High-dimensional consistencies of KOO methods in multivariate regression model and discriminant analysis

Yasunori Fujikoshi

Department of Mathematics, Hiroshima University, Higashi-Hiroshima, Japan

Abstract

In this paper, we review recent developments in high-dimensional consistencies of KOO methods for selection of variables in multivariate regression models and discriminant analysis models. The KOO methods considered are mainly based on general information criteria, but we also take up KOO methods based on some other selection methods. Some references are given for high-dimensional consistencies in some other multivariate models.

Key Words and Phrases: Discriminant analysis, General information criteria, High-dimensional consistency, KOO Methods, Multivariate regression model, Selection of variables.

2020 MSC: Primary 62H12, Secondary 62H30

1. Introduction

It is important to consider selection of variables in multivariate analysis. One of the approaches is to first consider variable selection models and then apply model selection criteria such as AIC, BIC, Cp. The AIC and BIC are to find the model which minimizes

$$\mathrm{GIC} = -2\log L(\widehat{\boldsymbol{\theta}}) + dg,$$

where $L(\widehat{\boldsymbol{\theta}})$ is the maximum likelihood, dg is the penalty term, and g is the number of unknown parameters. For AIC and BIC, d is defined as 2 and $\log n$, respectively, where n denotes the sample size. Cp is defined by using the mean squared error instead of $-2\log L(\widehat{\boldsymbol{\theta}})$.

In the seletion of k variables $x_1, \ldots, x_k$, we identify $\{x_1, \ldots, x_k\}$ with the index set $\{1, \ldots, k\} \equiv \boldsymbol{\omega}$, and denote GIC for subset $\boldsymbol{j} \subset \boldsymbol{\omega}$ by $\mathrm{GIC}_{\boldsymbol{j}}$. Then, the model selection based on GIC chooses the model

$$\widetilde{\boldsymbol{j}}_G = \arg\min_{\boldsymbol{j}} \mathrm{GIC}_{\boldsymbol{j}}.$$

Here the minimum is usually taken over all subsets. It has been pointed out that there are computational problems for GIC methods, like AIC, BIC and Cp methods, since we need to compute $2^k - 1$ statistics for the selection of k variables. To avoid this computational problem, Nishii et al. (1988) proposed a method which is essentially due to Zhao et al. (1988). The method, which was named the knock-one-out (KOO) method by Bai et al. (2018), determines "selection" or "no selection" for each variable by comparing the model aftr removing a variable and the full model. More precisely, the KOO method chooses the model or the set of variables given by

$$\widehat{\boldsymbol{j}}_G = \{j \in \boldsymbol{\omega} \mid \mathrm{GIC}_{\boldsymbol{\omega}\backslash j} > \mathrm{GIC}_{\boldsymbol{\omega}}\},$$

where $\boldsymbol{\omega}\backslash j$ is the set obtained by removing element j from the set $\boldsymbol{\omega}$.

In large-sample setting, Nishii et al. (1988) studied strong consistency of $\widehat{\boldsymbol{j}}_G$ and $\widetilde{\boldsymbol{j}}_G$ in discriminant analysis, canonical correlation analysis and multivariate calibration analysis. It is well known that BIC is consistent, but AIC is not consistent. On the other hand, from some recent work in multivariate regression models by Fujikoshi et al. (2014), Yanagihara et al. (2015) and Bai et al. (2018), it is known that if the BIC selection rule is consistent so is the AIC selection rule, but not vice versa. There are considerably many results on the KOO methods besides these high-dimensional consistency results, but the purpose of this paper is to review only the recent of the latter. Though we consider mainly the KOO methods based on general information

criteria, ones based on some other selection methods are also considered. Some results in multivariate regression models have been studied for strong consistency under nonnormality by Bai et al. (2018). However, those results are under revision, in the present paper we discuss only with results on weak consistency under normality.

The remainder of the present paper is organized as follows. In Section 2, we present KOO methods based on a general information criterion in a multivariate regression model, as well as KOO methods based on some other criteria. In Section 3, we present KOO methods based on a general information criterion for selection of variables in discriminant analysis. The methods are discussed in two-group and multiple group cases separately. In Section 4, we briefly discuss selection of vriables in some other models.

2. Multivariate regression model

2.1. KOO Methods based on GIC

Suppose that there are n observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ on p response variables $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ and n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ on k explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$. Here, $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ are observations of $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively, for the ith subject. Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$. The multivariate regression model is written as

$$\mathbf{Y} = \mathbf{X}\mathbf{\Theta} + \mathcal{E},$$

where $\mathbf{\Theta}$ is a $k \times p$ regression coefficient matrix, and $\mathcal{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$ is the error matrix. It is assumed that the $\boldsymbol{\epsilon}_i$'s are independently distributed as a $p-$variate normal distribution $\mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma})$. In order to explore a simpler linear structure, we consider how to select the explanatory variables. In general, the selection of x_i may be decided by whether or not the ith row

$\boldsymbol{\theta}_i$ of $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)^\top$ is the null vector. We consider such a selection problem, assuming that $\boldsymbol{\Sigma}$ is unknown positive definite, though in the last part of this section we consider three covariance structures. For notational simplicity, let us identify the set $\{x_1, \ldots, x_k\}$ with $\boldsymbol{\omega} = \{1, \ldots, k\}$, and let $k_{\boldsymbol{j}}$ be the cardinality of $\boldsymbol{j}$. Further, let $\boldsymbol{j}$ be a subset of $\boldsymbol{\omega}$, $\boldsymbol{x}_{\boldsymbol{j}} = (x_j, j \in \boldsymbol{j})^\top$, and $\mathbf{X}_{\boldsymbol{j}} = (\boldsymbol{x}_j, j \in \boldsymbol{j})$. Denote the model based on $\boldsymbol{x}_{\boldsymbol{j}}$ by

$$M_{\boldsymbol{j}} : \quad \mathbf{Y} = \mathbf{X}_{\boldsymbol{j}} \Theta_{\boldsymbol{j}} + \boldsymbol{\mathcal{E}}.$$

Then, GIC is expressed as

$$\mathrm{G}_{d,\boldsymbol{j}} = n \log |\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}| + d \left\{ k_{\boldsymbol{j}} p + p(p+1)/2 \right\} + np\{\log(2\pi) + 1\}. \tag{2.1}$$

Here, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$ is the MLE of $\boldsymbol{\Sigma}$ given by

$$n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}}) \mathbf{Y}, \ \mathbf{P}_{\boldsymbol{j}} = \mathbf{X}_{\boldsymbol{j}} (\mathbf{X}_{\boldsymbol{j}}^\top \mathbf{X}_{\boldsymbol{j}})^{-1} \mathbf{X}_{\boldsymbol{j}}^\top,$$

where $\mathbf{P}_{\boldsymbol{j}}$ is an orthogonal projection matrix onto to the space spanned by the colomn vectors of $\mathbf{X}_j$. It is assumed that the rank of $\mathbf{X}_{\boldsymbol{j}}$ is $k_{\boldsymbol{j}}$, and further, it is assumed that $n - k \geq p$ when we use (2.1). Using

$$\mathrm{T}_{d,j} = \mathrm{G}_{d,\boldsymbol{\omega}\setminus j} - \mathrm{G}_{d,\boldsymbol{\omega}} = n \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|/|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) - dp, \tag{2.2}$$

the KOO method chooses the model

$$\widehat{\boldsymbol{j}}_d = \{ j \in \boldsymbol{\omega} \mid \mathrm{T}_{d,j} > 0 \}. \tag{2.3}$$

Recently, high-dimensional consistency properties of KOO methods have been studied by Bai et al. (2018), Sakurai and Fujikoshi (2020) and Oda and Yanagihara (2020, 2021). These considered KOO methods based on the Cp criterion, as well as GIC. Bai et al. (2018) studied strong consistency properties under nonnormality. Sakurai and Fujikoshi (2020) studied the case when a certain covariance structure holds, specifically, an independent covariance structure, a uniform covariance structure, and an autoregressive covariance structure. In our model the sample size is not necessarily larger

than the number of response variables. Sufficient conditions for these criteria to be consistent are derived under a high-dimensional asymptotic framework such that the sample size and the dimensionality proceed to infinity together, with their ratio converging to a finite nonzero constant. However, we consider the case that the number of variables k is fixed.

