Electronic Journal of Statistics
Vol. 14 (2020) 1386–1412
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1701

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 739-8526, Japan
e-mail: oda.stat@gmail.com; yanagi-hiro@hiroshima-u.ac.jp

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 a generalized C_p criterion. The consistency of the method is provided by a high-dimensional asymptotic framework such that the sample size and the sum of the dimensions of response vectors and explanatory vectors divided by the sample size tend to infinity and some positive constant which are less than one, respectively. 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.

MSC 2010 subject classifications: Primary 62J05; secondary 62E20.
Keywords and phrases: Consistency, high-dimensional asymptotic framework, multivariate linear regression, variable selection.

Received March 2019.

Contents

1 Introduction 1387
2 Preliminaries 1390
3 Main results 1392
3.1 Proposed selection method 1392
3.2 Consistency of proposed selection method 1394
3.3 Extension of the ZKB selection method 1394
4 Numerical studies 1396
A Appendix A 1399
A.1 Proof of equations (2.4) and (2.5) 1399
A.2 Proof of equation (3.4) 1401
A.3 Proof of Lemma 3.1 1401

*Corresponding author.

A.4 Proof of Theorem 3.1 1402
A.5 Proof of Lemma 3.2 1404
A.6 Proof of Theorem 3.2 1404
A.7 Proof of Lemma A.1 1406
A.8 Proof of Lemma A.2 1408
A.9 Proof of Lemma A.3 1409
A.10 R file related to this article 1410
Acknowledgments 1410
References 1411

1. Introduction

Multivariate linear regression is a widely known method of inferential analysis. It features in many theoretical and applied textbooks (see, e.g., [21, chap 9], [24, 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 we assume that $\mathrm{rank}(\boldsymbol{X}) = k < n$ and (n, p, k) satisfies $N - 4 > 0$ 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 the full set $\omega = \{1, \ldots, k\}$ 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 fourth column vectors of $\boldsymbol{X}$. Using the notation j, the candidate model with k_j explanatory variables is expressed as follows:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_j \boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1.1}$$

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_ω and the explanatory matrix $\boldsymbol{X}_\omega$ in the full model ω 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. Without loss of generality, we sort column vectors of $\boldsymbol{X}$ as $\boldsymbol{X} = (\boldsymbol{X}_{j_*}, \boldsymbol{X}_{j_*^c})$, where set A^c denotes the compliment of set A. 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. The C_p criterion was proposed by [13, 14]. 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 [1], and the counterpart for a multivariate linear regression with multiple responses was proposed by [15]. The GC_p criterion can express a wide variety of variable selection criteria, e.g., the C_p criterion for multivariate contexts proposed by [20], and the modified C_p (MC_p) criterion proposed by [3].

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. A practicable selection method was proposed by [17, 31] 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 ω. 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 ω is greater than the criterion for the full set ω, 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 ω 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, where "ZKB" is formed from the initial letters of the authors in [31].

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_* because in general the probability of selecting the true subset is approximated by the asymptotic probability. To this end, let LS, LR, LE and LTE be the large-sample (LS), large-response vector (LR), large-explanatory vector (LE) and large-true explanatory vector

(LTE) asymptotic frameworks such that only n, p, k and k_* tend to infinity, respectively. Further, they are denoted by LS: $n \to \infty$, LR: $p \to \infty$, LE: $k \to \infty$ and LTE: $k_* \to \infty$. LS was used by [16, 17, 19, 31] under the ZKB selection method. However, it is not appropriate to use LS for high-dimensional data because approximate accuracy using LS deteriorates as p or k become large. Hence, criteria used by [16, 17, 19, 31] may not have consistency under the ZKB selection method when p or k tend to infinity. In the context for the consistency of variable selection criteria under the full search method, [4, 27, 28] used the following asymptotic frameworks as $(p+k)/n \to c \in [0,1)$:

[4]: LS and LR,
[27]: LS or (LS and LR),
[28]: (LS and LR) or (LS and LR and LE) or (LS and LR and LTE) as $k/n \to 0$.

Since as described above [4, 27, 28] used asymptotic frameworks such that not only n but also p, k or k_* tend to ∞, the probabilities of selecting the true subset will be high for high-dimensional data suited to the used asymptotic frameworks. However, the probabilities may become low for high-dimensional data not suited to the used asymptotic frameworks. Moreover, it is hard for us to judge whether p, k and k_* are large or not, and so we do not know which asymptotic framework is suitable to given data. Hence, to ensure the consistency, it is more desirable to use an asymptotic framework regardless of sizes of p, k and k_*.

In this paper, we consider the consistency of the GC_p criterion under the ZKB selection method and propose the new consistent ZKB selection method even in high-dimensional contexts. Moreover, we also propose the selection method which can perform group selections. To achieve this, we use the following high-dimensional (HD) asymptotic framework:

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

Importantly, the HD asymptotic framework can be rewritten as

HD: LS or (LS and LR) or (LS and LE) or (LS and LTE) or (LS and LR and LE) or (LS and LR and LTE) as $(p+k)/n \to c \in [0,1)$.

This means that n always tends to infinity, but p, k and k_* may tend to infinity as $(p+k)/n \to c \in [0,1)$. Hence, 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_*. Moreover, even when k is large under $N-4>0$, our proposed method will be very fast although the full search methods used in like [4, 27, 28] cannot be calculable. In resent years, regularization methods are often used for estimating the regression coefficients. The lasso is famous as one of methods estimating the regression coefficients and selecting explanatory variables simultaneously. In multivariate linear regression, it is possible to select explanatory variables by the group lasso proposed by [29], and several papers (e.g., [11, 18, 26, 30]) proposed regularization methods by extending the group lasso for multivariate linear regression case. Moreover, a generalized adaptive elastic-net was proposed and the consistency properties of the method were

obtained by [26]. The consistency properties were provided by using asymptotic frameworks such that $\log k/\log n \to \nu \in [0,1)$ or $\log k = o(n^{1-2\kappa})$ for some $\kappa \in (0,1/2)$. However, the properties are not ensured as p tends to infinity. Our method is consistent even when p tends to infinity as long as $N \to \infty$. Further, our method is faster than an adaptive group lasso even when p or k are large.

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 (1.1), 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 (1.1) is defined by

$$GC_p(j) = (n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1}) + \alpha p k_j, \tag{2.1}$$

where α is a positive constant. The first and second terms in (2.1) express the residual sum of squares with the weighted matrix $\boldsymbol{S}_\omega^{-1}$ and α times the strength of the penalty for the number of elements of $\boldsymbol{\Theta}_j$ in model (1.1), 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

$$\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2},\ \delta_j = \mathrm{tr}(\boldsymbol{\Delta}_j). \tag{2.2}$$

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 three assumptions are prepared:

Assumption A1. The true subset j_* is included in the full set ω, i.e., $j_* \subset \omega$.

