Hiroshima Math. J.
00 (0000), 1–32

Consistent variable selection criteria in multivariate linear regression even when dimension exceeds sample size

Ryoya Oda

(Received Xxx 00, 0000)

Abstract. This paper is concerned with the selection of explanatory variables in multivariate linear regression. The Akaike's information criterion and the C_p criterion cannot perform in high-dimensional situations such that the dimension of a vector stacked with response variables exceeds the sample size. To overcome this, we consider two variable selection criteria based on an L_2 squared distance with a weighted matrix, namely the scalar-type generalized C_p criterion and the ridge-type generalized C_p criterion. We clarify conditions for their consistency under a hybrid-ultra-high-dimensional asymptotic framework such that the sample size always goes to infinity but the number of response variables may go to infinity. Numerical experiments show that the probabilities of selecting the true subset by criteria satisfying consistency conditions are high even when the dimension is larger than the sample size. Finally, we illuminate the practical utility of these criteria using empirical data.

1. Introduction

Multivariate linear regression is an important and very widely used inferential statistical methodology. It is the cornerstone of many theoretical and applied statistics textbooks (see, e.g., Srivastava, 2002, chap 9; Timm, 2002, chap 4) and it has widespread applications in many fields. Let $\boldsymbol{Y} = (\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(n)})'$ be an $n \times p$ observation matrix stacking individual p response variables, and $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})'$ be an $n \times k$ observation matrix stacking individual non-stochastic k explanatory variables, where n is the sample size. Note that $\boldsymbol{X}$ may include the intercept term that the column vector is $\boldsymbol{1}_n$,where $\boldsymbol{1}_n$ is an n-dimensional vector of ones. Assume that $\mathrm{rank}(\boldsymbol{X}) = k < n$ to ensure the existence of variable selection criteria used in this paper. We consider linear regression for n samples of a vector of individual p response variables and k explanatory variables on $\{(\boldsymbol{y}'_{(i)}, \boldsymbol{x}'_{(i)})' |\ i = 1, \ldots, n\}$. Then, the multivariate linear regression is written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{\mathcal{E}},$$

The author is supported financially by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

2010 *Mathematics Subject Classification.* Primary 62J05; Secondary 62H12.

Key words and phrases. Hybrid-ultra-high-dimensional asymptotic framework, Multivariate linear regression, Non-normality, Selection consistency, Variable selection criterion.

where $\mathbf{\Theta}$ is a $k \times p$ unknown matrix of regression coefficients, and each row of an $n \times p$ error matrix $\boldsymbol{\mathcal{E}}$ is identically distributed with a mean vector $\mathbf{0}_p$, which is a p-dimensional vector of zeros, and a covariance matrix $\mathbf{\Sigma}$.

In actual data analysis contexts, it is important to specify salient explanatory variables affecting response variables. In multivariate linear regression, this is regarded as the problem of selecting the best subset of explanatory variables. Variable selection criteria are widely used in empirical contexts to choose the best subset of explanatory variables. The Akaike's information criterion (AIC) (Akaike, 1973; 1974) and the C_p criterion (Sparks *et al.*, 1983) which is a multivariate version of Mallows' C_p criterion (Mallows, 1973; 1995) are well-known examples in this respect. The AIC and C_p criterion are estimators of risk functions, the Kullback-Leibler loss function and the mean squared prediction error standardized by the true covariance matrix, respectively. Further, as extensions of the AIC and C_p criterion, the generalized information criterion (GIC) and the generalized C_p (GC_p) criterion were proposed by Nishii *et al.* (1988) and Nagai *et al.* (2012), respectively. The GIC and GC_p criterion were generalized from the AIC and C_p criterion by replacing "2" (the penalty term for model complexity) with any positive number. Note that the GIC includes the AIC, the Bayesian information criterion (BIC) proposed by Schwarz (1978), a consistent AIC (CAIC) proposed by Bozdogan (1987), and the Hannan-Quinn information criterion (HQC) proposed by Hannan and Quinn (1979). Further, the GC_p criterion includes the C_p criterion and the modified C_p (MC_p) criterion proposed by Fujikoshi and Satoh (1997).

Importantly, there are increasing demands in recent years vis-a-vis analyzing high-dimensional data such that p exceeds n (for an example, see Wille *et al.*, 2004). For high-dimensional cases, we need a variable selection criterion which can be operationalized even when $p > n$. However, note that the GIC consists of the logarithm of the determinant of the sample covariance matrix, and the GC_p criterion consists of the inverse matrix of the sample covariance matrix. Therefore, since the sample covariance matrix becomes singular when p is larger than n, more precisely $n - k < p$, the GIC always gives $-\infty$ and the GC_p criterion cannot be defined when $p > n$. However, criteria proposed by Fujikoshi *et al.* (2011), Yamamura *et al.* (2010), and Kubokawa and Srivastava (2012) are calculable even when $p > n$. Fujikoshi *et al.* (2011) proposed the prediction error (PE) criterion based on the mean squared prediction error. Yamamura *et al.* (2010) and Kubokawa and Srivastava (2012) proposed criteria using a ridge-type sample covariance matrix as an estimator of the true covariance matrix. Moreover, their criteria are exact or asymptotically unbiased estimators of risk functions under some conditions.

In this paper, we consider consistency as one of the asymptotic properties of variable selection criteria. In a given variable selection context, the desired outcome is to specify explanatory variables which substantively affect the response variable according to the nature and extent of available empirical data. In other words, it is hoped that the true subset of variables is identified as the

best subset by variable selection. Since we do not know the true subset, we use a variable selection criterion to maximize the probability of selecting the true subset. When the probability that the chosen subset is the true subset approaches 1, consistency is assured, i.e., the following equation holds:

$$P(\hat{j} = j_*) \to 1,$$

where $\hat{j}$ is the best subset according to a variable selection criterion and j_* is the true subset. It is expected that a consistent variable selection criterion has a high probability of selecting the true subset when the amount of data is sufficient. Therefore, consistency is an important property of a variable selection criterion. In the context of $n > p$, assuming that the true distribution of the error vector is the multivariate normal distribution, Fujikoshi *et al.* (2014) and Yanagihara *et al.* (2015) obtained the consistency properties of criteria such as the AIC and C_p criterion. They used a moderate-high-dimensional asymptotic framework such that both n and p go to ∞ but p does not exceed n. Moreover, Yanagihara *et al.* (2015) also used an asymptotic framework defined by adding $k/n \to 0$ to the moderate-high-dimensional asymptotic framework. Relaxing the normality assumption, Yanagihara (2015) dealt with conditions for consistency of the GIC under the moderate-high-dimensional asymptotic framework. Under the normality assumption, Yanagihara (2016) obtained conditions for consistency of the GC_p criterion under a hybrid-moderate-high-dimensional asymptotic framework such that n goes to ∞ and p may go to ∞ but p/n converges to some positive constant included in $[0, 1)$. Relaxing the normality assumption, Yanagihara (2019) focused on conditions for consistency of the GIC and GC_p criterion under the hybrid-moderate-high-dimensional asymptotic framework. As such, therein, p does not exceed n. On the other hand, in the context where $p > n$, Katayama and Imori (2014) considered variable selection criteria based on a lasso-type estimation for the inverse of the covariance matrix. Under the normality assumption, they showed that the criteria are consistent in a restricted-ultra-high-dimensional asymptotic framework such that both n and p go to infinity but p may exceed n and $\log p/n \to 0$ while $k/n \to 0$.

The aim of this paper is to obtain conditions for consistency of variable selection criteria (which are introduced in subsection 2.1) under non-normality and a high-dimensional asymptotic framework such that n goes to infinity but p may exceed n. To obtain conditions for consistency, the following hybrid-ultra-high-dimensional (HUHD) asymptotic framework is mainly used:

$$\text{HUHD} : n \to \infty,\ p/n \to c \in [0, \infty],\ k\text{: fixed},$$

where $c = \infty$ means that p/n goes to ∞. The HUHD asymptotic framework has two key characteristics. First, the divergence speed of p is not restricted, hence this asymptotic framework incorporates an asymptotic framework such that n and p go to ∞ but p may be larger than n, namely the ultra-high-dimensional (UHD) asymptotic framework, which is written as

$$\text{UHD} : (n, p) \to \infty,\ p/n \to c \in [0, \infty],\ k\text{: fixed}.$$

Second, the HUHD asymptotic framework also includes the large-sample asymptotic framework such that only n tends to ∞. From this, it is expected that consistent variable selection criteria under the HUHD asymptotic framework select the true subset with high probability regardless of the size of p.

The remainder of the paper is organized as follows. In section 2, we present the necessary notation and assumptions to clarify conditions for consistency. In section 3, we obtain conditions for consistency. In section 4, for the purposes of verification, we conduct numerical experiments and illuminate the practical utility of consistent criteria by using real data examples. Technical details are provided in the Appendix.

2. Preliminaries

2.1. Models and Criteria. Suppose that j denotes a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements, and $\boldsymbol{X}_j$ denotes an $n \times k_j$ matrix consisting of columns of $\boldsymbol{X}$ indexed by elements of j, where k_A is the number of elements in a set A denoted by $k_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth column vectors of $\boldsymbol{X}$. Then, the candidate model M_j with k_j explanatory variables from subset j is expressed as follows:

$$M_j : \boldsymbol{Y} = \boldsymbol{X}_j \boldsymbol{\Theta}_j + \boldsymbol{\mathcal{E}}_j, \tag{1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients, and each row of $\boldsymbol{\mathcal{E}}_j$ is identically distributed with a mean of $\boldsymbol{0}_p$ and a covariance matrix $\boldsymbol{\Sigma}_j$. Let j_* ($\subset \omega$) be a true subset, and assume that the data are generated from the following true model M_{j_*} with k_{j_*} true explanatory variables:

$$M_{j_*} : \boldsymbol{Y} = \boldsymbol{X}_{j_*} \boldsymbol{\Theta}_* + \boldsymbol{\mathcal{E}}_*,$$

where $\boldsymbol{\Theta}_*$ is a $k_{j_*} \times p$ unknown matrix of true regression coefficients and $\boldsymbol{\mathcal{E}}_* = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ is an $n \times p$ true error matrix. Assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are identically distributed according to a distribution of $\boldsymbol{\varepsilon}$ with

$$E[\boldsymbol{\varepsilon}] = \boldsymbol{0}_p,\ Cov[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma}_*,\ E[||\boldsymbol{\varepsilon}||^4] < \infty,$$

where $||\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true unknown covariance matrix. Although it is typical to assume independence of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$, here we assume a moment condition which relaxes independence; specifically, we assume that for any $i \neq j$, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are satisfied with the following moment condition:

$$\begin{aligned} E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j'] = E[\boldsymbol{\varepsilon}_i] E[\boldsymbol{\varepsilon}_j'],\ E[||\boldsymbol{\varepsilon}_i||^2 ||\boldsymbol{\varepsilon}_j||^2] = E[||\boldsymbol{\varepsilon}_i||^2] E[||\boldsymbol{\varepsilon}_j||^2], \\ E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_j'] = E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i'] E[\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_j']. \end{aligned}$$

Note that the above moment condition is similar to assuming independence. 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. Moreover, for expository purposes, we represent $\boldsymbol{X}_{j_*}$, $\boldsymbol{X}_\omega$, k_{j_*} and k_ω as $\boldsymbol{X}_*$, $\boldsymbol{X}$, k_*, and k, respectively.

We consider two variable selection criteria based on the following weighted L_2 squared distance:

$$d(\boldsymbol{A}, \boldsymbol{B}|\boldsymbol{G}) = \mathrm{tr}\{(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{G}^{-1}(\boldsymbol{A}-\boldsymbol{B})'\},$$

where $\boldsymbol{G}$ is a positive definite matrix. Let $\boldsymbol{S}_j$ be an estimator of $\boldsymbol{\Sigma}_j$ in the candidate model M_j, which is given by

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

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, and $\boldsymbol{P}_j$ 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 minimum value of $d(\boldsymbol{Y}, \boldsymbol{X}_j\boldsymbol{\Theta}_j|\boldsymbol{G})$ for $\boldsymbol{\Theta}_j$ is expressed as

$$\min_{\boldsymbol{\Theta}_j} d(\boldsymbol{Y}, \boldsymbol{X}_j\boldsymbol{\Theta}_j|\boldsymbol{G}) = \mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y}\boldsymbol{G}^{-1}\} = (n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{G}^{-1}). \quad (2)$$

The minimum value in (2) expresses a measurement about the goodness of fit for model M_j. Using (2) in the candidate model M_j, the following class of variable selection criteria is considered:

$$\mathcal{L}(j|\alpha, \boldsymbol{G}) = (n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{G}^{-1}) + \alpha p k_j, \quad (3)$$

where α is a positive constant which expresses the complexity of the model M_j. It is straightforward that (3) with $\alpha = 2$ and $\boldsymbol{G} = \boldsymbol{S}_\omega$ is the C_p criterion proposed by Sparks *et al.* (1983) when $n > p$. Moreover, (3) with $\boldsymbol{G} = \boldsymbol{S}_\omega$ is the GC_p criterion proposed by Nagai *et al.* (2012). However, the GC_p criterion cannot be defined when $p > n$. Therefore, we consider two criteria obtained by substituting one of two specific weighted matrices instead of $\boldsymbol{S}_\omega$ into $\boldsymbol{G}$ in (3). By substituting the scalar matrix $p^{-1}\mathrm{tr}(\boldsymbol{S}_\omega)\boldsymbol{I}_p$ into $\boldsymbol{G}$, we define the scalar-type generalized C_p (SGC_p) criterion as follows:

$$SGC_p(j|\alpha) = p^{-1}\mathcal{L}(j|\alpha, p^{-1}\mathrm{tr}(\boldsymbol{S}_\omega)\boldsymbol{I}_p) = (n-k_j)\frac{\mathrm{tr}(\boldsymbol{S}_j)}{\mathrm{tr}(\boldsymbol{S}_\omega)} + \alpha k_j. \quad (4)$$

Note that the $SGC_p(j|\alpha)$ criterion is obtained by dividing $\mathcal{L}(j|\alpha, p^{-1}\mathrm{tr}(\boldsymbol{S}_\omega)\boldsymbol{I}_p)$ by p because the divided p is redundant for variable selection. The SGC_p criterion with $\alpha = 2$ is essentially the same as the PE criterion proposed by Fujikoshi *et al.* (2011). Moreover, the value $\mathrm{tr}(\boldsymbol{S}_j)/\mathrm{tr}(\boldsymbol{S}_\omega)$ in (4) corresponds to the MANOVA test statistic in Fujikoshi *et al.* (2004). They applied the Dempster trace criterion when $p > n$ for tests about one and two sample mean vectors in Dempster (1958; 1960). Note that there is no inverse of the sample covariance matrix in the SGC_p criterion. Thus, this criterion is calculable even when $p > n$. Let $\boldsymbol{S}_\lambda$ be the ridge-type sample covariance matrix, which is defined by

$$\boldsymbol{S}_\lambda = \boldsymbol{S}_\omega + \frac{\mathrm{tr}(\boldsymbol{S}_\omega)}{\lambda}\boldsymbol{I}_p,$$

where λ is a positive ridge parameter. Then, by substituting $\boldsymbol{S}_\lambda$ into $\boldsymbol{G}$, we define the ridge-type generalized C_p (RGC_p) criterion as follows:

$$RGC_p(j|\alpha, \lambda) = \mathcal{L}(j|\alpha, \boldsymbol{S}_\lambda) = (n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\lambda^{-1}) + \alpha p k_j. \quad (5)$$

The first term in (5) is similar to that of the ridge-type C_p criterion used by Kubokawa and Srivastava (2012). If $\boldsymbol{S}_\omega$ is invertible and $\lambda = \infty$, then (5) coincides with the GC_p criterion. However, $\boldsymbol{S}_\omega$ is singular when $p > n$. The scalar matrix $\lambda^{-1}\mathrm{tr}(\boldsymbol{S}_\omega)\boldsymbol{I}_p$ keeps $\boldsymbol{S}_\lambda$ invertible even in such case. The best subsets are given by minimizing the SGC_p criterion and RGC_p criterion, i.e., defined by

$$\hat{j}_S = \arg\min_{j\in\mathcal{J}} SGC_p(j|\alpha),\ \hat{j}_R = \arg\min_{j\in\mathcal{J}} RGC_p(j|\alpha,\lambda), \tag{6}$$

where $\mathcal{J}$ is a family of subsets of ω denoted by $\mathcal{J} = \{j_1, \ldots, j_K\}$ and K is the number of candidate subsets.

