On model selection consistency using a kick-one-out method for selecting response variables in high-dimensional multivariate linear regression

Ryoya Oda* [†] **Hirokazu Yanagihara**[‡] **and Yasunori Fujikoshi**[§]

(Last Modified: October 8, 2021)

Abstract

This paper deals with the selection of non-redundant response variables in normality-assumed multivariate linear regression, where the redundancy of the response variables is defined by conditional expectation. A sufficient condition for model selection consistency is obtained using a kick-one-out method based on the generalized information criterion under a high-dimensional asymptotic framework such that the sample size tends to infinity, and the number of response variables and explanatory variables does not exceed the sample size but may tend to infinity. A consistent kick-one-out method using the obtained condition is proposed. Simulation results show that the proposed method has a high probability of selecting true non-redundant variables.

1 Introduction

The multivariate linear regression model is fundamental in inferential statistical analysis and is introduced in many statistics textbooks (e.g., [20, 22]). Suppose that $\boldsymbol{y}$ is a p-dimensional response vector and $\boldsymbol{x}$ is a k-dimensional explanatory vector. Then, the normality-assumed multivariate linear regression model with $\boldsymbol{y}$ and $\boldsymbol{x}$ is given by

$$\boldsymbol{y} \sim \mathcal{N}_p(\boldsymbol{\Theta}'\boldsymbol{x}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ unknown regression coefficients matrix and $\boldsymbol{\Sigma}$ is a $p \times p$ unknown covariance matrix that is positive definite. In actual empirical contexts, selecting variables for the model is of key interest. It is generally expected that the accuracy of prediction is improved and that interpretation of the model is made easier by proper variable selection. In the literature on multivariate linear regression, numerous papers have dealt with the variable selection problem as it relates to selecting explanatory variables. On the other hand, it is also important to consider

*Corresponding author. Email: ryoya-oda@hiroshima-u.ac.jp

[†]School of Informatics and Data Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

[‡]Graduate School of Advanced Science and Engineering, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

[§]Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

the problem of selecting appropriate response variables, as including redundant response variables in the model will almost certainly lead to an improper understanding of the essential structure of the model. [8] defined redundancy using conditional expectation. Specifically, a response variable is determined to be redundant when the conditional expectation of the response variable given the non-redundant variables does not rely on the explanatory variables. This implies that the relationship between the redundant variables and the explanatory variables can be expressed by the relationship between the non-redundant variables and the explanatory variables. Thus, removing such redundant variables will promote a better understanding of the essential structure of the model.

As a result of the rapid advances in information technology and database systems in recent years, efforts to analyze high-dimensional data where not only the sample size but also the number of variables is large have become commonplace. Although [8] derived a formal Akaike's information criterion (AIC) [1, 2] for selecting response variables, the behavior of the formal AIC is not apparent for high-dimensional data. Moreover, variable selection using criteria such as the AIC is often performed by minimizing the criterion for all combinations of response variables, i.e., all subsets of ω, where $\omega = \{1, \ldots, p\}$ is defined as the full set of response variable suffixes. This means that it may be impossible to apply the approach due to its high computational cost when the number of response variables is large. For this reason, we focus on the practicable selection method proposed by [12, 24], using the kick-one-out (KOO) method described in [3]. With the KOO method, if the criterion value for the subset with one variable removed from the full set ω is greater than the criterion value for ω, then the removed variable is selected. This makes it possible to apply the KOO method for high-dimensional data without the risk of incurring excessive computational costs.

In this paper, we propose the use of the generalized information criterion (GIC) developed in [11] as a variable selection criterion to address the redundancy issue described in [8]. Specifically, we propose a KOO method based on the GIC which has a model selection consistency property under a high-dimensional asymptotic framework such that the sample size always tends to infinity and the number of response variables and explanatory variables does not exceed the sample size but may tend to infinity. Consistency here means that the probability that the selected variables are identical to the true non-redundant variables tends to 1 under a high-dimensional asymptotic framework. Thus, it can be expected that the proposed method has a high probability of selecting true non-redundant variables when the sample size is large, regardless of the number of variables. In several previous works, a consistency property of the KOO method has been obtained for selection problems under a high-dimensional asymptotic framework in several multivariate models, e.g., selecting the explanatory variables in a multivariate linear regression model [3, 15, 16, 18] and selecting redundant variables in discriminant analysis [7, 10, 14, 17]. However, there is no consistent KOO method for selecting response variables in the sense of the redundancy described in [8].

The remainder of the paper is organized as follows: In section 2, we define the redundancy of response variables and introduce the KOO method based on the GIC. In section 3, we describe the model selection consistency property of the KOO method based on the GIC under a high-dimensional asymptotic framework. In section 4, we conduct numerical experiments for verification purposes. Technical details are relegated to the Appendix.

2 Framework for selecting response variables

2.1 Redundancy of response variables

We define the redundancy of response variables as in [8]. Let j denote a subset of $\omega = \{1, \ldots, p\}$ containing p_j elements and $\boldsymbol{y}_j$ denote the p_j-dimensional vector consisting of the components of $\boldsymbol{y}$ indexed by the elements of j. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{y}_j$ consists of the first, second, and fourth elements of $\boldsymbol{y}$. Without loss of generality, we sort $\boldsymbol{y}$ into $\boldsymbol{y} = (\boldsymbol{y}_j', \boldsymbol{y}_{\bar{j}}')'$ for a subset j, where $\boldsymbol{y}_{\bar{j}}$ is the $(p - p_j)$-dimensional vector and $\bar{A}$ denotes the complement of a set A. Similar to the division of $\boldsymbol{y}$, we can express the division of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\Theta} = (\boldsymbol{\Theta}_j, \boldsymbol{\Theta}_{\bar{j}}), \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{jj} & \boldsymbol{\Sigma}_{j\bar{j}} \\ \boldsymbol{\Sigma}_{j\bar{j}}' & \boldsymbol{\Sigma}_{\bar{j}\bar{j}} \end{pmatrix}, \tag{2}$$

where $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Theta}_{\bar{j}}$ are $k \times p_j$ and $k \times (p - p_j)$, $\boldsymbol{\Sigma}_{jj}$ and $\boldsymbol{\Sigma}_{\bar{j}\bar{j}}$ are $p_j \times p_j$ and $(p - p_j) \times (p - p_j)$, and $\boldsymbol{\Sigma}_{j\bar{j}}$ is $p_j \times (p - p_j)$. Then, from a property of a multivariate normal distribution (e.g., [21]), the conditional distribution of $\boldsymbol{y}_{\bar{j}}$ given $\boldsymbol{y}_j$ can be written as

