STRONG CONSISTENCY OF THE AIC, BIC, C_P AND KOO METHODS IN HIGH-DIMENSIONAL MULTIVARIATE LINEAR REGRESSION

By Zhidong Bai[*,§], Yasunori Fujikoshi[†,¶] and Jiang Hu[‡,§]

Northeast Normal University[§] and Hiroshima University[¶]

Variable selection is essential for improving inference and interpretation in multivariate linear regression. Although a number of alternative regressor selection criteria have been suggested, the most prominent and widely used are the Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's C_p, and their modifications. However, for high-dimensional data, experience has shown that the performance of these classical criteria is not always satisfactory. In the present article, we begin by presenting the necessary and sufficient conditions (NSC) for the strong consistency of the high-dimensional AIC, BIC, and C_p, based on which we can identify some reasons for their poor performance. Specifically, we show that under certain mild high-dimensional conditions, if the BIC is strongly consistent, then the AIC is strongly consistent, but not vice versa. This result contradicts the classical understanding. In addition, we consider some NSC for the strong consistency of the high-dimensional kick-one-out (KOO) methods introduced by Zhao et al. (1986) and Nishii et al. (1988). Furthermore, we propose two general methods based on the KOO methods and prove their strong consistency. The proposed general methods remove the penalties while simultaneously reducing the conditions for the dimensions and sizes of the regressors. A simulation study supports our consistency conclusions and shows that the convergence rates of the two proposed general KOO methods are much faster than those of the original methods.

1. Introduction. In multivariate statistical analysis, the most general and favorable model to investigate the relationship between a predictor matrix $\mathbf{X}$ and a response matrix $\mathbf{Y}$ is the multivariate linear regression (MLR) model. More specifically, let

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}, \tag{1.1}$$

[*]Supported by NSFC 11571067 and 11471140.

[†]Supported by NSFC 11771073.

[‡]Supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #16K00047, 2016-2018.

MSC 2010 subject classifications: Primary 62J05, 62H12; secondary 62E20

Keywords and phrases: AIC, BIC, C_p, KOO methods, Strong consistency, High-dimensional criteria, Multivariate linear regression, Variable selection

where $\mathbf{Y} = (y_{ij}) : n \times p$ (the responses), $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k) : n \times k$ (the predictors), $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)' : k \times p$ (the regression coefficients), and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_p) = (e_{ij}) : n \times p$ (the random errors). The goal in MLR analysis is to estimate the regression coefficients Θ. The estimates should be such that the estimated regression plane explains the variation in the values of the responses with great accuracy. The classical linear least-squares solution is to estimate the matrix of regression coefficients $\hat{\Theta}$ by

$$\hat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

However, model (1.1) (referred to hereinafter as the full model) is not always a good model because some of the predictors may be uncorrelated with the responses, i.e., the corresponding rows of Θ are zeros. The Akaike information criterion (AIC), Bayesian information criterion (BIC), and Mallows's C_p are among the most popular and versatile strategies for model selection among the predictors.

Let $\mathbf{j}$ be a subset of $\boldsymbol{\omega} = \{1, 2, \cdots, k\}$ and $\mathbf{X}_\mathbf{j} = (\mathbf{x}_j, j \in \mathbf{j})$ and $\Theta_\mathbf{j} = (\boldsymbol{\theta}_j, j \in \mathbf{j})'$. Denote model $\mathbf{j}$ by

$$M_\mathbf{j} : \quad \mathbf{Y} = \mathbf{X}_\mathbf{j}\Theta_\mathbf{j} + \mathbf{E}. \tag{1.2}$$

Akaike's seminal paper (Akaike, 1973) proposed the use of the Kullback-Leibler distance as a fundamental basis for model selection known as the AIC, which is defined as follows:

$$A_\mathbf{j} = n\log(|\widehat{\boldsymbol{\Sigma}}_\mathbf{j}|) + 2[k_\mathbf{j}p + \frac{1}{2}p(p+1)] + np(\log(2\pi) + 1), \tag{1.3}$$

where

$$n\widehat{\boldsymbol{\Sigma}}_\mathbf{j} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{j})\mathbf{Y}, \quad \mathbf{P}_\mathbf{j} = \mathbf{X}_\mathbf{j}(\mathbf{X}_\mathbf{j}'\mathbf{X}_\mathbf{j})^{-1}\mathbf{X}_\mathbf{j}',$$

and $k_\mathbf{j}$ is the cardinality of subset $\mathbf{j}$. The BIC, which is also known as the Schwarz criterion, was proposed by Schwarz (1978) in the form of a penalized log-likelihood function, in which the penalty is equal to the logarithm of the sample size times the number of estimated parameters in the model, i.e.,

$$B_\mathbf{j} = n\log(|\widehat{\boldsymbol{\Sigma}}_\mathbf{j}|) + \log(n)[k_\mathbf{j}p + \frac{1}{2}p(p+1)] + np(\log(2\pi) + 1). \tag{1.4}$$

A criterion with behavior similar to that of the AIC for variable selection in regression models is Mallows's C_p proposed by Mallows (1973), which is defined as follows:

$$C_\mathbf{j} = (n-k)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_\boldsymbol{\omega}^{-1}\widehat{\boldsymbol{\Sigma}}_\mathbf{j}) + 2pk_\mathbf{j}. \tag{1.5}$$

Refer to (Fujikoshi, 1983; Sparks et al., 1983; Nishii et al., 1988) for additional details of formulas (1.3), (1.4), and (1.5). Then, the AIC, BIC, and C_p rules are used to select

$$\hat{\mathbf{j}}_A = \arg\min A_{\mathbf{j}}, \quad \hat{\mathbf{j}}_B = \arg\min B_{\mathbf{j}} \quad \text{and} \quad \hat{\mathbf{j}}_C = \arg\min C_{\mathbf{j}}, \tag{1.6}$$

respectively.

If the data are generated from a model (referred to hereinafter as a true model) that is one of the candidate models, then we would apply a model selection method to identify the true model. Then, some optimality, such as consistency, is desirable for model selection. A model selection method is weakly consistent if, with probability tending to one, the selection method is able to select the true model from the candidate models. Strong consistency means that the true model tends to almost surely be selected. Strong consistency implies weak consistency but not vice versa. Thus, strong consistency can provide a deeper understanding of the selection methods. Under a large-sample asymptotic framework, i.e., dimension p is fixed and n tends to infinity, the AIC and C_p are not consistent (Fujikoshi, 1985; Fujikoshi and Veitch, 1979), but the BIC is strongly consistent (Nishii et al., 1988). However, in recent years, statisticians have increasingly noticed that these properties cannot be adapted to high-dimensional data. In particular, experience has shown that the classical model selection criteria tend to select more variables than necessary when k and p are large. For the case in which k is fixed, p is large but smaller than n, and $p/n \to c \in [0, 1)$, which is referred to as a large-sample and large-dimensional asymptotic framework, the BIC has been shown to be not consistent, but the AIC and C_p are weakly consistent under certain conditions (see, e.g., (Fujikoshi et al., 2014; Yanagihara et al., 2015; Yanagihara, 2015)).

To clarify the model selection methods, in the present paper, we focus on the strongly consistent properties under a large-model, large-sample, and large-dimensional (LLL) asymptotic framework, i.e., $\min\{k, p, n\}$ tends to infinity for the case in which $p/n \to c \in (0, 1)$, $k/n \to \alpha \in (0, 1 - c)$, and the true model size is fixed. We do not intend to judge the advantages and disadvantages of the existing selection methods in the MLR model beyond their consistency properties in the present paper, because these advantages and disadvantages depend on their intended applications and on the nature of the data. Our goal is to explain the theoretical insights of the classical selection methods and modified methods under an LLL framework, and we hope that this article will stimulate further research that will provide an even clearer understanding of high-dimensional variable selection. Here, we refer to three recent reviews comparing the variable selection methods (Anzanello

and Fogliatto, 2014; Blei et al., 2017; Heinze et al., 2018). In addition, note that in the present paper, we assume that $n - k > p$. A number of studies have examined sparse and penalized methods for high-dimensional data for which this condition is not satisfied, see, e.g., (Li et al., 2015; Zou and Hastie, 2005).

We now describe the four main contributions of the present paper.

