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 $y_1^{(i)}, \ldots, y_{n_i}^{(i)}$ from $p$-dimensional normal distributions $N_p(\mu^{(i)}, \Sigma)$, $i = 1, 2$. Let $Y$ be the total sample matrix defined by

$$Y = (y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)})'.$$

The coefficients of the population discriminant function are given by

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

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

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

respectively. Here $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ are the sample mean vectors, and $S$ 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 $y_j$ denotes the $p_j$ vector consisting of the elements of $y$ indexed by the elements of $j$. We use the notation $D_j$ and $D_\omega$ for $D$ based on $y_j$ and $y_\omega (= y)$, 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 $y_j$. Let $AIC_j$ be the AIC for $M_j$. Then, it is known (see, e.g., Fujikoshi 1985) that

$$A_j = AIC_j - AIC_\omega$$

$$= n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_j^2)}{n - 2 + g^2D_j^2} \right\} - 2(p - p_j),$$

where $g = \sqrt{(n_1n_2)/n}$. Similarly, let $BIC_j$ be the BIC for $M_j$, and we have

$$B_j = BIC_j - BIC_\omega$$

$$= n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_j^2)}{n - 2 + g^2D_j^2} \right\} - (\log n)(p - p_j).$$

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 $\min_j AIC_j$ and $\min_j BIC_j$, 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

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

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

$$T_{2,i} > 0 \iff AIC_{\omega(-i)} - AIC_\omega > 0.$$

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

$$TC_d = \{ i \in \omega \mid T_{d,i} > 0 \},$$

or $\{ y_i \in \{y_1, \ldots, y_p\} \mid T_{d,i} > 0 \}$. The notation $\hat{TC}_d$ is also used for $TC_d$.4
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, $TC_{\log n}$ is consistent. However, we note that $TC_{\log n}$ 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 $\omega$ is called overspecified model if $j$ include the true model. On the other hand, a subset $j$ of $\omega$ 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., $\mathcal{F}_- = \mathcal{F}_+ \cap \mathcal{F}$. It is said that a model selection criterion $\hat{j}$ has a high-dimensional consistency if

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

Here we list some of our main assumptions:

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

5
A2 (The high-dimensional asymptotic framework):
\[ p \to \infty, \quad n \to \infty, \quad p/n \to c \in (0, \infty), \quad n_i/n \to k_i > 0, \quad (i = 1, 2). \]

For the dimensionality \( p_* \) of the true model and the Mahalanobis distance \( \Delta \), the following two cases are considered:

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

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

B1 : \( d/n \to 0 \).

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

\[ TC_d = j_* \iff "T_{d,i} > 0 \text{ for } i \in j_*", \quad \text{and } "T_{d,i} \leq 0 \text{ for } i \notin j_*" \]

Therefore

\[
P(TC_d = j_*) = P \left( \bigcap_{i \in j_*} "T_{d,i} > 0" \bigcap_{i \notin j_*} "T_{d,i} < 0" \right)
= 1 - P \left( \bigcup_{i \in j_*} "T_{d,i} \leq 0" \bigcup_{i \notin j_*} "T_{d,i} \geq 0" \right)
\geq 1 - \sum_{i \in j_*} P(T_{d,i} \leq 0) - \sum_{i \notin j_*} P(T_{d,i} \geq 0).
\]

We shall consider to show

\[ [F1] \equiv \sum_{i \in j_*)} P(T_{d,i} \leq 0) \to 0. \quad (3.1) \]

\[ [F2] \equiv \sum_{i \notin j_*)} P(T_{d,i} \geq 0) \to 0. \quad (3.2) \]

[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, $h = O(n^{-1-r})$, and the condition $B2$ is satisfied. So, the test-based criteria with

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

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 $TC_{\log n}$ 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

$$\sum_{i\notin J^*} P(TC_d = i) = \sum_{i\notin J^*} P(T_{d,i} \geq 0) \to 0.$$ 

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

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

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 $y$ is decomposed as a sum of the squares of Mahalanobis distance of $y_{(-i)}$ and the conditional Mahalanobis distance of $y_{(i)}$ given $y_{(-i)}$ as follows:

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

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

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

Related to consistency of $TC_d$ under assumption A4, we consider the following assumption:
A5: For \( i \in j_s \),
\[
\Delta^2_{(-i)} = O(p), \Delta^2_{[-i),(-i]} = O(p^b), \quad 0 < b < 1.
\]

**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
"\( p_i \) 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 \( TC_{\log n} \) and \( TC_{\sqrt{n}} \) 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, \( TC_2 \), \( TC_{\log n} \), and \( TC_{\sqrt{n}} \). Note that \( TC_{\sqrt{n}} \) satisfies sufficient conditions B1 and B2 for its consistency, but \( TC_2 \) and \( TC_{\log n} \) do not satisfy them.

