Title	名義尺度型データ処理の一方法 : 多次元属性空間縮小の計量的手法
Sub Title	An approach to multivariate analysis of nominal scales : some statistical methods for reducing multidimensional attribute space
Author	堀内, 四郎(Horiuchi, Shiro)
Publisher	三田哲學會
Publication year	1973
Jtitle	哲學 No.61 (1973. 10) ,p.157- 185
JaLC DOI	
Abstract	In the area of sociological measurement, it is one of the most fundamental problems how data consisting of a large number of nominal scales can be analyzed and reduced. For data collected in social research are usually obtained by forcing respondents to select one out of categories listed on questionaire. It seems to the present writer that there are two kinds of approach to this problem. They are the followings : 1) This type of data can be summarized by introducing the composite variable α, [numerical formula] w_jk=the value given to the k-th category of the j-th variable. When a criterion function of w_jk, which is to be maximized or minimized in terms of the purpose of analysis, is considered, the value of w_jk is calculated so that the given criterion function is optimized. This type of method, "the Hayashi quantification theory," is well known among behavioral scientists in Japan, which is named after the founder of this methodology. 2) This type of data can also be summarized by means of the new nominal scale, which is constructed i) by selecting a number of variables out of all given nominal scales, ii) by grouping some of categories which belong to the same variable selected, and iii) by combining categories which belong to the different variables. This is equivalent to grouping ultimate classes included in a multiple contingency table which shows interrelationships among nominal scales. This procedure is named "reduction of multidimensional attribute space," which was formulated by P.P. Lazarsfeld and H. Barton in 1951. Though they practiced this operation in an intuitive manner, some statistical methods which reduce a multidimensional attribute space have been developed in the last ten years. They are PSA (Polarized Subgroup Analysis), AID (Automatic Interaction Detector), association analysis and so on. These methods may be appropriately included under the label of "Optimal Tree-Structure Analysis." Since, the results analyzed by one of these techniques, are normally presented in dendrogram style.
Notes	
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00150430-00000061-0157

名義尺度型データ処理の一方法

——多次元属性空間縮小の計量的手法——

堀　内　四　郎

1. 多次元属性空間の縮小

測定とデータ解析に関する技法の開発・整備は，経験科学の発展における重要な課題のひとつである．スティーブンス (S. S. Stevens) によれば，最広義には，測定 (measurement) とは，ある規則にしたがって，対象に数値をわりあてることである[1)]．測定の尺度には，いくつかの種類がある．名義尺度 (nominal scale)，順序尺度 (ordinal scale)，間隔尺度 (interval scale)，比例尺度 (ratio scale)，絶対尺度 (absolute scale) の区別は，広範に受容されているものである．

計量社会学ないし社会測定においては，名義尺度型変数のデータ処理が，重要な課題のひとつになっている．これは主として，つぎのような事情によるものである．社会学の研究においては，質問紙による調査により，データ収集が行なわれる場合が，きわめて多い．質問紙における設問形式には，いくつかの種類があるが，回答選択的質問 (check list question) が，もっとも広範に使用されている．すなわち，回答をあらかじめ定められ提示された2個以上の選択肢の中から，選ばせる形式の質問である．回答選択肢間の差異は，たんに分類的なカテゴリー間の相違にすぎないことが多い．この場合，研究者は，同一選択肢に回答した被調査者にたいして，その選択肢と一対一に対応する数値を付与し，名義尺度型のデータを得るわけである．

現在の行動科学的諸研究においては，データ解析のために，多変量解析

(multivariate analysis) の諸手法が，広範に利用されている．狭義の多変量解析法は，多変量正規分布モデルを前提とするものに限定されるであろうが，広義の多変量解析法は，互いに相関のある多変量（多種類の特性値）のデータのもつ特徴を要約し，かつ所与の目的に応じて総合するための手法である[2]．社会学においては，互いに関連のある，多くの名義尺度型変数についてのデータを，要約・総合するための手法[3]の開発が，とくに必要とされるのである．このような手法の開発に関しては，さしあたり，主として2つの方向が考えられる．第1には，多くの名義尺度型変数の各カテゴリーを，ダミー変数として扱い，それぞれのダミー変数の一次結合により，間隔尺度として処理できるような，合成変数を設定することである．林の数量化理論[4]が，これである．

数量化理論では，個体 i が，名義尺度型変数 x_j $(j=1, 2, \cdots\cdots, n)$ のカテゴリー k $(k=1, 2, \cdots\cdots, k_j)$ に反応するときにのみ1，他の (k_j-1) 個のカテゴリーのいずれかに反応したときは0の値をとる，$\delta_i(jk)$ なる量を導入する．いま，n 個のそれぞれの変数の，k_j 個のカテゴリーのそれぞれに対し，w_{jk} なる数値を与えるとき，個体 i にたいする新しい合成変数を，つぎのように定義する．

$$\alpha_i = \sum_{j=1}^{n} \sum_{k=j}^{k_j} \delta_i(jk) w_{jk}$$

多くの名義尺度型変数にたいする個体 i の反応を，α_i として要約・総合するわけである．この場合に，w_{jk} を付与する方法の相違によって，いわゆる数量化理論のⅠ類・Ⅱ類・Ⅲ類など，手法上の変異が生ずる．名義尺度型変数のデータ解析においては，きわめて有効な手法であり，とくに社会学者の間では，「計量社会学は，林の数量化理論の開発によってはじめて全面的に可能になった，といっても過言でない[5]」と，高く評価されている．

第2には，多数の名義尺度型変数についての，すべてのカテゴリーの中から，少数のカテゴリーを選択し，融合し，組合せることによって，新し

い1個の名義尺度を構成することが考えられる．林の数量化理論においては，間隔尺度として処理可能な変数αを，新らたに合成し，それによってデータを要約する．これにたいして，新しい名義尺度を作成し，それによってデータを要約することも可能であろう．

この考え方は，基本的には「多重分割表」(Multiple Contingency Table)の発想にもとづいている．多重分割表は，それぞれk_j個のカテゴリーをもつ，名義尺度型変数 $x_1, x_2, \cdots, x_j, \cdots, x_n$ が与えられたとき，カテゴリーのすべての組合せを考慮して，$\prod_{j=1}^{n} k_j$個のカテゴリーをもつ，1個の名義尺度型変数を構成することを，含意しているように思われる．なぜならば，多重分割表の各最小枡 (ultimate class) は，$\prod_{j=1}^{n} k_j$個の組合せパターンのそれぞれを，表わすものになっているからである．

多重分割表の限界のひとつは，変数の個数nの増大により，$\prod_{j=1}^{n} k_j$の値が，大きくなりすぎることである．実質上，4重分割表までが限界と思われる．それ以上に多くの変数を導入した場合，多重分割表は，①煩雑で解りにくい，②各枡の度数が，著しく減少する場合が多い，などの欠点をもつことになる．

したがって，変数の個数が多い場合には，以下の手続きによって，カテゴリーの組合せパターンを減少させることが，データ要約の見地から，有効であるように思われる．すなわち，所与の目的に応じて，①いくつかの変数を捨象すること，②同一変数のカテゴリーの中から，数個を選択して融合させ，新しい1個のカテゴリーを作成すること，の2通りの作業である．以上の作業について換言するならば，$\prod_{j=1}^{n} k_j$個のカテゴリーをもつ名義尺度型変数を，所与の目的に応じて，より少数のカテゴリーをもつ新しい名義尺度型変数に，変換することといえるであろう．

上述の作業は，かつてラザスフェルド (P.F. Lazarsfeld) とバートン (A.H. Barton) が，「多次元属性空間の縮小」(reduction of multidimensional attribute space) と呼んだ手続きに，ほぼ対応している[6]．「多次

元属性空間」という用語は，多変数に関する特性値のすべての組合せを要素とする，集合を指示している．名義尺度型変数も含めて考えられているので，この空間では「距離」を定義することが困難である.「多次元属性空間の縮小」とは，特性値のすべての組合せを，少数のパターンにまとめることである．

ラザスフェルドとバートンは,「多次元属性空間の縮小」に関して，以下のような分析事例を提示している．すなわち，アメリカにおける「社会的優位性」(social advantage) を分析するための戦略要因として,「人種」「出生地」「学歴」の３変数を選定する．各変数を〔白人—黒人〕〔アメリカ生れ—外国生れ〕〔大卒—非大卒〕のように，２分法名義尺度に操作化すれば，カテゴリーの組合せパターンは，８通りになる．８個のパターンの相互関係は，図１のような「樹形図」(dendrogram) として表現される．図１においては，{白人，アメリカ生れ，大卒} から {黒人，外国生れ，非大卒} までの，８通りのパターンが示されている．しかし，ラザスフェルドとバートンは，つぎのような理由で，８個のパターンを，よ少数のパターンにまとめることを主張する．第１に，黒人はきわめて社会的に不利であり，大卒であろうと，アメリカ生れであろうと，大差はない．第２に，白人の間

