HIROSHIMA MATH. J.
13 (1983), 203–214

A criterion for variable selection in multiple discriminant analysis

Yasunori FUJIKOSHI

(Received, September 17, 1982)

§ 1. Introduction

This paper deals with the problem of variable selection in discriminant analysis with $q+1$ populations and a multivariate linear model. The variable selection is important since there are situations where the deletion of some variables from the original variables may be preferable for the practical aim of statistical analysis. A number of stepwise procedures have been proposed for reducing the number of variables required to discriminate among the $q+1$ populations (e.g., see McCabe [7], Farmer and Freund [2]). McKay [8] has proposed a procedure for determining all subsets of variables that provide essentially as much separation among the $q+1$ populations as the original set of variables, based on a simultaneous test procedure in Gabriel's [6] sense.

In this paper we propose a criterion for determining the "best" subset of variables in the discriminant analysis whose aim is to interpret the differences among the $q+1$ populations in terms of only a few canonical discriminant variables. We obtain a creterion, based on a model fitting approach. We regard the problem of finding the "best" subset of variables as one of finding the "best" model, by introducing a family of parametric models. The parametric models are based on "no additional information hypotheses" due to Rao [10]. Our criterion is obtained by applying Akaike's information criterion (Akaike [1]) to choice of the models. The problem of finding the "best" subset of variables in a multivariate linear model is also discussed. This is a generalization of the problem of variable selection in the discriminant analysis. Asymptotic distributions of the criterion for variable selection in the multivariate linear model are obtained, resulting in generalizations of Fujikoshi [5] in the case of two-group discriminant analysis. The asymptotic distribution in the case when the original variables are ordered a priori can be reduced to a simple form.

§ 2. Multiple discriminant analysis

Consider $q+1$ p-variate normal population Π_α $(\alpha=1,\ldots, q+1)$ with means $\boldsymbol{\mu}_\alpha$ and the same covariance matrix Σ. Let $\boldsymbol{x}=(x_1,\ldots, x_p)'$ be the column vector of the p variables. Assume that N_α samples from Π_α are available. We will

identify a subset of the variables $x_1,\ldots, x_p$ by the corresponding subset of the set of subscripts $1, 2,\ldots, p$. If $j=\{j_1, j_2,\ldots, j_{k(j)}\}$ is such a subset of subscripts, $\boldsymbol{x}(j)$ will denote the vector variable whose components are specified by the elements of j. We can express $\boldsymbol{x}(j)$ as

$$\boldsymbol{x}(j) = G(j)\boldsymbol{x} \tag{2.1}$$

where $G(j)$ is a $k(j)\times p$ matrix whose (α, j_α) elements are all one for $\alpha=1,\ldots, k(j)$ and other elements are zero. Let J be the family of all possible subsets of the set of subscripts $\{1, 2,\ldots, p\}$. Then the problem of variable selection may be regarded as how to select the "best" subset j from J.

It is important to consider the criterion for variable selection such that the criterion relates as closely as possible to the practical aim of discriminant analysis. We consider variable selection in the case when we are interested in interpreting the differences among the $q+1$ populations in terms of only a few canonical discriminant variates. The discriminant analysis with this aim is called descriptive discriminant analysis. We shall first introduce a family of parametric models $M(j)$, $j\in J$ such that $M(j)$ means that $\boldsymbol{x}(j)$ is the "best" subset of variables for the descriptive discriminant analysis. Let Ω be the population between-groups covariance matrix defined by

$$\Omega = \textstyle\sum_{\alpha=1}^{q+1} (N_\alpha/N)(\boldsymbol{\mu}_\alpha - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_\alpha - \bar{\boldsymbol{\mu}})' \tag{2.2}$$

where $\bar{\boldsymbol{\mu}}=(1/N)\sum_{\alpha=1}^{q+1} N_\alpha\boldsymbol{\mu}_\alpha$ and $N=N_1+\cdots+N_{q+1}$. Then the coefficient vector $\boldsymbol{a}_\alpha$ of the α-th canonical variate $y_\alpha=\boldsymbol{a}_\alpha'\boldsymbol{x}$ is defined as the solution of