2.2. Consistency of KOO methods based on GIC

First we consider the distributional reduction of $\mathrm{T}_{d,j}$ in (2.2). Note that the first term of $\mathrm{T}_{d,j}$ is $-2\log\lambda_j$, where λ_j is the likelihood ratio crterion for testing the hypothesis $\boldsymbol{\theta}_j = \mathbf{0}$. More precisely,

$$-2\log\lambda_j = n\log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|/|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) = n\log|\mathbf{S}_e|/|\mathbf{S}_e + \mathbf{S}_{h,j}|.$$

Here, $\mathbf{S}_e = \mathbf{Y}^\top(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{Y} \sim \mathrm{W}_p(n-k, \boldsymbol{\Sigma})$, $\mathbf{S}_{h,j} = \mathbf{Y}^\top(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\mathbf{Y} \sim \mathrm{W}_p(1, \boldsymbol{\Sigma}; \boldsymbol{\Gamma}_j)$, and $\mathbf{S}_e$ and $\mathbf{S}_h$ are independent. The noncentrality matrix may be expressed as

$$\boldsymbol{\Gamma}_j = (\mathbf{X}\boldsymbol{\Theta})^\top(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\mathbf{X}\boldsymbol{\Theta} = \boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top,$$

where $\boldsymbol{\gamma}_j = \boldsymbol{\theta}_j\boldsymbol{x}_j^\top\widetilde{\boldsymbol{x}}_j(\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{x}}_j)^{-1/2}$ and $\widetilde{\boldsymbol{x}}_j = (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\boldsymbol{x}_j$. In this case, we may write $\mathbf{S}_{h,j} = \boldsymbol{u}_j\boldsymbol{u}_j^\top$, where $\boldsymbol{u}_j \sim \mathrm{N}_p(\boldsymbol{\gamma}_j, \boldsymbol{\Sigma})$. Then, using $|\mathbf{S}_e + \boldsymbol{u}_j\boldsymbol{u}_j^\top| = |\mathbf{S}_e|(1 + \boldsymbol{u}_j^\top\mathbf{S}_e^{-1}\boldsymbol{u}_j)$ and a well-known result (see, e.g., Fujikoshi et al. (2010), Theorem 3.1.1), we have

$$\mathrm{T}_{d,j} = n\log\left\{1 + \frac{\chi_j^2(p;\delta_j^2)}{\chi^2(m)}\right\} - dp, \tag{2.4}$$

where $\chi^2(m)$ denotes a random variable with the chi-square distribution with $m = n - p - k + 1$ degrees of freedom, $\chi_j^2(p;\delta_j^2)$ denotes the noncentral χ^2 distributed random variable with p degree of freedom and noncentrality $\delta_j^2 = \boldsymbol{\gamma}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_j$, and $\chi^2(m)$ and $\chi_j^2(p;\delta_j^2)$ are independent.

In this section, we examine the high-dimensional consistency of KOO method based on GIC which is expressed as (2.3) in term of $\mathrm{T}_{d,j}$. Consistency

here will be proved by showing the following two properties:

$$[\mathrm{F1}] \equiv \sum_{j \in \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,j} \leq 0) \to 0. \tag{2.5}$$

$$[\mathrm{F2}] \equiv \sum_{j \notin \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,j} \geq 0) \to 0. \tag{2.6}$$

The sufficiency of these properties can be shown by using the following inequalty:

$$\begin{aligned}
\Pr(\widehat{\boldsymbol{j}}_{\mathrm{G}} = \boldsymbol{j}_*) &= \Pr\left(\bigcap_{j \in \boldsymbol{j}_*} \text{“}\mathrm{T}_{d,j} > 0\text{”} \bigcap_{j \notin \boldsymbol{j}_*} \text{“}\mathrm{T}_{d,j} < 0\text{”} \right) \\
&= 1 - \Pr\left(\bigcup_{j \in \boldsymbol{j}_*} \text{“}\mathrm{T}_{d,j} \leq 0\text{”} \bigcup_{j \notin \boldsymbol{j}_*} \text{“}\mathrm{T}_{d,j} \geq 0\text{”} \right) \\
&\geq 1 - \sum_{j \in \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,j} \leq 0) - \sum_{j \notin \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,j} \geq 0).
\end{aligned}$$

Here, [F1] denotes the probability that the true variables are not selected, and [F2] denotes the probability that the non-true variables are selected. The same notation is used for other variable selection methods.

We make the following assumptions.

P1: The true subset $\boldsymbol{j}_*$ is included in the full set $\boldsymbol{\omega}$, i.e., $\boldsymbol{j}_* \subset \boldsymbol{\omega}$.

P2: A high-dimensional asymptotic framework holds:

$$p/n \to c_1 \in (0,1), \quad k/n \to c_2 \in [0,1), \text{ where } c_1 + c_2 < 1.$$

P3: The noncentrality parameters satisfy $\delta_j^2 = \mathrm{O}(n)$ for $j \in \boldsymbol{j}_*$.

The assumption $c_1 + c_2 < 1$ in P2 comes from $n - p - k > 0$. In P3, the noncentrality parameters may be expressed as

$$\delta_j^2 = \boldsymbol{\theta}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}_j (\boldsymbol{x}_j^\top \widetilde{\boldsymbol{x}}_j)^2 (\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{x}}_j)^{-1}.$$

Here we denote the true covariance matrix $\boldsymbol{\Sigma}^*$ by $\boldsymbol{\Sigma}$ simply.

When $j \notin \boldsymbol{j}_*$, from (2.4) we can write $\mathrm{T}_{d,j} = n \log\left\{1 + \chi_j^2(p)/\chi^2(m)\right\} - dp$, and therefore we have

$$\begin{aligned}
[\mathrm{F2}] &= \sum_{j \notin \boldsymbol{j}_*} \Pr(n \log\left\{1 + \chi_j^2(p)/\chi^2(m)\right\} \geq dp) \\
&= (k - k_{\boldsymbol{j}_*}) \Pr(U \geq h) \\
&\leq (k - k_{\boldsymbol{j}_*}) \Pr(U \geq h_0),
\end{aligned}$$

where

$$\begin{aligned}
U &= \frac{\chi^2(p)}{\chi^2(m)} - \frac{p}{m-2}, \\
h &= \mathrm{e}^{dp/n} - 1 - \frac{p}{m-2}, \quad h_0 = \frac{dp}{n} - \frac{p}{m-2}.
\end{aligned} \tag{2.7}$$

Note that $h_0 < h$. Then, under the assumption $h_0 > 0$ we have,

$$[\mathrm{F2}] \leq (k - k_{\boldsymbol{j}_*})\, h^{-2\ell} \mathrm{E}[U^{2\ell}] \leq (k - k_{\boldsymbol{j}_*})\, h_0^{-2\ell} \mathrm{E}[U^{2\ell}].$$

Under P2, it is easy to see that $\mathrm{E}[U^2] = \mathrm{O}(n^{-1}), \mathrm{E}[U^4] = \mathrm{O}(n^{-2})$. Assuming that $h^{-4} \leq \mathrm{O}(n^{1-\epsilon})$ for some $\epsilon > 0$, we have that $[\mathrm{F2}] \to 0$.

When $j \in \boldsymbol{j}_*$, we can write $\mathrm{T}_{d,j} = n \log\left\{1 + \chi_j^2(p; \delta_j^2)/\chi^2(m)\right\} - dp$, and therefore we have

$$[\mathrm{F1}] = \sum_{j \in \boldsymbol{j}_*} \Pr(n \log\left\{1 + \chi_j^2(p; \delta_j^2)/\chi^2(m)\right\} \leq dp) = \sum_{j \in \boldsymbol{j}_*} \Pr(\widetilde{U}_j \leq \widetilde{h}_j),$$

where for $j \in \boldsymbol{j}_*$,

$$\widetilde{U}_j = \frac{\chi_j^2(p; \delta_j^2)}{\chi^2(m)} - \frac{p + \delta_j^2}{m-2}, \quad \widetilde{h}_j = \mathrm{e}^{dp/n} - 1 - \frac{p + \delta_j^2}{m-2} = h - \frac{\delta_j^2}{m-2}. \tag{2.8}$$

Then, under the assumption $\widetilde{h}_j < 0$, or equivalently $h < \delta_j^2/(m-2)$, we have

$$[\mathrm{F1}] \leq k_{\boldsymbol{j}_*} |\widetilde{h}_j|^{-2\ell} \mathrm{E}[\widetilde{U}^{2\ell}].$$

Under P1 and P3, it is easy to see that $\mathrm{E}[\widetilde{U}_j^2] = \mathrm{O}(n^{-1}), \mathrm{E}[\widetilde{U}_j^4] = \mathrm{O}(n^{-2})$. Assuming that $|\widetilde{h}_j|^{-4} \leq \mathrm{O}(n^{1-\epsilon})$ for some $a < 1$, we can see that $[\mathrm{F2}] \to 0$. These imply the following theorem.

