Sankhyā : The Indian Journal of Statistics
San Antonio Conference: selected articles
2002, Volume 64, Series A, Pt.2, pp 256-267

SELECTION OF VARIABLES FOR DISCRIMINANT ANALYSIS IN A HIGH-DIMENSIONAL CASE

By YASUNORI FUJIKOSHI
Hiroshima University, Japan

SUMMARY. One way of proceeding with variable selection in discriminant analysis is to formulate it as a problem of selecting the best model from a family of variable selection models. The variable selection models considered are the ones based on no additional information, due to Rao (1948, 1973). We consider a selection criterion based on an approximately unbiased estimator (AIC, Akaike (1970)) for the expected log-predictive likelihood or equivalently the expected Kullback-Leibler information for a candidate model. Such a method has been proposed by Fujikoshi (1985), based on the asymptotic theory under the usual large sample framework. The purpose of the present paper is to explore the approach under a high-dimensional framework when the dimension is comparable to the sample size. For two-groups discriminant analysis, a modified criterion is proposed.

1. Introduction

One of the common statistical problems in discriminant analysis is to describe the difference of several different groups in terms of new transformed variables. A well known procedure is to use canonical variate analysis. Our goal is to remove a set of redundant variables, and to find the best subset of variables.

We formulate the problem as a model selection problem. We consider a family of variable selection models based on no additional information due to Rao (1948, 1970). The selection criterion is based on the idea of Akaike (1973). We use an approximately unbiased estimator (AIC) for the AIC type of risk defined by the expected log-predictive likelihood or equivalently the expected Kullback-Leibler information for a candidate model. Such a method has been proposed by Fujikoshi (1985), based on the usual large sample framework. On the other hand, in discriminant analysis we encounter

Paper received June 2000; revised November 2001.
AMS (1991) subject classification. Primary 62H12; secondary 62E30.
Keywords and phrases. AIC, bias correction, discriminant analysis, high dimensional case, modified criterion, selection of variables, two-groups case.

a high dimensional case, i.e., the case when the dimension is comparable to the sample size. There are some works on asymptotic approximations for the expected probabilities of misclassification in the two-groups discriminant analysis under a high dimensional framework. For these works see, e.g., Raudys (1972), Wyman et al. (1990), and Fujikoshi and Seo (1998), in which they point to the goodness of such approximations.

In this paper we consider the problem of estimating the AIC type of risk in a high dimensional case when the dimension is comparable to the sample size, and we attempt to derive a reduction for the bias term when we estimate the AIC type of risk by $-2\log$ likelihood. For two-groups discriminant analysis, we obtain an asymptotic expression which leads to a modified criterion.

The present paper is organized in the following way. In Section 2 we state our approach. In Section 3 we derive a reduction for the bias term in the estimation of the AIC type of risk. In Section 4 we restrict to the two-group case, and derive an asymptotic approximation of the bias term under a high dimensional framework. This leads to a modified criterion.

2. An Approach to Selection of Variables

Let $\boldsymbol{y} = (y_1, ..., y_p)'$ be a random vector measurable on the individuals of each of $q+1$ populations $\Pi_1, ..., \Pi_{q+1}$. Let

$$Y = [\boldsymbol{y}_1^{(1)}, ..., \boldsymbol{y}_{N_1}^{(1)}, ..., \boldsymbol{y}_1^{(q+1)}, ..., \boldsymbol{y}_{N_{q+1}}^{(q+1)}] \tag{2.1}$$

be an observation matrix, where $\boldsymbol{y}_1^{(i)}, ..., \boldsymbol{y}_{N_i}^{(i)}$ are random samples of sizes N_i from Π_i. Let $N = N_1 + \ldots + N_{q+1}$. Suppose that the $N \times p$ observation matrix Y has the probability density function $g(Y)$ under the true model M^*, and

$$M^* : \mathrm{E}^*(\boldsymbol{y}|\Pi_i) = \boldsymbol{\mu}^{(i)*}, \quad \mathrm{Var}^*(\boldsymbol{y}|\Pi_i) = \Sigma^*, \tag{2.2}$$

where E^* and Var^* denote the expectation and the covariance matrix under the true model M^*.