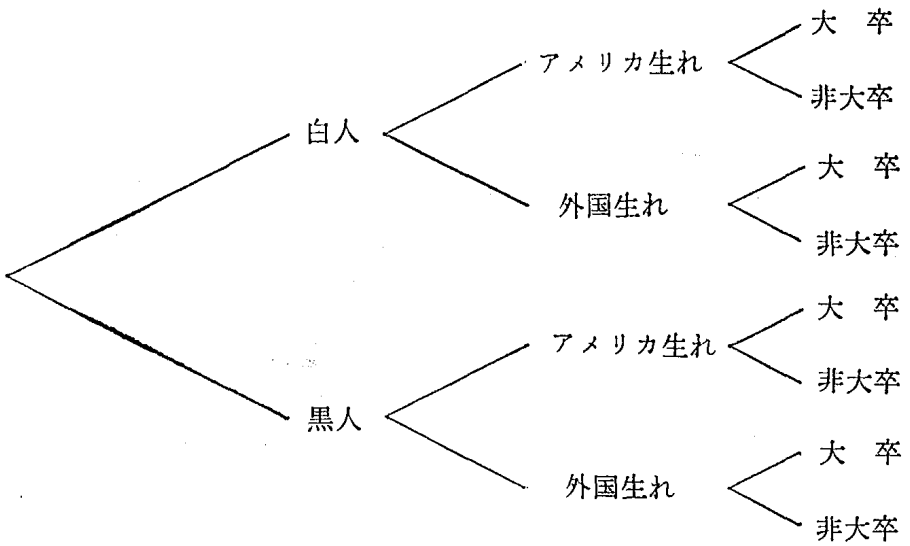

図１　社会的優位性に関連する多次元属性空間

では，アメリカ生れか否かということが，学歴よりも重要である．したがって，図2のように，{白人，アメリカ生れ，大卒}{白人，アメリカ生れ，非大卒}{白人，外国生れ}{黒人}の4パターンにまとめる方が，効率的である．

図1から図2への移行が，つまり多次元属性空間の縮小である．しかし，この作業は，標準化された手続きによって行なわれたものではない．むしろ，分析者自身の経験と直観に，おおいに依存している．仮に，他の研究者にたいして，図1の8パターンを，社会的優位性の分析に有効な結果をもたらすように，4パターンにまとめることを，指示したと仮定してみよう．彼は，図2とはまったく異なったまとめ方をするかもしれない．したがって，計量的な手法を開発して，多次元属性空間縮小の手続きを標準化することが，計量社会学者に要求される課題のひとつであるように思われる．

多次元属性空間縮小の計量的手法を論ずるに先立って，本論で使用される記号と用語を定義しておこう．多次元属性空間の縮小は，すべての個体を，与えられたカテゴリーの，すべての可能な組合せの数よりも，少数のクラスに分割することでもある．したがって「集団の分割」として解釈することも，また可能である．ここでは，集団の分割という考え方にもとづいて，用語を規定していくことにする．なお，集団の分割過程は，図2のような樹形図として表現できるので，「グラフ理論」における「木」(tree)

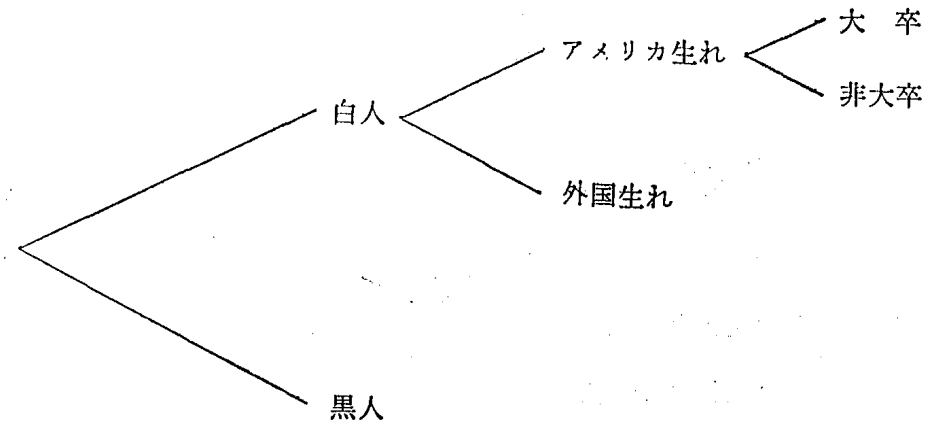

図2　縮小された多次元属性空間

概念と対応させて，論ずることができる．

集団mが複数の集団に分割されるとき，mを「親集団」(parent group)，分割によってmから生ずる集団を「下位集団」(subgroup) と定義する．すべての個体が所属する集団は，親集団をもたない．樹形図の中では，グラフ理論における「木」の「根」の位置を占める．これを，「根集団」(root group) と呼ぶことにする．また，それ以上分割されない集団は，下位集団をもたない．樹形図の中では，グラフ理論における「木」の「最終頂点」の位置を占める．これを「最終集団」(terminal group) と呼ぶことにする．図2においては，{白人，アメリカ生れ，大卒}{白人，アメリカ生れ，非大卒}{白人，外国生れ}{黒人} は，いずれも最終集団である．

根集団を r 回分割することによって，獲得される集団を，「第 r 段階の集団」(group at the r-th stage) と呼ぶ．図2においては，{黒人} は第1段階の集団であり，{白人，アメリカ生れ，大卒} は第3段階の集団である．なお，2つの付帯条件をつけ加えておこう．第1に，分割の基準として使用される変数の個数は，1回の分割について1変数に限定されるものとする．図3のように，2個以上の変数を用いて，1回の分割を行なうことは，さしあたり考慮しない．第2に，同一段階における2つ以上の集団を，それぞれ分割する場合に使用される変数は，必ずしも同一のものでなくてもよい．図4は，第1段階における2つの集団を，それぞれ異なった変数を用いて，分割する場合の例である．

根集団を u 個の下位集団に分割するとき，各集団をG(1)，G(2)，……，

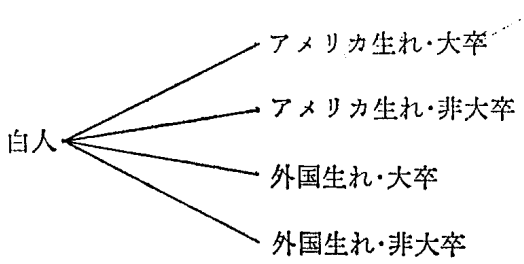

図3　集団分割形式（A）

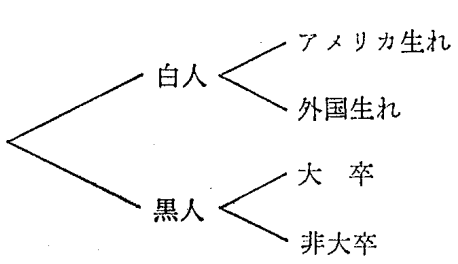

図4　集団分割形式（B）

G(u) と表わす．G(1) は，さらに v 個の下位集団G(1·1)，G(1·2)，……，G(1·v) に分割される．このようにして，第 r 段階の特定集団mは，Gの後のカッコに r 個の数字を並べた記号によって，表現されることになる．図5は，このような記号を用いて，図2を書きなおしたものである．

以上のように定義された用語の一部を用いて，多次元属性空間縮小の計量的手法に関する基本的アイディアを，定式化しておこう．多数の名義尺度型変数に関するデータを用いて，根集団を最終集団まで分割する．分割は，以下の条件を満足するものでなければならない．すなわち，①特定集団mに所属するすべての個体は，共通の反応パターンを示していなければならない．共通の反応パターンとは，いくつかの変数に関して，それぞれ特定の同一カテゴリーに反応することを意味している．②この反応パターンを示すすべての個体は，集団mに所属していなければならない．

また，分割の適切性に関して，いくつかの規準を設定することができる．外的基準のある場合の分割においては，適切性の規準は，ⓐ最終集団の数を少なくすること，ⓑ最終集団への分割の，外的基準にたいする説明力（F値，χ^2 値など）を大きくすること，の2点である．外的基準のない場合の分割においては，適切性の規準は，ⓐ最終集団の数を少なくすること，ⓑ最終集団への分割と，他の多くの変数との関連を高くすること，の2点