2.2. Assumptions for Consistency. We prepare assumptions for consistency. To describe several classes of j that express the column indexes of $\boldsymbol{X}$ in the candidate model (1), we separate $\mathcal{J}$ into two sets, one is a family of overspecified subsets that includes the true subset, i.e., $\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subset j\}$, and the other is a family of underspecified subsets that are not overspecified subsets, i.e., $\mathcal{J}_- = \mathcal{J}_+^c \cap \mathcal{J}$. Let a $p \times p$ non-centrality matrix and parameter be expressed by

$$\boldsymbol{\Delta}_j = \boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*,\ \delta_j^2 = \mathrm{tr}(\boldsymbol{\Delta}_j). \tag{7}$$

It should be noted that $\boldsymbol{\Delta}_j = \boldsymbol{O}_{p,p}$ and $\delta_j^2 = 0$ hold from properties of projection matrices if and only if $j \in \mathcal{J}_+$, where $\boldsymbol{O}_{p,p}$ is the $p \times p$ matrix with all elements as zero. Then, we prepare the following assumptions for consistency:

A1. The true subset j_* is included in $\mathcal{J}$, i.e., $j_* \in \mathcal{J}$.

A2. $\limsup_{p\to\infty} \frac{1}{p}\mathrm{tr}(\boldsymbol{\Sigma}_*) < \infty$.

A3. $\limsup_{p\to\infty} \frac{\kappa_4}{\mathrm{tr}(\boldsymbol{\Sigma}_*)^2} < \infty$, where $\kappa_4 = E[||\boldsymbol{\varepsilon}||^4] - \mathrm{tr}(\boldsymbol{\Sigma}_*)^2 - 2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)$.

A4. For every $j \in \mathcal{J}_-$, there exists $\ell \in j_* \cap j^c$ such that

$$\liminf_{n\to\infty} \frac{1}{n}\boldsymbol{x}_\ell'(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell})\boldsymbol{x}_\ell > 0,\ \liminf_{p\to\infty} \frac{1}{p}||\boldsymbol{\theta}_\ell||^2 > 0,$$

where $\omega_\ell = \{\ell\}^c$, and $\boldsymbol{x}_\ell$ and $\boldsymbol{\theta}_\ell$ are the ℓ-th column vectors of $\boldsymbol{X}_*$ and $\boldsymbol{\Theta}_*'$, respectively.

Assumption A1 is needed to consider consistency. From the definition of $\mathcal{J}_+$, the true subset j_* can be regarded as the smallest overspecified subset. Assumption A2 is a regularity assumption for the true covariance matrix $\boldsymbol{\Sigma}_*$. If the number of response variables whose variances are $O(p)$ is finite and the variances of the other response variables are $O(1)$, assumption A2 holds. Assumption A3 is the restriction for the fourth moment of $\boldsymbol{\varepsilon}$. From properties of the multivariate normal distribution (e.g., Magnus and Neudecker, 1979; Himeno and Yamada, 2014), $\kappa_4 = 0$ when $\boldsymbol{\varepsilon}$ is distributed according to the multivariate normal distribution. Moreover, some specific multivariate distributions such as the multivariate t-distribution or the multivariate contaminated normal distribution are satisfied with assumption A3. Assumption A4 concerns explanatory variables and true regression coefficients. In terms of explanatory variables, this means that a sample

covariance of residuals in the linear regression of $\boldsymbol{x}_\ell$ with the remaining $\boldsymbol{X}_{\omega_\ell}$ does not converge to 0. It is straightforward to show that this is weaker than assuming $\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) > 0$, where $\lambda_{\min}(\boldsymbol{A})$ is the minimum eigenvalue of a square matrix $\boldsymbol{A}$. The assumption for true regression coefficients is essentially used in Katayama and Imori (2014). For example, when all the elements of each $\boldsymbol{\theta}_\ell$ are non-zero constants not converging to 0, the assumption for true regression coefficients holds. Moreover, even when half of the elements of $\boldsymbol{\theta}_\ell$ are zeros and the remaining half are non-zero constants not converging to 0, the assumption is satisfied. Hence, the assumption for the true regression coefficients will be not unrealistic. Further, if p diverges as fast as n, i.e., $c \in [0, \infty)$ in the HUHD asymptotic framework, the assumption for true regression coefficients can become weaker such as $\liminf_{p\to\infty} q_p^{-1}||\boldsymbol{\theta}_\ell||^2 > 0$ for some $q_p \to \infty$ $(p \to \infty)$. Note that assumption A4 does not always have to hold for every $\ell \in j_*$. For example, if $\mathcal{J}$ is a set of nested subsets, i.e., $\mathcal{J} = \{\{1\}, \ldots, \{1, \ldots, k\}\}$, then assumption A4 needs to hold only for $\ell = k_*$. If assumption A4 is supported, for every $j \in \mathcal{J}_-$, the following inequality holds (the proof is given in Appendix A):

$$\inf_{n>k, p\geq 1} \frac{1}{np}\lambda_{\max}(\boldsymbol{\Delta}_j) > 0, \tag{8}$$

where $\lambda_{\max}(\boldsymbol{A})$ is the maximum eigenvalue of a square matrix $\boldsymbol{A}$.

Furthermore, we consider the following assumption that is regarded as a special case of assumption A3:

A3′. $\lim_{p\to\infty} \dfrac{\xi^2}{\mathrm{tr}(\boldsymbol{\Sigma}_*)^2} = 0$, where $\xi^2 = \max\{\kappa_4, \mathrm{tr}(\boldsymbol{\Sigma}_*^2)\}$.

Assumption A3′ is used under the UHD asymptotic framework, and this assumption is stronger than assumption A3. For example, assumption A3′ is satisfied if the following conditions hold:

$$\begin{aligned}
&\lim_{p\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_*)^2} = 0,\ \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{u},\ \boldsymbol{u} = (u_1, \ldots, u_p)',\\
&E[u_a] = 0,\ E[u_a^4] \leq r_u\ (a = 1, \ldots, p),\\
&E[u_a^2 u_b^2] = 1\ (a \neq b),\ E[u_a u_b u_c u_d] = 0\ (a \neq b, c, d),
\end{aligned} \tag{9}$$

where r_u is a positive constant not dependent on p. When $\boldsymbol{\varepsilon} = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{u}$, κ_4 is calculated as follows:

$$\kappa_4 = \sum_{a=1}^{p}\{(\boldsymbol{\Sigma}_*)_{aa}\}^2\left(E[u_a^4] - 3\right) \leq |r_u - 3|\mathrm{tr}(\boldsymbol{\Sigma}_*^2),$$

where $(\boldsymbol{A})_{ab}$ expresses the (a, b)-th element of a matrix $\boldsymbol{A}$. The condition about the true covariance matrix $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_*^2)/\mathrm{tr}(\boldsymbol{\Sigma}_*)^2 = 0$ is called the sphericity condition, and it is often used for $p \gg n$ setting (e.g., Aoshima *et al.*, 2018).

3. Main Results

3.1. Conditions for Consistency of the SGC_p Criterion. We obtain conditions for consistency of the SGC_p criterion (4). Recall that the best subset chosen by minimizing the SGC_p criterion is defined by (6). Then, the SGC_p criterion is consistent if $P(\hat{j}_S = j_*) \to 1$. The probability $P(\hat{j}_S = j_*)$ can be expressed as

$$P(\hat{j}_S = j_*) = P\left(\cap_{j \in \mathcal{J} \cap \{j_*\}^c} \{SGC_p(j|\alpha) > SGC_p(j_*|\alpha)\}\right).$$

We separate $\mathcal{J} \cap \{j_*\}^c$ into $\mathcal{J}_+ \cap \{j_*\}^c$ and $\mathcal{J}_-$ because the non-centrality matrix $\boldsymbol{\Delta}_j$ in (7) behaves differently for each of the cases of $j \in \mathcal{J}_+ \cap \{j_*\}^c$ and $j \in \mathcal{J}_-$. From this and the subadditivity of a measure, a lower bound of $P(\hat{j}_S = j_*)$ is written as

$$P(\hat{j}_S = j_*) \geq 1 - \overline{P}_S - \underline{P}_S,$$

where $\overline{P}_S$ and $\underline{P}_S$ are defined by

$$\overline{P}_S = P\left(\cup_{j \in \mathcal{J}_+ \cap \{j_*\}^c} \{SGC_p(j|\alpha) \leq SGC_p(j_*|\alpha)\}\right), \tag{10}$$

$$\underline{P}_S = P\left(\cup_{j \in \mathcal{J}_-} \{SGC_p(j|\alpha) \leq SGC_p(j_*|\alpha)\}\right). \tag{11}$$

To obtain conditions for consistency of the SGC_p criterion, we consider conditions such that $\overline{P}_S$ and $\underline{P}_S$ converge to 0. First, we prepare the results about the orders of several probabilities. For subsets $j, h \subset \omega$, let $\boldsymbol{W}$, $\boldsymbol{U}_j$, and $\boldsymbol{V}_{j,h}$ be random matrices defined by

$$\boldsymbol{W} = \boldsymbol{\mathcal{E}}_*'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{\mathcal{E}}_*,\ \boldsymbol{U}_j = \boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}_*,\ \boldsymbol{V}_{j,h} = \boldsymbol{\mathcal{E}}_*'(\boldsymbol{P}_j - \boldsymbol{P}_h)\boldsymbol{\mathcal{E}}_*. \tag{12}$$

Then, we derive the following lemma about the orders of the tail probabilities for functions of (12) (the proof is given in Appendix B).

LEMMA 1. *Let $\boldsymbol{W}$, $\boldsymbol{U}_j$, and $\boldsymbol{V}_{j,h}$ be given by* (12)*, and let $r_1 > 0$, $r_2 > 0$, $r_3 < 0$, $r_4 > 0$, $r_5 > 0$, and $r_6 > 0$. Then, under the HUHD asymptotic framework, the following results hold:*

(i) *If $r_1 > \mathrm{tr}(\boldsymbol{\Sigma}_*)$ and $r_2 < \mathrm{tr}(\boldsymbol{\Sigma}_*)$, then we have*

$$P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \geq r_1\right) = O\left(\xi^2 n^{-1}\{r_1 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2}\right),$$

$$P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \leq r_2\right) = O\left(\xi^2 n^{-1}\{\mathrm{tr}(\boldsymbol{\Sigma}_*) - r_2\}^{-2}\right),$$

where ξ^2 is given in assumption A3′.

(ii) *For $j \not\supseteq j_*$, we have*

$$P\left(\mathrm{tr}(\boldsymbol{U}_j) \leq r_3\right) = O\left(\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}_j)|r_3|^{-2}\right),$$

where $\boldsymbol{\Delta}_j$ is defined by (7)*.*

(iii) *For $j \supsetneq h$, if $r_4 > \mathrm{tr}(\boldsymbol{\Sigma}_*)$, then we have*

$$P\left(\mathrm{tr}(\boldsymbol{V}_{j,h}) \geq (k_j - k_h)r_4\right) = O\left(\xi^2\{r_4 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2}\right).$$

(iv) *For $j \supsetneq h$, if $r_6/r_5 \to 0$, then we have*

$$P\left(\mathrm{tr}(\boldsymbol{V}_{j,h}) - (k_j - k_h)\mathrm{tr}(\boldsymbol{\Sigma}_*) + r_5 \leq r_6\right) = O(\xi^2 r_5^{-2}).$$

By using Lemma 1, we give the orders of $\overline{P}_S$ and $\underline{P}_S$ (the proof is given in Appendix C).

LEMMA 2. *Suppose that assumptions A1, A2, and A4 hold, and for some constants τ_S satisfying $0 < \tau_S < 1$, the following equations hold:*

$$\lim_{n\to\infty, p/n\to c} \alpha\tau_S > 1, \quad \lim_{n\to\infty, p/n\to c} n^{-1}\alpha = 0, \tag{13}$$

under the HUHD asymptotic framework. Then, the orders of $\overline{P}_S$ and $\underline{P}_S$ defined in (10) and (11) are given by

$$\begin{aligned}
\overline{P}_S &= O\left(\xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2} \max\left\{(\alpha\tau_S - 1)^{-2}, n^{-1}(1-\tau_S)^{-2}\right\}\right), \\
\underline{P}_S &= O\left(\xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2} \max\left\{(\alpha\tau_S - 1)^{-2}, n^{-1}(1-\tau_S)^{-2}\right\}\right) \\
&\quad + O\left(\max\left\{\xi^2 n^{-2}p^{-2}, \xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2} n^{-1}, \lambda_{\max}(\mathbf{\Sigma}_*) n^{-1} p^{-1}\right\}\right),
\end{aligned}$$

where ξ^2 is defined in assumption A3′.

