応用統計学 Vol. 8, No. 2

研究論文

Single linkage 法と Complete linkage 法の特性とクラスター数評価基準

柳澤 幸雄・大隅 昇
日本科学技術研修所　統計数理研究所

要　旨　数多くのクラスタリングの技法のうち連結性または距離の順位性のみに依存する single linkage 法と complete linkage 法を取り上げ，技法とクラスターの関係を“凝塊性”の立場から測るいくつかの基準量を用意し，クラスター数の評価を行う問題をとりあげる．さらにそれを利用して技法の性質，あるいは技法ごとのクラスター化の過程を推し測る方法として“感度分析”の考え方を導入する．そして布置の明確な人工データを用いて，感度分析の効用を実験により具体的に検討する．この実験を通して，感度分析が技法の性質と安定性，データの凝塊性の検討に有効であることを示す．

1. まえがき

データ解析において，データに内在する傾向，共通性を摘出し把握することは，データより帰納して理論上のモデルを構成する際の有力な手がかりを与える．つまり，データの潜在的傾向を解明することは，データ解析の第一段階としてきわめて重要である．この目的に対し，クラスター分析は一つの有力な手掛りを与えるものである．

クラスター分析は，最近多くの分野で使われるようになり，数多くの手法が提案されてきたが，この多様性がかえって，この分析法のマイナス面となっていることは否めない．単に手法を積み上げるだけにとどまらず，各分析法のもつ特性，適用に当っての有効性を検討しておくことが必要である．それがあってはじめて，クラスター分析の実際面での適用が妥当性をもつこととなる．

本論文の目的は，もっともよく知られている２つの技法に焦点をしぼって，その理論的特性を利用した，技法間の有効性比較の方法としての感度分析の方法を提示することである．この感度分析の方法は，人工データによる実験から

1978年８月24日受付

確かめられた限りでは，与えられたデータに対する技法の挙動の解明に有用であり，とくにデータの“凝塊性”について有効な情報を与えることが明らかになる．ここで凝塊性とは，ある程度まとまりをもったコンパクトなクラスターがいくつか存在するという状況をいう．

つぎに，取り上げる対象の範囲を順を追って項目別に説明しておく．

A．データの形式

取り扱うデータは，個体数 n，変量数 m の $n\times m$ 次のデータ行列 $\boldsymbol{X}=(x_{ij})$ $(i=1,2,\cdots,n;\ =1,2,\cdots,m)$ とし，i 番目の観測ベクトルを $\boldsymbol{x}_i'=(x_{i1},\cdots,x_{im})$ とする．

B．個体間の類似度，あるいは距離の定義

ユークリッド距離を使用し，データにはとくに特定の確率分布を仮定しない．

C．クラスターの概念

クラスターを一義的に定義することは困難であるが，ここではクラスターとしてどのようなものを考えようとしているかを説明しておく．本論文を通じて我々がクラスターとみなすのは，データのうちで多次元の球，あるいは比較的それに近い集塊状をなすもので，クラスター内距離がクラスター間距離に比較し小さいものであるような，いわゆるコンパクトなクラスターの集まりを想定する．

たとえば図1のような2次元のデータの布置があるとき，a，b のデータに対して，㈠や㈡のような集塊はクラスターと考えるが，c，d のデータの場合に，我々は㈢，㈣のような形状の意味ではクラスターとみなさない（この手当ての一案として，MDS，Intrinsic Dimensionality などを事前に使用することが考えられる）．

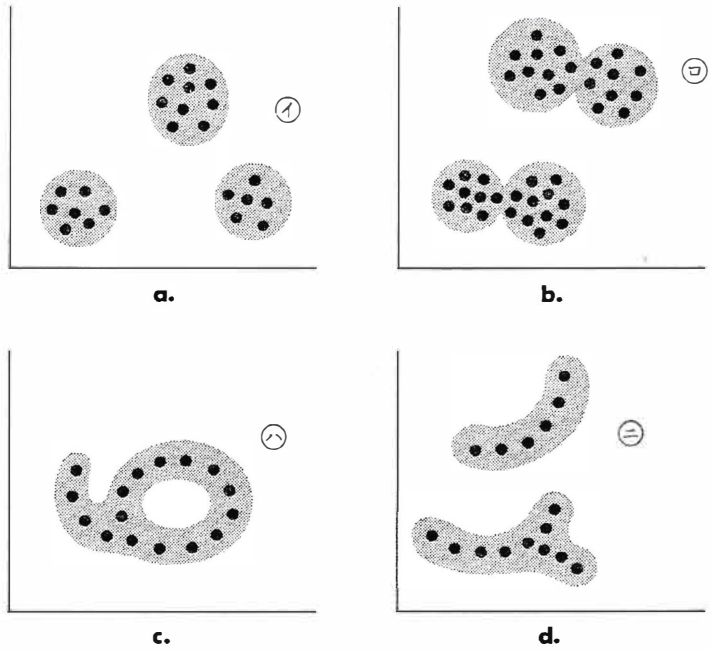


図1

D．クラスター化の技法

ここでは single linkage 法（SL 法と略す）および complete linkage 法（CL 法と略す）を取り上げる．

E．クラスター数評価の基準量

クラスター化の技法のみでは，クラスター数をきめがたい．それは同一データであっても，技法ごとにクラスター化の過程が異なるからである．したがって，クラスター数評価のために，我々は凝塊性を利用する．この場

合凝塊性を測る基準としては，2節に述べる4つの基準を用意する．

以上項目別にあげた範囲内でクラスター分析に対する検討を行うが，まず 2.1. ではクラスター化の技法に関連した問題を取り上げる．我々の対象とする SL 法および CL 法は，クラスター化にあたって個体またはクラスター間の連結性のみを利用してクラスターを生成する方法といえる．この種の方法には階層の構成のされ方にある種の特性があるが，これに関しては文末の〔補足〕で言及する．

2.2. では，クラスター数評価のために用いる4つの基準について，それらの持つ意味と狙いを要約する．クラスタリングの技法とクラスター数評価の基準を組合せてクラスター分析の手法を得るわけであるが，3節では，技法と基準量の双方の性質を調べ，さらに技法間の有効性の比較を行う上で有用な方法として感度分析を提案する．そして，人工データに感度分析を適用して，以上の検討とデータの凝塊性について具体的な比較検討を行う．

2. 手法とその性質

2.1. Single linkage 法と Complete linkage 法

大きさが n のデータを，k 個の空でない互いに排反な集合に分割する方法の総数は，第二種の Stirling 数で表わされるように多数あり，すべての分割について確かめることは事実上不可能である．そこで，ここでは個体間の距離の小さい個体から順次寄せ集めながらクラスターを作る“凝集型の階層的技法”を中心に考えをすすめる．

一般に，階層的技法とは次のように定義される．

集合族 $S_1, S_2, \cdots, S_{n-1}$ が階層的（hierarchical）とは，

$C=\{\boldsymbol{x}_l\}$ $l=1, \cdots, n$ （全データの集合）

$S_i=\{C_{ij}\}$ $j=1, \cdots, n-i$ （C を $n-i$ 個に分割して得られる集合族）

C_{ij} $(i=1, \cdots, n-1;\ j=1, \cdots, n-i)$ （集合族 S_i に属する第 i クラスター）

とするとき，以下の条件を満足するものをいう．

（条件1） 任意の i，p に対し，$C_{ip} \neq \phi$

（条件2） $p \neq q$ のとき $C_{ip} \cap C_{iq} = \phi$

（条件3） 任意の i，j に対し，$\bigcup_p C_{ip} = \bigcup_q C_{jq} = C$

（条件4） C_{ip} と $C_{i+1,q}$ の添字 p，q を適当につけかえると，

$$C_{ip}=C_{i+1,p} \qquad (p=1, \cdots, n-i-2)$$

$$C_{i,n-i-1} \cup C_{i,n-i} = C_{i+1,n-i-1}$$

ところで Lance と Williams は，階層的技法のうちいくつかの技法を総括し，次の統一的表現が可能であることを示した[3]．

クラスター C_p，C_q を結合し，新しいクラスター C_t を作るとき，C_t と C_r（$r \neq p, q$）との距離を

$$d_{tr}=\alpha_p d_{pr}+\alpha_q d_{qr}+\beta d_{pq}+\gamma |d_{pr}-d_{qr}| \qquad (1)$$

と定義する．ここで d_{ij} はクラスター C_i，C_j 間の距離で $(i, j=p, q, r)$，α_p，α_q，β，γ は，技法ごとに与えられる定数である．

この階層的技法としてもっとも良く使用され，利用分野の多様性をみても代表的なクラスタリングの技法と考えられるものに，次に述べる SL 法と CL 法がある．本報告では，クラスター化の技法をこの2つにしぼることにする．じつはこの2つの技法は，クラスタリングにおいて相反する性質をもつものであり，この点からも以

下比較をしてゆく上で，これらに限定することが許されよう．

SL 法および CL 法とは次のようなものである．

＜SL法＞ （1）式において，$\alpha_p=\alpha_q=\frac{1}{2}$，$\beta=0$，$\gamma=-\frac{1}{2}$ と置いて与えられるもので

$$d_{tr}=\frac{1}{2}(d_{pr}+d_{qr})-\frac{1}{2}|d_{pr}-d_{qr}|=\min\{d_{pr},\ d_{qr}\}$$

で与えられる．すなわち d_{tr} は，クラスター C_t 内の個体と C_r 内の個体をむすぶすべての距離のうちの最小値である．

＜CL法＞ （1）式において，$\alpha_p=\alpha_q=\gamma=\frac{1}{2}$，$\beta=0$ と置いてえられるもので，

$$d_{tr}=\frac{1}{2}(d_{pr}+d_{qr})+\frac{1}{2}|d_{pr}-d_{qr}|=\max\{d_{pr},\ d_{qr}\}$$

