第６講　実証研究への応用（Ｉ）

6.1　インドにおける結婚持参金の推計

- Vijayendra Rao. (1993) "The Rising Price of Husbands: A Hednic Analysis of Dowry Increases in Rural India," *Journal of Political Economy*, 101(4), pp.666-677.

- Lena Edlund. (2000) "The Marriage Squeeze Interpretation of Dowry Inflation: A Comment," *Journal of Political Economy*, 108(6), pp.1327-1333.

- Vijayendra Rao. (2000) "The Marriage Squeeze Interpretation of Dowry Inflation: Response," *Journal of Political Economy*, 108(6), pp.1334-1335.

（課題 1）結婚持参金（Dowry）の決定式を推計せよ。

（課題 2）プロビット、トービットを用いて、花嫁・花婿の就学年数の決定式を推計せよ。

（課題 3）このデータセットを用いて、追加的なクロス・セクション分析を行いなさい（トピックは自由）。

推計においては、その関数型の選択の根拠となる統計テストの結果も含めて報告すること。また、推計結果だけではなく、理論的議論や全体の内容を吟味するような総合的な議論も含めること。

6.2　都道府県別にみた公共投資の効率性

（課題 4）1990 年代の財政の効果がなかったということは事実か。1950-60 年代と比べて何が変わったのか。

データの所在

土居丈朗 (http://www.econ.keio.ac.jp/staff/tdoi/)、深尾京司 (http://www.ier.hit-u.ac.jp/ fukao/) のホームページ、内閣府「県民経済計算」、朝日新聞「民力」

6.3　マクロベースでみた時系列の乗数効果

- 中島・北村・木村・新保「テキストブック　経済統計」（東洋経済新報社）第７章参照。

（課題 5）国民経済計算の政府公的資本形成および政府消費がＧＤＰ成長に与えた効果の時系列推計をしてみよう。

6.4 都道府県別にみた一人当たり公共投資額ランキングの変遷

- 河野龍太郎「都道府県別にみた公共投資の効率性：90 年代の公共投資は景気対策ではなかった！」『金融ビジネス』2001 年 6 月　pp.76-80。

この論文によれば、90 年代に行われた公共工事は、マクロ安定化政策と呼ぶのは名ばかりで、現実には単なる所得分配の色彩が極めて強かった。90 年代の不況はバブル崩壊の不況であり、都市型不況であった。その対策として地方に公共投資を行う必要はあったのか？

一人当たり県民所得と県民総生産に占める公共投資額比率を図示すると、公共投資比率は、一人当たり所得の多い地域では低く、所得の少ない地域では高いという傾向が顕著に見られる。

とすれば、公共投資は所得分配政策であったとみるべきである。

（課題 6）河野論文　図 2 では 1998 年度のみで作られているが、90 年代の各都道府県の数値をプロットして、各都道府県別の公共投資比率と県民所得の関係をパネル推計する。

（課題 7）公共投資の拡大が地域の生産性の上昇に結びついているか。生産性の低い産業の一つである建設業を通じて経済資源を投入した結果、その地域の生産性上昇の機会を奪ってしまったということもある。すなわち、賃金を高止まりさせ、企業進出を阻んできたとは考えられないだろうか検討してみよう。

6.5 過疎的地域と都市的地域の判別分析

- 杉原敏夫・藤田渉『多変量解析』（牧野書店）、第 3 章参照

過疎的／都市的の指標として人口密度を採用する。すなわち、事前に過疎的地域 10 県（富山、島根、新潟、鹿児島、福島、青森、山形、岩手、秋田、北海道）、都市的地域 10 県（東京、大阪、神奈川、埼玉、愛知、京都、兵庫、福岡、奈良、千葉）を選んでおいた上に、

$$\begin{array}{rll} \text{地域特性指標} & z=1: & \text{その地域が過疎的地域であった場合} \\ & z=0: & \text{その地域が都市的地域であった場合} \end{array}$$

