相関比基準による系統クラスター化について

水　野　欽　司

いろいろな現象について解析を進める場合に，いわゆるクラスター分析（cluster analysis）はきわめて有用な方法である。

特に個体の分類と識別をその主要な内容とする研究領域においては，多くの観測対象を少数個の比較的等質なグループに分割する方法として，不可欠のものとさえいえる。

クラスター分析の具体的な技法については量的，質的データの両面に亘って従来いろいろ考案され検討されてきている。

主として取上げられている問題は，

1）観測対象間の類似性の測度定義と最適クラスター化の基準をどうするか

2）それを実現する手順すなわちアルゴリズムをどうするか

等の問題である。このうちクラスター化の基準の合理性（一般の多変量解析との関連が明確であるという意味での）は重要であるが，それにも増して実行上無視し得ない問題は，いかに迅速にクラスター化を完成するかの問題であろう。現状ではなお多くの実用上の欠点をもつというべきである。

いろいろ提案されている中で，いわゆる系統クラスター化（hierarchical clustering）は特殊な利点をもっていると考えられる。

まず第一に，この手順は演算量を小さくする。下位グループの逐次合併により最終的な分類を達成するこの方法では，はじめ全対象はそれぞれが1グループであり，グループの数は対象数Nに等しい。合併の繰返しによりグループを1個ずつ減らして少数個のグループを得る。これはグループ数を固定したとき，そこで用いるクラスター化の基準測度において最良性を保証しないが，対象のあらゆる可能な組合せの比較を避けるので演算量が小さい。

第二の利点は系統的な分類構造はむしろ合目的的な場合が多いことである。われわれが知りたいのは特定数のグルーピングだけでなく，その上位分類や細分類も同時に欲しいことが多い。

第三にはグループ数を固定的に仮定する必要がない利点である。これは上の事柄とも関連する。クラスター化が必要な多くの実際場面では潜在グループの数について事前に情報をもっていない。したがって，あらかじめ予想により数を指定し，その数に関して設定した基準を最適化したとしても，グループ数を少し変更しただけで再び最適化に向って大幅な組替えが行なわれるとすれば，それは著るしく不自然なものといわねばならない。非系統的な方法ではこの類が多い。最適化が優れば優るほどグループの数に対し過敏となり，グループの内容構成を不安定にする怖れがある。系統的な方法は一般に基準の最適化に対し鈍いが，下位が上位に包含される構造の記述であるから，われわれは得られた構造の内容を検討することにより，目的上適当な数のグループを採択することができる。

本稿は，この系統クラスター化に関し，“一般化された相関比”を基準とする系統合併の試みについて述べる。*特にこの基準を取り上げる理由は，グループ間の判別の測度として現状で最もふさわしく思われるからである。また，簡単な数値データを使用した計算例について，関連する類似の方法による場合との比較を行なう。

I　最適グループ合併

1. ‘一般化された’相関比

“グループ間のちらばりが大きく各グループ内でのちらばりが小さい”という意味でのグループの個性化を数値基準として相関比（一般化された相関比）η^2 で表わす。

いまp種の変数 X_1, X_2, ……, X_p の全平均を $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_p$ とする。またグループの種別をsで表わし，各グループの対象数を n_s，またそれぞれの平均を $\bar{x}_{1s}$, $\bar{x}_{2s}$, ……, $\bar{x}_{ps}$ とすると，相関比 η^2 は次のようになる。

$$(1)\eta^2 = 1 - \frac{|\mathbf{W}|}{|\mathbf{T}|}, \qquad (2)\eta^2 = \frac{|\mathbf{B}|}{|\mathbf{T}|}$$

ただし $\mathbf{T}$ は全体の偏差積和のp×pの行列で，そのj行k列の要素を t_{jk} とすれば，

＊　その概要の一部は既に引用文献(3)にて報告ずみである。

$$t_{jk}=\sum_s\sum_i(x_{ijs}-\bar{x}_j)(x_{iks}-\bar{x}_k)$$

また$\mathbf{W}$と$\mathbf{B}$は同じくp×pの行列で，それぞれ層内，層間の偏差積和行列である。$\mathbf{W}$，$\mathbf{B}$のj行k列の要素をw_{jk}，b_{jk}とすれば，

$$w_{jk}=\sum_s\sum_i(x_{ijs}-\bar{x}_{js})(x_{iks}-\bar{x}_{ks})$$

$$b_{jk}=\sum_s n_s(\bar{x}_{js}-\bar{x}_j)(\bar{x}_{ks}-\bar{x}_k)$$

ここでx_{ijs}，x_{iks}は各分類対象の変数x_j，x_kの値である。

'一般化された相関比'を$_{(1)}\eta^2$と$_{(2)}\eta^2$として2種類考慮するのは，$\mathbf{T}=\mathbf{W}+\mathbf{B}$であるが一般に$|\mathbf{T}|\neq|\mathbf{W}|+|\mathbf{B}|$であるという事情による。普通には前者$_{(1)}\eta^2$が多く用いられるが，ここでは後述の理由で両方を考慮する。なおp＝1のときは$_{(1)}\eta^2$と$_{(2)}\eta^2$は等しい。