$$\boldsymbol{y}_{\bar{j}} | \boldsymbol{y}_j \sim \mathcal{N}_{p-p_j} \left((\boldsymbol{\Theta}_{\bar{j}} - \boldsymbol{\Theta}_j \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\Sigma}_{j\bar{j}})' \boldsymbol{x} + \boldsymbol{\Sigma}_{j\bar{j}}' \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{y}_j, \boldsymbol{\Sigma}_{\bar{j}\bar{j} \cdot j} \right), \tag{3}$$

where $\boldsymbol{\Sigma}_{\bar{j}\bar{j} \cdot j} = \boldsymbol{\Sigma}_{\bar{j}\bar{j}} - \boldsymbol{\Sigma}_{j\bar{j}}' \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\Sigma}_{j\bar{j}}$. From (3), if the equation $\boldsymbol{\Theta}_{\bar{j}} - \boldsymbol{\Theta}_j \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\Sigma}_{j\bar{j}} = \boldsymbol{O}_{k,p-p_j}$ holds, then the conditional distribution of $\boldsymbol{y}_{\bar{j}}$ given $\boldsymbol{y}_j$ does not depend on the explanatory vector $\boldsymbol{x}$. In other words, the relationship between $\boldsymbol{y}_{\bar{j}}$ and $\boldsymbol{x}$ can be described by the relationship between $\boldsymbol{y}_j$ and $\boldsymbol{x}$. In this sense, the model such that $\boldsymbol{y}_{\bar{j}}$ is redundant in the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ may be expressed as

$$M_j : \ (3) \ s.t. \ \boldsymbol{\Theta}_{\bar{j}} - \boldsymbol{\Theta}_j \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\Sigma}_{j\bar{j}} = \boldsymbol{O}_{k,p-p_j}. \tag{4}$$

We note that (4) is also related to selecting response variables in a multivariate inverse regression (e.g., [5, 6, 13]). In fact, one of the purposes in a multivariate inverse regression is to estimate an unknown explanatory vector $\boldsymbol{x}_0$ corresponding to an observed response vector $\boldsymbol{y}_0$. Then, if $\boldsymbol{\Theta}$ is full row rank, the classical estimator $\hat{\boldsymbol{x}}_0$ of $\boldsymbol{x}_0$ can be expressed as

$$\hat{\boldsymbol{x}}_0 = \arg\min_{\boldsymbol{x}_0} (\boldsymbol{y}_0 - \boldsymbol{\Theta}' \boldsymbol{x}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_0 - \boldsymbol{\Theta}' \boldsymbol{x}_0) = (\boldsymbol{\Theta} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Theta}')^{-1} \boldsymbol{\Theta} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_0 = \boldsymbol{\Xi}_j \boldsymbol{y}_{0,j} + \boldsymbol{\Xi}_{\bar{j}} \boldsymbol{y}_{0,\bar{j}},$$

where $(\boldsymbol{\Xi}_j, \boldsymbol{\Xi}_{\bar{j}}) = (\boldsymbol{\Theta} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Theta}')^{-1} \boldsymbol{\Theta} \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{y}_0 = (\boldsymbol{y}_{0,j}', \boldsymbol{y}_{0,\bar{j}}')'$. Moreover, it holds that $\boldsymbol{\Xi}_{\bar{j}} = \boldsymbol{\Theta}_{\bar{j}} - \boldsymbol{\Theta}_j \boldsymbol{\Sigma}_{jj}^{-1} \boldsymbol{\Sigma}_{j\bar{j}}$. Hence, the redundancy model (4) can be also regarded as the redundancy of response variables for estimating the classical estimator $\hat{\boldsymbol{x}}_0$ in a multivariate inverse regression.

2.2 Kick-one-out method based on GIC

We first introduce the generalized information criterion (GIC) for the redundancy model (4). Let $\{(\boldsymbol{y}_{(i)}, \boldsymbol{x}_{(i)})\}$ $(i = 1, \ldots, n)$ be a set of *i.i.d.* copies from $(\boldsymbol{y}, \boldsymbol{x})$ and let $\boldsymbol{Y} = (\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(n)})'$ and $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})'$, where n is the sample size. We assume $\mathrm{rank}(\boldsymbol{X}) = k < n$ and $n - p - k - 1 > 0$ for the existence of the GIC and our proposed method. Suppose $\boldsymbol{Y}_j$ and $\boldsymbol{Y}_{\bar{j}}$ are the $n \times p_j$ and $n \times (p - p_j)$ partitioned matrices of $\boldsymbol{Y} = (\boldsymbol{Y}_j, \boldsymbol{Y}_{\bar{j}})$ for the redundancy

model (4). The model (4) can be regarded as a redundancy model as in estimating an unknown explanatory vector $\boldsymbol{x}_0$ by using the classical estimator $\hat{\boldsymbol{x}}_0$ in a multivariate inverse regression. Hence, in accordance with [13], the maximum log-likelihood of $\mathcal{N}_{n\times p}(\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma}\otimes\boldsymbol{I}_n)$ under the redundant model (4) is given by

$$np(\log 2\pi + 1) + n\log\left|\frac{1}{n}\boldsymbol{Y}_j'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_j\right| + n\log\left|\frac{1}{n}\boldsymbol{Y}_{\bar{j}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_j})\boldsymbol{Y}_{\bar{j}}\right|,$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ for a square matrix $\boldsymbol{A}$. Then, the GIC for the model (4) is written as

$$\begin{aligned}\mathrm{GIC}(j) = {} & np(\log 2\pi + 1) + n\log\left|\frac{1}{n}\boldsymbol{Y}_j'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_j\right| + n\log\left|\frac{1}{n}\boldsymbol{Y}_{\bar{j}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_j})\boldsymbol{Y}_{\bar{j}}\right| \\ & + \alpha\left\{\frac{1}{2}p(p+1) + kp_j\right\}, \end{aligned} \tag{5}$$

where α is a positive constant expressing a penalty for the complexity of the model. Various variable selection criteria are included in the GIC by specifying α. For example, the AIC [1, 2], HQC [9], BIC [19], and CAIC [4] are expressed by the GIC for the following α:

$$\alpha = \begin{cases} 2 & (\mathrm{AIC}) \\ 2\log\log n & (\mathrm{HQC}) \\ \log n & (\mathrm{BIC}) \\ 1 + \log n & (\mathrm{CAIC}) \end{cases}.$$

