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

Tetsuro Sakurai* and Yasunori Fujikoshi**

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

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

Abstract

This paper is concerned with selection of variables in two-group discriminant analysis with the same covariance matrix. We propose a distance-based criterion (DC) drawing on the distance of each variables. The selection method can be applied for high-dimensional data. Sufficient conditions for the distance-based criterion to be consistent are provided when the dimension and the sample are large. 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, Distance-based criterion.

1. Introduction

This paper is concerned with the variable selection problem in two-group discriminant analysis with the same covariance matrix. Some methods have been proposed for high-dimensional data as well as large-sample data. One of the methods are based on model selection criteria such as AIC and BIC, see, e.g., Fujikoshi (1984), Nishii et al. (1988). There are methods based on misclassification errors by MaLachlan (1976), Fujikoshi (1985), Hyodo and Kubokawa (2014), Yamada et al. (2017). For high-dimensional data, there are Lasso and other regularization methods by Clemmensen et al. (2011), Witten and Tibshirani (2011), etc.

In this paper we propose a distance-based criterion based on the distance of each variables, which is useful for high-dimensional data as well as large-sample data. Here, the distance of each variables is measured as the one except for the effect of other variables. The criterion involves a constant term which should be determined through some optimality. We propose a class of constants satisfying a consistency when the dimension and the sample size are large. The class depends on the parameters. With respect to the actual use, we also propose a class of estimators for the constant term which satisfies a consistency. 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 distance-based criterion. In Sections 3 we derive sufficient conditions for the distance-based criterion to be consistent under a high-dimensional case. In Section 4 we also propose a class of estimators for the constant term. The proposed distance-based criterion is also examined through a Monte Carlo simulation in Section 5. In Section 6, conclusions are offered. All proofs of our results are provided in the Appendix.

2. Distance-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

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

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. Let $\boldsymbol{\beta}$ be the coefficients of the population discriminant function given by

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

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.

One of the approaches for variable selection is to consider a set of models as follows. For each subset j of $\omega = \{1, \ldots, p\}$, consider a variable selection model M_j 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.

The variable selection methods based on AIC and BIC are given as $\min_j \mathrm{AIC}_j$ and $\min_j \mathrm{BIC}_j$, respectively. Therefor, such model selection methods have a computationally onerous problem when p is large, since the methods involve minimizing with respect to the subsets of 2^p. To circumvent this issue, we consider a distance-based criterion (DC) drawing on the distance of each variables. A contribution of y_i in the distance between two groups may be measured by

$$D^2 - D_{(-i)}^2 \tag{2.5}$$

where $(-i)$ is the subset of $\omega = \{1, \ldots, p\}$ obtained by omitting i from ω. If "$D^2 - D_{(-i)}^2$" is large, it may be considered that y_i has a large contribution in discriminant analysis. Conversely, if "$D^2 - D_{(-i)}^2$" is small, it may be considered that y_i has a small contribution in discriminant analysis. Let us define

$$\mathrm{D}_{d,i} = D^2 - D_{(-i)}^2 - d, \quad i = 1, \ldots, p, \tag{2.6}$$

where d is a positive constant. We propose a distance-based criterion for the selection of variables defined by selecting the set of suffixes or the set of variables given by

$$\mathrm{DC}_d = \{ i \in \omega \mid \mathrm{D}_{d,i} > 0 \}, \tag{2.7}$$

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

For a distance-based criterion, there is an important problem how to decide the constant term d. In Section 3 we derive a range of d which has a high-dimensional consistency. The range depends on the population parameters. For a practical usefulness, we also propose a class of estimators

for d given by

$$\widehat{d} = \frac{n-2+g^2D^2}{(n-p-1)g^2}\left\{a(np)^{1/3} + \frac{n-p-1}{n-p-3}\right\}, \tag{2.8}$$

where a is a constant satisfying $a = O(1)$. It is shown that $\mathrm{DC}_{\widehat{d}}$ is consistent under a high-dimensional framwork.

3. High-dimensional Consistency

For studying consistency properties of the variable selection criterion DC_d, it is assumed that the true model M_* is specified through the values of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ and Σ. For simplicity, we write the variable selection models in terms of $\boldsymbol{\beta}$, or M_j. Assume that the true model M_* is included in the full model M_ω. 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, defined by

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

Let $\mathcal{F}$ separate 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., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F}$.

