



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

Yasunori Fujikoshi

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

## ARTICLE INFO

### Article history:

Received 18 October 2021

Accepted 18 October 2021

Available online 2 November 2021

### AMS 2020 subject classifications:

primary 62H12

secondary 62H30

### Keywords:

Discriminant analysis

General information criteria

High-dimensional consistency

KOO method

Multivariate regression model

Selection of variables

## 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 take up also KOO methods based on some other selection methods. Some references are given for high-dimensional consistencies in some other multivariate models.

© 2021 Published by Elsevier Inc.

## 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

$$\text{GIC} = -2 \log L(\hat{\theta}) + dg,$$

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

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

$$\tilde{\mathbf{j}}_G = \arg \min_{\mathbf{j}} \text{GIC}_{\mathbf{j}}.$$

Here the minimum is usually taken over all subsets. It has been pointed out that there are computational problems for GIC methods, for 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. [11] proposed a method which is essentially due to Zhao et al. [26]. The method, which was named the knock-one-out (KOO) method by Bai et al. [1], determines “selection” or “no selection”

E-mail address: [fujikoshi\\_y@yahoo.co.jp](mailto:fujikoshi_y@yahoo.co.jp).

for each variable by comparing the model after removing a variable and the full model. More precisely, the KOO method chooses the model or the set of variables given by

$$\hat{\mathbf{j}}_G = \{j \in \omega \mid \text{GIC}_{\omega \setminus j} > \text{GIC}_\omega\},$$

where  $\omega \setminus j$  is the set obtained by removing element  $j$  from the set  $\omega$ .

In a large-sample setting, Nishii et al. [11] studied strong consistency of  $\hat{\mathbf{j}}_G$  and  $\tilde{\mathbf{j}}_G$  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. [7], Yanagihara et al. [25], and Bai et al. [1], 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, we consider also the ones based on some other selection methods. Some results in multivariate regression models have been studied for strong consistency under nonnormality by Bai et al. [1]. However, since their results are under revision, in the present paper we discuss only 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 variables in some other models.

## 2. Multivariate regression model

### 2.1. KOO methods based on GIC

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

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

where  $\Theta$  is a  $k \times p$  regression coefficient matrix, and  $\mathcal{E} = (\epsilon_1, \dots, \epsilon_n)^\top$  is the error matrix. It is assumed that the  $\epsilon_i$ 's are independently distributed as a  $p$ -variate normal distribution  $N_p(\mathbf{0}, \Sigma)$ . 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  $i$ th row  $\theta_i$  of  $\Theta = (\theta_1, \dots, \theta_k)^\top$  is the null vector. We consider such a selection problem, assuming that  $\Sigma$  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  $\{x_1, \dots, x_k\}$  with  $\omega \equiv \{1, \dots, k\}$ , and let  $k_j$  be the cardinality of  $\mathbf{j}$ . Further, let  $\mathbf{j}$  be a subset of  $\omega$ ,  $\mathbf{x}_j = (x_j, j \in \mathbf{j})^\top$ , and  $\mathbf{X}_j = (\mathbf{x}_j, j \in \mathbf{j})$ . Denote the model based on  $\mathbf{x}_j$  by

$$M_j : \mathbf{Y} = \mathbf{X}_j\Theta_j + \mathcal{E}.$$

Then, GIC is expressed as

$$\text{G}_{d,j} = n \log |\hat{\Sigma}_j| + d \{k_j p + (p(p+1)/2)\} + np \{\log(2\pi) + 1\}. \quad (1)$$

Here,  $\hat{\Sigma}_j$  is the MLE of  $\Sigma$  given by

$$n\hat{\Sigma}_j = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{P}_j) \mathbf{Y}, \quad \mathbf{P}_j = \mathbf{X}_j (\mathbf{X}_j^\top \mathbf{X}_j)^{-1} \mathbf{X}_j^\top,$$

where  $\mathbf{P}_j$  is an orthogonal projection matrix onto the space spanned by the column vectors of  $\mathbf{X}_j$ . It is assumed that the rank of  $\mathbf{X}_j$  is  $k_j$ , and further, it is assumed that  $n - k \geq p$  when we use (1). Using

$$T_{d,j} = \text{G}_{d,\omega \setminus j} - \text{G}_{d,\omega} = n \log(|\hat{\Sigma}_{\omega \setminus j}|/|\hat{\Sigma}_\omega|) - dp, \quad (2)$$

the KOO method chooses the model

$$\hat{\mathbf{j}}_d = \{j \in \omega \mid T_{d,j} > 0\}. \quad (3)$$

Recently, high-dimensional consistency properties of KOO methods have been studied by Bai et al. [1], Sakurai and Fujikoshi [21] and Oda and Yanagihara [15,16]. These considered KOO methods are based on the  $C_p$  criterion, as well as GIC. Bai et al. [1] studied strong consistency properties under nonnormality. Sakurai and Fujikoshi [21] 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 when the number of explanatory variables  $k$  is fixed.

