Consistency of Test-based Criterion for Selection of Variables in High-dimensional Two Group-Discriminant Analysis

Yasunori Fujikoshi[*] and Tetsuro Sakurai[**]

[*] *Department of Mathematics, Graduate School of Science,*
Hiroshima University, 1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

[**] *Center of General Education, Tokyo University of Science, Suwa,*
5000-1 Toyohira, Chino, Nagano 391-0292, Japan

Abstract

This paper is concerned with selection of variables in two-group discriminant analysis with the same covariance matrix. We propose a test-based criterion (TC) drawing on the significance of each variable. The selection method can be applied for high-dimensional data. Sufficient conditions for the test-based criterion to be consistent are provided when the dimension and the sample are large. For the case that the dimension is larger than the sample size, the regularized method is proposed. Our results, and tendencies therein are explored numerically through a Monte Carlo simulation.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: Consistency, Discriminant analysis, High-dimensional framework, Selection of variables, Test-based criterion.

1. Introduction

This paper is concerned with the variable selection problem in two-group discriminant analysis with the same covariance matrix. In a variable selection problem under such discriminant model, one of the goals is to find a subset of variables whose coefficients of the linear discriminant function are not zero. Several methods including model section criteria AIC and BIC have been developed. It is known (see, e.g., Fujikoshi 1984, Nishii et al.1988) that in a large sample framework, AIC is not consistent, but BIC is consistent. On the other hand, the selection methods are based on the minimization of the criteria, and become computationally onerous when the dimension is large. Though some stepwise methods have been proposed, their optimality is not known. In our discriminant model, there are methods based on misclassification errors by MaLachlan (1976), Fujikoshi (1985), Hyodo and Kubokawa (2014), Yamada et al. (2018) for a high-dimensional case as well as a large-sample case. It is known (see, e.g., Fujikoshi 1985) that two methods by misclassification error rate and Akaike's information criterion are asymptotically equivalent under a large-sample framework. For high-dimensional data, Lasso and other regularization methods have been extended. For such study, see, e.g., Clemmensen et al. (2011), Witten and Tibshirani (2011), etc.

In this paper we propose a test-based criterion based on significance test of each variable, which is useful for high-dimensional data as well as large-sample data. The criterion involves a constant term which should be determined by point of some optimality. We propose a class of constants satisfying a consistency when the dimension and the sample size are large. For the case when the dimension is larger than the sample size, a regularized method is numerically examined. Our results, and tendencies therein are explored numerically through a Monte Carlo simulation.

The remainder of the present paper is organized as follows. In Section 2, we present the relevant notation and the test-based method. In Sections

3 we derive sufficient conditions for the test-based criterion to be consistent under a high-dimensional case. In Section 4 we study the test-based criterion through a Monte Carlo simulation. In Section 5, we propose the ridge-type criteria, whose consistency properties are numerically examined. In Section 6, conclusions are offered. All proofs of our results are provided in the Appendix.

2. Test-based Criterion

In two-group discriminant analysis, suppose that we have independent samples $\boldsymbol{y}_1^{(i)}, \ldots, \boldsymbol{y}_{n_i}^{(i)}$ from p-dimensional normal distributions $\mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, 2$. Let $\mathbf{Y}$ be the total sample matrix defined by

$$\mathbf{Y} = (\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_{n_1}^{(1)}, \boldsymbol{y}_1^{(2)}, \ldots, \boldsymbol{y}_{n_2}^{(2)})'.$$

The coefficients of the population discriminant function are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) = (\beta_1, \ldots, \beta_p)'. \tag{2.1}$$

Let Δ and D be the population and the sample Mahalanobis distances defined by $\Delta = \left\{(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(1)})\right\}^{1/2}$, and

$$D = \left\{(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'\mathbf{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})\right\}^{1/2},$$

respectively. Here $\bar{\boldsymbol{x}}^{(1)}$ and $\bar{\boldsymbol{x}}^{(2)}$ are the sample mean vectors, and $\mathbf{S}$ be the pooled sample covariance matrix based on $n = n_1 + n_2$ samples.

Suppose that j denotes a subset of $\omega = \{1, \ldots, p\}$ containing p_j elements, and $\boldsymbol{y}_j$ denotes the p_j vector consisting of the elements of $\boldsymbol{y}$ indexed by the elements of j. We use the notation D_j and D_ω for D based on $\boldsymbol{y}_j$ and $\boldsymbol{y}_\omega(= \boldsymbol{y})$, respectively. Let M_j be a variable selection model defined by

$$M_j: \ \beta_i \neq 0 \text{ if } i \in j, \text{ and } \beta_i = 0 \text{ if } i \notin j. \tag{2.2}$$

