社会システム分析のための統合化プログラム２４

― 判別分析と数量化Ⅱ類の整合化 ―

福井正康・尾崎誠・朝日亮太

福山平成大学経営学部経営学科

概要

我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。今回はこれまで独立に扱われてきた判別分析と数量化Ⅱ類のプログラムについて、マハラノビスの距離を利用する方法と正準形式を利用する方法に分けて整理し、統一的にプログラムを再構成した。この論文ではその背景にある理論と実際のプログラムの動作について説明する。

キーワード

College Analysis，統計，判別分析，数量化Ⅱ類

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

１．はじめに

我々はこれまで College Analysis の多変量解析の中に判別分析と数量化Ⅱ類のプログラムを組み込んできた[1]。その際、判別分析はマハラノビスの平方距離（以後マハラノビス距離と略す）を用いた方法、数量化Ⅱ類は最大固有値に対する固有ベクトルだけを用いた正準形式の方法を取り扱った。その後、判別分析に正準形式を用いた方法（正準判別分析と呼ばれる）を加えたが、作成時期が異なったため、これらの連携については考えなかった。今回多変量解析の見直しを行うに当たり、これらのプログラムを整理し、それぞれマハラノビス距離を用いた方法（マハラノビス形式と略す）と正準形式を用いた方法（正準形式と略す）を整理し、それらの関連付けを行った[2]。もちろん判別結果については以前のものと違いはないが、判別関数の係数や定数に少しずつ変更を加え、２つの方法の類似性と相違性がより明確になるようにした。具体的な変更点は表 1.1 に与える通りである。

表 1.1 プログラムの変更点

	群数	マハラノビス形式	正準形式
判別分析	2 群	変更なし	係数と定数項の調整
	3 群以上	定数項の調整	係数と定数項の調整
数量化Ⅱ類	2 群	新規作成	係数と定数項の調整
	3 群以上	新規作成	係数と定数項の調整・多次元化

判別分析のマハラノビス形式には、分布関数の理論から、判別群の生起確率や誤判別損失などを加えていたが、正準形式ではこれらは考えない。また数量化Ⅱ類でもあまり考えることはない。しかし、これらを考えないことは、生起確率が等しく誤判別損失が等しい場合につながると考えると、マハラノビス形式で 3 群以上の判別分析の定数項に少し修正を加える必要が出てきた。また、数量化Ⅱ類の計算が第 1 カテゴリを除いた判別分析であることを示すために、判別分析と数量化Ⅱ類とでこれまで定義が異なっていた分散比について同じ定義にした。これによりこれまで比例していた係数が、完全に同じものとなった。また、数量化Ⅱ類の正準形式で、第 1 次元だけを利用してきた結果を多次元に拡張し、正準判別分析と同様の散布図を表示できるようにした。最後に今回数量化Ⅱ類について、新たにマハラノビス形式のプログラムも作成した。これらの変更と拡張により、判別分析、数量化Ⅱ類、及びマハラノビス形式、正準形式の関係が結果の上から読み取り易くなり、学習者にとって分かり易いプログラムになった。

この論文ではこれまでのものを含めて、理論を詳細に記述する。そのため、参考文献 [1] と重なる部分も多いが、理論を説明するために必要な個所はそのまま引用した。しかし、数量化Ⅱ類と判別分析の同等な部分については、判別分析に任せることにした。

分析の結果の中で、判別分析に標準化係数、数量化Ⅱ類に基準化係数があるが、これらは別物である。前者は標準化された変数を用いて同じ結果を出すための係数で、後者は各アイテムの第 1 カテゴリに 0 と異なる数値を与え、各カテゴリが判別関数に対して正負のどちらの方向に効いているのかを明らかにするための係数である。

２．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 2.1 のような形式で与えられる。

表 2.1　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 は群、i は変数、λ はレコード番号を表わす。

2.1　マハラノビス距離を用いた方法

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。

$$R_1 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h \ge 0\},$$
$$R_2 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h < 0\}$$

ここに、判別の分点は 0 である。

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

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

これらを成分とする平均ベクトル $\overline{\mathbf{x}}^{\alpha}$ 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。

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

これを判別関数に代入して以下の線形判別関数を得る。

$$z = \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h$$
$$= {}^t\mathbf{x}\mathbf{S}^{-1}(\overline{\mathbf{x}}^1 - \overline{\mathbf{x}}^2) - \frac{1}{2}\,{}^t(\overline{\mathbf{x}}^1 + \overline{\mathbf{x}}^2)\mathbf{S}^{-1}(\overline{\mathbf{x}}^1 - \overline{\mathbf{x}}^2) - \log h$$

$\mathbf{a} = \mathbf{S}^{-1}(\overline{\mathbf{x}}^1 - \overline{\mathbf{x}}^2)$ とすると、判別関数は以下のように書くことができる。

$$z = {}^t\mathbf{x}\mathbf{a} - \frac{1}{2}\,{}^t(\overline{\mathbf{x}}^1 + \overline{\mathbf{x}}^2)\mathbf{a} - \log h \qquad (2.1)$$

判別関数は、変数 x_i の標準化値 u_i と不偏分散 s_i^2 を用いて以下のように書くこともできる。

$$z = {}^t\mathbf{u}\mathbf{c} + {}^t\overline{\mathbf{x}}\mathbf{a} - \frac{1}{2}\,{}^t(\overline{\mathbf{x}}^1 + \overline{\mathbf{x}}^2)\mathbf{a} - \log h, \quad c_i = a_i s_i \qquad (2.2)$$

この係数 $\mathbf{c}$ を標準化係数と呼ぶ。標準化係数は変数の重要性をみるときに利用される。

判別関数 (2.1) は各群の平均 $\overline{\mathbf{x}}^{\alpha}$ から、$\mathbf{x}$ までのマハラノビスの平方距離 $D^{2(\alpha)}$ の差として以下のように定義することもできる。

$$z = \frac{1}{2}(D^{2(2)} - D^{2(2)}) - \log h \ , \ D^{2(\alpha)} = {}^t(\mathbf{x} - \overline{\mathbf{x}}^{\alpha})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}^{\alpha})$$

この z は $\log h = 0$ の場合、$\mathbf{x}$ が 2 つの群別平均の中央である $(\overline{\mathbf{x}}^1 + \overline{\mathbf{x}}^2)/2$ のとき、0 になっている。

変数 z の確率分布は、$\log h = 0$ の場合、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。

$$z \sim N(D^2/2, D^2) \qquad \mathbf{x} \in G_1 \text{の場合}$$
$$z \sim N(-D^2/2, D^2) \qquad \mathbf{x} \in G_2 \text{の場合}$$