The true model was assumed as follows: the true dimension is \( j_s = 3, 6 \), the true mean vectors;
\[
\mu_1 = \alpha(1, \ldots, 1, 0, \ldots, 0)', \quad \mu_2 = \alpha(-1, \ldots, -1, 0, \ldots, 0)'.
\]
and the true covariance matrix $\Sigma_0 = I_p$.

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 ~ 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 $\text{TC}_2$ 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 $\text{TC}_2$ has no consistency under a large sample case.

- The selection probabilities of the true model by $\text{TC}_{\log n}$ 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 $\text{TC}_{\sqrt{n}}$ 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 $\alpha$ presenting a distance between two groups becomes large, the selection probabilities of the true model by $\text{TC}_2$ and $\text{TC}_{\log n}$ 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 $\text{TC}_{\log n}$ 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 TC\(_{\log n}\) are not consistent as \(p\) increases, but 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 identified.

- When \(p_* = p/4\), the consistency of TC\(_2\) and TC\(_{\log n}\) can be not seen. The consistency of TC\(_{\sqrt{n}}\) can be seen when \(n\) is large.

<table>
<thead>
<tr>
<th>(p_* = 3), (\alpha = 1)</th>
<th>TC(_2)</th>
<th>TC(_{\log n})</th>
<th>TC(_{\sqrt{n}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n_1) (n_2) (p)</td>
<td>Under</td>
<td>True</td>
<td>Over</td>
</tr>
<tr>
<td>50 50 5</td>
<td>0.00</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>100 100 5</td>
<td>0.00</td>
<td>0.70</td>
<td>0.30</td>
</tr>
<tr>
<td>200 200 5</td>
<td>0.00</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>50 50 25</td>
<td>0.00</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>100 100 50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200 200 100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50 50 50</td>
<td>0.01</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>100 100 100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200 200 200</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(p_* = 3), (\alpha = 2)</th>
<th>TC(_2)</th>
<th>TC(_{\log n})</th>
<th>TC(_{\sqrt{n}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n_1) (n_2) (p)</td>
<td>Under</td>
<td>True</td>
<td>Over</td>
</tr>
<tr>
<td>50 50 5</td>
<td>0.00</td>
<td>0.27</td>
<td>0.73</td>
</tr>
<tr>
<td>100 100 5</td>
<td>0.00</td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>200 200 5</td>
<td>0.00</td>
<td>0.72</td>
<td>0.28</td>
</tr>
<tr>
<td>50 50 25</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>100 100 50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200 200 100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50 50 50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>100 100 100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200 200 200</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$p$</td>
<td>$TC_2$</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Under</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>25</td>
<td>0.00</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0.00</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>200</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4.2. Selection rates of $TC_2$, $TC_{log\ n}$ and $TC_{\sqrt{n}}$ for $p = 6$

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$p$</th>
<th>$TC_2$</th>
<th>$TC_{log\ n}$</th>
<th>$TC_{\sqrt{n}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Under</td>
<td>True</td>
<td>Over</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>25</td>
<td>0.08</td>
<td>0.01</td>
<td>0.92</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0.19</td>
<td>0.00</td>
<td>0.81</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>200</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$p$</th>
<th>$TC_2$</th>
<th>$TC_{log\ n}$</th>
<th>$TC_{\sqrt{n}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Under</td>
<td>True</td>
<td>Over</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>25</td>
<td>0.06</td>
<td>0.01</td>
<td>0.93</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0.12</td>
<td>0.00</td>
<td>0.88</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>200</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$p$</th>
<th>$TC_2$</th>
<th>$TC_{log\ n}$</th>
<th>$TC_{\sqrt{n}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Under</td>
<td>True</td>
<td>Over</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>25</td>
<td>0.03</td>
<td>0.01</td>
<td>0.96</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0.13</td>
<td>0.00</td>
<td>0.87</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>200</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 4.3. Selection rates of $TC_2$, $TC_{\log n}$ and $TC_{\sqrt{n}}$ for $p_*=p/4$