Let K be any subset of $P = \{1, 2, \ldots, p\}$, with $|K| = k$ members, and let $\boldsymbol{y}_K$ be the corresponding subvector of $\boldsymbol{y}$, containing only those elements whose indices are in K. For any subset K, we define a variable selection model M_K, and its AIC type of risk R_K. We also consider two estimators AIC_K and $MAIC_K$ for R_K. Suppose that we are interested in selecting the

best subset from a family $\mathcal{K}$ of subsets in P. Then the variable selection criterion based on AIC is defined to select $\hat{K}$ such that

$$\min_{K \in \mathcal{K}} AIC_K = AIC_{\hat{K}}. \tag{2.3}$$

The variable selection criterion based on $MAIC$ is defined similarly. The model corresponding to $K \in \mathcal{K}$ is called a candidate model. Note that the best subset of variables is the model minimizing the AIC type of risk in some sense, and our variable selection method is based on estimators of the risk. Therefore, it is important to construct a good estimator for R_K for any $K \in \mathcal{K}$, or more generally for any K in P. The purpose of the present paper is to construct a good estimator of the risk for any given candidate model. So, in the following we fix a subset K, and assume that $K = \{1, 2, \ldots, k\}$. This is for notational simplicity, and the results are applicable for any subset. Further, we shall identify K with k.

Now we consider a model M_k, which means that the first k variate $\boldsymbol{y}_1 = (y_1, ..., y_k)$ is sufficient, and the remainder variate $\boldsymbol{y}_2 = (y_{k+1}, ..., y_p)$ is redundant, i.e., the remainder $p-k$ variate $\boldsymbol{y}_2$ has no additional information in canonical variate analysis, in presence of $\boldsymbol{y}_1$. Here we assume that the random vector $\boldsymbol{y}$ under Π_i is distributed as $N(\boldsymbol{\mu}^{(i)}, \Sigma)$. In order to write the model M_k in a parametric form, let us consider the partitions

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{2.4}$$

and let

$$\boldsymbol{\mu}_{2\cdot 1}^{(i)} = \boldsymbol{\mu}_2^{(i)} - \Gamma \boldsymbol{\mu}_1^{(i)}, \quad i = 1, ..., q+1, \ \Gamma = \Sigma_{21}\Sigma_{11}^{-1}.$$

Then our candidate model M_k can be written as

$$M_k : \boldsymbol{y}|\Pi_i \sim N(\boldsymbol{\mu}^{(i)}, \Sigma), (i = 1, ..., q+1), \ \boldsymbol{\mu}_{2\cdot 1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot 1}^{(q+1)}. \tag{2.5}$$

Let $f(Y; \Theta)$ be the density function of Y when $\boldsymbol{y}|\Pi_i \sim N(\boldsymbol{\mu}^{(i)}, \Sigma), i = 1, ..., q+1$, where $\Theta = \{\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(q+1)}, \Sigma\}$. Then, we can write the risk defined by the expected log-predictive likelihood for a model M_k as

$$R_k = \mathrm{E}_Y^* \mathrm{E}_X^* [-2 \log f(X; \hat{\Theta}_k)], \tag{2.6}$$

where X is an $N \times p$ random matrix that has the same distribution as Y and is independent of Y, and $\hat{\Theta}_k$ is the maximum likelihood estimator of Θ under M_k. Following Akaike (1973) we have an estimator for R_k defined by

$$AIC_k = -2 \log f(Y; \hat{\Theta}_k) + b_{k0}, \tag{2.7}$$

where $b_{k0} = 2\times$ (the number of independent parameters under M_k). Note that the risk R_k can be expressed as

$$R_k = \mathrm{E}_Y^*[-2\log f(Y;\hat{\Theta}_k)] + b_k, \tag{2.8}$$

where

$$b_k = \mathrm{E}_Y^*\mathrm{E}_X^*[-2\log\{f(X;\hat{\Theta}_k)/f(Y;\hat{\Theta}_k)\}]. \tag{2.9}$$

This means that a naive estimator of R_k is $-2\log f(Y;\hat{\Theta}_k)$, and b_k is its bias term in the estimation of R_k. Further, b_{k0} is an estimator for b_k. Our interest is to examine the problem of evaluating and estimating the bias term b_k in a high dimensional framework.

In general, the bias correction problem has been studied in the usual large sample framework. For discriminant analysis, see Fujikoshi (1985). In multivariate analysis we encounter a high dimensional case, i.e., the case when the dimension p is comparable to the sample size N. Especially, in discriminant analysis it has been pointed out that some approximations under a high dimensional framework are more useful in the comparison with the ones in the usual large sample framework. This gives a strong motivation for studying the bias correction problem in a high dimensional framework.

