# 分類系統樹の性質とその比較評価方式

統計数理研究所　大隅　昇
〃　中村　隆

(1980年1月　受付)

## Some Properties of Monotone Hierarchical Dendrogram in Numerical Classification

Noboru Ohsumi and Takashi Nakamura
(The Institute of Statistical Mathematics)

The solution of an agglomerative hierarchical clustering problem is generally given as a dendrogram. The distance obtained from the dendrogram which is characterized by a hierarchical structure $H$ and a level function $h$ possesses the ultrametric properties.

In this paper, a relationship between a set of eigenvalues obtained by applying the principal coordinates analysis to the ultrametric distance matrix $\boldsymbol{D}^{+}=(\delta_{ij})$ on $\langle H, h\rangle$ and values of a level function on $\langle H, h\rangle$ is discussed. As a result, for example, formation of clusters on $\langle H, h\rangle$ is represented by the linear combination of the squared values of level function, which is corresponding to the sum of eigenvalues. Moreover we can deduce that arrangement of each element in the configuration matrix formed by the principal coordinates analysis is related to the fusion of each cluster on $\langle H, h\rangle$. Finally we propose a useful procedure of comparison among several dendrograms (i.e., configuration matrices) by using Gower's Generalized Procrustes analysis.

## 1. まえがき

現存する数多くの数値分類法（あるいはクラスター分析法）のうち，凝集型階層的手法（Agglomerative Hierarchical Clustering; AHC 手法）は計算用のプログラムが手軽に作成できることもあって多くの分野で利用されている．ふつう入力データとして分類対象間の非類似度または類似度行列が与えられると，これに適当な AHC 手法を適用して，出力情報である解の分類系統樹（デンドログラム）を作るというのが標準的な分析手順である．利用者はこの系統樹の上で分類対象間の関連性を観察し必要に応じて適当にクラスター化を行う．

多くの場合，AHC 手法は非類似度あるいは類似度としてノンメトリックな場合までも扱うことができるという特徴を持っている．事実，生物学，心理学，社会学を初め多くの分野で扱うデータはこうした性質を備えていることが多いので，AHC 手法が頻繁に利用される．

しかし，ここで得られた系統樹の上の分類対象間の関係は必ずしもノンメトリックではないという点に注意しなくてはならない．系統樹の上に現われた分類対象間の関連性はそのとき用いた算法によって生成された新たな関係であって，必ずしももとのそれと一致するとは限らないのである．しかし，分類対象間に何らかの類似性が内在しているのであれば系統樹の上に作られた新たな関係も，それを近似していると解釈される．実際，AHC 手法の利用者は系統樹を観察してクラスター構成を解釈するという手続きを踏んでいる．

このように系統樹は，多数の分類対象間の関係を2次元平面内の情報に要約するという利点もあるが，一般にその解釈は恣意的となりやすい．また，それが一種の近似表現であることから様々の問題が生じる．たとえば，次のような点である．

1) 系統樹の上にみられるクラスター生成過程をいかに評価すべきか，つまり系統樹上の分類対象の連結順位は同じであってもクラスターのまとまりの程度は様々である．この連結順位と階層構造に内在する特性をどのように評価するか．

2) 入力データとして与えられた非類似度（類似度）行列を，解である系統樹がどの程度近似しているのであろうか．(解の適合性の評価)．

3) 系統樹間の比較をいかに行うか．たとえば，同一データに複数の算法を用いて求めた複数の系統樹をどのように比較し解釈すべきか．

これらを解決するために，本研究では系統樹のもつ情報を主座標分析により別の情報に変換することを考えた．すなわち，系統樹上の強メトリック距離と主座標分析で得られる固有値，負荷量ベクトルとの間にみられる関係を求め，続いてこれを利用すると系統樹の適切な評価を行うことが可能となることを明らかにする．

## 2. 凝集型の階層的数値分類法

### 2.1 階層的構造と強メトリック距離

一般に，凝集型の階層的数値分類法（AHC 手法）とは分類結果が図1のような系統樹で与えられる手法をいう．いま分類対象として大きさが $n$ の個体集合 $E=\{1, 2, \cdots, i, \cdots, n\}$ が与えられたとき，適当な個体間の非類似度 $d_{ij}$ を約束し，非類似度行列 $\boldsymbol{D}=(d_{ij})$ $(i, j=1, 2, \cdots, n)$ として表現する（類似度を考えることも出来るが，ここでは非類似度について述べることにする）．前述のようにこの $d_{ij}$ は必ずしもメトリックである必要はない．次にこの $\boldsymbol{D}$ に対し適当な算法を適用して，類似した個体から逐次的に連結・併合を繰り返すと図1のような系統樹が生成されるので，これを適当に切断すればクラスター化が達成される．ここで，後の議論のために必要な記法と概念を準備しておく．

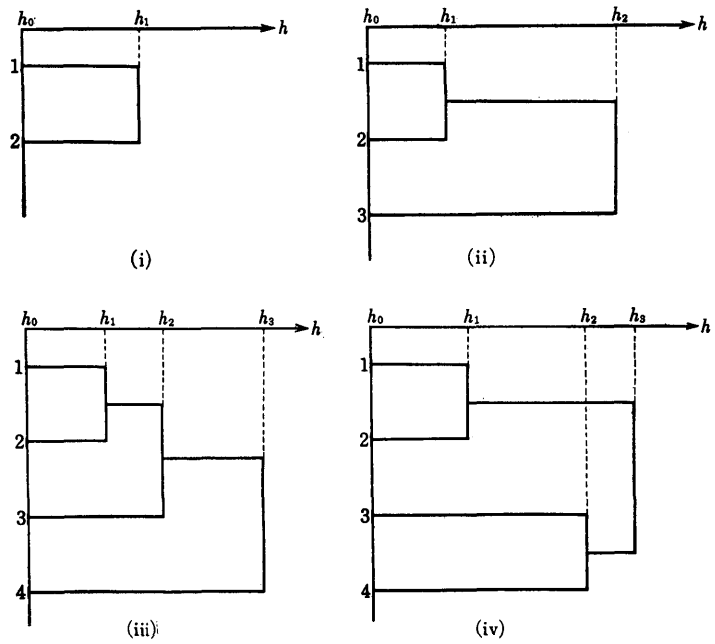



図1 基本系統樹の4つの型

**[1] 階層的構造とクラスター**

ここではとくに，**連結の順序に関して単調な階層的系統樹**を考える[1]．**階層的** (hierarchical) とは，個体あるいはクラスターの逐次連結によりえられる個体のすべての部分集合 $A_1, A_2, A_3, \cdots$ について，集合 $H=\{A_1, A_2, A_3, \cdots\}$ が次の条件を満たす場合をいう．

1) $H=\{A_1, A_2, \cdots\}$ 上の任意の2つの集合 $A_i, A_j$ $(A_i \neq A_j)$ が，$A_i \subset A_j$ 又は $A_j \subset A_i$ 又は $A_i \cap A_j = \phi$

2) $E = \bigcup_i A_i$, $A_i \in H$

を満たすとき，$H$ は階層的であるという．そして $H$ の各要素を**クラスター**と呼ぶ．

たとえば図1−(iv) を例にとれば，$E=\{1, 2, 3, 4\}$, $H=\{\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ である．そして $\{1\}, \{1, 2\}$ 等はクラスターである．

**[2] 水準関数 $h$**

1) 階層構造 $H$ は次の条件を満たす正の実数値関数 $h$ をもつ．これをクラスターの**結合水準の関数**あるいは**水準関数**と呼ぶ．

i) $\forall A, B \in H$, $A \subset B \Rightarrow h(A) < h(B)$

ii) $\forall x \in E$, $h(x) = 0$

2) 水準関数 $h(\cdot)$ のとりうる値を連結の順位に対応させて $h_\alpha (\alpha=0, 1, \cdots, n-1)$ と表わす．ここで，$h_\alpha$ は単調増加であり同位はないものとする (strictly ordered である)[2]．すなわち，

$$h_0 = 0 < h_1 < h_2 < \cdots < h_\alpha < \cdots < h_{n-1}$$

**[3] 系統樹 $<H, h>$**

以上から，系統樹とは階層構造 $H$ と水準関数 $h_\alpha$ によって特徴づけられる平面内の布置図であるから，これを記号 $<H, h>$ により表わす．