2. 系統クラスター化の場合

段階的に下位グループの合併を積み上げる系統クラスター化では出発時における相関比が最高値（個別対象から開始するときは1）である。グループ合併により相関比は減少しても増加することはない。したがって，ここでは最適性基準としての相関比は'順次高めて行く'基準ではなく，'減少を防ぐ'基準である。すなわち，合併によるその低下を最小に維持するための基準として相関比η^2を使用する。いいかえれば系統樹的構造の中の各段階で基準の最大化が計られるものとする。

3. グループ合併による$_{(1)}\eta^2$，$_{(2)}\eta^2$の変化

グループの改編によって相関比に影響を与えるのは$|\mathbf{T}|$が一定であるから$|\mathbf{W}|$または$|\mathbf{B}|$である。そこで$_{(1)}\eta^2$の場合は$|\mathbf{W}|$の増加を，$_{(2)}\eta^2$の場合は$|\mathbf{B}|$の減少を極力小さくするよう合併を進めればよい。

いまクラスター化がある段階まで進んだ状態として新たに二つのグループ（aとb）を合併する場合を考える。合併前の行列を$\mathbf{W}$，$\mathbf{B}$とし，合併後を$^*\mathbf{W}$，$^*\mathbf{B}$とする。ここで，

$$d_j=\sqrt{n_an_b/(n_a+n_b)}(\bar{x}_{ja}-\bar{x}_{jb})$$

とし，d_jを要素とするベクトルを$\mathbf{d}$とする。

$$\mathbf{d}'=\{d_1,\ d_2,\ \cdots\cdots,\ d_p\}$$

このとき，合併前と後では次の関係がある。

$$^*\mathbf{W}=\mathbf{W}+\mathbf{dd}',\quad ^*\mathbf{B}=\mathbf{B}-\mathbf{dd}'$$

また，グループaとbの合併により$\mathbf{W}$，$\mathbf{B}$の行列式は次のように変化することがいえる。

$$|^*\mathbf{W}|=|\mathbf{W}|(1+\mathbf{d}'\mathbf{W}^{-1}\mathbf{d})\qquad \text{ただし}\quad |\mathbf{W}|\neq 0$$

$$|^*\mathbf{B}|=|\mathbf{B}|(1-\mathbf{d}'\mathbf{B}^{-1}\mathbf{d})\qquad 〃\qquad |\mathbf{B}|\neq 0$$

したがって合併による相関比の減少を最小限にとどめるには，先の$_{(1)}\eta^2$ならば$\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}$を，$_{(2)}\eta^2$ならば$\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$をそれぞれ最小にするグループa，bを選び，それらを合併して新しいグループcを作ればよいことがいえる。

新グループcにおける対象数n_cと平均は，

$$n_c=n_a+n_b$$

$$\bar{x}_{jc}=(n_a\bar{x}_{ja}+n_b\bar{x}_{jb})/n_c\quad(j=1,\ \cdots\cdots,\ p)$$

となる。合併後は，$^*\mathbf{W}$，$^*\mathbf{B}$は，

$$^*\mathbf{W}=\mathbf{W}+\mathbf{dd}',\quad ^*\mathbf{B}=\mathbf{B}-\mathbf{dd}'$$

である。

合併が終了したら次の段階のために$^*\mathbf{W}^{-1}$，あるいは$^*\mathbf{B}^{-1}$を計算しなければならない。これは次の関係を利用して比較的簡単に求めることができる。

$$^*\mathbf{W}^{-1}=[\mathbf{W}+\mathbf{dd}']^{-1}=\mathbf{W}^{-1}-\frac{(\mathbf{W}^{-1}\mathbf{d})(\mathbf{W}^{-1}\mathbf{d})'}{1+\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}}$$

すなわち，右辺第二項を計算して$\mathbf{W}^{-1}$を調整すればよい。ベクトル$(\mathbf{W}^{-1}\mathbf{d})$は最小化グループ対を求める計算より得られているから，新たに必要な計算量はわずかですむ。

$^*\mathbf{B}^{-1}$の場合も同様である。

$$^*\mathbf{B}^{-1}=[\mathbf{B}-\mathbf{dd}']^{-1}=\mathbf{B}^{-1}+\frac{(\mathbf{B}^{-1}\mathbf{d})(\mathbf{B}^{-1}\mathbf{d})'}{1-\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}}$$

特殊な場合として一次元データでは以上の計算は不要で全く簡単なものとなる。

変数が1個の場合は，相関比を2種類考える必要はなく，通常の相関比を考えて，合併の各段階で次の$\triangle\mathbf{W}$が最小となるグループ対を求め，

$$\triangle\mathbf{W}=\frac{n_an_b}{n_a+n_b}(\bar{x}_a-\bar{x}_b)^2$$

合併して行けばよい。

数値x_iを最初に大小順に配列しておけば，隣合うグループ対を比較すればよく，相関比の変化も簡単な関係となる。さらに数値Xiを平均0，分散1に基準化しておけば，合併後相関比$^*\eta^2$は，

$$^*\eta^2=\eta^2-\min(\triangle\mathbf{W})$$

である。

4. 両相関比の併用

一般に系統クラスター化では合併の初期では全対象は多数のグループに細分化されており，行列$\mathbf{W}$のランクは落ちているのが通例である。各対象を1グループと考えて出発するときは必ずそうであるし，あらかじめ別の理由によってある程度グループ化されている状態から始める場合でもそれはありうる。一方クラスター化が進んでグループ数が行列の次数（変数の数）以下になれば$\mathbf{B}$のランクが落ちる。