We identify the selection of M_j with the selection of $\boldsymbol{y}_j$. Let AIC_j be the AIC for M_j. Then, it is known (see, e.g., Fujikoshi 1985) that

$$\begin{aligned} \mathrm{A}_j &= \mathrm{AIC}_j - \mathrm{AIC}_\omega \\ &= n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_j^2)}{n - 2 + g^2 D_j^2} \right\} - 2(p - p_j), \end{aligned} \tag{2.3}$$

where $g = \sqrt{(n_1 n_2)/n}$. Similarly, let BIC_j be the BIC for M_j, and we have

$$\begin{aligned} \mathrm{B}_j &= \mathrm{BIC}_j - \mathrm{BIC}_\omega \\ &= n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_j^2)}{n - 2 + g^2 D_j^2} \right\} - (\log n)(p - p_j). \end{aligned} \tag{2.4}$$

In a large sample framework, it is known (see, Fujikoshi 1985, Nishii et al. 1988) that AIC is not consistent, but BIC is consistent. On the other hand, the variable selection methods based on AIC and BIC are given as $\min_j \mathrm{AIC}_j$ and $\min_j \mathrm{BIC}_j$, respectively. Therefore, such criteria become computationally onerous when p is large. To circumvent this issue, we consider a test-based criterion (TC) drawing on the significance of each variable. A critical region for "$\beta_i = 0$" based on the likelihood ratio principle is expressed (see, e.g., Rao 1973, Fujikoshi et al. 2010) as

$$\mathrm{T}_{d,i} = n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_{(-i)}^2)}{n - 2 + g^2 D_{(-i)}^2} \right\} - d > 0, \tag{2.5}$$

where $(-i), i = 1, \ldots, p$ is the subset of $\omega = \{1, \ldots, p\}$ obtained by omitting the i from ω, and d is a positive constant which may depend on p and n. Note that

$$\mathrm{T}_{2,i} > 0 \iff \mathrm{AIC}_{\omega(-i)} - \mathrm{AIC}_\omega > 0.$$

We consider a test-based criterion for the selection of variables defined by selecting the set of suffixes or the set of variables given by

$$\mathrm{TC}_d = \{ i \in \omega \mid \mathrm{T}_{d,i} > 0 \}, \tag{2.6}$$

or $\{ y_i \in \{y_1, \ldots, y_p\} \mid \mathrm{T}_{d,i} > 0 \}$. The notation $\widehat{j}_{\mathrm{TC}_d}$ is also used for TC_d.

In general, if d is large, a small number of variables is selected. On the other hand, if d is small, a large number of variables is selected. Ideally, we want to select only the true variables whose discriminant coefficients are not zero. For a test-based criterion, there is an important problem how to decide the constant term d. Nishii et al. (1988) have used a special case with $d = \log n$. They noted that under a large sample framework, $\mathrm{TC}_{\log n}$ is consistent. However, we note that $\mathrm{TC}_{\log n}$ will be not consistent for a high-dimensional case, through a simulation experiment. We propose a class of d satisfying a high-dimensional consistency including $d = \sqrt{n}$ in Section 3.

3. Consistency of TC_d under High-dimensional Framework

For studying consistency of the variable selection criterion TC_d, it is assumed that the true model M_* is included in the full model. Let the minimum model including M_* be M_{j_*}. For a notational simplicity, we regard the true model M_* as M_{j_*}. Let $\mathcal{F}$ be the entire suite of candidate models, that is

$$\mathcal{F} = \{\{1\}, \ldots, \{k\}, \{1, 2\}, \ldots, \{1, \ldots, k\}\}.$$

A subset j of ω is called overspecified model if j include the true model. On the other hand, a subset j of ω is called underspecified model if j does not include the true model. Then, $\mathcal{F}$ is separated into two sets, one is a set of overspecified models, i.e., $\mathcal{F}_+ = \{j \in \mathcal{F} \mid j_* \subseteq j\}$ and the other is a set of underspecified models, i.e., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F}$. It is said that a model selection criterion $\widehat{j}$ has a high-dimensional consistency if

$$\lim_{p/n \to c \in (0,\infty)} \Pr(\widehat{j} = j_*) = 1.$$

Here we list some of our main assumptions:

A1 (The true model): $M_{j_*} \in \mathcal{F}$.

A2 (The high-dimensional asymptotic framework):
$p \to \infty,\ n \to \infty,\ p/n \to c \in (0, \infty),\ \ n_i/n \to k_i > 0,\ (i = 1, 2).$

For the dimensionality p_* of the true model and the Mahalanobis distance Δ, the following two cases are considered:

A3 : p_* is finite, and $\Delta^2 = \mathrm{O}(1)$.

For the constant d of test-based statistic $\mathrm{T}_{d,i}$ in (2.5), we consider the following assumptions:

B1 : $d/n \to 0$.

B2 : $h \equiv d/n - 1/(n - p - 3) > 0$, and $h = \mathrm{O}(n^{-a})$, where $0 < a < 1$.