3. Reduction of the Bias Term

In this section we attempt to give a reduction for the bias term b_k in (2.9). Let $\bar{\boldsymbol{y}}^{(i)}$ and $\bar{\boldsymbol{y}}$ be the sample mean vectors of the observations of the i-th groups and all the groups, respectively. Further, let W and B be the matrices of sums of squares and products due to within groups and between groups, respectively. Then, letting $T = W + B$ and expressing partitions of these matrices in the same way as in (2.4), we can write the MLE $\hat{\Theta}_k$ of Θ under M_k (see, e.g., Fujikoshi (1985)) as

$$\begin{aligned} \bar{\boldsymbol{\mu}}_1^{(i)} &= \bar{\boldsymbol{y}}_1^{(i)}, \quad \bar{\boldsymbol{\mu}}_{2\cdot1}^{(i)} = \bar{\boldsymbol{y}}_2 - \hat{\Gamma}\bar{\boldsymbol{y}}_1, \\ \hat{\Gamma} &= T_{21}T_{11}^{-1}, \quad N\hat{\Sigma}_{11} = W_{11}, \\ N\hat{\Sigma}_{22\cdot1} &= T_{22\cdot1} = T_{22} - T_{21}T_{11}^{-1}T_{12}, \end{aligned} \tag{3.1}$$

and hence

$$\begin{aligned} -2\log f(Y;\hat{\Theta}_k) &= -N\log|N^{-1}W_{11}| \\ &\quad + N\log|N^{-1}T_{22\cdot1}| + Np(1+\log 2\pi) \\ &= -N\log\{|W_{22\cdot1}|/|T_{22\cdot1}|\} \\ &\quad + N\log|N^{-1}W| + Np(1+\log 2\pi), \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} b_{k0} &= 2\{k(q+1)+p-k+\frac{1}{2}p(p+1)\} \\ &= -2(p-k)q+2\{p(q+1)+\frac{1}{2}p(p+1)\}. \end{aligned} \tag{3.3}$$

For simplicity, we denote the true parameters $\boldsymbol{\mu}^{(i)*}$ and Σ^* by $\boldsymbol{\mu}^{(i)}$ and Σ, respectively. We can write

$$\begin{aligned} b_k &= \mathrm{E}_Y^* \mathrm{E}_X^* \left[-\sum_{i=1}^{q+1}\sum_{j=1}^{N_i} \mathrm{tr}\hat{\Sigma}^{-1}(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})' \right. \\ &\quad \left. +\sum_{i=1}^{q+1}\sum_{j=1}^{N_i} \mathrm{tr}\hat{\Sigma}^{-1}(\boldsymbol{x}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{x}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})' \right] \\ &= \mathrm{E}_Y^* \left[-\sum_{i=1}^{q+1}\sum_{j=1}^{N_i} \mathrm{tr}\hat{\Sigma}^{-1}(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})' \right. \\ &\quad \left. +N\mathrm{tr}\hat{\Sigma}^{-1}\Sigma+\sum_{i=1}^{q+1} N_i(\boldsymbol{\mu}^{(i)}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{\mu}^{(i)}-\hat{\boldsymbol{\mu}}^{(i)})' \right]. \end{aligned} \tag{3.4}$$

Note that the following identities hold:

$$\begin{aligned} &\sum_{i=1}^{q+1}\sum_{j=1}^{N_i} \mathrm{tr}\hat{\Sigma}^{-1}(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}^{(i)})' = Np, \\ &\mathrm{tr}\hat{\Sigma}^{-1}\Sigma = \mathrm{tr}\hat{\Sigma}_{11}^{-1}\Sigma_{11}+\mathrm{tr}\hat{\Sigma}_{22\cdot 1}^{-1}(-\hat{\Gamma}\ I)\Sigma(-\hat{\Gamma}\ I)', \\ &\sum_{i=1}^{q+1} \mathrm{tr}\hat{\Sigma}^{-1}N_i(\boldsymbol{\mu}^{(i)}-\hat{\boldsymbol{\mu}}^{(i)})(\boldsymbol{\mu}^{(i)}-\hat{\boldsymbol{\mu}}^{(i)})' \\ &\quad = \mathrm{tr}\hat{\Sigma}_{11}^{-1}\sum_{i=1}^{q+1} N_i(\bar{\boldsymbol{y}}_{11}^{(i)}-\boldsymbol{\mu}_1^{(i)})(\bar{\boldsymbol{y}}_{11}^{(i)}-\boldsymbol{\mu}_1^{(i)})' \\ &\quad + N\mathrm{tr}\hat{\Sigma}_{22\cdot 1}^{-1}(-\hat{\Gamma}\ I)\{(\bar{\boldsymbol{y}}-\bar{\boldsymbol{\mu}})(\bar{\boldsymbol{y}}-\bar{\boldsymbol{\mu}})'+\Omega\}(-\hat{\Gamma}\ I)', \end{aligned}$$

