数量化理論II類の解説とその適用例

Hayashi's Quantification Method Type II and its Application

ネットワーク情報学部　永添めぐみ　村上明日香　佐藤 創

School of Network and Information　Megumi NAGASOE, Asuka MURAKAMI, Hajime SATO

Keywords: Hayashi's Quantification Method, Variance Ratio, Generalized Eigenvalue Problem

まえがき

本稿は本誌 No. 11 に掲載された文献 [1]（以下ではこれを「前稿」と呼ぶ）の続編である．本稿で述べるデータ解析法の適用例には前稿に引き続き，2006 年度「プロジェクト」で行った「メッセンジャーの使用状況に関するアンケート調査」のデータを用いる．

数量化 II 類と呼ばれる方法は，多変量解析法の一つとして世に知られているが，本学部では周知とは言えない．この点で数量化 I 類と“同類”である．この方法も大変に適用範囲が広く，多くの人に使ってもらいたいと思って紹介する．

数量化 II 類の解説は佐藤が，前回の反省点と今回の結果，および SPSS の利用については永添，村上が記す．

1　前回の説明と反省点

まず，前回の説明と反省から書き始める．

アンケートで得たデータの解析目標は，インターネットの「メッセンジャー」の使用頻度の違いが，電話の使用頻度，メールの使用頻度，性別，所属学部によってどれくらい説明できるかを明らかにすることであった．

これらの説明変量はいずれもカテゴリカルな（名義型の）変量であるが，計算過程が比較的簡単な数量化 I 類による解析を試みたために，被説明変量（外的基準と呼ばれる．変数 Z で表す）の値を次のように単純素朴に数量化したのであった（I 類法では非説明変量の型は数量である）．

Z	メッセンジャー使用頻度 (1 日)	人数	Z の値
Z_1	使ったことがない	111	0.0
Z_2	普段全く使わない	44	1.0
Z_3	1〜2 回	58	2.0
Z_4	3〜4 回	19	3.0
Z_5	5〜6 回	8	4.0
Z_6	毎日	30	5.0

その結果，各変量を数量化した値（カテゴリー値と呼ばれる）が次の表のように得られた（X_5 の値は訂正）．カテゴリー値の大きいほど，その選択肢を選ぶ人はメッセンジャーの使用頻度が高い傾向にあることを表す．

X	電話使用頻度 (1 日)	人数	カテゴリー値
X_1	使わない	2	0.495
X_2	0 分	5	−0.064
X_3	1〜10 分	142	−0.085
X_4	11〜30 分	42	0.348
X_5	それ以上	33	−0.005
Y	メール使用頻度 (1 日)	人数	カテゴリー値
Y_1	0 通	14	−0.615
Y_2	1〜5 通	90	0.098
Y_3	6〜10 通	85	0.125
Y_4	11〜20 通	52	0.018
Y_5	それ以上	29	−0.409
U	性別	人数	カテゴリー値
U_1	男	198	0.208
U_2	女	72	−0.573
V	所属学部	人数	カテゴリー値
V_1	経営	84	−1.049
V_2	ネットワーク情報	186	0.474

前回の分析結果は次の表に集約される．数量化 I 類法は回帰分析の一般化であるから，被説明変量 Z との相関係数が大きい説明変量ほど，説明力が強い．レンジはカテゴリー値の最大最小の差である．

変量	単相関	偏相関	レンジ	順位
X(電話)	0.074	0.109	0.581	(4)
Y(メール)	0.120	0.146	0.740	(3)
U(性別)	0.194	0.233	0.781	(2)
V(学部)	0.422	0.440	1.523	(1)

重相関係数　$r_{z,xyuv} = 0.494$

これより，メッセンジャーの使用頻度を最も説明する変量は所属学部，次いで性別であり，ネットワーク情報学部の男子学生がメッセンジャーを多用している実態が明らかにされた．しかし，4 つの説明変量の値から線形回帰式で得られる Z の予測値と実測値との相関係数は 0.494 程度であり，満足できる結果とは言えなかった．

前稿の中でも触れたが，本来カテゴリカルな変量である Z は，数量化 I 類ではなく II 類法で解析されるべきである．この反省のもとに，今回は同じデータに対して数量化 II 類を適用してみることになった．

2 判別分析

数量化法II類は，判別分析法の拡張と考えると容易に理解できるので，先に判別分析法を簡単に説明する（詳しい説明は成書に委ねる）．分散，共分散，相関係数などの基本概念は前稿で述べた．今回のキーワードは分散比である．説明は前稿より少し難しくなり，紙数を抑えるために，止むを得ず記号を多用し，表現も省略を含む．

2.1 1変量による判別分析

数量的な変量 X について，例えば次のように，3つの群に分類されたデータが与えられる場合を考える．

群	標本値	標本数
G_1	3.3, 2.5, 3.9	3
G_2	4.4, 5.2, 4.7, 3.8	4
G_3	6.2, 5.7, 4.1	3

一般に，**標本数**を n とし，**群の個数**を g とする．群 G_ℓ の標本数を n_ℓ $(\ell = 1,\cdots,g)$ とし，G_ℓ に属す k 番目の標本値を

$$x_k^{(\ell)} \qquad (k = 1,\cdots,n_\ell;\ \ell = 1,\cdots,g) \tag{1}$$

とする（$n = \sum_{\ell=1}^{g} n_\ell$ である）．群 G_ℓ の**平均**と**分散**を

$$m^{(\ell)} = \frac{1}{n_\ell}\sum_{k=1}^{n_\ell} x_k^{(\ell)}, \quad s^{(\ell)} = \frac{1}{n_\ell}\sum_{k=1}^{n_\ell}(x_k^{(\ell)} - m^{(\ell)})^2$$

で表す．さらに，**全平均**を

$$m = \frac{1}{n}\sum_{\ell=1}^{g}\sum_{k=1}^{n_\ell} x_k^{(\ell)} = \sum_{\ell=1}^{g}\frac{n_\ell}{n}m^{(\ell)},$$

で表し，次の概念を導入する．

群内分散 $S_J = \sum_{\ell=1}^{g}\frac{n_\ell}{n}s^{(\ell)}$，（通常は S_W で表す）

群間分散 $S_B = \frac{1}{n}\sum_{\ell=1}^{g} n_\ell (m^{(\ell)} - m)^2$,

全分散 $S_T = \frac{1}{n}\sum_{\ell=1}^{g}\sum_{k=1}^{n_\ell}(x_k^{(\ell)} - m)^2$.