Here we list some of our main assumptions:

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

A2 (The high-dimensional asymptotic framework):
$p \to \infty,\ n \to \infty,\ p/n \to c \in (0, \infty)$.

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

A4 : $n_i/n \to k_i > 0,\ i = 1, 2$.

Relating to our proof of a consistency of DC_d in (2.7), note that

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

Therefore,

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

Our result will be otained by proving the followings:

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

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

$[F1]$ is the probability that DC_d does not select the set of true variables. $[F2]$ is the probability that DC_d select the set of no true variables. Our consistency depends on

$$\tau_{\min} = \min_{i\in j_*} \tau_i, \tag{3.3}$$

where $\tau_i = \Delta^2 - \Delta^2_{(-i)},\ i \in \omega$.

Theorem 3.1. *Suppose that assumptions* A1, A2, A3 *and* A4 *are satisfied. Let* $b = \tau_{\min}/(1-c)$*, where* $\tau_{\min}$ *is defined by* (3.3)*. Then, if* $0 < d < b$*, the distance-based criterion* DC_d *satisfies* [F1], i.e.*,*

$$\sum_{i\in j_*} P(\mathrm{D}_{d,i} \leq 0) \to 0. \tag{3.4}$$

Related to a sufficient condition for (3.2), we use the following result: for $i \not\subset j_*$,

$$\begin{aligned} \mathrm{E}\left\{D^2 - D^2_{(-i)}\right\} &= \frac{n-2}{(n-p-3)(n-p-2)}\left\{g^{-2}(n-3) + \Delta^2\right\} \\ &\equiv q(p, n, \Delta^2). \end{aligned} \tag{3.5}$$

Theorem 3.2. *Suppose that assumptions* A1*,* A2*,* A3 *and* A4 *are satisfied. Then, if* $d = O(1)$ *and* $d > q(p, n, \Delta^2)$*, the distance-based criterion* DC_d *satisfies*

$$\sum_{i \not\in j_*} P(\mathrm{D}_{d,i} \geq 0) \to 0, \tag{3.6}$$

Combining Theorems 3.1 and 3.2, a sufficiency condition for DC_d to be consistent is given as follows:.

Theorem 3.3. *Suppose that assumptions* A1*,* A2*,* A3 *and* A4 *are satisfied. Then, if* $d = O(1)$ *and*

$$q(p, n, \Delta^2) < d < b = \tau_{\min}/(1-c), \tag{3.7}$$

the distance-based criterion DC_d *is consistent.*

For an actual use, it is necessary to get an estimator $\widehat{d}$ for d. This problem is discussed in Section 4.

4. A Method of Determining the Constant Term

In the previous section we have seen that the distance-based criterion DC_d is consistent if we use a constant term d such that $d = \mathrm{O}(1)$ and $q(p, n, \Delta) < d < \tau_{\min}$, where $q(p, n, \Delta)$ is given by (3.5), $\alpha_{\min} = \min_{i \in j_*} \alpha_i$ and $\alpha_i =$

$\Delta^2 - \Delta^2_{(-i)}$. However, such a d involves unknown parameters. We need to estimate d by using the estimators of Δ^2 and τ_{min}. Here, we consider to construct an estimator by using the fact that DC_d is based on the test of $\alpha_i = 0$ or $\beta_i = 0$. The likelihood test is based on (see, e.g., Rao 1973, Fujikoshi et al. 2010)

$$\frac{g^2 D^2_{\{i\}\cdot(-i)}}{n-2+g^2 D^2_{(-i)}}$$

whose null distribution is a ratio χ^2_1/χ^2_{n-p-1} of independent chi-squared variates χ^2_1 and χ^2_{n-p-1}, where $D^2_{\{i\}\cdot(-i)} = D^2 - D^2_{(-i)}$. Let us define $\hat{d}$ as follows:

$$\widehat{d} = \frac{n-2+g^2D^2}{(n-p-1)g^2}\left\{a(np)^{1/3} + \frac{n-p-1}{n-p-3}\right\}. \tag{4.1}$$

Here a is a costant sasisfying $a = O(1)$. Then, it is shown that $\mathrm{DC}_{\widehat{d}}$ has a consistency.

