

マルチスケールブートストラップ法による 3 次の漸近精度をもつ不偏な検定

統計数理研究所 下平英寿

1 はじめに

データ $\mathcal{X}_n = \{x_1, \dots, x_n\}$ の関数 $H_0(\mathcal{X}_n)$ で 0 か 1 を値に取るものを考える．(例 1) $\bar{x} = (x_1 + \dots + x_n)/n \leq 0$ の時 $H_0(\mathcal{X}_n) = 1$, $\bar{x} > 0$ の時 $H_0(\mathcal{X}_n) = 0$ ．(例 2) 哺乳類の DNA 配列データ $\mathcal{X}_n$ をクラスタ分析して得られた系統樹がウサギとマウスの単系統を支持するとき $H_0(\mathcal{X}_n) = 1$, そうでない時 $H_0(\mathcal{X}_n) = 0$ ．我々が本当に知りたいのは $n \rightarrow \infty$ の極限で $H_0(\mathcal{X}_n)$ の取る値 $H_0(\mathcal{X}_\infty)$ とする．ところがデータ $\mathcal{X}_n$ にはバラツキがあるから $H_0(\mathcal{X}_n)$ の値もバラツキがある．仮に $H_0(\mathcal{X}_n) = 0$ という観測値が得られたからといって必ずしも $H_0(\mathcal{X}_\infty) = 0$ とは限らない．そこで, $H_0(\mathcal{X}_\infty) = 1$ を帰無仮説, $H_0(\mathcal{X}_\infty) = 0$ を対立仮説とした検定を考え, 帰無仮説が棄却されたとき $H_0(\mathcal{X}_\infty) = 0$ であったと判断する．検定の確率値 $p(\mathcal{X}_n)$ は 0 から 1 の間の実数を値に取り,

$$\Pr\{p(\mathcal{X}_n) < \alpha\} \leq \alpha, \quad H_0(\mathcal{X}_\infty) = 1 \quad (1)$$

及び

$$\Pr\{p(\mathcal{X}_n) < \alpha\} \geq \alpha, \quad H_0(\mathcal{X}_\infty) = 0 \quad (2)$$

を少なくとも近似的に満たすものを考える．

現実のデータ解析では $H_0(\mathcal{X}_n)$ は手続き (計算機プログラム) として与えられていて解析的に表現することが困難な場合が多いが, それでも直ちに適用可能な確率値の計算法を以下で与える．この方法の特徴の一つは関数 $H_0(\mathcal{X}_n)$ と $\mathcal{X}_n$ のリサンプリングを通してのみデータにアクセスすることである．素朴なブートストラップ法やダブルブートストラップ法も同様な特徴をもつが, 本方法は素朴なブートストラップ法と同じオーダーの計算量にもかかわらずダブルブートストラップ法と同じオーダーの精度が得られる．

2 マルチスケールブートストラップ法

サイズ n のデータ $\mathcal{X}_n = \{x_1, \dots, x_n\}$ から復元抽出で n_1 個取り出した複製データを $\mathcal{X}_{n_1}^* = \{x_1^*, \dots, x_{n_1}^*\}$ と書く．普通のブートストラップ法では $n_1 = n$ とするが, ここでは n_1 を自由に変える．さらに $\mathcal{X}_{n_1}^*$ から復元抽出で n_2 個取り出した複製データを $\mathcal{X}_{n_2}^{**} = \{x_1^{**}, \dots, x_{n_2}^{**}\}$ と書く．続けて $\mathcal{X}_{n_2}^{**}$ から復元抽出で n_3 個取り出した複製データを $\mathcal{X}_{n_3}^{***} = \{x_1^{***}, \dots, x_{n_3}^{***}\}$ と書く． (n_1, n_2, n_3) の組み合わせを K 通り与え $(n_1^{(k)}, n_2^{(k)}, n_3^{(k)})$, $k = 1, \dots, K$ と書く．それぞれの組み合わせについて $(\mathcal{X}_{n_1}^*, \mathcal{X}_{n_2}^{**}, \mathcal{X}_{n_3}^{***})$ を $B^{(k)}$ 個生成する．このうち $H_0(\mathcal{X}_{n_1}^*) = 1$ となった回数を $C_1^{(k)}$, $H_0(\mathcal{X}_{n_2}^{**}) = 1$ となった回数を $C_2^{(k)}$, $H_0(\mathcal{X}_{n_3}^{***}) = 1$ となった回数を $C_3^{(k)}$ とする．ここで計算した複製の総数と $H_0(\cdot)$ の評価回数は $3 \sum_{k=1}^K B^{(k)}$ である．