$$\Omega\boldsymbol{a}_\alpha = \lambda_\alpha \Sigma\, \boldsymbol{a}_\alpha,\ \boldsymbol{a}_\alpha'\Sigma\,\boldsymbol{a}_\beta = \delta_{\alpha\beta} \tag{2.3}$$

where $\lambda_1\geq\cdots\geq\lambda_p\geq 0$ are the characteristic roots of $\Sigma^{-1}\Omega$ and $\delta_{\alpha\beta}$ is the Kronecker delta. Let m be the number of non-zero characteristic roots of $\Sigma^{-1}\Omega$. Then $m=\mathrm{rank}\,(\Omega)\leq \mathrm{Min}\,(p, q)$ and the differences among the $q+1$ populations can be summarized in terms of the first m canonical variates $y_1,\ldots, y_m$. It is natural to say that a variable x_γ is irrelevant for the description of the differences among the $q+1$ populations if the γ-th components of $\boldsymbol{a}_\alpha$, $\alpha=1,\ldots, m$ are all zero. This implies the following definition of $M(j)$.

$$\begin{aligned} M(j)\colon\ & G(j^*)\boldsymbol{a}_\alpha = 0,\ \alpha = 1,\ldots, m \quad \text{and} \\ & (\textstyle\sum_{\alpha=1}^{m} G(j)\boldsymbol{a}_\alpha\boldsymbol{a}_\alpha'G(j)')_{\gamma\gamma} > 0, \quad \text{for any } \gamma\in j \end{aligned} \tag{2.4}$$

where j^* is the complement of j with respect to $\{1, 2,\ldots, p\}$. It is known (Fujikoshi [4]) that

$$\begin{aligned} & G(j^*)\boldsymbol{a}_\alpha = 0, \quad \alpha = 1,\ldots, m \\ & \Longleftrightarrow \end{aligned} \tag{2.5}$$

$$H(j)\colon G(j^*)[\boldsymbol{\mu}_1 - \Sigma G(j)'\{G(j)\Sigma G(j)'\}^{-1}G(j)\boldsymbol{\mu}_1]$$
$$= \cdots = G(j^*)[\boldsymbol{\mu}_{q+1} - \Sigma G(j)'\{G(j)\Sigma G(j)'\}^{-1}G(j)\boldsymbol{\mu}_{q+1}].$$

The latter statement in (2.5) is equivalent to "no additional information hypothesis" due to Rao [10]. Since $H(j) \Leftrightarrow \operatorname{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)' = \operatorname{tr}\Sigma^{-1}\Omega$ (McKay [8]), we can write $M(j)$ as

(2.6) $M(j)$: $\operatorname{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)' = \operatorname{tr}\Sigma^{-1}\Omega$ and
$\operatorname{tr}\{G(i)\Sigma G(i)'\}^{-1}G(i)\Omega G(i)' < \operatorname{tr}\Sigma^{-1}\Omega$ for any proper subset i of j.

Since $\operatorname{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)'$ and $\operatorname{tr}\Sigma^{-1}\Omega$ are the distances among the $q+1$ populations based on $\boldsymbol{x}(j)$ and $\boldsymbol{x}$ respectively, we can say that if $M(j)$ is true, $\boldsymbol{x}(j)$ is a parsimoneous subset of variables that provides essentially the same information as $\boldsymbol{x}$ for the descriptive discriminant analysis.

Next we shall derive Akaike's information criterion (Akaike [1]), for choice of the models $\{M(j); j \in J\}$. The criterion is to choose the model for which

(2.7) $$\mathrm{AIC}(j) = -2\log L(\hat{\theta}(j)) + 2p(j)$$

is minimized, where $L(\theta)$ is the likelihood function of observations, $\hat{\theta}(j)$ is the maximum likelihood estimate of $\theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{q+1}, \Sigma\}$ under $M(j)$ and $p(j)$ is the dimensionality of θ under $M(j)$. This criterion was constructed to choose a model such that the model yields the best predictions for future observations with the same structure as the original observations. Let B and W be the matrices of sums of squares and products due to between groups and within groups, respectively, based on random samples of size N_α from Π_α ($\alpha = 1, \ldots, q+1$). Considering the conditional distribution of $\boldsymbol{x}(j^*)$ given $\boldsymbol{x}(j)$ and using (2.5), it is shown that

