High-Dimensional Consistencies of KOO Methods for Selection of Variables in Multivariate Linear Regression Models with Covariance Structures

Yasunori Fujikoshi* and Tetsuro Sakurai**

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

**School of General and Management Studies,
Suwa University of Science
5000-1 Toyohira, Chino, Nagano 391-0292, Japan

Abstract

In this paper, we consider high-dimensional consistencies of KOO methods for selection of variables in multivariate regression model with covariance structures. The covariance structures considered are (1) independent covariance structure with the same variance, (2) independent covariance structure with different variances, and (3) uniform covariance structure. Sufficient conditions for our KOO methods to be consistent are derived under a high-dimensional asymptotic framework such that the sample size n, the number p of response variables and the number k of explanatory variables are large as in the way $p/n \to c_1 \in (0, 1)$ and $k/n \to c_2 \in [0, 1)$, where $c_1 + c_2 < 1$.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: Consistency property, Covariance structures, High-dimensional asymptotic framework, KOO methods, Multivariate linear regression.

1. Introduction

We consider a multivariate linear regression of p response variables $y_1, \ldots, y_p$ on a subset of k explanatory variables $x_1, \ldots, x_k$. Suppose that there are n observations on $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_k)'$, and let $\mathbf{Y} : n \times p$ and $\mathbf{X} : n \times k$ be the observation matrices of $\boldsymbol{y}$ and $\boldsymbol{x}$ with the sample size n, respectively. The multivariate linear regression model including all the explanatory variables under normality is written as

$$\mathbf{Y} \sim \mathrm{N}_{n \times p}(\mathbf{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n), \tag{1.1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ unknown matrix of regression coefficients and $\boldsymbol{\Sigma}$ is a $p \times p$ unknown covariance matrix. The notation $\mathrm{N}_{n \times p}(\cdot, \cdot)$ means the matrix normal distribution such that the mean of $\mathbf{Y}$ is $\mathbf{X}\boldsymbol{\Theta}$ and the covariance matrix of $\mathrm{vec}\,(\mathbf{Y})$ is $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$, or equivalently, the rows of $\mathbf{Y}$ are independently normal with the same covariance matrix $\boldsymbol{\Sigma}$. Here, $\mathrm{vec}(\mathbf{Y})$ be the $np \times 1$ column vector obtained by stacking the columns of $\mathbf{Y}$ on top of one another. We assume that $\mathrm{rank}(\mathbf{X}) = k$.

It is important to consider selection of regression variables in multivariate analysis. One of the approaches is to first consider variable selection models and then apply model selection criteria such as AIC, BIC, etc. Such a criterion for the full model (1.1) is expressed as

$$\mathrm{GIC} = -2 \log L(\widehat{\boldsymbol{\Xi}}) + dg, \tag{1.2}$$

where $L(\widehat{\boldsymbol{\Xi}})$ is the maximum likelihood, $\Xi = \{\boldsymbol{\Theta}, \boldsymbol{\Sigma}\}$, $d > 0$ is the penaty term, and g is the number of unknown parameters given by $\{kp + \frac{1}{2}p(p+1)\}$. For AIC and BIC, d is defined as 2 and $\log n$, respectively. In the seletion of k variables $x_1, \ldots, x_k$, we identify $\{x_1, \ldots, x_k\}$ with the index set $\{1, \ldots, k\} \equiv \boldsymbol{\omega}$, and denote GIC for subset $\boldsymbol{j} \subset \boldsymbol{\omega}$ by $\mathrm{GIC}_{\boldsymbol{j}}$. Then, the model selection based on GIC chooses the model

$$\widetilde{\boldsymbol{j}} = \arg\min_{\boldsymbol{j}} \mathrm{GIC}_{\boldsymbol{j}}. \tag{1.3}$$

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

$$\widehat{\boldsymbol{j}} = \{j \in \boldsymbol{\omega} \mid \mathrm{GIC}_{\boldsymbol{\omega}\backslash j} > \mathrm{GIC}_{\boldsymbol{\omega}}\}, \tag{1.4}$$

where $\boldsymbol{\omega}\backslash j$ is a short expression for $\boldsymbol{\omega}\backslash\{j\}$ which is the set obtained by removing element j from the set $\boldsymbol{\omega}$.

When $\boldsymbol{\Sigma}$ is unknown positive definite, it has been pointed (see, e.g., Yanagihara et al. (2015), Fujikoshi et al. (2014), etc.) that in a high-dimensional case, AIC and C_p have consistency properties, but BIC is not necessarily consistent. The KOO methods in multivariate regression model has been studied by Bai et al. (2018), Bai et al. (2022), Oda and Yanagihara (2020, 2021). The KOO method in discriminant analysis, see Fujikoshi and Sakurai (2019), Oda et al. (2020). For a review, see Fujikoshi (2022).

In this paper we assume that the covariance structure is one of three covariance structures; (1) an independent covariance structure with the same variance, (2) an independent covariance structure with different variances and (3) a uniform covariance structure. Sufficient conditions for the KOO method given by (1.4) to be consistent are derived under a high-dimensional asymptotic framework such that the sample size n, the number p of response variables and the number k of explanatory variables are large as in the way $p/n \to c_1 \in (0,1)$ and $k/n \to c_2 \in [0,1)$, where $c_1 + c_2 < 1$. Sakurai and Fujikoshi (2020) has considered similar problems under covariance structures (1), (3) and (4) an autoregressive covariance structure, but do not consider under (2). More over, in the study of asymptotic consistencies they assumed that k is fixed, but in this paper k may tend to infity such that $k/n \to c_2 \in [0, 1)$.

The present paper is organized as follows. In section 2, we present notations and preliminaries. In Section 3, we state KOO methods with covariance structures (1), (2) and (3) in termes of key statistics. Further, an approach for their consistencies is stated in Section 3. In Sections 4, 5 and 6 we discuss consistency properties of KOO methods under the covariance structures (1), (2) and (3). In Section 7, our conclusions are discussed.

2. Notations and Preliminaries

Suppose that $\boldsymbol{j}$ denotes a subset of $\boldsymbol{\omega} = \{1, \ldots, k\}$ containing $k_{\boldsymbol{j}}$ elements, and $\mathbf{X}_{\boldsymbol{j}}$ denotes the $n \times k_{\boldsymbol{j}}$ matrix consisting the columns of $\mathbf{X}$ indexed by the elements of $\boldsymbol{j}$. Then, $\mathbf{X}_{\boldsymbol{\omega}} = \mathbf{X}$. Further, we assume that the covariance matrix $\boldsymbol{\Sigma}$ have a covariance structure $\boldsymbol{\Sigma}_c$. Then, we have a generic candidate model

$$M_{c,\boldsymbol{j}}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}}, \boldsymbol{\Sigma}_{c,\boldsymbol{j}} \otimes \mathbf{I}_n), \tag{2.1}$$

where $\boldsymbol{\Theta}_{\boldsymbol{j}}$ is a $k_{\boldsymbol{j}} \times p$ unknown matrix of regression coefficients. We assume that $\mathrm{rank}(\mathbf{X}) = k$.