## 2.2. Consistency of KOO methods based on GIC

First we consider the distributional reduction of  $T_{d,j}$  in (2). Note that the first term of  $T_{d,j}$  is  $-2 \log \lambda_j$ , where  $\lambda_j$  is the likelihood ratio criterion for testing the hypothesis  $\theta_j = \mathbf{0}$ . More precisely,

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

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

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

where  $\boldsymbol{\gamma}_j = \theta_j \mathbf{x}_j^\top \tilde{\mathbf{x}}_j (\tilde{\mathbf{x}}_j^\top \tilde{\mathbf{x}}_j)^{-1/2}$  and  $\tilde{\mathbf{x}}_j = (\mathbf{I}_n - \mathbf{P}_{\omega \setminus j}) \mathbf{x}_j$ . In this case, we may write  $\mathbf{S}_{h,j} = \mathbf{u}_j \mathbf{u}_j^\top$ , where  $\mathbf{u}_j \sim N_p(\boldsymbol{\gamma}_j, \Sigma)$ . Then, using  $|\mathbf{S}_e + \mathbf{u}_j \mathbf{u}_j^\top| = |\mathbf{S}_e|(1 + \mathbf{u}_j^\top \mathbf{S}_e^{-1} \mathbf{u}_j)$  and a well-known result (see, e.g., Fujikoshi et al. [9], Theorem 3.1.1), we have

$$T_{d,j} = n \log \left\{ 1 + \frac{\chi_j^2(p; \delta_j^2)}{\chi^2(m)} \right\} - dp, \quad (4)$$

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  $\chi^2$  distributed random variable with  $p$  degree of freedom and noncentrality  $\delta_j^2 = \boldsymbol{\gamma}_j^\top \Sigma^{-1} \boldsymbol{\gamma}_j$ , 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 (3) in terms of  $T_{d,j}$ . Consistency here will be proved by showing the following two properties:

$$[\text{F1}] \equiv \sum_{j \in \mathbf{j}_*} \Pr(T_{d,j} \leq 0) \rightarrow 0. \quad (5)$$

$$[\text{F2}] \equiv \sum_{j \notin \mathbf{j}_*} \Pr(T_{d,j} \geq 0) \rightarrow 0. \quad (6)$$

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

$$\begin{aligned} \Pr(\widehat{\mathbf{J}}_G = \mathbf{j}_*) &= \Pr \left( \bigcap_{j \in \mathbf{j}_*} "T_{d,j} > 0" \bigcap_{j \notin \mathbf{j}_*} "T_{d,j} < 0" \right) = 1 - \Pr \left( \bigcup_{j \in \mathbf{j}_*} "T_{d,j} \leq 0" \bigcup_{j \notin \mathbf{j}_*} "T_{d,j} \geq 0" \right) \\ &\geq 1 - \sum_{j \in \mathbf{j}_*} \Pr(T_{d,j} \leq 0) - \sum_{j \notin \mathbf{j}_*} \Pr(T_{d,j} \geq 0). \end{aligned}$$

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  $\mathbf{j}_*$  is included in the full set  $\omega$ , i.e.,  $\mathbf{j}_* \subset \omega$ .

P2: A high-dimensional asymptotic framework holds:

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

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

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

$$\delta_j^2 = \theta_j^\top \Sigma^{-1} \theta_j (\mathbf{x}_j^\top \tilde{\mathbf{x}}_j)^2 (\tilde{\mathbf{x}}_j^\top \tilde{\mathbf{x}}_j)^{-1}.$$

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

When  $j \notin \mathbf{j}_*$ , from (4) we can write  $T_{d,j} = n \log \{1 + \chi_j^2(p)/\chi^2(m)\} - dp$ , and therefore we have

$$[\text{F2}] = \sum_{j \notin \mathbf{j}_*} \Pr(n \log \{1 + \chi_j^2(p)/\chi^2(m)\} \geq dp) = (k - k_{j_*}) \Pr(U \geq h) \leq (k - k_{j_*}) \Pr(U \geq h_0),$$

where

$$U = \frac{\chi^2(p)}{\chi^2(m)} - \frac{p}{m-2}, \quad h = e^{dp/n} - 1 - \frac{p}{m-2}, \quad h_0 = \frac{dp}{n} - \frac{p}{m-2}. \quad (7)$$

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

$$[\text{F2}] \leq (k - k_{j_*}) h^{-2\ell} E[U^{2\ell}] \leq (k - k_{j_*}) h_0^{-2\ell} E[U^{2\ell}].$$

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

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

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

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

$$\tilde{U}_j = \frac{\chi_j^2(p; \delta_j^2)}{\chi^2(m)} - \frac{p + \delta_j^2}{m - 2}, \quad \tilde{h}_j = e^{dp/n} - 1 - \frac{p + \delta_j^2}{m - 2} = h - \frac{\delta_j^2}{m - 2}. \quad (8)$$

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

$$[F1] \leq k_{j_*} |\tilde{h}_j|^{-2\ell} E[\tilde{U}^{2\ell}].$$

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

**Theorem 1.** Suppose that assumptions P1 and P2 hold. Then, for the KOO method (3) based on GIC, the following asymptotic results hold. (i) Suppose that the quantity  $h_0$  in (7) is asymptotically positive or  $\infty$ , and  $h_0^{-4} \leq O(n^a)$  for some  $a < 1$ . Then,  $[F2] \rightarrow 0$ . (ii) Suppose that P3 holds, the quantity  $\tilde{h}_0$  in (8) is asymptotically negative, and  $|\tilde{h}_j|^{-4} = O(n^a)$  for some  $a < 1$ . Then,  $[F1] \rightarrow 0$ . (iii) Suppose that the assumptions in (i) and (ii) hold. Then, the KOO method (3) based on GIC is asymptotically 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 1 for AIC is satisfied if  $2 - 1/(1 - c_1 - c_2) > 0$ .