- First, in Section 3.2, we present the necessary and sufficient conditions (NSC) for the strong consistency of variable selection methods based on the AIC, BIC, and C_p under an LLL asymptotic framework, including the ranges of c and α, the moment condition of random errors (our results do not require the normality condition), and the convergence rate of the noncentrality matrix. Specifically, on the basis of these results, we conclude that under an LLL asymptotic framework, if the BIC is strongly consistent, then the AIC is strongly consistent, but not vise versa, which contradicts the classical understanding.
- Second, in Section 3.3, we examine the strongly consistent properties of the kick-one-out (KOO) methods based on the AIC, BIC, and C_p under an LLL asymptotic framework, which were introduced by Zhao et al. (1986) and Nishii et al. (1988) and followed by Fujikoshi and Sakurai (2018). The KOO methods, which were proposed for the computation problem in the classical AIC, BIC, and C_p, reduce the number of computational statistics from $2^k - 1$ to k. In addition, Nishii et al. (1988) showed that under a large-sample asymptotic framework, the KOO methods share the same conditions and strong consistency of the classical AIC and BIC. However, in the present paper, we find that under an LLL asymptotic framework, the KOO methods have higher costs for dimension conditions than do the AIC, BIC, and C_p for strong consistency.
- Third, on the basis of the KOO methods, in Section 3.4, we propose two general KOO methods that not only remove the penalty terms but also reduce the conditions for the dimensions and sizes of the predictors. Furthermore, the sufficient condition is given for their strong consistency. The proposed general KOO methods have considerable advantages, such as simplicity of expression, ease of computation, limited restrictions, and fast convergence.
- Fourth, random matrix theory (RMT) is introduced to model selection methods in high-dimensional MLR. The new theoretical results and the concepts behind their proofs are applicable to numerous other model selection methods, such as the modified AIC and C_p (Fujikoshi and Satoh, 1997; Bozdogan, 1987). Furthermore, the technical tool

developed in the present paper is applicable to future research, e.g., the growth curve model (Enomoto et al., 2015; Fujikoshi et al., 2013), multiple discriminant analysis (Fujikoshi, 1983; Fujikoshi and Sakurai, 2016a), principal component analysis (Fujikoshi and Sakurai, 2016b; Bai et al., 2018), and canonical correlation analysis (Nishii et al., 1988; Bao et al., 2018).

The remainder of this paper is organized as follows. In Section 2, we present the necessary notation and assumptions for the MLR model under an LLL asymptotic framework. The main results on the strong consistency of the AIC, BIC, C_p, KOO, and general KOO methods are stated in Section 3. Section 4 presents simulation studies to illustrate the performance of our results. The main theorems are proven in Section 5, and other potential applications are briefly discussed in Section 6. Finally, additional technical results are present in the Appendix.

2. Notation and Assumptions. In this section, we introduce the notation and assumptions required for our main results to hold. Recall the MLR model (1.1)

$$M: \quad \mathbf{Y} = \mathbf{X}\Theta + \mathbf{E},$$

$\mathbf{j}$ is a subset of $\boldsymbol{\omega} = \{1, 2, \cdots, k\}$, $k_{\mathbf{j}}$ is the cardinality of set $\mathbf{j}$, $\mathbf{X}_{\mathbf{j}} = \{\mathbf{x}_j, j \in \mathbf{j}\}$, and $\Theta_{\mathbf{j}} = \{\theta_j, j \in \mathbf{j}\}$. Model $\mathbf{j}$ is denoted by

$$M_{\mathbf{j}}: \quad \mathbf{Y} = \mathbf{X}_{\mathbf{j}}\Theta_{\mathbf{j}} + \mathbf{E}. \tag{2.1}$$

Denote the true model as $\mathbf{j}_*$, and

$$M_{\mathbf{j}_*}: \quad \mathbf{Y} = \mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*} + \mathbf{E}.$$

We first suppose that the model and the random errors satisfy the following conditions:

(A1): The true model $\mathbf{j}_*$ is a subset of set $\boldsymbol{\omega}$, and $k_* := k_{\mathbf{j}_*}$ is fixed.
(A2): $\{e_{ij}\}$ are i.i.d. with zero means, unit variances, and finite fourth moments.

Note that in a typical multivariate regression model, the rows of $\mathbf{E}$ are assumed to be i.i.d. from a p-variate distribution with zero mean and covariance matrix Σ. Since we consider the distribution of a statistic invariant under the transformation $\mathbf{Y} \to \mathbf{Y}\Sigma^{-1/2}$, without loss of generality, we may assume that $\Sigma = \mathbf{I}_p$ by replacing Θ with $\Theta\Sigma^{-1/2}$. Moreover, the finite fourth

moments condition is required only for the technical proof; we believe finite second moments are sufficient.

Let

$$\mathbf{P_j} = \mathbf{X_j}(\mathbf{X_j'X_j})^{-1}\mathbf{X_j'} \text{ and } \mathbf{Q_j} = \mathbf{I}_n - \mathbf{P_j};$$

then, we have

$$n\widehat{\mathbf{\Sigma}}_\mathbf{j} = \mathbf{Y'Q_jY}. \tag{2.2}$$

For the purpose of identifiability, we assume that

(A3): $\mathbf{X'X}$ is positive definite and $\mathbf{x}_j'\mathbf{X}_{\mathbf{j}_*} = \mathbf{0}$ for any $j \in \boldsymbol{\omega}\backslash\mathbf{j}_*$.

Note that if assumption (A3) is satisfied, then for any $\mathbf{j} \subset \boldsymbol{\omega}$, $\mathbf{X_j'X_j}$ is invertible because $\mathbf{X_j'X_j}$ is a submatrix of $\mathbf{X'X}$. In addition, if there exists some j that satisfies $\mathbf{x}_j'\mathbf{X}_{\mathbf{j}_*} \neq \mathbf{0}$, then we have

$$\mathbf{X}_{\boldsymbol{\omega}}\Theta_{\boldsymbol{\omega}} = \mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*} + \mathbf{X}_{\boldsymbol{\omega}\backslash\mathbf{j}_*}\Theta_{\boldsymbol{\omega}\backslash\mathbf{j}_*} = \mathbf{X}_{\mathbf{j}_*}\widetilde{\Theta}_{\mathbf{j}_*} + \widetilde{\mathbf{X}}_{\boldsymbol{\omega}\backslash\mathbf{j}_*}\Theta_{\boldsymbol{\omega}\backslash\mathbf{j}_*},$$

where

$$\widetilde{\Theta}_{\mathbf{j}_*} = \Theta_{\mathbf{j}_*} - (\mathbf{X}_{\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*})^{-1}\mathbf{X}_{\mathbf{j}_*}'\Theta_{\boldsymbol{\omega}\backslash\mathbf{j}_*} \text{ and } \widetilde{\mathbf{X}}_{\boldsymbol{\omega}\backslash\mathbf{j}_*} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{j}_*})\mathbf{X}_{\boldsymbol{\omega}\backslash\mathbf{j}_*}.$$

Then, it holds that

$$\widetilde{\mathbf{X}}_{\boldsymbol{\omega}\backslash\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*} = \mathbf{0}.$$

Therefore, by considering the parameter transformation $\Theta_{\mathbf{j}_*} \to \widetilde{\Theta}_{\mathbf{j}_*}$, we can assume that (A3) holds.

In the present paper, we focus primarily on an LLL asymptotic framework, which is specified as follows.

(A4): Assume that as $\{k, p, n\} \to \infty$, $p/n \to c \in (0, 1)$ and $k/n \to \alpha \in (0, 1-c)$, and denote $t/n \to \alpha_t$ for $0 \leq t \leq k$.

We assume that c and α are larger than 0 because we can take α and c as two unknown parameters, the consistent estimators of which are k/n and p/n, respectively, no information for the convergence of $\{k, p, n\}$ exists for any dataset at hand, and k/n and p/n are always positive. Therefore, the assumption that c and α are positive is reasonable.