復元抽出のバラツキにより各 $C_3^{(k)}$ は 2 項分布に従うが, $B^{(k)} \rightarrow \infty$ の時の $C_3^{(k)}/B^{(k)} \rightarrow \pi_3(n/n_1^{(k)}, n/n_2^{(k)}, n/n_3^{(k)})$ と書くことにする．同様に, $\pi_1(n/n_1^{(k)})$, $\pi_2(n/n_1^{(k)}, n/n_2^{(k)})$ を定義する．これらの関数 $\pi_i(\cdot)$ と確率値の関係が以下で与えられる．

3 近似的に不偏な検定

適当な未知の非線形変換 $y = f(\mathcal{X}_n)$ が存在して, r 次元ベクトル $y = (y_1, \dots, y_r)$ の確率分布が指数型分布族 $y \sim \exp(\theta^i y_i - \psi(\theta) - h(y))$ であると仮定する. 期待値パラメタ $\eta_i = \partial\psi/\partial\theta^i$ を用いて $H_0(\mathcal{X}_\infty) = 1$ となる領域が $H_0 = \{\eta = (\eta_1, \dots, \eta_r) \mid \eta_r \leq -d^{ab}\eta_a\eta_b - e^{abc}\eta_a\eta_b\eta_c\}$ と書けると仮定する. ただし添え字 a, b, c は $1, \dots, r-1$ を動く. ここでポテンシャル $\phi(\eta) = \max_\theta\{\theta^i\eta_i - \psi(\theta)\}$ が $\partial\phi/\partial\eta_i|_0 = 0$, $\partial^2\phi/\partial\eta_i\partial\eta_j|_0 = \delta_{ij}$ を満たし, データ $y = (0, \dots, 0, \lambda)$ を観測したと仮定しても一般性を失わない. このとき $O(n^{-3/2})$ の誤差で以下が成り立つ.

$$\begin{aligned} \pi_3(\tau_1, \tau_2, \tau_3) &= 1 - \Phi\left\{\gamma_1 T_0 (1 + \gamma_3 T_1 + 4\gamma_3^2 T_1^2 + \gamma_5 T_2 + \gamma_6 T_3) \right. \\ &\quad \left. - (\gamma_1 T_0)^{-1} (\gamma_2 + \gamma_3 T_1 + 7\gamma_3^2 T_1^2 + \gamma_4 T_1 + 3\gamma_5 T_2 + 3\gamma_6 T_3) \right\}. \end{aligned} \quad (3)$$

ただし $T_0 = (\tau_1 + \tau_2 + \tau_3)^{-1/2}$, $T_1 = (\tau_1\tau_2 + \tau_2\tau_3 + \tau_3\tau_1)T_0^4$, $T_2 = (\tau_1\tau_2\tau_3 + \tau_2^2\tau_3 + \tau_1^2(\tau_2 + \tau_3))T_0^6$, $T_3 = (\tau_1\tau_2\tau_3)T_0^6$. 係数 $\gamma_1, \dots, \gamma_6$ は $\phi^{ijk} = \partial^3\phi/\partial\eta_i\partial\eta_j\partial\eta_k|_0$, $\phi^{ijkl} = \partial^4\phi/\partial\eta_i\partial\eta_j\partial\eta_k\partial\eta_l|_0$ として