Bai et al. [1] have studied high-dimensional strong consistency of KOO methods based on AIC, BIC and  $C_p$  by examining their strong convergences. Further, Oda and Yanagihara [16] considered a transformation from  $d$  to  $b$  given by

$$d = \frac{n}{p} \log \left\{ 1 + \frac{p}{m - 2} (1 + b) \right\}, \quad (9)$$

with  $b > 0$ . Then, writing  $\rho = (pb)/(m - 2)$ , we have  $h = \rho$ ,  $\tilde{h}_j = \rho - \delta_j^2/(m - 2)$ . Using these expressions, we can write

$$[F2] = (k - k_{j_*}) \Pr(U \geq \rho) \leq k \rho^{-2\ell} E(U^{2\ell}), \quad (10)$$

$$[F1] = \sum_{j \in \mathbf{j}_*} \Pr \left( \tilde{U}_j \leq \rho - \frac{\delta_j^2}{m - 2} \right). \quad (11)$$

Instead of the assumption P3, we use more precise conditions P4 and P5 on the noncentrality parameter as the one in terms of  $\mathbf{x}_j^\top (\mathbf{I}_n - \mathbf{P}_{\omega \setminus j}) \mathbf{x}_j$  and  $\boldsymbol{\theta}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}_j$ ,  $j \in \mathbf{j}_*$ .

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

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

Evaluating (10) and (11), Oda and Yanagihara [16] described a class of  $b$  having a high-dimensional consistency as in the following theorem.

**Theorem 2.** Suppose that assumptions P1, P4 and P5 hold. Then, the KOO method (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:

$$\sqrt{pb}/k^{1/(2r)} \rightarrow \infty, \quad pb/n^{c_3} \rightarrow 0. \quad (12)$$

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

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

as an example of  $d$  satisfying (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 [8], Sparks et al. [22]). Its generalization may be defined (see, e.g., Yanagihara [24]) by

$$GC_{d,j} = (n - k) \text{tr} \hat{\boldsymbol{\Sigma}}_j \hat{\boldsymbol{\Sigma}}_\omega^{-1} + dpk_j.$$

The  $GC_{2,j}$ , with  $d = 2$ , is the usual  $C_p$  criterion. The KOO method based on  $GC_{d,j}$  is to select  $j$  such that  $T_{d,j;GC}$  is positive, where

$$T_{d,j;GC} = GC_{d,\omega \setminus j} - GC_{d,\omega}.$$

Using the same notation as for GIC, we can express

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

Therefore we have

$$\sum_{j \notin J_*} \Pr(T_{d,j;GC} \geq 0) = \sum_{j \notin J_*} \Pr((n-k) \chi_j^2(p) \{ \chi^2(m) \}^{-1} \geq dp) = (k - k_{j_*}) \Pr(U \geq \tilde{h}_0),$$

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

$$\sum_{j \in J_*} \Pr(T_{d,j;GC} \leq 0) = \sum_{j \in J_*} \Pr((n-k) \chi_j^2(p; \delta_j^2) \{ \chi^2(m) \}^{-1} \geq dp) = \sum_{j \in J_*} \Pr(\tilde{U}_j \geq \tilde{h}_j),$$

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

From the above reduction, we can get a result similar to [Theorem 1](#). Oda and Yanagihara [14] 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

$$GP_{d,j} = (n-k) \frac{\text{tr} \hat{\Sigma}_j}{\text{tr} \hat{\Sigma}_\omega} + dk_j.$$

The criterion in a special case  $d = 2$  was proposed by Fujikoshi et al. [5]. For its generalization, see Oda [12].

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

$$\hat{J}_{d,GP} = \{j \in \omega \mid T_{d,j,GP} > 0\},$$

where

$$T_{d,j,GP} = GP_{d,\omega \setminus j} - GP_{d,\omega} = (n-k) \frac{\text{tr} \mathbf{S}_{h,j}}{\text{tr} \mathbf{S}_\omega} - d.$$

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

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

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

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

$$(1/p) \text{tr} \Sigma \rightarrow O(1), \quad (1/p) \text{tr} \Sigma^2 \rightarrow O(1).$$

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

$$(1/p) \text{tr} \Sigma \Omega \rightarrow O(1), \quad (1/p) \text{tr} \Sigma^2 \Omega \rightarrow O(1).$$