これは上の基準$_{(1)}\eta^2$，$_{(2)}\eta^2$のどちらにせよ，単独で

は全過程を通して使用できないことを意味している。

そこで両基準を併用することによりこの困難を避けるとすると一つの解決案は次のようになる。はじめ基準 $_{(2)}\eta^2$ を適用し行列 $\mathbf{W}$ がフルランクとなり $_{(1)}\eta^2$ がある値を上廻るのを待ち基準を $_{(1)}\eta^2$ に移行する方法である。これは大局的には $_{(1)}\eta^2$ を使うが，$_{(2)}\eta^2$ をその補助として使用しようとする形式である。

本稿ではこのいわばリレー方式を検討する。

II　相関比基準の意味

ここで用いた相関比 $_{(1)}\eta^2$ の定義における $|\mathbf{W}|/|\mathbf{T}|$ はいわゆる S.S. Wilks (1962)[6] の Λ 基準にあたるものであり，H.P. Friedman and J. Rubin (1967)[2] が用いた最適クラスター化の基準 $|\mathbf{T}|/|\mathbf{W}|$ と逆数の意味で同等である。かれらはそれを最大化すべき量として定義している。かれらの場合は分析の出発においてグループ数を与件として試みのグループ化を行ない，山登り法 (hill-climbing pass) または強制法 (forcing pass) などの方法により個々の対象を移動させて修正する。

A.J. Scott and M.J. Symons (1971) によれば，結果的に $|\mathbf{W}|$ を最小にすることは，全対象グループがm個のp次元正規分布からの標本でありその共分散行列が等しいことを仮定した場合において，集団情報皆無のときの全対象のグループ帰属に対する最尤解である。[5]

本稿での基準はグループ合併による低下を最小に維持するためのものであるから，グループ数を先決条件とする方法の上の議論と同列には扱えないが基本的な性格は同じといってよい。

2グループの合併による層内積和行列 $\mathbf{W}$ の行列式の変化は，変化後を $|^*\mathbf{W}|$ として，

$$|^*\mathbf{W}|=|\mathbf{W}|(1+\mathbf{d}'\mathbf{W}^{-1}\mathbf{d})$$

であり，それ故 $\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}$ の最小となるグループ a，b を合併すればよいことは既に述べた。$\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}$ の値は，

$$\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}=\frac{n_a n_b}{n_a+n_b}(\bar{\mathbf{x}}_a-\bar{\mathbf{x}}_b)'\mathbf{W}^{-1}(\bar{\mathbf{x}}_a-\bar{\mathbf{x}}_b)$$

（ただし，$\bar{\mathbf{x}}_a$，$\bar{\mathbf{x}}_b$ はグループ a，b の平均ベクトル）

Hotelling の T^2 統計量に対応する量である。

しかし，クラスター化を行なう実際場面で分布型について種々の仮定をおくことは無理が多いから，純粋に記述的なものとしておくのがよいであろう。

方法の検討が進んで十分固まれば，標本論の立場からこれらを整理し計算手順に若干の修正を加えることは容易であろう。

いずれにしても相関比に基づく系統合併は，結局は Mahalanobis の距離の意味でグループ間の近接性を評価していることになる。

基準 $_{(1)}\eta^2$，$_{(2)}\eta^2$ に関連する他の方法で，変数間の共分散を考慮せず通常のユークリッドの距離の意味でグループ近接性を定義する仕方がある。これは層内積和行列 $\mathbf{W}$ のトレース $t_r(\mathbf{W})$ を最小化するクラスター化でよく使われる。[1]

普通にはグループ k における偏差積和行列を $\mathbf{W}_k$ とするとき，

$$t_r(\mathbf{W})=\sum_{k=1}^{m}t_r(\mathbf{W}_k)$$

が最小となるよう個々の対象をいずれかのグループに帰属させる方法をとる。$t_r(\mathbf{W})$ の最小化は $t_r(\mathbf{B})$ の最大化と同義であることはいうまでもない。

系統クラスター化でこれを行なうとすれば，2グループの合併による $t_r(\mathbf{W})$ の増加を最小に保つよう行なうことである。$t_r(\mathbf{W})$ の増加量を $\triangle t_r(\mathbf{W})$ とするとき両グループの合併により，

$$\triangle t_r(\mathbf{W})=t_r(\mathbf{dd}')=\mathbf{d}'\mathbf{d}$$

がいえる。$\mathbf{d}'\mathbf{d}$ が最小となるグループ a，b を求めればよい。この基準は1変数（p=1）のとき，基準 $_{(1)}\eta^2$，$_{(2)}\eta^2$ と同等のクラスター化となる。

$t_r(\mathbf{W})$ はp次元空間における各対象ベクトルの端点とその属するグループ重心との距離の2乗和を意味する。グループごとにこの2乗距離の平均をとればグループの凝集性に関する一つの測度——集中度D——を定義することができる。[4]

