多変量解析における変量の冗長性

藤 越 康 祝*

REDUNDANCY OF SOME VARIABLES IN MULTIVARIATE ANALYSIS

Yasunori FUJIKOSHI

The purpose of this paper is to review the problem of evaluating whether a given set of variables is redundant, in the presence of the rest variables. The redundancy problem is mainly discussed in discriminant analysis, canonical correlation analysis and principal component analysis. We will give some equivalent formulations for redundancy of a variable subset, which are appropriate for intuitive interpretation, inferential structure and mathematical treatment, respectively. The LR test and the AIC criterion for redundancy of a variable subset are also given. We give a brief discussion for the redundancy problem in other multivariate models and some topics related to the redundancy problem.

1. 序

多変量解析において、考慮されている変量の中に冗長の変量が含まれているかどうかを検証する問題を考える。この問題は、新しい変量の追加によって統計解析の目的に沿った付加情報が得られるかどうかを検証する問題でもある。冗長の変量を除くことの重要性は、推測の効率の向上とデータ収集に関する負担の軽減が得られ、さらには、データ解釈の単純化が期待されることにある。実際のデータ解析においては、現象解析の観点からの変量の重要性を考慮する必要があるが、ここでは、この点は考慮せず統計的情報の観点からの冗長性を問題にする。変量の冗長性問題は、変量選択問題とも密接に関係している。実際、変量の冗長性を定式化し、この定式化に基づくモデルを導入し、さらに、モデル選択基準を適用することによって、一つの変量選択法が提案される。

Rao (1946, 1948, 1970) は判別分析において、変量の冗長性の定式化を与え、その検定法を提案した。彼の研究以来、判別分析を含めた多くの分野でこの方面の研究がなされている (例えば, Fujikoshi (1985a, 1989), Siotani, Hayakawa & Fujikoshi (1985) の第 13 章)。

本論文では、主として判別分析、正準相関分析、主成分分析における変量の冗長性について解説する。また、その他のモデルにおける同様な研究や関連する研究についても触れる。

各モデルにおける冗長性問題は、次の方針に従って解説される。まず、 p 次元変量 $X = (X_1, X_2)'$ における部分変量 X_2 の冗長性、または、 $X_1: q \times 1$ の十分性の定式化を問題にする。この場合、統計解析の目的に沿う定式化が重要となる。 X_2 の冗長性の定式化によって導入されるモデルまたは仮説を M_q と表す。仮説 M_q に対して、いくつかの同値な仮説が与えられる。これらは、解釈、推測、数学的扱い、のそれぞれにおいて好都合なものである。次に、冗長性仮説 M_q 検定問題を考える。主として尤度比基準

$$\lambda = L(\hat{\theta}_p)/L(\hat{\theta}_q)$$

を与える。ここに、 $L(\theta)$ は大きさ N の標本に基づく尤度関数、 $\hat{\theta}_p, \hat{\theta}_q$ はそれぞれの M_p (制限のないモデル)、 M_q のもとでのパラメータ θ の最尤推定量である。特定な部分変量の冗長性だけでなく、種々の部分変量の冗長性に関心がある場合がしばしば生ずる。この場合は、解析の目的に沿った最良な部分変量を選ぶ変量選択問題となる。仮説 M_q の選択と対応する変量の選択を同一視し、さらに、モデル選択基準を適用す

* 広島大学理学部数学教室 (Dept. of Mathematics, Faculty of Science, Hiroshima university)

ることにより、一つの变量選択法が提案される。種々のモデル選択基準が提案されているが、ここでは尤度比基準と密接に関連している AIC 基準 (Akaike (1973))

$$AIC_q = -2 \log L(\hat{\Theta}_q) + 2m_q$$

を用いたときの方法を具体的に与える。ここに、 m_q は M_q のもとでの Θ の次元である。AIC_q が最小になる仮説 M_q 、または、対応する変量が合理的であるとされる。

2. 判別分析

2.1. 冗長性と付加情報

p 次元変数 $X = (X_1, \dots, X_p)'$ に基づく k 個の母集団 $\Pi_1, \dots, \Pi_k$ に関する判別問題を考える。判別問題には、新しい個体がどの母集団に属するかを判定することを目的にする場合と、母集団間の差異をできるだけ少数個の判別関数で説明することを目的にする場合がある。後者の判別分析は、正準判別分析ともよばれ、通常、各母集団での共分散行列が等しいことを前提にしている。これらの判別の目的に対して、Rao (1970) は変量の冗長性を次のように定義している。

定義 2.1. p 次元変数 X を $X = (X_1', X_2')'$, $X_1: q \times 1$ と分割する。 X_1 を与えたときの X_2 の条件つき分布が k 個の母集団において同一であるとき、 $X_2(X_1)$ は k 個の母集団間の判別情報に関して冗長 (十分) であるという。

X_2 の冗長性は、 X_1 に対して X_2 を追加しても判別に関する付加情報が得られないことを意味するものである。このことを、より明白にするため、まず、分布間の距離との関連をみることにする。 Π_i における X の pdf を $f_i(x)$ とするとき、 Π_i と Π_j の間の距離の代表的基準である Kullback 情報量は

$$d_p(i, j; x) = \int_{R^p} [f_i(x) - f_j(x)] \log \left(\frac{f_j(x)}{f_i(x)} \right) dx$$

と表される。これらの情報量の 1 次元化尺度として

$$d_p(X) = \sum_{1 \leq i < j \leq k} d_p(i, j; X)$$

を考える。このとき

$$d_p(i, j; X) = d_q(i, j; X_1) + d_{p-q}(i, j; X_2 | X_1)$$

が成り立つことから、次の結果が得られる。

定理 2.1. 次の 3 つの条件は同値である。

- (1) X の部分変数 X_2 が k 個の母集団間の判別情報に関して冗長である。
- (2) $d_p(i, j; X) = d_q(i, j; X_1) (1 \leq i < j \leq k)$,
- (3) $d_p(X) = d_q(X_1)$.

(1), (2) の同値性は Pynnönen (1988) によって示された。(2), (3) の同値性は、 $d_p(i, j; X) \geq d_q(i, j; X_1) (1 \leq i < j \leq k)$ に注意すると得られる。

定義 2.1 の冗長性は判別関数や判別法によっても特徴づけられる。正準判別関数を用いた特徴づけは第 2.3 節で与えることにし、ここでは、判別法との関係について述べる。 p 次元空間 R^p を互に素な k 個の領域 $R_1, \dots, R_k$ に分割し、 $x \in R_i \Rightarrow x \in \Pi_i$ と判定する判別方式を $R = \{R_1, \dots, R_k\}$ と表す。 Π_i に属する個体を Π_j に属すると判別することによる損失を $c(j|i)$ とする。ここに、 $c(j|i) = 0 (i=j), > 0 (i \neq j)$ 。また、母集団 Π_i の先験確率を $p_i (> 0)$ とする。このとき期待損失を最小にする判別方式 $R^* = \{R_1^*, \dots, R_k^*\}$ は

$$R_i^* = \{x | h^{(i)}(x) \leq h^{(j)}(x), j=1, \dots, k, j \neq i\}$$

で与えられる (例えば、Anderson (1984))。ここに

$$h^{(i)}(x) = \sum_{j=1, j \neq i}^k p_j f_j(x) c(i|j).$$

判別方式 R^* はベイズ判別方式とよばれる。次の結果が示されている (Pynnönen (1988))。

定理 2.2. p 次元変数に基づく k 個の母集団の判別におけるベイズ判別法を $R^* = \{R_1^*, \dots, R_k^*\}$ とする。また、部分変数 X_1 だけにに基づくベイズ判別法を $\tilde{R} = \{\tilde{R}_1, \dots, \tilde{R}_k\}$ とする。すなわち

$$\tilde{R}_i = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^p | h^{(i)}(x_1) \leq h^{(j)}(x_1) \right. \\ \left. j=1, \dots, k, j \neq i \right\}$$