A consistency of TC_d in (2.6) for some $d > 0$ shall be shown along the following outline: In general, we have

$$\mathrm{TC}_d = j_* \Leftrightarrow \text{"}\mathrm{T}_{d,i} > 0 \text{ for } i \in j_*\text{"},\ \text{ and "}\mathrm{T}_{d,i} \leq 0 \text{ for } i \notin j_*\text{"}$$

Therefore

$$\begin{aligned} P(\mathrm{TC}_d = j_*) &= P\left(\bigcap_{i\in j_*} \text{"}\mathrm{T}_{d,i} > 0\text{"} \bigcap_{i\notin j_*} \text{"}\mathrm{T}_{d,i} < 0\text{"}\right) \\ &= 1 - P\left(\bigcup_{i\in j_*} \text{"}\mathrm{T}_{d,i} \leq 0\text{"} \bigcup_{i\notin j_*} \text{"}\mathrm{T}_{d,i} \geq 0\text{"}\right) \\ &\geq 1 - \sum_{i\in j_*} P(\mathrm{T}_{d,i} \leq 0) - \sum_{i\notin j_*} P(\mathrm{T}_{d,i} \geq 0). \end{aligned}$$

We shall consider to show

$$[\mathrm{F1}] \equiv \sum_{i\in j_*} P(\mathrm{T}_{d,i} \leq 0) \to 0. \tag{3.1}$$

$$[\mathrm{F2}] \equiv \sum_{i\notin j_*} P(\mathrm{T}_{d,i} \geq 0) \to 0, \tag{3.2}$$

[F1] denotes the probability such that the true variables are not selected. [F2] denotes the probability such that the non true variables are selected.

Theorem 3.1. *Suppose that assumptions* A1, A2 *and* A3 *are satisfied. Then, the test-based criterion* TC_d *is consistent if* B1 *and* B2 *are satisfied.*

Let $d = n^r$, where $0 < r < 1$. Then, $h = \mathrm{O}(n^{-(1-r)})$, and the condition B2 is satisfied. So, the test-based criteria with

$$d = n^{3/4}, \quad n^{2/3}, \quad n^{1/2}, \quad n^{1/3} \quad \text{or} \quad n^{1/4}$$

have a high-dimensional consistency. Among of them, we have numerically seen that the one with $d = \sqrt{n}$ has a good behavior. Note that TC_2 and $\mathrm{TC}_{\log n}$ do not satisfy B2.

As a special case, under the assumptions of Theorem 3.1 we have seen that the probability of selecting overspecified models tends to zero, that is

$$\sum_{i \notin j_*} P(\mathrm{TC}_d = i) = \sum_{i \notin j_*} P(\mathrm{T}_{d,i} \geq 0) \to 0.$$

The proof given there is applicable also to the case replaced assumption A3 by assumption A4:

A4 : $p_* = \mathrm{O}(p)$, and $\Delta^2 = \mathrm{O}(p)$.

In other words, such a property holds regardless of whether the dimension of j_* is finite or not. Further, it does not depend on the order of the Mahalanobis distance. The square of Mahalanobis distance of $\boldsymbol{y}$ is decomposed as a sum of the squares of Mahalanobis distance of $\boldsymbol{y}_{(-i)}$ and the conditional Mahalanobis distance of $\boldsymbol{y}_{\{i\}}$ given $\boldsymbol{y}_{(-i)}$ as follows:

$$\Delta^2 = \Delta^2_{(-i)} + \Delta^2_{\{i\}\cdot(-i)}. \tag{3.3}$$

When $i \in j_*$, $(-i) \notin \mathcal{F}_+$ and hence

$$\Delta^2_{\{i\}\cdot(-i)} = \Delta^2 - \Delta^2_{(-i)} > 0.$$

Related to consistency of TC_d under assumption A4, we consider the following assumption:

A5 : For $i \in j_*$,

$$\Delta^2_{(-i)} = \mathrm{O}(p), \ , \Delta^2_{\{i\}\cdot(-i)} = \mathrm{O}(p^b), \quad 0 < b < 1.$$

Theorem 3.2. *Suppose that assumptions* A1, A2, A4 *and* A5 *are satisfied. Then, the test-based criterion* TC_d *is consistent if* B1, B2 *and* "$a < b, \quad 3/4 < b$" *are satisfied.*

It is conjectured that the condition "$3/4 < b$" can be replaced by "$1/2 < b$"

Theorem 3.3. *Suppose that assumption;* A1 *and*

$$"p; \text{ fixed}, \ n_i/n \to k_i(i = 1, 2), \ d \to \infty, \ d/n \to 0"$$

are satisfied. Then, the test-based criterion TC_d *is consistent when* n *tends to infty.*

From Theorem 3.3 we can see that $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ are consistent under a large-sample framework.

4. Numerical Study

In this section we numerically explore the validity of our claims through three test-based criteria, TC_2, $\mathrm{TC}_{\log n}$, and $\mathrm{TC}_{\sqrt{n}}$. Note that $\mathrm{TC}_{\sqrt{n}}$ satisfies sufficient conditions B1 and B2 for its consistency, but TC_2 and $\mathrm{TC}_{\log n}$ do not satisfy them.

The true model was assumed as follows: the true dimension is $j_* = 3, 6$, the true mean vectors;

$$\boldsymbol{\mu}_1 = \alpha(1, \dots, 1, 0, \dots, 0)', \quad \boldsymbol{\mu}_2 = \alpha(-1, \dots, -1, 0, \dots, 0)',$$

and the true covariance matrice $\mathbf{\Sigma}_* = \mathbf{I}_p$.

