社会システム分析のための統合化プログラム７

― 多変量解析 ―

福井正康・細川光浩

福山平成大学経営学部経営情報学科

概要

我々は教育での利用を主な目的に、社会システム分析に用いられる様々な手法を統一的に扱うプログラムを作成してきたが、今回は多変量解析のうち、重回帰分析、判別分析、主成分分析、数量化Ｉ類、数量化Ⅱ類、数量化Ⅲ類をシステムに組み込んだ。この論文では各分析について統計量の定義を示し、プログラムの操作法を説明している。

キーワード

社会システム分析，OR，統計，多変量解析，重回帰分析，判別分析，主成分分析，数量化理論，ソフトウェア，統合化プログラム

URL: http://www.heisei-u.ac.jp/~fukui/

１章　はじめに

我々はこれまで主に教育を目的に、様々な分析手法をプログラム化してきたが[1-5]、多変量解析は統計分析の基礎であり、社会システム分析に関する統合ソフトウェアを作成する際に避けて通ることはできない手法である。今回我々は、重回帰分析、判別分析、主成分分析及び、数量化Ⅰ，Ⅱ，Ⅲ類に関するプログラムをシステムに組み込んだ。これらの分析を選んだ理由は、量的データとカテゴリデータとの対比という意味で、重回帰分析と数量化Ⅰ類、判別分析と数量化Ⅱ類、主成分分析と数量化Ⅲ類について類似性が見られるからである。これらはいずれもよく知られており、テキスト類も豊富であることから、ここではプログラムで利用した統計量についての定義を中心に解説し、その後で分析画面と出力画面を例示することにする。

我々は初心者への対応を重視し、分析画面や出力画面について、要点は押さえつつ、できるだけ簡素化することを心掛けた。一度のクリックで結果が表示されることはこれまでの分析と同じである。しかし、多くの統計分析用のソフトウェアが世に出ている現在、利用者からの批判も多いと思われる。どの程度分析を取り入れて、分かり易さを残すか、難しい問題である。平成１３年度後期から一部の分析を卒業論文や大学院の講義で利用しており、一応の評価は得られている。今後平成１４年度からセミナーや学部の講義で本格的に活用する予定である。

２章　重回帰分析

重回帰分析は、目的変数を複数の説明変数の線形回帰式で予測する手法である。データは以下の表 2.1 の形式で与えられる。

表 2.1 重回帰分析のデータ

目的変数	説明変数 1	…	説明変数 p
y_1	x_{11}	…	x_{p1}
y_2	x_{12}	…	x_{p2}
:	:		:
y_n	x_{1n}	…	x_{pn}

実測値は以下のような 1 次式と正規分布する誤差 ε_λ で与えられるものと考える。

$$y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0 + \varepsilon_\lambda , \qquad \varepsilon_\lambda \sim N(0,\sigma^2)$$ 分布［異なる λ について独立］

線形回帰式は偏回帰係数 b_i，b_0 を用いて、以下の形で与えられる。

$$Y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0$$

これらの偏回帰係数は実測値と予測値のずれの 2 乗和 EV が最小になるように決定される。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \text{最小化}$$

即ち、b_i と b_0 についての EV の微係数を 0 とおいて以下の式を得る。

$$b_i = (\mathbf{S}^{-1}\mathbf{S}_y)_i , \quad b_0 = \bar{y} - \sum_{i=1}^{p} b_i \bar{x}_i$$

ここに、$\mathbf{S}^{-1}$ は説明変数の共分散行列 $\mathbf{S}$ の逆行列、$\mathbf{S}_y$ は目的変数と説明変数の共分散ベクトルである。

$$(\mathbf{S})_{ij} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j),\ (\mathbf{S}_y)_i = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})(x_{i\lambda} - \bar{x}_i)$$

偏回帰係数は変数の平均や分散によって影響を受け、係数の重要性が分かりにくいが、データを以下のように標準化して重回帰分析を行なうと変数の影響力の強さがはっきりと示される。ここに s_y^2, s_i^2 は目的変数及び説明変数 i の不偏分散である。

$$\tilde{y}_\lambda = \frac{y_\lambda - \bar{y}}{s_y},\ \tilde{x}_{i\lambda} = \frac{x_{i\lambda} - \bar{x}_i}{s_i}$$

これらの新しいデータ $\tilde{y}_\lambda$ と $\tilde{x}_{i\lambda}$ で作った重回帰式の偏回帰係数 $\tilde{b}_i$ を標準化偏回帰係数と言い、回帰式は以下のように表わされる。

$$\tilde{Y}_\lambda = \sum_{i=1}^{p}\tilde{b}_i\tilde{x}_{i\lambda}$$

標準化偏回帰係数と偏回帰係数との関係は $\tilde{b}_i = b_i\, s_i / s_y$ で与えられる。

重相関係数 R は実測値と予測値の相関係数であり、以下のように与えられる。

$$R = s_{yY} / (s_y s_Y)$$

ここに、s_{yY} は実測値 y と予測値 Y の共分散、s_y^2 と s_Y^2 は実測値と予測値の不偏分散である。

$$s_{yY} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})(Y_\lambda - \bar{Y}),\ s_y^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2,\ s_Y^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(Y_\lambda - \bar{Y})^2$$

実測値の全変動 SV は回帰変動 RV と残差変動 EV の和として表わされる。

$$SV = \sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n}(y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n}(Y_\lambda - \bar{Y})^2 = EV + RV$$

全変動に占める回帰変動の割合は、予測値が実測値を説明する割合を表わしていると考えられ、その値を寄与率という。寄与率は重相関係数の 2 乗に等しいことが示されるので、記号 R^2 で表わすことにする。

$$R^2 = RV / SV$$

寄与率や重相関係数の値は説明変数の数が増えれば大きくなることが知られており、これを緩和するために以下のような自由度調整済み重相関係数 $\bar{R}$ が考えられている。

$$\bar{R} = \sqrt{1 - \frac{EV/(n-p-1)}{SV/(n-1)}}$$

重回帰式の有効性は回帰変動と残差変動を比べて、回帰変動が十分大きいことが重要で、この検定には、以下の性質が利用される。

$$F = \frac{RV/p}{EV/(n-p-1)} \sim F_{p,n-p-1} \text{分布}$$

重回帰式全体の有効性とは別に、それぞれの偏回帰係数の有効性も検討される。これらは偏回帰係数が 0 と異なることを示して確かめられる。この検定には以下の性質が利用される。

$b_i = 0$ の検定　　$t_i = \dfrac{b_i}{\sqrt{a^{ii}EV/(n-p-1)}} \sim t_{n-p-1}$ 分布

$b_0 = 0$ の検定　　$t_0 = \dfrac{b_0}{\sqrt{\left(\dfrac{1}{n} + \sum_{i=1}^{p}\sum_{j=1}^{p} \bar{x}_i \bar{x}_j a^{ij}\right) EV \Big/ (n-p-1)}} \sim t_{n-p-1}$ 分布

ここに a^{ij} は $\mathbf{A} = (n-1)\mathbf{S}$ としたときの行列 $\mathbf{A}$ の逆行列 $\mathbf{A}^{-1}$ の i, j 成分である。

説明変数 i を除く他の説明変数で作った $x_{i\lambda}$ の予測回帰式を以下のように書く。

$$X_{i\lambda} = b_1^{(i)} x_{1\lambda} + \cdots + b_{i-1}^{(i)} x_{i-1\lambda} + b_{i+1}^{(i)} x_{i+1\lambda} + \cdots + b_p^{(i)} x_{p\lambda} + b_0^{(i)}$$

また、説明変数 i を除く他の説明変数で作った目的変数の予測回帰式を以下のように書く。

$$Y_{i\lambda} = b_1'^{(i)} x_{1\lambda} + \cdots + b_{i-1}'^{(i)} x_{i-1\lambda} + b_{i+1}'^{(i)} x_{i+1\lambda} + \cdots + b_p'^{(i)} x_{p\lambda} + b_0'^{(i)}$$

実測値からこれらの予測値を引いた値をそれぞれ $x'_{i\lambda}$，$y'_{i\lambda}$ として、

$$x'_{i\lambda} = x_{i\lambda} - X_{i\lambda}, \quad y'_{i\lambda} = y_\lambda - Y_{i\lambda},$$