ここに、D^2 は群平均 $\bar{\mathbf{x}}^1$ と $\bar{\mathbf{x}}^2$ のマハラノビスの平方距離で、以下のように定義される。

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

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。

以上のような理論では、線形判別関数で表わされる判別分析がうまく利用できる条件は、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（群の数を m）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。

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

但し、C_α は群 α を他の群と間違えた場合の損失である。定数項に含まれる m は、各群の生起確率が同じで誤判別損失が 1 の場合、これらを考えない理論と繋がるように、定数項を 0 にするための定数である。

$\mathbf{a}^\alpha = \mathbf{S}^{-1}\bar{\mathbf{x}}^\alpha$ として、この判別関数は以下のように書くこともできる。

$$z^{\alpha} = {}^{t}\mathbf{x}\mathbf{a}^{\alpha} - \frac{1}{2}{}^{t}\bar{\mathbf{x}}^{\alpha}\mathbf{a}^{\alpha} + \log C_{\alpha}P_{\alpha}m \qquad (2.3)$$

2 群の場合と同様に、判別関数は変数 x_i の標準化値 u_i と不偏分散 s_i^2 を用いて以下のように書くこともできる。

$$z^{\alpha} = {}^{t}\mathbf{u}\mathbf{c}^{\alpha} + {}^{t}\bar{\mathbf{x}}\mathbf{a}^{\alpha} - \frac{1}{2}{}^{t}\bar{\mathbf{x}}^{\alpha}\mathbf{a}^{\alpha} + \log C_{\alpha}P_{\alpha}m, \quad c_i^{\alpha} = a_i^{\alpha}s_i \qquad (2.4)$$

この係数 $\mathbf{c}^{\alpha}$ を標準化係数と呼ぶ。

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。

2.2 正準形式を用いた方法

正準形式の判別分析（正準判別分析と呼ばれる）は、判別関数の拡がりを最大化するように係数を求めるもので、特に 3 群以上の場合は、判別得点を複数次元の空間上に配置し、判別をより分かり易く表現する手法である。これまでのプログラムでは、数量化Ⅱ類でその中の主要な 1 次元を取り出して判別する方法を導入している。以下に正準判別分析の理論を示す。

正準判別分析は、判別群で分けられたデータについて、「群間分散／群内分散」を最大化するように線形判別関数の係数を決定する手法である。判別関数を以下のように表す。ここに z_0 は後に決める定数項である。

$$z = \sum_{i=1}^{p} a_i x_i + z_0$$

判別群を α，群別のデータの番号を λ，変数の番号を i，としてデータを $x_{i\lambda}^{\alpha}$（$\alpha = 1,\cdots,m$，$\lambda = 1,\cdots,n_{\alpha}$，$i = 1,\cdots,p$）と表す。このデータを用いて、群 α の λ 番目の判別関数の値 z_{λ}^{α} は以下のようになる。

$$z_{\lambda}^{\alpha} = \sum_{i=1}^{p} a_i x_{i\lambda}^{\alpha} + z_0$$

この z_{λ}^{α} による群間分散 s_B^2，群内分散 s^2 を以下のように定義する。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_{\alpha}\left(\bar{z}^{\alpha} - \bar{z}\right)^2, \quad s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}\left(z_{\lambda}^{\alpha} - \bar{z}^{\alpha}\right)^2$$

ここに、$\bar{z}^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} z_{\lambda}^{\alpha}$，$\bar{z} = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{z}^{\alpha}$，$n = \sum_{\alpha=1}^{m} n_{\alpha}$ である。

これより、$\bar{x}_i^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} x_{i\lambda}^{\alpha}$，$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{x}_i^{\alpha}$ として、s_B^2 と s^2 は以下のようになる。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha \left[\sum_{i=1}^{p} a_i\left(\overline{x}_i^{\alpha} - \overline{x}_i\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_{ij} a_j$$

$$s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}\left[\sum_{i=1}^{p} a_i\left(x_{i\lambda}^{\alpha} - \overline{x}^{\alpha}\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i s_{ij} a_j$$

ここに、

$$b_{ij} = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha\left(\overline{x}_i^{\alpha} - \overline{x}_i\right)\left(\overline{x}_j^{\alpha} - \overline{x}_j\right)$$

$$s_{ij} = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}\left(x_{i\lambda}^{\alpha} - \overline{x}_i^{\alpha}\right)\left(x_{j\lambda}^{\alpha} - \overline{x}_j^{\alpha}\right)$$

である。行列の成分として、$(\mathbf{B})_{ij} = b_{ij}$ ，$(\mathbf{S})_{ij} = s_{ij}$ ，$(\mathbf{a})_i = a_i$ とすると、s_B^2 と s^2 はこれらの行列を用いて次のように書ける。

$$s_B^2 = {}^t\mathbf{aBa} \quad , \quad s^2 = {}^t\mathbf{aSa}$$

ここに、$n \geq m$ の場合、一般に $rank(\mathbf{B}) = m-1$，$rank(\mathbf{S}) = n-m$ である。

群間分散を群内分散で割った分散比 ρ は以下のようになる。

$$\rho = s_B^2 / s^2 = {}^t\mathbf{aBa} / {}^t\mathbf{aSa}$$

この分散比を最大化するには、以下の解を求める。

$$\partial\rho / \partial\mathbf{a} = \frac{1}{\left(s^2\right)^2}\left[\partial s_B^2 / \partial\mathbf{a}\, s^2 - s_B^2\, \partial s^2 / \partial\mathbf{a}\right] = \mathbf{0}$$

$\partial s_B^2 / \partial\mathbf{a} = 2\mathbf{Ba}$ ，$\partial s_W^2 / \partial\mathbf{a} = 2\mathbf{Sa}$ であるので、上の式は以下となる。

$$\mathbf{Ba} = \rho\mathbf{Sa} \qquad (2.5)$$

これを対称行列の固有方程式にするために、適当な下三角行列 $\mathbf{F}$ を用いて対称行列 $\mathbf{S}$ を $\mathbf{S} = \mathbf{F}\,{}^t\mathbf{F}$ のように書いて、(2.5)式を以下のようにする。

$$\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\,{}^t\mathbf{Fa} = \rho\,{}^t\mathbf{Fa}$$