系統合併でこれを用いるときは2グループ a，b を合併したときの集中度D，

$$D=(t_r(\mathbf{W}_a)+t_r(\mathbf{W}_b)+\mathbf{d}'\mathbf{d})/(n_a+n_b)$$

が最小となるグループ対を選べばよい。これは $t_r(\mathbf{W})$ の最小と類似しているが，グループのちらばりの範囲をややよく反映する。

III　計算例

次に比較的簡単な数値データを用いた計算例について検討する。

比較のため，ここで取り上げる系統クラスター化の方法は次の4通りである。

C1 ……相関比 $_{(1)}\eta^2$ の減少を最低にする合併（ただし，$_{(1)}\eta^2\geq 1-\varepsilon$ のときは $_{(2)}\eta^2$ を基準として使用する）

C2 ……相関比 $_{(2)}\eta^2$ の減少を最低にする合併（ただ

表1　　計算例に用いた項目

（項　　目）	（平均）
Ⅰ　民主的秩序の破壊者である	3.031
Ⅱ　この事件は非常に時代ばなれした感じがする	3.438
Ⅲ　現代の'いらだたしさ'に正面から立ちむかった彼らの行動力には大いに共感できる	2.563
Ⅳ　思想と行動とを合致させようとする生き方には賛成である	2.250

（共分散行列）

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Ⅰ	.843			
Ⅱ	.580	1.371		
Ⅲ	−.174	−.309	1.496	
Ⅳ	−.258	−.141	.828	1.063

（相関行列）

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Ⅰ	1.000			
Ⅱ	.540	1.000		
Ⅲ	−.155	−.215	1.000	
Ⅳ	−.272	−.117	.657	1.000

し，$_{(2)}\eta^2 \leqq \delta$ のとき基準を $_{(1)}\eta^2$ に切換える）

C_3 ……行列 $\mathbf{W}$ のトレース $t_r(\mathbf{W})$ の増加を最小にする合併

C_4 ……集中度Dが最小となるグループ対の合併

以下，C_1，C_2，……はこれらの方法および結果の略号として用いる。上で ε，δ は正の小さい数である。C_1，C_2は結局，相関比基準の切換えをどこで行なうかの程度の差をいうに過ぎない。C_1は比較的早く，C_2は遅い場合となっている。

1. 試算用のデータ

データはいわゆる'三島由紀夫切腹事件'の直後に工科系男子大学生に行なった意見調査の結果の一部である。*これは事件に関する種々の感想項目に対する賛成，反対を5段階で評定したものである。ここではその中の4項目に対する回答結果を使用し，「まったく賛成」を1点，「まったく反対」を5点とする5点スケールの数値として扱うことにする。

回答者32名の4項目に対する回答内容は表2に掲げる通りである。回答パターンの等しい対象をかなり含んでいる。

また各項目の平均，分散共分散行列，および相関行列を表1にまとめておく。項目ⅠとⅡ，および項目ⅢとⅣはそれぞれやや高い相関を認めることができるが，前2項目と後の2項目の間では相関が低い。

なおこの分散共分散行列に関し成分分析を行ない個々の対象（回答者）の成分値を算出した。これは分類計算の結果を図示するときの利便を計ったものである。固有値 λ は，

$\lambda_1=2.42$，　$\lambda_2=1.46$，　$\lambda_3=.57$，　$\lambda_4=.32$

である。λ_1 と λ_2 で全体の81%となっている。成分値の分散は λ に基準化しておく。（図2）

2. 計算の具体的手順

逐次合併により所定のグループ数における分類を行なう手順は次の通りである。記号は既述のように使うとする。分析の開始時には全対象がそれぞれ1グループであり，グループ数は対象数Nと等しい。

(1)　C_1 および C_2

方法 C_1 による場合の手順の概略は以下の通り。

1）全体の偏差積和行列 $\mathbf{T}$ を作る。

2）$|\mathbf{T}|$ および $\mathbf{T}^{-1}$ を計算する，$\mathbf{W}$を$\mathbf{O}$とする。（出発時における $\mathbf{B}$ は $\mathbf{T}$ と等しいから，これらは $|\mathbf{B}|$ と $\mathbf{B}^{-1}$ の初期値である）

3）$_{(2)}\eta^2=1$ とおく。

4）$\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$ が最小となるグループの対を探す。（グループ数をsとすると，$_sC_2$ 通りの組合せについてこの値を計算し比較しなければならない）

5）グループ対の合併に伴う相関比 $_{(2)}\eta^2$ の計算

$$*_{(2)}\eta^2 = {}_{(2)}\eta^2(1-\mathbf{d}'\mathbf{B}^{-1}\mathbf{d})$$

6）逆行列 $\mathbf{B}^{-1}$ を $*\mathbf{B}^{-1}$ に，行列 $\mathbf{W}$ を $*\mathbf{W}$ に変更する。

7）グループ対の合併

$$n_a+n_b=n_c$$
$$(n_a\bar{x}_{ja}+n_b\bar{x}_{jb})/n_c=\bar{x}_{jc}$$
$$(j=1,\ 2,\ \cdots\cdots,\ p)$$