Assumption A2. There exists $c_1 > 0$ such that

$$\min_{\ell \in j_*} n^{-1}\boldsymbol{x}_{\{\ell\}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}})\boldsymbol{x}_{\{\ell\}} \geq c_1,$$

where $\boldsymbol{x}_{\{\ell\}}$ is the ℓ-th column vectors of $\boldsymbol{X}$.

Assumption A3. There exist $1/2 < c_A \leq 1$ and $c_2 > 0$ such that

$$n^{1-c_A} \min_{\ell \in j_*} \boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_{\{\ell\}} \geq c_2, \tag{2.3}$$

where $\boldsymbol{\theta}_{\{\ell\}}$ is the ℓ-th column vectors of $\boldsymbol{\Theta}'_*$.

Assumption A1 is needed to consider consistency because the probability of selecting the true subset becomes 0 if it does not hold. Assumption A2 means that the minimum value among the sample variances of residuals resulting from the linear regression of $\boldsymbol{x}_{\{\ell\}}$ with the remaining $\boldsymbol{X}_{\omega_{\{\ell\}}}$ for $\ell \in j_*$ is always positive and does not converge to 0. We often see an assumption for explanatory variables such that the inequality $n^{-1}\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) \geq c_1$, where $\lambda_{\min}(\boldsymbol{A})$ is the minimum eigenvalue of a square matrix $\boldsymbol{A}$. Assumption A2 is weaker than this assumption because the inequality $\min_{\ell \in j_*} \boldsymbol{x}'_{\{\ell\}}(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}})\boldsymbol{x}_{\{\ell\}} \geq \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})$ holds. Assumption A3 is a weak assumption for the true regression coefficients and the true covariance matrix. If $c_A < 1$, Assumption A3 allows $\min_{\ell \in j_*} \boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_{\{\ell\}}$ to converge to 0. Moreover, for all $\ell = 1, \ldots, k_*$, the following inequality holds (the proof is given in Appendix A.1):

$$\boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_{\{\ell\}} \geq \max_{a=1,\ldots,p} \frac{\theta_{*\ell a}^2}{\sigma_{*a}^2}, \tag{2.4}$$

where $\theta_{*\ell a}$ is the (ℓ, a)-th element of $\boldsymbol{\Theta}_*$ and σ_{*a}^2 is the a-th diagonal element of $\boldsymbol{\Sigma}_*$. From (2.4), Assumption A3 can be rewritten as the assumption which does not rely on the correlations of response variables by replacing $\boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_{\{\ell\}}$ with $\max_{a=1,\ldots,p} \theta_{*\ell a}^2 / \sigma_{*a}^2$ in (2.3). If Assumptions A1–A3 are supported, the following inequality holds (the proof is given in Appendix A.1):

$$n^{-c_A} \delta_{\min} \geq c_1 c_2, \tag{2.5}$$

where $\delta_{\min} = \min_{\ell \in j_*} \delta_{\{\ell\}}$. The above equation restricts the divergence order of the non-centrality parameter $\delta_{\{\ell\}}$. If k is fixed and $c_A = 1$, (2.5) is as per what was put forward in [27].

Finally, we identify the upper bound of the rank of the non-centrality 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:

$$m_j = \#(j), \ d_j = \operatorname{rank}(\boldsymbol{\Delta}_j). \tag{2.6}$$

In accordance with [28], it follows from Assumption A1 that the rank of $\boldsymbol{X}'_*(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*$ is calculated as

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

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

$$\begin{aligned} d_j &\leq \min\{\operatorname{rank}(\boldsymbol{X}'_*(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{X}_*), \operatorname{rank}(\boldsymbol{\Theta}_* \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\Theta}'_*)\} \\ &\leq \begin{cases} 0 & (j \subset j_*^c) \\ \min\{m_j, p\} & (j \subset j_*) \end{cases}. \end{aligned} \tag{2.7}$$

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 (2.1) satisfying*

$$\alpha = \frac{n-k}{N-2} + \beta,\ \beta > 0\ s.t.\ \frac{\sqrt{p}}{k^{1/2r}}\beta \to \infty,\ \frac{p}{n^{c_A}}\beta \to 0, \tag{3.1}$$

as $n \to \infty$, $(p+k)/n \to c \in [0,1)$, for some $r \in \mathbb{N}$, where c_A is defined in Assumption A3.

We now introduce the ZKB selection method using a variable selection criterion (SC). 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 ω will be greater than that for ω 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*

$$\hat{j} = \{\ell \in \omega \mid HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)\}. \tag{3.2}$$

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

$$\tilde{\alpha} = \frac{n-k}{N-2} + \tilde{\beta},\ \tilde{\beta} = \frac{(n-k)\sqrt{N+p-4}}{(N-2)\sqrt{N-4}} \cdot \frac{k^{1/4}\log n}{\sqrt{p}}. \tag{3.3}$$

This $\tilde{\alpha}$ is based on [27]. It is straightforward to observe that $\tilde{\beta}$ is satisfied with $(\sqrt{p}/k^{1/2r})\tilde{\beta} \to \infty$ and $(p/n^{c_A})\tilde{\beta} \to 0$ as $n \to \infty$, $(p+k)/n \to c \in [0,1)$ for $r \geq 3$ and $3/4 < c_A \leq 1$. Therefore, the GC_p criterion with $\alpha = \tilde{\alpha}$ is included in the class of the $HCGC_p$ criterion with $3/4 < c_A \leq 1$. 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 ℓ when k is large. To overcome this problem, we offer an efficient calculation of $GC_p(\omega_{\{\ell\}}) - GC_p(\omega)$. Let r_ℓ and $\boldsymbol{z}_\ell$ be the (ℓ, ℓ)-th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and the ℓ-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, respectively. Then, using r_ℓ and $\boldsymbol{z}_\ell$, we can express $\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_{\{\ell\}}}$ as follows (the proof of (3.4) is given in Appendix A.2):

$$\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_{\{\ell\}}} = \frac{1}{r_\ell}\boldsymbol{z}_\ell\boldsymbol{z}'_\ell. \tag{3.4}$$

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

$$GC_p(\omega_{\{\ell\}}) - GC_p(\omega) = \frac{1}{r_\ell}\boldsymbol{z}'_\ell\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell - p\alpha. \tag{3.5}$$

Note that (3.5) 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 ℓ-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:

$$\boldsymbol{z}'_\ell\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell = \boldsymbol{t}'_\ell\boldsymbol{t}_\ell. \tag{3.6}$$

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 (3.5) (and also use (3.6) when p is small) to perform the ZKB selection method using the GC_p criterion.