Theorem 2.1. *Suppose that assumptions* P1 *and* P2 *hold. Then, for the* KOO *method* (2.3) *based on* GIC*, the following asymptotic results hold.*
(i) *Suppose that the quantity* h_0 *in* (2.7) *is asymptotically positive or* ∞*, and* $h_0^{-4} \leq \mathrm{O}(n^a)$ *for some* $a < 1$*. Then,* $[\mathrm{F2}] \to 0$*.*
(ii) *Suppose that* P3 *holds, the quantity* $\widetilde{h}_0$ *in* (2.8) *is asymptotically negative, and* $|\widetilde{h}_j|^{-4} = \mathrm{O}(n^a)$ *for some* $a < 1$*. Then,* $[\mathrm{F1}] \to 0$*.*
(iii) *Suppose that the assumptions in* (i) *and* (ii) *hold. Then, the* KOO *method* (2.3) *based on* GIC *is asymptoticaly consistent.*

Note that

$$\lim h_0 = c_1 \left\{ \lim d - \frac{1}{1 - c_1 - c_2} \right\},$$

and so, the condition for h_0 in Theorem 2.1 for AIC is satisfied if $2 - 1/(1 - c_1 - c_2) > 0$.

Bai et al. (2018) have studied high-dimensional strong consistency of KOO methods based on AIC, BIC and C_p by examining their strong convergences. On the other hand, Oda and Yanagihara (2020, 2021) and Sakurai and Fujikoshi (2020) used the above method. Further, Oda and Yanagihara (2021) considered a transformation from d to b given by

$$d = \frac{n}{p} \log \left\{ 1 + \frac{p}{m-2}(1+b) \right\}, \tag{2.9}$$

with $b > 0$. Then, writing $\rho = (pb)/(m-2)$, we have

$$h = \rho, \quad \widetilde{h}_j = \rho - \delta_j^2/(m-2).$$

Using these expressions, we can write

$$[\mathrm{F2}] = (k - k_{\boldsymbol{j}_*})\Pr(U \geq \rho) \leq k\rho^{-2\ell}\mathrm{E}(U^{2\ell}), \tag{2.10}$$

$$[\mathrm{F1}] = \sum_{j \in \boldsymbol{j}_*} \Pr\left(\widetilde{U}_j \leq \rho - \frac{\delta_j^2}{m-2} \right). \tag{2.11}$$

Instead of the assumption P3, we use more precise conditions P4 and P5 on the noncentralty parameter as the one in terms of $\boldsymbol{x}_j^\top(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\boldsymbol{x}_j$ and

$\boldsymbol{\theta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_j,\ j \in \boldsymbol{j}_*$:

P4: There exists $c_1 > 0$ such that $n^{-1}\min_{j\in\boldsymbol{j}_*} \boldsymbol{x}_j^\top(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\boldsymbol{x}_j \geq c_1$.

P5: There exist $c_1 > 0$ and $c_3 \geq 1/2$ such that $n^{1-c_3}\min_{j\in\boldsymbol{j}_*} \boldsymbol{\theta}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_j \geq c_2$, where $\boldsymbol{\theta}_j$ and $\boldsymbol{\Sigma}$ are their true values.

Evaluating (2.10) and (2.11), Oda and Yanagihara (2021) described a class of b having a high-dimensional consistency as in the following theorem.

Theorem 2.2. *Suppose that assumptions* P1, P4 *and* P5 *hold. Then, the* KOO *method* (2.3) *based on* GIC *has the selection consistency under the high-dimensional asymptotic framework* P2*, if for some integer r the following conditions are satisfied:*

$$\sqrt{p}b/k^{1/(2r)} \to \infty, \text{ and } pb/n^{c_3} \to 0. \tag{2.12}$$

If we can take $r = 2$ in the assumption of Theorem 2.2, we have

$$d = \widetilde{d} = \frac{n}{p}\log\left(1 + \frac{p}{m-2} + \frac{k^{1/4}\sqrt{p}\log n}{m-2}\right),$$

as an example of d satisfying (2.12).

2.3. KOO Methods based on some other criteria

As a criterion with behavior similar to that of AIC, we have the C_p criterion (see Fujikoshi and Satoh (1997), Sparks et al. (1983)). Its generalization may be defined (see, e.g., Yanagihara (2016)) by

$$\mathrm{GC}_{d,\boldsymbol{j}} = (n-k)\mathrm{tr}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1} + dpk_j.$$

The $\mathrm{GC}_{2,\boldsymbol{j}}$ case, with $d = 2$, is the usual C_p criterion. The KOO method based on $\mathrm{GC}_{d,\boldsymbol{j}}$ is to select j such that $\mathrm{T}_{d,j;\mathrm{GC}}$ is positive, where

$$\mathrm{T}_{d,j;\mathrm{GC}} = \mathrm{GC}_{d,\boldsymbol{\omega}\setminus j} - \mathrm{GC}_{d,\boldsymbol{\omega}}.$$

Using the same notation as for GIC, we can express

$$\begin{aligned}\mathrm{T}_{d,j;\mathrm{GC}} &= (n-k)\mathrm{tr}\mathbf{Y}^\top(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\mathbf{Y}\left\{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{Y}\right\}^{-1} - dp \\ &= (n-k)\boldsymbol{u}_j^\top \mathbf{S}_e^{-1}\boldsymbol{u}_j - dp \\ &= (n-k)\chi_j^2(p;\delta_j^2)\left\{\chi^2(m)\right\}^{-1} - dp.\end{aligned}$$

Therefore we have

$$\begin{aligned}\sum_{j\notin\boldsymbol{j}_*}\Pr(\mathrm{T}_{d,j;\mathrm{GC}} \geq 0) &= \sum_{j\notin\boldsymbol{j}_*}\Pr((n-k)\chi^2(p)\{\chi^2(m)\}^{-1} \geq dp) \\ &= (k - k_{\boldsymbol{j}_*})\Pr(U \geq \widetilde{h}_0),\end{aligned}$$

where $\widetilde{h}_0 = dp(n-k)^{-1} - p(m-2)^{-1}$. Similarly, we have

$$\begin{aligned}\sum_{j\in\boldsymbol{j}_*}\Pr(\mathrm{T}_{d,j;\mathrm{GC}} \leq 0) &= \sum_{j\in\boldsymbol{j}_*}\Pr((n-k)\chi_j^2(p;\delta_j^2)\{\chi^2(m)\}^{-1} \geq dp) \\ &= \sum_{j\in\boldsymbol{j}_*}\Pr(\widetilde{U}_j \geq \bar{h}_j),\end{aligned}$$

where $\bar{h}_j = dp(n-k)^{-1} - (p+\delta_j^2)(m-2)^{-1} = \widetilde{h}_0 - \delta_j^2(m-2)^{-1}$.

From the above reduction, we can get a result similar to Thorem 2.1. Oda and Yanagihara (2020) considered the transformation $d = (n-k)(m-2)^{-1}+r$ and gave a condition for r having high-dimensional consistency.

Next we consider the generalized prediction error criterion defined by

$$\mathrm{GP}_{d,\boldsymbol{j}} = (n-k)\frac{\mathrm{tr}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}}{\mathrm{tr}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}} + dk_{\boldsymbol{j}}.$$

The criterion in a special case $d = 2$ was proposed by Fujikoshi et al. (2011). For its generalization, see Oda (2020).