$$\begin{aligned} \gamma_1 &= \lambda + \frac{1}{3}\lambda^2\phi^{rrr} + \lambda^3\left\{-\frac{1}{8}\phi^{arr}\phi^{arr} - \frac{1}{18}\phi^{rrr}\phi^{rrr} + \frac{1}{8}\phi^{rrrr}\right\} \\ \gamma_2 &= \lambda\left\{-d^{aa} - \frac{1}{6}\phi^{rrr}\right\} + \lambda^2\left\{d^{ab}d^{ab} - \frac{1}{2}d^{aa}\phi^{rrr} + \frac{1}{8}\phi^{arr}\phi^{arr} + \frac{1}{72}\phi^{rrr}\phi^{rrr} - \frac{1}{24}\phi^{rrrr}\right\} \\ \gamma_3 &= -\frac{1}{6}\lambda\phi^{rrr} + \lambda^2\left\{\frac{1}{4}\phi^{arr}\phi^{arr} + \frac{1}{9}\phi^{rrr}\phi^{rrr} - \frac{1}{8}\phi^{rrrr}\right\} \\ \gamma_4 &= \lambda^2\left\{-d^{ab}\phi^{abr} + \frac{1}{3}d^{aa}\phi^{rrr} + \frac{1}{2}\phi^{abr}\phi^{abr} + \frac{1}{2}\phi^{arr}\phi^{arr} + \frac{2}{9}\phi^{rrr}\phi^{rrr} - \frac{1}{4}\phi^{aarr} - \frac{1}{6}\phi^{rrrr}\right\} \\ \gamma_5 &= \lambda^2\left\{-\frac{1}{8}\phi^{arr}\phi^{arr} - \frac{1}{8}\phi^{rrr}\phi^{rrr} + \frac{1}{12}\phi^{rrrr}\right\} \\ \gamma_6 &= \lambda^2\left\{-\frac{1}{8}\phi^{arr}\phi^{arr} - \frac{1}{8}\phi^{rrr}\phi^{rrr} + \frac{1}{24}\phi^{rrrr}\right\}. \end{aligned}$$

一般に $\lambda = O(1)$, $d^{ab} = O(n^{-1/2})$, $\phi^{ijk} = O(n^{-1/2})$, $\phi^{ijkl} = O(n^{-1})$ であるから, 係数は $\gamma_1 = O(1)$, γ_2, γ_3 は $O(n^{-1/2})$, $\gamma_4, \gamma_5, \gamma_6$ は $O(n^{-1})$ である. マルチスケールブートストラップで求めた $C_3^{(k)}$, $k = 1, \dots, K$ に $\pi_3(\tau_1, \tau_2, \tau_3)$ の当てはめを行い, 係数 $\gamma_1, \dots, \gamma_6$ を推定することができる. そして

$$p(y) = 1 - \Phi\left\{\gamma_1 (1 + \gamma_3 + 4\gamma_3^2 + \gamma_6) + \gamma_1^{-1} (\gamma_2 + \gamma_3^2/2 + \gamma_4 + \gamma_5) \right\} \quad (4)$$

で与えられる確率値は (1) と (2) を $O(n^{-3/2})$ の誤差で満たす. なお $\pi_1(\tau_1) = \pi_3(\tau_1, 0, 0) = 1 - \Phi(\gamma_1/\sqrt{\tau_1} - (\gamma_2/\gamma_1)\sqrt{\tau_1})$, $\pi_2(\tau_1, \tau_2) = \pi_3(\tau_1, \tau_2, 0)$ から部分的な情報が得られる. 例えば $C_2^{(k)}$ を用いて $\gamma_1, \gamma_2, \gamma_3$ が推定できるので $O(n^{-1})$ の誤差の確率値が得られる. 特に y が正規なら $\gamma_3 = \dots = \gamma_6 = 0$ なので $C_1^{(k)}$ だけで $O(n^{-3/2})$ の誤差の確率値が得られる. この正規の場合では, 文献 [1] と [2] に $\pi_1(\tau_1)$ と $p(y)$ に結びつく結果があり, [3] でマルチスケールブートストラップを提案し, [4] で計算機プログラムを一般公開した.

参考文献 [1] EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72**, 45–58. [2] EFRON, B. AND TIBSHIRANI, R. (1998). The problem of regions. *Ann. Statist.* **26**, 1687–1718. [3] SHIMODAIRA, H. (2000). Another calculation of the p -value for the problem of regions using the scaled bootstrap resamplings. Technical Report No. 2000-35, Stanford University. [4] SHIMODAIRA, H. AND HASEGAWA, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17** in press.