(2.8) $$-2\log \mathrm{Max}_{M(j)} L(\theta) = Np\{1 + \log(2\pi/N)\} + N\log|G(j)WG(j)'|$$
$$+ N\log|G(j^*)[T - TG(j)'\{G(j)TG(j)'\}^{-1}G(j)T]G(j^*)'|$$

where $T = B + W$. Since the correspondence between θ and $\boldsymbol{\mu}_{1\alpha} = G(j)\boldsymbol{\mu}_\alpha$, $\boldsymbol{\mu}^*_{2\alpha} = G(j^*)[I_p - \Sigma G(j)'\{G(j)\Sigma(j)'\}^{-1}G(j)]\boldsymbol{\mu}_\alpha$, and $\Sigma^* = (G(j)', G(j^*)')'\Sigma(G(j)', G(j^*)')$ is is one-to-one, we have

(2.9) $$2p(j) = 2\{(q+1)k(j) + p - k(j) + \frac{1}{2}p(p+1)\}$$
$$= p(p+1) + 2p(q+1) - 2q(p-k(j)).$$

Therefore the criterion based on (2.7) is equivalent to choosing the model $M(j)$ to minimize

(2.10)

$$\begin{aligned} A(j) &= \mathrm{AIC}(j) - \mathrm{AIC}(\{1, 2, \ldots, p\}) \\ &= -N \log \frac{|G(j^*)[W - WG(j)'\{G(j)WG(j)'\}^{-1}G(j)W]G(j^*)'|}{|G(j^*)[T - TG(j)'\{G(j)TG(j)'\}^{-1}G(j)T]G(j^*)'|} \\ &\qquad - 2q(p - k(j)) \\ &= -N \log \left\{ \frac{|W|}{|T|} \Big/ \frac{|G(j)WG(j)'|}{|G(j)TG(j)'|} \right\} - 2q(p - k(j)) \end{aligned}$$

where $A(\{1, 2, \ldots, p\}) = 0$.

For the case of two populations, i.e., $q = 1$ we obtain

$$A(j) = N \log [1 + (D^2 - D(j)^2)/\{N(N-2)/(N_1 N_2) + D(j)^2\}] - 2(p - k(j)) \tag{2.11}$$

where D and $D(j)$ are the sample Mahalanobis distance between Π_1 and Π_2 based on $\boldsymbol{x}$ and $\boldsymbol{x}(j)$, respectively. The criterion in the special case was derived by Fujikoshi [5].

It may be noted that the first term of $A(j)$ in (2.10) is the likelihood ratio statistic for testing the no aditional information hypothesis $H(j)$ in (2.5). This test statistic was introduced by Rao [9].

§ 3. Extension to a multivariate linear model

In this section we deal with the problem of variable selection in a multivariate linear model which is a generalization of the case described in the previous section. Consider a matrix X whose rows are independent N observations of the p vector variate $\boldsymbol{x} = (x_1, \ldots, x_p)'$ being normally distributed with common covariance matrix Σ and expectation

$$\mathrm{E}(X) = Z\Xi \tag{3.1}$$

where Z is a known $N \times b$ matrix of rank b and Ξ is a $b \times p$ matrix of unknown parameters. Let

$$\delta^2 = \mathrm{tr}\, \Sigma^{-1}\Omega \tag{3.2}$$

where $N\Omega = (C\Xi)'\{C(Z'Z)^{-1}C'\}^{-1}C\Xi$ and C is a known $q \times b$ matrix of rank q. The quantity δ^2 can be regarded as a measure of deparatures from the nullity of the hypothesis "$C\Xi = 0$". If we put

$$C: q \times (q+1) = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ & & \cdots\cdots & & \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

in multiple discriminant analysis, then Ω is equal to the population between-groups covariance matrix in (2.2). We consider the variable selection in the case when we want to have a large δ^2. The quantity δ^2 for a subvector variate $\boldsymbol{x}(j) = G(j)\boldsymbol{x}$ may be defined by

$$\delta^2(j) = \mathrm{tr}\,\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)'. \tag{3.3}$$

In general, it holds that

$$\delta^2(i) \leq \delta^2(j) \leq \delta^2 \tag{3.4}$$

for any i, j such that $i \subseteq j$. Our problem is to find a parsimonious subset of variables that provides the same value as δ^2. For this purpose we consider the same model as in (2.6), i.e.,