ここに、 $h^{(i)}(x_1) = \sum_{j=1, j \neq i}^k p_j f_{j1}(x_1) c(i|j)$, $f_{i1}(x_1)$ は Π_i における X_1 の pdf である。このとき、 X_2 が k 個の母集団間の判別情報に関して冗長であるならば、2 つベイズ判別方式 R^* と $\tilde{R}$ は同一である。

損失 $c(i|j) (i \neq j)$ が一定である場合には、適当な正則条件のもとで、定理 2.2 の逆も成り立つことが示される。

この節では、次の記号を用いる。

$$E(X | \Pi_i) = \mu^{(i)} = \begin{bmatrix} \mu_1^{(i)} \\ \mu_2^{(i)} \end{bmatrix},$$

$$\text{Var}(X|\Pi_i) = \Sigma^{(i)} = \begin{bmatrix} \Sigma_{11}^{(i)} & \Sigma_{12}^{(i)} \\ \Sigma_{21}^{(i)} & \Sigma_{22}^{(i)} \end{bmatrix},$$

$$i=1, \dots, k.$$

Π_i からの大きさ N_i の標本を $\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{N_i}^{(i)}$ とし

$$W^{(i)} = \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})' = \begin{bmatrix} W_{11}^{(i)} & W_{12}^{(i)} \\ W_{21}^{(i)} & W_{22}^{(i)} \end{bmatrix}$$

$$W = W^{(1)} + \dots + W^{(k)} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix},$$

$$T = \sum_{i=1}^k \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}})' = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

とおく. ここに, $\bar{\mathbf{x}}^{(i)} = (1/N_i) \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}$, $\bar{\mathbf{x}} = (1/N) \sum_{i=1}^k N_i \bar{\mathbf{x}}^{(i)}$, $N = N_1 + \dots + N_k$ である.

2.2. $k=2$ & $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$

正規性を仮定すると, X_2 の冗長性条件は

$$E(X_2|X_1; \Pi_1) = E(X_2|X_1; \Pi_2)$$

と同等であり, これは

$$M_q; \mu_2^{(1)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(1)} = \mu_2^{(2)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(2)}$$

と表わされる. また, Kullback 情報量は

$$d_p(X) = d_p(1, 2; X)$$

$$= (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) = \Delta_p^2$$

であり, マハラノビスの距離の2乗に等しい.

ベイズ判別方式 $R^* = \{R_1^*, R_2^*\}$ は, 線形判別関数

$$U = \gamma' [X - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})]$$

と境界点 $c = \log [p_2 c(1|2) / \{p_1 c(2|1)\}]$ を用いて

$$R_1^*: U \geq c, R_2^*: U < c$$

と表わせる. ただし, $\gamma = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$.

線形判別関数 U を用いた判別は, 正規性の仮定とは無関係に Fisher (1936) によって提案された. 線形判別関数の係数ベクトル γ を

$$\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \begin{matrix} q \\ p-q \end{matrix}$$

と分割する. X_2 の冗長性は " $\gamma_2=0$ " と表わされる. これらの結果は次の定理としてまとめられる.

定理 2.3. 次の4つの条件は同値である.

- (1) $\Delta_p^2 = \Delta_q^2$,
- (2) $\gamma_2 = 0$,

$$(3) M_q: \mu_2^{(1)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(1)} = \mu_2^{(2)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(2)},$$

(4) X が p 次元正規分布に従うとき, X_2 が判別情報に関して冗長である.

定理 2.3 は Rao (1970) によって与えられた. 条件 (1)~(3) は, 正規性の仮定とは無関係に X_2 の冗長性の定式化として用いることができる.

正規性を仮定すれば, X_2 の冗長性仮説 M_q に対する尤度比統計量が求められる. X および X_1 に基づく標本マハラノビスの距離をそれぞれ D_p, D_q とする. すなわち

$$D_p^2 = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})' S^{-1} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}),$$

$$D_q^2 = (\bar{\mathbf{x}}_1^{(1)} - \bar{\mathbf{x}}_1^{(2)})' S_{11}^{-1} (\bar{\mathbf{x}}_1^{(1)} - \bar{\mathbf{x}}_1^{(2)}),$$

ただし, $n = N_1 + N_2 - 2$ とおくと

$$S = \frac{1}{n} W = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

定理 2.4. p 次元変量 X は, $X|\Pi_i \sim N_p(\mu^{(i)}, \Sigma)$, $i=1, 2$ とする.

(1) X_2 の冗長性仮説 M_q に対する尤度比検定の棄却域は

$$F = \frac{N-p-1}{p-q} \cdot \frac{D_p^2 - D_q^2}{n(N_1^{-1} + N_2^{-1}) + D_p^2} > c$$

で与えられる.

(2) 仮説 M_q のもとで, F は自由度 $p-q, N-p-1$ の F -分布に従う.

検定統計量 F は Rao (1948) によって提案された. 仮説 M_q に対する AIC 基準を AIC_q とすると

$$A_q = AIC_q - AIC_p$$

$$= N \log \left[1 + \frac{D_p^2 - D_q^2}{n(N_1^{-1} + N_2^{-1}) + D_q^2} \right] - 2(p-q)$$

となる. この基準量を用いた変量選択法は, 誤判別率に基づく変量選択法 (McLachlan (1980)) と漸近的に同等である (Fujikoshi (1985b)).