Note that the proposed selection method is calculable when $N - 4 > 0$. When $k > n$, we can formally combine the proposed selection method and screening methods by [8, 10, 12], which can apply to screening explanatory variables for a multivariate linear regression with multiple responses. However, we should pay attention to use their methods because the screening properties are ensured when p or k_* are fixed although the consistency of the proposed selection method is ensured even when p and k_* may diverge. On the other hand, a multivariate linear regression can be regarded as a perfunctory linear regression on a single response from the explanatory matrix $(\boldsymbol{I}_p \otimes \boldsymbol{X})$. However, notice that in generally we cannot directly apply several screening methods (e.g., [2]) for a linear regression with a single response to our variable selection problem. This is because our selection problem can be regarded as a group selection problem for explanatory variables corresponding to the p-dimensional regression coefficient vectors.

3.2. Consistency of proposed selection method

We ensure the consistency of the ZKB selection method using the $HCGC_p$ criterion (3.2). To do so, we present a lemma about sufficient conditions for consistency (the proof is given in Appendix A.3). 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*:

$$\sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega)) \to 0, \tag{3.7}$$

$$\sum_{\ell \in j_*} P(HCGC_p(\omega_{\{\ell\}}) < HCGC_p(\omega)) \to 0. \tag{3.8}$$

Then, the ZKB selection method using the $HCGC_p$ criterion (3.2) is consistent, that is $P(\hat{j} = j_*) \to 1$ holds.

By showing that the sufficient conditions (3.7) and (3.8) in Lemma 3.1 hold, the consistency of the ZKB selection method using the $HCGC_p$ criterion (3.2) can be obtained as follows (the proof is given in Appendix A.4):

Theorem 3.1. *Suppose that Assumptions A1–A3 hold. Then, the ZKB selection method using the $HCGC_p$ criterion (3.2) 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 (3.3) is also consistent under Assumptions A1, A2 and Assumption A3 with $3/4 < c_A \leq 1$.

3.3. Extension of the ZKB selection method

In the previous subsections, we proposed the ZKB selection method using the $HCGC_p$ criterion (3.2). However, when the full model ω 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 (2.6). 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 (3.2) when $m_j = 1$ $(\forall j \in \mathcal{J})$ or $q = k$. The EZKB selection method can accommodate the selection of grouped explanatory variables. 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*

$$\hat{j}_{\mathcal{J}} = \{j \in \mathcal{J} \mid HCGC_p(\omega_j) > HCGC_p(\omega)\}. \tag{3.9}$$

Next, we ensure the consistency of the EZKB selection method using the $HCGC_p$ criterion (3.9). 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 about sufficient conditions for consistency (the proof is given in Appendix A.5).

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

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

Then, the EZKB selection method using the $HCGC_p$ criterion (3.9) is consistent.

Using Lemma 3.2, the consistency of the EZKB selection method using the $HCGC_p$ criterion (3.9) can be obtained as follows (the proof is given in Appendix A.6):

Theorem 3.2. *Suppose that Assumptions A1–A3 hold. Then, the EZKB selection method using the $HCGC_p$ criterion (3.9) 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 (3.2). Hence, as an example of the consistent EZKB selection method under $3/4 < c_A \leq 1$ in Assumption A3, we can use the method using the $HCGC_p$ criterion with $\alpha = \tilde{\alpha}$ in (3.3).

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

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

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

$$GC_p(\omega_j) - GC_p(\omega) = \mathrm{tr}(\boldsymbol{R}_j^{-1}\boldsymbol{Z}_j'\boldsymbol{Y}\boldsymbol{S}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{Z}_j) - m_j p\alpha. \tag{3.10}$$

The proof of the above equation is omitted because it essentially mimics (3.4). Although (3.10) 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 (3.3) 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 (3.5) (and (3.6) if $p < 100$ and $k \geq p$). The explanatory matrix $\boldsymbol{X}$, the true coefficient matrix $\boldsymbol{\Theta}_*$ and the true covariance matrix $\boldsymbol{\Sigma}_*$ were determined as follows:

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

where $\boldsymbol{\Psi}$ is the $k\times k$ autoregressive matrix with the correlation ψ, 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 $\xi_2 = 0.8$. Although Theorems 3.1 and 3.2 were obtained by assuming that $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model, we also examine the probabilities under the non-normality in this simulation. Let $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n)'$ be a $n\times p$ random matrix, where $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n$ are independent and identically distributed according to the multivariate t-distribution with 10 degrees of freedom, mean $\boldsymbol{0}_p$ and covariance matrix $(5/4)\boldsymbol{\Sigma}_*$. Then, we constructed the following two true models:

- Multivariate normal distribution: $\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}(\boldsymbol{\Theta}_*', \boldsymbol{O}_{k-k_*,p}')', \boldsymbol{\Sigma}_*\otimes\boldsymbol{I}_n)$.
- Multivariate t-distribution: $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\Theta}_*', \boldsymbol{O}_{k-k_*,p}')' + (4/5)^{1/2}\boldsymbol{\mathcal{E}}$.

For comparison, we also calculated the probabilities of selecting the true subset and the CPU times using the adaptive group lasso (AGL) proposed by [25]. The estimator of $\boldsymbol{\Theta}$ by the AGL is written as

$$\begin{aligned}
\hat{\boldsymbol{\Theta}}_\tau &= \arg\min_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}|\tau),\\
f(\boldsymbol{\Theta}|\tau) &= \mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta})'\} + 2\tau\sum_{a=1}^{k} w_a||\boldsymbol{\theta}_a||,
\end{aligned} \tag{4.1}$$

where τ 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 (4.1) is centralized and standardized. To optimize (4.1), we used a coordinate descent algorithm based on [6]. The algorithm is given as follows. Let 100 candidates of τ be $\tau_t = \exp\{t\log(\tau_{\max}+1)/(100-1)\} - 1$ $(t \in \{0,1,2,\ldots,99\})$, where $\tau_{\max} = \max_{a=1,\ldots,k} w_a^{-1}||\boldsymbol{Y}'\boldsymbol{x}_{\{a\}}||$. Initialize $\hat{\boldsymbol{\Theta}}_{\tau_0} = \hat{\boldsymbol{\Theta}}^{\text{aft}}_{\tau_0} = (\hat{\boldsymbol{\theta}}_1^{(0)},\ldots,\hat{\boldsymbol{\theta}}_k^{(0)})' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. For $t = 1,\ldots,99$,