(3.5) $M(j)$: $\delta^2(j) = \delta^2$ and $\delta^2(i) < \delta^2$ for any proper subset i of j,

which means that $\boldsymbol{x}(j)$ is the "best" subset of variables. The other expressions for $M(j)$ are obtained by using the equivalence of the following three statements (3.6), and (3.7) and (3.8):

$$\delta^2(j) = \delta^2, \tag{3.6}$$

$$H(j): C[\Xi - \Xi G(j)'\{G(j)\Sigma G(j)'\}^{-1}G(j)\Sigma]G(j^*)' = O, \tag{3.7}$$

$$G(j^*)\boldsymbol{a}_\alpha = 0, \quad \alpha = 1,\ldots, m, \tag{3.8}$$

where $m = \mathrm{rank}\,(\Omega)$ and $\boldsymbol{a}_\alpha$ are the characteristic vectors of Ω with respect to Σ as in (2.3). McKay [8] has proved the equivalence of (3.6) and (3.7). The equivalence of (3.7) and (3.8) is obtained by the same argument as in multiple discriminant analysis due to Fujikoshi [4]. The hypothesis (3.7) can be interpreted as the hypothesis that $\boldsymbol{x}(j^*)$ supplies no additional information about deparatures from nullity of the hypothesis "$C\Xi = O$", independently of $\boldsymbol{x}(j)$. It is known (Rao [9]) that likelihood ratio statistic for testing $H(j)$ is

$$-N \log [\{|W|/|T|\}/\{|G(j)WG(j)'|/|G(j)TG(j)'|\}] \tag{3.9}$$

where $T = W + B$,

$$B = (C\hat{\Xi})'\{C(Z'Z)^{-1}C'\}^{-1}C\hat{\Xi},\ W = X'(I_N - Z(Z'Z)^{-1}Z')X, \tag{3.10}$$

and $\hat{\Xi}=(Z'Z)^{-1}Z'X$. This result is obtained by considering the conditional distribution of $\boldsymbol{x}(j^*)$ given $\boldsymbol{x}(j)$. The matrices W and B are the matrices of sums of squares and products due to error and departure from the hypothesis in the problem of testing "$C\Xi=O$" against "$C\Xi\neq O$". Using (3.9) it is easily seen that the selection criterion based on (2.7) is equivalent to choosing the model $M(j)$ to minimize

$$A(j)=\mathrm{AIC}(j)-\mathrm{AIC}(\{1,2,\ldots,p\}) \tag{3.11}$$
$$=-N\log\left\{\frac{|W|}{|T|}\Big/\frac{|G(j)WG(j)'|}{|G(j)TG(j)'|}\right\}-2q(p-k(j))$$

where $A(\{1,2,\ldots,p\})=0$.

§4. Asymptotic distributions

We denote the subset j to minimize the $A(j)$ in (3.11) by $\hat{j}(J)$, i.e.,

$$\mathrm{Min}_{j\in J}A(j)=A(\hat{j}(J)). \tag{4.1}$$

In the following we assume that the model $M(j_0)$ in (3.5) is true, where $j_0=\{1,\ldots,k\}$. Then we have that

A1: $\mathrm{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)'=\delta^2$ for $j\in J_1$

and

A2: $\mathrm{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega G(j)'<\delta^2$ for $j\in J_2$

where $\delta^2=\mathrm{tr}\,\Sigma^{-1}\Omega$, $J_1=\{j;\ j\supseteq j_0\}\cap J$ and $J_2=J_1^c\cap J$. This is easily seen by using the equivalence of (3.6) and (3.8). In general, we are interested in knowing how large is the probability of $\Pr(\hat{j}(J)=j_0)$. Apart from it, to derive the distribution of $\hat{j}(J)$ may be fundamental in studying the statistical properties of the selection method $\hat{j}(J)$. In the following we shall derive the asymptotic distribution when $M(j_0)$ is true.

The matrices W and B in (3.10) are independently distributed as a central Wishart distribution $W_p(N-b,\Sigma)$ and a noncentral Wishart distribution $W_p(q,\Sigma;N\Omega)$, respectively. The matrix Ω depends on N. It is natural to assume that $\Omega=O(1)$ with respect to N. For simplicity we make the following assumption for Ω:

A3: Ω is a fixed matrix.