$x_k^{(\ell)} - m = (x_k^{(\ell)} - m^{(\ell)}) + (m^{(\ell)} - m)$ の変形から，関係

$$S_T = S_J + S_B \tag{2}$$

（群内分散と群間分散の和は全分散に等しい）が導かれる．

$$\eta^2 = \frac{S_B}{S_T} \qquad (0 \le \eta^2 \le 1) \tag{3}$$

を**分散比**，または**相関比**という．分散比の大きいほど，群内分散が相対的に小さく，変量 X の値が g 個の群 $G_1,\cdots,G_g$ の違いをよく表現していることを意味する．

値 x の個体と群 G_ℓ との距離を

$$d(x, G_\ell) = \frac{|x - m^{(\ell)}|}{\sqrt{s^{(\ell)}}} \qquad (\ell = 1,2,\cdots,g) \tag{4}$$

によって定める（これを**マハラノビス距離**という．分散の大きい群への距離は相対的に小さい）．各個体はそれから最も近い群に属すものと判別する．

上の数値例では，

$g = 3,\ n_1 = 3, n_2 = 4, n_3 = 3,\ n = 10,$
$m^{(1)} = 3.233, m^{(2)} = 4.525, m^{(3)} = 5.333,$
$m = 4.380,$
$s^{(1)} = 0.329, s^{(2)} = 0.257, s^{(3)} = 0.802,$
$S_W = 0.442, S_B = 0.676, S_T = 1.118$

であるから，$S_T = S_J + S_B$ が確認され，分散比は $\eta^2 = 0.604$ $(\eta = 0.777)$ となる．

マハラノビス距離を用いて判別すると，G_2 の 5.2 は G_3 に，G_2 の 3.8 は G_1 に，G_3 の 4.1 は G_2 に，それぞれ「誤判別」される（正答率 70 %）．

変量 X に対して，X の属す群の番号を値にとるカテゴリカルな変量を Y とする．X と Y の相関係数が最大になるように Y を**数量化**するには，群 G_ℓ の値を

$$m^{(\ell)} - m \qquad (\text{平均 } 0,\ \text{分散 } S_B)$$

とすればよい．このときの**相関係数** r_{yx} が分散比の平方根 η に等しい（そのために分散比を η^2 で表した）．

2.2 多変量による判別分析

前項の考え方は p 個の変量の場合に一般化されるが，原理を理解しやすいように，$p = 2$ の場合で説明する．

n 個の各個体について，データとして数量的な2変量 X, Y の値と，その個体の属す群の番号 Z（$1,2,\cdots,g$ のいずれかの値）が与えられるものとする（変量 Z はカテゴリカル）．2変量 X, Y の値によって Z の値（群の番号）を判別する問題を考える．

煩雑を避けて，n 個の標本データ

$$\{(x_k, y_k, z_k),\ k = 1,\cdots,n\}$$

を予め z_k の値で分類して次のように表す．

G_1 (個数 n_1)		G_2 (個数 n_2)		$\cdots$	G_g (個数 n_g)	
X	Y	X	Y		X	Y
$x_1^{(1)}$	$y_2^{(1)}$	$x_1^{(2)}$	$y_2^{(2)}$		$x_1^{(g)}$	$y_2^{(g)}$
$x_2^{(1)}$	$y_2^{(1)}$	$x_2^{(2)}$	$y_2^{(2)}$	$\cdots$	$x_2^{(g)}$	$y_2^{(g)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$x_{n_1}^{(1)}$	$y_{n_1}^{(1)}$	$x_{n_2}^{(2)}$	$y_{n_2}^{(2)}$		$x_{n_g}^{(g)}$	$y_{n_g}^{(g)}$

群 G_ℓ に関する平均を $m_x^{(\ell)}, m_y^{(\ell)}$，共分散行列を

$$S^{(\ell)} = \begin{bmatrix} s_{xx}^{(\ell)} & s_{xy}^{(\ell)} \\ s_{xy}^{(\ell)} & s_{yy}^{(\ell)} \end{bmatrix} \qquad (\ell = 1,2,\cdots,g)$$

とする．さらに，各変量の全平均を m_x, m_y，全分散を s_{xx}, s_{yy}，全共分散を s_{xy} とし，群内共分散行列，群間共分散行列，全共分散行列をそれぞれ，

$$J=\sum_{\ell=1}^{g}\frac{n_\ell}{n}S^{(\ell)},\quad \text{(通常は } W \text{ (within group) で表す)}$$

$$B=\sum_{\ell=1}^{g}\frac{n_\ell}{n}\begin{bmatrix}(m_x^{(\ell)}-m_x)^2 & (m_x^{(\ell)}-m_x)(m_y^{(\ell)}-m_y)\\ (m_y^{(\ell)}-m_y)(m_x^{(\ell)}-m_x) & (m_y^{(\ell)}-m_y)^2\end{bmatrix},$$

$$T=\begin{bmatrix}s_{xx} & s_{xy}\\ s_{yx} & s_{yy}\end{bmatrix}$$

とおけば，(2) と同様の関係 $T=J+B$ が成り立つ．

値 $P=(x,y)$ をもつ個体と群 G_ℓ とのマハラノビス距離の 2 乗は，群 G_ℓ の共分散行列の逆行列 $S^{(\ell)^{-1}}$ を用いて，次のように表される：

$$D^2(P,G_\ell)=(x-m_x^{(\ell)},y-m_y^{(\ell)})\,S^{(\ell)^{-1}}\begin{bmatrix}x-m_x^{(\ell)}\\ y-m_y^{(\ell)}\end{bmatrix}.$$

判別法は，この距離により最も近い群を選ぶことになる．

説明変量 X,Y の合成変量

$$W=aX+bY$$

を考え，その分散比の a,b による最大化問題を解く．

標本値を 2.1 節の $x_k^{(\ell)}$ のかわりに

$$w_k^{(\ell)}=a\,x_k^{(\ell)}+b\,y_k^{(\ell)}$$

とする．w の平均値は群 G_ℓ と全体とで

$$m_w^{(\ell)}=a\,m_x^{(\ell)}+b\,m_y^{(\ell)},\qquad m_w=a\,m_x+b\,m_y$$

である．G_ℓ における分散は，

$$s_{ww}^{(\ell)}=\frac{1}{n_\ell}\sum_{k=1}^{n_\ell}(w_k^{(\ell)}-m_w^{(\ell)})^2=a^2s_{xx}^{(\ell)}+2ab\,s_{xy}^{(\ell)}+b^2s_{yy}^{(\ell)}$$