where

$$\bar{\boldsymbol{\mu}} = \sum_{i=1}^{q+1}(N_i/N)\boldsymbol{\mu}^{(i)}, \quad \Omega = \sum_{i=1}^{q+1}(N_i/N)(\boldsymbol{\mu}^{(i)}-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(i)}-\bar{\boldsymbol{\mu}})'.$$

Further, note that

$$\mathrm{tr}\hat{\Sigma}_{22\cdot 1}^{-1}(-\hat{\Gamma}\ I)\Omega(-\hat{\Gamma}\ I)' = N\{\mathrm{tr}T^{-1}\Omega-\mathrm{tr}T_{11}^{-1}\Omega_{11}\}.$$

These imply the following result.

THEOREM 3.1. *Let b_k be the bias term defined by* (2.9) *for the estimator* $-2\log f(X;\hat{\Theta}_k)$ *of the risk* R_k *in* (2.6). *Let* $T = W + B$, *where* W *and* B *are the matrices of sums of squares and products due to within groups and between groups, respectively. Then we can write* b_k *as*

$$\begin{aligned} b_k &= -Np \\ &+ \mathrm{E}_Y^* \left[N^2 \mathrm{tr} W_{11}^{-1} \left\{ \Sigma_{11} + \sum_{i=1}^{q+1} (N_i/N)(\bar{\boldsymbol{y}}_1^{(i)} - \boldsymbol{\mu}_1^{(i)})(\bar{\boldsymbol{y}}_1^{(i)} - \boldsymbol{\mu}_1^{(i)})' \right\} \right. \\ &+ N^2 \mathrm{tr} T^{-1} \{ \Sigma + \Omega + (\bar{\boldsymbol{y}} - \bar{\boldsymbol{\mu}})(\bar{\boldsymbol{y}} - \bar{\boldsymbol{\mu}})' \} \qquad (3.5) \\ &- \left. N^2 \mathrm{tr} T_{11}^{-1} \{ \Sigma_{11} + \Omega_{11} + (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{\mu}}_1)(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{\mu}}_1)' \} \right], \end{aligned}$$

where E^* *denotes the expectation under the true model* M^* *and the true parameters* $\boldsymbol{\mu}^{(i)*}$ *and* Σ^* *are simply denoted by* $\boldsymbol{\mu}^{(i)}$ *and* Σ, *respectively.*

THEOREM 3.2. *Suppose that the true model* M^* *is normal. Then, under the same notation as in Theorem* 3.1 *we can write* b_k *as*

$$\begin{aligned} b_k &= -Np + \frac{Nk(N+q+1)}{N-k-q-2} + \mathrm{E}_Y^* \left[N^2 \mathrm{tr} T^{-1} \left\{ (1+N^{-1})\Sigma + \Omega \right\} \right. \\ &\quad \left. - N^2 \mathrm{tr} T_{11}^{-1} \left\{ (1+N^{-1})\Sigma_{11} + \Omega_{11} \right\} \right]. \qquad (3.6) \end{aligned}$$

PROOF. From the assumption of normality, W is independent of $\{\bar{\boldsymbol{y}}^{(1)}, ..., \bar{\boldsymbol{y}}^{(q+1)}\}$, and T is independent of $\bar{\boldsymbol{y}}$. Further, since W_{11} is distributed as a Wishart distribution $W_k(N-q-1, \Sigma_{11})$, we have

$$\mathrm{E}(W_{11}^{-1}) = (N-k-q-2)^{-1}\Sigma_{11}^{-1}. \qquad (3.7)$$

Using these facts and (3.5) it is easy to get (3.6).

There are some works on bias corrections of AIC in other models (see, e.g., Sugiura (1978), Hurvich and Tsai (1989), Fujikoshi and Satoh (1997)). However, most of the corrections have been considered under assumption that a candidate model involves the true model. We note that this assumption is not imposed in Theorems 3.1 and 3.2.