Next, we obtain conditions for consistency of the SGC_p criterion (4). Note that the results in Lemma 2 are derived without assumptions A3 and A3′. We use assumption A3 or A3′ to obtain consistency conditions, although the UHD asymptotic framework is used when assumption A3′ is supported. It is straightforward that $\limsup_{p\to\infty} \xi \mathrm{tr}(\mathbf{\Sigma}_*)^{-1} < \infty$ holds under assumption A3, but $\lim_{p\to\infty} \xi \mathrm{tr}(\mathbf{\Sigma}_*)^{-1} = 0$ holds under assumption A3′. By using this fact and Lemma 2, we obtain consistency conditions about α (the proof is given in Appendix D).

THEOREM 1. *Suppose that assumptions A1, A2, A3, and A4 hold. Then, the SGC_p criterion is consistent under the HUHD asymptotic framework if the following conditions are satisfied:*

$$\lim_{n\to\infty, p/n\to c} \alpha = \infty, \quad \lim_{n\to\infty, p/n\to c} \frac{\alpha}{n} = 0. \tag{14}$$

Furthermore, when replacing assumption A3 with assumption A3′, the SGC_p criterion is consistent under the UHD asymptotic framework if the following conditions are satisfied:

$$\lim_{(n,p)\to\infty, p/n\to c} \alpha > 1, \quad \lim_{(n,p)\to\infty, p/n\to c} \frac{\alpha}{n} = 0. \tag{15}$$

From Theorem 1, if assumption A3′ is supported, the SGC_p criterion is consistent under the UHD asymptotic framework even when α is a constant not dependent on n and p such as $\alpha = 2$. When assumption A3′ is not supported but assumption A3 is, α should diverge to render the SGC_p criterion consistent. Moreover, if (14) holds, then (15) holds. It is difficult to verify whether assumption A3′ holds using empirical data. Hence, we recommend that (14) be used to render the SGC_p criterion consistent by deciding α. On the other hand, we also obtain conditions for inconsistency (the proof is given in Appendix E).

Theorem 2. *Suppose that assumptions A1, A2, A3, and A4 hold. Let conditions of α under the HUHD asymptotic framework be as follows:*

C1. $\lim_{n\to\infty,p/n\to c}\alpha<1$ *and there exists* $j\in\mathcal{J}_+\cap\{j_*\}^c$ *such that*

$$\lim_{n\to\infty,p/n\to c}\frac{\kappa_4 I(\kappa_4>0)+2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{(1-\alpha)^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^2}<k_j-k_*, \tag{16}$$

where $I(\kappa_4>0)$ is an indicator function, i.e., if $\kappa_4>0$ then $I(\kappa_4>0)=1$, otherwise $I(\kappa_4>0)=0$.

C2. *There exists* $j\subsetneq j_*$ *such that*

$$\lim_{n\to\infty,p/n\to c}\frac{\alpha\mathrm{tr}(\boldsymbol{\Sigma}_*)}{\delta_j^2}>(k_*-k_j)^{-1}.$$

Then, if either of the conditions C1 or C2 is satisfied, the SGC_p criterion is inconsistent, i.e., $\lim_{n\to\infty,p/n\to c}P(\hat{j}_S=j_*)<1$ holds under the HUHD asymptotic framework. Furthermore, when replacing assumption A3 with assumption A3′, (16) and $\lim_{(n,p)\to\infty,p/n\to c}P(\hat{j}_S=j_*)=0$ always hold under the UHD asymptotic framework if $\lim_{(n,p)\to\infty,p/n\to c}\alpha<1$.

We observe that the SGC_p criterion is inconsistent when α is too small from condition C1 or too large from condition C2. Although we cannot cover all the consistency or inconsistency conditions of α from only Theorems 1 and 2, these theorems nevertheless provide much information about the consistency or inconsistency of the SGC_p criterion.

3.2. Conditions for Consistency of the RGC_p Criterion. We obtain conditions for consistency of the RGC_p criterion (5). In the same way as subsection 3.1, a lower bound of $P(\hat{j}_R=j_*)$ is written as

$$P(\hat{j}_R=j_*)\geq 1-\overline{P}_R-\underline{P}_R,$$

where $\overline{P}_R$ and $\underline{P}_R$ are given by

$$\overline{P}_R=P\left(\cup_{j\in\mathcal{J}_+\cap\{j_*\}^c}\{RGC_p(j|\alpha,\lambda)\leq RGC_p(j_*|\alpha,\lambda)\}\right), \tag{17}$$

$$\underline{P}_R=P\left(\cup_{j\in\mathcal{J}_-}\{RGC_p(j|\alpha,\lambda)\leq RGC_p(j_*|\alpha,\lambda)\}\right). \tag{18}$$

First, we obtain the orders of $\overline{P}_R$ and $\underline{P}_R$. Then, we examine the orders by using moments of a statistic. It is difficult to calculate the moments of $\boldsymbol{a}'\boldsymbol{S}_\lambda^{-1}\boldsymbol{a}$ because of the existence of the inverse matrix of $\boldsymbol{S}_\lambda$, where $\boldsymbol{a}$ is a p-dimensional vector. Therefore, we do not evaluate $\boldsymbol{a}'\boldsymbol{S}_\lambda^{-1}\boldsymbol{a}$ directly, but evaluate the following lower and upper bounds:

$$||\boldsymbol{a}||^2\lambda_{\min}(\boldsymbol{S}_\lambda^{-1})\leq\boldsymbol{a}'\boldsymbol{S}_\lambda^{-1}\boldsymbol{a}\leq||\boldsymbol{a}||^2\lambda_{\max}(\boldsymbol{S}_\lambda^{-1}). \tag{19}$$

By using (19) and Lemma 1, we give the orders of $\overline{P}_R$ and $\underline{P}_R$ (the proof is given in Appendix F).

LEMMA 3. *Suppose that assumptions A1, A2, and A4 hold, and for some constants τ_R satisfying $0 < \tau_R < 1$ the following equations hold:*

$$\lim_{n\to\infty,p/n\to c} \lambda^{-1}p\alpha\tau_R > 1, \quad \lim_{n\to\infty,p/n\to c} n^{-1}(1+\lambda^{-1})p\alpha = 0,$$

under the HUHD asymptotic framework. Then, the orders of $\overline{P}_R$ and $\underline{P}_R$ defined in (17) and (18) are given by

$$\begin{aligned}\overline{P}_R &= O\left(\xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2} \max\left\{(\lambda^{-1}p\alpha\tau_R - 1)^{-2}, n^{-1}(1-\tau_R)^{-2}\right\}\right), \\ \underline{P}_R &= O\left(\xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2} \max\left\{(\lambda^{-1}p\alpha\tau_R - 1)^{-2}, n^{-1}(1-\tau_R)^{-2}\right\}\right) \\ &\quad + O\left(\max\left\{\xi^2 n^{-2}p^{-2}, \xi^2 \mathrm{tr}(\mathbf{\Sigma}_*)^{-2}n^{-1}, \lambda_{\max}(\mathbf{\Sigma}_*)n^{-1}p^{-1}\right\}\right),\end{aligned}$$

where ξ^2 is defined in assumption A3′.

By using Lemma 3, we obtain consistency conditions of the RGC_p criterion. Since the RGC_p criterion has the two parameters α and λ, the conditions are connected with α and λ.

THEOREM 3. *Suppose that assumptions A1, A2, A3, and A4 hold. Then, the RGC_p criterion is consistent under the HUHD asymptotic framework if the following conditions are satisfied:*

$$\lim_{n\to\infty,p/n\to c} \frac{p\alpha}{\lambda} = \infty, \quad \lim_{n\to\infty,p/n\to c} \frac{(1+\lambda^{-1})p\alpha}{n} = 0. \tag{20}$$

Furthermore, when replacing assumption A3 with assumption A3′, the RGC_p criterion is consistent under the UHD asymptotic framework if the following conditions are satisfied:

$$\lim_{(n,p)\to\infty,p/n\to c} \frac{p\alpha}{\lambda} > 1, \quad \lim_{(n,p)\to\infty,p/n\to c} \frac{(1+\lambda^{-1})p\alpha}{n} = 0. \tag{21}$$

The proof of Theorem 3 is omitted because the theorem can be proved in the same way as Theorem 1. From Theorem 3, if we set $\lambda = 1$ and $\alpha = \tilde{\alpha}/p$ ($\tilde{\alpha} > 0$), conditions (20) and (21) are the same as (14) and (15), respectively. Note that conditions (20) and (21) may be strong because they are derived using inequality (19). From Theorem 3, we observe that the larger λ becomes, the larger α should be, to satisfy conditions (20) and (21). Furthermore, we also obtain conditions for inconsistency (the proof is given in Appendix G).

THEOREM 4. *Suppose that assumptions A1, A2, A3, and A4 hold. Let conditions of α under the HUHD asymptotic framework be as follows:*

C3. $\lim_{n\to\infty,p/n\to c}(1+\lambda^{-1})p\alpha < 1$ *and there exists $j \in \mathcal{J}_+ \cap \{j_*\}^c$ such that*

$$\lim_{n\to\infty,p/n\to c} \frac{\kappa_4 I(\kappa_4 > 0) + 2\mathrm{tr}(\mathbf{\Sigma}_*^2)}{\{1-(1+\lambda^{-1})p\alpha\}^2 \mathrm{tr}(\mathbf{\Sigma}_*)^2} < k_j - k_*. \tag{22}$$

Table 1. Assumptions and asymptotic behaviors of n and p to ensure consistency of six criteria.

Criterion	Assumptions	Asymptotic behavior
1	A1, A2, A3′, A4	$p \to \infty$
2	A1, A2, A3, A4	free
3	A1, A2, A3, A4	$\log\log p / \log n \to 0$
4	A1, A2, A3′, A4	$p \to \infty$
5	A1, A2, A3, A4	free
6	A1, A2, A3, A4	$\log\log p / \log n \to 0$

C4. *There exists* $j \subsetneq j_*$ *such that*

$$\lim_{n\to\infty, p/n\to c} \frac{p\alpha \mathrm{tr}(\mathbf{\Sigma}_*)}{\lambda \delta_j^2} > (k_* - k_j)^{-1}.$$

Then, if either of the conditions C3 or C4 is satisfied, the RGC_p *criterion is inconsistent, i.e.,* $\lim_{n\to\infty,p/n\to c} P(\hat{j}_R = j_*) < 1$ *holds under the HUHD asymptotic framework. Furthermore, when replacing assumption A3 with assumption A3′,* (22) *and* $\lim_{(n,p)\to\infty,p/n\to c} P(\hat{j}_R = j_*) = 0$ *always hold under the UHD asymptotic framework if* $\lim_{(n,p)\to\infty,p/n\to c}(1+\lambda^{-1})p\alpha < 1$.

From Theorem 4, we observe that λ should be large in order not to satisfy conditions C3 and C4. However, if λ is large, $p\alpha\lambda^{-1}$ in (20) and (21) is small and then the condition of α to have consistency becomes restricted.

4. Numerical Experiments

4.1. Criteria for Numerical Experiments. To conduct numerical experiments, we use the following six criteria:

Criterion 1: the SGC_p criterion with $\alpha = 2$.
Criterion 2: the SGC_p criterion with $\alpha = \log n$.
Criterion 3: the SGC_p criterion with $\alpha = (\log n / \log\log p)^{1/2}$.
Criterion 4: the RGC_p criterion with $\alpha = 2p^{-1}$ and $\lambda = 1$.
Criterion 5: the RGC_p criterion with $\alpha = p^{-1}\log n$ and $\lambda = 1$.
Criterion 6: the RGC_p criterion with $\alpha = p^{-1}(n\log n / \log\log p)^{1/2}$ and $\lambda = n^{1/2}$.

Table 1 shows the assumptions and asymptotic behaviors of n and p to ensure the consistency of the above six criteria. We observe that to ensure consistency, p has to diverge for criteria 1 and 4, but p does not have to diverge for criteria 2, 3, 5, and 6. Further, criteria 3 and 6 are consistent when $\log\log p / \log n \to 0$. Since this slightly restricts the behavior of p, it may not be suitable where p increases dramatically. However, such a case is unrealistic, so this behavior is reasonable for empirical contexts. Note that the penalty terms $k_j\alpha$ or $k_j p\alpha$ in criteria 1, 2, 4, and 5 do not include p, but those in criteria 3 and 6 do.

For comparison, we also consider criteria in Katayama and Imori (2014) given by

$$\mathrm{HGIC}(j) = p + \log\left|(1 - k_j/n)\,\boldsymbol{D}_{\boldsymbol{S}_j}\right| + \beta p k_j,$$

where $\boldsymbol{D}_{\boldsymbol{S}_j} = \mathrm{diag}\{(\boldsymbol{S}_j)_{11}, \ldots, (\boldsymbol{S}_j)_{pp}\}$ and $\mathrm{diag}\{(\boldsymbol{A})_{11}, \ldots, (\boldsymbol{A})_{pp}\}$ is the diagonal matrix with diagonal elements corresponding to those of a $p \times p$ matrix $\boldsymbol{A}$. Especially, we use the following three HGICs from their paper:

Criterion 7: the HGIC with $\beta = n^{-1}(\log p)(\log\log p)^{1/2}$.
Criterion 8: the HGIC with $\beta = n^{-1}(\log p)(\log\log p)$.
Criterion 9: the HGIC with $\beta = n^{-1}(\log p)(\log\log p)^{3/2}$.

From Katayama and Imori (2014), criteria 7, 8, and 9 are consistent under several assumptions such as normality when $p \to \infty$ and $\log p/n \to 0$ for our numerical studies.

