CIRJE-F-872

A Variable Selection Criterion for Linear Discriminant Rule and its Optimality in High Dimensional Setting

Masashi Hyodo
Graduate School of Economcis, University of Tokyo

Tatsuya Kubokawa
University of Tokyo

December 2012; Revised in February 2013

A Variable Selection Criterion for Linear Discriminant Rule and its Optimality in High Dimensional Setting

Masashi Hyodo*and Tatsuya Kubokawa†
University of Tokyo

February 4, 2013

Abstract

In this paper, we suggest the new variable selection procedure, called MEC, for linear discriminant rule in the high-dimensional setup. MEC is derived as a second-order unbiased estimator of the misclassification error probability of the linear discriminant rule. It is shown that MEC not only decomposes into 'fitting' and 'penalty' terms like AIC and Mallows C_p, but also possesses an asymptotic optimality in the sense that MEC achieves the smallest possible conditional probability of misclassification in candidate variable sets. Through simulation studies, it is shown that MEC has good performances in the sense of selecting the true variable sets.

Key words and phrases: asymptotic optimality, high dimension, linear discriminant analysis, misclassification error, multivariate normal, second-order approximation, variable selection.

1 Introduction

In this paper, we consider the problem of classifying a future observation vector into one of the two population groups Π_1 and Π_2. For each $i = 1, 2$, Π_i denotes a population from a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_i, \Sigma)$, and it is supposed that $\boldsymbol{x}_{ij}$, $j = 1, \ldots, N_i$, are observed from the population Π_i. Here, $\boldsymbol{\mu}_i$, $i = 1, 2$, and Σ are unknown parameters, and they are estimated by the sample mean $\overline{\boldsymbol{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$, $i = 1, 2$, and the pooled sample covariance matrix $S = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)'$ for $n = N_1 + N_2 - 2$. When the dimension p is large, the model involves many unknown parameters, which causes a large misclassification error in the linear discriminant rule (LDR). Thus, it is desired to find an optimal subset of variables in LER. Variable selection methods for discriminant analysis have been studied by Fujikoshi (1985, 2002), Sakurai, Nakada and

*Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: hyodohh@yahoo.co.jp

†Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

Fujikoshi (2012), Wilbur, Ghosh, Nakatsu and Doerge (2002) and others. Related to this issue, multiple testing problems for no additional information have been discussed by Rao (1948, 1970, 1973) and Kshirsagar (1972). In this paper, we suggest a new variable selection procesure based on error of misclassification and establish the optimality in high-dimensional situation.

To explain the new variable selection procesure, consider the following linear discriminant rule. Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be a future observation in the full model. Let $\boldsymbol{j} = (j_1, \ldots, j_{k(j)})$ be a subset of the set $\{1, 2, \ldots, p\}$, and let $\boldsymbol{x}(\boldsymbol{j}) = (x_{j_1}, \ldots, x_{j_{k(j)}})$ be the corresponding subvector of $\boldsymbol{x}$. The model based on the variable $\boldsymbol{x}(\boldsymbol{j})$ is denoted by $\boldsymbol{j}$. Let $\mathcal{J}$ be a suitable family of subsets of $\{1, \ldots, p\}$. The LDR for classifying $\boldsymbol{x}$ based on the model j is that $\boldsymbol{x}$ is classified as coming from Π_1, if $W(\boldsymbol{j}) > \alpha$, and from Π_2, if $W(\boldsymbol{j}) < \alpha$, where α is cut off point for classification rule, and

$$W(\boldsymbol{j}) = (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \{\boldsymbol{x}(\boldsymbol{j}) - \tfrac{1}{2}(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) + \bar{\boldsymbol{x}}_2(\boldsymbol{j}))\}.$$

Here, $\bar{\boldsymbol{x}}_i(\boldsymbol{j})$, $i = 1, 2$, and $S(\boldsymbol{j})$ are the sample mean and the pooled sample covariance matrix in the model $\boldsymbol{j}$. Then the problem of variable selection in LDR is regarded as how to select the best subset $\boldsymbol{j}$ from $\mathcal{J}$. To this end, we consider the conditional error probabilities of misallocation $L_1(\boldsymbol{j}) = P[W(\boldsymbol{j}) < \alpha | \boldsymbol{x}(\boldsymbol{j}) \in \Pi_1, \overline{\boldsymbol{x}}_1(\boldsymbol{j}), \overline{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j})]$ and $L_2(\boldsymbol{j}) = P[W(\boldsymbol{j}) > \alpha | \boldsymbol{x}(\boldsymbol{j}) \in \Pi_2, \overline{\boldsymbol{x}}_1(\boldsymbol{j}), \overline{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j})]$, which can be expressed as

$$L_g(\boldsymbol{j}) = \Phi\left((-1)^g \frac{(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \{\boldsymbol{\mu}_g(\boldsymbol{j}) - (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) + \bar{\boldsymbol{x}}_2(\boldsymbol{j}))/2\}}{\sqrt{(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \Sigma(\boldsymbol{j}) S(\boldsymbol{j})^{-1} (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))}} \right), \tag{1.1}$$

for $g = 1, 2$, where $\Phi(\cdot)$ is the standard normal distribution function, and $\boldsymbol{\mu}_g(\boldsymbol{j})$ and $\Sigma(\boldsymbol{j})$ denote the population mean and covariance matrix in the model j. When π_i, $i = 1, 2$ is a prior probability of the group membership, the expected error rate is given by

$$R(\boldsymbol{j}) = \pi_1 R_1(\boldsymbol{j}) + \pi_2 R_2(\boldsymbol{j})$$

where $R_g(\boldsymbol{j})$ is the unconditional error of misallocation given by $R_g(\boldsymbol{j}) = \mathrm{E}[L_g(\boldsymbol{j})]$ for $g = 1, 2$. The variable selection procedure proposed in this paper is an asymptotically unbiased estimator of the misclassification error $R(\boldsymbol{j})$ in the high-dimansional setteing.

A naive procesure for selection of variables is to minimize

$$\Phi\left(-D(\boldsymbol{j})/2\right), \tag{1.2}$$

with respect to $\boldsymbol{j} \in \mathcal{J}$, where $D(\boldsymbol{j})$ is the sample Mahalanobis distance based on $\boldsymbol{x}(\boldsymbol{j})$, namely,

$$D(\boldsymbol{j}) = (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j})). \tag{1.3}$$

However, $\Phi\left(-D(\boldsymbol{j})/2\right)$ has the bias $R(\boldsymbol{j}) - \mathrm{E}[\Phi(-D(\boldsymbol{j})/2)]$ which is not negligible. MacLachlan (1976,1980) derived a second order asymptotic unbiased estimator of $R(\boldsymbol{j})$ under the large sample framework, namely,

(A0): $n \to \infty$, but p is bounded.

Fujikoshi (1985) applied the estimator given by MacLachlan (1976,1980) to the variable selection problem, and investigated the asymptotic properties and the relationship with AIC. In the high-dimensional setting On the other hand, Raudys (1972) and Wyman, Young and Turner (1990) derived the asymptotic approximations of the error probability under the high-dimensional setting given by

(A1) : $(n, p) \to \infty$ with $p/n \to c_0 \in [0, 1)$.

It is known that these approximations are also good in the large sample situation (A0) as seen from Fujikoshi, Ulyanov and Shimizu (2009).

In this paper, we derive a second-order unbiased estimator of $R(\boldsymbol{j})$ in the high-dimensional setting (A1). The unbiased estimator is here called the *Misclassification Error Criterion* (MEC), which is useful for selecting variables in linear discriminant rule. We show that MEC can be decomposed into the 'fitting' and 'penalty' terms, namely,

$$\text{MEC} = \Phi\left(-D(\boldsymbol{j})/2\right) + (\text{penalty}),$$

where the penalty term increases in the number of the unknown parameres. This is a desirable property that variable selection procedures like AIC and C_p should possess. We also show that MEC has an asymptotical optimality as a variable selection procedure in (A1). Such an optimality in the high-dimensional setting is not known as long as we know.

Recently, Kubokawa, Hyodo and Srivastava (2013) derived a second-order approximation of the error probability of misclassification (EPMC) for the ridge-type linear discriminant rule in high dimensional setting, and derived a second-order unbiased estimator of EPMC. Since the ridge-type linear discriminant rule is not invariant under scale transformations, their approach needs to calculate various kinds of fourth moments of the inverted Wishart matrix. It was hard to obtain such fourth moments, so that the approach used by Kubokawa, *et al.* (2013) cannot be used for developing an asymptotic optimality of MEC. Instead, the method used in this paper is to express $L_g(\boldsymbol{j})$ based on nine primitive random variables, namely four random variables having the standard normal distribution and five random variables having chi-square distributions. This approach not only makes it easier to derive the second-order approximation and the second-order unbiased estimator of $R(\boldsymbol{j})$, but also enables us to establish the asymptotic optimality of MEC as a varaible selection procedure in both high-dimensional and large-sample situations.

The organization of this paper is as follows. In Section 2.1, we determine the asymptotically optimal cut off point α based on the expected error rate $R(\boldsymbol{j})$. In Section 2.2, we derive the second order unbiased estimator of $R(\boldsymbol{j})$ in high dimensional setting and propose the new variable selection procedure MEC based on this estimator. In Section 3.1, we show that MEC can be decomposed into the "fitting term" and "dimensionality penalty term" like Mallows C_p and AIC. In Section 3.2, we prove the asymptotic optimality of MEC. In Section 4, we investigate performances of MEC through numerical studies. The conclusion of our study is summarized in Section 5. Some preliminary results are given in Appendix.

2 MEC: A Variable Selection Procedure for High Dimensional Data

In this Section, we derive the second order asymptotic unbiased estimator of $R(\boldsymbol{j})$ under the following high dimensional framework (A1)-(A3).

2.1 The optimal cut off point based on expected probability of misclassification

An optimal rule of allocation can be defined by taking it to be one that minimizes the expected error rate $R(\boldsymbol{j})$. However, the expected error rate $R(\boldsymbol{j})$ is not explicit formula. So, we determine the optimal cut off point based on limiting value of $R(\boldsymbol{j})$. We assume the following conditions, in order to derive limiting value of $R(\boldsymbol{j})$.

(A1) : $(n,\ p) \to \infty$ with $p/n \to c_0 \in [0, 1)$.

(A2) : $(n,\ N_1,\ N_2) \to \infty$ with $N_1/n \to \gamma_1,\ N_2/n \to \gamma_2$.

(A3) : $\lim_{p\to\infty} \Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \infty$.

Let $\sharp(\boldsymbol{j})$ be the cardinality of $\boldsymbol{j}$. In what follows, we treat the case that $(\sharp(\boldsymbol{j}),\ n) \to \infty$ with $\sharp(\boldsymbol{j})/n \to c \in [0, 1)$, but the derived results include the case of $\sharp(\boldsymbol{j})$ bounded. We primarily consider the asymptotic approximation for the expected error rate of $\mathrm{E}[L_1(\boldsymbol{j})]$, since the asymptotic approximations for the expected error rate of $\mathrm{E}[L_2(\boldsymbol{j})]$ can be constructed similarly.