Next, we present additional notation that is frequently used herein. Denote

$$\mathbf{J}_+ = \{\mathbf{j} : \mathbf{j} \supset \mathbf{j}_*\},\ \mathbf{J}_- = \{\mathbf{j} : \mathbf{j} \not\supset \mathbf{j}_*\} \text{ and } \mathbf{J} = \mathbf{J}_- \cup \mathbf{j}_* \cup \mathbf{J}_+.$$

In the following, we use the terms overspecified model and underspecified model to indicate whether a model $\mathbf{j}$ includes the true model (i.e., $\mathbf{j} \in \mathbf{J}_+$) or not (i.e., $\mathbf{j} \in \mathbf{J}_-$). If $\mathbf{j}$ is an overspecified model and $k_{\mathbf{j}} - k_{\mathbf{j}_*} = m > 0$, then subsets of models $\mathbf{j} = \mathbf{j}_0 = \mathbf{j}_{-1} \supset \cdots \supset \mathbf{j}_{-m} = \mathbf{j}_*$ exist such that each consecutive pair decreases one and only one index, i.e., $\mathbf{j}_t = \mathbf{j}_0 \backslash \{j_1, \cdots, j_{-t}\}$ for $t = 0, -1, \ldots, -m$, which means that j_{-t} is in $\mathbf{j}_{t+1}$ but not in $\mathbf{j}_t$. If $\mathbf{j}$ is an underspecified model, denote $\mathbf{j}_- = \mathbf{j} \cap \mathbf{j}_*$, $\mathbf{j}_+ = \mathbf{j} \cap \mathbf{j}_*^c$ and write the elements in $\mathbf{j}_* \cap \mathbf{j}_-^c$ as $i_1, \cdots, i_s$ and the elements in $\mathbf{j}_+$ as $j_1, \cdots, j_m$. Define the model index set $\mathbf{j}_t = \mathbf{j} \cup \{i_{t+1}, \cdots, i_s\}$ for $t = 0, 1, \cdots, s$ with convention that $\mathbf{j}_s = \mathbf{j}$, which also indicates that i_t is in $\mathbf{j}_{t+1}$ but not in $\mathbf{j}_t$. Moreover, we can define $\mathbf{j}_t = \mathbf{j}_0 \backslash \{j_t, \cdots, j_{-t}\}$ for $t = 0, -1, \ldots, -m$, but we should note that in this case, $\mathbf{j}_0 \neq \mathbf{j}$. The positive subscript of $\mathbf{j}$ indicates the addition of indexes, and, correspondingly, the negative subscript of $\mathbf{j}$ indicates the removal of indexes. In addition, we denote $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{i_t} / \|\mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{i_t}\|$ for $t > 0$ and $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{j_{-t}} / \|\mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{j_{-t}}\|$ for $t < 0$. Thus, for any integer t, the following two equations are straightforward:

$$\mathbf{P}_{\mathbf{j}_{t+1}} = \mathbf{P}_{\mathbf{j}_t} + \mathbf{a}_t\mathbf{a}_t' \tag{2.3}$$

and

$$\mathbf{Q}_{\mathbf{j}_{t+1}} = \mathbf{Q}_{\mathbf{j}_t} - \mathbf{a}_t\mathbf{a}_t'. \tag{2.4}$$

We hereinafter denote the spectral norm for a matrix by $\|\cdot\|$. For $\mathbf{j} \in \mathbf{J}_-$, we need additional notation for the main results. Let $t < 0$ and $\boldsymbol{\ell}_t = \{i_1, \cdots, i_{-t}\}$. Then, we denote

$$\begin{aligned}
\Delta_t :=& (\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_t})^{1/2}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'(\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_t})^{1/2};\\
\tilde{\mathbf{a}}_t' :=& \mathbf{a}_l'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_t}(\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{Q}_{\mathbf{j}_l}\mathbf{X}_{\boldsymbol{\ell}_t})^{-1/2};\\
\delta_t :=& \tilde{\mathbf{a}}_t'\left((1 - k_{\mathbf{j}_t}/n)\mathbf{I} + n^{-1}\Delta_t\right)^{-1}\tilde{\mathbf{a}}_t;\\
\eta_t :=& \frac{1}{n}\tilde{\mathbf{a}}_t'\Delta_t\tilde{\mathbf{a}}_t = \frac{1}{n}\mathbf{a}_t'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{a}_t;\\
\Phi_{\mathbf{j}} =& \frac{1}{n}\Theta_{\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*}'(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\mathbf{j}})\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}.
\end{aligned}$$

By basic calculation, we obtain that δ_t and η_t can both be expressed as functions of the noncentrality matrix $\Phi_{\mathbf{j}}$, as follows:

$$\prod_{l=1}^{-t}\tilde{\mathbf{a}}_l'((1 - \alpha_m)\mathbf{I} + \frac{1}{n}\Delta_l)^{-1}\tilde{\mathbf{a}}_l = (1 - \alpha_m)^{p+t}|(1 - \alpha_m)\mathbf{I} + \Phi_{\mathbf{j}}|^{-1}$$

and

$$\sum_{l=1}^{-t}\eta_l=\frac{1}{n}\mathrm{tr}[\Theta_{\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*}'(\sum_{l=1}^{-t}\mathbf{a}_l\mathbf{a}_l')\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}]=\mathrm{tr}(\Phi_{\mathbf{j}}).$$

For the underspecified model, our results require another assumption:

(A5): For all $\mathbf{j}\in\mathbf{J}_-$ with $k_{\mathbf{j}_+}=m>0$ and $k_{\mathbf{j}_*\cap\mathbf{j}_-^c}=s>0$, as $\{p,n\}\to\infty$, $(1-\alpha_m)^{s-p}|(1-\alpha_m)\mathbf{I}+\Phi_{\mathbf{j}}|\to\tau_{\mathbf{j}}<\infty$ and $\mathrm{tr}(\Phi_{\mathbf{j}})\to\kappa_{\mathbf{j}}<\infty$.

If assumption (A5) does not hold, then the following assumption is considered:

(A5'): For all $\mathbf{j}\in\mathbf{J}_-$ with $k_{\mathbf{j}_+}=m>0$ and $k_{\mathbf{j}_*\cap\mathbf{j}_-^c}=s>0$, as $\{p,n\}\to\infty$, the largest eigenvalue of $\Phi_{\mathbf{j}}$ tends to infinity.

Throughout the present paper, we use $o_{a.s}(1)$ to denote almost surely scalar negligible entries.

3. Main results. In this section, we present the main results of the present paper. First, we present some preliminary results, which are not only the basis of our theorems but also have meaning themselves and have many potential applications in other multivariate analysis problems.

3.1. *Preliminaries.* From the definition of the selection method based on the AIC, BIC, and C_p in (1.3)-(1.6), the strong consistency of the selection method based on the AIC (resp. the BIC and C_p) is equivalent to that for all $\mathbf{j}\in\mathbf{J}\backslash\mathbf{j}_*$, $A_{\mathbf{j}}>A_{\mathbf{j}_*}$ (resp. $B_{\mathbf{j}}>B_{\mathbf{j}_*}$ and $C_{\mathbf{j}}>C_{\mathbf{j}_*}$) almost surely for sufficiently large p and n. In addition, as $\mathbf{J}=\mathbf{J}_-\cup\mathbf{j}_*\cup\mathbf{J}_+$, we need to consider only the following two cases, i.e., the overspecified case ($\mathbf{j}\in\mathbf{J}_+$) and the underspecified case ($\mathbf{j}\in\mathbf{J}_-$), and investigate for each case the conditions that guarantee that the inequality $A_{\mathbf{j}}>A_{\mathbf{j}_*}$ (resp. $B_{\mathbf{j}}>B_{\mathbf{j}_*}$ and $C_{\mathbf{j}}>C_{\mathbf{j}_*}$) holds. We first consider the overspecified case, i.e., $\mathbf{j}\in\mathbf{J}_+$, and assume that $k_{\mathbf{j}}-k_{\mathbf{j}_*}=m>0$; then, we have

$$\frac{1}{n}(A_{\mathbf{j}}-A_{\mathbf{j}_*})=\frac{1}{n}\sum_{t=0}^{m-1}(A_{\mathbf{j}_{-t}}-A_{\mathbf{j}_{-t-1}}).$$

Based on the definition of $A_{\mathbf{j}}$ in (1.3), it follows that

$$A_{\mathbf{j}_{-t}}-A_{\mathbf{j}_{-t-1}}=\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|}\right)+\frac{2p}{n},$$

which implies

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \sum_{t=0}^{m-1}\left[\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|}\right) + \frac{2p}{n}\right]. \tag{3.1}$$

Analogously, we also have

$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = \sum_{t=0}^{m-1}\left[\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|}\right) + \log(n)\frac{p}{n}\right] \tag{3.2}$$

and

$$\begin{aligned}\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) =& (1-\frac{k}{n})\mathrm{tr}\widehat{[\mathbf{\Sigma}_{\boldsymbol{\omega}}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathbf{j}} - \widehat{\mathbf{\Sigma}}_{\mathbf{j}_*})]} + 2\frac{p}{n}m \\ =& \sum_{t=0}^{m-1}\left((1-\frac{k}{n})\mathrm{tr}[\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}} - \widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}})] + 2\frac{p}{n}\right).\end{aligned} \tag{3.3}$$

Then, we have the following lemma.