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

$$\left|1 - \frac{f(\hat{\boldsymbol{\Theta}}^{\text{aft}}_{\tau_t}|\tau_t)}{f(\hat{\boldsymbol{\Theta}}^{\text{bef}}_{\tau_t}|\tau_t)}\right| < \varepsilon_{\text{AGL}},$$

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

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

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

where $|\mathcal{A}_t|$ is the number of non-zero row vectors of $\hat{\boldsymbol{\Theta}}_{\tau_t}$, and $\alpha_1 = 2$, $\alpha_2 = 2\log\log n$ and $\alpha_3 = \log n$. We name the AGL using $\text{IC}(\tau_t|\alpha_i)$ $(i = 1,2,3)$ as AGL^1, AGL^2 and 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 AGL^i $(i = 1,2,3)$ when $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model. 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 AGL^1 are low as the sample size increases in many cases. The probabilities by AGL^2 reach 100% only when the sample size is large and the dimensions are small. The probabilities by AGL^3 seem to increase slowly in some cases, but are low when k_* is large. Table 2 shows the probabilities when $\boldsymbol{Y}$ is dis-

TABLE 1
True subset selection probabilities (%) when $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model

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	80.4	99.8	100.0	40.6	69.4	81.7
500	10	10	5	100.0	85.0	100.0	100.0	60.7	69.0	95.2
1000	10	10	5	100.0	85.5	100.0	100.0	74.4	76.9	97.7
2000	10	10	5	100.0	85.2	100.0	100.0	87.0	99.5	99.5
3000	10	10	5	100.0	83.9	100.0	100.0	11.0	100.0	100.0
200	160	10	5	99.9	0.0	0.0	0.3	0.0	0.0	0.3
500	400	10	5	100.0	0.0	0.0	37.4	0.0	0.0	30.5
1000	800	10	5	100.0	0.0	0.0	96.1	0.0	0.0	62.9
2000	1600	10	5	100.0	0.0	0.0	100.0	0.0	0.0	85.7
3000	2400	10	5	100.0	0.0	0.0	100.0	0.0	0.0	92.1
200	10	160	5	100.0	0.0	16.5	88.3	0.2	1.5	4.3
500	10	400	5	100.0	0.0	70.4	100.0	1.0	5.9	15.5
1000	10	800	5	100.0	0.0	90.5	100.0	15.7	26.0	36.6
2000	10	1600	5	100.0	0.0	93.7	100.0	44.0	84.1	86.7
3000	10	2400	5	100.0	0.0	95.4	100.0	24.1	24.3	35.1
200	10	160	80	100.0	0.2	34.0	93.5	0.0	0.0	0.1
500	10	400	200	100.0	0.1	83.1	99.9	0.0	0.0	3.1
1000	10	800	400	100.0	0.0	93.2	100.0	0.0	0.0	0.0
2000	10	1600	800	100.0	0.0	97.1	100.0	0.0	0.0	0.0
3000	10	2400	1200	100.0	0.0	97.5	100.0	0.0	0.0	0.0
200	80	80	5	100.0	0.0	0.0	31.5	0.0	0.0	2.5
500	200	200	5	100.0	0.0	0.0	99.4	0.0	5.3	19.4
1000	400	400	5	100.0	0.0	0.3	100.0	0.0	20.7	40.8
2000	800	800	5	100.0	0.0	78.2	100.0	0.0	41.3	66.1
3000	1200	1200	5	100.0	0.0	99.6	100.0	0.0	51.4	78.9
200	80	80	40	100.0	0.0	0.0	52.7	0.0	0.0	0.2
500	200	200	100	100.0	0.0	0.0	99.7	0.0	0.0	0.0
1000	400	400	200	100.0	0.0	3.7	100.0	0.0	0.0	1.3
2000	800	800	400	100.0	0.0	89.9	100.0	0.0	0.0	68.4
3000	1200	1200	600	100.0	0.0	100.0	100.0	0.0	0.0	96.0

tributed according to the multivariate t-distribution under the true model. We observe that the results in Table 2 are about the same as those in Table 1. Hence, it is expected that our results may hold even under the non-normality. The proofs of Theorems 3.1 and 3.2 in this paper are needed to calculate the moments of $GC_p(\omega_{\{\ell\}}) - GC_p(\omega)$ and $GC_p(\omega_j) - GC_p(\omega)$, and we calculated them by assuming that $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model. However, our results will be shown even under non-normality if another approach to the evaluation of the moments exists although we need to calculate the moments consisting of the inverse matrix $\boldsymbol{S}_\omega^{-1}$. Table 3 shows the CPU times by the ZKB selection method using the $HCGC_p$ criterion denoted by $HCGC_p$ and AGL^3 when $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model, 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 the AGLs.

TABLE 2
True subset selection probabilities (%) when $\boldsymbol{Y}$ is distributed according to the multivariate t-distribution under the true model

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	80.6	99.5	100.0	51.0	67.8	81.8
500	10	10	5	100.0	85.2	99.9	100.0	76.5	92.2	96.9
1000	10	10	5	100.0	86.9	100.0	100.0	78.9	95.6	97.9
2000	10	10	5	100.0	87.4	100.0	100.0	83.3	99.5	99.5
3000	10	10	5	100.0	84.0	99.9	100.0	0.0	99.8	100.0
200	160	10	5	99.8	0.0	0.0	0.2	0.0	0.0	0.2
500	400	10	5	100.0	0.0	0.0	33.8	0.0	0.0	23.8
1000	800	10	5	100.0	0.0	0.0	96.2	0.0	0.0	65.7
2000	1600	10	5	100.0	0.0	0.0	100.0	0.0	0.0	81.9
3000	2400	10	5	100.0	0.0	0.0	100.0	0.0	0.0	89.8
200	10	160	5	100.0	0.0	19.7	87.9	0.2	0.8	3.9
500	10	400	5	100.0	0.0	72.7	100.0	3.0	9.5	20.2
1000	10	800	5	100.0	0.0	88.4	100.0	4.8	22.2	49.5
2000	10	1600	5	100.0	0.0	94.9	100.0	12.4	50.2	66.9
3000	10	2400	5	100.0	0.0	96.0	100.0	33.3	36.0	65.4
200	10	160	80	100.0	0.3	33.1	93.1	0.0	0.0	0.0
500	10	400	200	100.0	0.0	84.7	100.0	0.0	0.0	0.0
1000	10	800	400	100.0	0.0	93.0	100.0	0.0	0.0	0.0
2000	10	1600	800	100.0	0.0	96.9	100.0	0.0	0.0	0.0
3000	10	2400	1200	100.0	0.0	97.8	100.0	0.0	0.0	0.0
200	80	80	5	100.0	0.0	0.0	34.4	0.0	0.2	3.4
500	200	200	5	100.0	0.0	0.0	99.6	0.0	4.8	21.9
1000	400	400	5	100.0	0.0	0.1	100.0	0.0	14.7	37.3
2000	800	800	5	100.0	0.0	79.0	100.0	0.0	43.6	66.7
3000	1200	1200	5	100.0	0.0	99.8	100.0	0.0	48.5	76.2
200	80	80	40	100.0	0.0	0.0	52.9	0.0	0.0	0.3
500	200	200	100	100.0	0.0	0.0	99.8	0.0	0.0	0.0
1000	400	400	200	100.0	0.0	4.4	100.0	0.0	0.0	2.2
2000	800	800	400	100.0	0.0	90.5	100.0	0.0	0.0	68.1
3000	1200	1200	600	100.0	0.0	100.0	100.0	0.0	0.0	94.0