4.2. Simulations. We verify the foregoing exposition by simulations. The probabilities of selecting the true subset j_* were evaluated by Monte Carlo simulations with 10,000 iterations. Ten subsets $j_m = \{1, \ldots, m\}$ $(m = 1, \ldots, 10)$, with several different values of n and p, were prepared for these simulations. We generated the explanatory matrix $\boldsymbol{X}$ as follows. We independently generated $s_1, \ldots, s_n$ from $U(-1, 1)$, where $U(a, b)$ denotes a uniform distribution on the range (a, b). Using $s_1, \ldots, s_n$, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$, where the (a, b)-th element is defined by s_a^{b-1} $(a = 1, \ldots, n;\ b = 1, \ldots, k)$. The true subset was determined by $j_* = \{1, 2, 3, 4, 5\}$. The true coefficient matrix $\boldsymbol{\Theta}_*$ adhered to the following structure:

$$\boldsymbol{\Theta}_* = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_*})',\ \boldsymbol{\theta}_a = \begin{cases} \left(a(-1)^{a-1}\mathbf{1}'_{\lfloor p/2 \rfloor}, \mathbf{0}'_{\lceil p/2 \rceil}\right)' & (a : \text{odd}) \\ \left(\mathbf{0}'_{\lfloor p/2 \rfloor}, a(-1)^{a-1}\mathbf{1}'_{\lceil p/2 \rceil}\right)' & (a : \text{even}) \end{cases},$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling functions, respectively. For these numerical simulations, we expressed $\boldsymbol{\mathcal{E}}_*$ as $\boldsymbol{Z}_*\boldsymbol{\Sigma}_*^{1/2}$, where $\boldsymbol{Z}_* = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are independent and identically distributed from $\boldsymbol{z} = (z_1, \ldots, z_p)'$ with mean $\mathbf{0}_p$ and covariance matrix $\boldsymbol{I}_p$. Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)'$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)' \sim$ *i.i.d.* $N_p(\mathbf{0}_p, \boldsymbol{I}_p)$, and $\tau \sim \chi^2(10)$ be mutually independent random vectors and variable. Then, $\boldsymbol{z}$ is generated from the following four distributions:

(D1) multivariate normal distribution: $\boldsymbol{z} = \boldsymbol{\nu}$.
(D2) multivariate t-distribution with 10 degrees of freedom: $\boldsymbol{z} = (8/\tau)^{1/2}\boldsymbol{\nu}$.
(D3) independent skew-normal distribution with shape parameter 10:

$$z_a = \left(1 - \frac{2}{\pi}\eta^2\right)^{-1/2}\left(\frac{\nu_a}{\sqrt{1 + 10^2}} + \eta\,|\zeta_a| - \sqrt{\frac{2}{\pi}}\eta\right)\ (a = 1, \ldots, p),$$

where $\eta = 10/\sqrt{1 + 10^2}$.

(D4) independent log-normal distribution:

$$z_a = \frac{\exp(\nu_a) - \sqrt{e}}{\sqrt{e(e-1)}} \ (a = 1, \ldots, p).$$

Note that distributions (D1)-(D4) are satisfied with $\kappa_4 = O(\mathrm{tr}(\boldsymbol{\Sigma}_*^2))$. The true covariance matrix $\boldsymbol{\Sigma}_*$ was set as the following two structures:

(S1) exchangeable structure with correlation 0.8:

$$\boldsymbol{\Sigma}_* = (1 - 0.8)\boldsymbol{I}_p + 0.8\mathbf{1}_p\mathbf{1}_p'.$$

(S2) autoregressive structure with correlation 0.8: $(\boldsymbol{\Sigma}_*)_{ab} = (0.8)^{|a-b|}$.

Note that assumption A3′ is not satisfied when the true covariance matrix $\boldsymbol{\Sigma}_*$ is (S1), but assumption A3′ is satisfied when the true covariance matrix $\boldsymbol{\Sigma}_*$ is (S2) under distributions (D1)-(D4). Under these settings, we used the 8 combinations of the four distributions and the two true covariance matrices (S1) and (S2). Tables 2-9 show the probabilities of selecting the true subset j_* using each of the nine criteria. In each table, the probabilities of selecting the true subset j_* were evaluated for distributions (D1)-(D4) and the two covariance matrices (S1) and (S2). When the true covariance matrix $\boldsymbol{\Sigma}_*$ has an exchangeable structure, i.e., in Tables 2, 4, 6, and 8, it appears that criteria 2, 5, and 6 are consistent for both cases where only n is large and where n and p are large, but criteria 1 and 4 are not consistent. This is because assumption A3 is satisfied for the cases of (S1) and distributions (D1)-(D4), but assumption A3′ is not satisfied for such cases. Moreover, although criterion 3 is consistent from Table 1, it looks inconsistent in Tables 2, 4, 6, and 8. This is because the penalty term in criterion 3 is smaller than that in criterion 1 for our numerical simulations. On the other hand, when the true covariance matrix $\boldsymbol{\Sigma}_*$ has an autoregressive structure, i.e., in Tables 3, 5, 7, and 9, we observe that criteria 1 and 4 also are consistent except for the case that only n is large because (S2) is satisfied with $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_*^2)/\mathrm{tr}(\boldsymbol{\Sigma}_*)^2 = 0$, so assumption A3′ is satisfied for the cases of (S2) and distributions (D1)-(D4). This result accords with Theorem 1 and Theorem 3. In Tables 2-9, criteria 7, 8, and 9 are consistent when n and p are large, but they are not consistent when only n is large. Further, we observe that the probabilities by criteria 7, 8, and 9 are low when $p/n = 10$ and $n \leq 100$. In sum, the probabilities by criterion 6 are the highest across Tables 2-9.

4.3. Empirical Examples. First, we verify the probabilities of selecting the true subsets by using real data. The dataset pertains to 8 groups ($g = 1, \ldots, 8$) of black cotton fibers dyed by Indigo and its derivative dyes. Each cotton fiber has 55 samples, and each sample has 541 variables, which are the absorbances for wavelengths from 240 nm to 780 nm in steps of 1 nm. Let the explanatory matrix be denoted as $\boldsymbol{X} = (\boldsymbol{T}, \mathbf{1}_9) \otimes \mathbf{1}_{25}$, where $\boldsymbol{T} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_8)$ and $\boldsymbol{e}_a$ $(a = 1, \ldots, 8)$ is a 9-dimensional vector such that the $(a+1)$-th element is one and the other elements are zeros, and the symbol $\otimes$ denotes the Kronecker product (see, e.g., Harville, 1997). Here, the 9-th column vector of $\boldsymbol{X}$ expresses the intercept term.

Table 2. True subset selection probabilities (%) for distribution (D1) and covariance matrix (S1).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	21.63	14.98	22.55	17.16	8.08	8.47	20.61	20.16	19.07
50	10	60.36	40.23	59.66	66.62	24.93	33.85	59.03	58.01	55.66
100	10	76.52	77.66	82.75	93.46	66.19	92.64	75.95	71.39	66.84
300	10	76.85	98.84	87.04	94.07	99.94	100.00	78.37	74.04	69.62
500	10	77.93	99.29	89.00	94.35	99.98	100.00	79.48	75.35	70.58
20	10	21.63	14.98	22.55	17.16	8.08	8.47	20.61	20.16	19.07
50	25	61.12	38.26	60.76	67.77	22.35	59.33	45.61	41.91	37.58
100	50	76.81	80.63	72.85	93.73	70.28	99.84	81.69	71.91	59.54
300	150	78.03	98.97	75.24	94.07	99.95	100.00	99.32	99.86	99.71
500	250	79.15	99.32	76.87	94.72	99.98	100.00	99.65	99.92	99.99
20	20	22.29	15.53	23.61	17.72	8.98	13.70	17.20	16.54	15.47
50	50	62.23	40.07	61.01	69.52	24.00	71.87	33.67	24.71	17.24
100	100	77.29	79.20	70.82	93.73	69.63	99.93	65.98	49.18	32.14
300	300	78.08	99.12	73.07	94.35	99.91	100.00	99.71	99.75	95.57
500	500	77.61	99.51	74.10	94.49	99.98	100.00	99.92	99.98	99.99
20	200	22.34	15.55	23.73	17.92	8.65	22.15	1.93	0.45	0.05
50	500	62.46	39.86	56.29	69.84	24.57	86.62	5.75	1.10	0.11
100	1000	78.29	79.10	64.59	94.62	69.38	100.00	23.71	6.37	0.71
300	3000	77.91	99.11	68.65	94.40	99.95	100.00	98.79	77.91	27.54
500	5000	78.15	99.37	70.10	94.78	99.96	100.00	100.00	99.97	88.23

Table 3. True subset selection probabilities (%) for distribution (D1) and covariance matrix (S2).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	30.50	14.80	32.68	28.33	10.36	22.29	30.09	28.72	25.72
50	10	82.05	52.56	83.80	91.24	45.42	89.56	78.53	73.66	67.77
100	10	83.71	98.18	89.67	94.43	98.45	99.99	83.28	78.35	72.95
300	10	84.68	99.73	93.09	94.52	99.96	100.00	85.88	82.02	76.97
500	10	84.49	99.85	94.33	95.03	100.00	100.00	86.26	82.18	77.18
20	10	30.50	14.80	32.68	28.33	10.36	22.29	30.09	28.72	25.72
50	25	90.56	52.56	87.53	94.82	47.00	98.20	75.27	65.29	53.06
100	50	97.02	99.78	95.13	98.42	99.74	99.98	99.86	98.53	91.47
300	150	99.84	100.00	99.71	99.88	100.00	100.00	100.00	100.00	100.00
500	250	99.99	100.00	99.96	100.00	100.00	100.00	100.00	100.00	100.00
20	20	36.12	11.95	43.92	32.64	8.51	39.09	19.76	16.49	13.49
50	50	96.34	60.40	91.27	97.75	56.98	99.25	37.88	11.56	1.64
100	100	99.44	99.81	97.78	99.74	99.80	99.98	97.21	74.30	14.13
300	300	99.99	100.00	99.98	99.99	100.00	100.00	100.00	100.00	100.00
500	500	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
20	200	42.48	2.60	78.26	41.12	2.31	79.96	0.00	0.00	0.00
50	500	99.80	63.28	99.88	99.79	62.75	99.95	0.00	0.00	0.00
100	1000	100.00	99.87	100.00	100.00	99.87	100.00	0.77	0.00	0.00
300	3000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.91	1.98
500	5000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99

Table 4. True subset selection probabilities (%) for distribution (D2) and covariance matrix (S1).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	22.29	15.96	22.52	18.23	9.30	10.22	20.60	20.22	19.13
50	10	61.48	40.40	60.74	67.76	24.75	34.53	60.41	58.71	56.19
100	10	77.39	78.92	83.05	93.94	66.97	92.39	76.78	72.65	67.66
300	10	77.70	99.01	87.88	94.55	99.95	100.00	79.01	74.94	70.17
500	10	77.41	99.21	88.80	94.35	99.98	100.00	79.13	75.02	70.73
20	10	22.29	15.96	22.52	18.23	9.30	10.22	20.60	20.22	19.13
50	25	61.17	38.43	60.62	68.15	23.01	59.65	46.28	42.45	38.38
100	50	78.41	78.98	74.38	94.00	69.74	99.83	80.51	71.61	59.57
300	150	78.17	99.06	75.18	94.21	99.96	100.00	99.40	99.88	99.60
500	250	78.43	99.23	76.29	94.37	99.97	100.00	99.61	99.94	99.99
20	20	22.07	15.90	23.70	18.16	9.62	14.41	17.21	16.40	15.53
50	50	62.04	40.12	60.64	69.32	25.68	71.64	33.99	26.04	18.39
100	100	77.57	78.97	71.01	93.83	69.61	99.92	66.47	49.38	31.81
300	300	78.03	99.05	73.13	94.44	99.95	100.00	99.75	99.74	95.35
500	500	77.96	99.43	74.18	94.53	99.98	100.00	99.89	99.99	100.00
20	200	22.95	15.90	24.15	18.60	9.56	22.99	2.07	0.55	0.12
50	500	61.84	40.02	56.49	69.89	24.87	85.74	6.26	1.12	0.09
100	1000	78.47	79.00	64.86	94.29	69.99	99.97	24.41	6.80	0.67
300	3000	78.29	99.01	69.30	94.41	99.96	100.00	98.81	78.31	28.53
500	5000	78.13	99.35	70.35	94.28	99.95	100.00	99.99	99.89	87.79

Table 5. True subset selection probabilities (%) for distribution (D2) and covariance matrix (S2).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	30.11	15.39	31.54	28.33	10.83	23.41	29.59	28.25	26.02
50	10	81.60	52.82	83.98	91.25	45.54	88.72	78.12	73.27	67.12
100	10	83.97	97.60	90.35	94.57	98.05	100.00	83.64	79.17	73.61
300	10	84.61	99.66	93.46	95.28	99.98	100.00	86.06	81.78	77.19
500	10	84.91	99.84	94.50	95.22	100.00	100.00	86.49	82.24	77.50
20	10	30.11	15.39	31.54	28.33	10.83	23.41	29.59	28.25	26.02
50	25	89.73	52.59	86.55	93.67	47.28	97.34	75.13	65.57	53.62
100	50	96.64	99.66	94.42	98.42	99.62	99.97	99.74	98.30	90.77
300	150	99.83	100.00	99.68	99.90	100.00	100.00	100.00	100.00	100.00
500	250	99.99	100.00	99.96	99.99	100.00	100.00	100.00	100.00	100.00
20	20	34.99	12.91	42.79	32.77	9.68	38.90	20.75	17.61	14.52
50	50	95.85	58.85	90.59	97.68	55.56	99.28	40.15	14.80	2.26
100	100	99.14	99.77	97.23	99.53	99.73	99.95	97.24	74.44	18.24
300	300	100.00	100.00	99.96	100.00	100.00	100.00	100.00	100.00	100.00
500	500	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
20	200	43.38	4.80	69.97	41.56	4.42	73.24	0.00	0.00	0.00
50	500	99.67	62.22	98.37	99.66	61.48	99.37	0.00	0.00	0.00
100	1000	100.00	99.78	99.77	100.00	99.78	99.87	2.37	0.00	0.00
300	3000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.76	3.27
500	5000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99

Table 6. True subset selection probabilities (%) for distribution (D3) and covariance matrix (S1).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	21.90	15.89	22.29	17.83	9.05	9.33	21.26	20.80	19.72
50	10	59.15	39.59	58.61	66.40	23.76	33.31	57.89	56.95	54.66
100	10	76.84	79.04	83.15	93.42	67.42	92.36	76.28	71.56	66.69
300	10	78.27	99.16	88.31	94.67	99.95	100.00	79.67	75.24	70.73
500	10	78.11	99.27	89.12	94.63	100.00	100.00	79.95	75.28	70.45
20	10	21.90	15.89	22.29	17.83	9.05	9.33	21.26	20.80	19.72
50	25	60.47	37.59	60.24	66.81	22.21	57.71	44.78	41.00	36.97
100	50	77.58	78.89	73.17	93.82	69.48	99.93	80.24	70.42	58.81
300	150	78.13	99.02	75.21	94.14	99.95	100.00	99.42	99.76	99.73
500	250	78.48	99.29	76.27	94.25	99.98	100.00	99.70	99.88	99.98
20	20	22.79	15.79	24.12	18.15	9.16	13.64	17.69	16.80	15.85
50	50	61.81	39.58	60.21	68.69	24.81	71.49	33.74	25.24	17.51
100	100	76.79	79.34	69.97	93.52	69.42	99.98	65.84	49.07	31.76
300	300	78.34	99.08	73.58	94.53	99.98	100.00	99.84	99.85	95.62
500	500	78.19	99.26	74.54	94.53	99.96	100.00	99.83	99.97	99.99
20	200	21.35	15.30	23.11	17.62	8.74	21.52	1.90	0.37	0.05
50	500	62.10	39.74	56.75	69.79	24.51	86.52	5.73	0.94	0.10
100	1000	77.68	79.05	64.83	93.55	69.59	99.98	23.94	6.41	0.62
300	3000	79.06	99.06	69.29	94.59	99.99	100.00	98.83	77.64	27.51
500	5000	78.27	99.33	70.53	94.64	99.97	100.00	99.98	99.94	88.55