この $x'_{i\lambda}$ と $y'_{i\lambda}$ の相関係数を偏相関係数と呼び、$\tilde{r}_{iy}$ で表わす。偏相関係数は他の変数の影響を除いた相関係数と見ることができ、以下のように表わすこともできる。

$$\tilde{r}_{iy} = -r^{iy} \Big/ \sqrt{r^{ii} r^{yy}}$$

ここに r^{iy}，r^{ii}，r^{yy} は、目的変数と説明変数を合せた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分である。

$$\mathbf{R} = \begin{pmatrix} 1 & r_{y1} & \cdots & r_{yp} \\ r_{1y} & 1 & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{py} & r_{p1} & \cdots & 1 \end{pmatrix}, \quad \mathbf{R}^{-1} = \begin{pmatrix} r^{yy} & r^{y1} & \cdots & r^{yp} \\ r^{1y} & r^{11} & \cdots & r^{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r^{py} & r^{p1} & \cdots & r^{pp} \end{pmatrix}$$

具体的な分析画面を図 2.1 に表わす。「相関行列」ボタンでは目的変数と説明変数を含んだ相関行列 $\mathbf{R}$ が表示される。その際、相関係数を 0 と比較する検定の確率値も表示される。「重回帰分析」ボタンでは、テキスト画面とグリッド画面の２つのウィンドウが開き、分析結果と分散分析表が表示される。これらは図 2.2 と図 2.3 に示される。「予測値と残差」ボタンでは、図 2.4 のように各レコード毎の実測値、予測値、残差が示される。また、「実測／予測値の散布図」ボタンでは、図 2.5 のように実測値と予測値の散布図が描かれる。

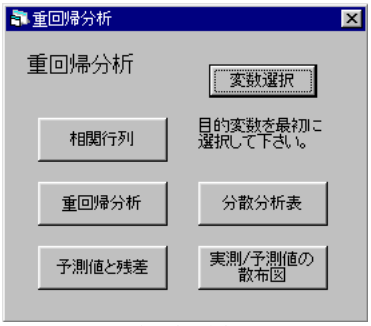


図 2.1 重回帰分析画面

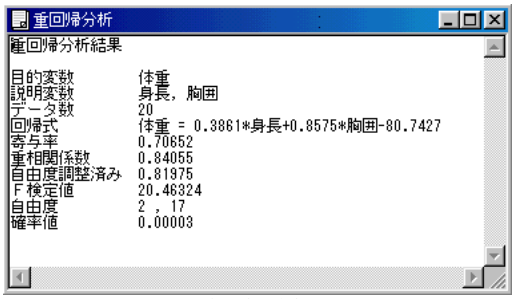


図 2.2 重回帰分析出力画面

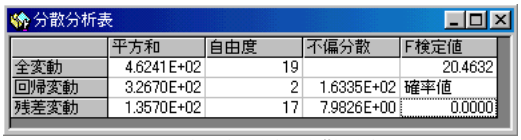


図 2.3 重回帰分析分散分析表

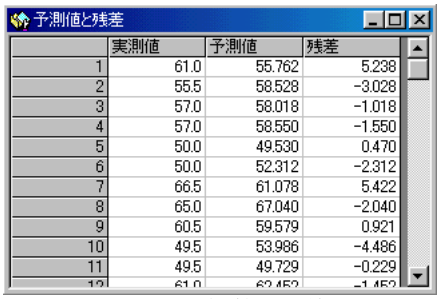


図 2.4 予測値と残差

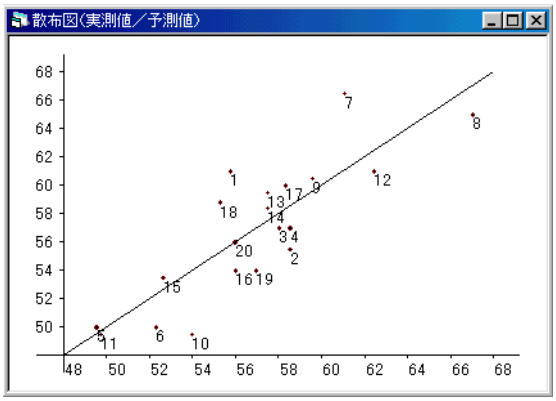


図 2.5 実測値と予測値の散布図

3章　判別分析

判別分析は外的基準によって群別に分類されたデータから、群を判別するための線形（場合によっては 2 次）関数を見出すことを目的としている。データは例えば 2 群の場合、表 3.1 のような形式で与えられる。

表 3.1 判別分析のデータ（2 群の場合）

群 1			群 2		
変数 1	…	変数 p	変数 1	…	変数 p
x_{11}^1	…	x_{p1}^1	x_{11}^2	…	x_{p1}^2
x_{12}^1	…	x_{p2}^1	x_{12}^2	…	x_{p2}^2
⋮		⋮	⋮		⋮
$x_{1n_1}^1$	…	$x_{pn_1}^1$	$x_{1n_2}^2$	…	$x_{pn_2}^2$

変数の一般的な表式 $x_{i\lambda}^{\alpha}$ において、α は外的基準（群）、i は変数、λ はレコード番号を表わす。ここでは、最初に 2 群の場合の理論について考える。

2 つの群 G_1 と G_2 について、群 $G_1 \cup G_2$ から、G_α（$\alpha = 1,2$）の要素を取り出す確率を P_α とし、G_α の要素を G_β（$\alpha \neq \beta$）と誤判別する損失を $C_{\beta\alpha}$ とする。また、群 α の確率密度関

数を $f_\alpha(\mathbf{x})$ とすると、G_α の要素を G_β と誤判別する確率 $Q_{\beta\alpha}$ は以下となる。

$$Q_{\beta\alpha} = \int_{R_\beta} f_\alpha(\mathbf{x})d\mathbf{x}$$

ここに領域 R_β は、R_β 内の要素を G_β の要素と判別する領域である。これから、誤判別による損失 L は以下のように与えられる。

$$\begin{aligned} L &= C_{21}P_1Q_{21} + C_{12}P_2Q_{12} \\ &= C_{21}P_1\int_{R_2} f_1(\mathbf{x})d\mathbf{x} + C_{12}P_2\int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\ &= C_{21}P_1\int_{R_1\cup R_2} f_1(\mathbf{x})d\mathbf{x} + \int_{R_1}[C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x})]d\mathbf{x} \end{aligned}$$

これより、損失を最小にするためには R_1 として第 2 項の被積分関数が負になる領域を選べばよい。即ち各群の領域として、以下のような領域を考えれば良いことが分かる。

$$R_1 = \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) \le 0\},$$
$$R_2 = \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) > 0\}$$

これを $h = C_{12}P_2 / C_{21}P_1$ として書き換えて、以下のような条件を得る。

$$R_1 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) \ge \log h\},$$
$$R_2 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) < \log h\}$$

ここに、$\log h$ を判別の分点という。

今、群 α の変数 i の平均 m_i^α と各群共通な共分散 s_{ij} をそれぞれ以下のように求め、

$$m_i^\alpha = \frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} x_{i\lambda}, \quad s_{ij} = \frac{1}{n_1 + n_2 - 1}\sum_{\alpha=1}^{2}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^\alpha - m_i^\alpha)(x_{j\lambda}^\alpha - m_j^\alpha),$$

これらを成分とする平均ベクトル $\mathbf{m}^\alpha$ と共分散行列 $\mathbf{S}$ を用いて、以下の多変量正規分布の確率密度関数を考える。

$$f_\alpha(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \mid \mathbf{S} \mid}} \exp\left[-\frac{1}{2}{}^t(\mathbf{x} - \mathbf{m}^\alpha)\mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}^\alpha)\right]$$

これを判別関数に代入して以下の線形判別関数を得る。

$$\begin{aligned} z &= \log f_1(\mathbf{x}) / f_2(\mathbf{x}) \\ &= {}^t\mathbf{x}\mathbf{S}^{-1}(\mathbf{m}^1 - \mathbf{m}^2) - \frac{1}{2}{}^t(\mathbf{m}^1 + \mathbf{m}^2)\mathbf{S}^{-1}(\mathbf{m}^1 - \mathbf{m}^2) \end{aligned}$$

これから、$z \ge \log h$ のとき群 1 と判定し、$z < \log h$ のとき群 2 と判定する。

変数 z の確率分布は、個体 $\mathbf{x}$ が群 1 に属するか、群 2 に属するかに応じて、以下のような正規分布に従うことが知られている。

$z \sim N(D^2/2, D^2)$ $\quad$ $\mathbf{x} \in G_1$ の場合

$z \sim N(-D^2/2, D^2)$ $\quad$ $\mathbf{x} \in G_2$ の場合

ここに、D^2 はマハラノビスの平方距離と呼ばれ、以下で定義される。

$$D^2 = {}^t(\mathbf{m}^1 - \mathbf{m}^2)\mathbf{S}^{-1}(\mathbf{m}^1 - \mathbf{m}^2)$$

この性質から誤判別の理論確率は以下で与えられることが分かる

$$Q_{21} = \int_{-\infty}^{\log h} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z - D^2/2)^2}{2D^2} \right] dz = Z\left(\frac{\log h - D^2/2}{D} \right)$$

$$Q_{12} = \int_{\log h}^{\infty} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z + D^2/2)^2}{2D^2} \right] dz = 1 - Z\left(\frac{\log h + D^2/2}{D} \right)$$