Theorem 4.1. *Suppose that assumptions* A1, A2, A3 *and* A4 *are satisfied. Then, the distance-based criterion* $\mathrm{DC}_{\widehat{d}}$ *with* $\widehat{d}$ *in* (4.1) *is consistent.*

As a special $\hat{d}$, we consider the followings:

$$\widehat{d}_1 = \frac{n-2+g^2D^2}{(n-p-1)g^2}\left\{(np)^{1/3} + \frac{n-p-1}{n-p-3}\right\}, \tag{4.2}$$

$$\widehat{d}_2 = \frac{n-2+g^2D^2}{(n-p-1)g^2}\left\{\left(1-\frac{p}{n}\right)^2(np)^{1/3} + \frac{n-p-1}{n-p-3}\right\}. \tag{4.3}$$

The estimators $\widehat{d}_1$ and $\widehat{d}_2$ are defined from $\widehat{d}$ by choosing a as $a = 1$ and $a = (1-p/n)^2$, respectively.

In Theorem 3.3, we have seen that DC_d is consistent under a high-dimensional framework if d satisfies $q(p,n,\Delta^2) < d < \tau_{\min}/(1-c)$. The expectation of $\widehat{d}$ is related to the range as follows:

$$\begin{aligned}
&(1)\ \mathrm{E}(\widehat{d}) > q(p,n,\Delta^2),\\
&(2)\ \lim_{p/n\to c}\mathrm{E}(\widehat{d}) = 0 \le \tau_{\min}/(1-c).
\end{aligned}$$

The above result follows from that

$$\begin{aligned}\mathrm{E}[\hat{d}] &= \left\{ a(np)^{1/3} + \frac{n-p-1}{n-p-3} \right\} \frac{n-2+g^2 \frac{p+g^2\Delta^2}{n-p-3}}{(n-p-1)g^2} \\ &= a(np)^{1/3} \left\{ \frac{(n-2)(n-3)}{g^2(n-p-3)(n-p-1)} + \frac{n-2}{(n-p-3)(n-p-1)} \Delta^2 \right\} \\ &\quad + \frac{(n-2)(n-3)}{g^2(n-p-3)^2} + \frac{n-2}{(n-p-3)^2} \Delta^2.\end{aligned}$$

5. Numerical Study

In this section we numerically explore the validity of our claims through three distance-based criteria, DC_{d_0}, $\mathrm{DC}_{\widehat{d}_1}$, and $\mathrm{DC}_{\widehat{d}_2}$. Here, d_0 is the midpoint in the interval $[q(p, n, \Delta^2),\ \tau_{\min}/(1-c)]$ in (3.7).

The true model was assumed as follows: the true dimension is $p_* = 4, 8$, the true mean vectors;

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

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

The selection rates associated with these criteria are given in Tables 5.1 to 5.4. "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 5.1 $\sim$ 5.4. From these tables, we can identify the following tendencies.

Table 5.1. Selection rates of DC_{d_0}, $\mathrm{DC}_{\widehat{d}_1}$ and $\mathrm{DC}_{\widehat{d}_2}$ for $p_* = 4$ and $\alpha = 1$

$p_* = 4$, $\alpha = 1$			d_0			$\widehat{d}_1$			$\widehat{d}_2$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	10	0.34	0.65	0.01	0.44	0.56	0.00	0.25	0.75	0.01
100	100	10	0.11	0.89	0.00	0.00	1.00	0.00	0.00	0.99	0.01
200	200	10	0.01	0.99	0.00	0.00	1.00	0.00	0.00	1.00	0.00
50	50	25	0.39	0.51	0.10	0.95	0.05	0.00	0.45	0.52	0.03
100	100	50	0.19	0.81	0.01	0.63	0.37	0.00	0.06	0.94	0.00
200	200	100	0.03	0.97	0.00	0.06	0.94	0.00	0.00	1.00	0.00
50	50	50	0.55	0.17	0.29	1.00	0.00	0.00	0.54	0.24	0.22
100	100	100	0.28	0.59	0.13	1.00	0.00	0.00	0.10	0.61	0.29
200	200	200	0.09	0.91	0.00	0.99	0.01	0.00	0.00	0.92	0.08