When $\boldsymbol{\Sigma}_{c,\boldsymbol{j}}$ is a $p \times p$ unknown covariance matrix, we can write the GIC in (1.2) as

$$\mathrm{GIC}_{c,\boldsymbol{j}} = n \log |\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}| + np(\log 2\pi + 1) + d \left\{ k_{\boldsymbol{j}} p + \frac{1}{2} p(p+1) \right\}, \tag{2.2}$$

where $n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y}$ and $\mathbf{P}_{\boldsymbol{j}} = \mathbf{X}_{\boldsymbol{j}}(\mathbf{X}_{\boldsymbol{j}}'\mathbf{X}_{\boldsymbol{j}})^{-1}\mathbf{X}_{\boldsymbol{j}}'$. When $\boldsymbol{j} = \boldsymbol{\omega}$, the model $M_{c,\boldsymbol{\omega}}$ is called the full model. Note that $\widehat{\boldsymbol{\Sigma}}_{c,\boldsymbol{\omega}}$ and $\mathbf{P}_{\boldsymbol{\omega}}$ are defined from $\widehat{\boldsymbol{\Sigma}}_{c,\boldsymbol{j}}$ and $\mathbf{P}_{\boldsymbol{j}}$ as $\boldsymbol{j} = \boldsymbol{\omega}$, $k_{\boldsymbol{\omega}} = k$ and $\mathbf{X}_{\boldsymbol{\omega}} = \mathbf{X}$.

In this paper, we consider the cases that the covariance matrix $\boldsymbol{\Sigma}_c$ belongs to each of the following three structures;

(1) Independent covariance structure with the same variance (ICSS);

$\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}_p$,

(2) Independent covariance structure with different variances (ICSD);

$\mathbf{\Sigma}_b = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$,

(3) Uniform covariance structure (UCS);

$\mathbf{\Sigma}_u = \sigma_u^2 (\rho_u^{1-\delta_{ij}})_{1 \le i,j \le p}$.

The models considered in this paper can be expressed as (2.1) with $\Sigma_{v,\boldsymbol{j}}$, $\Sigma_{b,\boldsymbol{j}}$, and $\mathbf{\Sigma}_{u,\boldsymbol{j}}$ for $\mathbf{\Sigma}_{c,\boldsymbol{j}}$. Let $f(\mathbf{Y}; \mathbf{\Theta}_{\boldsymbol{j}}, \mathbf{\Sigma}_{c,\boldsymbol{j}})$ be the density of $\mathbf{Y}$ in (2.1) with $\mathbf{\Sigma} = \mathbf{\Sigma}_{c,\boldsymbol{j}}$. In the derivation of the GIC under the covariance structute $\mathbf{\Sigma} = \mathbf{\Sigma}_{c,\boldsymbol{j}}$, we will use the following equality:

$$-2\log \max_{\mathbf{\Theta}_{\boldsymbol{j}}, \mathbf{\Sigma}_{c,\boldsymbol{j}}} f(\mathbf{Y}; \mathbf{\Theta}_{\boldsymbol{j}}, \mathbf{\Sigma}_{c,\boldsymbol{j}}) = np\log(2\pi) + \min_{\mathbf{\Sigma}_{c,\boldsymbol{j}}} \left\{ np\log|\mathbf{\Sigma}_{c,\boldsymbol{j}}| + \mathrm{tr}\mathbf{\Sigma}_{c,\boldsymbol{j}}^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y} \right\}. \tag{2.3}$$

Let $\widehat{\mathbf{\Sigma}}_{c,\boldsymbol{j}}$ be the quantity minimizing the right side of (2.3). Then, in our model, it satisfies $\mathrm{tr}\widehat{\mathbf{\Sigma}}_{c,\boldsymbol{j}}^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y} = np$, and we obtain

$$\begin{aligned} \mathrm{GIC}_{c,\boldsymbol{j}} &= -2\log f(\mathbf{Y}; \widehat{\mathbf{\Theta}}_{\boldsymbol{j}}, \widehat{\mathbf{\Sigma}}_c) + dm_{c,\boldsymbol{j}} \\ &= np\log|\widehat{\mathbf{\Sigma}}_{c,\boldsymbol{j}}| + np(\log 2\pi + 1) + dm_{c,\boldsymbol{j}}, \end{aligned} \tag{2.4}$$

where $m_{c,\boldsymbol{j}}$ is the number of independent unknown parameters under $M_{c,\boldsymbol{j}}$, and d is a positive constant which may depend on n. For AIC and BIC, d is defined by 2 (Akaike (1973)) and $\log n$ (Schwarz (1978)), respectively.

3. Approach to Consistencies of KOO Methods

Our KOO method is based on

$$T_{c,j;d} = \mathrm{GIC}_{c,\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{c,\boldsymbol{\omega}}. \tag{3.1}$$

In fact, the KOO method chooses the model

$$\widehat{\boldsymbol{j}}_{c;d} = \{j \mid T_{c,j;d} > 0\}. \tag{3.2}$$

Its consistency can be proved by showing the following two properties:

$$\text{Q1}: \quad [\text{F1}] \equiv \sum_{j \in \boldsymbol{j}_*} \Pr(T_{c,j;d} \le 0) \to 0, \tag{3.3}$$

$$\text{Q2}: \quad [\text{F2}] \equiv \sum_{j \notin \boldsymbol{j}_*} \Pr(T_{c,j;d} \ge 0) \to 0, \tag{3.4}$$

as in Fujikoshi (2022). The result can be shown by using the following inequalty:

$$\begin{aligned}
\Pr(\widehat{\boldsymbol{j}}_{c;d} = \boldsymbol{j}_*) &= \Pr\left(\bigcap_{j \in \boldsymbol{j}_*} \text{``}T_{c,j;d} > 0\text{''} \bigcap_{j \notin \boldsymbol{j}_*} \text{``}T_{c,j;d} < 0\text{''} \right) \\
&= 1 - \Pr\left(\bigcup_{j \in \boldsymbol{j}_*} \text{``}T_{c,j;d} \le 0\text{''} \bigcup_{j \notin \boldsymbol{j}_*} \text{``}T_{c,j;d} \ge 0\text{''} \right) \\
&\ge 1 - \sum_{j \in \boldsymbol{j}_*} \Pr(T_{c,j;d} \le 0) - \sum_{j \notin \boldsymbol{j}_*} \Pr(T_{c,j;d} \ge 0).
\end{aligned}$$

Here, [F1] denotes the probability that the true variables are not selected, and [F2] denotes the probability that the non-true variables are selected. Such notations will be used for other variable selection methods. We call "x_j is included in the true set of variables" if $\boldsymbol{\theta}_j \neq \mathbf{0}$.

Here we list some of our main assumptions:

A1: The set $\boldsymbol{j}_*$ of the true explanatory variables is included in the full subset, i.e., $\boldsymbol{j}_* \subset \boldsymbol{\omega}$. and the set $\boldsymbol{j}_*$ is finite.

A2: The high dimensional asymptotic framework:
$p \to \infty,\ n \to \infty,\ k \to \infty,\ p/n \to c_1 \in (0,1),\ k/n \to c_2 \in [0,1),$
where $0 < c_1 + c_2 < 1$.