It may be noted that the criterion can be used in the case of $(n-k) < p$. The KOO method based on $\mathrm{GP}_{d,\boldsymbol{j}}$ is to select a subset given by

$$\widehat{\boldsymbol{j}}_{d,\mathrm{GP}} = \{j \in \boldsymbol{\omega} \mid \mathrm{T}_{d,j,\mathrm{GP}} > 0\},$$

where

$$\mathrm{T}_{d,j;\mathrm{GP}} = \mathrm{GP}_{d,\boldsymbol{\omega}\setminus j} - \mathrm{GP}_{d,\boldsymbol{\omega}} = (n-k)\frac{\mathrm{tr}\mathbf{S}_{h,j}}{\mathrm{tr}\mathbf{S}_{\boldsymbol{\omega}}} - d.$$

One of our interests is to examine a sufficient condition for $\widehat{\boldsymbol{j}}_{d,\mathrm{GP}}$ to be asymptotically consistent. Some results has been obtained under P1, P3 and the following additional assumptions:

$\widetilde{\mathrm{P}}2$: The high-dimensional asymptotic framework is as follows:

$$p/n \to \widetilde{c}_1 \in (0,\infty), \quad k/n \to c_2 \in [0,1).$$

$\widetilde{\mathrm{P}}3$: The covariance matrix $\boldsymbol{\Sigma}$ satisfies the following conditions;

$$(1/p)\mathrm{tr}\boldsymbol{\Sigma} \to \mathrm{O}(1), \quad (1/p)\mathrm{tr}\boldsymbol{\Sigma}^2 \to \mathrm{O}(1).$$

$\widetilde{\mathrm{P}}4$: The covariance matrix $\boldsymbol{\Sigma}$ and the noncentrality matrix satisfy the following conditions;

$$(1/p)\mathrm{tr}\boldsymbol{\Sigma}\boldsymbol{\Omega} \to \mathrm{O}(1), \quad (1/p)\mathrm{tr}\boldsymbol{\Sigma}^2\Omega \to \mathrm{O}(1).$$

3. Discriminant analysis

3.1. Two-group case

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 $\Pi_i : \mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i \in \{1,2\}$. Let $\mathbf{Y}$ be the total sample matrix defined by

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

The coefficients of the population discriminant function are given by

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

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

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

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

In this section, $\boldsymbol{\omega}$ is used for $\{1, \ldots, p\}$, and we let $\boldsymbol{j}$ denote a subset of $\boldsymbol{\omega}$ containing $p_{\boldsymbol{j}}$ elements, and $\boldsymbol{y}_{\boldsymbol{j}}$ denote the $p_{\boldsymbol{j}}$ vector consisting of the elements of $\boldsymbol{y}$, indexed by the elements of $\boldsymbol{j}$. We use the notation $D_{\boldsymbol{j}}$ and $D_{\boldsymbol{\omega}}$ for D based on $\boldsymbol{y}_{\boldsymbol{j}}$ and $\boldsymbol{y}_{\boldsymbol{\omega}} (= \boldsymbol{y})$, respectively. Let $M_{\boldsymbol{j}}$ be the variable selection model, defined by

$$M_{\boldsymbol{j}} : \ \beta_j \neq 0 \text{ if } j \in \boldsymbol{j}, \quad \beta_j = 0, j \notin \boldsymbol{j}.$$

Model $M_{\boldsymbol{j}}$ assume that $\Delta_{\boldsymbol{j}} = \Delta_{\boldsymbol{\omega}}$, i.e., that the Mahalanobis distance based on $\boldsymbol{y}_{\boldsymbol{j}}$ is the same as the one based on the full set of variables, $\boldsymbol{y}$. We identify the selection of $M_{\boldsymbol{j}}$ with the selection of $\boldsymbol{y}_{\boldsymbol{j}}$. Let $\mathrm{GIC}_{\boldsymbol{j}}$ be the GIC for $M_{\boldsymbol{j}}$ in two-group discriminant analysis, and $\mathrm{T}_{d,j} = \mathrm{GIC}_{\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{\boldsymbol{\omega}}$. Then, we have (see, e.g., Fujikoshi and Sakurai (2019)) that

$$\mathrm{T}_{d,j} = n \log \left\{ 1 + \frac{g^2 (D^2_{\boldsymbol{\omega}} - D^2_{\boldsymbol{\omega}\backslash j})}{n - 2 + g^2 D^2_{\boldsymbol{\omega}\backslash j}} \right\} - d, \tag{3.1}$$

where $g = \sqrt{(n_1 n_2)/n}$ and d is a positive constant that may depend on p and n. Our KOO method is defined by selecting the set of suffixes or the set of variables given by

$$\widehat{\boldsymbol{j}}_d = \{j \in \boldsymbol{\omega} \mid \mathrm{T}_{d,j} > 0\}, \quad \mathrm{T}_{d,j} \text{ given by (3.1)}. \tag{3.2}$$

Consistency of $\widehat{\boldsymbol{j}}_d$ for some $d > 0$ shall be shown following a similar outline as in the multivariate regression model. Denote a subset $\boldsymbol{\omega}\backslash j$ by $\{-j\}$ simply. The square of the Mahalanobis distance of $\boldsymbol{y}$ is decomposed as a sum of the squares of the Mahalanobis distance of $\boldsymbol{y}_{\{-j\}}$ and the conditional Mahalanobis distance of $\boldsymbol{y}_{\{j\}}$ given $\boldsymbol{y}_{\{-j\}}$

$$\Delta^2 = \Delta^2_{\{-j\}} + \Delta^2_{\{j\}\cdot\{-j\}}, \tag{3.3}$$

and hence $\Delta^2_{\{j\}\cdot\{-j\}} = \Delta^2 - \Delta^2_{\{-j\}}$.

Note that in the use of the KOO method, we have assumed $n - 2 \geq p$. For high-dimensional data such that $p > n$, Lasso and other regularization methods have been extended. For such studies, see, e.g., Clemmensen et al. (2011), Witten and Tibshirani (2011), and Hao and Dong (2015).

For a distributional reduction of $\mathrm{T}_{d,j}$ in (3.1), we use the following Lemma (see, e.g., Fujikoshi et al. (2010)).

Lemma 3.1. *Let D_1 and D be the sample Mahalanobis distances based on $\boldsymbol{y}_1; p_1 \times 1$ and $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top)^\top$, respectively, and let $D^2_{2\cdot1} = D^2 - D^2_1$. Similarly, the corresponding population quantities are expressed as Δ_1, Δ and $\Delta^2_{2\cdot1}$. Then, it holds that*

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

(ii) $D^2_{2\cdot1} = (n-2)g^{-2}\chi^2\left(p_2; g^2\Delta^2_{2\cdot1}(1+R)^{-1}\right)\left\{\chi^2(n-p-1)\right\}^{-1}(1+R)$.

(iii) $\dfrac{g^2(D^2 - D_1^2)}{n-2+g^2D_1^2} = \chi^2(p_2; g^2\Delta^2_{2\cdot1}(1+R)^{-1})\{\chi^2(n-p-1)\}^{-1}$.

Here, $\chi^2(p_1;\cdot)$, $\chi^2(n-p_1-1)$, $\chi^2(p_2;\cdot)$ and $\chi^2(n-p-1)$ denote independent chi-square variables.

Using Lemma 3.1 (iii), we have

$$\mathrm{T}_{d,j} = n\log\left\{1 + \frac{\chi^2(1; g^2\Delta^2_{\{j\}\cdot\{-j\}}(1+R_j)^{-1})}{\chi^2(n-p-1)}\right\} - d, \tag{3.4}$$

where $R_j = \chi^2(p-1; g^2\Delta^2_{\{-j\}})\left\{\chi^2(n-p)\right\}^{-1}$. When $j \notin \boldsymbol{j}_*$, $\Delta^2_{\{j\}\cdot\{-j\}} = 0$, and we have

$$\mathrm{T}_{d,j} = n\log\left\{1 + \frac{\chi^2(1)}{\chi^2(n-p-1)}\right\} - d. \tag{3.5}$$

Here, we list some of our main assumptions. Assumptions A1 and A2 are assumed under our high-dimensional asymptotic framework.

A1(The true model): The true subset $\boldsymbol{j}_*$ is included in the full set $\boldsymbol{\omega}$, i.e., $\boldsymbol{j}_* \subset \boldsymbol{\omega}$;

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

For a proof of "[F1] $\to 0$", we use the following two assumptions:

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

A4: For $j \in \boldsymbol{j}_*$, $\lim \Delta^2_{\{j\}\cdot\{-j\}} > 0$ and $\lim \Delta^2_{\{-j\}} > 0$, under $\Delta^2 = \mathrm{O}(1)$.

For the constant d in (3.1), we consider the following conditions:

B1: $d/n \to 0$;

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

The following results were given by Fujikoshi and Sakurai (2019).