Table 5.2. Selection rates of DC_{d_0}, $\mathrm{DC}_{\widehat{d}_1}$ and $\mathrm{DC}_{\widehat{d}_2}$ for $p_* = 4$ and $\alpha = 10$

$p_* = 4$, $\alpha = 10$			d_0			$\widehat{d}_1$			$\widehat{d}_2$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	10	0.28	0.71	0.01	0.18	0.82	0.00	0.10	0.90	0.01
100	100	10	0.07	0.93	0.00	0.00	1.00	0.00	0.00	1.00	0.00
200	200	10	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
50	50	25	0.34	0.62	0.04	0.76	0.24	0.00	0.19	0.78	0.03
100	100	50	0.10	0.90	0.00	0.18	0.82	0.00	0.01	0.98	0.01
200	200	100	0.01	0.99	0.00	0.00	1.00	0.00	0.00	1.00	0.00
50	50	50	0.49	0.29	0.22	1.00	0.00	0.00	0.30	0.36	0.34
100	100	100	0.22	0.72	0.06	1.00	0.01	0.00	0.02	0.69	0.29
200	200	200	0.05	0.95	0.00	0.81	0.19	0.00	0.00	0.90	0.10

Table 5.3. Selection rates of DC_{d_0}, $\mathrm{DC}_{\widehat{d}_1}$ and $\mathrm{DC}_{\widehat{d}_2}$ for $p_* = 8$ and $\alpha = 1$

$p_* = 8$, $\alpha = 1$			d_0			$\widehat{d}_1$			$\widehat{d}_2$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	10	0.81	0.18	0.01	1.00	0.00	0.00	1.00	0.00	0.00
100	100	10	0.52	0.48	0.00	0.79	0.21	0.00	0.65	0.35	0.00
200	200	10	0.20	0.80	0.00	0.03	0.97	0.00	0.02	0.98	0.00
50	50	25	0.86	0.05	0.09	1.00	0.00	0.00	1.00	0.00	0.00
100	100	50	0.68	0.26	0.06	1.00	0.00	0.00	0.97	0.03	0.00
200	200	100	0.30	0.69	0.01	1.00	0.00	0.00	0.53	0.47	0.00
50	50	50	0.93	0.00	0.07	1.00	0.00	0.00	1.00	0.00	0.00
100	100	100	0.79	0.05	0.16	1.00	0.00	0.00	0.96	0.02	0.02
200	200	200	0.52	0.40	0.09	1.00	0.00	0.00	0.53	0.43	0.04

Table 5.4. Selection rates of DC_{d_0}, $\mathrm{DC}_{\widehat{d}_1}$ and $\mathrm{DC}_{\widehat{d}_2}$ for $p_* = 8$ and $\alpha = 10$

$p_* = 8$, $\alpha = 1$			d_0			$\widehat{d}_1$			$\hat{d}_2$		
n_1	n_2	p	Under	True	Over	Under	True	Over	Under	True	Over
50	50	10	0.79	0.19	0.02	1.00	0.00	0.00	0.99	0.01	0.00
100	100	10	0.48	0.52	0.00	0.61	0.40	0.00	0.48	0.52	0.00
200	200	10	0.14	0.86	0.00	0.00	1.00	0.00	0.00	1.00	0.00
50	50	25	0.84	0.08	0.08	1.00	0.00	0.00	1.00	0.00	0.00
100	100	50	0.60	0.33	0.06	1.00	0.00	0.00	0.91	0.09	0.00
200	200	100	0.25	0.75	0.00	1.00	0.00	0.00	0.27	0.73	0.00
50	50	50	0.91	0.00	0.09	1.00	0.00	0.00	0.99	0.00	0.01
100	100	100	0.74	0.08	0.18	1.00	0.00	0.00	0.89	0.06	0.04
200	200	200	0.45	0.49	0.06	1.00	0.00	0.00	0.31	0.63	0.06

From Tables 5.1 $\sim$ 5.4, we can identify the following tendncies.

