A consistent variable selection method in high-dimensional canonical discriminant analysis

Ryoya Oda[*] , **Yuya Suzuki**,
Hirokazu Yanagihara and Yasunori Fujikoshi

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

(Last Modified: February 3, 2019)

Abstract

In this paper, we obtain the sufficient conditions to determine the consistency of a variable selection method based on a generalized information criterion in canonical discriminant analysis. To examine the consistency property, we use a high-dimensional asymptotic framework such that as the sample size n goes to infinity, then the ratio of the length of the observation vector p to the sample size, p/n, converges to a constant that is less than one even if the dimension of the observation vector also goes to infinity. Using the derived conditions, we propose a consistent variable selection method. From numerical simulations, we show that the probability of selecting the true model by our proposed method is high even when p is large.

1 Introduction

Canonical discriminant analysis (CDA) is a statistical method for classifying observations in a p-dimensional random vector $\boldsymbol{x}$ into one of $q+1$ populations Π_i $(i = 1, \ldots, q+1)$, and to describe the differences with a reduced number of dimensions. Assume that $q \le p$ and each Π_i is a p-dimensional normal population with mean vector $\boldsymbol{\mu}^{(i)}$ and the common positive definite covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$\Pi_i : N_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}) \ (i = 1, \ldots, q+1).$$

Suppose that we have n_i samples from each Π_i, and let the $n_i \times p$ matrices of the observations from Π_i be denoted by $\boldsymbol{X}_i$. These matrices can be expressed as follows:

$$\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_{q+1}')' \sim N_{n\times p}(\boldsymbol{G}\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n).$$

Here, n is the total number of samples, i.e., $n = \sum_{i=1}^{q+1} n_i$, and $\boldsymbol{G}$ and $\boldsymbol{\mathcal{M}}$ are $n \times (q+1)$ and $(q+1) \times p$ matrices given by

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{0}_{n_1} & \cdots & \boldsymbol{0}_{n_1} \\ \boldsymbol{0}_{n_2} & \boldsymbol{1}_{n_2} & \cdots & \boldsymbol{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{n_{q+1}} & \boldsymbol{0}_{n_{q+1}} & \cdots & \boldsymbol{1}_{n_{q+1}} \end{pmatrix}, \ \boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(q+1)})',$$

[*] Corresponding author. Email: oda.stat@gmail.com

where $\mathbf{1}_{n_i}$ and $\mathbf{0}_{n_i}$ are n_i-dimensional vectors of ones and zeros, respectively. We also assume $n - p - q - 2 > 0$ when proposing our method. Let $\bar{\boldsymbol{\mu}}$ be the population overall mean vector given by $\bar{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{q+1} n_i \boldsymbol{\mu}^{(i)}$, and let $\boldsymbol{\Omega}$ be the population between-groups covariance matrix defined by

$$\boldsymbol{\Omega} = \frac{1}{n}\sum_{i=1}^{q+1} n_i(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})' = \frac{1}{n}\boldsymbol{\mathcal{M}}'\boldsymbol{G}'(\boldsymbol{P}_{\boldsymbol{G}} - \boldsymbol{P}_{\mathbf{1}_n})\boldsymbol{G}\boldsymbol{\mathcal{M}},$$

where $\boldsymbol{P}_{\boldsymbol{A}}$ is the projection matrix to the subspace spanned by the columns of a matrix $\boldsymbol{A}$, i.e., $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$. In CDA, the linear discriminant functions play an important role. Let the population linear discriminant functions be denoted by $f_a(\boldsymbol{x}|\boldsymbol{\beta}^{(a)}) = \boldsymbol{\beta}^{(a)'}\boldsymbol{x}$ $(a = 1, \ldots, q)$. Here, the coefficient vectors $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(q)}$ are given as the solutions of

$$\boldsymbol{\Omega}\boldsymbol{\beta}^{(a)} = \lambda_a \boldsymbol{\Sigma}\boldsymbol{\beta}^{(a)},\ \boldsymbol{\beta}^{(a)'}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(b)} = \delta_{ab}\ (a = 1, \ldots, q;\ b = 1, \ldots, q),$$

where δ_{ab} is the Kronecker delta, i.e., if $a = b$ then $\delta_{ab} = 1$, otherwise $\delta_{ab} = 0$, and λ_a is the a-th maximum eigenvalue of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}$ satisfying $\lambda_1 \geq \cdots \geq \lambda_q > 0$.

It is important to specify the factors affecting the classification in CDA. In this paper we consider variable selection methods for a redundancy model. Suppose that j denotes a subset of $\omega = \{1, \ldots, p\}$ containing p_j elements, and $\boldsymbol{x}_j$ denotes the p_j-dimensional random vector consisting of the components of $\boldsymbol{x}$ indexed by the elements of j. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{x}_j$ consists of the first, second, and fourth elements of $\boldsymbol{x}$. Without loss of generality, we express $\boldsymbol{x}$ as $\boldsymbol{x} = (\boldsymbol{x}_j', \boldsymbol{x}_{\bar{j}}')'$, where $\boldsymbol{x}_{\bar{j}}$ is the $(p - p_j)$-dimensional random vector and $\bar{A}$ denotes the compliment of a set A. Similar to $\boldsymbol{x}$, we define $\boldsymbol{\beta}^{(a)}$ as $\boldsymbol{\beta}^{(a)} = (\boldsymbol{\beta}_j^{(a)'}, \boldsymbol{\beta}_{\bar{j}}^{(a)'})'$. Then, for a candidate model such that $\boldsymbol{x}_{\bar{j}}$ is redundant, we consider the following model (see, Fujikoshi, 1982):

$$\boldsymbol{\beta}_{\bar{j}}^{(1)} = \cdots = \boldsymbol{\beta}_{\bar{j}}^{(q)} = \mathbf{0}_{p-p_j}. \tag{1}$$

We call a candidate model (1) as model j or redundancy model j. Here, for model j, $\boldsymbol{x}_{\bar{j}}$ does not contribute to the population linear discriminant functions $f_1, \ldots, f_q$. Thus, $\boldsymbol{x}_{\bar{j}}$ is regarded as the redundant vector in CDA for model j. To select the optimal model among all such candidate models, some information criteria (IC) have been proposed. Fujikoshi (1983) applied Akaike's information criterion (AIC) (Akaike, 1973; 1974), and a modified version was also proposed by Fujikoshi (1985). Nishii *et al.* (1988) considered a generalized information criterion (GIC) by replacing AIC's penalty, 2, with any positive constant. It is noted that the GIC includes several well-known ICs, for example, the AIC, Bayesian information criterion (BIC) and Hannan-Quinn information criterion (HQC), where the BIC and HQC were proposed by Schwarz (1978) and Hannan and Quinn (1979), respectively. The usual selection method based on an IC regards the best model as the model that has the minimum IC among all models (1) included in the set of candidate models $\mathcal{J}$, that is the best model is written as

$$\tilde{j} = \arg\min_{j \in \mathcal{J}} \mathrm{IC}(j), \tag{2}$$

where $\mathrm{IC}(j)$ denotes the value of the IC for model j.

If there is a true model j_*, which is the minimum subset including the true vector $\boldsymbol{x}_{j_*}$ from models (1), then $\boldsymbol{x}_{j_*}$ can be considered as the vector that determines the classification. Consistency is well-known as a desirable property for a variable selection method. A variable selection method is said to be consistent if the probability of selecting the true model j_* converges to 1. In the context of CDA, Fujikoshi (1984) showed that (2) based on the AIC is not consistent, and Nishii *et al.* (1988) obtained the sufficient conditions for (2), based on the GIC, to be strongly consistent, which is a stricter property than standard consistency. However, these results were obtained under the large-sample (LS) asymptotic framework such that only n goes to infinity. In general, the LS asymptotic framework is not suitable for a high-dimensional case such that not only n but also the dimension p are large, and asymptotic results may cause a non-negligible bias. Moreover, it is usually considered that the number of candidate models is huge in the high-dimensional case. Thus, it is not practical to use the selection method (2). To overcome this problem, we consider the method proposed in Zhao *et al.* (1986) and Nishii *et al.* (1988), which was developed for the LS situation. The method is as follows. Let ℓ be a subset of ω satisfying $\#(\ell) = p - 1$, and let the elements of $\bar{\ell}$ be denoted as e_ℓ, that is ℓ and e_ℓ satisfy $\ell = \omega \backslash \{e_\ell\}$, where $\#(A)$ denotes the number of elements of a set A. Then, the best model under a criterion, IC, is written as

$$\hat{j} = \{e_\ell \in \omega \mid \mathrm{IC}(\ell) > \mathrm{IC}(\omega)\}. \tag{3}$$

This selection method is useful for the high-dimensional case because the number of calculations required to compute IC is only $p + 1$. Nishii *et al.* (1988) showed under the LS asymptotic framework that the consistency conditions for the method (2) based on the GIC are the same as for the method (3). Recently, Fujikoshi and Sakurai (2018) obtained the sufficient conditions for consistency for the method (3) based on the GIC in two-group discriminant analysis (i.e., $q = 1$) as both n and p go to infinity. Furthermore, such variable selection methods have been used in multivariate regression, e.g., by Sakurai and Fujikoshi (2017), Bai *et al.* (2018), and Oda and Yanagihara (2019). In this paper, following Bai *et al.* (2018), the method (3) is called the Kick-One-Out (KOO) method.