### 3. Discriminant analysis

#### 3.1. Two-group case

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

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

The coefficients of the population discriminant function are given by

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

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

$$D = \{(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)})^\top \mathbf{S}^{-1}(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)})\}^{1/2},$$

respectively. Here,  $\bar{\mathbf{y}}^{(1)}$  and  $\bar{\mathbf{y}}^{(2)}$  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,  $\omega$  is used for  $\{1, \dots, p\}$ , and we let  $\mathbf{j}$  denote a subset of  $\omega$  containing  $p_j$  elements, and  $\mathbf{y}_j$  denote the  $p_j$  vector consisting of the elements of  $\mathbf{y}$ , indexed by the elements of  $\mathbf{j}$ . We use the notation  $D_j$  and  $D_\omega$  for  $D$  based on  $\mathbf{y}_j$  and  $\mathbf{y}_\omega (= \mathbf{y})$ , respectively. Let  $M_j$  be the variable selection model, defined by

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

Model  $M_j$  implies that  $\Delta_j = \Delta_\omega$ , i.e., that the Mahalanobis distance based on  $\mathbf{y}_j$  is the same as the one based on the full set of variables,  $\mathbf{y}$ . We identify the selection of  $M_j$  with the selection of  $\mathbf{y}_j$ . Let  $\text{GIC}_j$  be the GIC for  $M_j$  in two-group discriminant analysis, and  $T_{d,j} = \text{GIC}_{\omega \setminus j} - \text{GIC}_\omega$ . Then, we have (see, e.g., Fujikoshi and Sakurai [6]) that

$$T_{d,j} = n \log \left\{ 1 + \frac{g^2(D_\omega^2 - D_{\omega \setminus j}^2)}{n - 2 + g^2 D_{\omega \setminus j}^2} \right\} - d, \quad (13)$$

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

$$\hat{\mathbf{j}}_d = \{j \in \omega \mid T_{d,j} > 0\}. \quad (T_{d,j} \text{ is provided in (13)}). \quad (14)$$

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

$$\Delta^2 = \Delta_{\{-j\}}^2 + \Delta_{\{j\}|\{-j\}}^2, \quad (15)$$

and hence  $\Delta_{\{j\}|\{-j\}}^2 = \Delta^2 - \Delta_{\{-j\}}^2$ .

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. [2], Witten and Tibshirani [23], and Hao and Dong [10].

For a distributional reduction of  $T_{d,j}$  in (13), we use the following Lemma (see, e.g., Fujikoshi et al. [9]).

**Lemma 1.** Let  $D_1$  and  $D$  be the sample Mahalanobis distances based on  $\mathbf{y}_1; p_1 \times 1$  and  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top)^\top$ , respectively, and 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

- (i)  $D_1^2 = (n - 2)g^{-2}R$ ,  $R = \chi^2(p_1; g^2 \Delta_1^2) \{\chi^2(n - p_1 - 1)\}^{-1}$ ;
- (ii)  $D_{2,1}^2 = (n - 2)g^{-2}\chi^2(p_2; g^2 \Delta_{2,1}^2(1 + R)^{-1}) \{\chi^2(n - p - 1)\}^{-1}(1 + R)$ ;
- (iii)  $g^2(D^2 - D_1^2) \{n - 2 + g^2 D_1^2\}^{-1} = \chi^2(p_2; g^2 \Delta_{2,1}^2(1 + R)^{-1}) \{\chi^2(n - p - 1)\}^{-1}$ .

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 1(iii), we have

$$T_{d,j} = n \log \left\{ 1 + \frac{\chi^2(1; g^2 \Delta_{\{j\}|\{-j\}}^2(1 + R_j)^{-1})}{\chi^2(n - p - 1)} \right\} - d, \quad (16)$$

where  $R_j = \chi^2(p - 1; g^2 \Delta_{\{-j\}}^2) \{\chi^2(n - p)\}^{-1}$ . When  $j \notin \mathbf{j}_*$ ,  $\Delta_{\{j\}|\{-j\}}^2 = 0$ , and we have

$$T_{d,j} = n \log \left\{ 1 + \frac{\chi^2(1)}{\chi^2(n - p - 1)} \right\} - d. \quad (17)$$

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  $\mathbf{j}_*$  is included in the full set  $\omega$ , i.e.,  $\mathbf{j}_* \subset \omega$ ;

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

For a proof of "[F1]  $\rightarrow 0$ ", which is defined by (5) with  $T_{d,j}$  in (16), we use the following two assumptions:

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

A4: For  $j \in \mathbf{j}_*$ ,  $\lim_{n \rightarrow \infty} \Delta_{\{j\}|\{-j\}}^2 > 0$  and  $\lim_{n \rightarrow \infty} \Delta_{\{-j\}}^2 > 0$ , under  $\Delta^2 = O(1)$ .

For the constant  $d$  in (13), we consider the following conditions:

B1:  $d/n \rightarrow 0$ ;

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

The following results were given by Fujikoshi and Sakurai [6].

**Theorem 3.** Let [F1] and [F2] be the quantities defined in (5) and (6) with  $T_{d,j}$  in (16). Suppose that assumptions A1 and A2 hold. Then, for the KOO method (14), the following asymptotic results hold.