Appendix A

A.1. Proof of equations (2.4) *and* (2.5)

First, we show (2.5). For an arbitrary $\ell \in j_*$, we have the following equation:

$$(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}})\boldsymbol{x}_{\{\ell_1\}} \begin{cases} = \boldsymbol{0}_n & (\ell_1 \neq \ell) \\ \neq \boldsymbol{0}_n & (\ell_1 = \ell) \end{cases} .$$

From the above equation, $\boldsymbol{\Delta}_{\{\ell\}}$ which is defined in (2.2) can be expressed as follows:

$$\begin{aligned} \boldsymbol{\Delta}_{\{\ell\}} &= \boldsymbol{\Sigma}_*^{-1/2} \left(\sum_{\ell \in j_*} \boldsymbol{\theta}_{\{\ell\}} \boldsymbol{x}'_{\{\ell\}} \right) (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}}) \left(\sum_{\ell \in j_*} \boldsymbol{x}_{\{\ell\}} \boldsymbol{\theta}'_{\{\ell\}} \right) \boldsymbol{\Sigma}_*^{-1/2} \\ &= \boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\theta}_{\{\ell\}} \boldsymbol{x}'_{\{\ell\}} (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}}) \boldsymbol{x}_{\{\ell\}} \boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1/2} \\ &= \boldsymbol{x}'_{\{\ell\}} (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}}) \boldsymbol{x}_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\theta}_{\{\ell\}} \boldsymbol{\theta}'_{\{\ell\}} \boldsymbol{\Sigma}_*^{-1/2}. \end{aligned}$$

TABLE 3
CPU times (s) when $\boldsymbol{Y}$ is distributed according to the multivariate normal distribution under the true model

n	p	k	k_*	$HCGC_p$	AGL[3]
200	10	10	5	0.0010	0.0091
500	10	10	5	0.0046	0.0108
1000	10	10	5	0.0124	0.0208
2000	10	10	5	0.0346	0.0428
3000	10	10	5	0.0606	0.0679
200	160	10	5	0.0036	0.0611
500	400	10	5	0.0438	0.8003
1000	800	10	5	0.2996	4.9315
2000	1600	10	5	2.1194	33.0587
3000	2400	10	5	6.6364	99.3778
200	10	160	5	0.0029	0.1048
500	10	400	5	0.0167	0.4377
1000	10	800	5	0.0734	1.3389
2000	10	1600	5	0.3844	4.7584
3000	10	2400	5	1.1095	11.4628
200	10	160	80	0.0024	0.0934
500	10	400	200	0.0146	0.4503
1000	10	800	400	0.0774	1.3550
2000	10	1600	800	0.4126	4.8661
3000	10	2400	1200	1.1370	11.8354
200	80	80	5	0.0045	0.1386
500	200	200	5	0.0327	1.5399
1000	400	400	5	0.5254	57.9087
2000	800	800	5	4.9127	492.7444
3000	1200	1200	5	15.5871	1724.8028
200	80	80	40	0.0059	0.1409
500	200	200	100	0.0312	1.5835
1000	400	400	200	0.5342	59.2730
2000	800	800	400	4.8972	516.9926
3000	1200	1200	600	15.7256	1730.3650

Hence, we have

$$\delta_{\{\ell\}} = \mathrm{tr}(\boldsymbol{\Delta}_{\{\ell\}}) = \boldsymbol{x}'_{\{\ell\}}(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}})\boldsymbol{x}_{\{\ell\}}\boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\theta}_{\{\ell\}}.$$

The above equation leads to the following inequality:

$$\begin{aligned} n^{-c_A}\delta_{\min} &\geq \left\{ n^{-1} \min_{\ell \in j_*} \boldsymbol{x}'_{\{\ell\}}(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_{\{\ell\}}})\boldsymbol{x}_{\{\ell\}} \right\} \left\{ n^{1-c_A} \min_{\ell \in j_*} \boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\theta}_{\{\ell\}} \right\} \\ &\geq c_1 c_2. \end{aligned}$$

Next, we show (2.4). Let $\lambda_{\max}(\boldsymbol{A})$ be the maximum eigenvalue of the square matrix $\boldsymbol{A}$. Then, we have

$$\boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\theta}_{\{\ell\}} = \lambda_{\max}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\theta}_{\{\ell\}}\boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1/2}) = \max_{||\boldsymbol{e}||=1} \boldsymbol{e}'\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\theta}_{\{\ell\}}\boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{e}.$$

By using $(\boldsymbol{e}'_a\boldsymbol{\Sigma}_*\boldsymbol{e}_a)^{-1/2}\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{e}_a$ as $\boldsymbol{e}$, we have

$$\lambda_{\max}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\theta}_{\{\ell\}}\boldsymbol{\theta}'_{\{\ell\}}\boldsymbol{\Sigma}_*^{-1/2}) \geq \max_{a=1,\ldots,p} \frac{(\boldsymbol{e}'_a\boldsymbol{\theta}_{\{\ell\}})^2}{\sigma_{*a}^2},$$

where $\boldsymbol{e}_a$ is the p-dimensional vector such that the a-th element is one and the other elements are zero. The above equation completes the proof of (2.4). □

A.2. Proof of equation (3.4)

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_ℓ 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., [7, Theorem 8.5.11]), $\boldsymbol{P}_\omega$ and $\boldsymbol{P}_{\omega_{\{\ell\}}}$ can be expressed as follows:

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

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 $\boldsymbol{X}(\boldsymbol{r}'_\ell, r_\ell)'$ is the k-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Therefore, (3.4) can be derived. □

A.3. Proof of Lemma 3.1

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

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

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. □

A.4. 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 A.7):

Lemma A.1. *For $j \subset \omega$ $(j \neq \omega)$, suppose that $\delta_{j,i}$ $(i = 1, \ldots, m_j)$ are constants satisfying $\mathrm{tr}(\boldsymbol{\Delta}_j) = \sum_{i=1}^{m_j} \delta_{j,i}$ and $\delta_{j,i} \geq m_j^{-1} \lambda_{\max}(\boldsymbol{\Delta}_j)$, where $\boldsymbol{\Delta}_j$ and m_j are defined by (2.2) and (2.6), 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 $v_i \sim \chi^2(n-p-k+1)$ $(i = 1, \ldots, m_j)$, where u_i and $u_{j,i}$ are independent of v_i for each i. Then, under Assumption A1, we have*

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

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 A.8).

Lemma A.2. *Let δ be a positive constant. And 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, and $N = n - p - k + 1$. Then, for $N - 4r > 0$ $(r \in \mathbb{N})$, we have*

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = O(p^r n^{-2r}),$$
$$E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = O(\max\{(p+\delta)^r n^{-2r}, (p+\delta)^{2r} n^{-3r}\}),$$

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

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

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

where u and u_ℓ 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 (A.2), we have

$$\sum_{\ell \notin j_*} P(HCGC_p(\omega_{\{\ell\}}) > HCGC_p(\omega))$$
$$= (k - k_*) P\left(\frac{u}{v} > \frac{p}{n-k}\alpha\right) = (k-k_*) P\left(\frac{u}{v} - \frac{p}{N-2} > \rho\right)$$

$$\leq (k-k_*)P\left(\left|\frac{u}{v}-\frac{p}{N-2}\right|\geq\rho\right), \tag{A.3}$$

where $\rho=\{p/(n-k)\}\beta$. Moreover, since $(N-2)\rho\delta_{\min}^{-1}=O(p\beta n^{-c_A})=o(1)$ from (2.5) and (3.1), when N is sufficiently large, we have

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

Applying Markov's inequality to (A.3) and (A.4), the following upper bounds can be derived:

$$\begin{aligned}
&(k-k_*)P\left(\left|\frac{u}{v}-\frac{p}{N-2}\right|\geq\rho\right)\leq(k-k_*)\rho^{-2r}E\left[\left(\frac{u}{v}-\frac{p}{N-2}\right)^{2r}\right],\\
&\sum_{\ell\in j_*}P\left(\left|\frac{u_\ell}{v}-\frac{p+\delta_{\{\ell\}}}{N-2}\right|\geq\frac{\delta_{\{\ell\}}}{N-2}-\rho\right)\\
&\leq\sum_{\ell\in j_*}\left(\frac{\delta_{\{\ell\}}}{N-2}-\rho\right)^{-2\tilde{r}}E\left[\left(\frac{u_\ell}{v}-\frac{p+\delta_{\{\ell\}}}{N-2}\right)^{2\tilde{r}}\right],
\end{aligned}$$

where r is a natural number defined by (3.1) and $\tilde{r}$ are any natural numbers. From the above equations and Lemma A.2, the following equation can be derived:

$$\sum_{\ell\notin j_*}P(HCGC_p(\omega_{\{\ell\}})>HCGC_p(\omega))=O(kp^{-r}\beta^{-2r})=o(1).$$

Moreover, for sufficiently large $\tilde{r}$, since $k_*p^{\tilde{r}}n^{-2\tilde{r}c_A}=o(1)$, $k_*n^{-\tilde{r}c_A}=o(1)$, $k_*p^{2\tilde{r}}n^{-2\tilde{r}c_A-\tilde{r}}=o(1)$ and $k_*n^{-\tilde{r}}=o(1)$, we have

$$\begin{aligned}
&\sum_{\ell\in j_*}P(HCGC_p(\omega_{\{\ell\}})<HCGC_p(\omega))\\
&=\sum_{\ell\in j_*}O\left((p+\delta_{\{\ell\}})^{\tilde{r}}\delta_{\{\ell\}}^{-2\tilde{r}}+(p+\delta_{\{\ell\}})^{2\tilde{r}}\delta_{\{\ell\}}^{-2\tilde{r}}n^{-\tilde{r}}\right)\\
&=O\left(k_*(p+\delta_{\min})^{\tilde{r}}\delta_{\min}^{-2\tilde{r}}+k_*(p+\delta_{\min})^{2\tilde{r}}\delta_{\min}^{-2\tilde{r}}n^{-\tilde{r}}\right)=o(1).
\end{aligned}$$

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

A.5. Proof of Lemma 3.2

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

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

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. □

A.6. Proof of Theorem 3.2

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

$$HCGC_p(\omega_j) - HCGC_p(\omega) = \begin{cases} (n-k) \displaystyle\sum_{i=1}^{m_j} \frac{u_i}{v_i} - m_j p \alpha & (j \in \mathcal{J}_-) \\ (n-k) \displaystyle\sum_{i=1}^{m_j} \frac{u_{j,i}}{v_i} - m_j p \alpha & (j \in \mathcal{J}_+) \end{cases}, \quad \text{(A.5)}$$

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

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

Here, $\delta_{j,i}$ $(i = 1, \ldots, m_j)$ are constants satisfying $\sum_{i=1}^{m_j} \delta_{j,i} = \mathrm{tr}(\boldsymbol{\Delta}_j)$ and $\delta_{j,i} \geq m_j^{-1} \lambda_{\max}(\boldsymbol{\Delta}_j)$, where $\boldsymbol{\Delta}_j$ is given by (2.2). When $j \in \mathcal{J}_+$, let ℓ 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:

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

In addition, let $d_j = \mathrm{rank}(\boldsymbol{\Delta}_j)$ which is defined by (2.6). From (2.7), we observe that d_j is bounded. Since $d_j \lambda_{\max}(\boldsymbol{\Delta}_j) \geq \mathrm{tr}(\boldsymbol{\Delta}_j)$ holds, the following inequality is obtained:

$$\delta_{j,i} \geq m_j^{-1} \lambda_{\max}(\boldsymbol{\Delta}_j) \geq (m_j d_j)^{-1} \mathrm{tr}(\boldsymbol{\Delta}_j) \geq (m_j d_j)^{-1} \delta_{\{\ell\}}. \quad \text{(A.6)}$$

