日本数学会
２００４年度秋季総合分科会
２００４年9月１９日～9月２２日
北海道大学（札幌）

統計数学分科会
特別講演

ブートストラップ法の幾何学とスケール変換

下平英寿 （東工大・情報理工）

講演予稿
１２ページ

shimo@is.titech.ac.jp
http://www.is.titech.ac.jp/~shimo/

ブートストラップ法の幾何学と スケール変換

下平英寿 （東工大・情報理工）

1 はじめに

近年ゲノム科学など様々な分野で膨大なデータが急速に蓄積されている．これから知識発見を行うために，データマイニングなどのデータ解析手法によって非常に多くの仮説が同時に探索されることがある．このような状況では，データに内在する確率的な揺らぎ，すなわちバラツキの影響が増幅されて誤った発見に導かれやすくなる．この影響を正しく評価することが重要である．

リサンプリング法はデータからのサンプリングによってバラツキを評価する一般的な手法であり，一種の確率シミュレーション技法である．代表的なリサンプリング法であるブートストラップ法 (Efron 1979) が発表されて 20 年以上が経ち，様々なデータ解析に利用されているが，その精度が必ずしも十分でないことが分かってきた．より精度の高い方法が必要である．

現実のデータ解析は手続きが複雑で出力も多様である．そこで広範な応用を可能にするために，興味のある仮説を支持したか否かという二者択一の結果だけを計算に利用する方法を考える．ブートストラップ法によって生成した多数の複製データが仮説を支持した頻度はブートストラップ確率と呼ばれ，バイオインフォマティクスの応用で以前から利用されている (Felsenstein 1985)．ブートストラップ確率は，仮説検定の確率値として解釈できる．一般にサンプルサイズ n の増加に対して検定のバイアスが $n^{-k/2}$ に比例して小さくなるとき，k 次の精度の近似的に不偏な検定という．ブートストラップ確率の精度は 1 次であり，十分とはいえない．

データの背後にある確率モデルや興味のある仮説を反映して距離や曲率といった幾何学的な量が定義できる．この幾何学的な量を用いてブートストラップ確率を表現することが可能であり，さらに

不偏な確率値を表現することも可能である (Efron and Tibshirani 1998)．複製データのサンプルサイズを n から n' に変化させると，バラツキのスケールが $\tau = \sqrt{n/n'}$ 倍されるが，このときのブートストラップ確率の変化は τ の多項式で表現できて，その係数が幾何学的な量に対応する (Shimodaira 2002)．この事実を利用したのがマルチスケール・ブートストラップ法であり，ブートストラップ確率の変化率から 3 次の精度の確率値が計算できる．この新しい原理によって，p^*-formula (Barndorff-Nielsen 1986) やダブルブートストラップ法 (Hall 1992) と数学的に等価な確率値を容易に計算することが可能になった (Shimodaira, in press)．

データの確率モデルは，なんらかの非線形変換によって指数型分布族として表現できるものと仮定する．しかしその変換を実際に知る必要はない．つまり確率モデルや仮説の曲面を解析的に与える必要はない．手法の導出には Weyl のチューブ公式と同様の座標系を用いる．Edgeworth 漸近展開の計算を 3 次の精度で行うのは煩雑であり，数式処理ソフトウエア (*Mathematica* および *MathTensor*) を用いて証明を行った (Shimodaira 2004)．

以下ではまず 4 節まで用語の定義など行い，5 節から 8 節までは例を用いながら提案手法を説明する．9 節から 11 節はテクニカルな側面を説明する．

2　ブートストラップ法

データ $x_1, \ldots, x_n$ の各要素は独立同分布に従う確率変数の実現値とする．その期待値 μ の推定値は平均 $\bar{x} = (x_1 + \cdots + x_n)/n$ によって与えられる．$\bar{x}$ のバラツキを調べるために，データと同じ確率分布に従う確率変数の実現値を擬似乱数を使って生成したものを，複製データ $x_1^*, \ldots, x_n^*$ とする．ただし μ は未知なので，μ の役割を $\bar{x}$ で置き換えて生成する．この手続きを多数回繰り返し実行して，$\bar{x}^* = (x_1^* + \cdots + x_n^*)/n$ のバラツキを観測することにより，$\bar{x}$ のバラツキが推測できる．

ここではデータの確率モデルを仮定してそれを利用しているの

で，この手法はパラメトリック･ブートストラップ法と呼ばれる．一方，通常のブートストラップ法はモデルを仮定せずに，$x_1, \ldots, x_n$ からランダムに重複を許して n 個の要素を取り出して $x_1^*, \ldots, x_n^*$ を生成する．以降の議論はすべてパラメトリック･ブートストラップ法を用いるが，これは現実のデータ解析に通常のブートストラップ法を適用する一連の手続きを，簡略にモデル化していると考える．

3　確率モデル

データに何らかの変換を適用したものを p 次元ベクトル y として表現する．y は確率変数 Y の実現値であり，Y の期待値を η と書く．$Y \sim f(y;\eta)$ は η を未知パラメタとする指数型分布族であり，特に，$y = \sqrt{n}\bar{x}$，$\eta = \sqrt{n}\mu$ の形に書けると仮定する．$n \to \infty$ とする漸近的な議論に関して，ここでは $\eta = O(1)$ とするので，$\mu = O(n^{-1/2})$ のいわゆる local alternatives に相当する．

