# A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables

**Ryoya Oda*** **and Hirokazu Yanagihara**

*Department of Mathematics, Graduate School of Science, Hiroshima University*

1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

(Last Modified: January 6, 2019)

**Abstract**

We put forward a variable selection method for selecting explanatory variables in a normality-assumed multivariate linear regression. It is cumbersome to calculate variable selection criteria for all subsets of explanatory variables when the number of explanatory variables is large. Therefore, we propose a fast and consistent variable selection method based on Zhao *et al.* (1986) and Nishii *et al.* (1988). The consistency of the method is provided by a high-dimensional asymptotic framework such that the dimensions of response vectors and explanatory vectors $p$ and $k$ may tend to infinity with sample size $n$ but $(p+k)/n$ converges to a constant within $[0, 1)$. Through numerical simulations, it is shown that the proposed method has a high probability of selecting the true subset of explanatory variables and is fast under a moderate sample size even when the number of dimensions is large.

## 1 Introduction

Multivariate linear regression is a widely known method of inferential analysis. It features in many theoretical and applied textbooks (see, e.g., Srivastava, 2002, chap 9; Timm, 2002, chap 4) and it is used by researchers in many fields. Let $\boldsymbol{Y}$ be an $n \times p$ observation matrix of $p$ response variables and $\boldsymbol{X}$ be an $n \times k$ observation matrix of $k$ non-stochastic explanatory variables, where $n$ is the sample size, and $p$ and $k$ are the numbers of response variables and explanatory variables, respectively. Let $N = n - p - k + 1$ and $D = \{(n, p, k) \in \mathbb{N}^3 \mid N - 4 > 0\}$. Further, we assume that $\mathrm{rank}(\boldsymbol{X}) = k$ and $(n, p, k) \in D$ in proposing our method.

In actual empirical contexts, it is important to specify the factors affecting response variables. In multivariate linear regression, this is regarded as the problem of selecting a subset of explanatory variables. Suppose that $j$ denotes a subset of $\omega = \{1, \ldots, k\}$ containing $k_j$ elements, and $\boldsymbol{X}_j$ denotes the $n \times k_j$ matrix consisting of columns of $\boldsymbol{X}$ indexed by the elements of $j$, where $k_A$ denotes the number of elements in a set $A$, i.e., $k_A = \#(A)$. Next, $j$ expresses the subset of explanatory variables. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second and

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

fourth column vectors of $\boldsymbol{X}$. Using the notation $j$, the candidate model with $k_j$ explanatory variables is expressed as follows:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients and $\boldsymbol{\Sigma}_j$ is a $p\times p$ unknown covariance matrix. In particular, the total number of explanatory variables $k_\omega$ and the explanatory matrix $\boldsymbol{X}_\omega$ in the full model $\omega$ express $k$ and $\boldsymbol{X}$, respectively. Herein, we assume that the data are generated from the following true model with $k_{j_*}$ explanatory variables:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_{j_*}\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n),$$

where $\boldsymbol{\Theta}_*$ is a $k_{j_*} \times p$ true unknown matrix of regression coefficients and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true unknown covariance matrix assuming that $\boldsymbol{\Sigma}_*$ is positive definite. For expository purposes, we represent $k_{j_*}$ and $\boldsymbol{X}_{j_*}$ as $k_*$ and $\boldsymbol{X}_*$, respectively.

To systematize and optimize the configuration of models, variable selection criteria have been widely used. Mallows (1973; 1995) proposed the $C_p$ criterion. In this paper, we focus on a generalized variable selection criterion based on the $C_p$ criterion, termed the Generalized $C_p$ ($GC_p$) criterion. The $GC_p$ criterion for a linear regression with a single response was proposed by Atkinson (1980), and the counterpart for a multivariate linear regression with multiple responses was proposed by Nagai *et al.* (2012). The $GC_p$ criterion can express a wide variety of variable selection criteria, e.g., the $C_p$ criterion for multivariate contexts proposed by Sparks *et al.* (1983), and the modified $C_p$ ($MC_p$) criterion proposed by Fujikoshi and Satoh (1997).

The best subset chosen by a variable selection criterion is usually defined as the subset of explanatory variables which minimizes the value of that criterion among all candidate subsets. The basic approach to identifying the best subset involves searching over all candidate subsets. We call this method the "full search method". To elaborate, assuming a full search method is used, variable selection criteria for $2^k - 1$ subsets need to be calculated. Recently, increasing attention has been paid to investigating statistical methods for high-dimensional data, in which the dimension of response vectors $p$ or the number of explanatory variables $k$ is large. However, in high-dimensional data contexts, particularly where $k$ is large, it may be impossible to apply the full search method because the total number of subsets of explanatory variables exponentially increases when $k$ becomes large. For example, if $k = 40$ and the time taken to calculate a variable selection criterion for a subset is 0.01 seconds, then the time required to implement the full search method will be $(2^{40} - 1) \times 0.01$ seconds, i.e., about 35 years. Thus, for practical reasons, we need another search method when $k$ is large. Zhao *et al.* (1986) and Nishii *et al.* (1988) proposed a practicable selection method when $k$ is large. This method is based on the behavior of variable selection criteria for the subset where a variable is removed from the full set $\omega$. In that selection method, the best subset $\hat{j}$ is determined as follows. For each explanatory variable, if the criterion for the subset where a variable is removed from $\omega$ is greater than the criterion for the full set $\omega$, then the removed variable is regarded as the element of the best subset. Since this method is needed to calculate variable selection criteria for only $k$ subsets and $\omega$ for searching the best subset $\hat{j}$, we expect that the method is faster than the full search method, and it is practical for high-dimensional data contexts. We call this method the "ZKB selection method" and consider it using a class of the $GC_p$ criterion.

An important property of a variable selection criterion is its consistency. Consistency is achieved where the probability of selecting the true subset $j_*$ converges to 1, i.e., $P(\hat{j} = j_*) \to 1$. However, since we do not know the true subset $j_*$, we often hope to specify $j_*$ by variable selection. Then, we should use a variable selection criterion that maximizes the probability of selecting the true subset. It is expected that a consistent variable selection criterion has a high-probability of selecting the true subset $j_*$. Hence, it is important to ensure the consistency of the selection method using a variable selection criterion. To this end, Zhao *et al.* (1986), Nishii *et al.* (1988), Rao and Wu (1989), and Nishii (1988) used the large-sample (LS) asymptotic framework such that only $n$ tends to infinity. However, it is not appropriate to use the LS asymptotic framework for high-dimensional data because approximate accuracy using the LS asymptotic framework deteriorates as $p$ or $k$ become large.

The aim of this paper is to propose the ZKB selection method using a class of the $GC_p$ criterion, which is consistent even in high-dimensional contexts. To achieve this, we use the following high-dimensional (HD) asymptotic framework:

$$n \to \infty, \ \frac{p+k}{n} \to c \in [0, 1).$$

Importantly, the HD asymptotic framework includes the following six asymptotic frameworks:

- $n \to \infty, \ p, k$: fixed,
- $(n, p) \to \infty, \ p/n \to c \in [0, 1), \ k$: fixed,
- $(n, k) \to \infty, \ k/n \to c \in [0, 1), \ p, k_*$: fixed,
- $(n, k, k_*) \to \infty, \ k/n \to c \in [0, 1), \ p$: fixed,
- $(n, p, k) \to \infty, \ (p + k)/n \to c \in [0, 1), \ k_*$: fixed,
- $(n, p, k, k_*) \to \infty, \ (p + k)/n \to c \in [0, 1)$.

Hence, our proposed method is consistent under all the above situations. Thus it is expected that our proposed method will have a high probability of selecting the true subset where $n$ is large regardless of the sizes of $p$, $k$ and $k_*$.

The remainder of the paper is organized as follows. In section 2, we present the necessary notation and assumptions to ensure consistency of our method. In section 3, we put forward the proposed method, explicate its consistency, and present a fast algorithm. We also propose an extended ZKB selection method. In section 4, we conduct numerical experiments for verification purposes. Technical details are relegated to the Appendix.

## 2 Preliminaries

First, we present the $GC_p$ criterion. Let $\boldsymbol{S}_j$ be the unbiased estimator of $\boldsymbol{\Sigma}_j$ in model (1), which is defined by

$$\boldsymbol{S}_j = \frac{1}{n - k_j} \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y},$$

where $\boldsymbol{P}_j$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_j$, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. Then, the $GC_p$ criterion in model (1) is defined by

$$GC_p(j) = (n - k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1}) + \alpha p k_j, \tag{2}$$

where $\alpha$ is a positive constant. The first and second terms in (2) express the residual sum of squares with the weighted matrix $\boldsymbol{S}_\omega^{-1}$ and $\alpha$ times the strength of the penalty for the number of elements of $\boldsymbol{\Theta}_j$ in model (1), respectively.

Next, we present notation and assumptions to ensure consistency of our method. For a subset $j \subset \omega$, let a $p \times p$ non-centrality matrix and parameter be denoted by

$$\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2},\ \delta_j = \mathrm{tr}(\boldsymbol{\Delta}_j). \tag{3}$$

where $\omega_j = j^c$ and $j^c$ denotes as $\omega\backslash j$. It should be emphasized that $\boldsymbol{\Delta}_j = \boldsymbol{O}_{p,p}$ and $\delta_j = 0$ hold if and only if $j \subset j_*^c$, where $\boldsymbol{O}_{p,p}$ is a $p \times p$ matrix of zeros. To ensure the consistency of our method, the following two assumptions are prepared:

Assumption A1. $j_* \subset \omega$.
Assumption A2. $\forall \ell \in j_*,\ \inf_{(n,p,k)\in D} \frac{1}{n}\delta_{\{\ell\}} > 0.$

Assumption A1 is needed to consider consistency because the probability of selecting the true subset becomes 0 if it does not hold. Assumption A2 restricts the divergence order of the non-centrality parameter $\delta_{\{\ell\}}$. If $k$ is fixed, Assumption A2 is as per what was put forward in Yanagihara (2016).