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 (2.5), (3.1), (A.5) and (A.6), when N is sufficiently large, we have

$$\begin{aligned}
&\sum_{j\in\mathcal{J}_-} P(HCGC_p(\omega_j) > HCGC_p(\omega)) \\
&= \sum_{j\in\mathcal{J}_-} P\left(\sum_{i=1}^{m_j}\frac{u_i}{v_i} > \frac{m_j p}{n-k}\alpha\right) \le \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\frac{u_i}{v_i} > \frac{p}{n-k}\alpha\right) \\
&= \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\frac{u_i}{v_i} - \frac{p}{N-2} > \rho\right) \le \sum_{j\in\mathcal{J}_-}\sum_{i=1}^{m_j} P\left(\left|\frac{u_i}{v_i} - \frac{p}{N-2}\right| \ge \rho\right), \quad \text{(A.7)} \\
&\sum_{j\in\mathcal{J}_+} P(HCGC_p(\omega_j) < HCGC_p(\omega)) \\
&= \sum_{j\in\mathcal{J}_+} P\left(\sum_{i=1}^{m_j}\frac{u_{j,i}}{v_i} < \frac{m_j p}{n-k}\alpha\right) \le \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\frac{u_{j,i}}{v_i} < \frac{p}{n-k}\alpha\right) \\
&= \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2} < \rho - \frac{\delta_{j,i}}{N-2}\right) \\
&\le \sum_{j\in\mathcal{J}_+}\sum_{i=1}^{m_j} P\left(\left|\frac{u_{j,i}}{v_i} - \frac{p+\delta_{j,i}}{N-2}\right| \ge \frac{\delta_{j,i}}{N-2} - \rho\right), \quad \text{(A.8)}
\end{aligned}$$

where $\rho = \{p/(n-k)\}\beta$. Then, by applying Markov's inequality to (A.7) and (A.8), their following upper bounds can be derived:

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

where $\tilde{r}$ are any natural numbers. Hence, from the above equations and Lemma A.2, the following equations can be derived:

$$\sum_{j\in\mathcal{J}_-} m_j\rho^{-2r} E\left[\left(\frac{u_1}{v_1} - \frac{p}{N-2}\right)^{2r}\right] = O(kp^{-r}\beta^{-2r}) = o(1).$$

Note that m_j is bounded and $\#(\mathcal{J}_+) \le k_*$, and it follows from (A.6) that $\delta_{j,i}^{-1} \le m_j d_j \delta_{\{\ell\}}^{-1}$. Hence, for sufficiently large $\tilde{r}$, we have

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

$$= \sum_{j \in \mathcal{J}_+} \sum_{i=1}^{m_j} O\left((p+\delta_{j,i})^{\tilde{r}} \delta_{j,i}^{-2\tilde{r}} + (p+\delta_{j,i})^{2\tilde{r}} \delta_{j,i}^{-2\tilde{r}} n^{-\tilde{r}}\right) = o(1).$$

Therefore, from Lemma 3.2, Theorem 3.2 can be shown. □

A.7. Proof of Lemma A.1

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

$$\begin{aligned} & GC_p(\omega_j) - GC_p(\omega) \\ &= \sum_{i=1}^{m_j} \{GC_p(j_{-,i-1}) - GC_p(j_{-,i})\} \\ &= (n-k) \sum_{i=1}^{m_j} \mathrm{tr}[\boldsymbol{Y}'(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})\boldsymbol{Y}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\}^{-1}] - m_j p\alpha. \end{aligned} \tag{A.9}$$

Let

$$\boldsymbol{W}_{j,i} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{P}_{j_{-,i}} - \boldsymbol{P}_{j_{-,i-1}})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}, \ \boldsymbol{W} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}.$$

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., [5, chap 2, Theorem 2.4.2]), 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, (A.9) is expressed as

$$GC_p(\omega_j) - GC_p(\omega) = (n-k) \sum_{i=1}^{m_j} \mathrm{tr}(\boldsymbol{W}_{j,i}\boldsymbol{W}^{-1}) - m_j p\alpha. \tag{A.10}$$

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 (A.10), we can derive (A.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

$$GC_p(\omega_j) - GC_p(\omega) = (n-k)\mathrm{tr}[\boldsymbol{Y}'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_j})\boldsymbol{Y}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\}^{-1}] - m_j p\alpha. \tag{A.11}$$

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, (A.11) is expressed as

$$GC_p(\omega_j) - GC_p(\omega) = (n-k)\mathrm{tr}(\boldsymbol{W}_j\boldsymbol{W}^{-1}) - m_j p\alpha. \tag{A.12}$$

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., $\boldsymbol{\Lambda}_j = \mathrm{diag}(\lambda_{j,1}, \ldots, \lambda_{j,p})$ $(\lambda_{j,1} \geq \cdots \geq \lambda_{j,p})$. 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 (2.6). It is obvious that $\lambda_{j,d_j+1} = \cdots = \lambda_{j,p} = 0$. Since it holds that $d_j \leq m_j$ from (2.7), let $\boldsymbol{\Gamma}_j$ be as follows:

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

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 $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{m_j})$ be a $m_j \times m_j$ orthogonal matrix satisfying $\boldsymbol{h}_1 = m_j^{-1/2}\boldsymbol{1}_{m_j}$, and we express $\boldsymbol{H}\boldsymbol{\Gamma}_j$ as $\boldsymbol{H}\boldsymbol{\Gamma}_j = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{m_j})'$. Then, we have

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

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

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

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 (A.12), we can derive (A.1) for the case of $j\subset j_*$. □

A.8. Proof of Lemma A.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 A.2 (the proof is given in Appendix A.9).