- (i) Suppose that A3 and A4 hold. Then,  $[F1] \rightarrow 0$ ;  
(ii) Suppose that B1 and B2 hold. Then,  $[F2] \rightarrow 0$ ;  
(iii) Suppose that the assumptions in (i) and (ii) hold. Then, the KOO method (3) is asymptotically consistent.

**Proof of Theorem 3.** First we show “ $[F1] \rightarrow 0$ ”. Let  $j \in \mathbf{j}_*$ . Then,  $\Delta_{\{-j\}}^2 < \Delta^2$ , and  $\Delta_{\{j\} \cdot \{-j\}}^2 > 0$ . Using (16) and Lemma 1(iii),

$$T_{d,j} = n \log \left\{ 1 + \frac{\chi^2(1; g^2 \Delta_{\{j\} \cdot \{-j\}}^2 (1 + R_j)^{-1})}{\chi_{n-p-1}^2} \right\} - d,$$

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

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

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

$$\frac{1}{n} T_{d,j} \rightarrow \log \left( 1 + \frac{k_1 k_2 \lim \Delta_{\{j\} \cdot \{-j\}}^2}{1 + k_1 k_2 \lim \Delta_{\{-j\}}^2} \right) > 0,$$

which implies our assertion.

Next we show “ $[F2] \rightarrow 0$ ”. Using (17), we can write

$$[F2] = (p - p_*) \Pr(U > e^{d/n} - 1 - (n - p - 3)^{-1}) \leq (p - p_*) \Pr(U \geq h) \leq (p - p_*) h^{-2\ell} E(U^{2\ell}), \quad \ell \in \{1, 2, \dots\},$$

where  $U = \chi^2(1)/\chi^2(n-p-1) - 1/(n-p-3)$ . Noting  $h = O(n^{-a})$  and  $E[U^{2\ell}] = O(n^{-2\ell})$  (see, e.g., Theorem 16.2.2 in Fujikoshi et al. [9]), we have

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

Choosing  $\ell > 1/\{2(1-a)\}$ , we have “ $[F2] \rightarrow 0$ ”.  $\square$

Let  $d = n^r$ , where  $0 < r < 1$ . Then,  $h = O(n^{-(1-r)})$ , and condition B2 is satisfied. So, the KOO method (14) with

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

has a high-dimensional consistency. Among these, it has been pointed out in Sakurai and Fujikoshi [21] that the one with  $d = \sqrt{n}$  has numerically good behavior. Note that  $\hat{\mathbf{j}}_2$  and  $\hat{\mathbf{j}}_{\log n}$  as in (14) do not satisfy B2.

From our discussions, a consistency result may be noted under the following large-sample framework

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

That is,  $\hat{\mathbf{j}}_{\log n}$  and  $\hat{\mathbf{j}}_{\sqrt{n}}$  are consistent under a large-sample framework, if  $d \rightarrow \infty$  and  $d/n \rightarrow 0$ . A similar result was given by Nishii et al. [11] in the multiple-group case.

### 3.2. Multiple-group case

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

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

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

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

where  $n = n_1 + \dots + n_{q+1}$  and  $\bar{\boldsymbol{\mu}} = n^{-1}(n_1 \boldsymbol{\mu}^{(1)} + \dots + n_{q+1} \boldsymbol{\mu}^{(q+1)})$ . Let  $\lambda_1 \geq \dots \geq \lambda_q$  be the non-zero characteristic roots of  $\boldsymbol{\Omega} \boldsymbol{\Sigma}^{-1}$ . Then, the corresponding characteristic vectors are denoted by  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q$  with normalizations  $\boldsymbol{\beta}_i^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_i = 1$ ,  $i \in \{1, \dots, q\}$ . These 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  $i$ th population linear discriminant function. For simplicity, assume that  $p > q$  and  $\lambda_q > 0$ . Related to a partition of  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top)^\top$ ,  $\mathbf{y}_i : p_i \times 1$ , 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  $\beta_i$  and  $\Omega$  be partitioned as

$$\beta_i = \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

in the same way as  $\mu^{(i)}$  and  $\Sigma$ . The sufficiency hypothesis of  $\mathbf{y}_1$  or the redundancy hypothesis of  $\mathbf{y}_2$  was introduced by Rao [19,20] as

$$H_{2,1}: \mu_{2,1}^{(1)} = \cdots = \mu_{2,1}^{(q+1)},$$

where  $\mu_{2,1}^{(i)} = \mu_2^{(i)} - \Sigma_{21}\Sigma_{11}^{-1}\mu_1^{(i)}$ ,  $i \in \{1, \dots, q+1\}$ . Then the statement  $H_{2,1}$  is equivalent to one of the following equalities:

$$(i) \quad \beta_{2i} = 0, \quad i \in \{1, \dots, q\}, \quad (ii) \quad \text{tr}\Sigma^{-1}\Omega = \text{tr}\Sigma_{11}^{-1}\Omega_{11}.$$