It is said that a general model selection criterion $\widehat{\boldsymbol{j}}_{c;d}$ is high-dimensional consistent if

$$\lim \Pr(\widehat{\boldsymbol{j}}_{c;d} = \boldsymbol{j}_*) = 1,$$

under a high-dimensional asymptotic framework. Here, "lim" means the limit under A2.

4. Asymptotic Cosistency under an Independent Covariance Structure

In this section we shall show an asymptotic consistency of KOO method based on a general information criterion under an independent covariance structure. A generic candidate model when the set of explanatory vaiables is $\boldsymbol{j}$ can be expressed as

$$M_{v,\boldsymbol{j}}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}}, \boldsymbol{\Sigma}_{v,\boldsymbol{j}} \otimes \mathbf{I}_n), \tag{4.1}$$

where $\boldsymbol{\Sigma}_{v,\boldsymbol{j}} = \sigma_{v,\boldsymbol{j}}^2 \mathbf{I}_p$ and $\sigma_{v,\boldsymbol{j}}^2 > 0$. Let us denote the density of $\mathbf{Y}$ under (4.1) by $f(\mathbf{Y}; \Theta_{\boldsymbol{j}}, \sigma_{v,\boldsymbol{j}})$. Then, we have

$$\begin{aligned} -2\log f(\mathbf{Y}; \boldsymbol{\Theta}_{\boldsymbol{j}}, \sigma_{v,\boldsymbol{j}}^2) &= np\log(2\pi) + np\log \sigma_{v,\boldsymbol{j}}^2 \\ &\quad + \frac{1}{\sigma_{v,\boldsymbol{j}}^2}\mathrm{tr}(\mathbf{Y} - \mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}})'(\mathbf{Y} - \mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}}). \end{aligned}$$

Therefore, it is easily seen that the maximum estimators of $\boldsymbol{\Theta}_{\boldsymbol{j}}$ and $\sigma_{v,\boldsymbol{j}}^2$ under $M_{v,\boldsymbol{j}}$ are given as

$$\widehat{\boldsymbol{\Theta}}_{\boldsymbol{j}} = (\mathbf{X}_{\boldsymbol{j}}'\mathbf{X}_{\boldsymbol{j}})^{-1}\mathbf{X}_{\boldsymbol{j}}'\mathbf{Y}, \quad \widehat{\sigma}_{v,\boldsymbol{j}}^2 = \frac{1}{np}\mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}. \tag{4.2}$$

The general information criterion (2.4) is given by

$$\mathrm{GIC}_{v,\boldsymbol{j}} = np\log \widehat{\sigma}_{v,\boldsymbol{j}}^2 + np(\log 2\pi + 1) + dm_{v,\boldsymbol{j}}, \tag{4.3}$$

where d is a positive constant and $m_{v,j} = k_j p + 1$. Using (3.1) and (4.3) we have

$$\begin{aligned} T_{v,j;d} &\equiv \mathrm{GIC}_{v,\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{v,\boldsymbol{\omega}} \\ &= np\log\left(1 + U_{2j}U_1^{-1}\right) - dp, \end{aligned} \tag{4.4}$$

where

$$U_1 = \mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{Y} = \sum_{\ell=1}^{p} \boldsymbol{y}_\ell'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\boldsymbol{y}_\ell,$$

$$U_{2j} = \mathrm{tr}\mathbf{Y}'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\mathbf{Y} = \sum_{\ell=1}^{p} \boldsymbol{y}_\ell'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\boldsymbol{y}_\ell.$$

It is seen that $U_1/\sigma^2_{v,\boldsymbol{j}_*}$ and $U_{2j}/\sigma^2_{v,\boldsymbol{j}_*}$ are independently distributed as a central chi-square distribution and a nocentral chi-square distribution, respectively. More precisely, assume that

$$\mathrm{E}(\boldsymbol{Y}) = \mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*}, \tag{4.5}$$

and let $\sigma^2_{v,*} = \sigma^2_{v,\boldsymbol{j}_*}$. Then, using basic distributional propertites (see, Fujikoshi et al. (2010)) on quadratic forms of normal variates and Wishart matrices, we have the following results:

$$\begin{aligned} &(1)\ U_1/\sigma^2_{v,*} \sim \chi^2_{(n-k)p}, \\ &(2)\ U_{2j}/\sigma^2_{v,*} \sim \chi^2_p(\delta^2_{v,j}), \\ &(3)\ U_1 \perp U_{2j}, \end{aligned} \tag{4.6}$$

where the noncentrality parameter $\tau^2_{v,j}$ is defined by

$$\delta^2_{v,j} = \frac{1}{\sigma^2_{v,*}}\mathrm{tr}(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*})'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*}.$$

It may be noted that if $j \notin \boldsymbol{j}_*$, $\delta^2_{v,j} = 0$, and if $j \in \boldsymbol{j}_*$, in general, $\tau^2_{v,j} \neq 0$. For a sufficient condition for consistency of the KOO method based on $\mathrm{GIC}_{v,\boldsymbol{j}}$, we assume

$$\text{A3v : For any } j \in \boldsymbol{j}_*,\ \delta^2_{v,j} = \mathrm{O}(np), \text{and } \lim_{p/n\to c_1} \frac{1}{np}\delta^2_{v,j} = \eta^2_{v,j} > 0. \tag{4.7}$$

Now we consider high-dimensional asymptotic consistency of the KOO method based on $\mathrm{GIC}_{v,\boldsymbol{j}}$ in (4.3), whose selection method is given by $\widehat{\boldsymbol{j}}_{v,j;d} = \{j \mid T_{v,j;d} > 0\}$. When $j \notin \boldsymbol{j}_*$, from (4.4) we can write

$$\mathrm{T}_{v,j;d} = np\log\left\{1 + \chi^2_p/\chi^2_m\right\} - dp, \quad m = (n-k)p.$$

Therefore we have

$$\begin{aligned}[\mathrm{F2}] &= \sum_{j \notin \boldsymbol{j}_*} \Pr(np \log \{1 + \chi_p^2/\chi_m^2\} \geq dp) \\ &= (k - k_{\boldsymbol{j}_*}) \Pr(U \geq h) \\ &\leq (k - k_{\boldsymbol{j}_*}) \Pr(U \geq h_0),\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}U &= \frac{\chi_p^2}{\chi_m^2} - \frac{p}{m-2}, \\ h &= \mathrm{e}^{d/n} - 1 - \frac{p}{m-2}, \quad h_0 = \frac{d}{n} - \frac{p}{m-2}.\end{aligned} \tag{4.9}$$

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

$$[\mathrm{F2}] \leq (k - k_{\boldsymbol{j}_*})\, h^{-2\ell} \mathrm{E}[U^{2\ell}] \leq (k - k_{\boldsymbol{j}_*})\, h_0^{-2\ell} \mathrm{E}[U^{2\ell}]. \tag{4.10}$$

Related to the assumption $h_0 > 0$, we assume

$$\mathrm{A4v}: \ d > \frac{np}{m-2} \to \frac{1}{1-c_2}, \text{and } d = \mathrm{O}(n^a), \quad 0 < a < 1. \tag{4.11}$$

The first part in A4v implies $h_0 > 0$. It is easy to see that

$$\mathrm{E}[U^2] = \frac{2p(m+p-2)}{(m-2)^2(m-4)} = \mathrm{O}((n^2 p)^{-1}).$$

Further, $h_0^{-2} = \mathrm{O}(n^{2(1-a)})$. Therefore, from (4.10) with we have that [F2] $\to 0$.