2.3. $\Sigma^{(1)} = \dots = \Sigma^{(k)} = \Sigma$ の場合

まず, X について正規性を仮定する. 定義 2.1 から X_2 の冗長性条件は

$$E(X_2|X_1; \Pi_1) = \dots = E(X_2|X_1; \Pi_k)$$

と同値であって

$$M_q; \mu_2^{(1)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(1)} = \dots = \mu_2^{(k)} - \Sigma_{21} \Sigma_{11}^{-1} \mu_1^{(k)}$$

と表わされる。また、 Π_i と Π_j の間の Kullback 情報量は $d_p(i, j; \mathbf{X}) = (\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)})$ である。従って、 k 個の母集団に関する Kullback 情報量は、 $\bar{\boldsymbol{\mu}} = (1/k)(\boldsymbol{\mu}^{(1)} + \dots + \boldsymbol{\mu}^{(k)})$ とおくと

$$\begin{aligned} d_p(\mathbf{X}) &= \sum_{1 \leq i < j \leq k} (\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)}) \\ &= \text{tr} \left\{ [\boldsymbol{\mu}^{(1)} - \bar{\boldsymbol{\mu}}, \dots, \boldsymbol{\mu}^{(k)} - \bar{\boldsymbol{\mu}}] \right. \\ &\quad \times [\boldsymbol{\mu}^{(1)} - \bar{\boldsymbol{\mu}}, \dots, \boldsymbol{\mu}^{(k)} - \bar{\boldsymbol{\mu}}]' \boldsymbol{\Sigma}^{-1} \left. \right\} \end{aligned}$$

となる。

次に、変量の冗長性の定式化を次元縮少を伴う判別関数、すなわち、正準判別変量を用いて行うを試みる。この判別関数は Fisher (1936) によって導入されたものであって、群間変動行列と群内変動行列に基づいている。母集団における群間変動行列は

$$\mathcal{Q} = \sum_{i=1}^k g_i (\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})'$$

として定義される。ここに、 $\{g_i\}$ はウェイトの集合であって、 $g_i > 0 (i=1, \dots, k)$ かつ $\sum_{i=1}^k g_i = 1$ をみたし、 $\bar{\boldsymbol{\mu}} = \sum_{i=1}^k g_i \boldsymbol{\mu}^{(i)}$ である。このとき、正準判別変量 $\boldsymbol{\gamma}_j' \mathbf{X}$ の係数ベクトル $\boldsymbol{\gamma}_j$ は

$$\mathcal{Q} \boldsymbol{\gamma}_j = \lambda_j \boldsymbol{\Sigma} \boldsymbol{\gamma}_j, \quad \boldsymbol{\gamma}_i' \boldsymbol{\Sigma} \boldsymbol{\gamma}_j = \delta_{ij}$$

の解として与えられる。ここに、 $\lambda_1 \geq \dots \geq \lambda_m > \lambda_{m+1} = \dots = \lambda_p = 0$ は $\boldsymbol{\Sigma}^{-1} \mathcal{Q}$ の固有根であって $-m$ は $\text{rank}(\mathcal{Q})$ に等しい。正準判別変量として意味のあるものは最初の m 個であって、残りの正準判別変量は判別解析の目的に対して意味のないものである。 $\mathbf{X}$ の分割に対応して

$$\boldsymbol{\gamma}_j = \begin{bmatrix} \boldsymbol{\gamma}_{1j} \\ \boldsymbol{\gamma}_{2j} \end{bmatrix}, \quad \mathcal{Q} = \begin{bmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22} \end{bmatrix}$$

と分割する。定理 2.4 は次のように一般化される (Fujikoshi (1982))。

定理 2.5. 次の 4 つの条件は同値である。

- (1) $\text{tr} \mathcal{Q} \boldsymbol{\Sigma}^{-1} = \text{tr} \mathcal{Q}_{11} \boldsymbol{\Sigma}_{11}^{-1}$,
- (2) $\boldsymbol{\gamma}_{2j} = \mathbf{0} \quad (j=1, \dots, m = \text{rank}(\mathcal{Q}))$,
- (3) $M_q: \boldsymbol{\mu}_2^{(1)} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1^{(1)} = \dots = \boldsymbol{\mu}_2^{(k)} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1^{(k)}$,
- (4) $\mathbf{X}$ が p 次元正規分布に従うとき、 $\mathbf{X}_2$ が判別情報に関して冗長である。

定理 2.5 (1) における $\text{tr} \mathcal{Q} \boldsymbol{\Sigma}^{-1}$ は、正規性の仮定の

もとにおける Kullback 情報量と密接に関連している。すなわち、 $g_i = 1/k (i=1, \dots, k)$ のとき

$$d_p(\mathbf{X}) = \text{tr} \mathcal{Q} \boldsymbol{\Sigma}^{-1}$$

が成り立つ。

定理 2.6. p 次元変数 $\mathbf{X}$ は、 $\mathbf{X} | \Pi_i \sim N_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i=1, \dots, k$ とする。

(1) $\mathbf{X}_2$ の冗長性仮説 M_q に対する尤度比基準 λ は

$$\begin{aligned} \lambda^{2/N} &= \frac{|W|}{|W_{11}|} \bigg/ \frac{|T|}{|T_{11}|} \\ &= |W_{22.1}| / |T_{22.1}| \end{aligned}$$

与えられる。

(2) 仮説 M_q のもとで、 $\wedge$ はラムダ分布 $\wedge_{p-q, k-1, N-k-q}$ に従う。

定理 2.6 の尤度比統計量は Rao (1948) によって与えられた。一般に、統計量 U がラムダ分布 $\wedge_{p, g, n}$ に従うとき

$$\begin{aligned} P(-m \log U \leq x) &= P(\chi_{pq}^2 \leq x) \\ &+ \frac{pq}{48m^2} (p^2 + q^2 - 5) \left[P(\chi_{pq+4}^2 \leq x) - P(\chi_{pq}^2 \leq x) \right] \\ &+ O(m^{-4}) \end{aligned}$$

が成り立つ (例えば, Anderson (1984))。ここに、 $m = n - \frac{1}{2}(p+1-q)$ 。仮説 M_q に対する AIC 基準を AIC_q とおくと

$$\begin{aligned} A_q &= AIC_q - AIC_p \\ &= -N \log \{ |W_{22.1}| / |T_{22.1}| \} - 2(p-q)(k-1) \end{aligned}$$

となる。変量選択基準としての A_q の漸近的挙動については、Fujikoshi (1983) を参照されたい。

2.4. 共分散行列が異なる場合

p 次元変数 $\mathbf{X}$ は、 $\mathbf{X} | \Pi_i \sim N_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}^{(i)})$, $i=1, \dots, k$ に従うものとする。このとき、 $\mathbf{X}_2$ の冗長性条件は

$$\begin{aligned} E(\mathbf{X}_2 | \mathbf{X}_1; \Pi_i) &= E(\mathbf{X}_2 | \mathbf{X}_1; \Pi_j) \\ &\quad (i, j=1, \dots, k) \end{aligned}$$

$$\text{Var}(\mathbf{X}_2 | \mathbf{X}_1; \Pi_i) = \text{Var}(\mathbf{X}_2 | \mathbf{X}_1; \Pi_j)$$

と同値である。この条件は次のように表わせる (Pynnönen (1986))。

定理 2.7. $\mathbf{X} | \Pi_i \sim N_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}^{(i)})$, $i=1, \dots, k$ とし、 $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$, $\mathbf{X}_1: q \times 1$ とする。このとき、 $\mathbf{X}_2$ が判別情報に関して冗長であるための必要十分条件は

$$\begin{aligned}\mu_2^{(1)} - \Sigma_{21}^{(1)} \Sigma_{11}^{(1)-1} \mu_1^{(1)} &= \dots = \mu_2^{(k)} - \Sigma_{21}^{(k)} \Sigma_{11}^{(k)-1} \mu_1^{(k)}, \\ M_q: \Sigma_{21}^{(1)} \Sigma_{11}^{(1)-1} &= \dots = \Sigma_{21}^{(k)} \Sigma_{11}^{(k)-1}, \\ \Sigma_{22,1}^{(1)} &= \dots = \Sigma_{22,1}^{(k)}\end{aligned}$$