The aim of this paper is to obtain sufficient conditions such that the KOO method (3) based on the GIC is consistent when n goes to infinity and p may go to infinity in canonical discriminant analysis. To achieve this, the following high-dimensional (HD) asymptotic framework is used:

$$n \to \infty,\ \frac{n_i}{n} \to \rho_i \in (0, 1)\ (i = 1, \ldots, q + 1),\ \frac{p}{n} \to c \in [0, 1),\ p_{j_*}, q\text{: fixed}. \tag{4}$$

Using our consistency conditions, we propose a consistent variable selection method. Since the HD asymptotic framework includes the LS asymptotic framework, our proposed method is consistent even when only n is large, so the consistency of the method does not rely on the divergence order of p as long as $c < 1$.

The remainder of the paper is organized as follows. In section 2, we present the KOO method based on the GIC and the necessary assumptions for deriving our results. In section 3, we obtain the sufficient conditions for the method to be consistent and propose a consistent variable selection method. In section 4, we present the results of numerical simulations and compare the performance of our proposed method with that of the KOO method (3) based on existing criteria. Technical details are relegated to the Appendix.

2 Preliminaries

In this section, we present the KOO method (3) based on the GIC and the necessary assumptions for deriving our results. Hereafter, since we consider (3) not (2), we define ℓ ($\subset \omega$) as a model satisfying $\#(\ell) = p - 1$, and e_ℓ denotes the elements of $\bar{\ell}$. First, we present the KOO method based on the GIC. Let $\boldsymbol{W}$ and $\boldsymbol{B}$ be the matrices of the sums of squares and products within groups and between groups given by

$$\boldsymbol{W} = \boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{P_G})\boldsymbol{X},\ \boldsymbol{B} = \boldsymbol{X}'(\boldsymbol{P_G} - \boldsymbol{P}_{\boldsymbol{1}_n})\boldsymbol{X}.$$

We express the partitions of $\boldsymbol{\mu}^{(i)}$ $(i = 1, \dots, q+1)$, $\boldsymbol{\Sigma}$,$\boldsymbol{W}$ and $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$ corresponding to the division of $\boldsymbol{x} = (\boldsymbol{x}'_\ell, x_{\bar{\ell}})'$ as follows:

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

From Fujikoshi (1982), the model ℓ is equivalent to

$$\mu_{\bar{\ell}\cdot\ell}^{(1)} = \cdots = \mu_{\bar{\ell}\cdot\ell}^{(q+1)}, \tag{5}$$

where $\mu_{\bar{\ell}\cdot\ell}^{(i)} = \mu_{\bar{\ell}}^{(i)} - \boldsymbol{\sigma}'_{\ell\bar{\ell}}\boldsymbol{\Sigma}_{\ell\ell}^{-1}\boldsymbol{\mu}_\ell^{(i)}$ $(i = 1, \dots, q+1)$, and $\mu_{\bar{\ell}\cdot\ell}^{(i)}$ expresses the value after removing the random term from the i-th group's conditional mean of $x_{\bar{\ell}}$ giving $\boldsymbol{x}_\ell$. The expression (5) was introduced by Rao (1948; 1973). Let $f(\boldsymbol{X}|\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma})$ be the probability density function of $N_{n\times p}(\boldsymbol{G}\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Then, the maximum negative twofold log-likelihood under (5) is expressed as follows (see e.g., Fujikoshi *et al.*, 2010, chap 9.4.3):

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\{-2\log f(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\ s.t.\ \mu_{\bar{\ell}\cdot\ell}^{(1)} = \cdots = \mu_{\bar{\ell}\cdot\ell}^{(q+1)}\} = np(1 + \log 2\pi) - n\log\frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{t_{\bar{\ell}\bar{\ell}\cdot\ell}} + n\log|n^{-1}\boldsymbol{W}|,$$

where $w_{\bar{\ell}\bar{\ell}\cdot\ell} = w_{\bar{\ell}\bar{\ell}} - \boldsymbol{w}'_{\ell\bar{\ell}}\boldsymbol{W}_{\ell\ell}^{-1}\boldsymbol{w}_{\ell\bar{\ell}}$ and $t_{\bar{\ell}\bar{\ell}\cdot\ell} = t_{\bar{\ell}\bar{\ell}} - \boldsymbol{t}'_{\ell\bar{\ell}}\boldsymbol{T}_{\ell\ell}^{-1}\boldsymbol{t}_{\ell\bar{\ell}}$. Then, the GIC in the model (5) is defined by

$$\mathrm{GIC}(\ell) = np(1 + \log 2\pi) - n\log\frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{t_{\bar{\ell}\bar{\ell}\cdot\ell}} + n\log|n^{-1}\boldsymbol{W}| + \alpha h_\ell,$$

where α is a positive constant and h_ℓ is the number of parameters in (5), i.e., $h_\ell = -q + p(q+1) + p(p+1)/2$. In particular, the GIC in ω is given by

$$\mathrm{GIC}(\omega) = np(1 + \log 2\pi) + n\log|n^{-1}\boldsymbol{W}| + \alpha h_\omega,$$

where $h_\omega = p(q+1) + p(p+1)/2$. By choosing α, we can express the following specific criteria as a special case of the GIC:

$$\alpha = \begin{cases} 2 & \text{(AIC)} \\ \log n & \text{(BIC)} \\ 2\log\log n & \text{(HQC)} \end{cases}. \tag{6}$$

The optimal model $\hat{j}$ obtained by the KOO method based on the GIC is defined in the same way as (3), i.e.,

$$\hat{j} = \{e_\ell \in \omega \mid \mathrm{GIC}(\ell) > \mathrm{GIC}(\omega)\}. \tag{7}$$

Let $\mathcal{J}_+$ denote the set of overspecified models, which is defined by

$$\mathcal{J}_+ = \{j \subset \omega \mid \boldsymbol{\beta}_{\bar{j}}^{(1)} = \cdots = \boldsymbol{\beta}_{\bar{j}}^{(q+1)} = \mathbf{0}_{p-p_j}\}.$$

The true model j_* is expressed as the overspecified model having the smallest number of elements, i.e., $j_* = \arg\min_{j \in \mathcal{J}_+} p_j$.

Next, we present the necessary assumptions for deriving our results. Let

$$\delta = \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}),\ \delta_\ell = \mathrm{tr}(\boldsymbol{\Omega}_{\ell\ell}\boldsymbol{\Sigma}_{\ell\ell}^{-1}),$$

where $\boldsymbol{\Omega}_{\ell\ell}$ is the $(p-1)\times(p-1)$ matrix resulting from deleting the p-th row and column vectors from $\boldsymbol{\Omega}$. Note that δ is the square of the Mahalanobis distance among multiple groups except for the case of constant multiplication. For example, if $q = 1$ then $(n_1 n_2/n^2)\delta$ is equivalent to the square of the Mahalanobis distance between two groups. Note that (1) and (5) are equivalent to (see, Fujikoshi, 1982)

$$\delta = \delta_\ell. \tag{8}$$

The above equation means that the value of the Mahalanobis distance does not change even if redundant variables are removed. Let the minimum eigenvalue of a square matrix $\boldsymbol{A}$ be denoted by $\lambda_{\min}(\boldsymbol{A})$. To examine the sufficient conditions for consistency, we introduce the following two assumptions:

A1. There exists $c_1 > 0$ such that $\lambda_{\min}(\boldsymbol{\Sigma}) > c_1$.

A2. For all $\ell \supset \bar{j}_*$ $(\#(\ell) = p-1)$, there exists $c_2 > 0$ such that $\delta - \delta_\ell > c_2$.

Assumption A1 ensures the covariance matrix $\boldsymbol{\Sigma}$ is positive definite asymptotically. Moreover, from the general formulas for the determinant of the partitioned matrix (e.g., Lütkepohl, 1997, 4.2.2 (6); 9.12.2 (5)), if Assumption A1 is true then the following equation holds:

$$\sigma_{\bar{\ell}\bar{\ell}\cdot\ell} = \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_{\ell\ell}|} \geq \frac{\prod_{a=1}^{p}\lambda_a(\boldsymbol{\Sigma})}{\prod_{a=1}^{p-1}\lambda_a(\boldsymbol{\Sigma})} = \lambda_{\min}(\boldsymbol{\Sigma}) > c_1\ (\forall \ell \subset \omega\ (\#(\ell) = p-1)), \tag{9}$$

where $\sigma_{\bar{\ell}\bar{\ell}\cdot\ell} = \sigma_{\bar{\ell}\bar{\ell}} - \boldsymbol{\sigma}_{\ell\bar{\ell}}'\boldsymbol{\Sigma}_{\ell\ell}^{-1}\boldsymbol{\sigma}_{\ell\bar{\ell}}$ and $\lambda_a(\boldsymbol{\Sigma})$ is the a-th maximum eigenvalue of $\boldsymbol{\Sigma}$ satisfying $\lambda_1(\boldsymbol{\Sigma}) \geq \cdots \geq \lambda_p(\boldsymbol{\Sigma})$. Assumption A2 is related to the behavior of the difference between two Mahalanobis distances. From the definition of the true model and (8), we note that $\delta - \delta_\ell > 0$ holds under the finite case. Assumption A2 keeps $\delta - \delta_\ell$ positive asymptotically.

