統計数理（2002）
第 50 巻 第 1 号 33–44

特集「DNA 配列の統計解析」
［研究詳解］

ブートストラップ法によるクラスタ分析のバラツキ評価

下平 英寿†

（受付 2001 年 10 月 1 日；改訂 2002 年 1 月 25 日）

要　　旨

クラスタリングにおけるバラツキを確率値（p-value）として定量的に評価する方法を解説する．もし仮に母集団からデータを何回もサンプルできるとすると，それをクラスタ分析した結果は観測値毎に異なる可能性がある．つまりクラスタ分析の結果得られる樹状図やそれから導かれる群（クラスタ）はデータや特徴量のサンプリングによるバラツキの影響を受けている．そこで観測値から得られた結果がどれほど信頼できるのかを 0 から 1 の範囲の実数を値にとる確率値として表現する．これはクラスタ分析という手法の性能評価をしているのではなく，データが本来持っている情報の不確実性を定量的に評価している．この方法はデータが仮説を支持するかしないかを示す二値関数とブートストラップ法によるリサンプリングだけを使っているので，クラスタ分析に限らずかなり広いクラスの問題に適用可能である．仮説を表す母数空間の領域の近似的に不偏な検定から確率値は計算される．基礎となっているのは Efron（1985）と Efron and Tibshirani （1998）による符号付距離と曲率の理論である．これを実用的な手法にするためのアイデアが Shimodaira（2000, 2002）のマルチスケールブートストラップ法である．生物の DNA から進化を推定する分子系統樹の問題を例題として取り上げる．

キーワード：クラスタ分析，ブートストラップ法，マルチスケールブートストラップ法，近似的に不偏な検定，分子系統樹．

1. はじめに

クラスタ分析（例えば竹内（1989）, p. 381）では分類対象の類似性の情報を用いてその個体をいくつかの階層的な群（クラスタ）に分けることが行われる．すなわち互いに似たものは同じ群に含まれ，さらにこれらの群を幾つかの群に分けるということを階層的に行い，分類の結果は樹状図を用いて表される．応用研究ではしばしばクラスタ分析の結果得られたひとつの樹状図が示され，これをもとに分類対象間の関係が議論される．ところがもし仮に母集団からデータを何回もサンプルできるとすると，それをクラスタ分析した結果は観測値毎に異なる可能性がある．すなわちわれわれが観測したデータから得られた樹状図は，データのサンプリングに関するバラツキや分類に利用した特徴量の選択に関するバラツキに影響されている．本稿では

† 統計数理研究所（現 東京工業大学 情報理工学研究科：〒152–8552 東京都目黒区大岡山 2–12–1; shimo@is.titech.ac.jp）

このバラツキの影響を定量的に評価し，得られた分類結果がどれほど信頼出来るのかを議論する．これはクラスタ分析という手法の性能評価をしているのではなく，データが本来持っている情報の不確実性を定量的に評価する試みである．

分類対象の個数を M とし，それぞれの個体が長さ N からなるデータを考える．これは $M \times N$ の大きさのデータ行列

$$X = (X_{it}\,;\ i = 1, \ldots, M,\, t = 1, \ldots, N)$$

として表される．X_{it} は個体 i の t 番目のデータを表す．一般的なクラスタ分析では，X_{it} は t 番目の特徴量であり，個体 i と個体 j の間の類似度を例えば

$$\sum_{t=1}^{N} (X_{it} - X_{jt})^2$$

によって定める．個体間の類似度から階層的なクラスタを構成する方式（計算機ソフトウエア，アルゴリズム）には様々なものが提案されているが，いずれにしても樹状図をひとつ出力する．どのような類似度を用いどのような方式を使うかは個々の応用では大変重要な問題であるがここではそれには立ち入らず，問題に応じた適切な類似度と方式を採用していると仮定する．

あらかじめ候補となる樹状図があり，それが「本来」の樹状図であるかどうかを検定したいという場合を考える．その樹状図を T で表す．もし複数の候補 $T_1, T_2, \ldots$ がある場合には，以下の議論をそれぞれの候補で繰り返し行う．候補 T は「仮説」を表している．一方，データ X のクラスタ分析の結果得られた樹状図を $T(X)$ で表す．バラツキの影響で $T(X) = T$ の場合もあるだろうしそうでない場合もある．もしたまたま $T(X) = T$ であったなら仮説 T がもっともらしいと判断し，$T(X) \neq T$ ならば仮説 T が疑わしいと考えるのがごく自然である．本稿ではさらに進めて，仮説 T が真実であるかどうかを定量的に評価し，0 から 1 の範囲の値をとる確率値（p-value）を計算する方法を述べる．