ここで $\mathbf{A} = \mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}$ ，$\mathbf{u} = {}^t\mathbf{Fa}$ （$\mathbf{a} = {}^t\mathbf{F}^{-1}\mathbf{u}$）とすると、上式は以下のような対称行列の固有方程式となる。

$$\mathbf{Au} = \rho\mathbf{u} \qquad (2.6)$$

${}^t\mathbf{uu} = 1$ の規格化条件を付けて r 番目の固有値 $\rho^{(r)}$ について方程式を解いた答えを、$\mathbf{u}^{(r)}$ とすると、正準判別関数の係数は以下で与えられる。

$$\mathbf{a}^{(r)} = {}^t\mathbf{F}^{-1}\mathbf{u}^{(r)}$$

以上より、第 r 番目の固有値に対応する判別関数 $z^{(r)}$ は以下のようになる。

$$z^{(r)} = {}^t\mathbf{xa}^{(r)} - {}^t\tilde{\mathbf{x}}^{\alpha}\mathbf{a}^{(r)} \qquad (2.7)$$

ここに $\tilde{\mathbf{x}}^{\alpha} = \frac{1}{m}\sum_{\alpha=1}^{m}\bar{\mathbf{x}}^{\alpha}$ である。定数項については、後に述べる２群の場合のマハラノビス形式と正準形式の同一性から、各固有ベクトルに対応する判別関数の群別平均の単純平均が 0 になるように決めた。

マハラノビス形式と同様、変数 x_i の標準化値 u_i と不偏分散 s_i^2 を用いて判別関数は以下のように書くこともできる。

$$z^{(r)} = {}^t\mathbf{u}\mathbf{c}^{(r)} + {}^t\bar{\mathbf{x}}\mathbf{a}^{(r)} - {}^t\tilde{\mathbf{x}}\mathbf{a}^{(r)}, \qquad c_i^{(r)} = a_i^{(r)}s_i \tag{2.8}$$

この係数 $\mathbf{c}^{(r)}$ を標準化係数と呼ぶ。

(2.6) 式から、

$$\rho^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{A}\mathbf{u}^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\mathbf{u}^{(r)} = {}^t\mathbf{a}^{(r)}\mathbf{B}\mathbf{a}^{(r)} = s_B^{(r)2}$$

となり、r 番目の固有値は群間分散の第 r 成分に等しくなる。この性質を用いて、r 番目の固有値に対する変動の寄与率 $P^{(r)}$ を以下で与える。

$$P^{(r)} = \rho^{(r)} \Big/ \sum_{k=1}^{m-1}\rho^{(k)}$$

2.3　２群におけるマハラノビス形式と正準形式の同等性

さて、ここで述べてきた従来の理論とマハラノビスの距離を用いた判別分析とはどのような関係にあるのだろうか。(2.5)式について再考する。ここに方程式を再度挙げておく。

$$\mathbf{B}\mathbf{a} = \rho\mathbf{S}\mathbf{a}$$

行列 $\mathbf{B}$ は成分を用いて書くと以下のように表される。

$$\begin{aligned} b_{ij} &= \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha\left(\bar{x}_i^{\alpha} - \bar{x}_i\right)\left(\bar{x}_j^{\alpha} - \bar{x}_j\right) \\ &= \frac{1}{n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta\left(\bar{x}_i^{\alpha}\bar{x}_j^{\alpha} - \bar{x}_i^{\alpha}\bar{x}_j^{\beta}\right) \\ &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta\left(\bar{x}_i^{\alpha} - \bar{x}_i^{\beta}\right)\left(\bar{x}_j^{\alpha} - \bar{x}_j^{\beta}\right) \end{aligned}$$

これより、$(\mathbf{S}_B\mathbf{a})_{ij}$ は以下のように書ける。

$$\begin{aligned} (\mathbf{S}_B\mathbf{a})_i &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m}\sum_{j=1}^{p} n_\alpha n_\beta\left(\bar{x}_i^{\alpha} - \bar{x}_i^{\beta}\right)\left(\bar{x}_j^{\alpha} - \bar{x}_j^{\beta}\right)a_j \\ &= \sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} c_{\alpha\beta}\left(\bar{x}_i^{\alpha} - \bar{x}_i^{\beta}\right) \end{aligned}$$

$$c_{\alpha\beta} = \frac{n_\alpha n_\beta}{2n(n-m)} \sum_{j=1}^{p} \left(\bar{x}_j^\alpha - \bar{x}_j^\beta \right) a_j$$

特に 2 群の判別の場合、方程式(2.5)は以下となる。

$$\rho \mathbf{Sa} = \mathbf{S}_B \mathbf{a} = c(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

$$c = 2c_{12} = -2c_{21} = \frac{n_1 n_2}{n(n-2)} \sum_{j=1}^{p} \left(\bar{x}_j^1 - \bar{x}_j^2 \right) a_j$$

これより、解 $\mathbf{a}$ を求めると以下となる。

$$\mathbf{a} = \frac{c}{\rho} \mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

これは、(2.1)式で与えられたマハラノビス形式の判別関数の係数の定数倍である。よって、判別の分点を 0 にするような判別関数は以下となる。

$$z = \frac{c}{\rho} {}^t\mathbf{x}\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \frac{c}{2\rho} {}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

これは、判別関数全体が定数倍となっただけで、判別結果は $-\log h$ の項を除いて同等である。

2.4　ソフトウェアの利用法

メニュー［分析－多変量解析等－判別分析］をクリックすると、図 2.1 のような判別分析メニュー画面が表示される。

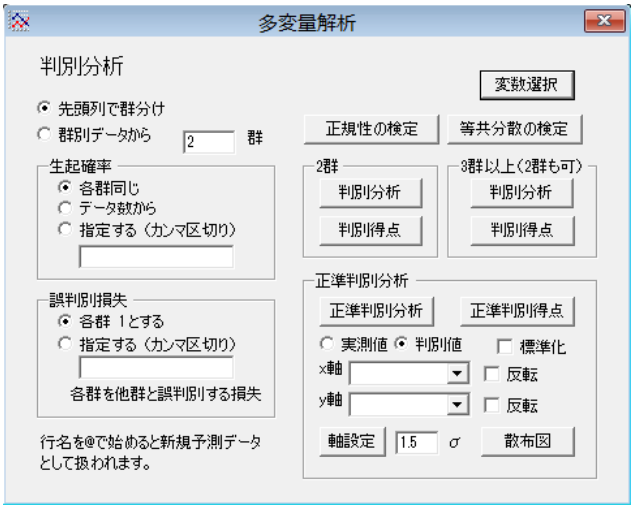


図 2.1　判別分析メニュー画面

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、ラジオボタンの「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。これらのデフォルト値は生起確率が「各群同じ」、誤判別損失が「各群 1 とする」である。