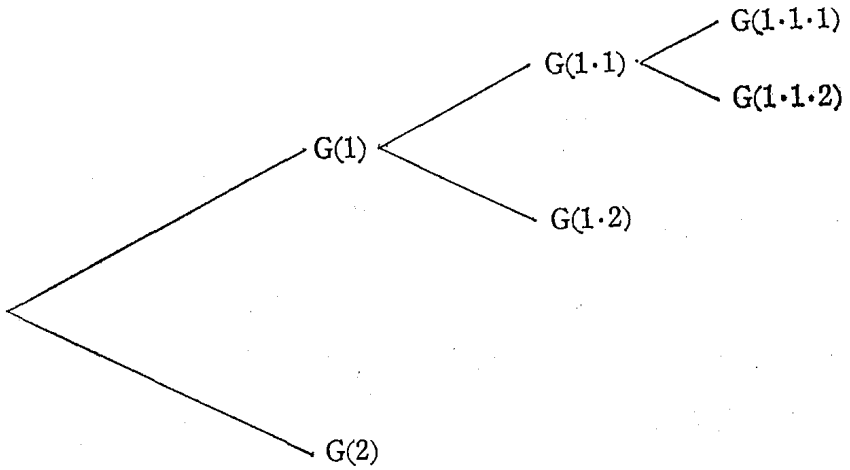

図5　多次元属性空間の記号的表現

である.

なお，外的基準のある場合，ない場合のいずれにおいても，ⓐとⓑは，相反する性格をもった規準である．さしあたり具体的には，つぎのようなステップをとるのが，有効と思われる．すなわち，まず規準ⓐをある程度以上に満足するための制約条件を，なんらかの形で——一例を挙げれば，計算の打切り基準として——課しておき，その制約条件のもとでⓑを最適化する，という手順である.

以上が,「多次元属性空間縮小の計量的手法」の，基本的アイディアである．これに該当するものとしては，関連分析，AID, PSA など，萌芽的ないくつかの手法が考案されている．しかし，それらを関連づけて整理する試みは，ほとんど行なわれていないように思われる．本論では，すでに示したように，外的基準のある場合とない場合の区別を軸にして，いくつかの手法を整理してみたい.

2. 外的基準のある場合： その1 — PSA

変数増加法によるダミー変数重回帰分析を利用して，多重分割を行なう試みが，レヴィ (S. G. Levy) によって，精力的に繰りかえされている[7]. 彼はこれを PSA (Polarized Subgroup Analysis) と呼び，人種問題の研究に活用している．本節では，PSA の手順を説明し，分析事例を紹介し，さらにその限界について論ずることにしたい.

PSA の手順は，以下の通りである.

1) 従属変数（外的基準）として，1個の間隔尺度型変数 y，独立変数として，n 個の2分法名義尺度型変数 $x_1, x_2, \cdots\cdots, x_n$ を指定する．x_j ($j=1, 2, \cdots, n$) は，ダミー変数として処理可能である.

2) 変数増加法を用いて，逐次的重回帰分析を行なう.

2・1) n 個の変数 $x_1, x_2, \cdots, x_n$ のうち，y との単相関係数が最大のものを選び，これを $x_{(1)}$ とする．y の $x_{(1)}$ にたいする回帰式を求める.

2・2)　$x_{(1)}$ 以外の（n—1）個の変数のそれぞれと，$x_{(1)}$ との2つを独立変数とする（n—1）個の重回帰式を計算し，y にたいする決定係数の最大なものを選ぶ．選ばれた重回帰式に用いられた変数を $x_{(1)}$, $x_{(2)}$ とする．

2・3)　$x_{(1)}$, $x_{(2)}$ のほかに，さらに1個の変数を，残りの（n—2）個の変数のなかから選び，この3変数にたいする重回帰式を（n—2）個つくる．このなかで，決定係数の最大なものに用いられた変数を $x_{(1)}$, $x_{(2)}$, $x_{(3)}$ とする．

2・4)　なんらかの打切り基準（stopping rule）が働いて，計算が終了するまで，以下同様の手順をくり返し，p（$p \leqq n$）個の独立変数を含む重回帰式を求める．

打切り基準は，計算をどこで終了するかを定める規則であり，つぎの3通りがある．

①あらかじめ p を指定し，取り込まれた変数の個数が，p に達すれば中止する．

②決定係数の値を，あらかじめ指定しておく．変数を加えていくにつれて決定係数が大きくなり，指定した値に達すれば中止する．

③変数を1個追加することによって得られる決定係数の増分に注目し，これがあらかじめ指定された値以下になれば中止する．

以上のような，2・1）から2・4）までの手順により，1番から p 番まで順序づけられた，p 個のダミー変数 $x_{(1)}$, $x_{(2)}$, …, $x_{(p)}$ を得ることができる．

3)　変数増加法によるダミー変数重回帰分析の結果を用いて，集団を分割・融合し，樹形構造を形成する．

3・1)　つぎのような手順で，集団分割を行なう．

①根集団を，$x_{(1)}$ の値が1であるか0であるかにより，下位集団G(1)，G(2) に分割する．G(1) と G(2) は，第1段階の集団である．

②G(1)，G(2) のそれぞれを親集団として，$x_{(2)}$ の値に応じて，下位集団G(1・1) と G(1・2) および G(2・1) と G(2・2) に分割する．G(1・1)，G(1・2)，G(2・1)，G(2・2) は第2段階の集団である．

③同様にして，2^p 個の第 p 段階集団が獲得されるまで，集団の分割を反復する．

3・2) 分割された下位集団のいくつかを，つぎのようにして融合する[8]．

①従属変数 y に関して，値ϵを指定する．

②第1段階から第 p 段階までの，それぞれの集団 m について，y の平均値 $\bar{y}_m$ および分散 σ_m^2 を計算する．

③第 (p—1) 段階における同一の親集団から分割された，第 p 段階下位集団の対のそれぞれについて，y の級間分散を計算し，ϵ 以下であれば両者を融合する．なお，特定の第 (r—1) 段階集団から分割された，すべての（この場合は，2個の）第 r 段階集団が，再度1個に融合した場合，もとの第 (r—1) 段階集団を「融合集団」と呼ぶことにしよう．③第 (p—2) 段階における同一の親集団から分割された，第 (p—1) 段階集団の対のなかで，融合集団をふくむものに注目する．このような対のそれぞれについて，y の級間分散を計算し，ϵ 以下であれば両者を融合する．

④同様にして，融合ができなくなるまで，以上の手順を反復する．

3・3) 分割と融合の結果，獲得された＜親集団—下位集団＞関係の，樹形構造を図示する．この際，各集団を表わす点の位置が，各集団の $\bar{y}_m$ の値を表わすように配慮する．

PSA の分析事例として，ここではレヴィによる，人種問題にたいする態度の研究[9]を紹介しよう．図6は，レヴィのデータ解析結果を示したものである．縦軸には従属変数として，人種統合の進展と努力に関する態度尺度値が，目盛られている．上部には，ダミー変数型の独立変数が，変数増加法によって選定された順序に，左から右に並べられている．図6から，