が成り立つことである。

Π_i と Π_j の間の Kullback 情報量は

$$\begin{aligned}d_p(i, j; X) &= \frac{1}{2} (\mu^{(i)} - \mu^{(j)})' (\Sigma^{(i)-1} + \Sigma^{(j)-1}) \\ &\quad (\mu^{(i)} - \mu^{(j)}) + \frac{1}{2} \text{tr}(\Sigma^{(i)} \Sigma^{(j)-1} - 2I_p)\end{aligned}$$

で与えられる。また、損失 $c(i|j)(i \neq j)$ が一定である場合のベイズ判別法は

$$x \in \Pi_i \Leftrightarrow Q_p(i, j; x) \geq c_{ij} \quad (j=1, \dots, k, j \neq i)$$

である。ここに、 $c_{ij} = \log(p_j/p_i)$,

$$\begin{aligned}Q_p(i, j; x) &= \frac{1}{2} (x - \mu^{(i)})' \Sigma^{(i)-1} (x - \mu^{(j)}) \\ &\quad - \frac{1}{2} (x - \mu^{(j)})' \Sigma^{(j)-1} (x - \mu^{(j)}) \\ &\quad + \frac{1}{2} \log \frac{|\Sigma^{(j)}|}{|\Sigma^{(i)}|}\end{aligned}$$

定理 2.8. $X|\Pi_i \sim N_p(\mu^{(i)}, \Sigma^{(i)})$, $i=1, \dots, k$ とする。

(1) X_2 の冗長性仮説 M_q に対する尤度比基準は

$$\lambda = \frac{\prod_{i=1}^k |W_{22,1}^{(i)}|^{N_i/2}}{|T_{22,1}|^{N/2}} \cdot \frac{N^{N(p-q)/2}}{\prod_{i=1}^k N_i^{N_i(p-q)/2}}$$

で与えられる。

(2) 仮説 M_q のもとで、 $-2\log\lambda$ は $N_i (i=1, \dots, k)$ が大のとき漸近的に自由度 $\frac{1}{2}(k-1)(p-q)(p+q+3)$ の χ^2 -分布に従う。

上の定理は Pynnönen (1988) によって与えられた。冗長性仮説 M_q に対する AIC 基準を AIC_q とすると

$$\begin{aligned}A_q &= AIC_q - AIC_p \\ &= -2 \log \lambda - (k-1)(p-q)(p+q+3)\end{aligned}$$

である。

3. 正準相関分析

3.1. 冗長性と付加情報

2組の変量 $X_u: p \times 1$, $X_v: q \times 1$ の間の線形関係は Hotelling (1936) によって導入された正準相関係数と正準相関変量によって説明される。 X_u, X_v の各々についての冗長性を定式化するため、次の分割を考える。

$$X_u = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad X_v = \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} \begin{matrix} q_1 \\ q_2 \end{matrix}$$

これに対応して、 $X = (X_u', X_v')'$ の平均 μ , 共分散行列 Σ を次のように分割する。

$$\begin{aligned}\mu &= \begin{bmatrix} \mu_u \\ \mu_v \end{bmatrix}, \quad \mu_u = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mu_v = \begin{bmatrix} \mu_3 \\ \mu_4 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{bmatrix}, \quad \Sigma_{uu} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\ \Sigma_{uv} &= \begin{bmatrix} \Sigma_{13} & \Sigma_{14} \\ \Sigma_{23} & \Sigma_{24} \end{bmatrix}, \quad \Sigma_{vv} = \begin{bmatrix} \Sigma_{33} & \Sigma_{34} \\ \Sigma_{43} & \Sigma_{44} \end{bmatrix}\end{aligned}$$

X_1 と X_3 との線形関係において、 X_1, X_3 にそれぞれ X_2, X_4 が追加されたとき、付加情報が得られるかどうかを問題にする。 X_u と X_v との間の第 j 正準相関係数 ρ_j , 第 j 正準相関変量 $a'_{uj}X_u, a'_{vj}X_v$ の係数ベクトルは、固有方程式

$$\begin{aligned}\Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu} a_{uj} &= \rho_j^2 \Sigma_{uu} a_{uj}, \quad a'_{uj} \Sigma_{uu} a_{uj} = 1 \\ \Sigma_{vu} \Sigma_{uu}^{-1} \Sigma_{uv} a_{vj} &= \rho_j^2 \Sigma_{vv} a_{vj}, \quad a'_{vj} \Sigma_{vv} a_{vj} = 1\end{aligned}$$

の解として定義される。ここに、 $\rho_1 \geq \dots \geq \rho_m > \rho_{m+1} = \dots = \rho_{\max(p,q)} = 0$, ρ_j^2 は $\Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu} \Sigma_{uu}^{-1}$ または $\Sigma_{vu} \Sigma_{uu}^{-1} \Sigma_{uv} \Sigma_{vv}^{-1}$ の固有根である。ゼロでない正準相関係数の個数 m は $\text{rank}(\Sigma_{uv})$ に等しく、正準相関の次元とよばれる。係数ベクトル a_{uj}, a_{vj} についても

$$a_{uj} = \begin{bmatrix} a_{1j} \\ a_{2j} \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad a_{vj} = \begin{bmatrix} a_{3j} \\ a_{4j} \end{bmatrix} \begin{matrix} q_1 \\ q_2 \end{matrix}$$

のように分割する。 X_1 と X_3 との間の第 j 正準相関係数を ρ_j^* とし、 $\rho_1^* \geq \dots \geq \rho_{m^*}^* > \rho_{m^*+1}^* = \dots = \rho_{\max(p_1, p_3)}^* = 0$ とする。このとき、 $m^* = \text{rank}(\Sigma_{12})$ である。

定理 3.1. X_u と X_v との正準相関係数を $\rho_1 \geq \dots \geq \rho_m > \rho_{m+1} = \dots = 0$, X_1 と X_3 との正準相関係数を $\rho_1^* \geq \dots \geq \rho_{m^*}^* > \rho_{m^*+1}^* = \dots = 0$ とする。このとき、 $m \geq m^*$ であって

$$\rho_i \geq \rho_i^* \quad (i=1, \dots, m^*)$$

が成り立つ。

定理 3.1 より

$$\begin{aligned}\sum_{i=1}^m \rho_i^2 &\geq \sum_{i=1}^{m^*} \rho_i^{*2}, \\ \prod_{i=1}^m (1 - \rho_i^2) &\leq \prod_{i=1}^{m^*} (1 - \rho_i^{*2})\end{aligned}$$

が成り立つ。Laha (1954), Siotani (1957) は、一方の変量の組に新しい変量が追加された場合に上の特別な

結果を示した。この場合の定理 3.1 は Chen (1971), Mckay (1977a) によって示された。定理 3.1 は Chen (1971) の結果を 2 回用いることによって示されるが、直接的にも証明できる (Fujikoshi (1982))。