- In all the DC criteria with d_0, $\widehat{d}_1$ and $\widehat{d}_1$, there are cases which their selection rates are high. Especially, when the dimension p is small as in $p = 10$, their selection rates are near to 1. However, when n and p are near, their selection rates do not near to one when n and p are not large. On the other hand, for the case which a consistency can not be seen in Tables, it will be expected to be consistent as the values of n and p are large.

- In all the DC criteria with d_0, $\widehat{d}_1$ and $\widehat{d}_1$, their selection rates of the true model are incresing as n and p are large.

- In all the DC criteria with d_0, $\widehat{d}_1$ and $\widehat{d}_1$, their selection rates of the true model are decreasing as n and p are closer.

- In all the DC criteria with d_0, $\widehat{d}_1$ and $\widehat{d}_1$, their selection rates of the true model are decreasing as the dimension p_* is increasing.

- In all the DC criteria with d_0, $\widehat{d}_1$ and $\widehat{d}_1$, their selection rates of the true model are increasing as the distance between groups or α is increasing.

- The DC criteria with d_0 has a tendency of not selecting overspecified models as p is small in comparison with n, and has a tendency of selecting overspcified models as p and n are near.

- The DC criteria with $\widehat{d}_1$ does not select overspesified models, and has a tendancy of choosing the variables strictly.
- The DC criteria with $\widehat{d}_2$ has a tendency of not selectiong overspecified models when p is small, and has a tendency of selecting overspesified models as n and p are near, as in the DC criteria with d_0.

6. Concluding Remarks

In this paper we propose a distance-based criterion (DC) for the variable selection problem, based on drawing on the significance of each the variables except for the effect due to other variables. The criterion involves a constant term d, and is denoted by DC_d. DC_d need not to examine all the subsets, but need to examine only the p subsets $\omega_{(-i)}$. The circumvent computational complexities associated with all the subsets selection procedures been resolved. Further, it was identified that a range of d and a class of its estimators such that DC_d has a high-dimensional consistency property. Especially the criteria DC_d with $d = d_0$, $d = \widehat{d}_1$ and $d = \widehat{d}_2$ were numerically examined. However, an optimality problem on $\widehat{d}$ is left for future research. A study of high-dimensional consistency properties when p_* is infinite and $\Delta^2 = \mathrm{O}(p)$ are also left.

Appendix: Proofs of Theorems 3.1 and 3.2

A1 Preliminary Lemmas

First we summarize the distributional results related to the Mahalanobis distance and the statistics $\mathrm{D}_{d,i}$ in (2.5). For a notational simplicity, consider a decomposition of $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2')', \ \boldsymbol{y};\ p_1 \times 1,\ \boldsymbol{p}_2;\ p_2 \times 1$. Let D_1 and D be the

sample Mahalanobis distances based on $\boldsymbol{y}_1$ and $\boldsymbol{y}$, respectively. Let $D^2_{2\cdot1} = D^2 - D^2_1$. Similarly, the corresponding population quantities are expressed as Δ_1, Δ and $\Delta^2_{2\cdot1}$. Let the coefficient vector $\boldsymbol{\beta}$ of the linear discriminant function in (2.1) decompose as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\boldsymbol{\beta}_1;\ p_1 \times 1$, $\boldsymbol{\beta}_2;\ p_2 \times 1$. The following lemmas (see, e.g., Fujikoshi et al. 2010) are used.

Lemma A1.1. *For the population Mahalanobis distances Δ and Δ_1, it holds that*

$$(1)\ \Delta^2_{2\cdot1} = \Delta^2 - \Delta^2_1 \geq 0.$$
$$(2)\ \Delta^2_{2\cdot1} = 0 \Leftrightarrow \boldsymbol{\beta}_2 = \mathbf{0}.$$

Note that

$$\Delta^2 - \Delta^2_{(-i)} = 0 \Leftrightarrow \beta_i = 0. \tag{A1.1}$$

Lemma A1.2. *For the sample Maharanobis distances D_1 and D, it holds that*

(1) $D^2_1 = (n-2)g^{-2}\chi^2_{p_1}(g^2\Delta^2_1)\left\{\chi^2_{n-p_1-1}\right\}^{-1}$.

(2) If $\Delta^2_{2\cdot1} = 0,\ g^2(D^2 - D^2_1)\{n - 2 + g^2D^2_1\}^{-1} = \chi^2_{p_2}\{\chi^2_{n-p-1}\}^{-1}$.