Finally, we identify the upper bound of the rank of the non-centrality parameter matrix $\boldsymbol{\Delta}_j$, which is used to ensure consistency. For a subset $j \subset \omega$ ($j \neq \omega$), let $m_j$ and $d_j$ be the number of elements of $j$ and the rank of $\boldsymbol{\Delta}_j$ as follows:

$$m_j = \#(j),\ d_j = \mathrm{rank}(\boldsymbol{\Delta}_j). \tag{4}$$

In accordance with Yanagihara *et al.* (2015), it follows from Assumption A1 that the rank of $\boldsymbol{X}_*'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*$ is calculated as

$$\mathrm{rank}(\boldsymbol{X}_*'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*) = \begin{cases} 0 & (j \subset j_*^c) \\ m_j & (j \subset j_*) \end{cases}.$$

It is straightforward that $\mathrm{rank}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*') \leq \min\{p, k_*\}$. Since $m_j \leq k_*$ holds when $j \subset j_*$, the following equation can be derived:

$$d_j \leq \min\{\mathrm{rank}(\boldsymbol{X}_*'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*), \mathrm{rank}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*')\} \leq \begin{cases} 0 & (j \subset j_*^c) \\ \min\{m_j, p\} & (j \subset j_*) \end{cases}. \tag{5}$$

# 3 Main Results

## 3.1 Proposed Selection Method

We define a class of the $GC_p$ criterion, denoted as the high-dimensionality-adjusted consistent generalized $C_p$ ($HCGC_p$) criterion:

**Definition 3.1.** *The $HCGC_p$ criterion is defined by the $GC_p$ criterion* (2) *satisfying*

$$\alpha = \frac{n-k}{N-2} + \beta,\ \beta > 0 \ s.t. \ \frac{\sqrt{p}}{\sqrt[2r_1]{k}}\beta \to \infty,\ \frac{\sqrt[2r_2]{k}p}{n}\beta \to 0, \tag{6}$$

*as* $n \to \infty,\ (p+k)/n \to c \in [0,1)$, *for some* $r_1 \in \mathbb{N}$ *and* $r_2 \in \mathbb{N}\backslash\{1\}$.

We now introduce the ZKB selection method using a variable selection criterion (SC). Let $\ell$ be an element of $\omega$. The best subset chosen by the ZKB selection method using an SC is written as

$$\{\ell \in \omega \mid \mathrm{SC}(\omega_{\{\ell\}}) > \mathrm{SC}(\omega)\},$$

where $\omega_{\{\ell\}}$ expresses $\{\ell\}^c$ or $\omega\backslash\{\ell\}$. The ZKB selection method is based on the idea that the value of the SC for the subset where a true variable is removed from $\omega$ will be greater than that for $\omega$ asymptotically. We define the following best subset chosen by the ZKB selection method using the $HCGC_p$ criterion:

**Definition 3.2.** *The best subset chosen by the ZKB selection method using the $HCGC_p$ criterion is defined by*

$$\hat{j} = \{\ell \in \omega \mid HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)\}. \tag{7}$$

Next, to use this method in actual empirical contexts we have to decide the value of $\alpha$ because the $HCGC_p$ criterion is expressed as the class of criteria. Hence, we show the following value of $\alpha$:

$$\tilde{\alpha} = \frac{n-k}{N-2} + \tilde{\beta},\ \tilde{\beta} = \frac{(n-k)\sqrt{N+p-4}}{(N-2)\sqrt{N-4}} \cdot \frac{\sqrt[4]{k}\log n}{\sqrt{p}}. \tag{8}$$

This $\tilde{\alpha}$ is based on Yanagihara (2016). It is straightforward to observe that $\tilde{\beta}$ is satisfied with $(\sqrt{p}/\sqrt[6]{k})\tilde{\beta} \to \infty$ and $(\sqrt[6]{k}p/n)\tilde{\beta} \to 0$ as $n \to \infty,\ (p+k)/n \to c \in [0,1)$. Therefore, the $GC_p$ criterion with $\alpha = \tilde{\alpha}$ is included in the class of the $HCGC_p$ criterion. In practice, regardless of whether there is the constant value $\{(n-k)\sqrt{N+p-4}\}/\{(N-2)\sqrt{N-4}\}$ in $\tilde{\beta}$, the criterion belongs to the class of the $HCGC_p$ criterion. However, the constant value plays a role in terms of stabilizing the behavior of $p^{-1/2}\{HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega)\}$ for $\ell \in j_*^c$.

Since the ZKB selection method using the $GC_p$ criterion only necessitates calculating the differences $GC_p(\omega_{\{\ell\}}) - GC_p(\omega)$ for $\ell = 1, \ldots, k$, it can be expected that the calculation time associated with this method will be shorter than that for the full search method. However, it is important that $GC_p(\omega_{\{\ell\}})$ consists of the projection matrix $\boldsymbol{P}_{\omega_{\{\ell\}}} = \boldsymbol{X}_{\omega_{\{\ell\}}}(\boldsymbol{X}'_{\omega_{\{\ell\}}}\boldsymbol{X}_{\omega_{\{\ell\}}})^{-1}\boldsymbol{X}'_{\omega_{\{\ell\}}}$ and the calculation time of an inverse matrix costs about the cube of the size of the matrix. Hence, it is not advisable to calculate $(\boldsymbol{X}'_{\omega_{\{\ell\}}}\boldsymbol{X}_{\omega_{\{\ell\}}})^{-1}$ for each $\ell$ when $k$ is large. To overcome this problem, we offer an efficient calculation of $GC_p(\omega_{\{\ell\}}) - GC_p(\omega)$. Let $r_\ell$ and $\boldsymbol{z}_\ell$ be the $(\ell, \ell)$-th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and the $\ell$-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, respectively. Then, using $r_\ell$ and $\boldsymbol{z}_\ell$, we can express $\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_{\{\ell\}}}$ as follows (the proof of (9) is given in Appendix A):

$$\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_{\{\ell\}}} = \frac{1}{r_\ell}\boldsymbol{z}_\ell\boldsymbol{z}'_\ell. \tag{9}$$

Using the above equation, $GC_p(\omega_{\{\ell\}}) - GC_p(\omega)$ can be expressed as

$$GC_p(\omega_{\{\ell\}}) - GC_p(\omega) = \frac{1}{r_\ell}\boldsymbol{z}_\ell'\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell - p\alpha. \tag{10}$$

Note that (10) does not need to calculate $(\boldsymbol{X}_{\omega_{\{\ell\}}}'\boldsymbol{X}_{\omega_{\{\ell\}}})^{-1}$ if only $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ can be calculated. Moreover, the calculation cost of the product of each $\boldsymbol{Y}'\boldsymbol{z}_\ell$ relies on $n$. Hence, we also present an efficient calculation of $\boldsymbol{z}_\ell'\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell$ when $p$ is small. Let $\boldsymbol{t}_\ell$ be the $\ell$-th column vector of $\boldsymbol{S}_\omega^{-1/2}\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then, the following equation can be derived:

$$\boldsymbol{z}_\ell'\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell = \boldsymbol{t}_\ell'\boldsymbol{t}_\ell. \tag{11}$$

Since $\boldsymbol{t}_\ell$ is a $p$-dimensional vector, the calculation cost of $\boldsymbol{t}_\ell'\boldsymbol{t}_\ell$ does not rely on $n$. Therefore, we propose to use (10) (and also use (11) when $p$ is small) to perform the ZKB selection method using the $GC_p$ criterion.

## 3.2 Consistency of Proposed Selection Method

We ensure the consistency of the ZKB selection method using the $HCGC_p$ criterion (7). To do so, we present a lemma for the sufficient conditions for consistency (the proof is given in Appendix B). Importantly, Lemma 3.1 does not rely on a specific asymptotic framework, indeed any such framework could be applied here.

**Lemma 3.1.** *Suppose that Assumption A1 and the following equations hold*:

$$\sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) \to 0, \sum_{\ell \in j_*} P(HCGC_p(\omega_{\{\ell\}}) < HCGC_p(\omega)) \to 0. \tag{12}$$

*Then, the ZKB selection method using the $HCGC_p$ criterion* (7) *is consistent, that is $P(\hat{j} = j_*) \to 1$ holds.*

By showing that the sufficient conditions (12) in Lemma 3.1 hold, the consistency of the ZKB selection method using the $HCGC_p$ criterion (7) can be obtained as follows (the proof is given in Appendix C):

**Theorem 3.1.** *Suppose that Assumptions A1 and A2 hold. Then, the ZKB selection method using the $HCGC_p$ criterion* (7) *is consistent as $n \to \infty,\ (p+k)/n \to c \in [0, 1)$.*

From Theorem 3.1, the ZKB selection method using the $HCGC_p$ criterion with $\alpha = \tilde{\alpha}$ given by (8) is also consistent under Assumptions A1 and A2.