仮説は樹状図として与えられるとは限らず，群として与えられるほうが一般的かもしれない．樹状図 T に含まれる階層的な群を

$$T = \{G_1, G_2, \ldots, G_g\}$$

と表す．ただし $G_i \subset \{1, \ldots, M\}$ はひとつの群であり，$G_i \subset G_j$ または $G_j \subset G_i$ または $G_i \cap G_j = \emptyset$ を満たす必要がある．g は T に含まれる群の数であり，一般的な樹状図では $g = M - 1$ である．仮説となるある群 $G \subset \{1, \ldots, M\}$ が，データから得られた樹状図 $T(X)$ にたまたま含まれていて $G \in T(X)$ ならば，我々は仮説 G をもっともらしいと考えるだろうし，逆に $G \notin T(X)$ ならば仮説 G を疑うだろう．しかし単に G が $T(X)$ に含まれているかどうかだけで判断するより，信頼性の程度を定量的に確率値で与えたほうがより望ましい．

仮説を樹状図 T とするか群 G とするかいずれにしても，我々はデータ X から得られた樹状図 $T(X)$ が仮説を支持したときにその仮説をもっともらしいと考え，支持しなかったときに疑わしいと考える．そこでデータ X が仮説を支持するときに値 1 をとり，そうでないときに値 0 をとるような関数 $S(X)$ を考えることにする．つまり仮説が T の場合には，$T(X) = T$ のとき $S(X) = 1$，$T(X) \neq T$ のとき $S(X) = 0$ である．仮説が G の場合には，$G \in T(X)$ のとき $S(X) = 1$，$G \notin T(X)$ のとき $S(X) = 0$ である．仮説が具体的にどのような形式をとるにせよ，この関数 $S(X)$ のみを使って我々は確率値の計算法を与える．したがってこの方法はクラスタ分析だけに限らず，実はかなり広いクラスの問題を扱える．

確率値の計算の基礎は Efron（1979）のブートストラップ法によってデータを複製することである．具体的には 3 節で説明するが簡単に言えば，データ X の複製を乱数を用いたリサンプリングによって多数生成し，それがどのくらいの頻度で仮説を支持するかを数えて確率値を計

算する方法である．このような確率値の計算法は Felsenstein（1985）によって提案されて以来広く用いられている．ところがこの素朴な方法だと確率値のバイアスが一次の精度（漸近的に $O(N^{-1/2})$ のオーダ）しかない．我々が与える計算法は確率値のバイアスが三次の精度（漸近的に $O(N^{-3/2})$ のオーダ）である．理論的な基礎は Efron（1985）と Efron and Tibshirani（1998）であり，仮説を支持する母数空間の領域に関する符号付距離と境界の曲率というものが関わっている．これらの理論に基づき Efron et al.（1996）では二次の精度（漸近的に $O(N^{-1})$ のオーダ）の計算法が与えられていたが，我々の方法はそれより実装が簡単でかつ精度が良い．この近似的に不偏な検定（approximately unbiased test; AU test）の確率値を容易に計算可能にしたアイデアの中心は 4 節で述べるマルチスケールブートストラップ法である．

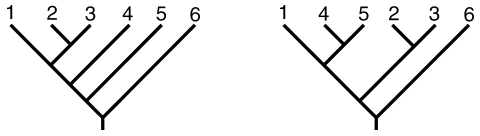
2. 分子系統樹の推定

クラスタ分析の具体例として，DNA から生物の進化を推定する問題を取り上げる．データは Shimodaira and Hasegawa（1999）で用いた図 1 にあるような 6 種の哺乳類のアミノ酸シーケンスである．したがってデータ行列 $X = (X_{it})$ の各要素はアミノ酸に対応した 20 種のアルファベットを値にとる．$M = 6$ 種の哺乳類について長さ $N = 3414$ のシーケンスを用いたので，行列の大きさは 6×3414 である．

ここでは類似度からクラスタ分析する方法ではなく，進化の確率モデルに基づいた最尤法を用いてクラスタ分析を行う（Cavalli-Sforza and Edwards（1967), Felsenstein（1981), 長谷川・岸野（1996））．これを簡単に述べると以下のようになる．最尤法では各樹状図 T 毎に対数尤度 $L(T, X)$ を計算する．対数尤度というのは，もっともらしさをあらわす量だと思えばよい．M 個の分類対象の可能な樹状図の数を n とする．候補となるすべての樹状図 $T_1, T_2, \ldots, T_n$ に対して対数尤度 $L(T_1, X), \ldots, L(T_n, X)$ を計算し，その中で対数尤度を最大にする樹状図を採用する．手続きは一見複雑だがデータ X から樹状図 $T(X)$ を一つ選ぶことには変わりない．いずれ

```
          2         3         4         5         6         7         8         9
          01234567890123456789012345678901234567890123456789012345678901234567890123456789
human     ERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYITAPTLALTIALLLWAPLPMPNPLVNLNLGLL
seal      ERKVLGYMQLRKGPNIVGPYGLLQPIADAVKLFTKEPLRPLTSSTTMFIMAPILALALALTMWVPLPMPYPLINMNLGVL
cow       ERKVLGYMQLRKGPNVVGPYGLLQPIADAIKLFIKEPLRPATSSASMFILAPIMALGLALTMWIPLPMPYPLINMNLGVL
rabbit    ERKILGYMQLRKGPNIVGPYGLLQPIADAIKLFTKEPLRPLTSSPLLFIIAPTLALTLALSMWLPIPMPYPLVNLNMGIL
mouse     ERKILGYMQLRKGPNIVGPYGILQPFADAMKLFMKEPMRPLTTSMSLFIIAPTLSLTLALSLWVPLPMPHPLINLNLGIL
opossum   ERKVLGYMQFRKGPNVIGPYGILQPFADALKLFIKEPLRPMTSSISMFTIAPTLALTLAFTIWTPLPMPNALLDLNLGLL
```