ブートストラップ法は $Y^* \sim f(y^*;y)$ によって Y^* の実現値 y^* を多数生成することに相当する．もし複製データの要素を n' 個に変更すると，$\bar{x}^* = (x_1^* + \cdots + x_{n'}^*)/n'$ のバラツキのスケールは τ 倍される．したがって，$f(y;\eta)$ のポテンシャル関数 $\phi(\eta)$ を $\phi(\eta,\tau) = \phi(\eta)/\tau^2$ で置き換えて密度関数 $f(y;\eta,\tau)$ を定義すると，$Y^* \sim f(y^*;y,\tau)$ となる．

例 (正規モデル). Y が p 次元正規分布に従うとき，$f(y;\eta,\tau)$ は $N_p(\eta,\tau^2 I_p)$，ポテンシャル関数は $\phi(\eta,\tau) = \|\eta\|^2/(2\tau^2)$ である．

4　不偏検定

パラメタ空間の正の体積を持つ領域 $\mathcal{R}$ は滑らかな境界 $\partial\mathcal{R}$ を持つと仮定し，興味のある仮説を「$\eta \in \mathcal{R}$」とする．仮説の領域を μ で表現したときのサイズを $O(1)$ とするので，η で表現すると $O(n^{1/2})$ で大きくなる．

例 (球体). $\mathcal{R} = \{\eta : \|\eta\| \le n^{1/2}\}$，球面 $\partial\mathcal{R} = \{\eta : \|\eta\| = n^{1/2}\}$.

スケールτのブートストラップ確率は

$$\tilde{\alpha}_1(y,\tau) = \Pr\{Y^* \in \mathcal{R}; y, \tau\}$$

である．通常は$\tau = 1$なので，これを特に$\hat{\alpha}_0(y) = \tilde{\alpha}_1(y,1)$と書く．$y$が$\mathcal{R}$から遠く離れるほど$\hat{\alpha}_0(y)$が0に近づき，仮説が正しくないことを示唆する．あらかじめ決めた有意水準$0 < \alpha < 1$に対して$\hat{\alpha}_0(y) < \alpha$ならば，仮説を棄却する．

仮説を検定するための一般の確率値を$\hat{\alpha}(y)$と書く．棄却確率は$\beta(\alpha,\eta) = \Pr\{\hat{\alpha}(Y) < \alpha; \eta\}$である．仮説検定では任意の$\eta \in \mathcal{R}$に対して$\beta(\alpha,\eta) \leq \alpha$を要求する．さらに任意の$\eta \notin \mathcal{R}$に対して$\beta(\alpha,\eta) \geq \alpha$ならば，この検定は不偏であるという．$\beta(\alpha,\eta)$は$\eta$に関して連続であるから，不偏ならば

$$\beta(\alpha,\eta) = \alpha, \quad \eta \in \partial\mathcal{R}$$

である．この等式が厳密には成り立たない場合，その誤差がバイアスであり，これを小さくする確率値の計算法を与えたい．バイアスが$O(n^{-k/2})$であれば，k次の精度の近似的に不偏な検定であり，kが大きいほど望ましいと考える．$\hat{\alpha}(y)$として$\hat{\alpha}_0(y)$を用いると$k = 1$であるが，以下に述べる方法では$k = 3$になり，より精度の高い方法といえる．

5　マルチスケール・ブートストラップ法

数値例1. 正規モデル$Y \sim N_p(\eta, \tau^2 I_p)$．球体$\mathcal{R} = \{\eta : \|\eta\|^2 \leq n\}$．ただし$p = 4, n = 10$．観測データを$\|y\|^2 = 26.80$とすると，$y \notin \mathcal{R}$である．$\|y\| - \sqrt{n} = 5.177 - 3.162 = 2.015$が仮説$\eta \in \mathcal{R}$を棄却するほど十分に大きいといえるか？

このとき$n_1 = 3, 6, 10, 15, 21$とサンプルサイズを変化させるとスケール$\tau_1 = \sqrt{n/n_1}$のブートストラップ確率は

$$\tilde{\alpha}_1(y,\tau_1) = 0.0359,\ 0.0205,\ 0.0085,\ 0.0028,\ 0.0008$$

と変化する．特に$\hat{\alpha}_0(y) = 0.0085$は1%以下であり，仮説を棄却する十分な証拠があると判断する．

この簡単な例では$\|Y\|^2$が非心カイ二乗分布に従うことを利

用して厳密に不偏な確率値が計算できる．この値は $\hat{\alpha}_\infty(y) = \Pr\{\|Y\| \geq \|y\|; \eta \in \partial\mathcal{R}\} = 0.05$ であり，先ほどの $\hat{\alpha}_0(y)$ の値は厳密値から大きく異なっていたことになる．