2 群の判別の場合、「等共分散の検定」ボタンで等共分散性を調べることができる。図 2.2 に「等共分散の検定」の出力結果を示す。

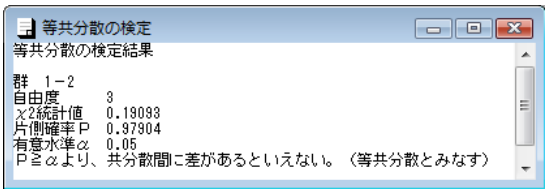


図 2.2　等共分散の検定

図 2.3 と図 2.4 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 0 と比較して決定される。

判別分析

	勉強時間	平均点	定数項
▶ 判別関数	2.2461	0.2007	-23.0187
標準化係数	2.6210	2.2787	-0.3788
F検定値	19.8822	15.0274	
自由度	1,27	1,27	
確率	0.0001	0.0006	
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	
理論から	0.1167	0.1167	
実測から	0.0769	0.0588	
判別数（実＼予）	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 2.3　判別分析実行結果（2 群の形式）

標準化係数の定数項は、重回帰分析などでは 0 になるが、判別分析では、判別の分点を 2 つの群の群別平均のデータ数による加重平均ではなく、単純平均にしていることから、2 つの群のデータ数が異なる場合、一般に 0 にならない。

判別得点

	所属群	判別得点	判別群
6	1	0.3280	1
7	1	−0.7743	2
8	1	4.9054	1
9	1	1.3153	1
10	1	1.8934	1
11	1	1.0704	1
12	1	4.0450	1
13	1	2.2301	1
14	2	−4.8682	2
15	2	−0.0469	2
16	2	−0.9540	2
17	2	−2.1784	2

図 2.4　判別得点（2 群の形式）

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を図 2.5 と図 2.6 に示す。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。

判別分析

	勉強時間	平均点	定数項
1群判別関数	8.7369	1.0833	−61.8513
2群判別関数	6.4908	0.8826	−38.8327
1群標準化係数	10.1951	12.2975	47.1974
2群標準化係数	7.5741	10.0189	47.5762
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	0.0769	0.0588	
判別数（実＼予）	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 2.5　判別分析実行結果（3 群以上の形式）

判別得点：既存データの判別

	所属群	1群	2群	判別群
6	1	53.3136	52.9857	1
7	1	43.6412	44.4156	2
8	1	72.6009	67.6956	1
9	1	58.3039	56.9886	1
10	1	54.6531	52.7597	1
11	1	50.2115	49.1412	1
12	1	65.9266	61.8816	1
13	1	62.2250	59.9949	1
14	2	23.2397	28.1079	2
15	2	48.9213	48.9682	2
16	2	44.3643	45.3183	2
17	2	37.7561	39.9345	2

図 2.6　判別得点（3 群以上の形式）

次に我々は正準形式に基づく判別の結果を示す。これは正準判別分析とも呼ばれている。正準判別分析における判別関数は、変数の数≧群の数、の場合は、群の数－１個作られる。同じデータを用いた結果を図 2.7 に示す。

正準判別分析

	勉強時間	平均点	定数項
▶ 判別1	0.9423	0.0842	-9.6565
標準化1	1.0995	0.9559	-0.1589
	固有値	寄与率	累積寄与率
判別1	1.4950	1.0000	1.0000
判別の分点	0		
	1群を他群と	2群を他群と	
誤判別確率	0.0769	0.0588	

図 2.7　正準判別分析実行結果

生起確率が同じで誤判別損失が 1 の場合、2 群のマハラノビス形式と正準形式の同等性から、判別関数の係数は比例している。また、判別の分点は２つの形式とも 0 に設定している。

正準判別分析の判別得点では、図 2.8 のように最後に群別得点平均が付く。これは 3 群以上の場合でも同様である。

判別得点

	所属群	判別得点1	判別群
25	2	-1.8352	2
26	2	-2.3991	2
27	2	-2.4203	2
28	2	-1.8778	2
29	2	-0.4873	2
30	2	-2.0510	2
群別得点平均	1	1.1919	
	2	-1.1919	

図 2.8　正準判別分析の判別得点

次に 3 群以上の正準判別分析の結果を図 2.9 に示す。

正準判別分析

	がくの長さ	がくの幅	花弁の長さ	花弁の幅	定数項
▶ 判別1	0.8294	1.5345	-2.2012	-2.8105	2.1051
判別2	0.0241	2.1645	-0.9319	2.8392	-6.6615
標準化1	0.6868	0.6688	-3.8858	-2.1422	0.0000
標準化2	0.0200	0.9434	-1.6451	2.1641	0.0000
	固有値	寄与率	累積寄与率		
判別1	32.1919	0.9912	0.9912		
判別2	0.2854	0.0088	1.0000		

図 2.9　正準判別分析実行結果

ここでは標準化係数が 0 になっているが、これは 3 つの群のデータ数がすべて同じであることにより、一般には 0 と異なる。3 群の判別得点は 2 つの固有値に対応して図 2.10 のように 2 種類出力される。

判別得点

	所属群	判別得点1	判別得点2
1	1	8.0618	0.3004
2	1	7.1287	-0.7867
3	1	7.4898	-0.2654
4	1	6.8132	-0.6706
5	1	8.1323	0.5145
6	1	7.7019	1.4617
7	1	7.2126	0.3558
8	1	7.6053	-0.0116
9	1	6.5606	-1.0152
10	1	7.3431	-0.9473

図 2.10　正準判別分析の判別得点

これは 2 次元上の点であるので、「軸設定」を行い、「散布図」ボタンをクリックすることにより、図 2.11 のような散布図が表示される。

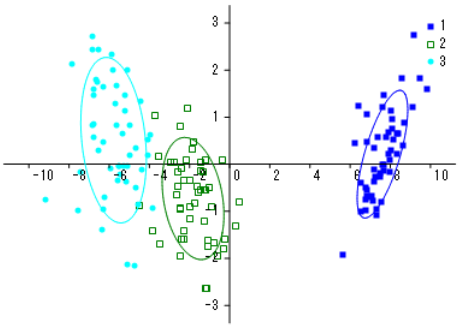


図 2.11　判別得点散布図

ここには、各群の分布を 2 変量正規分布とみなした場合の、1.5σ の確率楕円が示されている。確率楕円の大きさ、座標軸の反転等はメニューで変更できる。

この 2 変量正規分布の密度関数式は、グラフメニュー「設定－正規楕円半径－密度関数数式」で図 2.12 のように表示される。

群別正規分布密度関数

```
0.2897*exp(-0.9543*(1.4208*(x-(7.6076))^2+1.2220*(y-(0.2151))^2+(-1.8184)*(x-(7.6076))*(y-(0.2151))))
0.1863*exp(-0.5387*(0.9504*(x-(-1.8251))^2+1.3374*(y-(-0.7279))^2+(0.6045)*(x-(-1.8251))*(y-(-0.7279))))
0.1280*exp(-0.5264*(0.8446*(x-(-5.7826))^2+0.7278*(y-(0.5128))^2+(0.3514)*(x-(-5.7826))*(y-(0.5128))))
```