定義 3.1. X_u と X_v との間の正準相関分析において、 X_2 と X_4 (X_1 と X_3) が正準相関情報に関して冗長 (十分) であるとは

$$m = m^*, \rho_i = \rho_i^* (i=1, \dots, m^*)$$

が成り立つことである。

この定義は個々の正準相関係数の増加の有無に基づいており、冗長性の意味は明白であるが数学的に取り扱いにくいと云う欠点がある。このため、正準相関係数のある種の関数を正準相関情報基準として用いることを考える。一般に、 X_u と X_v の正準相関基準として

$$\Delta_{p,q}(X_u, X_v; g) = g(\rho_1, \dots, \rho_k)$$

を考える。ここに、 $k = \min(p, q)$, $g(x_1, \dots, x_k)$ は、 $x_j \geq 0 (j=1, \dots, k)$ において定義され

- (i) $g(x_1, \dots, x_k) \geq 0$,
- (ii) 各成分について単調増加である

をみたすとする。

定理 3.2. X_u と X_v の正準相関分析において、 X_2 と X_4 が正準相関情報に関して冗長である (すなわち、 X_1 と X_3 が正準相関情報に関して十分である) ための必要十分条件は

$$\Delta_{p,q}(X_u, X_v; g) = \Delta_{p_1, q_1}(X_1, X_3; g)$$

が成り立つことである。

定理 3.2 における冗長性条件は、基準関数 g の選び方に依存しないものであることを注意しておく。従って、以下では取り扱いやすい基準関数として、 $g(x_1, \dots, x_k) = \{x_1^2 + \dots + x_k^2\}^{1/2}$ を用い、 $\Delta_{p,q}(X_u, X_v; \{x_1^2 + \dots + x_k^2\}^{1/2})$ を単に $\Delta_{p,q}(X_u, X_v)$ と書く。このとき

$$\Delta_{p,q}^2(X_u, X_v) = \rho_1^2 + \dots + \rho_k^2 = \text{tr} \Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu} \Sigma_{uu}^{-1}$$

である。冗長性に関する推測問題は、追加される変量が一方の変量の組に対してであるか、または、両方の組に対してであるかによって、かなり異なっている。このため、以下では、それぞれ別個に扱う。(p+q) 次元変量 $X = (X_u, X_v)'$ についての大きさ $N = n+1$ の標本 $x_1, \dots, x_N$ に基づく標本共分散行列を S とする、すなわち

$$S = \frac{1}{n} \sum_{j=1}^N (x_j - \bar{x})(x_j - \bar{x})'$$

S についても、 Σ の場合と同様な分割を考える。また、記号 $\Sigma_{2v,1}$, $\Sigma_{24,13}$ 等を用いる。例えば

$$\begin{aligned} \Sigma_{2v,1} &= \Sigma_{2v} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{1v} \\ \Sigma_{24,13} &= \Sigma_{24} - [\Sigma_{21} \quad \Sigma_{23}] \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{14} \\ \Sigma_{34} \end{bmatrix} \end{aligned}$$

である。

3.2. X_2 の冗長性

X_u と X_v との正準相関分析において、 X_u の部分変量 X_2 の冗長性は、定理 3.2 より

$$\text{tr} \Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu} \Sigma_{uu}^{-1} = \text{tr} \Sigma_{1v} \Sigma_{vv}^{-1} \Sigma_{1v} \Sigma_{11}^{-1}$$

と同値である。左辺は $\text{tr} \Sigma_{1v} \Sigma_{vv}^{-1} \Sigma_{v1} \Sigma_{11}^{-1} + \text{tr} \Sigma_{2v,1} \Sigma_{vv}^{-1} \Sigma_{v2,1} \Sigma_{22,1}^{-1}$ と表わされる。従って、 X_2 の冗長性条件は “ $\Sigma_{2v,1} = O$ ” となる。この他、回帰係数や正準相関変量を用いた冗長性条件を与えることができる。 X_v の X_u への回帰における回帰係数行列を

$$\Gamma_{vu} = \Sigma_{vu} \Sigma_{uu}^{-1} = [\Gamma_{v1} \quad \Gamma_{v2}]$$

と分割する。また、 X_2 の $(X_1', X_v')'$ への回帰における回帰係数行列を

$$\begin{aligned} B_{2(1v)} &= [\Sigma_{21} \quad \Sigma_{2v}] \begin{bmatrix} \Sigma_{11} & \Sigma_{1v} \\ \Sigma_{v1} & \Sigma_{vv} \end{bmatrix}^{-1} \\ &= [B_{21} \quad B_{2v}] \end{aligned}$$

と分割する。このとき、次の結果が示される (Siotani (1957), Mckay (1977a), Fujikoshi (1982))。

定理 3.3. X_u と X_v との正準相関分析において、 X_u の部分変量 X_2 の冗長性条件は、次の 4 つの条件の各々と同値である。

- (1) $\Sigma_{2v,1} = O$,
- (2) $\Gamma_{v2} = O$,
- (3) $B_{2v} = O$,
- (4) $a_{2j} = 0 (j=1, \dots, m = \text{rank}(\Sigma_{uv}))$.

X_2 の冗長性仮説 (条件) を $M_{p,q}$ と表わす。正規性を仮定すると、定理 3.3 の (1), (2), (3) は、それぞれ、条件つきモデル

$$(i) \begin{bmatrix} X_2 \\ X_v \end{bmatrix} | X_1, (ii) X_v | X_u, (iii) X_2 | \begin{bmatrix} X_1 \\ X_v \end{bmatrix}$$

において好都合な表示である。この結果を利用するこ

とによって、尤度比基準を与えることができる。

定理 3.4. $(p+q)$ 次元変量 $X=(X'_u, X'_v)'$ は正規分布に従うとする。 $X_u=(X'_1, X'_2)'$, $X_1: p_1 \times 1$ とし, X_u と X_v との正準相関分析における X_2 の冗長性仮説を $M_{p,q}$ とする。

(1) $M_{p,q}$ に対する尤度基準を λ とすると, $\lambda^{2/N} = \Lambda$ は次のように表わされる。

$$\begin{aligned}\Lambda &= \left| \frac{S_{22.1} \quad S_{2v.1}}{S_{v2.1} \quad S_{vv.1}} \right| / \{|S_{22.1}| |S_{vv.1}|\} \\ &= |S_{vv.u}| / |S_{vv.1}| \\ &= |S_{22.1v}| / |S_{22.1}|.\end{aligned}$$

(2) 仮説 $M_{p,q}$ のもとで, Λ はラムダ分布 $\Lambda_{p-p_1, q, n-p_1-q}$ に従う。

尤度比統計 Λ に対する上記の 3 つの表示は, 条件つきモデル (i), (ii), (iii) での導出に対応したものになっている。仮説 $M_{p,q}$ に対する AIC 基準を $AIC_{p,q}$ とすると

$$\begin{aligned}A_{p,q} &= AIC_{p,q} - AIC_{p,q} \\ &= -N \log \Lambda - 2(p-p_1)q\end{aligned}$$

となる。

X_u と X_v との正準相関分析において, とくに, X_v が 1 次元変量 ($q=1$) の場合は, 説明変量がランダムな場合の重回帰分析であることを注意しておく。