で与えられる．すなわち d_{tr} は，クラスター C_t 内の個体と C_r 内の個体をむすぶすべての距離のうちの最大値である．

定義から明らかなように，これらの技法は，いずれも個体間の連結性だけに注目しており，クラスター内の個体数（クラスター・サイズ）やバラツキは考慮していない．このように連結性だけに依存する技法では，クラスター数を決定することは通常むづかしい．

たとえばクラスター数が異なる2つのデータに対し，上述の技法のもとで同じ性質（正確には，次に述べる global order equivalence の意味）をもつデンドログラムが対応することがありうる．この場合，技法によって作りだされたデンドログラムにもとづいてクラスター数を評価するためには，他の手がかり（指標）が必要となる．我々はそのためにクラスター数評価の基準量の概念を導入する．これらについては 2.2. に述べる．

一方また，同じ性質のデンドログラムを生じるようなデータの布置の間には，どのような関係があるかという問題がある．この意味での技法の特性については〔補足〕に述べた．つぎにクラスター数評価の基準量の有効性を検討するうえに必要な global order equivalence* の概念を導入しておく．

定義 階層的な集合族 $S_1, S_2, \cdots, S_{n-1}$ と $S_1', S_2', \cdots, S_{n-1}'$ $(S_i=\{C_{ij}\}\ j=1,\cdots,n-i,\ S'_i=\{C_{ij}'\}\ j=1,\cdots,n-i)$ に対し，集合 $C=\bigcup_j C_{ij}$ と集合 $C'=\bigcup_j C_{ij}'$ が global order equivalent set であるとは，$f(C_{ih},\ C_{ij})$ を C_{ih} と C_{ij} の距離とするとき任意の i に対し常に，

$$f(C_{ih},\ C_{ij})\leqq f(C_{ik},\ C_{il}) \quad\Longleftrightarrow\quad f(C_{ih}',\ C_{ij}')\leqq f(C_{ik}',\ C_{il}') \qquad (2)$$

が成立していることである．

global order equivalent set である場合には，両者の個体の連結順序が同じになり，またデンドログラムとしても同じ性質のものを生ずることとなる（global order equivalent dendrogram については 2.2. の図2および説明参照）．

この global order equivalence に関して，SL 法および CL 法の特性を調べておけば，いろいろ利用できる．これらについては〔補足〕に1つの結果を与えておいたが，これを要約すると，つぎのようになる．

SL 法または CL 法にもとづくクラスタリングにおいて，個体の移動のもとに global order equivalent な集合となるようなデータの布置における個々の点(データ)の移動許容範囲を幾何学的に定めることができる．

2.2. クラスター数評価の基準量

* global order equivalence という術語は，文献[7]において使われているが，我々の概念は，それとはやや異なるものである．[7]では(2)式を $f_1(C_{ih}, C_{ij})\leqq f_1(C_{ik}, C_{il}) \Longleftrightarrow f_2(C_{ih}, C_{ij})\leqq f_2(C_{ik}, C_{il})$ としたとき，f_1 と f_2 を global order equivalence と定義している。

2.1. に述べたように，一般にはクラスター数の決定にあたって，クラスター化の技法とともに別の基準量が必要となる．基準量を使用するとクラスター数を定量的に決める目安がえられる．

ここでは，クラスターを凝塊性の観点からとらえ，これをはかる基準量をいくつか用意し，それらの定義とその意味，およびクラスター数決定の手順について説明をおこなう．

いまデータを k 個の互いに排反な空でないクラスターに分割したとき，$\boldsymbol{S}_T$：全平方和・積和行列，$\boldsymbol{S}_W$：群内平方和・積和行列，$\boldsymbol{S}_W$：群間平方和・積和行列とおくと，以下の関係が知られている．

$$\boldsymbol{S}_T=\boldsymbol{S}_W+\boldsymbol{S}_B \tag{3}$$

この式は，k を固定したとき $\boldsymbol{S}_W$ の行列式あるいは $\boldsymbol{S}_W$ のトレース（群内平方和）の大きさにより，データの凝塊性をはかる目安となる．しかし，ここでは k を可変として，データの凝塊性をはかりたい．そこで（3）式より導かれる関係式で，群内平方和および行列式を基本とした基準量を考える．以下それらの基準量を個別に説明する．

（基準1） Calinski と Harabasz の基準[2]

Calinski と Harabasz (1971) が，いわゆる dendrite method と関連して提案した分散比基準である．

$$VRC(k)=\frac{\mathrm{tr}(\boldsymbol{S}_B)}{k-1}\Big/\frac{\mathrm{tr}(\boldsymbol{S}_W)}{n-k} \tag{4}$$

ここで $\mathrm{tr}(\boldsymbol{A})$ は，行列 $\boldsymbol{A}$ のトレースを表わす．

VRC 値が最大値をとるか，あるいはクラスター数がある $(k-1)$ から k への変化に対して急激に増加するとき，その k の値をクラスター数の目安として選ぶ．Calinski と Harabasz は，計算を節約するという立場から，VRC 値のいくつかの極大値があれば，そのときの k の値の最小値をもって，選出すべきクラスター数とすることを薦めている．しかし，我々は上記のようないくつかの k の値から一つ選出するときは，k の最大値をとることを薦めたい．これはデータを分割したのち，いくつかのクラスターを再び併合させて分割数を減少させることは容易であるが，一度集約したデータをさらに細分割することが，前者に比較して難かしいという実質的な立場をとるからである．

（基準2） Beale の基準[1]

E. M. L. Beale (1969) が提案した統計量である．

$$F(k_2,k_1)=\frac{R(k_1)-R(k_2)}{a\cdot R(k_2)} \tag{5}$$

$$\text{ここで，}\ a=\frac{n-k_1}{n-k_2}\left(\frac{k_2}{k_1}\right)^{2/m}-1$$

$R(k)$ はクラスター数が k のときの群内平方和の総和，k_1 はクラスター数，k_2 はデータを k_1 個に分割した後，さらに細分割したときのクラスター数で，ここでは $k_2=k_1+1$ とした．

Beale の F 値は，クラスター数が k_1 から k_2 に増加したときの群内平方和の総和の変化量と，クラスター数が k_2 のときの群内平方和の総和の比であり，さらにその比をデータの変量数 m で調整したものである．クラスター数決定にあたっては，クラスター数がある $k_1=k_2-1$ から k_2 への変化に対して急激に増加するとき，その k_2 の値をクラスター数の目安として選ぶ．

（基準3） Marriott の基準[5]

Marriott (1971) が提案した基準である．

$$C(k)=k^2|\boldsymbol{W}|/|\boldsymbol{T}| \tag{6}$$

ここで $\boldsymbol{W}$, $\boldsymbol{T}$ は，それぞれ群内分散・共分散行列，全分散・共分散行列で，$\boldsymbol{W}=\frac{1}{n-k}\boldsymbol{S}_W$，$\boldsymbol{T}=\frac{1}{n-1}\boldsymbol{S}_T$

である．また $|\boldsymbol{W}|$，$|\boldsymbol{T}|$ は，それぞれ行列 $\boldsymbol{W}$，$\boldsymbol{T}$ の行列式である．

クラスター数 k の増加に伴い単調に減少する統計量 $|\boldsymbol{W}|$ と，単調に増加する値 k_2 との積を基準量とし，これを $|\boldsymbol{T}|$ で規準化することにより，データ数による変動と相対的なクラスターの拡がりを調節したものである．

C 値が最小値，もしくは急激に減少すればそのときの k の値を，またいくつかの極小値があれば，そのときの k の値の最大値をクラスター数の目安として選ぶ．

（基準4） Maronna と Jacovkis の基準[4]

Maronna と Jacovkis (1974) が提案した基準である．

$$\psi^*=\psi/m(n-k) \qquad (7)$$

$$\text{ここで，}\ \psi=m\sum_{i=1}^{k}(n_i-1)\,|\boldsymbol{W}_i|^{1/m}$$

n_i はデータを k 群に分割したとき，第 i 群（第 i クラスター）のクラスター・サイズで，$\boldsymbol{W}_i$ は第 i 群の分散・共分散行列である．

ψ は k の増加に伴い単調に減少することが彼らにより示されており，ψ^* は，ψ のデータ数による変動を調整したもので，クラスター化によるデータの分離度を示す基準とされている．

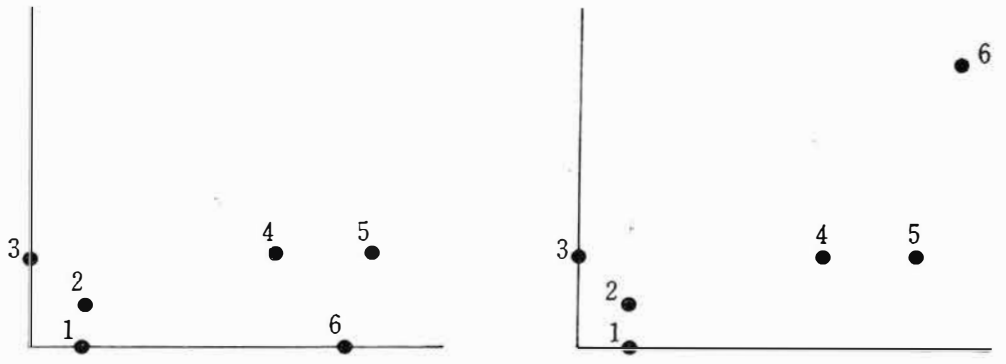
クラスター数の目安としては，ψ^* 値が k の増加に伴いもっとも急激に減少したときの k の値を選ぶ．

つぎに，簡単な例を使って，基準量からクラスター数の目安がどのように決定されるか，そのようすを調べることとする．

図2の a（データ(1)と呼ぶ），b（データ(2)と呼ぶ）について SL 法，CL 法によりクラスタリングをおこなうと，図2の c～f のデンドログラムを得る．ここで図2の c と d，あるいは e と f は，global order equivalent dendrogram であり，クラスター数をいくつとしても得られるデータの分割状態は同じである．このような場合には，クラスター数決定が主観的になりがちである（たとえば，デンドログラムの見栄えで決めてしまうなど）．