Table 7. True subset selection probabilities (%) for distribution (D3) and covariance matrix (S2).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	30.45	14.76	32.27	28.07	10.16	23.14	30.34	29.17	26.12
50	10	81.52	52.70	83.44	90.82	45.01	90.16	78.40	73.37	67.27
100	10	84.10	98.11	90.46	94.78	98.23	100.00	83.70	78.96	73.43
300	10	84.42	99.71	93.04	94.73	99.99	100.00	85.64	81.46	76.40
500	10	84.96	99.88	94.16	95.04	100.00	100.00	86.56	82.52	77.86
20	10	30.45	14.76	32.27	28.07	10.16	23.14	30.34	29.17	26.12
50	25	91.01	52.23	87.82	95.06	46.94	98.17	76.08	65.60	53.29
100	50	96.60	99.71	94.45	98.18	99.68	99.99	99.79	98.55	91.70
300	150	99.89	100.00	99.69	99.92	100.00	100.00	100.00	100.00	100.00
500	250	100.00	100.00	99.99	100.00	100.00	100.00	100.00	100.00	100.00
20	20	34.51	11.45	42.51	31.57	7.84	37.78	19.87	16.62	13.61
50	50	95.68	60.97	91.13	97.35	57.68	99.19	39.87	12.94	2.02
100	100	99.37	99.71	97.79	99.63	99.69	99.96	97.49	75.85	14.72
300	300	99.99	100.00	99.97	99.99	100.00	100.00	100.00	100.00	100.00
500	500	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
20	200	42.35	2.47	78.67	40.77	2.29	79.88	0.00	0.00	0.00
50	500	99.78	63.15	99.81	99.77	62.60	99.93	0.00	0.00	0.00
100	1000	100.00	99.84	100.00	100.00	99.84	100.00	0.97	0.00	0.00
300	3000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.90	1.69
500	5000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99

Table 8. True subset selection probabilities (%) for distribution (D4) and covariance matrix (S1).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	24.34	18.28	24.92	21.26	12.14	14.48	23.71	22.97	21.98
50	10	60.32	43.80	60.30	67.20	30.36	43.29	60.05	58.69	55.96
100	10	75.85	77.48	81.35	92.46	67.97	88.79	75.24	71.37	66.70
300	10	78.01	98.91	87.99	94.37	99.80	100.00	79.23	74.98	70.45
500	10	77.40	99.47	89.01	94.43	99.95	100.00	79.05	75.17	70.67
20	10	24.34	18.28	24.92	21.26	12.14	14.48	23.71	22.97	21.98
50	25	59.68	40.15	58.88	67.84	26.31	61.64	50.03	46.20	42.21
100	50	76.63	78.70	73.07	93.02	69.83	99.55	81.24	73.02	61.75
300	150	79.18	98.99	76.09	94.54	99.97	100.00	99.32	99.82	99.72
500	250	78.87	99.47	76.67	94.77	99.97	100.00	99.71	99.95	99.98
20	20	23.65	17.89	24.85	20.57	11.35	17.57	20.81	19.84	19.00
50	50	61.52	40.95	60.03	69.55	26.75	71.89	36.77	28.51	20.89
100	100	77.85	77.94	71.18	93.93	68.29	99.85	67.17	51.20	33.10
300	300	78.72	98.95	74.09	94.34	99.99	100.00	99.64	99.82	95.78
500	500	77.95	99.16	74.17	94.37	99.97	100.00	99.82	99.99	100.00
20	200	21.99	16.18	23.77	18.22	9.37	22.62	2.48	0.52	0.09
50	500	62.30	39.45	57.04	69.65	24.20	85.51	6.97	1.42	0.10
100	1000	77.91	79.46	64.73	94.00	70.21	99.98	25.04	6.58	0.55
300	3000	78.44	99.15	68.10	94.53	99.94	100.00	98.87	79.62	29.35
500	5000	79.02	99.36	70.49	94.82	99.96	100.00	99.99	99.91	88.51

Table 9. True subset selection probabilities (%) for distribution (D4) and covariance matrix (S2).

		Criterion								
n	p	1	2	3	4	5	6	7	8	9
20	10	32.63	20.03	33.69	32.97	16.82	30.72	34.48	32.00	28.36
50	10	77.75	57.62	79.44	87.20	52.67	85.82	76.31	71.51	65.27
100	10	83.87	94.52	89.47	94.32	93.98	99.53	83.45	78.79	73.54
300	10	84.60	99.66	92.98	94.95	99.98	100.00	85.69	81.74	76.93
500	10	83.69	99.82	93.65	94.73	100.00	100.00	85.08	81.05	76.19
20	10	32.63	20.03	33.69	32.97	16.82	30.72	34.48	32.00	28.36
50	25	87.57	55.58	85.15	92.23	51.24	95.57	84.33	78.54	70.73
100	50	96.08	99.33	93.67	97.82	99.18	99.89	99.92	99.33	95.79
300	150	99.77	100.00	99.58	99.88	100.00	100.00	100.00	100.00	100.00
500	250	99.98	100.00	99.98	99.99	100.00	100.00	100.00	100.00	100.00
20	20	35.49	16.77	39.99	34.43	13.66	40.56	33.45	30.82	27.57
50	50	94.34	60.21	88.51	96.00	57.19	98.38	64.54	38.32	15.21
100	100	98.78	99.60	96.46	99.32	99.56	99.89	99.13	89.61	46.20
300	300	99.98	100.00	99.95	99.99	100.00	100.00	100.00	100.00	100.00
500	500	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
20	200	43.46	4.89	69.53	41.61	4.55	73.07	0.00	0.00	0.00
50	500	99.67	62.63	99.05	99.69	62.00	99.67	0.00	0.00	0.00
100	1000	100.00	99.90	99.97	100.00	99.90	99.98	14.76	0.00	0.00
300	3000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.98	14.35
500	5000	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Moreover, let the family of candidate subsets be all of the subsets included in the intercept term, i.e., $\mathcal{J} = \{j \in \mathfrak{P}(\{1,\ldots,9\}) |\ j \cap \{9\} \neq \emptyset\}$, where $\mathfrak{P}(A)$ is the power set of a set A. Then, for each group $b = 1,\ldots,8$, we carried out the following two steps:

Step 1. Let $\boldsymbol{U}_g$ $(g = 1,\ldots,8)$ be the 25×541 response matrices by random sampling without replacement from group g. Further, let $\boldsymbol{U}_{9,b}$ be the 25×541 response matrices by random sampling without replacement from the remaining samples in group b. Then, the response matrix is constructed as $\boldsymbol{Y}_b = (\boldsymbol{U}_1', \ldots, \boldsymbol{U}_8', \boldsymbol{U}_{9,b}')'$.

Step 2. Let the coefficient matrix $\boldsymbol{\Theta}_b$ given by $\boldsymbol{\Theta}_b = (\boldsymbol{\theta}_{1,b}, \ldots, \boldsymbol{\theta}_{8,b}, \boldsymbol{\theta}_{9,b})'$. Then, apply multivariate linear regression with $\boldsymbol{X}$ and $\boldsymbol{\Theta}_b$ to the response matrix $\boldsymbol{Y}_b$, and choose the best subset by performing variable selection from the explanatory variables excepting the intercept, i.e., from the elements of $\mathcal{J}$.

From steps 1 and 2, we have $n = 225$, $p = 541$, and $k = 9$ in this example. Note that $\boldsymbol{\theta}_{b,b}$ should be $\boldsymbol{0}_p$ and the remainder should not be $\boldsymbol{0}_p$, because $\boldsymbol{U}_{9,b}$ is extracted from the same group as $\boldsymbol{U}_b$. Hence, we know that the true subset is $j_{*,b} = \{1,\ldots,9\} \cap \{b\}^c$ when $\boldsymbol{Y}_b$ is used as the response matrix. Moreover, to increase calculation speed, instead of a variable selection method such as (6), we used the best subset $\tilde{j}$ by the following method:

$$\tilde{j} = \{\ell \in \omega |\ \mathrm{SC}(\omega_\ell) > \mathrm{SC}(\omega)\}, \tag{23}$$

where $\mathrm{SC}(j)$ expresses the value of a variable selection criterion (SC) for model M_j, and ω_ℓ is defined in assumption A4. The selection method as per (23) was proposed by Zhao *et al.* (1986). From Nishii *et al.* (1988), it is known that when k is fixed, a criterion under (23) is consistent if the criterion under the selection method such as (6) is consistent. For these settings, we iterated steps 1 and 2 $10,000$ times for each group $b = 1,\ldots,8$. Table 10 shows the probabilities of selecting the true subset by the nine criteria for each group $b = 1,\ldots,8$. We observe that the probabilities by criterion 6 are highest except where $b = 5, 6$. However, all nine criteria have very low probabilities where $b = 5, 6$. This is because groups 5 and 6 are very similar. Actually, letting $\bar{\boldsymbol{y}}_g$ be the sample mean vector of group g, we have $||\bar{\boldsymbol{y}}_5 - \bar{\boldsymbol{y}}_6|| \fallingdotseq 0.46$ but $||\bar{\boldsymbol{y}}_g - \bar{\boldsymbol{y}}_h|| \geq 1.60$ for the cases of $g, h \neq 5, 6$ $(g \neq h)$. Hence, groups 5 and 6 will be very similar on average. Moreover, criterion 6 selected $\{1,\ldots,9\} \cap \{5,6\}^c$ as the best subset for many iterations when $b = 5, 6$.

Next, we provide an example of variable selection using empirical data from Wille *et al.* (2004) as well as Yamamura *et al.* (2010). There are 795 genes which may exhibit associations with 39 genes from two biosynthesis pathways in Arabidopsis thaliana. All variables were logarithmically transformed. We configured the former 795 genes to response variables ($p = 795$) with the latter 39 genes and an intercept as explanatory variables ($k = 40$). The sample size is $n = 118$. We searched for the best subset of these models by using the selection method (23). Table 11 shows the explanatory variables selected by each criterion and the

Table 10. True subset selection probabilities (%) for each group $b = 1, \ldots, 8$ in the black cotton fibers dataset

	Criterion								
b	1	2	3	4	5	6	7	8	9
1	79.96	97.09	76.19	90.82	99.55	99.98	56.07	4.63	0.04
2	84.12	98.33	80.43	94.15	99.84	100.00	99.88	99.96	99.29
3	97.94	100.00	96.79	99.80	100.00	100.00	92.85	16.50	0.47
4	86.62	98.75	83.16	95.37	99.86	100.00	32.92	3.48	0.03
5	5.65	0.11	8.41	1.66	0.00	0.00	0.00	0.00	0.00
6	12.14	0.42	16.45	4.31	0.01	0.00	0.00	0.00	0.00
7	72.52	92.94	68.48	85.56	91.70	98.86	90.40	60.48	21.15
8	99.57	100.00	98.98	99.96	100.00	100.00	100.00	100.00	100.00

number of elements of the best subsets. From Table 11, we observe that criteria 7, 8, and 9 selected zero explanatory variables, and criteria 2 and 5 selected few variables. On the other hand, criteria 3 and 6 selected about half of the variables.

5. Conclusions and Discussions

We obtained the conditions for consistency of the SGC_p criterion and RGC_p criterion under the HUHD and UHD asymptotic frameworks. Importantly, consistency is established under non-normality and does not rely on the divergence speed of the dimension of the vector stacked with response variables p. Numerical studies suggest that criterion 6 has the highest probabilities of selecting the true subset, although consistency of criterion 6 holds when $\log\log p/\log n \to 0$.

Herein, the scalar matrix $p^{-1}\mathrm{tr}(\boldsymbol{S}_\omega)\boldsymbol{I}_p$ and the ridge-type sample covariance matrix $\boldsymbol{S}_\lambda$ were used as $\boldsymbol{G}$ in the weighted L_2 squared distance $d(\boldsymbol{A}, \boldsymbol{B}|\boldsymbol{G})$. The SGC_p criterion and RGC_p criterion are invariant under transformation by a scalar times orthogonal matrices of $\boldsymbol{Y}$, i.e., $\boldsymbol{Y} : \boldsymbol{Y} \to a\boldsymbol{Y}\boldsymbol{F}$, where $\boldsymbol{F}$ satisfies $\boldsymbol{F}\boldsymbol{F}' = \boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_p$ and $a \in \mathbb{R}$. However, they are not invariant under transformation by nonsingular matrices of $\boldsymbol{Y}$, so their consistency is affected by the elements of $\boldsymbol{\Sigma}_*$ even for overspecified subsets. This is often the case in high-dimensional contexts such that $p > n$. On the other hand, using $\mathrm{diag}\{(\boldsymbol{S}_\omega)_{11}, \ldots, (\boldsymbol{S}_\omega)_{pp}\}$ or $\boldsymbol{S}_\omega + \lambda^{-1}\mathrm{diag}\{(\boldsymbol{S}_\omega)_{11}, \ldots, (\boldsymbol{S}_\omega)_{pp}\}$ as $\boldsymbol{G}$ may eradicate the influence of the diagonal elements of $\boldsymbol{\Sigma}_*$. Hence, it is also important to examine consistency in such cases. To do so would require assuming normality of the error vector and this represents fruitful terrain for future research.