LEMMA 3.1. *Suppose that assumptions (A1) through (A4) hold. For all overspecified models* $\mathbf{j}$ *with* $k_{\mathbf{j}} - k_{\mathbf{j}_*} = m > 0$*, we have*

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) - \sum_{t=1}^{m}\log\left(\frac{1-\alpha_{m-t}-c}{1-\alpha_{m-t}}\right) - 2mc = o_{a.s.}(m), \tag{3.4}$$

$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) - \sum_{t=1}^{m}\log\left(\frac{1-\alpha_{m-t}-c}{1-\alpha_{m-t}}\right) - mc\log(n) = o_{a.s.}(m) \tag{3.5}$$

and

$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) - \frac{mc(\alpha-1)}{1-\alpha-c} - 2cm = o_{a.s.}(m). \tag{3.6}$$

Moreover, if $m/n \to \alpha_m > 0$*, then we have*

$$\frac{1}{n^2}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \phi(\alpha_m, c) + o_{a.s.}(1), \tag{3.7}$$

$$\frac{1}{n^2}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) - (\log(n)-2)c\alpha_m = \phi(\alpha_m, c) + o_{a.s.}(1), \tag{3.8}$$

and

$$\frac{1}{n^2}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = \alpha_m\psi(\alpha, c) + o_{a.s.}(1), \tag{3.9}$$

where $\phi(\alpha_m, c) = 2c\alpha_m + \log\left(\frac{(1-c)^{1-c}(1-\alpha_m)^{1-\alpha_m}}{(1-c-\alpha_m)^{1-c-\alpha_m}}\right)$ *and* $\psi(\alpha, c) = \frac{c(\alpha-1)}{1-\alpha-c} + 2c$.

The proof of this lemma is presented in Section 5.

REMARK 3.2. *Note that, taking the AIC for example, this lemma indicates that for some fixed* $m > 0$ *such that* $\sum_{t=1}^{m} \log\left(\frac{1-\alpha_{m-t}-c}{1-\alpha_{m-t}}\right) + 2mc > 0$, *for all* $\mathbf{j} \in \mathbf{J}_+$ *satisfying* $k_{\mathbf{j}} - k_{\mathbf{j}_*} = m$ *and for sufficiently large* p *and* n, $\mathbf{j}$ *almost surely cannot be selected by the AIC. On the other hand, if* $\sum_{t=1}^{m} \log\left(\frac{1-\alpha_{m-t}-c}{1-\alpha_{m-t}}\right) + 2mc < 0$, *then for sufficiently large* p *and* n, $\mathbf{j}_*$ *almost surely cannot be selected by the AIC, which means that, in this case, the AIC is almost surely inconsistent. The BIC and* C_p *for the case of* $m/n \to \alpha_m > 0$ *are analogous.*

REMARK 3.3. *3D plots are presented in Figure 1 to illustrate the ranges of* α *and* c *such that* $\phi(\alpha, c) > 0$ *and* $\psi(\alpha, c) > 0$. *This figure shows that large* α *and* c *both result in overestimation of the true model. Moreover, Fujikoshi et al. (2014); Yanagihara et al. (2015) proved that for the fixed-*k *case, the consistency ranges of* c *for the AIC and* C_p *are* $[0, 0.797)$ *and* $[0, 1/2)$, *respectively, which coincide with our results in Lemma 3.1 when* $\alpha \to 0$.

Next, we consider the underspecified case, i.e., $\mathbf{j} \in \mathbf{J}_-$. Analogously, we have

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \sum_{t=-s}^{m-1} \mathcal{A}_{-t},$$

$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = \sum_{t=-s}^{m-1} \mathcal{B}_{-t} \quad \text{and} \quad \frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = \sum_{t=-s}^{m-1} \mathcal{C}_{-t},$$

where

$$\mathcal{A}_{-t} = \begin{cases} \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|) - \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|) + 2p/n & \text{if } t \geq 0, \\ \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|) - \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|) - 2p/n & \text{if } t < 0, \end{cases}$$

$$\mathcal{B}_{-t} = \begin{cases} \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|) - \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|) + \log(n)p/n & \text{if } t \geq 0, \\ \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|) - \log(|\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|) - \log(n)p/n & \text{if } t < 0, \end{cases}$$

and

$$\mathcal{C}_{-t} = \begin{cases} (1 - k/n)\mathrm{tr}[\widehat{\mathbf{\Sigma}}_{\omega}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}} - \widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}})] + 2p/n & \text{if } t \geq 0, \\ (1 - k/n)\mathrm{tr}[\widehat{\mathbf{\Sigma}}_{\omega}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}} - \widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}})] - 2p/n, & \text{if } t < 0. \end{cases}$$

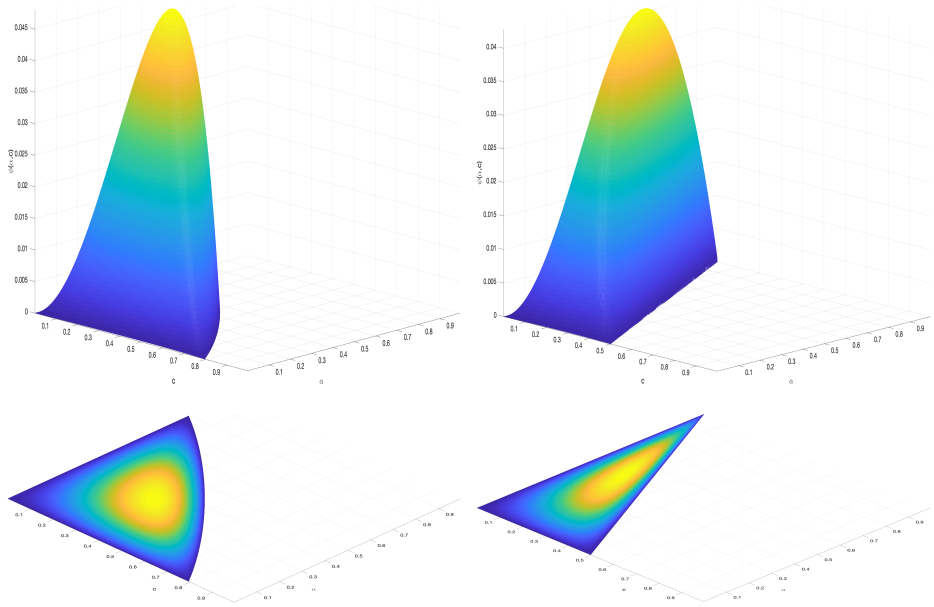


FIG 1. *3D plots for* $\phi(\alpha, c) > 0$ *and* $\psi(\alpha, c) > 0$. *The left two figures are a wireframe mesh and a contour plot for* $\phi(\alpha, c) > 0$. *The right two figures are a wireframe mesh and a contour plot for* $\psi(\alpha, c) > 0$.

Since $\mathbf{j}_0 \supset \mathbf{j}_*$, from Lemma 3.1,

$$\sum_{t=0}^{m-1} \mathcal{A}_{-t} - \sum_{t=1}^{m} \log\left(\frac{1-\alpha_{m+s-t}-c}{1-\alpha_{m+s-t}}\right) - 2mc = o_{a.s.}(m),$$

$$\sum_{t=0}^{m-1} \mathcal{B}_{-t} - \sum_{t=1}^{m} \log\left(\frac{1-\alpha_{m+s-t}-c}{1-\alpha_{m+s-t}}\right) - mc\log(n) = o_{a.s.}(m)$$

and

$$\sum_{t=0}^{m-1} \mathcal{C}_{-t} - m\psi(\alpha, c) = o_{a.s.}(m).$$

Therefore, we need to consider only the case in which $t < 0$. Then, we have the following lemma.

LEMMA 3.4. *Suppose assumptions* (A1) through (A5) *hold. If* $\mathbf{j} \in \mathbf{J}_-$ *and* $t < 0$, *then we have*

$$\mathcal{A}_{-t} + \log(\delta_t) + \log(1-\alpha_m - c) + 2c = o_{a.s.}(1), \tag{3.10}$$

$$\mathcal{B}_{-t} + \log(\delta_t) + \log(1 - \alpha_m - c) + c\log(n) = o_{a.s.}(1), \tag{3.11}$$

and

$$\mathcal{C}_{-t} - \frac{(1-\alpha)(\eta_t + c)}{1 - c - \alpha} + 2c = o_{a.s.}(1). \tag{3.12}$$

The proof of this lemma is presented in Section 5. By combining Lemmas 3.1 and 3.4, we directly obtain the following lemma.