The selection rates associated with these criteria are given in Tables 4.1 to 4.3. "Under", "True", and "Over" denote the underspecified models, the true model, and the overspecified models, respectively. We focused on selection rates for 10^3 replications in Tables 4.1 $\sim$ 4.2, and for 10^2 replications in Tables 4.3.

From Tables 4.1, we can identify the following tendencies.

- The selection probabilities of the true model by TC_2 are relatively large when the dimension is small as in the case $p = 5$. However, the values do not approach to 1 as n increases, and it seems that TC_2 has no consistency under a large sample case.

- The selection probabilities of the true model by $\mathrm{TC}_{\log n}$ are near to 1, and has a consistency under a large sample case. However, the probabilities are decreasing as p increases, and so will not a consistency in a high-dimensional case.

- The selection probabilities of the true model by $\mathrm{TC}_{\sqrt{n}}$ approach 1 as n is large, even if p is small. Further, if p is large, but under a high-dimensional framework such that n is also large, then it has a consistency. However, the probabilities decrease as the ratio p/n approaches 1.

- As the quantity α presenting a distance between two groups becomes large, the selection probabilities of the true model by TC_2 and $\mathrm{TC}_{\log n}$ increase in a large sample se as in the case $p = 5$. However, the effect becomes small when p is large. On the other hand, the selection probabilities of the true model by $\mathrm{TC}_{\log n}$ increase in a sense both in large-sample and high-dimensional cases in select

In Table 4.2, we examine the case where the dimension p_* of the true model is larger that the one in Table 4.1. The following tendencies can be identified

- As is the case with Table 4.1, TC_2 and $\mathrm{TC}_{\log n}$ are not consistent as p increases, but $\mathrm{TC}_{\sqrt{n}}$ is consistent. In general, the probability of selecting the true model decreases as the dimension of the true model is large.

In Table 4.3, we examine the case where the dimension p_* of the true model is relatively large, and especially in the case of $p_* = p/4$. The following tendencies can be idetified.

- When $p_* = p/4$, the consistency of TC_2 and $\mathrm{TC}_{\log n}$ can be not seen. The consistency of $\mathrm{TC}_{\sqrt{n}}$ can be seen when n is large.

Table 4.1. Selection rates of TC_2, $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ for $p_* = 3$

$p_* = 3$, $\alpha = 1$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	5	0.00	0.66	0.34	0.00	0.93	0.07	0.04	0.96	0.00
100	100	5	0.00	0.70	0.30	0.00	0.95	0.05	0.00	1.00	0.00
200	200	5	0.00	0.71	0.29	0.00	0.95	0.05	0.00	1.00	0.00
50	50	25	0.00	0.01	0.98	0.01	0.28	0.71	0.07	0.80	0.13
100	100	50	0.00	0.00	1.00	0.00	0.13	0.87	0.00	0.94	0.06
200	200	100	0.00	0.00	1.00	0.00	0.06	0.94	0.00	0.99	0.01
50	50	50	0.01	0.00	0.99	0.04	0.01	0.95	0.15	0.34	0.52
100	100	100	0.00	0.00	1.00	0.00	0.00	1.00	0.01	0.53	0.46
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.74	0.26

$p_* = 3$, $\alpha = 2$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	5	0.00	0.27	0.73	0.00	0.85	0.15	0.00	1.00	0.00
100	100	5	0.00	0.67	0.33	0.00	0.95	0.05	0.00	1.00	0.00
200	200	5	0.00	0.72	0.28	0.00	0.97	0.04	0.00	1.00	0.00
50	50	25	0.00	0.00	1.00	0.00	0.27	0.72	0.01	0.86	0.13
100	100	50	0.00	0.00	1.00	0.00	0.16	0.85	0.00	0.95	0.05
200	200	100	0.00	0.00	1.00	0.00	0.05	0.96	0.00	0.99	0.01
50	50	50	0.00	0.00	1.00	0.00	0.01	0.98	0.03	0.37	0.59
100	100	100	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.53	0.47
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.75	0.25

$p_* = 3$, $\alpha = 3$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	5	0.00	0.70	0.31	0.00	0.92	0.08	0.00	0.99	0.01
100	100	5	0.00	0.71	0.29	0.00	0.95	0.05	0.00	1.00	0.00
200	200	5	0.00	0.70	0.30	0.00	0.97	0.03	0.00	1.00	0.00
50	50	25	0.00	0.01	0.99	0.00	0.26	0.74	0.01	0.87	0.12
100	100	50	0.00	0.00	1.00	0.00	0.13	0.87	0.00	0.94	0.06
200	200	100	0.00	0.00	1.00	0.00	0.04	0.96	0.00	0.99	0.01
50	50	50	0.00	0.00	1.00	0.00	0.01	0.99	0.03	0.35	0.62
100	100	100	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.51	0.49
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.73	0.27

Table 4.2. Selection rates of TC_2, $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ for $p_* = 6$