4. High Dimensional Case

It is possible to derive an asymptotic expansion of the b_k in (3.6) under the usual large sample framework. Such results have been essentially obtained in Fujikoshi (1985). However, there is a difficulty in deriving an asymptotic behaviour for multiple groups case under a high dimensional framework. In this section we consider the two-groups case where the true model is normal. Then, from Theorem 3.2 we can write the bias term as

$$b_k = -Np + \frac{Nk(N+2)}{N-k-3} + N^2\{Q(N,p,\tau^2) - Q(N,k,\tau_1^2)\}, \tag{4.1}$$

where

$$Q(N,p,\tau^2) = \mathrm{E}[\mathrm{tr}(S+\boldsymbol{u}\boldsymbol{u}')^{-1}\{(1+N^{-1})I_p + \boldsymbol{\nu}\boldsymbol{\nu}'\}],$$
$$Q(N,k,\tau_1^2) = \mathrm{E}[\mathrm{tr}(S_{11}+\boldsymbol{u}_1\boldsymbol{u}_1')^{-1}\{(1+N^{-1})I_k + \boldsymbol{\nu}_1\boldsymbol{\nu}_1'\}].$$

Here $\boldsymbol{u}$ and S are independent and have a normal distribution $N_p(\boldsymbol{\nu}, I_p)$ and a Wishart distribution $W_p(N-2, I_p)$, respectively, $\boldsymbol{\nu} = (\sqrt{N_1N_2}/N)(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}), \tau^2 = \boldsymbol{\nu}'\boldsymbol{\nu} = \{(N_1N_2)/N^2\}\Delta^2$ and

$$\Delta^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}).$$

Similarly $S_{11}, \boldsymbol{u}_1$ and $\boldsymbol{\nu}_1$ denote the first k parts of $S, \boldsymbol{u}$ and $\boldsymbol{\nu}$, respectively, and $\tau_1^2 = \boldsymbol{\nu}_1'\boldsymbol{\nu}_1$. Since $Q(N,k,\tau_1^2)$ is obtained from $Q(N,p,\tau^2)$, we consider to evaluate $Q(N,p,\tau^2)$. Note that

$$(S+\boldsymbol{u}\boldsymbol{u}')^{-1} = S^{-1} - (1+\boldsymbol{u}'S^{-1}\boldsymbol{u})^{-1}S^{-1}\boldsymbol{u}\boldsymbol{u}'S^{-1}.$$

Using this identity and (3.7) we have

$$\begin{aligned} Q(N,p,\tau^2) &= \frac{1}{N-p-3}\{(1+N^{-1})p + \tau^2\} \\ &- (1+N^{-1})\mathrm{E}\left[(1+\boldsymbol{u}'S^{-1}\boldsymbol{u})^{-1}\boldsymbol{u}'S^{-2}\boldsymbol{u}\right] \\ &- \mathrm{E}\left[(1+\boldsymbol{u}'S^{-1}\boldsymbol{u})^{-1}(\boldsymbol{\nu}'S^{-1}\boldsymbol{u})^2\right]. \end{aligned} \tag{4.2}$$

In order to evaluate the expectations in (4.2), we use the following stochastic expressions.

LEMMA 4.1. *Suppose that $\boldsymbol{u}$ and S are independently distributed as a normal distribution $N_p(\boldsymbol{\nu}, I_p)$ and a Wishart distribution $W_p(N-2, I_p)$, respectively. Then, the following stochastic expressions hold*:

$$\begin{aligned} \boldsymbol{u}'S^{-1}\boldsymbol{u} &= v_1/v_2, \\ \boldsymbol{u}'S^{-2}\boldsymbol{u} &= (v_1/v_2^2)\{1 + v_3/v_4\}, \\ \boldsymbol{\nu}'S^{-1}\boldsymbol{u} &= (\tau/v_2)\left[z_1 + \tau + \{v_0v_3/v_4\}^{1/2} z_2/\sqrt{v_5+z_2^2}\right], \end{aligned} \tag{4.3}$$

where z_1, z_2 are standard normal variates, $v_i, i = 0, 2, 3, 4$ are chi-squared variates with f_i degrees of freedom, they are all independent, and $v_1 = v_0 + (z_1 + \tau)^2$, so v_1 is distributed as a noncentral chi-squared variate with f_1 degrees of freedom and non-centrality parameter τ^2. Here $f_i, i = 0, 1, ..., 5$ are given by

$$f_0 = p - 1, \quad f_1 = p, \quad f_2 = N - p - 1,$$
$$f_3 = p - 1, \quad f_4 = N - p, \quad f_5 = p - 2.$$