図 1. 哺乳類（ヒト，アザラシ，ウシ，ウサギ，マウス，オポッサム）のミトコンドリア DNA のアミノ酸シーケンス．解析に用いた長さ $N = 3414$ のデータのうち $t = 20$ から $t = 99$ の部分までを示した．



(a) (b)

図 2. 哺乳類（ヒト=1，アザラシ=2，ウシ=3，ウサギ=4，マウス=5，オポッサム=6）の系統樹．（a）最尤法で選ばれた系統樹．(b）最新のデータが支持している系統樹．

にしても最尤法を自動的に行う計算機プログラム（Adachi and Hasegawa (1996), Yang (1997), Swofford (1998)）が開発されているので，手続きが複雑かどうかということはさほど問題ではない．

哺乳類のデータから計算した $T(X)$ は図 2(a) に示した．DNA から推定した生物種の樹状図は分子系統樹とも呼ばれる．$T(X)$ は生物が進化の過程で分化していった順序を表している．ところが，この $T(X)$ をそのまま真実として受け入れるのは危険である．データ X は進化の確率モデルで定義される確率変数の実現値であり，サンプリングによるバラツキがある．したがって X から計算される $T(X)$ にもバラツキがある．実際，その後新たに得られたデータや生物学的な知識を動員すると，どうやら図 2(b) が真実ではないかと現在では考えられている（長谷川政美（私信), Cao et al. (2000), Madsen et al. (2001), Murphy et al. (2001)）．この最新データを入手する以前に戻って考えると，図 1 のデータから推定された樹状図をそのまま信じていたら誤った結論に導かれていた可能性が高い．このような早まった結論を避けるためには，バラツキを考慮して $T(X)$ の信頼性を評価することが必要になる．

3. ブートストラップ法

データ X から計算する樹状図 $T(X)$ のバラツキを見るには，母集団からデータを何回もサンプルして樹状図がどのように分布するかを見ればよい．ところが実際には母集団から得られるのはひとつのデータ X だけである．そこで Efron (1979) はデータからのリサンプリングによって X の複製を何回でも生成できる一般的な方法を考え，ブートストラップ法と名づけた．これを以下で説明する．

データ行列 $X = (X_{it})$ を

$$X = (x_1, x_2, \ldots, x_N)$$

と書く．ただし，$x_t = (X_{it}\,;\, i = 1, \ldots, M)$ はデータ行列の t 列目をあらわす．X の複製 X^* は

$$X^* = (x_{t_1}, x_{t_2}, \ldots, x_{t_N})$$

と書かれる．ここで $t_1, \ldots, t_N$ は $1, \ldots, N$ のどれかの値を重複を許してランダムに取ることにするので，X^* の $1, 2, \ldots$ 列目は X の $t_1, t_2, \ldots$ 列目を取り出したものである（図 3）．

$t_1, \ldots, t_N$ は次のように乱数を使って作られる．まず $1, \ldots, N$ からランダムに数を選びそれを t_1 とする．同様に $1, \ldots, N$ からランダムに数を選びそれを t_2 とする．ここで重複を許すので t_1 と t_2 がたまたま同じ数になることもある．これを N 回繰り返して $t_1, \ldots, t_N$ を生成する．

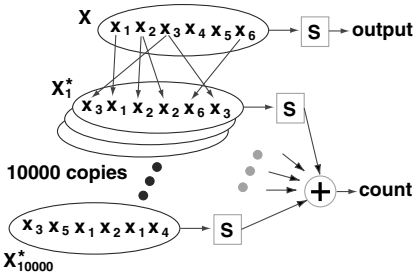


図 3. ブートストラップ法．ここでは $N = 6$ のデータから複製を $B = 10000$ 個生成している．複製が仮説を何回支持したかがカウントされる．

表 1. 上位 15 個の樹状図と確率値.

番号	$\tilde{p}$	$\hat{p}$	樹状図
1	**0.579**	**0.792**	((((1(23))4)5)6)
2	**0.312**	**0.517**	(((1((23)4))5)6)
3	0.035	**0.131**	((1((23)(45)))6)
4	0.036	**0.115**	((((14)(23))5)6)
5	0.017	**0.103**	(((1(45))(23))6)
6	0.013	**0.076**	(((1(23))(45))6)
7	0.005	0.030	((1(((23)4)5))6)
8	0.003	0.028	((((15)4)(23))6)
9	0.001	0.009	(((15)((23)4))6)
10	0.000	0.005	(((14)((23)5))6)
11	0.000	0.003	((((14)5)(23))6)
12	0.000	0.002	(((1((23)5))4)6)
13	0.000	0.001	((((15)(23))4)6)
14	0.000	0.001	((1(((23)5)4))6)
15	0.000	0.001	(((((13)2)4)5)6)

番号 16–105 の樹状図は確率値 0.000.