Finally, we consider the influence of increasing p on consistency. To do so, another expression of multivariate linear regression is given by

$$\mathrm{vec}(\boldsymbol{Y}) = (\boldsymbol{I}_p \otimes \boldsymbol{X})\mathrm{vec}(\boldsymbol{\Theta}) + \mathrm{vec}(\boldsymbol{\mathcal{E}}),$$

where $\mathrm{vec}(\boldsymbol{A})$ is the np-dimensional vector consisting of the columns of an $n \times p$ matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ and is defined by $\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_n')'$ (see, e.g., Harville, 1997). From the above expression, multivariate linear regression is regarded as univariate linear regression with the np-dimensional response vector $\mathrm{vec}(\boldsymbol{Y})$ and the explanatory matrix $\boldsymbol{I}_p \otimes \boldsymbol{X}$ formally. From this, at first glance

Table 11. Selected explanatory variables based on the Arabidopsis thaliana dataset

	Criterion								
Name	1	2	3	4	5	6	7	8	9
Intercept	1	1	1	1	1	1	0	0	0
AACT1	1	0	1	1	0	1	0	0	0
AACT2	0	0	1	0	0	1	0	0	0
CMK	0	0	1	0	0	0	0	0	0
DPPS1	0	0	0	0	0	0	0	0	0
DPPS2	1	0	1	1	0	1	0	0	0
DPPS3	0	0	0	0	0	0	0	0	0
DXPS1	0	0	0	0	0	0	0	0	0
DXPS2(cla1)	1	0	1	1	0	1	0	0	0
DXPS3	0	0	1	0	0	0	0	0	0
DXR	1	0	1	1	0	1	0	0	0
FPPS1	0	0	0	0	0	0	0	0	0
FPPS2	0	0	0	0	0	0	0	0	0
GGPPS1mt	0	0	0	0	0	0	0	0	0
GGPPS2	0	0	0	0	0	0	0	0	0
GGPPS3	0	0	0	0	0	0	0	0	0
GGPPS4	0	0	0	0	0	0	0	0	0
GGPPS5	0	0	0	0	0	0	0	0	0
GGPPS6	1	0	1	1	0	1	0	0	0
GGPPS8	0	0	0	0	0	0	0	0	0
GGPPS9	0	0	0	0	0	0	0	0	0
GGPPS10	0	0	0	0	0	0	0	0	0
GGPPS11	0	0	1	0	0	0	0	0	0
GGPPS12	1	0	1	1	0	1	0	0	0
GPPS	1	0	1	1	0	1	0	0	0
HDR	1	0	1	1	0	1	0	0	0
HDS	1	0	1	1	0	1	0	0	0
HMGR1	1	0	1	1	0	1	0	0	0
HMGR2	0	0	1	0	0	1	0	0	0
HMGS	0	0	1	0	0	0	0	0	0
IPPI1	1	0	1	1	0	1	0	0	0
IPPI2	0	0	1	0	0	1	0	0	0
MCT	0	0	1	0	0	0	0	0	0
MECPS	0	0	1	0	0	1	0	0	0
MK	0	0	0	0	0	0	0	0	0
MPDC1	0	0	0	0	0	0	0	0	0
MPDC2	0	0	1	0	0	0	0	0	0
PPDS1	0	0	0	0	0	0	0	0	0
PPDS2mt	0	0	0	0	0	0	0	0	0
UPPS1	1	0	1	1	0	1	0	0	0
$\#(j)$	13	1	23	13	1	17	0	0	0

(1: selected variable, 0: non-selected variable.)

it seems that the dimension p has a role in increasing the sample size. However, from the results in Lemma 2 and Lemma 3, the probabilities of selecting j_* by the consistent criteria in this paper always approach 1 by diverging n, but do not always approach 1 by diverging only p. Moreover, increasing p leads to fast convergence of the probability of selecting the true subset under assumption A3$'$, but this is not always the case under assumption A3. This difference depends on the assumption about $\mathbf{\Sigma}_*$ and κ_4 since $\xi \mathrm{tr}(\mathbf{\Sigma}_*)^{-1} = o(1)$ holds under assumption A3$'$ not A3. This may also be verified from our simulations. Hence, to ensure fast convergence of the probability of selecting the true subset, a small sample size may be sufficient under assumption A3$'$ when p is large. As per subsection 2.2, assumption A3$'$ holds when (9) is supported. Since the sphericity condition $\lim_{p\to\infty} \mathrm{tr}(\mathbf{\Sigma}_*^2)/\mathrm{tr}(\mathbf{\Sigma}_*)^2 = 0$ is equivalent to $\lim_{p\to\infty} \lambda_{\max}(\mathbf{\Sigma}_*)/\mathrm{tr}(\mathbf{\Sigma}_*) = 0$, note that this condition implies that the maximum eigenvalue of $\mathbf{\Sigma}_*$ is not particularly large in the sense that $\lambda_{\max}(\mathbf{\Sigma}_*) = o(p)$ under assumption A2. However, in general $\lambda_{\max}(\mathbf{\Sigma}_*)$ tends to be very large for high-dimensional cases. Thus, it may not be suitable to assume the sphericity condition for high-dimensional cases. Aoshima and Yata (2018; 2019) considered methods to translate statistics under the strongly spiked model $\liminf_{p\to\infty} \lambda_{\max}(\mathbf{\Sigma}_*)^2/\mathrm{tr}(\mathbf{\Sigma}_*^2) > 0$ into those under the non-strongly spiked model $\lim_{p\to\infty} \lambda_{\max}(\mathbf{\Sigma}_*)^2/\mathrm{tr}(\mathbf{\Sigma}_*^2) = 0$. By applying their idea to criteria for multivariate linear regression used in this paper, fast convergence of the probability of selecting the true subset can be ensured even under assumption A3, and, again, this should be explored in future research.

Appendix

A. Proof of equation (8)**.** Let $j \in \mathcal{J}_-$. From properties of projection matrices, for any $\ell \in j_* \cap j^c$, we have the following equation:

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

Using the above equation, $\mathbf{\Theta}_*' \boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{X}_* \mathbf{\Theta}_*$ can be expressed as follows:

$$\begin{aligned} \mathbf{\Theta}_*' \boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{X}_* \mathbf{\Theta}_* &= \left(\sum_{\ell \in j_*} \boldsymbol{\theta}_\ell \boldsymbol{x}_\ell' \right) (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \left(\sum_{\ell \in j_*} \boldsymbol{x}_\ell \boldsymbol{\theta}_\ell' \right) \\ &= \boldsymbol{\theta}_\ell \boldsymbol{x}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{x}_\ell \boldsymbol{\theta}_\ell' \\ &= \boldsymbol{x}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{x}_\ell \boldsymbol{\theta}_\ell \boldsymbol{\theta}_\ell'. \end{aligned}$$

Since we have

$$\boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_j) \boldsymbol{X}_* - \boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{X}_* = \boldsymbol{X}_*' (\boldsymbol{P}_{\omega_\ell} - \boldsymbol{P}_j) \boldsymbol{X}_*,$$

and $\boldsymbol{X}_*' (\boldsymbol{P}_{\omega_\ell} - \boldsymbol{P}_j) \boldsymbol{X}_*$ is positive-semidefinite, the following equation can be derived:

$$\lambda_{\max}(\mathbf{\Delta}_j) \geq \lambda_{\max}(\mathbf{\Theta}_*' \boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{X}_* \mathbf{\Theta}_*) = \boldsymbol{x}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{x}_\ell \boldsymbol{\theta}_\ell' \boldsymbol{\theta}_\ell.$$

Hence, equation (8) can be derived from assumption A4. □

B. Proof of Lemma 1. We need a lemma to prove Lemma 1. To derive the upper bounds of probabilities, we use the variances of $(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})$, $\mathrm{tr}(\boldsymbol{U}_j)$, and $\mathrm{tr}(\boldsymbol{V}_{j,h})$. The results for the variances are as follows (the proof is given in Appendix H):

LEMMA B.1. *Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix and $\boldsymbol{B}$ be a $p \times n$ matrix. Then, the following results hold:*

(i) $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)] = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{\Sigma}_*)$.

(ii) $E[\mathrm{tr}(\boldsymbol{B}\boldsymbol{\mathcal{E}}_*)^2] = \mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{B}\boldsymbol{B}')$.

(iii) $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2] = \left(\sum_{i=1}^n\{(\boldsymbol{A})_{ii}\}^2\right)\kappa_4 + \mathrm{tr}(\boldsymbol{A})^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^2 + 2\mathrm{tr}(\boldsymbol{A}^2)\mathrm{tr}(\boldsymbol{\Sigma}_*^2)$, *where $\kappa_4 = E[||\boldsymbol{\varepsilon}||^4] - \mathrm{tr}(\boldsymbol{\Sigma}_*)^2 - 2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)$, which is defined in assumption A3.*

Let $j \supsetneq h$. Since $\boldsymbol{I}_n - \boldsymbol{P}_\omega$ and $\boldsymbol{P}_j - \boldsymbol{P}_h$ are symmetric idempotent matrices, we can identify that

$$\sum_{i=1}^n\{(\boldsymbol{I}_n - \boldsymbol{P}_\omega)_{ii}\}^2 \le \sum_{i=1}^n(\boldsymbol{I}_n - \boldsymbol{P}_\omega)_{ii} = \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{P}_\omega) = n - k,$$
$$\sum_{i=1}^n\{(\boldsymbol{P}_j - \boldsymbol{P}_h)_{ii}\}^2 \le \sum_{i=1}^n(\boldsymbol{P}_j - \boldsymbol{P}_h)_{ii} = \mathrm{tr}(\boldsymbol{P}_j - \boldsymbol{P}_h) = k_j - k_h.$$

From the above equations and Lemma B.1, we can evaluate the expectations and variances of $(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})$, $\mathrm{tr}(\boldsymbol{U}_j)$, and $\mathrm{tr}(\boldsymbol{V}_{j,h})$ as follows:

$$E[(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})] = \mathrm{tr}(\boldsymbol{\Sigma}_*),\ Var[(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})] \le 3(n-k)^{-1}\xi^2,$$
$$E[\mathrm{tr}(\boldsymbol{U}_j)^2] = \mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}_j),$$
$$E[\mathrm{tr}(\boldsymbol{V}_{j,h})] = (k_j - k_h)\mathrm{tr}(\boldsymbol{\Sigma}_*),\ Var[\mathrm{tr}(\boldsymbol{V}_{j,h})] \le 3(k_j - k_h)\xi^2.$$

Then, we obtain the results of Lemma 1 by using Chebyshev's inequality. First, we derive the results of (i), (ii), and (iii) as follows:

$$\begin{aligned}
&P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \ge r_1\right)\\
&= P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{\Sigma}_*) \ge r_1 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\right)\\
&\le P\left(|(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{\Sigma}_*)| \ge r_1 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\right)\\
&\le Var[(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})]\{r_1 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2} = O\left(\xi^2 n^{-1}\{r_1 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2}\right),\\
&P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \le r_2\right)\\
&= P\left((n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{\Sigma}_*) \le r_2 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\right)\\
&\le P\left(|(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{\Sigma}_*)| \ge \mathrm{tr}(\boldsymbol{\Sigma}_*) - r_2\right)\\
&\le Var[(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})]\{\mathrm{tr}(\boldsymbol{\Sigma}_*) - r_2\}^{-2} = O\left(\xi^2 n^{-1}\{\mathrm{tr}(\boldsymbol{\Sigma}_*) - r_2\}^{-2}\right),\\
&P\left(\mathrm{tr}(\boldsymbol{U}_j) \le r_3\right)\\
&\le P\left(|\mathrm{tr}(\boldsymbol{U}_j)| \ge |r_3|\right)\\
&\le E[\mathrm{tr}(\boldsymbol{U}_j)^2]|r_3|^{-2} = O\left(\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}_j)|r_3|^{-2}\right),
\end{aligned}$$

$$
\begin{aligned}
&P\left(\mathrm{tr}(\boldsymbol{V}_{j,h}) \geq (k_j - k_h) r_4\right) \\
&= P\left(\mathrm{tr}(\boldsymbol{V}_{j,h}) - (k_j - k_h)\mathrm{tr}(\boldsymbol{\Sigma}_*) \geq (k_j - k_h)\{r_4 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}\right) \\
&\leq Var[\mathrm{tr}(\boldsymbol{V}_{j,h})](k_j - k_h)^{-2}\{r_4 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2} = O\left(\xi^2\{r_4 - \mathrm{tr}(\boldsymbol{\Sigma}_*)\}^{-2}\right).
\end{aligned}
$$

Next, we obtain result (iv). When n is sufficiently large or both n and p are sufficiently large, we have

$$
-r_5 + r_6 < 0,\ (r_5 - r_6)^{-1} = O(r_5^{-1}).
$$

Hence, result (iii) can be derived as follows:

$$
\begin{aligned}
&P\left(\mathrm{tr}(\boldsymbol{V}_{j,h}) - (k_j - k_{\tilde{j}})\mathrm{tr}(\boldsymbol{\Sigma}_*) + r_5 \leq r_6\right) \\
&\leq P\left(|\mathrm{tr}(\boldsymbol{V}_{j,h}) - (k_j - k_h)\mathrm{tr}(\boldsymbol{\Sigma}_*)| \geq r_5 - r_6\right) \\
&\leq Var[\mathrm{tr}(\boldsymbol{V}_{j,h})]\,(r_5 - r_6)^{-2} = O(\xi^2 r_5^{-2}).
\end{aligned}
$$

□

C. Proof of Lemma 2. First, we obtain the order of $\overline{P}_S$. For $j \in \mathcal{J}_+ \cap \{j_*\}^c$, let $\boldsymbol{W} = \boldsymbol{\mathcal{E}}_*'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{\mathcal{E}}_*$ and $\boldsymbol{V}_{j,j_*} = \boldsymbol{\mathcal{E}}_*'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}_*$ defined by (12). It is straightforward that the equation $(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{X}_* = (\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{X}_* = \boldsymbol{O}_{n,k_*}$ holds. Then, we have

$$
\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\} = \mathrm{tr}(\boldsymbol{W}),\ \mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{Y}\} = \mathrm{tr}(\boldsymbol{V}_{j,j_*}).
$$

Using the above equations, $SGC_p(j|\alpha) - SGC_p(j_*|\alpha)$ is calculated as

$$
\begin{aligned}
SGC_p(j|\alpha) - SGC_p(j_*|\alpha) &= -(n-k)\frac{\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{Y}\}}{\mathrm{tr}(\boldsymbol{W})} + (k_j - k_*)\alpha \\
&= -(n-k)\frac{\mathrm{tr}(\boldsymbol{V}_{j,j_*})}{\mathrm{tr}(\boldsymbol{W})} + (k_j - k_*)\alpha. \qquad \text{(C.1)}
\end{aligned}
$$

Let E_S be an event defined by

$$
E_S = \{(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \geq \tau_S \mathrm{tr}(\boldsymbol{\Sigma}_*)\}. \qquad \text{(C.2)}
$$