Suppose that $\boldsymbol{x} \in \Pi_1$. Then, a conditional distribution given $(\overline{\boldsymbol{x}}_1(\boldsymbol{j}), \overline{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j}))$ is written as $W(\boldsymbol{j})|(\overline{\boldsymbol{x}}_1(\boldsymbol{j}), \overline{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j})) \sim \mathcal{N}_p(-U, V)$, where

$$\begin{aligned} U =& (\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \overline{\boldsymbol{x}}_2(\boldsymbol{j}))'S(\boldsymbol{j})^{-1}(\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \boldsymbol{\mu}_1(\boldsymbol{j})) - 2^{-1}D(\boldsymbol{j}), \\ V =& (\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \overline{\boldsymbol{x}}_2(\boldsymbol{j}))'S(\boldsymbol{j})^{-1}\Sigma(\boldsymbol{j})S(\boldsymbol{j})^{-1}(\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \overline{\boldsymbol{x}}_2(\boldsymbol{j})), \end{aligned}$$

where $D(\boldsymbol{j})$ is given in (1.3). Then, the expected error rate of $W(\boldsymbol{j})$ can be expressed as

$$\mathrm{E}[L_1(\boldsymbol{j})] = \mathrm{E}\left[\Phi\left(\frac{U+\alpha}{\sqrt{V}}\right)\right],$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable.

As given in (6.1) and (6.2), U and V can be expanded as $U = U_0 + n^{-1/2}U_1 + n^{-1}U_2$ and $V = V_0 + n^{-1/2}V_1 + n^{-1}V_2$ with

$$U_0 = -\frac{1}{2(1-c)}\left(\Delta(\boldsymbol{j})^2 + \frac{c(\gamma_1-\gamma_2)}{\gamma_1\gamma_2}\right), \quad V_0 = \frac{1}{(1-c)^3}\left(\frac{c}{\gamma_1\gamma_2} + \Delta(\boldsymbol{j})^2\right),$$

$U_1 = O_p(1),\ U_2 = O_p(1),\ V_1 = O_p(1)$ and $V_2 = O_p(1)$. Then we observe that

$$\frac{U+\alpha}{V^{1/2}} = \{U_0 + \frac{U_1}{\sqrt{n}} + \frac{U_2}{n} + \alpha\}\frac{1}{V_0^{1/2}}\Big\{1 - \frac{1}{2V_0}\left(\frac{V_1}{\sqrt{n}} + \frac{V_2}{n}\right) + \frac{3V_1^2}{8nV_0^2}\Big\} + O_p(n^{-3/2}), \quad (2.1)$$

which gives the expansion

$$(U+\alpha)/V^{1/2} = w_0 + w_1 + w_2 + O_p(n^{-3/2}),$$

where

$$\begin{aligned}
w_0 =& V_0^{-1/2}(U_0+\alpha),\\
w_1 =& \frac{1}{\sqrt{n}V_0^{1/2}}\Big\{U_1 - \frac{U_0+\alpha}{2V_0}V_1\Big\},\\
w_2 =& \frac{1}{nV_0^{1/2}}\Big\{U_2 - \frac{U_0+\alpha}{2V_0}V_2 + \frac{3(U_0+\alpha)}{8V_0^2}V_1^2 - \frac{1}{2V_0}U_1V_1\Big\}.
\end{aligned}$$

Using the Taylor series expansion again, we can approximate $E[L_1(\boldsymbol{j})]$ as

$$\begin{aligned}
\mathrm{E}[L_1(\boldsymbol{j})] =& \mathrm{E}[\Phi(w_0 + (w_1+w_2))]\\
=& \Phi(w_0) + \phi(w_0)E[w_1 + w_2 - \frac{1}{2}w_0w_1^2] + O(n^{-3/2}),
\end{aligned}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. Letting $H = \mathrm{E}[w_1+w_2-\frac{1}{2}w_0w_1^2]$, we can write H as

$$\begin{aligned}
H =& \frac{1}{V_0^{1/2}}\left\{\frac{\mathrm{E}[U_1]}{\sqrt{n}} + \frac{\mathrm{E}[U_2]}{n}\right\} - \frac{U_0+\alpha}{2V_0^{3/2}}\left\{\frac{\mathrm{E}[V_1]}{\sqrt{n}} + \frac{\mathrm{E}[V_2]}{n}\right\} - \frac{U_0+\alpha}{2V_0^{3/2}}\frac{\mathrm{E}[U_1^2]}{n}\\
&+ \frac{U_0+\alpha}{8V_0^{5/2}}\Big(3 - \frac{(U_0+\alpha)^2}{V_0}\Big)\frac{\mathrm{E}[V_1^2]}{n} - \frac{1}{2V_0^{3/2}}\Big(1 - \frac{(U_0+\alpha)^2}{V_0}\Big)\frac{\mathrm{E}[U_1V_1]}{n}. \qquad (2.2)
\end{aligned}$$

Since U_1, U_2, V_1 and V_2 are given around (6.1) and (6.2), we can calculate the moments in (2.2), which yields the following theorem. Define H_U, H_V, H_1, H_2 and H_{12} by

$$\begin{aligned}
H_U(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& -\frac{1}{2(1-c)^2}\left(\Delta(\boldsymbol{j})^2 + \frac{c(\gamma_1-\gamma_2)}{\gamma_1\gamma_2}\right),\\
H_V(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& \frac{(c+3)\Delta(\boldsymbol{j})^2\gamma_1\gamma_2 + (5-c)c}{(c-1)^4\gamma_1\gamma_2},\\
H_1(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& \frac{\Delta(\boldsymbol{j})^4}{2(1-c)^3} + \frac{1}{(1-c)^3\gamma_2}\left(\frac{c}{\gamma_1} + \Delta(\boldsymbol{j})^2\right) + \frac{c(\gamma_1-\gamma_2)^2}{2(1-c)^3\gamma_1^2\gamma_2^2},\\
H_2(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& \frac{2(c+4)\Delta(\boldsymbol{j})^4}{(1-c)^7} + \frac{4\left\{(c+1)^2+c\right\}\Delta(\boldsymbol{j})^2}{(1-c)^7\gamma_1\gamma_2} + \frac{2c\left\{(c+1)^2+c\right\}}{(1-c)^7\gamma_1^2\gamma_2^2},\\
H_{12}(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& -\frac{2\Delta(\boldsymbol{j})^4}{(1-c)^5} - \frac{2(c+1)\Delta(\boldsymbol{j})^2}{(1-c)^5\gamma_2} - \frac{c(c+1)\left(\gamma_1^2-\gamma_2^2\right)}{(1-c)^5\gamma_1^2\gamma_2^2},
\end{aligned}$$

for $\Delta(\boldsymbol{j}) = \{\boldsymbol{\mu}_1(\boldsymbol{j}) - \boldsymbol{\mu}_2(\boldsymbol{j})\}'\Sigma(\boldsymbol{j})^{-1}\{\boldsymbol{\mu}_1(\boldsymbol{j}) - \boldsymbol{\mu}_2(\boldsymbol{j})\}$.

Theorem 2.1 *Assume the conditions* (A1)-(A3). *The second order approximation of* $\mathrm{E}[L_1(\boldsymbol{j})]$ *is given by*

$$\mathrm{E}[L_1(\boldsymbol{j})] = \Phi\left(\frac{U_0+\alpha}{\sqrt{V_0}}\right) + \frac{1}{n}\phi\left(\frac{U_0+\alpha}{\sqrt{V_0}}\right)H(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) + O(n^{-3/2}), \qquad (2.3)$$

where, $H(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c)$ *is given by*

$$\begin{aligned} H(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) =& \frac{1}{V_0^{1/2}} \Big\{ H_U(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) - \frac{U_0+\alpha}{2V_0} H_V(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) \\ & - \frac{U_0+\alpha}{2V_0} H_1(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) \\ & + \frac{U_0+\alpha}{8V_0^2} \Big(3 - \frac{(U_0+\alpha)^2}{V_0}\Big) H_2(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) \\ & - \frac{1}{2V_0} \Big(1 - \frac{(U_0+\alpha)^2}{V_0}\Big) H_{12}(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c) \Big\}. \end{aligned} \tag{2.4}$$

It is noted that the second order approximation of $\mathrm{E}[L_2(\boldsymbol{j})]$ can be also obtained by replacing $(\gamma_1, \gamma_2, \alpha)$ in (2.3) and (2.4) with $(\gamma_2, \gamma_1, -\alpha)$. Thus, it follows from Theorem 2.1 that the limiting value of $R(\boldsymbol{j})$ is given by

$$\pi_1 \Phi\left(-\frac{\Delta(\boldsymbol{j})^2 + \frac{c(\gamma_1-\gamma_2)}{\gamma_1\gamma_2} - (1-c)\alpha}{2\sqrt{1-c}\sqrt{\Delta^2(\boldsymbol{j}) + c/(\gamma_1\gamma_2)}} \right) + \pi_2 \Phi\left(-\frac{\Delta(\boldsymbol{j})^2 - \frac{c(\gamma_1-\gamma_2)}{\gamma_1\gamma_2} + (1-c)\alpha}{2\sqrt{1-c}\sqrt{\Delta^2(\boldsymbol{j}) + c/(\gamma_1\gamma_2)}} \right),$$

which can be minimized at

$$\alpha = \frac{c(\gamma_1-\gamma_2)}{2\gamma_1\gamma_2(1-c)} + \left(\frac{1}{(1-c)^2} + \frac{c}{(1-c)^2 \Delta(\boldsymbol{j})^2 \gamma_1\gamma_2} \right) \log\left(\frac{\pi_2}{\pi_1}\right). \tag{2.5}$$

It is hereafter assumed that $\pi_1 = \pi_2 = 1/2$. Then, the asymptotically optimal classification rule is given by

$$\begin{aligned} & W(\boldsymbol{j}) = (\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \overline{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \{ \boldsymbol{x}(\boldsymbol{j}) - \frac{1}{2}(\overline{\boldsymbol{x}}_1(\boldsymbol{j}) + \overline{\boldsymbol{x}}_2(\boldsymbol{j})) \} > (\text{resp.}<) \frac{c(\gamma_1-\gamma_2)}{2\gamma_1\gamma_2(1-c)} \\ & \implies \quad \boldsymbol{x} \in \Pi_1 (\text{resp.}\Pi_2). \end{aligned} \tag{2.6}$$

2.2 Derivation of MEC

We now derive a new variable selection procedure, called MEC, based on the misclassification error probability of the linear discriminant rule. As variable selection methods, Mallow's C_p and AIC are well known. Mallow's C_p is an unbiased estimator of the prediction error for a future observation relative to a quadratic loss, and AIC is motivated from minimization of prediction error relative to the Kullback-Leibler information. Recall that $L_1(\boldsymbol{j})$ given in (1.1) can be interpreted as a conditional prediction error that a future observation $\boldsymbol{x}$ from Π_1 is misclassified into Π_2. Thus, the expected error rate $R(\boldsymbol{j}) = (R_1(\boldsymbol{j}) + R_2(\boldsymbol{j}))/2$ for $\pi_1 = \pi_2 = 1/2$ is regarded as a predictive probability of misclassification, which suggests that an unbiased estimator of $R(\boldsymbol{j})$ possesses the ability to work as a variable selection procedure. In this subsection, we derive the second order unbiased estimator of $R(\boldsymbol{j})$ in the high dimensional setting (A1) or the large sample setting (A0).