Theorem 3.1. *Suppose that assumptions* A1 *and* A2 *hold. Then, for the* KOO *method* (3.2)*, the following asymptotic results hold.*
(i) *Suppose that* A3 *and* A4 *hold. Then,* [F1] $\to 0$;
(ii) *Suppose that* B1 *and* B2 *hold. Then,* [F2] $\to 0$;
(iii) *Suppose that the assumptions in* (1) *and* (2) *hold. Then, the* KOO *method* (2.3) *is asymptoticaly consistent.*

Proof of Theorem 3.1: First we show "[F1] $\to 0$". Let $j \in \boldsymbol{j}_*$. Then, $\Delta^2_{\{-j\}} < \Delta^2$, and $\Delta^2_{\{j\}\cdot\{-j\}} > 0$. Using (3.4) and Lemma 3.1 (iii)

$$\mathrm{T}_{d,j} = n \log \left\{ 1 + \frac{\chi^2(1; g^2 \Delta^2_{\{j\}\cdot\{-j\}}(1 + R_j)^{-1})}{\chi^2_{n-p-1}} \right\} - d,$$

where $R_j = \chi^2(p-1; g^2\Delta^2_{\{-j\}})\left\{\chi^2_{n-p}\right\}^{-1}$. Since $\boldsymbol{j}_*$ is finite, it is sufficient to show $\mathrm{T}_{d,j} \xrightarrow{p} t_j > 0$ or $\mathrm{T}_{d,j} \xrightarrow{p} \infty$. It is easily seen that $R_j \approx (p + g^2\Delta^2_{\{-j\}})(n-p)^{-1}$ and hence

$$(1+R_j)^{-1} \approx \frac{n-p}{n+g^2\Delta^2_{\{-j\}}},$$

where " $\approx$ " means asymptotic equivalence. Therefore, we obtain

$$\frac{1}{n}\mathrm{T}_{d,j} \to \log\left(1 + \frac{k_1 k_2 \lim \Delta^2_{\{j\}\cdot\{-j\}}}{1 + k_1 k_2 \lim \Delta^2_{\{-j\}}}\right) > 0,$$

which implies our assertion.

Next we show "[F2] $\to$ 0". Using (3.5), we can write

$$\begin{aligned}
[\mathrm{F2}] &= (p-p_*)\Pr(U > \mathrm{e}^{d/n} - 1 - (n-p-3)^{-1}) \\
&\le (p-p_*)\Pr(U \ge h) \\
&\le (p-p_*)h^{-2\ell}\mathrm{E}(U^{2\ell}), \quad \ell \in \{1,2,\ldots,\},
\end{aligned}$$

where $U = \chi^2(1)/\chi^2(n-p-1) - 1/(n-p-3)$. Noting $h = \mathrm{O}(n^{-a})$ and $\mathrm{E}[U^{2\ell}] = \mathrm{O}(n^{-2\ell})$ (see, e.g., Theorem 16.2.2 in Fujikoshi et al. (2010)), we have

$$[\mathrm{F2}] \le \mathrm{O}(n^{1-2\ell(1-a)}).$$

Choosing $\ell > 1/\{2(1-a)\}$, we have "[F2] $\to$ 0". □

Let $d = n^r$, where $0 < r < 1$. Then, $h = \mathrm{O}(n^{-(1-r)})$, and condition B2 is satisfied. Thus, the KOO method (3.2) with

$$d \in \left\{n^{3/4}, \quad n^{2/3}, \quad n^{1/2}, \quad n^{1/3}, \quad n^{1/4}\right\}$$

has a high-dimensional consistency. Among these, it has been pointed in Fujikoshi and Sakurai (2019) that the one with $d = \sqrt{n}$ has numerically a good behavior. Note that $\widehat{\boldsymbol{j}}_2$ and $\widehat{\boldsymbol{j}}_{\log n}$ as in (3.2) do not satisfy B2.

From our discussions, a consistency result under a large sample framework

$$\text{"}p;\ \text{fixed},\ \ n_i/n \to k_i > 0,\ \ i \in \{1,2\}.$$

may be noted. That is, $\hat{\boldsymbol{j}}_{\log n}$ and $\hat{\boldsymbol{j}}_{\sqrt{n}}$ are consistent under a large-sample framework, since $d \to \infty$ and $d/n \to 0$. Similar results was given by Nishii et al. (1988) in multiple-group case.

3.2. Multiple-group case

In multiple-group discriminant analysis, suppose that we have independent samples $\boldsymbol{y}_1^{(i)}, \ldots, \boldsymbol{y}_{n_i}^{(i)}$ from p-dimensional normal distributions $\Pi_i : \mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i \in \{1, \ldots, q+1\}$ of $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$. Let

$$\mathbf{Y} = (\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_{n_1}^{(1)}, \ldots, \boldsymbol{y}_1^{(q+1)}, \ldots, \boldsymbol{y}_{n_{q+1}}^{(q+1)})^\top$$

be the total sample matrix. The population between groups matrix is defined by

$$\boldsymbol{\Omega} = \sum_{i=1}^{q+1} \frac{n_i}{n} (\boldsymbol{\mu}^{(i)} - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(i)} - \overline{\boldsymbol{\mu}})^\top,$$

where $n = n_1 + \cdots + n_{q+1}$ and $\overline{\boldsymbol{\mu}} = n^{-1}(n_1 \boldsymbol{\mu}^{(1)} + \cdots + n_{q+1} \boldsymbol{\mu}^{(q+1)})$. Let $\lambda_1 \geq \cdots \geq \lambda_q$ be the non-zero characteristic roots of $\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}$. Then, the corresponding characteristic vectors are denoted by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ with normalizations $\boldsymbol{\beta}_i^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_i = 1$, $i \in \{1, \ldots, q\}$. They are the solutions of

$$\boldsymbol{\Omega}\boldsymbol{\beta}_i = \lambda_i \boldsymbol{\Sigma}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_j = \delta_{ij}.$$

Then, $\boldsymbol{\beta}_i$ is the coefficient vector of the ith population linear discriminant function. For simplicity, assume that $p > q$ and $\lambda_q > 0$. Related to a partition of $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top)^\top$, $\boldsymbol{y}_i : p_i \times 1$, let $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$ be partitioned as

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

Similarly, let $\boldsymbol{\beta}_i$ and $\boldsymbol{\Omega}$ be partitioned as

$$\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_{1i} \\ \boldsymbol{\beta}_{2i} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}$$

in the same way as $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$. The sufficiency hypothesis of $\boldsymbol{y}_1$ or the redundancy hypothesis of $\boldsymbol{y}_2$ was introduced by Rao (1970, 1973)) as

$$H_{2\cdot 1}: \ \boldsymbol{\mu}_{2\cdot 1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot 1}^{(q+1)},$$

where $\boldsymbol{\mu}_{2\cdot 1}^{(i)} = \boldsymbol{\mu}_2^{(i)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1^{(i)}$, $i \in \{1, \ldots, q+1\}$. Then the statement $H_{2\cdot 1}$ is equivalent to one of the following equalifications:

$$\text{(i)} \quad \boldsymbol{\beta}_{2i} = 0, \quad i \in \{1, \ldots, q\}, \quad \text{(ii)} \quad \mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega} = \mathrm{tr}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Omega}_{11}.$$

Related to a selection of variables, we need to consider slight modifications of $H_{2\cdot 1}$, (i) or (ii) themselves. The modification of (i) may be defined by

$$M_{\{1,\ldots,p_1\}}: \ \boldsymbol{\beta}_{2j} = \mathbf{0}, \ \beta_{ij} \neq 0, \ i \in \{1, \ldots, p_1\}, \ j \in \{1, \ldots, q\}.$$

Let $\mathbf{B}$ and $\mathbf{W}$ be the matrices of sums of squares and products due to between-groups and within-groups, respectively, i.e.,

$$\mathbf{B} = \sum_{i=1}^{q+1} n_i(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}})^\top, \quad \mathbf{W} = \sum_{i=1}^{q+1}(\boldsymbol{y}_j^{(i)} - \overline{\boldsymbol{y}})(\boldsymbol{y}_j^{(i)} - \overline{\boldsymbol{y}})^\top.$$