## 3.3 Extension of the ZKB selection method

In the previous sub sections, we proposed the ZKB selection method using the $HCGC_p$ criterion (7). However, when the full model $\omega$ includes several explanatory variables such as multinomial variables, it will be not appropriate to use the ZKB selection method because whether such explanatory variables should be chosen or not should be decided simultaneously. To overcome this problem, we extend the ZKB selection method. Let $\mathcal{J}$ be a family of sets of some explanatory variables denoted by $\mathcal{J} = \{j_1, \ldots, j_q\}$, where $q$ is the number of these sets. Since we suppose dummy variables or non-dummy variables as explanatory variables, we assume that $m_{j_a}$ is finite, $j_a$ is satisfied with $j_a \subset j_*$ or $j_a \subset j_*^c$ and $j_a \cap j_b = \emptyset$ $(a \neq b)$ for $j_a, j_b \in \mathcal{J}$, where $m_{j_a}$ is defined by (4). Then, it is clear that $\cup_{a=1}^q j_a = \omega$ holds. For example, if $k = 7$ and the sets of explanatory variables are $\{1\}$, $\{2\}$, $\{3, 5\}$ and $\{4, 6, 7\}$ then $\mathcal{J} = \{\{1\}, \{2\}, \{3, 5\}, \{4, 6, 7\}\}$, $q = 4$, and the subsets $\{3, 5\}$ and $\{4, 6, 7\}$ express the subsets of binomial and trinomial dummy variables, respectively. Using this notation, we consider the following best subset chosen by the extended ZKB (EZKB) selection method using an SC:

$$\{j \in \mathcal{J} \mid \mathrm{SC}(\omega_j) > \mathrm{SC}(\omega)\}.$$

We observe that the EZKB selection method is equivalent to the ZKB selection method (7) when $m_j = 1$ $(\forall j \in \mathcal{J})$ or $q = k$. Moreover, since the EZKB selection method can accommodate the selection of grouped explanatory variables, the method is similar to Group Lasso as proposed by Yuan and Lin (2006). We define the following best subset chosen by the EZKB selection method using the $HCGC_p$ criterion:

**Definition 3.3.** *The best subset chosen by the EZKB selection method using the $HCGC_p$ criterion is defined by*

$$\hat{j}_{\mathcal{J}} = \{j \in \mathcal{J} \mid HCGC_p(\omega_j) > HCGC_p(\omega)\}. \tag{13}$$

Next, we ensure the consistency of the EZKB selection method using the $HCGC_p$ criterion (13). Let $\mathcal{J}_+ = \{j \in \mathcal{J} \mid j \subset j_*\}$ and $\mathcal{J}_- = \{j \in \mathcal{J} \mid j \subset j_*^c\}$. Then, as with Lemma 3.1, we present the following lemma for the sufficient conditions for consistency (the proof is given in Appendix D).

**Lemma 3.2.** *Suppose that Assumption A1 and the following equations hold:*

$$\sum_{j \in \mathcal{J}_+} P(HCGC_p(\omega_j) < HCGC_p(\omega)) \to 0, \quad \sum_{j \in \mathcal{J}_-} P(HCGC_p(\omega_j) > HCGC_p(\omega)) \to 0.$$

*Then, the EZKB selection method using the $HCGC_p$ criterion* (13) *is consistent.*

Using Lemma 3.2, the consistency of the EZKB selection method using the $HCGC_p$ criterion (13) can be obtained as follows (the proof is given in Appendix E):

**Theorem 3.2.** *Suppose that Assumptions A1 and A2 hold. Then, the EZKB selection method using the $HCGC_p$ criterion* (13) *is consistent as $n \to \infty$, $(p + k)/n \to c \in [0, 1)$.*

From Theorem 3.2, we can observe that the EZKB selection method using the $HCGC_p$ criterion is also consistent as with the ZKB selection method (7). Hence, as an example of the consistent EZKB selection method, we can use the method using the $HCGC_p$ criterion with $\alpha = \tilde{\alpha}$ in (8).

Finally, we provide an efficient calculation of $GC_p(\omega_j) - GC_p(\omega)$. Let $\boldsymbol{R}_j$ and $\boldsymbol{Z}_j$ be the $m_j \times m_j$ and $n \times m_j$ matrices consisting of the row and column elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and the column vectors of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ indexed by the elements of $j$, respectively. For example, if $j = \{2, 5\}$, then $\boldsymbol{R}_j$ and $\boldsymbol{Z}_j$ are expressed as

$$\boldsymbol{R}_j = \begin{pmatrix} \tilde{x}_{22} & \tilde{x}_{25} \\ \tilde{x}_{52} & \tilde{x}_{55} \end{pmatrix}, \ \boldsymbol{Z}_j = (\tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_5),$$

where $\tilde{x}_{ab}$ is the $(a, b)$-element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\tilde{\boldsymbol{z}}_a$ is the $a$-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then, using $\boldsymbol{R}_j$ and $\boldsymbol{Z}_j$, $GC_p(\omega_j) - GC_p(\omega)$ can be expressed as

$$GC_p(\omega_j) - GC_p(\omega) = \mathrm{tr}(\boldsymbol{R}_j^{-1}\boldsymbol{Z}_j'\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{Z}_j) - m_j p\alpha. \tag{14}$$

The proof of the above equation is omitted because it essentially mimics (9). Although (14) requires the calculation of the inverse matrix of $\boldsymbol{R}_j$, it will not be computationally onerous because the size is finite.

# 4 Numerical studies

We present numerical results to explore the validity of our claim based on Monte Carlo simulations with $1,000$ iterations executed in MATLAB 9.3.0 on a Panasonic CF-SV7UFKVS with an Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz and 16 GB of RAM. The probabilities of selecting the true subset and the CPU times are presented for the ZKB selection methods using the $HCGC_p$ criterion with $\alpha = \tilde{\alpha}$ given in (8) and the three $GC_p$ criteria with $\alpha = 2$, $2\log\log n$ and $\log n$ (named $GC_p^{(1)}$, $GC_p^{(2)}$ and $GC_p^{(3)}$). The calculations were performed using (10) (and (11) if $p < 100$ and $k \geq p$). We constructed the true model: $\boldsymbol{Y} \sim N_{n\times k}(\boldsymbol{X}(\boldsymbol{\Theta}_*', \boldsymbol{O}_{k-k_*,p}')', \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n)$. The explanatory matrix $\boldsymbol{X}$, the true coefficient matrix $\boldsymbol{\Theta}_*$ and the true covariance matrix $\boldsymbol{\Sigma}_*$ were determined as follows:

$$\boldsymbol{X} \sim N_{n\times k}(\boldsymbol{O}_{n,k}, \boldsymbol{\Psi} \otimes \boldsymbol{I}_n),\ \boldsymbol{\Theta}_* \sim N_{k_*\times p}(\boldsymbol{O}_{k_*,p}, \boldsymbol{I}_p \otimes \boldsymbol{I}_{k_*}),\ \boldsymbol{\Sigma}_* = \xi_1\{(1-\xi_2)\boldsymbol{I}_p + \xi_2\boldsymbol{1}_p\boldsymbol{1}_p'\},$$

where $\boldsymbol{\Psi}$ is the $k \times k$ autoregressive matrix with the correlation $\psi$, i.e., $(\boldsymbol{\Psi})_{ab} = \psi^{|a-b|}$, and $\boldsymbol{1}_p$ is a $p$-dimensional vector of ones. Further, we set $\psi = 0.5$, $\xi_1 = 0.4$ and $\xi_2 = 0.8$.

For comparison, we also calculated the probabilities of selecting the true subset and the CPU times using Adaptive Group Lasso (AGL) proposed by Wang and Leng (2008). The estimator of $\boldsymbol{\Theta}$ by AGL is written as

$$\hat{\boldsymbol{\Theta}}_\tau = \arg\min_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}|\tau),\ f(\boldsymbol{\Theta}|\tau) = \mathrm{tr}\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta})'\} + \tau\sum_{a=1}^{k} w_a||\boldsymbol{\theta}_a||, \tag{15}$$

where $\tau$ is a turning parameter, $w_a$ is the weight for the norm $||\boldsymbol{\theta}_a|| = (\boldsymbol{\theta}_a'\boldsymbol{\theta}_a)^{1/2}$, and $\boldsymbol{\theta}_a$ is the $a$-th column vector of $\boldsymbol{\Theta}'$. Each column vector of $\boldsymbol{Y}$ and $\boldsymbol{X}$ in (15) is centralized and standardized. To optimize (15), we used a coordinate descent algorithm based on Friedman *et al.* (2010). The algorithm is given as follows. Let 100 candidates of $\tau$ be $\tau_t = \exp\{t\log(\tau_{\max}+1)/(100-1)\} - 1$ $(t \in \{0, 1, 2, \ldots, 99\})$, where $\tau_{\max} = \max_{1\leq a\leq k} w_a^{-1}||\boldsymbol{Y}'\boldsymbol{X}_{\{a\}}||$. Initialize $\hat{\boldsymbol{\Theta}}_{\tau_0} = \hat{\boldsymbol{\Theta}}_{\tau_0}^{\mathrm{aft}} = (\hat{\boldsymbol{\theta}}_1^{(0)}, \ldots, \hat{\boldsymbol{\theta}}_k^{(0)})' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. For $t = 1, \ldots, 99$,

1. Update $\hat{\boldsymbol{\Theta}}_{\tau_t}^{\mathrm{bef}} \leftarrow \hat{\boldsymbol{\Theta}}_{\tau_{t-1}}^{\mathrm{aft}}$ and $(\hat{\boldsymbol{\theta}}_1^{(t)}, \ldots, \hat{\boldsymbol{\theta}}_k^{(t)})' \leftarrow \hat{\boldsymbol{\Theta}}_{\tau_{t-1}}^{\mathrm{aft}}$. For each $a \in \{1, \ldots, k\}$,
   (1). Calculate $\boldsymbol{c}_a = \boldsymbol{Y}'\boldsymbol{X}_{\{a\}} - \sum_{i\neq a}^{k}(\boldsymbol{X}'\boldsymbol{X})_{ai}\hat{\boldsymbol{\theta}}_i^{(t)}$.
   (2). If $\tau_t w_a \leq ||\boldsymbol{c}_a||$, then update $\hat{\boldsymbol{\theta}}_a^{(t)} \leftarrow \{(||\boldsymbol{c}_a|| - \tau_t w_a)/((\boldsymbol{X}'\boldsymbol{X})_{aa}||\boldsymbol{c}_a||)\}\boldsymbol{c}_a$, otherwise $\hat{\boldsymbol{\theta}}_a^{(t)} \leftarrow \boldsymbol{0}_p$.