8）グループ数の判定

合併後のグループ数を $*S$ とし，$*S=S-1$ に変更する。

＊　引用文献(3)，(4)におけるデータと同類のもの。

表2　分類の結果（5グループと3グループ）

	相関比 C_1					C_2		トレースW C_3		集中度 C_4	
グループ(1)	グループ	3)	2	1	5	3	グループ(1)	グループ	3)	グループ(1)	グループ	3)
		6)	3	1	4	3			6)			6)
		15)	1	1	5	5			15)			15)
	①	28)	4	2	4	2		①			①	
(2)	②	10)	2	2	1	1		②	10)	(2)	②	10)
		26)	2	2	1	2			26)			26)
		32)	1	2	1	1			32)			32)
	③	5)	2	4	4	2		③	1)	(3)	③	1)
		16)	2	4	2	2			2)			2)
		29)	2	5	3	2			7)			7)
(3)	④	7)	4	4	3	2			9)			9)
		21)	4	4	3	2			21)			21)
		9)	4	4	1	1			22)			22)
		22)	4	4	1	1			23)			23)
		23)	4	5	2	1			24)			24)
		1)	4	4	2	2			25)			25)
		24)	4	4	2	2			31)			27)
		25)	4	4	2	2		④	8)			31)
		31)	4	4	2	1			11)			5)
		8)	3	3	2	2			12)	(3)	④	8)
		11)	3	3	2	2			13)			11)
		13)	3	3	2	2			14)			12)
		17)	3	3	4	2			16)			13)
		18)	3	3	3	2			17)			14)
		30)	3	4	4	3			18)			16)
		2)	4	5	2	2			19)			18)
		12)	3	4	2	3			20)			19)
		14)	3	4	2	2			29)			20)
		20)	3	4	1	2			30)			29)
		19)	2	3	2	4			5)			17)
	⑤	4)	3	5	5	5			28)			28)
		27)	4	5	3	4		⑤	4)			30)
									27)		⑤	4)

数字は対象番号，C_1については回答の内容を付してある。

① $*S = m$（目標グループ数）ならば終了

② $*S > m$ ならば，$|*\mathbf{W}|$を計算する。

(イ) $|*\mathbf{W}| \leqq \varepsilon |\mathbf{T}|$ ならば，4）～8）の手順を反復する。

(ロ) $|*\mathbf{W}| > \varepsilon |\mathbf{T}|$ ならば9）へ移行

9）新基準 ${}_{(1)}\eta^2$ の計算と $\mathbf{W}^{-1}$ の設定

${}_{(1)}\eta^2 = 1 - |\mathbf{W}| / |\mathbf{T}|$

10) $\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}$ が最小となるグループの対を探す。

11）グループ対の合併に伴う相関比 ${}_{(1)}\eta^2$ の計算

$1 - {}^*{}_{(1)}\eta^2 = (1 - {}_{(1)}\eta^2)(1 + \mathbf{d}'\mathbf{W}^{-1}\mathbf{d})$

12）逆行列 $\mathbf{W}^{-1}$ を $*\mathbf{W}^{-1}$ に変更する。

13）グループ対の合併（7）と同じ）

14）グループ数の判定

① $*S = S - 1 = m$ ならば終了

② $*S > m$ ならば 10）～14）の手順を反復する。

方法 C_2 では上の手順のうち，8）を次の8′）に取り替えたものとなる。

8′）(1) $|*\mathbf{B}| > \delta |\mathbf{T}|$ ならば4）～8′）の手順を反復する。

(2) $|*\mathbf{B}| \leq \delta|\mathbf{T}|$ ならば 9）へ移行

この計算では ε を0.0001，δ を0.2として計算している。

(2) C_3 および C_4

方法 C_3 の手順は，

1）行列 $\mathbf{W}$ のトレース $t_r(\mathbf{W})$ を0としておく。

2）$\mathbf{d'd}$ が最小となるグループの対を探す。

3）$t_r(\mathbf{W})$ を $t_r(*\mathbf{W})$ に変更する。

$$t_r(*\mathbf{W})=\mathbf{d'd}+t_r(\mathbf{W})$$

以下，グループ対を合併し一つのグループとし，グループ数（$*S=S-1$）の判定を行なう。

終了でなければ 2）からの手順を反復する。結果的な基準値 $t_r(\mathbf{W})$ の最小値は3）で求めた累和である。

方法 C_4 の手順は C_3 と異なりグループ別の $t_r(\mathbf{W}_g)$ を用意する。

1）$t_r(\mathbf{W}_g)$ を0とおく。（$g=1, 2, \cdots\cdots N$）

2）集中度Dが最小となるグループ対を求める。グループをa，bとすると，

$$D=(t_r(\mathbf{W}_a)+t_r(\mathbf{W}_b)+\mathbf{d'd})/(n_a+n_b)$$

以下，グループ対の合併を行なう。新グループCにつき，$t_r(\mathbf{W}_c)$ を作って，グループ数の判定を行ない，終了に至らぬときは2）に戻って反復する。

3. 分類の結果

各方法による分類の模様は表2に示す通りである。いずれもグループ数5，および3の二つの場合を示してある。方法 C_1 以外は4項目に対する回答内容を省略し対象番号のみを挙げてある。全般に配分された対象の数の不揃いが目立っている。