したがって，凝集型の階層的数値分類法 (AHC 手法) とは，与えられた個体間に約束した適当な非類似度の行列 $\boldsymbol{D}$ を適当な算法により系統樹 $<H, h>$ に変換する方法である，と要約できる．

AHC 手法は与えられた非類似度がメトリックであるか否かにかかわりなく，比較的広い範囲の $\boldsymbol{D}$ に対して利用できるという柔軟性がある．しかし，系統樹 $<H, h>$ 上の個体間の関係は次に述べるように必らずメトリックな関係として表わすことができる．

(1) $$\delta_{ij} \triangleq h(A_{ij}) = \min\{h(A) \mid A \in H, i \in A, j \in A\}$$

つまり $<H, h>$ 上の関係とはこの $\delta_{ij}$ を個体 $i$ と $j$ との間の非類似度とすることである．このとき明らかに次の関係がある．

i) $\delta_{ii} = 0$ $\quad \forall i \in E$

ii) $\delta_{ij} = \delta_{ji}$ $\quad \forall i, j \in E$

iii) $\delta_{ij} \leq \max\{\delta_{ik}, \delta_{jk}\}$ $\quad \forall i, j, k \in E$

$\delta_{ij}$ がメトリックな非類似度と異なる点は iii) にある．これは通常の三角不等式を，さらに強めた二等辺三角不等式により置きかえたものであるから**強メトリック距離** (ultrametric distance) と呼んでいる．

したがって，階層構造をもつ系統樹 $<H, h>$ 上の関係とは，与えられた $E$ の上に強メトリ

---

1) 算法によっては，必ずしも単調な階層構造をとらない解を与えることがあるが，ここでは上のような場合のみを考える．

2) かりに同位の場合があっても，後述する主座標分析法の結果で，固有値が等根となるだけである．しかし，固有ベクトルが一意に定まらないので，念のために，本報告の議論では同位の場合を除外しておく．

ックな関係 $\boldsymbol{D}^+=(\delta_{ij})$ を埋めこむことである，あるいは任意の非類似度行列 $\boldsymbol{D}=(d_{ij})$ から強メトリックな距離行列 $\boldsymbol{D}^+=(\delta_{ij})$ を生成することである，と言い表わすことができる．

以上の準備のもとに系統樹の性質を調べよう．

### 2.2 強メトリック距離の性質

系統樹の観察から分類結果を分析する際，客観的規準がないと，どうしても恣意的な解釈におち入り易い．そこで $\delta_{ij}$ がメトリック距離であることに注目して，$<H, h>$ を別の情報に置きかえることを考える．換言すると $<H, h>$ 上の $n$ 個の個体の関係を $p$ 次元 $(p\leq n-1)$ ユークリッド空間内に布置して，この変換過程でみられる性質を利用して系統樹の評価を行う方式を考えてみよう．こうした目的に適した方法として容易に連想されるものに，いわゆる主座標分析法（Torgerson-Gower）法；以下 T-G 法と略す）がある．

ではこれらの手法をどのように利用するか，その論旨を要約しておこう．

まず，強メトリック距離の関係にある行列 $\boldsymbol{D}^+$ に T-G 法を適用すれば明らかに $<H, h>$ 上の情報はすべてユークリッド空間内に布置できる．しかもこのとき，$<H, h>$ 上の距離と，T-G 法でえられる固有値および布置行列との間に，ある密接な関係がある．そこでこの特性を利用して系統樹の評価を行うことができる．

T-G 法はメトリックな距離空間を前提にした手法であるために，ノンメトリックな構造のデータに対して適用することが難かしい．しかし，現実にはノンメトリックな非類似度データを扱うことが多い．一方，AHC 手法は，前述のように多様な種類の非類似度行列を扱うことが出来るが，系統樹の解釈が主観的となりやすいという難点がある．そこで $<H, h>$ 上の非類似度は必ずメトリックであるという点に注目して，任意の非類似度行列 $\boldsymbol{D}$ に対して AHC 手法を適用して個体間の関連をメトリックな情報として近似表現した上で，さらにその情報の解釈を客観的に行うために，T-G 法によりユークリッド空間内に布置する．以上の手続きにより2つの手法の不足を互いに補うとともに，その一連の処理過程で現われる規則的な性質を用いた系統樹の評価方式まで考えることが可能となる．

以上から明らかなように T-G 法が重要な役割をはたすので，次にその要点と性質をまとめておく．

〔主座標分析法（Torgerson-Gower 法）〕

(1) $\boldsymbol{D}=(d_{ij})$ から $\boldsymbol{A}=(a_{ij})$ を作る $(i, j=1, 2, \cdots, n)$．ここで $a_{ij}=\dfrac{1}{2}d_{ij}^2$．

(2) $\boldsymbol{A}$ から $\boldsymbol{B}=(b_{ij})$ を作る．

ここで，$b_{ij}=\bar{a}_{i\cdot}+\bar{a}_{\cdot j}-a_{ij}-\bar{a}_{\cdot\cdot}$

$\bar{a}_{i\cdot}=\sum_j a_{ij}/n$，$\bar{a}_{\cdot j}=\sum_i a_{ij}/n$，$\bar{a}_{\cdot\cdot}=\sum_{i,j} a_{ij}/n^2$

(3) $\boldsymbol{B}$ の固有値 $\lambda_\alpha(\alpha=1, 2, \cdots, p;\ \lambda_1>\lambda_2>\cdots>\lambda_p;\ p\leq n-1)$ を求める．

(4) $\boldsymbol{x}_{(\alpha)}'\boldsymbol{x}_{(\alpha)}=\lambda_\alpha$ となるように調整した固有ベクトル $\boldsymbol{x}_{(\alpha)}$（負荷量ベクトル）を列とする布置行列 $\boldsymbol{X}=(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \cdots, \boldsymbol{x}_{(p)})$ を求める．

(5) $\boldsymbol{X}$ の第 $i$ 行ベクトル $\boldsymbol{x}_i=(x_{i1}, x_{i2}, \cdots, x_{ip})'$ が個体 $i$ の座標を与える．

上のことから $\boldsymbol{XX}'=\boldsymbol{B}$，$\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{\Lambda}$ $(\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$，$\lambda_n=0)$ であり，より具体的には次の性質がある．

① $\bar{\boldsymbol{x}}=\sum_i \boldsymbol{x}_i/n=\boldsymbol{0}$（原点が平均ベクトル）

② $\boldsymbol{x}_{(\alpha)}'\boldsymbol{x}_{(\alpha)}=\lambda_\alpha \quad (\alpha=1, 2, \cdots, p;\ p\leq n-1)$

$\boldsymbol{x}_{(\alpha)}'\boldsymbol{x}_{(\beta)}=0 \quad (\alpha\neq\beta$；直交条件）

③ 任意の個体 $i, j$ に対して

$$(\boldsymbol{x}_i-\boldsymbol{x}_j)'(\boldsymbol{x}_i-\boldsymbol{x}_j) = b_{ii}-2b_{ij}+b_{jj}$$
$$= -a_{ii}+2a_{ij}-a_{jj} = 2a_{ij}=d_{ij}^2$$

④ $b_{ij}=\boldsymbol{x}_i'\boldsymbol{x}_j$ とくに，$b_{ii}=\boldsymbol{x}_i'\boldsymbol{x}_i=\bar{a}_{i.}+\bar{a}_{.j}-\bar{a}_{..}$

⑤ ③, ④から，$\sum_{i,j}^{n} d_{ij}^2/2 = n\,\mathrm{tr}\,\boldsymbol{B} = n\,\mathrm{tr}\,\boldsymbol{\Lambda}$