これは判別分析の有効性を示している。

判別分析では、判別関数の係数についてもその有効性を検定できる。変数 i の係数が 0 であるかどうかの検定は、以下の性質を利用する。

$$F_i = \frac{(n_1 + n_2 - p - 1) n_1 n_2 (D^2 - D_i^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_i^2} \sim F_{1, n_1 + n_2 - p - 1} \text{ 分布}$$

ここに、D_i^2 は両群の変数 i を除いたマハラノビスの平方距離である。

以上のように線形判別関数で表わされる判別分析が実行可能な条件は、分布が多変量正規分布に従うことに加えて 2 群の共分散が等しいことである。この検定には以下の性質が利用される。

$$\chi^2 = \left[1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2} \right) \frac{2p^2 + 3p - 1}{6(p+1)} \right] \log \frac{|\mathbf{S}|^{n_1 + n_2 - 2}}{|\mathbf{S}^1|^{n_1 - 1} |\mathbf{S}^2|^{n_2 - 1}} \sim \chi^2_{p(p+1)/2} \text{ 分布}$$

ここに、$\mathbf{S}^\alpha$ は群 α の共分散行列である。

３群以上の判別には以下の判別関数を考え、z_α が最大になる群 α に属するものと判定する。

$$z_\alpha = {}^t\mathbf{x}\mathbf{S}^{-1}\mathbf{m}^\alpha - \frac{1}{2} {}^t\mathbf{m}^\alpha \mathbf{S}^{-1}\mathbf{m}^\alpha + \log C_\alpha P_\alpha$$

但し、C_α は群 α を他の群と間違えた場合の損失である。上で与えた 2 群の場合の判別関数はこの判別関数を用いて、$z = z_1 - z_2$ として求めることができる。

具体的な判別分析画面を図 3.1 に示す。データの形式は、先頭列で群分けする場合と最初から群分けされている場合が扱える。但し後者の場合、予め群の数を入力しておかなければならない。各群の生起確率や誤判別損失の値は、オプションボタンの「指定する」を選び、テキストボックス内に値をカンマ区切りで入力することによって、自由に設定することができる。但し、確率の値は合計が 1 になることが必要であるので、無限小数の場合は 1/3 のように、分数で入力する。また 2 群の判別の場合、「等共分散の検定」で等共分散性を調べることができる。

図 3.2 に「等共分散の検定」の出力結果を示す。図 3.3 と図 3.4 に 2 群の判別分析と判別得点の出力結果を示す。判定は判別得点を判別の分点と比較して決定される。比較のために同じデータを用いて 3 群以上の判別のプログラムを実行した出力結果が図 3.5 と図 3.6 である。本来は 3 群以上で利用すべきであるが、2 群の判別で用いても問題はない。

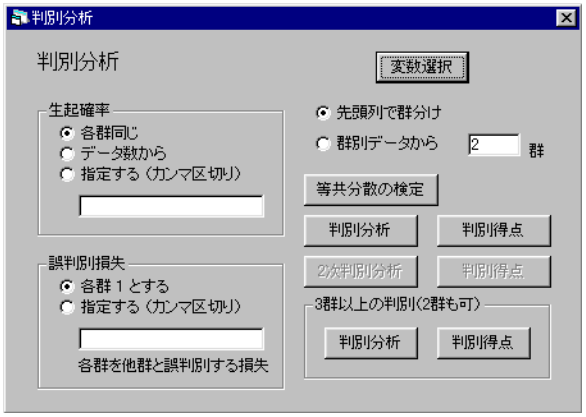


図 3.1 判別分析画面

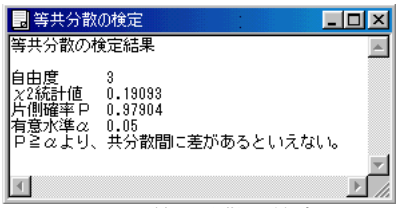
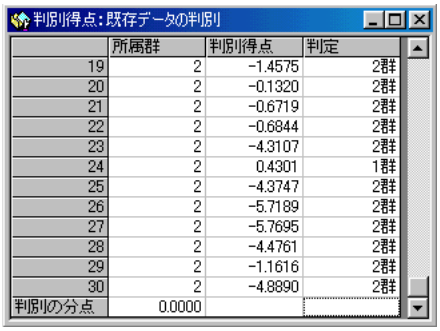


図 3.2 等共分散の検定

	勉強時間	平均点	定数項	判別の分点
判別関数	2.2461	0.2007	-23.0187	0.0000
F検定値	19.8822	15.0274		
自由度	1,27	1,27		
確率	0.00013	0.00061		
マハラノビスの距離	5.6823			
誤判別確率	1群を2群と	2群を1群と		
実測から	0.07692	0.05882		
理論から	0.11665	0.11665		

図 3.3 判別分析実行画面（2 群形式）



	所属群	判別得点	判定
19	2	-1.4575	2群
20	2	-0.1320	2群
21	2	-0.6719	2群
22	2	-0.6844	2群
23	2	-4.3107	2群
24	2	0.4301	1群
25	2	-4.3747	2群
26	2	-5.7189	2群
27	2	-5.7695	2群
28	2	-4.4761	2群
29	2	-1.1616	2群
30	2	-4.8890	2群
判別の分点	0.0000		

図 3.4 判別得点（2 群形式）

	勉強時間	平均点	定数項	判別の分点
1群判別関数	8.7369	1.0833	-61.8513	0.6931
2群判別関数	6.4908	0.8826	-38.8327	0.6931
マハラノビスの距離	1群	2群		
1群	0.0000	5.6823		
2群	5.6823	0.0000		
誤判別確率	1群を他群と	2群を他群と		
実測から	0.07692	0.05882		

図 3.5 判別分析実行画面（3 群以上形式）

	所属群	1群	2群	判定
1	1	62.4309	58.7798	1群
2	1	69.3853	64.2573	1群
3	1	52.0338	51.2501	1群
4	1	73.9872	68.5083	1群
5	1	81.8883	75.0082	1群
6	1	52.6205	52.2925	1群
7	1	42.9481	43.7224	2群
8	1	71.9078	67.0024	1群
9	1	57.6108	56.2954	1群
10	1	53.9600	52.0666	1群
11	1	49.5184	48.4480	1群
12	1	65.2335	61.1885	1群

図 3.6 判別得点（3 群以上形式）

4 章　主成分分析

主成分分析は、変数の 1 次結合により、新しい意味付けのできる特徴的な変数を作り出すことを目的としている。この新しい変数を主成分と呼ぶ。主成分分析のデータ形式は表 4.1 で与えられる。

表 4.1 主成分分析のデータ

変数 1	変数 2	…	変数 p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
:	:	…	:
x_{1n}	x_{2n}	…	x_{pn}

我々は新しい変数として以下の 1 次式を考える。

$$y_\lambda = \sum_{i=1}^{p} u_i x_{i\lambda}$$

特徴的な変数とは、データの変化に最も敏感であることと考え、係数 u_i は変数 y の不偏分散 s^2 が

最大になるように求める。但し、スケールの自由度を無くすため係数に ${}^t\mathbf{uu}=1$ の制約を付ける。ここに $\mathbf{u}$ は成分が u_i の縦ベクトルである。

不偏分散 s^2 は係数ベクトル $\mathbf{u}$ と共分散行列 $\mathbf{S}$ を用いて以下のように与えられる。

$$s^2=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})^2={}^t\mathbf{uSu}\,,\qquad (\mathbf{S})_{ij}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda}-\bar{x}_i)(x_{j\lambda}-\bar{x}_j)$$

この制約付き最大化問題は、Lagrange の未定定数法を用いて以下の量 L の極値問題となり、解は行列 $\mathbf{S}$ の固有方程式で与えられる。

$$L={}^t\mathbf{uSu}-\lambda({}^t\mathbf{uu}-1)\qquad\rightarrow\qquad\mathbf{Su}=\lambda\mathbf{u}$$

この最大固有値に対する固有ベクトル $\mathbf{u}$ を用いて作られた変数 y を第 1 主成分といい、順次固有値の大きい方から第 2 主成分、第 3 主成分と呼ぶ。一般に p 変数の場合、第 p 主成分まで選ぶことができる。

係数 u_i は変数の平均や分散から影響を受けるので、変数を標準化して分析を実行する場合も多い。この場合固有方程式は相関行列 $\mathbf{R}$ を用いて上と同様に与えられる。

$$\mathbf{Ru}=\lambda\mathbf{u}$$

正規化された固有ベクトルを求めることは、線形変換における座標回転の角度を決めることを意味する。即ち、主成分分析は、座標回転によって最も分散の大きな主軸を選び、さらにその主軸に直交し、分散が最大になるような軸を次々と定めてゆく方法である。