を定義する。さらに各変量を次のように記号化する。標本数は対象とする上記の 20 県である。

$$\begin{array}{lll} \text{地域特性指標} & z_i & (i=1,\cdots,20) \\ \text{人口増加率} & x_{i1} & (i=1,\cdots,20) \\ \text{老齢人口割合} & x_{i2} & (i=1,\cdots,20) \end{array}$$

ここでが変量とについてどのように分布しているかをみる。判別分析では、新しい座標軸を定義して、2 群の重なり合う部分を最小とする座標軸の設定を行う。

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2 \end{aligned} \tag{6.1}$$

この原理に従って係数 a_{ij} が決定される。実際に x_1, x_2 を説明変量として**判別関数**（discriminant function）z を構成すれば

$$z = 0.1026x_1 + 1.7118x_2 \tag{6.2}$$

となる。
それぞれの判別基準は次のようになる。
人口増加率（x_1）の場合

過疎的地域の標本平均： 0.050
都市的地域の標本平均： 2.346
判別基準値＝ 1.198

老齢人口割合（x_2）の場合

過疎的地域の標本平均： 18.322
都市的地域の標本平均： 12.658
判別基準値＝ 15.490

判別関数による基準値＝ 26.638

人口増加率単独ではとくに、都市的地域において誤判別率が高い（40% ）、これに対して老齢人口割合はかなり高い判別能力を有している。

判別分析の原理（Ⅰ）

（1）1変量による判別
変量 x についての母集団平均と分散がそれぞれ $(\mu_1, \sigma^2), (\mu_2, \sigma^2)$ である2つの群 $G_{1,}G_2$ があるものとする。2つの群の分散は等しいと仮定する。

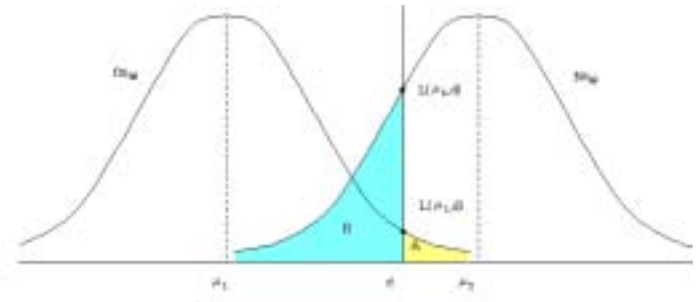


図6．1 境界による区分

G_1 と G_2 が重複しているため、境界を引くことにより標本を群別する場合に誤判別（discrimination error）が生じる可能性がある。すなわち

A： G_1に属している標本が G_2の領域として群別される。
B： G_2に属している標本が G_1の領域として群別される。

この誤判別の確率を最小にする境界を引くことが判別分析の主題であり、境界を設定する関数を**判別関数**という。

変量 x の分布は正規分布であると仮定する。標本 X についての変量 x が G_1 に属している確率密度関数を $f^{(1)}(x)$ とすると、

$$f^{(1)}(x) = \left(1/\sqrt{2\pi}\sigma\right)^{-1} \exp\left(-(x-\mu^{(1)})^2/2\sigma^2\right) \tag{6.3}$$

変量 x が G_2 に属している確率密度関数を $f^{(2)}(x)$ とすると、

$$f^{(2)}(x) = \left(1/\sqrt{2\pi}\sigma\right)^{-1} \exp\left(-(x-\mu^{(2)})^2/2\sigma^2\right) \tag{6.4}$$

図 6.1 のように、x の特定の値 d を定め、分布の平均 μ を変化させれば、d に対する分布の尤度（likelihood）$L(\mu, d)$ が定義される。

$$L(\mu, d) = \left(1/\sqrt{2\pi}\sigma\right)^{-1} \exp\left(-(d-\mu)^2/2\sigma^2\right) \tag{6.5}$$

d に対する G_1,G_2 の尤度は $L(\mu^{(1)}, d), L(\mu^{(2)}, d)$ であり、$x = d$ は尤度の大きい方の群に属すると考えられる。すなわち、

$$\begin{array}{l} L(\mu^{(1)}, d) \geq L(\mu^{(2)}, d) \text{ の場合}: \quad x = d \text{ は } G_1 \text{ に属する。} \\ L(\mu^{(1)}, d) < L(\mu^{(2)}, d) \text{ の場合}: \quad x = d \text{ は } G_2 \text{ に属する。} \end{array}$$