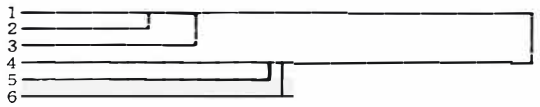
次にこれらのデータに Beale の基準 F を使用して得た結果が，図2の g，h である．これによると，SL 法，CL 法とも，データ（1）ではクラスター数2，データ（2）ではクラスター数3と観測される．このように基準量を使用することで，技法とクラスター数の関係について，その基準量による客観的，数値的な判断がある程度可能となる．

つまり，個体またはクラスター間の連結性だけに依存する技法によりデータを定量的にとらえ，次にクラスターを塊状のものとして評価する基準量からクラスター数を定量的にとらえるのである．また，これらを併用する



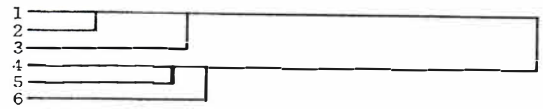
a. データ(1)の散布図　　**b.** データ(2)の散布図

図2　SL 法，CL 法を適用したとき global order eguivalent dendrogram となるようなデータの例

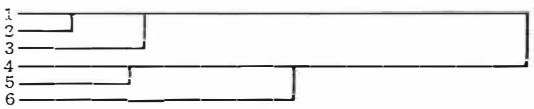


c. データ(1)にSL法を適用したときのデンドログラム

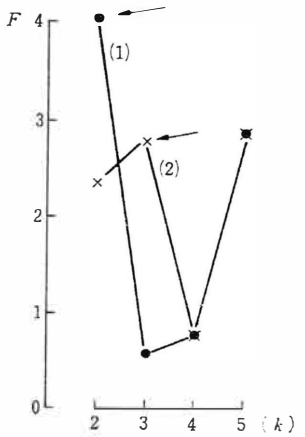
d. データ(2)にSL法を適用したときのデンドログラム



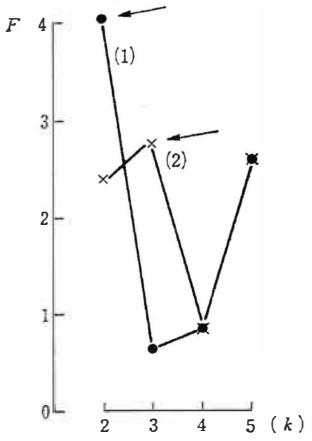
e. データ(1)にCL法を適用したときのデンドログラム



f. データ(2)にCL法を適用したときのデンドログラム



g. データ(1), (2)にSL法を適用したときのクラスター数kの変化と基準量Fの関係



h. データ(1), (2)にCL法を適用したときのクラスター数kの変化と基準量Fの関係

図2 （つづき）

ことにより，技法による分割の相違，あるいはクラスター化の過程を基準量の変化として観察できる．我々は技法と基準量とを併用するこの方法を“ハイブリッドな方法”と呼ぶこととする．

3. 感度分析によるクラスターの評価

さて次に，ハイブリッドな方法に，感度分析の考え方を取り入れて，与えられたデータに対する手法の感度あるいは頑健性を調べる手順を述べる．まず，3.1.では感度分析の目的とその必要性を説明し，また感度分析の具体的な手順を説明する．次に 3.2. では人工データを使って感度分析をおこない，その有効性を検討する．

3.1. 感度分析とその方式

2.2.では，クラスタリングに基準量を使用することで，クラスター数を定量的に決める一つの方式を述べた．さらにこの問題を掘りさげてゆくと，データの布置や性質は，技法や基準量によってどの程度掌握できるか（与えられたデータの布置の掌握），また逆に技法や基準量は，データの布置の変化に対しどの程度安定したものであるかという問題（技法の安定性，基準量の安定性）がある．このような点を分析するために，我々は感度分析を導入する．

感度分析とは，与えられたデータの布置をある程度保存するように，個々のデータに微小の乱数を加え，その

結果得られたデータに対してクラスタリングを行い，その結果と乱数を加える前のクラスタリングの結果を比べ，それに基づいて技法，基準量，データの検討をおこなう一連の体系をいうものである．

感度分析をおこなうことによって，技法と基準量がデータの布置の変化による影響の程度の観察を可能とすると同時に，技法間，基準量間の違いを比較できる．そのような結果は，技法や基準量の性質の解明につながるであろう．一方データについては，乱数を加えた影響がクラスター数にどのような変化を与えるかをみることにより，データの凝塊性の程度について，単に基準量だけを使用したときに比べ，克明に分析できることとなる．

以下クラスター化の技法としてSL法またはCL法を使用する場合の感度分析についてその手順を述べるが，それに先立って，感度分析に関連して，データの布置の保存度の計測について説明しておく．本来データにある乱れが加わったとき，この乱れにもかかわらずデータの凝塊性が保たれるならば，我々はその乱れのもとでもデータの布置が保存されたと言ってよいであろう．しかし乱数によって乱れを与えるときよほど小さい乱れでない限り，データの布置のパターンはある程度はくずれてくることになる．この点に関し，なんらかの方法でデータの布置のくずれの程度（逆にいえば保存の程度）を計測することが必要となる．我々はこの計測を，はじめのデータの距離行列の要素と乱数を加えたデータの距離行列の要素間の“順位相関係数”で行うことにする．順位相関係数の値が1に近い程データの布置の保存の程度は高く，逆にもとのデータの布置のパターンがくずれれば順位相関係数の値は1より離れるものとみられる．

ここで，とくに順位相関係数を使用する理由は次のとおりである．SL法，CL法では個体間の連結情報，距離行列の要素間の順位のみによりクラスタリングをおこなう．したがって，もとのデータの距離行列の要素間の順位が変化後のデータにおいても保存されている場合には，〔補足〕に述べるSL法とCL法の特性により，クラスター化の過程は2組のデータの間で不変で，クラスター数をいくつにしてもその分割結果は一致する．すなわち両データ間の順位相関係数の値が高いことが，データの布置の保存と対応している．

次に感度分析の手順を述べることにする．

（手順1）　1節で定義したデータ行列 $\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$ を使ってクラスタリングをおこない，k の変化に伴う各基準量を計算する．

（手順2）（手順1）のデータの各個体 $\boldsymbol{x}_i$ に多次元正規乱数 $\boldsymbol{e}_i \sim N(0, c\boldsymbol{I})$ を加えて，$\boldsymbol{y}_i$ $(i=1, \cdots, n)$ を作成する．ここで

$$\boldsymbol{Y}=(\boldsymbol{y}_1, \cdots \boldsymbol{y}_n)'=\boldsymbol{X}+(\boldsymbol{e}_1 \cdots, \boldsymbol{e}_n)'$$

なおバラツキに関する定数 c はデータの布置の保存度と関連する量であるが，我々はこの値を，$\boldsymbol{X}$ にもとづく距離行列 (d_{ij}) と $\boldsymbol{Y}$ にもとづく距離行列 (d_{ij}^*) の順位相関係数によって制御するようにする．またここで $\boldsymbol{0}$ はゼロベクトルであり $\boldsymbol{I}$ は単位行列である．

（手順3）　$\boldsymbol{y}_i$ $(i=1, \cdots, n)$ を使ってクラスタリングし，各基準量を計算する．

（手順4）（手順2），（手順3）を繰り返す．

（手順5）（手順3）で計算した基準量を，各クラスター数，各基準量ごとにまとめ，その挙動を観察，検討する．

3.2. 人工データによる実験

3.2.1. 実験の目的

クラスター化の技法，基準量がデータの潜在的傾向をとらえる程度は，技法間，また基準量間で差異があるであろう．それは与えられたデータの布置のちがいとも関連してくる．この場合，どのようなデータに対してどの

表1

個体番号	データ(A)		データ(B)	
	x_1	x_2	x_1	x_2
1	0	28	56	31
2	55	1	56	37
3	13	19	30	9
4	43	39	22	1
5	64	33	13	16
6	10	3	3	18
7	21	10	29	27
8	26	30	26	24
9	61	18	49	43
10	77	45	61	32
11	68	36	55	37
12	6	27	63	45
13	77	1	54	47
14	73	37	63	39
15	8	44	30	4
16	10	15	27	9
17	74	25	8	11
18	34	46	10	16
19	16	39	37	18
20	33	28	43	15
21	29	5	50	38
22	7	23	56	32
23	46	10	55	43
24	63	16	54	43
25	4	7	25	5
26	59	17	18	5
27	5	27	10	19
28	21	9	3	19
29	24	4	30	27
30	55	7	26	24
31	35	29	45	11
32	54	12	42	17
33	17	40	49	43
34	56	41	55	48
35	76	39	13	17
36	49	44	2	12
37	57	18	26	20
38	4	5	31	27
39	27	24	26	28
40	2	42	24	26
41	16	6	61	44
42	18	49	53	33
43	48	28	38	11
44	26	20	37	12
45	12	41	55	38
46	31	43	49	43
47	61	37	24	4
48	12	33	25	4
49	52	1	37	18
50	60	46	36	18

技法および基準量が有効かという点の検討が重要な問題となる．我々の意図する実験は，1つにはこのような点の検討のためのものである．

現実のデータのパターンは多様であるから，それに応じてさまざまなものを考えるということはかえって分析を混乱させてしまう．ここでは2つのはっきり違った種類のデータに限定して検討することとする．これらは 1) クラスターと呼べる領域があいまいなデータ， 2) 明確な領域をもつクラスターの集まりが存在すると考えられるデータ，の2種類のデータである．