3 Main Results

In this section, we obtain sufficient conditions for the consistency of the KOO method (7) based on the GIC and propose a consistent variable selection method. To obtain these conditions, we first introduce some notations. Let $\boldsymbol{\Gamma}_\ell$ be a $p \times p$ non-singular matrix given by

$$\boldsymbol{\Gamma}_\ell = \begin{pmatrix} \boldsymbol{\Sigma}_{\ell\ell}^{-1/2} & \mathbf{0}_{p-1} \\ -\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2}\boldsymbol{\sigma}_{\ell\bar{\ell}}'\boldsymbol{\Sigma}_{\ell\ell}^{-1} & \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2} \end{pmatrix},$$

and let a $p \times p$ transformation of the population between-groups covariance matrix $\boldsymbol{\Omega}$ be denoted by

$$\boldsymbol{\Psi}_\ell = n\boldsymbol{\Gamma}_\ell \boldsymbol{\Omega} \boldsymbol{\Gamma}'_\ell, \tag{10}$$

which is called the non-centrality matrix. To examine the consistency conditions, it is important to understand the behavior of the non-centrality matrix $\boldsymbol{\Psi}_\ell$. To better understand the characteristics of $\boldsymbol{\Psi}_\ell$, we present another expression for $\boldsymbol{\Psi}_\ell$. Let $\boldsymbol{d}$ and $\boldsymbol{D}$ be a $(q+1)$-dimensional vector and $(q+1) \times (q+1)$ diagonal matrix consisting of $(n_1/n)^{1/2}, \ldots, (n_{q+1}/n)^{1/2}$, respectively, i.e.,

$$\boldsymbol{d} = (d_1, \ldots, d_{q+1})', \ \boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_{q+1}), \ d_i = \sqrt{\frac{n_i}{n}} \ (i = 1, \ldots, q+1).$$

From $\boldsymbol{d}$ and $\boldsymbol{D}$, we can derive the following expression for $\boldsymbol{\Psi}_\ell$:

$$\boldsymbol{\Psi}_\ell = n\boldsymbol{\Gamma}_\ell \boldsymbol{\mathcal{M}}' \boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}\boldsymbol{\mathcal{M}}\boldsymbol{\Gamma}'_\ell. \tag{11}$$

Note that $\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}}$ is symmetric and idempotent, and its rank is $\mathrm{rank}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}}) = q$. These facts imply $\mathrm{rank}(\boldsymbol{\Psi}_\ell) \le \min(p, q) = q$. Hence, $\boldsymbol{\Psi}_\ell$ can be decomposed into

$$\boldsymbol{\Psi}_\ell = \boldsymbol{\Theta}'_\ell \boldsymbol{\Theta}_\ell, \tag{12}$$

where $\boldsymbol{\Theta}_\ell$ is a $q \times p$ matrix. The result $\mathrm{rank}(\boldsymbol{\Psi}_\ell) \le q$ can also be seen from

$$\mathrm{rank}(\boldsymbol{\Psi}_\ell) = \mathrm{rank}(\boldsymbol{\Omega}) \le \mathrm{rank}(\boldsymbol{P}_{\boldsymbol{G}} - \boldsymbol{P}_{\boldsymbol{1}_n}) = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{G}} - \boldsymbol{P}_{\boldsymbol{1}_n}) = q.$$

Let us split $\boldsymbol{\Theta}_\ell$ into a sub-matrix and a sub-vector of $\boldsymbol{\Theta}_\ell = (\boldsymbol{\Theta}_{\ell,1}, \boldsymbol{\theta}_{\ell,2})$, where $\boldsymbol{\Theta}_{\ell,1}$ is a $q \times (p-1)$ matrix and $\boldsymbol{\theta}_{\ell,2}$ is a q-dimensional vector. Then, we can obtain some properties of the non-centrality matrix (the proof is given in Appendix A).

Lemma 3.1. *Let ℓ be a subset of ω satisfying $\#(\ell) = p-1$. The non-centrality matrix $\boldsymbol{\Psi}_\ell$ defined in* (10) *and $\boldsymbol{\theta}_{\ell,2}$, which is the p-th column vector of $\boldsymbol{\Theta}_\ell$ given by* (12)*, has the following properties:*

(i) *For all $\ell \supset j_*$, $\boldsymbol{\theta}_{\ell,2} = \boldsymbol{0}_q$.*

(ii) *For all $\ell \supset \bar{j}_*$, $\inf_{n>p, p\ge 1} n^{-1}\boldsymbol{\theta}'_{\ell,2}\boldsymbol{\theta}_{\ell,2} > 0$ under the HD asymptotic framework* (4) *when Assumptions A1 and A2 hold.*

(iii) *There is a $(q+1) \times (q+1)$ positive semi-definite matrix $\boldsymbol{\Delta}$ such that for all $\ell \supset \bar{j}_*$,*

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}_\ell) \to \mathrm{tr}(\boldsymbol{\Delta}),$$

in the HD asymptotic framework (4).

To obtain sufficient conditions for consistency, we rewrite α as

$$\alpha = \frac{n}{q}\log(1+\beta), \ \beta > 0. \tag{13}$$

Using Lemma 3.1 and (13), the conditions of β for (7) based on the GIC in the HD asymptotic framework can be derived from Theorem 3.1 (the proof is given in Appendix B).

Theorem 3.1. *Suppose that Assumptions A1 and A2 hold. Then the KOO method* (7) *based on the GIC is consistent in the HD asymptotic framework, if the following conditions C1 and C2 are satisfied* :

C1. $p^{-1/2r} n\beta \to \infty$ *for some* $r \in \mathbb{N}$,

C2. $\beta \to 0$,

where β *is defined by* $\alpha = (n/q)\log(1+\beta)$ *in* (13).

Although the conditions in Theorem 3.1 are expressed in terms of β, they can also be written in terms of α. Alternate expressions for the conditions are given by Corollary 3.1:

Corollary 3.1. *The conditions* C1 *and* C2 *in Theorem 3.1 are equivalent to the following conditions* C1′ *and* C2′:

C1′. $p^{-1/2r}\alpha \to \infty$ *for some* $r \in \mathbb{N}$,

C2′. $\dfrac{\alpha}{n} \to 0$.

From Corollary 3.1, we observe that the divergence of α should be expressible as a polynomial in order to satisfy the conditions C1′ and C2′ when both n and p go to infinity. Hence, we derive the following two properties for the specific criteria (6):

- When p is fixed, the AIC satisfies C2′ but not C1′, and the BIC and HQC satisfy both C1′ and C2′ as $n \to \infty$.
- When p goes to ∞, the AIC, BIC and HQC satisfy C2′ but not C1′ as $(n, p) \to \infty$.

These facts imply that the KOO methods (7) based on the AIC, BIC and HQC may not be consistent in the HD asymptotic framework. Therefore, we propose the KOO method based on an example criterion that is always consistent in the HD asymptotic framework. We define the criterion, named the high-dimensionality-adjusted consistent information criterion (HCIC), as follows:

$$\mathrm{HCIC}(\ell) = np(1+\log 2\pi) - n\log\frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{t_{\bar{\ell}\bar{\ell}\cdot\ell}} + n\log|n^{-1}\boldsymbol{W}| + \frac{n}{q}\log\left(1+\frac{\log n}{\sqrt{n}}\right)h_\ell. \quad (14)$$

It is straightforward to check that the HCIC satisfies conditions C1 and C2 in Theorem 3.1. Therefore, the KOO method (7) based on the HCIC is consistent in the HD asymptotic framework. For the two group case, our derived sufficient conditions are essentially the same as those obtained by Fujikoshi and Sakurai (2018) although the assumptions are different. Note that the GIC where $\alpha = n^{1/2}$ satisfies conditions C1′ and C2′. Fujikoshi and Sakurai (2018) mentioned that the KOO method (7) based on the criterion where $\alpha = n^{1/2}$ performed well in numerical studies for the two group case. The penalty of the HCIC is larger than $n^{1/2}$ because the penalty of the HCIC is $\alpha = O(n^{1/2}\log n)$. However, the HCIC also performed well in the numerical studies in section 4.