2. Update $\hat{\mathbf{\Theta}}_{\tau_t}^{\mathrm{aft}} \leftarrow (\hat{\boldsymbol{\theta}}_1^{(t)}, \ldots, \hat{\boldsymbol{\theta}}_k^{(t)})'$. If

$$\left| 1 - \frac{f(\hat{\mathbf{\Theta}}_{\tau_t}^{\mathrm{aft}} | \tau_t)}{f(\hat{\mathbf{\Theta}}_{\tau_t}^{\mathrm{bef}} | \tau_t)} \right| < \varepsilon,$$

then define $\hat{\mathbf{\Theta}}_{\tau_t} = \hat{\mathbf{\Theta}}_{\tau_t}^{\mathrm{aft}}$ , otherwise go back to step 1.

In our setting, we used $\varepsilon = 0.01$, and $w_a$ was given by $||\hat{\boldsymbol{\theta}}_a^{\mathrm{LSE}}||^{-1}$, where $\hat{\boldsymbol{\theta}}_a^{\mathrm{LSE}}$ is the least square estimator (LSE) of $\boldsymbol{\theta}_a$, i.e., $(\hat{\boldsymbol{\theta}}_1^{\mathrm{LSE}}, \ldots, \hat{\boldsymbol{\theta}}_k^{\mathrm{LSE}})' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. To choose the best turning parameter, we used three criteria as follows:

$$\begin{aligned} \hat{\tau}^{(i)} &= \arg \min_{\tau_0, \ldots, \tau_{99}} \mathrm{IC}^{(i)}(\tau_t), \\ \mathrm{IC}^{(i)}(\tau_t) &= \frac{1}{p} \mathrm{tr}\{(\boldsymbol{Y} - \boldsymbol{X}\hat{\mathbf{\Theta}}_{\tau_t})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\mathbf{\Theta}}_{\tau_t})\boldsymbol{S}_\omega^{-1}\} + |\mathcal{A}_t|\alpha_i \ (i = 1, 2, 3), \end{aligned}$$

where $|\mathcal{A}_t|$ is the number of non-zero row vectors of $\hat{\mathbf{\Theta}}_{\tau_t}$, and $\alpha_1 = 2$, $\alpha_2 = 2 \log \log n$ and $\alpha_3 = \log n$. We name the AGL using $\mathrm{IC}^{(i)}(\tau_t)$ $(i = 1, 2, 3)$ as $\mathrm{AGL}^{(1)}$, $\mathrm{AGL}^{(2)}$ and $\mathrm{AGL}^{(3)}$, respectively. Table 1 shows the probabilities of selecting the true subset by the ZKB selection methods using the $HCGC_p$, $GC_p^{(i)}$ $(i = 1, 2, 3)$ denoted by $HCGC_p$, $GC_p^{(i)}$ $(i = 1, 2, 3)$ and $\mathrm{AGL}^{(i)}$ $(i = 1, 2, 3)$. From Table 1, we observe that the selection method using the $HCGC_p$ criterion always exhibits high probabilities of selecting the true subset for all combinations of $n$, $p$, $k$ and $k_*$ in Table 1. Although the probabilities by the method using the $GC_p^{(3)}$ criterion also achieve 100%, the performance by the method using the $HCGC_p$ criterion is better than those when the $GC_p^{(3)}$ criterion is used when the sample size is moderate. On the other hand, the probabilities by $\mathrm{AGL}^{(1)}$ are low as the sample size increases in many cases. The probabilities by $\mathrm{AGL}^{(2)}$ reach 100% only when the sample size is large and the dimensions are small. The probabilities by $\mathrm{AGL}^{(3)}$ seem to increase slowly in some cases, but are low when $k_*$ is large. Table 2 shows the CPU times by the ZKB selection method using the $HCGC_p$ criterion denoted by $HCGC_p$ and $\mathrm{AGL}^{(3)}$, and the former is faster than the latter. The difference is particularly clear when the dimensions are large. In sum, the ZKB selection method using the $HCGC_p$ criterion with $\alpha = \tilde{\alpha}$ exhibits the highest probabilities of selecting the true subset and is faster than AGLs.

Table 1: True subset selection probabilities (%)

| $n$ | $p$ | $k$ | $k_*$ | $HCGC_p$ | $GC_p^{(1)}$ | $GC_p^{(2)}$ | $GC_p^{(3)}$ | AGL$^{(1)}$ | AGL$^{(2)}$ | AGL$^{(3)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 200 | 10 | 10 | 5 | 100.0 | 80.2 | 99.6 | 100.0 | 38.9 | 57.9 | 72.8 |
| 500 | 10 | 10 | 5 | 100.0 | 83.8 | 100.0 | 100.0 | 63.9 | 88.7 | 92.7 |
| 1000 | 10 | 10 | 5 | 100.0 | 85.5 | 100.0 | 100.0 | 87.6 | 89.6 | 99.3 |
| 2000 | 10 | 10 | 5 | 100.0 | 85.9 | 100.0 | 100.0 | 87.4 | 99.5 | 99.5 |
| 3000 | 10 | 10 | 5 | 100.0 | 86.6 | 100.0 | 100.0 | 0.0 | 100.0 | 100.0 |
| 200 | 160 | 10 | 5 | 99.9 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.4 |
| 500 | 400 | 10 | 5 | 100.0 | 0.0 | 0.0 | 34.1 | 0.0 | 0.0 | 29.6 |
| 1000 | 800 | 10 | 5 | 100.0 | 0.0 | 0.0 | 95.7 | 0.0 | 0.0 | 66.4 |
| 2000 | 1600 | 10 | 5 | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 | 86.5 |
| 3000 | 2400 | 10 | 5 | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 | 92.6 |
| 200 | 10 | 160 | 5 | 100.0 | 0.1 | 20.1 | 86.3 | 1.6 | 5.6 | 12.4 |
| 500 | 10 | 400 | 5 | 100.0 | 0.0 | 73.3 | 99.9 | 12.1 | 22.6 | 40.4 |
| 1000 | 10 | 800 | 5 | 100.0 | 0.0 | 88.4 | 100.0 | 20.5 | 31.5 | 52.0 |
| 2000 | 10 | 1600 | 5 | 100.0 | 0.0 | 95.0 | 100.0 | 27.5 | 40.8 | 50.1 |
| 3000 | 10 | 2400 | 5 | 100.0 | 0.0 | 95.5 | 100.0 | 10.4 | 14.6 | 52.1 |
| 200 | 10 | 160 | 80 | 99.8 | 0.4 | 35.8 | 93.5 | 0.0 | 0.0 | 0.0 |
| 500 | 10 | 400 | 200 | 100.0 | 0.1 | 82.6 | 100.0 | 0.0 | 0.0 | 10.2 |
| 1000 | 10 | 800 | 400 | 100.0 | 0.0 | 93.9 | 100.0 | 0.0 | 0.0 | 0.0 |
| 2000 | 10 | 1600 | 800 | 100.0 | 0.0 | 96.8 | 100.0 | 0.0 | 0.0 | 0.0 |
| 3000 | 10 | 2400 | 1200 | 100.0 | 0.0 | 98.2 | 100.0 | 0.0 | 0.0 | 0.0 |
| 200 | 80 | 80 | 5 | 100.0 | 0.0 | 0.0 | 34.4 | 0.0 | 0.1 | 5.3 |
| 500 | 200 | 200 | 5 | 100.0 | 0.0 | 0.0 | 99.7 | 0.0 | 5.5 | 21.9 |
| 1000 | 400 | 400 | 5 | 100.0 | 0.0 | 0.3 | 100.0 | 0.0 | 22.2 | 44.3 |
| 2000 | 800 | 800 | 5 | 100.0 | 0.0 | 79.6 | 100.0 | 0.0 | 41.7 | 66.6 |
| 3000 | 1200 | 1200 | 5 | 100.0 | 0.0 | 99.7 | 100.0 | 0.0 | 53.3 | 78.9 |
| 200 | 80 | 80 | 40 | 100.0 | 0.0 | 0.0 | 52.7 | 0.0 | 0.0 | 0.3 |
| 500 | 200 | 200 | 100 | 100.0 | 0.0 | 0.1 | 100.0 | 0.0 | 0.0 | 0.0 |
| 1000 | 400 | 400 | 200 | 100.0 | 0.0 | 3.0 | 100.0 | 0.0 | 0.0 | 2.0 |
| 2000 | 800 | 800 | 400 | 100.0 | 0.0 | 89.3 | 100.0 | 0.0 | 0.0 | 66.5 |
| 3000 | 1200 | 1200 | 600 | 100.0 | 0.0 | 99.8 | 100.0 | 0.0 | 0.0 | 95.0 |