The matrix $\mathbf{T} = \mathbf{B} + \mathbf{W}$ is called the matrix of sums of squares and products due to the total variation. We use the same partions for $\mathbf{B}, \mathbf{W}$ and $\mathbf{T}$ as in the partitions of $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$, etc. Then, GIC for $M_{2\cdot 1}$ is given (see, e.g., Fujikoshi et al. (2010)) as

$$\begin{aligned}\mathrm{GIC}_{\{1,\ldots,p_1\}} &= -n\log\{|\mathbf{W}_{22\cdot 1}|/|\mathbf{T}_{22\cdot 1}|\} + n\log|n^{-1}\mathbf{W}| \\ &\quad + np(1+\log 2\pi) + d\left\{qp_1 + p + \frac{1}{2}p(p+1)\right\}. \qquad (3.6)\end{aligned}$$

In general, let us denote the GIC for model $M_{\boldsymbol{j}}$ by $\mathrm{GIC}_{d,\boldsymbol{j}}$ or simply $\mathrm{GIC}_{\boldsymbol{j}}$. For expressing the KOO method based on GIC_d in (3.6) in a simple way, we use the following notation: For $\ell \in \boldsymbol{\omega}$, $\boldsymbol{\ell}$ is used for any subset of $p-1$ elements in $\boldsymbol{\omega}$ such that $\boldsymbol{\ell} = \{-\ell\}$. We arrange the elements of $\boldsymbol{y}$ as $\boldsymbol{y} = (\boldsymbol{y}_{\boldsymbol{\ell}}^\top, y_\ell)^\top$. Corresponding to this partition, we express the partitions of $\boldsymbol{\mu}^{(i)}$, $1 \leq i \leq q+1$, $\boldsymbol{\Sigma}$,$\mathbf{W}$ and $\mathbf{T} = \mathbf{W} + \mathbf{B}$ as follows:

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\ell}}^{(i)} \\ \mu_\ell^{(i)} \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\ell}\boldsymbol{\ell}} & \boldsymbol{\sigma}_{\boldsymbol{\ell}\ell} \\ \boldsymbol{\sigma}_{\boldsymbol{\ell}\ell}' & \sigma_{\ell\ell} \end{pmatrix}, \ \mathbf{W} = \begin{pmatrix} \mathbf{W}_{\boldsymbol{\ell}\boldsymbol{\ell}} & \boldsymbol{w}_{\boldsymbol{\ell}\ell} \\ \boldsymbol{w}_{\boldsymbol{\ell}\ell}' & w_{\ell\ell} \end{pmatrix}, \ \mathbf{T} = \begin{pmatrix} \mathbf{T}_{\boldsymbol{\ell}\boldsymbol{\ell}} & \boldsymbol{t}_{\boldsymbol{\ell}\ell} \\ \boldsymbol{t}_{\boldsymbol{\ell}\ell}' & t_{\ell\ell} \end{pmatrix}.$$

The sufficiency condition of $\boldsymbol{y}_{\boldsymbol{\ell}}$ may be written as $\mu_{\ell\cdot\boldsymbol{\ell}}^{(1)} = \cdots = \mu_{\ell\cdot\boldsymbol{\ell}}^{(q+1)}$, where $\mu_{\ell\cdot\boldsymbol{\ell}}^{(i)} = \mu_{\ell}^{(i)} - \boldsymbol{\sigma}_{\ell\ell}'\boldsymbol{\Sigma}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\ell}}^{(i)}$ $(1 \leq i \leq q+1)$. Then, we have

$$\mathrm{T}_{d,\ell} \equiv \mathrm{GIC}_{\boldsymbol{\ell}} - \mathrm{GIC}_{\boldsymbol{\omega}} = -n \log \frac{w_{\ell\ell\cdot\boldsymbol{\ell}}}{t_{\ell\ell\cdot\boldsymbol{\ell}}} - dq, \tag{3.7}$$

where $w_{\ell\ell\cdot\boldsymbol{\ell}} = w_{\ell\ell} - \boldsymbol{w}_{\ell\ell}'\mathbf{W}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\boldsymbol{w}_{\boldsymbol{\ell}\ell}$ and $t_{\ell\ell\cdot\boldsymbol{\ell}} = t_{\ell\ell} - \boldsymbol{t}_{\ell\boldsymbol{\ell}}'\mathbf{T}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1}\boldsymbol{t}_{\boldsymbol{\ell}\ell}$. The KOO method based on $\mathrm{T}_{d,\ell}$ is to select

$$\widehat{\boldsymbol{j}}_d = \{j \in \boldsymbol{\omega} \mid \mathrm{T}_{d,j} > 0\}, \quad \mathrm{T}_{d,j} \text{ given by (3.7)}. \tag{3.8}$$

The following results on consistency of $\widehat{\boldsymbol{j}}_d$ are mainly based on Oda et al. (2020). For a distributional reduction of $\mathrm{T}_{d,\ell}$, we use the following Lemma.

Lemma 3.2. *When dealing with $w_{\ell\ell\cdot\boldsymbol{\ell}}/t_{\ell\ell\cdot\boldsymbol{\ell}}$, we may assume without loss of generality that $\mathbf{W}$ and $\mathbf{B}$ are independent and $\mathbf{W} \sim W_p(n-q-1, \mathbf{I}_p)$ and $\mathbf{B} \sim W_p(q, \mathbf{I}_p; \boldsymbol{\Psi})$, where*

$$\boldsymbol{\Psi} = n\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\top} = \boldsymbol{\Theta}\boldsymbol{\Theta}^{\top}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1/2} & \mathbf{0}_{p-1} \\ \sigma_{\ell\ell\cdot\boldsymbol{\ell}}^{-1/2}\boldsymbol{\sigma}_{\ell\ell}'\boldsymbol{\Sigma}_{\boldsymbol{\ell}\boldsymbol{\ell}}^{-1} & \sigma_{\ell\ell\cdot\boldsymbol{\ell}}^{-1/2} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \Theta_{\boldsymbol{\ell}} \\ \boldsymbol{\theta}_{\ell}^{\top} \end{pmatrix},$$

and $\boldsymbol{\Theta}_{\boldsymbol{\ell}}; (p-1) \times q$, $\boldsymbol{\theta}_{\ell}; q \times 1$. We can express $w_{\ell\ell\cdot\boldsymbol{\ell}}/t_{\ell\ell\cdot\boldsymbol{\ell}}$ and $\mathrm{T}_{d,\ell}$ as

$$\begin{aligned} \frac{w_{\ell\ell\cdot\boldsymbol{\ell}}}{t_{\ell\ell\cdot\ell}} &= \frac{w_{\ell\ell\cdot\boldsymbol{\ell}}}{w_{\ell\ell\cdot\boldsymbol{\ell}} + (t_{\ell\ell\cdot\boldsymbol{\ell}} - w_{\ell\ell\cdot\boldsymbol{\ell}})} = \frac{s_e}{s_e + s_h}, \\ \mathrm{T}_{d,\ell} &= n \log\left(1 + s_h s_e^{-1}\right) - dq, \end{aligned}$$

where s_e and s_h are conditionally independent given $\mathbf{U}$ and $\mathbf{Z}$, and

$$s_e \sim \chi^2(n-p-q), \quad s_h \mid \mathbf{U}, \mathbf{Z} \sim \chi^2(q; \gamma_{\ell}^2).$$

Here,

$$\gamma_{\boldsymbol{\ell}}^2 = \boldsymbol{\theta}_{\ell}^{\top}\{\mathbf{I}_q + \mathbf{Z}^{\top}(\mathbf{U}^{\top}\mathbf{U})^{-1}\mathbf{Z}\}^{-1}\boldsymbol{\theta}_{\ell},$$

where $\mathbf{U}; p-1 \times (n-q-1)$ and $\mathbf{Z}; q \times (p-1)$ are independent random matrices whose elements are independent normal variabes with the same variance 1, $\mathrm{E}(\mathbf{U}) = \mathbf{O}$, and $\mathrm{E}(\boldsymbol{Z}) = \Theta_{\boldsymbol{\ell}}$.

In the special case of $q = 1$, $s_e \sim \chi^2(n-p-1)$, $s_h \sim \chi^2(1;\gamma_\ell^2)$, $\gamma_\ell^2 = g^2\Delta_{\ell\cdot\boldsymbol{\ell}}^2(1+R)^{-1}$, $R = \chi^2(p-1;g^2\Delta_{\boldsymbol{\ell}}^2)/\chi^2(n-p)$. Here, we use that

$$\boldsymbol{\theta}_\ell^\top\boldsymbol{\theta}_\ell = g^2\boldsymbol{\Delta}_{\ell\cdot\boldsymbol{\ell}}^2, \quad \mathrm{tr}\Theta_{\boldsymbol{\ell}}^\top\Theta_{\boldsymbol{\ell}} = g^2\Delta_{\boldsymbol{\ell}}^2.$$

These distributional results are coincident with the results in Lemma 3.1.