Let Σ and Ω be partitioned as

$$\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix},\ \Sigma_{11}:k\times k \text{ and } \Omega=\begin{pmatrix}\Omega_{11}&\Omega_{12}\\ \Omega_{21}&\Omega_{22}\end{pmatrix},\ \Omega_{11}:k\times k. \tag{4.2}$$

The following Lemma is useful in reducing our distribution problem.

LEMMA 1. *Assume that* $\operatorname{tr}\Sigma_{11}^{-1}\Omega_{11}=\operatorname{tr}\Sigma^{-1}\Omega$. *Then there exists a lower triangular matrix*

$$A=\begin{pmatrix} A_{11} & O \\ A_{21} & A_{22} \end{pmatrix},\quad A_{11}: k\times k \tag{4.3}$$

such that A_{22} *is a lower triangular matrix,*

$$A\Omega A'=\Lambda \quad \text{and} \quad A\Sigma A'=I_p \tag{4.4}$$

where $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_k,0,\ldots,0)$ *and* $\lambda_1\geq\cdots\geq\lambda_k$ *are the possible non-zero roots of* $\Sigma^{-1}\Omega$.

PROOF. Let $\Sigma_{22.1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Then there exists a lower triangular matrix A_{22} such that $A_{22}\Sigma_{22\cdot 1}A_{22}'=I_{p-k}$. We define A_{11} and A_{21} by $H\Sigma_{11}^{-1/2}$ and $-A_{22}\Sigma_{21}\Sigma_{11}^{-1}$, respectively, where H is an orthogonal matrix. Then $A\Sigma A'=I_p$. We use that

$$\begin{aligned} &\operatorname{tr}\Sigma_{11}^{-1}\Omega_{11}=\operatorname{tr}\Sigma^{-1}\Omega \\ \Longleftrightarrow\quad &(-\Sigma_{21}\Sigma_{11}^{-1}, I_{p-k})\Omega=O \\ \Longleftrightarrow\quad &\lambda_\alpha=\operatorname{ch}_\alpha(\Sigma_{11}^{-1}\Omega_{11}),\quad \alpha=1,\ldots,k, \end{aligned} \tag{4.5}$$

where $\operatorname{ch}_\alpha(M)$ is th α-th largest characteristic root of M. The equivalences follows from

$$\operatorname{tr}\Sigma^{-1}\Omega=\operatorname{tr}\Sigma_{11}^{-1}\Omega_{11}+\operatorname{tr}(-\Sigma_{21}\Sigma_{11}^{-1}, I_{p-k})\Omega(-\Sigma_{21}\Sigma_{11}^{-1}, I_{p-k})'\Sigma_{22\cdot 1}$$

and

$$\lambda_\alpha\geq\operatorname{ch}_\alpha(\Sigma_{11}^{-1}\Omega_{11}),\quad \alpha=1,\ldots,k \qquad \text{(e.g., see Gabriel [6])}.$$

Using the above properties we have

$$A\Omega A'=\begin{pmatrix} H'\Sigma_{11}^{-1/2}\Omega_{11}\Sigma_{11}^{-1/2}H & O \\ O & O \end{pmatrix},$$

and hence we can choose H satisfying $A\Omega A'=\Lambda$.

LEMMA 2. *Assume that* A1 $\sim$ A3 *hold. Then*

$$\lim_{N\to\infty}N^{-1}A(i)=0<\lim_{N\to\infty}N^{-1}A(j). \tag{4.6}$$

for any i, j *such that* $i\in J_1$ *and* $j\in J_2$.