3.3. X_2 と X_4 の冗長性

定理 3.2 より, X_2 と X_4 の冗長性条件は

$$\Delta_{p,q}^2(X_u, X_v) = \Delta_{p_1,q_1}^2(X_1, X_3)$$

すなわち

$$\text{tr} \Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu} \Sigma_{uu}^{-1} = \text{tr} \Sigma_{13} \Sigma_{33}^{-1} \Sigma_{31} \Sigma_{11}^{-1}$$

と同値である。後者の条件より, “ $\Sigma_{2v.1}=0, \Sigma_{4u.3}=0$ ” が得られる。 $(X'_2, X'_4)'$ の $(X'_1, X'_3)'$ への回帰における回帰係数行列を

$$\begin{bmatrix} B_{21} & B_{23} \\ B_{41} & B_{43} \end{bmatrix} = \begin{bmatrix} \Sigma_{21} & \Sigma_{23} \\ \Sigma_{41} & \Sigma_{43} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix}^{-1}$$

と置く。また, X_1 と X_3 を与えたときの X_2 と X_4 の偏共分散行列を

$$\begin{bmatrix} \Sigma_{22.13} & \Sigma_{24.13} \\ \Sigma_{42.13} & \Sigma_{44.13} \end{bmatrix} = \begin{bmatrix} \Sigma_{22} & \Sigma_{24} \\ \Sigma_{42} & \Sigma_{44} \end{bmatrix}$$

$$- \begin{bmatrix} \Sigma_{21} & \Sigma_{23} \\ \Sigma_{41} & \Sigma_{43} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{12} & \Sigma_{14} \\ \Sigma_{32} & \Sigma_{34} \end{bmatrix}$$

とする。このとき, 次の結果が得られる。

定理 3.5. $X_u=(X'_1, X'_2)'$ と $X_v=(X'_3, X'_4)'$ との正準相関分析において, 次の 3 つの条件の各々は, X_2 と X_4 が正準相関情報に関して冗長であるための必要十分条件である。

- (1) $\Sigma_{2v.1}=0, \Sigma_{4u.3}=0$ (または $\Sigma_{41.3}=0$),
- (2) $B_{23}=0, B_{41}=0, \Sigma_{24.13}=0$,
- (3) $\begin{bmatrix} a_{2j} \\ a_{4j} \end{bmatrix} = 0$ ($j=1, \dots, m=\text{rank}(\Sigma_{uv})$).

正規性を仮定する。 $(X'_1, X'_3)'$ が与えられたときの $(X'_2, X'_4)'$ の条件つき分布は, 正規分布であって, その平均, 分散行列は次のように表わされる。

$$\begin{aligned}E \left(\begin{bmatrix} X_2 \\ X_4 \end{bmatrix} \middle| \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \right) &= \begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix} + \begin{bmatrix} B_{21} & B_{23} \\ B_{41} & B_{43} \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_3 - \mu_3 \end{bmatrix}, \\ \text{Var} \left(\begin{bmatrix} X_2 \\ X_4 \end{bmatrix} \middle| \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \right) &= \begin{bmatrix} \Sigma_{22.13} & \Sigma_{24.13} \\ \Sigma_{42.13} & \Sigma_{44.13} \end{bmatrix}.\end{aligned}$$

定理 3.5 における冗長性条件 (2) は, 上記条件つきモデルに関して好都合な表示である。この条件つきモデルは 3.2 節における条件つきモデル (iii) の拡張である。他の条件つきモデル (i), (ii) の拡張では, 好都合な冗長性条件が得られないことを注意しておく。

定理 3.6. $(p+q)$ 次元変量 $X=(X'_u, X'_v)'$ は正規分布に従うものとする。 $X_u=(X'_1, X'_2)'$ と $X_v=(X'_3, X'_4)'$ との間の正準相関分析における X_2 と X_4 の冗長性仮説 (X_1 と X_3 の十分性仮説) を M_{p_1, q_1} とする。ここに, $X_1: p_1 \times 1, X_3: q_1 \times 1$ 。

(1) 仮説 M_{p_1, q_1} に対する尤度比基準を λ とすると

$$\lambda^{2/N} = \Lambda = \frac{\begin{vmatrix} \Sigma_{22.13} & \Sigma_{24.13} \\ \Sigma_{42.13} & \Sigma_{44.13} \end{vmatrix}}{\begin{vmatrix} \Sigma_{22.1} & \Sigma_{44.3} \end{vmatrix}} = \Lambda_{(1)} \Lambda_{(2)}$$

で与えられる。ここに

$$\begin{aligned}\Lambda_{(1)} &= |\Sigma_{22.13}| / |\Sigma_{22.1}|, \\ \Lambda_{(2)} &= |\Sigma_{44.123}| / |\Sigma_{44.3}|.\end{aligned}$$

(2) 仮説 M_{p_1, q_1} のもとで, $\Lambda_{(1)}, \Lambda_{(1)}, \Lambda_{(2)}$ は互に独立で, それぞれ, ラムダ分布 $\Lambda_{p-p_1, q_1, n-p_1-q_1}, \Lambda_{q-q_1, p, n-p-q_1}$ に従う。

定理 3.5. (2) より, $-N \log \Delta$ は N が大のとき漸近的に自由度 $pq - p_1q_1$ の χ^2 -分布に従う. 高次のオーダーまでの近似については, Fujikoshi (1982) を参照されたい. 仮説 M_{p_1, q_1} に対する AIC 基準を AIC_{p_1, q_1} とすると

$$\begin{aligned} A_{p_1, q_1} &= AIC_{p_1, q_1} - AIC_{p, q} \\ &= -N \log \Delta - 2(pq - p_1q_1) \end{aligned}$$

となる.

4. 主成分分析

p 次元変量 X は, 平均 μ , 共分散行列 Σ をもつとする. 主成分分析は, X の変動を少数個の線形変量で説明することを目的としている. 用いられる線形変量は主成分とよばれ

$$y_j = \gamma_j'(X - \mu), j = 1, \dots, p$$

で定義される. ここに, γ_j は Σ の固有値 λ_j に対応する長さ 1 の固有ベクトルである. すなわち

$$\Sigma \gamma_j = \lambda_j \gamma_j, \gamma_i' \gamma_j = \delta_{ij}$$

の解で, $\lambda_1 \geq \dots \geq \lambda_p > 0$ とする.

変量 X を $X = (X_1, X_2)$, $X_1: q \times 1$ と分割し, X_2 の冗長性 (X_1 の十分性) 問題を考える. X の分割に対応して, γ_j, Σ を

$$\gamma_j = \begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \end{bmatrix}, j = 1, \dots, p, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

と分割する. 判別分析や正準相関分析の場合と同様に考えれば, X_2 の冗長性を

$$“\gamma_{2j} = 0, j = 1, \dots, p”$$

として定義することになる. しかし, Σ の正則性を前提にする限り, 上のような定式化は無意味であることが分る. 実際, $\gamma_{2j} = 0 (j = 1, \dots, p)$ であること, $\Sigma_{11} \gamma_{1j} = \lambda_j \gamma_{1j}, j = 1, \dots, p$ が成り立つ. これより, $\lambda_{q+1} = \dots = \lambda_p = 0$ となり, Σ の正定値性に矛盾する. この方法による定式化の問題点は, すべての主成分を対象にしていることにある. しかし, いくつかの特定な主成分に限定することによってこの問題点を解消することができる. 一方, 応用上の観点から, いくつかの主成分だけに着目した変量の冗長性問題も重要である. 関心ある主成分の集合を

$$\{y_i | i \in V\}$$

とする. ここで, V は $\{1, \dots, p\}$ の部分集合で, その要

素の個数 $m (= \#(V))$ は q 以下であるとする. $m > q$ の場合は, 仮説の設定が無意味になる. また, Σ の固有値の重複度に関連して, $W = V^c \cap \{1, \dots, p\}$ とおくと

$$|\lambda_i - \lambda_j| > 0 (i \in V, j \in W)$$

を仮定する.

