カテゴリカルなデータにおける正準相関分析

——特性グループ間の相関に基く数量化——

統計数理研究所　仁　木　直　人

(1980年8月　受付)

Canonical Correlation Analysis of Categorical Data

—Quantification Based on the Correlation between Two Sets of Categorical Variates—

Naoto Niki
(The Institute of Statistical Mathematics)

A method of "quantification" based on the correlation between two sets of categorical variates is described. We consider "scores" x and y of categories in the first and second sets of variates and consider a particular pair of variates as "scores" of samples given by $u=D^{(1)}x$ and $v=D^{(2)}y$, where $D^{(t)}$ $(t=1, 2)$ are the response patterns of samples to the categories in the t-th set. Under the conditions $\bar{u}=\bar{v}=0$ the correlation coefficient between u and v attains its maximum ρ_1 for $x=x_1$ and $y=y_1$ provided that ρ_1^2 is the second largest eigenvalue of the equations $F^+GH^+G'x=\rho^2x$ and $H^+G'F^+Gy=\rho^2y$ with the eigenvectors x_1 and y_1, where $F=D^{(1)\prime}D^{(1)}$, $G=D^{(1)\prime}D^{(2)}$, $H=D^{(2)\prime}D^{(2)}$ and X^+ signifies a generalized inverse of X. The largest eigenvalue is always one but the associated eigenvectors do not satisfy the conditions $\bar{u}=\bar{v}=0$. We can also determine the optimum multi-dimensional "scores" x_1—x_S and y_1—y_S which are the successive eigenvectors associated with the second to $(S+1)$-th largest eigenvalues.

1. 序

カテゴリカルなデータの解析においては，2元クロス表を取扱うことが非常に多く，また数多くの解析手法が知られている．

この中で，Guttman [2] や林 [3] に始まる数量化は，「分析目的に応じて操作的に測度を定義し，その測度に基いて各区分（カテゴリ）に数値を割りつける」という非常に柔軟な考え方をするもので，クロス表の取扱いに関しても応用範囲は広大である．

いま，多くの特性項目（カテゴリカルな変数）について全ての組合せでの2元クロス表が与えられているとしよう．そして，これらの特性項目を2群に分け，2つの群間の関係を解析することを考える．典型的な群の分け方は「説明項目群」と「基準項目群」の2群とすることであるが，とくにそのような意識がなくとも2群に分けて考える方が合理的な場合も多いであろう．たとえば，「政治意識に関する項目群」と「教育問題に関する項目群」との相関を見たいとする場合などである．

項目を数群に分けて解析する試みは林 [4] によってすでに行なわれているが，これは分割された各群間の分離を一般化相関比を測度として最大化しようとするものである．

本稿では，これと異なり，2群に分かれた各群ごとにある一次結合を作り，その一次結合間

の相関を最も高くするように各区分に与える数値を決めようとするものである．いわば，カテゴリカルなデータにおける正準相関分析である．

なおごく最近，岩坪 [9] が「正準相関係数最大化の数量化法」を発表している．これは本稿の方法と基本的に同一のものである．しかし，根本となる発想には異なる面があり，解くべき方程式も本稿の形の方が多次元化した場合や既存の数量化法との関連をみる上で理解しやすいだろう．また，Okamoto and Endo [11] がまとめを行なった「正準相関を測度とした数量化」は普通「クロス表の数量化」と呼ばれるもの（形式的に通常の正準相関分析に一致する）であって，のちに述べるように本稿の特別な場合に相当する．

第2節では，数量化の基準となる測度の定義を行なったのち，解の満たすべき固有方程式を導く．第3節では，その固有方程式の基本的な性質について考察する．次に第4節では，各区分に与える数値を多次元化する拡張について述べる．第5節では，本稿の特別な場合に相当する，数量化II類，「クロス表の数量化」，数量化III類との関連について議論する．最後に第6節では，簡単な実例を用いて具体的な解法の説明を行なう．

2. カテゴリカルな変数における正準相関分析

n 個の対象について，ある1群の特性に関する $m^{(1)}$ 項目の測定データ，および，別の1群の特性に関する $m^{(2)}$ 項目の測定データがあるものとする．各測定項目はいくつかの区分に分割されていて，各対象は必ずそのうちのどれかひとつだけに該当することとする．

すなわち，データとしては，

$$\delta_{ijk}{}^{(t)} = \delta_i{}^{(t)}(l) = \begin{cases} 1 & \text{第 } i \text{ 対象が第 } t \text{ 群の第 } j \text{ 項目第 } k \text{ 区分に該当するとき，} \\ 0 & \text{該当しないとき，} \end{cases}$$

$$\left[\begin{array}{l} t = 1, 2;\ i = 1, 2, \cdots, n; \\ j = 1, 2, \cdots, m^{(t)};\ k = 1, 2, \cdots, c_j{}^{(t)} \\ l = 1, 2, \cdots, p^{(t)} \end{array}\right]$$