4 Numerical studies

In this section, we numerically compare the probabilities of selecting the true model by the KOO methods (7) based on the HCIC in (14) and the AIC, BIC and HQC in (6). The probabilities of selecting the true model j_* were evaluated by Monte Carlo simulations with $10,000$ iterations. In this numerical experiment, we set $p_{j_*} = 4$, $q = 3$ and $n_i = n/(q+1)$ $(i = 1, \ldots, q+1)$, and the exchangeable matrix was used as the covariance matrix, i.e., $\boldsymbol{\Sigma} = (1-\xi)\boldsymbol{I}_p + \xi\boldsymbol{1}_p\boldsymbol{1}_p'$ with $\xi = 0.8$. The mean vectors $\boldsymbol{\mu}^{(i)}$ $(i = 1, \ldots, 4)$ were constructed as follows. Sub-vectors of $\boldsymbol{\mu}^{(i)}$ are expressed as $\boldsymbol{\mu}^{(i)} = (\boldsymbol{\mu}_1^{(i)'}, \boldsymbol{\mu}_2^{(i)'})'$, where $\boldsymbol{\mu}_1^{(i)}$ and $\boldsymbol{\mu}_2^{(i)}$ are p_{j_*}- and $(p - p_{j_*})$-dimensional vectors, respectively. The elements of $\boldsymbol{\mu}_1^{(i)}$ $(i = 1, \ldots, 4)$ were defined as follows:

$$\boldsymbol{\mu}_1^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \ \boldsymbol{\mu}_1^{(2)} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \ \boldsymbol{\mu}_1^{(3)} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \ \boldsymbol{\mu}_1^{(4)} = -\sum_{i=1}^{3} \boldsymbol{\mu}_1^{(i)}.$$

We set $\boldsymbol{\mu}_2^{(2)} = \boldsymbol{\mu}_2^{(3)} = \boldsymbol{0}_{p-p_{j_*}}$. Then, based on $\boldsymbol{\mu}_1^{(i)}$ $(i = 1, \ldots, 4)$, $\boldsymbol{\mu}_2^{(1)}$ and $\boldsymbol{\mu}_2^{(4)}$ were constructed as

$$\boldsymbol{\mu}_2^{(1)} = \frac{\xi p_{j_*}}{1 + \xi(p_{j_*} - 1)} \boldsymbol{1}_{p-p_{j_*}}, \ \boldsymbol{\mu}_2^{(4)} = -\frac{5\xi}{1 + \xi(p_{j_*} - 1)} \boldsymbol{1}_{p-p_{j_*}}.$$

The model that has the above mean vectors and exchangeable matrix satisfies $\boldsymbol{\beta}_{\bar{j}_*}^{(1)} = \cdots = \boldsymbol{\beta}_{\bar{j}_*}^{(3)} = \boldsymbol{0}_{p-p_{j_*}}$ and the definition of the true model. Under these settings, the data $\boldsymbol{X}$ were generated from $N_{n\times p}(\boldsymbol{G}\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Table 1 shows the probabilities of selecting the true model j_* by the KOO methods (7) based on each of the four criteria. In this table, the left column shows the results when p is fixed, and the right column shows the results when p increases with n keeping p/n constant. From the Table, we observe that the method based on the HCIC has the highest probabilities among the methods based on the four criteria in all cases. However, it seems that the probabilities increase slowly when p approaches n because we used an asymptotic framework such that p/n converges to a constant that is less than 1. On the other hand, the probabilities obtained with the method based on the AIC are low even when p is small. The reason is that the AIC does not satisfy the conditions C1 and C2. The probabilities for the methods based on the BIC and HQC increase as only n increases, but they do not tend to 100.00 % as n and p increase except in the BIC's case for $p/n = 0.1$. The results also suggest that the BIC and HQC satisfy the conditions C1 and C2 as $n \to \infty$ but do not satisfy them as $(n, p) \to \infty$.

5 Conclusions

In this paper, we consider the variable selection problem in canonical discriminant analysis, and provide sufficient conditions to determine the consistency of the KOO method (7) based on the GIC in the HD asymptotic framework such that n goes to ∞ and p may also go to ∞ but p/n converges to a constant that is less than 1. From Corollary 3.1, we observe that the AIC, BIC and HQC do not satisfy the sufficient conditions for consistency in the HD asymptotic framework and so the KOO methods based on them may not be consistent. Therefore, we proposed the

Table 1: Probabilities of selecting the true model j_* with each of the four criteria; the left column shows the results when p is fixed, and the right column shows the results when p increases with n keeping p/n some constants, which are $0.1, 0.3, 0.5, 0.8$

		Selection Probability (%)					Selection Probability (%)			
n	p	AIC	BIC	HQC	HCIC	p	AIC	BIC	HQC	HCIC
100	10	38.55	95.72	76.34	95.31	10	37.25	95.98	75.81	95.20
200	10	44.13	98.85	86.61	99.95	20	8.78	95.76	62.33	99.69
500	10	47.43	99.79	92.03	100.00	50	0.11	96.24	40.87	100.00
1000	10	48.61	99.91	94.62	100.00	100	0.00	96.81	23.08	100.00
100	30	0.25	55.03	9.62	83.78	30	0.22	56.25	9.68	83.88
200	30	1.56	90.15	38.15	99.35	60	0.00	54.01	2.55	98.52
500	30	2.99	98.69	65.76	100.00	150	0.00	51.01	0.11	100.00
1000	30	3.54	99.47	75.68	100.00	300	0.00	50.63	0.00	100.00
100	50	0.00	5.37	0.08	73.79	50	0.00	5.94	0.04	73.99
200	50	0.01	70.05	7.86	98.93	100	0.00	2.40	0.00	95.96
500	50	0.10	96.10	39.84	100.00	250	0.00	0.56	0.00	100.00
1000	50	0.27	99.00	58.75	100.00	500	0.00	0.12	0.00	100.00
100	80	0.00	0.00	0.00	3.83	80	0.00	0.00	0.00	3.63
200	80	0.00	19.09	0.07	97.71	160	0.00	0.00	0.00	39.84
500	80	0.00	89.75	13.43	100.00	400	0.00	0.00	0.00	98.57
1000	80	0.00	98.04	35.71	100.00	800	0.00	0.00	0.00	99.99

KOO method (7) based on the high-dimensionality-adjusted consistent information criterion (HCIC), which always has consistency under the HD asymptotic framework. The simulation results showed the validity of the sufficient conditions and the high-probability of selecting the true model by the method based on the HCIC.

In order to obtain the sufficient conditions for consistency, we used only two simple assumptions under the HD asymptotic framework. However, it is also important to consider the case such that q and p_{j_*} may go to infinity in order to improve the accuracy of the approximations when q and p_{j_*} are not small. In such situations, more complex assumptions are required, but we leave this as a future work. For the high-dimensional case such that p is larger than n, the GIC is not defined because the inverse matrix of $\boldsymbol{W}$ does not exist. However, some papers deal with such high-dimensional cases by applying regularization methods (e.g., Hastie *et al.*, 1995; Clemmensen *et al.*, 2011) and screening methods (e.g., Cheng *et al.*, 2017). We can extend our method to the high-dimensional case by using such methods, but this is left as a future work.

Acknowledgments

Ryoya Oda was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. Hirokazu Yanagihara and Yasunori Fujikoshi were both partially supported by the Ministry of Education, Science, Sports, and Culture through Grants-in-Aid for Scientific Research (C), #18K03415 and #16K00047, respectively.

Appendix

A Proof of Lemma 3.1

We calculate $\boldsymbol{\Psi}_\ell$ in (11) to show (i) and (ii). From the definitions of $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\Gamma}_\ell$, we can calculate $\boldsymbol{\mathcal{M}}\boldsymbol{\Gamma}_\ell'$ as follows:

$$\boldsymbol{\mathcal{M}}\boldsymbol{\Gamma}_\ell' = \begin{pmatrix} \boldsymbol{\mu}_\ell^{(1)'}\boldsymbol{\Sigma}_{\ell\ell}^{-1/2} & -\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2}\mu_{\bar{\ell}\cdot\ell}^{(1)} \\ \vdots & \vdots \\ \boldsymbol{\mu}_\ell^{(q+1)'}\boldsymbol{\Sigma}_{\ell\ell}^{-1/2} & -\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2}\mu_{\bar{\ell}\cdot\ell}^{(q+1)} \end{pmatrix}.$$

Since $\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}$ is the (p,p)-th element of $\boldsymbol{\Psi}_\ell$, $\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}$ is expressed as

$$\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2} = n\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1}(\mu_{\bar{\ell}\cdot\ell}^{(1)},\ldots,\mu_{\bar{\ell}\cdot\ell}^{(q+1)})\boldsymbol{D}(\boldsymbol{I}_{q+1}-\boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}(\mu_{\bar{\ell}\cdot\ell}^{(1)},\ldots,\mu_{\bar{\ell}\cdot\ell}^{(q+1)})'. \tag{A.1}$$

It should be noted that equation $(\boldsymbol{I}_{q+1}-\boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{a} = \boldsymbol{0}_{q+1}$ holds if and only if $\boldsymbol{a} = a_0\boldsymbol{D}\boldsymbol{1}_{q+1}$ for some $a_0 \in \mathbb{R}$. First, we show (i). When $\ell \supset j_*$, it follows from the equivalence of (1) and (5) that $\mu_{\bar{\ell}\cdot\ell}^{(1)} = \cdots = \mu_{\bar{\ell}\cdot\ell}^{(q+1)}$ holds. Then, the following equation can be derived:

$$\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2} = n\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1}(\mu_{\bar{\ell}\cdot\ell}^{(1)})^2\boldsymbol{1}_{q+1}'\boldsymbol{D}(\boldsymbol{I}_{q+1}-\boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}\boldsymbol{1}_{q+1} = 0.$$

The above equation implies that $\boldsymbol{\theta}_{\ell,2} = \boldsymbol{0}_q$.