一般に α 値から z 値 $= -\Phi^{-1}(\alpha)$ を定義する．ただし $\Phi^{-1}(\cdot)$ は標準正規分布関数の逆関数である．11節の議論によって，$\tilde{z}_1(y,\tau_1) = -\Phi^{-1}(\tilde{\alpha}_1(y,\tau_1))$ と τ_1 の関係は次式で表現されることが示される．

$$\tilde{z}_1(y,\tau_1) \approx \hat{v}/\tau_1 + \hat{c}\tau_1$$

ただし $\approx$ は両辺の誤差が $O(n^{-3/2})$ であることを表す．係数は幾何学的な量を反映しており，$\hat{v}$ は y と $\partial\mathcal{R}$ の距離を表し，$\hat{c}$ は $\partial\mathcal{R}$ の曲率を表す．この理論式を実際に得られたブートストラップ確率に当てはめて，$\hat{v}$ と $\hat{c}$ を回帰係数として推定すると，$\hat{v} = 2.002, \hat{c} = 0.385$ が得られる．

実は不偏な確率値の z 値 $\hat{z}_\infty(y) = -\Phi^{-1}(\hat{\alpha}_\infty(y))$ は

$$\hat{z}_\infty(y) \approx \hat{v} - \hat{c}$$

を満たすことが示せるので，$\hat{z}_1(y) = \hat{v} - \hat{c}$，$\hat{\alpha}_1(y) = \Phi(-\hat{z}_1(y))$ と定義して，先ほど推定した $\hat{v}, \hat{c}$ を代入すると，

$$\hat{\alpha}_1(y) = \Phi(-2.002 + 0.385) = 0.0529$$

が得られる．以上のようにスケールを変化させた複数個のブートストラップ確率から高精度の確率値を計算する手続きをマルチスケール・ブートストラップ法という (Shimodaira 2002)．正規モデルを仮定すると，任意の $\mathcal{R}$ に対して $\hat{\alpha}_1(y)$ は3次の精度になる．

$\tilde{z}_1(y,\tau)$ を縦軸，$1/\tau$ を横軸にプロットすると，次式に示されるように，$\tau = 1$ における曲線の傾きが $\hat{z}_1(y)$ になる．

$$\left.\frac{\partial\tilde{z}_1(y,\tau)}{\partial(1/\tau)}\right|_{\tau=1} \approx \hat{z}_1(y)$$

したがって，この方法はスケールを変化させたときのブートストラップ確率の変化率から高精度確率値を計算しているといえる．

6 加速定数

数値例 2. 各 X_i が期待値 μ の指数分布に従うとき，$Y = \sqrt{n}\bar{X}$

はガンマ分布 ($p=1$，指数 n) に従い，期待値 $\eta=\sqrt{n}\mu$ である．仮説は $\eta\le\sqrt{n}$ とする．$n=10$，$y=4.968$ とすれば，$y-\sqrt{n}=1.806>0$ であり $y\notin\mathcal{R}$ となる．実はちょうど $\hat{\alpha}_\infty(y)=0.05$ となるように値を選んである．

数値例 1 と同様に計算を行うと，

$$\tilde{\alpha}_1(y,\tau_1)=0.2990,\ 0.1875,\ 0.1115,\ 0.0622,\ 0.0322$$

から $\hat{v}=1.328, \hat{c}=-0.110$ が推定される．確率値は

$$\hat{\alpha}_1(y)=\Phi(-1.328-0.110)=0.0753$$

である．$\hat{\alpha}_0(y)=0.1115$ に比べれば $\hat{\alpha}_\infty(y)=0.05$ に近いものの，必ずしも精度が高くないことが分かる．

前節では正規モデルを仮定していたが，一般の指数型分布族では

$$\tilde{z}_1(y,\tau_1)\doteq\frac{\hat{v}-2\hat{a}\hat{v}^2}{\tau_1}+(\hat{c}-\hat{a})\tau_1,$$

と書ける．ただし $\doteq$ は両辺の誤差が $O(n^{-1})$ であることを表す．$\hat{a}$ は加速定数と呼ばれ，確率モデルの性質と $\partial\mathcal{R}$ から定まる量である．正規モデルでは Y の分散行列が単位行列に固定されていたことが本質的であり $\hat{a}=0$ であった．一般の指数型分布族では，$\partial\mathcal{R}$ の法ベクトル方向に関して，η を動かしたときの分散の変化率が $\hat{a}$ に反映されている．

Efron (1987) の ABC 法によれば，3 個の幾何学的な量 $\hat{v},\hat{c},\hat{a}$ を用いて不偏な確率値は

$$\hat{z}_\infty(y)\doteq\hat{v}-\hat{c}+\hat{a}(1-\hat{v}^2)$$

と表現される．一方，前節の方法で計算した確率値は

$$\hat{z}_1(y)\doteq\hat{v}-\hat{c}+\hat{a}(1-2\hat{v}^2)$$

であり，$\hat{a}\hat{v}^2=O(n^{-1/2})$ だけ誤差がある．これを補正するために，$\hat{a}$ を推定する必要がある．

7　2ステップ＝マルチスケール法

各y^*からスケールτ_2のブートストラップ法によって複製y^{**}を一個だけ生成する．つまり$Y^* \sim f(y^*; y, \tau_1)$と$Y^{**} \sim f(y^{**}; y^*, \tau_2)$によって，$(y^*, y^{**})$を多数生成し，スケール$(\tau_1, \tau_2)$のブートストラップ確率

$$\tilde{\alpha}_2(y, \tau_1, \tau_2) = \Pr\{Y^{**} \in \mathcal{R}; y, \tau_1, \tau_2\}$$