PROOF. Since $N^{-1}W$ and $N^{-1}B$ converge to Σ and Ω, respectively, we have

$$\begin{aligned}\lim_{N\to\infty} N^{-1}A(j) &= -\log(|\Sigma|/|\Sigma+\Omega|)\\ &\quad +\log(|G(j)\Sigma(j)'|/|G(j)(\Sigma+\Omega)G(j)'|)\\ &= \log|I_p+\Sigma^{-1}\Omega|\\ &\quad -\log|I_{k(j)}+(G(j)\Sigma G(j)')^{-1}G(j)\Omega G(j)'|.\end{aligned}$$

Using (4.5) it is seen that for $i\in J_1$

$$\mathrm{ch}_\alpha(\Sigma^{-1}\Omega)=\mathrm{ch}_\alpha(\{G(i)\Sigma G(i)'\}^{-1}G(i)\Omega G(i)') \qquad \text{for any } \alpha,$$

and for $j\in J_2$

$$\mathrm{ch}_\alpha(\Sigma^{-1}\Omega)\geq \mathrm{ch}_\alpha(\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega(j)') \qquad \text{for any } \alpha,$$

and the inequality holds strictly for some α. This implies (4.6).

Using the same argument as in Shibata [11] or Fujikoshi [5] from Lemma 2 we have the following reduction for the asymptotic distribution of $\hat{j}(J)$:

$$\begin{aligned}\lim_{N\to\infty}\Pr(\hat{j}(J)=j) &= r(j\mid \hat{j}_0, J)\\ &= \begin{cases}\lim_{N\to\infty}\Pr(A(j)\leq A(m),\ m\in J_1), & j\in J_1,\\ 0, & j\in J_2.\end{cases}\end{aligned} \tag{4.7}$$

Now we study asymptotic behaviour of $A(j), j\in J_1$. For $j\in J_1$, we can write $G(j)$ as

$$G(j)=\begin{pmatrix} I_k & O\\ O & L(j)\end{pmatrix}, \quad L(j)\colon (k(j)-k)\times(p-k). \tag{4.8}$$

Then we have

$$\begin{aligned}A(j) &= N\log\left\{\begin{vmatrix} W_{11} & W_{12}L(j)'\\ L(j)W_{21} & L(j)W_{22}L(j)'\end{vmatrix}\Big/\begin{vmatrix} T_{11} & T_{12}L(j)'\\ L(j)T_{21} & L(j)T_{22}L(j)'\end{vmatrix}\right\}\\ &\quad -N\log(|W|/|T|)-2q(p-k(j))\\ &= A(j_0)+N\log\{|L(j)W_{22\cdot 1}L(j)'|/|L(j)T_{22\cdot 1}L(j)'|\}\\ &\quad -2q(k-k(j))\end{aligned} \tag{4.9}$$

where $W_{22\cdot 1}=W_{22}-W_{21}W_{11}^{-1}W_{12}$, $T_{22\cdot 1}=T_{22}-T_{21}T_{11}^{-1}T_{12}$, and $W_{\alpha\beta}$ and $T_{\alpha\beta}$ are the submatrices of W and T, respectively, partitioned as in (4.2). Let

$$W^*=AWA' \quad \text{and} \quad B^*=ABA' \tag{4.10}$$

which are independently distributed as a central Wishart distribution $W_p(N-b,$

I_p) and a noncentral Wishart distribution $W_p(q, I_p; N\Lambda)$, respectively. We can write B^* as $B^* = U^{*\prime}U^*$, where

$$U^* = Z + \begin{pmatrix} \sqrt{N}\Lambda_1^{1/2} & O \\ O & O \end{pmatrix}, \tag{4.11}$$

$Z=(z_{\alpha\beta})$: $q \times p$ and $z_{\alpha\beta}$'s are independent identically random variables with the standard normal distribution. Let

$$U^* = (U_1^*, U_2^*),\ U_1^*: q \times k,\ Z = (Z_1, Z_2),\ Z_1: q \times k,$$
$$U = U^*(A^{-1})' = (U_1, U_2),\ U_1: q \times k.$$

Since $T_{22\cdot1} - W_{22\cdot1} = (U_2 - U_1 W_{11}^{-1} W_{12})'(I_q + U_1 W_{11}^{-1} U_1')^{-1}(U_2 - U_1 W_{11}^{-1} W_{12})$ (Fujikoshi [3]), it holds that

$$T_{22\cdot1} - W_{22\cdot1} = A_{22}^{-1}(U_2^* - U_1^* W_{11}^{*-1} W_{12}^*)'(I_q + U_1^* W_{11}^{*-1} U_1^*)^{-1} \cdot (U_2^* - U_1^* W_{11}^{*-1} W_{12}^*)(A_{22}^{-1})', \tag{4.12}$$