Next, we show (ii). By applying the general formula for the inverse of a partitioned matrix (e.g., Harville, 1997, Theorem 8.5.11) to $\boldsymbol{\Sigma}^{-1}$, we can obtain the following equation:

$$\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}(\delta - \delta_\ell) = \frac{1}{n}\sum_{i=1}^{q+1} n_i \left(\mu_{\bar{\ell}\cdot\ell}^{(i)} - \frac{1}{n}\sum_{k=1}^{q+1} n_k \mu_{\bar{\ell}\cdot\ell}^{(k)} \right)^2 .$$

From the above equation, Assumption A2 and (9), we can observe that $\max_{i_1 \neq i_2} |\mu_{\bar{\ell}\cdot\ell}^{(i_1)} - \mu_{\bar{\ell}\cdot\ell}^{(i_2)}|$ does not converge to 0 under the HD asymptotic framework. On the other hand, in the HD asymptotic framework, the following equations hold:

$$\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}} \to \boldsymbol{R},\ \boldsymbol{D} \to \boldsymbol{L},$$

where $\boldsymbol{R}$ is a symmetric and idempotent matrix satisfying $\mathrm{rank}(\boldsymbol{R}) = q$, and $\boldsymbol{L}$ is a diagonal matrix whose diagonal elements are positive. Moreover, $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{0}_{q+1}$ holds if and only if $\boldsymbol{a} = a_0 \boldsymbol{L}\boldsymbol{1}_{q+1}$ for some $a_0 \in \mathbb{R}$. These facts lead to the following equation:

$$\inf_{n>p, p\geq 1} (\mu_{\bar{\ell}\cdot\ell}^{(1)}, \ldots, \mu_{\bar{\ell}\cdot\ell}^{(q+1)}) \boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}(\mu_{\bar{\ell}\cdot\ell}^{(1)}, \ldots, \mu_{\bar{\ell}\cdot\ell}^{(q+1)})' > 0.$$

Therefore, using (A.1) and equation $\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \geq \sigma_{\bar{\ell}\bar{\ell}}^{-1}$, we can derive the following equation:

$$\inf_{n>p, p\geq 1} \frac{1}{n}\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2} \geq \sigma_{\bar{\ell}\bar{\ell}}^{-1} \inf_{n>p, p\geq 1} (\mu_{\bar{\ell}\cdot\ell}^{(1)}, \ldots, \mu_{\bar{\ell}\cdot\ell}^{(q+1)}) \boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}(\mu_{\bar{\ell}\cdot\ell}^{(1)}, \ldots, \mu_{\bar{\ell}\cdot\ell}^{(q+1)})' > 0.$$

This completes the proof of (ii).

Finally, we show (iii). When we consider the case of $\ell \supset \bar{j}_*$, we can express $\boldsymbol{x}$ as $\boldsymbol{x} = (\boldsymbol{x}_{\bar{j}_*}', \boldsymbol{x}_{j_*}')' = (\boldsymbol{x}_{\bar{j}_*}', \boldsymbol{x}_{j_*\cap\ell}', x_{\bar{\ell}})'$ without loss of generality. We give the expressions for $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\Sigma}$ corresponding to the partition of $\boldsymbol{x} = (\boldsymbol{x}_{\bar{j}_*}', \boldsymbol{x}_{j_*}')'$ as follows:

$$\boldsymbol{\mathcal{M}} = (\boldsymbol{\mathcal{M}}_{\bar{j}_*}, \boldsymbol{\mathcal{M}}_{j_*}),\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\bar{j}_*\bar{j}_*} & \boldsymbol{\Sigma}_{\bar{j}_*j_*} \\ \boldsymbol{\Sigma}_{\bar{j}_*j_*}' & \boldsymbol{\Sigma}_{j_*j_*} \end{pmatrix}.$$

Let $\boldsymbol{\Gamma}_{\bar{j}_*}$ be a $p \times p$ non-singular matrix given by

$$\boldsymbol{\Gamma}_{\bar{j}_*} = \begin{pmatrix} \boldsymbol{\Sigma}_{\bar{j}_*\bar{j}_*\cdot j_*}^{-1/2} & -\boldsymbol{\Sigma}_{\bar{j}_*\bar{j}_*\cdot j_*}^{-1/2}\boldsymbol{\Sigma}_{\bar{j}_*j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1} \\ \boldsymbol{O}_{p_*, p-p_{j_*}} & \boldsymbol{\Sigma}_{j_*j_*}^{-1/2} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\bar{j}_*\bar{j}_*\cdot j_*} = \boldsymbol{\Sigma}_{\bar{j}_*\bar{j}_*} - \boldsymbol{\Sigma}_{\bar{j}_*j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{\bar{j}_*j_*}'$. And let $\boldsymbol{\Psi}_{\bar{j}_*}$ be a $p \times p$ matrix denoted by

$$\boldsymbol{\Psi}_{\bar{j}_*} = n\boldsymbol{\Gamma}_{\bar{j}_*}\boldsymbol{\mathcal{M}}'\boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}\boldsymbol{\mathcal{M}}\boldsymbol{\Gamma}_{\bar{j}_*}'.$$

From the definition of the true model j_* and the equivalence between (1) and (5), we observe that equation $\boldsymbol{\mathcal{M}}_{\bar{j}_*} - \boldsymbol{\mathcal{M}}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{\bar{j}_*j_*}' = \boldsymbol{O}_{q+1, p-p_{j_*}}$ holds. Then, $n^{-1}\mathrm{tr}(\boldsymbol{\Psi}_{\bar{j}_*})$ can be calculated as

$$\begin{aligned} \frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}_{\bar{j}_*}) &= \mathrm{tr}\left\{ (\boldsymbol{O}_{q+1,p-p_{j_*}}, \boldsymbol{\mathcal{M}}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1/2})'\boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}(\boldsymbol{O}_{q+1,p-p_{j_*}}, \boldsymbol{\mathcal{M}}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1/2}) \right\} \\ &= \mathrm{tr}\left\{ \boldsymbol{\Sigma}_{j_*j_*}^{-1/2}\boldsymbol{\mathcal{M}}_{j_*}'\boldsymbol{D}(\boldsymbol{I}_{q+1} - \boldsymbol{P}_{\boldsymbol{d}})\boldsymbol{D}\boldsymbol{\mathcal{M}}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1/2} \right\} \\ &\to \mathrm{tr}\left(\boldsymbol{\Sigma}_{j_*j_*}^{-1/2}\boldsymbol{\mathcal{M}}_{j_*}'\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}\boldsymbol{\mathcal{M}}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1/2} \right). \end{aligned} \tag{A.2}$$

Using $\boldsymbol{\Psi}_{\bar{j}_*}$, $\mathrm{tr}(\boldsymbol{\Psi}_\ell)$ can be written as

$$\mathrm{tr}(\boldsymbol{\Psi}_\ell) = \mathrm{tr}\left\{\boldsymbol{\Gamma}_\ell \boldsymbol{\Gamma}_{\bar{j}_*}^{-1} \boldsymbol{\Psi}_{\bar{j}_*} (\boldsymbol{\Gamma}'_{\bar{j}_*})^{-1} \boldsymbol{\Gamma}'_\ell\right\}.$$

From the definitions of $\boldsymbol{\Gamma}_\ell$ and $\boldsymbol{\Gamma}_{\bar{j}_*}$, it is straightforward to observe that $\boldsymbol{\Gamma}_\ell \boldsymbol{\Sigma} \boldsymbol{\Gamma}'_\ell = \boldsymbol{\Gamma}_{\bar{j}_*} \boldsymbol{\Sigma} \boldsymbol{\Gamma}'_{\bar{j}_*} = \boldsymbol{I}_p$. Therefore, this fact and (A.2) lead to the following results:

$$\begin{aligned}
\frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}_\ell) &= \frac{1}{n}\mathrm{tr}\left\{\boldsymbol{\Psi}_{\bar{j}_*} (\boldsymbol{\Gamma}'_{\bar{j}_*})^{-1} \boldsymbol{\Gamma}'_\ell \boldsymbol{\Gamma}_\ell \boldsymbol{\Gamma}_{\bar{j}_*}^{-1}\right\} \\
&= \frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}_{\bar{j}_*}) \\
&\to \mathrm{tr}\left(\boldsymbol{\Sigma}_{j_* j_*}^{-1/2} \boldsymbol{\mathcal{M}}'_{j_*} \boldsymbol{L}\boldsymbol{R}\boldsymbol{L} \boldsymbol{\mathcal{M}}_{j_*} \boldsymbol{\Sigma}_{j_* j_*}^{-1/2}\right).
\end{aligned}$$