のように a,b の 2 次形式となるが，これを行列 $S^{(\ell)}$ を用いて $(a,b)\,S^{(\ell)}\begin{bmatrix}a\\ b\end{bmatrix}$ と表現する．同様に，w の群間分散 S_B，全分散 S_T，分散比 η^2 を，行列 B,T を用いて

$$S_B=(a,b)\,B\begin{bmatrix}a\\ b\end{bmatrix},\quad S_T=(a,b)\,T\begin{bmatrix}a\\ b\end{bmatrix},\quad \eta^2=\frac{S_B}{S_T}$$

と表す．η^2 が最大となる a,b を求めるには，a,b による η^2 の偏微分係数を 0 とおいて得られる方程式

$$B\begin{bmatrix}a\\ b\end{bmatrix}=\eta^2T\begin{bmatrix}a\\ b\end{bmatrix}\qquad(5)$$

（これを**一般固有値問題**という）を解き，最大固有値 η^2 と対応する固有ベクトル (a,b) を求めればよい．固有値は代数方程式 $|B-\eta^2T|=0$ の解で，今の場合，2 個存在する．固有ベクトルは連立 1 次方程式 (5) の不定解で，条件 $S_T=1$ を課すことができる．

第 2 固有値 η'^2 に対応する固有ベクトル (a',b') による合成変量 $W'=a'X+b'Y$ を導入すると，変量 W と W' は無相関となる．W,W' を各座標とする平面上に標本値をプロットした散布図では，同じ群の個体は近づき，異なる群の個体は離れるように配置される（ただし，$p=2$ の場合は単なる座標変換にすぎない．主成分分析との類似性にも注目）．

変量 Z は固有ベクトル (a,b), (a',b') によって

$$\begin{array}{l}Z=\ell\ \text{のとき}\\ (\ell=1,\cdots,g)\end{array}\quad\begin{cases}c_\ell=a\,(m_x^{(\ell)}-m_x)+b\,(m_y^{(\ell)}-m_y),\\ c'_\ell=a'(m_x^{(\ell)}-m_x)+b'(m_y^{(\ell)}-m_y)\end{cases}$$

と数量化される（平均 0，分散 S_B）．これは W,W' と Z の相関が極大となる数量化に相当し，その相関係数の値は，それぞれ分散比の平方根 η,η' に等しい．

この結果，3 変量 X,Y,Z に関する共分散行列，単相関行列，偏相関行列が決定する．Z との単相関係数，偏相関係数は，変量 X,Y の判別への寄与度を比べる指標となる．η,η' は Z に対する X,Y の重相関係数 $r_{z,xy}$, $r_{z',xy}$ に等しい．

数値例　前項のデータを 2 変量に拡張する．

G_1 $(n_1=3)$

X	Y
3.3	14.2
2.5	12.8
3.9	15.3

G_2 $(n_2=4)$

X	Y
4.4	10.9
5.2	10.3
4.7	10.5
3.8	10.2

G_3 $(n_3=3)$

X	Y
6.2	11.5
5.7	12.1
4.1	14.4

群間共分散行列 B と全共分散行列 T は

$$B=\begin{bmatrix}0.676 & -0.620\\ -0.620 & 2.338\end{bmatrix},\quad T=\begin{bmatrix}1.118 & -0.777\\ -0.777 & 3.150\end{bmatrix}$$

で，固有値，固有ベクトル，Z の数量化は次の通り．

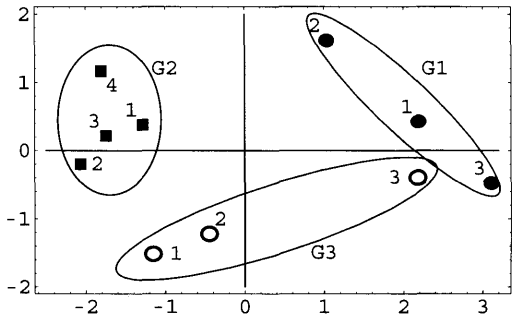
(1)

η^2 =	0.746
a =	−0.251
b =	0.968
c_1 =	1.324
c_2 =	−1.084
c_3 =	0.121

(2)

η'^2 =	0.549
a' =	−0.953
b' =	−0.303
c'_1 =	0.762
c'_2 =	0.571
c'_3 =	−1.523

参考までに点 $(w_k^{(\ell)},w'^{(\ell)}_k)$ の散布図を示す．



マハラノビス距離による判別はすべて正答となった.

$\ell\, k$	$w_k^{(\ell)}$	$w'^{(\ell)}_k$	距離 1, 2, 3	判別	正誤
1 1	2.188	0.428	2. 中 大	1	○
1 2	1.034	1.615	2. 中 大	1	○
1 3	3.102	−0.477	2. 大 中	1	○
2 1	−1.283	0.381	大 2.6 中	2	○
2 2	−2.064	−0.199	大 2.3 中	2	○
2 3	−1.745	0.217	大 0.1 中	2	○
2 4	−1.809	1.165	大 2.9 中	2	○
3 1	−1.154	−1.516	大 中 2.	3	○
3 2	−0.448	−1.221	大 中 2.	3	○
3 3	2.180	−0.395	大 中 2.	3	○

各固有値に対応する相関行列を示す（それぞれ，右上が単相関，左下が偏相関である）.

(1) X Y Z

$$\begin{bmatrix} & -0.414 & -0.458 \\ -0.048 & & 0.856 \\ -0.219 & 0.824 & \end{bmatrix}$$

重相関係数 $r_{z,xy} = 0.864$

(2) X Y Z

$$\begin{bmatrix} & -0.414 & -0.629 \\ -0.614 & & -0.097 \\ -0.738 & -0.505 & \end{bmatrix}$$

重相関係数 $r_{z,xy} = 0.741$

2.3 正準変量，p 変量への一般化

上のように被説明変量の数量化を行い，相関分析と一体化させた判別分析は，**正準判別分析**，**重判別分析**，**正準分析**などと呼ばれることがある．この意味で，2.2 節の最適化された合成変量 W, W' は**正準変量**（固有値の大きい順に，第 1，第 2，$\cdots$）と呼ばれる.

2 変量 X, Y の場合は $p\,(\geq 3)$ 個の変量 $X_1, X_2, \cdots, X_p$ の場合に自然に一般化される．このとき，一般固有値問題の固有値 η^2 の個数は $\min\{g-1, p\}$ であり，固有値の大きい固有ベクトル $a_1, a_2, \cdots, a_p$ から，判別力の強い正準変量 $\sum_{j=1}^{p} a_j X_j$ が得られる.

固有値の計算では，B の方を対角化して $\Lambda = PBP^{-1}$ とすると，方程式は $\left|(1/\eta^2)\Lambda - PTP^{-1}\right| = 0$ となる.