Next, we introduce the kick-one-out method based on the GIC. Denote by ℓ the subset of ω satisfying $\#(\ell) = p - 1$ and, moreover, denote by $\bar{\ell}$ the compliment of ℓ. Then, the best subset decided by the kick-one-out method based on the GIC can be written as

$$\hat{j} = \{\bar{\ell} \in \omega \mid \mathrm{GIC}(\ell) > \mathrm{GIC}(\omega)\}, \tag{6}$$

where $\mathrm{GIC}(\omega)$ is the GIC for the full set ω and is defined for the model such that all the response variables are non-redundant, i.e., $\mathrm{GIC}(\omega)$ is written as

$$\mathrm{GIC}(\omega) = np(\log 2\pi + 1) + n\log\left|\frac{1}{n}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\right| + \alpha\left\{\frac{1}{2}p(p+1) + kp\right\}.$$

3 Model selection consistency of the KOO method based on GIC

In this section, we demonstrate the model selection consistency property of the KOO method based on the GIC defined by (6). First, we define the notation used in deriving the consistency property. Let $\mathcal{J}_+$ be the family of sets consisting of the overspecified subsets; that is, $\mathcal{J}_+$ is given by

$$\mathcal{J}_+ = \{j \subset \omega \mid \boldsymbol{\Theta}_{\bar{j}} - \boldsymbol{\Theta}_j\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{j\bar{j}} = \boldsymbol{O}_{k,p-p_j}\}.$$

Then, the true model j_* is defined as the subset such that the number of elements is smallest for all the subsets in $\mathcal{J}_+$; that is, j_* is given by $j_* = \arg\min_{j\in\mathcal{J}_+} p_j$. The following assumptions are made:

Assumption A1. $n \to \infty,\ p/n \to c_1 \in [0,1),\ k/n \to c_2 \in [0, 1-c_1)$.

Assumption A2. *The true subset j_* is included in the full set ω, i.e., $j_* \subset \omega$.*

Assumption A1 is our high-dimensional asymptotic framework. This assumption means that n always tends to infinity, but p and k can be fixed or can tend to infinity. Note that p_{j_*} is finite because we suppose that the number of non-redundant variables is small. Assumption A2 is needed to consider the model's selection consistency $P(\hat{j} = j_*) \to 1$.

Next, we re-form our expression of (6) to obtain the consistency property. For $j = \ell$ $(\#(\ell) = p-1)$, the divisions of $\boldsymbol{Y}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ are given by

$$\boldsymbol{Y} = (\boldsymbol{Y}_\ell, \boldsymbol{y}_{\bar{\ell}}),\ \boldsymbol{\Theta} = (\boldsymbol{\Theta}_\ell, \boldsymbol{\theta}_{\bar{\ell}}),\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\ell\ell} & \boldsymbol{\sigma}_{\ell\bar{\ell}} \\ \boldsymbol{\sigma}'_{\ell\bar{\ell}} & \sigma_{\bar{\ell}\bar{\ell}} \end{pmatrix}.$$

By considering a distributional reduction of $\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega)$, the best subset $\hat{j}$ can be rewritten as follows (the proof is given in Appendix 1).

Lemma 1. *For $\ell \subset \omega$ $(\#(\ell) = p-1)$, let s_ℓ be a random variable distributed according to the chi-squared distribution $\chi^2(n-p-k+1)$. Further, let h_ℓ be a random variable such that it is conditionally independent of s_ℓ given $\boldsymbol{Y}_\ell$, and the conditional distribution given $\boldsymbol{Y}_\ell$ is the non-central chi-squared distribution $\chi^2(k; \delta_\ell)$. Here, the non-centrality parameter δ_ℓ is defined by*

$$\delta_\ell = \sigma^{-1}_{\bar{\ell}\bar{\ell}\cdot\ell} \boldsymbol{\xi}'_\ell \boldsymbol{X}' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{X} \boldsymbol{\xi}_\ell,$$

where $\sigma_{\bar{\ell}\bar{\ell}\cdot\ell} = \sigma_{\bar{\ell}\bar{\ell}} - \boldsymbol{\sigma}'_{\ell\bar{\ell}} \boldsymbol{\Sigma}^{-1}_{\ell\ell} \boldsymbol{\sigma}_{\ell\bar{\ell}}$ and $\boldsymbol{\xi}_\ell = \boldsymbol{\theta}_{\bar{\ell}} - \boldsymbol{\Theta}_\ell \boldsymbol{\Sigma}^{-1}_{\ell\ell} \boldsymbol{\sigma}_{\ell\bar{\ell}}$. Then, the best subset $\hat{j}$ can be expressed as

$$\hat{j} = \left\{ \bar{\ell} \in \omega \mid \frac{h_\ell}{s_\ell} > \exp\left(\frac{k}{n}\alpha\right) - 1 \right\}.$$

From Lemma 1, we see that the behavior of the non-centrality parameter δ_ℓ plays an important role in variable selection. Some of the properties of δ_ℓ are described below (the proof is given in Appendix 4).

Proposition 1. *For $\ell \subset \omega$ $(\#(\ell) = p-1)$, the following properties of δ_ℓ hold.*

(i) *If $\bar{\ell} \notin j_*$, then $\boldsymbol{\xi}_\ell = \boldsymbol{0}_k$, and hence $\delta_\ell = 0$.*

(ii) *If $\bar{\ell} \in j_*$, then we can express δ_ℓ under Assumption A1 as follows:*

$$\delta_\ell = O_p\left(\mathrm{tr}(\boldsymbol{\Theta}_{j_*} \boldsymbol{\Sigma}^{-1}_{j_* j_*} \boldsymbol{\Theta}'_{j_*}) \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})\right),$$

where $\lambda_{\max}(\boldsymbol{A})$ is the maximum eigenvalue of a square matrix $\boldsymbol{A}$, and $\boldsymbol{\Theta}_{j_*}$ and $\boldsymbol{\Sigma}_{j_* j_*}$ are the partition matrices $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Sigma}_{jj}$ in (2) corresponding to $j = j_*$.