前者をデータ（A），後者をデータ（B）（ともに $n=50$, $m=2$ で表1にそれらのデータを挙げた）となづける．

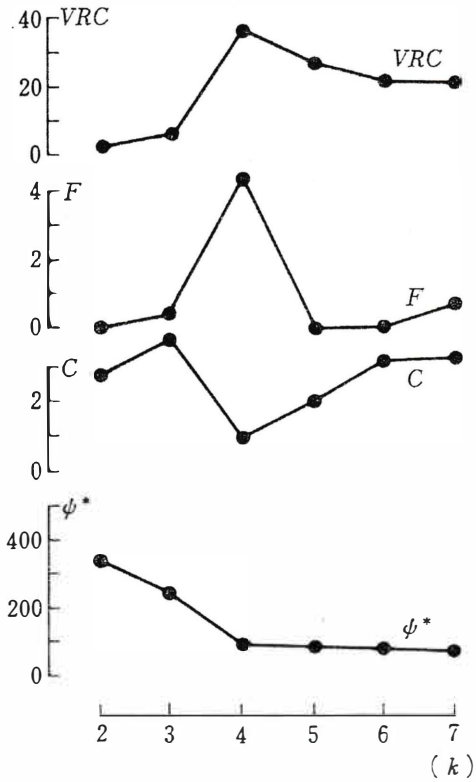
これらのデータに対して感度分析を行い，データに乱数を加えたときおこる変化を分析することにより，取り上げる技法間および基準量間の効力や限界をはっきりと比較できる．

感度分析の方法の今一つの目的は，データの凝塊性について情報を与えることである．凝塊性の強い（我々の考える意味でのクラスターが明確な）データほど，乱数を加えることによる乱れの影響を受けにくく安定度が高い，逆に凝塊性が弱くクラスターのはっきりしないデータほど，乱数による乱れに対し敏感である．すなわち凝塊性と感度の間には，ある種の関係がある．したがって感度分析は凝塊性の解析に利用できるわけである．

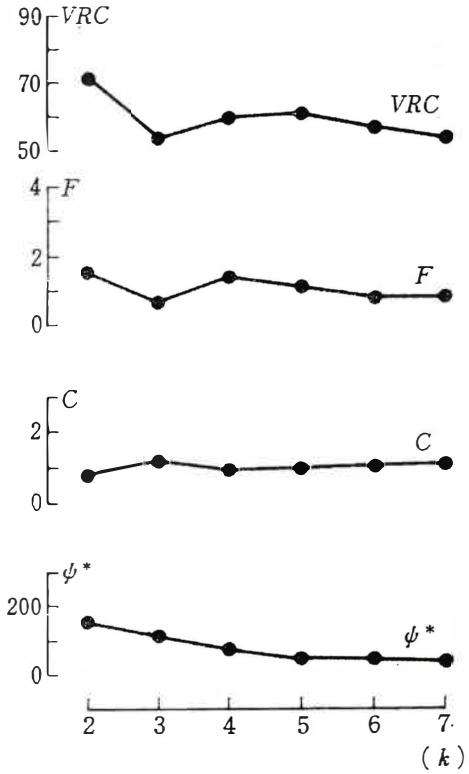
ここで実験結果の考察に入る前に，上述のデータ（A），(B) を用いて実際に実験を行うときの手順について若干の補足をする．

第1に乱数（多変量正規乱数）を加えるとき，データの布置がある程度保存されている場合と，それが次第にくずれていく場合とを観測したい．前に述べたように，データの布置が完全に保存されている場合には，順位相関係数は1である．しかしもとのデータの距離行列に1つでも同位の距離があれば，乱数を加えたデータの距離行列との順位相関係数は1とはならない．そこで，この場合やむをえず順位相関係数を 1と $1-\varepsilon$ の間にあるように乱数を加えることとし，十分小さな値 ε を実験から推量して与えた．またデータの布置がくずれた場合を調べるためには，この ε を段階的に大きくしていくこととした．

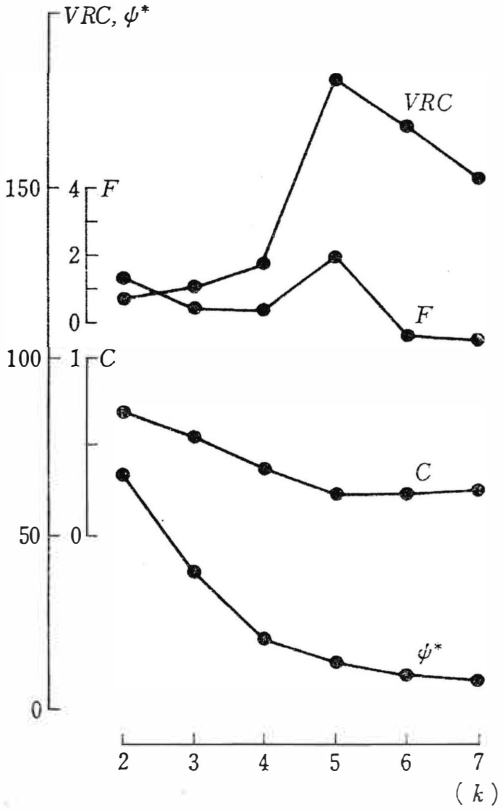
第2に（手順4）の繰り返し回数であるが，現用の大型計算機の能力を考えて100回前後とした．



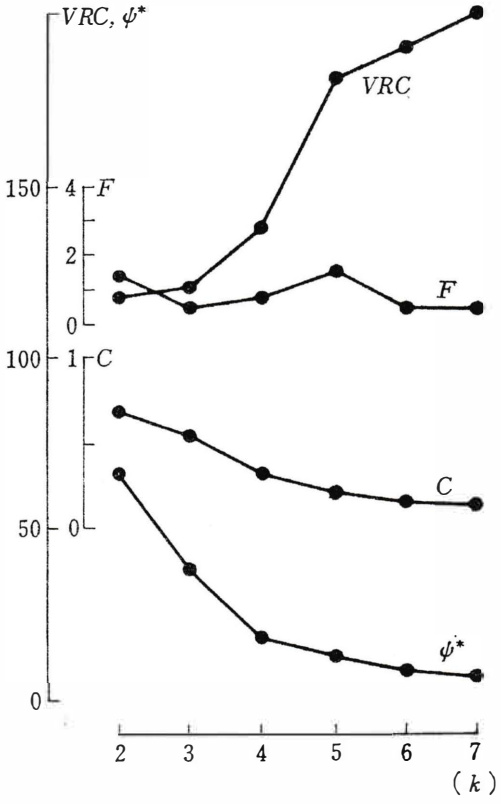
a. データ(A)にSL法を適用したとき



b. データ(A)にCL法を適用したとき



c. データ(B)にSL法を適用したとき



d. データ(B)にCL法を適用したとき

図3　クラスター数 k の変化と基準量の関係

3.2.2. 実験結果と考察

まず図3は，データ（A），（B）について，クラスター数kの変化に伴う基準量の挙動を示したものである．

既述のように VRC 値と F 値ではその値が急激に増加したところを，C 値と ϕ^* 値ではその値が急激に減少したところを，クラスター数の目安とするわけである．したがってデータ（A）に SL 法を使用したときは，どの基準量によっても4群となる．一方 CL 法を使用したときは SL 法を使用したときに比べ，基準量の挙動はそれほど明確でない．すなわち実験結果から見る限りでは，ここでいうクラスターと呼べるような領域があまりはっきりしないデータに対するクラスター化の技法としては，SL 法の方が CL 法よりも適しているということになる．しかし実際問題として，独自の目的をもつ場合は，杓子定規な技法の選択は危険であり，この問題は技法間の関連性として捕えねばならないが，これについてはあとで述べる．

データ（B）に SL 法を使用したときは，基準量 VRC 値，F 値，C 値で5群，ϕ^* 値で4群，CL 法を使用したときは，基準量 VRC 値，F 値で5群，C 値，ϕ^* 値で4群となり，どの基準量もその挙動はかなり明確である．すなわちこの場合，どの基準量も有効である．これはデータ（B）の性格から，ある意味では当然の結果である．

以上からクラスター数の決定と同時に，クラスターの構成は技法ごとに自動的にきまる．データ（A）を SL 法で4群に分割し，分類散布図をかいたものが図4の a である．なお後の分析のため，データ（A）を CL 法で4群に分割した結果も図4の b にあげた．データ（B）を5群に分割するときは，SL 法，CL 法とも同じ結果を示し，その分類散布図は図4の c となる．

つぎに感度分析を行うが，乱数の付加によりデータの布置がどの程度影響を受けたか，その保存の度合をクラスターの“一致率”で調べる．ここでクラスターの一致率とは以下のものをいう．

データを $\boldsymbol{x}_l$ $(l=1,\cdots,n)$ とし，$\boldsymbol{x}_l$ に乱数を加えたものを $\boldsymbol{y}_l$ とする．$\boldsymbol{x}_l$ $(l=1,\cdots,n)$ および $\boldsymbol{y}_l$ $(l=1,\cdots,n)$ を互いに排反な k 個のクラスターに分割したとき，それぞれのクラスターを C_i，C_i' $(i=1,\cdots,k)$ とする．いま

$$S_{ij}=\{l \mid C_i \ni \boldsymbol{x}_l,\ C_j' \ni \boldsymbol{y}_l\}$$

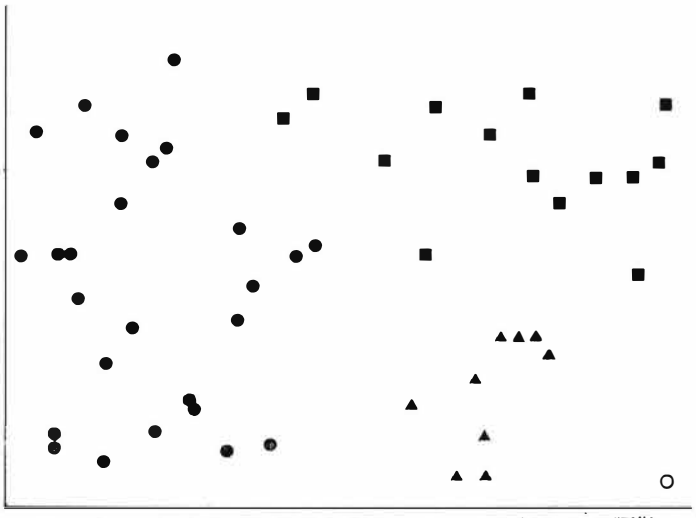
とし，$\bar{S}_{ij}$ を S_{ij} の要素の数とするとき，$\sum_{i=1}^{k}\bar{S}_{ii}$ が最大になるように C_i' の添字 i を適当につけかえる．このとき $\sum_{i=1}^{k}\bar{S}_{ii}/n$ をクラスターの一致率とした．この一致率を使用し，データの布置の保存の度合を調べる．