なお方法 C_1 は，相関比 ${}_{(2)}\eta^2$ を基準として出発し，$|\mathbf{W}|$ が境界値 $\varepsilon\cdot|\mathbf{T}|$ を越えたときに基準 ${}_{(1)}\eta^2$ に切換えたものである。$\varepsilon=0.0001$ は ${}_{(1)}\eta^2$ で表わすと0.9999に当る。境界値を越えたのはグループ数13のときである。移行が比較的遅いのは回答内容が同一の対象をかなり含んでいることによる。2種の相関比の合併の進行に伴う値の変化を図示したのが図1である。

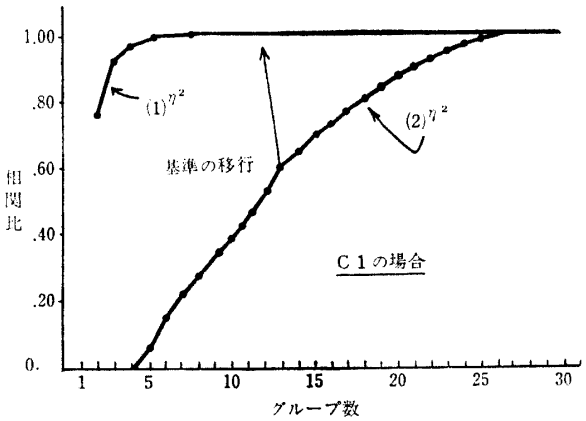


図1　グループ合併による相関比の減少

また方法 C_2 は最初の基準 ${}_{(2)}\eta^2$ を継続して，$|\mathbf{B}|$ が $\delta\cdot|\mathbf{T}|$ を下廻ったとき新基準 ${}_{(1)}\eta^2$ に移行したもの。この計算では $\delta=0.2$ でこれは ${}_{(2)}\eta^2=0.2$ となる。実際にはグループ数6まで最初の基準で5グループから新基準に移っている。

表3では各方法の結果的な分類効果を表わす測度として両相関比および $\mathbf{W}$ のトレース（$t_r(\mathbf{W})$）を5グループ，4グループ，3グループの場合について示した。

方法 C_1 では ${}_{(1)}\eta^2$ がどのグループ数の場合においても他の方法のそれより優っている。また方法 C_2 では ${}_{(2)}\eta^2$ が，方法 C_3 では行列 $\mathbf{W}$ のトレースが最小となっている。いずれもそれぞれの方法の個性が活かされた結果と

表3　グループ化の効果の比較

基準と手順		相関比		トレースW	集中度
		C_1	C_2	C_3	C_4
3グループ	相関比 ${}_{(1)}\eta^2$	.922	.917	.852	.852
	〃 ${}_{(2)}\eta^2$	.000	.000	.000	.000
	トレース $t_r(W)$	90.73	107.20	83.40	83.40
4グループ	相関比 ${}_{(1)}\eta^2$	.967	.969	.951	.919
	〃 ${}_{(2)}\eta^2$	.000	.000	.000	.000
	トレース $t_r(W)$	70.85	76.95	64.21	67.31
5グループ	相関比 ${}_{(1)}\eta^2$	.988	.985	.986	.964
	〃 ${}_{(2)}\eta^2$	.053	.110	.011	.004
	トレース $t_r(W)$	55.85	61.96	47.60	55.95