Table 2: CPU times (s)

| $n$ | $p$ | $k$ | $k_*$ | $HCGC_p$ | $\mathrm{AGL}^{(3)}$ |
|---|---|---|---|---|---|
| 200 | 10 | 10 | 5 | 0.0012 | 0.0184 |
| 500 | 10 | 10 | 5 | 0.0028 | 0.0184 |
| 1000 | 10 | 10 | 5 | 0.0094 | 0.0233 |
| 2000 | 10 | 10 | 5 | 0.0272 | 0.0490 |
| 3000 | 10 | 10 | 5 | 0.0635 | 0.0851 |
| 200 | 160 | 10 | 5 | 0.0036 | 0.0985 |
| 500 | 400 | 10 | 5 | 0.0476 | 1.1419 |
| 1000 | 800 | 10 | 5 | 0.3290 | 6.9375 |
| 2000 | 1600 | 10 | 5 | 2.1253 | 40.4359 |
| 3000 | 2400 | 10 | 5 | 6.8453 | 118.6481 |
| 200 | 10 | 160 | 5 | 0.0061 | 0.5672 |
| 500 | 10 | 400 | 5 | 0.0129 | 2.9384 |
| 1000 | 10 | 800 | 5 | 0.0562 | 10.8056 |
| 2000 | 10 | 1600 | 5 | 0.3902 | 44.1574 |
| 3000 | 10 | 2400 | 5 | 1.0536 | 103.2526 |
| 200 | 10 | 160 | 80 | 0.0026 | 0.6110 |
| 500 | 10 | 400 | 200 | 0.0131 | 2.8939 |
| 1000 | 10 | 800 | 400 | 0.0795 | 12.2046 |
| 2000 | 10 | 1600 | 800 | 0.3588 | 44.4453 |
| 3000 | 10 | 2400 | 1200 | 1.1123 | 90.9889 |
| 200 | 80 | 80 | 5 | 0.0114 | 0.3176 |
| 500 | 200 | 200 | 5 | 0.0322 | 3.1167 |
| 1000 | 400 | 400 | 5 | 0.4416 | 44.6930 |
| 2000 | 800 | 800 | 5 | 3.9170 | 560.0503 |
| 3000 | 1200 | 1200 | 5 | 11.8998 | 2256.8923 |
| 200 | 80 | 80 | 40 | 0.0101 | 0.3437 |
| 500 | 200 | 200 | 100 | 0.0290 | 3.3121 |
| 1000 | 400 | 400 | 200 | 0.4313 | 45.2645 |
| 2000 | 800 | 800 | 400 | 3.9815 | 552.0320 |
| 3000 | 1200 | 1200 | 600 | 12.1984 | 2252.4657 |

# Acknowledgments

Ryoya Oda was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. Hirokazu Yanagihara was partially supported by a Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Science, Sports, and Culture #18K03415.

# Appendix

## A Proof of equation (9)

Without loss of generality, let $\boldsymbol{X} = (\boldsymbol{X}_{\omega_{\{\ell\}}}, \boldsymbol{X}_{\{\ell\}})$ for an $\ell \in \omega$. Further, let $\boldsymbol{R}_\ell$, $\boldsymbol{r}_\ell$ and $r_\ell$ be satisfied with

$$\begin{pmatrix} \boldsymbol{R}_\ell & \boldsymbol{r}_\ell \\ \boldsymbol{r}_\ell' & r_\ell \end{pmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Then, using the general formula for the inverse of a block matrix (e.g., Harville, 1997, Theorem 8.5.11), $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{P}_{\omega_{\{\ell\}}}$ can be expressed as follows:

$$\begin{aligned} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' &= \boldsymbol{X}_{\omega_{\{\ell\}}}\boldsymbol{R}_\ell\boldsymbol{X}_{\omega_{\{\ell\}}}' + \boldsymbol{X}_{\omega_{\{\ell\}}}\boldsymbol{r}_\ell\boldsymbol{X}_{\{\ell\}}' + \boldsymbol{X}_{\{\ell\}}\boldsymbol{r}_\ell'\boldsymbol{X}_{\omega_{\{\ell\}}}' + r_\ell\boldsymbol{X}_{\{\ell\}}\boldsymbol{X}_{\{\ell\}}', \\ \boldsymbol{P}_{\omega_{\{\ell\}}} &= \boldsymbol{X}_{\omega_{\{\ell\}}}\boldsymbol{R}_\ell\boldsymbol{X}_{\omega_{\{\ell\}}}' + r_\ell^{-1}\boldsymbol{X}_{\omega_{\{\ell\}}}\boldsymbol{r}_\ell\boldsymbol{r}_\ell'\boldsymbol{X}_{\omega_{\{\ell\}}}'. \end{aligned}$$

From the above equations, we have

$$\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_{\{\ell\}}} = \frac{1}{r_\ell}\boldsymbol{X}\begin{pmatrix} \boldsymbol{r}_\ell \\ r_\ell \end{pmatrix}\begin{pmatrix} \boldsymbol{r}_\ell \\ r_\ell \end{pmatrix}'\boldsymbol{X}'.$$

Note that $r_\ell$ is the $(\ell, \ell)$-th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $\boldsymbol{X}(\boldsymbol{r}_\ell', r_\ell)'$ is the $\ell$-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Therefore, (9) can be derived. □

## B Proof of Lemma 3.1

We can express $P(\hat{j} = j_*)$ as follows:

$$\begin{aligned} &P(\hat{j} = j_*) \\ &= P\left(\left(\bigcap_{\ell \in j_*}\left\{HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega) > 0\right\}\right)\bigcap\left(\bigcap_{\ell \notin j_*}\left\{HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega) \leq 0\right\}\right)\right). \end{aligned}$$

Then, the following lower bound of $P(\hat{j} = j_*)$ can be derived:

$$\begin{aligned} &P(\hat{j} = j_*) \\ &\geq 1 - \sum_{\ell \in j_*} P\left(HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega) < 0\right) - \sum_{\ell \notin j_*} P\left(HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega) > 0\right). \end{aligned}$$

This completes the proof of Lemma 3.1. □

## C Proof of Theorem 3.1

We first describe two lemmas. The first lemma gives another expression of $GC_p(\omega_j) - GC_p(\omega)$ for $j \subset \omega$ $(j \neq \omega)$ (the proof is given in Appendix F):

**Lemma C.1.** *For $j \subset \omega$ ($j \neq \omega$), suppose that $\delta_{j,i}$ ($1 \le i \le m_j$) are constants satisfying* $\mathrm{tr}(\boldsymbol{\Delta}_j) = \sum_{i=1}^{m_j} \delta_{j,i}$ *and* $\delta_{j,i} \ge m_j^{-1}\lambda_{\max}(\boldsymbol{\Delta}_j)$*, where $\boldsymbol{\Delta}_j$ and $m_j$ are defined by* (3) *and* (4)*, and $\lambda_{\max}(\boldsymbol{\Delta}_j)$ is the maximum eigenvalue of $\boldsymbol{\Delta}_j$. Let $u_i$, $u_{j,i}$, and $v_i$ be random variables distributed according to $u_i \sim \chi^2(p)$, $u_{j,i} \sim \chi^2(p; \delta_{j,i})$ and $v_i \sim \chi^2(n-p-k+1)$ ($1 \le i \le m_j$), where $u_i$ and $u_{j,i}$ are independent of $v_i$ for each $i$. Then, under Assumption A1, we have*

$$GC_p(\omega_j) - GC_p(\omega) = \begin{cases} (n-k)\displaystyle\sum_{i=1}^{m_j} \frac{u_i}{v_i} - m_j p\alpha & (j \subset j_*^c) \\ (n-k)\displaystyle\sum_{i=1}^{m_j} \frac{u_{j,i}}{v_i} - m_j p\alpha & (j \subset j_*) \end{cases}. \tag{C.1}$$

The following lemma is needed to evaluate the divergence orders of the moments of $GC_p(\omega_j) - GC_p(\omega)$ (the proof is given in Appendix G).

**Lemma C.2.** *Let $D = \{(n,p,k) \in \mathbb{N}^3 \mid N-4 > 0\}$, where $N = n-p-k+1$. Suppose that $\delta$ is a constant satisfying $\inf_{(n,p,k)\in D} n^{-1}\delta > 0$ and $N - 4r > 0$ for $r \in \mathbb{N}$. Let $u_1$, $u_2$ and $v$ be random variables distributed according to $\chi^2(p)$, $\chi^2(p;\delta)$ and $\chi^2(N)$, where $u_1$ and $u_2$ are independent of $v$. Then, we have*

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = O(p^r n^{-2r}),\ E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = O(\delta^r n^{-2r}),$$

*as $n - p - k \to \infty$.*

Applying the results of Lemma C.1 for $m_j = 1$ to $HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega)$, we have

$$HCGC_p(\omega_{\{\ell\}}) - HCGC_p(\omega) = \begin{cases} (n-k)\dfrac{u}{v} - p\alpha & (\ell \notin j_*) \\ (n-k)\dfrac{u_\ell}{v} - p\alpha & (\ell \in j_*) \end{cases}, \tag{C.2}$$

where $u$ and $u_\ell$ are independent of $v$, and $u \sim \chi^2(p)$, $u_\ell \sim \chi^2(p; \delta_{\{\ell\}})$ and $v \sim \chi^2(N)$. From (C.2), we have

$$\begin{aligned} \sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) &= (k-k_*)P\left(\frac{u}{v} > \frac{p}{n-k}\alpha\right) \\ &= (k-k_*)P\left(\frac{u}{v} - \frac{p}{N-2} > \rho\right) \\ &\le (k-k_*)P\left(\left|\frac{u}{v} - \frac{p}{N-2}\right| \ge \rho\right), \end{aligned} \tag{C.3}$$

$$\begin{aligned} \sum_{\ell \in j_*} P(HCGC_p(\omega_{\{\ell\}}) < HCGC_p(\omega)) &= \sum_{\ell \in j_*} P\left(\frac{u_\ell}{v} < \frac{p}{n-k}\alpha\right) \\ &= \sum_{\ell \in j_*} P\left(\frac{u_\ell}{v} - \frac{p+\delta_{\{\ell\}}}{N-2} - \rho < -\frac{\delta_{\{\ell\}}}{N-2}\right) \\ &\le \sum_{\ell \in j_*} P\left(\left|\frac{u_\ell}{v} - \frac{p+\delta_{\{\ell\}}}{N-2} - \rho\right| \ge \frac{\delta_{\{\ell\}}}{N-2}\right), \end{aligned} \tag{C.4}$$

where $\rho = \{p/(n-k)\}\beta$. Applying Markov's inequality to (C.3) and (C.4), the following upper bounds can be derived:

$$(k-k_*)P\left(\left|\frac{u}{v}-\frac{p}{N-2}\right| \geq \rho\right) \leq (k-k_*)\rho^{-2r_1}E\left[\left(\frac{u}{v}-\frac{p}{N-2}\right)^{2r_1}\right],$$

$$\sum_{\ell \in j_*} P\left(\left|\frac{u_\ell}{v}-\frac{p}{N-2}-\rho\right| \geq \frac{\delta_{\{\ell\}}}{N-2}\right) \leq \sum_{\ell \in j_*}\left(\frac{\delta_{\{\ell\}}}{N-2}\right)^{-2r_2} E\left[\left(\frac{u_\ell}{v}-\frac{p+\delta_{\{\ell\}}}{N-2}-\rho\right)^{2r_2}\right],$$

where $r_1$ and $r_2$ are natural numbers defined by (6). From the above equations and Lemma C.2, the following equations can be derived:

$$\sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) = O(kp^{-r_1}\beta^{-2r_1}),$$
$$\sum_{\ell \in j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) = \sum_{\ell \in j_*} O(\max\{p^{2r_2}\beta^{2r_2}\delta_{\{\ell\}}^{-2r_2}, \delta_{\{\ell\}}^{-r_2}\}).$$