Note that

$$\mathrm{T}_{d,\ell} > 0 \quad \Leftrightarrow \quad \frac{\chi^2(q;\gamma_\ell^2)}{\chi^2(n-p-q)} > \mathrm{e}^{dq/n} - 1,$$

and when $\ell \notin \boldsymbol{j}_*$, $\gamma_{\boldsymbol{\ell}}^2 = 0$. In this section we use the penalty term β given by

$$\beta = \mathrm{e}^{dq/n} - 1 - \frac{q}{n-p-q-2}, \tag{3.9}$$

instead of d, assuming $\beta > 0$.

For the multiple-group case, assumptions A2, A3 and A4 are generalized as follows:

$\widetilde{\mathrm{A}}2$: $p \to \infty$, $n \to \infty$, $p/n \to c \in (0,1)$, $n_i/n \to k_i > 0$, $i \in \{1,\ldots,q+1\}$;

$\widetilde{\mathrm{A}}3$: The number p_* of true variables is finite, and $\lim n^{-1}\mathrm{tr}\boldsymbol{\Theta}^\top\boldsymbol{\Theta} > 0$;

$\widetilde{\mathrm{A}}4$: For $\ell \in \boldsymbol{j}_*$, $\lim n^{-1}\boldsymbol{\theta}_\ell^\top\boldsymbol{\theta}_\ell > 0$, and $\lim n^{-1}\mathrm{tr}\boldsymbol{\Theta}_{\boldsymbol{\ell}}^\top\boldsymbol{\Theta}_{\boldsymbol{\ell}} > 0$.

The assumption B2 is changed to

$\widetilde{\mathrm{B}}2$: $h_q \equiv q\{dn^{-1} - (n-p-q-2)^{-1}\} > 0$ and $h_q = \mathrm{O}(n^{-a})$, for $0 < a < 1$.

Theorem 3.2. *Let* [F1] *and* [F2] *be the quantities defined in* (2.5) *and* (2.6) *with* $\mathrm{T}_{d,j}$ *in* (3.7). *Suppose that assumptions* A1 *and* A2 *hold. Then, for the* KOO *method* (3.8)*, the following asymptotic results hold.*
(i) *Suppose that* $\widetilde{\mathrm{A}}3$ *and* $\widetilde{\mathrm{A}}4$ *hold. Then,* $[\mathrm{F1}] \to 0$*;*

(ii) *Suppose that* B1 *and* $\widetilde{\text{B}}2$ *hold. Then,* [F2] $\to 0$*;*
(iii) *Suppose that the assumptions in* (i) *and* (ii) *hold. Then, the* KOO *method* (3.8) *is asymptotically consistent.*

Proof of Theorem 3.2: First, we consider the case of $\ell \notin \boldsymbol{j}_*$. Then, $\boldsymbol{\theta}_\ell = \mathbf{0}$, and s_h is distributed according to $\chi^2(q)$, and we have

$$\begin{aligned}
[\text{F2}] &= (p-p_*)\Pr(U_q > \mathrm{e}^{(dq)/n} - 1 - q(n-p-q-2)^{-1}) \\
&\leq (p-p_*)\Pr(U_q \geq h_q) \\
&\leq (p-p_*)h_q^{-2\ell}\mathrm{E}(U_q^{2\ell}), \quad \ell \in \{1,2,\ldots,\},
\end{aligned}$$

where $U_q = \chi^2(q)/\chi^2(n-p-q) - q/(n-p-q-2)$. Noting $h_q = \mathrm{O}(n^{-a})$ and $\mathrm{E}[U_q^{2\ell}] = \mathrm{O}(n^{-2\ell})$ (see, e.g., Theorem 16.2.2 in Fujikoshi et al. [**?**]), we have

$$[\text{F2}] \leq \mathrm{O}(n^{1-2\ell(1-a)}).$$

Therefore, choosing $\ell > 1/\{2(1-a)\}$, we have "[F2] $\to 0$". Next, we consider the case of $\ell \in \boldsymbol{j}_*$. Note that [F1] $= \sum_{\ell \in \boldsymbol{j}_*} \Pr(\widetilde{\mathrm{T}}_{d,\ell} \leq 0)$, where $\widetilde{\mathrm{T}}_{d,\ell} = (1/n)\mathrm{T}_{d,\ell}$. In this case, we assume that $k_{\boldsymbol{j}_*}$ is finite. To show that [F1] $\to 0$, it is sufficient to show that there exists a positive number $\widetilde{t}_\ell$ such that $\lim \widetilde{\mathrm{T}}_{d,\ell} \geq \widetilde{t}_\ell$, or $\lim \widetilde{\mathrm{T}}_{d,j} \to \infty$. Note that

$$\widetilde{\mathrm{T}}_{d,\ell} \approx \log\left(1 + \frac{\gamma_\ell^2}{n-p-q}\right),$$

and $\gamma_\ell^2 \geq \boldsymbol{\theta}_\ell^\top \boldsymbol{\theta}_\ell \left[1 + \mathrm{tr}\left\{\mathbf{Z}^\top(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{Z}\right\}\right]^{-1}$. Further, we can write

$$\mathrm{tr}\left\{\mathbf{Z}^\top(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{Z}\right\} = \mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^\top),$$

where $\mathbf{V} = (\mathbf{Z}^\top\mathbf{Z})^{1/2}\left\{\mathbf{Z}^\top(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{Z}\right\}^{-1}(\mathbf{Z}^\top\mathbf{Z})^{1/2}$. Since $\mathbf{V} \sim \mathrm{W}_q(n-p, \mathbf{I}_q)$ (see, e.g., Fujikoshi et al. [**?**], Theorem 2.3.3), $\mathbf{V} \approx (n-p)\mathbf{I}_q$. Similarly, $\mathbf{Z}^\top\mathbf{Z} \sim \mathrm{W}_q(p-1, \mathbf{I}_q; \boldsymbol{\Theta}_{\boldsymbol{\ell}}^\top\boldsymbol{\Theta}_{\boldsymbol{\ell}})$, $\mathbf{Z}^\top\mathbf{Z} \approx (p-1)\mathbf{I}_q + \boldsymbol{\Theta}_{\boldsymbol{\ell}}^\top\boldsymbol{\Theta}_{\boldsymbol{\ell}}$. From these results we can get the required result. For more details, see Outline of Proof of Theorem 3.3 and also Oda et al. (2020). □

In the following, we give some further results due to Oda et al. (2020). Without loss of generality, the penalty term d may be rewritten as

$$d = nq^{-1}\log(1+b), \text{ or } b = \mathrm{e}^{dq/n} - 1, \tag{3.10}$$

assuming $b > 0$. Then, we have

$$\begin{aligned}[\mathrm{F2}] \equiv \sum_{\ell \notin \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,\ell} \geq 0) &= (p-p_*)\Pr\left(\frac{s_h}{s_e} > b\right) \\ &\leq (p-p_*)b^{-2r}\mathrm{E}\left[\left\{\frac{\chi^2(q)}{\chi^2(n-p-q)}\right\}^{2r}\right]. \end{aligned} \tag{3.11}$$

The above last inequality is derived by the Markov inequality. The order of the expectation in the above last equation is $\mathrm{O}(n^{-2r})$ (see, e.g., Theorem 16.2.2 in Fujikoshi et al. (2010)). Now we make the following assumption for the penalty term b:

B3: $p^{-1/(2r)}nb \to \infty$, for some positive integer r, and $b \to 0$.

Therefore, from B3, we have

$$\sum_{\ell \notin \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,\ell}0) \leq \mathrm{O}(pn^{-2r}b^{-2r}) \to 0. \tag{3.12}$$

Here, we note that B3 is equivalent to

$\widetilde{\mathrm{B}}3$: $p^{-1/(2r)}d \to \infty$, for some positive integer r, and $d/n \to 0$.