が与えられる．ただし，$c_j{}^{(t)}$ は第 t 群の第 j 項目の区分数であり，

$$\sum_{k=1}^{c_j{}^{(t)}} \delta_{ijk}{}^{(t)} = 1 \tag{1}$$

を満たす．また l は各群における区分の通し番号で，

$$l = k \quad (j = 1), \qquad l = k + \sum_{h=1}^{j-1} c_h{}^{(t)} \quad (j > 1),$$

$$p^{(t)} = \sum_{h=1}^{m^{(t)}} c_h{}^{(t)} \qquad \text{（各群での総区分数）}$$

である．

ここで，2群の特性の間になんらかの関係が想定される場合を考えてみよう．

いま，第1群の第 j 項目第 k 区分（通し番号で第 l 区分）に数値 $x_{jk} = x(l)$ を与え，第2群の第 j 項目第 k 区分（通し番号第 l 区分）に $y_{jk} = y(l)$ を与えることにする．そして，第 i 対象に対し，その該当する区分に与えた数値の合計をそれぞれ u_i, v_i すなわち，

$$u_i = \sum_{l=1}^{P^{(1)}} \delta_i{}^{(1)}(l)\, x(l),$$

$$v_i = \sum_{l=1}^{p^{(2)}} \delta_i{}^{(2)}(l)\, y(l)$$

としたとき，この u_i と v_i の相関係数を測度として，両者の相関が最も高くなるように，逆に $x(l)$, $y(l)$ の値をもとめようと考える．

これは，カテゴリカルな変数における正準相関分析とみることができよう．

計算の便利のため，各変数をマトリクス表示しておこう．

$x(l)$ $(l=1, 2, \cdots, p^{(1)})$, $y(l)$ $(l=1, 2, \cdots, p^{(2)})$; u_i, v_i $(i=1, 2, \cdots, n)$ を要素とする縦ベクトルをそれぞれ $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}$ で表わし，$\delta_i{}^{(t)}(l)$ $(t=1, 2;\ i=1, 2, \cdots, n;\ l=1, 2, \cdots, p^{(t)})$ を要素とする n 行 $p^{(t)}$ 列行列を $\boldsymbol{D}^{(t)}$ で表わす．明らかに

$$\boldsymbol{u} = \boldsymbol{D}^{(1)}\boldsymbol{x}$$

$$\boldsymbol{v} = \boldsymbol{D}^{(2)}\boldsymbol{y}$$

である．

いま，$\boldsymbol{x}, \boldsymbol{y}$ は u_i, v_i の平均がともに 0，すなわち，要素が全て 1 の n 次の縦ベクトルを $\mathbf{1}_n$ と書くことにすれば，$\mathbf{1}_n' = (1\ 1\ \cdots\ 1)$ で ($'$ は転置を表わす)，

$$n\bar{u} = \mathbf{1}_n'\boldsymbol{u} = \mathbf{1}_n'\boldsymbol{D}^{(1)}\boldsymbol{x} = 0$$

$$n\bar{v} = \mathbf{1}_n'\boldsymbol{v} = \mathbf{1}_n'\boldsymbol{D}^{(2)}\boldsymbol{y} = 0 \qquad (2)$$

となるように選ぶことにする．

すると，u_i および v_i の分散 $V^{(1)}, V^{(2)}$ および共分散 C は，

$$\boldsymbol{F} = \boldsymbol{D}^{(1)\prime}\boldsymbol{D}^{(1)}$$

$$\boldsymbol{H} = \boldsymbol{D}^{(2)\prime}\boldsymbol{D}^{(2)}$$

$$\boldsymbol{G} = \boldsymbol{D}^{(1)\prime}\boldsymbol{D}^{(2)}$$

とおけば，

$$V^{(1)} = \frac{1}{n}\boldsymbol{u}'\boldsymbol{u} = \frac{1}{n}\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}$$

$$V^{(2)} = \frac{1}{n}\boldsymbol{v}'\boldsymbol{v} = \frac{1}{n}\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}$$

$$C = \frac{1}{n}\boldsymbol{u}'\boldsymbol{v} = \frac{1}{n}\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}$$

よって，u_i と v_i の相関係数 ρ は

$$\rho = \frac{\frac{1}{n}\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}}{\sqrt{\frac{1}{n}\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}\cdot\frac{1}{n}\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}}}$$

と表わされる．なお，$\boldsymbol{F}$ は第 1 群の特性相互の，$\boldsymbol{H}$ は第 2 群の特性相互の，また $\boldsymbol{G}$ は第 1 群と第 2 群の特性間のクロス表になっていることを注意しておく．

さて，u_i, v_i のスケールは自由にとることができるが，それぞれの分散が

$$V^{(1)} = V^{(2)} = 1$$

となるようにするのが便利であろう．すなわち，

$$\frac{1}{n}\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}=1,$$

(3)

$$\frac{1}{n}\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}=1$$

なる条件のもとで,

$$\rho=\frac{1}{n}\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}$$

を最大とするような $\boldsymbol{x}$, $\boldsymbol{y}$ を決めよう. (2) の条件についてはのちに考慮する.

ラグランジェの未定定数法を適用する. 未定定数 λ, μ を導入して,

$$Q=\frac{1}{n}\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}-\frac{1}{2}\lambda\left(\frac{1}{n}\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}-1\right)-\frac{1}{2}\mu\left(\frac{1}{n}\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}-1\right)$$

とし, この Q を $\boldsymbol{x}, \boldsymbol{y}$ の各要素で偏微分したものをマトリクス表示すれば, 形式的に

$$\frac{\partial Q}{\partial \boldsymbol{x}}=\frac{1}{n}\boldsymbol{G}\boldsymbol{y}-\frac{1}{n}\lambda\boldsymbol{F}\boldsymbol{x}$$

$$\frac{\partial Q}{\partial \boldsymbol{y}}=\frac{1}{n}\boldsymbol{G}'\boldsymbol{x}-\frac{1}{n}\mu\boldsymbol{H}\boldsymbol{y}$$

と書ける. ゆえに,

$$\boldsymbol{G}\boldsymbol{y}=\lambda\boldsymbol{F}\boldsymbol{x}$$
$$\boldsymbol{G}'\boldsymbol{x}=\mu\boldsymbol{H}\boldsymbol{y}$$

を解けば求める解が得られる.