参考　$g = 2$ の場合，一般固有値問題は簡単に解ける．変量 X_j の平均値を $m_j^{(1)}, m_j^{(2)}$ $(j = 1, \cdots, p)$，その差を $d_j = m_j^{(1)} - m_j^{(2)}$, $\boldsymbol{d} = (d_1, \cdots, x_p)^T$ とおけば，$B = \frac{n_1 n_2}{n^2} \boldsymbol{d}\boldsymbol{d}^T$ となるから，$\boldsymbol{a} = (a_1, \cdots, a_p)^T$ とすれば

$$\eta^2 = \frac{n_1 n_2}{n^2}\, \boldsymbol{d}^T T^{-1} \boldsymbol{d}, \qquad \boldsymbol{a} = C T^{-1} \boldsymbol{d}$$

である．ここに，C は任意定数，肩付き ^{T} は転置を表す.

しかし，この特殊性は例外処理の必要性を意味しない．一般の計算法で同じ結果が得られる．$g = 2$ の特殊性は，外的基準が 2 値をとる重回帰分析との共通性があることである．これらを理解しておくとよい.

3 数量化 II 類の方法

数量化 II 類とよばれるデータ解析法は，説明変量もカテゴリカルである場合に，判別分析（正準判別分析）法を拡張したものにあたる.

3.1 ダミー変数

カテゴリカルな変量の扱い方は，前稿と同じである．すなわち，例えば，変量 X は s 個のカテゴリー $X_1, X_2, \cdots, X_s$ への分類，変量 Y は t 個のカテゴリー $Y_1, Y_2, \cdots, Y_t$ への分類であるとき，分類結果は 0 と 1 からなるベクトル

$$x = (x_1, x_2, \cdots, x_s), \qquad y = (y_1, y_2, \cdots, y_t)$$

で表す．変数 x_i, y_j の値はそれぞれ，1 個だけ 1 で，それ以外はすべて 0 であるので，**ダミー変数**と呼ばれる.

3.2 数量化の原理

群 G_ℓ に属する n_ℓ 個のデータは，ダミー変数を用いて

$$(x_{k1}^{(\ell)}, \cdots, x_{ks}^{(\ell)}), \quad (y_{k1}^{(\ell)}, \cdots, y_{kt}^{(\ell)}) \qquad (k = 1, \cdots, n_\ell)$$

と表すことにする $(\ell = 1, \cdots, g)$.

基本方針は，カテゴリー値

$$a = (a_1, \cdots, a_s), \qquad b = (b_1, \cdots, b_t)$$

を適当に定め，$X = i, Y = j$ のときそれぞれを a_i, b_j に置き換える数量化をしたときに，分散比を最大にすることである.

そのために，変量が $p = s + t$ 個のときの判別分析（2.3 節）の手順を進める．群 G_ℓ の相対頻度と全相対頻度を

$$\bar{x}_i^{(\ell)} = \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} x_{ki}^{(\ell)}, \qquad \bar{x}_i = \sum_{\ell=1}^{g} \frac{n_\ell}{n} \bar{x}_i^{(\ell)},$$

$$\bar{y}_j^{(\ell)} = \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} y_{kj}^{(\ell)}, \qquad \bar{y}_j = \sum_{\ell=1}^{g} \frac{n_\ell}{n} \bar{y}_j^{(\ell)}$$

で表し，ii', ij', ji', jj' 要素を代表的に示した行列 B, T $(s+t$ 次対称$)$ を

$$B = \begin{bmatrix} \displaystyle\sum_\ell \frac{n_\ell}{n} \bar{x}_\ell^{(\ell)} \bar{x}_{i'}^{(\ell)} - \bar{x}_\ell \bar{x}_{i'} & \displaystyle\sum_\ell \frac{n_\ell}{n} \bar{x}_\ell^{(\ell)} \bar{y}_{j'}^{(\ell)} - \bar{x}_\ell \bar{y}_{j'} \\ \displaystyle\sum_\ell \frac{n_\ell}{n} \bar{y}_j^{(\ell)} \bar{x}_{i'}^{(\ell)} - \bar{y}_j \bar{x}_{i'} & \displaystyle\sum_\ell \frac{n_\ell}{n} \bar{y}_j^{(\ell)} \bar{y}_{j'}^{(\ell)} - \bar{y}_j \bar{y}_{j'} \end{bmatrix},$$

$$T = \begin{bmatrix} \dfrac{\delta_{ii'}}{n} \bar{x}_\ell - \bar{x}_\ell \bar{x}_{i'} & \dfrac{1}{n} \displaystyle\sum_{\ell,k} x_{ki}^{(\ell)} y_{kj}^{(\ell)} - \bar{x}_\ell \bar{y}_{j'} \\ \dfrac{1}{n} \displaystyle\sum_{\ell,k} y_{kj}^{(\ell)} x_{ki'}^{(\ell)} - \bar{y}_j \bar{x}_{i'} & \dfrac{\delta_{jj'}}{n} \bar{y}_j - \bar{y}_j \bar{y}_{j'} \end{bmatrix}$$

$(\delta_{ii'}$ は，$i = i'$ のとき 1, $i \neq i'$ のとき 0 である$)$

とすれば，分散比最大化問題は，一般固有値問題

$$B\begin{bmatrix} a \\ b \end{bmatrix} = \eta^2 T\begin{bmatrix} a \\ b \end{bmatrix}$$

の最大固有値 η^2 と固有ベクトル (a,b) を求めることに帰着する．固有値 η^2 $(0 \leq \eta \leq 1)$ は通常，$g-1$ 個 $(g-1 < s+t$ だから$)$ 存在する．

ポイント　数量化II類においては，行列 B,T から各変量に対応する行と列から1行1列ずつ除去した部分行列の一般固有値問題を解けばよいことが知られている．

各固有値 η^2 に対応する固有ベクトルを求めるには，連立1次方程式 $\left(B-\eta^2 T\right)\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ を解くが，ダミー変数に関する制約条件 $\sum_{i=1}^{s} x_i = 0$, $\sum_{j=1}^{t} y_j = 0$ を考慮して，$\sum_{i=1}^{s} \bar{x}_i\, a_i = 0$, $\sum_{j=1}^{t} \bar{y}_j\, b_j = 0$ $(X,Y$ の平均を0$)$ のほかに，例えば条件 $S_T = (a,b)T\begin{bmatrix} a \\ b \end{bmatrix} = 1$ を課すことができる（依然，符号 $\pm$ の自由度が残る）．