$p_* = 6$, $\alpha = 1$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	25	0.08	0.01	0.92	0.30	0.24	0.46	0.86	0.12	0.02
100	100	50	0.00	0.00	1.00	0.01	0.16	0.83	0.30	0.66	0.04
200	200	100	0.00	0.00	1.00	0.00	0.06	0.94	0.01	0.98	0.01
50	50	50	0.19	0.00	0.81	0.49	0.00	0.51	0.90	0.03	0.07
100	100	100	0.00	0.00	1.00	0.05	0.00	0.95	0.48	0.28	0.25
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.03	0.73	0.24

$p_* = 6$, $\alpha = 2$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	25	0.06	0.01	0.93	0.23	0.24	0.53	0.75	0.21	0.04
100	100	50	0.00	0.00	1.00	0.01	0.16	0.83	0.14	0.81	0.05
200	200	100	0.00	0.00	1.00	0.00	0.06	0.94	0.00	0.99	0.01
50	50	50	0.12	0.00	0.88	0.36	0.01	0.63	0.82	0.07	0.11
100	100	100	0.00	0.00	1.00	0.02	0.00	0.98	0.33	0.32	0.35
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.02	0.74	0.24

$p_* = 6$, $\alpha = 3$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	25	0.03	0.01	0.96	0.17	0.24	0.59	0.71	0.25	0.05
100	100	50	0.00	0.00	1.00	0.00	0.17	0.83	0.13	0.83	0.04
200	200	100	0.00	0.00	1.00	0.00	0.07	0.93	0.00	0.99	0.01
50	50	50	0.13	0.00	0.87	0.35	0.01	0.65	0.82	0.05	0.13
100	100	100	0.00	0.00	1.00	0.03	0.00	0.97	0.30	0.37	0.33
200	200	200	0.00	0.00	1.00	0.00	0.00	1.00	0.01	0.75	0.24

Table 4.3. Selection rates of TC_2, $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ for $p_* = p/4$

$p = 100$, $p_* = 25$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	α	Under	True	Over	Under	True	Over	Under	True	Over
100	100	1	0.99	0.00	0.01	1.00	0.00	0.00	1.00	0.00	0.00
200	200	1	0.28	0.00	0.72	0.94	0.01	0.05	1.00	0.00	0.00
500	500	1	0.00	0.00	1.00	0.01	0.36	0.63	1.00	0.00	0.00
1000	1000	1	0.00	0.00	1.00	0.00	0.58	0.42	0.43	0.57	0.00
2000	2000	1	0.00	0.00	1.00	0.00	0.73	0.27	0.00	1.00	0.00

$p = 200$, $p_* = 50$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	α	Under	True	Over	Under	True	Over	Under	True	Over
200	200	1	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
500	500	1	0.19	0.00	0.81	0.96	0.01	0.03	1.00	0.00	0.00
1000	1000	1	0.00	0.00	1.00	0.02	0.17	0.81	1.00	0.00	0.00
2000	2000	1	0.00	0.00	1.00	0.00	0.45	0.55	1.00	0.00	0.00
5000	5000	1	0.00	0.00	1.00	0.00	0.70	0.30	0.00	1.00	0.00
10000	10000	1	0.00	0.00	1.00	0.00	0.75	0.25	0.00	1.00	0.00

5. Ridge-type criteria

When $p > n - 2$, $\mathbf{S}$ becomes singular, and so we cannot use the TC_d. One way to overcome this problem is to use the ridge-type estimator of $\boldsymbol{\Sigma}$ defined by

$$\hat{\boldsymbol{\Sigma}}_\lambda = \frac{1}{n}\{(n-2)\mathbf{S} + \lambda\mathbf{I}_p\}, \tag{5.1}$$

where $\lambda = (n-2)(np)^{-1}\mathrm{tr}\mathbf{S}$. The estimator was used in multivariate regression model, by Kubokawa and Srivastava (2012), Fujikoshi and Sakurai (2016), etc.

The numerical experiment was done for $j_* = 3$, $\boldsymbol{\mu}_1 = \alpha(1,1,1,0,\ldots,0)$, $\boldsymbol{\mu}_2 = \alpha(-1,-1,-1,0,\ldots,0)$, $\boldsymbol{\Sigma} = \mathbf{I}_p$. We focused on selection rates for 10^2 replications in Tables 5.1. From Table 5.1, we can identify the following tendencies.

- TC_2 has not consistency. On the other hand, it seems that $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ have consistency when the dimension p and the total sample size n are separated.

Table 5.1. Selection rates of TC_2, $\mathrm{TC}_{\log n}$ and $\mathrm{TC}_{\sqrt{n}}$ for $p_* = p/4$

$\alpha = 1$			TC_2			$\mathrm{TC}_{\log n}$			$\mathrm{TC}_{\sqrt{n}}$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
15	15	40	0.31	0.00	0.69	0.48	0.00	0.52	0.73	0.04	0.23
25	25	60	0.15	0.00	0.85	0.30	0.00	0.70	0.52	0.00	0.48
50	50	110	0.07	0.00	0.93	0.14	0.00	0.87	0.35	0.00	0.65
15	15	90	0.14	0.29	0.57	0.48	0.48	0.04	0.91	0.09	0.00
25	25	150	0.01	0.13	0.86	0.10	0.83	0.07	0.60	0.40	0.00
50	50	300	0.00	0.02	0.98	0.00	0.95	0.05	0.08	0.93	0.00