と判別する。$L(\mu^{(1)}, d) = L(\mu^{(2)}, d)$ の場合は A、B の相互の誤判別の確率の和は最小となる。このときは $d = (\mu^{(1)} + \mu^{(2)})/2$ となる。

一般に、x が分散 σ^2 を持つ正規分布を仮定した場合、標本 X と G_1,G_2 との距離は次のように定義される。

$$D_1^2 = \left((x-\mu^{(1)})/\sigma\right) \tag{6.6}$$

$$D_2^2 = \left((x-\mu^{(2)})/\sigma\right)^2 \tag{6.7}$$

この場合、標本 X は距離の近い方の群に属すると考えられるから、

$$\begin{array}{l} D_1^2 \geq D_2^2 \text{の場合}: \quad X \text{ は } G_2 \text{に属する。} \\ D_1^2 < D_2^2 \text{の場合}: \quad X \text{ は } G_1 \text{に属する。} \end{array}$$

距離の 2 乗の差をとると、

$$\begin{aligned} D^2 &= D_1^2 - D_2^2 \\ &= (1-\sigma^2)(-2(\mu^{(1)} - \mu^{(2)})x + (\mu^{(1)} - \mu^{(2)})(\mu^{(1)} + \mu^{(2)})) \end{aligned} \tag{6.8}$$

ここで

$$d = -2(\mu^{(1)} - \mu^{(2)})x + (\mu^{(1)} - \mu^{(2)})(\mu^{(1)} + \mu^{(2)}) \tag{6.9}$$

とすれば、d が判別関数となる。d は x について 1 次式であり、**線形判別関数**(linear discriminant function) という。

（2）多変量による判別

判別として2群判別を考える。G_1、G_2 はおのおの m 個の変量で構成されている。G_1、G_2 の平均を次のベクトルで定義する。

$$\mu^{(1)\prime} = \left[\mu_1^{(1)}, \mu_2^{(1)}, \cdots, \mu_m^{(1)}\right] \tag{6.10}$$

$$\mu^{(2)\prime} = \left[\mu_1^{(2)}, \mu_2^{(2)}, \cdots, \mu_m^{(2)}\right] \tag{6.11}$$

分散は共分散行列として表される。

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2m} \\ \vdots & & \ddots & \\ \sigma_{m1} & \sigma_{m2} & & \sigma_m^2 \end{bmatrix} \tag{6.12}$$

判別される標本 X は m 次元の要素を持つので次のように定義する。

$$x' = [x_1, x_2, \cdots, x_n] \tag{6.13}$$

X と G_1 との距離を D_1 とし、G_2 との距離を D_2 とすると

$$D_1^2 = (\mathbf{x} - \mu^{(1)})'\Sigma^{-1}(\mathbf{x} - \mu^{(1)}) \tag{6.14}$$

$$D_2^2 = (\mathbf{x} - \mu^{(2)})'\Sigma^{-1}(\mathbf{x} - \mu^{(2)}) \tag{6.15}$$

この距離は**マハラノビス距離**と呼ばれ、構成する変量間の相関を考慮した距離となっている。

$$\begin{aligned} D^2 &= D_1^2 - D_2^2 \\ &= -2(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mathbf{x} - (\mu^{(1)} + \mu^{(2)})/2) \\ &= -2(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}\mathbf{x} + 2(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}((\mu^{(1)} + \mu^{(2)})/2) \\ &= -2Z + 2d \end{aligned} \tag{6.16}$$

Z は線形判別関数、d は判別基準値である。

$$Z = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}\mathbf{x} \tag{6.17}$$

$$d = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}((\mu^{(1)} + \mu^{(2)})/2) \tag{6.18}$$