前述の結果より，データ（A）を4群，データ（B）を5群とし，乱数を加えたデータとの対応を求め，データ（A）について9回，データ（B）について10回の実験合計を表2にした．また表にはあらわさなかったが，データ（A）において，9回の実験の順位相関係数の平均が 0.9998 のとき，一致率は SL 法で1，CL 法で 0.98 であり，データ（B）において，10回の実験の順位相関係数の平均が 0.9994 のとき，一致率は SL 法，CL 法とも1であった．

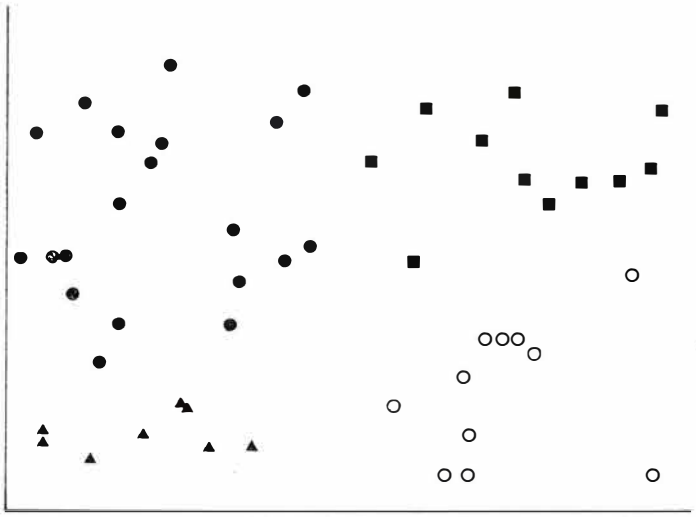
これと表2より，順位相関係数が1よりわずかに小さくなっただけで，一致率は急激に減少していることがわかる．したがって，データの布置のパターンが保存されるのは順位相関係数がかなり1にちかい場合である．

この性質を考慮して，データの布置がある程度保たれていると思われる場合，およびそれを徐々に崩していった場合について感度分析をおこなう．

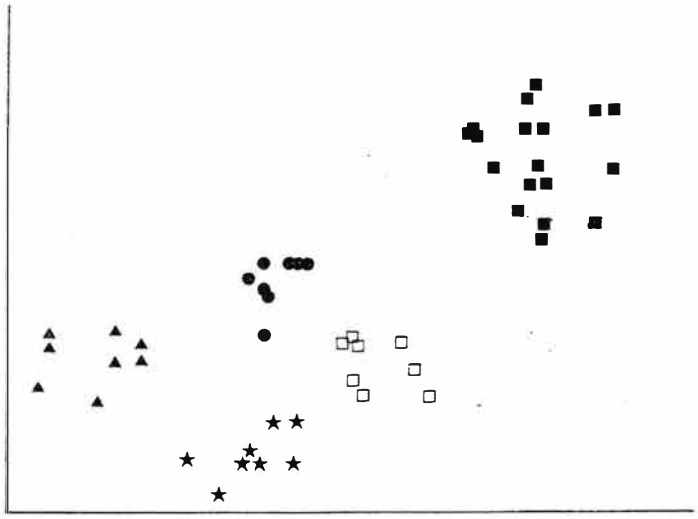
基準量の性質，感度を調べるため，SL 法を使用してデータ（B）をクラスタリングしたときの，基準量の平均値の挙動を図5，図6の c に示した．これらは表3の記号を用い，それぞれの順位相関係数で制御しながら



a. データ(A)に SL 法を使用して4群に分割した分類散布図



b. データ(A)に CL 法を使用して4群に分割した分類散布図



c. データ(B)に SL 法または CL 法を使用して5群に分割した分類散布図

図4

表 2　データに乱数を加えたとき，もとのデータと乱数を加えたデータとの順位相関係数の平均 $\bar{r}$ と，両者をクラスタリングしたときのクラスターの一致率

データ（A）

	C_1'	C_2'	C_3'	C_4'	計
C_1	211	8	15		234
C_2	18	63			81
C_3	81	9	36		126
C_4				9	9
計	310	80	51	9	450

SL 法を適用したとき
$\bar{r}=0.9842$，一致率＝0.7089

	C_1'	C_2'	C_3'	C_4'	計
C_1	234				234
C_2		81			81
C_3	38		88		126
C_4				9	9
計	272	81	88	9	450

SL 法を適用したとき
$\bar{r}=0.9993$，一致率＝0.9155

	C_1'	C_2'	C_3'	C_4'	計
C_1	123	51		6	180
C_2	8	64			72
C_3	1		93	5	99
C_4	11	4	6	78	99
計	143	119	99	89	450

CL 法を適用したとき
$\bar{r}=0.9842$，一致率＝0.7956

	C_1'	C_2'	C_3'	C_4'	計
C_1	159	21			180
C_2		72			72
C_3			96	3	99
C_4	4			95	99
計	163	93	96	98	450

CL 法を適用したとき
$\bar{r}=0.9993$，一致率＝0.9378

データ（B）

	C_1'	C_2'	C_3'	C_4'	C_5'	計
C_1	168	2		8	2	180
C_2		50	3	10	17	80
C_3	1	24	45	6	4	80
C_4		41		22	17	80
C_5		40	2	9	29	80
計	169	157	50	55	69	500

SL 法を適用したとき
$\bar{r}=0.9527$，一致率＝0.6280

	C_1'	C_2'	C_3'	C_4'	C_5'	計
C_1	180					180
C_2		58	1		21	80
C_3			74		6	80
C_4				79	1	80
C_5		46			34	80
計	180	104	75	79	62	500

SL 法を適用したとき
$\bar{r}=0.9877$，一致率＝0.8500

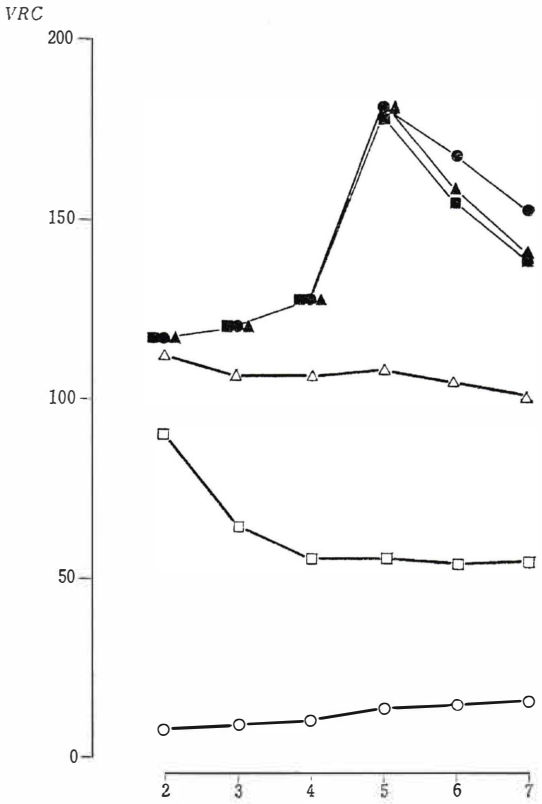
	C_1'	C_2'	C_3'	C_4'	C_5'	計
C_1	173				7	180
C_2		72		3	5	80
C_3		8	68	4		80
C_4		3		74	3	80
C_5		8		11	61	80
計	173	91	68	92	76	500

CL 法を適用したとき
$\bar{r}=0.9527$，一致率＝0.8960

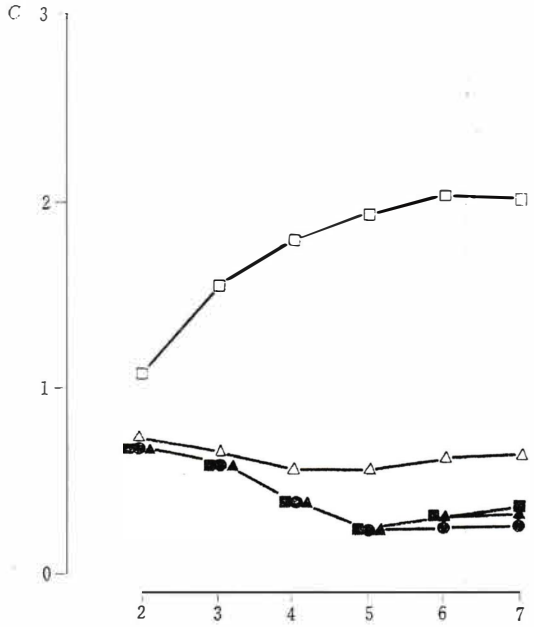
	C_1'	C_2'	C_3'	C_4'	C_5'	計
C_1	180					180
C_2		78			2	80
C_3			80			80
C_4		1		79		80
C_5		4		3	73	80
計	180	83	80	82	75	500