6. Concluding Remarks

In this paper we propose a test-based criterion (TC) for the variable selection problem, based on drawing on the significance of each variable. The criterion invoves a constant term d, and is denoted by TC_d. When $d = 2$ and $d = \log n$, the corresponding TC's are related to AIC and BIC, respectively. However, the usual model selection criteria such as AIC and BIC need to examine all the subsets. However, TC_d need not to examine all the subsets, but need to examine only the p subsets $(-i),\ i = 1, \ldots, p$. This circumvents computational complexities associated with AIC and BIC has been resolved. Further, it was identified that TC_d has a high-dimensional consistency property for some d including $d = \sqrt{n}$, when (i) p_* is finite and $\Delta^2 = \mathrm{O}(1)$, and (ii) p_* is infinite and $\Delta^2 = \mathrm{O}(p)$, The problem of determining an optimum d is left as a future work. Further, an extension to nonnormality is left.

Appendix: Proofs of Theorems 3.1, 3.2 and 3.3

A1. Preliminary Lemmas

First we study distributional results related to the test statistics $\mathrm{T}_{d,i}$ in (2.5). For a notational simplicity, consider a decomposition of $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)'$, $\boldsymbol{y}$; $p_1 \times 1$, $\boldsymbol{p}_2$; $p_2 \times 1$. Similarly, decompose $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, and

$$\mathsf{S} = \begin{pmatrix} \mathsf{S}_{11} & \mathsf{S}_{12} \\ \mathsf{S}_{21} & \mathsf{S}_{22} \end{pmatrix}, \quad \mathsf{S};\ p_1 \times p_2$$

Let λ be the likelihood ratio criterion for testing a hypothesis $\boldsymbol{\beta}_2 = \mathbf{0}$, then

$$-2 \log \lambda = n \log \left\{ 1 + \frac{g^2(D^2 - D_1^2)}{n - 2 + g^2 D_1^2} \right\} \tag{A.1}$$

The following lemma (see, e.g., Fujikoshi et al. 2010) is used.

Lemma A1. *Let D_1 and D be the sample Mahalanobis distances based on $\boldsymbol{y}_1$ and $\boldsymbol{y}$, respectively. Let $D_{2\cdot 1}^2 = D^2 - D_1^2$. Similarly, the corresponding population quantities are expressed as Δ_1, Δ and $\Delta_{2\cdot 1}^2$. Then, it holds that*

(1) $D_1^2 = (n-2)g^{-2}R, \quad R = \chi_{p_1}^2(g^2\Delta_1^2)\left\{\chi_{n-p_1-1}^2\right\}^{-1}.$

(2) $D_{2\cdot 1}^2 = (n-2)g^{-2}\chi_{p_2}^2\left(g^2\Delta_{2\cdot 1}^2 \cdot \dfrac{1}{1+R}\right)\left\{\chi_{n-p-1}^2\right\}^{-1}(1+R).$

(3) $\dfrac{g^2(D^2 - D_1^2)}{n - 2 + g^2 D_1^2} = \chi_{p_2}^2(g^2\Delta_{2\cdot 1}^2(1+R)^{-1})\{\chi_{n-p-1}^2\}^{-1}$

Here, $\chi_{p_1}^2(\cdot)$, $\chi_{n-p_1-1}^2$, $\chi_{p_2}^2(\cdot)$, and χ_{n-p-1}^2 are independent chi-square variates.

Related to the conditional distribution of the righthand side of (3) with $p_2 = 1$ and $m = n-p-1$ in Lemma A1, consider the random variable defined by

$$V = \frac{\chi_1^2(\lambda^2)}{\chi_m^2} - \frac{1+\lambda^2}{m-2}, \tag{A.2}$$

where $\chi_1^2(\lambda^2)$ and χ_m^2 are independent. We can express V as

$$V = U_1U_2 + (m-2)^{-1}U_1 + (1+\lambda^2)U_2, \tag{A.3}$$

in terms of the centralized variables U_1 and U_2 defined by

$$U_1 = \chi_1^2(1+\lambda^2) - (1+\lambda^2), \quad U_2 = \frac{1}{\chi_m^2} - \frac{1}{m-2}. \tag{A.4}$$

It is well known (see, e.g., Tiku 1985) that

$$\begin{aligned} \mathrm{E}(U_1) &= 0, \\ \mathrm{E}(U_1^2) &= 2(1+2\lambda^2), \\ \mathrm{E}(U_1^3) &= 8(1+3\lambda^2), \\ \mathrm{E}(U_1^4) &= 48(1+4\lambda^2) + 4(1+3\lambda^2)^2. \end{aligned}$$

Further,

$$\begin{aligned} \mathrm{E}\left(U_2^k\right) &= \sum_{i=0}^{k} {}_kC_i \mathrm{E}\left\{\left(\frac{1}{\chi_m^2}\right)^i\right\}\left(-\frac{1}{m-2}\right)^{k-i} \\ &= \sum_{i=1}^{k} {}_kC_i \frac{1}{(m-2)\cdots(m-2i)}\left(-\frac{1}{m-2}\right)^{k-i} + \left(-\frac{1}{m-2}\right)^{k}. \end{aligned}$$

These give the first four moments of V. In particular, we use the following results.