これらの式にそれぞれ $\boldsymbol{x}'$ および $\boldsymbol{y}'$ を左からかければ,

$$\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}=\lambda\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}=n\lambda$$
$$\boldsymbol{y}'\boldsymbol{G}'\boldsymbol{x}=\mu\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}=n\mu$$

となり, 未定定数 λ, μ はともに u_i と v_i の相関係数

$$\rho=\frac{1}{n}\boldsymbol{x}'\boldsymbol{G}\boldsymbol{y}=\frac{1}{n}\boldsymbol{y}'\boldsymbol{G}'\boldsymbol{x}$$

に等しいことがわかる. よって

$$\boldsymbol{G}\boldsymbol{y}=\rho\boldsymbol{F}\boldsymbol{x}$$

(4)

$$\boldsymbol{G}'\boldsymbol{x}=\rho\boldsymbol{H}\boldsymbol{y}$$

を解くことになる.

いま議論を簡単にするため, 2つの対称行列 $\boldsymbol{F}$ および $\boldsymbol{H}$ の一般化逆行列のうち, 各群の第1項目を除き, 他の項目の第1区分に対応する行および列の要素が全て0であるものを, それぞれ $\boldsymbol{F}^+$, $\boldsymbol{H}^+$ と書くことにする. 一般にこれらは一意に定まり,

$$\mathrm{rank}\,(\boldsymbol{F}^+)=\mathrm{rank}\,(\boldsymbol{F})$$
$$\mathrm{rank}\,(\boldsymbol{H}^+)=\mathrm{rank}\,(\boldsymbol{H})$$

を満たす反射型一般化逆行列である.

このような $\boldsymbol{F}^+$, $\boldsymbol{H}^+$ を選んだことは,

$$x_{j1}=0 \qquad (j=2,3,\cdots,m^{(1)})$$
$$y_{j1}=0 \qquad (j=2,3,\cdots,m^{(2)})$$

なる条件をつけることにより，(4) の特解が

$$\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{y}=\rho\boldsymbol{x} \tag{5}$$
$$\boldsymbol{H}^{+}\boldsymbol{G}'\boldsymbol{x}=\rho\boldsymbol{y}$$

により一意に定まることを意味する．

(5) を互いに代入しあうことにより，結局

$$\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{H}^{+}\boldsymbol{G}'\boldsymbol{x}=\rho^{2}\boldsymbol{x}$$
$$\boldsymbol{H}^{+}\boldsymbol{G}'\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{y}=\rho^{2}\boldsymbol{y}$$

なる固有値問題に帰着する．

3. 固有値問題の基本的性質

$$\boldsymbol{A}=\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{H}^{+}\boldsymbol{G}'$$
$$\boldsymbol{B}=\boldsymbol{H}^{+}\boldsymbol{G}'\boldsymbol{F}^{+}\boldsymbol{G}$$

と書くことにする．$\boldsymbol{A}$ と $\boldsymbol{B}$ の固有値は共通であり，一般に $\mathrm{rank}\,(\boldsymbol{G})=\min(p^{(1)}-m^{(1)}+1,\ p^{(2)}-m^{(2)}+1)$ 個の 0 でない値を持つ．

このうち，$\rho^{2}=1$ なる固有値が含まれていることを述べる．

いま，始めから $c_1{}^{(1)}$, $c_1{}^{(2)}$ 個の要素が 1 で他の全要素が 0 である $p^{(1)}$, $p^{(2)}$ 次の縦ベクトルをそれぞれ $\boldsymbol{x}_*$, $\boldsymbol{y}_*$ とする．すなわち，

$$\boldsymbol{x}_*=\begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\end{pmatrix}\begin{matrix}\left.\begin{matrix}\\ \\ \\ \end{matrix}\right\}c_1{}^{(1)},\\ \\ \\ \\ \\ \end{matrix} \qquad \boldsymbol{y}_*=\begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\end{pmatrix}\begin{matrix}\left.\begin{matrix}\\ \\ \\ \end{matrix}\right\}c_1{}^{(2)}\\ \\ \\ \\ \\ \end{matrix}$$

である．

(1) より明らかに

$$\boldsymbol{F}\boldsymbol{x}_*=n\,\boldsymbol{1}_{p^{(1)}}$$
$$\boldsymbol{G}'\boldsymbol{x}_*=n\,\boldsymbol{1}_{p^{(2)}}$$
$$\boldsymbol{G}\boldsymbol{y}_*=n\,\boldsymbol{1}_{p^{(1)}}$$
$$\boldsymbol{H}\boldsymbol{y}_*=n\,\boldsymbol{1}_{p^{(2)}}$$

であるから，

$$\boldsymbol{A}\boldsymbol{x}=\rho^{2}\boldsymbol{x} \tag{6}$$
$$\boldsymbol{B}\boldsymbol{y}=\rho^{2}\boldsymbol{y}$$

は，$\rho^{2}=1$, $\boldsymbol{x}=\boldsymbol{x}_*$, $\boldsymbol{y}=\boldsymbol{y}_*$ なる解を持つ．

しかし，この解は (3) の条件

$$\frac{1}{n}\,\boldsymbol{x}_*\boldsymbol{F}\boldsymbol{x}_*=1$$

$$\frac{1}{n}\boldsymbol{y}_*\boldsymbol{H}\boldsymbol{y}_* = 1$$

を満たすが，(2) の条件は

$$\mathbf{1}_n'\boldsymbol{D}^{(1)}\boldsymbol{x}_* = n \neq 0$$
$$\mathbf{1}_n'\boldsymbol{D}^{(2)}\boldsymbol{y}_* = n \neq 0$$

より満たすことができない．すなわち，$\rho^2=1$, $\boldsymbol{x}=\boldsymbol{x}_*$, $\boldsymbol{y}=\boldsymbol{y}_*$ は排除されるべき形式解である．