さて，系統樹の構造は，2つの個体（またはクラスター）の逐次連結あるいは逆に逐次二分割から成り立っている．このことから系統樹の基本をなす単位は，図1の (i)～(iii) の3つの型で代表される．$n=4$ の場合には図1-(iv) の型も考えられるので，これを加えてとりあえず4通りの型について考察しよう．図1の4つの型のうち (iv) は (i) の合成，(iii) は (ii) の拡張と考えてよい．そして一般に $n$ が大きい場合でも，その系統樹はすべて (i)～(iv) を基本単位として分解できるので，この4つの型を**基本系統樹**と名づける．

**A.** 図1-(i) の場合

基本系統樹のうち図1-(i) は

$$\boldsymbol{D}^+ = \begin{pmatrix} 0 & h_1 \\ h_1 & 0 \end{pmatrix}$$

となるから T-G 法を用いるまでもなく，$\boldsymbol{X}=(a,-a)'$ $(a>0)$ の形の布置を得る．そして，T-G 法からえられる固有値 $\lambda_1$ と $h_1$, $a$ の間に明らかに次の関係がある．

(2) $$h_1^2 = 2\lambda_1 \quad \text{または} \quad \lambda_1 = 2a^2$$

**B.** 図1-(ii) の場合

次に図1-(ii) を考えよう．ここで $\delta_{11}=\delta_{22}=\delta_{33}=h_0=0$, $\delta_{12}=h_1$, $\delta_{13}=\delta_{23}=h_2$ $(0<h_1<h_2)$ であるから，

$$\boldsymbol{D}^+ = \begin{pmatrix} 0 & h_1 & h_2 \\ h_1 & 0 & h_2 \\ h_2 & h_2 & 0 \end{pmatrix}$$

となる．これに T-G 法を適用すると布置 $\boldsymbol{X}$ は，符号を除いて一意に定まり

(3) $$\boldsymbol{X} = \begin{pmatrix} -a & b \\ -a & -b \\ 2a & 0 \end{pmatrix} \qquad (a,\ b>0)$$

で与えられる．これは2次元ユークリッド空間内に，3点の重心を原点として等辺が他の一辺より長い二等辺三角形 $(\delta_{12}<\delta_{13}=\delta_{23})$ を布置したことに相当する．ここで，$\delta_{21}^2=h_1^2=4b^2$, $\delta_{31}^2=\delta_{32}^2=h_2^2=9a^2+b^2$ である．一方，T-G 法の性質から，2つの固有値は $\lambda_1=6a^2$, $\lambda_2=2b^2$ $(\lambda_1>\lambda_2)$ で与えられる．以上の関係を整理すると次の式がえられる．

(4.1) $$h_1^2 = 2\lambda_2$$

(4.2) $$h_2^2 = \frac{3}{2}\lambda_1 + \frac{1}{2}\lambda_2$$

**C.** 図1-(iii) の場合

続いて図1-(iii) を調べる．$\boldsymbol{D}^+$ の各要素は $\delta_{ii}=h_0=0$ $(i=1,2,3,4)$, $\delta_{12}=h_{12}$, $\delta_{31}=\delta_{32}=h_2$, $\delta_{41}=\delta_{42}=\delta_{43}=h_3$ $(0<h_1<h_2<h_3)$ である．このときの布置行列は，

(5) $$\boldsymbol{X} = \begin{pmatrix} a & d & g \\ a & d & -g \\ b & e & 0 \\ c & f & 0 \end{pmatrix}$$ （ここで各記号は実数，とくに $g>0$）

となる（各面が $\delta_{12}<\delta_{31}=\delta_{32}<\delta_{41}=\delta_{42}=\delta_{43}$ を満たす二等辺三角形であるような四面体を考えればよい）．この場合，図1の (i)，(ii) の拡張となっている．これについてやや詳しく説明を加えておく．

まず T-G 法の性質 ①，② により，

$$2a+b+c=0,\quad 2d+e+f=0,\quad 2ad+be+cf=0 \tag{6}$$

である．同じく ② から，固有値について次の関係がある．

$$\lambda_1=2a^2+b^2+c^2,\quad \lambda_2=2d^2+e^2+f^2,\quad \lambda_3=2g^2 \qquad (\text{ここで } \lambda_1>\lambda_2>\lambda_3) \tag{7}$$

一方，距離の条件から，

$$\begin{cases} h_1^2=\delta_{12}^2=4g^2 \\ h_2^2=\delta_{31}^2=\delta_{32}^2=(a-b)^2+(d-e)^2+g^2 \\ h_3^2=\delta_{41}^2=\delta_{42}^2=(a-c)^2+(d-f)^2+g^2 \\ h_3^2=\delta_{43}^2=(b-c)^2+(e-f)^2 \end{cases} \tag{8}$$

式 (8) の第1式と式 (7) の $\lambda_3=2g^2$ から容易に次式を得る．

$$h_1^2=2\lambda_3 \tag{9}$$

続いて，式 (8) の第2式から式 (6)，(7) を用いて $ab$, $de$ を消去すると，

$$h_2^2=4(a^2+d^2)+2(b^2+e^2)+\frac{\lambda_3}{2}-\frac{1}{2}(\lambda_1+\lambda_2) \tag{10}$$

となる．一方，T-G 法の性質 ④ から，

$$2(a^2+d^2+g^2)=b_{11}=b_{22}=-\frac{\lambda_3}{4}+\frac{1}{8}(2h_2^2+h_3^2)$$

$$2(b^2+e^2)=b_{33}=-\frac{\lambda_3}{4}+\frac{1}{8}(6h_2^2+h_3^2)$$

であるから，$g^2=\lambda_3/2$ を第1式に代入しさらに，$4(a^2+d^2)$，$2(b^2+e^2)$ を求めて式 (10) に代入すると次の関係を得る．

$$2h_2^2+3h_3^2=4(\lambda_2+\lambda_1)+2\lambda_3 \tag{11}$$

こうして，$h^2$ を $\lambda$ で表わすことが出来る．なお式 (11) の関係を求めるために式 (8) の第2式から出立したが，式 (8) の第3，4式を用いても式 (11) と同じ関係式がえられる．したがって T-G 法の性質だけから，$h_2^2$ と $h_3^2$ とをこれ以上分解することは出来ない．

**D.** 図1-(iv) の場合

基本系統樹の最後の一つについて，上と同じ方式で計算を行うと布置 $\boldsymbol{X}$ として次の関係を得る（この場合には各面が，$\delta_{21}=h_1<\delta_{43}=h_2<\delta_{31}=\delta_{32}=\delta_{41}=\delta_{42}=h_3$ を満たす二等辺三角形となるような四面体を考えればよい）．

$$\boldsymbol{X}=\begin{pmatrix} -a & 0 & c \\ -a & 0 & -c \\ a & -b & 0 \\ a & b & 0 \end{pmatrix} \qquad (a,\ b,\ c>0) \tag{12}$$

このとき水準と固有値との対応は次のようになる．

(13.1) $h_1^2 = 2\lambda_3$

(13.2) $h_2^2 = 2\lambda_2$

(13.3) $h_3^2 = \lambda_1 + \frac{1}{2}(\lambda_2+\lambda_3)$

以上で基本系統樹について，水準関数と T-G 法による固有値，負荷量ベクトルとの関係が明らかになった．

ところで，式 (4.2)，(11)，(13.3) はそれぞれ次のように書きかえることが出来る．

(14.1) $h_1^2+2h_2^2 = 3(\lambda_2+\lambda_1)$ （式 (4.2)）

(14.2) $h_1^2+2h_2^2+3h_2^2 = 4(\lambda_3+\lambda_2+\lambda_1)$ （式 (11)）

(14.3) $h_1^2+h_2^2+4h_3^2 = 4(\lambda_3+\lambda_2+\lambda_1)$ （式 (13.3)）

さて，以上の結果を観察すると次の性質がある．

(P1) $<H, h>$ 上の距離行列 $\boldsymbol{D}^+$ に T-G 法を用いて得られる布置行列 $\boldsymbol{X}$ は必らず $(n-1)$ 次元ユークリッド空間内に布置できる（これについては，Lefkovitch [5]，Holman [4] を引用すれば明らかである）．