Lemma A2. *Let V be the random variable defined by* (A.3). *Suppose that* $\lambda^2 = \mathrm{O}(m)$. *Then*

$$\begin{aligned} &\mathrm{E}(V) = 0, \quad \mathrm{E}(V^2) = \frac{2(m-3-2\lambda^2+\lambda^4)}{(m-2)^2(m-4)} = \mathrm{O}(m^{-1}), \\ &\mathrm{E}(V^3) = \mathrm{O}(m^{-2}), \quad \mathrm{E}(V^4) = \mathrm{O}(m^{-2}). \end{aligned}$$

A2. Proof of Theorems 3.1

First we show "[F1] → 0". Let $i \in j_*$. Then, $(-i) \notin \mathcal{F}_+$, and hence

$$\Delta_{(-i)}^2 < \Delta^2, \quad \Delta_{\{i\}\cdot(-i)}^2 > 0.$$

Using (A.1) and Lemma A1 (3)

$$\mathrm{T}_{d,i} = n\log\left\{1 + \frac{\chi_1^2(g^2\Delta_{\{i\}\cdot(-i)}^2(1+R)^{-1})}{\chi_{n-p-1}^2}\right\} - d,$$

where $R = \chi^2_{p-1}(g^2\Delta^2_{(-i)})\left\{\chi^2_{n-p}\right\}^{-1}$. Here, since j_* is finite, by showing

$$\mathrm{T}_{d,i} \xrightarrow{p} t > 0 \quad \text{or} \quad \mathrm{T}_{d,i} \xrightarrow{p} \infty,$$

we obtain $P(\mathrm{T}_{d,i} \leq 0) \to 0$, and hence "[F1] $\to$ 0". It is easily seen that

$$R \to \frac{c + k_1k_2\Delta^2_{(-i)}}{1-c},$$

and hence

$$(1+R)^{-1} \to \frac{1-c}{1+k_1k_2\Delta^2_{(-i)}}.$$

Therefore, we obtain

$$\frac{1}{n}\mathrm{T}_{d,i} \to \log\left(1 + \frac{k_1k_2\Delta^2_{\{i\}\cdot(-i)}}{1+k_1k_2\Delta^2_{\{i\}\cdot(-i)}}\right) > 0,$$

which implies our assertion.

Next, consider to show "[F2] $\to$ 0". For any $i \notin j_*$, $\Delta^2 = \Delta^2_{(-i)}$. Therefore, using Lemma A1(3) we have

$$\mathrm{T}_{d,i} = n\log\left(1 + \frac{\chi^2_1}{\chi^2_{n-p-1}}\right) - d, \tag{A.5}$$

whose distribution does not depend on i. Here, χ^2_1 and χ^2_{n-p-1} are independent chi-square variates with 1 and $n-p-1$ degrees of freedom. This implies that

$$\mathrm{T}_{d,i} > 0 \Leftrightarrow \frac{\chi^2_1}{\chi^2_{n-p-1}} > e^{d/n} - 1.$$

Noting that $\mathrm{E}[\chi^2_1/\chi^2_{n-p-1}] = (n-p-3)^{-1}$, let

$$U = \frac{\chi^2_1}{\chi^2_{n-p-1}} - \frac{1}{n-p-3}.$$

Then, since $e^{d/n} - 1 - \frac{1}{n-p-3} > h$,

$$\begin{aligned} P(\mathrm{T}_{d,i} > 0) &= P\left(U > e^{d/n} - 1 - \frac{1}{n-p-3}\right) \\ &\leq P\left(U > h\right). \end{aligned}$$

Further, using Markov inequality, we have

$$\begin{aligned} P(\mathrm{T}_{d,i} > 0) &\le P(|U| > h) \\ &\le h^{-2\ell}\mathrm{E}(U^{2\ell}), \quad \ell = 1, 2, \ldots \end{aligned}$$

Further, it is easily seen that

$$\mathrm{E}(U^{2\ell}) = \mathrm{O}(n^{-2\ell}),$$

by using e.g., Theorem 16.2.2 in Fujikoshi et al. (2010), When $h = O(n^{-a})$,

$$h^{-2\ell}\mathrm{E}(U^{2\ell}) = \mathrm{O}(n^{-2(1-a)\ell}).$$

Choosing ℓ such that $\ell > (1-a)^{-1}$, we have "[F2] $\to$ 0".

A3. Proof of Theorem 3.2

First, note that in the proof of "[F2] $\to$ 0" in Theorem 3.1, assumption A3 is not used. This implies the assertion in Theorem 3.2.