Consider the linear discriminant rule (2.6) with the asymptotically optimal cut-off point α given in (2.5) for $\pi_1 = \pi_2 = 1/2$. For notational covenience, let

$$U_\alpha = U_0 + \alpha = -\frac{1}{2(1-c)} \Delta(\boldsymbol{j})^2.$$

Note that $V_0 = (1-c)^{-3}\{\Delta(\boldsymbol{j})^2 + c/(\gamma_1\gamma_2)\}$. Since the second-order expansion given in Theorem 2.1 is a function of $\Delta(\boldsymbol{j})^2$, we begin by obtaining a consistent estimator of $\Delta(\boldsymbol{j})^2$. Define $\widehat{\Delta}(\boldsymbol{j})^2$ by

$$\widehat{\Delta}(\boldsymbol{j})^2 = (1-c)D(\boldsymbol{j}) - \frac{c}{\gamma_1\gamma_2}, \tag{2.7}$$

for $D(\boldsymbol{j})$ given in (1.3). Then, the estimator $\widehat{\Delta}(\boldsymbol{j})^2$ is expanded as

$$\widehat{\Delta}(\boldsymbol{j})^2 = \Delta(\boldsymbol{j})^2 + \frac{D_1}{\sqrt{n}} + \frac{D_2}{n} + O_p(n^{-3/2}),$$

where

$$\begin{aligned}
D_1 =& -\frac{\sqrt{2}v_2\left(c + \Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)}{\sqrt{1-c}\gamma_1\gamma_2} + \frac{\sqrt{2}\sqrt{c}v_1}{\gamma_1\gamma_2} + \frac{2\Delta(\boldsymbol{j})u_1}{\sqrt{\gamma_1\gamma_2}},\\
D_2 =& -\frac{1}{(1-c)\gamma_1\gamma_2} - \frac{\Delta(\boldsymbol{j})^2}{1-c} + \frac{2\left(\frac{c}{\gamma_1\gamma_2} + \Delta(\boldsymbol{j})^2\right)}{1-c}v_2^2 + \frac{1}{\gamma_1\gamma_2}u_1^2\\
&- \frac{2\sqrt{2}\Delta(\boldsymbol{j})}{\sqrt{1-c}\sqrt{\gamma_1\gamma_2}}v_2u_1 - \frac{2\sqrt{c}}{\sqrt{1-c}\gamma_1\gamma_2}v_1v_2
\end{aligned}$$

for u_1 and u_2 given in Lemma 6.1 and v_1 and v_2 defined above (6.1).

We consider to substitute the consistent estimator into the limiting term $\Phi((U_\alpha)V_0^{-1/2})$ or $\Phi(U_\alpha V_0^{-1/2})$ in Theorem 2.1. Let $\widehat{U}_\alpha = -2^{-1}(1-c)^{-1}\widehat{\Delta}(\boldsymbol{j})^2$ and $\widehat{V}_0 = (1-c)^{-3}\{\widehat{\Delta}(\boldsymbol{j})^2 + c/(\gamma_1\gamma_2)\}$. For the term $\Phi(U_\alpha V_0^{-1/2})$, however, the estimator $\Phi(\widehat{U}_\alpha\widehat{V}_0^{-1/2})$ is not a second order unbiased estimator of $\Phi(U_\alpha V_0^{-1/2})$, since $\Phi(U_\alpha V_0^{-1/2}) = O(1)$. Since $\widehat{\Delta}(\boldsymbol{j})^2 = \Delta(\boldsymbol{j})^2 + D_1/\sqrt{n} + D_2/n + O_p(n^{-3/2})$, it is noted that

$$\begin{aligned}
\widehat{U}_\alpha =& U_\alpha + c_1\left(\frac{D_1}{\sqrt{n}} + \frac{D_2}{n}\right) + O_p(n^{-3/2}),\\
\widehat{V}_0 =& V_0 + c_2\left(\frac{D_1}{\sqrt{n}} + \frac{D_2}{n}\right) + O_p(n^{-3/2}),
\end{aligned}$$

for $c_1 = -\{2(1-c)\}^{-1}$ and $c_2 = (1-c)^3$. Then, it follows from (2.1) that

$$\begin{aligned}
\widehat{U}_\alpha\widehat{V}_0^{-1/2} =& U_\alpha V_0^{-1/2} + V_0^{-1/2}\left(c_1\frac{D_1}{\sqrt{n}} - \frac{U_\alpha}{2V_0}c_2\frac{D_1}{\sqrt{n}}\right)\\
&+ V_0^{-1/2}\left\{c_1\frac{D_2}{n} - \frac{U_\alpha}{2V_0}c_2\frac{D_2}{n} + \frac{3U_\alpha}{8V_0^2}c_2^2\frac{D_1^2}{n} - \frac{1}{2V_0}c_1c_2\frac{D_1^2}{n}\right\} + O_p(n^{-3/2}),
\end{aligned}$$

which imples that

$$\mathrm{E}[\Phi(\widehat{U}_\alpha\widehat{V}_0^{-1/2})] = \Phi(U_\alpha V_0^{-1/2}) + \frac{1}{n}\phi(U_\alpha V_0^{-1/2})K(\Delta(\boldsymbol{j})) + O(n^{-3/2}), \tag{2.8}$$

where

$$\begin{aligned}
K(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) =& \frac{1}{V_0^{1/2}}c_1E[\sqrt{n}D_1 + D_2] - \frac{U_0}{2V_0^{3/2}}c_2E[\sqrt{n}D_1 + D_2] - \frac{U_\alpha}{2V_0^{3/2}}c_1^2E[D_1^2]\\
&+ \frac{U_\alpha}{8V_0^{5/2}}\left(3 - \frac{U_\alpha^2}{V_0}\right)c_2^2E[D_1^2] - \frac{1}{2V_0^{3/2}}\left(1 - \frac{U_\alpha^2}{V_0}\right)c_1c_2E[D_1^2].
\end{aligned}$$

Combining (2.3) and (2.8), we can see that the approximation of $\mathrm{E}[L_1(\boldsymbol{j})]$ is expressed as

$$\begin{aligned}\mathrm{E}[L_1(\boldsymbol{j})] =&\Phi(U_\alpha V_0^{-1/2}) + \frac{1}{n}\phi(U_\alpha V_0^{-1/2})H(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) + O(n^{-3/2})\\ =&\mathrm{E}[\Phi(\widehat{U}_\alpha \widehat{V}_0^{-1/2})] + \frac{1}{n}\phi(U_\alpha V_0^{-1/2})\{H(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) - K(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c)\} + O(n^{-3/2}).\end{aligned}$$

To calculate the moments in $K(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c)$, note that $\mathrm{E}[D_1] = 0$,

$$\begin{aligned}\mathrm{E}[D_2] =&\frac{\Delta(\boldsymbol{j})^2}{1-c} + \frac{c}{(1-c)\gamma_1\gamma_2},\\ \mathrm{E}[D_1^2] =&\frac{2\Delta(\boldsymbol{j})^4}{1-c} + \frac{4\Delta(\boldsymbol{j})^2}{(1-c)\gamma_1\gamma_2} + \frac{2c}{(1-c)\gamma_1^2\gamma_2^2}.\end{aligned}$$

Replacing the unknown parameter $\Delta(\boldsymbol{j})$ in $H(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c) - K(\Delta(\boldsymbol{j}),\gamma_1,\gamma_2,c)$ with the consistent estimator $\widehat{\Delta}(\boldsymbol{j})$, we obtain the second order asymptotically unbiased estimator of $\gamma_1(\boldsymbol{j})$ given by

$$\widehat{R}_1(\boldsymbol{j}) = \Phi(\widehat{U}_0\widehat{V}_0^{-1/2}) + \frac{1}{n}\phi(\widehat{U}_0\widehat{V}_0^{-1/2})\{H(\widehat{\Delta}(\boldsymbol{j}),\gamma_1,\gamma_2,c) - K(\widehat{\Delta}(\boldsymbol{j}),\gamma_1,\gamma_2,c)\}. \tag{2.9}$$

Since the second order asymptotically unbiased estimator of $R_2(\boldsymbol{j})$ can be obtained by interchanging γ_1 with γ_2 in (2.9), we get

$$\widehat{R}_2(\boldsymbol{j}) = \Phi(\widehat{U}_\alpha\widehat{V}_0^{-1/2}) + \frac{1}{n}\phi(\widehat{U}_\alpha\widehat{V}_0^{-1/2})\{H(\widehat{\Delta}(\boldsymbol{j}),\gamma_2,\gamma_1,c) - K(\widehat{\Delta}(\boldsymbol{j}),\gamma_1,\gamma_2,c)\}. \tag{2.10}$$

Combining (2.9) and (2.10), we obtain the second order asymptotically unbiased estimator of $R(\boldsymbol{j})$ given by

$$MEC = \widehat{R}(\boldsymbol{j}) = \frac{1}{2}\left(\widehat{R}_1(\boldsymbol{j}) + \widehat{R}_2(\boldsymbol{j})\right), \tag{2.11}$$

which can be used for variable selection. We here call $\widehat{R}(\boldsymbol{j})$ the *Misclassification Error Criterion* (MEC). The best subset of variables among $\mathcal{J}$ is suggested as $\boldsymbol{j}$ which minimizes $\widehat{R}(\boldsymbol{j})$ among $\mathcal{J}$.

3 Asymptotic Properties of MEC

3.1 Decomposition into fitting and penalty terms

A feature of variable selection procedures like AIC and C_p is that they are decomposed into the two terms: 'fitting' term and 'penalty' term for model complexity. It is interesting to investigate whether MEC has a similar feature.

It is noted that $\Phi\left(-D(\boldsymbol{j})/2\right)$ decreases as the cardinality of variable set $\sharp(\boldsymbol{j})$ increases, namely, $\Phi\left(-D(\boldsymbol{j}_1)/2\right) > \Phi\left(-D(\boldsymbol{j}_2)/2\right)$ if $\sharp(\boldsymbol{j}_2) > \sharp(\boldsymbol{j}_1)$. This means that $\Phi\left(-D(\boldsymbol{j})/2\right)$ is regarded as a fitting term. Thus, MEC is decomposed as

$$MEC(\boldsymbol{j}) = \Phi\left(-D(\boldsymbol{j})/2\right) + \hat{b}(\boldsymbol{j}), \tag{3.1}$$

where $\hat{b}(\boldsymbol{j}) = \widehat{R}(\boldsymbol{j}) - \Phi\left(-D(\boldsymbol{j})/2\right)$. Let $b(\boldsymbol{j}) = R(\boldsymbol{j}) - \mathrm{E}[\Phi\left(-D(\boldsymbol{j})/2\right)]$. Then it is noted that $\hat{b}(\boldsymbol{j})$ is a second order asymptotically unbiased estimator of $b(\boldsymbol{j})$, namely $\mathrm{E}[\hat{b}(\boldsymbol{j})] = R(\boldsymbol{j}) - \mathrm{E}[\Phi\left(-D(\boldsymbol{j})/2\right)] + O(n^{-3/2})$.

