らずモデル $Z_i|\xi \sim N(\xi_i, 1)$, $i = 1, \ldots, K$ を前提にすれば一般に適用可能な方法であった．ところが1節の正規モデルを前提にした領域の問題に限って言えば，ブートストラップ確率は仮説の事後確率（事前分布はmatching priorではなく一様分布）と解釈できて，領域 $\mathcal{H}_i$ のブートストラップ確率 α_i からFDRを近似計算するほうが簡便であろう．添え字の集合 $A(\alpha) = \{i \in 1, \ldots, K : p_i \geq \alpha\}$ と定義すると，1-FDRは，$\hat{\pi}_i = \sum_{j \in A(p_i)} \alpha_j / \sum_{j \in A(p_i)} 1$ である．ここでは p_i として不偏な p-値を用いる必要もなく，$p_i = \alpha_i$, $i = 1, \ldots, K$ とするのがベイズ判別の意味で最適になる．

5 参考文献

J. R. Cook and L. A. Stefanski (1994) "Simulation-extrapolation estimation in parametric measurement error models." *J. Amer. Statist. Assoc.* **89** 1314–1328.

B. Efron, R. Tibshirani (1998) "The problem of regions." *Ann. Statist.* **26** 1687–1718

E. L. Lehmann (1952) "Testing multiparameter hypotheses." *Ann. Math. Statistics* **23** 541–552.

H. Shimodaira (2008) "Testing Regions with Nonsmooth Boundaries via Multiscale Bootstrap." *Journal of Statistical Planning and Inference* **138** 1227-1241. http://dx.doi.org/10.1016/j.jspi.2007.04.001

H. Shimodaira (submitted) "Frequentist and Bayesian measures of confidence via multiscale bootstrap for testing three regions."

H. Shimodaira (in prep) マルチスケール法によるFDR計算 （タイトル未定）

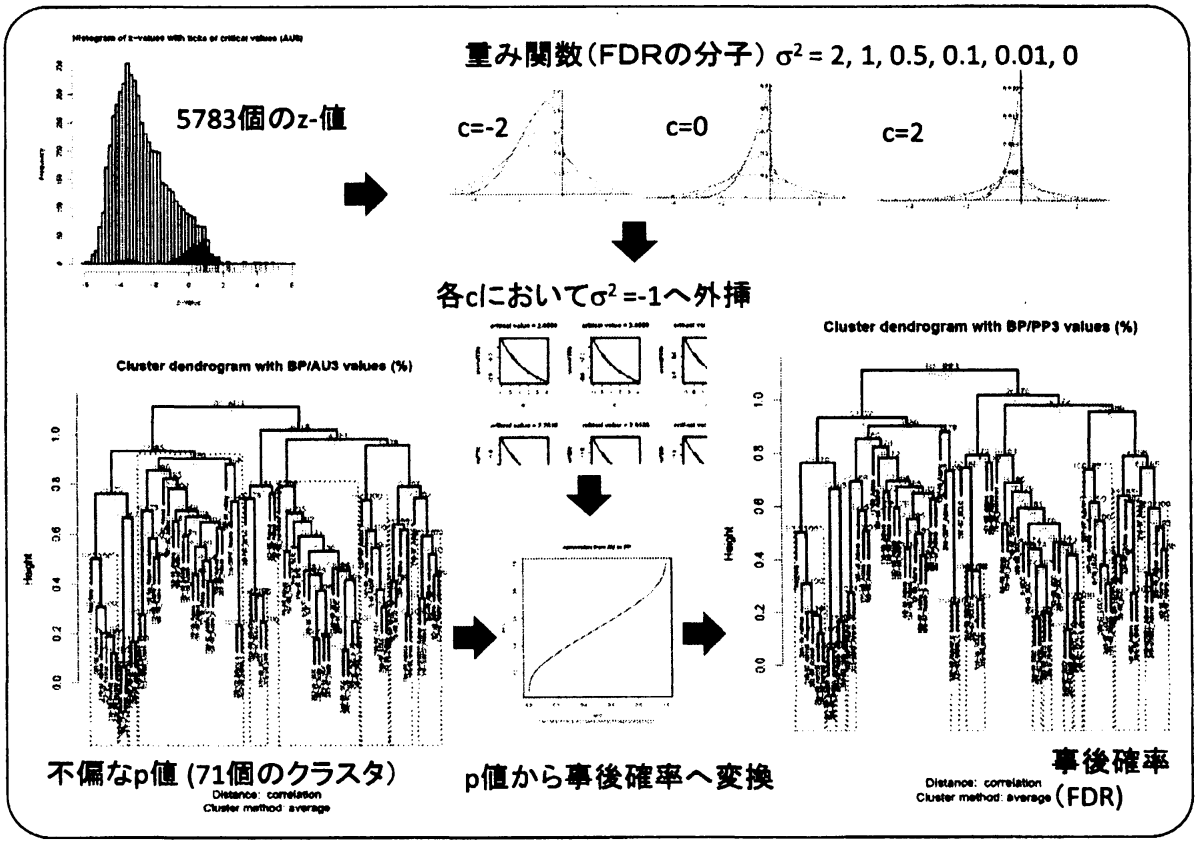

図 1: FDR 計算の概略．左下の木の赤い枠は $p_i > 0.95$ となるクラスタ，右下の木の赤い枠は $\pi_i > 0.95$ となるクラスタを示す．木に示されているのは 71 個の p_i または π_i であるが，ブートストラップ法で出現した 5783 個のクラスタに対して p_i が計算されている．