From (i) in Proposition 1, the non-centrality parameter δ_ℓ is 0 when $\bar{\ell} \notin j_*$. Moreover, from (ii) it may be expected that δ_ℓ diverges at about the same speed as the maximum eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$ when $\bar{\ell} \in j_*$. In fact, if $\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) = O(n)$ and $\lambda_{\max}(\boldsymbol{\Theta}'_{j_*} \boldsymbol{\Theta}_{j_*}) = O(1)$, then $\delta_\ell = O_p(n)$. However, we cannot determine the exact divergence speed of δ_ℓ because δ_ℓ includes the inverse of a non-central Wishart matrix. Hence, we require the following additional assumption.

Assumption A3. *There exists* $\tau > 0$ *such that* $P\left(\min_{\bar{\ell}\in j_*} \delta_\ell > \tau n\right) \to 1$.

Assumption A3 means that $n^{-1}\delta_\ell$ does not tend to 0 asymptotically in the sense of the convergence in probability, and this will be natural from (ii) in Proposition 1. In the selection contexts for explanatory variables (e.g., [3, 23]), it is assumed that the corresponding non-central parameters have $O(n)$ and do not converge to 0. Even from that point of view, we consider that Assumption A3 will be appropriate. Moreover, it is possible to relax Assumption A3 and express it as $P\left(\min_{\bar{\ell}\in j_*} \delta_\ell > \tau n^\kappa\right) \to 1$, where κ is a constant satisfying $1/2 < \kappa < 1$. However, we do not consider such a relaxation, as it makes conditions for the consistency of the KOO method based on the GIC stricter; that is, the relaxation and the conditions are related to the transactions.

Finally, we can demonstrate the consistency property of the KOO method $\hat{j}$ based on the GIC under Assumptions A1-A3. To do so, we use the following expression for α:

$$\alpha = \frac{n}{k}\log\left(1 + \frac{k}{N-2} + \beta\right),\ \beta > 0, \tag{7}$$

where $N = n - p - k + 1$. Then, from Lemma 1 and the definition of the method (6), the lower bound of the probability $P(\hat{j} = j_*)$ can be derived as

$$\begin{aligned} P(\hat{j} = j_*) &= P\left(\left(\bigcap_{\bar{\ell}\notin j_*} \{\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) \le 0\}\right) \bigcap \left(\bigcap_{\bar{\ell}\in j_*} \{\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) > 0\}\right)\right) \\ &\ge 1 - P\left(\bigcup_{\bar{\ell}\notin j_*} \left\{\frac{h_\ell}{s_\ell} > \beta + \frac{k}{N-2}\right\}\right) - P\left(\bigcup_{\bar{\ell}\in j_*} \left\{\frac{h_\ell}{s_\ell} \le \beta + \frac{k}{N-2}\right\}\right), \end{aligned} \tag{8}$$

where h_ℓ and s_ℓ are defined in Lemma 1. Therefore, we look for the condition on β such that the two probabilities in the last expression of (8) tend to 0, in order to have $\hat{j}$ manifest the consistency property. Such results are derived in Theorem 1 (the proof is given in Appendix 3).

Theorem 1. *Suppose that Assumptions A1-A3 hold. Then, the KOO method based on the GIC defined by* (6) *exhibits model selection consistency, i.e.,* $P(\hat{j} = j_*) \to 1$ *holds, if for some* $r \in \mathbb{N}$ *the following condition on* α *is satisfied:*

$$\alpha = \frac{n}{k}\log\left(1 + \frac{k}{N-2} + \beta\right),\ \beta > 0\ s.t.\ \frac{n}{p^{1/2r}k^{1/2}}\beta \to \infty,\ \beta \to 0. \tag{9}$$

Theorem 1 specifies the α condition necessary for the KOO method based on the GIC to have the consistency property. Thus, we propose a KOO method based on the GIC with α that satisfies (9). However, we need to decide on a value of α in order to perform variable selection. An example of α satisfying (9) is given as follows:

$$\alpha = \tilde{\alpha} = \frac{n}{k}\log\left(1 + \frac{k}{N-2} + \beta\right),\ \beta = \frac{p^{1/4}k^{1/2}(\log p)(\log n)}{n}. \tag{10}$$

Note that $\tilde{\alpha}$ satisfies condition (9) for $r \ge 2$. The value β in (10) may satisfy both $n\beta/(p^{1/4}k^{1/2}) \to \infty$ and $\beta \to 0$ in a well-balanced manner. Although the consistency property in Theorem 1 is obtained by rewriting α as β, we can derive the consistency property without β if k is fixed.

Corollary 1. *Suppose that Assumptions A1-A3 hold and assume that k is fixed. Then, the KOO method based on the GIC defined by* (6) *exhibits consistency, i.e., $P(\hat{j} = j_*) \to 1$ holds, if for some $r \in \mathbb{N}$ the following condition of α is satisfied:*

$$\frac{1}{p^{1/2r}}\alpha \to \infty, \ \frac{1}{n}\alpha \to 0. \tag{11}$$

Moreover, α satisfying condition (9) *meets* (11) *if k is fixed.*

The proof of Corollary 1 is omitted since it is similar to the proof of Theorem 1. From Corollary 1, we see that the AIC, HQC, BIC, and CAIC do not satisfy condition (11) unless p is fixed or p diverges at a very slow speed.

4 Numerical Studies

In this section, we report the numerical results of a simulation experiment in which the probabilities of selecting the true subset j_* using the KOO method based on the GIC were determined. Let $\tilde{\alpha}$ be the α given by (10), and note that $\tilde{\alpha}$ satisfies the condition of consistency (9). The probabilities are calculated using Monte Carlo simulations with 10,000 iterations. For comparison, we calculated the results for several KOO methods based on GICs for specific α values, including the AIC ($\alpha = 2$), HQC ($\alpha = 2\log\log n$), BIC ($\alpha = \log n$), and CAIC ($\alpha = 1 + \log n$). Moreover, we implemented the KOO method based on the GIC for $\alpha = n^{1/2}$, which, from Corollary 1, is consistent when k is fixed. We set the true subset and the number of true response variables as $j_* = \{1, 2, 3, 4, 5\}$ and $p_{j_*} = 5$, respectively. The response matrix $\boldsymbol{Y}$ is generated from $\mathcal{N}_{n\times p}(\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$, where $\boldsymbol{X}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ are determined as shown below. The explanatory matrix $\boldsymbol{X}$ is generated from $\mathcal{N}_{n\times k}(\boldsymbol{O}_{n,k}, \boldsymbol{\Phi} \otimes \boldsymbol{I}_n)$, where the (a, b)-th element of the $k \times k$ matrix $\boldsymbol{\Psi}$ is $(0.5)^{|a-b|}$. The covariance matrix $\boldsymbol{\Sigma}$ is given by $\boldsymbol{\Sigma} = 0.4\{(1 - 0.8)\boldsymbol{I}_p + 0.8\boldsymbol{1}_p\boldsymbol{1}_p'\}$. Let the partition matrices of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ be as (2) corresponding to $j = j_*$; then the partition matrices of $\boldsymbol{\Theta}$ are given by