CL 法を適用したとき
$\bar{r}=0.9877$，一致率＝0.9800

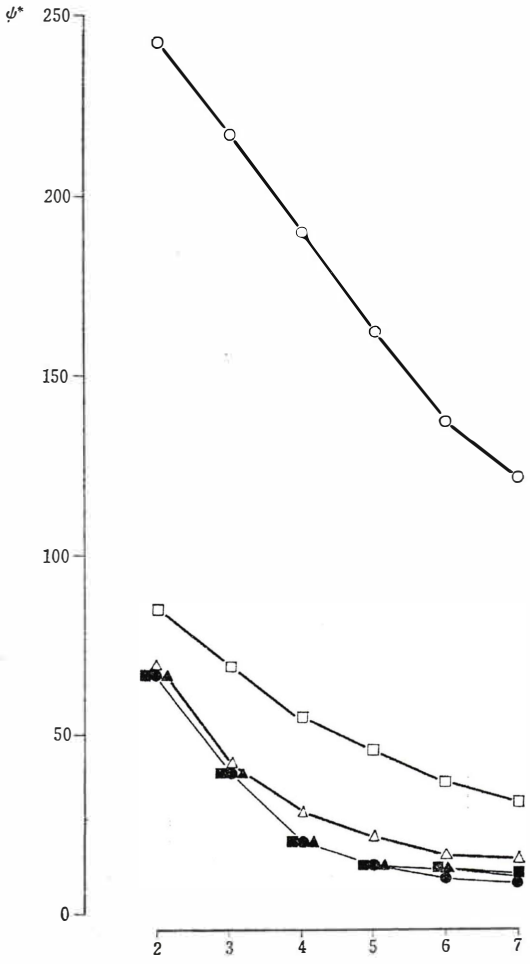
図5　順位相関係数に対する基準量の挙動について



a. データ(B)にSL法を適用したとき



b. データ(B)にSL法を適用したとき



c. データ(B)にSL法を適用したとき

表3　$\bar{r}$　;順位相関係数の平均
$D(r)$;順位相関係数の標準偏差
(表中rはrである)

データ	記　号	$\bar{r}$	$D(r)$
A	●	——	——
	■	.9992	.0001
	△	.9820	.0028
	□	.9317	.0101
	○	.6773	.0415
B	●	——	——
	▲	.9998	.0000
	■	.9994	.0001
	△	.9869	.0022
	□	.9498	.0083
	○	.7419	.0407

［注］　記号●は乱数を加えない場合

おこなったものである．さらに表4は，基準量の平均値と標準偏差をまとめたものである．これらにもとづいて基準量の感度を検討する．

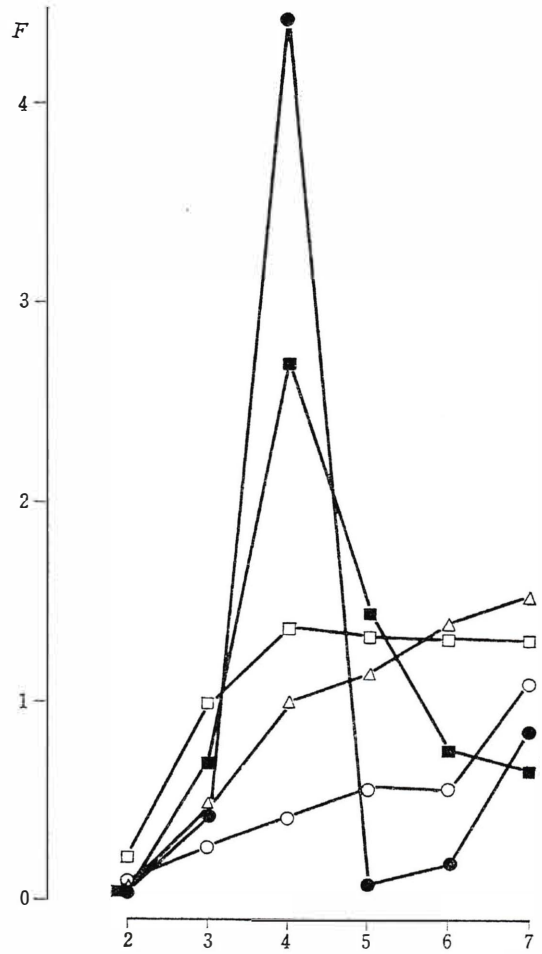
VRC 値の挙動を図5の a と表4でみると，順位相関係数が 0.9994 までは，もとのデータにおける挙動との差異はほとんどみられない．これを基準量の標準偏差でみると，クラスター数2，3，4に比較し，クラスター数5で急激な増加がみられる．これは *VRC* 値がこの基準で決定されるクラスター数のところで，データの凝塊性に応じて敏感に反応しているためである．

F 値の方は，データの凝塊性に対し，*VRC* 値よりもさらに敏感に反応している．これを図5の a と図6の c を使用して説明する．図5の a において，各順位相関係数別に k の増加に伴う *VRC* 値の挙動を観測すると，

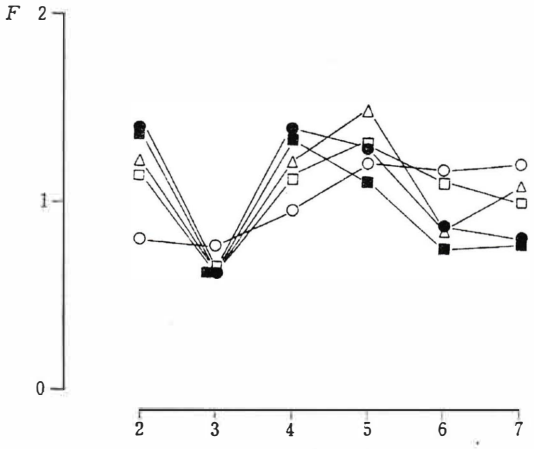
表4　データ（B）に乱数を加えたものを，クラスタリングしたとき基準量の挙動の統計量と順位相関係数の関係について（ＳＬ法の場合）（表中γはγである）

	$\bar{\gamma}$ / クラスター数	乱数を加えない場合 基準量の値	0.9994 C.M	0.9994 C.S.D	0.9869 C.M	0.9869 C.S.D	0.9498 C.M	0.9498 C.S.D	0.7419 C.M	0.7419 C.S.D
VRC	2	117.158	116.194	1.225	112.396	5.576	89.852	32.820	7.169	18.283
	3	120.287	119.922	1.663	106.542	15.987	64.837	23.756	9.033	16.367
	4	127.644	127.083	1.987	106.395	19.886	55.428	19.261	10.360	15.626
	5	181.316	178.383	11.001	108.111	32.347	55.606	23.077	13.456	20.162
	6	167.874	154.653	13.022	104.894	31.026	54.361	21.899	14.664	18.493
	7	152.979	138.839	10.100	100.725	26.127	54.812	21.197	15.603	17.437
F	2	2.343	2.338	0.024	2.248	0.115	1.797	0.656	0.143	0.365
	3	1.463	1.460	0.027	1.234	0.371	0.707	0.966	0.366	0.669
	4	1.446	1.440	0.034	1.256	0.821	0.804	1.316	0.615	1.161
	5	3.008	2.934	0.377	1.250	1.176	1.123	1.412	1.151	2.001
	6	0.761	0.446	0.230	1.047	1.193	1.011	1.195	1.191	2.173
	7	0.583	0.535	0.629	0.997	1.206	1.165	1.322	1.098	1.806
C	2	0.686	0.686	0.008	0.709	0.042	1.083	0.904	3.411	0.630
	3	0.556	0.558	0.011	0.641	0.164	1.546	1.287	6.272	1.951
	4	0.369	0.372	0.011	0.585	0.273	1.813	1.131	8.934	3.785
	5	0.223	0.231	0.041	0.595	0.341	1.943	1.406	10.504	5.849
	6	0.235	0.286	0.067	0.639	0.383	2.024	1.314	11.406	7.275
	7	0.257	0.331	0.054	0.644	0.319	2.016	1.439	12.714	9.508
ψ^*	2	66.294	66.431	0.697	69.238	3.569	88.268	32.010	243.653	43.810
	3	39.474	39.595	0.537	42.324	5.632	69.041	24.792	217.141	54.609
	4	19.783	19.898	0.388	26.725	8.038	55.411	17.264	190.130	57.551
	5	13.776	13.946	0.646	20.790	5.983	44.407	16.882	160.201	60.428
	6	10.417	12.323	1.436	17.813	5.195	36.655	14.316	137.155	58.541
	7	9.130	11.057	1.627	15.175	3.549	30.292	12.674	121.406	57.606

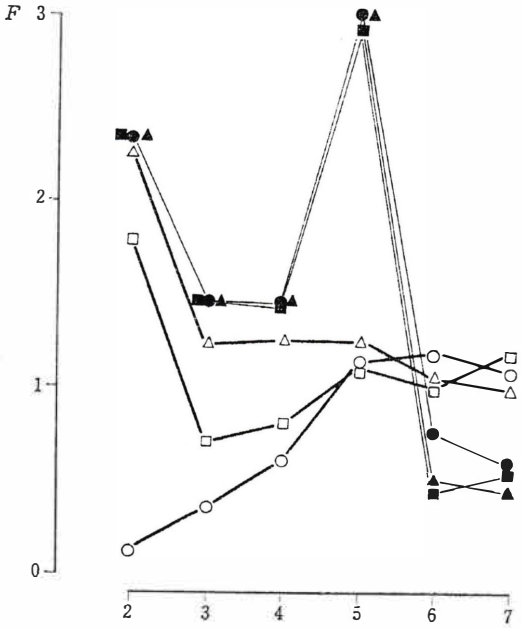
$\bar{\gamma}$：順位相関係数の平均　　C.M：基準量の平均　　C.S.D：基準量の標準偏差