図 2.12　2 変量正規分布密度関数式

この式をコピーし、分析メニュー「数学－2 変量関数グラフ」のテキストボックスに貼り付けて（[Shift+Ins] または [Ctrl+v]）、（範囲を設定、分割数を増加、色を指定に）表示させると、図 2.13 のように 3 つの密度関数グラフを重ね合わせて視覚化することもできる。これによってどの程度分離ができているのか直感的に見ることもできる。

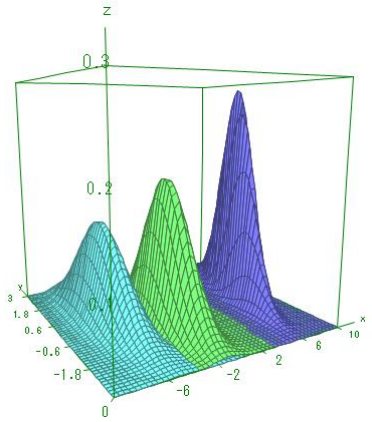


図 2.13　確率密度関数の視覚化

３．数量化Ⅱ類

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。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 3.1 のデータ形式を元にする。

表 3.1　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。

$$\sum_{j=1}^{r_i} x_{ij\lambda}^{\alpha} = 1 \qquad (3.1)$$

このため、アイテムごとに独立なカテゴリの数は 1 つ少なくなる。通常は第 1 カテゴリを除いた変数を用いて分析を実行する。

ここで、$x_{ij\lambda}^{\alpha}$ の表式を判別分析と類似のものとするため、新しい表記として $x_{I\lambda}^{\alpha}$ を導入する。この大文字の I はアイテム i 、その中のカテゴリ $j\,(=2,\cdots,r_i)$ について、順番にアイテム 1 から並べた数で、$I \equiv \sum_{k=1}^{i-1}(r_k - 1) + (j-1)$ で定義される。変数 I の範囲は $I = 1,2,\cdots,P \equiv \sum_{k=1}^{p}(r_k - 1)$ である。この変数表記法を用いると第 1 カテゴリを除いた数量化Ⅱ類は判別分析と同等であることが理解し易い。以後は

$$\sum_{I=1}^{P} f_I \quad \Leftrightarrow \quad \sum_{i=1}^{p}\sum_{j=1}^{r_i} f_{ij}$$

と置き換えることによって、両者の表記を使い分けることにする。

3.1 マハラノビスの距離に基づく方法

新しい変数表記法 $x_{I\lambda}^{\alpha}$ でデータを見ると 0,1 型のデータであっても、判別分析と同等に扱うことができる。よってデータの判別はマハラノビスの距離に基づく方法を用いて、判別分析と同じように行うことができる。但し、データの分布は正規分布ではないので、判別分析の最初のところで述べた分布関数による判別の理由付けはできない。しかし、2.3 節で述べたように、2 群の場合は正準形式と同等であるので、判別関数による群間分散の最大化の方法による理由付けは説得力がある。3 群以上の場合は、群間の 1 対比較によって判別を行うものと解釈すると、判別の問題は判別分析と全く同等に考えることができる。

2 群の場合、判別分析と同じように作られた係数を用いて判別関数は以下のように与えられる。ここでは判別関数との類似性を強調するため、新しい変数表記法を用いている。

$$z = \sum_{I=1}^{P} a_I x_I - \frac{1}{2}\sum_{I=1}^{P}(\bar{x}_I^1 + \bar{x}_I^2)a_I \quad , \quad a_I = \sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}(\bar{x}_J^1 - \bar{x}_J^2) \qquad (3.2)$$

また、3 群以上の場合、群 α の判別関数は以下のように与えられる。

$$z^{\alpha} = \sum_{I=1}^{P} a_I^{\alpha} x_I - \frac{1}{2}\sum_{I=1}^{P}\bar{x}_I^{\alpha} a_I^{\alpha} \quad , \quad a_I^{\alpha} = \sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}\bar{x}_J^{\alpha} \qquad (3.3)$$

2 群の場合も 3 群以上の場合も、係数ベクトル a_{ij} 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 d_{ij} $(i = 1,\cdots, p,\ j = 1,2,\cdots,r_i)$ も計算しておく。

2 群の場合　　$d_{ij} = \hat{a}_{ij} - \sum_{k=1}^{r_i} \tilde{x}_{ik}\hat{a}_{ik}$,　　$\hat{a}_{ij} = \begin{cases} 0 & j=1 \\ a_{ij} & j \neq 1 \end{cases}$

3 群以上の場合　　$d_{ij}^{\alpha} = \hat{a}_{ij}^{\alpha} - \sum_{k=1}^{r_i} \tilde{x}_{ik}\hat{a}_{ik}^{\alpha}$,　　$\hat{a}_{ij}^{\alpha} = \begin{cases} 0 & j=1 \\ a_{ij}^{\alpha} & j \neq 1 \end{cases}$

ここに基準化ウェイトの意味がカテゴリの影響が判別に正に働くか負に働くかを見ることであると考えて、以下のように、$\tilde{x}_{ik}$ はアイテム i、カテゴリ k における群平均の単純平均とした。

$$\tilde{x}_{ik} = \frac{1}{m}\sum_{\alpha=1}^{m} \overline{x}_{ik}^{\alpha}$$

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。

2 群の場合　　$z = \sum_{i=1}^{p}\sum_{j=1}^{r_p} d_{ij}x_{ij}$　　(3.4)

3 群以上の場合　　$z^{\alpha} = \sum_{i=1}^{p}\sum_{j=1}^{r_i} d_{ij}^{\alpha}x_{ij} + \sum_{i=1}^{p}\sum_{j=1}^{r_i} \tilde{x}_{ij}\hat{a}_{ij}^{\alpha} - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{r_i} \overline{x}_{ij}^{\alpha}\hat{a}_{ij}^{\alpha}$　　(3.5)