$$\boldsymbol{\Theta}_{j_*} \sim \mathcal{N}_{k\times p_{j_*}}(\boldsymbol{O}_{k,p_{j_*}}, \boldsymbol{I}_{p_{j_*}} \otimes \boldsymbol{I}_k), \ \boldsymbol{\Theta}_{\bar{j}_*} = \boldsymbol{\Theta}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*\bar{j}_*}.$$

Tables 1 and 2 show the probabilities for selecting the true subset j_* by KOO methods based on the six criteria, AIC, HQC, BIC, CAIC, GIC with $\alpha = n^{1/2}$, and GIC with $\alpha = \tilde{\alpha}$, where $\tilde{\alpha}$ is given by (10). It is evident from the tables that the probabilities for the KOO method based on the GIC with $\alpha = \tilde{\alpha}$ are high in all cases. This supports our assertion that the method with $\alpha = \tilde{\alpha}$ is consistent under a high-dimensional asymptotic framework. The KOO methods based on the AIC, HQC, BIC, and CAIC, which do not satisfy (9), have low probabilities in at least some instances. The probabilities associated with the KOO method based on the GIC with $\alpha = n^{1/2}$ are high when k is small, as $\alpha = n^{1/2}$ satisfies condition (11). Based on these results, we recommend using $\alpha = \tilde{\alpha}$ to find the true subset.

Table 1: True subset selection probabilities (%) by the KOO methods based on the six criteria.

n	p	k	AIC	HQC	BIC	CAIC	$\alpha = n^{1/2}$	$\alpha = \tilde{\alpha}$
100	10	10	68.78	98.25	99.97	100.00	100.00	100.00
300	10	10	81.54	99.86	100.00	100.00	100.00	100.00
500	10	10	84.22	99.93	100.00	100.00	100.00	100.00
1000	10	10	85.15	100.00	100.00	100.00	100.00	100.00
3000	10	10	85.84	100.00	100.00	100.00	100.00	100.00
5000	10	10	85.56	100.00	100.00	100.00	100.00	100.00
10000	10	10	86.18	100.00	100.00	100.00	100.00	100.00
100	40	10	0.00	9.98	84.49	97.62	99.79	99.92
300	120	10	0.00	7.19	96.98	99.70	100.00	100.00
500	200	10	0.00	4.61	98.51	99.88	100.00	100.00
1000	400	10	0.00	2.01	99.63	99.94	100.00	100.00
3000	1200	10	0.00	0.13	99.91	100.00	100.00	100.00
5000	2000	10	0.00	0.03	99.93	100.00	100.00	100.00
10000	4000	10	0.00	0.00	99.99	100.00	100.00	100.00
100	10	40	71.86	99.96	100.00	97.78	0.00	99.54
300	10	120	99.80	100.00	100.00	47.78	0.00	100.00
500	10	200	100.00	100.00	99.48	0.82	0.00	100.00
1000	10	400	100.00	100.00	23.00	0.00	0.00	100.00
3000	10	1200	100.00	100.00	0.00	0.00	0.00	100.00
5000	10	2000	100.00	100.00	0.00	0.00	0.00	100.00
10000	10	4000	100.00	100.00	0.00	0.00	0.00	100.00
100	20	20	19.46	96.12	99.99	100.00	37.67	100.00
300	60	60	25.69	100.00	100.00	100.00	0.00	100.00
500	100	100	56.37	100.00	100.00	100.00	0.00	100.00
1000	200	200	95.81	100.00	100.00	100.00	0.00	100.00
3000	600	600	100.00	100.00	100.00	99.98	0.00	100.00
5000	1000	1000	100.00	100.00	100.00	75.95	0.00	100.00
10000	2000	2000	100.00	100.00	100.00	0.00	0.00	100.00

Table 2: True subset selection probabilities (%) by the KOO methods based on the six criteria.

n	p	k	AIC	HQC	BIC	CAIC	$\alpha = n^{1/2}$	$\alpha = \tilde{\alpha}$
100	80	10	0.00	0.00	0.00	0.00	2.21	34.80
300	240	10	0.00	0.00	0.00	0.00	91.30	99.94
500	400	10	0.00	0.00	0.00	0.00	99.74	100.00
1000	800	10	0.00	0.00	0.00	0.00	100.00	100.00
3000	2400	10	0.00	0.00	0.00	0.00	100.00	100.00
5000	4000	10	0.00	0.00	0.00	0.00	100.00	100.00
10000	8000	10	0.00	0.00	0.00	0.00	100.00	100.00
100	10	80	0.00	27.91	83.72	4.83	0.00	28.62
300	10	240	0.07	100.00	0.00	0.00	0.00	50.04
500	10	400	0.08	100.00	0.00	0.00	0.00	61.44
1000	10	800	0.28	100.00	0.00	0.00	0.00	73.37
3000	10	2400	0.52	100.00	0.00	0.00	0.00	87.57
5000	10	4000	0.47	100.00	0.00	0.00	0.00	91.32
10000	10	8000	0.38	100.00	0.00	0.00	0.00	94.46
100	40	40	0.00	0.10	81.39	85.35	0.00	70.28
300	120	120	0.00	5.61	98.03	24.26	0.00	99.82
500	200	200	0.00	67.13	87.81	0.52	0.00	100.00
1000	400	400	0.00	99.98	7.78	0.00	0.00	100.00
3000	1200	1200	0.00	100.00	0.00	0.00	0.00	100.00
5000	2000	2000	0.00	100.00	0.00	0.00	0.00	100.00
10000	4000	4000	0.00	100.00	0.00	0.00	0.00	100.00