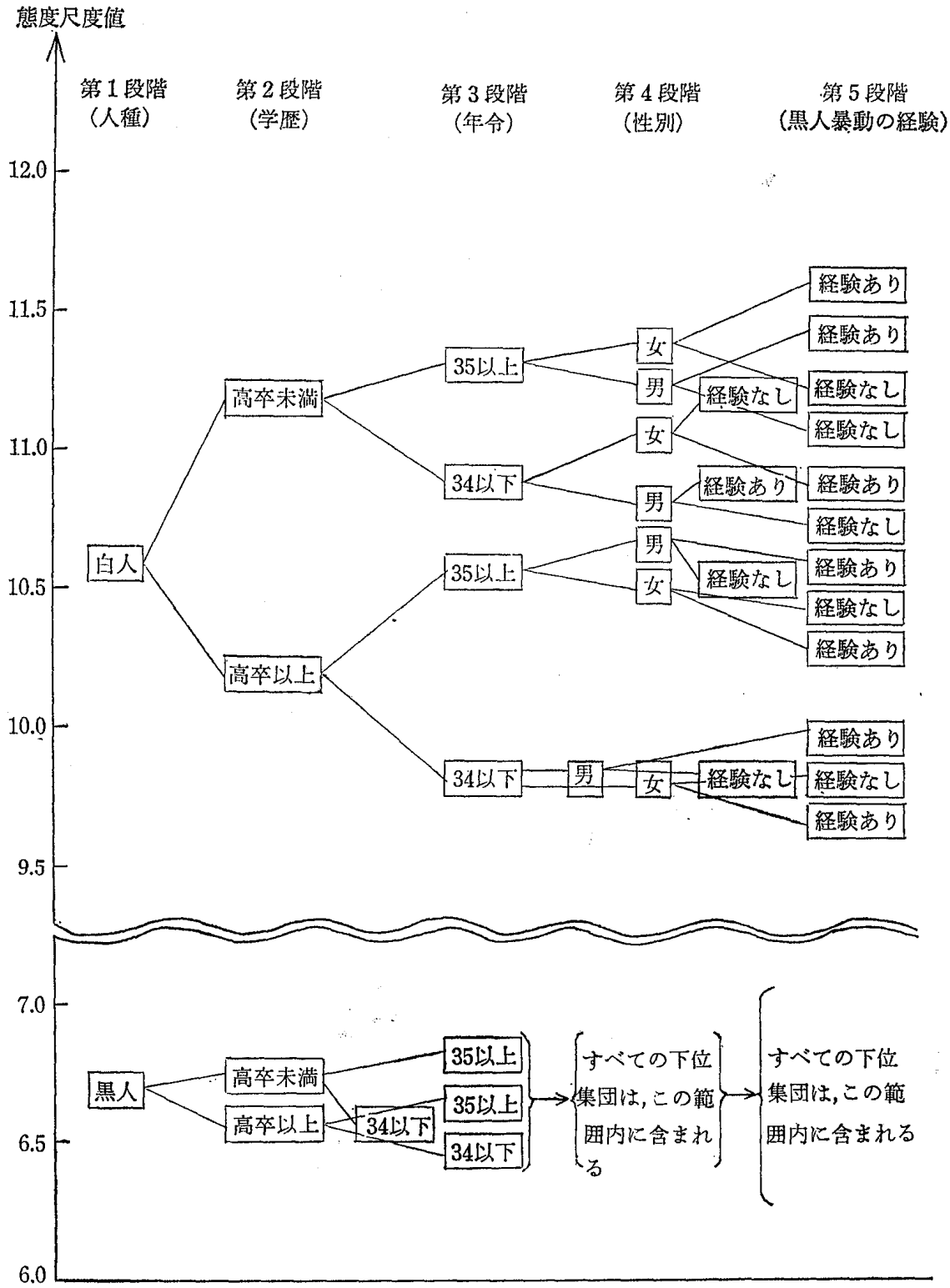

図6　ＰＳＡの分析事例：人種統合の進展と努力に関する態度の差異

各集団の態度尺度の平均値と，集団間の＜親集団―下位集団＞関係を，理解することができる．

さいごに，多次元属性空間縮小の計量的手法という視点から，ＰＳＡの限界のいくつかを指摘しておこう．

①２分法名義尺度型変数しか処理できない．３分法以上の名義尺度型変数を処理する場合は，解析に先立って，いくつかのカテゴリーをまとめ，２分法形式に変換しておかなければならない．

②同一段階における２つ以上の集団を，それぞれ分割する場合に使用される変数は，単一のものに限定される．図４のように，それぞれ異なった変数を用いて，分割することはできない．

③従属変数に関する集団間の差異を確認する作業においては，独立変数間の交互作用を明確化しているにもかかわらず，変数選択の過程では，交互作用を無視している．ＰＳＡにおけるダミー変数重回帰分析は，交互作用項を考慮していない．回帰方程式に交互作用項を含めた場合には，変数選択のために重回帰分析を用いることが，できなくなるからである．したがって，単純な加法的モデル，すなわち２つ以上の独立変数が，従属変数にたいして，加法的に作用していると想定するモデルを，採用せざるをえないのである．

④変数増加法によるダミー変数重回帰分析においては，変数選択のための適切性の指標として，決定係数あるいは重相関係数が使用される．しかし，ＰＳＡの目的は，合成変数の構成ではなく，集団の分割である．したがって，変数選択のための適切性の指標としては，級間分散ないし級間変動の方が妥当であろう。ここでいう級間分散は，選択されるダミー変数のカテゴリーを組合せることにより，形成される集団間における，級間分散を指示している．なお，限界④は限界③と，きわめて関連している．

⑤変数増加法によって獲得された，p 個の独立変数による重回帰式は，n 個のなかから選ばれる p 個の変数の，いかなる組合せにたいする重回帰式

よりも，その決定係数は大きい，という保証がない．すなわち，全体としての最適化は，各ステップごとの最適化によっては，必ずしも達成されないのである．これは，変数増加法それ自体の限界といえよう．

以上は，ＰＳＡの限界のなかで，とくに重要と思われる5点である．本節では，外的基準が間隔尺度型変数である場合のみを想定して，分析の手続きを説明してきた．しかし，従属変数が2分法名義尺度の場合にも，原則として適用可能である．すなわち，手順の解説のなかの「重回帰分析」を「判別分析」に，「決定係数」を「判別効率」に代置して考えればよいのである．

3. 外的基準のある場合：その 2—AID

AID (Automatic Interaction Dectector) は，ミシガン大学サーベイリサーチセンターにおいて，モルガン (J.N. Morgan) とゾンクィスト (J.A. Sonquist) によって，開発された手法である[10]。はじめに，本節で使用する記号を定義しておこう．集団 m の，変数 y についての平方和を TSS_m とする．集団 m を下位集団に分割したときの，下位集団の級間平方和を BSS_m とする．なお，とくに根集団の平方和を TSS_t，根集団を分割したときの級間平方和を BSS_t と書くことにする．

ＡＩＤの手順は，以下のとおりである．

1)　間隔尺度型またはダミー変数型の従属変数 y と，名義尺度型の独立変数 $x_1, x_2, \cdots, x_n$ を指定する．ＰＳＡの場合と異なり，x_j $(j=1, 2, \cdots, n)$ のカテゴリーは，2個以上いくつあってもよい．

2)　従属変数 y についての BSS_t/TSS_t が，もっとも大きくなるような，x_j による根集団の2分割を求める．

2・1)　変数 x_j が k_j 個のカテゴリーをもつ場合，k_j 個のカテゴリーを2組に分割するすべての組合せを求める．それぞれの組合せについて BSS_t/TSS_t を計算し，この値がもっとも大きくなるような2分割を，明らかにする．

例をあげて考えてみよう．独立変数としての「職業」を，「事務職」「労務職」「自営業」「その他」の4カテゴリーからなる形式に，操作化したと想定する．この場合｛事務職 *vs.* 労務職，自営業，その他｝｛労務職 *vs.* 事務職，自営業，その他｝｛自営業 *vs.* 事務職，労務職，その他｝｛その他 *vs.* 事務職，労務職，自営業｝｛事務職，労務職 *vs.* 自営業，その他｝｛事務職，自営業 *vs.* 労務職，その他｝｛事務職，その他 *vs.* 労務職，自営業｝という，7通りの2分割がある．この7通りの中で，BSS_t/TSS_t が最大になるような分割を明らかにする．

なお，x_j が順序尺度の場合の2分割法は，(k_j-1) 通りしかない．例として，「年令」という独立変数を考え，それが｛20才未満，20代，30代 40代，50才以上｝という，5個の順序づけられたカテゴリーから，成り，立っていると想定してみよう．2分割のしかたは，｛20才未満 *vs.* 20代，30代，40代，50才以上｝｛20才未満，20代 *vs.* 30代，40代，50才以上｝｛20才未満，20代，30代 *vs.* 40代，50才以上｝｛20才未満，20代，30代，40代 *vs.* 50才以上｝の4通りである．

2・2) n 個の独立変数それぞれについて，以上の計算を行い，すべての2分割のなかで，BSS_t/TSS_t が最大になるような分割を求める．これにより，根集団を2個の下位集団に分割する．

3) 分割によって獲得されたそれぞれの下位集団 m を，BSS_m/TSS_m が最大になるように，2) と同様の手続きにより，2分割する．

4) 以上のような集団分割を，つぎの打切り基準のいずれかが作用して，計算が中止されるまで反復する．

打切り基準は，①集団 m の個体数，②TSS_m/TSS_t，③BSS_m/TSS_t の3通りである．分析に先立って，それぞれについて特定の値を定めておく．すなわち集団 m の個体数が一定数以下になるか，集団 m の TSS_m が TSS_t の一定割合以下になるか，あるいは集団 m を2分割するときの最大の BSS_m が，TSS_t の一定割合以下になったならば，集団 m は分割されないがわけである．