(P2) 階層構造 $H$ 上の入れ子の関係が水準の平方 $h_\alpha^2$ に関して成り立つ．あるいは $h_\alpha^2$ の一次結合で系統樹は合成できる．これは上で求めた水準と固有値との各関係式をみれば明らかである．たとえば，図1の基本系統樹はそれぞれ次のように解釈できる．

a) 図1−(ii) は (i) の大きさが2のクラスター $\{1,2\}$ に，大きさが1のクラスター（個体）$\{3\}$ を追加してえられる．したがって (i) でえた $h_1^2$ に，クラスター・サイズの重み係数で調整した $(1\times2)h_2^2$ を加えて式 (14.1) の左辺となる．

b) (ii) でえられる大きさ3のクラスター $\{1,2,3\}$ に対して $\{4\}$ を加えて (iii) となるので，式 (14.2) がえられる．

c) 図1−(iv) を考える．図1−(i) の単位を2個合成する．これが $h_1^2+h_2^2$ に相当する．次はクラスター・サイズがそれぞれ2である2つのクラスター $\{1,2\}$，$\{3,4\}$ の連結により，重み係数を $2\times2=4$ と求めて，$4h_3^2$ とする．これらを合成して，式 (14.3) を得る．

ここで，性質 (P2) あるいは式 (2) および (14.1)〜(14.3) の関係は，簡単に，一般の $n$ に対して成り立つことがわかる．これは次の2つの規則として表わすことができる．

**＜規則 I＞**

すべての系統樹 $<H, h>$ は基本系統樹の合成により構成されている，あるいは基本系統樹の単位に分解できる．したがって個体数が $n$ である任意の $<H, h>$ から求めた $\boldsymbol{D}^+$ に T-G 法を適用して得られる $(n-1)$ 個の固有値 $\lambda_{n-\alpha}$ $(\alpha=1, \cdots, n-1)$ について，その総和 $\sum_{\alpha=1}^{n-1}\lambda_{n-\alpha}$ は必らず水準の平方の一次結合により表現できる．すなわち，

(15) $$\sum_{\alpha=1}^{n-1} w_\alpha h_\alpha^2 = n\sum_{\alpha=1}^{n-1}\lambda_{n-\alpha} \qquad (n \geq 2)$$

（ここで，$w_\alpha$ は正の重み係数；$h_1<h_2<\cdots<h_{n-1}$；$\lambda_{n-1}<\lambda_{n-2}<\cdots<\lambda_1$）

この規則の証明は T-G 法の性質 ⑤ を用いれば明らかである．すなわち ⑤ から，

(16) $$\frac{1}{2}\sum_{i,j}^{n}\delta_{ij}^2 = n\sum_{i=1}^{n-1}\lambda_i$$

であるが，$\delta_{ij}^2$ は必らずいずれかの $h_\alpha^2$ に等値である．とくに最終の連結の水準 $h_{n-1}$ を考える．

この水準で結合する2つのクラスターを $C_1, C_2$ $(C_1, C_2 \in H)$ とかき，そのクラスター・サイズをそれぞれ $n_1$, $n_2$ $(n_1+n_2=n)$ とおけば，式 (16) の左辺は必ず $2n_1n_2h_{n-1}^2$ の項を含まねばならない ($\boldsymbol{D}^+$ 中に $2n_1n_2$ 個の $h_{n-1}^2$ がある)．次に，水準 $h_{n-2}$ を考えると，$h_{n-2}=\max\{h(C_1), h(C_2)\}$ であるから，かりに $h_{n-2}=h(C_1)$ として $C_1$ を $n_{11}, n_{12}$ 個 $(n_{11}+n_{12}=n_1)$ に，水準 $h_{n-2}$ で分割することを考えると $2n_{11}n_{12}h_{n-2}^2$ の項を含まねばならない．以下同じことを繰り返せば，すなわち $\boldsymbol{D}^+$ の要素のうち等値であるものの個数を調べ，これを対応する水準の平方の係数 $w_\alpha$ とする操作を繰り返すと，式 (16) の左辺は必ず $h_\alpha^2$ の一次結合となる．これを逆順に考えて，水準 $h_1^2, h_2^2, \cdots$ の順に基本系統樹を単位として合成しても同じ結果がえられる．以上から式 (15) の成り立つことがわかる．

いま <規則 I> の特別な場合として，図1の (i)～(iii) の拡張を考えよう．すなわち，2個に1個が追加，3個に1個が追加，…と逐次1個の個体を連結し増殖させる，いわゆる“連鎖性のある系統樹”の場合である．このとき次の関係が成り立つ．

**<規則 II>**

連鎖性のある系統樹 $<H, h>$ について，その水準関数 $h_\alpha$ と $<H, h>$ 上の距離 $\boldsymbol{D}^+$ に T-G 法を適用してえられる固有値 $\lambda_{n-\alpha}(\alpha=1, 2, \cdots, n-1)$ との間に次の関係が成立する．

$$\sum_{\alpha=1}^{n-1} \alpha h_\alpha^2 = n \sum_{\alpha=1}^{n-1} \lambda_{n-\alpha} \qquad (n \geq 2) \tag{17}$$

さてここで，(P1)，(P2) に加えてさらに以下の性質が観察される．

(P3) 水準関数と固有値との間には，順位に関して形式的に逆順の対応がある．

a) とくに，2個の個体が初めて連結する場合を考えると，そのときの水準 $h_\alpha$ は固有値 $\lambda_{n-\alpha}$ と次の関係にある．

$$h_\alpha^2 = 2\lambda_{n-\alpha} \tag{18}$$

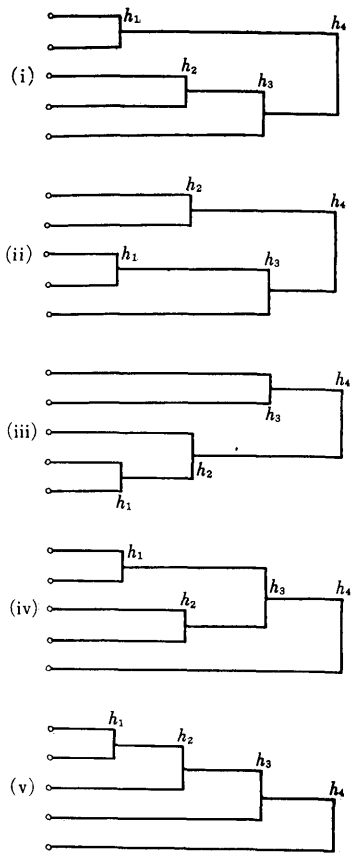
b) 一般に，ある水準 $h_\alpha$ で連結するクラスターのうち少くとも一方が，クラスター・サイズが2以上のクラスターの場合，$h_\alpha^2$ は必ずしも $\lambda_{n-\alpha}$ だけで表わすことは出来ないが，性質の (P2) を考慮すると形式的に $h_\alpha^2$ は $\lambda_{n-\alpha}$ と対応する．

(P4) 布置行列 $\boldsymbol{X}$ の要素の配列がクラスターの分割 (あるいは逆に連結) に対応した規則性をもつ．

たとえば式 (3)，(5)，(12) の $\boldsymbol{X}$ を観察すると負荷量ベクトルの符号と数値から，系統樹上の連結または分割の情報を読みとることができる．式 (3) では $h_1^2$ つまり $\lambda_2$ に対応するベクトル $\boldsymbol{x}_{(2)}$ で {1}，{2} と分割される．また $\boldsymbol{x}_{(1)}'=(-a, -a, 2a)$ であるからここでも符号と数値から {1, 2} と {3} に分割される．式 (5)，(12) についても全く類似の関係がみられる．

(P5) (P3) との関連で，水準 $h_\alpha$ での結合は形式的に負荷量ベクトル $\boldsymbol{x}_{(n-\alpha)}$ に対応すると考えてよい．しかも，一旦結合した2個の個体については，そのときの結合水準より値の大きい水準に対応する負荷量ベクトルの要素はすべて同値となる．たとえば式 (5) で $h_1^2$ すなわち $\lambda_3$ に相当するベクトル $\boldsymbol{x}_{(3)}$ を除く (つまり個体1, 2を併合する) と，$\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}$ の要素について個体1と2を区別することは出来ない．