次に $\rho^2\neq 1$ なる固有値に対応する固有ベクトル $\boldsymbol{x}$, $\boldsymbol{y}$ が条件 (2) を満たすことを述べる．

$$\begin{aligned}\mathbf{1}_{p^{(1)}}'\boldsymbol{F}\boldsymbol{x} &= \mathbf{1}_{p^{(1)}}'\boldsymbol{D}^{(1)\prime}\boldsymbol{D}^{(1)}\boldsymbol{x}\\ &= m_1\mathbf{1}_n'\boldsymbol{D}^{(1)}\boldsymbol{x}\\ &= m_1 n\bar{u}\\ \mathbf{1}_{p^{(1)}}'\boldsymbol{G}\boldsymbol{y} &= m_1\mathbf{1}_n'\boldsymbol{D}^{(2)}\boldsymbol{y}\\ &= m_1 n\bar{v}\\ \mathbf{1}_{p^{(2)}}'\boldsymbol{G}'\boldsymbol{x} &= m_2\mathbf{1}_n'\boldsymbol{D}^{(1)}\boldsymbol{x}\\ &= m_2 n\bar{u}\\ \mathbf{1}_{p^{(2)}}'\boldsymbol{H}\boldsymbol{y} &= m_2\mathbf{1}_n'\boldsymbol{D}^{(2)}\boldsymbol{y}\\ &= m_2 n\bar{v}\end{aligned}$$

であるから，(4) の両辺にそれぞれ $\mathbf{1}_{p^{(1)}}'$ および $\mathbf{1}_{p^{(2)}}'$ を左からかけることにより，

$$m_1 n\bar{v} = \rho m_1 n\bar{u}$$
$$m_2 n\bar{u} = \rho m_2 n\bar{v}$$

を得る．ゆえに

$$(1-\rho^2)\bar{u} = 0$$
$$(1-\rho^2)\bar{v} = 0$$

より，$\rho^2\neq 1$ ならば $\bar{u}=0$ かつ $\bar{v}=0$ となって条件 (2) を満たす．

したがって $\boldsymbol{A}, \boldsymbol{B}$ の最大固有値1に次ぐ，2番目に大きい固有値 $\rho_1{}^2$ を求めれば，それが u_i, v_i の相関係数の最大値（正準相関係数に相当する）を与え，また対応する固有ベクトル $\boldsymbol{x}_1$, $\boldsymbol{y}_1$ はそれぞれ第1群，第2群の各区分に付する最適な数値を与える．

実際に (6) を解く際には，$\boldsymbol{x}, \boldsymbol{y}$ の0でない要素数 $(p^{(1)}-m^{(1)}+1)$, $(p^{(2)}-m^{(2)}+1)$ の小さい方について先に解を求め，次いで (5) によりもう一方を求めるという手続きに従う．

なお，区分に与えた数値を各項目ごとの平均が0となるように修正しておくと結果が見やすい．それには，各項目ごとにある値 $\xi_j(j\neq 1, 2, \cdots, m^{(1)})$, $\eta_j(j=1, 2, \cdots, m^{(2)})$ を各区分に与えた数値 x_{jk}, y_{jk} に加え，

$$\sum_{i=1}^{n}\sum_{k=1}^{c_j{}^{(1)}}\delta_{ijk}{}^{(1)}(x_{jk}+\xi_j) = 0$$

$$\sum_{i=1}^{n}\sum_{k=1}^{c_j{}^{(2)}}\delta_{ijk}{}^{(2)}(y_{jk}+\eta_j) = 0$$

が満足されれば，x_{jk}, y_{jk} の代わりに $x_{jk}+\xi_j, y_{jk}+\eta_j$ を採用することにする．このような修正が可能なことは，ξ_j, η_j を $c_j{}^{(1)}, c_j{}^{(2)}$ 個ずつ順に並べた $p^{(1)}, p^{(2)}$ 次の縦ベクトル $\boldsymbol{\xi}, \boldsymbol{\eta}$ が，

$$\boldsymbol{F}\boldsymbol{x} = \mathbf{0}_{p^{(1)}} = 0\cdot\mathbf{1}_{p^{(1)}}$$
$$\boldsymbol{H}\boldsymbol{y} = \mathbf{0}_{p^{(2)}} = 0\cdot\mathbf{1}_{p^{(2)}}$$

の解であることから明らかで，$\boldsymbol{F}$, $\boldsymbol{H}$ の一般化逆行列を適当に選び直すことに対応する．

4. 多次元への拡張

次に各区分に多次元の数値を付すことを考えよう．

第1群および第2群の各区分に付すS次元の数値をベクトルとして，

$$\begin{matrix} {}^1\boldsymbol{x},\ {}^2\boldsymbol{x},\ \cdots,\ {}^S\boldsymbol{x} \\ {}^1\boldsymbol{y},\ {}^2\boldsymbol{y},\ \cdots,\ {}^S\boldsymbol{y} \end{matrix} \qquad (S \leq S_0 = \mathrm{rank}\,(\boldsymbol{G}) - 1)$$

で表わす．またそれらによって第 i 対象に与えられる数値 ${}^su_i$, ${}^sv_i$ $(s=1, 2, \cdots, S;\ i=1, 2, \cdots, n)$ をベクトル化して，

$$\begin{matrix} {}^s\boldsymbol{u} = \boldsymbol{D}^{(1)}\,{}^s\boldsymbol{x} \\ {}^s\boldsymbol{v} = \boldsymbol{D}^{(2)}\,{}^s\boldsymbol{y} \end{matrix} \qquad (s = 1, 2, \cdots, S)$$