これらの固有方程式の第 a 固有値 λ_a に対する固有ベクトル $\mathbf{u}^a$ の成分を以下のように表わす。

$${}^t\mathbf{u}^a=(u_1^a\quad u_2^a\quad\cdots\quad u_p^a)$$

固有値 λ_a は第 a 主成分の分散を表わすことが知られている。このことから、全分散 s^2 に対する第 a 主成分の分散の割合 c_a は以下で与えられ、寄与率と呼ばれる。

$$c_a=\lambda_a\Big/\sum_{i=1}^{p}\lambda_i$$

因子負荷量 r_{ai} は第 a 主成分と変数 i の相関係数として与えられるが、これは共分散行列と相関行列を元にした場合に分けて、それぞれ以下のような形に表わされる。

$$r_{ai}=\frac{\sqrt{\lambda_a}\,u_i^a}{s_i}\quad\text{（共分散行列から），}\qquad r_{ai}=\sqrt{\lambda_a}\,u_i^a\quad\text{（相関行列から）}$$

ここで s_i^2 は変数 i の不偏分散である。

主成分得点 y_λ^a は個体毎の第 a 主成分の値として以下のように定義される。

$$y_\lambda^a=\sum_{i=1}^{p}u_i^a x_{i\lambda}$$

主成分分析において主成分を区別するためには、その固有値の大きさに差がなければならない。そこで固有値を $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r$ とした場合、大きいほうから r 個だけ値が異なり、残りは

$\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p$ となるかどうかの Anderson による sphericity の検定を行なう。この検定には以下の性質が利用される。

$$\chi^2 = -n\sum_{a=r+1}^{p}\log\lambda_a + n(p-r)\log\left(\sum_{a=r+1}^{p}\lambda_a \Big/ (p-r)\right) \sim \chi^2_{(p-r-1)(p-r+2)/2} \text{ 分布}$$

実際の主成分分析のメニュー画面を図 4.1 に与える。主成分分析は、表 4.1 に与えたデータの形から実行する場合に加え、それを集計した共分散行列や相関行列から実行する場合も想定される。それ故データの形式としてこれら 3 つの場合が含まれている。等固有値の検定にはデータ数も必要になることから、集計結果からの計算ではデータ数を入力する必要もある。計算を実行するモデルには、通常のデータから計算する「共分散行列から」と標準化されたデータから計算する「相関行列から」の 2 種類がある。勿論、データ形式で相関行列を選んだ場合は共分散行列からの計算はできない。

計算結果の表示としては「共分散行列」や「相関行列」も必要と思われるので加えてある。主成分分析は「主成分分析」ボタンで実行され、出力例は、図 4.2 に示される。

等固有値の検定結果は図 4.3 に示される。ここに表示された第 i 主成分の χ^2 値は、固有値を大きさの順番に並べた場合、第 i 主成分以降の固有値がすべて等しいとみなせるかどうかの検定値であり、等固有値確率はその確率値を表わす。それゆえ等固有確率が有意水準より大きい主成分以降が利用に適さないことを示している。極端な例として、第 1 主成分の等固有値確率が有意水準より小さい場合、主成分分析自体があまり意味を持たない。

「主成分得点」の出力は各主成分毎に図 4.4 に与えられ、2 つの主成分に関する主成分得点の散布図は図 4.5 に与えられる。これによって主成分で見た場合の個体の類似度を把握することが容易となる。

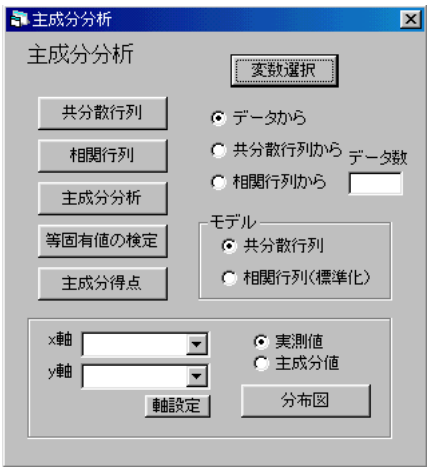
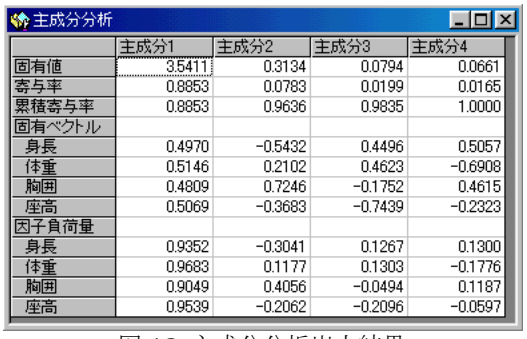


図 4.1 主成分分析のメニュー



	主成分1	主成分2	主成分3	主成分4
固有値	3.5411	0.3134	0.0794	0.0661
寄与率	0.8853	0.0783	0.0199	0.0165
累積寄与率	0.8853	0.9636	0.9835	1.0000
固有ベクトル				
身長	0.4970	−0.5432	0.4496	0.5057
体重	0.5146	0.2102	0.4623	−0.6908
胸囲	0.4809	0.7246	−0.1752	0.4615
座高	0.5069	−0.3683	−0.7439	−0.2323
因子負荷量				
身長	0.9352	−0.3041	0.1267	0.1300
体重	0.9683	0.1177	0.1303	−0.1776
胸囲	0.9049	0.4056	−0.0494	0.1187
座高	0.9539	−0.2062	−0.2096	−0.0597

図 4.2 主成分分析出力結果

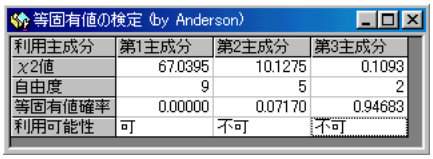


図 4.3 等固有値の検定結果

図 4.4 主成分得点出力結果

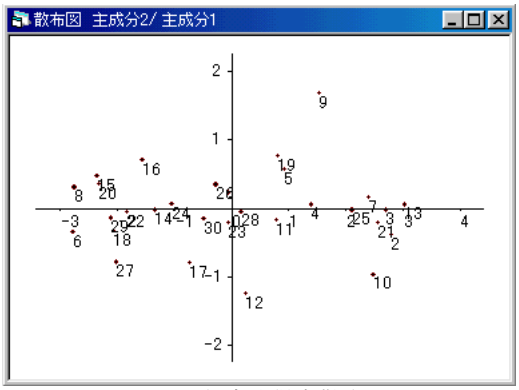


図 4.5 主成分得点散布図

５章　数量化理論

数量化理論はカテゴリデータを用いる分析で、各カテゴリに数値を与えてデータを数量化し、その構造や特徴を探る手法である。今回のプログラムでは数量化Ⅰ類からⅢ類まで分析に組み込んだ。数量化Ⅰ類は、目的変数をカテゴリデータから推測する手法で、量的データの重回帰分析に相当する。数量化Ⅱ類はカテゴリデータに関する線形判別関数を定義し、個体を分類することが狙いであり、判別分析に相当する。数量化Ⅲ類は 0/1 データによる主成分分析に類似の分析法である。

5.1 数量化Ⅰ類

数量化Ⅰ類の変数は目的変数とアイテム毎に複数個含まれるカテゴリ変数からなる。データの基本的な形は表 5.1.1 に示される。カテゴリデータは各アイテム中の 1 つのカテゴリを選択するようになっており、選択された値が 1 で、他の値が 0 であるように定められている。これはデータの一般的な書式 $x_{ij\lambda}$ を用いて以下のように表わすこともできる。

$$x_{ij\lambda} \in \{0,1\},\quad \sum_{j=1}^{r_i} x_{ij\lambda} = 1$$

表 5.1.1 数量化Ⅰ類のデータ

目的変数	アイテム 1				アイテム p		
	カテゴリ 1	…	カテゴリ r_1	…	カテゴリ 1	…	カテゴリ r_p
y_1	x_{111}	…	x_{1r_11}	…	x_{p11}	…	x_{pr_p1}
y_2	x_{112}	…	x_{1r_12}	…	x_{p12}	…	x_{pr_p2}
:	:		:		:		:

y_n	x_{11n}	$\cdots$	x_{1r_1n}	$\cdots$	x_{p1n}	$\cdots$	x_{pr_pn}

目的変数は第 2 アイテム以降の第 1 カテゴリを除いた、以下の式で予測される。

$$Y_\lambda = \sum_{j=1}^{r_1} \hat{a}_{1j} x_{1j\lambda} + \sum_{i=2}^{p} \sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda}$$