これにより、次のような判別ができる。

$D^2 \geq 0$ すなわち $Z \leq d$: X は G_2に属する。
$D^2 < 0$ すなわち $Z > d$: X は G_1に属する。

実際の議論では、母集団のパラメータはわからないので、標本平均や標本共分散行列を用いればよい。

（3）分散が不均一の場合

(6.14)、(6.15) において、G_1,G_2 の母集団の共分散行列を Σ_1,Σ_2 とする。ある標本 X と双方の群までの距離は次のように定義される。

$$D_1^2 = (\mathbf{x} - \mu^{(1)})' \Sigma_1^{-1} (\mathbf{x} - \mu^{(1)}) \tag{6.19}$$
$$D_2^2 = (\mathbf{x} - \mu^{(2)})' \Sigma_2^{-1} (\mathbf{x} - \mu^{(2)}) \tag{6.20}$$

従って、

$$D^2 = D_1^2 - D_2^2 = (\mathbf{x} - \mu^{(1)})' \Sigma_1^{-1} (\mathbf{x} - \mu^{(1)}) - (\mathbf{x} - \mu^{(2)})' \Sigma_2^{-1} (\mathbf{x} - \mu^{(2)}) \tag{6.21}$$

X については、次のように判別される。

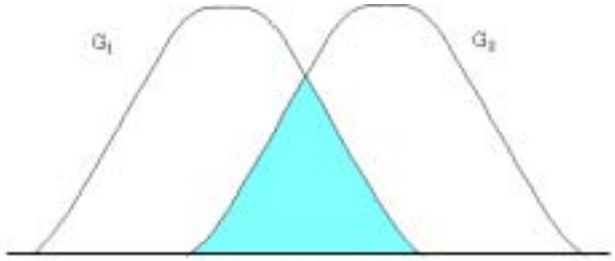
$D^2 \geq \log(\Sigma_1 / \Sigma_2)$: X は G_2に属する。
$D^2 < \log(\Sigma_1 / \Sigma_2)$: X は G_1に属する。

この場合、判別関数は x についての 2 次形式となり、線形とはならない。

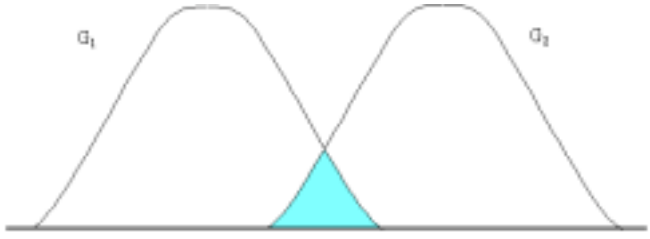
判別分析の原理（II）

（1）相関比に基づく判別関数の構成
母集団の分布が重複した 2 つの群 G_1,G_2 を判別しようとするとき、図 6.2 のような 2 つの場合を考える。

図 6.2　G_1,G_2 の分布



（1）重複部分が大きい



（1）重複部分が小さい

図 6.2(1) は重複する部分が大きく、判別の基準値を設けるのが困難。図 3.2.(2) は重複する部分が小さく、誤判別の確率は低い。とすれば、逆に、各群の母集団の分布の重複する部分が最小となるような変量を探すことが有効である。

各群の分布の重複度合いを表す指標として、**相関比**を用いることができる。

１変量の場合を考えよう。

G_1,G_2 の２つの群を考え、前者に属する n_1 個の標本を $x_i^{(1)}$、後者に属する n_2 個の標本を $x_i^{(2)}$ とする。また $n = n_1 + n_2$ とする。すなわち、

$$\begin{aligned} x_i^{(1)}(i = 1, 2, \cdots n_1): &\quad G_1\text{に属する標本} \\ x_i^{(2)}(i = 1, 2, \cdots n_2): &\quad G_2\text{に属する標本} \end{aligned}$$

このとき、G_1 においては

$$\text{標本平均：}\quad \bar{x}^{(1)} = \sum_{i=1}^{n_1} x_i^{(1)} \diagup n_1 \tag{6.22}$$

$$\text{標本分散：}\quad S_1 = \sum_{i=1}^{n_1} (x_i^{(1)} - \bar{x}_i^{(1)})^2 \diagup (n_1 - 1) \tag{6.23}$$