LEMMA 3.5. *Suppose that assumptions* (A1) through (A5) *hold. For all underspecified models* $\mathbf{j}$ *with* $k_{\mathbf{j}_-} = s > 0$ *and* $k_{\mathbf{j}_+} = m \geq 0$, *we have*

$$\begin{aligned}\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = &\sum_{t=1}^{m} \log\left(\frac{1 - \alpha_{m+s-t} - c}{1 - \alpha_{m+s-t}}\right) \\ &+ \log(\tau_{\mathbf{j}}) - s\log(1 - \alpha_m - c) + 2(m-s)c + o_{a.s.}(m),\end{aligned}$$

$$\begin{aligned}\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = &\sum_{t=1}^{m} \log\left(\frac{1 - \alpha_{m+s-t} - c}{1 - \alpha_{m+s-t}}\right) \\ &+ \log(\tau_{\mathbf{j}}) - s\log(1 - \alpha_m - c) + (m-s)c\log(n) + o_{a.s.}(m)\end{aligned}$$

and

$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = m\psi(\alpha, c) + \frac{(1-\alpha)(\kappa_{\mathbf{j}} + sc)}{1 - c - \alpha} - 2sc + o_{a.s.}(m).$$

REMARK 3.6. *Lemmas 3.1 through 3.5 are the fundamental results for the model selection criteria discussed herein. These lemmas have numerous potential applications in multivariate analysis problems, such as the growth curve model (Enomoto et al., 2015; Fujikoshi et al., 2013), multiple discriminant analysis (Fujikoshi, 1983; Fujikoshi and Sakurai, 2016a), principal component analysis (Fujikoshi and Sakurai, 2016b; Bai et al., 2018), and canonical correlation analysis (Nishii et al., 1988; Bao et al., 2018).*

3.2. *Strong consistency of the AIC, BIC, and C_p.* Finally, we are in position to present our main results concerning the strong consistency of the AIC, BIC, and C_p using the results shown in the previous subsection.

THEOREM 3.7. *Suppose assumptions* (A1) through (A5) *hold.*

(1) *If* $\phi(\alpha, c) > 0$ *and for any* $\mathbf{j} \in \mathbf{J}_-$ *with* $m - s < 0$, *if* $\log(\tau_{\mathbf{j}}) > (s-m)(\log(1-c)+2c)$, *then the variable selection method based on the AIC is strongly consistent. If for some* $\mathbf{j} \in \mathbf{J}_-$ *with* $m - s < 0$ *and* $\log(\tau_{\mathbf{j}}) < (s-m)(\log(1-c)+2c)$, *then the variable selection method based on the AIC is almost surely underspecified. Otherwise, if* $\phi(\alpha, c) < 0$, *then the variable selection method based on the AIC is almost surely overspecified.*

(2) *The variable selection method based on the BIC is almost surely underspecified.*

(3) *If* $\psi(\alpha, c) > 0$ *and for any* $\mathbf{j} \in \mathbf{J}_-$ *with* $m - s < 0$, *if* $\tau_{\mathbf{j}} > (s-m)(1-2\alpha-2c)c - cm\alpha$, *then the variable selection method based on* C_p *is strongly consistent. If for some* $\mathbf{j} \in \mathbf{J}_-$ *with* $m - s < 0$, *and* $\kappa_{\mathbf{j}} < (s-m)(1-2\alpha-2c)c - cm\alpha$, *then the variable selection method based on* C_p *is almost surely underspecified. Otherwise, if* $\psi(\alpha, c) < 0$, *then the variable selection method based on* C_p *is almost surely overspecified.*

PROOF. We first prove (1). For the case $\mathbf{j} \supset \mathbf{j}_*$, from the definition of $\phi(\alpha, c)$, we know that if $\phi(\alpha, c) > 0$, then we have $\log(1-c)+2c > 0$, and for any $0 < \alpha_m \le \alpha$, $\psi(\alpha_m, c) > 0$. In addition, if $\alpha_m = 0$, by (3.7), then we have

$$\frac{1}{nm}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) \xrightarrow{a.s.} \log(1-c)+2c > 0.$$

Thus, according to Lemma 3.1, for $\mathbf{j} \supset \mathbf{j}_*$ and for sufficiently large p and n, we have

$$A_{\mathbf{j}} > A_{\mathbf{j}_*} \quad a.s.$$

This result indicates that, in this case, the AIC asymptotically selects $\mathbf{j}_*$.

Next, we consider the case of $\mathbf{j} \in \mathbf{J}_-$. Note that $s \le k_*$, $\Phi_{\mathbf{j}} > 0$ and $\log(\tau_{\mathbf{j}}) > s\log(1-\alpha_m)$. If $m \ge s$, by Lemma 3.5 and $\phi(\alpha, c) > 0$, for sufficiently large p and n, $A_{\mathbf{j}} > A_{\mathbf{j}_*}$ almost surely. Thus, we need to consider only the case in which $m < s$. In this case, since k_* is fixed, $\alpha_m = 0$ and $\tau_{\mathbf{j}} > 1$. Then, under the condition $\log(\tau_{\mathbf{j}}) > (s-m)(\log(1-c)+2c)$, for sufficiently large p and n, we have

$$A_{\mathbf{j}} > A_{\mathbf{j}_*} \quad a.s..$$

Therefore, the variable selection method based on the AIC is strongly consistent.

For some $\mathbf{j} \in \mathbf{J}_-$ with $m - s < 0$, $\log(\tau_{\mathbf{j}}) < (s-m)(\log(1-c)+2c)$, then from Lemmas 3.1 and 3.5, we know that for sufficiently large p and n,

$$A_{\mathbf{j}} < A_{\mathbf{j}_*}, \quad a.s.,$$

which means that, in this case, the AIC asymptotically selects the underspecified model $\mathbf{j}$. On the other hand, if $\phi(\alpha, c) < 0$ and $\log(\tau_{\mathbf{j}}) > (s-m)(\log(1-c)+2c)$, by the same discussion, the AIC asymptotically selects the overspecified model. Thus, we obtain (1).

The proofs of (2) and (3) are analogous; thus, the details are not presented herein. This complete the proof of this theorem. □

REMARK 3.8. *In this theorem, k/n and p/n, respectively, are typically used instead of α and c because, for a real dataset, we do not have information regarding their limits.*

If assumption (A5) does not hold, we instead consider assumption (A5') and have the following theorem.

THEOREM 3.9. *Suppose that assumptions* (A1) through (A4) *and* (A5') *hold.*

(1) *If $\phi(\alpha, c) > 0$, then the variable selection method based on the AIC is strongly consistent.*
(2) *If $\phi(\alpha, c) > 0$ and for any $\mathbf{j} \in \mathbf{J}_-$ with $m - s < 0$, if*

$$\lim_{n,p\to\infty} \Big(\log(|\mathbf{I} + \Phi_{\mathbf{j}}|) - c(s-m)\log(n) \Big) > (s-m)\log(1-c),$$

then the variable selection method based on the BIC is strongly consistent.
(3) *If $\psi(\alpha, c) > 0$, then the variable selection method based on the C_p is strongly consistent.*

REMARK 3.10. *Theorem 3.9 can be obtained using a proof procedure similar to that for Theorem 3.7; thus, the details are not presented herein. Note that although Lemma 3.5 holds under assumption (A5), the lemma still holds under assumption (A5') when the equations are normalized by the orders of $\log(\tau_{\mathbf{j}})$ and $\kappa_{\mathbf{j}}$. This result is easily obtained by the proof of Lemma 3.5.*

REMARK 3.11. *Combining Theorems 3.7 and 3.9, under an LLL asymptotic framework, if the BIC is strongly consistent, then the AIC is strongly consistent but not vice versa. This result contradicts the classical understanding that under a large-sample asymptotic framework, the AIC and C_p are not consistent, but the BIC is strongly consistent.*