ここに、係数 $\hat{a}_{ij}$ は以下の残差変動 EV を最小化するように求める。残差変動 EV の係数 $\hat{a}_{ij}$ についての微係数を 0 として、以下の解を得る。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \rightarrow \qquad \hat{\mathbf{a}} = ({}^t\mathbf{XX})^{-1\,t}\mathbf{Xy}$$

ここに、各行列やベクトルは以下のように定義されるが、第 2 アイテム以降の第 1 カテゴリを外しているのは、行列 ${}^t\mathbf{XX}$ の正則性を失わせないためである。

$${}^t\hat{\mathbf{a}} = (\hat{a}_{11} \quad \cdots \quad \hat{a}_{1r_1} \quad \hat{a}_{22} \quad \cdots \quad \hat{a}_{2r_2} \quad \cdots \quad \hat{a}_{p2} \quad \cdots \quad \hat{a}_{pr_p})$$

$${}^t\mathbf{y} = (y_1 \quad y_2 \quad \cdots \quad y_n)$$

$$\mathbf{X} = \begin{pmatrix} x_{111} & \cdots & x_{1r_11} & x_{221} & \cdots & x_{2r_21} & \cdots & x_{p21} & \cdots & x_{pr_p1} \\ x_{112} & \cdots & x_{1r_12} & x_{222} & \cdots & x_{2r_22} & \cdots & x_{p22} & \cdots & x_{pr_p2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{11n} & \cdots & x_{1r_1n} & x_{22n} & \cdots & x_{2r_2n} & \cdots & x_{p2n} & \vdots & x_{pr_pn} \end{pmatrix}$$

さて、係数 $\hat{a}_{ij}$ について第 1 カテゴリがないことに違和感を感じる場合は、以下のような基準化された係数 a_{ij}（$i = 1, 2, \cdots, p$，$j = 1, 2, \cdots, r_i$）を導入する。

$$a_{ij} = \tilde{a}_{ij} - \sum_{k=1}^{r_i} \tilde{a}_{ik} \bar{x}_{ik}\ ,\quad \tilde{a}_{ij} = \begin{cases} 0 & i \neq 1, j = 1 \\ \hat{a}_{ij} & else \end{cases}$$

ここに、$\bar{x}_{ik}$ はアイテム i、カテゴリ k に関するデータの平均である。パラメータ $\tilde{a}_{ij}$ をカテゴリウェイト、a_{ij} を基準化されたカテゴリウェイトという。

基準化されたカテゴリウェイト a_{ij} を用いて予測値は以下の形で与えられる。

$$Y_\lambda = \bar{y} + \sum_{i=1}^{p} \sum_{j=1}^{r_i} a_{ij} x_{ij\lambda}$$

分析の寄与率 R^2 と重相関係数 R は、以下のように全変動 SV に占める、回帰変動 RV の割合とその平方根で与えられる。

$$SV = \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n} (Y_\lambda - \bar{y})^2 = EV + RV$$

$$R^2 = RV/SV\ ,\quad R = \sqrt{RV/SV}$$

各アイテムと目的変数の共分散行列 s_{ij}, s_{iy}, s_{yy} を以下で定義する。

$$s_{ij}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(X_{i\lambda}-\overline{X}_i)(X_{j\lambda}-\overline{X}_j),\quad s_{iy}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(X_{i\lambda}-\overline{X}_i)(y_\lambda-\overline{y}),$$

$$s_{yy}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\overline{y})^2$$

ここに、アイテム i の予測値 $X_{i\lambda}$ 及びその平均 $\overline{X}_i$ は以下で与えられる。

$$X_{i\lambda}=\sum_{j=1}^{r_i}\tilde{a}_{ij}x_{ij\lambda},\quad \overline{X}_i=\frac{1}{n}\sum_{\lambda=1}^{n}X_{i\lambda}$$

上で定義した共分散行列を用いた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分 r^{ij},r^{iy},r^{yy} から、アイテム i と目的変数との偏相関係数 $\tilde{r}_{iy}$ は以下のように求められる。

$$\tilde{r}_{iy}=-r^{iy}\Big/\sqrt{r^{ii}r^{yy}}$$

実際の分析メニュー画面は図 5.1.1 に与える。入力にはアイテム毎にカテゴリ名が記されているものとアイテム内をカテゴリ数に分け 0/1 で回答を表わしたものの 2 種類のデータが利用できる。もちろん 0/1 で表わされたデータには、アイテム毎のカテゴリ数を与える必要があり、テキストボックス内にカンマ区切りで入力する。コマンドボタン「0/1 型への変換」ではカテゴリ名データからもう 1 つの入力型である 0/1 型データに変換する。出力結果を図 5.1.2 に示す。

カテゴリウェイトと基準化されたカテゴリウェイトの値はコマンドボタン「カテゴリウェイト」をクリックすることによって得られる。また、これらの値による予測値から得られる重相関係数と寄与率も与えられる。出力画面は図 5.1.3 に示す。

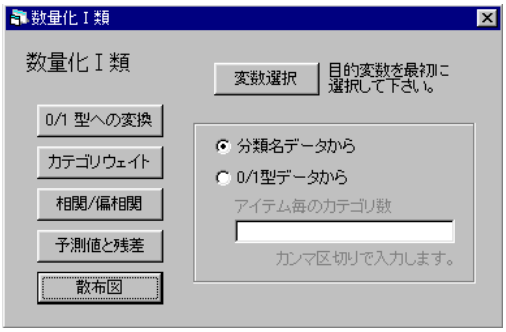
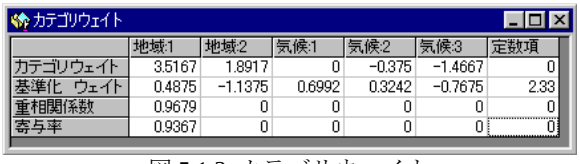


図 5.1.1 数量化Ｉ類メニュー画面


カテゴリウェイト

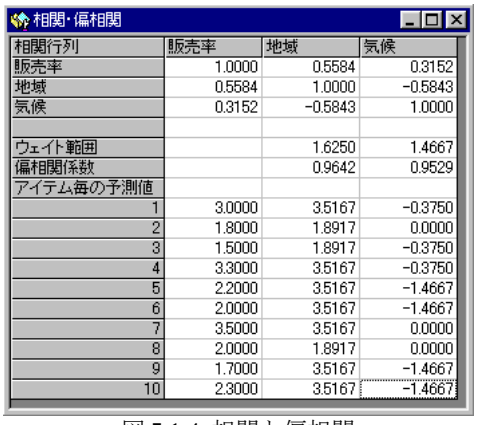
	地域:1	地域:2	気候:1	気候:2	気候:3	定数項
カテゴリウェイト	3.5167	1.8917	0	-0.375	-1.4667	0
基準化　ウェイト	0.4875	-1.1375	0.6992	0.3242	-0.7675	2.33
重相関係数	0.9679	0	0	0	0	0
寄与率	0.9367	0	0	0	0	0

図 5.1.3 カテゴリウェイト


データ基本型

	販売率	地域:1	地域:2	気候:1	気候:2	気候:3
1	3	1	0	0	1	0
2	1.8	0	1	1	0	0
3	1.5	0	1	0	1	0
4	3.3	1	0	0	1	0
5	2.2	1	0	0	0	1
6	2	1	0	0	0	1
7	3.5	1	0	1	0	0
8	2	0	1	1	0	0
9	1.7	1	0	0	0	1
10	2.3	1	0	0	0	1

図 5.1.2 0/1 型データへの変換


相関・偏相関

相関行列	販売率	地域	気候
販売率	1.0000	0.5584	0.3152
地域	0.5584	1.0000	-0.5843
気候	0.3152	-0.5843	1.0000
ウェイト範囲		1.6250	1.4667
偏相関係数		0.9642	0.9529
アイテム毎の予測値			
1	3.0000	3.5167	-0.3750
2	1.8000	1.8917	0.0000
3	1.5000	1.8917	-0.3750
4	3.3000	3.5167	-0.3750
5	2.2000	3.5167	-1.4667
6	2.0000	3.5167	-1.4667
7	3.5000	3.5167	0.0000
8	2.0000	1.8917	0.0000
9	1.7000	3.5167	-1.4667
10	2.3000	3.5167	-1.4667

図 5.1.4 相関と偏相関

目的変数とアイテム間の相関行列、目的変数とアイテム間の偏相関係数及び、個体毎のアイテムの予測値は「相関/偏相関」ボタンで得られ、図 5.1.4 にその出力結果を示す。目的変数に対する予測値と残差は「予測値と残差」ボタンで図 5.1.5 のように与えられ、その「散布図」を図 5.1.6 に示す。