判別分析は変数 1 つ 1 つが独立であったが、数量化Ⅱ類の場合は、1 つのアイテムが判別分析の 1 つの変数に対応する。その中にはいくつかのカテゴリが含まれているために、アイテムの重要性は複数のカテゴリをまとめた重要性と解釈される。そのため、アイテムの重要性をみるには、カテゴリによる判別関数値の変化幅であるウェイト範囲や以下に述べるアイテムと判別関数値との相関係数、アイテムと判別関数値との偏相関係数の値などが参照される。

アイテムと判別関数間の相関係数を次のように与える。

$$r_{ij} = s_{ij} \Big/ \sqrt{s_{ii}s_{jj}}\ ,\quad r_{iz} = s_{iz} \Big/ \sqrt{s_{ii}s_{zz}}$$

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

$$s_{ij} = \frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha} - \overline{x}_i)(x_{j\lambda}^{\alpha} - \overline{x}_j)\ ,\quad s_{iz} = \frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha} - \overline{x}_i)(z_{\lambda}^{\alpha} - \overline{z})\ ,$$

$$s_{zz} = \frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(z_{\lambda}^{\alpha} - \overline{z})^2$$

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

変更点を明らかにするために、プログラム変更以前の定義も与えておく。

$$s_{iz} = \frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha} - \overline{x}_i)(\overline{z}^{\alpha} - \overline{z})\ ,\quad s_{zz} = \frac{1}{n-1}\sum_{\alpha=1}^{m} n_\alpha(\overline{z}^{\alpha} - \overline{z})^2\ ,\quad \overline{z}^{\alpha} = \frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} z_{\lambda}^{\alpha}$$

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^{iz}, r^{zz} を用いて、以下のように与えられる。

$$\tilde{r}_{iz} = -r^{iz} \Big/ \sqrt{r^{ii} r^{zz}}$$

数量化Ⅱ類では 2 群の判別の場合、各アイテムについて判別分析と同様にその有効性の F 値を求めることができる。アイテム i の有効性の F 値は以下となる。最後の分布形は仮に変数の正規性が成り立つ場合の性質であるが、当然数量化Ⅱ類のデータでは成り立たない。参考までの仮の表示である。

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

ここに、D_i^2 は両群のカテゴリ i を除いたマハラノビス距離である。

3.2 正準形式に基づく方法

マハラノビス形式と同様に、判別関数は係数 a_{ij} $(i = 1, \cdots, p,\ j = 2, \cdots, r_i)$ と定数 z_0 を用いて以下のように与える。

$$z_\lambda = \sum_{i=1}^{p} \sum_{j=2}^{r_i} a_{ij} x_{ij\lambda} + z_0$$

この判別関数は新しい変数表記法では以下となる。

$$z_\lambda = \sum_{I=1}^{P} a_I x_I + z_0$$

この表記法では、第 1 カテゴリを除いた数量化Ⅱ類と判別分析が同等である。

我々は z_λ^α の群間の変動 s_B^2 と群別変動の合計 s^2 を以下のように定義し、群間の変動を際立たせるために、これらの分散比 $\rho = s_B^2 / s^2$ を最大化することを考える。

$$s_B^2 = \sum_{\alpha=1}^{m} n_\alpha (\bar{z}^\alpha - \bar{z})^2, \quad s^2 = \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} (z_\lambda^\alpha - \bar{z}^\alpha)^2$$

ここに、$\bar{z}^\alpha = \dfrac{1}{n_\alpha} \sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$, $\bar{z} = \dfrac{1}{n} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$, $n = \sum_{\alpha=1}^{m} n_\alpha$ である。

この分散比を係数で微分することにより、判別分析と同様に以下の方程式が得られる。

$$\mathbf{Ba} = \rho \mathbf{Sa} \tag{3.6}$$

この方程式はデータを以下のようにまとめ、

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

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

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

方程式中の行列を以下のように定義することによって得られる。

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

$$\mathbf{S}=\frac{1}{n-m}\,{}^t(\mathbf{X}-\bar{\mathbf{X}}_B)(\mathbf{X}-\bar{\mathbf{X}}_B),\quad \mathbf{B}=\frac{1}{n-m}\,{}^t(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})$$

ここに n はすべての群のデータ数の合計、m は群の数である。

方程式 (3.6) は正準判別分析と同様の方法で変形され、以下となる。

$$\mathbf{Au}=\rho\mathbf{u} \tag{3.7}$$

ここに、$\mathbf{A}=\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}$, $\mathbf{u}={}^t\mathbf{Fa}$、また $\mathbf{F}$ は $\mathbf{S}=\mathbf{F}\,{}^t\mathbf{F}$ となる下三角行列である。

(3.7) 式の第 r 固有値に対する規格化された固有ベクトル $\mathbf{u}^{(r)}$ を使って、係数は $\mathbf{a}^{(r)}={}^t\mathbf{F}^{-1}\mathbf{u}^{(r)}$ となり、これにより判別関数は以下となる。

$$z^{(r)}=\sum_{I=1}^{P}a_I^{(r)}x_I-\sum_{I=1}^{P}a_I^{(r)}\tilde{x}_I \tag{3.8}$$

ここで定数項については、正準判別分析と同様に、各固有値に対応する判別関数の群別平均の単純平

均が 0 になるようにしている。

係数 $a_{ij}^{(r)}$ 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した係数 $d_{ij}^{(r)}$ $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$ も計算しておく。