In this subsection, we express the bias $b(\boldsymbol{j})$ with an explicit formula, and show that $b(\boldsymbol{j})$ works as dimensionality penalty under the high dimensional or large sample situation. We evaluate $\mathrm{E}[\Phi(-D(\boldsymbol{j})/2)]$ since the explicit formula of $R(\boldsymbol{j})$ is provided in subsection 2.1. Using Lemma 5.1, we can approximate $D(\boldsymbol{j})$ stochastically as

$$D(\boldsymbol{j})^2 = E_0 + \frac{E_1}{\sqrt{n}} + \frac{E_2}{n}, \tag{3.2}$$

where

$$\begin{aligned}
E_0 &= \frac{1}{(1-c)}\left(\Delta(\boldsymbol{j})^2 + \frac{c}{\gamma_1\gamma_2}\right), \\
E_1 &= -\frac{\sqrt{2}v_2\left(c + \Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)}{(1-c)^{3/2}\gamma_1\gamma_2} + \frac{\sqrt{2}\sqrt{c}v_1}{\gamma_1\gamma_2(1-c)} + \frac{2\Delta(\boldsymbol{j})u_1}{\sqrt{\gamma_1\gamma_2}(1-c)}, \\
E_2 &= -\frac{1}{(1-c)^2\gamma_1\gamma_2} - \frac{\Delta(\boldsymbol{j})^2}{(1-c)^2} + \frac{2\left(\frac{c}{\gamma_1\gamma_2} + \Delta(\boldsymbol{j})^2\right)}{(1-c)^2}v_2^2 + \frac{1}{\gamma_1\gamma_2(1-c)}u_1^2 \\
&\quad -\frac{2\sqrt{2}\Delta(\boldsymbol{j})}{(1-c)^{3/2}\sqrt{\gamma_1\gamma_2}}v_2u_1 - \frac{2\sqrt{c}}{(1-c)^{3/2}\gamma_1\gamma_2}v_1v_2
\end{aligned}$$

Using the stochastic expansion (3.2), we can see that

$$\mathrm{E}\left[\Phi\left(-\frac{D(\boldsymbol{j})}{2}\right)\right] = \Phi(m_0) + \frac{1}{n}\phi(m_0)\mathrm{E}\left[m_1 + m_2 - \frac{1}{2}m_0m_1^2\right] + O(n^{-3/2}),$$

where

$$m_0 = -\frac{1}{2}E_0^{1/2},\ m_1 = -\frac{1}{4}E_0^{-1/2}\frac{E_1}{\sqrt{n}},\ m_2 = \frac{1}{16}E_0^{-3/2}\frac{E_1^2}{n} - \frac{1}{4}E_0^{-1/2}\frac{E_2}{n}.$$

We can obtain the expectations of m_1, m_2 and m_1^2 using the moments $\mathrm{E}[E_1] = 0$,

$$\begin{aligned}
\mathrm{E}[E_2] &= \frac{(3-2c)c + \Delta(\boldsymbol{j})^2\gamma_1\gamma_2}{(1-c)^2\gamma_1\gamma_2}, \\
\mathrm{E}[E_1^2] &= \frac{2\left(c + \Delta(\boldsymbol{j})^2\gamma_1\gamma_2\left(\Delta(\boldsymbol{j})^2\gamma_1\gamma_2 + 2\right)\right)}{(1-c)^3\gamma_1^2\gamma_2^2}.
\end{aligned}$$

Thus, it is observed that

$$\mathrm{E}\left[\Phi\left(-\frac{D(\boldsymbol{j})}{2}\right)\right] = \Phi(m_0) + \frac{1}{n}\phi(m_0)M(\Delta(\boldsymbol{j})^2, \gamma_1, \gamma_2, c) + O(n^{-3/2}),$$

where

$$M(\Delta(\boldsymbol{j})^2, \gamma_1, \gamma_2, c) = \frac{1}{16}E_0^{-3/2}\mathrm{E}[E_1^2] - \frac{1}{4}E_0^{-1/2}\mathrm{E}[E_2] + \frac{1}{64}E_0^{-1/2}\mathrm{E}[E_1^2].$$

Then, we can get a second-order approximation of the bias term $b(\boldsymbol{j})$. Define $b^{(g_1,g_2)}(\boldsymbol{j})$ by

$$\begin{aligned} b^{(g_1,g_2)}(\boldsymbol{j}) &= \Phi\left(\Delta_1(\boldsymbol{j})\right) - \Phi\left(\Delta_2(\boldsymbol{j})\right) \\ &\quad + \frac{1}{n}\left\{\phi(\Delta_1(\boldsymbol{j}))H(\Delta(\boldsymbol{j}), r_{g_1}, r_{g_2}, c) - \phi\left(\Delta_2(\boldsymbol{j})\right) M(\Delta(\boldsymbol{j}), \gamma_1, \gamma_2, c)\right\}, \end{aligned}$$

where

$$\begin{aligned} \Delta_1(\boldsymbol{j}) &= -\frac{\sqrt{1-c}\Delta^2(\boldsymbol{j})}{2\sqrt{\Delta^2(\boldsymbol{j}) + c/(\gamma_1\gamma_2)}}, \\ \Delta_2(\boldsymbol{j}) &= -\frac{1}{2\sqrt{1-c}}\sqrt{\Delta^2(\boldsymbol{j}) + c/(\gamma_1\gamma_2)}. \end{aligned}$$

Theorem 3.1 *Under the assumptions* (A1)-(A3)*,*

$$b(\boldsymbol{j}) = \frac{1}{2}\left(b^{(1,2)}(\boldsymbol{j}) + b^{(2,1)}(\boldsymbol{j})\right) + O(n^{-3/2}).$$

From Theorem 3.1, an expression of the bias $b(\boldsymbol{j})$ implies that $b(\boldsymbol{j})$ depends on $\sharp(\boldsymbol{j})$ through $c = \lim_{n,p\to\infty} \sharp(\boldsymbol{j})/n$. Concedring the limiting term of $b(\boldsymbol{j})$, namely, $\Phi\left(\Delta_1(\boldsymbol{j})\right) - \Phi\left(\Delta_2(\boldsymbol{j})\right)$, it can be seen that it increases in $\sharp(\boldsymbol{j})$ through c for $\boldsymbol{j} \in \{\boldsymbol{j}|\Delta^2(\boldsymbol{j}) = \Delta^2\}$, since $\Delta_1(\boldsymbol{j})$ is increases in c, $\Delta_1(\boldsymbol{j})$ is decreases in c.

To make it clear that $b(\boldsymbol{j})$ works as a penalty of the cardinality $\sharp(\boldsymbol{j})$, we consider the large sample framework (A0). Then, it is observed

$$\begin{aligned} \mathrm{E}[L_1(\boldsymbol{j})] &= \Phi\left(-\frac{\Delta(\boldsymbol{j})}{2}\right) + \frac{1}{n}\phi\left(-\frac{\Delta(\boldsymbol{j})}{2}\right)\left\{\frac{1}{4\Delta(\boldsymbol{j})\gamma_1}\left(\frac{\Delta(\boldsymbol{j})^2}{4} + 3(\sharp(\boldsymbol{j}) - 1)\right)\right. \\ &\quad \left. + \frac{1}{4\Delta(\boldsymbol{j})\gamma_2}\left(\frac{\Delta(\boldsymbol{j})^2}{4} - \sharp(\boldsymbol{j}) + 1\right) + \frac{1}{4}\Delta(\boldsymbol{j})(\sharp(\boldsymbol{j}) - 1)\right\} + o(n^{-1}), \\ \mathrm{E}\left[\Phi\left(-\frac{D(\boldsymbol{j})}{2}\right)\right] &= \Phi\left(-\frac{\Delta(\boldsymbol{j})}{2}\right) + \frac{1}{n}\phi\left(-\frac{\Delta(\boldsymbol{j})}{2}\right)\left(\frac{\Delta(\boldsymbol{j})^3}{32} - \frac{\sharp(\boldsymbol{j})}{4\Delta(\boldsymbol{j})\gamma_1\gamma_2} - \frac{\Delta(\boldsymbol{j})\sharp(\boldsymbol{j})}{4}\right. \\ &\quad \left. + \frac{\Delta(\boldsymbol{j})}{16\gamma_1\gamma_2} + \frac{1}{4\Delta(\boldsymbol{j})\gamma_1\gamma_2} - \frac{\Delta(\boldsymbol{j})}{8}\right) + o(n^{-1}). \end{aligned}$$

Then, we get the following proposition.

Proposition 3.1 *Under the assumptions* (A0)*,* (A2) *and* (A3)*,*

$$b(\boldsymbol{j}) = \frac{1}{n}\phi\left(-\frac{\Delta(\boldsymbol{j})}{2}\right)\left\{\frac{\sharp(\boldsymbol{j})}{\Delta(\boldsymbol{j})r_1} + \frac{\Delta(\boldsymbol{j})\sharp(\boldsymbol{j})}{2} - \frac{1}{\Delta(\boldsymbol{j})r_1} - \left(\frac{\Delta^3(\boldsymbol{j})}{32} + \frac{\Delta(\boldsymbol{j})}{8}\right)\right\} + o(n^{-1}).$$

Thus, the limiting term of $b(\boldsymbol{j})$*,* $\boldsymbol{j} \in \{\boldsymbol{j}|\Delta^2(\boldsymbol{j}) = \Delta^2\}$ *increases as* $\sharp(\boldsymbol{j})$ *increases.*

The arguments given in this subsection shows that MEC $\widehat{R}(\boldsymbol{j})$ is decomposed into the "fitting term" and the "penalty term", which is the same feature as in variable selection procedures like AIC and C_p.

3.2 Asymptotic optimality of MEC

In this subsection, we show that MEC has an asymptotic optimality in the high-dimensional setting (A1). The optimality is related to Li (1987), who showed that Mallows' C_p is asymptotically equivalent to the squared error loss in a linear regression model, and that the estimator selected by C_p asymptotically achieves the smallest possible squared error loss in the class of model average estimators. The squared error loss corresponds to the conditional misclassification error rate $L(\boldsymbol{j})$ in the classification problem, where

$$L(\boldsymbol{j}) = 2^{-1}\{L_1(\boldsymbol{j}) + L_2(\boldsymbol{j})\}.$$

Thus, we shall verify that $\widehat{R}(\boldsymbol{j})$ is asymptotically equivalent to the conditional misclassification error rate $L(\boldsymbol{j})$, and that MEC asymptotically achieves the smallest possible conditional misclassification error rate.

The primary goal of this section is to demonstrate that under reasonable conditions, MEC is asymptotically optimal in the sense that

$$\left| \frac{L(\hat{\boldsymbol{j}})}{\inf_{\boldsymbol{j}\in\mathcal{J}} L(\boldsymbol{j})} \right| \xrightarrow{p} 1, \tag{3.3}$$

where $\hat{\boldsymbol{j}}$ is the best selection satisfying $MEC(\hat{\boldsymbol{j}}) = \min_{\boldsymbol{j}\in\mathcal{J}} MEC(\boldsymbol{j})$. It can be seen that sufficient conditions for (3.3) are given by

$$\begin{aligned} &\text{(i)} \quad \sup_{\boldsymbol{j}\in\mathcal{J}} \left| \frac{L(\boldsymbol{j})}{R(\boldsymbol{j})} - 1 \right| \xrightarrow{p} 0, \\ &\text{(ii)} \quad \sup \left| \frac{\widehat{R}(\boldsymbol{j}) - L(\boldsymbol{j})}{R(\boldsymbol{j})} \right| \xrightarrow{p} 0. \end{aligned} \tag{3.4}$$

We shall check conditions (i) and (ii) using chebyshev's inequality. For any $\varepsilon > 0$,

$$\begin{aligned} & \Pr\left(\sup_{\boldsymbol{j}\in\mathcal{J}} \left| \frac{L(\boldsymbol{j})}{R(\boldsymbol{j})} - 1 \right| > \varepsilon \right) \\ \leq & \sum_{\boldsymbol{j}\in\mathcal{J}} \frac{\mathrm{E}[|L(\boldsymbol{j}) - R(\boldsymbol{j})|^{2m}]}{R(\boldsymbol{j})^{2m}\varepsilon^{2m}} \\ \leq & \sum_{\boldsymbol{j}\in\mathcal{J}} \left(\mathrm{E}[|L_1(\boldsymbol{j}) - R_1(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}} + \mathrm{E}[|L_2(\boldsymbol{j}) - R_2(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}} \right)^{2m} \left(\frac{1}{2R(\boldsymbol{j})\varepsilon} \right)^{2m} \\ \leq & \sup_{\boldsymbol{j}\in\mathcal{J}} \left(\mathrm{E}[|L_1(\boldsymbol{j}) - R_1(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}} + \mathrm{E}[|L_2(\boldsymbol{j}) - R_2(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}} \right)^{2m} \sum_{\boldsymbol{j}\in\mathcal{J}} \left(\frac{1}{2R(\boldsymbol{j})\varepsilon} \right)^{2m} \end{aligned}$$

and

$$\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{j}\in\mathcal{J}}\left|\frac{\widehat{R}(\boldsymbol{j})-L(\boldsymbol{j})}{R(\boldsymbol{j})}-1\right|>\varepsilon\right)\\
\leq\ &\sum_{\boldsymbol{j}\in\mathcal{J}}\frac{\mathrm{E}[|\widehat{R}(\boldsymbol{j})-L(\boldsymbol{j})|^{2m}]}{R(\boldsymbol{j})^{2m}\varepsilon^{2m}}\\
\leq\ &\sum_{\boldsymbol{j}\in\mathcal{J}}\left(\mathrm{E}[|\widehat{R}_1(\boldsymbol{j})-L_1(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}}+\mathrm{E}[|\widehat{R}_2(\boldsymbol{j})-L_2(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}}\right)^{2m}\left(\frac{1}{2R(\boldsymbol{j})\varepsilon}\right)^{2m}\\
\leq\ &\sup_{\boldsymbol{j}\in\mathcal{J}}\left(\mathrm{E}[|\widehat{R}_1(\boldsymbol{j})-L_1(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}}+\mathrm{E}[|\widehat{R}_2(\boldsymbol{j})-L_2(\boldsymbol{j})|^{2m}]^{\frac{1}{2m}}\right)^{2m}\sum_{\boldsymbol{j}\in\mathcal{J}}\left(\frac{1}{2R(\boldsymbol{j})\varepsilon}\right)^{2m}.
\end{aligned}$$

Thus, we need to evaluate $\mathrm{E}[|L_i(\boldsymbol{j})-R_i(\boldsymbol{j})|^{2m}]$ and $\mathrm{E}[|\widehat{R}_i(\boldsymbol{j})-L_i(\boldsymbol{j})|^{2m}]$ for $i=1,2$.