定義 4.1. $m (\leq q < p)$ 個の主成分 $\{y_i | i \in V\}$ に関して, $(p - q)$ の次元変量 X_2 が冗長である (q 次元変量 X_1 が十分である) とは,

$$\gamma_{2i} = 0 \quad (i \in V)$$

が成り立つことである. この仮説を $M_q(V)$ と表す.

Tyler (1982, 1983) は, より一般的な形での仮説 $M_q(V)$ の検定問題を扱っている. Flury (1985, 1986) は $\Sigma_1^{-1} \Sigma_2$ の固有ベクトルに関して同様な仮説を考え, その検定統計量を提案している.

行列 A, B を

$$A = \begin{bmatrix} I_q \\ O \end{bmatrix}, \quad B = \begin{bmatrix} O \\ I_{p-q} \end{bmatrix}$$

と定義する. また, 部分空間

$$\Gamma_V = R\{\gamma_i | i \in V\}, \Gamma_W = R\{\gamma_i | i \in W\}$$

への射影行列をそれぞれ P_V, P_W とする.

定理 4.1. 次の 4 つの条件は同値である.

- (1) $M_q(V): \gamma_{2i} = 0 \quad (i \in V),$
- (2) $R[A] \supset \Gamma_V,$
- (3) $R[B] \subset \Gamma_W,$
- (4) $P_W B = B.$

Tyler (1982, 1983) は一般的な行列 A, B に対して, 定理 4.1 の (3) または (4) を検定するための漸近的 χ^2 統計量を提案している. X についての大きさ $N = n + 1$ の標本に基づく標本共分散行列を S とする. S の固有値, 固有ベクトルを l_i, h_i とし

$$\begin{aligned} l_1 > \dots > l_p > 0, h_i' h_j = \delta_{ij} \\ h_i = \begin{bmatrix} h_{1i} \\ h_{2i} \end{bmatrix}, i = 1, \dots, p \end{aligned}$$

とする. このとき, 提案されている統計量は

$$T_n = n \sum_{i \in V} h_{2i}' \left\{ \sum_{j \in W} \frac{l_i l_j}{(l_i - l_j)^2} h_{2j} h_{2j}' \right\}^{-1} h_{2i}$$

である. 正規性の仮定のもとで, T_n は n が大のとき漸

近的に自由度 $m(p-q)$ の χ^2 -分布に従う。仮説 $M_q(V)$ はいくつかの固有ベクトルに制約を課すことによって得られる仮説であって、一般に、尤度比基準を求めることは困難である。

5. その他のモデルにおける変量の冗長性と関連する問題

Rao (1973) は多変量線形モデルにおける線形仮説の検定に関連して、付加情報仮説を導入し、その検定問題を扱っている。この付加情報仮説は変量の冗長性と関連するものであるが、その関連は必ずしも明白ではなかった。これに対して、Fujikoshi (1983) は Rao の付加情報仮説と同値な条件を与え、如何なる目的に対する変量の冗長性であるかを明白にしている。この付加情報仮説は多変量逆回帰モデルにおける変量の冗長性とも関連している (Fujikoshi & Nishii (1986))。多変量逆回帰モデルについては、Brown (1982)、Osborne (1991) を参照されたい。

共変量がある場合の判別問題に対しても、変量の冗長性の定式化が与えられ、その検定が研究されている (Fujikoshi & Khatri (1990))。尤度基準が求められているが、その分布は複雑で求められていない。共変量がある場合の判別分析における冗長性問題は、成長曲線モデル (Potthoff & Roy (1964)) における共変量に関する冗長性問題と類似している。後者の問題に対して、Fujikoshi & Rao (1991) は検出力および推定の効率の立場からの冗長性の定式化を与えている。

用いられている変量自身の冗長性の他に、線形変換変量の冗長性も重要である。正準判別変量の冗長性問題は別判解析の次元として扱われている。正準相関変量についても同様である。種々の次元推定法が研究されているが、その推定法の一つとして、冗長性仮説に対して情報基準量を適用する方法も提案されている (Fujikoshi & Veitch (1979), Fujikoshi (1985c))。主成分の冗長性問題は、正準判別変量等の場合とかなり異なっている。主成分の数の定め方については、例えば、Eastment & Krzanowski (1982), Krzanowski (1987), Wold (1978) 等を参照されたい。

冗長性仮説の検定統計量としては、尤度比基準の他に、Lawley-Hotelling 型基準と Bartlett-Nandi-Pillai 型基準が考えられる (Fujikoshi (1989))。これらの検定の性質に関する研究もあるが (Fujikoshi (1982), Kariya, Fujikoshi & Krishnaiah (1987), Pynnönen (1988) 等)、十分には研究されていない。

冗長性仮説を定式化し、モデル選択基準を適用することにより、一つの変量選択法が提案される。一方、別な方法による多くの変量選択法が提案されている。これらについては、例えば、Jolliffe (1972, 1973), Krzanowski (1987), McCabe (1984), McKay (1976, 1977a, b, 1979), McKay & Campeall (1972a, b), McLachlan (1980), Schaafsma (1982) 等も参照されたい。

冗長性仮説が棄却されると、付加情報量の大きさに関心がある。この推測に関連した研究として、Fujikoshi, Krishnaiah & Schmidhammer (1987), Wijsman (1986) 等がある。一方、冗長性仮説をモデルとして考慮することになれば、対応する冗長の変量を除くことになる。これによる推定の精度の向上に関する研究としては、Althem (1984), Siotani, Hayakawa & Fujikoshi (1985, p. 587 problems 13.4.1, 13.4.2) 等を参照されたい。

参考文献

- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov and F. Czaki, eds., *2nd International Symposium on Information Theory* (Akademiai Kiadó, Budapest), 267-281.
- Althem, P.M.E. (1984) Improving the precision of estimation by fitting a model. *J. R. Stat. Soc., B* **46**, 118-119.
- Anderson, T.W. (1984) *An Introduction to Multivariate Analysis*. 2nd ed., Wiley, New York.
- Brown, P. (1982) Multivariate calibration. *J.R. Statist. Soc., B* **44**, 287-321.
- Chen, C.W. (1971) On some problems on canonical correlation analysis. *Biometrika*, **58**, 399-400.
- Eastment, H.T. & Krzanowski (1982) Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, **24**, 73-77.
- Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179-188.
- Flury, B.N. (1985) Analysis of linear combinations with extreme ratios of variance. *J. Amer. Statist. Assoc.*, **80**, 915-922.
- Flury, B.N. (1986) An asymptotic test for redundancy of variables in the comparison of two covariance matrix. *Statist. Prob. Letters.*, **4**, 123-126.
- Fujikoshi, Y. & Veitch, L.G. (1979) Estimation of