ただし $t_r(T)=152.72$

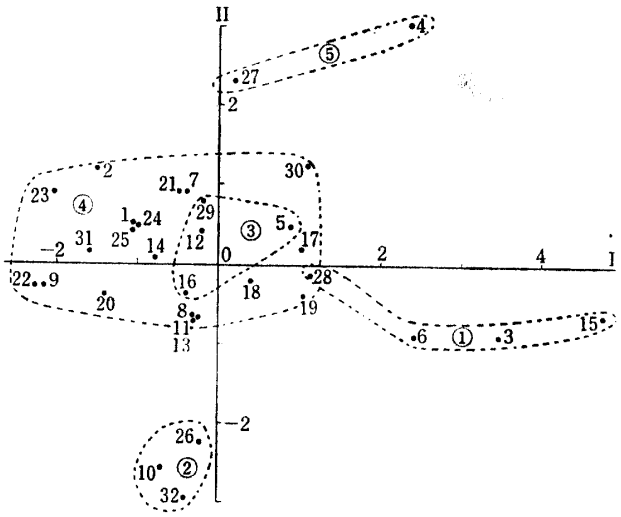


図2－1　主成分における対象のグループ化
―C₁(相関比)による場合―

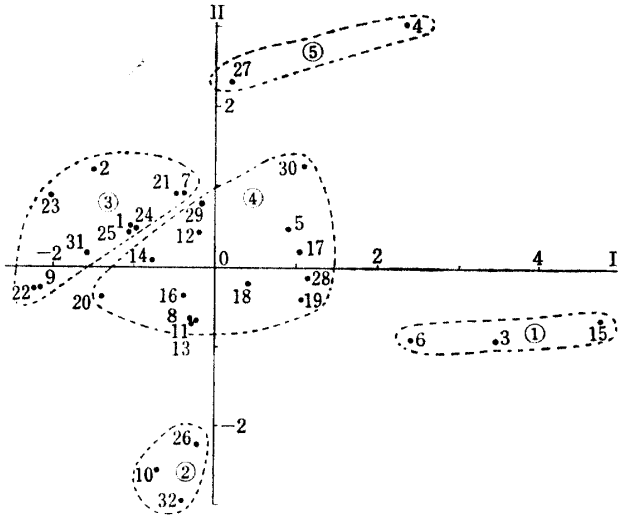


図2－2　主成分における対象のグループ化
―C₃(t(W))による場合―

なっている。方法C_4は集中度D最小のグループを作る合併である。上の3種の測度に関してはやや劣っているが3グループではC_3と同一の結果となった。

採用された測度に関する限り，4種の方法間の差は概して小さいといわなければならない。

分類された内容については，C_1とC_2がやや似ており，C_3とC_4は良く似ているといえる。これは「距離」の考え方の相違からみて当然であろう。

4. 分類結果の図示

方法C_1とC_3の場合に限って対象の分類の模様を4項目の主成分の空間に表わしてみる。

図2は第一成分，第二成分における5グループ分類の結果を示す。図2－1はC_1，図2－2はC_3による分類である。

おおむね似ているが，C_3は第一，第二主成分の空間でよく分離している。C_1はその傾向においてやや劣っている。これは変数間の相関性が考慮されるために，主要な成分の方向で同じ距離でも過小に評価することを意味している。

5. 判別分析との関係

次に得られたグループ分類に基づきあらためて判別分析を実施した結果を図3に示そう。

3グループ，5グループの場合に限り，判別変量（主要2軸）の値で各対象の位置を図示してある。各判別変量は平均0，分散1に基準化してある。

ここで第一判別変量はp次元の変数ベクトルを$\mathbf{x}$とし重みベクトルを$\mathbf{v}$とする一次結合$y=\mathbf{v}'\mathbf{x}$であり，それはyに関する相関比（μ^2と記す），

$$\mu^2=\frac{\mathbf{v}'\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{T}\mathbf{v}}$$

を最大にする解である。第二判別変量以下は他と直交するという条件の下で同様にして得られる。これらの解は固有方程式，

$$\mathbf{B}\mathbf{v}=\mu^2\mathbf{T}\mathbf{v}$$

を解いて，0でない固有値μ^2の大きい順に対応する固有ベクトル$\mathbf{v}$をとればよい。

一方，C_1の基準${}_{(1)}\eta^2$に関して$|\mathbf{T}|/|\mathbf{W}|$は，

$$|\mathbf{T}|/|\mathbf{W}|=|\mathbf{W}^{-1}\mathbf{T}|=|\mathbf{I}+\mathbf{W}^{-1}\mathbf{B}|=\prod_r(1+\lambda_r)$$

がいえる[(2)(6)]。ここでλは固有方程式

$$\mathbf{B}\mathbf{v}=\lambda\mathbf{W}\mathbf{v}$$

の解である。したがって$\mathbf{W}^{-1}\mathbf{B}$の固有値$\lambda$と$\mathbf{T}^{-1}\mathbf{B}$の固有値$\mu^2$には