PROOF. Deev (1970) has given similar results, but his third expression is slightly different and involves an error (see Remark 4.1). Further, he gives only an outline of the proof. Here we give a complete proof. This proof uses the method based on a random orthogonal matrix H whose first column is proportional to $\boldsymbol{u}$. Note that $W = H'SH$ has the same distribution with S, and is independent of $\boldsymbol{u}$. Let W be partitioned as

$$W = \begin{bmatrix} w_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{21} & W_{22} \end{bmatrix},$$

where $w_{11} : 1 \times 1$. Then we can see that

$$\begin{aligned} \boldsymbol{u}'S^{-1}\boldsymbol{u} &= \boldsymbol{u}'\boldsymbol{u}/w_{11\cdot 2}, \\ \boldsymbol{u}'S^{-2}\boldsymbol{u} &= \boldsymbol{u}'\boldsymbol{u}(1 + \boldsymbol{w}_{12}W_{22}^{-2}\boldsymbol{w}_{21})/w_{11\cdot 2}^2, \end{aligned}$$

where $w_{11\cdot 2} = w_{11} - \boldsymbol{w}_{12}W_{22}^{-1}\boldsymbol{w}_{21}$. The first two results are obtained by using the fact (see, e.g., Siotani et al. (1985)) that $w_{11\cdot 2} \sim \chi^2_{N-p-1}$, $W_{22} \sim W_{p-1}(N - 2, I_{p-1})$, $\boldsymbol{W}_{22}^{-1/2}\boldsymbol{w}_{21} \sim N_{p-1}(0, I_{p-1})$ and they are independent. Let H be decomposed as $H = [(\boldsymbol{u}'\boldsymbol{u})^{-1/2}\boldsymbol{u} \quad H_2]$. Then

$$\boldsymbol{\nu}'S^{-1}\boldsymbol{u} = \tau\{z_1 + \tau + \boldsymbol{x}_1'\boldsymbol{x}_2\}/w_{11\cdot 2},$$

where

$$\begin{aligned} z_1 &= \tilde{\boldsymbol{\nu}}'(\boldsymbol{u} - \boldsymbol{\nu}), \quad \tilde{\boldsymbol{\nu}} = \tau^{-1}\boldsymbol{\nu}, \\ \boldsymbol{x}_1 &= -(\boldsymbol{u}'\boldsymbol{u})^{1/2}H_2'\tilde{\boldsymbol{\nu}}, \quad \boldsymbol{x}_2 = W_{22}^{-1}\boldsymbol{w}_{21}. \end{aligned}$$

Further, we have

$$v_1 = \boldsymbol{u}'\boldsymbol{u} = (\boldsymbol{u} - \boldsymbol{\nu})'\{I_p - \tilde{\boldsymbol{\nu}}\tilde{\boldsymbol{\nu}}'\}(\boldsymbol{u} - \boldsymbol{\nu}) + (z_1 + \tau)^2 = v_0 + (z_1 + \tau)^2.$$

Here v_0 and z_1 are independent, and $v_0 \sim \chi^2_{p-1}$. Note that

$$\boldsymbol{x}_1'\boldsymbol{x}_1 = v_0, \quad \boldsymbol{x}_2'\boldsymbol{x}_2 = \boldsymbol{w}_{12}W_{22}^{-2}\boldsymbol{w}_{21} = v_3/v_4.$$

Let

$$r = \boldsymbol{x}_1'\boldsymbol{x}_2/\{\boldsymbol{x}_1'\boldsymbol{x}_1 \cdot \boldsymbol{x}_2'\boldsymbol{x}_2\}^{1/2}. \tag{4.4}$$

Since the distribution of $\boldsymbol{x}_2$ is spherical, from Theorem 1.5.7 in Muirhead (1982) we get that $\sqrt{p-2}r/\sqrt{1-r^2}$ is distributed as a t-distribution with $p-2$ degrees of freedom, and r is independent of $\boldsymbol{x}_1'\boldsymbol{x}_1$ and $\boldsymbol{x}_2'\boldsymbol{x}_2$. Therefore we can write r as

$$r = z_2/\sqrt{v_5 + z_2^2},$$

which completes the proof.