Then, by using (C.1) and (C.2), we have

$$
\begin{aligned}
\overline{P}_S &= P\left(\cup_{j \in \mathcal{J}_+ \cap \{j_*\}^c}\left\{\mathrm{tr}(\boldsymbol{V}_{j,j_*}) \geq (n-k)^{-1}\mathrm{tr}(\boldsymbol{W})(k_j - k_*)\alpha\right\}\right) \\
&= P\left(\left\{\cup_{j \in \mathcal{J}_+ \cap \{j_*\}^c}\left\{\mathrm{tr}(\boldsymbol{V}_{j,j_*}) \geq (n-k)^{-1}\mathrm{tr}(\boldsymbol{W})(k_j - k_*)\alpha\right\}\right\} \cap (E_S \cup E_S^c)\right) \\
&\leq P\left(\cup_{j \in \mathcal{J}_+ \cap \{j_*\}^c}\left\{\mathrm{tr}(\boldsymbol{V}_{j,j_*}) \geq (k_j - k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\tau_S\right\}\right) + P(E_S^c) \\
&\leq \sum_{j \in \mathcal{J}_+ \cap \{j_*\}^c} P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) \geq (k_j - k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\tau_S\right) + P(E_S^c). \qquad \text{(C.3)}
\end{aligned}
$$

From (i) and (iii) of Lemma 1, the orders of two terms in (C.3) are as follows:

$$
\begin{aligned}
&\sum_{j \in \mathcal{J}_+ \cap \{j_*\}^c} P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) \geq (k_j - k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\tau_S\right) \\
&= O\left(\xi^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2}(\alpha\tau_S - 1)^{-2}\right), \\
P(E_S^c) &= O\left(\xi^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2}n^{-1}(1 - \tau_S)^{-2}\right).
\end{aligned}
$$

From the above equations and (C.3), we have

$$\overline{P}_S = O\left(\xi^2 \mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2} \max\left\{(\alpha\tau_S - 1)^{-2}, n^{-1}(1-\tau_S)^{-2}\right\}\right). \tag{C.4}$$

Next, we obtain the order of $\underline{P}_S$. For $j \in \mathcal{J}_-$, let

$$j_+ = j \cup j_*,\ E_{S,j} = \{SGC_p(j_+|\alpha) - SGC_p(j_*|\alpha) \geq 0\}.$$

Using j_+ and $E_{S,j}$, we have

$$\begin{aligned}
\underline{P}_S &= P\left(\cup_{j\in\mathcal{J}_-}\{SGC_p(j|\alpha) - SGC_p(j_+|\alpha) + SGC_p(j_+|\alpha) - SGC_p(j_*|\alpha) \leq 0\}\right) \\
&= P\Big(\cup_{j\in\mathcal{J}_-}\{SGC_p(j|\alpha) - SGC_p(j_+|\alpha) + SGC_p(j_+|\alpha) - SGC_p(j_*|\alpha) \leq 0\} \\
&\quad \cap (E_{S,j} \cup E_{S,j}^c)\Big) \\
&\leq P\left(\cup_{j\in\mathcal{J}_-}\{SGC_p(j|\alpha) - SGC_p(j_+|\alpha) \leq 0\}\right) + P(\cup_{j\in\mathcal{J}_-} E_{S,j}^c).
\end{aligned} \tag{C.5}$$

Since $j_+ \in \mathcal{J}_+$ is the same as (C.4), the order of $P(\cup_{j\in\mathcal{J}_-} E_{S,j}^c)$ is calculated as

$$P(\cup_{j\in\mathcal{J}_-} E_{S,j}^c) = O\left(\xi^2 \mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2} \max\left\{(\alpha\tau_S - 1)^{-2}, n^{-1}(1-\tau_S)^{-2}\right\}\right). \tag{C.6}$$

Notice that

$$\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j)\boldsymbol{Y}\} = \mathrm{tr}(\boldsymbol{V}_{j_+,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2,$$

where δ_j^2 and $\boldsymbol{U}_j = \boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}_*$ are defined by (7) and (12), respectively. From this, $SGC_p(j|\alpha) - SGC_p(j_+|\alpha)$ is calculated as

$$\begin{aligned}
&SGC_p(j|\alpha) - SGC_p(j_+|\alpha) \\
&= (n-k)\frac{\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j)\boldsymbol{Y}\}}{\mathrm{tr}(\boldsymbol{W})} - (k_{j_+} - k_j)\alpha \\
&= (n-k)\mathrm{tr}(\boldsymbol{W})^{-1}\left\{\mathrm{tr}(\boldsymbol{V}_{j_+,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2\right\} - (k_{j_+} - k_j)\alpha,.
\end{aligned} \tag{C.7}$$

Let E_1 and $E_{2,j}$ be events defined by

$$E_1 = \left\{(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \leq \frac{3}{2}\mathrm{tr}(\boldsymbol{\Sigma}_*)\right\},\ E_{2,j} = \left\{\mathrm{tr}(\boldsymbol{U}_j) \geq -\frac{1}{4}\delta_j^2\right\}. \tag{C.8}$$

Then, by using (C.7) and (C.8), we have

$$\begin{aligned}
&P\left(\cup_{j\in\mathcal{J}_-}\{SGC_p(j|\alpha) - SGC_p(j_+|\alpha) \leq 0\}\right) \\
&= P\left(\cup_{j\in\mathcal{J}_-}\left\{\mathrm{tr}(\boldsymbol{V}_{j_+,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2 \leq (n-k)^{-1}\mathrm{tr}(\boldsymbol{W})(k_{j_+} - k_j)\alpha\right\}\right) \\
&= P\Big(\cup_{j\in\mathcal{J}_-}\left\{\mathrm{tr}(\boldsymbol{V}_{j_+,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2 \leq (n-k)^{-1}\mathrm{tr}(\boldsymbol{W})(k_{j_+} - k_j)\alpha\right\} \\
&\quad \cap (E_1 \cup E_1^c)\Big) \\
&\leq P\left(\bigcup_{j\in\mathcal{J}_-}\left\{\mathrm{tr}(\boldsymbol{V}_{j_+,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2 \leq \frac{3}{2}(k_{j_+} - k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\right\}\right) \\
&\quad + P(E_1^c)
\end{aligned}$$

$$
\begin{aligned}
&= P\left(\bigcup_{j\in\mathcal{J}_-}\left\{\mathrm{tr}(\boldsymbol{V}_{j_+,j})+2\mathrm{tr}(\boldsymbol{U}_j)+\delta_j^2\le\frac{3}{2}(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\right\}\cap(E_{2,j}\cup E_{2,j}^c)\right)\\
&\quad+P(E_1^c)\\
&\le\sum_{j\in\mathcal{J}_-}P\left(\mathrm{tr}(\boldsymbol{V}_{j_+,j})+\frac{1}{2}\delta_j^2\le\frac{3}{2}(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\right)\\
&\quad+P(E_1^c)+\sum_{j\in\mathcal{J}_-}P(E_{2,j}^c). \qquad \text{(C.9)}
\end{aligned}
$$

Notice that

$$
\frac{\mathrm{tr}(\boldsymbol{\Sigma}_*)}{np}\left(\frac{3}{2}\alpha-1\right)\to 0,\ \mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}_j)\le\lambda_{\max}(\boldsymbol{\Sigma}_*)\delta_j^2.
$$

Hence, by using (8) and (i), (ii), and (iii) of Lemma 1, the orders of three terms in (C.9) can be derived as follows:

$$
\begin{aligned}
&\sum_{j\in\mathcal{J}_-}P\left(\mathrm{tr}(\boldsymbol{V}_{j_+,j})+\frac{1}{2}\delta_j^2\le\frac{3}{2}(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\alpha\right)\\
&=\sum_{j\in\mathcal{J}_-}P\left(\mathrm{tr}(\boldsymbol{V}_{j_+,j})-(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)+\frac{1}{2}\delta_j^2\le(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\left(\frac{3}{2}\alpha-1\right)\right)\\
&\le\sum_{j\in\mathcal{J}_-}P\left(\frac{\mathrm{tr}(\boldsymbol{V}_{j_+,j})-(k_{j_+}-k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)}{np}+\frac{1}{2}\tilde{\delta}\le(k_{j_+}-k_j)\frac{\mathrm{tr}(\boldsymbol{\Sigma}_*)}{np}\left(\frac{3}{2}\alpha-1\right)\right)\\
&=O(\xi^2n^{-2}p^{-2}), \qquad \text{(C.10)}\\
&P(E_1^c)=O(\xi^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2}n^{-1}), \qquad \text{(C.11)}\\
&\sum_{j\in\mathcal{J}_-}P(E_{2,j}^c)=\sum_{j\in\mathcal{J}_-}O(\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}_j)\delta_j^{-4})=O(\lambda_{\max}(\boldsymbol{\Sigma}_*)n^{-1}p^{-1}), \qquad \text{(C.12)}
\end{aligned}
$$

where $\tilde{\delta}$ is a positive constant satisfying $0<\tilde{\delta}<\min_{j\in\mathcal{J}_-}\inf_{n>k,p\ge1}(np)^{-1}\delta_j^2$. From (C.5), (C.6), (C.9), (C.10), (C.11), and (C.12), we have

$$
\begin{aligned}
\underline{P}_S&=O\left(\xi^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2}\max\left\{(\alpha\tau_S-1)^{-2},n^{-1}(1-\tau_S)^{-2}\right\}\right)\\
&\quad+O\left(\max\left\{\xi^2n^{-2}p^{-2},\xi^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2}n^{-1},\lambda_{\max}(\boldsymbol{\Sigma}_*)n^{-1}p^{-1}\right\}\right). \qquad \text{(C.13)}
\end{aligned}
$$

Therefore, (C.4) and (C.13) complete the proof of Lemma 2. □

D. Proof of Theorem 1. First, we obtain the consistency conditions under assumptions A1, A2, A3, and A4. Note that under assumptions A2 and A3, the following equations hold:

$$
\frac{\xi}{\mathrm{tr}(\boldsymbol{\Sigma}_*)}=O(1),\ \frac{\xi}{p}=O(1),\ \frac{\lambda_{\max}(\boldsymbol{\Sigma}_*)}{p}=O(1).
$$

Let us take $\tau_S = 1/2$ in Lemma 2. By using Lemma 2 and the above equations, the orders of $\overline{P}_S$ and $\underline{P}_S$ are as follows:

$$\overline{P}_S = O\left(\max\left\{(\alpha/2-1)^{-2}, n^{-1}\right\}\right),$$
$$\underline{P}_S = O\left(\max\left\{(\alpha/2-1)^{-2}, n^{-1}\right\}\right) + O(n^{-1}).$$

The above equations and (13) give the consistency conditions in (14).

Next, we obtain the consistency conditions under assumptions A1, A2, A3′, and A4. Let us take $\tau_S = 1 - n^{-1/2}$ in Lemma 2. Then, using (13), we have

$$(\alpha\tau_S - 1)^{-2} = (\alpha-1)^{-2}\left\{1 - \frac{\alpha}{\sqrt{n}(\alpha-1)}\right\}^{-2} = O\left((\alpha-1)^{-2}\right),$$
$$n^{-1}(1-\tau_S)^{-2} = 1.$$

Note that under assumptions A2 and A3′, the following equations hold:

$$\frac{\xi}{\mathrm{tr}(\boldsymbol{\Sigma}_*)} = o(1),\ \frac{\xi}{p} = o(1),\ \frac{\lambda_{\max}(\boldsymbol{\Sigma}_*)}{p} = o(1).$$

Hence, the orders of $\overline{P}_S$ and $\underline{P}_S$ are as follows:

$$\overline{P}_S = o\left((\alpha-1)^{-2}\right) + o(1),\ \underline{P}_S = o\left((\alpha-1)^{-2}\right) + o(1).$$

The above equations and (13) give the consistency conditions in (15). □

E. Proof of Theorem 2. First, we show the inconsistency under condition C1. Let $\boldsymbol{W}$ and $\boldsymbol{V}_{j,j_*}$ be defined by (12) and let $E_3 = \{(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \le (1+n^{-1/4})\mathrm{tr}(\boldsymbol{\Sigma}_*)\}$. For any $j \in \mathcal{J}_+ \cap \{j_*\}^c$, we have

$$\begin{aligned}
&P(\hat{j}_S = j_*)\\
&= P\left(\cap_{h\in\mathcal{J}\cap\{j_*\}^c}\{SGC_p(h|\alpha) > SGC_p(j_*|\alpha)\}\right)\\
&\le P\left(SGC_p(j|\alpha) > SGC_p(j_*|\alpha)\right)\\
&= P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) < \alpha(k_j-k_*)(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})\right)\\
&\le P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) - (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*) < (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\{(1+n^{-1/4})\alpha - 1\}\right)\\
&\quad + P(E_3^c). \qquad \text{(E.1)}
\end{aligned}$$

Moreover, when n is sufficiently large or n and p are sufficiently large, we have

$$\begin{aligned}
&P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) - (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*) < (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\{(1+n^{-1/4})\alpha - 1\}\right)\\
&\le P\left(|\mathrm{tr}(\boldsymbol{V}_{j,j_*}) - (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)| \ge (k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)\{1-(1+n^{-1/4})\alpha\}\right)\\
&\le \frac{Var[\mathrm{tr}(\boldsymbol{V}_{j,j_*})]}{(k_j-k_*)^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^2\{1-(1+n^{-1/4})\alpha\}^2}\\
&\le \frac{\kappa_4 I(\kappa_4 > 0) + 2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{(k_j-k_*)\mathrm{tr}(\boldsymbol{\Sigma}_*)^2\{1-(1+n^{-1/4})\alpha\}^2}
\end{aligned}$$

$$= (k_j - k_*)^{-1}(1-\alpha)^{-2}\left(1 - \frac{n^{-1/4}\alpha}{1-\alpha}\right)^{-2}\left\{\frac{\kappa_4 I(\kappa_4 > 0) + 2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_*)^2}\right\}. \quad \text{(E.2)}$$

Further, by using (i) in Lemma 1, the order of $P(E_3^c)$ is as follows:

$$P(E_3^c) = O(\xi^2 \mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2} n^{-1/2}). \quad \text{(E.3)}$$

From (E.1), (E.2), and (E.3), condition C1 gives the following inequality:

$$\lim_{n\to\infty, p/n\to c} P(\hat{j}_S = j_*)$$
$$\leq (k_j - k_*)^{-1}\left\{\lim_{n\to\infty, p/n\to c} \frac{\kappa_4 I(\kappa_4 > 0) + 2\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{(1-\alpha)^2 \mathrm{tr}(\boldsymbol{\Sigma}_*)^2}\right\} < 1.$$