表 2. 上位 9 個の群と確率値.

番号	$\tilde{p}$	$\hat{p}$	群
1	**1.000**	**1.000**	{23}
2	**0.927**	**0.954**	{1234}
3	**0.592**	**0.749**	{123}
4	**0.318**	**0.469**	{234}
5	0.036	**0.111**	{14}
6	0.040	**0.088**	{2345}
7	**0.065**	**0.075**	{45}
8	0.019	**0.069**	{145}
9	0.004	0.015	{15}

番号 10–25 の群は確率値 0.000.

つまり，N 通りの目があるサイコロを N 回振ってその出た目を記録するのと同じである．そして出た目の X の列を順に取り出して X^* が作られる．

ブートストラップ法では X の複製 X^* を作る手続きを B 回繰り返し，B 個の複製

$$X_1^*, X_2^*, \ldots, X_B^*$$

を生成する．ただし B は十分に大きな数（例えば $B = 10000$）とする．この多数の複製のバラツキは，母集団における X のバラツキを近似的に表していると考えられる．従って

$$T(X_1^*), T(X_2^*), \ldots, T(X_B^*)$$

のバラツキを調べることによって，$T(X)$ がどれほど信頼できるかが評価できる．

ここで 1 節で説明したように問題を少し一般化して仮説の支持または不支持を表す関数 $S(X)$ を用いる．そして

$$S(X_1^*), S(X_2^*), \ldots, S(X_B^*)$$

のうち値が 1 になった回数を C とする．つまり，

$$C = S(X_1^*) + \cdots + S(X_B^*)$$

と書いても良い．Felsenstein（1985）はブートストラップ確率を

$$\tilde{p} = \frac{C}{B}$$

と定義し，これが 1 に近いほど仮説はもっともらしく，0 に近いほど仮説は疑わしいと考えた．

哺乳類のデータにこのブートストラップ確率を計算した結果を表 1 と表 2 に示した．6 種の哺乳類のラベルは図 2 で示したものである．表 1 では 105 通りの樹状図をそれぞれ仮説として確率値を計算し，表 2 では 25 通りの群をそれぞれ仮説としている．ここではオポッサムは常に一番外側に置いて群 {1,2,3,4,5} が常に正しいと仮定して分析してある．少しテクニカルな話になるがオポッサムはこの場合アウトグループと呼ばれ，reversible なマルコフ過程を進化のモデルに採用した最尤法では「無根系統樹」しか推定できないという制約上必要な処置である．

図 2(a) の樹状図は表 1 の 1 行目に対応し，図 2(b) は 5 行目に対応する．ブートストラップ確率 $\tilde{p}$ を見ると，1 行目では $\tilde{p} \geq 0.05$ であり有意水準 $\alpha = 0.05$ では棄却されない．ところが 5 行目では $\tilde{p} < 0.05$ であり，この樹状図は棄却される．2 節で述べたように 5 行目の樹状図が最新のデータで支持されており，解析に用いた古いデータでは残念ながらそれと矛盾する結論を導いてしまったことになる．これは表 2 にも反映されていて，8 行目の群 {1,4,5} は $\tilde{p} < 0.05$ で棄却されている．この群は図 2(b) に含まれるので，やはり最新データと矛盾した結論になる．すなわちバラツキを評価していても，必ずしも正しい結論に導かれるわけではない．例えば Graur et al. (1996) では同じような哺乳類のデータを分析して，ウサギとマウスからなる群 {4,5} を有意に棄却したが，これも後になってみるとおかしいと考えられている．どのような方式でバラツキを評価してもこのような事態は起こりえるが，問題なのは素朴なブートストラップ確率にバイアスがあり，誤った結論に導かれる可能性が必要以上に高いということである．そこで次の節で述べるような改良が必要になる．

4. マルチスケールブートストラップ法

ブートストラップ法では X からランダムに N 個の列を取り出して複製 X^* を作った．もし取り出す個数（つまり複製の長さ）を変えて N' とすると複製は

$$X^* = (x_{t_1}, x_{t_2}, \ldots, x_{t_{N'}})$$

となる．普通のブートストラップ法では $N' = N$ であるが，もし $N' \neq N$ とすると複製のバラツキの程度（標準偏差）が変化する．例えば $N' = 2N$ とすればバラツキの程度は $1/\sqrt{2}$ 倍になり，逆に $N' = N/2$ とすればバラツキの程度は $\sqrt{2}$ 倍になる．一般に $N' = rN$ とすればバラツキの程度は $1/\sqrt{r}$ 倍になる．$1/\sqrt{r}$ を複製のスケールと呼び，それはデータ長の比 r によって制御できる．ブートストラップ確率 $\tilde{p}$ も N'，つまり r に依存して変化する．実際 5 節で説明するように，$N' = rN$ の時のブートストラップ確率の「理論値」は

$$\pi(r; d, c) = 1 - \Phi(d\sqrt{r} + c/\sqrt{r}) \tag{4.1}$$

で与えられる．ただし $\Phi(\cdot)$ は標準正規分布関数，d は符号付距離，c は境界の曲率に関係した量であり，詳細は後ほど説明される．Shimodaira (2000, 2002) は $\tilde{p}$ の変化からより精度の高い AU test の確率値を求める方法を提案し，これをマルチスケールブートストラップ法と名づけた．この手続きは以下のようになる（図 4）．