Appendix 1: Proof of Lemma 1

From (5), the GIC for $j = \ell$ ($\#(\ell) = p - 1$) is written as

$$\begin{aligned}\mathrm{GIC}(\ell) =& np(\log 2\pi + 1) + n \log \left| \frac{1}{n} \boldsymbol{Y}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{Y}_\ell \right| + n \log \left\{ \frac{1}{n} \boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{y}_{\bar{\ell}} \right\} \\ &+ \alpha \left\{ \frac{1}{2} p(p+1) + k(p-1) \right\}.\end{aligned}$$

On the other hand, the GIC for the full set ω can also be expressed as

$$\begin{aligned}\mathrm{GIC}(\omega) =& np(\log 2\pi + 1) + n \log \left| \frac{1}{n} \boldsymbol{Y}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{Y}_\ell \right| + n \log \left\{ \frac{1}{n} \boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{y}_{\bar{\ell}} \right\} \\ &+ \alpha \left\{ \frac{1}{2} p(p+1) + kp \right\}.\end{aligned}$$

Hence, $\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega)$ is expressed by

$$\begin{aligned}\mathrm{GIC}(\ell) - \mathrm{GIC}(\omega) &= n \log \frac{\boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{y}_{\bar{\ell}}}{\boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{y}_{\bar{\ell}}} - k\alpha \\ &= n \log \left\{ 1 + \frac{\boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{y}_{\bar{\ell}}}{\boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{y}_{\bar{\ell}}} \right\} - k\alpha.\end{aligned}$$

Let $\tilde{s}_\ell = \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{y}_{\bar{\ell}}$ and $\tilde{h}_\ell = \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{y}_{\bar{\ell}}' (\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{y}_{\bar{\ell}}$. We can show that $\tilde{s}_\ell$ and $\tilde{h}_\ell$ satisfy the requirements imposed on s_ℓ and h_ℓ in Lemma 1, respectively.

First, we consider $\tilde{s}_\ell$. From a property of a conditional distribution of a multivariate normal distribution (e.g., [21]), $\boldsymbol{y}_{\bar{\ell}}$ can be expressed as

$$\boldsymbol{y}_{\bar{\ell}} = \boldsymbol{Y}_\ell \boldsymbol{\Sigma}_{\ell\ell}^{-1} \boldsymbol{\sigma}_{\ell\bar{\ell}} + \boldsymbol{X} \boldsymbol{\xi}_\ell + \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{1/2} \boldsymbol{u}_\ell,$$

where $\boldsymbol{u}_\ell$ is independent of $\boldsymbol{Y}_\ell$ and $\boldsymbol{u}_\ell \sim \mathcal{N}_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$. Since $(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{Y}_\ell \boldsymbol{\Sigma}_{\ell\ell}^{-1} \boldsymbol{\sigma}_{\ell\bar{\ell}} = \boldsymbol{0}_n$ and $(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{X} \boldsymbol{\xi}_\ell = \boldsymbol{0}_n$ hold from a property of the projection matrix, we have

$$\tilde{s}_\ell = \boldsymbol{u}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})}) \boldsymbol{u}_\ell. \tag{12}$$

From Cochran's Theorem (e.g., [8]), we have $\tilde{s}_\ell \sim \chi^2(n - p - k + 1)$.

Next, we consider $\tilde{h}_\ell$. From a property of the projection matrix, $(\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{Y}_\ell \boldsymbol{\Sigma}_{\ell\ell}^{-1} \boldsymbol{\sigma}_{\ell\bar{\ell}} = \boldsymbol{0}_n$ holds. Hence, $\tilde{h}_\ell$ can be expressed as

$$\tilde{h}_\ell = \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} (\boldsymbol{X} \boldsymbol{\xi}_\ell + \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{1/2} \boldsymbol{u}_\ell)' (\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) (\boldsymbol{X} \boldsymbol{\xi}_\ell + \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{1/2} \boldsymbol{u}_\ell). \tag{13}$$

Since $\boldsymbol{Y}_\ell$ and $\boldsymbol{u}_\ell$ are independent, the conditional distribution of $\tilde{h}_\ell$ given $\boldsymbol{Y}_\ell$ is the non-central chi-squared distribution $\chi^2(k; \delta_\ell)$ from Cochran's Theorem, where δ_ℓ is expressed as

$$\delta_\ell = \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{\xi}_\ell' \boldsymbol{X}' (\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{X} \boldsymbol{\xi}_\ell = \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{\xi}_\ell' \boldsymbol{X}' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) \boldsymbol{X} \boldsymbol{\xi}_\ell.$$

Finally, we see that $\tilde{s}_\ell$ and $\tilde{h}_\ell$ are conditionally independent given $\boldsymbol{Y}_\ell$ from (12), (13) and Cochran's Theorem because $(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})})(\boldsymbol{P}_{(\boldsymbol{Y}_\ell, \boldsymbol{X})} - \boldsymbol{P}_{\boldsymbol{Y}_\ell}) = \boldsymbol{O}_{n,n}$ holds. Therefore, the proof of Lemma 1 is completed. □

Appendix 2: Proof of Proposition 1

First, we show (i). Without loss of generality, we can rewrite $\boldsymbol{y}$ by $\boldsymbol{y} = (\boldsymbol{y}_{j_*}', \boldsymbol{y}_{\bar{j}_*}')' = (\boldsymbol{y}_{j_*}', \boldsymbol{y}_{\bar{j}_* \cap \ell}', y_{\bar{\ell}})'$ for $\bar{\ell} \notin j_*$. Then, the proof of (i) is completed by letting $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_{j_*}$ in the following lemma (the proof is given in Appendix 4).

Lemma A.1. *Suppose that the divisions of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ are given by*

$$\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3) = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_{(23)}) = (\boldsymbol{\Theta}_{(12)}, \boldsymbol{\Theta}_3),$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{13}' & \boldsymbol{\Sigma}_{23}' & \boldsymbol{\Sigma}_{33} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{1(23)} \\ \boldsymbol{\Sigma}_{1(23)}' & \boldsymbol{\Sigma}_{(23)(23)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{(12)(12)} & \boldsymbol{\Sigma}_{(12)3} \\ \boldsymbol{\Sigma}_{(12)3}' & \boldsymbol{\Sigma}_{33} \end{pmatrix}.$$

If $\boldsymbol{\Theta}_{(23)} = \boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{1(23)}$, then we have $\boldsymbol{\Theta}_3 = \boldsymbol{\Theta}_{(12)} \boldsymbol{\Sigma}_{(12)(12)}^{-1} \boldsymbol{\Sigma}_{(12)3}$.