This completes the proof of (iii). □

B Proof of Theorem 3.1

The probability $P(\hat{j} = j_*)$ can be expressed as

$$P(\hat{j} = j_*) = P\left(\left(\bigcap_{\ell \supset \bar{j}_*} \{\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\}\right) \bigcap \left(\bigcap_{\ell \supset j_*} \{\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) \leq 0\}\right)\right).$$

From basic probability theories, we obtain

$$P(\hat{j} = j_*) \geq 1 - \sum_{\ell \supset \bar{j}_*} P\left(\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) < 0\right) - \sum_{\ell \supset j_*} P\left(\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\right). \tag{B.1}$$

From the basic properties of a multivariate normal distribution and Cochran's Theorem (e.g., Fujikoshi *et al.*, 2010, Theorem 2.4.2), $\boldsymbol{W}$ and $\boldsymbol{B}$ are independent and $\boldsymbol{W} \sim W_p(n-q-1, \boldsymbol{\Sigma})$, $\boldsymbol{B} \sim W_p(q, \boldsymbol{\Sigma}; n\boldsymbol{\Omega})$. Hence, from Lemma C.1 in Appendix C, we can express the distributions of $\boldsymbol{W}$ and $\boldsymbol{B}$ as $\boldsymbol{W} \sim W_p(n-q-1, \boldsymbol{I}_p)$ and $\boldsymbol{B} \sim W_p(q, \boldsymbol{I}_p; \boldsymbol{\Psi}_\ell)$ when $w_{\bar{\ell}\bar{\ell}\cdot\ell}/t_{\bar{\ell}\bar{\ell}\cdot\ell}$. Moreover, from expression (12), we can apply Lemma C.2 in Appendix C to $w_{\bar{\ell}\bar{\ell}\cdot\ell}/t_{\bar{\ell}\bar{\ell}\cdot\ell}$, that is we can express $w_{\bar{\ell}\bar{\ell}\cdot\ell}/t_{\bar{\ell}\bar{\ell}\cdot\ell}$ as follows:

$$\frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{t_{\bar{\ell}\bar{\ell}\cdot\ell}} = \frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{w_{\bar{\ell}\bar{\ell}\cdot\ell} + (t_{\bar{\ell}\bar{\ell}\cdot\ell} - w_{\bar{\ell}\bar{\ell}\cdot\ell})} = \frac{s_e}{s_e + s_h}, \tag{B.2}$$

where s_e and s_h are conditionally independent given $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$, and

$$s_e \sim \chi^2(n-p-q),\ s_h|\boldsymbol{U}_1, \boldsymbol{Z}_1 \sim \chi^2(q; \gamma_\ell),\ \gamma_\ell = \boldsymbol{\theta}'_{\ell,2}\{\boldsymbol{I}_q + \boldsymbol{Z}_1(\boldsymbol{U}'_1\boldsymbol{U}_1)^{-1}\boldsymbol{Z}'_1\}^{-1}\boldsymbol{\theta}_{\ell,2}.$$

Here, $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$ are independent random matrices distributed according to

$$\boldsymbol{U}_1 \sim N_{(n-q-1)\times(p-1)}(\boldsymbol{O}_{n-q-1,p-1}, \boldsymbol{I}_{p-1} \otimes \boldsymbol{I}_{n-q-1}),\ \boldsymbol{Z}_1 \sim N_{q\times(p-1)}(\boldsymbol{\Theta}_{\ell,1}, \boldsymbol{I}_{p-1} \otimes \boldsymbol{I}_q),$$

where $\boldsymbol{\Theta}_{\ell,1}$ and $\boldsymbol{\theta}_{\ell,2}$ are the partitioned matrix and vector, respectively, of $\boldsymbol{\Theta}_\ell$ defined in (12). From (B.2), $\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega)$ can be expressed as

$$\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) = -n\log\frac{w_{\bar{\ell}\bar{\ell}\cdot\ell}}{t_{\bar{\ell}\bar{\ell}\cdot\ell}} - \alpha q = -n\log\frac{s_e}{s_e + s_h} - \alpha q. \tag{B.3}$$

First, we consider the case of $\ell \supset j_*$. Then, $\boldsymbol{\theta}_{\ell,2} = \mathbf{0}_q$ holds from Lemma 3.1-(i). Hence, we observe that s_h is distributed according to $\chi^2(q)$ from Lemma C.2. From expression (B.3), for all $r \in \mathbb{N}$ the following equation can be derived:

$$\begin{aligned}\sum_{\ell \supset j_*} P\left(\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\right) &= (p - p_*)P\left(\frac{s_h}{s_e} > e^{q\alpha/n} - 1\right) \\ &= (p - p_*)P\left(\frac{s_h}{s_e} > \beta\right) \\ &\leq (p - p_*)\beta^{-2r} E\left[\left(\frac{s_h}{s_e}\right)^{2r}\right].\end{aligned}$$

The last inequality is derived by Markov's inequality when $n-p$ is sufficiently large. From the r-th moments of the chi-squared distribution and inverse-chi-squared distribution, it is straightforward to observe that the divergence order of the expectation in the last of the above equations is $O(n^{-2r})$. Therefore, from condition C1, we have

$$\sum_{\ell \supset j_*} P\left(\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\right) = O(pn^{-2r}\beta^{-2r}) \to 0. \tag{B.4}$$

Next, we consider the case of $\ell \supset \bar{j}_*$. Since s_h is conditionally distributed according to $\chi^2(q; \gamma_\ell)$ given $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$, we can express s_h as follows:

$$s_h = \gamma_\ell + t + 2\varepsilon\sqrt{\gamma_\ell},$$

where t and ε are random variables satisfying $t|\boldsymbol{U}_1, \boldsymbol{Z}_1 \sim \chi^2(q)$ and $\varepsilon|\boldsymbol{U}_1, \boldsymbol{Z}_1 \sim N(0,1)$, respectively. Then, the following equation is easily verified:

$$\frac{s_e}{n-p-q} = 1 + o_p(1), \ \frac{t}{n} = o_p(1). \tag{B.5}$$

From Lemma 3.1-(iii), the following equation can be derived:

$$\frac{\gamma_\ell}{n} \leq \frac{\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}}{n} \leq \frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}_\ell) = O(1). \tag{B.6}$$

Hence, from the Cauchy-Schwarz inequality, we derive

$$E\left[\left|\frac{1}{n}\varepsilon\sqrt{\gamma_\ell}\right|\right] \leq \sqrt{E\left[\frac{1}{n}\varepsilon^2\right]} \cdot \sqrt{E\left[\frac{1}{n}\gamma_\ell\right]} = \frac{1}{\sqrt{n}}\sqrt{E\left[\frac{1}{n}\gamma_\ell\right]} = O(n^{-1/2}).$$

This leads to the following equation:

$$\frac{\varepsilon\sqrt{\gamma_\ell}}{n} = o_p(1). \tag{B.7}$$

Let us examine the lower bound of γ_ℓ/n. It is straightforward to observe that

$$\frac{\gamma_\ell}{n} \geq \frac{\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}}{n} \cdot \frac{1}{1 + \mathrm{tr}\{\boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\}}.$$

Then, applying Lemma C.3 to the above equation gives

$$\frac{\gamma_\ell}{n} \geq \frac{\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}}{n}\left\{\frac{n-p}{n-p+qp+\mathrm{tr}(\boldsymbol{\Theta}_{\ell,1}\boldsymbol{\Theta}_{\ell,1}')} + O_p(n^{-1/2})\right\}.$$

From Lemma 3.1-(ii), (iii) and (B.6), the above equation can be expressed as

$$\frac{\gamma_\ell}{n} \ge \left\{ \frac{n-p}{n-p+qp+\mathrm{tr}(\boldsymbol{\Theta}_{\ell,1}\boldsymbol{\Theta}_{\ell,1}')} \right\} \left(\inf_{n>p,p\ge 1} \frac{\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}}{n} \right) + O_p(n^{-1/2}). \tag{B.8}$$

Therefore, by using (B.5)-(B.8), the following equation can be derived:

$$\begin{aligned} \frac{s_h}{s_e} &= \frac{n}{n-p-q} \cdot \frac{n-p-q}{s_e} \cdot \left(\frac{\gamma_\ell}{n} + \frac{t}{n} + \frac{2\varepsilon\sqrt{\gamma_\ell}}{n} \right) \\ &= \frac{n}{n-p-q} \{1+o_p(1)\} \left\{ \frac{\gamma_\ell}{n} + o_p(1) \right\} \\ &\ge \left\{ \frac{n}{n-p+qp+\mathrm{tr}(\boldsymbol{\Theta}_{\ell,1}\boldsymbol{\Theta}_{\ell,1}')} \right\} \left(\inf_{n>p,p\ge 1} \frac{\boldsymbol{\theta}_{\ell,2}'\boldsymbol{\theta}_{\ell,2}}{n} \right) + o_p(1). \end{aligned} \tag{B.9}$$