3.3 相関分析

各固有値に対応する固有ベクトル

$$a = (a_1, \cdots, a_s), \qquad b = (b_1, \cdots, b_t)$$

をカテゴリー値として，説明変量 X,Y はそれぞれ

$$aX = \sum_{i=1}^{s} a_i X_i, \qquad bY = \sum_{j=1}^{t} b_j Y_j$$

として数量化される．被説明変量 $Z = (Z_1, \cdots, Z_g)$ の数量化 $cZ = \sum_{\ell=1}^{g} c_\ell Z_\ell$ の係数は，

$$c_\ell = \sum_{i=1}^{s} a_i\,(\bar{x}_i^{(\ell)} - \bar{x}_i) + \sum_{j=1}^{t} b_j\,(\bar{y}_j^{(\ell)} - \bar{y}_j) \qquad (6)$$

である（平均0，分散 S_B）．このとき，$aX+bY$ と cZ の相関係数は極大化されている（再び合成変量 $\alpha(aX)+\beta(bY)$ を最適化しても，すでに変量 aX, bY が最適化されているので，$\alpha=\beta$ となるにすぎない）．

この数量化により，X, Y, Z に関する単相関行列，偏相関行列が求まり，重相関係数 $r_{z,xy}$ は η と一致する．

数値例　$p=2$, $s=t=2$, $g=3$ であって，データが

G_1 $(n_1=3)$

X	Y
1	1
1	1
1	2

G_2 $(n_2=4)$

X	Y
2	1
2	1
2	2
1	2

G_3 $(n_3=3)$

X	Y
2	2
2	2
2	2

であるとする．相対頻度を X について示せば，

$\bar{x}_1^{(1)} = 1$, $\bar{x}_1^{(2)} = 1/4$, $\bar{x}_1^{(3)} = 0$, $\bar{x}_1 = 2/5$,

$\bar{x}_2^{(1)} = 0$, $\bar{x}_2^{(2)} = 3/4$, $\bar{x}_2^{(3)} = 1$, $\bar{x}_2 = 3/5$,

である．群間共分散行列 B と全共分散行列 T は

$$B = \begin{bmatrix} 0.165 & -0.165 & 0.090 & -0.090 \\ -0.165 & 0.165 & -0.090 & 0.090 \\ 0.090 & -0.090 & 0.073 & -0.073 \\ -0.090 & 0.090 & -0.073 & 0.073 \end{bmatrix},$$

$$T = \begin{bmatrix} 0.24 & -0.24 & 0.04 & -0.04 \\ -0.24 & 0.24 & -0.04 & 0.04 \\ 0.04 & -0.04 & 0.24 & -0.24 \\ -0.04 & 0.04 & -0.24 & 0.24 \end{bmatrix}$$

である．B, T から第2行第2列，第4行第4列を除去して

$$B_0 = \begin{bmatrix} 0.165 & 0.090 \\ 0.090 & 0.073 \end{bmatrix}, \qquad T_0 = \begin{bmatrix} 0.24 & 0.04 \\ 0.04 & 0.24 \end{bmatrix}$$

とし，$B_0 - \eta^2 T_0$ の B_0 を対角化すると η^2 に関する方程式

$$\begin{vmatrix} 0.440/\eta^2 - 0.551 & -0.036 \\ -0.036 & 0.036/\eta^2 - 0.409 \end{vmatrix} = 0$$

を得る．(B を対角化してから全成分が0の2行2列を除去する方法によっても，同じ結果が得られる．)

固有値，固有ベクトル，(6)式の数量化係数は次の通りである．

(1)

η^2	$= 0.804$
a_1	$= -1.037$
a_2	$= 0.691$
b_1	$= -0.501$
b_2	$= 0.334$
c_1	$= -1.260$
c_2	$= 0.176$
c_3	$= 1.026$

(2)

η'^2	$= 0.089$
a'_1	$= 0.684$
a'_2	$= -0.456$
b'_1	$= -1.136$
b'_2	$= 0.758$
c'_1	$= 0.179$
c'_2	$= -0.360$
c'_3	$= 0.302$

第1カテゴリー値 a_i, b_j と c_ℓ により，最初のデータは

X	Y
−1.037	−0.501
−1.037	−0.501
−1.037	0.334

X	Y
0.691	−0.501
0.691	−0.501
0.691	−0.791
−1.037	0.334

X	Y
0.691	0.334
0.691	0.334
0.691	0.334

と数量化されることになり，単相関行列と偏相関行列

$$\begin{bmatrix} 1 & 0.167 & 0.820 \\ 0.167 & 1 & 0.494 \\ 0.820 & 0.494 & 1 \end{bmatrix}, \quad \begin{bmatrix} -1 & -0.479 & 0.861 \\ -0.479 & -1 & 0.633 \\ 0.861 & 0.633 & -1 \end{bmatrix}$$

が求まる．結果は，次の表に集約される．

	単相関係数	偏相関係数	カテゴリー値のレンジ
X	0.820	0.861	1.728
Y	0.494	0.633	0.835

重相関係数　$r_{z,xy} = \eta = 0.897$

標本数が少ないため群毎の分散行列 $S^{(\ell)}$ は正則でない．共通に群内分散行列 S_J を用いたマハラノビス距離による判別を行うと，G_2 の $(2,2)$ は G_3 に，$(1,2)$ は G_1 に誤判別される．

4 数量化II類の適用

4.1 定式化

アンケートの概要やデータについては，前稿 [1] を参照されたい．変量 X,Y,U,V,Z は，1節で説明した通り，

すべてカテゴリカルである．Z を $g=6$ である被説明変量として扱い，数量化II類を適用する．ダミー変数の総数は $5+5+2+2=14$ である．

カテゴリー値の変数を，前稿と同様に，

$$a=(a_1,\cdots,a_5),\quad b=(b_1,\cdots,b_5),$$
$$p=(p_1,p_2),\qquad q=(q_1,q_2)$$

とする．14次元ベクトル $\boldsymbol{e}=(a,b,p,q)$ を用いると，この例における一般固有値問題は $B\boldsymbol{e}=\eta^2 T\boldsymbol{e}$ と表される．この B,T は14行14列の対称行列である．