Next, we show (ii). Without loss of generality, we can rewrite $\boldsymbol{y}$ by $\boldsymbol{y} = (\boldsymbol{y}_{\bar{j}_*}', \boldsymbol{y}_{j_*}')' = (\boldsymbol{y}_{\bar{j}_*}', \boldsymbol{y}_{j_* \cap \ell}', y_{\bar{\ell}})'$ for $\bar{\ell} \in j_*$. Let the partition matrix of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ corresponding to the true subset j_* be as follows:

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

Then, we set the matrices $\boldsymbol{\Psi}_\ell$, $\boldsymbol{\Gamma}_\ell$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Gamma}$ as follows:

$$\boldsymbol{\Psi}_\ell = \boldsymbol{\Gamma}_\ell' \boldsymbol{\Theta}' \boldsymbol{\Theta} \boldsymbol{\Gamma}_\ell, \ \boldsymbol{\Gamma}_\ell = \begin{pmatrix} \boldsymbol{\Sigma}_{\ell\ell}^{-1/2} & -\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2} \boldsymbol{\Sigma}_{\ell\ell}^{-1} \boldsymbol{\sigma}_{\ell\bar{\ell}} \\ \boldsymbol{0}_{p-1}' & \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1/2} \end{pmatrix},$$
$$\boldsymbol{\Psi} = \boldsymbol{\Gamma}' \boldsymbol{\Theta}' \boldsymbol{\Theta} \boldsymbol{\Gamma}, \ \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\bar{j}_* \bar{j}_* \cdot j_*}^{-1/2} & \boldsymbol{O}_{p-p_{j_*}, p_{j_*}} \\ -\boldsymbol{\Sigma}_{j_* j_*}^{-1} \boldsymbol{\Sigma}_{j_* \bar{j}_*} \boldsymbol{\Sigma}_{\bar{j}_* \bar{j}_* \cdot j_*}^{-1/2} & \boldsymbol{\Sigma}_{j_* j_*}^{-1/2} \end{pmatrix}.$$

Note that the (p,p)-th element of $\boldsymbol{\Psi}_\ell$ is $\sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{\xi}_\ell' \boldsymbol{\xi}_\ell$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}_\ell')^{-1} \boldsymbol{\Gamma}_\ell^{-1} = (\boldsymbol{\Gamma}')^{-1} \boldsymbol{\Gamma}^{-1}$ holds. Moreover, $\boldsymbol{\Psi}$ can be calculated as

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{O}_{p-p_{j_*}, p-p_{j_*}} & \boldsymbol{O}_{p-p_{j_*}, p_{j_*}} \\ \boldsymbol{O}_{p_{j_*}, p-p_{j_*}} & \boldsymbol{\Sigma}_{j_* j_*}^{-1/2} \boldsymbol{\Theta}_{j_*}' \boldsymbol{\Theta}_{j_*} \boldsymbol{\Sigma}_{j_* j_*}^{-1/2} \end{pmatrix}.$$

Then, we have

$$\begin{aligned} \delta_\ell &\leq \sigma_{\bar{\ell}\bar{\ell}\cdot\ell}^{-1} \boldsymbol{\xi}_\ell' \boldsymbol{\xi}_\ell \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) \\ &\leq \mathrm{tr}(\boldsymbol{\Psi}_\ell) \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) \\ &= \mathrm{tr}\{\boldsymbol{\Psi} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_\ell \boldsymbol{\Gamma}_\ell' (\boldsymbol{\Gamma}')^{-1}\} \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) \\ &= \mathrm{tr}\{\boldsymbol{\Psi} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Gamma}')^{-1}\} \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) \\ &= \mathrm{tr}(\boldsymbol{\Psi}) \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}) \\ &= \mathrm{tr}(\boldsymbol{\Theta}_{j_*} \boldsymbol{\Sigma}_{j_* j_*}^{-1} \boldsymbol{\Theta}_{j_*}') \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}). \end{aligned}$$

Therefore, the proof of (ii) in Proposition 1 is completed. □

Appendix 3: Proof of Theorem 1

First, we show $P(\cup_{\bar{\ell}\notin j_*}\{h_\ell s_\ell^{-1} > \beta + k/(N-2)\}) \to 0$ in (8). Since $\delta_\ell = 0$ for $\bar{\ell} \notin j_*$ from Proposition 1, we see that $h_\ell \sim \chi^2(k)$ and, moreover, s_ℓ is independent of h_ℓ. Then, for any $r \in \mathbb{N}$ we have

$$P\left(\bigcup_{\bar{\ell}\notin j_*}\left\{\frac{h_\ell}{s_\ell} > \beta + \frac{k}{N-2}\right\}\right) \leq \sum_{\bar{\ell}\notin j_*} P\left(\frac{h_\ell}{s_\ell} - \frac{k}{N-2} > \beta\right)$$
$$\leq (p - p_{j_*})\beta^{-2r} E\left[\left(\frac{h_\ell}{s_\ell} - \frac{k}{N-2}\right)^{2r}\right].$$

Then, the order of the moment of the above equation is $O(k^r n^{-2r})$ from [15, Lemma A.2]. Hence, we have $P(\cup_{\bar{\ell}\notin j_*}\{h_\ell s_\ell^{-1} > \beta + k/(N-2)\}) = O(pk^r\beta^{-2r}n^{-2r}) \to 0$.

Next, we show $P(\cup_{\bar{\ell}\in j_*}\{h_\ell s_\ell^{-1} \leq \beta + k/(N-2)\}) \to 0$ in (8). From a property of the non-central chi-squared distribution, h_ℓ can be expressed as follows:

$$h_\ell = \delta_\ell + t_\ell + 2\delta_\ell^{1/2} v_\ell,$$

where t_ℓ and v_ℓ are random variables, and the conditional distributions of those given $\boldsymbol{Y}_\ell$ are $t_\ell|\boldsymbol{Y}_\ell \sim \chi^2(k)$ and $v_\ell|\boldsymbol{Y}_\ell \sim \mathcal{N}(0,1)$. Let $E = \{\min_{\bar{\ell}\in j_*}\delta_\ell > \tau n\}$, where τ is given by Assumption A3. Then, we have

$$P\left(\bigcup_{\bar{\ell}\in j_*}\left\{\frac{h_\ell}{s_\ell} \leq \beta + \frac{k}{N-2}\right\}\right) = P\left(\left\{\bigcup_{\bar{\ell}\in j_*}\left\{\frac{h_\ell}{s_\ell} \leq \beta + \frac{k}{N-2}\right\}\right\} \cap (E \cup E^c)\right)$$
$$\leq P\left(\left\{\bigcup_{\bar{\ell}\in j_*}\left\{\frac{h_\ell}{s_\ell} \leq \beta + \frac{k}{N-2}\right\}\right\} \cap E\right) + P(E^c)$$
$$\leq \sum_{\bar{\ell}\in j_*} P\left(\tau + n^{-1}t_\ell + 2\tau^{1/2}n^{-1/2}v_\ell \leq n^{-1}s_\ell\left(\beta + \frac{k}{N-2}\right)\right)$$
$$+ P(E^c).$$