予測値と残差

	観測値	予測値	残差
1	3.0000	3.1417	-0.1417
2	1.8000	1.8917	-0.0917
3	1.5000	1.5167	-0.0167
4	3.3000	3.1417	0.1583
5	2.2000	2.0500	0.1500
6	2.0000	2.0500	-0.0500
7	3.5000	3.5167	-0.0167
8	2.0000	1.8917	0.1083
9	1.7000	2.0500	-0.3500
10	2.3000	2.0500	0.2500

図 5.1.5 予測値と残差

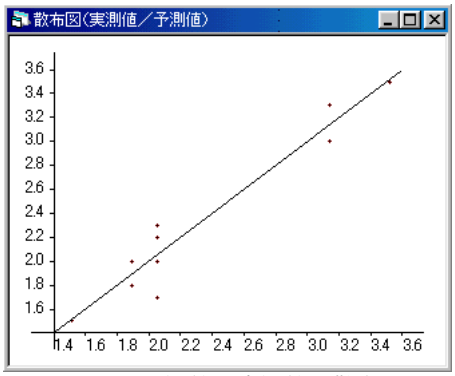


図 5.1.6 予測値と実測値の散布図

5.2 数量化Ⅱ類

カテゴリデータで群分類を行なう数量化Ⅱ類は、群の数を m、群 α のデータ数を n_α、アイテム数を p、アイテム i のカテゴリ数を r_i として、表 5.2.1 のデータ形式を元にする。

表 5.2.1 数量化Ⅱ類のデータ

	アイテム 1				アイテム p		
	カテゴリ 1	…	カテゴリ r_1	…	カテゴリ 1	…	カテゴリ r_p
群 1	x^1_{111}	…	$x^1_{1r_11}$	…	x^1_{p11}	…	$x^1_{pr_p1}$
	:		:		:		:
	$x^1_{11n_1}$	…	$x^1_{1r_1n_1}$		$x^1_{p1n_1}$	…	$x^1_{pr_pn_1}$
:	:		:		:		:
群 m	x^m_{111}	…	$x^m_{1r_11}$	…	x^m_{p11}	…	$x^m_{pr_p1}$
	:		:		:		:
	$x^m_{11n_m}$	…	$x^m_{1r_1n_m}$		$x^m_{p1n_m}$	…	$x^m_{pr_pn_m}$

一般にデータを $x^\alpha_{ij\lambda} \in \{0,1\}$ の形で表わすと、$\alpha\,(1,2,\cdots,m)$ は群、$\lambda\ \ (1,2,\cdots,n_\alpha)$ は個体、$i\,(1,2,\cdots,p)$ はアイテム、$j\,(1,2,\cdots,r_i)$ はアイテム毎のカテゴリである。各変数には次の関係がある。

$$\sum_{j=1}^{r_i} x^\alpha_{ij\lambda} = 1, \quad \sum_{i=1}^{p}\sum_{j=1}^{r_i} x^\alpha_{ij\lambda} = p$$

判別関数は係数 $\hat{a}_{ij}\ (i=1,\cdots,p,\ j=2,\cdots,r_i)$ を用いて以下のように与えられる。

$$y_\lambda^\alpha = \sum_{i=1}^{m}\sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda}^\alpha$$

この係数を求めるために、群間の分散 s_B^2 と全分散 s^2 を以下のように定義し、群間の分散の比率である分散比 $\eta^2 = s_B^2 / s^2$ を最大化することを考える。

$$s_B^2 = \frac{1}{n-1}\sum_{\alpha=1}^{m} n_\alpha (\bar{y}^\alpha - \bar{y})^2, \quad s^2 = \frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha} (y_\lambda^\alpha - \bar{y})^2$$

ここに、$\bar{y}^\alpha$ は群 α における判別関数値の平均で、$\bar{y}$ は判別関数値の全平均である。

準備として、表 5.2.1 から各アイテムの第 1 カテゴリを除いたデータについて、以下のような行列を定義しておく。

$$\mathbf{X} = \begin{pmatrix} x_{121}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p21}^1 & \cdots & x_{pr_p 1}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{12n_1}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p2n_1}^1 & \cdots & x_{pr_p n_1}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{121}^m & \cdots & x_{121}^m & \cdots & x_{p21}^m & \cdots & x_{pr_p 1}^m \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{12n_m}^m & \cdots & x_{12n_m}^m & \cdots & x_{p2n_m}^m & \cdots & x_{pr_p n_m}^m \end{pmatrix}$$

$$\overline{\mathbf{X}}_B = \begin{pmatrix} \bar{x}_{12}^1 & \cdots & \bar{x}_{1r_1}^1 & \cdots & \bar{x}_{p2}^1 & \cdots & \bar{x}_{pr_p}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^1 & \cdots & \bar{x}_{1r_1}^1 & \cdots & \bar{x}_{p2}^1 & \cdots & \bar{x}_{pr_p}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^m & \cdots & \bar{x}_{12}^m & \cdots & \bar{x}_{p2}^m & \cdots & \bar{x}_{pr_p}^m \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^m & \cdots & \bar{x}_{12}^m & \cdots & \bar{x}_{p2}^m & \cdots & \bar{x}_{pr_p}^m \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}n_1 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}n_m \end{matrix}$$

$$\overline{\mathbf{X}} = \begin{pmatrix} \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \end{pmatrix} \left.\vphantom{\begin{matrix}a\\a\end{matrix}}\right\}n$$

ここに、n はすべての群のデータ数の合計である。

分散比 η^2 の $\hat{a}_{ij}$ についての微係数を 0 とすると解くべき方程式は以下となる。

$$(\mathbf{S}_B - \eta^2 \mathbf{S})\hat{\mathbf{a}} = 0$$

ここに、$\hat{\mathbf{a}}$, $\mathbf{S}$, $\mathbf{S}_B$ は上で定義した行列 $\mathbf{X}$, $\overline{\mathbf{X}}_B$, $\overline{\mathbf{X}}$ を用いて以下で与えられる。

$${}^t\hat{\mathbf{a}} = (\hat{a}_{12} \quad \cdots \quad \hat{a}_{1r_1} \quad \cdots \quad \hat{a}_{p2} \quad \cdots \quad \hat{a}_{pr_p}),$$

$$\mathbf{S}={}^t(\mathbf{X}-\overline{\mathbf{X}})(\mathbf{X}-\overline{\mathbf{X}}),\quad \mathbf{S}_B={}^t(\overline{\mathbf{X}}_B-\overline{\mathbf{X}})(\overline{\mathbf{X}}_B-\overline{\mathbf{X}})$$

これを解くために、行列$\mathbf{S}$を下三角行列$\mathbf{F}$を用いて$\mathbf{S}=\mathbf{F}^t\mathbf{F}$のように表わし、$\mathbf{u}={}^t\mathbf{F}\hat{\mathbf{a}}$とすると以下の固有方程式を得る。

$$(\mathbf{F}^{-1}\mathbf{S}_B{}^t\mathbf{F}^{-1})\mathbf{u}=\eta^2\mathbf{u}$$

最大固有値η^2に対する固有ベクトル$\mathbf{u}$を用いて、係数ベクトル$\hat{\mathbf{a}}$は以下のように与えられる。

$$\hat{\mathbf{a}}={}^t\mathbf{F}^{-1}\mathbf{u}$$

係数ベクトル$\hat{\mathbf{a}}$は各アイテムの第1カテゴリを除いたものであるので、以下のような基準化された係数a_{ij} $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$も計算しておく。

$$a_{ij}=\tilde{a}_{ij}-\sum_{k=1}^{r_i}\tilde{a}_{ik}\bar{x}_{ik},\qquad \tilde{a}_{ij}=\begin{cases}0 & j=1\\ \hat{a}_{ij} & j\neq 1\end{cases}$$

ここに、$\bar{x}_{ik}$はアイテムi、カテゴリkにおけるデータの平均である。係数$\tilde{a}_{ij}$をカテゴリウェイト、a_{ij}を基準化されたカテゴリウェイトという。

基準化されたカテゴリウェイトを用いると、判別関数値は以下のように与えられる。

$$y_\lambda^a=\bar{y}+\sum_{i=1}^{p}\sum_{j=1}^{r_p}a_{ij}x_{ij\lambda}^{\alpha}$$

アイテムと判別関数間の相関係数を次のように与える。

$$r_{ij}=s_{ij}\Big/\sqrt{s_{ii}s_{jj}},\quad r_{iy}=s_{iy}\Big/\sqrt{s_{ii}s_{yy}}$$