REMARK 4.1. In Deev (1970) the term r in (4.4) is denoted by $\sin\theta$. However, the probability density function of θ should be read as

$$\frac{\Gamma(\frac{1}{2}(p-1))}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}(p-2))}(1-\sin^2\theta)^{\frac{1}{2}(p-3)} \quad \left(-\frac{\pi}{2} \le \theta \le \frac{\pi}{2}\right).$$

Fujikoshi and Seo (1998) has derived a stochastic expression when $\boldsymbol{\nu}$ is an independent normal variate.

Applying Lemma 4.1 to the expectations in (4.2) we can get

$$\begin{aligned} \mathrm{E}\left[\frac{\boldsymbol{u}'S^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'S^{-1}\boldsymbol{u}}\right] &= \mathrm{E}\left[\left\{\frac{1}{v_2} - \frac{1}{v_1+v_2}\right\}\left\{1+\frac{v_3}{v_4}\right\}\right] \\ &= \frac{N-3}{N-p-2}\left\{\frac{1}{N-p-3} - \mathrm{E}\left[\frac{1}{v_1+v_2}\right]\right\}, \end{aligned} \tag{4.5}$$

$$\begin{aligned} \mathrm{E}\left[\frac{(\boldsymbol{\nu}'S^{-1}\boldsymbol{u})^2}{1+\boldsymbol{u}'S^{-1}\boldsymbol{u}}\right] &= \tau^2 E\left[\frac{(z_1+\tau)^2}{v_2(v_1+v_2)}\right] \\ &+ \frac{\tau^2}{N-p-2}\mathrm{E}\left[\frac{v_0}{v_2(v_1+v_2)}\right]. \end{aligned} \tag{4.6}$$

Now we need to evaluate the three expectations involved in (4.5) and (4.6). For this, we assume the following high dimensional framework:

$$\lim_{p\to\infty}\frac{N_i}{p} = c_i \ (i=1,2), \quad \lim_{p\to\infty}\tau^2 = c_0, \tag{4.7}$$

where $c_i (i=1,2)$ are positive constants and c_0 is a nonnegative constant. Then,

$$\mathrm{E}\left[\frac{1}{v_1+v_2}\right] = \frac{1}{N-3}\left\{1-\frac{\tau^2}{N-3}\right\} + O_3,$$

$$\begin{aligned} \mathrm{E}\left[\frac{(z_1+\tau)^2}{v_2(v_1+v_2)}\right] &= \frac{1}{(N-p-1)(N-1)} + O_3, \\ \mathrm{E}\left[\frac{v_0}{v_2(v_1+v_2)}\right] &= \frac{p-1}{(N-p-1)(N-1)} + O_2, \end{aligned} \tag{4.8}$$

where O_j denotes the term of the jth order with respect to $(N_1^{-1}, N_2^{-1}, p^{-1})$. These results can be easily extended to higher order term. For the first result, see, e.g., Lieftinck-Koeijers (1988). Substituting (4.5), (4.6) and (4.8) to (4.2), we obtain

$$Q(N,p,\tau^2) = Q_1(N,p,\tau^2) + O_3, \tag{4.9}$$

where

$$\begin{aligned} Q_1(N,p,\tau^2) &= \frac{p(N+1)}{N(N-p-2)} \\ &+ \tau^2\left\{\frac{1}{N-p-3} - \frac{N+1}{N(N-3)(N-p-2)}\right\} \\ &- \frac{\tau^2}{(N-p-1)(N+1)}\left\{\tau^2 + \frac{N-2}{N-p-2}\right\}. \end{aligned}$$

Similarly we can get an asymptotic expansion of $Q(N,k,\tau_1^2)$ which is obtained from the one for $Q(N,p,\tau^2)$ by substituting k and τ_1^2 to p and τ^2, respectively. Summarizing these results we have the following Theorem.

THEOREM 4.1. *Let b_k be the bias term defined by* (2.9) *in two-groups case, i.e., $q=1$. Suppose that the true model M^* is normal. Then, under the assumption of* (4.7) *we can obtain an asymptotic expansion of b_k with an error of O_1 given by*

$$b_k = b_{k1} + O_1, \tag{4.10}$$

where

$$\begin{aligned} b_{k1} &= \frac{N}{N-k-3}\{-(p-k)N + pk + 3p + 2k\} \\ &+ N^2\{Q_1(N,p,\tau^2) - Q_1(N,k,\tau_1^2)\}. \end{aligned}$$

and $Q_1(N,p,\tau^2)$ is given by (4.9).