Now we consider to show "[F1] $\to$ 0" when $p_* = \mathrm{O}(p)$ and $\Delta^2 = \mathrm{O}(p)$. In this case, b tends to ∞. Based on the proof in Theorem 3.1, we can express $\mathrm{T}_{d,i}$ for $i \in \{1, \ldots, b\}$ as

$$\mathrm{T}_{d,i} = n \log \left\{ 1 + \frac{\chi_1^2(\widehat{\lambda}_i^2)}{\chi_{n-p-1}^2} \right\} - d,$$

where $\widehat{\lambda}_i^2 = g^2\Delta^2_{\{i\}\cdot(-i)}(1 + R_i)^{-1}$ and $R_i = \chi^2_{p-1}(g^2\Delta^2_{(-i)}) \left\{\chi^2_{n-p}\right\}^{-1}$. Note that χ_1^2 and χ^2_{n-p-1} are independent of R_i, and hence $\widehat{\lambda}_i^2$. Then, we have

$$P(T_{d.i} \le 0) = P(\widehat{V} \le \widehat{h}) \tag{A.6}$$

where

$$\begin{aligned} \widehat{V} &= \frac{\chi_1^2(\widehat{\lambda}_i^2)}{\chi_m^2} - \frac{1 + \widehat{\lambda}_i^2}{m-2}, \\ \widehat{h} &= e^{d/n} - 1 - (1 + \widehat{\lambda}_i^2)/(m-2), \end{aligned}$$

and $m = n-p-1$. Considering the conditional distribution of the right-hand side in (A.6), we have

$$P(\widehat{V} \leq \widehat{h}) = \mathrm{E}_{\widehat{\lambda}_i^2}\left\{Q(\widehat{\lambda}_i^2)\right\}, \tag{A.7}$$

where

$$Q(\lambda_i^2) = P(V \leq h).$$

Here,

$$\begin{aligned} V &= \frac{\chi_1^2(\lambda_i^2)}{\chi_m^2} - \frac{1+\lambda_i^2}{m-2}, \\ h &= e^{d/n} - 1 - (1+\lambda_i^2)/(m-2). \end{aligned}$$

From assumption A5, let

$$\Delta_{(-i)}^2 = p\Gamma_{(-i)}^2. \quad \Delta_{\{i\}\cdot(-i)}^2 = p^b\Gamma_{\{i\}\cdot(-i)}^2.$$

Then, $\Gamma_{(-i)}^2 = \mathrm{O}(1)$ and $\Gamma_{\{i\}\cdot(-i)}^2 = \mathrm{O}(1)$. We can easily see that

$$\widehat{\lambda}_i^2 \sim p^b\theta_i^2 - \mathrm{O}(p^b), \quad \theta_i^2 = \{(1-c)\Gamma_{\{i\}\cdot(-i)}^2\}/\{c\Gamma_{(-i)}^2\}.$$

Now we consider the probability $P(V \leq h)$ when $\lambda_i^2 = \mathrm{O}(p^b)$. From assumptions $a < b$, for large n, $h < 0$. In that case

$$\begin{aligned} p(V \leq h) &\leq P(|V| \geq |h|) \\ & h^{-4}\mathrm{E}(V^4), \end{aligned}$$

whose order is $\mathrm{O}(n^{-(1+\epsilon)})$ with $\epsilon > 0$. Noting that $\widehat{\lambda}_i^2 = \widehat{\lambda}_i^2 + \mathrm{O}(n^{-1/2})$. we have $P(\mathrm{T}_{d,i} \leq 0) = \mathrm{O}(n^{-(1+\epsilon)})$, which implies "[F1] → 0".

A4. Proof of Theorem 3.3

The assortion "[F1] → 0" follows from the proof of "[F1] → 0" in Theorem 3.1. For a proof of "[F2] → 0", it is enough to show that

$$\text{for } i \notin j_*, \quad \mathrm{T}_{d,i} \to \ -\infty.$$

since p has been fixed. From (A.5), the limiting distribution of $\mathrm{T}_{d,i}$ is "$\chi_1^2 - d$". This implies "[F2] → 0".

Acknowledgements

The first author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), 16K00047, 2016-2018.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd. International Symposium on Information Theory* (eds. B. N. Petrov and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Clemmensen, L., Hastie, T., Witten, D. M. and Ersbell, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

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

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

[5] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **144**, 184–200.

[6] Fujikoshi, Y. and Sakurai, T. (2016). High-dimensional consistency of rank estimation criteria in multivariate linear model. *J. Multivariate Anal.*, **149**, 199–212.

[7] Hyodo, M. and Kubokawa, T. (2014). A variable selection criterion for linear discriminant rule and its optimality in high dimensional and large sample data. *J. Multivariate Anal.*, **123**, 364–379.

[8] Kubokawa, T. and Srivastava, M. S. (2012). Selection of variables in multivariate regression models for large dimensions. *Communication in Statistics-Theory and Methods*, **41**, 2465–2489.

[9] McLachlan, G. J. (1976). A criterion for selecting variables for the linear discriminant function. *Biometrics*, bf 32, 529-534.

[10] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.

[11] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New-York.

[12] Sakurai, T., Nakada, T. and Fujikoshi, Y. (2013). High-dimensional AICs for selection of variables in discriminant analysis. *Sankhya, Ser. A*, **75**, 1–25.

[13] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.*, **9**, 869–897. Witten, D. W. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc. B*, **73**, 753–772.