(3) If $\Delta^2_{2\cdot1} = 0,\ (D^2 - D^2_1)\{n - 2 + g^2D^2_1\}^{-1}$and D^2_1 are independent,

(4) $D^2_{2\cdot1} = (n-2)g^{-2}\chi^2_{p_2}\left(g^2\Delta^2_{2\cdot1} \cdot \dfrac{1}{1+R}\right)\left\{\chi^2_{n-p-1}\right\}^{-1}(1+R)$,

where $R = \chi^2_{p_1}(g^2\Delta^2_1)\left\{\chi^2_{n-p_1-1}\right\}^{-1}$. Here, $\chi^2_{p_1}(\cdot)$, $\chi^2_{n-p_1-1}$, $\chi^2_{p_2}(\cdot)$, and χ^2_{n-p-1} are independent chi-square variates.

From Lemma A1.1, D^2 and $D^2_{(-i)}$ can be expressed as

$$D^2 = (n-2)g^{-2}\frac{\chi^2_p(g^2\Delta^2)}{\chi^2_{n-p-1}}, \quad D^2_{(-i)} = (n-2)g^{-2}\frac{\chi^2_{p-1}(g^2\Delta^2_{(-i)})}{\chi^2_{n-p-2}}, \tag{A1.2}$$

where $g = \sqrt{(n_1n_2)/n}$. It is easy to see that

$$\frac{1}{m}\chi^2_m \xrightarrow{p} 1, \quad \frac{1}{m}\chi^2_m(\delta^2) \xrightarrow{p} 1 + \eta^2, \tag{A1.3}$$

if $\eta^2 = \lim_{m\to\infty}\delta^2/m$.

A2 Proofs of Theorems 3.1 and 3.2

Without loss of generality, we may $j_* = \{1, \ldots, s\}$ and $p_* = s$. Then, [F1] and [F2] are expressed as

$$[\mathrm{F1}] = \sum_{i=1}^{s} P(\mathrm{D}_{d,i} \leq 0), \quad [\mathrm{F2}] = \sum_{i=s+1}^{p} P(\mathrm{D}_{d,i} \geq 0).$$

Proof of Theorem 3.1

Using (A1.2) and (A1.3),

$$\begin{aligned} D^2 &= (n-2)g^{-2}\frac{\chi_p^2(g^2\Delta^2)}{\chi_{n-p-1}^2} \\ &= (n-2)\frac{n}{n_1 n_2} \cdot \frac{\chi_p^2(g^2\Delta^2)}{p} \cdot \frac{n-p-1}{\chi_{n-p-1}^2} \cdot \frac{p}{n-p-1} \\ &\xrightarrow{p} \frac{c}{k_1 k_2(1-c)} + \frac{1}{1-c}\Delta^2. \end{aligned}$$

Similarly,

$$D_{(-i)}^2 \xrightarrow{p} \frac{c}{k_1 k_2(1-c)} + \frac{1}{1-c}\Delta_{(-i)}^2.$$

Therefor,

$$D^2 - D_{(-i)}^2 \xrightarrow{p} \frac{1}{1-c}\left(\Delta^2 - \Delta_{(-i)}^2\right)$$

If $i \in j_* = \{1, \ldots, s\}$, then, $\tau_i = \Delta^2 - \Delta_{(-i)}^2 > 0,\ i = 1, \ldots, s$, and

$$D^2 - D_{(-i)}^2 \xrightarrow{p} \frac{1}{1-c}\tau_i.$$

Therefor, if $d < \frac{1}{1-c}\tau_{\min}$, where $\tau_{\min} = \min_{i=1,\ldots,s} \tau_i$, $\mathrm{D}_{d,i} = D^2 - D_{(-i)}^2 - d$ is positive, and

$$\begin{aligned} &P\left(\mathrm{D}_{d,i} \leq 0\right) \to 0 \\ \Rightarrow\ &[\mathrm{F1}] = \sum_{i=1}^{s} P\left(\mathrm{D}_{d,i} \leq 0\right) \to 0, \end{aligned}$$

since s is finite.