Lemma A.3. *Let $X_1\sim\chi^2(t)$ and $X_2\sim\chi^2(t;\psi)$, where ψ is a positive constant. Then, we have*

$$E[(X_1-t)^h] = \begin{cases} 1 & (h=0)\\ 0 & (h=1)\\ O(t^{\lfloor h/2\rfloor}) & (h\geq 2)\end{cases}, \tag{A.13}$$

$$E[\{X_2-(t+\psi)\}^h] = \begin{cases} 1 & (h=0)\\ 0 & (h=1)\\ O((t+\psi)^{\lfloor h/2\rfloor}) & (h\geq 2)\end{cases}. \tag{A.14}$$

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

$$E\left[\left(\frac{1}{X_1}-\frac{1}{t-2}\right)^h\right] = \begin{cases} 1 & (h=0)\\ 0 & (h=1)\\ O(t^{-2h+\lfloor h/2\rfloor}) & (h\geq 2)\end{cases}, \tag{A.15}$$

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

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = \sum_{\substack{a+b+c=2r \\ 0\le a,b,c\le 2r}} \frac{(2r)!}{a!b!c!} p^b \xi^c E[(u_1 - p)^{a+c}] E[(v^{-1} - \xi)^{a+b}], \tag{A.16}$$

$$E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = \sum_{\substack{a+b+c=2r \\ 0\le a,b,c\le 2r}} \frac{(2r)!}{a!b!c!} \xi_\delta^b \xi^c E[(u_2 - \xi_\delta)^{a+c}] E[(v^{-1} - \xi)^{a+b}]. \tag{A.17}$$

From (A.13) and (A.15), the divergence order in (A.16) is maximized when $a = b = 0, c = 2r$ because of $pn^{-1} = O(1)$. Moreover, from (A.14) and (A.15), the divergence order in (A.17) is maximized when either $a = b = 0, c = 2r$ or $a = c = 0, b = 2r$. Therefore, we can derive the divergence orders as follows:

$$E\left[\left(\frac{u_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = O(p^r n^{-2r}),$$
$$E\left[\left(\frac{u_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = O(\max\{(p+\delta)^r n^{-2r}, (p+\delta)^{2r} n^{-3r}\}).$$

□

A.9. Proof of Lemma A.3

We elaborate only on the case of $h \ge 2$ because it is straightforward when $h = 0, 1$. First, we derive (A.13) and (A.14). Let $h_1, \ldots, h_d$ be natural numbers satisfying $\sum_{i=1}^d h_i = h$ and $2 \le h_1, \ldots, h_d$. From [22], 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 [9, 23], 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, 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 (A.13) and (A.14).

Next, we derive (A.15). From the multinomial theorem, we have

$$E\left[\left(\frac{1}{X_1} - \frac{1}{t-2}\right)^h\right]$$
$$= \sum_{i=0}^h \frac{h!}{i!(h-i)!}\left(-\frac{1}{t-2}\right)^{h-i} E\left[\left(\frac{1}{X_1}\right)^i\right]$$

$$= \left(-\frac{1}{t-2}\right)^h + \sum_{i=1}^{h} \frac{h!}{i!(h-i)!} \left(-\frac{1}{t-2}\right)^{h-i} \prod_{d=1}^{i} \frac{1}{t-2d}$$

$$= \left(-\frac{1}{t-2}\right)^h$$

$$\cdot \prod_{d=1}^{h} \frac{1}{t-2d} \left[\{-(t-2)\}^h + \sum_{i=0}^{h-1} \frac{h!}{i!(h-i)!} \{-(t-2)\}^i \prod_{d=1}^{h-i} \{t-2h+2(d-1)\}\right].$$

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

$$\begin{aligned} & E\left[\left(\frac{1}{X_1} - \frac{1}{t-2}\right)^h\right] \\ &= \left(-\frac{1}{t-2}\right)^h \prod_{d=1}^{h} \frac{1}{t-2d} \left\{\sum_{i=0}^{h} \frac{h!}{i!(h-i)!} \{-(t-2)\}^i E[T^{h-i}]\right\} \\ &= sE[\{T-(t-2)\}^h] \\ &= s \sum_{i=0}^{h} \frac{h!}{i!(h-i)!} \{-2(h-1)\}^i E[\{T-(t-2h)\}^{h-i}]. \end{aligned} \tag{A.18}$$

Note that $s = O(t^{-2h})$ and it follows from (A.13) that

$$E[\{T-(t-2h)\}^{h-i}] = \begin{cases} 1 & (i=h) \\ 0 & (i=h-1) \\ O(t^{\lfloor (h-i)/2 \rfloor}) & (i \leq h-2) \end{cases}. \tag{A.19}$$

The equations (A.18) and (A.19) complete (A.15). □

A.10. R *file related to this article*

The R file "hcgcp.R" to perform the EZKB selection method using the $HCGC_p$ criterion can be found online at https://github.com/roda6288/hcgcp_function.

Acknowledgments

The authors would like to thank two reviewers and an Associate Editor for many constructive comments. The first author's research is supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. The second author's research is partially supported by a Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Science, Sports, and Culture #18K03415.

References

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

[2] Chudik, A., Kapetanios, G. and Pesaran, M. H. (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrika* **86** 1479–1512. MR3843496

[3] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84** 707–716. MR1603952

[4] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.* **123** 184–200. MR3130429

[5] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* John Wiley & Sons, Inc., Hoboken, New Jersey. MR2640807

[6] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33(1)** 1–22. MR1082147

[7] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York. MR1467237

[8] He, Y., Jiang, T., Wen, J. and Xu, G. (2018). Likelihood ratio test in multivariate linear regression: from low to high dimension. arXiv:1812.06894.

[9] 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. MR0719028

[10] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

[11] Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240

[12] Luo, S. (2018). Variable selection in high-dimensional sparse multiresponse linear regression models. *Stat. Pap.* https://doi.org/10.1007/s00362-018-0989-x.

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

[14] Mallows, C. L. (1995). More comments on C_p. *Technometrics* **37** 362–372. MR1365719

[15] Nagai, I., Yanagihara, H. and Satoh, K. (2012). Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods. *Hiroshima Math. J.* **42** 301–324. MR3050124

[16] Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.* **27** 392–403. MR0970962

[17] Nishii, R., Bai, Z. D. and Krishnaiah, P. R. (1988). Strong consistency

of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.* **18** 451–462. MR0991240

[18] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R. and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4(1)** 53–77. MR2758084

[19] Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76** 369–374. MR1016028

[20] Sparks, R. S., Coutsourides, D. and Troskie, L. (1983). The multivariate C_p. *Comm. Statist. A – Theory Methods* **12** 1775–1793. MR0704853

[21] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, New York. MR1915968

[22] Stuart, A. and 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. MR0246399

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

[24] Timm, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag, New York. MR1908225

[25] Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Comput. Stat. Data An.* **52** 5277–5286. MR2526593

[26] Xin, X., Hu, J. and Liu, L. (2017). On the oracle property of a generalized adaptive elastic-net for multivariate linear regression with a diverging number of parameters. *J. Multivariate Anal.* **162** 16–32. MR3719332

[27] 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.

[28] 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. Statist.* **9** 869–897. MR3338666

[29] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B* **68** 49–67. MR2212574

[30] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Stat. Soc. B* **69** 329–346. MR2323756

[31] Zhao, L. C., Krishnaiah, 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. MR0862239