When $j \in \boldsymbol{j}_*$, we can write $\mathrm{T}_{v,j;d} = np \log \{1 + \chi_p^2(\delta_{v,j}^2)/\chi_m^2\} - dp$. Therefore we can express have [F1] as

$$[\mathrm{F1}] = \sum_{j \in \boldsymbol{j}_*} \Pr(\widetilde{T}_{v,j;d} \leq 0),$$

where

$$\widetilde{T}_{v,j;d} = \frac{p}{n} \log \left\{1 + \frac{\chi_p^2(\delta_{v,j}^2)}{\chi_m^2}\right\} - \frac{d}{n}.$$

From the assumptions A3v and A4v it is easily seen that

$$\widetilde{T}_{v,j;d} \to c_1 \log(1 + \eta_{v,j}^2) > 0.$$

This implies that $\Pr(\widetilde{T}_{v,j;d} \le 0) \to 0$.

These imply the following theorem.

Theorem 4.1. *Suppose that the assumptions* A1, A2 A3v, and A4v *are satisfied. Then, the* KOO *method based on general information criteria* $\mathrm{GIC}_{v,\boldsymbol{j}}$ *defined by* (4.3) *is asymptotically consistent.*

An alternative apporoach for "[F1] $\to$ 0". When $j \in \boldsymbol{j}_*$, we can write $\mathrm{T}_{v,j;d} = np \log\left\{1 + \chi_p^2(\delta_{v,j}^2)/\chi_m^2\right\} - dp$. Therefore we have

$$\begin{aligned}[\mathrm{F1}] &= \sum_{j \in \boldsymbol{j}_*} \Pr(np \log\left\{1 + \chi_p^2(\delta_{v,j}^2)/\chi_m^2\right\} \le dp) \\ &= \sum_{j \in \boldsymbol{j}_*} \Pr(\widetilde{U}_j \le \widetilde{h}_j),\end{aligned}$$

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

$$\widetilde{U}_j = \frac{\chi_p^2(\delta_{v,j}^2)}{\chi_m^2} - \frac{p + \delta_{v,j}^2}{m-2}, \quad \widetilde{h}_j = \mathrm{e}^{d/n} - 1 - \frac{p + \delta_{v,j}^2}{m-2} = h - \frac{\delta_{v,j}^2}{m-2}.$$

Then, under $d = \mathrm{O}(n^a)(0 < a < 1)$, A3v in (4.7) and the assumption $\widetilde{h}_j < 0$ (or equivalently $h < \delta_j^2/(m-2)$) we have

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

It is easily seen that

$$\mathrm{E}[\widetilde{U}_j^2] = \frac{2(p + 2\delta_{v,j}^2)(m + p - 2 + \delta_{v,j}^2)}{(m-2)^2(m-4)} = \mathrm{O}((np)^{-1}),$$

and under $d = n^a(0 < a < 1)$ and A3v,

$$|\widetilde{h}_j|^2 \to \frac{\eta_{v,j}^2}{c_1(1 - c_2)}.$$

These imply that [F1] $\to$ 0. In this approach, it has been assumed that $\widetilde{h}_j < 0$ (or equivalently $h < \delta_j^2/(m-2)$).

5. Asymptotic Cosistency under an Independent Covariance Structure with Different Variances

In this section we assume that the covariance matrix $\boldsymbol{\Sigma}$ has an independent covariance matrix with different variances, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_b = \mathrm{diag}(\sigma_{b1}^2, \ldots, \sigma_{bp})$. First, let us consider to derive a key statistic $T_{b,j;d} = \mathrm{GIC}_{b,\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{b,\boldsymbol{\omega}}$. Consider a candidate model with $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\Theta}$,

$$M_{b,\boldsymbol{\omega}}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma}_b \otimes \mathbf{I}_n). \tag{5.1}$$

Let the density in the full model express as $f(\mathbf{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}_b)$. Then, we have

$$-2\log f(\mathbf{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}_b) = np\log(2\pi) + \sum_{\ell=1}^{p} \left\{ n\log\sigma_{b\ell}^2 + \frac{1}{\sigma_{b\ell}^2}(\boldsymbol{y}_\ell - \mathbf{X}\boldsymbol{\theta}_\ell)'(\boldsymbol{y}_\ell - \mathbf{X}\boldsymbol{\theta}_\ell) \right\}.$$

It holds that

$$-2\log \max_{\boldsymbol{\Theta},\boldsymbol{\Sigma}_b} f(\mathbf{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}_b) = np\left(\log 2\pi + 1\right) + \sum_{\ell=1}^{p} n\log\frac{1}{n}\boldsymbol{y}_\ell'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\boldsymbol{y}_\ell. \tag{5.2}$$

Next, consider the model removing the jth explanatory variable from the full model $M_{b,\boldsymbol{\omega}}$, which is denoted by $M_{b,\boldsymbol{\omega}\backslash j}$ or $M; b, \boldsymbol{\omega}\backslash j$. Similarly, it is shown that

$$-2\log \max_{M;b,\boldsymbol{\omega}\backslash j} f(\mathbf{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}_b) = np\left(\log 2\pi + 1\right) + \sum_{\ell=1}^{p} n\log\frac{1}{n}\boldsymbol{y}_\ell'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}\backslash j})\boldsymbol{y}_\ell. \tag{5.3}$$

Using (5.2) and (5.3), we can obtain a general information criterion (2.4) for two models $M_{b,\boldsymbol{\omega}}$ and $M_{b,\boldsymbol{\omega}\backslash j}$, and we have

$$T_{b,j;d} \equiv \mathrm{GIC}_{b,\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{b,\boldsymbol{\omega}} = \sum_{\ell=1}^{p} n\log\left(1 + U_{2\ell}U_{1\ell}^{-1}\right) - dp, \tag{5.4}$$

where

$$U_{1\ell} = \boldsymbol{y}_\ell'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\boldsymbol{y}_\ell, \quad \ell = 1, \ldots, p,$$
$$U_{2\ell} = \boldsymbol{y}_\ell'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})\boldsymbol{y}_\ell, \quad \ell = 1, \ldots, p.$$

Let us assume that

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*} \text{ and } \sigma_{b,*}^2 = \sigma_{b,\boldsymbol{j}_*}^2 \tag{5.5}$$