3.3. *KOO methods based on the AIC, BIC, and C_p.* The AIC, BIC, and C_p become computationally complex as k becomes large because we must compute a minimum of $2^k - 1$ statistics. An alternate procedure, which was introduced by Zhao et al. (1986) and Nishii et al. (1988) and implemented by Fujikoshi and Sakurai (2018), is available to avoid this problem. In the following, we examine the performance of this procedure under an LLL framework. Denote

$$\tilde{A}_j := \frac{1}{n}(A_{\boldsymbol{\omega}\setminus j} - A_{\boldsymbol{\omega}}) = \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) - 2p/n,$$
$$\tilde{B}_j := \frac{1}{n}(B_{\boldsymbol{\omega}\setminus j} - B_{\boldsymbol{\omega}}) = \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) - \log(n)p/n,$$
$$\tilde{C}_j := \frac{1}{n}(C_{\boldsymbol{\omega}\setminus j} - C_{\boldsymbol{\omega}}) = (1 - k/n)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}) - (n - k + 2)p/n.$$

Choose the model

$$\tilde{\mathbf{j}}_A = \{j \in \boldsymbol{\omega} | \tilde{A}_j > 0\}, \quad \tilde{\mathbf{j}}_B = \{j \in \boldsymbol{\omega} | \tilde{B}_j > 0\}$$
$$\tilde{\mathbf{j}}_C = \{j \in \boldsymbol{\omega} | \tilde{C}_j > 0\}.$$

These methods are based on the comparison of two models, models $M_{\boldsymbol{\omega}\setminus j}$ and $M_{\boldsymbol{\omega}}$; therefore, selection methods $\tilde{\mathbf{j}}_A$, $\tilde{\mathbf{j}}_B$ and $\tilde{\mathbf{j}}_C$ are referred to as kick-one-out (KOO) methods based on the AIC, BIC and C_p, respectively.

Note that the $-2\log$ likelihood ratio statistic for testing $\boldsymbol{\theta}_j = \mathbf{0}$ under normality can be expressed as

$$n\left\{\log(|\hat{\Sigma}_{\boldsymbol{\omega}}|) - \log(|\hat{\Sigma}_{\boldsymbol{\omega}/j}|)\right\}.$$

Similarly, $(n-k)\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{\omega}}^{-1}\hat{\Sigma}_{\boldsymbol{\omega}\setminus j})$ is the Lawley-Hotelling trace statistic for testing $\boldsymbol{\theta}_j = \mathbf{0}$. Here, $\tilde{A}_j$ ($\tilde{B}_j$, $\tilde{C}_j$) is regarded as a measure that expresses the degree of contribution of $\mathbf{x_j}$ based on A_j (B_j, C_p). As such, the KOO methods may also be referred to as test-based methods, as in Fujikoshi and Sakurai (2018).

Therefore, we have the following theorem for the KOO methods.

THEOREM 3.12. *Suppose assumptions* (A1) through (A4) *hold.*

(1) *If* $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c$ *and assumption* (A5) *holds, then under the condition that for any* $j \in \mathbf{j}_*$, $\log(\tau_{\boldsymbol{\omega}\setminus j}) > \log(1-\alpha-c) + 2c$, *the KOO method based on the AIC is strongly consistent and the KOO method based on the BIC is not consistent.*

(2) *If* $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c$ *and assumption* (A5') *holds, then the KOO method based on the AIC is strongly consistent.*

(3) *If* $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c$ *and for any* $j \in \mathbf{j}_*$,

$$\lim_{p,n}[\log((1-\alpha)^{1-p}|(1-\alpha)\mathbf{I} + \Phi_{\boldsymbol{\omega}\backslash j}|) - \log(n)c] > \log(1-\alpha-c),$$

then the KOO method based on the BIC is strongly consistent.

(4) *If assumption* (A5') *holds or if assumption* (A5) *holds but for any* $j \in \mathbf{j}_*$, $\kappa_{\boldsymbol{\omega}\backslash j} > \frac{c(1-\alpha-2c)}{1-\alpha}$, *then under the condition that* $(1-\alpha) < 2(1-\alpha-c)$, *the KOO method based on the* C_p *is strongly consistent.*

PROOF. We consider only the case in which, under assumptions (A1) through (A5), the other cases are analogous. If j does not exist in the true model $\mathbf{j}_*$, then $\boldsymbol{\omega}\backslash j$ includes $\mathbf{j}_*$. By Lemma 3.1, we obtain

$$\tilde{A}_j \xrightarrow{a.s.} \log\Big(\frac{1-\alpha}{1-\alpha-c}\Big) - 2c, \tag{3.13}$$

$$\tilde{B}_j + \log(n)c \xrightarrow{a.s.} \log\Big(\frac{1-\alpha}{1-\alpha-c}\Big), \quad \tilde{C}_j \xrightarrow{a.s.} \frac{(1-\alpha)c}{1-\alpha-c} - 2c.$$

If j lies in the true model $\mathbf{j}_*$, by Lemmas 3.5, we have that

$$\tilde{A}_j \xrightarrow{a.s.} \log(\tau_{\boldsymbol{\omega}\backslash j}) - \log(1-\alpha-c) - 2c, \tag{3.14}$$

$$\tilde{B}_j \xrightarrow{a.s.} \log(\tau_{\boldsymbol{\omega}\backslash j}) - \log(1-\alpha-c) - \log(n)c, \quad \tilde{C}_j \xrightarrow{a.s.} \frac{(1-\alpha)(\kappa_{\boldsymbol{\omega}\backslash j}+c)}{1-\alpha-c} - 2c.$$

Thus, we complete the proof based on a discussion similar to that for the proof of Theorem 3.7. □

REMARK 3.13. *When the dimension* p *and model size* k *are fixed but the sample size* $n \to \infty$, *the asymptotic performance of the KOO methods and the classical AIC and BIC procedures are the same, as described by Nishii et al. (1988). However, according to Theorems 3.7 and 3.12, when* p *and* k *are large, the conditions for the KOO methods based on the AIC, BIC, and* C_p *are stronger than those based on the classical AIC, BIC, and* C_p. *The reason is that the KOO methods are compared with the full model, whereas the classical AIC, BIC, and* C_p *are compared with the true model. When the full model size is large and the true model size is small, the methods have different properties.*

3.4. *General KOO methods.* In the classical AIC, BIC, and C_p, including the KOO methods based on the AIC, BIC, and C_p, the penalty terms are important and modified by many researchers. For the classical information criteria under a large-sample asymptotic framework, Nishii et al. (1988)

proved that the strong consistency must be on the order of the penalty larger than $O(\log\log n)$ and smaller than $O(n)$, which coincides with the fact that the AIC is not consistent and the BIC is strongly consistent. However, on the basis of the above results, under an LLL framework, a large penalty may cause incorrect selection (actually, a constant penalty is sufficient for strong consistency), and the ranges of α and c may be crucial for the consistency. Thus, we consider a new criterion that is independent of the penalty and reduces the conditions for α and c. Therefore, in this subsection, we propose two general KOO methods based on the likelihood ratio statistic and the Lawley-Hotelling trace statistic.

Comparison of (3.13) and (3.14) indicates that the differences in the limits of $\tilde{A}_j$ for j that exist and do not exist in the true model $\mathbf{j}_*$ are the two terms $\log(\tau_{\boldsymbol{\omega}\setminus j})$ and $\log(1-\alpha)$. More specially, $\log(1-\alpha)<0$ and $\log(\tau_{\boldsymbol{\omega}\setminus j})>0$ (in most cases). Then, we can imagine that the k values of $\tilde{A}_j$ should be separated on both sides of the critical point $-\log(1-\alpha-c)-2c$. Thus, on the basis of these properties, we propose the following methods. Denote

$$\breve{A}_j := \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) \quad \text{and} \quad \breve{C}_j := \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}).$$

Choose the model

$$\breve{\mathbf{j}}_A = \{j\in\boldsymbol{\omega}|\breve{A}_j > -\log(1-\alpha-c)\}, \quad \breve{\mathbf{j}}_C = \{j\in\boldsymbol{\omega}|\breve{C}_j > \frac{\alpha+c}{1-\alpha-c}+p\}.$$

We provide a numerical example in Figure 2 to illustrate our idea more clearly. In this case, the use of Z_2 and Z_4 as the critical points is more reasonable and more intuitive than the use of Z_1 and Z_3.

Then, we have the following theorem.