を計算する．正規モデルでは

$$\tilde{\alpha}_2(y, \tau_1, \tau_2) = \tilde{\alpha}_1(y, \sqrt{\tau_1^2 + \tau_2^2})$$

となるが，一般の指数型分布族では，

$$\tilde{z}_2(y, \tau_1, \tau_2) - \tilde{z}_1(y, \sqrt{\tau_1^2 + \tau_2^2}) \doteq \frac{\hat{a}\tau_1^2\tau_2^2(\hat{v}^2 - (\tau_1^2 + \tau_2^2))}{(\tau_1^2 + \tau_2^2)^{5/2}}$$

であることを利用して$\hat{a}$が推定できる．数値例2において$n_2 = 6, 15$, $\tau_2 = \sqrt{n/n_2}$とすると，たとえば，

$$\tilde{\alpha}_2(y, \sqrt{\tfrac{10}{6}}, \sqrt{\tfrac{10}{6}}) = 0.3063 \neq \tilde{\alpha}_1(y, \sqrt{\tfrac{10}{3}}) = 0.2990$$

$$\tilde{\alpha}_2(y, \sqrt{\tfrac{10}{10}}, \sqrt{\tfrac{10}{15}}) = 0.1866 \neq \tilde{\alpha}_1(y, \sqrt{\tfrac{10}{6}}) = 0.1875$$

のように，極わずかの差があることが分かる．

$\hat{z}_1(y)$と$\tilde{z}_2(y, \tau_1, \tau_2) - \tilde{z}_1(y, \sqrt{\tau_1^2 + \tau_2^2})$の式をまとめると，

$$\tilde{z}_2(y, \tau_1, \tau_2) \doteq s_1\hat{\gamma}_1(1 + s_2\hat{\gamma}_3) - \frac{\hat{\gamma}_2 + s_2\hat{\gamma}_3}{s_1\hat{\gamma}_1}$$

ただし$s_1 = (\tau_1^2 + \tau_2^2)^{-1/2}$, $s_2 = \tau_1^2\tau_2^2 s_1^4$である．ここでは$\hat{v}, \hat{c}, \hat{a}$の代わりに$\hat{\gamma}_1 \doteq \hat{v} - 2\hat{a}\hat{v}^2$, $\hat{\gamma}_2 \doteq \hat{v}(\hat{a} - \hat{c})$, $\hat{\gamma}_3 \doteq \hat{v}\hat{a}$を用いている．この理論式を数値例2に当てはめると回帰係数として

$$\hat{\gamma}_1 = 1.328, \quad \hat{\gamma}_2 = 0.144, \quad \hat{\gamma}_3 = 0.137$$

と推定される．一方，不偏な確率値は

$$\hat{z}_\infty(y) \doteq \hat{\gamma}_1(1 + \hat{\gamma}_3) + \frac{\hat{\gamma}_2}{\hat{\gamma}_1}$$

と書けるので，右辺を$\hat{z}_2(y)$と置いて，推定した係数を代入すると，

$$\hat{\alpha}_2(y) = 1 - \Phi\left\{1.328(1 + 0.137) + \frac{0.144}{1.328}\right\} = 0.0528$$

が得られる．$\hat{\alpha}_0(y) = 0.1115$ や $\hat{\alpha}_1(y) = 0.0753$ に比べて，$\hat{\alpha}_\infty(y) = 0.05$ に近く，精度が高いことが分かる．任意の指数型分布族，任意の $\mathcal{R}$ に対して，この 2 ステップ＝マルチスケール法は 2 次の精度である．

8　3 ステップ＝マルチスケール法

各 y^{**} からスケール τ_3 のブートストラップ法によって複製 y^{***} を一個だけ生成し，スケール (τ_1, τ_2, τ_3) のブートストラップ確率

$$\tilde{\alpha}_3(y, \tau_1, \tau_2, \tau_3) = \Pr\{Y^{***} \in \mathcal{R}; y, \tau_1, \tau_2, \tau_3\}$$

を計算する．前節と同様の議論を繰り返すと次式を得る．

$$\begin{aligned}\tilde{z}_3(y, \tau_1,\tau_2, \tau_3) \approx\ & \hat{\gamma}_1 s_1 \left(1 + \hat{\gamma}_3 s_2 + 4\hat{\gamma}_3^2 s_2^2 + \hat{\gamma}_5 s_3 + \hat{\gamma}_6 s_4\right) \\ & - (\hat{\gamma}_1 s_1)^{-1} \left(\hat{\gamma}_2 + \hat{\gamma}_3 s_2 + 7\hat{\gamma}_3^2 s_2^2 + \hat{\gamma}_4 s_2 + 3\hat{\gamma}_5 s_3 + 3\hat{\gamma}_6 s_4\right)\end{aligned}$$

ただし $s_1 = (\tau_1^2 + \tau_2^2 + \tau_3^2)^{-1/2}$, $s_2 = (\tau_1^2\tau_2^2 + \tau_2^2\tau_3^2 + \tau_3^2\tau_1^2)s_1^4$, $s_3 = (\tau_1^2\tau_2^2\tau_3^2 + \tau_2^4\tau_3^2 + \tau_1^4(\tau_2^2 + \tau_3^2))s_1^6$, $s_4 = (\tau_1^2\tau_2^2\tau_3^2)s_1^6$ である．