Proof of Theorem 3.2

Next we shall show [F2] under the assumptions in Theorem 3.2. Assume that $i \in \{s+1, \ldots, p\}$. Then, we have $\Delta^2 = \Delta^2_{(-i)}$. In general, note that $D^2 - D^2_{(-i)} > 0$. Using Lemma A1.2 (1),

$$\mathrm{E}(D^2) = \frac{(n-2)np}{n_1 n_2 (n-p-3)} + \frac{n-2}{n-p-3}\Delta^2,$$

and hence

$$E\left\{D^2 - D^2_{(-i)}\right\} = q(p, n, \Delta^2),$$

where $q(p, n, \Delta^2)$ is given by (3.5). Suppose that.

$$d > q(p, n, \Delta^2).$$

Then, we have

$$\begin{aligned}
&P\left(D^2 - D^2_{(-i)} \geq d\right) \\
&= P\left(D^2 - D^2_{(-i)} - \mathrm{E}[D^2 - D^2_{(-i)}] \geq d - \mathrm{E}[D^2 - D^2_{(-i)}]\right) \\
&\leq P\left(\left|D^2 - D^2_{(-i)} - \mathrm{E}[D^2 - D^2_{(-i)}]\right| \geq d - \mathrm{E}[D^2 - D^2_{(-i)}]\right) \\
&\leq \frac{1}{(d - \mathrm{E}[D^2 - D^2_{(-i)}])^2}\mathrm{Var}(D^2 - D^2_{(-i)}).
\end{aligned}$$

The last inequality follows from Chebyshev's inequality. It is easy to see that $\mathrm{E}[D^2 - D^2_{(-i)}] = \mathrm{O}(n^{-1})$. Since $d = \mathrm{O}(1)$ from our assumption,

$$(d - \mathrm{E}[D^2 - D^2_{(-i)}])^2 = \mathrm{O}(1).$$

Using Lemma A1.2 (4), the denominator is expressed as

$$\begin{aligned}
\mathrm{Var}(D^2 - D^2_{(-i)}) &= \mathrm{E}\{(D^2 - D^2_{(-i)})^2\} - \{\mathrm{E}(D^2 - D^2_{(-i)})\}^2 \\
&= (n-2)^2 g^{-4} \mathrm{E}\left\{\left(\frac{\chi^2_1}{\chi^2_{n-p-1}}\right)^2 \left(1 + \frac{\chi^2_{p-1}(g^2\Delta^2)}{\chi^2_{n-p}}\right)^2\right\} \\
&\quad - (n-2)^2 g^{-4} \frac{1}{(n-p-3)^2}\left(1 + \frac{p-1+g^2\Delta^2}{n-p-2}\right)^2.
\end{aligned}$$

The expectation of the first term can be computed as

$$
\begin{aligned}
&\mathrm{E}\left\{\left(\frac{\chi_1^2}{\chi_{n-p-1}^2}\right)^2\left(1+\frac{\chi_{p-1}^2(g^2\Delta^2)}{\chi_{n-p}^2}\right)^2\right\} \\
&= \frac{1}{(n-p-3)(n-p-5)}\mathrm{E}\left\{1+2\frac{\chi_{p-1}^2(g^2\Delta^2)}{\chi_{n-p}^2}+\left(\frac{\chi_{p-1}^2(g^2\Delta^2)}{\chi_{n-p}^2}\right)^2\right\} \\
&= \frac{1}{(n-p-3)(n-p-5)}\Bigg\{1+2\frac{p-1+g^2\Delta^2}{n-p-2} \\
&\quad + \frac{(p-1)^2+2(p-1)+2(p-1)g^2\Delta^2+4g^2\Delta^2+g^4\Delta^4}{(n-p-2)(n-p-4)}\Bigg\} \\
&= \frac{1}{(n-p-3)(n-p-5)}\Big\{1+O(1)+O(1)\Big\} = O(n^{-2}).
\end{aligned}
$$

The second term is evaluated as

$$
\begin{aligned}
&(n-2)^2 g^{-4}\frac{1}{(n-p-3)^2}\left(1+\frac{p-1+g^2\Delta^2}{n-p-2}\right)^2 \\
&= O(1)\times\frac{1}{(n-p-3)^2}\left(1+O(1)\right)^2 = O(n^{-2}).
\end{aligned}
$$

Therefore, we have

$$
\mathrm{Var}(D^2 - D_{(-i)}^2) = (n-2)^2 g^{-4} O(n^{-2}) - O(n^{-2}) = O(n^{-2}).
$$

These imply the followings:

$$
\begin{aligned}
&P\left(D^2 - D_{(-i)}^2 \geq d\right) \leq O(n^{-2}), \text{ and hence} \\
&\sum_{i=s+1}^{p} P\left(D^2 - D_{(-i)}^2 \geq d\right) \to 0,
\end{aligned}
$$

which proves "[F2] $\to 0$".