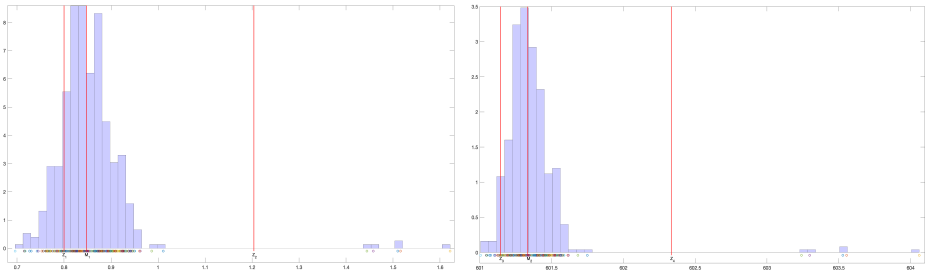
THEOREM 3.14. *Suppose assumptions* (A1) through (A4) *hold and that for any* $j\in\mathbf{j}_*$, $\lim\operatorname{tr}(\Phi_{\boldsymbol{\omega}\setminus j})>\alpha$. *Then, the general KOO methods are strongly consistent, i.e.,*

$$\lim_{n,p\to\infty}\breve{\mathbf{j}}_A \overset{a.s.}{\to} \mathbf{j}_* \quad \textit{and} \quad \lim_{n,p\to\infty}\breve{\mathbf{j}}_C \overset{a.s.}{\to} \mathbf{j}_*.$$

PROOF. Since the rank of matrix $\Phi_{\boldsymbol{\omega}\setminus j}$ is one, we have

$$\log((1-\alpha)^{1-p}|(1-\alpha)\mathbf{I}+\Phi_{\boldsymbol{\omega}\setminus j}|) = \log(1-\alpha+\operatorname{tr}(\Phi_{\boldsymbol{\omega}\setminus j})),$$

which, together with (3.13) and (3.14), directly implies this theorem. Thus, we complete the proof. □



(a) Histogram of $\breve{A}_j$ (b) Histogram of $\breve{C}_j$

FIG 2. *We chose a Gaussian sample with* $p = 600$, $n = 1500$, $k = 450$ *and* $k_* = 5$. *Hence,* $c = 0.4$ *and* $\alpha = 0.3$. *The histograms represent the distributions of the* k *values of* $\breve{A}_j$ *and* $\breve{C}_j$ *respectively. In (a),* $Z_1 = 2c$ *(resp.* $Z_2 = -\log(1-\alpha-c)$*) represents the critical point of the KOO method (resp. general KOO method) based on AIC. In (b),* $Z_3 = p + 2c/(1-\alpha)$ *(resp.* $Z_4 = p + (\alpha+c)/(1-\alpha-c)$*) represents the critical point of the KOO method (resp. general KOO method) based on* C_p. $M_1 = \log((1-\alpha)/(1-\alpha-c))$ *(resp.* $M_2 = p + c/(1-\alpha-c)$*) is the limit of* $\breve{A}_j$ *(resp.* $\breve{C}_j$*) when* j *does not lie in the true model.*

REMARK 3.15. *Note that the condition in this theorem is much weaker than that in the AIC, BIC, and* C_p *and in the KOO methods based on the AIC, BIC, and* C_p. *In addition, the general KOO methods are conservative approaches and are likely to overestimate the true model because, in practice,* $\mathrm{tr}(\Phi_{\boldsymbol{\omega}\setminus j})$ *are always large. In addition, if we have information regarding the order of* $\mathrm{tr}(\Phi_{\boldsymbol{\omega}\setminus j})$, *we can easily modify the general KOO methods by adding lower-order terms to obtain better convergence rates for* $\breve{\mathbf{j}}_A$ *and* $\breve{\mathbf{j}}_C$. *However, we do not pursue this direction in the present study.*

4. Simulation studies. In this section, we numerically examine the validity of our claims. More precisely, we attempt to examine the consistency properties of the KOO methods and the general KOO methods based on the AIC, BIC, and C_p in an LLL framework with different settings. The classical AIC, BIC, and C_p procedures are not considered herein because of their computational challenges. We conduct a number of simulation studies to examine the effects of assumptions (A2), (A3), and (A5) on the consistency of estimators $\tilde{\mathbf{j}}_A$, $\tilde{\mathbf{j}}_B$, $\tilde{\mathbf{j}}_C$, $\breve{\mathbf{j}}_A$, and $\breve{\mathbf{j}}_C$. Moreover, we are interested in gaining insight into the rate of convergence.

We consider the following two settings:

> Setting I: Fix $k_* = 5$, $p/n = \{0.2, 0.4, 0.6\}$ and $k/n = \{0.1, 0.2\}$ with several different values of n. We generate an $n \times n$ random matrix with independent $N(0,1)$ entries. By QR decomposition, we construct

an $n \times n$ orthogonal matrix and choose the first k columns as $\mathbf{X}$. Set $\Theta_{\mathbf{j}_*} = \sqrt{n}\mathbf{1}_5\boldsymbol{\theta}_*$ and $\Theta = (\Theta_{\mathbf{j}_*}, \mathbf{0})$, where $\mathbf{1}_5$ is a five-dimensional vector of ones and $\boldsymbol{\theta}_* = ((-0.5)^0, \ldots, (-0.5)^{p-1})$.

Setting II: This setting is the same as Setting I, except $\Theta_{\mathbf{j}_*} = n\mathbf{1}_5\boldsymbol{\theta}_*$.

We consider three cases for the distribution of $\mathbf{E}$: (i) a standard normal distribution; (ii) a standardized t distribution with three degrees of freedom, i.e., $e_{ij} \sim t_3/\sqrt{Var(t_3)}$; and (iii) a standardized chi-square distribution with two degrees of freedom, i.e., $e_{ij} \sim \chi^2_2/\sqrt{Var(\chi^2_2)}$.

We highlight some salient features of our settings and the distributions. For Setting I, convergent values in the conditions for consistency are presented in Table 1. From these values and Theorem 3.12, we know that $\tilde{\mathbf{j}}_A$ is strongly consistent in cases where $\{\alpha = 0.1, c = 0.2, 0.4, 0.6\}$ and $\{\alpha = 0.2, c = 0.2\}$. $\tilde{\mathbf{j}}_C$ is strongly consistent in cases where $\{\alpha = 0.1, 0.2, c = 0.2\}$, and in other cases, $\tilde{\mathbf{j}}_A$, $\tilde{\mathbf{j}}_B$ and $\tilde{\mathbf{j}}_C$ are inconsistent. By contrast, $\breve{\mathbf{j}}_A$ and $\breve{\mathbf{j}}_C$ are consistent in all cases. For Setting II, $\log(\tau_{\boldsymbol{\omega}\setminus\{1\}}) = \log(n) + O(1)$ and $\kappa_{\boldsymbol{\omega}\setminus\{1\}} = O(n)$, which satisfy assumption (A5'). Under this setting, whether $\tilde{\mathbf{j}}_B$ is strongly consistent depends on the values of c and α.

	$c = .2$				$c = .4$				$c = .6$			
	V_1	V_2	V_3	V_4	V_1	V_2	V_3	V_4	V_1	V_2	V_3	V_4
$\alpha = .1$	.15	.50	.76	1.22	.21	.10	.70	1.29	.10	-.30	.81	1.53
$\alpha = .2$	.11	.40	.87	1.23	.11	0	.97	1.33	-.17	-.40	1.17	1.63

TABLE 1

Values of $V_1 := 2c - \log(\frac{1-\alpha}{1-\alpha-c})$, $V_2 := 2(1-\alpha-c) - (1-\alpha)$, $V_3 := \log(\tau_{\boldsymbol{\omega}\setminus\{1\}}) - \log(1-\alpha-c) - 2c$, *and* $V_4 := \kappa_{\boldsymbol{\omega}\setminus\{1\}} - \frac{c(1-\alpha-2c)}{1-\alpha}$.

To illustrate the performance of these estimators, the selection percentages of belonging to $\mathbf{J}_-$, $\{\mathbf{j}_*\}$ and $\mathbf{J}_+$ were computed by Monte Carlo simulations with 1,000 repetitions. We first considered the standard normal distribution case. Since the sum of the three selection percentages is 1, for the sake of clarity of the plots, we display only the selection percentages of belonging to $\mathbf{J}_-$ and $\{\mathbf{j}_*\}$ (see Figure 3). Moreover, in some cases, when the selection percentages of belonging to $\mathbf{J}_+$ are close to 1, the selected model sizes are indicators of the consistency of the estimators, as presented in Figure 4. On the basis of these results, we have the following conclusions: (1) Under Setting I, the performances of the general KOO methods $\breve{\mathbf{j}}_A$ and $\breve{\mathbf{j}}_C$ are much better than those of the KOO methods $\tilde{\mathbf{j}}_A$, $\tilde{\mathbf{j}}_B$, and $\tilde{\mathbf{j}}_C$, and the sufficient conditions for the consistency of the KOO methods are satisfied. The convergence of $\tilde{\mathbf{j}}_C$ is the fastest among the five estimators. (2) If c is large or close to the boundary of the sufficient conditions for consistency, i.e., V_1

and V_2 are small, then the convergence rate of the selection probabilities is slow, i.e., only sufficiently large samples can guarantee their selection accuracy. However, despite the low selection accuracy in this case, these methods usually overestimate the true model, and the selection sizes are also under control. An overspecified model is more acceptable than an underspecified model. (3) The KOO method based on the BIC performs the best among the three methods under Setting II, and when its sufficient conditions for consistency are satisfied. The reasons is that overestimating the true model by the BIC is difficult compared to overestimating the true model by the other criteria.