<table>
<thead>
<tr>
<th>$p = 100$, $p_*=25$</th>
<th>$TC_2$</th>
<th>$TC_{\log n}$</th>
<th>$TC_{\sqrt{n}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$\alpha$</td>
<td>Under</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>1</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p = 200$, $p_*=50$</th>
<th>$TC_2$</th>
<th>$TC_{\log n}$</th>
<th>$TC_{\sqrt{n}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$\alpha$</td>
<td>Under</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>5000</td>
<td>5000</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>10000</td>
<td>10000</td>
<td>1</td>
<td>0.00</td>
</tr>
</tbody>
</table>

5. Ridge-type criteria

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

$$\hat{\Sigma}_\lambda = \frac{1}{n} \{(n-2)S + \lambda I_p\},$$

(5.1)

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

The numerical experiment was done for $j_*=3$, $\mu_1 = \alpha(1,1,1,0,\ldots,0)$, $\mu_2 = \alpha(-1,-1,-1,0,\ldots,0)$, $\Sigma = I_p$. 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 $TC_{\log n}$ and $TC_{\sqrt{n}}$ have consistency when the dimension $p$ and the total sample size $n$ are separated.
Table 5.1. Selection rates of $TC_2$, $TC_{\log n}$ and $TC_{\sqrt{n}}$ for $p_\ast = p/4$

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$p$</th>
<th>Under</th>
<th>True</th>
<th>Over</th>
<th>Under</th>
<th>True</th>
<th>Over</th>
<th>Under</th>
<th>True</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>15</td>
<td>40</td>
<td>0.31</td>
<td>0.00</td>
<td>0.69</td>
<td>0.48</td>
<td>0.00</td>
<td>0.52</td>
<td>0.73</td>
<td>0.04</td>
<td>0.23</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>60</td>
<td>0.15</td>
<td>0.00</td>
<td>0.85</td>
<td>0.30</td>
<td>0.00</td>
<td>0.70</td>
<td>0.52</td>
<td>0.00</td>
<td>0.48</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>110</td>
<td>0.07</td>
<td>0.00</td>
<td>0.93</td>
<td>0.14</td>
<td>0.00</td>
<td>0.87</td>
<td>0.35</td>
<td>0.00</td>
<td>0.65</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>90</td>
<td>0.14</td>
<td>0.29</td>
<td>0.57</td>
<td>0.48</td>
<td>0.48</td>
<td>0.04</td>
<td>0.91</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>150</td>
<td>0.01</td>
<td>0.13</td>
<td>0.86</td>
<td>0.10</td>
<td>0.83</td>
<td>0.07</td>
<td>0.60</td>
<td>0.40</td>
<td>0.00</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>300</td>
<td>0.00</td>
<td>0.02</td>
<td>0.98</td>
<td>0.00</td>
<td>0.95</td>
<td>0.05</td>
<td>0.08</td>
<td>0.93</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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 involves 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) $p_\ast$ is finite and $\Delta^2 = O(1)$, and (ii) $p_\ast$ is infinite and $\Delta^2 = O(p)$. 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 $T_{d,i}$ in (2.5). For a notational simplicity, consider a decomposition of $y = (y_1, y_2)'$, $y; p_1 \times 1$, $p_2; p_2 \times 1$. Similarly, decompose $\beta = (\beta_1', \beta_2')'$, and

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

Let $\lambda$ be the likelihood ratio criterion for testing a hypothesis $\beta_2 = 0$, then

$$-2 \log \lambda = n \log \left(1 + \frac{g^2(D^2 - D_1^2)}{n - 2 + g^2D_1^2}\right)$$  \hspace{1cm} (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 $y_1$ and $y$, respectively. Let $D_{2,1}^2 = D^2 - D_1^2$. Similarly, the corresponding population quantities are expressed as $\Delta_1$, $\Delta$ and $\Delta_{2,1}^2$. Then, it holds that

1. $D_1^2 = (n - 2)g^{-2}R$, \hspace{1cm} $R = \chi_{p_1}^2(g^2\Delta_1^2)\left\{\chi_{n-p_1-1}^2\right\}^{-1}$.
2. $D_{2,1}^2 = (n - 2)g^{-2}\chi_{p_2}^2\left(g^2\Delta_{2,1}^2 \cdot \frac{1}{1 + R}\right)\left\{\chi_{n-p-1}^2\right\}^{-1}(1 + R)$.
3. $\frac{g^2(D^2 - D_1^2)}{n - 2 + g^2D_1^2} = \chi_{p_2}^2(g^2\Delta_{2,1}^2(1 + R)^{-1})\left\{\chi_{n-p-1}^2\right\}^{-1}$