4.2 計算プロセス

まず，データから3.2節のように行列 B,T を定める．

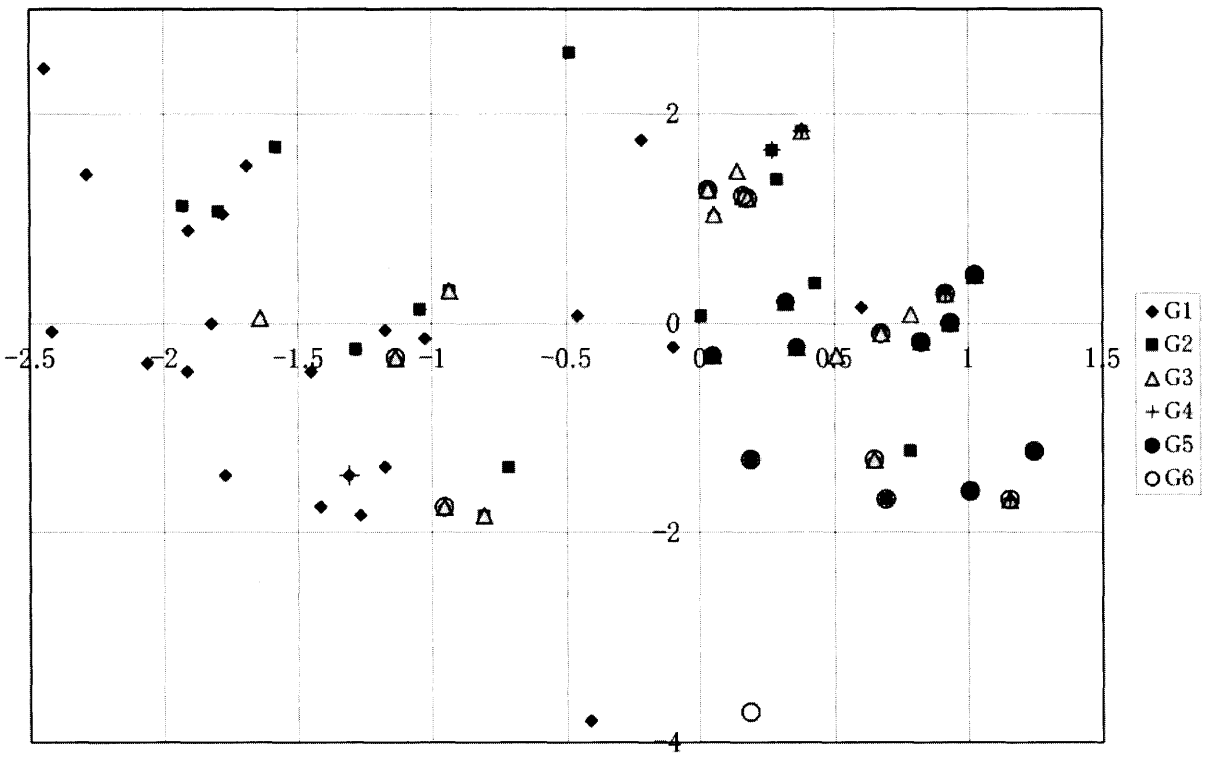
固有値 η^2 に関する方程式 $|B-\eta^2 T|=0$ を解くと，次の5（$=g-1$）個の解が得られる．

	固有値 η^2	相関係数 η
(1)	0.323	0.568
(2)	0.054	0.232
(3)	0.030	0.173
(4)	0.020	0.140
(5)	0.006	0.080

固有値をいくつまで考慮するかが問題となるが，ここでは最大固有値と第2固有値に限ることにし，それぞれに対応する相関係数を η,η'，固有ベクトルを $\boldsymbol{e},\boldsymbol{e}'$ で表す．

固有ベクトルは，連立1次方程式を解いて，右の表のように求まる．これらがカテゴリー値である．式(6)と同様に計算される Z の数量化係数 c_ℓ, c'_ℓ も記す．

変量	標本数	(1) $\boldsymbol{e}$	(2) $\boldsymbol{e}'$
X_1	2	−0.780	−3.582
X_2	51	0.059	0.593
X_3	142	−0.051	0.413
X_4	42	0.280	−1.091
X_5	33	−0.182	−1.088
Y_1	14	−0.680	0.943
Y_2	90	0.184	0.188
Y_3	85	0.092	−0.273
Y_4	52	−0.056	−0.193
Y_5	29	−0.413	0.106
U_1	198	0.172	−0.366
U_2	72	−0.473	1.005
V_1	84	−1.351	−0.104
V_2	186	0.610	0.047
変量	標本数	(1) c_ℓ	(2) c'_ℓ
Z_1	111	−0.657	−0.057
Z_2	44	0.159	0.344
Z_3	58	0.525	0.187
Z_4	19	0.663	−0.391
Z_5	8	0.667	−0.229
Z_6	30	0.585	−0.348



第1カテゴリー値 $\boldsymbol{e}$ による合成変量（正準変量）

$$w_k^{(\ell)} = \sum_{i=1}^{5} a_i x_{ki}^{(\ell)} + \sum_{j=1}^{5} b_j y_{kj}^{(\ell)} + \sum_{i=1}^{2} p_i u_{ki}^{(\ell)} + \sum_{j=1}^{2} q_j v_{kj}^{(\ell)}$$

と，同様に第2カテゴリー値 $\boldsymbol{e}'$ による $w'^{(\ell)}_k$ の2値によって定まる点

$$(w_k^{(\ell)},\, w'^{(\ell)}_k) \qquad (k = 1, \cdots, n_\ell;\ \ell = 1, \cdots, 6)$$

をすべてプロットした散布図を前のページに示した．

カテゴリー値 $\boldsymbol{e}, \boldsymbol{e}'$ に対応して，変量間の単相関行列や偏相関行列が2つずつ定まるが，省略する．

被説明変量 Z と各変量との単相関係数，偏相関係数，およびカテゴリー値のレンジを以下の表に示す．

(1)

変量	単相関	偏相関	レンジ
X(電話)	0.099	0.103	1.060
Y(メール)	0.157	0.162	0.864
U(性別)	0.156	0.193	0.645
V(学部)	0.519	0.531	1.961

重相関係数 $r_{z,xyuv} = 0.568\,(\eta)$

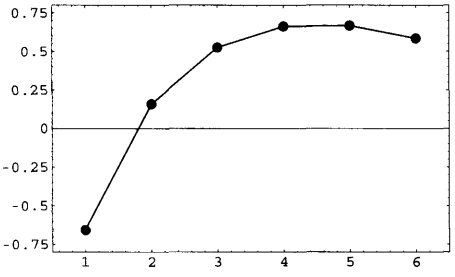
(2)

変量	単相関	偏相関	レンジ
X(電話)	0.178	0.175	4.175
Y(メール)	0.082	0.069	1.215
U(性別)	0.114	0.141	1.371
V(学部)	0.039	0.017	0.151

重相関係数 $r'_{z,xyuv} = 0.232\,(\eta')$

4.3 結果の考察

まず，前回に比べて重相関係数が少しながら増加したことに注目する（0.494 から 0.568）．最大固有値に対応する数量化係数 c_ℓ のグラフ（下図）の形状が興味深い．