A3 Proof of Theorem 4.1

As in the proofs of Theorems 3.1 and 3.2, assume that $j_* = \{1, \ldots, s\}$, and show that

$$[\mathrm{F1}] \equiv \sum_{i=1}^{s} P\left(D^2 - D^2_{(-i)} \leq \widehat{d}\right) \to 0,$$
$$[\mathrm{F2}] \equiv \sum_{i=s+1}^{p} P\left(D^2 - D^2_{(-i)} \geq \widehat{d}\right) \to 0.$$

First consider [F1], and assume that $i \in \{1, \ldots, p_*\}$. In the proof of Theorem 3.1, it has been shown that

$$D^2 \xrightarrow{p} \frac{c}{k_1 k_2 (1-c)} + \frac{1}{1-c}\Delta^2.$$

Using this result, we have

$$\widehat{d} \xrightarrow{p} 0.$$

These imply that

$$D^2 - D^2_{(-i)} - \widehat{d} \xrightarrow{p} \frac{1}{1-c}\tau_i > 0,$$

and hence

$$P(D^2 - D^2_{(-i)} \leq \widehat{d}) = 0,$$

which implies "[F1] $\to$ 0".

Next, consider to show "[F2] $\to$ 0". Assume that $i \in \{s+1, \ldots, p\}$. Then, $\Delta^2 = \Delta^2_{(-i)}$. Denote that $D^2 - D^2_{(-i)} = D_{\{i\}\cdot(-i)}$. Using Lemma A1.2(2) and (3), we can write as

$$\begin{aligned} P\left(D^2 - D^2_{(-i)} \geq \widehat{d}\right) &= P\left(\frac{\chi^2_1}{\chi^2_{n-p-1}} \geq \frac{g^2\widehat{d}}{n-2+g^2 D^2_{(i)}}\right) \\ &\leq P\left(\frac{\chi^2_1}{\chi^2_{n-p-1}} \geq \frac{g^2\widehat{d}}{n-2+g^2 D^2}\right) \\ &= P\left(F_{1,n-p-1} - \frac{n-p-1}{n-p-3} \geq a(np)^{1/3}\right). \end{aligned}$$

Noting that the mean of $F_{1,n-p-1}$ is $(n-p-1)(n-p-3)^{-1}$,

$$\begin{aligned} P\left(D^2 - D^2_{(-i)} \geq \widehat{d}\right) &\leq P\left(|F_{1,n-p-1} - \mathrm{E}[F_{1,n-p-1}]| \geq a(np)^{1/3}\right) \\ &\leq \frac{1}{a^2(np)^{2/3}}\mathrm{Var}(F_{1,n-p-1}) \\ &= \frac{1}{a^2(np)^{2/3}}\frac{2n^2(n-1)}{(n-2)^2(n-4)}. \end{aligned}$$

This implies that

$$\begin{aligned} [\mathrm{F2}] &\leq \sum_{i=j_*+1}^{p} \frac{1}{a^2(np)^{2/3}}\frac{2n^2(n-1)}{(n-2)^2(n-4)} \\ &\leq \frac{1}{a^2}\frac{p}{(np)^{2/3}}\frac{2n^2(n-1)}{(n-2)^2(n-4)} = \mathrm{O}(n^{-1/3}), \end{aligned}$$

which proves "[F2] $\to 0$".

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] Yamada, T., Sakurai, T. and Fujikoshi, Y. (2017). High-dimensional asymptotic results for EPMCs of W- and Z- rules. Hiroshima Statistical Research Group, TR 17-11.

[8] Witten, D. W. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc. B*, **73**, 753–772.