Next, we show the inconsistency under condition C2. For $j \subsetneq j_*$, let $E_4 = \{(n-k)^{-1}\mathrm{tr}(\boldsymbol{W}) \geq (1 - n^{-1/4})\mathrm{tr}(\boldsymbol{\Sigma}_*)\}$ and $E_{5,j} = \{\mathrm{tr}(\boldsymbol{U}_j) \leq n^{-1/4}\delta_j^2\}$, where $\boldsymbol{U}_j$ is defined by (12). Then ,we have

$$\begin{aligned}
&P(\hat{j}_S = j_*) \\
&\leq P\left(SGC_p(j|\alpha) > SGC_p(j_*|\alpha)\right) \\
&= P\left(\mathrm{tr}(\boldsymbol{V}_{j_*,j}) + 2\mathrm{tr}(\boldsymbol{U}_j) + \delta_j^2 > \alpha(k_* - k_j)(n-k)^{-1}\mathrm{tr}(\boldsymbol{W})\right) \\
&\leq P\left(\mathrm{tr}(\boldsymbol{V}_{j_*,j}) > (k_* - k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)(1 - n^{-1/4})\alpha - (1 + 2n^{-1/4})\delta_j^2\right) \\
&\quad + P(E_4^c) + P(E_{5,j}^c). \quad \text{(E.4)}
\end{aligned}$$

From condition (C2), it is straightforward to identify that

$$\lim_{n\to\infty, p/n\to c} \frac{(k_* - k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\{(1 - n^{-1/4})\alpha - 1\}}{(1 + 2n^{-1/4})\delta_j^2} > 1.$$

Hence, when n is sufficiently large or n and p are sufficiently large, we have

$$\begin{aligned}
&P\left(\mathrm{tr}(\boldsymbol{V}_{j,j_*}) > (k_* - k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)(1 - n^{-1/4})\alpha - (1 + 2n^{-1/4})\delta_j^2\right) \\
&\leq \frac{Var[\mathrm{tr}(\boldsymbol{V}_{j,j_*})]}{[(k_* - k_j)\mathrm{tr}(\boldsymbol{\Sigma}_*)\{(1 - n^{-1/4})\alpha - 1\} - (1 + 2n^{-1/4})\delta_j^2]^2} = O(n^{-2}). \quad \text{(E.5)}
\end{aligned}$$

Further, by using (i) and (ii) in Lemma 1, the orders of $P(E_4^c)$ and $P(E_{5,j}^c)$ are as follows:

$$P(E_4^c) = O(\xi^2 \mathrm{tr}(\boldsymbol{\Sigma}_*)^{-2} n^{-1/2}),\ P(E_{5,j}^c) = O(\lambda_{\max}(\boldsymbol{\Sigma}_*) p^{-1} n^{-1/2}). \quad \text{(E.6)}$$

Equations (E.4), (E.5), and (E.6) give $\lim_{n\to\infty, p/n\to c} P(\hat{j}_S = j_*) = 0$.

Finally, when we replace assumption A3 with assumption A3′, the results in this case can be derived from (E.1), (E.2), and (E.3) because of $\xi \mathrm{tr}(\boldsymbol{\Sigma}_*)^{-1} = o(1)$. □

F. Proof of Lemma 3. For $j \in \mathcal{J}_+ \cap \{j_*\}^c$, using (19), we have

$$\begin{aligned}
&RGC_p(j|\alpha,\lambda) - RGC_p(j_*|\alpha,\lambda) \\
&= -\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{Y}\boldsymbol{S}_\lambda^{-1}\} + (k_j - k_*)p\alpha \\
&\geq -\mathrm{tr}(\boldsymbol{V}_{j,j_*})\lambda_{\max}(\boldsymbol{S}_\lambda^{-1}) + (k_j - k_*)p\alpha \\
&\geq -\lambda(n-k)\frac{\mathrm{tr}(\boldsymbol{V}_{j,j_*})}{\mathrm{tr}(\boldsymbol{W})} + (k_j - k_*)p\alpha \\
&= \lambda\{SGC_p(j|\alpha) - SGC_p(j_*|\alpha)\} + (k_j - k_*)(p-\lambda)\alpha, \qquad \text{(F.1)}
\end{aligned}$$

where $\boldsymbol{V}_{j,j_*}$ and $\boldsymbol{W}$ are given by (12). Moreover, for $j \in \mathcal{J}_-$, using (19), we have

$$\begin{aligned}
&RGC_p(j|\alpha,\lambda) - RGC_p(j_+|\alpha,\lambda) \\
&= \mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j)\boldsymbol{Y}\boldsymbol{S}_\lambda^{-1}\} - (k_{j_+} - k_j)p\alpha \\
&\geq \lambda_{\min}(\boldsymbol{S}_\lambda^{-1})\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j)\boldsymbol{Y}\} - (k_{j_+} - k_j)p\alpha \\
&\geq (1+\lambda^{-1})^{-1}(n-k)\mathrm{tr}(\boldsymbol{W})^{-1}\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j)\boldsymbol{Y}\} - (k_{j_+} - k_j)p\alpha \\
&= (1+\lambda^{-1})^{-1}\{SGC_p(j|\alpha) - SGC_p(j_+|\alpha)\} + (k_{j_+} - k_j)\{(1+\lambda^{-1})^{-1} - p\}\alpha, \\
&\qquad \text{(F.2)}
\end{aligned}$$

where $j_+ = j \cup j_*$. From (F.1) and (F.2), we can replace $RGC_p(j|\alpha,\lambda) - RGC_p(j_*|\alpha,\lambda)$ and $RGC_p(j|\alpha,\lambda) - RGC_p(j_+|\alpha,\lambda)$ with $SGC_p(j|\alpha) - SGC_p(j_*|\alpha)$ and $SGC_p(j|\alpha) - SGC_p(j_+|\alpha)$, respectively. Therefore, in the same way as the proof of Lemma 2, the results of Lemma 3 can be derived. □

G. Proof of Theorem 4. For $j \in \mathcal{J}_+ \cap \{j_*\}^c$, using (19), we have

$$\begin{aligned}
&RGC_p(j|\alpha,\lambda) - RGC_p(j_*|\alpha,\lambda) \\
&\leq -\mathrm{tr}(\boldsymbol{V}_{j,j_*})\lambda_{\min}(\boldsymbol{S}_\lambda^{-1}) + (k_j - k_*)p\alpha \\
&\leq -(1+\lambda^{-1})^{-1}(n-k)\mathrm{tr}(\boldsymbol{W})^{-1}\mathrm{tr}(\boldsymbol{V}_{j,j_*}) + (k_j - k_*)p\alpha \\
&= (1+\lambda^{-1})^{-1}\{SGC_p(j|\alpha) - SGC_p(j_*|\alpha)\} + (k_j - k_*)\{p - (1+\lambda^{-1})^{-1}\}\alpha. \\
&\qquad \text{(G.1)}
\end{aligned}$$

For $j \subsetneq j_*$, using (19), we have

$$\begin{aligned}
&RGC_p(j|\alpha,\lambda) - RGC_p(j_*|\alpha,\lambda) \\
&\leq \lambda_{\max}(\boldsymbol{S}_\lambda^{-1})\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_*} - \boldsymbol{P}_j)\boldsymbol{Y}\} - (k_* - k_j)p\alpha \\
&\leq \lambda(n-k)\mathrm{tr}(\boldsymbol{W})^{-1}\mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{j_*} - \boldsymbol{P}_j)\boldsymbol{Y}\} - (k_* - k_j)p\alpha \\
&= \lambda\{SGC_p(j|\alpha) - SGC_p(j_*|\alpha)\} - (k_* - k_j)(\lambda - p\alpha). \qquad \text{(G.2)}
\end{aligned}$$

By using (G.1) and (G.2), in the same way as the proof of Theorem 2, the results of Theorem 4 can be derived. □

H. Proof of Lemma B.1. First, we calculate the expectation $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)]$ to prove (i). It is straightforward that

$$E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)] = \sum_{i,j}^{n}(\boldsymbol{A})_{ij}E[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_j] = \sum_{i=1}^{n}(\boldsymbol{A})_{ii}E[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i] = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{\Sigma}_*),$$

where the summation $\sum_{i,j}^{n}$ is defined by $\sum_{i=1}^{n}\sum_{j=1}^{n}$.

Next, we calculate the expectation $E[\mathrm{tr}(\boldsymbol{B}\boldsymbol{\mathcal{E}}_*)^2]$ in (ii). Let $\boldsymbol{b}_i$ be the i-th column vector of $\boldsymbol{B}$. Then, we have

$$E[\mathrm{tr}(\boldsymbol{B}\boldsymbol{\mathcal{E}}_*)^2] = \sum_{i,j}^{n}\boldsymbol{b}_i'E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j']\boldsymbol{b}_j = \sum_{i=1}^{n}\boldsymbol{b}_i'E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i']\boldsymbol{b}_i = \mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{B}\boldsymbol{B}').$$

Finally, we calculate the expectation $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2]$ in (ii). The expectation $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2]$ can be expressed as follows:

$$\begin{aligned}
E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2] &= \sum_{i,j,k,\ell}^{n}(\boldsymbol{A})_{ij}(\boldsymbol{A})_{k\ell}E[(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_j)(\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_\ell)] \\
&= \sum_{i=1}^{n}\{(\boldsymbol{A})_{ii}\}^2E[(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i)^2] + \sum_{i\neq j}^{n}(\boldsymbol{A})_{ii}(\boldsymbol{A})_{jj}E[(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i)(\boldsymbol{\varepsilon}_j'\boldsymbol{\varepsilon}_j)] \\
&\quad + 2\sum_{i\neq j}^{n}\{(\boldsymbol{A})_{ij}\}^2E[(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_j)^2] \\
&= \left(\sum_{i=1}^{n}\{(\boldsymbol{A})_{ii}\}^2\right)E[||\boldsymbol{\varepsilon}||^4] + \left(\sum_{i\neq j}^{n}(\boldsymbol{A})_{ii}(\boldsymbol{A})_{jj}\right)\mathrm{tr}(\boldsymbol{\Sigma}_*)^2 \\
&\quad + 2\left(\sum_{i\neq j}^{n}\{(\boldsymbol{A})_{ij}\}^2\right)\mathrm{tr}(\boldsymbol{\Sigma}_*^2),
\end{aligned}$$

where the summation $\sum_{i\neq j}^{n}$ is defined by $\sum_{j=1}^{n}\sum_{i:i\neq j}^{n}$. Hence, given that

$$\sum_{i\neq j}^{n}(\boldsymbol{A})_{ii}(\boldsymbol{A})_{jj} = \mathrm{tr}(\boldsymbol{A})^2 - \sum_{i=1}^{n}\{(\boldsymbol{A})_{ii}\}^2,\ \sum_{i\neq j}^{n}\{(\boldsymbol{A})_{ij}\}^2 = \mathrm{tr}(\boldsymbol{A}^2) - \sum_{i=1}^{n}\{(\boldsymbol{A})_{ii}\}^2,$$

we can calculate $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2]$ as follows:

$$E[\mathrm{tr}(\boldsymbol{\mathcal{E}}_*'\boldsymbol{A}\boldsymbol{\mathcal{E}}_*)^2] = \left(\sum_{i=1}^{n}\{(\boldsymbol{A})_{ii}\}^2\right)\kappa_4 + \mathrm{tr}(\boldsymbol{A})^2\mathrm{tr}(\boldsymbol{\Sigma}_*)^2 + 2\mathrm{tr}(\boldsymbol{A}^2)\mathrm{tr}(\boldsymbol{\Sigma}_*^2).$$

□

Acknowledgement

I wish to express my deepest gratitude to Prof. Hirokazu Yanagihara at Hiroshima University for his valuable advice and encouragement and introducing

me to various fields of mathematical statistics during the academic years 2014–2020. I also got a lot of advices about not only the personal manners as a researcher but also my private life from him, so I could not have come this far without his helps. In addition, I would like to thank Prof. Yasunori Fujikoshi at Hiroshima University for many helpful comments and suggestions about new research themes, Prof. Hirofumi Wakaki at Hiroshima University for his advice and help and Dr. Mariko Yamamura at Radiation Effects Research Foundation for her encouragement. Also, I thank to Dr. Shinpei Imori, Dr. Shintaro Hashimoto and Dr. Heewon Park at Hiroshima University for their encouragements, especially, Dr. Shinpei Imori for his valuable comments for numerical studies in this paper. Moreover, I am also grateful to Dr. Hiromi Itamiya at National Research Institute of Police Science, for providing me with the black cotton fiber dataset used in one of the empirical examples. I thank to my colleagues, seniors and juniors for their helps. I would also like to thank the referee for valuable comments.

References

[1] Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y. H. & Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Aust. Nz. J. Stat.*, **60**, 4–19.

[2] Aoshima, M. & Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Stat. Sinica*, **28**, 43–62.

[3] Aoshima, M. & Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Ann. I. Stat. Math.*, **71**, 473–503.

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

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

[6] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[7] Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.*, **29**, 995–1010.

[8] Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, **16**, 41–50.

[9] Fujikoshi, Y., Kan, T., Takahashi, S. & Sakurai, T. (2011). Prediction error criterion for selecting variables in a linear regression model. *Ann. I. Stat. Math.,* **63**, 387–403.

[10] Fujikoshi, Y., Himeno, T. & Wakaki, H. (2004). Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size. *J. Japan Statist. Soc.,* **34**, 19–26.

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

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

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

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

[15] Himeno, T. & Yamada, T. (2014). Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. *J. Multivariate Anal.*, **130**, 27–44.
[16] Katayama, S. & Imori, S. (2014). Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *J. Multivariate Anal.*, **132**, 138–150.
[17] Kubokawa, T. & Srivastava, M. S. (2012). Selection of variables in multivariate regression models for large dimensions. *Comm. Statist. A-Theory Methods*, **41**, 2465–2489.
[18] Magnus, J. R. & Neudecker, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.*, **7**, 381–894.
[19] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.
[20] Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.
[21] 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.
[22] Nishii, R., Bai, Z. D. & Krishnaiah, P. R. (1988). Strong consistency information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.
[23] Sparks, R. S., Coutsourides, D. & Troskie, L. (1983). The multivariate C_p. *Comm. Statist. A-Theory Methods*, **12**, 1775–1793.
[24] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
[25] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York.
[26] Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York.
[27] Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. & Bühlmenn, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.*, **5**, 1–13.
[28] Yamamura, M., Yanagihara, H. & Srivastava, M. S. (2010). Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japan. J. Appl. Stat.*, **39**, 1–19.
[29] Yanagihara, H. (2015). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of the normality assumption. *J. Japan Statist. Soc.*, **45**, 21–56.
[30] 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.
[31] Yanagihara, H. (2019). Evaluation of consistency of model selection criteria in multivariate linear regression models by large-sample and high-dimensional asymptotic theory under nonnormality. *J. Jpn. Stat. Soc. Jpn. Issue*, **48**, 1–13.
[32] 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.
[33] 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.

Ryoya Oda
Department of Mathematics
Graduate School of Science
Hiroshima University
Higashi-Hiroshima 739-8526, JAPAN
E-mail: ryoya-oda@hiroshima-u.ac.jp