and

$$W_{22\cdot1} = A_{22}^{-1} W_{22\cdot1}^* (A_{22}^{-1})' \tag{4.13}$$

where $W_{22\cdot1}^* = W_{22}^* - W_{21}^* W_{11}^{*-1} W_{12}^*$ and $W_{\alpha\beta}^*$ are the submatrices of W^* partitioned as in (4.2). Let

$$(1/N)W^* = I_p + (1/\sqrt{N})V. \tag{4.14}$$

Substituting (4.11) and (4.14) into (4.12) and (4.13), we obtain the following expressions:

$$T_{22\cdot1} - W_{22\cdot1} = A_{22}^{-1} Y' Y (A_{22}^{-1})' + O_p(N^{-1/2}), \tag{4.15}$$

$$(1/N)W_{22\cdot1} = \Sigma_{22\cdot1} + O_p(N^{-1/2}), \tag{4.16}$$

where $\Sigma_{22\cdot1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$,

$$Y = \begin{pmatrix} I_k + \Lambda_1 & O \\ O & I_{q+k} \end{pmatrix}^{-1/2} (Z_2 - \begin{pmatrix} \Lambda_1^{1/2} \\ O \end{pmatrix} V_{12}), \tag{4.17}$$

$\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $V_{\alpha\beta}$ are the submatrices of V partitioned as in (4.2). When N tends to infinity, the elements of Y: $q \times (p-k)$ are independently distributed as $N(0, 1)$. Using (4.15) and (4.16) we obtain

$$(4.18)\qquad N\log\{|L(j)W_{22\cdot 1}L(j)'|/|L(j)T_{22\cdot 1}L(j)'|\}$$
$$= -N\log|I_{k(j)-k} + \frac{1}{N}\{L(j)\frac{1}{N}W_{22\cdot 1}L(j)'\}^{-1}$$
$$\cdot L(j)(T_{22\cdot 1} - W_{22\cdot 1})L(j)'|$$
$$= -\operatorname{tr} K(j)Y'Y + O_p(N^{-1/2})$$

where

$$(4.19)\qquad K(j) = (A_{22}^{-1})'L(j)'\{L(j)\Sigma_{22\cdot 1}L(j)'\}^{-1}L(j)A_{22}^{-1}.$$

Therefore it holds that for any j, $m\in J_1$

$$(4.20)\quad A(j) - A(m)$$
$$= \operatorname{tr}\{K(m) - K(j)\}Y'Y - 2\{k(m) - k(j)\} + O_p(N^{-1/2}).$$

From (4.7) and (4.20) we have the following Theorems 1 and 2:

THEOREM 1. *Suppose that the model $M(j_0)$ in* (3.5) *is true and the assumption* A3 *is satisfied, where $j_0=\{1, 2,\ldots, k\}$. Then*

$$(4.21)\quad \lim_{N\to\infty}\Pr(\hat{j}(J) = j) = r(j\mid j_0, J)$$
$$= \begin{cases} \Pr(\operatorname{tr}\{K(m) - K(j)\}Y'Y \le 2q\{k(m) - k(j)\},\ m\in J_1), & j\in J_1, \\ 0, & j\in J_2, \end{cases}$$

where $J_1=\{j;\ j\supseteq j_0\}\cap J$, $J_2=J_1^c\cap J$, $K(j)$ is defined by (4.19), *$k(j)$ is the number of the elements of j, $Y=(y_{\alpha\beta})$; $q\times(p-k)$ and $y_{\alpha\beta}$'s are independent identically random variables with the standard normal distribution.*

THEOREM 2. *Let $\tilde{J}$ be a subfamily of J, and $\hat{j}(\tilde{J})$ the selection method defined by $\operatorname{Min}_{j\in\tilde{J}}A(j)=A(\hat{j}(\tilde{J}))$. Then under the same assumptions as in Theorem* 1 *and the assumption of $\tilde{J}_1\neq\emptyset$ it holds that*

$$(4.22)\qquad \lim_{N\to\infty}\Pr(\hat{j}(\tilde{J}) = j) = r(j\mid j_0, \tilde{J})$$
$$= \begin{cases} \Pr(\operatorname{tr}\{K(m) - k(j)\}Y'Y \le 2q\{k(m) - k(j)\},\ m\in\tilde{J}_1), & j\in\tilde{J}_1, \\ 0, & j\in\tilde{J}_2, \end{cases}$$

where $\tilde{J}_1=J_1\cap\tilde{J}$ and $\tilde{J}_2=J_2\cap\tilde{J}$.