Then, similarly as in (4.6), we have the following results:

$$\begin{aligned} &(1)\ U_{1\ell}/\sigma_{b,*}^2 \sim \chi_{n-k}^2, \quad \ell = 1, \ldots, p, \\ &(2)\ U_{2\ell}/\sigma_{b,*}^2 \sim \chi_1^2(\delta_{b,j;\ell}^2), \quad \ell = 1, \ldots, p, \\ &(3)\ U_{1\ell}, U_{2\ell}, (\ell = 1, \ldots, p) \text{ are independent,} \end{aligned} \tag{5.6}$$

where the noncentral parameters $\delta_{b,j;\ell}^2$ are defined by

$$\delta_{b,j;\ell}^2 = \frac{1}{\sigma_{b,*}^2}(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\theta}_*^{(\ell)})'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\theta}_*^{(\ell)}),$$

with $\boldsymbol{\Theta}_* = (\boldsymbol{\theta}_*^{(1)}, \ldots, \boldsymbol{\theta}_*^{(p)})$. It may be noted that if $j \notin \boldsymbol{j}_*$, $\delta_{b,j;\ell}^2 = 0$. and if $j \in \boldsymbol{j}_*$, $\delta_{b,j;\ell}^2 \neq 0$. For a sufficient condition for consistency of the KOO method based on $\mathrm{GIC}_{b,\boldsymbol{j}}$, we assume

$$\begin{aligned} \text{A3b : For any } j \in \boldsymbol{j}_*,\ &\lim (n-k)^{-1}\delta_{b,j;\ell}^2 = \eta_{b,j;\ell}^2 > 0, \text{ and} \\ &\lim \frac{1}{p}\sum_{\ell=1}^p \log\left\{1 + \frac{1}{n-k}\delta_{b,j;\ell}^2\right\} \to \eta_{b,j}^2 > 0. \end{aligned} \tag{5.7}$$

Now we consider high-dimensional asymptotic consistency of the KOO method based on $\mathrm{T}_{b,j;d}$ in (3.1), whose selection method is given by $\widehat{\boldsymbol{j}}_{v,j;d} = \{j \mid T_{b,j;d} > 0\}$. When $j \notin \boldsymbol{j}_*$, we have

$$\begin{aligned} [\text{F2}] &= \sum_{j \notin \boldsymbol{j}_*} \Pr(\sum_{\ell=1}^p n \log\left\{1 + U_{2\ell}U_{1\ell}^{-1}\right\} \geq d) \\ &\leq \sum_{j \notin \boldsymbol{j}_*}\sum_{\ell=1}^p \Pr(n \log\left\{1 + U_{2\ell}U_{1\ell}^{-1}\right\} \geq d). \end{aligned}$$

This implies that

$$\begin{aligned}[\mathrm{F2}] &\leq p(k-k_{\boldsymbol{j}_*})\Pr(n\log\left\{1+\chi_1^2/\chi_{n-k}^2\right\}\geq d)\\ &= p(k-k_{\boldsymbol{j}_*})\Pr(V\geq r), \end{aligned} \tag{5.8}$$

where

$$\begin{aligned} V &= \frac{\chi_1^2}{\chi_{n-k}^2}-\frac{1}{n-k-2},\\ r &= \mathrm{e}^{d/n}-1-\frac{1}{n-k-2}, \quad r_0=\frac{d}{n}-\frac{1}{n-k-2}. \end{aligned} \tag{5.9}$$

Note that $r_0 < r$. Then, under the assumption $r_0 > 0$ we have,

$$[\mathrm{F2}] \leq p\,(k-k_{\boldsymbol{j}_*})\,r^{-2\ell}\mathrm{E}[V^{2\ell}] \leq p\,(k-k_{\boldsymbol{j}_*})\,r_0^{-2\ell}\mathrm{E}[V^{2\ell}]. \tag{5.10}$$

Related to the assumption $r_0 > 0$, we assume

$$\text{A4b}: \ d > \frac{n}{n-k-2} \to \frac{1}{1-c_2}, \text{and } d=\mathrm{O}(n^a), \quad 0<a<1. \tag{5.11}$$

The first part in A4b implies $r_0 > 0$. It is easy to see that

$$\mathrm{E}[V^2] = \frac{2(n-k-1)}{(n-k-2)^2(n-k-4)} = \mathrm{O}((n^2)^{-1}).$$

Further, $r_0^{-2} = \mathrm{O}(n^{2(1-a)})$. Therefore, from (5.10) we have that $[\mathrm{F2}] \to 0$.

When $j \in \boldsymbol{j}_*$, we can write $T_{b,j;d} = n\sum_{\ell=1}^{p}\log\{1+U_{2\ell}U_{1\ell}^{-1}\}-dp$. Therefore we can express [F1] as

$$[F1] = \sum_{j\in\boldsymbol{j}_*}\Pr(\widetilde{T}_{b,j;d}\leq 0),$$

where

$$\widetilde{T}_{b,j;d} = \frac{1}{p}\sum_{\ell=1}^{p}\log\left\{1+\frac{\chi_{1;\ell}^2(\delta_{b,j;\ell}^2)}{\chi_{n-k;\ell}^2}\right\}-\frac{d}{n}.$$

From the assumptions A3b and A4b it is easily seen that

$$\widetilde{T}_{b,j;d} \to \eta_{b,j}^2 > 0.$$

This implies that $\Pr(\widetilde{T}_{b,j;d}\leq 0) \to 0$.

These imply the following theorem.

Theorem 5.1. *Suppose that the assumptions* A1, A2, A3b and A4b *are satisfied. Then, the* KOO *method based on* $T_{b,j:d}$ *in* (5.4) *is asymptotically consistent.*

Let us consider an alternative apporoach for "[F1] $\to 0$" as in the case of independent covariance structure. When $j \in \boldsymbol{j}_*$, we can write

$$\begin{aligned}
[\mathrm{F1}] &= \sum_{j \in \boldsymbol{j}_*} \Pr\left(\sum_{\ell=1}^{p} \left\{ n \log\left(1 + \frac{\chi^2_{1;\ell}(\delta^2_{b,j;\ell})}{\chi^2_{n-k;\ell}}\right) - d \right\} \le 0 \right) \\
&\le \sum_{j \in \boldsymbol{j}_*} \sum_{\ell=1}^{p} \Pr\left(n \log\left(1 + \frac{\chi^2_{1;\ell}(\delta^2_{b,j;\ell})}{\chi^2_{n-k;\ell}}\right) - d \le 0 \right) \\
&= \sum_{j \in \boldsymbol{j}_*} \sum_{\ell=1}^{p} \Pr\left(\widetilde{V}_{j,\ell} \le \tilde{r}_{j,\ell} \right).
\end{aligned}$$

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

$$\begin{aligned}
\widetilde{V}_{j,\ell} &= \frac{\chi^2_{1;\ell}(\delta^2_{b,j;\ell})}{\chi^2_{n-k;\ell}} - \frac{1+\delta^2_{b,j;\ell}}{n-k-2}, \quad \ell = 1, \ldots, p, \\
\widetilde{r}_{j,\ell} &= \mathrm{e}^{d/n} - 1 - \frac{1+\delta^2_{b,j;\ell}}{n-k-2} = r - \frac{\delta^2_{b,j}}{n-k-2}, \quad \ell = 1, \ldots, p,
\end{aligned}$$

where r is the same one as in (5.9). Note that $\chi^2_{1;\ell}(\delta^2_{b,j;\ell}), \ell = 1, \ldots, p$ are distributed as a noncentral distribution $\chi^2_1(\delta^2_{b,j;\ell})$, and they are independent. Then, under the assumption $\widetilde{r}_j < 0$ (or equivalently $r < \delta^2_{bj;\ell}/(n-k-2)$) we have

$$[\mathrm{F1}] \le k_{\boldsymbol{j}_*} \sum_{\ell=1}^{p} |\widetilde{r}_{j,\ell}|^{-2s} \mathrm{E}[\widetilde{V}^{2s}_{j,\ell}], \quad s = 1, 2, \ldots. \tag{5.12}$$