COROLLARY 4.1. *If M_k involves the true model M^*, i.e., $\tau^2 = \tau_1^2$, then*

$$\begin{aligned} b_{k1} &= b_{k0} + (p-k)\tau^2 + \mathrm{O}(N^{-1}) \\ &= b_{k0} + \mathrm{O}(N^{-1}), \quad \text{if } \tau^2 = \mathrm{O}(N^{-1}), \end{aligned} \tag{4.11}$$

where $b_{k0} = 2\{2k + (p-k) + \frac{1}{2}p(p+1)\}$.

The constant b_{k0} in Corollary 4.1 is the same as in (2.7). Therefore, it may be noted that b_{k1} is useful in the usual large sample situation as well as the high dimensional situation. There are many equivalent expressions for $Q_1(N, p, \tau^2)$ and $Q_1(N, k, \tau_1^2)$. However, we do not attempt to obtain the best choice. The quantity b_{k1} in (4.10) involves the unknown parameters τ^2 and τ_1^2, or equivalently Δ^2 and Δ_1^2. In practice, we suggest to replace these by the following unbiased estimators (see, e.g., Siotani et al. (1985)):

$$\hat{\Delta}^2 = \frac{N-p-3}{N-2}D^2 - \frac{Np}{N_1N_2},$$

$$\hat{\Delta}_1^2 = \frac{N-k-3}{N-2}D_1^2 - \frac{Nk}{N_1N_2},$$

where D^2 and D_1^2 are the sample Mahalanobis squared distance between two populations based on the full variate $\boldsymbol{y}$ and the first k variate $\boldsymbol{y}_1$, respectively. Let $\hat{b}_{k1}$ be the one ontained from b_{k1} by substituting $\hat{\Delta}^2$ and $\hat{\Delta}_1^2$ to Δ^2 and Δ_1^2, respectively. As a modification of AIC_k we suggest

$$MAIC_k = -2\log f(Y; \hat{\Theta}_k) + \hat{b}_{k1}. \tag{4.12}$$

From (3.2) we have

$$\begin{aligned} -2\log f(Y; \hat{\Theta}_k) &= -N\log\left\{1 + \frac{D_1^2 - D^2}{N(N-2)(N_1N_2)^{-1} + D^2}\right\} \\ &+ \log|N^{-1}W| + Np(1 + \log 2\pi), \end{aligned} \tag{4.13}$$

where $(N-2)^{-1}W$ is the usual pooled sample covariance matrix. The expression (4.11) is defined for $K = \{1, 2.\ldots.k\}$. However, it is easy to see that the expression (4.11) can be extended for any subset K in P.

Acknowledgement. The author would like to thank two referees for several valuable comments which lead to highlighting the main purpose of the paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B.N. Petrov and F. Csáki, eds., Akadémia Kiado, Budapest, 267-281.

Deev, A.D. (1970). Representation of statistics of discriminant analysis and asymptotic expansions when space dimensions are comparable with sample size. *Soviet Math. Dokl.* **11**, 1547-1550.

Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In *Multivariate Analysis-VI*, P.R. Krishnaiah, ed., Elsevier Science Publishers, Amsterdam, 219-236.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear model. *Biometrika*, **84**, 707-716.

Fujikoshi, Y. and Seo, T. (1998). Asymptotic approximations for EPMC's of the linear and the quadratic discriminant functions when the samples sizes and the dimension are large. *Statist. Anal. Random Arrays* **6** , 269-280.

Hurvich, C.M. and Tsai, C.L. (1989). Regression and times series model selection in small samples. *Biometrika* **76**, 297-307.

Lieftinck-Koeijers, C.A.J. (1988). Multivariate calibration: A generalization of the classical estimator. *J. Multivariate Anal.* **25**, 31-44.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

Rao, C.R. (1948). Tests of significance in multivariate analysis. *Biometrika* **35**, 58-79.

Rao, C.R. (1970). Inference on discriminant function coefficients. In *Essays in Prob. and Statist.* R.C. Bose, ed., Univ. of North Carolina Press, Chapel Hill, 537-602.

Raudys, S. (1972). On the amount of priori information in designing the classification algorithm. *Engrg. Cybernetics* **10**, 711-718.

Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Analysis: A Graduate Course and Handbook.* American Sciences Press, Ohio.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods***7**, 13-26.

Wyman, F.J., Young, D.M. and Turner, D.W. (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition* **23**, 775-783.

Yasunori Fujikoshi
Department of Mathematics
Graduate School of Science
Hiroshima University
Higashi-Hiroshima 739-8526
Japan
E-mail: fuji@math.sci.hiroshima-u.ac.jp