これから係数 $\hat{\gamma}_1, \ldots, \hat{\gamma}_6$ を推定し，

$$\hat{z}_3(y) = \hat{\gamma}_1 \left(1 + \hat{\gamma}_3 + 4\hat{\gamma}_3^2 + \hat{\gamma}_6\right) + \hat{\gamma}_1^{-1} \left(\hat{\gamma}_2 + \hat{\gamma}_3^2/2 + \hat{\gamma}_4 + \hat{\gamma}_5\right)$$

に代入して $\hat{\alpha}_3(y)$ を計算すると，任意の指数型分布族，任意の $\mathcal{R}$ に対して，この 3 ステップ＝マルチスケール法は 3 次の精度である．

数値例 2 に適用すると，$\hat{\gamma}_1 = 1.328, \hat{\gamma}_2 = 0.145, \hat{\gamma}_3 = 0.127, \hat{\gamma}_4 = -0.018, \hat{\gamma}_5 = -0.0004, \hat{\gamma}_6 = -0.036$ となり，

$$\begin{aligned}\hat{\alpha}_3(y) = 1 - \Phi\Bigg\{ & 1.328(1 + 0.127 + 0.065 - 0.036) \\ & + \frac{0.145 + 0.008 - 0.018 - 0.0004}{1.328} \Bigg\} = 0.0509\end{aligned}$$

が得られる．$\hat{\alpha}_2(y) = 0.0528$ よりもさらに良いことが分かる．

9　チューブ座標系

指数型分布族の密度関数は $\exp\left(\theta^i y_i - \psi(\theta) - h(y)\right)$ の形に書ける．自然パラメタ θ^i の代わりに期待値パラメタ $\eta_i = \partial\psi/\partial\theta^i$,

$i = 1,\ldots,p$ を用いた密度関数を $f(y;\eta)$ と書く．ポテンシャル関数は $\phi(\eta) = \max_\theta\{\theta^i\eta_i - \psi(\theta)\}$ である．計量行列 $\phi^{ij}(\eta) = \partial^2\phi(\eta)/\partial\eta_i\partial\eta_j$ の $\eta = 0$ における値を ϕ^{ij} と書く．同様に 1 階微分を ϕ^i，高階微分を ϕ^{ijk}, ϕ^{ijkl} などと書くと，$\phi^{ijk} = O(n^{-1/2})$, $\phi^{ijkl} = O(n^{-1})$ である．適当に座標系を取ることにより，$\phi^i = 0$, $\phi^{ij} = \delta^{ij}$(単位行列) とできる．

$\eta = 0$ の近傍で $\partial\mathcal{R}$ の曲面をパラメタ表示すると，$\eta_a(u) = u_a, a = 1,\ldots,p-1$, $\eta_p(u) \approx -d^{ab}u_au_b - e^{abc}u_au_bu_c$ とテーラ展開できる．ここで $d^{ab} = O(n^{-1/2})$, $e^{abc} = O(n^{-1})$ である．$\mathcal{R}$ は $\eta_p \le \eta_p(u)$ である．$\partial\mathcal{R}$ の単位法線ベクトルを $\mathcal{R}$ の外向きに取り $B^p(u)$ と書く．すると $\eta = 0$ の近傍で η は $\eta_i(u,v) = \eta_i(u) + B_i^p(u)v$, $i = 1,\ldots,p$ と書ける．v は実数であり符号付距離と呼ばれる．(u,v) は法ベクトル空間が 1 次元のチューブ座標系であり，座標変換 $\eta \leftrightarrow (u,v)$ が一対一となる近傍だけを考える．ヤコビアンは $\partial\eta/\partial(u,v) \approx \exp\big[-\frac{1}{2}\phi^{cpp}u_c + (2d^{aa} - \phi^{aap})v - \big\{2(d^{ab})^2 - 2d^{ab}\phi^{abp} + \frac{1}{2}(\phi^{abp})^2\big\}v^2 + \big\{\frac{1}{2}d^{cd}\phi^{ppp} - \frac{1}{4}\phi^{cdpp} + \frac{1}{4}\phi^{cpp}\phi^{dpp} + \frac{1}{2}\phi^{acp}\phi^{adp} + 2d^{ac}(d^{ad} - \phi^{adp})\big\}u_cu_d + \big\{6e^{aac} + d^{aa}\phi^{cpp} + 4d^{ac}\phi^{app} - \phi^{aacp} + \phi^{aad}\phi^{cdp} + \frac{1}{2}\phi^{aap}\phi^{cpp} - 2d^{cd}\phi^{aad} - (2d^{ad} - \phi^{adp})\phi^{acd}\big\}u_cv\big]$ である．これを用いて確率密度関数を (u,v) 座標系で表現する．スケール変換によって u, v が τ^{-1} 倍，ϕ^{ijk}, d^{ab} が τ 倍，ϕ^{ijkl}, e^{abc} が τ^2 倍されることを考慮すると次の結果を得る．