By calculating the variances of t_ℓ, v_ℓ and $(N-2)^{-1}s_\ell$, we see that $t_\ell = k + O_p(k^{1/2})$, $v_\ell = O_p(1)$, and $(N-2)^{-1}s_\ell = 1 + O_p(n^{-1/2})$ hold. Moreover, $P(E^c) = o(1)$ holds from Assumption A3. Hence, we have $P(\cup_{\bar{\ell}\in j_*}\{h_\ell s_\ell^{-1} \leq \beta + k/(N-2)\}) \to 0$. Therefore, the proof of Theorem 1 is completed. □

Appendix 4: Proof of Lemma A.1

From the assumption of Lemma A.1, the following equations hold:

$$\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12},\ \boldsymbol{\Theta}_3 = \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13}. \tag{14}$$

By using (14) and the general formula for the inverse of a block matrix, $\boldsymbol{\Theta}_{(12)}\boldsymbol{\Sigma}_{(12)(12)}^{-1}\boldsymbol{\Sigma}_{(12)3}$ is expanded as

$$\begin{aligned}
\boldsymbol{\Theta}_{(12)}\boldsymbol{\Sigma}_{(12)(12)}^{-1}\boldsymbol{\Sigma}_{(12)3} &= (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)\begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1} \\ -\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{\Sigma}_{22\cdot1}^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{23} \end{pmatrix} \\
&= \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} + \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} - \boldsymbol{\Theta}_2\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} \\
&\quad - \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{23} + \boldsymbol{\Theta}_2\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{23} \\
&= \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} + \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} - \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} \\
&\quad - \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{23} + \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22\cdot1}^{-1}\boldsymbol{\Sigma}_{23} \\
&= \boldsymbol{\Theta}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} \\
&= \boldsymbol{\Theta}_3,
\end{aligned}$$

where $\boldsymbol{\Sigma}_{22\cdot1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Therefore, the proof of Lemma A.1 is completed. □

Acknowledgments

This work was supported by funding from JSPS KAKENHI (grant numbers JP19K21672, JP20K14363 and JP20H04151 to Ryoya Oda; and JP16H03606, JP18K03415 and JP20H04151 to Hirokazu Yanagihara).

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), pp. 267–281. Akadémiai Kiadó, Budapest. doi:10.1007/978-1-4612-1694-0_15

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723. doi:10.1109/TAC.1974.1100705

[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. 18–09, *Statistical Research Group*, Hiroshima University.

[4] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370. doi:10.1007/BF02294361

[5] Brown, P. J. (1982). Multivariate calibration. *J. R. Statist. Soc. Ser. B*, **44**, 287–321. doi:10.1111/j.2517-6161.1982.tb01209.x

[6] Fujikoshi, Y. & Nishii, R. (1986). Selection of variables in a multivariate inverse regression problem. *Hiroshima Math. J.*, **13**, 269–277. doi:10.32917/hmj/1206130428

[7] Fujikoshi, Y. & Sakurai, T. (2019). Consistency of test-based method for selection of variables in high-dimensional two-group discriminant analysis. *Jpn. J. Stat. Data Sci.*, **2**, 155–171. doi:10.1007/s42081-019-00032-4

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

[9] Hannan, E. J. & Quinn, B. G (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B*, **26**, 270–273. doi:10.1111/j.2517-6161.1979.tb01072.x

[10] Nakagawa, I., Watanabe, H. & Hyodo, M. (2021). Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings. *J. Multivariate Anal.*, **184**, 104756. doi:10.1016/j.jmva.2021.104756

[11] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765. doi:10.1214/aos/1176346522

[12] Nishii, R., Bai, Z. D. & Krishnaiah, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462. doi:10.32917/hmj/1206129611

[13] Oda, R., Mima, Y., Yanagihara, H. & Fujikoshi, Y. (2020). A high-dimensional bias-corrected AIC for selecting response variables in multivariate calibration. *Comm. Statist. Theory Methods*, **50**, 3453–3476. doi:10.1080/03610926.2019.1705978

[14] Oda, R., Suzuki, Y., Yanagihara, H. & Fujikoshi, Y. (2020). A consistent variable selection method in high-dimensional canonical discriminant analysis. *J. Multivariate Anal.*, **175**, 104561. doi:10.1016/j.jmva.2019.104561

[15] Oda, R. & Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron. J. Statist.*, **14**, 1386-1412. doi:10.1214/20-EJS1701

[16] Oda, R. & Yanagihara, H. (2021). A consistent likelihood-based variable selection method in normal multivariate linear regression. *Smart Innov. Syst. Tec.*, **238**, 391–401. doi:10.1007/978-981-16-2765-1_33

[17] Sakurai, T. & Fujikoshi, Y. (2018). Consistency of distance-based criterion for selection of variables in high-dimensional two-group discriminant analysis. TR No. 18–05, *Statistical Research Group*, Hiroshima University.

[18] Sakurai, T. & Fujikoshi, Y. (2020). Exploring consistencies of information criterion and test based criterion for high dimensional multivariate regression models under three covariance structures. *Recent Developments in Multivariate and Random Matrix Analysis*, 313–334. doi:10.1007/978-3-030-56773-6_18

[19] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464. doi:10.1214/aos/1176344136

[20] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, New York.

[21] Srivastava, M. S. & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics.* North-Holland, New York.

[22] Timm, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag, New York.

[23] Yanagihara, H. (2016). A high-dimensionality-adjusted consistent C_p-type statistic for selecting variables in a normality-assumed linear regression with multiple responses. *Procedia Comput. Sci.*, **96**, 1096–1105. doi:10.1016/j.procs.2016.08.151

[24] 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. doi:10.1016/0047-259X(86)90017-5