Since p_{j_*} is finite, the condition C2 and (B.9) lead to

$$\sum_{\ell \supset j_*} P\left(\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\right) = \sum_{\ell \supset j_*} P\left(\frac{s_h}{s_e} < \beta \right) \to 0. \tag{B.10}$$

The equations (B.1), (B.4) and (B.10) complete the proof of Theorem 3.1. □

C Lemma C.1, C.2, C.3 and their proofs

C.1 Lemma C.1 and it's proof

Lemma C.1. *Suppose that $n-p-1>0$. Let $\boldsymbol{W}$ and $\boldsymbol{B}$ be independent random matrices satisfying $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$ and $\boldsymbol{B} \sim W_p(q, \boldsymbol{\Sigma}; \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is a $p\times p$ symmetric matrix and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix. Let $\boldsymbol{\Gamma}$ and the partition of $\boldsymbol{\Sigma}$ be denoted by*

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1/2} & \boldsymbol{0}_{p-1} \\ -\sigma_{22\cdot 1}^{-1/2}\boldsymbol{\sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} & \sigma_{22\cdot 1}^{-1/2} \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{12}' & \sigma_{22} \end{pmatrix},$$

where the size of $\boldsymbol{\Sigma}_{11}$ is $(p-1)\times(p-1)$ and $\sigma_{22\cdot 1} = \sigma_{22} - \boldsymbol{\sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_{12}$. Also, let $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$ and the partitions of $\boldsymbol{W}$ and $\boldsymbol{T}$ corresponding to the division of $\boldsymbol{\Sigma}$ be as follows:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}' & w_{22} \end{pmatrix}, \ \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{t}_{12} \\ \boldsymbol{t}_{12}' & t_{22} \end{pmatrix}.$$

Then, we can regard the distributions of $\boldsymbol{W}$ and $\boldsymbol{B}$ as $\boldsymbol{W} \sim W_p(n, \boldsymbol{I}_p)$ and $\boldsymbol{B} \sim W_p(q, \boldsymbol{I}_p; \boldsymbol{\Gamma\Omega\Gamma}')$ when we consider the distribution of $w_{22\cdot 1}/t_{22\cdot 1}$, where $w_{22\cdot 1} = w_{22} - \boldsymbol{w}_{12}'\boldsymbol{W}_{11}^{-1}\boldsymbol{w}_{12}$ and $t_{22\cdot 1} = t_{22} - \boldsymbol{t}_{12}'\boldsymbol{T}_{11}^{-1}\boldsymbol{t}_{12}$.

Proof. Let

$$\widetilde{\boldsymbol{W}} = \boldsymbol{\Gamma W \Gamma}', \ \widetilde{\boldsymbol{B}} = \boldsymbol{\Gamma B \Gamma}', \ \widetilde{\boldsymbol{T}} = \widetilde{\boldsymbol{W}} + \widetilde{\boldsymbol{B}}.$$

Then, $\widetilde{\boldsymbol{W}}$ and $\widetilde{\boldsymbol{B}}$ are independent, and it follows from $\boldsymbol{\Gamma\Sigma\Gamma}' = \boldsymbol{I}_p$ that $\widetilde{\boldsymbol{W}} \sim W_p(n, \boldsymbol{I}_p)$, $\widetilde{\boldsymbol{B}} \sim W_p(q, \boldsymbol{I}_p; \boldsymbol{\Gamma\Omega\Gamma}')$. Also, let the partitions of $\widetilde{\boldsymbol{W}}$ and $\widetilde{\boldsymbol{T}}$ corresponding to the divisions of $\boldsymbol{W}$ and $\boldsymbol{T}$ be written follows:

$$\widetilde{\boldsymbol{W}} = \begin{pmatrix} \widetilde{\boldsymbol{W}}_{11} & \widetilde{\boldsymbol{w}}_{12} \\ \widetilde{\boldsymbol{w}}_{12}' & \widetilde{w}_{22} \end{pmatrix}, \ \widetilde{\boldsymbol{T}} = \begin{pmatrix} \widetilde{\boldsymbol{T}}_{11} & \widetilde{\boldsymbol{t}}_{12} \\ \widetilde{\boldsymbol{t}}_{12}' & \widetilde{t}_{22} \end{pmatrix}.$$

Note that $\widetilde{\boldsymbol{W}}_{11} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{W}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}$ and $\widetilde{\boldsymbol{T}}_{11} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{T}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}$. Then, by using the general formula for the determinant of a partitioned matrix (e.g., Harville, 1997, Theorem 13.3.8), we have

$$\frac{w_{22\cdot 1}}{t_{22\cdot 1}} = \frac{|\boldsymbol{W}|}{|\boldsymbol{W}_{11}|} \cdot \frac{|\boldsymbol{T}_{11}|}{|\boldsymbol{T}|} = \frac{|\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}'|}{|\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{W}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}|} \cdot \frac{|\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{T}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}|}{|\boldsymbol{\Gamma}\boldsymbol{T}\boldsymbol{\Gamma}'|} = \frac{\widetilde{w}_{22\cdot 1}}{\widetilde{t}_{22\cdot 1}}.$$

This completes the proof of Lemma C.1. □

C.2 Lemma C.2 and it's proof

Lemma C.2. *Suppose that $n-p-q-2 > 0$. Let $\boldsymbol{W}$ and $\boldsymbol{B}$ be independent random matrices satisfying $\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{U}$, $\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{Z}$, $\boldsymbol{U} \sim N_{(n-q-1)\times p}(\boldsymbol{O}_{n-q-1,p}, \boldsymbol{I}_p \otimes \boldsymbol{I}_{n-q-1})$ and $\boldsymbol{Z} \sim N_{q\times p}(\boldsymbol{\Theta}, \boldsymbol{I}_p \otimes \boldsymbol{I}_q)$, where $\boldsymbol{\Theta}$ is a $q \times p$ matrix. Let $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$ then the partitions of $\boldsymbol{U}$, $\boldsymbol{Z}$, $\boldsymbol{\Theta}$, $\boldsymbol{W}$ and $\boldsymbol{T}$ are as follows:*

$$\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{u}_2),\ \boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{z}_2),\ \boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\theta}_2),\ \boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}' & w_{22} \end{pmatrix},\ \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{t}_{12} \\ \boldsymbol{t}_{12}' & t_{22} \end{pmatrix},$$

where the sizes of $\boldsymbol{U}_1$, $\boldsymbol{Z}_1$, $\boldsymbol{\Theta}_1$, $\boldsymbol{W}_{11}$, and $\boldsymbol{T}_{11}$ are $(n-q-1) \times (p-1)$, $q \times (p-1)$, $q \times (p-1)$, $(p-1) \times (p-1)$, and $(p-1) \times (p-1)$, respectively. Also, let $w_{22\cdot 1} = w_{22} - \boldsymbol{w}_{12}'\boldsymbol{W}_{11}^{-1}\boldsymbol{w}_{12}$ and $t_{22\cdot 1} = t_{22} - \boldsymbol{t}_{12}'\boldsymbol{T}_{11}^{-1}\boldsymbol{t}_{12}$. Then, given $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$, $w_{22\cdot 1}$ and $t_{22\cdot 1} - w_{22\cdot 1}$ are conditionally independent, and

$$w_{22\cdot 1} \sim \chi^2(n-p-q),\ t_{22\cdot 1} - w_{22\cdot 1} | \boldsymbol{U}_1, \boldsymbol{Z}_1 \sim \chi^2(q; \gamma),$$

where $\gamma = \boldsymbol{\theta}_2'\{\boldsymbol{I}_q + \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\}^{-1}\boldsymbol{\theta}_2$. Moreover, if $\boldsymbol{\theta}_2 = \boldsymbol{0}_q$, then $t_{22\cdot 1} - w_{22\cdot 1}$ is distributed according to $\chi^2(q)$.