In the above upper bound, it holds that

$$|\widetilde{r}_{j,\ell}| \sim \delta^2_{b,j;\ell}/(n-k) \to \eta^2_{b,j;\ell}. \tag{5.13}$$

Useful bounds will be obtained by giving the first few moments of $\widetilde{V}_{j;\ell}$. For example, note that

$$\begin{aligned}
\mathrm{E}[\widetilde{V}^2_{j,\ell}] &= \frac{2(1+2\delta^2_{v,j;\ell})(n-k-1+\delta^2_{v,j;\ell})}{(n-k-2)^2(n-k-4)} = \mathrm{O}(n^{-1}), \\
\mathrm{E}[\widetilde{V}^4_{j,\ell}] &= \mathrm{O}(n^{-2}).
\end{aligned}$$

Then, the bound (5.12) with $s = 2$ can be asymptotically expressed as

$$k_{\boldsymbol{j}*} \sum_{\ell=1}^{p} \eta_{b,j;\ell}^{-4} \mathrm{E}[\widetilde{V}_{j,\ell}^4] = k_{\boldsymbol{j}*} p \left(\frac{1}{p} \sum_{\ell=1}^{p} \eta_{b,j;\ell}^{-4} \right) \times \mathrm{O}(n^{-2}).$$

The above expression shall be $\mathrm{O}(n^{-1})$ under the assumption that $\frac{1}{p}\sum_{\ell=1}^{p} \eta_{b,j;\ell}^{-4}$ tends to a quantity.

6. Asymptotic Cosistency under a Uniform Covariance Structure

In this section we shall show an asymptotic consistency of KOO method based on a general information criterion under a uniform covariance structure. First, following to Sakurai and Fujikoshi (2020), we shall derive a $\mathrm{GIC}_{u,\boldsymbol{j}}$ as in (2.2) and a key statistic $T_{u,j;d}$ as in (3.1). A uniform covariance structure is givn by

$$\boldsymbol{\Sigma}_u = \sigma_u^2(\rho_u^{1-\delta_{ij}}) = \sigma_u^2\{(1-\rho_u)\mathbf{I}_p + \rho_u \mathbf{1}_p \mathbf{1}_p'\}, \tag{6.1}$$

with Kronecker delta δ_{ij}. The covariance structure is expressed as

$$\boldsymbol{\Sigma}_u = \alpha \left(\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p \right) + \beta \frac{1}{p}\mathbf{G}_p,$$

where

$$\alpha = \sigma_u^2(1-\rho_u), \quad \beta = \sigma_u^2\{1+(p-1)\rho_u\}, \quad \mathbf{G}_p = \mathbf{1}_p\mathbf{1}_p',$$

and $\mathbf{1}_p = (1, \ldots, 1)'$. Noting that the matrices $\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p$ and $\frac{1}{p}\mathbf{G}_p$ are orthogonal idempotent matrices, we have

$$|\boldsymbol{\Sigma}_u| = \beta\alpha^{p-1}, \quad \boldsymbol{\Sigma}_u^{-1} = \frac{1}{\alpha}\left(\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p \right) + \frac{1}{\beta} \cdot \frac{1}{p}\mathbf{G}_p.$$

Now we consider the multivariate regression model $M_{u,\boldsymbol{j}}$ given by

$$M_{u,\boldsymbol{j}} : \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}}, \boldsymbol{\Sigma}_{u,\boldsymbol{j}} \otimes \mathbf{I}_n), \tag{6.2}$$

where $\boldsymbol{\Sigma}_{u,\boldsymbol{j}} = \alpha_{\boldsymbol{j}} \left(\mathbf{I}_p - p^{-1}\mathbf{G}_p\right) + \beta_{\boldsymbol{j}} p^{-1}\mathbf{G}_p$. Let $\mathbf{H} = (\boldsymbol{h}_1, \mathbf{H}_2)$ be an orthognal matrix where $\boldsymbol{h}_1 = p^{-1/2}\mathbf{1}_p$, and let

$$\mathbf{W}_{\boldsymbol{j}} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y} \text{ and } \mathbf{U}_{\boldsymbol{j}} = \mathbf{H}'\mathbf{W}_{\boldsymbol{j}}\mathbf{H}.$$

Here, $\boldsymbol{h}_1$ is a characteristic vector of $\boldsymbol{\Sigma}_{u,\boldsymbol{j}}$, and each column vectors of $\mathbf{H}_2$ are the characteristic vectors of $\boldsymbol{\Sigma}_{u,\boldsymbol{j}}$. Let the density function of $\mathbf{Y}$ under $M_{u,\boldsymbol{j}}$ denote by $f(\mathbf{Y}; \boldsymbol{\Theta}_{\boldsymbol{j}}, \alpha_{\boldsymbol{j}}, \beta_{\boldsymbol{j}})$. Then, we have

$$\begin{aligned} g(\alpha_{\boldsymbol{j}}, \beta_{\boldsymbol{j}}) &= -2 \log \max_{\boldsymbol{\Theta}_{\boldsymbol{j}}} f(\mathbf{Y}; \boldsymbol{\Theta}_{\boldsymbol{j}}, \alpha_{\boldsymbol{j}}, \beta_{\boldsymbol{j}}) \\ &= np \log(2\pi) + n(p-1)\log \alpha_{\boldsymbol{j}} + n \log \beta_{\boldsymbol{j}} + \mathrm{tr}\boldsymbol{\Psi}_{\boldsymbol{j}}^{-1}\mathbf{U}_{\boldsymbol{j}}, \end{aligned}$$

where $\boldsymbol{\Psi}_{\boldsymbol{j}} = \mathrm{diag}(\beta_{\boldsymbol{j}}, \alpha_{\boldsymbol{j}}, \ldots, \alpha_{\boldsymbol{j}})$. Therefore, the maximum likelihood estimators of $\alpha_{\boldsymbol{j}}$ and $\beta_{\boldsymbol{j}}$ under $M_{u,\boldsymbol{j}}$ are given by

$$\begin{aligned} \widehat{\alpha}_{\boldsymbol{j}} &= \frac{1}{n(p-1)} \mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y}\mathbf{H}_2, \\ \widehat{\beta}_{\boldsymbol{j}} &= \frac{1}{n}\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{j}})\mathbf{Y}\boldsymbol{h}_1. \end{aligned}$$

The number of independent parameters under $M_{u,\boldsymbol{j}}$ is $m_{\boldsymbol{j}} = k_{\boldsymbol{j}}p + 2$. Noting that $\boldsymbol{\Psi}_{\boldsymbol{j}}$ is diagonal, we can get the general information criterion GIC in (2.4) for $\mathbf{Y}$ in (6.2) as

$$\mathrm{GIC}_{u,\boldsymbol{j}} = n(p-1)\log \widehat{\alpha}_{\boldsymbol{j}} + n \log \widehat{\beta}_{\boldsymbol{j}} + np(\log 2\pi + 1) + d(k_{\boldsymbol{j}}p + 2). \tag{6.3}$$