Note that $\#(j_*) \leq k_*$. Hence, if $(\sqrt[2r_2]{k}p/n)\beta \to 0$ then $\sqrt[2r_2]{k_*}p\beta/\delta_{\{\ell\}} = o(1)$ holds, and if $r_2 \in \mathbb{N}\backslash\{1\}$ then $k_*/\delta_{\{\ell\}}^{r_2} \to 0$ holds from Assumption A2. This gives the following equations for $r_2 \in \mathbb{N}\backslash\{1\}$:

$$\sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) = o(1), \quad \sum_{\ell \in j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) = o(1).$$

These equations and Lemma 3.1 complete the proof of Theorem 3.1. □

# D Proof of Lemma 3.2

We can express $P(\hat{j}_{\mathcal{J}} = j_*)$ as follows:

$$P(\hat{j}_{\mathcal{J}} = j_*)$$
$$= P\left(\left(\bigcap_{j \in \mathcal{J}_+}\{HCGC_p(\omega_j) - HCGC_p(\omega) > 0\}\right) \bigcap \left(\bigcap_{j \in \mathcal{J}_-}\{HCGC_p(\omega_j) - HCGC_p(\omega) \leq 0\}\right)\right).$$

Then, the following lower bound of $P(\hat{j}_{\mathcal{J}} = j_*)$ can be derived:

$$P(\hat{j}_{\mathcal{J}} = j_*)$$
$$\geq 1 - \sum_{j \in \mathcal{J}_+} P\left(HCGC_p(\omega_j) - HCGC_p(\omega) < 0\right) - \sum_{j \in \mathcal{J}_-} P\left(HCGC_p(\omega_j) - HCGC_p(\omega) > 0\right).$$

Therefore, Lemma 3.2 can be derived. □

# E Proof of Theorem 3.2

We can apply the results of Lemma C.1 to this proof, i.e., we can express the following distribution forms of $HCGC_p(\omega_j) - HCGC_p(\omega)$:

$$HCGC_p(\omega_j) - HCGC_p(\omega) = \begin{cases} (n-k)\displaystyle\sum_{i=1}^{m_j}\frac{u_i}{v_i} - m_j p\alpha & (j \in \mathcal{J}_-) \\ (n-k)\displaystyle\sum_{i=1}^{m_j}\frac{u_{j,i}}{v_i} - m_j p\alpha & (j \in \mathcal{J}_+) \end{cases}, \tag{E.1}$$

where $u_i$ and $u_{j,i}$ are independent of $v_i$, and

$$u_i \sim \chi^2(p),\ u_{j,i} \sim \chi^2(p;\delta_{j,i}),\ v_i \sim \chi^2(N)\ (1 \le i \le m_j).$$

Here, $\delta_{j,i}$ $(1 \le i \le m_j)$ are constants satisfying $\sum_{i=1}^{m_j} \delta_{j,i} = \mathrm{tr}(\boldsymbol{\Delta}_j)$ and $\delta_{j,i} \ge m_j^{-1}\lambda_{\max}(\boldsymbol{\Delta}_j)$, where $\boldsymbol{\Delta}_j$ is given by (3). When $j \in \mathcal{J}_+$, let $\ell$ be an element of $j$, i.e., $\ell \in j$. Then, since $\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}}$ and $\boldsymbol{P}_{\omega_{\{\ell\}}} - \boldsymbol{P}_{\omega_j}$ are semi-positive definite, the following equation can be derived:

$$\mathrm{tr}(\boldsymbol{\Delta}_j) = \delta_{\{\ell\}} + \mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{P}_{\omega_{\{\ell\}}} - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2}\} \ge \delta_{\{\ell\}}.$$

In addition, let $d_j = \mathrm{rank}(\boldsymbol{\Delta}_j)$ be defined by (4). From (5), we observe that $d_j$ is bounded. Since $d_j\lambda_{\max}(\boldsymbol{\Delta}_j) \ge \mathrm{tr}(\boldsymbol{\Delta}_j)$ holds, the following equation is obtained:

$$\delta_{j,i} \ge m_j^{-1}\lambda_{\max}(\boldsymbol{\Delta}_j) \ge (m_j d_j)^{-1}\mathrm{tr}(\boldsymbol{\Delta}_j) \ge (m_j d_j)^{-1}\delta_{\{\ell\}}. \tag{E.2}$$

Now, we derive the divergence orders of $\sum_{j\in\mathcal{J}_-} P(HCGC_p(\omega_j) > HCGC_p(\omega))$ and $\sum_{j\in\mathcal{J}_+} P(HCGC_p(\omega_j) < HCGC_p(\omega))$. From (E.1), we have

$$\begin{aligned}
\sum_{j\in\mathcal{J}_-} P(HCGC_p(\omega_j) > HCGC_p(\omega)) &= \sum_{j\in\mathcal{J}_-} P\left(\sum_{i=1}^{m_j}\frac{u_i}{v_i} > \frac{m_j p}{n-k}\alpha\right)\\
&\le \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\frac{u_i}{v_i} > \frac{p}{n-k}\alpha\right)\\
&= \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\frac{u_i}{v_i} - \frac{p}{N-2} > \rho\right)\\
&\le \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\left|\frac{u_i}{v_i} - \frac{p}{N-2}\right| \ge \rho\right), \qquad \text{(E.3)}\\
\sum_{j\in\mathcal{J}_+} P(HCGC_p(\omega_j) < HCGC_p(\omega)) &= \sum_{j\in\mathcal{J}_+} P\left(\sum_{i=1}^{m_j}\frac{u_{j,i}}{v_i} < \frac{m_j p}{n-k}\alpha\right)\\
&\le \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\frac{u_{j,i}}{v_i} < \frac{p}{n-k}\alpha\right)\\
&= \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} - \rho < -\frac{\delta_{j,i}}{N-2}\right)\\
&\le \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\left|\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} - \rho\right| \ge \frac{\delta_{j,i}}{N-2}\right), \qquad \text{(E.4)}
\end{aligned}$$

where $\rho = \{p/(n-k)\}\beta$. Then, by applying Markov's inequality to (E.3) and (E.4), their following upper bounds can be derived:

$$\begin{aligned}
&\sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\left|\frac{u_i}{v_i} - \frac{p}{N-2}\right| \ge \rho\right) \le \sum_{j\in\mathcal{J}_-} m_j\rho^{-2r_1} E\left[\left(\frac{u_1}{v_1} - \frac{p}{N-2}\right)^{2r_1}\right],\\
&\sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\left|\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} - \rho\right| \ge \frac{\delta_{j,i}}{N-2}\right) \le \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j}\left(\frac{\delta_{j,i}}{N-2}\right)^{-2r_2} E\left[\left(\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} - \rho\right)^{2r_2}\right].
\end{aligned}$$

Note that $\inf_{(n,p,k)\in D} n^{-1}\delta_{j,i} > 0$ from (E.2). Hence, from the above equations and Lemma C.2, the following equations can be derived:

$$\sum_{j\in\mathcal{J}_-} m_j\rho^{-2r_1} E\left[\left(\frac{u_1}{v_1} - \frac{p}{N-2}\right)^{2r_1}\right] = O(kp^{-r_1}\beta^{-2r_1}),$$

$$\sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j}\left(\frac{\delta_{j,i}}{N-2}\right)^{-2r_2} E\left[\left(\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} - \rho\right)^{2r_2}\right] = \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} O(\max\{p^{2r_2}\beta^{2r_2}\delta_{j,i}^{-2r_2}, \delta_{j,i}^{-r_2}\}).$$

Note that $m_j$ is bounded and $\#(\mathcal{J}_+) \le k_*$, and it follows from (E.2) that $\delta_{j,i}^{-1} \le m_j d_j \delta_{\{\ell\}}^{-1}$. Therefore, from Lemma 3.2, Theorem 3.2 can be shown. □

# F Proof of Lemma C.1

First, we derive results for the case of $j \subset j_*^c$. Let the elements of $j$ be $a_1, \dots, a_{m_j}$ ($a_s \neq a_t$ $(s \neq t)$), i.e., $j = \{a_1, \dots, a_{m_j}\}$. Further, let $j_{-,0} = \omega_j$ and $j_{-,i} = j_{-,i-1} \cup \{a_i\}$ $(1 \le i \le m_j)$. Then, it holds that $j_{-,m_j} = \omega$, and we can express $GC_p(\omega_j) - GC_p(\omega)$ as follows:

$$\begin{aligned} GC_p(\omega_j) - GC_p(\omega) &= \sum_{i=1}^{m_j}\{GC_p(j_{-,i-1}) - GC_p(j_{-,i})\} \\ &= (n-k)\sum_{i=1}^{m_j}\mathrm{tr}[\boldsymbol{Y}'(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})\boldsymbol{Y}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\}^{-1}] - m_j p\alpha. \quad \text{(F.1)} \end{aligned}$$