G_2 においては

$$\text{標本平均：}\quad \bar{x}^{(2)} = \sum_{i=1}^{n_2} x_i^{(2)} \diagup n_2 \tag{6.24}$$

$$\text{標本分散：}\quad S_2 = \sum_{i=1}^{n_2} (x_i^{(2)} - \bar{x}_i^{(2)})^2 \diagup (n_2 - 1) \tag{6.25}$$

この S_1, S_2 は G_1, G_2 の郡内における平均変動を表しており、$(n_1-1)S_1, (n_2-1)S_2$ は G_1, G_2 が各群内の変動の総和を表していることから、**級内変動**（sum of squares within classes）と呼ばれる。

標本を全体に広げて、標本平均と分数を求めれば、標本を x_i として、

$$\text{標本平均：}\quad \bar{x} = \sum_{i=1}^{n} x_i \diagup n_1 \tag{6.26}$$

$$\text{標本分散：}\quad S = \sum_{i=1}^{n} (x_i - \bar{x})^2 \diagup (n - 1) \tag{6.27}$$

ここで $(n-1)S$ は**全体変動**（total sum of squares）と呼ばれる。全体変動、G_1 の級内変動、G_2 の級内変動をそれぞれ T, W_1, W_2 とすれば

$$\begin{aligned} T &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=n_1+1}^{n} (x_i - \bar{x})^2 \\ &= \sum_{i=1}^{n_1} (x_i - \bar{x}^{(1)} + \bar{x}^{(1)} - \bar{x})^2 + \sum_{i=n_1+1}^{n} (x_i - \bar{x}^{(2)} + \bar{x}^{(2)} - \bar{x})^2 \\ &= W_1 + B_1 + W_2 + B_2 \end{aligned} \tag{6.28}$$

ここで $B_1 = \sum_{i=1}^{n_1} (\bar{x}_i - \bar{x}), \quad B_2 = \sum_{i=n_1+1}^{n_1} (\bar{x}_i^{(2)} - \bar{x})$

は、G_1, G_2 の標本平均の全体の平均に対する変動和であり、G_1, G_2 の級間変動（sum of squares between classes）と呼ばれる。図 6.2 で明らかなように、級間変動が大きい方が２群を分離しやすい。次の相関比 η^2 を定義する。

$$\eta^2 = (B_1 + B_2) \diagup T \tag{6.29}$$

相関比 η^2 は全体変動 (T) に対する級間変動の割合を表しており、G_1, G_2 がいくつかの変量から構成されている場合には、相関比が最大になる変量を選べば判

別効率が最も高くなる。また変量が複数個存在する場合は、それらの合成変量について、相関比を最大化すればよい。

m 個の変量から合成変量を構成し、相関比を最大化することを考えよう。変量を $x_j(1,2,\cdots,m)$ として、合成変量 z を構成する。この場合、x_j は標準化させておく。すなわち、全体の標本を x_{ij} と表記し、平均を x_j、標準偏差を s_j とすると、標準化した標本は次のように表される。

$$y_{ij} = (x_{ij} - \bar{x}_j) \diagup s_j \quad (i = 1,2,\cdots,n) \tag{6.30}$$

標準化した標本について合成変量 z を次のように構成する。

$$z = a_1 y_1 + a_2 y_2 + \cdots + a_m y_m = \mathbf{a}'\mathbf{y} \tag{6.31}$$

ここで、$\mathbf{a}$ は係数のベクトルである。

$$\mathbf{a}' = [a_1, a_2, \cdots, a_m] \tag{6.32}$$

標本 z_i について、(6.31) を書き換えると、

$$z_i = a_1 y_{i1} + a_2 y_{i2} + \cdots + a_m y_{im} = \sum_{k=1}^{m} y_{ik} a_k \tag{6.33}$$

z の標本ベクトルを

$$z' = [z_1, z_2, \cdots, z_n] \tag{6.34}$$

と表すと、z は行列 $\mathbf{Y}$ とベクトル a を用いて次のように表せる。

$$z = \mathbf{Y}\mathbf{a} \tag{6.35}$$

ここで、$\mathbf{Y}$ は n 行 m 列の $\mathbf{y}$ の標本行列であり、次のように表せる。

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{mm} \end{bmatrix} \tag{6.36}$$