ステップ 1．K 組のブートストラップを考える．データ長の比 $r_1, r_2, \ldots, r_K$，複製の個数 $B_1, B_2, \ldots, B_K$ を定める．以下の数値例では，$K = 10$, $r_1 = 0.5, r_2 = 0.6, \ldots, r_{10} = 1.4$, $B_1 = \cdots = B_{10} = 10,000$ を用いた．

ステップ 2．各 $k = 1, \ldots, K$ について，B_k 個の複製を $N' = r_k N$ を使って生成する．これを

$$X_1^*(r_k), X_2^*(r_k), \ldots, X_{B_k}^*(r_k)$$

と書く．そして複製が仮説を支持するかしないかを

$$S(X_1^*(r_k)), S(X_2^*(r_k)), \ldots, S(X_{B_k}^*(r_k))$$

によって調べる．仮説が支持された回数は

$$C(r_k) = S(X_1^*(r_k)) + S(X_2^*(r_k)) + \cdots + S(X_{B_k}^*(r_k))$$

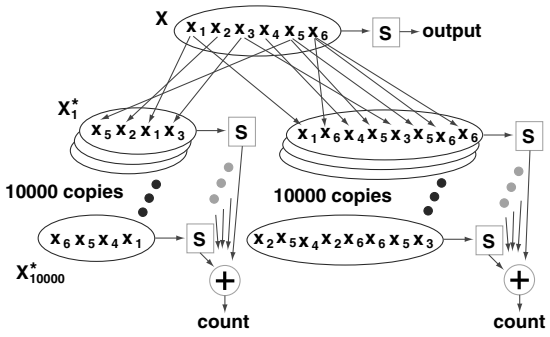


図 4. マルチスケールブートストラップ法．ここでは $N = 6$ のデータから， $N' = 4$ と $N' = 8$ の複製を $B_1 = B_2 = 10000$ 個ずつ生成している．各々のブートストラップ法で，複製が仮説を何回支持したかがカウントされる．

である．これからブートストラップ確率を

$$\tilde{p}(r_k) = \frac{C(r_k)}{B_k}$$

と計算する．

ステップ 3．計算された $\tilde{p}(r_k)$ をその理論値 $\pi(r_k; d, c)$ の曲線に当てはめ，回帰係数 d と c を推定する．具体的には重みつき最小二乗法（WLS）を使って

$$\mathrm{RSS}(d, c) = \sum_{k=1}^{K} (\Phi^{-1}(\pi(r_k; d, c)) - \Phi^{-1}(\tilde{p}(r_k)))^2 / v_k \,,$$

を最小にするような d と c を計算する．ここで $\Phi^{-1}(\cdot)$ は $\Phi(\cdot)$ の逆関数であり，分散 v_k は

$$v_k = \tilde{p}(r_k)(1 - \tilde{p}(r_k))/(\phi(\Phi^{-1}(\tilde{p}(r_k)))^2 B_k)$$

で与えられる．$\phi(\cdot)$ は標準正規密度関数である．

ステップ 4．推定した d と c を使い，補正した確率値を

$$\hat{p} = 1 - \Phi(d - c)$$

で計算する．

ただしステップ 3 における WLS をやめて，$B_k \tilde{p}(r_k)$ が母数 $\pi(r_k; d, c)$ の二項分布に従うことを利用した最尤法（MLE）で d と c を推定してもよい．これには WLS で推定した d と c を初期値として，ニュートン法などで数値的に

$$L(d, c) = \sum_{k=1}^{K} B_k \{\tilde{p}(r_k) \log \pi(r_k; d, c) + (1 - \tilde{p}(r_k)) \log(1 - \pi(r_k; d, c))\} \,,$$

を最大化して d と c を計算する．マルチスケールブートストラップ法は WLS と MLE の両方とも計算機ソフトウエア CONSEL（Shimodaira and Hasegawa（2001））に実装されている．CONSEL は分子系統樹のソフトウエアとの連携を意識して作られているが，その他の問題にも使える．ユーザは $\tilde{p}(r_k)$ を自分の問題にあわせて計算すればよい．

補正した確率値 $\hat{p}$ を哺乳類のデータで計算した結果は表 1 と表 2 に示されている．同じデータセットを用いた同様の計算は Shimodaira（2002）で行った．全般的に $\hat{p}$ は $\tilde{p}$ より大きめの値