以上で基本系統樹の上からえられる強メトリック距離あるいは水準関数 $h_\alpha$ と T-G 法からえられる固有値 $\lambda_{n-\alpha}$ 及び負荷量ベクトル (あるいは布置 $\boldsymbol{X}$) との対応を知ることが出来た．T-G 法とは距離 (平方ユークリッド距離) の関係を分散・共分散行列に置きかえた上で主成分分析により情報の縮約を行う，あるいは個体をユークリッド空間内に布置して観察するという点に特徴があるが，上の事実も，$h_\alpha^2$ と $\lambda_{n-\alpha}$ との対応において，この特性を反映したものといえよう．




$$\boldsymbol{X}^{(1)}=\begin{pmatrix} a & d & 0 & h \\ a & d & 0 & -h \\ -b & e & g & 0 \\ -b & e & -g & 0 \\ -c & -f & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{X}^{(2)}=\begin{pmatrix} -a & d & g & 0 \\ -a & d & -g & 0 \\ b & e & 0 & -h \\ b & e & 0 & h \\ c & -f & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{X}^{(3)}=\begin{pmatrix} a & d & e & 0 \\ a & -d & e & 0 \\ -b & 0 & -f & 0 \\ -c & 0 & g & -h \\ -c & 0 & g & h \end{pmatrix}$$

$$\boldsymbol{X}^{(4)}=\begin{pmatrix} a & -d & 0 & h \\ a & -d & 0 & -h \\ b & e & -g & 0 \\ b & e & g & 0 \\ -c & f & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{X}^{(5)}=\begin{pmatrix} a & -d & e & h \\ a & -d & e & -h \\ b & -f & -u & 0 \\ c & g & v & 0 \\ -x & -y & w & 0 \end{pmatrix}$$

図2　$n=5$ に対する系統樹と得られる布置行列のパターン

基本系統樹で調べた関係が確かに成立することを個体数が 5 ($n=5$) の場合と，いくつかの数値例により検証しておこう．

$n=5$ の場合，水準 $h_\alpha$ の連結順位に対して相異なる階層構造を生成する系統樹の組み合わせは図2の5通りがすべてである．かりに個体番号の付与の順が異なっていても，それは系統樹からえられる $\boldsymbol{D}^+$ の行と列の適当な入れ換えにより必ず図2の (i)～(v) のいずれかに対応する．

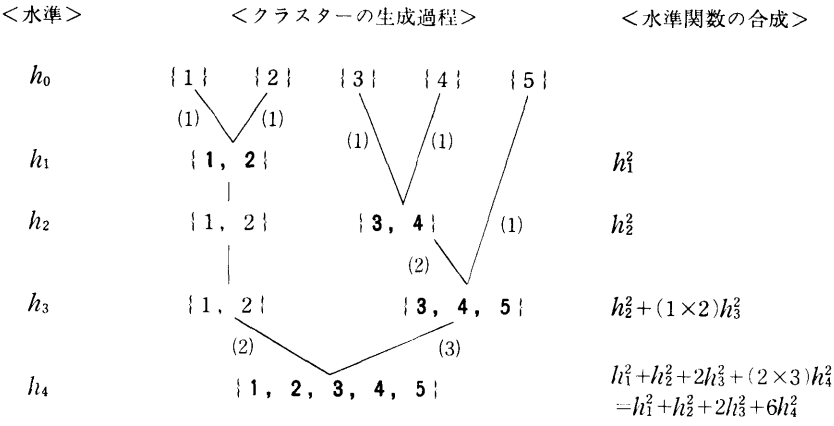
さて，(i)～(v) に対してそれぞれ行列 $\boldsymbol{D}^+$ を作り，T-G 法を用いると図2の右側に掲げた布置行列がえられる．各組を識別するために添字をつけて $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(5)}$ と書くことにする．まず $\boldsymbol{X}^{(1)}$ について前述の方式で計算を行うと，水準関数と固有値について次の関係を得る．

$$\text{(19)}\qquad \begin{cases} h_1^2 = 2\lambda_4 \\ h_2^2 = 2\lambda_3 \\ 2h_3^2+6h_4^2 = 5(\lambda_2+\lambda_1)+3(\lambda_3+\lambda_4) \end{cases}$$

第3式は＜規則 I＞から次のように書き換えられる．

$$\text{(19)}'\qquad h_1^2+h_2^2+2h_3^2+6h_4^2 = 5(\lambda_4+\lambda_3+\lambda_2+\lambda_1)$$

ここで性質 (P2) を用いると各連結の水準と生成されるクラスターとの間に確かに次の対応がある．




(*) ( ) 内はクラスター・サイズ

こうして式 (19)′ の左辺を水準関数の合成により求めることが出来る．ここで $n=5$ であるから最終式を $n\sum_{\alpha=1}^{n}\lambda_{n-\alpha}=5(\lambda_4+\lambda_3+\lambda_2+\lambda_1)$ に等価とおけばよい．

また他の性質についても $\boldsymbol{X}^{(1)}$ の布置のパターンからその成立は明らかである (5つの布置のそれぞれについて，クラスター化の特徴を強調するために仕切り線を入れてみた．これにより系統樹との対応がはっきりする)．

なお，基本系統樹の場合も含めて，$n\leq5$ に対する9通りの系統樹について水準関数と固有値との対応を求め表1として一覧にした．

簡単な数値例により以上の関係を確かめよう．

＜数値例 1＞

図1−(iii) に対応する例 ($n=4$ の連鎖性のある系統樹) を調べる．

$E=\{1,2,3,4\}$, $H=\{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{1,2,3\},\{1,2,3,4\}\}$, $h_0=0$, $h_1=0.5$, $h_2=1.7$, $h_3=3.0$ と与える．このとき，$\lambda_1=6.2696$, $\lambda_2=1.8629$, $\lambda_3=0.1250$

$$\boldsymbol{X}=\begin{pmatrix} 0.7934 & -0.5279 & 0.25 \\ 0.7934 & -0.5279 & -0.25 \\ 0.5763 & 1.1395 & 0 \\ -2.1630 & -0.0837 & 0 \end{pmatrix}$$

となる．まず $h_1^2=2\lambda_3$, $2h_2^2+3h_3^2=4(\lambda_2+\lambda_1)+2\lambda_3$ は明らかである．勿論，式 (14.2) についても，左辺，右辺とも 33.03 となり一致する．

＜数値例 2＞

$E=\{1,2,3,4,5\}$; $H=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{4,5\},\{3,4,5\},\{1,2\},\{1,2,3,4,5\}\}$; $h_0=0$, $h_1=0.1$, $h_2=0.8$, $h_3=0.9$, $h_4=1.0$ と与える．これは図2−(iii) に相当する．計算結果は次の通りである．

$$\lambda_1=0.8080,\quad \lambda_2=0.4050,\quad \lambda_3=0.4020,\quad \lambda_4=0.0050$$