Proof. From the definitions of $w_{22\cdot 1}$ and $t_{22\cdot 1}$, we can express $w_{22\cdot 1}$ and $t_{22\cdot 1}$ as follows:

$$w_{22\cdot 1} = \boldsymbol{u}_2'(\boldsymbol{I}_{n-q-1} - \boldsymbol{P}_{\boldsymbol{U}_1})\boldsymbol{u}_2, \tag{C.1}$$

$$t_{22\cdot 1} = \boldsymbol{u}_2'\boldsymbol{u}_2 + \boldsymbol{z}_2'\boldsymbol{z}_2 - (\boldsymbol{U}_1'\boldsymbol{u}_2 + \boldsymbol{Z}_1'\boldsymbol{z}_2)'(\boldsymbol{U}_1'\boldsymbol{U}_1 + \boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}(\boldsymbol{U}_1'\boldsymbol{u}_2 + \boldsymbol{Z}_1'\boldsymbol{z}_2). \tag{C.2}$$

Since $\boldsymbol{U}_1$ and $\boldsymbol{u}_2$ are independent, we observe that $w_{22\cdot 1} \sim \chi^2(n-p-q)$ from Cochran's Theorem. On the other hand, by the general formula for the inverse of the sum of matrices (e.g., Lütkepohl, 1997, 3.5.2 (2)), the following equation holds:

$$(\boldsymbol{U}_1'\boldsymbol{U}_1 + \boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1} = (\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1} - (\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\{\boldsymbol{I}_q + \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\}^{-1}\boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}.$$

By using the above equation and (C.1) and (C.2), $t_{22\cdot 1} - w_{22\cdot 1}$ is calculated as

$$\begin{aligned} & t_{22\cdot 1} - w_{22\cdot 1} \\ & = \{\boldsymbol{z}_2 - \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\boldsymbol{u}_2\}'\{\boldsymbol{I}_q + \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\}^{-1}\{\boldsymbol{z}_2 - \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\boldsymbol{u}_2\}. \end{aligned} \tag{C.3}$$

From the above equation, given $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$, we can observe that $t_{22\cdot 1} - w_{22\cdot 1}$ is distributed according to $\chi^2(q;\gamma)$. In particular, $t_{22\cdot 1} - w_{22\cdot 1} \sim \chi^2(q)$ holds when $\boldsymbol{\theta}_2 = \boldsymbol{0}_q$.

To show the conditional independence of $w_{22\cdot 1}$ and $t_{22\cdot 1} - w_{22\cdot 1}$, we express (C.1) and (C.3) as follows:

$$w_{22\cdot 1} = (\boldsymbol{z}_2', \boldsymbol{u}_2')\boldsymbol{E}(\boldsymbol{z}_2', \boldsymbol{u}_2')',\ t_{22\cdot 1} - w_{22\cdot 1} = (\boldsymbol{z}_2', \boldsymbol{u}_2')\boldsymbol{F}(\boldsymbol{z}_2', \boldsymbol{u}_2')',$$

where $\boldsymbol{E}$ and $\boldsymbol{F}$ are given by

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{O}_{q,q} & \boldsymbol{O}_{q,n-q-1} \\ \boldsymbol{O}_{n-q-1,q} & \boldsymbol{I}_{n-q-1} - \boldsymbol{P}_{\boldsymbol{U}_1} \end{pmatrix},$$
$$\boldsymbol{F} = \left(\boldsymbol{I}_q, -\boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\right)' \{\boldsymbol{I}_q + \boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{Z}_1'\}^{-1} \left(\boldsymbol{I}_q, -\boldsymbol{Z}_1(\boldsymbol{U}_1'\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\right).$$

It is straightforward to observe that $\boldsymbol{E}$ and $\boldsymbol{F}$ are symmetric and idempotent matrices satisfying $\boldsymbol{E}\boldsymbol{F} = \boldsymbol{O}_{n-1,n-1}$. These imply that $w_{22\cdot 1}$ and $t_{22\cdot 1} - w_{22\cdot 1}$ are conditionally independent given $\boldsymbol{U}_1$ and $\boldsymbol{Z}_1$ from Cochran's Theorem. □

C.3 Lemma C.3 and it's proof

Lemma C.3. *Suppose that $n-p-q-1 > 0$ and $q \le p$. Let $\boldsymbol{\Theta}$ be a $q \times (p-1)$ matrix satisfying* $\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}') = O(n)$. *And let $\boldsymbol{W}$ and $\boldsymbol{Z}$ be independent random matrices distributed according to $\boldsymbol{W} \sim W_{p-1}(n-q-1, \boldsymbol{I}_p)$ and $\boldsymbol{Z} \sim N_{q\times(p-1)}(\boldsymbol{\Theta}, \boldsymbol{I}_{p-1} \otimes \boldsymbol{I}_q)$. Then, we have*

$$\frac{1}{1 + \mathrm{tr}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}')} = \frac{n-p}{n-p+qp+\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}')} + O_p(n^{-1/2}),$$

as $n \to \infty,\ p/n \to c \in [0,1)$.

Proof. Let

$$\boldsymbol{V} = (\boldsymbol{Z}\boldsymbol{Z}')^{1/2}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}')^{-1}(\boldsymbol{Z}\boldsymbol{Z}')^{1/2}.$$

From a property of the Wishart distribution (e.g., Fujikoshi *et al.*, 2010, Theorem 2.3.3), we observe that $\boldsymbol{V}$ is independent of $\boldsymbol{Z}$ and $\boldsymbol{V} \sim W_q(n-p, \boldsymbol{I}_q)$. Then, we have

$$\frac{1}{1 + \mathrm{tr}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}')} = \frac{1}{1 + \mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{Z}')}.$$

We expand $\boldsymbol{V}^{-1}$ and $n^{-1}\boldsymbol{Z}\boldsymbol{Z}'$. Let

$$\boldsymbol{T} = \frac{1}{\sqrt{n-p}}\{\boldsymbol{V} - (n-p)\boldsymbol{I}_q\}.$$

Then, it is straightforward to observe that $\boldsymbol{T} = O_p(1)$, so $\boldsymbol{V}^{-1}$ is expanded as follows:

$$\boldsymbol{V}^{-1} = \frac{1}{n-p}\left(\boldsymbol{I}_q + \frac{1}{\sqrt{n-p}}\boldsymbol{T}\right)^{-1} = \frac{1}{n-p}\boldsymbol{I}_q + O_p(n^{-3/2}). \tag{C.4}$$

From basic properties of a matrix normal distribution, we have

$$E[\boldsymbol{Z}\boldsymbol{Z}'] = (p-1)\boldsymbol{I}_q + \boldsymbol{\Theta}\boldsymbol{\Theta}',$$
$$E[||\boldsymbol{Z}\boldsymbol{Z}' - E[\boldsymbol{Z}\boldsymbol{Z}']||^2] = 2(q+1)\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}') + q(q+1)(p-1) = O(n),$$

where $||\boldsymbol{A}||$ is the Frobenius norm for a matrix $\boldsymbol{A}$ defined by $||\boldsymbol{A}|| = \sqrt{\mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})}$. Hence, $n^{-1}\boldsymbol{Z}\boldsymbol{Z}'$ is expanded as follows:

$$\frac{1}{n}\boldsymbol{Z}\boldsymbol{Z}' = \frac{p}{n}\boldsymbol{I}_q + \frac{1}{n}\boldsymbol{\Theta}\boldsymbol{\Theta}' + O_p(n^{-1/2}). \tag{C.5}$$

From (C.4) and (C.5), $\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{Z}')$ can be expanded as

$$\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{Z}') = \frac{qp}{n-p} + \frac{1}{n-p}\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}') + O_p(n^{-1/2}).$$

Note that

$$\frac{qp}{n-p} + \frac{1}{n-p}\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}') \geq 0.$$

Therefore, we can expand $\{1 + \mathrm{tr}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}')\}^{-1}$ as follows:

$$\frac{1}{1 + \mathrm{tr}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}')} = \frac{n-p}{n-p+qp+\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Theta}')} + O_p(n^{-1/2}).$$

□

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), 995–1010. Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* **AC − 19**, 716–723.

[3] Bai, Z. D., Fujikoshi, Y. & Hu, J. (2018). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. TR No. 18–9, *Statistical Research Group*, Hiroshima University.

[4] Cheng, G., Li, X., Lai, P., Song, F. & Yu, J. (2017). Robust rank screening for ultrahigh dimensional discriminant analysis. *Stat. Comput.*, **27**, 535–545.

[5] Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

[6] Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis. *Ann. Inst. Statist. Math.*, **34**, 523–530.

[7] Fujikoshi, Y. (1983). A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.*, **13**, 203–214.

[8] Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In *Multivariate Analysis VI* (ed. P. R. Krishnaiah), 219–236, NorthHolland, Amsterdam.

[9] Fujikoshi, Y. & Sakurai, T. (2018). Consistency of test-based criterion for selection of variables in high-dimensional two group-discriminant analysis. *J. J. S. D.* (to appear in).

[10] Fujikoshi, Y., Ulyanov, V. V. & Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* John Wiley & Sons, Inc., Hoboken, New Jersey.

[11] Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B*, **26**, 270–273.

[12] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York.

[13] Hastie, T., Buja, A. & Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.

[14] Lütkepohl, H. (1997). *Handbook of Matrices.* Wiley, Chichester.

[15] Nishii, R., Bai, Z. D. & Krishnaiah, P. R. (1988). Strong consistency information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.

[16] Oda, R. & Yanagihara, H. (2019). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. TR No. 19–1, *Statistical Research Group*, Hiroshima University.

[17] Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.

[18] Rao, C. R. (1973). *Linear statistical inference and its applications,* (2nd ed.). Wiley, New York.

[19] Sakurai, T. & Fujikoshi, Y. (2017). High-dimensional properties of information criteria and their efficient criteria for multivariate linear regression models with covariance structures. TR No. 17–13, *Statistical Research Group*, Hiroshima University.

[20] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

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