になり各仮説は棄却されにくくなる．これは $c \geq 0$ となっていることから理解できるのだが，じつは次節で述べる仮説の領域 H が凸であり曲率が正であることが関係している．表 1 の 5 行目を見ると $\hat{p} \geq 0.05$ となって，この樹状図はもう棄却されなくなる．結果として最新データと矛盾しない結論を導いている．このことは表 2 の 8 行目にも反映されており，群 {1,4,5} はもう棄却されない．データのバラツキによって誤った仮説（表 1 の 1 行目，表 2 の 1, 2 行目）が最も高く支持されていたが，最新データによって最も高く支持されるようになった仮説（表 1 の 5 行目，表 2 の 7, 8 行目）も否定されていなかったわけである．

5. 近似的に不偏な検定

このようにして補正した確率値 $\hat{p}$ は，素朴なブートストラップ確率 $\tilde{p}$ より AU test としては一般的にずっと精度が良く，すなわち検定のバイアスが小さい．このことを以下で説明する．まず X の適当な関数

$$Y = f(X)$$

を考える．ただし Y はベクトルでその次元を m とする．そして

$$Y \sim N_m(\mu, I_m) \tag{5.1}$$

のように未知の平均ベクトル μ，共分散が単位行列の m 次元多変量正規分布に従っていると仮定する．逆にいうと，少なくとも近似的に（5.1）式が適当な m で成り立つような関数 $f(X)$ の存在を仮定する．そして m 次元空間の領域 H を考え，$f(X) \in H$ なら $S(X) = 1$，$f(X) \notin H$ なら $S(X) = 0$ とする．つまり，$Y \in H$ ならデータは仮説を支持し，$Y \notin H$ なら支持しない（図 5）．領域 H の境界を ∂H と書き，境界上で Y へ最も近い点を $\hat{\mu}$ と書くことにする．そして境界 ∂H は滑らかであると仮定する．

さて確率値 p が領域 H の検定に関して不偏であるとは，任意の有意水準 $0 < \alpha < 1$ に対して

$$\Pr\{p < \alpha \mid \mu\} \leq \alpha, \quad \mu \in H$$
$$\Pr\{p < \alpha \mid \mu\} \geq \alpha, \quad \mu \notin H$$

が成り立つことを言う．従って

$$\Pr\{p < \alpha \mid \mu\} = \alpha, \quad \mu \in \partial H \tag{5.2}$$

が成り立つ．一般に $p < \alpha$ のとき仮説は棄却される．不偏な検定では未知パラメタ μ がちょうど仮説の境界上 ∂H にあるとき，仮説を棄却する確率が α になる．そして μ が H の外側に出て

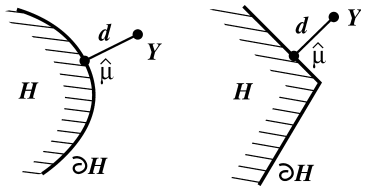


図 5. データベクトル Y が仮説を支持する領域 H; Shimodaira (2002)．領域の境界 ∂H 上の点で Y に最も近い点 $\hat{\mu}$ から Y までの符号付距離が d. (a) 境界 ∂H が滑らか. (b) 境界 ∂H が尖っている．

離れていくほど棄却確率は α より大きくなり，逆に μ が H の内側に入っていくほど棄却確率は α より小さくなる．結論から言うと，ブートストラップ確率 $\tilde{p}$ は（5.2）の誤差が $O(N^{-1/2})$ であるが，補正した確率値 $\hat{p}$ は（5.2）の誤差が $O(N^{-3/2})$ になる．適当に大きな数を N に代入するとわかるが，$N^{-1/2}$ より $N^{-3/2}$ のほうが小さな値になる．つまり $\tilde{p}$ より $\hat{p}$ のほうが誤差が小さい．

Efron（1985）の「補題」もしくは Efron and Tibshirani（1998）の（2.16）式によれば，三次の精度をもつ補正した確率値は

$$\hat{p} = 1 - \Phi(d - c) \tag{5.3}$$

と書ける．ただし，$d = \pm\|Y - \hat{\mu}\|$ で定義し，符号は $Y \in H$ のとき負，$Y \notin H$ のとき正とする．c は ∂H の曲率に関係した量である．実際 $c = c_1 - dc_2$ と書けて，$c_1 = \lambda_1 + \cdots + \lambda_{M-1}$ と $c_2 = \lambda_1^2 + \cdots + \lambda_{M-1}^2$ は $\hat{\mu}$ における ∂H の曲率を表す $(M-1) \times (M-1)$ 行列の固有値 $\lambda_1, \ldots, \lambda_{M-1}$ から計算できる．