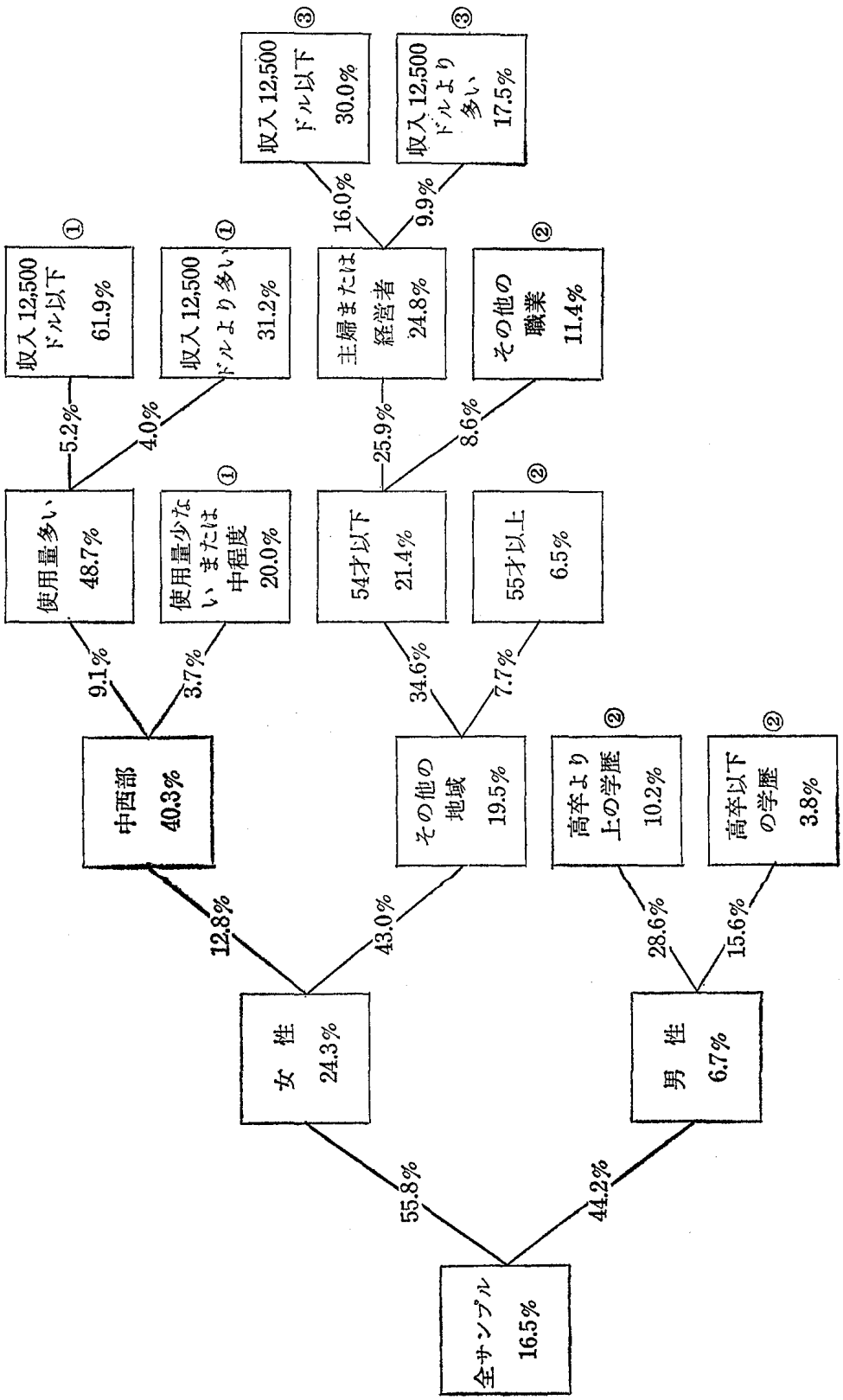

図7　AIDの分析事例：銘柄Xの通常購入者

図7は，ＡＩＤの分析事例[11]を示したものである．従属変数としては，ある銘柄[12]の商品の購入有無（ダミー変数），また独立変数としては，8個の社会経済的変数が使用された．各枡は特定の集団を表わしている．枡の中の数字は，特定集団における購入率であり，枡と枡を結ぶ線上の数字は，各集団の全サンプルにおける構成比率を示している．また，最終集団を表わす枡の右側には，①②③などの数字が記入されている．これは，上述の打切り基準のなかのいずれが作用して，計算が中止されたかを指示するものである．この事例では，①②③のそれぞれについて，25, 0.05, 0.02というパラメーターが設定されている．

4. 外的基準のある場合：その3—AID の修正

本節では，まずＡＩＤの限界を考察し，その論議にもとづいてＡＩＤを修正し，より有効な手法を検討してみたい．

はじめに，本節で使用される記号と用語を定義しておこう．ＡＩＤにおいて，根集団を BSS_t/TSS_t が最大になるように2分割するとき，y の平均値が高い方の下位集団をＧ(Ｈ)，低い方の下位集団をＧ(Ｌ)と表わす．Ｇ(Ｈ)をさらに2分割することによって獲得される下位集団を，y の平均値の高低に応じて，Ｇ(ＨＨ) およびＧ(ＬＨ) と表わす．同様にして，Ｇ(Ｌ) はＧ(ＬＨ) とＧ(ＬＬ) に分割される．このような記号法を，各段階の集団に適用することにより，図7を図8のように変換することができる．

第 r 段階における各集団は，Ｇの後のカッコ内に，‘Ｈ’または‘Ｌ’を r 個並べた記号によって指示される．ここで，r 個の記号がすべて‘Ｈ’であるか，またはすべて‘Ｌ’である集団を，第 r 段階の「極集団」(polar group) と呼ぼう．図8のＧ(ＬＬ) や Ｇ(ＨＨＨ) は，いずれも極集団である．極集団が最終集団である場合，それを「最終極集団」と呼ぶことにしたい．図8においては，Ｇ(ＬＬ) と Ｇ(ＨＨＨＨ) の2個だけが，最終極

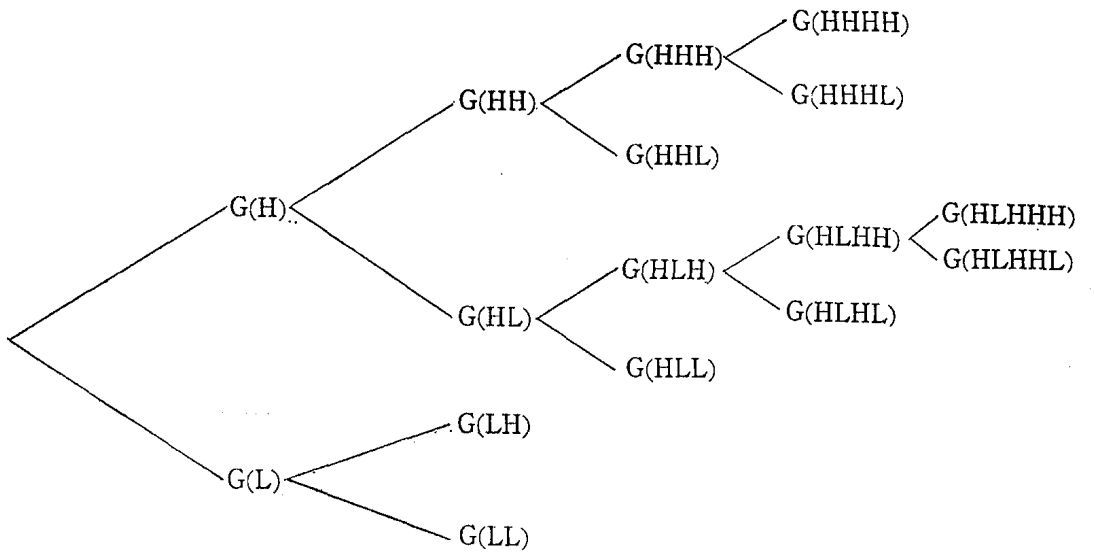

図8　ＡＩＤ分析結果の記号的表示

集団になっている.

以上のような記号法を用いて，ＡＩＤの限界を検討してみよう．その第１としては，ＡＩＤにおいては２分割のみが許容されており，親集団を３個以上の下位集団に分割することができない．これは，ＰＳＡにも共通する問題点である.

第２の限界は，ＡＩＤによって根集団をM個の最終集団に分割した場合，それが必ずしも，一定の制約条件下で級間分散を最大にするような，根集団のM分割ではないことである．これには，いくつかの理由が考えられる．第１の理由は，全体としての分割の最適化が，各ステップにおける２分割の最適化によっては，必ずしも達成されないことである.