群 G_1（メッセンジャーを使ったことがない）の値 c_1 が最小で，使用頻度の増加にともない大きくなるが，増加量は減少しながら G_5 でピークに達する．ここまでは穏当だが，G_6（毎日使う）の値 c_6 は c_4 より小さくなる．このことから G_4, G_5, G_6 の3つの群には，意味のある差はないということになる．

散布図については，6個の群が重なり合って，一見して分離がよいとは言えない．点の個数が270よりかなり少ないのは，同じ座標の点が多いからである．よく見ると，G_3, G_4, G_5, G_6 は右半面に多く分布し，G_1, G_2 はこの平面全体に分布するという特徴を観察することができる．

この分布の様子からは，マハラノビス距離による判別の正答率はあまり期待できないので，計算で確かめることはしなかった．逆に，分散比が0.3程度の分離はこの程度であることを学ぶべきであろう．

次に，第1カテゴリー値を見る．

変量 X では，電話を使わない層 X_1 のカテゴリー値 a_1 が最小で，前回の結果と逆になったが，その標本数はわずか2で大勢に影響を与えない（調べると，一人は G_1 に属してメールをよく使い $(Y=5)$，もう一人は G_6 に属してメールを使わない $(Y=1)$ という逆の傾向にあり，それが不安定要因になり得ることがわかった）．

X の中で最大のカテゴリー値は，前回同様 a_4 であり，層 X_4 は電話とメッセンジャーを併用してコミュニケーションをはかっていて，電話の使用頻度はコミュニケーション活動の活発さと考えられる．しかし，電話を最もよく使う層 X_5 のカテゴリー値が小さいのも前回と同じで，この層の人は電話ですべて済ませるので，メッセンジャーは使用しないと考えられる．

変量 Y については，前回とほぼ同じである．メールを使わない層 Y_1 とメールを最もよく使う層 Y_5 のカテゴリー値が特に小さく，カテゴリー値が最も大きいのは層 Y_2 である．このことから，一日に数通しかメールを使わない人は，その分メッセンジャーでコミュニケーションをはかっていると考えられる．

変量 U, V についても前回とさほど変わらなかった．

カテゴリー値のレンジでも前回同様，学部（V）が一番大きく，所属学部がメッセンジャーの使用頻度に対して影響力が強いと言える．

偏相関に注目すると，学部・性別・メール・電話の順となる．学部が0.531と抜きん出て大きな値をとるため，メッセンジャーの使用頻度を決定するとき，やはり学部が一番大きく影響していると言える．また，性別とメールが近い値をとっており，影響力は学部と比べるとだいぶ落ちるが，性別やメールの使用量によってもメッセンジャーの使用頻度に影響があると考えられる．

単相関も同じく学部の値が特に大きく，性別とメールが同程度，その後に電話と続く．

第2カテゴリー値については，重相関係数が0.232でそれほど大きくないので，説明力がないと言える．

今回の分析では，重相関係数が若干の改善を見せた．しかし，やはり強く相関があると言えるほどの値ではない．これを改善するためには，今回変量 Z を6つのカテゴリーに分けたが，メッセンジャーの使用頻度を3つほどのカテゴリーにまとめて分析を行うのが良いのではないかと考えられる．しかし，群をまとめて判別をよくす

るということは，細かな判別をあきらめるということであり，分析の目標を引き下げることにもなる．

そもそも，学部や性別がメッセンジャーの使用頻度に大きな影響を与えていることは，アンケートで得られた内容よりも，ネットワーク情報学部か経営学部かなどの外的要因によって分析結果が左右されてしまうことを意味する．これは，アンケート自体の内容が十分に練られていないことや，対象者が偏っていたことが原因と考えられる．学部ごとにわけて分析したら，どういう結果になるか，時間があれば試みてみたい．

5 学内で利用可能な計算ソフト

5.1 Mathematica と Excel

今回の計算は Mathematica で行ない，一部 Excel を利用した．Mathematica は，手軽に幅広い計算ができて，アルゴリズムを理解するには優れたツールである．統計関係の関数も用意されている（行列の対角化なども）．Excel にも便利な関数があり，BASIC でマクロを書けばかなりのことができる．

5.2 SPSS

SPSS とは，Statistical Package for the Social Sciences の略である．メニュー方式の統計解析ツールで，データさえ整えば容易に結果を得ることができる．

通常の SPSS には数量化の計算プログラムは含まれてないが，最近，1 号館マルチメディア実習室の端末 PI No.2115, 2116 の 2 台のみではあるが，それらが利用可能になった．今回それを利用して，我々の得た結果（部分）を確かめることができたので，その使い方を記す．

(1) まず，分析に使うデータを SPSS 形式（ファイル名は ‥.sav）で用意する．既存のデータファイルがあれば，SPSS 形式に変換できる．SPSS に共通だから，知っている人に教えてもらうのが早い．

最初から入力する場合は，SPSS のデータ・エディタを使う．データエディタは Excel と同じセル方式なので直感的に使いやすい．データ本体は「データビュー」部に入力し，細かい設定は「変数ビュー」部で行う．

データビュー

X	Y	U	V	Z
1	5	1	2	1
2	1	1	1	1
2	1	1	2	1
⋮	⋮	⋮	⋮	⋮
2	3	2	2	6

変数ビュー

名前	型	‥	測定
X	数値	‥	名義
Y	数値	‥	名義
U	数値	‥	名義
V	数値	‥	名義
Z	数値	‥	名義

(2) 次に，データを開いている状態で，hayasi.exe を実行する．これは，SPSS のフォルダの中にある（通常，C ドライブ→ Program Files → SPSS フォルダ→ GUI 版数量化プログラム）．すると数量化の方法を選択するウィンドウが開くので，「数量化 II 類」のボタンをクリックし，変数の指定のウィンドウに進む．ここで，説明変量の欄に X,Y,U,V の 4 つの変量を投入し（これはボタン操作），各変量のカテゴリー数の最小値と最大値（例えば，電話では 1 と 5）を入力する．非説明変量には Z を投入し，同様に最小値 1 と最大値 6 を入力する．

「インクルージョンレベル」にも説明変量の 4 つを投入し，分析に変量を取り入れる順番を指定する．今回は電話，メール，性別，学部をすべて 1 としてよい．必要に応じて「オプション」（12 項目）と「追加統計」（8 項目）を指定する．準備ができたら，「OK」ボタンをクリックして実行する．「貼り付け」をクリックしてコマンド言語によるシンタックス（‥.sps）を保存しておき，それを実行すれば反復しても変数の指定を繰り返えす必要がない（シンタックス・エディタで変更する）．