たしかに式（5.3）は精度の高い確率値を与えているが，実際の応用では d や c を解析的に与えることは非常に困難であるから，このままでは（5.3）は役に立たない．そこで 4 節のマルチスケールブートストラップ法が開発され，d と c を現実の問題で数値的に計算することが可能になった．その仕組みを理解するために，ブートストラップ確率が十分大きな B で

$$\tilde{p} = 1 - \Phi(d + c) \tag{5.4}$$

と書けるという Efron and Tibshirani（1998）の（2.19）式を利用する．（5.3）と（5.4）の違いは c の符号だけである．したがってもし境界 ∂H が平坦で $c = 0$ ならば $\tilde{p} = \hat{p}$ となる．ところが ∂H が曲がってくると $\tilde{p}$ と $\hat{p}$ は逆の方向へ変化してしまうので，$\tilde{p}$ は $\hat{p}$ の推定値としては精度が悪くなる．

マルチスケールブートストラップでデータ長を $N' = rN$ とすると（5.4）も変化する．Y の複製のバラツキが $1/\sqrt{r}$ 倍となってしまうので，これを元に戻して（5.4）を利用するには，Y の代わりに $\sqrt{r}Y$ を考えればよい．つまり図 5 の絵全体を $\sqrt{r}$ 倍拡大するのと同じである．こうすると同時に d が $\sqrt{r}\,d$ になり c が $c/\sqrt{r}$ になってしまう．こうして（4.1）で与えたブートストラップ確率の理論値 $\pi(r; d, c)$ が出てくる．これで 4 節の方法が理解できたことになる．

仮定したモデル（5.1）は制約が強すぎるように見えるかもしれない．例えば共分散行列が単位行列というのはとても強い制約である．ところが共分散行列が一般の場合でもそれを単位行列にするような Y の線形変換が存在する．従ってその線形変換も $f(\cdot)$ に含めてしまえば結局（5.1）に帰着できる．任意の滑らかな非線形変換を $f(\cdot)$ として使えるので，かなり広い範囲の問題が（5.1）に帰着できる．そして変換 $f(\cdot)$ におけるブートストラップ確率の不変性より，4 節の方法はそのまま使えることになる．

しかしまったく問題が無いわけではない．領域の境界はしばしば滑らかでなく，図 5(b) のように尖っている．滑らかな変換 $f(\cdot)$ ではこの特異性を消すことができず，結局図 5(b) を図 5(a) で近似することになる．これが検定のバイアスにつながり，特異性が無視できない場合には補正した確率値は必ずしも高い精度をもつわけではない．

6. おわりに

データが仮説を支持する $(S(X) = 1)$ かしない $(S(X) = 0)$ かという情報と，ブートストラップ法によるリサンプリングだけを使って AU test として精度の良い確率値を計算する方法を解説した．問題設定のシンプルさから，この方法はクラスタ分析に限らず，かなり広いクラスの問

題に適用可能であろう．基礎となっているのは Efron（1985）と Efron and Tibshirani（1998）による符号付距離と曲率の理論である．これを実用的な手法にするためのアイデアが Shimodaira（2000, 2002）のマルチスケールブートストラップ法である．

マルチスケールブートストラップ法の構成要素として本稿では単純なブートストラップ法によるリサンプリングを用いたが，応用では問題ごとにサンプリング法の工夫が必要である．例えば時系列データではブロックブートストラップ法を用いる．クラスタリングの特徴量を直接サンプリングするような場合は，特徴量の影響度に応じた重み付けしたリサンプリングが意味を持つだろう．いずれにしても形式的にブートストラップ法を適用するのではなく，何に関するバラツキを評価したいのかを適切に考慮する必要がある．

問題によっては計算量を減らす工夫が必要である．例えば分子系統樹の解析では，樹状図の対数尤度 $L(T_i, X_b^*)$ をすべての $i = 1, \ldots, n$, $b = 1, \ldots, B$ に対して計算するのは非常にコストがかかる．そこで，一種の線形近似である RELL 法（Kishino et al.（1990））を用いて近似的に対数尤度を計算した．Hasegawa and Kishino（1994）の数値例や Shimodaira（2001）の補題 1 で示したように，N が大きい問題ではこの近似の精度は十分に良い．

分子系統樹の解析では対数尤度を最大にする樹状図を選択しているので，これは統計的モデル選択の一例になる．下平（1993, 1999），Shimodaira（1998）はリサンプリングを利用して対数尤度の多重比較を行いモデル選択の信頼性を確率値として評価する方法を提案した．これは Shimodaira-Hasegawa（SH）test（Shimodaira and Hasegawa（1999）, Goldman et al.（2000））として分子系統樹の分野で利用されはじめている．この分野では Kishino-Hasegawa（KH）test（Kishino and Hasegawa（1989））がブートストラップ確率に並ぶ標準手法として広く利用されているが，SH test は KH test で見落とされていた「選択バイアス」を多重比較法で補正したものである．本稿の AU test における補正した確率値 $\hat{p}$ は，実は SH test と定性的には同じ役割を果たしている．つまり選択バイアスの大きさと境界の曲率とは定性的には同じものである．ただし多重比較では図 5(b）の尖りの先端に μ があると仮定することによって「最悪ケース」を想定した補正を行っているのに対して，本稿の AU test では図 5(a）のように滑らかな境界を仮定したうえで $\hat{\mu}$ 周辺の「典型的ケース」を想定した補正を行っている．これらの方法から計算される確率値は想定したケースの違いを反映して定量的には異なってくる．SH test のほうが AU test より保守的な結果になる．

計算機ソフトウエア CONSEL（Shimodaira and Hasegawa（2001））の DOS バイナリと UNIX ソースコードは著者より無償で入手可能である．これは AU test, KH test, SH test, ブートストラップ確率などを同時に計算する．

参 考 文 献

Adachi, J. and Hasegawa, M.（1996）. MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood, *Comput. Sci. Monographs*, No. 28, The Institute of Statistical Mathematics, Tokyo.

Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. and Hasegawa, M.（2000）. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, *Gene*, **259**, 149–158.

Cavalli-Sforza, L. L. and Edwards, A. W. F.（1967）. Phylogenetic analysis: Models and estimation procedures, *Evolution*, **32**, 550–570.

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

Efron, B.（1985）. Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, **72**,

45–58.

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

Efron, B., Halloran, E. and Holmes, S.（1996）. Bootstrap confidence levels for phylogenetic trees, *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 13429–13434.

Felsenstein, J.（1981）. Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J.（1985）. Confidence limits on phylogenies: An approach using the bootstrap, *Evolution*, **39**, 783–791.

Goldman, N., Anderson, J. P. and Rodrigo, A. G.（2000）. Likelihood-based tests of topologies in phylogenetics, *Systematic Biology*, **49**, 652–670.

Graur, D., Duret, L. and Gouy, M.（1996）. Phylogenetic position of the order Lagomorpha（rabbits, hares and allies）, *Nature*, **379**, 333–335.

Hasegawa, M. and Kishino, H.（1994）. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree, *Molecular Biology and Evolution*, **11**, 142–145.

長谷川政美，岸野洋久（1996）.『分子系統学』，岩波書店，東京.

Kishino, H. and Hasegawa, M.（1989）. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea, *Journal of Molecular Evolution*, **29**, 170–179.

Kishino, H., Miyata, T. and Hasegawa, M.（1990）. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *Journal of Molecular Evolution*, **30**, 151–160.

Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. and Springer, M. S.（2001）. Parallel adaptive radiations in two major clades of placental mammals, *Nature*, **409**, 610–614.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. and O'Brien, S. J.（2001）. Molecular phylogenetics and the origins of placental mammals, *Nature*, **409**, 614–618.

下平英寿（1993）. モデルの信頼集合と地図によるモデル探索，統計数理，**41**, 131–147.

Shimodaira, H.（1998）. An application of multiple comparison techniques to model selection, *Ann. Inst. Statist. Math.*, **50**, 1–13.

下平英寿（1999）. モデル選択理論の新展開，統計数理，**47**, 3–27.

Shimodaira, H.（2000）. Another calculation of the *p*-value for the problem of regions using the scaled bootstrap resamplings, Tech. Report, No. 2000-35, Stanford University, California.

Shimodaira, H.（2001）. Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection, *Comm. Statist. Theory Methods*, **30**, 1751–1772.

Shimodaira, H.（2002）. An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, **51**, 492–508.

Shimodaira, H. and Hasegawa, M.（1999）. Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution*, **16**, 1114–1116.

Shimodaira, H. and Hasegawa, M.（2001）. CONSEL: For assessing the confidence of phylogenetic tree selection, *Bioinformatics*, **17**, 1246–1247.

Swofford, D. L.（1998）. *PAUP*. Phylogenetic Analysis Using Parsimony* (**and Other Methods*), Version 4, Sinauer Associates, Sunderland, Massachusetts.

竹内　啓 編（1989）.『統計学辞典』，東洋経済新報社，東京.

Yang, Z.（1997）. PAML: A program package for phylogenetic analysis by maximum likelihood, *CABIOS*, **13**, 555–556.

Assessing the Uncertainty of the Cluster Analysis Using the Bootstrap Resampling

Hidetoshi Shimodaira

(The Institute of Statistical Mathematics)

This paper reviews the method of calculating the p-value for assessing the uncertainty of cluster analysis. Considering that the dendrogram as well as the derived clusters obtained by the cluster analysis is subject to change due to the fluctuation of the sampling of the data or that of the characters, the reliability of the result is represented as the p-value, between 0 and 1. This method is applicable to a wide class of problems, and is not limited to cluster analysis, since it uses only bootstrap resampling and the 0/1-value function to indicate whether the data supports the hypothesis. The p-value is calculated from the approximately unbiased test of the region in the parameter space representing the hypothesis. The method is based on the theory of "signed distance" and "curvature" by Efron (1985) and Efron and Tibshirani (1998). The key idea to convert the theory into a practical algorithm is the multiscale bootstrap resampling of Shimodaira (2000, 2002). The issue is illustrated by the phylogeny analysis to infer the history of evolution from the DNA sequences.

Key words: Cluster analysis, bootstrap, multiscale bootstrap, approximately unbiased test, molecular phylogeny.