Let $\boldsymbol{W}_{j,i} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$ and $\boldsymbol{W} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$. Note that $\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}}$ and $\boldsymbol{I}_n - \boldsymbol{P}_\omega$ are symmetric idempotent matrices, and it holds that $(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})(\boldsymbol{I}_n - \boldsymbol{P}_\omega) = \boldsymbol{O}_{n,n}$ and $(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})\boldsymbol{X}_* = (\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{X}_* = \boldsymbol{O}_{n,k_*}$. Then, from a property of the Wishart distribution and Cochran's Theorem (e.g. Fujikoshi *et al.*, 2010, chap 2), we can state that $\boldsymbol{W}_{j,i}$ and $\boldsymbol{W}$ are independent, and $\boldsymbol{W}_{j,i} \sim W_p(1, \boldsymbol{I}_p)$ and $\boldsymbol{W} \sim W_p(n-k, \boldsymbol{I}_p)$. Thus, (F.1) is expressed as

$$GC_p(\omega_j) - GC_p(\omega) = (n-k)\sum_{i=1}^{m_j}\mathrm{tr}(\boldsymbol{W}_{j,i}\boldsymbol{W}^{-1}) - m_j p\alpha. \quad \text{(F.2)}$$

From a property of the Wishart distribution, $\boldsymbol{W}_{j,i}$ can be expressed as $\boldsymbol{W}_{j,i} = \boldsymbol{z}_i\boldsymbol{z}_i'$, where $\boldsymbol{z}_i$ is independent of $\boldsymbol{W}$, and $\boldsymbol{z}_i \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. Then, we express $\boldsymbol{z}_i'\boldsymbol{W}^{-1}\boldsymbol{z}_i$ as

$$\boldsymbol{z}_i'\boldsymbol{W}^{-1}\boldsymbol{z}_i = \frac{\boldsymbol{z}_i'\boldsymbol{z}_i}{\{(\boldsymbol{z}_i'\boldsymbol{z}_i)^{-1/2}\boldsymbol{z}_i'\boldsymbol{W}^{-1}\boldsymbol{z}_i(\boldsymbol{z}_i'\boldsymbol{z}_i)^{-1/2}\}^{-1}}.$$

Let $u_i = \boldsymbol{z}_i'\boldsymbol{z}_i$ and $v_i = \{(\boldsymbol{z}_i'\boldsymbol{z}_i)^{-1/2}\boldsymbol{z}_i'\boldsymbol{W}^{-1}\boldsymbol{z}_i(\boldsymbol{z}_i'\boldsymbol{z}_i)^{-1/2}\}^{-1}$. Then, from a property of the Wishart distribution, we can state that $u_i$ and $v_i$ are independent, and $u_i \sim \chi^2(p)$ and $v_i \sim \chi^2(n-p-k+1)$. Therefore, $\mathrm{tr}(\boldsymbol{W}_{j,i}\boldsymbol{W}^{-1})$ is expressed as

$$\mathrm{tr}(\boldsymbol{W}_{j,i}\boldsymbol{W}^{-1}) = \frac{u_i}{v_i}.$$

From the above equation and (F.2), we can derive (C.1) for the case of $j \subset j_*^c$.

Next, we derive results for the case of $j \subset j_*$. Then, $GC_p(\omega_j) - GC_p(\omega)$ is expressed as

$$GC_p(\omega_j) - GC_p(\omega) = (n-k)\mathrm{tr}[\boldsymbol{Y}'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{Y}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\}^{-1}] - m_j p\alpha. \quad \text{(F.3)}$$

Let $\boldsymbol{W}_j = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$. Note that $\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j}$ is symmetric and idempotent, and it holds that $(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})(\boldsymbol{I}_n - \boldsymbol{P}_\omega) = \boldsymbol{O}_{n,n}$. Then, from a property of the non-central Wishart distribution and Cochran's Theorem, we can state that $\boldsymbol{W}_j$ and $\boldsymbol{W}$ are independent, and $\boldsymbol{W}_j \sim W_p(m_j, \boldsymbol{I}_p; \boldsymbol{\Delta}_j)$ and $\boldsymbol{W} \sim W_p(n-k, \boldsymbol{I}_p)$. Thus, (F.3) is expressed as

$$GC_p(\omega_j) - GC_p(\omega) = (n-k)\mathrm{tr}(\boldsymbol{W}_j\boldsymbol{W}^{-1}) - m_j p\alpha. \tag{F.4}$$

Let the spectral decomposition of $\boldsymbol{\Delta}_j$ be $\boldsymbol{\Delta}_j = \boldsymbol{Q}_j\boldsymbol{\Lambda}_j\boldsymbol{Q}_j'$, where $\boldsymbol{Q}_j$ is the $p\times p$ orthogonal matrix and $\boldsymbol{\Lambda}_j$ is the $p \times p$ diagonal matrix whose $a$-th diagonal element is an eigenvalue $\lambda_{j,a}$, i.e., $\boldsymbol{\Lambda}_j = \mathrm{diag}(\lambda_{j,1},\dots,\lambda_{j,p})$ $(\lambda_{j,1} \geq \cdots \geq \lambda_{j,p})$. Let $\boldsymbol{B}_{j,1} = \boldsymbol{Q}_j'\boldsymbol{W}_j\boldsymbol{Q}_j$ and $\boldsymbol{B}_{j,2} = \boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j$. Then, from a property of the non-central Wishart distribution, we can state that $\boldsymbol{B}_{j,1}$ and $\boldsymbol{B}_{j,2}$ are independent and $\boldsymbol{B}_{j,1} \sim W_p(m_j, \boldsymbol{I}_p; \boldsymbol{\Lambda}_j)$ and $\boldsymbol{B}_{j,2} \sim W_p(n-k, \boldsymbol{I}_p)$. Let $d_j = \mathrm{rank}(\boldsymbol{\Delta}_j)$ be defined in (4). It is obvious that $\lambda_{j,d_j+1} = \cdots = \lambda_{j,p} = 0$. Since it holds that $d_j \leq m_j$ from (5), let $\boldsymbol{\Gamma}_j$ be as follows:

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} \boldsymbol{\Lambda}_{j,0}^{1/2} & \boldsymbol{O}_{d_j,p-d_j} \\ \boldsymbol{O}_{m_j-d_j,d_j} & \boldsymbol{O}_{m_j-d_j,p-d_j} \end{pmatrix}, \ \boldsymbol{\Lambda}_{j,0} = \mathrm{diag}(\lambda_{j,1},\dots,\lambda_{j,d_j}).$$

By using $\boldsymbol{\Gamma}_j$, we can express $\boldsymbol{B}_{j,1}$ as $\boldsymbol{B}_{j,1} = (\boldsymbol{\mathcal{E}}_j + \boldsymbol{\Gamma}_j)'(\boldsymbol{\mathcal{E}}_j + \boldsymbol{\Gamma}_j)$, where $\boldsymbol{\mathcal{E}}_j \sim N_{m_j\times p}(\boldsymbol{O}_{m_j,p}, \boldsymbol{I}_p \otimes \boldsymbol{I}_{m_j})$ and $\boldsymbol{\mathcal{E}}_j$ is independent of $\boldsymbol{B}_{j,2}$. Let $\boldsymbol{H} = (\boldsymbol{h}_1,\dots,\boldsymbol{h}_{m_j})$ be a $m_j \times m_j$ orthogonal matrix satisfying $\boldsymbol{h}_1 = m_j^{-1/2}\boldsymbol{1}_{m_j}$, and let $(\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_{m_j})' = \boldsymbol{H}\boldsymbol{\Gamma}_j$. Then, we have

$$(\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_{m_j})' = \boldsymbol{H}\begin{pmatrix} \boldsymbol{\Lambda}_{j,0}^{1/2} & \boldsymbol{O}_{d_j,p-d_j} \\ \boldsymbol{O}_{m_j-d_j,d_j} & \boldsymbol{O}_{m_j-d_j,p-d_j} \end{pmatrix} = (\sqrt{\lambda_{j,1}}\boldsymbol{h}_1,\dots,\sqrt{\lambda_{j,d_j}}\boldsymbol{h}_{d_j}, \boldsymbol{O}_{m_j,p-d_j}).$$

Now, we put $\delta_{j,i} = ||\boldsymbol{\eta}_i||^2$ $(1 \leq i \leq m_j)$. Then, from the above equation, it is straightforward that $\delta_{j,i} \geq m_j^{-1}\lambda_{j,1}$ $(1 \leq i \leq m_j)$ and $\mathrm{tr}(\boldsymbol{\Delta}_j) = \sum_{i=1}^{m_j}\delta_{j,i}$. Let $(\boldsymbol{z}_{j,1},\dots,\boldsymbol{z}_{j,m_j})' = \boldsymbol{H}\boldsymbol{\mathcal{E}}_j$. Since $\boldsymbol{z}_{j,1},\dots,\boldsymbol{z}_{j,m_j} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$, $\boldsymbol{B}_{j,1}$ can be expressed as

$$\boldsymbol{B}_{j,1} = (\boldsymbol{\mathcal{E}}_j + \boldsymbol{\Gamma}_j)'\boldsymbol{H}'\boldsymbol{H}(\boldsymbol{\mathcal{E}}_j + \boldsymbol{\Gamma}_j) = (\boldsymbol{H}\boldsymbol{\mathcal{E}}_j + \boldsymbol{H}\boldsymbol{\Gamma}_j)'(\boldsymbol{H}\boldsymbol{\mathcal{E}}_j + \boldsymbol{H}\boldsymbol{\Gamma}_j) = \sum_{i=1}^{m_j}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'.$$