$\mathbf{y}_i$ は $N(0,1)$ に従うから、$\mathbf{z}$ の平均は 0 である。このことより、

$$\mathbf{T} = \sum_{i=1}^{n} (z_i - \bar{z})^2 = \sum_{i=1}^{n} z_i^2 = (\mathbf{Y}\mathbf{a})'(\mathbf{Y}\mathbf{a}) = \mathbf{a}'\mathbf{Y}'\mathbf{Y}\mathbf{a} \tag{6.37}$$

$\bar{z}$ は平均。$\mathbf{Y}'\mathbf{Y}$ は (6.36) より、その要素が標準化してあることから、z についての相関行列である。これを $\mathbf{R}$ とおけば

$$\mathbf{Y}'\mathbf{Y} = n\mathbf{R} \tag{6.38}$$

したがって、(6.37) より

$$\mathbf{T} = n\mathbf{a}'\mathbf{R}\mathbf{a} \tag{6.39}$$

となる。

級間変動は z の G_1,G_2 における平均を $\bar{z}_1, \bar{z}_2$ と表すと、

$$W_1 = \sum_{i=1}^{n_1} \bar{z}_1^2 = n_1 \bar{z}_1^2 \tag{6.40}$$

$$W_2 = \sum_{i=n_1+1}^{n_2} \bar{z}_2^2 = n_2 \bar{z}_2^2 \tag{6.41}$$

従って、

$$W_1 + W_2 = n_1 \bar{z}_1^2 + n_2 \bar{z}_2^2 \tag{6.42}$$

ここで、(6.33) より

$$\bar{z}_1 = \sum_{k=1}^{m} a_k \bar{y}_{k_1} = \mathbf{a}' \bar{\mathbf{y}}_1 \tag{6.43}$$

$$\bar{z}_2 = \sum_{k=1}^{m} a_k \bar{y}_{k_2} = \mathbf{a}' \bar{\mathbf{y}}_2 \tag{6.44}$$

ただし、$\bar{y}_{k_1}$ は $\mathbf{y}_k$ の G_1 における平均、$\bar{y}_{k_2}$ は G_2 における平均を表す。
(6.42) より

$$\begin{aligned} W_1 + W_2 &= n_1(\mathbf{a}'\bar{\mathbf{y}}_1)'(\mathbf{a}'\bar{\mathbf{y}}_1) + n_2(\mathbf{a}'\bar{\mathbf{y}}_2)'(\mathbf{a}'\bar{\mathbf{y}}_2) \\ &= n_1 \bar{\mathbf{y}}_1' \mathbf{a}\mathbf{a}' \bar{\mathbf{y}}_1 + n_2 \bar{\mathbf{y}}_2' \mathbf{a}\mathbf{a}' \bar{\mathbf{y}}_2 \\ &= \mathbf{a}'(n_1 \bar{\mathbf{y}}_1 \bar{\mathbf{y}}_1' + n_2 \bar{\mathbf{y}}_2 \bar{\mathbf{y}}_2')\mathbf{a} \end{aligned} \tag{6.45}$$

ここで

$$P = (n_1 \bar{\mathbf{y}}_1 \bar{\mathbf{y}}_1' + n_2 \bar{\mathbf{y}}_2 \bar{\mathbf{y}}_2') \tag{6.46}$$

と定義すれば、

$$W_1 + W_2 = n\mathbf{a}'\mathbf{P}\mathbf{a} \tag{6.47}$$

相関比 η^2 は (6.29) により次のように表せる。

$$\eta^2 = \mathbf{a}'\mathbf{P}\mathbf{a} / (\mathbf{a}'\mathbf{R}\mathbf{a}) \tag{6.48}$$

次に、η^2 を最大化するために、ラグランジェ乗数法を用いて関数 L を定義する。

$$L = \mathbf{a}'\mathbf{P}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{R}\mathbf{a} - C) \tag{6.49}$$