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

ここに基準化ウェイトの意味を考えて、$\tilde{x}_{ik}$ は判別関数のときと同様に、アイテム i、カテゴリ k における群平均の単純平均とした。

$$\tilde{x}_{ik} = \frac{1}{m} \sum_{\alpha=1}^{m} \bar{x}_{ik}^{\alpha}$$

基準化されたカテゴリウェイトを用いると、判別関数は以下のように与えられる。

$$z^{(r)} = \sum_{i=1}^{p} \sum_{j=1}^{r_i} d_{ij}^{(r)} x_{ij} \tag{3.9}$$

3.3 ソフトウェアの利用

メニュー［分析－多変量解析等－数量化理論－数量化Ⅱ類］を選択すると、数量化Ⅱ類のメニュー画面が図 3.1 のように表示される。

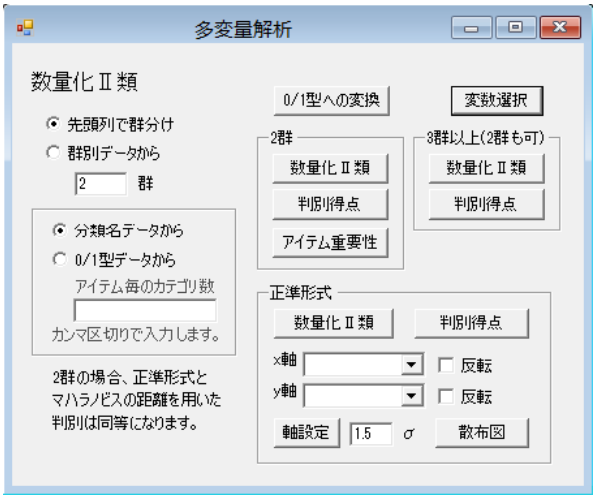
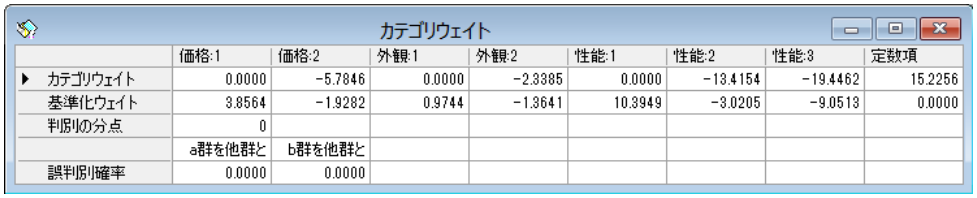


図 3.1　数量化Ⅱ類メニュー画面

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。また、計算方式としては、上側に示されたマ

ハラノビス形式と下側に示された正準形式のどちらかを選択できる。正準形式は、これまでの計算結果を踏襲するものであるが、定義の違いから、係数について定数倍の違いがある。しかし、判別結果については同じである。マハラノビス形式は、2 群の場合、判別分析のところで示したように、正準形式と定数倍の違いを除いて同じである。しかし、3 群以上の場合では大きく異なり、判別分析と同様の結果を出力する。マハラノビス形式の結果は、各カテゴリの第 1 アイテムを除いた変数で判別分析を行った結果と一致する。我々はまず、2 群の場合の結果を比較して、3 群の場合の違いを見ることにする。

「数量化Ⅱ類」コマンドボタンをクリックした結果を比較する。マハラノビス形式の結果を図 3.2a に、正準形式の結果を図 3.2b に与える。



	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ カテゴリウェイト	0.0000	-5.7846	0.0000	-2.3385	0.0000	-13.4154	-19.4462	15.2256
基準化ウェイト	3.8564	-1.9282	0.9744	-1.3641	10.3949	-3.0205	-9.0513	0.0000
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

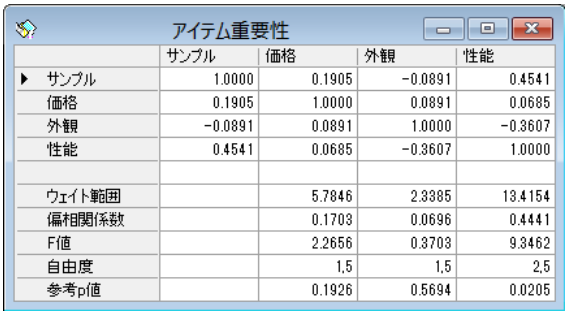
図 3. 2a　マハラノビス形式のカテゴリウェイト

	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ 判別1	0.0000	-1.4636	0.0000	-0.5917	0.0000	-3.3943	-4.9202	3.8524
基準化1	0.9757	-0.4879	0.2465	-0.3451	2.6301	-0.7642	-2.2901	0.0000
	固有値	寄与率	累積寄与率					
判別1	4.6862	1.0000	1.0000					
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

図 3.2b　正準形式のカテゴリウェイト

ここではカテゴリウェイト、基準化されたカテゴリウェイト、判別の分点、誤判別確率が表示される。2 群の判別の場合、判別の分点は 0 にしている。2 つのカテゴリウェイトはそれぞれ比例している。正準形式の場合は、固有値と寄与率、累積寄与率が表示されるが、2 群の場合、寄与率と累積寄与率は定義より 1 になる。

2 群の場合、2 つの方法は同等であるので、以後はマハラノビス形式の結果のみを表示する。「アイテム重要性」ボタンをクリックすると、図 3.3 のような結果が表示される。



アイテム重要性

	サンプル	価格	外観	性能
▶ サンプル	1.0000	0.1905	-0.0891	0.4541
価格	0.1905	1.0000	0.0891	0.0685
外観	-0.0891	0.0891	1.0000	-0.3607
性能	0.4541	0.0685	-0.3607	1.0000
ウェイト範囲		5.7846	2.3385	13.4154
偏相関係数		0.1703	0.0696	0.4441
F値		2.2656	0.3703	9.3462
自由度		1,5	1,5	2,5
参考p値		0.1926	0.5694	0.0205

図 3.3　アイテム重要性

ここでは、相関行列とそれを基に計算される偏相関係数及び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ウェイト範囲が表示される。ウェイト範囲は各アイテムの重要性を見るのに用いられる。またアイテムの重要性を示す F 値等も表示される。データに正規性がないために、F 値の確率は参考 p 値として表示してある。

図 3.4 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や 2 群の場合は判別の分点も示されている。

判別得点

	所属群	判別得点	予測群
▶ 1	a	1.8103	a
2	a	9.4410	a
3	a	12.8872	a
4	a	7.1026	a
5	b	-4.2205	b
6	b	-12.3436	b
7	b	-6.3128	b
8	b	-10.0051	b
9	b	-3.9744	b
10	b	-10.0051	b
群別得点平均	a	7.8103	
	b	-7.8103	
判別の分点		0	

図 3.4　判別得点

以後は 3 群以上の場合を扱う。3 群の場合、正準形式とマハラノビス形式ではかなり異なる。マハラノビス形式では群別の判別関数が出力されるのに対して、正準形式では固有値に対応する判別関数が出力される。前者はどの判別関数の値が大きいかによって判別結果を決めるが、後者は判別結果を多次元上に表示するためのものである。結果を比較して示しておく。それぞれ、図 3.5a と図 3.5b のように結果が表示される。

カテゴリウェイト

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
1群判別関数	0.0000	3.2656	3.7813	0.0000	2.0625	1.7188	−0.5328
2群判別関数	0.0000	15.9844	20.5759	0.0000	8.9375	10.0670	−12.7114
3群判別関数	0.0000	16.3281	22.0491	0.0000	10.3125	13.3080	−15.8739
1群基準化関数	−2.5495	0.7161	1.2318	−1.2432	0.8193	0.4755	3.2599
2群基準化関数	−13.0993	2.8850	7.4766	−6.3913	2.5462	3.6757	6.7792
3群基準化関数	−13.6903	2.6379	8.3589	−8.0233	2.2892	5.2847	5.8397
	1群を他群と	2群を他群と	3群を他群と				
誤判別確率	0.2000	0.4000	0.2500				