Lemma 3.1 *Assume either condition* (A0) *or condition* (A1). *Under conditions* (A2) *and* (A3), *it holds that*

$$\mathrm{E}[|\sqrt{n}(L_i(\boldsymbol{j})-R_i(\boldsymbol{j}))|^{2m}]<\infty,\ \mathrm{E}[|\sqrt{n}(\widehat{R}_i(\boldsymbol{j})-L_i(\boldsymbol{j}))|^{2m}]<\infty.$$

(Proof) Using stochastic expansions of U, V and $\widehat{\Delta}(\boldsymbol{j})$, it follows that

$$\begin{aligned}
\sqrt{n}(L_1(\boldsymbol{j})-R_1(\boldsymbol{j})) &= \phi\left(\Delta_1(\boldsymbol{j})\right)\left(\frac{U_1}{\sqrt{V_0}}-\frac{U_\alpha V_1}{2V_0^{3/2}}\right)+o_p(1),\\
\sqrt{n}(\widehat{R}_i(\boldsymbol{j})-R_i(\boldsymbol{j})) &= -\phi\left(\Delta_1(\boldsymbol{j})\right)\frac{1}{V_0}\left(\frac{1}{2(1-c)}+\frac{U_\alpha}{V_0(1-c)^3}\right)D_1+o_p(1).
\end{aligned}$$

We begin by expanding the statistic $\sqrt{n}(L_1(\boldsymbol{j})-R_1(\boldsymbol{j}))$ stochastically. Using (5.1) and (5.2), we can expand the statistic $\sqrt{n}(L_1(\boldsymbol{j})-R_1(\boldsymbol{j}))$ as

$$\begin{aligned}
\sqrt{n}(L_1(\boldsymbol{j})-R_1(\boldsymbol{j})) &= \phi\left(\Delta_1(\boldsymbol{j})\right)\left(\sqrt{\frac{(1-c)c}{\gamma_1\gamma_2}}\frac{\{c(4\gamma_2-2)+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2(4\gamma_2-1)\}}{2\sqrt{2}\left(c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)^{3/2}}v_1\right.\\
&\quad+\sqrt{\frac{1}{\gamma_1\gamma_2}}\frac{c(1-2\gamma_2)}{\sqrt{2}\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}}v_2+\frac{\Delta(\boldsymbol{j})^2\sqrt{(1-c)c\gamma_1\gamma_2}}{2\sqrt{2}\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}}v_3\\
&\quad-\frac{c\Delta(\boldsymbol{j})^2\sqrt{\gamma_1\gamma_2}}{2\sqrt{2}\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}}v_4+\frac{\Delta(\boldsymbol{j})\left\{2c(\gamma_2-1)+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2(2\gamma_2-1)\right\}}{2(1-c)^{-1/2}\left(c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)^{3/2}}u_1\\
&\quad\left.-\frac{\sqrt{c}\Delta(\boldsymbol{j})\gamma_2}{\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}}u_2+\sqrt{1-c}u_3-\sqrt{c}u_4\right)+o_p(1). \qquad (3.5)
\end{aligned}$$

It should be noted that the above statistic is a linear combination of independent random variables u_i and v_i with finite $2m$-th moment under the assumptions (A1)-(A3). Thus, for any $m\in\mathbb{R}$, $\mathrm{E}[|\sqrt{n}(L_i(\boldsymbol{j})-R_i(\boldsymbol{j}))|^{2m}]<\infty$ under the assumptions (A1)-(A3).

We here treat the case that p is a fixed constant and p/n tends to zero, i.e. (A0). Substitute $c = 0$ in (3.5), we obtain a stochastic expansion of the statistics $\sqrt{n}(L_1(\boldsymbol{j}) - R_1(\boldsymbol{j}))$ given by

$$\sqrt{n}(L_1(\boldsymbol{j}) - R_1(\boldsymbol{j})) \;=\; \frac{\gamma_2 - \gamma_1}{2\sqrt{\gamma_1\gamma_2}}u_1 + u_3 + o_p(1).$$

This shows that the above statistic is a linear combination of independent random variables u_1 and u_3 with finite $2m$-th moment under the large sample framework (A0). Thus, for any $m \in \mathbb{R}$, $\mathrm{E}[|\sqrt{n}(L_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))|^{2m}] < \infty$ under the large sample framework (A0).

We next evaluate the statistic $\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))$, which can be carried out similarly. Using (5.1) and (5.2), we can expand the statistic $\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))$ as

$$\begin{aligned}\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j})) \;=\; & -\phi\left(\Delta_1(\boldsymbol{j})\right)\frac{1}{V_0}\left(\frac{1}{2(1-c)} + \frac{U_\alpha}{V_0(1-c)^3}\right)\left(-\frac{\sqrt{2}\left(c + \Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)}{\sqrt{1-c}\gamma_1\gamma_2}v_2\right. \\ & \left. + \frac{\sqrt{2}\sqrt{c}}{\gamma_1\gamma_2}v_1 + \frac{2\Delta(\boldsymbol{j})}{\sqrt{\gamma_1\gamma_2}}u_1\right) + o_p(1). \qquad (3.6)\end{aligned}$$

It should be noted that above statistic is linear combination of independent random variables u_1, v_1 and v_2 with finite $2m$-th moment under the assumptions (A1)-(A3). Thus, for any $m \in \mathbb{R}$, $\mathrm{E}[|\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))|^{2m}] < \infty$ under the assumptions (A1)-(A3). In addition, consider the case of the large sample framework (A0). Substitutint $c = 0$ in (3.6), we obtain the stochastic expansion of the statistic $\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))$ given by

$$\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j})) \;=\; \frac{\gamma_2 - \gamma_1}{2\sqrt{\gamma_1\gamma_2}}u_1 + u_3 + o_p(1).$$

the above statistic is a linear combination of independent random variables u_1 and u_3 with finite $2m$-th moment under the large sample framework (A0). Thus, for any $m \in \mathbb{R}$, $\mathrm{E}[|\sqrt{n}(\widehat{R}_i(\boldsymbol{j}) - R_i(\boldsymbol{j}))|^{2m}] < \infty$ under the large sample framework (A0). Therefore, the proof of Lemma 3.1 is complete. ■

Combining Lemma 3.1 and (3.4) gives the sufficient condition for the asymptotic optimality (3.3), given by

$$\text{(C1)} \quad \sum_{\boldsymbol{j}\in\mathcal{J}}\left(\frac{1}{2\sqrt{n}R(\boldsymbol{j})}\right)^{2m} \to 0.$$

Hence, we obtain the following theorem.

Theorem 3.2 *Assume either condition* (A0) *or condition* (A1)*. Under conditions* (A2)*,* (A3) *and* (C1)*, MEC is asymptotically optimal in the sense that*

$$\left|\frac{L(\hat{\boldsymbol{j}})}{\inf_{\boldsymbol{j}\in\mathcal{J}} L(\boldsymbol{j})}\right| \xrightarrow{p} 1.$$

It is noted that the condition (C1) holds if the cardinality of the family $\mathcal{J}$ satisfies $\sharp(\mathcal{J}) = o(n^m)$. For example, for the family of the subsets $\mathcal{J} = \{\{1\}, \{1,2\}, \{1,2,3\}, \ldots, \{1,2,3,\ldots,p\}\}$, we have $\sharp(\mathcal{J}) = p$, so that the condition $\sharp(\mathcal{J}) = o(n^m)$ is satisfied under (A1). However, when $\mathcal{J}$ consists of all the subsets of $\{1,\ldots,p\}$, we have $\sharp(\mathcal{J}) = 2^p$, which does not satisfy the condition. Thus, Theorem 3.2 implies that we should pay heed to cardinality of the family $\mathcal{J}$ of candidate variable sets in the high-dimensional setting (A1).

4 Numerical Results

In this section, we investigate numerical properties of MEC by Monte Carlo simulations. The frequencies of selecting the true variables by selection procedures are calculated by simulation with 10,000 iterations.

The mean vector of Π_1 is set up by $\boldsymbol{\mu} = ((1/2)\mathbf{1}'_{10}, \mathbf{0}'_{p-10})'$, the mean of Π_2 is set up by zero vector and the covariance matrix of both groups is set up by I_p since the expected error rate $R(\boldsymbol{j})$ only depends on Mahalanobis distance $\Delta(\boldsymbol{j})$ and (n,p). The data sets are generated as follows:

$$\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1N_1} \sim \mathcal{N}(\boldsymbol{\mu}, I_p),\ \boldsymbol{x}_{21},\ldots,\boldsymbol{x}_{2N_2} \sim \mathcal{N}(\mathbf{0}, I_p)$$

in each step. In our numerical study, we consider the ten candidate variable sets $\boldsymbol{j}_i$ given by

$$\begin{aligned} \boldsymbol{j}_i &= \{1,2,\cdots,(10-i+1)\} \text{ for } i = 1,\cdots,5, \\ \boldsymbol{j}_i &= \{1,2,\cdots,(p-i+6)\} \text{ for } i = 6,\cdots,10. \end{aligned}$$

In this setting, the variable set which minimizes $R(\boldsymbol{j})$ is $\boldsymbol{j}_1$ in all the cases as seen from Table 1.

Table 1. Values of $R(\boldsymbol{j})$.