定理 (修正符号付距離の分布関数). $v \approx c_0 + c_1w + c_2w^2 + c_3w^3 - u_cb^c(w)$ によって座標変換 $(u,v) \leftrightarrow (u,w)$ を考える．ただし c_0, c_2 は $O(n^{-1/2})$，$c_1, c_3, b^c(w)$ は $O(n^{-1})$ である．w を修正符号付距離と呼ぶ．真のパラメタ値 η が (u,v) 座標系で $u = 0, v = \lambda$ と仮定しても一般性を失わない．Y は (u,w) 座標系で $(\hat{U},\hat{W})$ と書く．このとき $\hat{W}$ の分布関数を $\Pr\{\hat{W} \le \hat{w}\} = \Phi(z_c(\hat{w};\lambda,\tau))$ と書くと，

$$z_c(\hat{w};\lambda,\tau) \approx \tau^{-1}g_-(\hat{w},\lambda) + \tau g_+(\hat{w},\lambda)$$

である．ただし，$g_-(\hat{w},\lambda) = (\hat{w} - \lambda) - c_0 - \frac{1}{3}\phi^{ppp}\lambda^2 + \frac{1}{6}\phi^{ppp}\lambda\hat{w} + (\frac{1}{6}\phi^{ppp} - c_2)\hat{w}^2 - \frac{1}{6}c_0\phi^{ppp}\lambda - \big\{c_1 + \frac{1}{3}c_0\phi^{ppp}\big\}\hat{w} + \big\{\frac{1}{8}(\phi^{app})^2 + \frac{1}{18}(\phi^{ppp})^2 - \frac{1}{8}\phi^{pppp}\big\}\lambda^3 + \big\{-\frac{1}{8}(\phi^{app})^2 + \frac{1}{24}\phi^{pppp}\big\}\lambda^2\hat{w} + \big\{-\frac{1}{24}(\phi^{ppp})^2 + \frac{1}{24}\phi^{pppp} - \frac{1}{6}c_2\phi^{ppp}\big\}\lambda\hat{w}^2 + \big\{-\frac{1}{72}(\phi^{ppp})^2 + \frac{1}{24}\phi^{pppp} - \frac{1}{3}c_2\phi^{ppp} - c_3\big\}\hat{w}^3$ そして

$g_+(\hat{w},\lambda) = -(d^{aa}+\frac{1}{6}\phi^{ppp})+\{(d^{ab})^2-d^{ab}\phi^{abp}+\frac{1}{6}d^{aa}\phi^{ppp}+\frac{1}{2}(\phi^{abp})^2+\frac{1}{2}(\phi^{app})^2+\frac{13}{72}(\phi^{ppp})^2-\frac{1}{4}\phi^{aapp}-\frac{1}{8}\phi^{pppp}\}\hat{w}+\{(d^{ab})^2-\frac{1}{6}d^{aa}\phi^{ppp}+\frac{1}{8}(\phi^{app})^2+\frac{5}{72}(\phi^{ppp})^2-\frac{1}{24}\phi^{pppp}\}\lambda$ である．

上式の $z_c(\hat{w};\lambda,\tau)$ はパラメタが $(u,v)=(0,\lambda)$ と仮定しているが，一般の (u,v) の場合は $z_c(\hat{w};u,v,\tau)$ と書く．ブートストラップ確率の計算では $z_c(\hat{w}^*;\hat{u},\hat{v},\tau)$ が必要になるが，これは $z_c(\hat{w};\lambda,\tau)$ において $\hat{w}$ を $\hat{w}^*$ で，λ を $\hat{v}$ で置き換え，さらに d^{ab},ϕ^{ijk} を $\eta(\hat{u},0)$ で評価した $\hat{d}^{ab}=d^{ab}+O(n^{-1}),\hat{\phi}^{ijk}=\phi^{ijk}+O(n^{-1})$ に置き換えて得られる．加速定数は $\hat{a}=-\hat{\phi}^{ppp}/6$ である．$\hat{e}^{abc}\approx e^{abc},\hat{\phi}^{ijkl}\approx\phi^{ijkl}$ 等の $O(n^{-1})$ の量に関しては置き換えをする必要はない．$\eta(\hat{u},0)$ は y から $\partial\mathcal{R}$ への射影であり，これを $\hat{\eta}(y)$ と書く．

10　近似的に不偏な確率値の分布関数

数値例 1 や 2 では厳密に不偏な確率値が容易に得られたが，一般には次のように確率値を与えると，3 次の精度になることが分かる．まず $Y^*\sim f(y^*;\hat{\eta}(y),1)$ によって複製を生成し，Y^* を (u,v) 座標系で $(\hat{U}^*,\hat{V}^*)$ と書く．数値例 1 や 2 では y の v 座標 $\hat{v}$ より $\hat{V}^*$ が大きくなる確率が不偏な確率値になっていた．ここでも同様に