で表わす．

ここで，1次元の場合と同じく，${}^s\boldsymbol{u}$, ${}^s\boldsymbol{v}$ の平均に

$$\begin{matrix} n\,{}^s\bar{u} = \boldsymbol{1}_n'\,{}^s\boldsymbol{u} = 0 \\ n\,{}^s\bar{v} = \boldsymbol{1}_n'\,{}^s\boldsymbol{v} = 0 \end{matrix} \tag{7}$$

なる条件を，さらに，分散および共分散について

$$\begin{aligned} \frac{1}{n}\,{}^r\boldsymbol{u}'\,{}^s\boldsymbol{u} &= \frac{1}{n}\,{}^r\boldsymbol{x}'\boldsymbol{F}\,{}^s\boldsymbol{x} = \begin{cases} 1 & (r = s) \\ 0 & (r \neq s) \end{cases} \\ \frac{1}{n}\,{}^r\boldsymbol{v}'\,{}^s\boldsymbol{v} &= \frac{1}{n}\,{}^r\boldsymbol{y}'\boldsymbol{H}\,{}^s\boldsymbol{y} = \begin{cases} 1 & (r = s) \\ 0 & (r \neq s) \end{cases} \\ \frac{1}{n}\,{}^r\boldsymbol{u}'\,{}^s\boldsymbol{v} &= \frac{1}{n}\,\boldsymbol{x}'\boldsymbol{G}\,{}^s\boldsymbol{y} = 0 \qquad (r \neq s) \end{aligned} \tag{8}$$

なる条件を課すことにする．すなわち，${}^s\boldsymbol{u}$, ${}^s\boldsymbol{v}$ は次元間で互いに直交する．

このとき，各対象に与えた2つのS次元ベクトル

$$\begin{matrix} ({}^1u_i\ {}^2u_i \cdots {}^Su_i) \\ ({}^1v_i\ {}^2v_i \cdots {}^Sv_i) \end{matrix}$$

のベクトル相関を考える．すなわち，${}^s\boldsymbol{u}$, ${}^s\boldsymbol{v}$ $(s=1, 2, \cdots, S)$ を列ベクトルとする $n \times S$ 行列をそれぞれ $\boldsymbol{U}, \boldsymbol{V}$ と書き，回帰モデル

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{C} + \boldsymbol{E} \tag{9}$$

を想定する．誤差の一般化分散

$$V_\varepsilon = \left| \frac{1}{n}\boldsymbol{E}'\boldsymbol{E} \right|$$

を最小化するような係数行列は，(7) (8) より

$$\begin{aligned} \hat{\boldsymbol{C}} &= (\boldsymbol{U}'\boldsymbol{U})^{-1}(\boldsymbol{U}'\boldsymbol{V}) \\ &= \begin{pmatrix} {}^1\rho & & 0 \\ & {}^2\rho & \\ & \ddots & \\ 0 & & {}^S\rho \end{pmatrix} \end{aligned} \qquad \left(\text{ただし，}\ {}^s\rho = \frac{1}{n}\,{}^s\boldsymbol{x}'\boldsymbol{G}\,{}^s\boldsymbol{y}\right)$$

となり，また，$\hat{\boldsymbol{C}}$ を (9) に与えて最小となった V_ϵ は

$$\min V_\epsilon = \left| \frac{1}{n} (\boldsymbol{V}-\boldsymbol{U}\hat{\boldsymbol{C}})'(\boldsymbol{V}-\boldsymbol{U}\hat{\boldsymbol{C}}) \right|$$

$$= |\boldsymbol{I}-\hat{\boldsymbol{C}}^2| \qquad (\boldsymbol{I} \text{は単位行列})$$

$$= \prod_{s=1}^{S} (1-{}^s\rho^2)$$

と表わすことができる．よって，$\boldsymbol{U}$ と $\boldsymbol{V}$ の相関係数 ${}^SR$ としては，

$${}^SR^2 = 1-(\min V_\epsilon) \Big/ \left(\frac{1}{n}\boldsymbol{V}'\boldsymbol{V}\right)$$

$$= 1-\prod_{s=1}^{S} (1-{}^s\rho^2)$$

で定義するのが自然である．

この ${}^SR$ あるいは ${}^SR^2$ を最大化するような ${}^s\boldsymbol{x}$, ${}^s\boldsymbol{y}$ は，(6) において $\boldsymbol{A}$, $\boldsymbol{B}$ の 1 を除く固有値のうち大きいものから S 個とって

$$\rho_1{}^2, \rho_2{}^2, \cdots, \rho_S{}^2$$

とし，それらに対応する固有ベクトルを

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_S$$
$$\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_S$$

としたとき，

$${}^s\boldsymbol{x} = \boldsymbol{x}_s, \qquad (s = 1, 2, \cdots, S)$$
$${}^s\boldsymbol{y} = \boldsymbol{y}_s$$

とすればよいことはただちにわかる．このとき ${}^s\rho^2=\rho_s{}^2$ $(s=1, 2, \cdots, S)$ であって，

$${}^SR^2 = 1-\prod_{s=1}^{S} (1-\rho_s{}^2)$$

である．

なお，${}^SR^2$ の代わりの測度として，Wilks [13] の Λ,

$$\Lambda = 1-{}^SR^2$$

$$= \prod_{s=1}^{S} (1-{}^s\rho^2) \qquad \text{（最小化）}$$

や Hotelling [6] の一般化 $T_0{}^2$,

$$T_0{}^2 = \sum_{s=1}^{S} \frac{{}^s\rho^2}{1-{}^s\rho^2} \qquad \text{（最大化）}$$

Pillai [12] の V,

$$V = \sum_{s=1}^{S} {}^s\rho^2 \qquad \text{（最大化）}$$

などを用いても同じ解を得る．

5. 特別な場合

ここでは，既存の数量化手法と本稿の方法との関連について主に議論する．

a. 数量化 II 類

いま，2群に分かれた項目群の一方が，ただひとつの項目から成る場合を考える．すなわち $m^{(2)}=1$ とする．(もちろん $m^{(1)}=1$ としても全く同じ議論ができる．)