$(p,\ N_1,\ N_2)$	$R(\boldsymbol{j}_1)$	$R(\boldsymbol{j}_2)$	$R(\boldsymbol{j}_3)$	$R(\boldsymbol{j}_4)$	$R(\boldsymbol{j}_5)$	$R(\boldsymbol{j}_6)$	$R(\boldsymbol{j}_7)$	$R(\boldsymbol{j}_8)$	$R(\boldsymbol{j}_9)$	$R(\boldsymbol{j}_{10})$
$(50,50,50)$	**0.15**	0.17	0.18	0.19	0.21	0.25	0.25	0.25	0.24	0.24
$(50,100,100)$	**0.14**	0.15	0.17	0.18	0.20	0.19	0.19	0.19	0.18	0.18
$(100,100,100)$	**0.14**	0.15	0.17	0.18	0.20	0.25	0.25	0.25	0.25	0.25
$(100,200,200)$	**0.14**	0.15	0.163	0.18	0.20	0.19	0.19	0.19	0.19	0.19

We investigate the frequencies of selecting the true variable set minimizing $R(\boldsymbol{j})$ with several variable selection procedures. In this experiment, we compare the performances of AIC, BIC, Mc and MEC, where Mc is the large sample unbiased estimator suggested by MacLachlan (1976,1980). The AIC, BIC and Mc for the model $\boldsymbol{j}$ are defined by

$$\begin{aligned} AIC(\boldsymbol{j}) &= -N\log\frac{1+(N_1N_2)/(Nn)D(\boldsymbol{j})^2}{1+(N_1N_2)/(Nn)D^2} + N\log|N^{-1}W| + Np(1+\log 2\pi) + b_A, \\ BIC(\boldsymbol{j}) &= -N\log\frac{1+(N_1N_2)/(Nn)D(\boldsymbol{j})^2}{1+(N_1N_2)/(Nn)D^2} + N\log|N^{-1}W| + Np(1+\log 2\pi) + b_B, \\ Mc(\boldsymbol{j}) &= \Phi\left(-\frac{D(\boldsymbol{j})}{2}\right) + \phi\left(-\frac{D(\boldsymbol{j})}{2}\right)\left(\frac{\sharp(\boldsymbol{j})-1}{N_1D(\boldsymbol{j})} + 4(4\sharp(\boldsymbol{j})-1)D(\boldsymbol{j}) - \frac{D(\boldsymbol{j})^3}{32n}\right), \end{aligned}$$

where $D^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' S^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$, $b_A = 2(2\sharp(\boldsymbol{j}) + p - \sharp(\boldsymbol{j}) + p(p+1)/2)$ and $b_B = \log N(2\sharp(\boldsymbol{j}) + p - \sharp(\boldsymbol{j}) + p(p+1)/2)$ for $N = N_1 + N_2$.

Tables 2-5 reports frequencies of selecting each variable sets $\boldsymbol{j}_i$ by the variable selection procedures AIC, BIC, Mc and MEC. As seen from Tables 2-5, MEC is superior to the other criteria. Concerning AIC and MEC, their frequencies of selecting the true variable set get larger as the dimension p and sample size $N(= N_1 + N_2)$ are larger. Tables 2 and 4 treat the case of $p/N = 0.5$, which means that N is relatively small. In this case, the frequencies of selecting $\boldsymbol{j}_1$ by Mc is small. However, those values for Mc in Tables 3 and 5 are large. This arises from the reason that Mc is the variable selection criterion derived under the large sample framework (A0). However, MEC is excellent in all the cases, because it is derived under the high-dimensional setting.

Table 2. Comparison of frequencies of selecting $\boldsymbol{j}_i$ for $p = 50,\ N_1 = N_2 = 50$.

$\boldsymbol{j}_i$	$\boldsymbol{j}_1$	$\boldsymbol{j}_2$	$\boldsymbol{j}_3$	$\boldsymbol{j}_4$	$\boldsymbol{j}_5$	$\boldsymbol{j}_6$	$\boldsymbol{j}_7$	$\boldsymbol{j}_8$	$\boldsymbol{j}_9$	$\boldsymbol{j}_{10}$
AIC	72.2	13.7	3.2	0.7	0.2	2.6	1.7	1.6	1.4	2.7
BIC	50.1	22.5	13.3	8.0	6.1	0.0	0.0	0.0	0.0	0.0
Mc	12.6	2.2	0.6	0.1	0.0	41.1	18.0	10.2	6.7	8.5
MEC	**79.6**	15.3	3.5	0.8	0.2	0.1	0.1	0.1	0.1	0.2

Table 3. Comparison of frequencies of selecting $\boldsymbol{j}_i$ for $p = 50,\ N_1 = N_2 = 100$.

$\boldsymbol{j}_i$	$\boldsymbol{j}_1$	$\boldsymbol{j}_2$	$\boldsymbol{j}_3$	$\boldsymbol{j}_4$	$\boldsymbol{j}_5$	$\boldsymbol{j}_6$	$\boldsymbol{j}_7$	$\boldsymbol{j}_8$	$\boldsymbol{j}_9$	$\boldsymbol{j}_{10}$
AIC	96.6	2.3	0.1	0.0	0.1	0.2	0.0	0.1	0.2	0.4
BIC	85.1	12.1	2.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0
Mc	84.0	1.9	0.0	0.0	0.0	3.4	2.5	2.3	2.0	3.9
MEC	**97.1**	2.5	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2

Table 4. Comparison of frequencies of selecting $\boldsymbol{j}_i$ for $p = 100,\ N_1 = N_2 = 100$.

$\boldsymbol{j}_i$	$\boldsymbol{j}_1$	$\boldsymbol{j}_2$	$\boldsymbol{j}_3$	$\boldsymbol{j}_4$	$\boldsymbol{j}_5$	$\boldsymbol{j}_6$	$\boldsymbol{j}_7$	$\boldsymbol{j}_8$	$\boldsymbol{j}_9$	$\boldsymbol{j}_{10}$
AIC	96.2	2.2	0.1	0.0	0.0	0.5	0.3	0.2	0.2	0.3
BIC	84.2	12.5	2.6	0.6	0.1	0.0	0.0	0.0	0.0	0.0
Mc	13.3	0.4	0.0	0.0	0.0	42.0	18.7	10.6	6.9	8.1
MEC	**97.5**	2.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5. Comparison of frequencies of selecting $\boldsymbol{j}_i$ for $p = 100,\ N_1 = N_2 = 200$.

$\boldsymbol{j}_j$	$\boldsymbol{j}_1$	$\boldsymbol{j}_2$	$\boldsymbol{j}_3$	$\boldsymbol{j}_4$	$\boldsymbol{j}_5$	$\boldsymbol{j}_6$	$\boldsymbol{j}_7$	$\boldsymbol{j}_8$	$\boldsymbol{j}_9$	$\boldsymbol{j}_{10}$
AIC	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BIC	99.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mc	96.9	0.1	0.0	0.0	0.0	0.6	0.6	0.5	0.4	0.9
MEC	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

We next investigate the relationship of $MEC(\boldsymbol{j})$ and the conditional error probability $L(\boldsymbol{j})$ by comparing the probabilities of selecting the true variable sets. Table 6 reports the frequencies of selecting the true variable sets with $MEC(\boldsymbol{j})$ and $L(\boldsymbol{j})$. This shows confirms that both $L(\boldsymbol{j})$ and $MEC(\boldsymbol{j})$ select the true variable sets with high frequencies, and this observation is related to the optimality (3.3).

Table 6. Comparison of frequencies of selecting the true variable sets

(p, N_1, N_2)	$(50, 50, 50)$	$(50, 100, 100)$	$(100, 100, 100)$	$(100, 200, 200)$
$L(\boldsymbol{j})$	88.8	97.7	97.7	100.0
$MEC(\boldsymbol{j})$	79.7	97.1	97.5	100.0

Finaly, we check unbiasedness of MEC and Mc. For each subset $\boldsymbol{j}_i,\ i = 1, \cdots, 10$, we compute the averages of $L(\boldsymbol{j})$, MEC and Mc by Monte Carlo simulations. In Figures 1-4, the averages of $L(\boldsymbol{j})$, $MEC(\boldsymbol{j})$ and $Mc(\boldsymbol{j})$ for each $\boldsymbol{j}$ are plotted as "◯", "●" and "×", respectively. As seen from the figures, MEC and Mc perform well for the subsets $\boldsymbol{j}_i,\ i = 1, \cdots, 5$, while Mc is poor for the subsets $\boldsymbol{j}_i,\ i = 6, \cdots, 10$. It is noted that the dimensions of the subsets $\boldsymbol{j}_i,\ i = 1, \cdots, 5$ are not so large, but the dimensions of $\boldsymbol{j}_i,\ i = 6, \cdots, 10$ are large. This is why Mc was the variable selection procedure derived in the large sample framework. MEC has good accuracies in both cases. Thus we conclude that MEC is more flexible than Mc concerning the sample size and dimension.

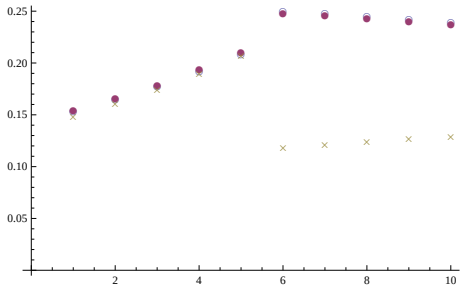

Fig.1. $(p, N_1, N_2) = (50, 50, 50)$.

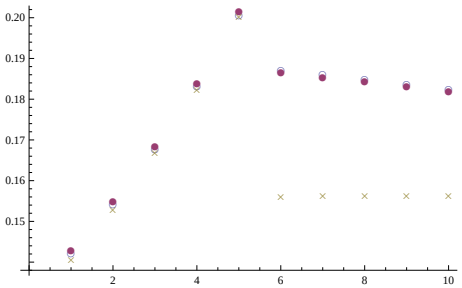

Fig.2. $(p, N_1, N_2) = (50, 100, 100)$.

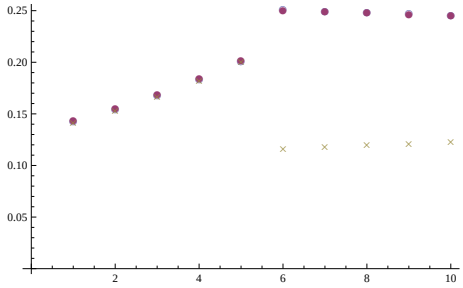

Fig.3. $(p, N_1, N_2) = (100, 100, 100)$.

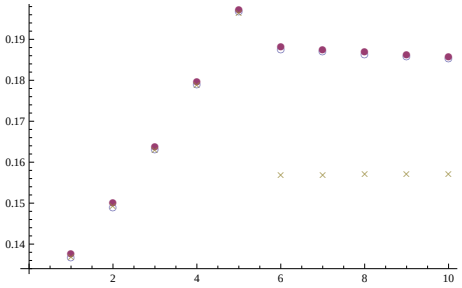

Fig.4. $(p, N_1, N_2) = (100, 200, 200)$.

5 Concluding Remarks

In this paper, we have suggested the new variable selection criterion MEC in linear discriminant analysis for high dimensional data. This is derived as a second-order unbiased estimator of the misclassification error probability. We have confirmed that MEC is decomposed into the "fitting term" and the "dimensionality penalty term" like AIC and Mallows' C_p. Moreover, we have shown that MEC is asymptotically optimal in the sense of achieving the smallest possible conditional probability of misclassification in candidate variable sets. Also, the superiority of MEC has been verified in the sense of selecting the true variable sets by simulation.

It may be important to point out that the optimality (3.3) given in Theorem 3.2 is guaranteed by condition (C1), or $\sharp(\mathcal{J}) = o(n^m)$. This codition always holds in the large sample framework (A0), but it is not necessarily satisfied in the high-dimensional situation (A1). For example, for the family of the subsets $\mathcal{J} = \{\{1\}, \{1,2\}, \{1,2,3\}, \ldots, \{1,2,3,\ldots,p\}\}$, we have $\sharp(\mathcal{J}) = p$, so that the condition $\sharp(\mathcal{J}) = o(n^m)$ is satisfied under (A1). However, when $\mathcal{J}$ consists of all the subsets of $\{1,\ldots,p\}$, we have $\sharp(\mathcal{J}) = 2^p$, which does not satisfy the condition. Thus, Theorem 3.2 implies that we should pay heed to cardinality of the family $\mathcal{J}$ of candidate variable sets in the high-dimensional setting (A1).