ここに、アイテムと判別関数間の共分散s_{ij}，s_{iy}，s_{yy}は以下のように定義される。

$$s_{ij}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha}-\bar{x}_i)(x_{j\lambda}^{\alpha}-\bar{x}_j),\quad s_{iy}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha}-\bar{x}_i)(y^{\alpha}-\bar{y}),$$

$$s_{yy}=\frac{1}{n-1}\sum_{\alpha=1}^{m}n_\alpha(y^{\alpha}-\bar{y})^2$$

但し、$x_{i\lambda}^{\alpha}=\sum_{j=1}^{r_i}\hat{a}_{ij}x_{ij\lambda}^{\alpha}$，$\bar{x}_i=\frac{1}{n}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}x_{i\lambda}^{\alpha}$，$y^{\alpha}=\frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha}y_\lambda^{\alpha}$，$\bar{y}=\frac{1}{n}\sum_{\alpha=1}^{m}n_\alpha y^{\alpha}$ である。

アイテムiと判別関数との偏相関係数$\tilde{r}_{iy}$は、上の相関係数を用いた相関行列$\mathbf{R}$の逆行列$\mathbf{R}^{-1}$の成分r^{ij},r^{iy},r^{yy}を用いて、以下のように与えられる。

$$\tilde{r}_{iy}=-r^{iy}\Big/\sqrt{r^{ii}r^{yy}}$$

数量化Ⅱ類のメニュー画面を図 5.2.1 に示す。データは先頭列で群分けを行なう場合と既に群別になっている場合が取り扱えるが、群別データからの場合は群の数を入力する

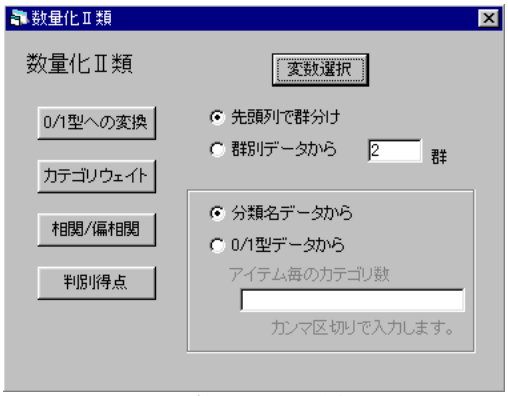


図 5.2.1 数量化Ⅱ類分析画面

必要がある。データの形式は各アイテムについてカテゴリ名を与える場合とカテゴリが既に 0/1 データとして分けられている場合があるが、0/1 データの場合各アイテム毎のカテゴリ数をカンマ区切りで入力しなければならない。

「カテゴリウェイト」コマンドボタンをクリックした結果は図 5.2.2 に与えられる。ここではカテゴリウェイトと基準化されたカテゴリウェイトが表示される。「相関/偏相関」コマンドボタンの結果は、図 5.2.3 に示される。ここでは、相関行列とそれを元に計算される偏相関係数及び各アイテム毎のカテゴリウェイトの最大と最小の差である範囲を表示する。図 5.2.4 は「判別得点」をクリックした場合の結果を表わしている。各個体が元々所属する群とその個体の数量化された値が示される。判別の助けとなるように各群の判別得点の平均も示されている。

カテゴリウェイト

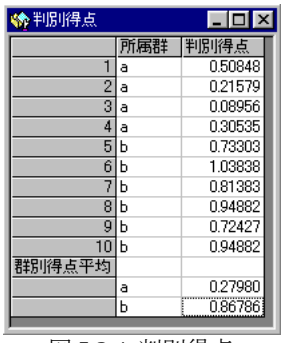
	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	切片
カテゴリウェイト	0	0.21579	0	0.08956	0	0.50848	0.73303	0
基準化ウェイト	-0.15105	0.06474	-0.03582	0.05373	-0.44576	0.06273	0.28727	0.63263

図 5.2.2 カテゴリウェイトと基準化されたカテゴリウェイト

相関・偏相関

相関行列	サンプル	価格	外観	性能
サンプル	1.0000	0.3563	-0.1667	0.8495
価格	0.3563	1.0000	0.0891	0.0682
外観	-0.1667	0.0891	1.0000	-0.3609
性能	0.8495	0.0682	-0.3609	1.0000
ウェイト範囲		0.2158	0.0896	0.7330
偏相関係数		0.5592	0.2630	0.8882

図 5.2.3 アイテムの相関係数と偏相関係数



判別得点

	所属群	判別得点
1	a	0.50848
2	a	0.21579
3	a	0.08956
4	a	0.30535
5	b	0.73303
6	b	1.03838
7	b	0.81383
8	b	0.94882
9	b	0.72427
10	b	0.94882
群別得点平均		
	a	0.27980
	b	0.86786

図 5.2.4 判別得点

5.3 数量化Ⅲ類

数量化Ⅲ類はカテゴリと個体にそれぞれ数値を与えて、特徴的な量を作りだし、データの持つ構造を解明しようとするものである。データの形式は表 5.3.1 で与えられる。

表 5.3.1 数量化Ⅲ類のデータ

カテゴリ 1	カテゴリ 2	…	カテゴリ p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
:	:		:
x_{1n}	x_{2n}	…	x_{pn}

個々のデータはカテゴリに反応した場合 1、反応しない場合は 0 で与えられる。

$$x_{i\lambda} \in \{0,1\}$$

ここに、i はカテゴリ、λ は個体を表わす。

カテゴリと個体に対して特徴的な係数 u_i と v_λ を得るために、まず以下のような u_i と v_λ の分散と共分散を考える。

$$s_u^2 = \frac{1}{T}\sum_{i=1}^{p} m_i (u_i - \bar{u})^2 , \quad s_v^2 = \frac{1}{T}\sum_{\lambda=1}^{n} n_\lambda (v_\lambda - \bar{v})^2 , \quad s_{uv} = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}(u_i - \bar{u})(v_\lambda - \bar{v})$$

ここに、m_i，n_λ，T 及び、$\bar{u}$，$\bar{v}$ は以下のように定義される。

$$m_i = \sum_{\lambda=1}^{n} x_{i\lambda}\ ,\quad n_\lambda = \sum_{i=1}^{p} x_{i\lambda}\ ,\quad T = \sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}\ ,$$

$$\bar{u} = \frac{1}{T}\sum_{i=1}^{p} m_i u_i\ ,\quad \bar{v} = \frac{1}{T}\sum_{\lambda=1}^{n} n_\lambda v_\lambda$$

これからカテゴリと個体の相関係数を $\rho = s_{uv}/(s_u s_v)$ と表わし、この相関係数を最大にする係数 u_i と v_λ を求める。但し、相関係数に関して係数 u_i，v_λ には定数を加減する任意性があることから、$\bar{u} = 0$，$\bar{v} = 0$ となるものを選ぶことにする。

これから相関係数の u_i，v_λ に関する微係数を 0 として、以下の式が導かれる。

$$\sum_{j=1}^{p}\sum_{\lambda=1}^{n} \frac{1}{n_\lambda} x_{i\lambda} x_{j\lambda} u_j = \rho^2 \sum_{\lambda=1}^{n} x_{i\lambda} u_i \qquad (1)$$

$$v_\lambda = \frac{1}{\rho}\frac{s_v}{s_u}\frac{1}{n_\lambda}\sum_{i=1}^{p} x_{i\lambda} u_i \qquad (2)$$

(2) 式とパラメータ v_λ のスケールの自由度を用いて、v_λ について以下のように決める。

$$v_\lambda = \frac{1}{\rho n_\lambda}\sum_{i=1}^{p} x_{i\lambda} u_i$$

また、以下の定義を用いると、

$$(\mathbf{z})_i = z_i = \sqrt{m_i}\,u_i\ ,\quad (\mathbf{C})_{ij} = c_{ij} = \frac{1}{\sqrt{m_i m_j}}\sum_{\lambda=1}^{n}\frac{1}{n_\lambda} x_{i\lambda} x_{j\lambda}\ ,$$

(1) 式は以下のように表わされる。

$$\mathbf{Cz} = \rho^2 \mathbf{z} \qquad (3)$$

この固有方程式は $\rho^2 = 1$，$z_i = 1$ の解を持つことが知られているが、これは $\bar{u} = 0$ を満たさないので、1 以外の固有値 ρ^2 の固有ベクトルを求める必要がある。これらの固有ベクトルは $\bar{u} = 0$，$\bar{v} = 0$ を満たすことが証明できる。ここで、固有ベクトル $\mathbf{z}$ に対して ${}^t\mathbf{z}\mathbf{z} = 1$ の規格化を行なうと、パラメータ u_i と v_λ については以下の規格化となる。

$$\sum_{i=1}^{p} m_i u_i^2 = 1,\quad \sum_{\lambda=1}^{n} n_\lambda v_\lambda^2 = 1$$