表1　水準関係と固有値の関係

| 個体数 ($n$) | 関　　係　　式 | 対応する系統樹 |
|---|---|---|
| 2 | $h_1^2 = 2\lambda_1$ | 図 1-(i) |
| 3 | $h_1^2 = 2\lambda_2$<br>$2h_2^2 = 3\lambda_1 + \lambda_2$<br>$h_1^2 + 2h_2^2 = 3(\lambda_1 + \lambda_2)$ | 図 1-(ii) |
| 4 | $h_1^2 = 2\lambda_3$<br>$2h_2^2 + 3h_3^2 = 4(\lambda_2 + \lambda_1) + 2\lambda_3$<br>$h_1^2 + 2h_2^2 + 3h_3^2 = 4(\lambda_3 + \lambda_2 + \lambda_1)$ | 図 1-(iii) |
| | $h_1^2 = 2\lambda_3$<br>$h_2^2 = 2\lambda_2$<br>$2h_3^2 = 2\lambda_1 + \lambda_2 + \lambda_3$<br>$h_1^2 + h_2^2 + 4h_3^2 = 4(\lambda_3 + \lambda_2 + \lambda_1)$ | 図 1-(iv) |
| 5 | $h_1^2 = 2\lambda_4$<br>$h_2^2 = 2\lambda_3$<br>$2h_3^2 + 6h_4^2 = 5(\lambda_2 + \lambda_1) + 3(\lambda_3 + \lambda_4)$<br>$h_1^2 + h_2^2 + 2h_3^2 + 6h_4^2 = 5(\lambda_4 + \lambda_3 + \lambda_2 + \lambda_1)$ | 図 2-(i), (ii) |
| | $h_1^2 = 2\lambda_4$<br>$h_3^2 = 2\lambda_2$<br>$2h_2^2 + 6h_4^2 = 5(\lambda_3 + \lambda_1) + 3(\lambda_2 + \lambda_4)$<br>$h_1^2 + h_3^2 + 2h_2^2 + 6h_4^2 = 5(\lambda_4 + \lambda_2 + \lambda_3 + \lambda_1)$ | 図 2-(iii) |
| | $h_1^2 = 2\lambda_4$<br>$h_2^2 = 2\lambda_3$<br>$4h_3^2 + 4h_4^2 = 5(\lambda_2 + \lambda_1) + 3(\lambda_3 + \lambda_4)$<br>$h_1^2 + h_2^2 + 4h_3^2 + 4h_4^2 = 5(\lambda_4 + \lambda_3 + \lambda_2 + \lambda_1)$ | 図 2-(iv) |
| | $h_1^2 = 2\lambda_4$<br>$h_1^2 + 2h_2^2 + 3h_3^2 + 4h_4^2 = 5(\lambda_4 + \lambda_3 + \lambda_2 + \lambda_1)$ | 図 2-(v) |

*)　表の中で，$h_1 < h_2 < h_3 < h_4$，$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$ である．

$$\boldsymbol{X} = \begin{pmatrix} 0.4782 & 0.45 & 0.0827 & 0 \\ 0.4782 & -0.45 & 0.0827 & 0 \\ -0.1440 & 0 & -0.5579 & 0 \\ -0.4062 & 0 & 0.1963 & 0.05 \\ -0.4062 & 0 & 0.1963 & -0.05 \end{pmatrix}$$

$h_1^2 = 2\lambda_4$，$h_2^2 = 2\lambda_2$，$h_3^2 + 6h_4^2 = 5(\lambda_4 + \lambda_3) + 2(\lambda_2 + \lambda_1)$ となり (P3) の成立は明らかである．さらに ＜規則 I＞により $h_1^2 + 2h_2^2 + h_3^2 + 6h_4^2 = 5(\lambda_4 + \lambda_3 + \lambda_2 + \lambda_1)$ となるが，左辺，右辺とも 8.1 となり一致が確かめられる．

またいずれの例も，$\boldsymbol{X}$ の要素の配置が系統樹上のクラスターの分割（あるいは連結順位）に対応している．その他の性質についてもその成立は明らかであろう．

ここで要点をまとめると次の通りである．

1°)　系統樹上の $n$ 個の個体間の関係は強メトリック距離であるから，情報を失うことなく，T-G 法により必ず $(n-1)$ 次元ユークリッド空間内に布置できる．

2°)　このとき，系統樹の水準関数と T-G 法でえられる固有値との間に規則性がある．

3°)　任意の系統樹は，必ず基本系統樹に分解することができる．しかもこの分解は水準関数

の合成として表現できる．

4°)　クラスター化の過程，あるいはクラスターの構成は布置行列の要素の配置に現われる．

## 3.　分類系統樹の比較

次に前節までの結果を利用して系統樹の比較・評価を行う方式，すなわち以下の問題の解決に手掛りを与えるような評価手順を考える．

1)　$<H, h>$ 上に生成されるクラスターの強度の評価

2)　$\boldsymbol{D}^+$ と $\boldsymbol{D}$ の適合性の評価，すなわち系統樹 $<H, h>$ 上の距離が，$\boldsymbol{D}$ のそれをどの程度近似しているかをあらわす方法

3)　複数の系統樹の比較方式

ここで1) はクラスター化の程度を測る規準，いいかえるとクラスター数の目安を与える規準を工夫することである．また2)，3) は，系統樹間の比較あるいはT-G法でえられる布置行列間の比較の方式を考えることに相当する．これらの問題に対して，前節までに調べた系統樹の性質を応用した評価方式を検討することが本節の目的である．

### 3.1　クラスター化の強度の尺度

系統樹を比較するとき，クラスター化の程度を十分考慮することが重要である．これは，図3－(i)，(ii)の2つの系統樹を考えてみれば明らかであろう．この2つの系統樹は連結の順位がまったく同じであるから，えられる階層構造もまったく同等である（いわゆる**順位に関して同型の系統樹**である）．したがってどの水準で切断してもえられるクラスター構成は同じである．しかし図から明らかなように水準関数の値が異なるうえ，(i) は (ii) にくらべてクラスターのまとまりがよい（これを**クラスター強度が高い**と呼ぼう）と解釈するのが妥当であろう．このようなクラスター化の強度を知り系統樹をくらべる指標として，次の寄与率あるいは累積寄与率を利用する．

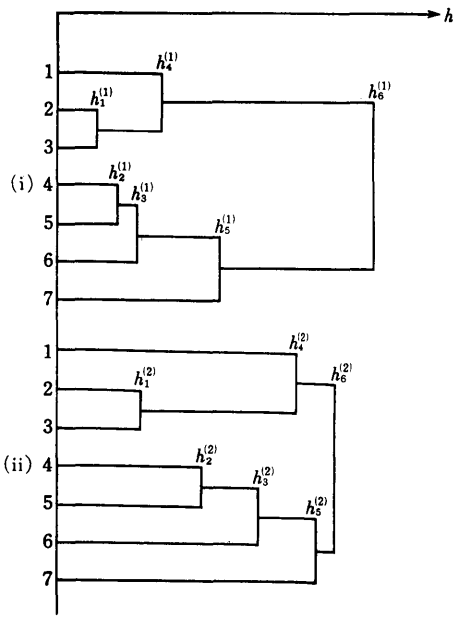



図3　連結順位に関して同型の2つの系統樹

(20)

$$\begin{cases} \text{寄　与　率}: \eta_\alpha = \left(\lambda_\alpha \Big/ \sum_{\alpha=1}^{n-1} \lambda_\alpha\right) \times 100\,(\%) \\ \text{累積寄与率}: \sum_{\alpha=1}^{k} \eta_\alpha = \left(\sum_{\alpha=1}^{k} \lambda_\alpha \Big/ \sum_{\alpha=1}^{n-1} \lambda_\alpha\right) \times 100\,(\%) \end{cases}$$

ここで，$\lambda_\alpha$ $(\alpha=1, 2, \cdots, n-1;\ \lambda_1>\lambda_2>\cdots>\lambda_{n-1})$ はT-G法でえられる固有値である．また $k$ は，第 $(n-k)$ 番目の連結 $h_{n-k}$ に対応する．つまり，クラスター数が $(k+1)$ の場合に相当する．

T-G法が，与えられた個体の布置の重心を原点とする主成分分析であるという事実と，水準関数の平方が固有値と逆順に対応するという性質からこの指標を用いることは自然である．またその有効性は次の例をみれば十分理解できる．

＜数値例 3＞

図3の (i)，(ii) に対して，次のように水準関数を与えてみる．

(i)　に対して：　$h_1^{(1)}=0.1,\ h_2^{(1)}=0.15,\ h_3^{(1)}=0.2,\ h_4^{(1)}=0.25,$
$h_5^{(1)}=0.4,\ h_6^{(1)}=0.8$

(ii)　に対して：　$h_1^{(2)}=0.2,\ h_2^{(2)}=0.35,\ h_3^{(2)}=0.5,\ h_4^{(2)}=0.6,$
$h_5^{(2)}=0.65,\ h_6^{(2)}=0.7$

ここで (i), (ii) を区別するために水準関数に上添字をつけてある．

このとき，T-G 法でえられる固有値から寄与率，累積寄与率を求め一覧にすると表 2 を得る．かりに累積寄与率が 80% を満たすことを目安に，水準の位置すなわちクラスター数を見積ると，(i) では $\lambda_1$, (ii) では $\lambda_3$ までとなるので，クラスター数は，(i) が 2 群，(ii) が 4 群と求められる．すなわち，