第２の理由は，極集団以外の集団mを，BSS_m/TSS_mが最大になるように分割することが，全体としての最終集団の級間分散を大きくする方向に，必ずしも貢献しないことである．単純な例として，図９のように根集団を分割して，４個の最終集団を獲得する場合を想定してみよう．この例では，明らかにＧ(ＨＨ)とＧ(ＬＬ)の，yについての平均値の差が大きくなるように，集団分割が行なわれている．ここで，根集団からＧ(ＨＨ)とＧ

（ＬＬ）を除いた，残余の部分に焦点を当ててみよう．この残余部分は，Ｇ（ＨＬ）とＧ（ＬＨ）に分割されているということができる．しかし，このような分割法が，４個の最終集団の級間分散を大きくするために，適切であるかどうかは，きわめて疑わしい．なぜならば，Ｇ（ＨＬ）とＧ（ＬＨ）それぞれの平均値は，近い値をとる可能性が，必ずしも小さくないのである．

したがって，４個の最終集団の級間分散をより大きくするためには，つぎのような分割手順をとることが，有効と思われる．すなわち，まず通常のＡＩＤにより，Ｇ（ＨＨ）とＧ（ＬＬ）を獲得する．つぎにＧ（ＨＬ）とＧ（ＬＨ）とを融合し，ひとつの新しい親集団を形成する．この親集団 m を，BSS_m/TSS_m が最大になるように，ＡＩＤを用いて２分割する．こうして獲得された２個の集団を，それぞれＧ′（Ｈ），Ｇ′（Ｌ）と表わすならば，図10のような形で，分割結果を示すことができる．｛Ｇ（ＨＨ），Ｇ′（Ｈ），Ｇ′（Ｌ），Ｇ（ＬＬ）｝の級間分散は，｛Ｇ（ＨＨ），Ｇ（ＨＬ），Ｇ（ＬＨ），Ｇ（ＬＬ）｝の級間分散よりも，大きくなるものと思われる．

しかし，ＡＩＤを開発したモルガンとゾンクィントの主たる狙いは，必ずしも，できるだけ級間分散が大きくなるように，根集団を最終集団に分割することではないようである．むしろ，なんらかの統計解析に先立つ準備的作業として，変数間の交互作用の存在を検出しておくことが，ＡＩＤ本来の目的であったと，考えるべきであろう．したがって，外的基準がある場合の，多次元属性空間縮小の計量的手法のひとつとして，ＡＩＤを活用するためには，技法上の修正が必要であるように思われる．

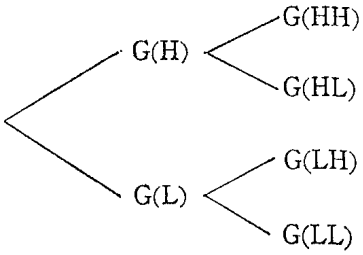

図９　ＡＩＤによる４分割

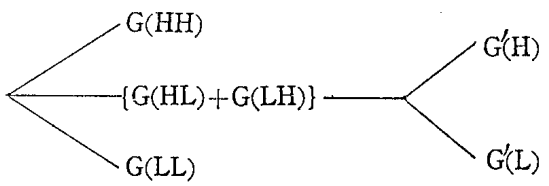

図10 「ＡＩＤによる４分割」の修正

ここでは，ＡＩＤを修正するささやかな試みのひとつを，提示するにとどめよう．これは，図10に示される前述の分割手続きを，より一般的な形に拡張したものである．その手順は，以下の通りである．

1)　間隔尺度型またはダミー変数型の従属変数 y と，名義尺度型の独立変数 $x_1, x_2, \cdots, x_n$ を指定する．

2)　ＡＩＤと同様にして，根集団を G(H) と G(L) に分割する．

3)　同様にして，G(H) と G(L) のそれぞれを，G(HH) と G(HL)，および G(LH) と G(LL) に分割する．

4)　第２段階における４個の集団のなかから，極集団である G(HH) と G(LL) のみを選び出し，ＡＩＤと同様な手続きで，それぞれ G(HHH) と G(HHL)，G(LLH) と G(LLL) に分割する．

5)　第３段階における４個の集団のなかから，２個の極集団，すなわち G(HHH) と G(LLL) を選び出し，同様にして，それぞれを２分割する．

6)　このようにして，各段階における極集団の分割を行なう．打切り基準が作用するまで，分割の作業を継続する．獲得された極集団を，'H' 記号のみで表わされるものと，'L' 記号のみで表わされるものとの２群に分け，それぞれの群で，もっとも高次の段階に属するものを，「第１次最終極集団」と呼ぶ．

7)　はじめの根集団から，２個の「第１次最終極集団」を除いた残余を，新たな根集団と考える．この根集団を「第２次根集団」，はじめの根集団を「第１次根集団」と呼ぶことにする．「第２次根集団」について，2) から 6) までの手続きを繰り返し，新しい２個の最終極集団を求める．これらを「第２次最終極集団」と呼ぶ．

8)　以下同様にして，第3次，第4次……の最終極集団を求めていく．すべての計算が中止されるのは，つぎの２通りの場合のいずれかである．①第 q 次最終極集団が，第 q 次根集団を２分割したものである場合．換言すれば，第 q 次最終極集団のそれぞれが，第 q 次根集団にたいして，第１段階の下

位集団になっている場合である．この場合，分割されるべき第$(q+1)$次根集団が，存在しないことになる．図11は，このようにして，第3次最終極集団を求めた時点で，計算が中止されたときの樹形図を示したものである．