このとき $\boldsymbol{H}$ は対角要素以外は全て0である対角行列となり，逆行列 $\boldsymbol{H}^{-1}$ をもつ．よって $\boldsymbol{y}$ について解くべき式は，

$$\boldsymbol{H}^{-1}\boldsymbol{G}'\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{y}=\rho^2\boldsymbol{y}$$

となるが，これは数量化 II 類の場合と全く一致する ([7], [10])．

小さな違いとしては，固有値が相関係数の2乗の代わりに相関比と意味づけされていること $\boldsymbol{x}$ や $\boldsymbol{y}$ に与えるスケールが (3) に従わないこと，があげられるがともに本質的なものではない

また岩坪 [8] の「重相関係数による数量化」は $\boldsymbol{G}'\boldsymbol{F}^{+}\boldsymbol{G}$ のひとつを具体的に与える方法を述べたものと解される．

b. クロス表の数量化 ([1], [11] など)

第1群，第2群のどちらもがただひとつの項目から成る $m^{(1)}=m^{(2)}=1$ の場合を考える．もちろん数量化 II 類の特別な場合でもある．

このとき，$\boldsymbol{x}$ および $\boldsymbol{y}$ の満たす固有方程式は

$$\boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}\boldsymbol{G}'\boldsymbol{x}=\rho^2\boldsymbol{x}$$
$$\boldsymbol{H}^{-1}\boldsymbol{G}'\boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{y}=\rho^2\boldsymbol{y}$$

であって，いわゆる「クロス表の数量化」の式に一致する．

c. 数量化 III 類

数量化 III 類 [3] は，形式的には $\boldsymbol{G}$ の要素が0または1のみから成る「クロス表の数量化」と考えることができる．しかし，問題の設定やデータの与え方には独特なものがあり，一概にこの系列に入れてしまうのは少し無理があるかもしれない．

一方，数量化 III 類の適用例には，各測定対象についての数種の特性項目からなる第1群と「測定対象」という単一項目からなる第2群との間の相関を扱っているとみられるものが多いことも事実である．この場合，第1群の特性項目間の直接の相関は無視されていることに注意を払う必要がある．

6. 簡単な計算例

表1はある本 [5] の11人の共著者のプロフィールを彼らの神秘感に関する設問に対する解答パターンとして示したものである．いずれも，肯定的な解答を1，否定的な解答を2で表現してある．

各解答（すなわち区分）間のクロス集計を行なうと，次の表2を得る．

ここで「信仰」「神」を第1群の項目，「UFO」「ネッシー」を第2群の項目とし，両群の相関関係を最も高くするように数量化を行なってみよう．すなわち，この11人の共著者に関して，第1群の「宗教観」と第2群の「科学的側面を持つ神秘感」との関わりを知ろうとする．

表 1 共著者のプロフィール

	宗教を信仰	神の存在	UFOの存在	ネッシーの存在
1	1	1	1	2
2	1	1	2	1
3	2	1	1	1
4	2	1	1	1
5	2	1	1	2
6	2	1	2	1
7	2	2	1	1
8	2	2	1	1
9	2	2	2	1
10	2	2	2	1
11	2	2	2	2

表2 クロス表

		信仰		神		UFO		ネッシー	
		1	2	1	2	1	2	1	2
信仰	1	2	0	2	0	1	1	1	1
	2	0	9	4	5	5	4	7	2
神	1	2	4	6	0	4	2	4	2
	2	0	5	0	5	2	3	4	1
UFO	1	1	5	4	2	6	0	4	2
	2	1	4	2	3	0	5	4	1
ネッシー	1	1	7	4	4	4	4	8	0
	2	1	2	2	1	2	1	0	3

本稿の記述法に従い,

$$\boldsymbol{F}=\begin{pmatrix}2&0&2&0\\0&9&4&5\\2&4&6&0\\0&5&0&5\end{pmatrix},\quad \boldsymbol{H}=\begin{pmatrix}6&0&4&2\\0&5&4&1\\4&4&8&0\\2&1&0&3\end{pmatrix},\quad \boldsymbol{G}=\begin{pmatrix}1&1&1&1\\5&4&7&2\\4&2&4&2\\2&3&4&1\end{pmatrix}$$

とする. $\boldsymbol{F}$ および $\boldsymbol{H}$ の一般化逆行列 $\boldsymbol{F}^{+}, \boldsymbol{H}^{+}$ としては, どちらも第2行の要素が全て0であるものを選べば,

$$\boldsymbol{F}^{+}\boldsymbol{G}=\begin{pmatrix}-1/4&1/4&-1/4&1/4\\0&0&0&0\\3/4&1/4&3/4&1/4\\2/5&3/5&4/5&1/5\end{pmatrix},$$

$$\boldsymbol{H}^{+}\boldsymbol{G}'=\begin{pmatrix}-1/16&1/16&1/4&-1/4\\0&0&0&0\\5/32&27/32&3/8&5/8\\3/8&5/8&1/2&1/2\end{pmatrix}$$

となるから，

$$\boldsymbol{A}=\boldsymbol{F}^{+}\boldsymbol{G}\boldsymbol{H}^{+}\boldsymbol{G}'$$
$$=\begin{pmatrix} 9/128 & -9/128 & -1/32 & 1/32 \\ 0 & 0 & 0 & 0 \\ 21/128 & 107/128 & 19/32 & 13/32 \\ 7/40 & 33/40 & 1/2 & 1/2 \end{pmatrix}$$