(i)　に対して，　$\{\{1,2,3\},\ \{4,5,6,7\}\}$

(ii)　に対して，　$\{\{1\},\{2,3\},\ \{4,5,6\},\{7\}\}$

とクラスター化することが適当であるとなるが，これは図を観察してえられる実感ともよく合っている．このように，寄与率を用いて系統樹の形相を計量化することができる．なお念のために，この例について＜規則 I＞が成立することを確かめておこう．基本系統樹の合成から得られる関係式は，

$$(h_1^2+2h_4^2)+(h_2^2+2h_3^2+3h_5^2)+(3\times4)h_6^2=7\sum_{\alpha=1}^{6}\lambda_{7-\alpha}$$

である．上の (i), (ii) で与えた水準関数と，求めた固有値について数値を求めるとこの等式が成立していることが確かめられる．さらに，$h_1^2=2\lambda_6$, $h_2^2=2\lambda_5$ の成立は容易にわかる．

表 2　〈数値例 3〉の結果

| $\alpha$ | (i) の場合 | | | (ii) の場合 | | |
|---|---|---|---|---|---|---|
| | $\lambda_\alpha$ | 寄与率 | 累積寄与率 | $\lambda_\alpha$ | 寄与率 | 累積寄与率 |
| 1 | 1.01054 | 84.23 | 83.23 | 0.51982 | 42.66 | 42.66 |
| 2 | 0.11009 | 9.18 | 93.41 | 0.25650 | 21.05 | 63.71 |
| 3 | $0.39909\times10^{-1}$ | 3.33 | 96.74 | 0.21696 | 17.80 | 81.51 |
| 4 | $0.22866\times10^{-1}$ | 1.90 | 98.64 | 0.14407 | 11.82 | 93.33 |
| 5 | $0.11250\times10^{-1}$ | 0.94 | 99.58 | $0.61252\times10^{-1}$ | 5.03 | 98.36 |
| 6 | $0.5\times10^{-2}$ | 0.42 | 100.00 | $0.2\times10^{-1}$ | 1.64 | 100.00 |

### 3.2　複数の系統樹の比較

ここでは，本節の初めに挙げた問題のうち 2), 3) について議論する．すなわち複数の系統樹の情報を同時に比較し，系統樹間の類似性や個体間の類似性を評価する方式を検討する．この問題は次のように考えれば実は今までの議論の拡張であるから，前節までの結果を応用してこれらを解決する手掛りを得ることができる．

いま，系統樹がその階層構造と連結の順位（水準関数）の大きさにおいて良く類似したものであれば，T-G 法からえられる布置行列の配置のパターンも類似しているはずである．すなわち複数の系統樹の比較は，そのまま複数の布置行列の比較の問題に置き換えられる．こうした分析に適した方法として容易に連想されるものにプロクラスタス法がある．ここでは多数個の布置行列の整合を行う方法として Gower により提案された，**一般化プロクラスタス法** (Generalized Procrustes Analysis；以下 G-P 法と略す) を用いることにする．

G-P 法は，与えられた複数組の行列に対し，適当な変換（回転，反転，尺度調整（伸縮））

を行うとともに，行列間の整合を最小二乗法を利用して達成するというものである．結果として，与えた行列の平均的布置行列（いわゆるコンセンサス布置行列）$\boldsymbol{Y}$ が得られ，それと同時に各行列がこの $\boldsymbol{Y}$ に対してどの程度適合しているか，あるいは各点（個体）の $\boldsymbol{Y}$ に対する寄与はどの程度であるのか，といった情報を一種の分散分析表の形として求めた上，残差分析により適合性を量的に評価できるという利点を備えている（詳細は文献 [1], [3], [8], [9] を参照）．

G-P 法による系統樹の比較の手順はきわめて簡単で次のように行えばよい．

(手順 1) 与えられた非類似度行列 $\boldsymbol{D}=(d_{ij})$ $(i, j=1, 2, \cdots, n)$ に複数の AHC 手法 $M_l$ $(l=1, 2, \cdots, m)$ を適用して $m$ 組の系統樹 $<H^{(l)}, h^{(l)}>$ $(l=1, 2, \cdots, m)$ を作る．

(手順 2) $<H^{(l)}, h^{(l)}>$ $(l=1, 2, \cdots, m)$ から $m$ 組の $\boldsymbol{D}_l^+$ $(l=1, 2, \cdots, m)$ を作り T-G 法によりこれを $m$ 組の布置行列 $\boldsymbol{X}_l$ $(l=1, 2, \cdots, m)$ に変換する．

(手順 3) $\boldsymbol{X}_l$ $(l=1, 2, \cdots, m)$ を G-P 法により整合し，コンセンサス行列 $\boldsymbol{Y}$ を求める．同時に，$m$ 組の布置および $n$ 個の個体についての分散分析表を作成する．これに基づいて系統樹および個体に関する類似性の評価分析を行う．

$\boldsymbol{D}$ からえられた $m$ 組の系統樹の比較は上の手順で行えばよい．さらに，$m$ 組の系統樹が $\boldsymbol{D}$（すなわちもとの個体間の関連性）に対して，それぞれどの程度適合しているかを知りたい，つまり“解の適合性”の評価分析を行う場合には，上の方式を拡張して次の手順を加えればよい．

(手順 1′) 与えられた非類似度行列 $\boldsymbol{D}=(d_{ij})$ に対して，T-G 法を適用する．ここで $d_{ij}$ はメトリックとは限らないので固有値に負の根が現われることがある．このときには正の根とそれに対応する負荷量ベクトルを求める．あるいは寄与率を目安に適当な次元数までの固有値に対応するベクトルを求める（このとき寄与率は，式 (18) の分母を $\sum_{\alpha=1}^{n-1}|\lambda_{n-\alpha}|$ とする．[6] を参照）．こうして $\boldsymbol{D}$ に対する布置行列 $\boldsymbol{X}_0$ を作る．この $\boldsymbol{X}_0$ を上の手順の $m$ 組の $\boldsymbol{X}_l$ に加えて，$(m+1)$ 組の布置行列とし，これに上の（手順 3）を行えばよい．

ところで上に述べたように $\boldsymbol{X}_0$ の次元数は系統樹から出立した場合と異なり必ずしも $(n-1)$ 次元とはならない．しかし，G-P 法によれば次元数の異なる行列の整合の場合には行列のランクは $\min_l\{\mathrm{rank}(\boldsymbol{X}_l)\}$ $(l=0, 1, \cdots, m)$ として調整される．あるいは，あらかじめ次元数 $q$ を指定した上でこれにそろえて $\boldsymbol{X}_l$ $(l=1, 2, \cdots, m)$ も $q$ 次元ベクトルまでを採用するという方式も考えられる．このことは第 $q$ 主成分までの負荷量ベクトルを用いることであるから，各系統樹について，次元数 $q$ に相当する水準で切断を行ったことに相当する．したがって，切断の位置より数値の小さい水準において連結した個体は合併されて 1 つのクラスターと考えることになる．これは，系統樹の上の特徴的な連結のみを強調し，類似した個体またはクラスターは併合して考えることに相当する．

上の方式を数値例により確かめよう．

＜数値例 4＞

図 4 の 4 つの系統樹の比較を考える．図を観察すると (i), (ii) が類似し (iii) はそれらと同型ではあるが個体の連結順位（すなわち階層構造）が異なるので，個体番号まで考慮すると，類似しているとはいえない．また (iv) は個体番号の並びは (i), (ii) と同じであっても，連結のパターンが異なる．

まず (i)～(iv) に対応する 4 組の $\boldsymbol{D}_l^+$ $(l=1, 2, 3, 4)$ を作る．続いて T-G 法により 4 組の布置行列 $\boldsymbol{X}_l$ $(l=1, 2, 3, 4)$ に変換する．これに G-P 法を適用した結果，次のコンセンサス行列を得た．