図 3.5a　マハラノビス距離を用いたカテゴリウェイト

カテゴリウェイト

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
判別1	0.0000	−2.9339	−4.0012	0.0000	−1.7342	−2.3017	3.8454
判別2	0.0000	−2.0006	−1.4483	0.0000	0.3109	2.2173	0.3633
基準化1	2.4717	−0.4622	−1.5295	1.3737	−0.3605	−0.9280	0.0000
基準化2	1.3014	−0.6992	−0.1469	−0.9381	−0.6272	1.2793	0.0000
	固有値	寄与率	累積寄与率				
判別1	5.5682	0.9778	0.9778				
判別2	0.1263	0.0222	1.0000				

図 3.5b　正準形式を用いたカテゴリウェイト

それぞれの方法の「判別得点」をクリックした結果を図 3.6a と図 3.6b に示す。

判別得点

	所属群	1群判別得	2群判別得	3群判別得	予測群
1	1	1.5297	−3.7739	−5.5614	1
2	1	2.7328	3.2730	0.4542	2
3	1	−0.5328	−12.7114	−15.8739	1
4	1	−0.5328	−12.7114	−15.8739	1
5	1	−0.5328	−12.7114	−15.8739	1
6	2	4.4516	13.3400	13.7623	3
7	2	3.2484	7.8645	6.1752	2
8	2	4.7953	12.2105	10.7667	2
9	2	4.4516	13.3400	13.7623	3
10	2	5.3109	16.8020	16.4877	2
11	3	4.4516	13.3400	13.7623	3
12	3	4.9672	17.9315	19.4833	3
13	3	4.7953	12.2105	10.7667	2
14	3	4.9672	17.9315	19.4833	3

図 3.6a　マハラノビス距離を用いた判別得点

判別得点

	所属群	判別得点1	判別得点2
1	1	2.1112	0.6742
2	1	0.9115	-1.6372
3	1	3.8454	0.3633
4	1	3.8454	0.3633
5	1	3.8454	0.3633
6	2	-1.3902	0.5801
7	2	-0.1558	-1.0850
8	2	-0.8227	-1.3263
9	2	-1.3902	0.5801
10	2	-1.8900	-0.7741
11	3	-1.3902	0.5801
12	3	-2.4575	1.1324
13	3	-0.8227	-1.3263
14	3	-2.4575	1.1324
群別得点平均	1	2.9118	0.0254
	2	-1.1298	-0.4050
	3	-1.7820	0.3796

図 3.6b　従来の方法による判別得点

マハラノビス形式では、判別関数の値の最も大きい群に判別されることが示されているが、正準形式では判別結果は明確に示されていない。正準形式では複数の次元の判別点を見て判断を下すため、2 次元上に散布図を描画する機能が付けられている。メニューの「軸設定」で表示する次元を設定し、「散布図」ボタンにより、図 3.7 のように判別得点を平面上に表示する。図中の楕円は1.5σ を表す楕円である。重なった点が多いため、散布図はあまり見易いとは言えない。

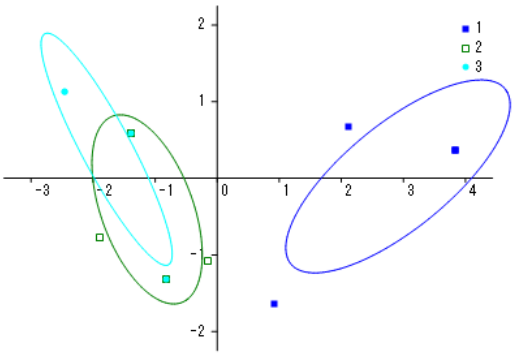


図 3.7　判別得点による散布図

４．おわりに

我々は、以前 College Analysis の中で判別分析と数量化Ⅱ類のプログラムを作成したが[1]、２つの分析の類似性を議論することはなかった。今回これらの分析を再検討するに当たり、以前のプログラムを見直し、2 つの分析間に係数の違いや不足する機能を見出した。この違いは結果や解釈に影響を与えるものでないが、これらの分析の学習者が類似性を理解するためには是正すべきものである。ま

た、不足する機能は２つの分析の対比のために補っておくべきものである。そのため、我々はこれらの分析のプログラムの大半を、同一性という視点で作り変えた。

基本的に判別分析は量的データ、数量化Ⅱ類は質的データの分析であるが、質的データを 0, 1 データに変更し、アイテムごとに第 1 カテゴリを除いて、判別分析で統合化できる。この方法だと量的データと質的データは区別されず、2 種類のデータを混在させて分析を実行することもできる。しかし、アイテムの重要性などの評価は見えにくくなるので、直接数量化Ⅱ類の中に混在するデータを取り込んで、プログラムによりどちらのデータかを判定し、分析することを考えてもよい。

この方法だとデータの種類を見極める手段を考えなければならないが、例えば、量的データには変数名の後ろに「#」、質的データには変数名の後ろに「&」、特に指定しない場合は何も付けない等の方法を考えればよい。これは判別分析と数量化Ⅱ類に限ったことではなく、他の分析でも量的データと質的データの誤用の防止などに役に立つ。今後これまでのプログラムに影響を与えないように組み込んでいきたい。

この論文では特に判別分析と数量化Ⅱ類について考えたが、我々は今回多変量解析全体について見直しを行い、新しいプログラムもいくつか追加した。次の論文ではこれらのプログラムについても説明したいと考えている。

参考文献

[1] 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, 福山平成大学経営情報研究, 7 号, (2002) 85-106.

[2] 多変量解析法入門, 永田靖, 棟近雅彦, サイエンス社, 2001.

Multi-purpose Program for Social System Analysis 24

- Integration of Discriminant Analysis and Quantification Method Type II -

Masayasu FUKUI, Makoto OZAKI and Ryota ASAHI

Department of Business Administration, Faculty of Business Administration,
Fukuyama Heisei University

Abstract

We have been constructing a unified program on the social system analysis for the purpose of education. In this paper discriminant analysis and quantization type II, which have been treated independently, are reconfigured in a unified manner by the method to use Mahalanobis distance and canonical format. We describe the mathematical theory and operation of our program.

Keywords

College Analysis, social system analysis, statistics, discriminant analysis, quantification method

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