Therefore we have

$$\begin{aligned} T_{u,j;d} &\equiv \mathrm{GIC}_{u,\boldsymbol{\omega}\backslash j} - \mathrm{GIC}_{u,\boldsymbol{\omega}} \\ &= n(p-1)\log\left\{\widehat{\alpha}_{\boldsymbol{\omega}\backslash j}\left(\widehat{\alpha}_{\boldsymbol{\omega}}\right)^{-1}\right\} + n\log\left\{\widehat{\beta}_{\boldsymbol{\omega}\backslash j}\left(\widehat{\beta}_{\boldsymbol{\omega}}\right)^{-1}\right\} - dp \qquad (6.4) \\ &= Z_{1j} + Z_{2j}. \end{aligned}$$

Here Z_{1j} and Z_{2j} are defined as follows:

$$\begin{aligned} Z_{1j} &= n(p-1)\log\left\{1 + V_{2j}^{(1)}\left(V_1^{(1)}\right)^{-1}\right\} - d(p-1), \\ Z_{2j} &= n\log\left\{1 + V_{2j}^{(2)}\left(V_1^{(2)}\right)^{-1}\right\} - d, \qquad (6.5) \end{aligned}$$

using the following $V_1^{(i)}$, $V_{2j}^{(i)}$, $i = 1, 2$:

$$V_1^{(1)} = \mathrm{tr}\mathbf{H}_2'\mathbf{Y}(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{Y}\mathbf{H}_2, \quad V_{2j}^{(1)} = \mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\backslash j})\mathbf{Y}\mathbf{H}_2,$$
$$V_1^{(2)} = \boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{Y}\boldsymbol{h}_1, \quad V_{2j}^{(2)} = \boldsymbol{h}_1'\mathbf{Y}'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\backslash j})\mathbf{Y}\boldsymbol{h}_1.$$

Related to the distributional reductions of Z_{1j}, Z_{2j}, $j = 1, \ldots, k$, we use the following Lemma frequently.

Lemma 6.1. *Let* $\mathbf{W}$ *have a noncentral Whishart distribution* $\mathrm{W}_p(m, \Sigma; \mathbf{\Omega})$. *Let the covariance matrix* $\mathbf{\Sigma}$ *be decomposed into the characteristic roots and vectors as*

$$\begin{aligned}\mathbf{\Sigma} &= \mathbf{H}\mathbf{\Lambda}\mathbf{H}' \\ &= (\mathbf{H}_1, \ldots, \mathbf{H}_h)\mathrm{diag}(\lambda_1\mathbf{I}_{q_1}, \ldots, \lambda_h\mathbf{I}_{q_h})(\mathbf{H}_1, \ldots, \mathbf{H}_h)',\end{aligned}$$

where $\lambda_1 > \ldots > \lambda_h > 0$ *and* $\mathbf{H}$ *is an orthogonal matrix. Then,* $\mathrm{tr}\mathbf{H}_j'\mathbf{W}_j\mathbf{H}_j$, $i = 1, \ldots, h$, *are independently distributed to noncentral chi-square distributions with* mk_j *degrees of freedom and noncentrality parameters* $\delta_j^2 = \mathrm{tr}\mathbf{H}_j'\mathbf{\Omega}\mathbf{H}_j$.

Proof. The result may be proved by considering the characteristic function of $(\mathrm{tr}\mathbf{H}_1'\mathbf{W}\mathbf{H}_1, \ldots, \mathrm{tr}\mathbf{H}_q'\mathbf{W}\mathbf{H}_q)$ which is expressed as (see, Fujikoshi et al. (2010, Theorem 2.1.2))

$$\begin{aligned}&\mathrm{E}\left[\mathrm{e}^{it_1\mathrm{tr}\mathbf{H}_1'\mathbf{W}\mathbf{H}_1+\cdots+it_h\mathrm{tr}\mathbf{H}_h'\mathbf{W}\mathbf{H}_h}\right] \\ &= \mathrm{E}\left[\mathrm{etr}(\mathbf{K})\right] \\ &= |\mathbf{I}_p - 2\mathbf{\Sigma}\mathbf{K}|^{-m/2}\mathrm{etr}\left\{\mathbf{\Omega}\mathbf{K}(\mathbf{I}_p - 2\mathbf{\Sigma}\mathbf{K})^{-1})\right\},\end{aligned}$$

where $\mathbf{K} = it_1\mathbf{H}_1\mathbf{H}_1' + \cdots + it_1\mathbf{H}_q\mathbf{H}_q'$. The result can easily obtained by checking that the above larst expression equals to

$$\prod_{j=1}^{q}(1 - 2it_j)^{-nk_j/2}\exp\left\{\frac{it_j}{1 - 2it_j}\mathrm{tr}\mathbf{H}_j'\mathbf{\Omega}\mathbf{H}_j\right\}.$$

□

Assume that the true model is expressed as

$$M_{u,\boldsymbol{j}_*} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*}, \boldsymbol{\Sigma}_{u,*} \otimes \mathbf{I}_n), \tag{6.6}$$

where $\boldsymbol{\Sigma}_{u,*} = \alpha_* (\mathbf{I}_p - p^{-1}\mathbf{G}_p) + \beta_* p^{-1}\mathbf{G}_p$. Using Lemma 6.1, we have the following lemma.

Lemma 6.2. *Under the true model* (6.6), *it holds that*
(1) $V_1^{(1)}/\alpha_*$ *and* $V_{2j}^{(1)}/\alpha_*$ *are independently distributed to a central chi-square distributin* $\chi^2_{(p-1)(n-k)}$ *and a noncentral chi-square distribution* $\chi^2_{p-1}(\delta^2_{1j})$, *respectively.*
(2) $V_1^{(2)}/\beta_*$ *and* $V_{2j}^{(2)}/\beta_*$ *are independently distributed to a central chi-square distributin* χ^2_{n-k} *and a noncentral chi-square distribution* $\chi^2_1(\delta^2_{2j})$, *respectively.*
(3) *The noncentrality parameters* δ^2_{1j} *and* δ^2_{2j} *are defined as follows:*

$$\delta^2_{1j} = \frac{1}{\alpha_*}\mathrm{tr}\mathbf{H}_2'(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*})'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*})\mathbf{H}_2$$
$$\delta^2_{2j} = \frac{1}{\beta_*}\boldsymbol{h}_1'(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*})'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\boldsymbol{\omega}\setminus j})(\mathbf{X}_{\boldsymbol{j}_*}\boldsymbol{\Theta}_{\boldsymbol{j}_*})\boldsymbol{h}_1.$$

Here, if $j \notin \boldsymbol{j}_*$, *then* $\delta^2_{1j} = 0$ *and* $\delta^2_{2j} = 0$.