(3) 出力は「SPSS ビューア」に表示される．画面で確かめ，必要ならば印刷する．出力ファイル（‥.spo）をセーブすれば，他のソフトで編集することもできる．

5.3 SAS

SAS は，Statistical Analysis System の略である．学内で利用可能なもう一つの統計解析ツールで，メニュー方式ではないが標準的な関数が備わっている．ただし，数量化法は含まれていない．

5.4 R

最近，ベル研究所で開発された統計計算とグラフィックスのためのフリーウェアである．SPSS や SAS の使えない環境でも支障なく利用できる点が最大の魅力で，将来性がある．名前 R の由来はわからない．日本語版が

http://cran.md.tsukuba.ac.jp

からインストールできる．大学の端末はすでにインストールずみである．S 言語環境に類似のコマンド型インターフェースで，入門書を見ながら容易にプログラミングできるようになる．専門家の作成した関数がインターネットに公開されていて，それを加工すると自分用のパッケージを整えることができる．この環境でも数量化法を実行できるであろう．

附録 一般固有値問題について

n 次正方行列 A, B に関して，

$$A\boldsymbol{x} = \lambda B\boldsymbol{x}, \quad \boldsymbol{x} \neq \boldsymbol{0}$$

を満たすベクトル $\boldsymbol{x}$ とスカラー λ を求める問題を，一般固有値問題という（λ を一般固有値，$\boldsymbol{x}$ を一般固有ベクトルという）．とくに B が単位行列の場合は，普通の

固有値問題で，線形代数学の教科書でなじみがある．B が正則の（逆行列をもつ）場合は，普通の固有値問題に帰着する．

B が正則でない場合の例を示す．

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x} = \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x}.$$

明らかに，$\lambda = 1, \boldsymbol{x} = (1,0,0)^T$ は“通常の”解である．$\boldsymbol{x} = (0,0,1)^T$ は，任意の値を許す“特異な”固有値 λ に対応する“一般”固有ベクトルである．$\lambda = 1/\varepsilon, \boldsymbol{x} = (0,1,0)^T$ も解であるが，数値計算上 $|\varepsilon|$ が 0 に近いとき，通常の解と特異な解の区別が困難になる．

例えば筆者らの計算プログラムでは，対角化で現れる対角要素は，絶対値が 10^{-8} 以下ならば 0 とみなした．本文で論じた一般固有値問題では，予め A, B の階数がわかるので特異な解を除外できる（3.2 節の末尾）．統計計算で扱う行列 A, B は実対称，正定値（または非負定値）であるから固有値は実数になるが，固有ベクトルの直交性は成り立たない．

$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \boldsymbol{x} = \lambda \begin{bmatrix} 3 & 2 \\ 6 & 4 \end{bmatrix} \boldsymbol{x}$．この例では，特異な解のみ現れ，任意の λ に対して $\boldsymbol{x} = \begin{bmatrix} -2\lambda + 2 \\ 3\lambda - 1 \end{bmatrix}$ がその固有ベクトルとなる（普通の固有値問題とは著しく相違する）．

一般固有値問題の解法としては，1973 年に発表された Moler と Stewart による **QZ 法**（正則行列 Q, Z により A, B をともに上三角行列 QAZ, QBZ に変換し，固有値を対応する対角要素の比として求める方法）があるが，筆者らは Q, Z の構成法をまだ把握していない．

（この付録では，Virginia 大学コンピュータ科学科が公開している講義資料を引用した．）

あとがき

前回行った数量化I類の結果ではどうも心残りになる部分があった．しかし「カテゴリカルな変数を数量化する」という考えはわかっても具体的な分析方法を理解するのに大変時間がかかった．また研究を続けるうちに，Mathematica で一からプログラムを組むのではなく，SPSS や R など簡単に統計できるソフトの存在を次々と知り，より手軽にやりやすくなるのではないかと感じた．大学の授業だけでは身に付けることのできなかった手法をより深く学べて，大変勉強になった．(永添)

今回は，前回試みなかった数量化II類を使って分析を行った．数量化だけでなく様々な分析の知識が必要であり，I類のときよりも理解するまでに時間がかかり苦労した．しかし，この分析方法を知ったことで，より幅広い分析ができるようになったと感じた．今回，学生が手軽に使用できるよう，SPSS で数量化を行う方法を紹介した．できるだけ詳しく記述したので，機会があればぜひ試みて欲しいと思う．(村上)

今回も一から Mathematica でプログラムを作り，それを実行させながら，数量化II類の分かりやすい解説を手がけた．前稿と同様に考え方と計算法を述べ，確率モデルにもとづく推定・検定については全く触れてない．世に周知の技法だから参考書を要約すればよかろうと予断していたが，細かい点になると意外にもきちんと書かれていない．時間をかけて自分なりに疑問点を解決したが，まだ書きそろっていない部分もある．

多変量解析の世界は「線形数学」の活躍舞台であり，その素晴らしい活躍には改めて感嘆するとともに，この感慨を学生諸君と共有したい気持ちが深まる．本学部では，アンケート調査の分析などデータ解析の必要性が十分に認識されていて，それを支援するハード・ソフトの環境が整っている．しかしながら，実際にそれを使いこなすには先達による丁寧な手ほどきが不可欠である．学生諸君が一人でも多く，データ解析の知識と技術を身につけ，社会に貢献してほしいと願っているのは筆者だけではない．

本稿の執筆にあたり，統計数理研究所の石黒真木夫氏，本学部の田中，江原，伊東，石鎚，斉藤諸先生方から貴重な助言と情報をいただいた．また，情報科学センターの支援を受けた．ここに感謝の気持ちを記す．(佐藤)

参考文献

[1] 永添めぐみ, 村上明日香, 佐藤 創,「数量化理論I類の解説とその適用例」, ネットワーク & インフォメーション No. 11, 専修大学ネットワーク情報学部, 2006.
[2] 竹内 啓, 柳井晴夫,「多変量解析の基礎」, 東洋経済新報社, 1972.
[3] 奥野忠一ほか,「多変量解析法」, 日科技連, 1985.
[4] 田中 豊, 垂水共之, 脇本和昌,「パソコン統計解析ハンドブック II, 多変量解析編」, 共立出版, 1986.
[5] 小林道正, 小林厚子,「Mathematica による多変量解析」, 現代数学社, 1996.
[6] 柳井晴夫, 緒方裕光,「SPSS による統計データ解析」, 現代数学社, 2006.
[7] 田栗正章, 藤越康祝, 柳井晴夫, C.R. ラオ,「やさしい統計入門」, ブルーバックス, 講談社, 2007.
[8] 山田剛史, 杉澤武俊, 村井潤一郎,「R によるやさしい統計学」, オーム社, 2008.