(3) 式の固有値 $\lambda_a (\neq 1)$ に対する固有ベクトルを ${}^t\mathbf{z}^a = (z_1^a \quad z_2^a \quad \cdots \quad z_p^a)$ として、係数 u_i^a と v_λ^a についてその表式をまとめておく。

$$u_i^a = z_i^a \big/ \sqrt{m_i}\ ,\quad v_\lambda^a = \frac{1}{\sqrt{\lambda_a} n_\lambda}\sum_{i=1}^{p} x_{i\lambda} u_i^a$$

ここに、u_i^a をカテゴリウェイト、v_λ^a を個体ウェイトと呼ぶ。

このカテゴリウェイトと個体ウェイトを用いてカテゴリ得点 y_i^a と個体得点 w_λ^a をそれぞれ以下のように定義する。

$$y_i^a = \sum_{\lambda=1}^{n} x_{i\lambda} v_\lambda^a, \quad w_\lambda^a = \sum_{i=1}^{p} x_{i\lambda} u_i^a$$

実際の分析画面を図 5.3.1 に示す。表示されるものは係数 u_i^a の値を示す「カテゴリウェイト」、その値を用いた「カテゴリ得点」、係数 v_λ^a の値を示す「個体ウェイト」及び「個体得点」である。また、カテゴリ得点や個体得点を視覚化して分類に役立てるために、２つの固有値に対する固有ベクトル（次元）を選んで表示する「散布図」も加えた。

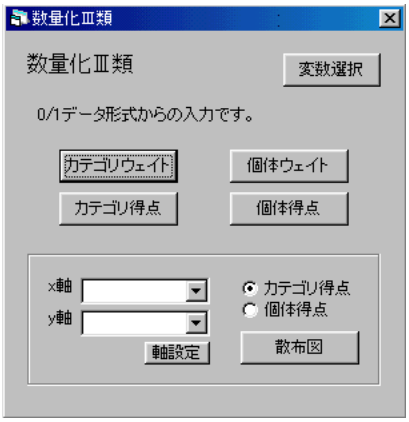


図 5.3.1 数量化Ⅲ類メニュー

図 5.3.2 にカテゴリウェイトの出力結果を図 5.3.3 にカテゴリ得点の出力結果を示す。ここでは、参考のために、通常表示しない $\rho^2 = 1$ に対する固有ベクトルから導かれる値を第 0 次元として表示している。個体ウェイトと個体得点については表示形式が同じなので省略する。また、散布図については図 5.3.4 に示す。

カテゴリウェイト

	第0次元	第1次元	第2次元	第3次元	第4次元	第5次元
固有値	1	0.35054	0.15565	0.09482	0.06045	0.02525
ご飯	0.13736	0.18786	-0.00236	0.10361	0.18384	0.03613
パン	0.13736	-0.16487	-0.00845	0.11717	0.04337	-0.22221
うどん	0.13736	0.17506	0.05945	-0.02856	-0.24309	-0.10974
そば	0.13736	-0.11344	0.2649	-0.01225	-0.00616	0.15133
ラーメン	0.13736	-0.02764	-0.0854	-0.25973	0.07363	-0.01392
スパゲッティ	0.13736	-0.07642	-0.20515	0.10414	-0.12137	0.18064

図 5.3.2 カテゴリウェイト

カテゴリ得点

	第0次元	第1次元	第2次元	第3次元	第4次元	第5次元
ご飯	1.37361	1.11223	-0.00929	0.31904	0.45201	0.05741
パン	1.23625	-0.87853	-0.03002	0.32473	0.09597	-0.31779
うどん	1.09888	0.82917	0.18765	-0.07036	-0.47816	-0.1395
そば	1.09888	-0.5373	0.83608	-0.03017	-0.01212	0.19238
ラーメン	1.37361	-0.16362	-0.33692	-0.79978	0.18104	-0.02212
スパゲッティ	1.09888	-0.36195	-0.64749	0.25654	-0.23874	0.22963

図 5.3.3 カテゴリ得点

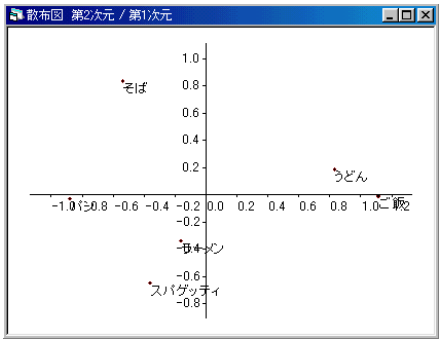


図 5.3.4 カテゴリ得点による散布図

６章　おわりに

今回は多変量解析を中心にプログラムを作成してきたが、手法が完全に固まっている点で独自性を打ち出すことはなかなか難しい。特に多くの統計ソフトウェアが作られている今、個々のプログラム自体に価値を見出すことはできないため、統合ソフトウェアとしてのバランスが重要になる。基本的に初心者を対象としているため、1 分析 1 画面とすることを条件としてプログラム

化を行った。それ故、必要と思われる機能を省くこともあったし、著者の都合で作り得なかったものもあった。ここでは分析ごとに問題点を考えてみる。

重回帰分析の大きな問題点は最良回帰式の決定手法が含まれていないことである。今後必要に応じて、変数増加法、変数減少法、変数増減法等を加えることになろう。また、残差の図的表示機能などで一般的な統計ソフトウェアに比べて機能が劣っている。

判別分析では等分散性が保証されない場合の 2 次判別関数やその判別得点も、今後の予定としてコマンドボタンだけが表示されている。また、多次元の判別については触れられていない。これらの点は今後の課題である。

主成分分析については、授業に利用する上で特に気掛かりな点はないが、散布図についてデータが、お互いに重なった場合や軸と重なった場合にはラベルが大変見にくくなる。これを是正する機能も考えなければならない。

数量化 I 類についても機能的に余り不足を感じることはない。ただ、分析実行中の逆行列の計算では正則でない場合もあり、その原因として誤差の可能性も完全には捨てきれず、多少不安が残る。この辺りの頑健性はその他の分析でも問題となるので、今後の重要な課題であろう。

数量化Ⅱ類には多次元の数量化を加えていない。判別は複数次元で行なわれることもあるので、判別分析と同様今後考える必要がある。

数量化Ⅲ類では、ある 1 例で固有値 1 が 2 つ現れている場合があり、その際に Jacobi 法を用いて求めた固有値の出現順序に問題が生じた。後にこれは行列が完全に 2 つの小行列に分離されることが原因であると分かったが、我々は用心のために、第 0 次元として固有値 1 の固有ベクトルも表示させることにした。

このプログラムは主に文献 6) から 9) を参考にしている。中でも主な理論は 6) と 7) 、プログラムの構成や理論的補足は 9) から学んだところが大きい。ここに感謝の意を表わす。

参考文献

1) 福井正康・田口賢士, 社会システム分析のための統合化プログラム, 福山平成大学経営情報研究, 3 号, 109-127, 1998.

2) 福井正康・田口賢士, 社会システム分析のための統合化プログラム２ －産業連関分析・KSIM・AHP－, 福山平成大学経営情報研究, 3 号, 129-144, 1998.

3) 福井正康・増川純一, 社会システム分析のための統合化プログラム３ －線形計画法・待ち行列シミュレーション－, 福山平成大学経営情報研究, 4 号, 99-115, 1999.

4) 福井正康, 社会システム分析のための統合化プログラム４ －基本統計－, 福山平成大学経営情報研究, 5 号, 89-100, 2000.

5) 福井正康, 社会システム分析のための統合化プログラム５ －システムの改良・ISM－, 福山平成大学経営情報研究, 6 号, 91-104, 2001.

6) 丹後俊郎, 新版医学への統計学, 朝倉書店, 1993.
7) 河口至商, 多変量解析入門Ⅰ, 森北出版, 1973.
8) 河口至商, 多変量解析入門Ⅱ, 森北出版, 1978.
9) 田中豊・垂水共之編, Windows 版 統計解析ハンドブック 多変量解析, 共立出版社, 1995.
10) 田中豊・脇本和昌, 多変量統計解析法, 現代数学社, 1983.

Multi-purpose Program for Social System Analysis 7

- Multivariate Analysis -

Masayasu FUKUI and Mitsuhiro HOSOKAWA

Department of Management Information, Faculty of Management,
Fukuyama Heisei University

Abstract

We have been constructing a unified program on social system analysis for the purpose of education. Now we added multiple regression analysis, discriminat analysis, principal component analysis and quantification method of the first, second and third type, to our system. The purpose of this paper is to explain definitions of the statistics and operation of our program.

Keywords

social system analysis, statistics, multivariate analysis, multiple regression analysis, discriminant analysis, principal component analysis, quantification method, software, unified program

URL: http://www.heisei-u.ac.jp/~fukui/