Here, $\chi_{p_1}^2(\cdot)$, $\chi_{n-p_1-1}^2$, $\chi_{p_2}^2(\cdot)$, and $\chi_{n-p-1}^2$ 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

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

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

$$V = U_1 U_2 + (m - 2)^{-1}U_1 + (1 + \lambda^2)U_2,$$  \hspace{1cm} (A.3)
in terms of the centralized variables $U_1$ and $U_2$ defined by
\[ U_1 = \chi_1^2(1 + \lambda^2) - (1 + \lambda^2), \quad U_2 = \frac{1}{\chi_m^2} - \frac{1}{m-2}. \]  
(A.4)

It is well known (see, e.g., Tiku 1985) that
\[ E(U_1) = 0, \]
\[ E(U_1^2) = 2(1 + 2\lambda^2), \]
\[ E(U_1^3) = 8(1 + 3\lambda^2), \]
\[ E(U_1^4) = 48(1 + 4\lambda^2) + 4(1 + 3\lambda^2)^2. \]

Further,
\[ E(U_k^2) = \sum_{i=0}^{k} kC_i E\left(\left\{\frac{1}{\chi_m^2}\right\}^i\right) \left(-\frac{1}{m-2}\right)^{k-i} \]
\[ = \sum_{i=1}^{k} kC_i \frac{1}{(m-2)\cdots(m-2i)} \left(-\frac{1}{m-2}\right)^{k-i} + \left(-\frac{1}{m-2}\right)^k. \]

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 $\lambda^2 = O(m)$. Then
\[ E(V) = 0, \quad E(V^2) = \frac{2(m - 3 - 2\lambda^2 + \lambda^4)}{(m-2)^2(m-4)} = O(m^{-1}), \]
\[ E(V^3) = O(m^{-2}), \quad E(V^4) = O(m^{-2}). \]

**A2. Proof of Theorems 3.1**

First we show "[F1] $\rightarrow 0$". Let $i \in j_x$. Then, $(-i) \notin \mathcal{F}_+$, and hence
\[ \Delta_{i(-i)}^2 < \Delta^2, \quad \Delta_{i(+i)}^2 > 0. \]

Using (A.1) and Lemma A1 (3)
\[ T_{d,i} = n \log \left\{ 1 + \frac{\chi_i^2(g^2\Delta_{i(-i)}^2(1 + R)^{-1})}{\chi_{n-p-1}^2} \right\} - d, \]
15
where \( R = \chi_{p-1}^2(g^2 \Delta^2_{(-i)}) \{ \chi_{n-p}^2 \}^{-1} \). Here, since \( j_\ast \) is finite, by showing
\[
T_{d,i} \xrightarrow{p} t > 0 \quad \text{or} \quad T_{d,i} \xrightarrow{p} \infty,
\]
we obtain \( P(T_{d,i} \leq 0) \to 0 \), and hence ”[F1] \to 0”. It is easily seen that
\[
R \to \frac{c + k_1 k_2 \Delta^2_{(-i)}}{1 - c},
\]
and hence
\[
(1 + R)^{-1} \to \frac{1 - c}{1 + k_1 k_2 \Delta^2_{(-i)}}.
\]
Therefore, we obtain
\[
\frac{1}{n} T_{d,i} \to \log \left( 1 + \frac{k_1 k_2 \Delta^2_{(i)}(-i)}{1 + k_1 k_2 \Delta^2_{(i)}(-i)} \right) > 0,
\]
which implies our assertion.

Next, consider to show ”[F2] \to 0”. For any \( i \notin j_\ast \), \( \Delta^2 = \Delta^2_{(-i)} \). Therefore, using Lemma A1(3) we have
\[
T_{d,i} = n \log \left( 1 + \frac{\chi^2_{n-p-1}}{\chi_{n-p-1}^2} \right) - d,
\]
whose distribution does not depend on \( i \). Here, \( \chi^2_{1} \) and \( \chi^2_{n-p-1} \) are independent chi-square variates with 1 and \( n-p-1 \) degrees of freedom. This implies that
\[
T_{d,i} > 0 \iff \frac{\chi^2_{1}}{\chi^2_{n-p-1}} > e^{d/n} - 1.
\]
Noting that \( E[\chi^2_{1}/\chi^2_{n-p-1}] = (n-p-3)^{-1} \), let
\[
U = \frac{\chi^2_{1}}{\chi^2_{n-p-1}} - \frac{1}{n-p-3}.
\]
Then, since \( e^{d/n} - 1 - \frac{1}{n-p-3} > h \),
\[
P(T_{d,i} > 0) = P \left( U > e^{d/n} - 1 - \frac{1}{n-p-3} \right) \leq P (U > h).
\]
Further, using Markov inequality, we have
\[
P(T_{d,i} > 0) \leq P(|U| > h) \\
\leq h^{-2\ell}E(U^{2\ell}), \quad \ell = 1, 2, \ldots
\]