Now we consider high-dimensional asymptotic consistency of the KOO method based on $\mathrm{T}_{b,j;d}$ in (5.4), whose selection method is given by $\widehat{\boldsymbol{j}}_{v,j;d} = \{j \mid T_{b,j;d} > 0\}$. For a sufficient condition for consistency of $\widehat{\boldsymbol{j}}_{v,j;d}$, we assume

A3u: For any $j \in \boldsymbol{j}_*$, $\delta^2_{1j} = \mathrm{O}(np)$, $\delta^2_{2j} = \mathrm{O}(n)$ and

$$\lim \frac{1}{np}\delta^2_{1j} = \eta^2_{1j} > 0, \quad \lim \frac{1}{n}\delta^2_{2j} = \eta^2_{2j} > 0, \tag{6.7}$$

When $j \notin \boldsymbol{j}_*$, we have

$$\begin{aligned}[\mathrm{F2}] &= \sum_{j\notin \boldsymbol{j}_*} \{\Pr(Z_{1j} + Z_{2j} \geq 0)\} \\ &\leq \sum_{j\notin \boldsymbol{j}_*} \{\Pr(Z_{1j} \geq 0) + \Pr(Z_{2j} \geq 0)\} \\ &= (k - k_{\boldsymbol{j}_*}) \left\{\Pr(Z^{(1)} \geq s_0^{(1)}) + \Pr(Z^{(2)} \geq s_0^{(2)})\right\}.\end{aligned}$$

Here,

$$\begin{aligned}
Z^{(1)} &= \frac{\chi^2_{p-1}}{\chi^2_{(p-1)(n-k)}} - \frac{p-1}{(p-1)(n-k)-2},\\
s^{(1)} &= \mathrm{e}^{d/n} - 1 - \frac{p-1}{(p-1)(n-k)-2}, \quad s_0^{(1)} = \frac{d}{n} - \frac{p-1}{(p-1)(n-k)-2},\\
Z^{(2)} &= \frac{\chi^2_1}{\chi^2_{n-k}} - \frac{1}{n-k-2},\\
s^{(2)} &= \mathrm{e}^{d/n} - 1 - \frac{1}{n-k-2}, \quad s_0^{(2)} = \frac{d}{n} - \frac{1}{n-k-2}.
\end{aligned}$$

Note that $s_0^{(1)} < s^{(1)}$ and $s_0^{(2)} < s^{(2)}$. Then, under the assumption that $s_0^{(1)} > 0$ and $s_0^{(2)} > 0$ we have

$$[\mathrm{F2}] \le (k - k_{\boldsymbol{j}_*}) \left[\left(s_0^{(1)}\right)^{-2\ell} \mathrm{E}\left[(Z^{(1)})^{2\ell}\right] + \left(s_0^{(2)}\right)^{-2\ell} \mathrm{E}\left[(Z^{(2)})^{2\ell}\right] \right]. \tag{6.8}$$

Related to the assumptions $s_0^{(1)} > 0$ and $s_0^{(2)} > 0$, we assume

$$\begin{aligned}
\mathrm{A4u} : d > \frac{n(p-1)}{(p-1)(n-k)-2} \to \frac{1}{1-c_2}, \quad d > \frac{n}{n-k-2} \to \frac{1}{1-c_2},\\
\text{and } d = \mathrm{O}(n^a), \quad 0 < a < 1.
\end{aligned} \tag{6.9}$$

The first part in A4u implies $s_0^{(1)} > 0$ and $s_0^{(2)} > 0$. It is easy to see that

$$\begin{aligned}
\mathrm{E}[(Z^{(1)})^2] &= \frac{2(p-1)^2(n-k+1)}{\{(p-1)(n-k)-2\}^2\{(p-1)(n-k)-4\}} = \mathrm{O}((n^3)^{-1}),\\
\mathrm{E}[(Z^{(2)})^2] &= \frac{2(n-k-1)}{(n-k-2)^2(n-k-4)} = \mathrm{O}((n^2)^{-1}).
\end{aligned}$$

Further, $(s_0^{(1)})^{-2} = \mathrm{O}(n^{2(1-a)})$ and $(s_0^{(2)})^{-2} = \mathrm{O}(n^{2(1-a)})$. Therefore, from (6.8) we have that $[\mathrm{F2}] \to 0$.

When $j \in \boldsymbol{j}_*$, we can write $T_{b,j;d} = n \sum_{\ell=1}^p \log\{1 + U_{2\ell} U_{1\ell}^{-1}\} - dp$. Therefore we can express [F1] as

$$[F1] = \sum_{j \in \boldsymbol{j}_*} \Pr(\widetilde{T}_{b,j;d} \le 0),$$

where

$$\widetilde{T}_{b,j;d} = \frac{1}{p}\sum_{\ell=1}^{p} \log\left\{1 + \frac{\chi^2_{1;\ell}(\delta^2_{b,j;\ell})}{\chi^2_{n-k;\ell}}\right\} - \frac{d}{n}.$$

From the assumptions A3b and A4b it is easily seen that

$$\widetilde{T}_{v,j;d} \to \log(1+\gamma^2_{v,j}) > 0.$$

This implies that $\Pr(\widetilde{T}_{v,j;d} \leq 0) \to 0$, and [F1] $\to 0$.

These imply the following theorem.

Theorem 6.1. *Suppose that the assumptions* A1, A2, A3u and A4u *are satisfied. Then, the* KOO *method based on* $T_{u,j:d}$ *in* (6.4) *is asymptotically consistent.*

7. Concluding Remarks

In this paper, we consider to select regression variables in p variate regression model with one of three covariance structures; (1) ICSS (an independent covariance structure with the same variance), (2) ICSD (an independent covariance structure with different variances), (3) UCS (a uniform covariance structure). It was proposed to use a KOO method based on a general information criterion with a penality term d. We point high-dimensional consistencies of the KOO methods with $d = \mathrm{O}(n^a), 0 < a < 1$. In Sakurai and Fujikoshi (2020), they studied asymptotic consistencies of KOO methods in (1) and (3). However, in their approach, the number of explanatory variables was fixed, but in this paper the number of explanatory variables may be tend to infinity.

It may noted that KOO methods may be feasible in computation. The idear goes back to Nishii et al. (1988) and Zhao et. al. (1988). However, high dimensional properties have been recently studied in Fujikoshi and Sakurai (2019), Oda and Yanagihara (2021, 2022), Fujikoshi (2022).

A high-dimensional study KOO method under AUTO (autoregressive covariance structure) is left. It is also left to extend our results to the case of non-normality.

Acknowledgements

The authors would like to express their gratitute to Professor Vldimir V. Ulyanov for his valuable comments and suggestions.

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 and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

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

[3] Bedrick, E. J. and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

[4] Fujikoshi, Y. (2022). High-dimensional consistencies of KOO methods in multivariate regression model and discriminant analysis. *J. Multivariate Anal.*, **188**, 104860.

[5] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

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

[7] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **149**, 199–212.

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

[9] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

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

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

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

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

[14] Sakurai, T. and Fujikoshi, Y. (2020). Exploring consistencies of information criterion and test-based criterion for high-dimensional

multivariate regression models under three covariance structures. In Festschrift in honor of Professr Dietrich von Rosen's 65th birthday (eds, T. Holgerson and M. Singnull), 313-334, Springer.

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

[16] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.*, **9**, 869–897.

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