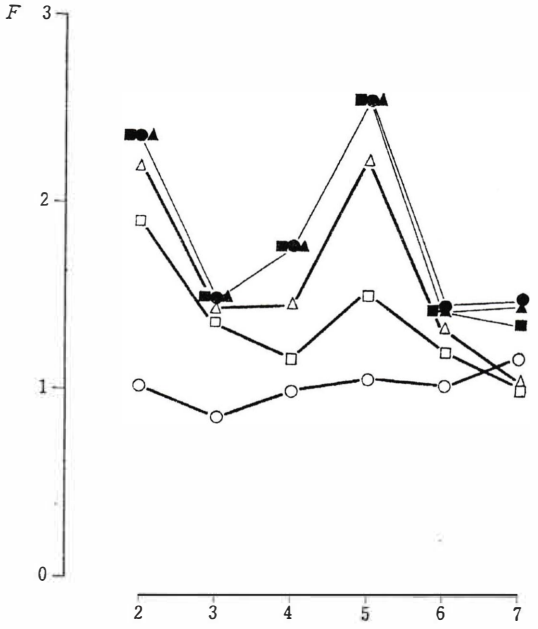
a. データ(A)にSL法を適用したとき



b. データ(A)にCL法を適用したとき



c. データ(B)にSL法を適用したとき



d. データ(B)にCL法を適用したとき

図 6　順位相関係数に対する基準量の挙動について

順位相関係数の平均値が 0.9869（図中の記号：△）まで，クラスター数5と判定される．ところが図6の c の F 値について同様の観測をおこなうと，クラスター数が5と判定できるのは，順位相関係数の平均値が 0.9994（図中の記号：□）までである．以上より F 値は VRC 値に比較して，データの凝塊性に一層敏感に反応していることがわかる．

同様に C 値，ϕ^* 値の挙動を図5の b，c でみると，ともにデータの凝塊性に対する反応はしているが，上述の F 値，VRC 値に比較し，それほど敏感ではない．

つぎにデータの凝塊性にとくに敏感な基準量 F 値のみを使用し，SL 法と CL 法を比較検討する．

SL 法について図6の a，c をみると，順位相関係数の平均値が，データ（A）において 0.9992 まで，データ（B）において 0.9994 までが，それぞれもとのデータにおける基準量の変化の傾向を残しているが，順位相関係数の平均値を，データ(A）において 0.9820，データ（B）において 0.9869 にすると，その類似性がくずれてしまう．CL 法について同様な比較をおこなうと，順位相関係数の平均値は，データ（A）では SL 法と同じであるが，データ（B）では 0.9869 の場合まで，もとのデータにおける基準量の変化の傾向が残っている．

これは，SL 法が CL 法に比較し，データ間の距離の変化に敏感で乱数の影響を強く受けるためであり，また距離のはなれたデータが1つあっても（つまり異常値のようなデータ），それに影響され，そのデータ1つを1つのクラスターとして検出する傾向があるためである．

それに対し CL 法は，データを塊状のクラスターがあるかの如くとらえる傾向がみられ，真に塊状のクラスターが存在するときには，SL 法より CL 法の方がクラスタリングの技法として良い．しかし，このことは単に SL 法より CL 法がすぐれているという意味ではない．むしろ，2技法の性質をみると，データの大勢をまとめ少数の異常値を検出することを目的にするなら SL 法を使用することが適切であることを，実験は顕著に示しているといえる．これはデータをどのように分割したいか，どのようなクラスターを目的としているかにより，技法の選択をおこなうべきであることに他ならないが，感度分析は，これに対する目安を与えるものである．

技法と基準量によりデータを分割し，その分割数の決定の手掛りがえられたが，ここでは，さらに一歩すすめて，感度分析をおこなうことで，データの凝塊性についてどの程度の情報がひきだせるかを検討する．

技法として CL 法を使用し，データの凝塊性にとくに敏感な基準量 F 値を使用する．

図6の b，d をみると，順位相関係数の平均値が，データ（B）では 0.9869 までもとのデータにおける基準量の変化の傾向を残しているが，データ（A）では 0.9992 までしか，その傾向を残していない．これは，データ（B）が明確なクラスターに分割されていると思われるデータのため，乱数による影響は，クラスター内にとどまり，クラスター単位で考えるときは，その影響があらわれないためである．それに対し，明確なクラスターの集まりと考えられないデータ（A）は，乱数の影響により，各個体がクラスター間の遷移を生じたり，ときにはクラスター数まで変化するなどの現象がおこるからである．

結局，感度分析をおこなって基準量の変化の傾向がどの程度保存されるかを順位相関係数で追跡することにより，データの凝塊性についての検討が可能となる．

4. むすび

以上の結果をまとめると，つぎの知見を得る．

（1） クラスター数決定の目安としての基準量は，データのある種のくせ（塊状の構造か否か）を把握する方法

として有効である．

（2） SL 法はデータの布置の変化に敏感である．

（3） SL 法はデータの大勢をまとめ，異常値とみられるデータの検出などに有効である．

（4） CL 法はデータの凝塊性からのみだれに対し安定している．

（5） ここでいうクラスターを検出するには，SL 法より CL 法の方が適している．

（6） 感度分析は，基準量の凝塊性に対する感度，技法の性質と安定性，データの凝塊性等の検討に有効である．

〔補足〕 技法によるデータの分割の差異を検討することは，技法の性質を把握するうえにおいて重要である．この目的の一端を担うものに解の同定性がある．ここでは与えられたデータの解の同定性の範囲（つまり，データの布置を中心に個々のデータを移動させるとき，分割の状態またはクラスターの連結状態が同じになるようなデータの移動許容範囲）を求めることで，これについてつぎに1つの結果を与えておく．

n個の2次元データのうち，1個体のみを移動させ，SL 法または CL 法を使ってクラスタリングをおこなうとき，global order equivalent set となるような1個体の存在領域は，円と2個体の垂直二等分線により分割できる．

ただし，個体間の距離に同位（tie）がないものとする．

上記は，つぎのように証明できる．

n個の個体を $P_1, P_2, \cdots, P_n$ とし，P_i と P_j の距離を d_{ij}，個体間の距離の順位行列を $\{O_{ij}\}$ とする．

$$d_{ij} < d_{kl} \iff O_{ij} < O_{kl}$$

d_{ij} に同位の距離がないので，SL 法，CL 法の定義から，$\{d_{ij}\}$，$\{O_{ij}\}$ にこれらの技法を使ってクラスタリングをおこなった場合，個体間の連結順序は双方とも同じになる．

したがって題意を満足する領域は，個体間の距離行列の順位が同じになるような1個体の存在領域と同じになる．

いま $P_1, P_2, \cdots, P_{n-1}$ を固定すれば，

$$d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_k j_k} \qquad 1 \leqq i_l,\ j_l \leqq n-1,\ (1 \leqq l \leqq k),\ k=(n-1)(n-2)/2$$

が成立している．これと d_{ni} $(i=1, 2\cdots, n-1)$ の順位がきまれば，そのときの領域が求めるものであり，この順位の比較はつぎの (イ) (ロ) (ハ) の3通りを考えればよい．

(イ) d_{ni} と d_{nj} の比較

P_i と P_j の垂直二等分線により領域を分ければ，その領域に P_i，P_n があるか，P_j，P_n があるかにより，d_{ni} と d_{nj} の大小が決定される．

(ロ) d_{ni} と d_{ij} $(j \neq n)$ の比較

P_i を中心とし半径 d_{ij} の円により領域を分割すれば，P_n が円の内側，あるいは外側にあるかにより，d_{ni} と d_{ij} の大小が決定される．

(ハ) d_{ni} と d_{jl} $(j, l \neq i, n)$ の比較

P_i を中心とし半径 d_{jl} の円により領域を分割すれば，P_n が円の内側，あるいは外側にあるかにより，d_{ni} と d_{jl} の大小が決定される．

以上 (イ)，(ロ)，(ハ)による領域の分割は，円と2個体の垂直二等分線を使用しており，この分割による領域内に P_n を入れれば，その位置にかかわらず $\{d_{ij}\}$ の順位は一定となる．