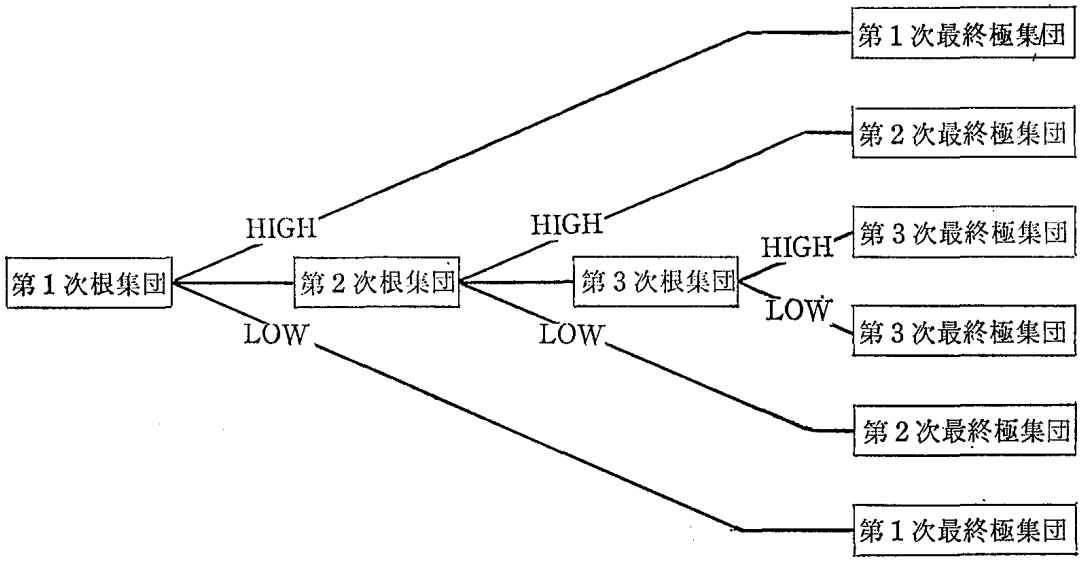

図11 「モミの木」型ＡＩＤのアウトプット：その1

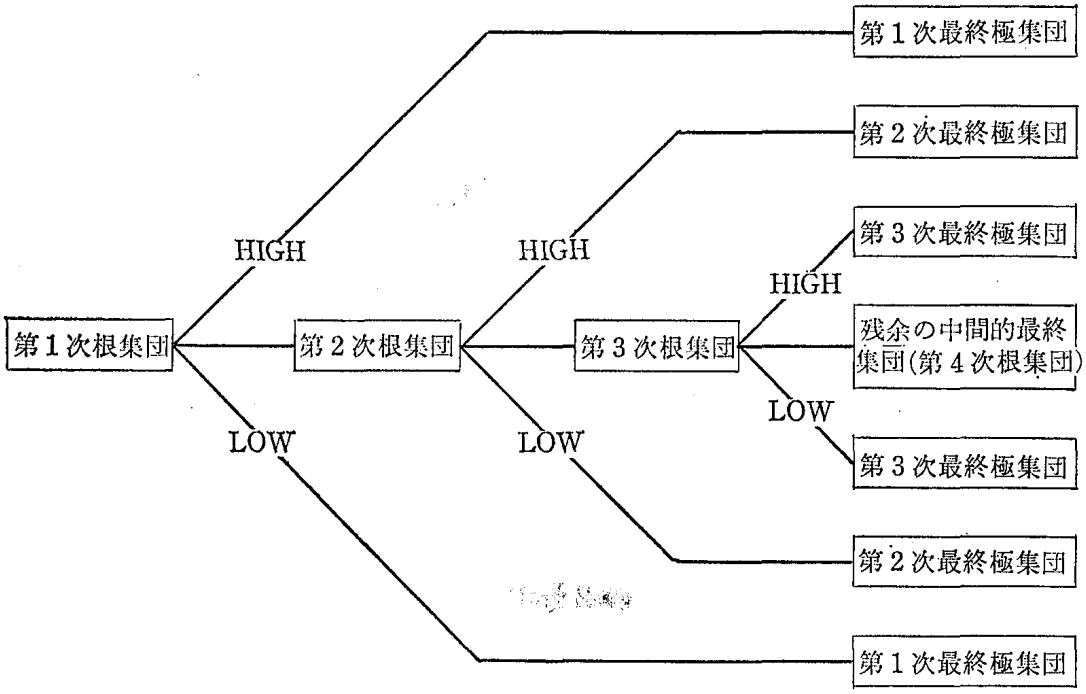

図12 「モミの木」型ＡＩＤのアウトプット：その2

②第（$q+1$）次根集団に打切り基準が適用され，その分割が行なわれない場合．図12は，第4次根集団が分割されないときの樹形図である．

以上が，多次元属性空間縮小における分割の適切性の規準を，より満足させるように，ＡＩＤを修正してみた場合の，ひとつの手法例である．このようにして獲得された最終集団の級間分散は，同一の根集団を同一の打切り基準のもとで，ＡＩＤを用いて分割した場合の，最終集団の級間分散よりも，大きくなるものと思われる．なお，図11および図12の形状を考慮するならば，この手法は「モミの木」型ＡＩＤ（'fir tree' type AID）とでも呼ばれるべきであろう[13]．

5. 外的基準のない場合—関連分析

「関連分析」(association analysis) は，オーストラリアにおいて，ランス (G. N. Lance) とウィリアムズ (W. T. Williams) によって開発された手法である[14]．その手順は，以下の通りである．

1)　n 個のダミー変数 x_j ($j=1, 2, \cdots, n$) について，その四分積率相関係数行列を作成する．

2)　n 個の変数のなかから，$\sum_{j \neq l} r_{jl}^2$ が最大であるような変数 x_l を選ぶ．x_l の値が1であるか0であるかに応じて，根集団を2個の下位集団に分割する．なお，$\sum_{j \neq l} r_{jl}^2$ の代りに，$\sum_{j \neq l} |r_{jl}|$ が使用されることもある．

3)　それぞれの下位集団について，手順 1) 2) を反復する．

4)　以下同様にして，何らかの打切り基準が作用するまで，集団分割を繰り返す．打切り基準は数多く考慮されているが，いずれが適切かという問題に関しては，定説がない．ここでは代表的なものとして，①分割される集団 m の個体数 N_m，および② $N_m \sum_{j \neq l} r_{jl}^2$ の2基準を指摘するにとどめよう．

関連分析と主成分分析を比較することは，興味深い作業である．すなわち，主成分分析においては，与えられた諸変数の一次結合という形をとる

合成変数を設定し，その合成変数と与えられた諸変数の，相関係数の平方和が最大になるようにする．これにたいして関連分析では，他の諸変数との相関係数の平方和が，最大になるような変数を，与えられた変数群のなかから選択するのである．

なお，関連分析を以下のように，位置づけることも可能である．すなわち，少数個のカテゴリーからなる1個の名義尺度型変数を新たに構成し，これによって，たがいに関連のある多くの名義尺度型変数についてのデータを，要約することを考える．この際，要約による情報の損失を，なんらかの意味で最小にすることを目的とする．このような特性をもつ手法としては，つぎの3通りが考えられる．

①名義尺度型データを，林の数量化理論第Ⅲ類や，ダミー変数因子分析などの手法を用いて処理し，間隔尺度として処理可能なデータ（サンプル数量，因子得点など）を作成する．このデータを，与えられる数値が間隔尺度型変数についてのものであることを，前提とするクラスター分析を援用して解析する．分析の結果，相互に近似性の高い個体同志は同一集団に，相互に近似性の低い個体同志は，異なった集団に所属するように，全個体が数個の集団に分割される．各集団が，新しい名義尺度のカテゴリーになるわけである．このような場合のクラスター分析としては，フリードマン (H. P. Friedman) とルービン (J. Rubin) の手法[15]が代表的なものであろう．

②名義尺度データから，なんらかの個体間関連係数[16]行列を作成し，個体間の類似性の指標にもとづくクラスター分析[17]を用いて，全個体を数個の集団に分割する．

③関連分析により，全個体を数個の集団に分割する．

以上の①②③のいずれも，名義尺度型変数についてのデータが与えられた場合に，同質的な個体によって各集団が構成されるように，全個体を数個の集団に分割する手法である．この意味では，クラスター分析の一種として，関連分析を特徴づけることもできる．

クラスター分析の手法は数多く考案されており，とくに階層的諸手法(hierarchical methods)は，広範に使用されている．階層的手法は，大別して，2つの系列に整理される．一方は，すべてのクラスターが，それぞれ1個ずつの個体によって構成されている状態から出発し，クラスターの融合を繰り返していく手法である．他方は，すべての個体が1個のクラスターに所属している状態から出発し，クラスターの分割を反復していく手法である．一般に前者は「凝集法」(agglomerative method)，後者は「分割法」(divisive method)と呼ばれている．分割法の一種として，関連分析を位置づけることも可能である．

関連分析と，その他の多くのクラスター分析の相違点のなかで，もっとも重要なものは，つぎの点であろうと思われる．すなわち，関連分析は，集団内部の同質性に関する，2つの厳しい条件を満足している．第1に，関連分析によって形成された集団に所属する個体は，すべてなんらかの，同一な反応パターンを示している．第2に，この反応パターンを示すすべての個体は，当該集団に所属している．

関連分析以外のクラスター分析においては，この2条件が，必ずしも満足されているわけではない．それらの手法においては，同一集団に所属する個体の反応パターンが，少なくともいくつかの変数に関してまったく同一であることを，要求してはいないのである．要求されているのは，集団内部における反応パターンの「同一性」ではなく，「相対的近似性」にすぎない．

したがって，分析によって形成された集団の特性を叙述する際に，関連分析とその他多くのクラスター分析の間には，大きな差異が生ずる．関連分析の場合は，その集団に所属する個体の，共通特性を論ずることができる．その他のクラスター分析の場合は，形成された集団における，平均値的ないし最頻値的特性を，述べうるのみである．

さいごに，関連分析の限界のいくつかを，指摘しておこう．第1に，2分法名義尺度型変数しか処理できないことである．これは，ＰＳＡにも共

通する問題点である．

第2には，関連分析によって獲得された，M個の集団への分割が，x_j による根集団のM分割のすべての組合せのなかで，必ずしも最適なものとは限らないことである．なお，ここでいう最適性とは，集団のM分割が，与えられたn個の変数に対して，より多くの変数と，もっとも高い関連を示すことを意味している．

6. 結　　論

本論では，「多次元属性空間縮小の計量的手法」として，ＰＳＡ，ＡＩＤ，関連分析など，4種類の技法を紹介してきた．これらの手法は，いくつかの利点を持っている．ここでは，そのなかから，以下の4点を指摘しておくことにしよう．

1)　多次元属性空間縮小の計量的手法は，「多重分割表」による分析を，発展させたものということができる．それは，つぎの点で，多重分割表と共通している．すなわち，いくつかの名義尺度型変数に関して，それらのカテゴリーを組合せることにより，データを叙述しようとする点である．

しかし，すでに述べたように，多重分割表においては，数多くの変数を同時に考慮することが困難であり，せいぜい4変数程度にとどまる．多次元属性空間縮小の計量的手法は，このような多重分割表の限界を，克服しようと試みるものである．すなわち，多くの名義尺度型変数に関するデータが与えられたとき，所与の目的に応じて，いくつかの変数を捨象し，いくつかのカテゴリーを融合するのである．