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

$$M_{\{1, \dots, p_1\}}: \beta_{2j} = \mathbf{0}, \quad \beta_{ij} \neq 0, \quad i \in \{1, \dots, p_1\}, \quad j \in \{1, \dots, q\}.$$

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.,

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

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 partitions for  $\mathbf{B}$ ,  $\mathbf{W}$  and  $\mathbf{T}$  as in the partitions of  $\Sigma$ ,  $\Omega$ , etc. Then, GIC for  $M_{2,1}$  is given (see, e.g., Fujikoshi et al. [9]) as

$$\text{GIC}_{\{1, \dots, p_1\}} = -n \log\{|\mathbf{W}_{22,1}|/|\mathbf{T}_{22,1}|\} + n \log |n^{-1}\mathbf{W}| + np(1 + \log 2\pi) + d \left\{ qp_1 + p + \frac{1}{2}p(p+1) \right\}. \quad (18)$$

In general, let us denote the GIC for model  $M_j$  by  $\text{GIC}_{d,j}$  or simply  $\text{GIC}_j$ . For expressing the KOO method based on  $\text{GIC}_d$  in (18) in a simple way, we use the following notation: For  $\ell \in \omega$ ,  $\ell$  is used for any subset of  $p-1$  elements in  $\omega$  such that  $\ell = \omega \setminus \{\ell\}$ . We arrange the elements of  $\mathbf{y}$  as  $\mathbf{y} = (\mathbf{y}_\ell^\top, y_\ell)^\top$ . Corresponding to this partition, we express the partitions of  $\mu^{(i)}$ ,  $1 \leq i \leq q+1$ ,  $\Sigma$ ,  $\mathbf{W}$  and  $\mathbf{T} = \mathbf{W} + \mathbf{B}$  as follows:

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

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

$$\text{T}_{d,\ell} \equiv \text{GIC}_\ell - \text{GIC}_\omega = -n \log \frac{w_{\ell\ell,\ell}}{t_{\ell\ell,\ell}} - dq, \quad (19)$$

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

$$\hat{\mathbf{j}}_d = \{j \in \omega \mid \text{T}_{d,j} > 0\}, \quad \text{T}_{d,j} \text{ given by (19)}. \quad (20)$$

For a distributional reduction of  $\text{T}_{d,\ell}$ , we use the following Lemma.

**Lemma 2.** When dealing with  $w_{\ell\ell,\ell}/t_{\ell\ell,\ell}$ , 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  $\mathbf{B} \sim W_p(q, \mathbf{I}_p; \Psi)$ , where

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

and  $\Theta_\ell; (p-1) \times q$ ,  $\theta_\ell; q \times 1$ . We can express  $w_{\ell\ell,\ell}/t_{\ell\ell,\ell}$  and  $\text{T}_{d,\ell}$  as

$$\frac{w_{\ell\ell,\ell}}{t_{\ell\ell,\ell}} = \frac{w_{\ell\ell,\ell}}{w_{\ell\ell,\ell} + (t_{\ell\ell,\ell} - w_{\ell\ell,\ell})} = \frac{s_e}{s_e + s_h}, \quad \text{T}_{d,\ell} = n \log(1 + s_h s_e^{-1}) - dq,$$

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,

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

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 variables with the same variance 1,  $E(\mathbf{U}) = \mathbf{O}$ , and  $E(\mathbf{Z}) = \Theta_\ell$ .



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

$$\boldsymbol{\theta}_\ell^\top \boldsymbol{\theta}_\ell = g^2 \Delta_{\ell,\ell}^2, \quad \text{tr} \boldsymbol{\Theta}_\ell^\top \boldsymbol{\Theta}_\ell = g^2 \Delta_\ell^2.$$

These distributional results are coincident with the results in Lemma 1.

Note that

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

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

$$\beta = e^{dq/n} - 1 - \frac{q}{n - p - q - 2}, \quad (21)$$

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

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

$\tilde{A}2$ :  $p \rightarrow \infty$ ,  $n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, 1)$ ,  $n_i/n \rightarrow k_i > 0$ ,  $i \in \{1, \dots, q + 1\}$ ;

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

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

The assumption B2 is changed to

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

**Theorem 4.** Let [F1] and [F2] be the quantities defined in (5) and (6) with  $T_{d,j}$  in (19). Suppose that assumptions A1 and A2 hold. Then, for the KOO method (20), the following asymptotic results hold. (i) Suppose that  $\tilde{A}3$  and  $\tilde{A}4$  hold. Then, [F1]  $\rightarrow 0$ . (ii) Suppose that B1 and  $\tilde{B}2$  hold. Then, [F2]  $\rightarrow 0$ . (iii) Suppose that the assumptions in (i) and (ii) hold. Then, the KOO method (20) is asymptotically consistent.