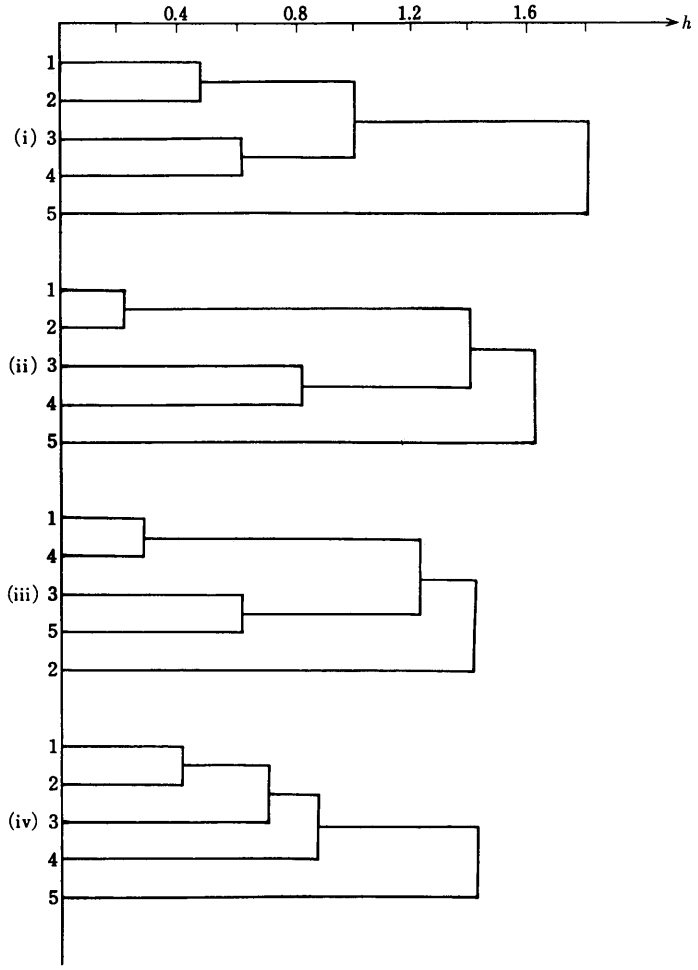
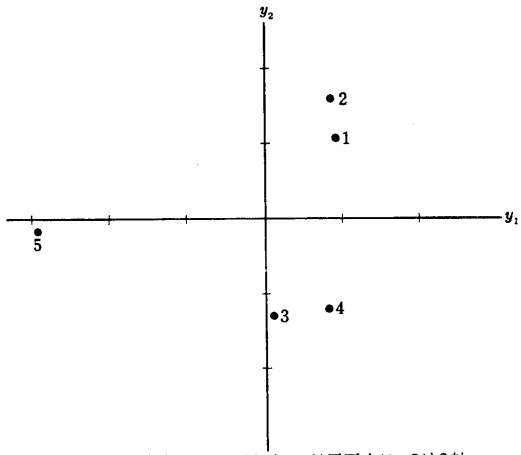



図4　4つの系統樹の比較

$$\boldsymbol{Y}=\begin{array}{c}\begin{matrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 & \boldsymbol{y}_4\end{matrix}\\ \begin{pmatrix} 0.1794 & 0.2063 & -0.0372 & 0.1425 \\ 0.1677 & 0.3195 & 0.0370 & -0.1540 \\ 0.0878 & -0.2506 & -0.2398 & -0.0122 \\ 0.1611 & -0.2478 & -0.2504 & 0.0235 \\ -0.5961 & -0.0274 & 0.0108 & 0.0002 \end{pmatrix}\end{array}$$

$\boldsymbol{Y}$ の2次元までを使って散布図を作ると図5となる．これをみると確かに系統樹の上の個体間の関係を良く反映し，しかもクラスター化の状態をうまく表わしている．また分散分析表は表3の (a)，(b) で与えられる．ここで (a) は個体についての分散分析表，(b) は布置行列の組 (つまり系統樹の組) についての残差分析の表である．(a) の ② 欄から個体番号 1, 5 はコンセンサスに対して適合が良いとはいえない．しかし① 欄をみると番号5の個体のコンセンサスに対する寄与の程度は高い (40.2%) ので系統樹を特徴づける重要な点 (個体) であることがわかる．これに対し個体番号1はそれほどでもない．一方 (b) をみると，(iii) の系統樹の適合がき




(*) 1,2,5はほぼ1,2軸平面上に，3は3軸の正，4は3軸の負の側に布置される

図5

わだって悪い（残差が0.2679）が，残りの3つの系統樹は類似しているようである．その他群内平方和に関する情報((a)の③欄，(b)の②欄）からも上の傾向がみえる．

こうして，複数組の系統樹の比較分析を効果的に行うことができる．

## 4. むすび

AHC 手法の解すなわち系統樹の解釈や評価は，これまでこれといった極め手となる方法がなく，往々にして恣意的判断となり易いという難点があった．ここではこうした問題に客観的な手掛りを与える一つの発見的方法を提案した．

表3 G-P 法の結果

(a) 個体についての分散分析表

| 個体番号 | ①コンセンサスへの寄与 (%) | ②残差 (%) | ③計（個体別群内平方和） | ④=①/③×100 (%) |
|---|---|---|---|---|
| 1 | 0.38568 (10.9) | **0.12806** (28.2) | 0.51374 | 75.1 |
| 2 | 0.62115 (17.5) | 0.05836 (12.9) | 0.67951 | 91.4 |
| 3 | 0.51266 (14.5) | 0.09801 (21.6) | 0.61068 | 84.0 |
| 4 | 0.60251 (17.0) | 0.06687 (14.7) | 0.66938 | 90.0 |
| 5 | **1.42460** (**40.2**) | **0.10207** (22.5) | **1.52669** | 93.3 |
| | 3.5466 (100) | 0.45338 (100) | 4.0000 | 88.7 |

(b) 系統樹についての残差分析表

| 系統樹の番号 | ①残差 | ②計（組別群内平方和） |
|---|---|---|
| (i) | 0.0621 | 1.0663 |
| (ii) | 0.0728 | 1.0525 |
| (iii) | **0.2679** | **0.8002** |
| (iv) | 0.0506 | 1.0810 |
| | 0.4534 | 4.0000 |

なお，解の適合性，すなわち $\boldsymbol{D}^+$ と $\boldsymbol{D}$ との適合性の評価の問題についてはその方式の提案に留めたが，いくつかの数値実験では良い結果を得ているので改めて報告したい．また，一般に個体数が $n$ の場合，系統樹の水準関数を“個別に”分解してこれとT-G法で得た固有値との対応関係を調べることや，G-P法の改良（重みつき最小二乗法とすること）などが今後の課題として残されている．

最後に，適切な助言を与えて下さった水野欽司氏，ならびに査読者に心から謝意を表します

## 参　考　文　献

[1] Berge, M.F.T. (1977). Orthogonal Procrustes rotation for two or more matrices, *Psychometrika*, **42**, 267–276.
[2] Gower, J.C. (1971). Statistical methods of comparing different multivariate analysis of the same data, *Mathematics in the Archaeological and Historical Sciences* (eds. Hodson et al.), Edinburgh Univ. Press.
[3] Gower, J.C. (1975). Generalized Procrustes analysis, *Psychometrika*, **40**, 33–51.
[4] Holman, E.W. (1972). The relation between hierarchical and Euclidean models for psychological distances, *Psychometrika*, **37**, 413–423.
[5] Lefkovitch, L.P. (1976). A loss function minimization strategy for grouping from dendrograms, *Systematic Zoology*, **25**, 41–48.
[6] Mardia, K.V. and others (1979). *Multivariate Analysis*, Academic Press.
[7] Ohsumi, N. (1980). Evaluation procedure of agglomerative hierarchical clustering methods by Fuzzy relations, *Data Analysis and Informatics* (eds. E. Diday et al.), North-Holland.
[8] Torgerson, W.S. (1958). *Theory and Methods of Scaling*, John Wiley.
[9] Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics, *J. R. Statist. Soc.*, B, **40**, 234–238.
[10] 芝　祐順 (1978). 因子分析法 (第2版), 東京大学出版会.