上の結果は，任意の次元の場合に対しつぎのように拡張される．

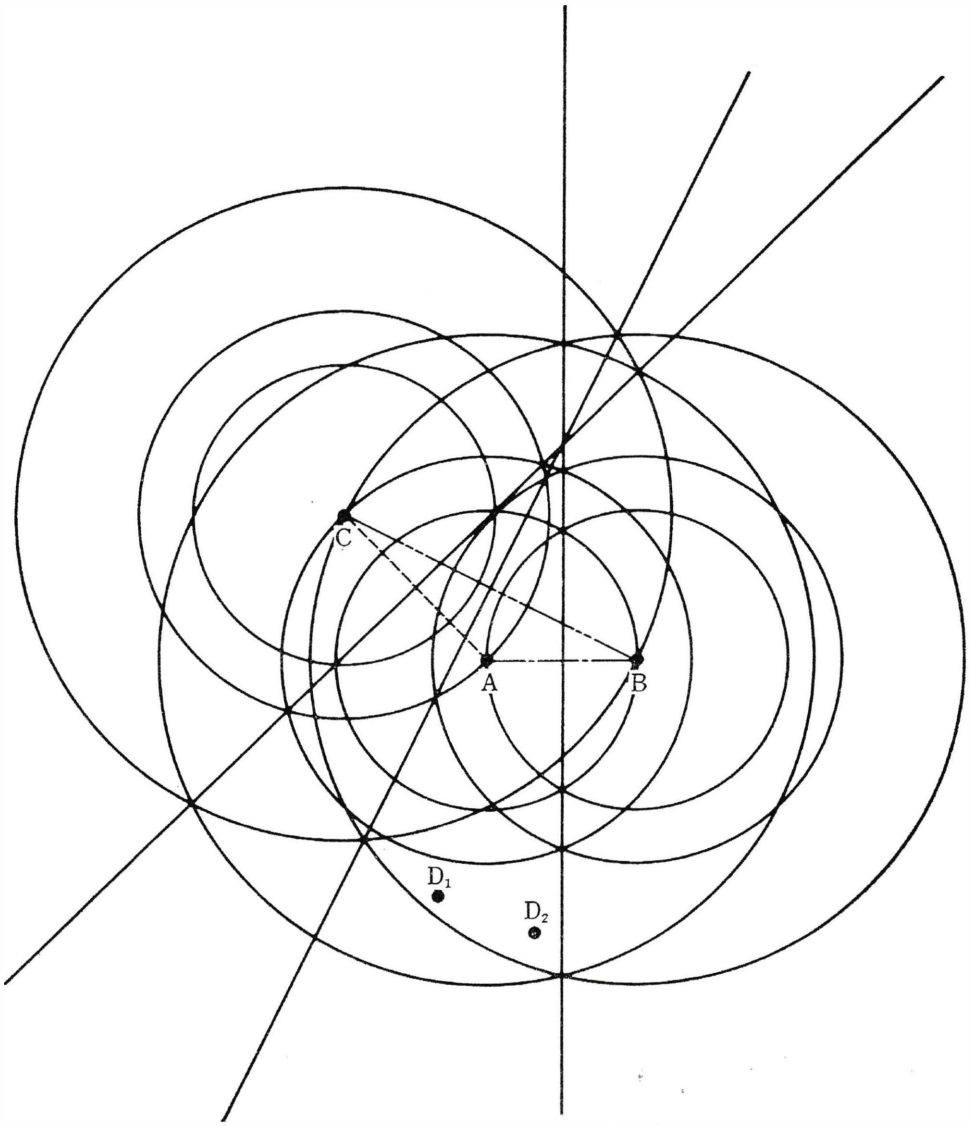
n個の多次元データのうち，1個体のみを移動させ，SL 法または CL 法を使ってクラスタリングをおこなうとき，global order equivalent set となるような1個体の存在領域は，多次元の球と2個体を垂直二等分する超平面により分割できる．

ただし，個体間の距離に同位がないものとする．

これは2次元データの場合の証明において，円を多次元の球，垂直二等分線を2個体を垂直二等分する超平面とすれば明らかである．

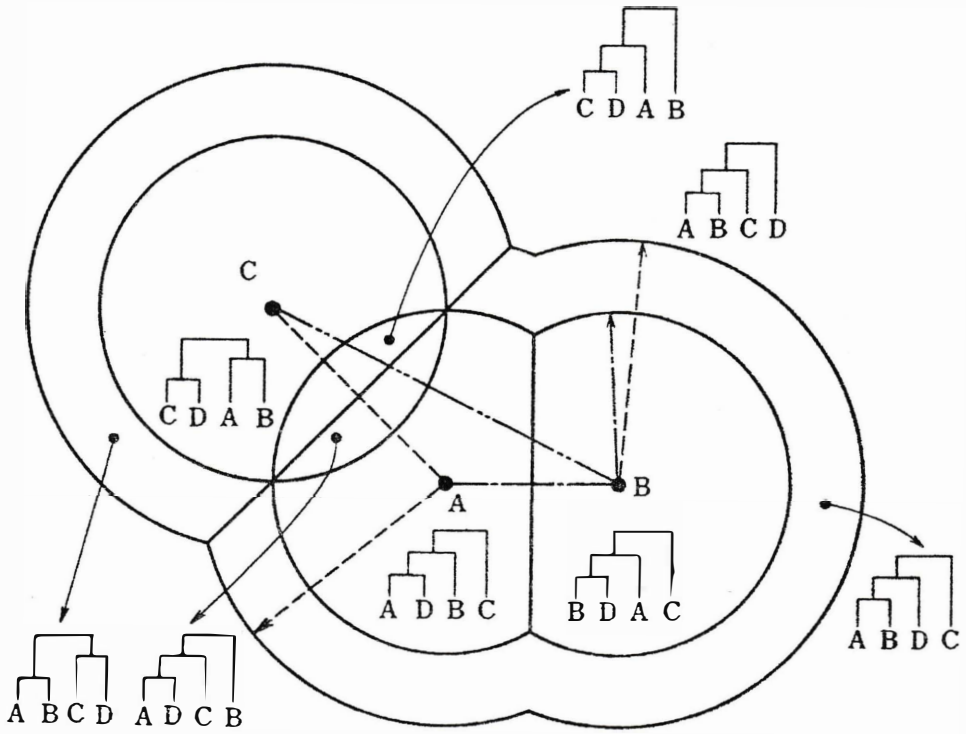
つぎに，2次元の簡単な例を使用して上述の結果を図示しておく．

図7の a は $n=4$ の場合で，個体 A，B，C を固定したとき，個体間の距離の順位が一定となるような個体 D の存在許容範囲である．さらに SL 法，および CL 法を使用してクラスタリングを行なった場合，global

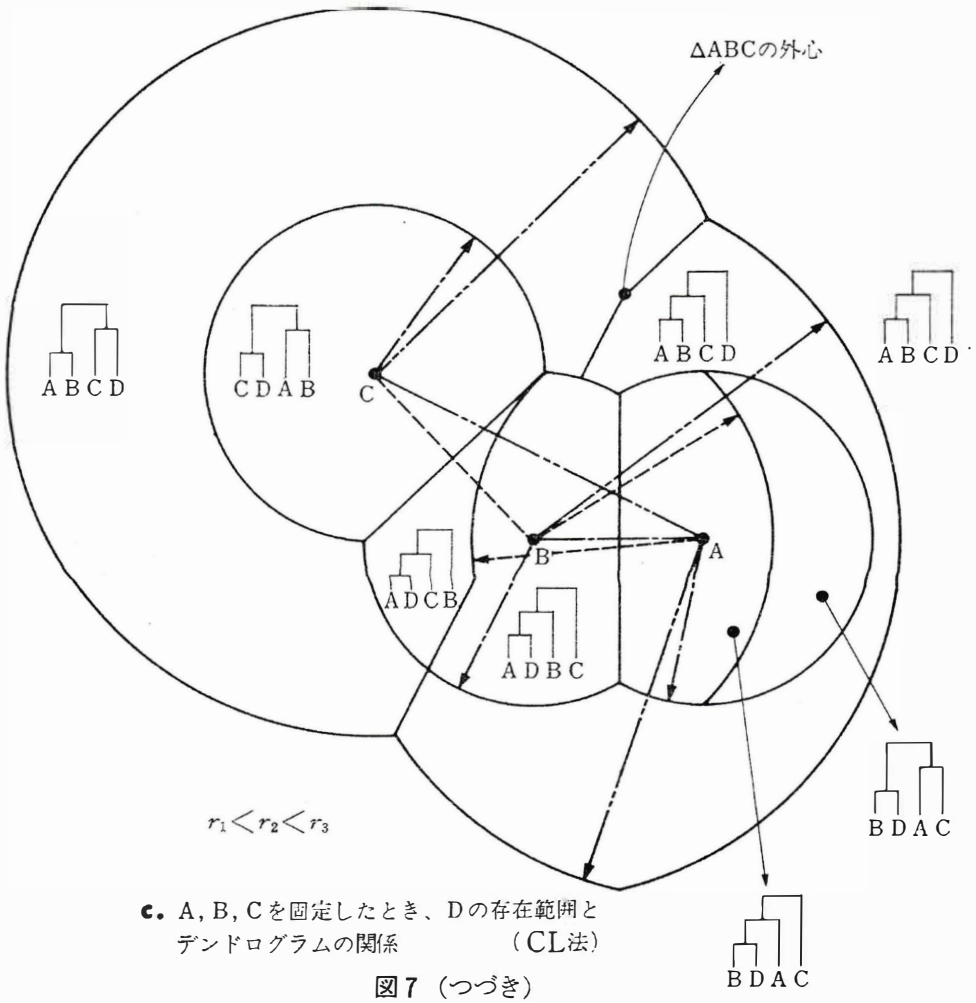


a. A, B, Cを固定したとき，固体間の距離の順位が一定となるようなDの存在許容範囲（線によって囲まれた範囲）．たとえば A，B，C，D_1 の固体間の距離の順位と，A，B，C，D_2 の固体間の距離の順位は一致する。

図 7



b. A, B, Cを固定したとき，Dの存在範囲とデンドログラムの関係 (SL法)



c. A, B, Cを固定したとき、Dの存在範囲とデンドログラムの関係 (CL法)

図7 (つづき)

order equivalent dendrogram となる個体 D の存在許容範囲を，図7の b, c に掲載した．ただし，c では r_1:点 A, B 間の距離．r_2:点 B, C 間の距離，r_3:点 C, A 間の距離である．

参 考 文 献

〔1〕 Beale, E. M. L. (1969) : Euclidean cluster analysis, *Bull. I. S. I.*, Vol. 43, Book 2, 92-94.

〔2〕 Calinski, T. and Harabasz, J. (1974) : A dendrite method for cluster analysis, *Communication in Statistics*, Vol. 3, No. 1, 1-27.

〔3〕 Lance, G. N. and Williams, W. T. (1967) : A general theory of classificatory sorting strategies : I. hierarchical system, *The Computer Journal*, Vol. 9, 373-380.

〔4〕 Maronna, R. and Jacovkis, P. M. (1974) : Multivariate clustering procedures with variable metrics, *Biometrics*, Vol. 30, 499-505.

〔5〕 Marriott, F. H. C. (1971) : Practical problems in a method of cluster analysis, *Biometrics*, Vol. 27, 501-514.

〔6〕 Matusita, K. and Ohsumi, N. (1978) : Evaluation procedure of clustering techniques, France-Japan Seminar, Paris.

〔7〕 Sibson, R. (1972) : Order invariant method for data analysis, *Journal of the Royal Statistical Society*, Series B, Vol. 34, 311-349.