6 Appendix

In this section, we derive the stochastic expansions for the statistics U and V. For the purpose, we prepare some stochastic expressions of required quadratic forms.

6.1 Stochastic expression of quadratic forms

Lemma 6.1 *Let $\boldsymbol{z} \sim \mathcal{N}_p(\tau \boldsymbol{e}_1, I_p)$, $\boldsymbol{g} \sim \mathcal{N}_p(\boldsymbol{0}, I_p)$ and $W \sim \mathcal{W}_p(n, I_p)$. Then, it holds that*

$$\text{(i) } \boldsymbol{z}'W^{-1}\boldsymbol{z} = \frac{(u_1+\tau)^2 + \tilde{v}_1}{\tilde{v}_2},$$

$$\text{(ii) } \boldsymbol{z}'W^{-2}\boldsymbol{z} = \frac{(u_1+\tau)^2 + \tilde{v}_1}{\tilde{v}_2^2}\left(1 + \frac{\tilde{v}_3}{\tilde{v}_4}\right),$$

$$\text{(iii) } \tau \boldsymbol{e}_1'W^{-1}\boldsymbol{z} = \frac{\tau}{\tilde{v}_2}\left\{u_1 + \tau + \left(\frac{\tilde{v}_1\tilde{v}_3}{\tilde{v}_4}\right)^{\frac{1}{2}} \frac{u_2}{\sqrt{\tilde{v}_5 + u_2^2}}\right\},$$

$$\text{(iv) } \boldsymbol{z}'W^{-1}\boldsymbol{g} = \frac{\{(u_1+\tau)^2 + \tilde{v}_1\}^{\frac{1}{2}}}{\tilde{v}_2}\left\{u_3 - u_4\left(\frac{\tilde{v}_3}{\tilde{v}_4}\right)^{\frac{1}{2}}\right\},$$

where, $\boldsymbol{e}_1 = (1,0,0,\cdots,0)$. Here, $u_i \sim \mathcal{N}(0,1)$, $i = 1,2,3,4$, $\tilde{v}_1 \sim \chi^2_{p-1}$, $\tilde{v}_2 \sim \chi^2_{N-p-1}$, $\tilde{v}_3 \sim \chi^2_{p-1}$, $\tilde{v}_4 \sim \chi^2_{N-p}$ and $\tilde{v}_5 \sim \chi^2_{p-2}$ and these variables are mutually independent.

(Proof) Fujikoshi (2000) has derived the stochastic expressions (i)-(iii). In this proof, we prove the stochastic expression (iv). Let P be the random orthogonal matrix with its

first column proportional to $\boldsymbol{z}$ i.e. $PP' = I_p$. Then $PWP' \sim \mathcal{W}_p(n, I_p)$ and

$$(PWP')^{-1} = \begin{pmatrix} \{\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y}\}^{-1} & * \\ -\{\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y}\}^{-1}(Y'Y)^{-1}Y'\boldsymbol{y} & Y'(I-\Pi_{\boldsymbol{y}})Y \end{pmatrix},$$

where $\Pi_Y = Y(Y'Y)^{-1}Y'$ and $\Pi_{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{y}'\boldsymbol{y})^{-1}\boldsymbol{y}'$. At first, we obtain the stochastic expressions (i)-(iii) by using Fujikoshi (2000). The parts (i)-(iii) can be expressed as

$$\begin{aligned}
\boldsymbol{z}'W^{-1}\boldsymbol{z} &= \frac{\boldsymbol{z}'\boldsymbol{z}}{\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y}}, \\
\boldsymbol{z}'W^{-2}\boldsymbol{z} &= \frac{\boldsymbol{z}'\boldsymbol{z}}{(\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y})^2}\left(1 + \frac{\boldsymbol{y}'Y(Y'Y)^{-2}Y'\boldsymbol{y}}{\boldsymbol{y}'\Pi_Y\boldsymbol{y}}\boldsymbol{y}'\Pi_Y\boldsymbol{y}\right), \\
\tau\boldsymbol{e}_1'W^{-1}\boldsymbol{z} &= \frac{\tau}{\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y}}\left\{\tau + \boldsymbol{e}_1'\boldsymbol{u} + \left(-(\boldsymbol{z}_2'\boldsymbol{z}_2)^{1/2}P_2'\boldsymbol{e}_1\right)'(Y'Y)^{-1}Y'\boldsymbol{y}\right\}.
\end{aligned}$$

From Fujikoshi (2002), we have

$$\frac{(-(\boldsymbol{z}_2'\boldsymbol{z}_2)^{1/2}P_2'\boldsymbol{e}_1)'(Y'Y)^{-1}Y'\boldsymbol{y}}{(\boldsymbol{z}_2'\boldsymbol{z}_2)\boldsymbol{e}_1'P_2P_2'\boldsymbol{e}_1\boldsymbol{y}'Y(Y'Y)^{-2}Y'\boldsymbol{y}} = \frac{u_2}{\sqrt{u_2^2+v_5}},$$

where

$$\begin{aligned}
u_1 &= \boldsymbol{e}_1'\boldsymbol{u} \sim \mathcal{N}(0,1),\ u_2 \sim \mathcal{N}(0,1), \tilde{v}_1 = \boldsymbol{u}'\left(I_{p-1} - (\boldsymbol{w}'\boldsymbol{w})^{-1}\boldsymbol{w}\boldsymbol{w}'\right)\boldsymbol{u} \sim \chi^2_{p-1}, \\
\tilde{v}_2 &= \boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y} \sim \chi^2_{N-p-1}, \tilde{v}_3 = \boldsymbol{y}'\Pi_Y\boldsymbol{y} \sim \chi^2_{p-1}, \\
\tilde{v}_4 &= \boldsymbol{y}'\Pi_Y\boldsymbol{y}(\boldsymbol{y}'Y(Y'Y)^{-2}Y'\boldsymbol{y})^{-1} \sim \chi^2_{N-p}, \tilde{v}_5 = \boldsymbol{y}'\Pi_Y\boldsymbol{y}(\boldsymbol{y}'Y(Y'Y)^{-2}Y'\boldsymbol{y})^{-1} \sim \chi^2_{p-2}.
\end{aligned}$$

Next, we prove (iv). The part (iv) denotes

$$\begin{aligned}
\boldsymbol{z}'W^{-1}\boldsymbol{g} = \frac{(\boldsymbol{z}'\boldsymbol{z})^{\frac{1}{2}}}{\boldsymbol{y}'(I-\Pi_Y)\boldsymbol{y}}\Bigg(&u_3 - \frac{\boldsymbol{y}'Y(Y'Y)^{-1}\boldsymbol{g}_2}{(\boldsymbol{y}'Y(Y'Y)^{-2}Y\boldsymbol{y})^{\frac{1}{2}}} \\
&\times \left(\frac{\boldsymbol{y}'Y(Y'Y)^{-2}Y\boldsymbol{y}}{\boldsymbol{y}'Y(Y'Y)^{-1}Y'\boldsymbol{y}}\right)^{\frac{1}{2}}(\boldsymbol{y}'Y(Y'Y)^{-1}Y'\boldsymbol{y})^{\frac{1}{2}}\Bigg),
\end{aligned}$$

where $\boldsymbol{g} = (u_3, \boldsymbol{g}_2')'$. On the other hand, we note that

$$u_3 = \boldsymbol{e}_1'\boldsymbol{g} \sim \mathcal{N}(0,1),\ u_4 = (\boldsymbol{y}'Y(Y'Y)^{-2}Y\boldsymbol{y})^{-\frac{1}{2}}\boldsymbol{y}'Y(Y'Y)^{-1}\boldsymbol{g_2} \sim \mathcal{N}(0,1).$$

Then, the variables $u_i,\ i = 1,2,3,4$ and $v_i,\ i = 1,2,3,4,5$ are mutually independent. ■

6.2 Stochastic expansions of U and V

To expand U and V stochastically, define random variables $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ by

$$\begin{aligned}
\boldsymbol{z}_1 =& \sqrt{\frac{N_1N_2}{N}}\Sigma(\boldsymbol{j})^{-1/2}(\overline{\boldsymbol{x}}_1(\boldsymbol{j}) - \overline{\boldsymbol{x}}_2(\boldsymbol{j})), \\
\boldsymbol{z}_2 =& \frac{1}{\sqrt{N}}\Sigma(\boldsymbol{j})^{-1/2}(N_1\overline{\boldsymbol{x}}_1(\boldsymbol{j}) + N_2\overline{\boldsymbol{x}}_2(\boldsymbol{j}) - N_1\boldsymbol{\mu}_1(\boldsymbol{j}) - N_2\boldsymbol{\mu}_2(\boldsymbol{j})), \\
W =& n\Sigma(\boldsymbol{j})^{-1/2}S(\boldsymbol{j})\Sigma(\boldsymbol{j})^{-1/2}.
\end{aligned}$$

It is seen that $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and W are mutually independently distributed as $\boldsymbol{z}_1 \sim \mathcal{N}_p(\boldsymbol{\tau}(\boldsymbol{j}), I_{\sharp(\boldsymbol{j})})$, $\boldsymbol{z}_2 \sim \mathcal{N}_p(\boldsymbol{0}, I_{\sharp(\boldsymbol{j})})$ and $W \sim \mathcal{W}_p(n, I_{\sharp(\boldsymbol{j})})$, respectively, where $\boldsymbol{\tau}(\boldsymbol{j}) = \sqrt{(N_1N_2)/N}\Sigma(\boldsymbol{j})^{-1/2}(\boldsymbol{\mu}_1(\boldsymbol{j}) - \boldsymbol{\mu}_2(\boldsymbol{j}))$. Using these variables, we can rewrite U and V as

$$
\begin{aligned}
U =& -\frac{(N_1 - N_2)n}{2N_1N_2}\boldsymbol{z}_1'W^{-1}\boldsymbol{z}_1 + \frac{n}{\sqrt{N_1N_2}}\boldsymbol{z}_1'W^{-1}\boldsymbol{z}_2 - \frac{n}{N_1}\boldsymbol{\tau}(\boldsymbol{j})'W^{-1}\boldsymbol{z}_1, \\
V =& \frac{n^2N}{N_1N_2}\boldsymbol{z}_1'W^{-2}\boldsymbol{z}_1.
\end{aligned}
$$

Let Γ be an orthogonal matrix with its first column proportional to $\boldsymbol{\tau}(\boldsymbol{j})$ i.e. $\Gamma\Gamma' = I_p$. Then $\Gamma W\Gamma' \sim \mathcal{W}_p(n, I_p)$, $\Gamma\boldsymbol{z}_1 \sim \mathcal{N}_p(\tau\boldsymbol{e}_1, I_{\sharp(\boldsymbol{j})})$, $\Gamma\boldsymbol{z}_2 \sim \mathcal{N}_p(\boldsymbol{0}, I_{\sharp(\boldsymbol{j})})$ and $W \sim \mathcal{W}_p(n, I_{\sharp(\boldsymbol{j})})$, respectively, where $\tau^2 = (N_1N_2)/N\Delta(\boldsymbol{j})^2$. For simplicity, we denote the statistics $\Gamma W\Gamma'$, $\Gamma\boldsymbol{z}_1$ and $\Gamma\boldsymbol{z}_2$ by W, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, respectively. We can rewrite the statistics U and V as

$$
\begin{aligned}
U =& -\frac{(N_1 - N_2)n}{2N_1N_2}\boldsymbol{z}_1'W^{-1}\boldsymbol{z}_1 + \frac{n}{\sqrt{N_1N_2}}\boldsymbol{z}_1'W^{-1}\boldsymbol{z}_2 - \frac{n}{N_1}\tau\boldsymbol{e}_1'W^{-1}\boldsymbol{z}_1, \\
V =& \frac{n^2N}{N_1N_2}\boldsymbol{z}_1'W^{-2}\boldsymbol{z}_1.
\end{aligned}
$$