C は $\mathbf{a}'\mathbf{R}\mathbf{a} = C$ という条件のもとに η^2 を最大化するという制約条件の定数。未定乗数 $\mathbf{a}$ で L を偏微分すると、

$$\partial L / \partial \mathbf{a} = 2\mathbf{P}\mathbf{a} - 2\lambda\mathbf{R}\mathbf{a} = 0 \tag{6.50}$$

$$\Leftrightarrow \mathbf{R}^{-1}\mathbf{P}\mathbf{a} = \lambda\mathbf{a} \tag{6.51}$$

次に (6.51) について固有値 λ を求め、固有ベクトル $\mathbf{a}$ を求める。また、(6.48) より

$$\eta^2 = \mathbf{a}'\mathbf{P}\mathbf{a} / (\mathbf{a}'\mathbf{R}\mathbf{a}) = \mathbf{a}'\lambda\mathbf{R}\mathbf{a} / (\mathbf{a}'\mathbf{R}\mathbf{a}) = \lambda \tag{6.52}$$

すなわち相関比は固有値に等しい。

固有ベクトル $\mathbf{a}$ を用いて、次の判別関数を構成する。

$$z = a_1\mathbf{y}_1 + a_2\mathbf{y}_2 + \cdots + a_m\mathbf{y}_m = \sum_{i=1}^{m} a_i\mathbf{y}_i \tag{6.53}$$

ある標本 X について、G_1,G_2 いずれかの判別を行う場合は X の各変量 $\mathbf{x}_j(j = 1, 2, \cdots, m)$ について標準化し、標準化された $\mathbf{y}_j(j = 1, 2, \cdots, m)$ について (6.53) に基づいて z を構成する。

$$
\begin{aligned}
&| \mathbf{z} - \bar{x}_1 | \geq | \mathbf{z} - \bar{x}_2 | \text{ ならば、} X \text{ は } G_2 \text{に属する。} \\
&| \mathbf{z} - \bar{x}_1 | < | \mathbf{z} - \bar{x}_2 | \text{ ならば、} X \text{ は } G_1 \text{に属する。}
\end{aligned}
$$

（２）多群判別について

相関比の方法は２群判別を多群判別に拡張することができる。今、l 群に判別する場合、群を $G_k(k = 1, 2, \cdots, l)$ とする。各群における標本平均を $\bar{x}^{(k)}$ とすると、全体変動は次のように記述できる。

$$
\begin{aligned}
T &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{k=1}^{l} \sum_{i=1}^{nk} (x_i - x^{(k)})^2 \\
&= \sum_{k=1}^{l} \sum_{i=1}^{n_k} (x_i - x^{(k)} + x^{(k)} - \bar{x})^2 = \sum_{k=1}^{l} W_k + \sum_{k=1}^{l} B_k
\end{aligned} \tag{6.54}
$$

ここで

$$
W_k = \sum_{k=1}^{l} (x_i - \bar{x}^{(k)})^2 \tag{6.55}
$$

$$
B_k = \sum_{i=1}^{n_k} (\bar{x}^{(k)} - \bar{x})^2 \tag{6.56}
$$

W_k, B_k は G_k の級内変動と級間変動である。相関比は次のように表せる。

$$
\eta^2 = \sum_{k=1}^{l} B_k \diagup T = \mathbf{a}'\mathbf{P}\mathbf{a} \diagup (\mathbf{a}'\mathbf{R}\mathbf{a}) \tag{6.57}
$$

η^2 を最大化するためにラグランジェ乗数法を用い、固有値とその固有ベクトルを求める。標本 X について標準化されたその変量 $y_j(j = 1, 2, \cdots, m)$ について、(6.53) より z の値を求め、$min\left\{| x - \bar{x}^{(k)} |\right\}$ とする G_k を X が属する群と決める。

（課題 8）人口増加率、老齢人口割合、一人当たり県民所得、一人当たり商店販売額、消費者物価指数の５つの変量を用いて標本を過疎的地域（1）と都会的地域（0）に判別せよ。

（注）北海道は人口密度は低いが、他の変量の動きによって都会的地域と判別される傾向があるが、これは北海道の特異性を表している。