**Proof of Theorem 4.** First, we consider the case of  $\ell \notin \mathbf{j}_*$ . Then,  $\boldsymbol{\theta}_\ell = \mathbf{0}$ , and  $s_h$  is distributed according to  $\chi^2(q)$ , and we have

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

where  $U_q = \chi^2(q) / \chi^2(n - p - q) - q / (n - p - q - 2)$ . Noting  $h_q = O(n^{-a})$  and  $E[U_q^{2\ell}] = O(n^{-2\ell})$  (see, e.g., Theorem 16.2.2 in Fujikoshi et al. [9]), we have

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

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

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

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

$\mathbf{V} = (\mathbf{Z}^\top \mathbf{Z})^{1/2} \{\mathbf{Z}^\top (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{Z}\}^{-1} (\mathbf{Z}^\top \mathbf{Z})^{1/2}$ . Since  $\mathbf{V} \sim W_q(n - p, \mathbf{I}_q)$  (see, e.g., Fujikoshi et al. [9], Theorem 2.3.3),  $\mathbf{V} \approx (n - p) \mathbf{I}_q$ . Similarly,  $\mathbf{Z}^\top \mathbf{Z} \sim W_q(p - 1, \mathbf{I}_q; \boldsymbol{\Theta}_\ell^\top \boldsymbol{\Theta}_\ell)$ ,  $\mathbf{Z}^\top \mathbf{Z} \approx (p - 1) \mathbf{I}_q + \boldsymbol{\Theta}_\ell^\top \boldsymbol{\Theta}_\ell$ . From these results we can get the required result. For more details, see Outline of Proof of Theorem 5 and also Oda et al. [14].  $\square$

In the following, we give some further results due to Oda et al. [14]. Without loss of generality, the penalty term  $d$  may be rewritten as

$$d = nq^{-1} \log(1 + b), \quad \text{or} \quad b = e^{dq/n} - 1, \quad (22)$$

assuming  $b > 0$ . Then, we have

$$[\text{F2}] \equiv \sum_{\ell \notin \mathbf{j}_*} \Pr(T_{d,\ell} \geq 0) = (p - p_*) \Pr\left(\frac{s_h}{s_e} > b\right) \leq (p - p_*) b^{-2r} E\left[\left\{\frac{\chi^2(q)}{\chi^2(n - p - q)}\right\}^{2r}\right]. \quad (23)$$

The above last inequality is derived by the Markov inequality. The order of the expectation in the above last equation is  $O(n^{-2r})$  (see, e.g., Theorem 16.2.2 in Fujikoshi et al. [9]). Now we make the following assumption for the penalty term  $b$ :

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

Therefore, from B3, we have

$$\sum_{\ell \notin \mathbf{j}_*} \Pr(T_{d,\ell} \leq 0) \leq O(pn^{-2r}b^{-2r}) \rightarrow 0. \quad (24)$$

Here, we note that B3 is equivalent to

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

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

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

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

The following result was given by Oda et al. [14].

**Theorem 5.** Suppose that assumptions A1, A2, B3, C1 and C2 hold. Then, the KOO method (20) based on GIC is consistent if the number  $p_*$  of true variables is finite.

**Outline of Proof of Theorem 5.** The result can be shown by proving (i);  $[F1] \rightarrow 0$  and (ii);  $[F2] \rightarrow 0$ . The result (ii) has been pointed out by (23) and (24). For a proof of (i), first we can see that assumptions C1 and C2 imply the assumptions  $\tilde{A}3$  and  $\tilde{A}4$ . Therefore, the result (i) will follow from the proof of Theorem 4. On the other hand, instead of using the proof of Theorem 4, we can prove by using

$$[F1] = \sum_{\ell \in \mathbf{j}_*} \Pr(T_{d,\ell} \leq 0) = \sum_{\ell \in \mathbf{j}_*} \Pr(L_\ell \leq 0), \quad L_\ell = \frac{s_h}{s_e} - \frac{q}{n-p-q-2} - \beta,$$

where  $\beta = b - q(n-p-q-2)^{-1}$ . It is assumed that  $\beta > 0$ . Here,  $s_h$  and  $s_e$  depend on  $\ell$ , but we may assume that  $s_e \sim \chi^2(n-p-q)$ , and  $s_h \sim \chi^2(q; \gamma_\ell^2)$  when  $\gamma_\ell^2$  is given in Lemma 2. The result (ii) will be shown by noting that  $L_\ell$  converges to a positive value.  $\square$

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 the high-dimensional consistency of KOO methods based on general information criteria in a multivariate regression model and discriminant analysis. Consistency of the KOO method based on some other model selection criteria is also examined in the multivariate regression model. Our results are restricted to the case of normality and weak consistency. Some results in the case of non-normality and strong consistency have been discussed in Bai et al. [1] in a multivariate regression model. However, since their results are under revision, we do not discuss them here. The consistency results in this paper are obtained by proving two properties [F1] and [F2] in (5) and (6). Sufficient conditions for each of [F1] and [F2] are 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. [18] for selection of variables in canonical correlation analysis, Fujikoshi et al. [4] and Enomoto [3] for selection of within-individual variables in the growth curve model, Oda et al. [13] for selection of response variables in multivariate calibration, and Oda et al. [17] for selection of response variables in multivariate regression model. Enomoto [3] has discussed consistency of the KOO method in the growth curve model, assuming that the covariance matrix has a uniform covariance structure. It is expected that high-dimensional properties of the KOO methods in these models will be further studied.