Then, we can express $(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'\boldsymbol{B}_{j,2}^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)$ as

$$(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'\boldsymbol{B}_{j,2}^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i) = \frac{||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^2}{\{||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'\boldsymbol{B}_{j,2}^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^{-1}\}^{-1}}.$$

Let $u_{j,i} = ||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^2$ and $v_i = \{||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'\boldsymbol{B}_{j,2}^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)||\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i||^{-1}\}^{-1}$. Then, from a property of the Wishart distribution, we can state that $u_{j,i}$ and $v_i$ are independent, and $u_{j,i} \sim \chi^2(p; \delta_{j,i})$ and $v_i \sim \chi^2(n-p-k+1)$. Therefore, $\mathrm{tr}(\boldsymbol{W}_j\boldsymbol{W}^{-1})$ is expressed as

$$\begin{aligned}\mathrm{tr}(\boldsymbol{W}_j\boldsymbol{W}^{-1}) &= \mathrm{tr}(\boldsymbol{Q}_j'\boldsymbol{W}_j\boldsymbol{Q}_j\boldsymbol{Q}_j'\boldsymbol{W}^{-1}\boldsymbol{Q}_j) = \mathrm{tr}(\boldsymbol{B}_{j,1}\boldsymbol{B}_{j,2}^{-1}) = \sum_{i=1}^{m_j}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i)'\boldsymbol{B}_{j,2}^{-1}(\boldsymbol{z}_{j,i} + \boldsymbol{\eta}_i) \\ &= \sum_{i=1}^{m_j}\frac{u_{j,i}}{v_i}.\end{aligned}$$

From the above equation and (F.4), we can derive (C.1) for the case of $j \subset j_*$. $\square$

## G Proof of Lemma C.2

We first describe a lemma concerning the central moments of chi-square and non-central chi-square random variables; this is required for proving Lemma C.2 (the proof is given in Appendix H).

**Lemma G.1.** *Let $X_1 \sim \chi^2(t)$ and $X_2 \sim \chi^2(t;\psi)$, where $\psi$ is a constant satisfying $t/\psi = O(1)$. Then, we have*

$$E[(X_1 - t)^h] = \begin{cases} 1 & (h=0) \\ 0 & (h=1) \\ O(t^{\lfloor h/2 \rfloor}) & (h \geq 2) \end{cases}, \tag{G.1}$$

$$E[\{X_2 - (t+\psi)\}^h] = \begin{cases} 1 & (h=0) \\ 0 & (h=1) \\ O(t^{\lfloor \psi/2 \rfloor}) & (h \geq 2) \end{cases}. \tag{G.2}$$

*Moreover, when $t - 2h > 0$, we have*

$$E\left[\left(\frac{1}{X_1} - \frac{1}{t-2}\right)^h\right] = \begin{cases} 1 & (h=0) \\ 0 & (h=1) \\ O(t^{-2h+\lfloor h/2 \rfloor}) & (h \geq 2) \end{cases}, \tag{G.3}$$

*where $\lfloor h \rfloor$ is the floor function defined by $\lfloor h \rfloor = \max\{m \in \mathbb{Z} \mid m \leq h\}$.*

Let $\xi = 1/(N-2)$ and $\xi_\delta = p + \delta$. Then, we have

$$\frac{u_1}{v} - \frac{p}{N-2} = (u_1 - p)(v^{-1} - \xi) + p(v^{-1} - \xi) + \xi(u_1 - p),$$
$$\frac{u_2}{v} - \frac{p+\delta}{N-2} = (u_2 - \xi_\delta)(v^{-1} - \xi) + \xi_\delta(v^{-1} - \xi) + \xi(u_2 - \xi_\delta).$$

Hence, from the multinomial theorem, we have

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = \sum_{\substack{a+b+c=2r \\ 0 \leq a,b,c \leq 2r}} \frac{(2r)!}{a!b!c!} p^b \xi^c E[(u_1 - p)^{a+c}] E[(v^{-1} - \xi)^{a+b}], \tag{G.4}$$

$$E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = \sum_{\substack{a+b+c=2r \\ 0 \leq a,b,c \leq 2r}} \frac{(2r)!}{a!b!c!} \xi_\delta^b \xi^c E[(u_2 - \xi_\delta)^{a+c}] E[(v^{-1} - \xi)^{a+b}]. \tag{G.5}$$

From the assumption $\inf_{(n,p,k)\in D} n^{-1}\delta > 0$, it follows that $p/\delta = O(1)$. Therefore, from (G.1), (G.2), and (G.3), the divergence orders in (G.4) and (G.5) are maximized when $a = b = 0, c = 2r$. Therefore, we can derive the divergence orders as follows:

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = O(p^r n^{-2r}),\ E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = O(\delta^r n^{-2r}).$$

□

## H Proof of Lemma G.1

We elaborate only on the case of $h \geq 2$ because it is straightforward when $h = 0, 1$. First, we derive (G.1) and (G.2). Let $h_1, \ldots, h_d$ be natural numbers satisfying $\sum_{i=1}^{d} h_i = h$ and $2 \leq h_1, \ldots, h_d$. From Stuart and Ord (1994), we can state that $h$-th central moments can be expressed as the linear combination of the products of $h_1, \ldots, h_d$-th cumulants. From Lancaster (1982) and Tiku (1985), $h$-th cumulants of $X_1 - t$ and $X_2 - (t + \psi)$ can, respectively, be expressed as follows:

$$\kappa_{h,1} = 2^{h-1}(h-1)!t, \ \kappa_{h,2} = 2^{h-1}(h-1)!(t + h\psi).$$

Then, it follows from $t/\psi = O(1)$ that $\kappa_{h,2} = O(\psi)$. Therefore, we observe that the maximum order term of each $h$-th central moment is $\kappa_{2,i}^{h/2}$ if $h$ is even and $\kappa_{2,i}^{(h-1)/2-1}\kappa_{3,i}$ if $h$ is odd $(i = 1, 2)$. This completes (G.1) and (G.2).

Next, we derive (G.3). From the multinomial theorem, we have

$$\begin{aligned}
&E\left[\left(\frac{1}{X_1} - \frac{1}{t-2}\right)^h\right] \\
&= \sum_{i=0}^{h} \frac{h!}{i!(h-i)!}\left(-\frac{1}{t-2}\right)^{h-i} E\left[\left(\frac{1}{X_1}\right)^i\right] \\
&= \left(-\frac{1}{t-2}\right)^h + \sum_{i=1}^{h} \frac{h!}{i!(h-i)!}\left(-\frac{1}{t-2}\right)^{h-i} \prod_{d=1}^{i} \frac{1}{t-2d} \\
&= \left(-\frac{1}{t-2}\right)^h \prod_{d=1}^{h} \frac{1}{t-2d}\left[\{-(t-2)\}^h + \sum_{i=0}^{h-1} \frac{h!}{i!(h-i)!}\{-(t-2)\}^i \prod_{d=1}^{h-i}\{t-2h+2(d-1)\}\right].
\end{aligned}$$

Let $T \sim \chi^2(t-2h)$, then it is known that

$$E[T^{h-i}] = \begin{cases} 1 & (i = h) \\ \prod_{d=1}^{h-i}\{t-2h+2(d-1)\} & (i \leq h-1) \end{cases}.$$

Hence, by letting $s = \{-(t-2)\}^{-h}\prod_{d=1}^{h}(t-2d)^{-1}$, we have

$$\begin{aligned}
E\left[\left(\frac{1}{X_1} - \frac{1}{t-2}\right)^h\right] &= \left(-\frac{1}{t-2}\right)^h \prod_{d=1}^{h} \frac{1}{t-2d}\left\{\sum_{i=0}^{h} \frac{h!}{i!(h-i)!}\{-(t-2)\}^i E[T^{h-i}]\right\} \\
&= sE[\{T-(t-2)\}^h] \\
&= s\sum_{i=0}^{h} \frac{h!}{i!(h-i)!}\{-2(h-1)\}^i E[\{T-(t-2h)\}^{h-i}]. \qquad \text{(H.1)}
\end{aligned}$$

Note that $s = O(t^{-2h})$ and it follows from (G.1) that

$$E[\{T-(t-2h)\}^{h-i}] = \begin{cases} 1 & (i = h) \\ 0 & (i = h-1) \\ O(t^{\lfloor (h-i)/2 \rfloor}) & (i \leq h-2) \end{cases}. \qquad \text{(H.2)}$$

The equations (H.1) and (H.2) complete (G.3). □

## References


[1] Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.

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

[3] Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**(1), 1–22.

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

[5] Lancaster, H. O. (1982). Chi-square distribution. In *Encyclopedia of Statistical Sciences, Vol. 1* (eds. S. Kotz & N. L. Johson), 439–442, John Wiley & Sons, New York.

[6] Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics*, **15**, 661–675.

[7] Mallows, C. L. (1995). More comments on $C_p$. *Technometrics*, **37**, 362–372.

[8] Nagai, I., Yanagihara, H. & Satoh, K. (2012). Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods. *Hiroshima Math. J.*, **42**, 301–324.

[9] Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, **27**, 392–403.

[10] 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.

[11] Rao, C. R. & Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.

[12] Sparks, R. S., Coutsourides, D. & Troskie, L. (1983). The multivariate $C_p$. *Comm. Statist. A-Theory Methods*, **12**, 1775–1793.

[13] Stuart, A. & Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics. Vol. 1. Distribution Theory* (6th ed.). Edward Arnold, London; distributed in the United States of America by Oxford University press, New York.

[14] Tiku, M. (1985). Noncentral chi-square distribution. In *Encyclopedia of Statistical Sciences, Vol. 6* (eds. S. Kotz & N. L. Johson), 276–280, John Wiley & Sons, New York.

[15] Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, **68**, 49–67.

[16] Wang, H. & Leng, C. (2008). A note on adaptive group lasso. *Comput. Stat. Data An.*, **52**, 5277–5286.

[17] 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.

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

[19] 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.