$$\mu^2=\frac{\lambda}{1+\lambda},\qquad \lambda=\frac{\mu^2}{1-\mu^2}$$

の関係がある。結局，基準${}_{(1)}\eta^2$は0でない固有値λ，μ^2により

$${}_{(1)}\eta^2=1-1/\prod_r(1+\lambda_r)=1-\prod_r(1-\mu_r^2)$$

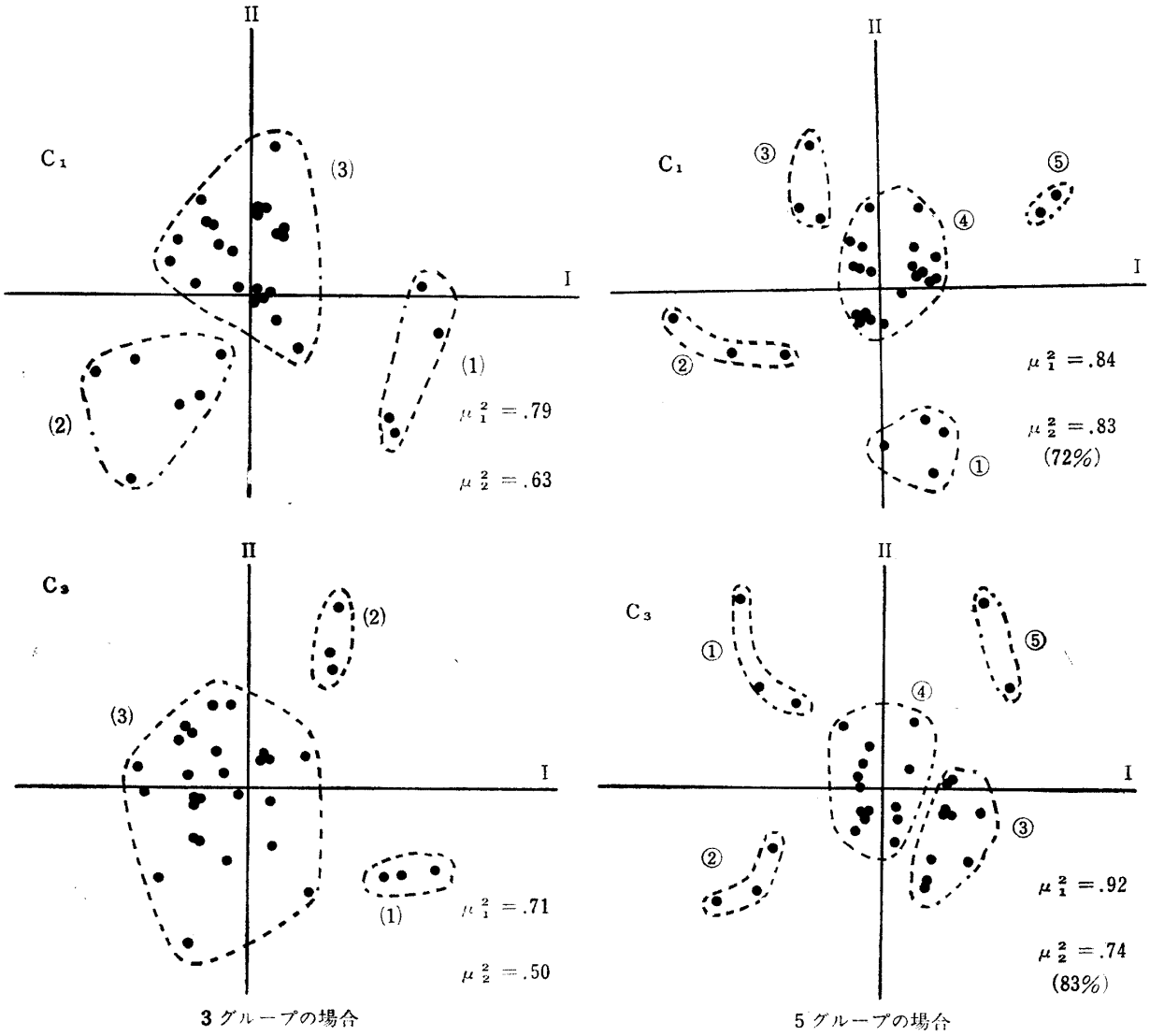
がいえる。

図3の3グループの場合の例でいえば，C_1では判別変量は2個で，$\mu_1^2=.79$，$\mu_2^2=.63$であり，

$${}_{(1)}\eta^2=1-(1-0.79)(1-0.63)=0.922$$

は既に分類計算の際，得られている値（表3）である。C_3では$\mu_1^2=.71$，$\mu_2^2=.50$で，${}_{(1)}\eta^2=.852$である。計算上，C_1がC_2より優っているが，図に見る分離状況からもやややC_1の優る傾向が認められる。

5グループの場合は判別変量は4種あるが，図にはその上位2種しか示していない。もちろんC_1は${}_{(1)}\eta^2$に



3グループの場合　　5グループの場合

図3　分類グループとその判別分析

おいて C_3 に優っているが上位2軸まではそうもいえない。少なくとも固有値 μ^2 の和では C_3 が C_1 に優る。

一般に C_1 は各軸の判別力を均等化する方向に働き，C_3 は上位の軸に集中させる傾向がある。

IV 結　び

小さな計算例であるため，相関比を用いる系統クラスター化の有効性について十分な保証を与えることはできない。

しかし，グループの個性化という意味での特徴を比較的よく発揮することは認められると思う。今後，より大きなデータに関して数多く検討する必要がある。

その際，まず第一に考慮すべきは計算節約の工夫と他のクラスター化法との有効な併用形式であろう。

今回の検討で行列 **W** のトレースの増加を最小にする合併（C_3），あるいは集中度最小のグループを作る合併（C_4）は決してわるくなかった。C_3，C_4 の手順は相関比による場合に比べ演算量が大幅に小さい。一方分析の当初は合併候補のグループの組合せ数が多い。

したがって，ここでは初期状態で相関比 $_{(2)}\eta^2$ を用いたが，それよりも単純にトレース $t_r(\mathbf{W})$ に関してクラスター化を行ない，その後に相関比 $_{(1)}\eta^2$ に移行するのがよいかも知れない。

その他，検討すべき課題は多い。たとえば，

1）系統合併の解の安定性の条件

2）変数（項目）側における分類との関連構造（変数と対象の同時的分類を意図するときどうするか）

3）多次元解析の他の諸方法との関係

4）適用が効果的な現象領域，あるいはデータの条件

などである。

本稿はその試みの段階の報告にとどまっている。

（1971年11月22日）

引用文献

(1) Edwards, A.W. and Cavalli-Sforza, L.L. 1965
A method for cluster analysis.
Biometrics, 21, 362—375.

(2) Friedman, H.P. and Rubin, J. 1967
On some invariant criteria for grouping data.
J. Amer. Statis. Assoc., 62, 1159—1178.

(3) 水野欽司 1971
多次元数値データの系統的自動分類
日本心理学会35回大会論文集, 745—746.

(4) 水野欽司 1970
系統的項目分類の一方法
名古屋大学教育学部紀要—教育心理学科—17, 117—124.

(5) Scott, A.J. and Symons, M.J. 1971
Clustering methods based on likelihood ratio criteria.
Biometrics, 27, 387—397.

(6) Wilks, S.S. 1962
Mathematical statistics.
John Wiley & Sons; N.Y.