## Acknowledgments

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

## References

- [1] Z. Bai, Y. Fujikoshi, J. Hu, Strong Consistency of the AIC, BIC, Cp and KOO Methods in High-Dimensional Multivariate Linear Regression, Hiroshima Statistical Research Group TR 18-09, 2018.
- [2] L. Clemmensen, T. Hastie, D.M. Witten, B. Ersbell, Sparse discriminant analysis, *Technometrics* 53 (2011) 406–413.
- [3] R. Enomoto, Model selection criteria in growth curve model with a uniform covariance, in: Reported in the 2019 Statistical Meeting at Chuo University, 2019.
- [4] Y. Fujikoshi, R. Enomoto, T. Sakurai, High-dimensional AIC in the growth curve mode, *J. Multivariate Anal.* 122 (2013) 239–250.

- [5] Y. Fujikoshi, T. Kan, S. Takahashi, T. Sakurai, Prediction error criterion for selecting variables in a linear regression model, *Ann. Inst. Stat. Math.* 63 (2011) 387–403.
- [6] Y. Fujikoshi, T. Sakurai, Consistency of test-based method for selection of variables in high-dimensional two group-discriminant analysis, *Japanese J. Stat. Data Sci.* 2 (2019) 155–171.
- [7] Y. Fujikoshi, T. Sakurai, H. Yanagihara, Consistency of high-dimensional AIC-type and  $C_p$ -type criteria in multivariate linear regression, *J. Multivariate Anal.* 123 (2014) 184–200.
- [8] Y. Fujikoshi, K. Satoh, Modified AIC and  $C_p$  statistic in multivariate regression models, *Biometrika* 84 (1997) 707–716.
- [9] Y. Fujikoshi, V.V. Ulyanov, R. Shimizu, *Multivariate Statistics: High-Dimensional and Large-Sample Approximations* Wiley, Hoboken, N.J., 2010.
- [10] N. Hao, B. Bin, Sparsifying the Fisher linear discriminant by rotation, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 77 (2015) 827–851.
- [11] R. Nishii, Z.D. Bai, P.R. Krishnaiah, Strong consistency of the information criterion for model selection in multivariate analysis, *Hiroshima Math. J.* 18 (1988) 451–462.
- [12] R. Oda, Consistent variable selection criteria in multivariate linear regression even when dimension exceeds sample size, *Hiroshima Math. J.* 59 (2020) 339–374.
- [13] R. Oda, Y. Mima, H. Yanagihara, Y. Fujikoshi, A high-dimensional bias-corrected AIC for selecting response variables in multivariate calibration, *Comm. Statist. Theory Methods* 50 (2019) 3453–3476.
- [14] R. Oda, Y. Suzuki, H. Yanagihara, Y. Fujikoshi, A consistent variable selection method in high-dimensional canonical discriminant analysis, *J. Multivariate Anal.* 175 (2020) 1–13.
- [15] R. Oda, H. Yanagihara, A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables, *Electron. J. Stat.* 14 (2020) 1386–1412.
- [16] R. Oda, H. Yanagihara, A consistent likelihood-based variable selection method in normal multivariate linear regression, in: I. Czarnowski, et al. (Eds.), *Intelligent Decision Technologies*, 2021, pp. 391–401.
- [17] R. Oda, H. Yanagihara, Y. Fujikoshi, On Model Selection Consistency of Kick-One-Out Method for Selecting Response Variables in High-Dimensional Multivariate Linear Regression, *Hiroshima Statistical Research Group TR* 21-06, 2021.
- [18] R. Oda, H. Yanagihara, Y. Fujikoshi, Strong consistency of log-likelihood-based information criterion in high-dimensional canonical correlation analysis, *Sankhya A* 83 (2021) 109–127.
- [19] C.R. Rao, Inference on discriminant function coefficients, in: R.C. Bose, et al. (Eds.), *Essays in Prob. and Statist.*, 1970, pp. 587–602.
- [20] C.R. Rao, *Linear Statistical Inference and its Applications*, second ed., John Wiley & Sons, New York, 1973.
- [21] T. Sakurai, Y. Fujikoshi, 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*, T. Holgersson and M. Singnull, Springer, 2020.
- [22] R.S. Sparks, D. Coutsourides, L. Troskie, The multivariate  $C_p$ , *Comm. Statist. Theory Methods* 12 (1993) 1775–1793.
- [23] D.W. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant, *J. R. Stat. Soc. Ser. B Stat. Methodol.* (2011) 753–772.
- [24] H. Yanagihara, A high-dimensionality-adjusted consistent  $C_p$ -type for selecting variables in a normality-assumed linear regression with multiple responses, *Procedia Comput. Sci.* 96 (2016) 1096–1105.
- [25] H. Yanagihara, H. Wakaki, Y. Fujikoshi, A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large, *Electron. J. Stat.* 9 (2015) 869–897.
- [26] L.C. Zhao, P.R. Krishnaiah, Z.D. Bai, On determination of the number of signals in presence of white noise, *J. Multivariate Anal.* 20 (1986) 1–25.