が得られる．

$\boldsymbol{A}$ が固有値 1 を持つことは

$$\boldsymbol{A}\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

より確かめられる．

さて，

(10) $$\boldsymbol{A}\boldsymbol{x}=\rho^2\boldsymbol{x}$$

なる固有値問題を解くことになるが，$x_{12}=x(2)=0$ は明らかであるから，$\boldsymbol{A}$ の第 2 行および第 2 列を取り除いた行列を $\boldsymbol{A}_1$ として，

$$\boldsymbol{A}_1\boldsymbol{z}=\rho^2\boldsymbol{z}$$

なる固有値問題を解けば充分である．さらに，この式は

$$\rho^2=1, \qquad \boldsymbol{z}_*'=(0\ 1\ 1)$$

なる形式解を持っていることから，単位行列の第 3 列を $\boldsymbol{z}_*$ でおきかえた行列を $\boldsymbol{T}$ とし，$\boldsymbol{T}$ による $\boldsymbol{A}_1$ の相似変換

$$\boldsymbol{A}_2=\boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}$$

を考える．ここに

$$\boldsymbol{T}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{T}^{-1}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$$

である．

すると，$\boldsymbol{A}_2$ の第 3 列は $(0\ 0\ 1)'$ となり，

$$\boldsymbol{A}_2=\begin{pmatrix} 9/128 & -1/32 & 0 \\ 7/640 & 3/32 & 0 \\ 7/40 & 1/2 & 1 \end{pmatrix}$$

である．ゆえに $\boldsymbol{A}_2$ の第 3 行および第 3 列をとり除いた行列を

$$\boldsymbol{A}_3=\begin{pmatrix} 9/128 & -1/32 \\ 7/640 & 3/32 \end{pmatrix}$$

とおけば，(10) の固有値問題は $\rho^2=1$ の解を含まない

(11) $$\boldsymbol{A}_3\boldsymbol{w}=\rho^2\boldsymbol{w}$$

まで簡単化される．$\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}$ の関係を整理しておけば，

$$\begin{aligned}\boldsymbol{x}' &= (x(1)\ \ 0\ \ x(3)\ \ x(4))\\ \boldsymbol{z}' &= (x(1)\ \ x(3)\ \ x(4))\\ \boldsymbol{w}' &= (x(1)\ \ x(3) - x(4))\end{aligned}$$

である．

(11) の固有値 ρ^2 は2次方程式の2根として与えられるから，計算により，

$$\begin{aligned}\rho_1^2 &= 0.104, \qquad \rho_2^2 = 0.0601\\ \rho_1 &= 0.322, \qquad \rho_2 = 0.245\end{aligned}$$

またこれらに対する固有ベクトルは，

$${}^1\boldsymbol{w} = w_1 \begin{pmatrix} 1 \\ -1.075 \end{pmatrix}, \qquad {}^2\boldsymbol{w} = w_2 \begin{pmatrix} 1 \\ 0.325 \end{pmatrix}$$

と求まる（w_1, w_2 は0でない任意定数）．

$\boldsymbol{w}$ から $\boldsymbol{x}$ を求める際には，(2) の $\bar{u}=0$ の条件より ${}^sx(4)/w_s\ (s=1,2)$ を決定し，次いで (3) の $V^{(1)}=1$ の条件より w_1, w_2 を定める．このようにして

$${}^1\boldsymbol{x} = \begin{pmatrix} 1.970 \\ 0 \\ -1.321 \\ 0.797 \end{pmatrix}, \qquad {}^2\boldsymbol{x} = \begin{pmatrix} 2.091 \\ 0 \\ -0.071 \\ -0.751 \end{pmatrix}$$

を得る．次いで (5) より $\boldsymbol{y} = (1/\rho)\boldsymbol{H}^+\boldsymbol{G}'\boldsymbol{x}$ として

$${}^1\boldsymbol{y} = \begin{pmatrix} -2.025 \\ 0 \\ 0.964 \\ 1.479 \end{pmatrix}, \qquad {}^2\boldsymbol{y} = \begin{pmatrix} 0.161 \\ 0 \\ -0.691 \\ 1.521 \end{pmatrix}$$

が求まる．このとき $\boldsymbol{x} = (1/\rho)\boldsymbol{F}^+\boldsymbol{G}\boldsymbol{y}$ となっていることも確められる．

さらに結果を見やすくするために，各項目ごとに平均を0としておくと，

$${}^1\boldsymbol{x} = \begin{pmatrix} 1.612 \\ -0.358 \\ -0.963 \\ 1.155 \end{pmatrix}, \qquad {}^2\boldsymbol{x} = \begin{pmatrix} 1.710 \\ -0.380 \\ 0.309 \\ -0.371 \end{pmatrix},$$

$${}^1\boldsymbol{y} = \begin{pmatrix} -0.920 \\ 1.104 \\ -0.141 \\ 0.375 \end{pmatrix}, \qquad {}^2\boldsymbol{y} = \begin{pmatrix} 0.073 \\ -0.088 \\ -0.603 \\ 1.608 \end{pmatrix}$$

となる．結果は図示すると見やすいだろう．図1〜3はその例で，図1は ${}^1\boldsymbol{x}, {}^1\boldsymbol{y}$ を横軸，${}^2\boldsymbol{x}, {}^2\boldsymbol{y}$ を縦軸として各区分に与えた数値をプロットしたものであり，図2および図3は各共著者に与えた ${}^1\boldsymbol{u}$ と ${}^1\boldsymbol{v}$ の組合せおよび ${}^2\boldsymbol{u}$ と ${}^2\boldsymbol{v}$ の組合せをプロットしたものである．

なお以上の計算には数量化 II 類について述べた [10] が参考になるだろう．ただし，計算機のプログラムを組む上からは，上記のような手続きは必ずしも得策とはいえない．$\boldsymbol{H}^+$ および $\boldsymbol{G}'\boldsymbol{F}^+\boldsymbol{G}$ などの対称性を積極的に使った方法を用いるべきであろう．