2)　前述のように，多次元属性空間縮小の計量的手法は，いくつかの名義尺度型変数のカテゴリーを組合せて，データを叙述しようとするものである．したがって，変数間の交互作用を，明らかにすることができる．

3)　社会調査において，簡略化ないし効率化のために，活用することができる．これは，つぎのような手順によって，行なわれる．まず，特定の問

題領域に関連する多くの変数を選定し，回答選択的設問形式にして，調査票を作成する．この調査票を用いて収集された，名義尺度型の多変量データを，多次元属性空間縮小の計量的手法のいずれかを用いて処理する．データ解析の結果，多数の質問項目のなかから，少数の質問項目を選び出し，それらのカテゴリーを融合し，組合せて，新しい1個の名義尺度を構成することになる．この新しい変数で，もとの多数の変数を代表させる．つぎの調査からは，当該問題領域に関しては，この新しい変数のみを使用することにする．このようにして，多次元属性空間縮小の計量的手法は，質問項目数と集計表の分量を節約するために，貢献することができるわけである．

4)　分析結果が，きわめて理解しやすい．正準変量を使用しないこと，結果が樹形構造として，明快に図示されること，などの理由によるものであろう．

以上が，多次元属性空間縮小の諸手法のもつ，利点のいくつかである．しかし，すでに指摘されたように，これらの手法は，いくつかの限界をもっている．とくに，つぎの2点は，各手法に共通している．すなわち，第1には，集団分割の各ステップにおいては，2分割しかできないことであり，第2には，各ステップにおける分割の最適化が，根集団を最終集団に分割する際の，全体としての結果の最適化を，必ずしも意味しないことである．このような限界を克服する方向で，より有効な手法を開発し，整備・統合していくことは，計量社会学における重要な課題のひとつであろう．

最後に，名称の問題に関して，一言しておこう．本論では，ラザスフェルドとバートンの用語法にもとづき，一貫して「多次元属性空間縮小の計量的手法」なる語を使用してきた．しかし，「多次元属性空間の縮小」という概念は，主成分分析や正準相関分析における「次元の減少」(reduction of dimensionality) 概念と，混同される危険が大きい．むしろ，分析結果が樹形図として表現される事情を考慮するならば，「最適樹形構造解析」

(optimal tree-structure analysis) なる名称の方が，より適切であるように思われる．

注

1) S. S. Stevens, "Mathematics, Measurement, and Psychophysics" (in S. S. Stevens, ed., Handbook of Experimental Psychology, Wiley, 1951) 〔吉田正昭訳「数学，測定，精神物理学」，吉田編『計量心理学』誠信書房，1968所収，P．73〕.
2) 奥野忠一，芳賀敏郎，久米均，吉沢正共著「多変量解析法」(日科技連, 1971) pp. 2-3.
3) 名義尺度型データの解析手法を論ずる際には，次書を欠くことはできない。安田三郎『社会統計学』(丸善，1969).
4) 林の数量化理論に関しては，林知己夫，樋口伊佐夫，駒沢勉共著『情報処理と統計数理』(産業図書，1970) pp. 223-350 参照.
5) 富永健一『産業社会の動態』(東洋経済，1973) p. 237.
6) Paul F. Lazarsfeld and Allen H. Barton, "Quantitative Measurement in the Social Sciences: Classification, Typologies, and Indices" (in D. Lerner and H. D. Lasswell, ed., The Policy Sciences, Stanford University Press, 1951, pp. 152-192.
7) Sheldon G. Levy, "Polarization in Racial Attitudes," Public Opinion Quarterly, Vol. 36, 1972, pp. 221-234.
8) なお，融合の手続きに関しては，レヴィ自身は，必ずしも明確に定式化していない。図6の右下部（第4・5段階の黒人集団）からも推測できるように，彼の作業は，直観に依存したものと思われる。この項は，筆者があえてレヴィの直観的操作を定式化し，補充することを試みたものである。
9) Levy, op. cit.
10) ＡＩＤに関しては，以下の文献を参照のこと。① James N. Morgan and John A. Sonquist, "Problems in the Analysis of Survey Data and a Proposal," Journal of American Statistical Association, Vol. 58, June 1963, pp. 415-35. ② John A. Sonquist, "Finding Variables That Work," Public Opinion Quarterly, Vol. 33, 1969, pp. 83-95. ③ Henry Assael, "Segmenting Markets by Group Purchasing Behavior: An Application of the AID Technique," Journal of Marketing Research, Vol. 7, May

1970, pp. 153-158. ④ Richard Staelin, "Another Look at A.I.D.," Journal of Advertising Research, Vol. 11, No. 5, Oct. 1971, pp. 23-28.

11) Assael, op. cit.

12) 銘柄名は明らかにされておらず，たんに 'X' と表わされているのみである。

13) なお，外的基準のある場合の分析としては，以上3手法の他に，つぎのような手順からなる方法も考えられる。①外的基準が間隔尺度であるか名義尺度であるかに応じて，林の数量化理論第Ⅰ類または第Ⅱ類により，データ処理を行なう。なお，この際の独立変数の選択にあたっては，関連の高いもの同志が含まれることのないよう，十分に配慮する．②独立変数のなかから，偏相関係数の高い順に，上位 p 個までを選び出す。③ＰＳＡの3) と同様の手続きにより，樹形図を作成する。

この手法は，計量的手法とは呼び難いかもしれない。数量化理論を利用するとはいえ，樹形構造の形成は，まったくの手作業だからである。しかし，このようにして獲得された，最終集団への分割と，外的基準とのクロス集計を行なうことは，影響力の大きい要因間の交互作用を検出するための簡便法としても，有用であるように思われる。

14) G.N. Lance and W.T. Williams, "Computer Programs for Monothetic Classification ("Association Analysis")," Computer Journal, Vol. 8, 1965, pp. 246-249.

15) H.P. Friedman and J. Rubin, "On Some Invariant Criteria for Grouping Data," Journal of American Statistical Association, Vol. 62, 1967, pp. 1159-1178.

16) 個体間関連係数に関しては，次書を参照のこと。Robert R. Sokal and Peter H.A. Sneath, Principles of Numerical Taxonomy, Freeman & Company, 1963, pp. 125-140.

17) 個体間類似性の指標を用いるクラスター分析の手法は，きわめて多い．名義尺度型変数にもとづく個体間関連係数を使用する場合には，以下に述べられている手法が適切であろう。Stephen C. Johnson, "Hierarchical Clustering Schemes," Psychometrika, Vol. 32, No. 3, September, 1967, pp. 241-254.

An Approach to Multivariate Analysis of Nominal Scales: Some Statistical Methods for Reducing Multidimensional Attribute Space

Shiro Horiuchi

Résumé

In the area of sociological measurement, it is one of the most fundamental problems how data consisting of a large number of nominal scales can be analyzed and reduced. For data collected in social research are usually obtained by forcing respondents to select one out of categories listed on questionaire. It seems to the present writer that there are two kinds of approach to this problem. They are the followings:

1) This type of data can be summarized by introducing the composite variable α,

$$\alpha_i = \sum_{j=1}^{n} \sum_{k=1}^{k_j} \delta_j(jk) w_{jk}$$

where $\delta_i(jk) = \begin{cases} 1 \text{ (when the } i\text{-th individual responds to the } k\text{-th} \\ \quad \text{category of the } j\text{-th variable)} \\ 0 \text{ (otherwise)} \end{cases}$

w_{jk}=the value given to the k-th category of the j-th variable.

When a criterion function of w_{jk}, which is to be maximized or minimized in terms of the purpose of analysis, is considered, the value of w_{jk} is calculated so that the given criterion function is optimized. This type of method, "the Hayashi quantification theory," is well known among behavioral scientists in Japan, which is named after the founder of this methodology.

2) This type of data can also be summarized by means of the

new nominal scale, which is constructed i) by selecting a number of variables out of all given nominal scales, ii) by grouping some of categories which belong to the same variable selected, and iii) by combining categories which belong to the different variables. This is equivalent to grouping ultimate classes included in a multiple contingency table which shows interrelationships among nominal scales.

This procedure is named " reduction of multidimensional attribute space," which was formulated by P. F. Lazarsfeld and H. Barton in 1951. Though they practiced this operation in an intuitive manner, some statistical methods which reduce a multidimensional attribute space have been developed in the last ten years. They are PSA (Polarized Subgroup Analysis), AID (Automatic Interaction Detector), association analysis and so on.

These methods may be appropriately included under the label of " Optimal Tree-Structure Analysis." Since, the results analyzed by one of these techniques, are normally presented in dendrogram style.