Relating to evaluation of "[F1] $= \sum_{\ell \in \boldsymbol{j}_*} \Pr(T_{d,\ell} \leq 0)$", let us denote a compliment of $\{\ell\}$ with respect to $\boldsymbol{\omega}$ by $\bar{\ell}$ or $\boldsymbol{\ell}$. Then, the number of elements in $\sharp(\bar{\ell}) = \sharp(\boldsymbol{\ell})$ is $p-1$. Let $\delta^2 = \mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$, which is an information quantity in discriminant analysis. This quantity for $\boldsymbol{y}_{\boldsymbol{j}}$ is denoted by $\delta^2_{\boldsymbol{j}}$, which is $\mathrm{tr}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{jj}}\boldsymbol{\Omega}_{\boldsymbol{jj}}$. Instead of $\widetilde{\mathrm{A}}3$ and $\widetilde{\mathrm{A}}4$, we consider the following assumptions.

C1: For all $\boldsymbol{\ell} \subset \boldsymbol{\omega}$ such that $\sharp(\boldsymbol{\ell}) = p - 1$ and $\ell = \overline{\boldsymbol{\ell}} \in \boldsymbol{j}_*$, there exists $c_1 > 0$ such that for all n and p, $\delta^2 - \delta_{\boldsymbol{\ell}}^2 > c_1$.

C2: There exists $c_2 >$ such that for all p, $\lambda_{\min}(\boldsymbol{\Sigma}) > c_2 > 0$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ is the minimum characteristic root of $\boldsymbol{\Sigma}$.

The following result was given by Oda et al. [**?**].

Theorem 3.3. *Suppose that assumptions* A1*,* A2*,* B3*,* C1 *and* C2 *hold. Then, the* KOO *method* (3.8) *based on* GIC *is consistent if the number* p_* *of true variables is finite.*

Outline of Proof of Theorem 3.3: The result can be shown by proving (i); [F1] $\to 0$ and (ii); [F2] $\to 0$. The result (ii) has been pointed out by (3.11) and (3.12). For a proof of (i), first we can see that assumptions $\widetilde{\mathrm{C}}1$ and $\widetilde{\mathrm{C}}2$ imply the assumptions $\widetilde{\mathrm{A}}3$ and $\widetilde{\mathrm{A}}4$. Therefore, the result (i) will follow from the proof of Theorem 3.2. On the other hand, instead of using the proof of Theorem 3.2, we can prove by using

$$[\mathrm{F1}] = \sum_{\ell \in \boldsymbol{j}_*} \Pr(\mathrm{T}_{d,\ell} \leq 0) = \sum_{\ell \in \boldsymbol{j}_*} \Pr\left(\mathrm{L}_\ell \leq 0\right),$$

where

$$\mathrm{L}_\ell = \frac{s_h}{s_e} - \frac{q}{n-p-q-2} - \beta, \quad \beta = b - q(n-p-q-2)^{-1}.$$

It is assumed that $\beta > 0$. Here, s_h and s_e depend on ℓ, but we may assume that $s_e \sim \chi^2(n-p-q)$, and $s_h \sim \chi^2(q; \gamma_\ell^2)$ when γ_ℓ^2 is given in Lemma 3.2. The result (ii) will be shown by noting that L_ℓ converges to a positive value. □

As in the case $q = 1$, it has been pointed out that the KOO method with $d = \sqrt{n}$ has numerically good behavior.

4. Conclusions

In this paper we overview of high-dimensional consistency of KOO methods based on general information criteria in multivariate regression model and discriminant analysis. Consistency of KOO method based on some other model section criteria are also examined in multivariate regression model. Our results have been restricted in the case of normality and weak consistency. For some results in the case of non-normality and strong consistency have been discussed in Bai et al. (2018) in multivariate regression model. However, since their results are under revision, we have not discussed here. The consistency results in this paper are obtained by prooving two properties [F1] and [F2] in (2.5) and (2.6). Sufficient conditions for each of [F1] and [F2] were given. In a high-dimensional asymptotic framework in discriminant analysis, we assume that the number of groups, $q+1$, is fixed. However, it would also be important to consider the case of q being large.

Some high-dimensional consistency results on information criteria of selection of variables in some other multivariate models have been reported. For examples, see Oda et al. (2019) for selection of variables in canonical correlation analysis, Fujikoshi et al. (2013) and Enomoto (2019) for selection of within individual variables in growth curve model, Oda et al. (2019) for selection of response variables in multivariate calibration, and Oda et al. (2021) for selection of response variables in multivariate regression model. Enomoto (2019) has dicussed consistency of KOO method in growth curve model, assuming that the covariance matrix has a uniform structure. It is expected that high-dimensional properties of KOO methods in these models shall be further studied.

Acknowledgements

The author would like to express his gratitute to Professor Dietrich von Rosen, Professor Tõnu Kollo, and Dr. Tetsuro Sakurai for their valuable comments and suggestions.

References

[1] Bai, Z., Fujikoshi, Y. and Hu, J. (2018). Strong consistency of the AIC, BIC, Cp and KOO methods in high-dimensional multivariate linear regression. *Hiroshima Statistical Research Group*, TR; 18-09.

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

[3] Enomoto, R. (2019). Model selection criteria in growth curve model with a uniform covariance. Reported in the 2019 Statistical Meeting at Chuo University.

[4] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p statistic in multivariate regression models. *Biometrika*, **84**, 707-716.

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

[6] Fujikoshi, Y., Kan, T. Takahashi, S. and Sakurai,T. (2011). Prediction error citerion for selecting variables in a linear regression model. *Ann. Inst. Statist. Math.*, **63**, 387-403.

Fujikoshi, Y., Enomoto, R. and Sakurai, T. (2013). High-dimensional *AIC* in the growth curve mode. *Journal of Multivariate Analysis*, **122**, 239-250.

[7] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional *AIC*-type and C_p-typ criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184–200.

[8] Fujikoshi, Y. and Sakurai, T. (2019). Consistency of test-based method for selection of variables in high-dimensional two group-discriminant analysis. *Japanese Journal of Statistics and Data Science*, **2**, 155–171.

[9] Hao, N. and Bin, B. (2015). Sparcifying the Fisher linear discriminant by rotation. *Journal of the Royal Statistical Society: Series B*, **77**, 827–851.

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

[11] Oda, R. (2020). Consistent variable selection criteria in multivariate linea regression even when dimension exceeds sample size. *Hiroshima Math. J.*, **59**, 339–374.

[12] Oda, R., Suzuki, Y., Yanagihara, H. and Fujikoshi, Y. (2020). A consistent variable selection method in high-dimensional canonical discriminant analysis. *J. Multivariate Anal.*, **175**, 1–13.

[13] Oda, R., Yanagihara, H. and Fujikoshi, Y. (2019). Strong consistency of log-likelihood-based information criterion in high-dimensional canonical correlation analysis. *Sankhya*. **A-83**, 109–127.

[14] Oda, R., and Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron J. Statist.*, **14**, 1386–1412.

[15] Oda, R., and Yanagihara, H. (2021). A consistent likelihood-based variable selection method in normal multivariate linear regression. *Intelligent Decision Technologies*, I. Czarnowski et al. (eds.), **238**, 391–401,

[16] Oda, R., Yanagihara, H. and Fujikoshi, Y. (2021). On model selection consistency of kick-one-out method for selecting response variables in high-dimensional multivariate linear regression. TR No.21-06, *Hiroshima Statistical Research Group*, Hiroshima University.

[17] Rao, C. R. (1970). Inference on discriminant function coefficients. In *Essays in Prob. and Statist.* (R. C. Bose et al., eds.), 587–602.

[18] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*(2nd ed.). John Wiley & Sons, New York.

[19] Sakurai, T. and Fujikoshi, Y. (2020). Exploring consistencies of information criterion and test-based criterion for high-dimensional multivariate regression models under three covariance structures. In Festschrift in honor of Professr Dietrich von Rosen's 65th birthday (eds, T. Holgerson and M. Singnull). Springer.

[20] Sparks, R. S., Coutsourides, D. and Troskie, L. (1993). The multivariate Cp. *Communications in Statistics - Theory and Methods*, **12**, 1775-1793.

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

[22] 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. *Electronic Journal of Statistics*, **9**, 869–897.

[23] Yanagihara, H. (2016). A high-dimensionality-adjusted consistent C_p-type for selecting variables in a normality-assumed linear regression with multiple responses. *Procedia Comput. Sci.*, **96**, 1096–1105.

[24] Zhao, L. C. , Krishnaiah, P. R. and Bai, Z. D. (1986). On determination of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.