$$\hat{\alpha}_\infty(y)=\Pr\{\hat{V}^*\geq\hat{v};\hat{\eta}(y),1\}$$

によって $\hat{\alpha}_\infty(y)$ を定義する．この z 値は $\hat{z}_\infty(y)=z_c(\hat{v};\hat{u},0,1)$ である．（係数 c_r はすべて 0 である．）つまり，$\hat{z}_\infty(y)\approx\hat{v}-(\hat{d}^{aa}+\frac{1}{6}\hat{\phi}^{ppp})+\frac{1}{6}\hat{\phi}^{ppp}\hat{v}^2+\{(d^{ab})^2-d^{ab}\phi^{abp}+\frac{1}{6}d^{aa}\phi^{ppp}+\frac{1}{2}(\phi^{abp})^2+\frac{1}{2}(\phi^{app})^2+\frac{13}{72}(\phi^{ppp})^2-\frac{1}{4}\phi^{aapp}-\frac{1}{8}\phi^{pppp}\}\hat{v}+\{-\frac{1}{72}(\phi^{ppp})^2+\frac{1}{24}\phi^{pppp}\}\hat{v}^3$ である．

より一般的に $\hat{z}_q(y)\approx\hat{z}_\infty(y)+q_0+q_1\hat{v}+q_2\hat{v}^2+q_3\hat{v}^3+\hat{u}_c g^c(\hat{v})$, $q_0=O(n^{-1/2})$, $q_1=O(n^{-1})$, $q_2=O(n^{-1/2})$, $q_3=O(n^{-1})$, $g^c(\hat{v})=O(n^{-1})$ を考えると，この $\hat{z}_q(y)$ 自身が再び $\hat{w}$ として表現され，係数は $c_0=-d^{aa}-\frac{1}{6}\phi^{ppp}+q_0$, $c_1=(d^{ab})^2-d^{ab}\phi^{abp}+\frac{1}{2}d^{aa}\phi^{ppp}+\frac{1}{2}(\phi^{abp})^2+\frac{1}{2}(\phi^{app})^2+\frac{17}{72}(\phi^{ppp})^2-\frac{1}{4}\phi^{aapp}-\frac{1}{8}\phi^{pppp}+\frac{1}{3}\phi^{ppp}(q_2-q_0)+q_1+2d^{aa}q_2-2q_0q_2$, $c_2=\frac{1}{6}\phi^{ppp}+q_2$, $c_3=-\frac{5}{72}(\phi^{ppp})^2+\frac{1}{24}\phi^{pppp}-\frac{2}{3}\phi^{ppp}q_2-2q_2^2+q_3$ である．したがって，この係数を $z_c(\hat{w};\lambda,1)$ に代入すると，$\hat{z}_q(Y)$ の分布関数が得られる．特に $\lambda=0$ とおくと，$\Pr\{\hat{z}_q(Y)\leq x;0\}\approx$

$\Phi\big[x - q_0 - q_2x^2 + \big\{-q_1 - 2q_2(d^{aa} + \frac{1}{6}\phi^{ppp} - q_0)\big\}x + \big\{\frac{1}{3}\phi^{ppp}q_2 + 2q_2^2 - q_3\big\}x^3\big]$ であるから，$\hat{\alpha}_q(y) = \Phi(-\hat{z}_q(y))$ が 3 次の精度の近似的に不偏な確率値になることと，$q_0 \approx q_1 \approx q_2 \approx q_3 \approx 0$ は同値である．つまり，ここで考えた確率値のクラスで 3 次の精度を持つのは，$\hat{\alpha}_\infty(y)$ との違いが $O(n^{-3/2})$ の確率値だけである．この意味で，p^*-formula (Barndorff-Nielsen 1986) やダブルブートストラップ法 (Hall 1992) から得られる確率値も $\hat{\alpha}_\infty(y)$ に等価である．

11　ブートストラップ確率の分布関数

スケール τ のブートストラップ法 $Y^* \sim f(y^*; y, \tau)$ によって複製を生成する．この Y^* を (u, v) 座標系で $(\hat{U}^*, \hat{V}^*)$ と書く．ブートストラップ確率は $\tilde{\alpha}_1(y, \tau) = \Pr\{\hat{V}^* \le 0; y, \tau\}$ なので，z 値は $\tilde{z}_1(y, \tau) = -z_c(0; \hat{u}, \hat{v}, \tau)$ である．（ここで係数 c_r はすべて 0 である．）つまり，$\tilde{z}_1(y, \tau) \approx \tau^{-1}\big[\hat{v} + \frac{1}{3}\hat{\phi}^{ppp}\hat{v}^2 - \big\{\frac{1}{8}(\phi^{app})^2 + \frac{1}{18}(\phi^{ppp})^2 - \frac{1}{8}\phi^{pppp}\big\}\hat{v}^3\big] + \tau\big[(\hat{d}^{aa} + \frac{1}{6}\hat{\phi}^{ppp}) - \big\{(d^{ab})^2 - \frac{1}{6}d^{aa}\phi^{ppp} + \frac{1}{8}(\phi^{app})^2 + \frac{5}{72}(\phi^{ppp})^2 - \frac{1}{24}\phi^{pppp}\big\}\hat{v}\big]$ になる．マルチスケール法は，この τ^{-1} と τ の係数を推定する．特に正規モデルならば $\phi^{ijk} = \phi^{ijkl} = 0$ なので，係数は $\hat{v}$ と $\hat{c} = \hat{d}^{aa} - (d^{ab})^2\hat{v}$ である．これより計算した $\hat{z}_1(y) = \hat{v} - \hat{c}$ が 10 節の $\hat{z}_\infty(y) \approx \hat{v} - \hat{d}^{aa} + (d^{ab})^2\hat{v}$ の右辺に等しく，$\hat{\alpha}_1(y)$ が 3 次の精度であることが分かる．