執筆にあたり，水野欽司氏から貴重な助言をいただいた．ここに記して感謝いたします．

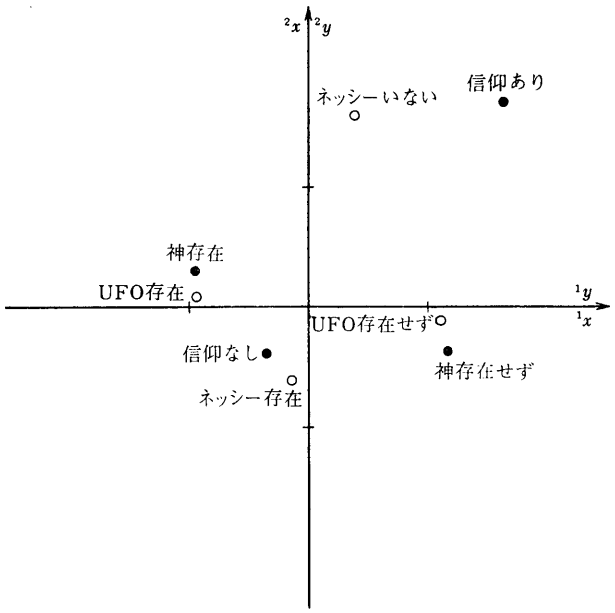
^2x ^2y
ネッシーいない
信仰あり
神存在
UFO存在
^1y
^1x
UFO存在せず
信仰なし
神存在せず
ネッシー存在

図 1

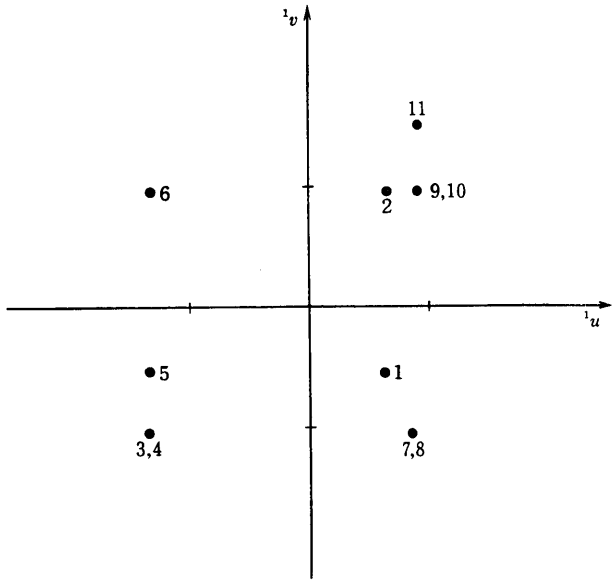
^1v
11
6
9,10
2
^1u
5
1
3,4
7,8

図 2

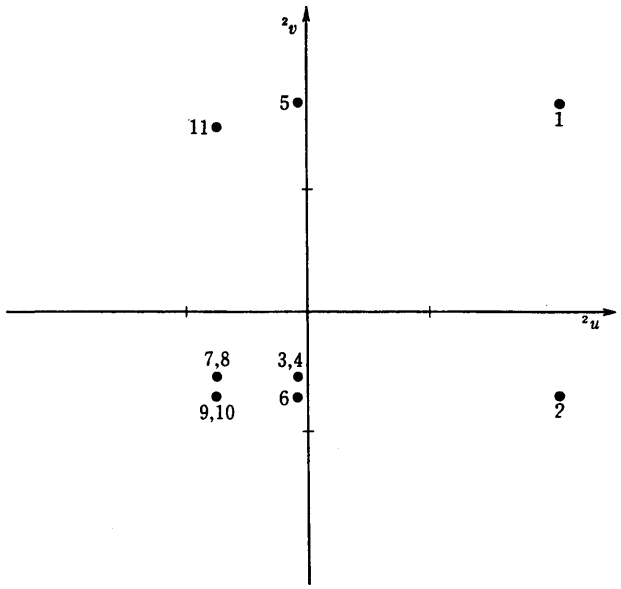

図 3

参 考 文 献

[1] Benzécri, J.P. (1973). *L'Analyse des Données*, Tom 2 L'Analyse des Correspondances, Dunod.

[2] Guttman, I. (1946). An approach for quantifying paired comparison and rank order, *Ann. Math. Statist.*, **17**, 144–163.

[3] Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, *Ann. Inst. Statist. Math.*, **3**, 69–98.

[4] Hayashi, C. (1954). Multidimensional quantification — with the applications to analysis of social phenomena —, *Ann. Inst. Statist. Math.*, **5**, 121–143.

[5] 林知己夫，飽戸 弘（編）(1976)．多次元尺度解析法，サイエンス社．

[6] Hotelling, H. (1951). A generarized T-test and measure of multivariate dispersion, *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, 23–41.

[7] 石田正次 (1979)．データ解析の基礎，森北出版．

[8] 岩坪秀一 (1971)．多次元質的データ構造の解析 ——重相関係数による数量化——，電子技術総合研究所彙報，**35**，622-632．

[9] 岩坪秀一 (1979)．n-way 質的データの多変量解析的手法の研究，電子技術総合研究所研究報告，No 801．

[10] 仁木直人 (1979)．数量化II類の解法について，統計数理研究所彙報，**26**，115-123．

[11] Okamoto, M. and Endo, H. (1973). Basic properties of categorical canonical correlation analysis, *J. Japan Statist. Soc.* **4**, 15–23.

[12] Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis, *Ann. Math. Statist.*, **26**, 117–121.

[13] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika*, **24**, 471–494.