Using Lemma 6.1, we express U and V as

$$
\begin{aligned}
U =& -\frac{(N_1 - N_2)n}{2N_1N_2}\frac{(u_1 + \tau)^2 + \tilde{v}_1}{\tilde{v}_2} + \frac{n}{\sqrt{N_1N_2}}\frac{\{(u_1 + \tau)^2 + \tilde{v}_1\}^{\frac{1}{2}}}{\tilde{v}_2}\left\{u_3 - u_4\left(\frac{\tilde{v}_3}{\tilde{v}_4}\right)^{\frac{1}{2}}\right\} \\
& - \frac{n}{N_1}\frac{\tau}{\tilde{v}_2}\left\{u_1 + \tau + \left(\frac{\tilde{v}_1\tilde{v}_3}{\tilde{v}_4}\right)^{\frac{1}{2}}\frac{u_2}{\sqrt{\tilde{v}_5 + u_2^2}}\right\}, \\
V =& \frac{n^2N}{N_1N_2}\frac{(u_1 + \tau)^2 + \tilde{v}_1}{\tilde{v}_2^2}\left(1 + \frac{\tilde{v}_3}{\tilde{v}_4}\right).
\end{aligned}
$$

Define variables v_1, v_2, v_3 v_4 and v_5 by

$$
\begin{aligned}
v_1 &= \frac{\tilde{v}_1 - (p-1)}{\sqrt{2(p-1)}},\ v_2 = \frac{\tilde{v}_2 - (N-p-1)}{\sqrt{2(N-p-1)}},\ v_3 = \frac{\tilde{v}_3 - (p-1)}{\sqrt{2(p-1)}},\ v_4 = \frac{\tilde{v}_4 - (N-p)}{\sqrt{2(N-p)}}, \\
v_5 &= \frac{\tilde{v}_5 - (p-2)}{\sqrt{2(p-2)}}.
\end{aligned}
$$

Note that v_i is asymptotically distributed as $\mathcal{N}(0,1)$ under condition (A1). Using Taylor series expansion based on these variables, we can expand U stochastically as

$$
U = U_0 + \frac{1}{\sqrt{n}}U_1 + \frac{1}{n}U_2, , \tag{6.1}
$$

where

$$
\begin{aligned}
U_0 &= -\frac{1}{2(1-c)}\left(\Delta(\boldsymbol{j})^2+\frac{c(\gamma_1-\gamma_2)}{\gamma_1\gamma_2}\right),\\
U_1 &= \frac{\sqrt{\gamma_1\gamma_2\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right)}}{(1-c)\sqrt{\gamma_1\gamma_2}}u_3-\frac{\sqrt{c}\sqrt{\gamma_1\gamma_2\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right)}}{(1-c)^{3/2}\sqrt{\gamma_1\gamma_2}}u_4\\
&\quad+\left(\frac{\Delta(\boldsymbol{j})^2}{\sqrt{2}(1-c)^{3/2}}+\frac{c}{\sqrt{2}(1-c)^{3/2}\gamma_1\gamma_2}-\frac{\sqrt{2}c}{(1-c)^{3/2}\gamma_1}\right)v_2\\
&\quad+\left(-\frac{\Delta(\boldsymbol{j})\gamma_2\sqrt{\frac{1}{\gamma_1\gamma_2}}}{1-c}+\frac{2\Delta(\boldsymbol{j})\sqrt{\gamma_1\gamma_2}}{(1-c)\gamma_1}-\frac{\Delta(\boldsymbol{j})\sqrt{\gamma_1\gamma_2}}{(1-c)\gamma_1\gamma_2}\right)u_1\\
&\quad-\frac{\sqrt{c}\Delta(\boldsymbol{j})\gamma_2\sqrt{\frac{1}{\gamma_1\gamma_2}}}{(1-c)^{3/2}}u_2+\left(\frac{\sqrt{2}\sqrt{c}}{(1-c)\gamma_1}-\frac{\sqrt{c}}{\sqrt{2}(1-c)\gamma_1\gamma_2}\right)v_1,\\
U_2 &= \left(\frac{\sqrt{c}\left(u_3-\frac{\sqrt{c}u_4}{\sqrt{1-c}}\right)}{\sqrt{2}(1-c)\gamma_1\gamma_2\sqrt{\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2}}-\frac{\Delta(\boldsymbol{j})u_2\sqrt{\frac{\gamma_2}{\gamma_1}}}{\sqrt{2}(1-c)^{3/2}}\right.\\
&\qquad\left.+\frac{\sqrt{c}v_2}{(1-c)^{3/2}\gamma_1\gamma_2}-\frac{2\sqrt{c}v_2}{(1-c)^{3/2}\gamma_1}\right)v_1\\
&\quad+\left(-\frac{\sqrt{2}\left(u_3-\frac{\sqrt{c}u_4}{\sqrt{1-c}}\right)\sqrt{\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2}}{(1-c)^{3/2}}-\frac{\sqrt{2}\Delta(\boldsymbol{j})u_1\sqrt{\frac{\gamma_2}{\gamma_1}}}{(1-c)^{3/2}}\right.\\
&\qquad\left.+\frac{\sqrt{2}\Delta(\boldsymbol{j})u_1}{(1-c)^{3/2}\sqrt{\gamma_1\gamma_2}}+\frac{\sqrt{2}\sqrt{c}\Delta(\boldsymbol{j})u_2\sqrt{\frac{\gamma_2}{\gamma_1}}}{(1-c)^2}\right)v_2\\
&\quad+\frac{\Delta(\boldsymbol{j})\left(\sqrt{1-c}u_3-\sqrt{c}u_4\right)u_1}{(1-c)^{3/2}\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}}-\frac{\left(u_4\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}+\Delta(\boldsymbol{j})\gamma_2u_2\right)v_3}{\sqrt{2-2c}(1-c)\sqrt{\gamma_1\gamma_2}}\\
&\quad+\frac{\sqrt{c}\left(u_4\sqrt{c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2}+\Delta(\boldsymbol{j})\gamma_2u_2\right)v_4}{\sqrt{2}(1-c)^2\sqrt{\gamma_1\gamma_2}}-\frac{v_2^2\left(-2c\gamma_2+c+\Delta(\boldsymbol{j})^2\gamma_1\gamma_2\right)}{(c-1)^2\gamma_1\gamma_2}\\
&\quad+\frac{\Delta(\boldsymbol{j})v_5u_2\sqrt{\frac{\gamma_2}{\gamma_1}}}{\sqrt{2}(1-c)^{3/2}}-\frac{(1-2\gamma_2)u_1^2}{2(1-c)\gamma_1\gamma_2}+\frac{1-2\gamma_2}{2\gamma_1\gamma_2-2c\gamma_1\gamma_2},
\end{aligned}
$$

Here, $u_i,\ i=1,\dots 4$ and $v_j,\ j=1,\dots,5$ are mutually independent and they are asymptotically distributed as $\mathcal{N}(0,1)$.

Using similar arguments, we can expand Vstochastically as

$$V=V_0+\frac{V_1}{\sqrt{n}}+\frac{V_2}{n}, \tag{6.2}$$

where

$$
\begin{aligned}
V_0 &= \frac{1}{(1-c)^3}\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right), \\
V_1 &= \frac{\frac{\sqrt{2}\sqrt{c}v_1}{\gamma_1\gamma_2}+2u_1\sqrt{\frac{\Delta(\boldsymbol{j})^2}{\gamma_1\gamma_2}}}{(1-c)^3}+\frac{c\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right)\left(-\frac{2\sqrt{2}v_2}{\sqrt{1-cc}}+\frac{\sqrt{2}v_3}{\sqrt{c}}-\frac{\sqrt{2}v_4}{\sqrt{1-c}}\right)}{(1-c)^3}, \\
V_2 &= \frac{1}{(1-c)^3}v_2^2\left(\frac{6\Delta(\boldsymbol{j})^2}{1-c}+\frac{6c}{(1-c)\gamma_1\gamma_2}\right) \\
&\quad -\frac{1}{(1-c)^3}\left(\frac{4\sqrt{2}\Delta(\boldsymbol{j})v_2u_1\sqrt{\frac{1}{\gamma_1\gamma_2}}}{\sqrt{1-c}}+\frac{4\sqrt{c}v_1v_2}{\sqrt{1-c}\gamma_1\gamma_2}-\frac{u_1^2}{\gamma_1\gamma_2}+\frac{1}{\gamma_1\gamma_2}\right) \\
&\quad +c\left(\frac{\sqrt{2}v_3}{(1-c)^3\sqrt{c}}-\frac{\sqrt{2}v_4}{(1-c)^{7/2}}\right)\left(\frac{\sqrt{2}\sqrt{c}v_1}{\gamma_1\gamma_2}-\frac{2\sqrt{2}v_2\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right)}{\sqrt{1-c}}\right. \\
&\quad \left.+2u_1\sqrt{\frac{\Delta(\boldsymbol{j})^2}{\gamma_1\gamma_2}}\right)+\frac{c}{(1-c)^3}\left(\frac{2v_4^2}{1-c}-\frac{2v_3v_4}{\sqrt{1-c}\sqrt{c}}\right)\left(\frac{c}{\gamma_1\gamma_2}+\Delta(\boldsymbol{j})^2\right).
\end{aligned}
$$

It can be seen that the stochastic expansions of U and V under the large sample framework (A0) is provided by replacing c in the limiting terms U_0 and V_0 with p/n, and by replacing c in first and second terms U_1, U_2, V_1 and V_2 with 0.

Acknowledgments.

Research of the second author was supported in part by Grant-in-Aid for Scientific Research (21540114 and 23243039) from Japan Society for the Promotion of Science.

References

[1] Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criterion. *J. Multivariate. Anal.*, **17**, 27-37.

[2] Fujikoshi, Y. (2002). Selection of variables for discriminant analysis in a high-dimensional case. *Sankhya Ser.* A, **64**, 256-257.

[3] Fujikoshi, Y., Ulyanov, and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, Hobeken, N. J.

[4] Fujikoshi, Y. and Seo, T. (1998). Asymptotic approximations for EPMC's of the linear and the quadratic discriminant function when the sample size and the dimension are large. *Random Operators and Stochastic Equations*, **6**, 269–280.

[5] Kshirsagar, A. M. (1972). *Multivariate Anaiysis*. New York Marcel Dekker.

[6] Kubokawa, T, Hyodo, M., and Srivastava, M.S. (2013). Asymptotic expansion and estimation of EPMC for linear classification rules in high dimension. *J. Multivariate Analysis*, to appear.

[7] McLachlan, G. J. (1976). A criterion for selecting variables for the linear discriminant function. *Biomerrics*, **32**, 529-515.

[8] McLachlan, G. J. (198O). Selection of variables in discriminant analysis (Letter to the Editor). *Biometrics*, **36**, 554.

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

[10] Rao, C. R. (1973). *Linear Starisrical Inference and Its Applicatwns*. Second Edition. New York Wiley.

[11] Raudys, S. (1972). On the amount of a priori information in designing the classification algorithm. *Technical Cybernetics*, **4**, 168-174.

[12] Sakurai, T., Nakada, T., and Fujikoshi, Y. (2012). High-Dimensional AICs for Selection of Redundancy Models in Discriminant Analysis. Technical Report 12-13, Hiroshima Statistical Research Group, Hiroshima University.

[13] Wilbur, J.D., Ghosh, J.K., Nakatsu, C.H., and Doerge, R.W. (2002). Variable selection in high-dimensional multivariate binary data with application to analysis of microbial community DNA finger prints. *Biometrics*. **58**, 378-386.

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