$\tau\tilde{z}_1(y, \tau)$ は再び $\hat{w}$ 又は $\hat{z}_q(y)$ として表現できて，係数は $q_0 = (1 + \tau^2)(d^{aa} + \frac{1}{6}\phi^{ppp})$, $q_1 = -(1 + \tau^2)(d^{ab})^2 + d^{ab}\phi^{abp} + \frac{1}{4}\phi^{aapp} - \frac{1}{2}(\phi^{abp})^2 - \frac{1}{8}(4+\tau^2)(\phi^{app})^2 + \frac{1}{6}(-1+\tau^2)d^{aa}\phi^{ppp} - \frac{1}{72}(13+5\tau^2)(\phi^{ppp})^2 + \frac{1}{24}(3+\tau^2)\phi^{pppp}$, $q_2 = \frac{1}{6}\phi^{ppp}$, $q_3 = -\frac{1}{8}(\phi^{app})^2 - \frac{1}{24}(\phi^{ppp})^2 + \frac{1}{12}\phi^{pppp}$ である．これよりブートストラップ確率 $\hat{\alpha}_0(y)$ は一般に 1 次の精度しかないことが分かる．

2 ステップ法の z 値は次の積分より得られる．

$$\tilde{z}_2(y, \tau_1, \tau_2) = \Phi^{-1}\Big\{\int \Phi(\tilde{z}_1(y^*, \tau_2))f(y^*; y, \tau_1)\, dy^*\Big\}$$

これをもう一度繰り返すと，3 ステップ法の z 値も得られる．結果として，8 節の $\tilde{z}_3(y, \tau_1, \tau_2, \tau_3)$ の右辺は $O(n^{-3/2})$ の誤差で正しく，その係数は $\hat{\gamma}_1 = \hat{v} + \frac{1}{3}\hat{v}^2\hat{\phi}^{ppp} + \hat{v}^3\big\{-\frac{1}{8}(\phi^{app})^2 - \frac{1}{18}(\phi^{ppp})^2 + \frac{1}{8}\phi^{pppp}\big\}$,

$\hat{\gamma}_2 = \hat{v}\{-\hat{d}^{aa} - \frac{1}{6}\hat{\phi}^{ppp}\} + \hat{v}^2\{(d^{ab})^2 - \frac{1}{2}d^{aa}\phi^{ppp} + \frac{1}{8}(\phi^{app})^2 + \frac{1}{72}(\phi^{ppp})^2 - \frac{1}{24}\phi^{pppp}\}$, $\hat{\gamma}_3 = -\frac{1}{6}\hat{v}\hat{\phi}^{ppp} + \hat{v}^2\{\frac{1}{4}(\phi^{app})^2 + \frac{1}{9}(\phi^{ppp})^2 - \frac{1}{8}\phi^{pppp}\}$, $\hat{\gamma}_4 = \hat{v}^2\{-d^{ab}\phi^{abp} + \frac{1}{3}d^{aa}\phi^{ppp} + \frac{1}{2}(\phi^{abp})^2 + \frac{1}{2}(\phi^{app})^2 + \frac{2}{9}(\phi^{ppp})^2 - \frac{1}{4}\phi^{aapp} - \frac{1}{6}\phi^{pppp}\}$, $\hat{\gamma}_5 = \hat{v}^2\{-\frac{1}{8}(\phi^{app})^2 - \frac{1}{8}(\phi^{ppp})^2 + \frac{1}{12}\phi^{pppp}\}$, $\hat{\gamma}_6 = \hat{v}^2\{-\frac{1}{8}(\phi^{app})^2 - \frac{1}{8}(\phi^{ppp})^2 + \frac{1}{24}\phi^{pppp}\}$ である．この係数を8節の $\hat{z}_3(y)$ の右辺に代入すると，$\hat{z}_\infty(y)$ に等価であることが分かる．つまり3ステップ法は3次の精度である．

なお，2 ステップ法，1 ステップ法の結果を確かめるには，$\tilde{z}_2(y, \tau_1, \tau_2) = \lim_{\tau_3 \to 0} \tilde{z}_3(y, \tau_1, \tau_2, \tau_3)$, $\tilde{z}_1(y, \tau_1) = \lim_{\tau_2 \to 0} \tilde{z}_2(y, \tau_1, \tau_2)$ とすればよい．

参考文献

[1] O. E. Barndorff-Nielsen. Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73(2):307–322, 1986.

[2] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7:1–26, 1979.

[3] B. Efron. Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, 82:171–185, 1987.

[4] B. Efron and R. Tibshirani. The problem of regions. *Ann. Statist.*, 26:1687–1718, 1998.

[5] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[6] P. Hall. *The bootstrap and Edgeworth expansion*. Springer-Verlag, New York, 1992.

[7] H. Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51:492–508, 2002.

[8] H. Shimodaira. Technical details of the multistep-multiscale bootstrap resampling. Research report B-403, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan, 2004.

[9] H. Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, in press.