Further, it is easily seen that
\[
E(U^{2\ell}) = O(n^{-2\ell}),
\]
by using e.g., Theorem 16.2.2 in Fujikoshi et al. (2010). When \( h = O(n^{-a}) \),
\[
h^{-2\ell}E(U^{2\ell}) = O(n^{-2(1-a)\ell}).
\]
Choosing \( \ell \) such that \( \ell > (1 - a)^{-1} \), we have \([F2] \rightarrow 0\).

**A3. Proof of Theorem 3.2**

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

Now we consider to show \([F1] \rightarrow 0\) when \( p_* = O(p) \) and \( \Delta^2 = O(p) \). In this case, \( b \) tends to \( \infty \). Based on the proof in Theorem 3.1, we can express \( T_{d,i} \) for \( i \in \{1, \ldots, b\} \) as
\[
T_{d,i} = n \log \left\{ 1 + \frac{\chi_i^2(\tilde{\lambda}_i^2)}{\chi_{n-p-1}^2} \right\} - d,
\]
where
\[
\tilde{\lambda}_i^2 = g^2 \Delta_{i,-i}(1 + R_i)^{-1} \quad \text{and} \quad R_i = \chi_{p-1}^2(g^2 \Delta_{i,-i}^2) \left\{ \chi_{n-p}^2 \right\}^{-1}.
\]
Note that \( \chi_i^2 \) and \( \chi_{n-p-1}^2 \) are independent of \( R_i \), and hence \( \tilde{\lambda}_i^2 \). Then, we have
\[
P(T_{d,i} \leq 0) = P(\hat{V} \leq \hat{h}) \quad \text{(A.6)}
\]
where
\[
\hat{V} = \frac{\chi_i^2(\tilde{\lambda}_i^2)}{\chi_{m}^2} - \frac{1 + \tilde{\lambda}_i^2}{m - 2},
\]
\[
\hat{h} = e^{d/n} - 1 - (1 + \tilde{\lambda}_i^2)/(m - 2),
\]

17
and $m = n - p - 1$. Considering the conditional distribution of the right-hand side in (A.6), we have

$$P(\tilde{V} \leq \hat{h}) = E_{\lambda_i^2} \{Q(\lambda_i^2)\},$$

where

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

Here,

$$V = \frac{\chi_i^2(\lambda_i^2)}{\chi_m} - \frac{1 + \lambda_i^2}{m - 2},$$

$$h = e^{d/n} - 1 - \frac{(1 + \lambda_i^2)/(m - 2)}{}.$$

From assumption A5, let

$$\Delta^2_{(=-i)} = p\Gamma^2_{(=-i)}; \quad \Delta^2_{(i)-i} = p^b\Gamma^2_{(i)-i}.$$  

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

$$\hat{\lambda}_i^2 \sim p^b\theta_i^2 = O(p^b), \quad \theta_i^2 = \{(1 - c)\Gamma^2_{(i)-i})/\{c\Gamma^2_{(i)}\}.$$

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

$$p(V \leq h) \leq P(|V| \geq |h|)$$

$$h^{-4}E(V^4),$$

whose order is $O(n^{-(1+\epsilon)})$ with $\epsilon > 0$. Noting that $\hat{\lambda}_i^2 = \lambda_i^2 + O(n^{-1/2})$. we have $P(T_{d,i} \leq 0) = O(n^{-(1+\epsilon)})$, which implies ”$[F1] \rightarrow 0$”.

A4. Proof of Theorem 3.3

The assertion ”$[F1] \rightarrow 0$” follows from the proof of ”$[F1] \rightarrow 0$” in Theorem 3.1. For a proof of ”$[F2] \rightarrow 0$”, it is enough to show that

$$\text{for } i \notin j_s, \quad T_{d,i} \rightarrow -\infty.$$ 

since $p$ has been fixed. From (A.5), the limiting distribution of $T_{d,i}$ is ”$\chi_1^2 - d^\epsilon$. This implies ”$[F2] \rightarrow 0$”.

18
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