We note that similar results are also obtained if we replace the assumption A3 in Theorems 1 and 2 by a weak assumption

A3′: There exists a positive semi-define matrix Ω_0 such that $\lim_{N\to\infty}\Omega=\Omega_0$ and $\operatorname{tr}\{G(j)\Sigma G(j)'\}^{-1}G(j)\Omega_0G(j)<\operatorname{tr}\Sigma^{-1}\Omega_0$ for $j\in J_2$.

In this case we need only to change the matrix A used in the definition of $K(j)$ by A_0, where A_0 is the matrix A in Lemma 1 with $\Omega=\Omega_0$.

We can obtain a further reduction of (4.22) when an ordering of variables is given a priori. As such a subfamily, consider the family of p models, $M(j)$, $j\in J_0$, where

$$J_0 = \{\{1\}, \{1, 2\},\ldots, \{1, 2,\ldots, p\}\}. \tag{4.23}$$

We denote the subset $\{1,\ldots, j\}$ by j. Then, since A_{22} is a lower triangular matrix, we have

$$K(j) = \begin{pmatrix} I_{j-k} & O \\ O & O \end{pmatrix}, \quad k \leq j \leq p. \tag{4.24}$$

This implies

$$\begin{aligned} \lim_{N\to\infty} \Pr(\hat{j}(J_0) = j) &= r(j \mid j_0, J_0) \\ &= \begin{cases} s_q(j-k)t_q(p-j), & k \leq j \leq p, \\ 0, & j \leq k-1, \end{cases} \end{aligned} \tag{4.25}$$

where

$$s_q(k) = \Pr\left(\cap_{\alpha=1}^{k}(U_\alpha > 0)\right),\ t_q(k) = \Pr\left(\cap_{\alpha=1}^{k}(U_\alpha \leq 0)\right), \tag{4.26}$$

$s_q(0)=t_q(0)=1$, $U_\alpha=(W_1-2q)+\cdots+(W_\alpha-2q)$ and $\{W_\alpha\}$ is a sequence of independent χ^2_q random variables. For a reduction of $s_q(k)$ and $t_q(k)$, see Spitzer [12] and Shibata [11]. We note that the asymptotic distribution of $\hat{j}(J_0)$ depends only on q, p and k, but not on the values of Ξ and Σ.

References

[1] H. Akaike, A new look at the statistical model indentification, I. E. E. E. Trans. Auto. Control, AC-**19** (1974), 716–723.

[2] J. H. Farmer and R. J. Freund, Variable selection in the multivariate analysis of variance (MANOVA), Communications in Statistics, **4** (1975), 87–98.

[3] Y. Fujikoshi, The power of the likelihood ratio test for additional infromation in a multivariate linear model, Ann. Inst. Statist. Math., Part A, **33** (1981), 279–285.

[4] Y. Fujikoshi, A test for additional information in canonical correlation analysis. To appear in Ann. Inst. Statist. Math., Part A, **34** (1982).

[5] Y. Fujikoshi, Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria, Submitted for publication.

[6] K. R. Gabriel, Simultaneous test procedures in multivariate analysis of variance, Biometrika, **55** (1968), 489–504.

[7] G. P. McCave, Computations for variable selection in discriminant analysis, Technometrics, **17** (1975), 103–110.

[8] R. J. McKay, Simultaneous procedures for variable selection in multiple discriminant analysis, Biometrika, **64** (1977), 283–290.
[9] C. R. Rao, Tests of significance in multivariate analysis. Biometrika, **35** (1948), 58–79.
[10] C. R. Rao, Linear Statistical Inference and its Applications, Wiley, New York (1965).
[11] R. Shibata, Selection of the order of an autoregessive model by Akaike's information criterion, Biometrika, **63** (1976), 117–126.
[12] F. Spitzer, A combinatorial lemma and its application to probability theory, Trans. Amer. Math. Soc., **82** (1956), 323–339.

Department of Mathematics,
Faculty of Science,
Hiroshima University