- dimensionality in canonical correlation analysis. *Biometrika*, **66**, 345-351.
- Fujikoshi, Y. (1981) The power of the likelihood ratio test for additional information in a multivariate linear model. *Ann. Inst. Statist. Math.*, **33**, 279-285.
- Fujikoshi, Y. (1982) A test for additional information in canonical correlation analysis. *Ann. Inst. Statist. Math.*, **34**, 137-144.
- Fujikoshi, Y. (1983) A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.*, **13**, 203-214.
- Fujikoshi, Y. (1985a) Selection of variables in discriminant analysis and canonical correlation analysis. In: P.R. Krishnaiah, ed., *Multivariate Analysis- VI* (Elsevier Science Publishers B.V., Amsterdam), 219-236.
- Fujikoshi, Y. (1985b) Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate Anal.*, **17**, 27-37.
- Fujikoshi, Y. (1985c) Two methods for estimation of dimensionality in canonical correlation analysis and the multivariate linear model. In: K. Matusita, ed., *Statistical Theory and Data Analysis* (Elsevier Science Publishers B.V. North-Holland), 233-240.
- Fujikoshi, Y. and Nishii, R. (1986) Selection of variables in a multivariate inverse regression. *Hiroshima Math. J.*, **16**, 269-277.
- Fujikoshi, Y., Krishnaiah, P.R. & Schmidhammer, J. (1987) Effect of additional variables in principal component analysis and canonical correlation analysis. In: A.K. Gupta, ed., *Advances in Multivariate Statistical Analysis* (D. Reidel Publishing Company, Dordrecht), 45-61.
- Fujikoshi, Y. (1989) Tests for redundancy of some variables in multivariate analysis. In: Y. Dodge, ed., *Statistical Data Analysis and Inference* (North-Holland, Amsterdam), 141-161.
- Fujikoshi, Y. & Khatri, C.G. (1990) A study of redundancy of some variables in covariate discriminant analysis. *Ann. Inst. Statist. Math.*, **42**, 769-782.
- Fujikoshi, Y. & Rao, C.R. (1991) Selection of covariables in the growth curve model. *Biometrika*, **78**, 779-785.
- Hetelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321-377.
- Jolliffe, I.T. (1972) Discarding variables in a principal component analysis I: Artificial data. *Appl. Statist.*, **21**, 160-173.
- Jolliffe, I.T. (1973) Discarding variables in a principal component analysis II: Real data, *Appl. Statist.*, **22**, 21-31.
- Kariya, T., Fujikoshi, Y. & Krishnaiah, P.R. (1987) On tests for selection of variables and independence under multivariate regression models. *J. Multiv. Anal.*, **21**, 207-237.
- Krzanowski, W.J. (1987) Cross-Validation in principal component analysis. *Biometrics*, **43**, 575-584.
- Krzanowski, W.J. (1987) Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, **36**, 22-33.
- Laha, R.G. (1954) On some problems in canonical correlations. *Sankhyā*, **14**, 61-66.
- McCabe, G.P. (1984) Principal variables. *Technometrics*, **26**, 137-144.
- McKay, R.J. (1976) Simultaneous procedures in discriminant analysis involving two groups. *Technometrics*, **18**, 47-53.
- McKay, R.J. (1977a) Variable selection in multivariate regression: An application of simultaneous test procedures. *J. Roy. Statist. Soc.*, **B 39**, 371-380.
- McKay, R.J. (1977b) Simultaneous procedures for variable selection in multiple discriminant analysis. *Biometrika*, **64**, 283-290.
- McKay, R.J. (1979) The adequacy of variable subsets in multivariate regression. *Technometrics*, **21**, 475-479.
- McKay, R.J. & Campbell, N.A. (1982a) Variable selection techniques in discriminant analysis: I. Description. *British J. Math. Statist. Psychology*, **35**, 1-29.
- McKay, R.J. & Campbell, N.A. (1982b) Variable selection techniques in discriminant analysis: II. Allocation. *British J. Math. Statist. Psychology*, **35**, 40-41.
- McLachlan, G.J. (1980) On the relationship between the F test and the overall error rate for variable selection in two-group discriminant analysis. *Biometrics*, **36**, 501-510.
- Osborne, C. (1991) Statistical calibration: a review. *Inst. Statist. Rev.*, **59**, 309-336.
- Potthoff, R.R. & Roy, S.N. (1964) A generalized multivariate analysis of variance model useful specially for growth curve problems. *Biometrika*, **51**, 313-326.
- Pynnönen, S. (1987) Selection of variables in nonlinear discriminant analysis by information criteria. In: T. Pukkila and S. Puntanen, eds., *Proceedings of the Second International Tampere Conference in Statistics*. (University of Tampere, Finland), 627-636.
- Pynnönen, S. (1988) Testing for additional information in variables in Multivariate Normal classification with unequal covariance matrices. *Acta Wasaensia No. 23: Statistics No. 2*, Universitas wasaensis, Finland.

- Rao, C.R. (1946) Tests with discriminant functions in multivariate analysis. *Sankhyā*, **7**, 407-414.
- Rao, C.R. (1948) Tests of significance in multivariate analysis. *Biometrika*, **35**, 58-79.
- Rao, C.R. (1965) The theory of least squares when parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447-458.
- Rao, C.R. (1966) Covariance adjustment and related problems in multivariate analysis. In: P.R. Krishnaiah, ed., *Multivariate Analysis* (Academic Press, New York), 87-103.
- Rao, C.R. (1970) Inference on discriminant function coefficients. In: R.C. Bose and Others, eds., *Essays in probability and Statistics* (Univ. of North Carolina Press, Chapel Hill. N.C.), 587-602.
- Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*, Wiley, New York.
- Schaafsma, W. (1982) Selecting variable in discriminant analysis for improving upon classical procedures. In: P.R. Krishnaiah & L.N. Kanal, eds., *Handbook of Statistics*, Vol. 2 857-883, North-Holland, Amsterdam.
- Siotani, M. (1957) Effect of the additional variates on the canonical correlation coefficients. *Proc. Inst. Statist. Math.*, **5**, 52-57.
- Siotani, M. Hayakawa, T. and Fujikoshi, Y. (1985) *Modern Multivariate Statistical Analysis: A Graduate Course Handbook*. American Sciences Press, INC., Ohio.
- Tyler, D.E. (1981) Asymptotic inference for eigenvectors. *Ann. Statist.*, **9**, 725-736.
- Tyler, D.E. (1983) A class of asymptotic tests for principal component vectors. *Ann. Statist.*, **11**, 1243-1250.
- Wijsman, R.A. (1986) Asymptotic distribution of the increase of the largest canonical correlation when one of the vectors is augmented. *J. Multivariate Anal.*, **18**, 169-177.
- Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397-407.