The results under non-normal distributions are similar to those under normal distributions. Please see Figures 5 through 8 in the Appendix, which verify our conclusion that the consistency of these estimators is independent of the distribution.

5. Technical proofs. In this section, we present the technical proofs of Lemma 3.1 and Lemma 3.4. We first briefly describe our proof strategy and the main tools of RMT. Recall equations (2.2) and (2.4). From Sylvester's determinant theorem, we obtain that

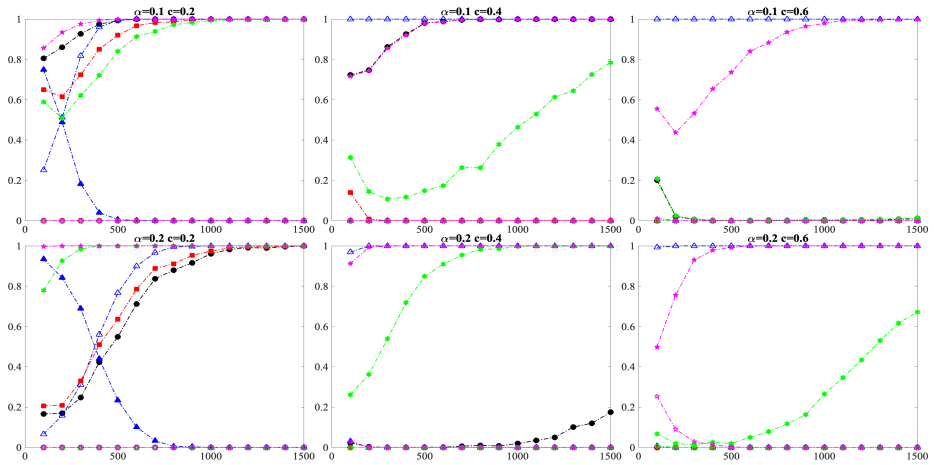
$$
\begin{aligned}
(5.1)\qquad |n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}| =& |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}| = |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{Y} - \mathbf{Y}'\mathbf{a}_{t+1}\mathbf{a}_{t+1}'\mathbf{Y}| \\
=& |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{Y}|(1 - \mathbf{a}_{t+1}'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_{t+1}) \\
=& |n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|(1 - \mathbf{a}_{t+1}'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_{t+1}).
\end{aligned}
$$

Thus, to prove Lemma 3.1 and Lemma 3.4, we need to obtain only the almost sure limits of $\mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t$ or similar expressions with different $\mathbf{j}_{-t}$. We define a function

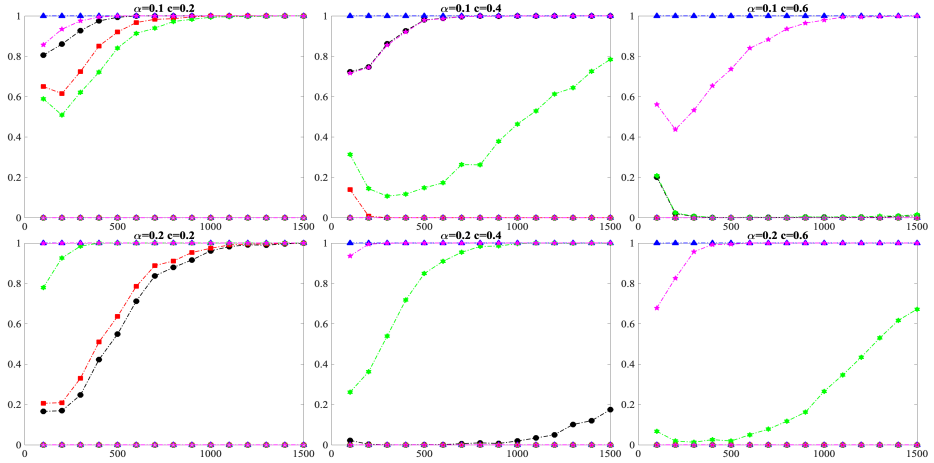
$$
\hbar_n(z) := n^{-1}\mathbf{a}_t'\mathbf{Y}(n^{-1}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y} - z\mathbf{I})^{-1}\mathbf{Y}'\mathbf{a}_t : \mathbb{C}^+ \longmapsto \mathbb{C}^+,
$$

where $\mathbb{C}^+ = \{z \in \mathbb{C}^+ : \Im z > 0\}$. Next, we prove that outside a null set independent of $\mathbf{j}_{-t}$, for every $z \in \mathbb{C}^+$, $\hbar_n(z)$ has a limit $\hbar(z) \in \mathbb{C}^+$. Note that by Vitali's convergence theorem (see Lemma 2.14 in (Bai and Silverstein, 2010), for example) it is sufficient to prove for any fixed $z \in \mathbb{C}^+$, $\hbar_n(z) \overset{a.s.}{\to} \hbar(z)$. Finally, we let $z \downarrow 0 + 0i$ and obtain almost surely $\hbar_n(0) \to \hbar(0)$.

We remark that this proof approach is common in RMT to obtain the limiting special distribution (LSD) of random matrices. Thus, the present paper can be viewed as an application of RMT in multivariate statistical analysis. Moreover, since the type of matrix $\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y})^{-1}\mathbf{Y}'$ is special, and to the best of our knowledge, no known conclusions in RMT can be applied directly to obtain the limit of $\mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t$, we have to derive some new theoretical results for our theorems.

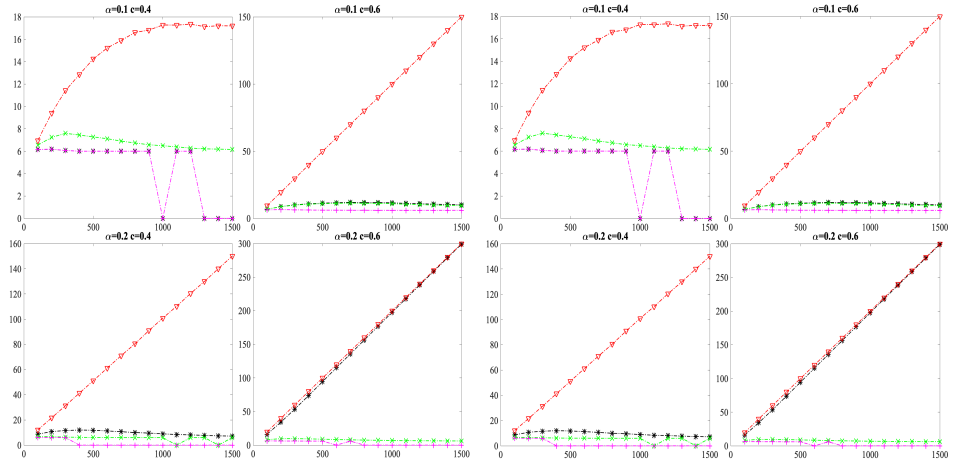


(a) Setting I



(b) Setting II

FIG 3. *Selection percentages under AIC, BIC and C_p for Settings I and II with a standard normal distribution. The horizontal axes represent the sample size n, and the vertical axes represent the selection percentages. Black solid circles, blue solid triangles, red solid squares, green solid hexagrams and magenta solid pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A = \mathbf{j}_*$, $\tilde{\mathbf{j}}_B = \mathbf{j}_*$, $\tilde{\mathbf{j}}_C = \mathbf{j}_*$, $\breve{\mathbf{j}}_A = \mathbf{j}_*$ and $\breve{\mathbf{j}}_C = \mathbf{j}_*$, respectively. Correspondingly, black circles, blue triangles, red squares, green hexagrams, and magenta pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_B \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_-$, $\breve{\mathbf{j}}_A \in \mathbf{J}_-$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_-$, respectively.*



(a) Setting I (b) Setting II

FIG 4. *Overspecified model sizes of AIC and C_p for Settings I and II with a standard normal distribution. The horizontal axes represent the sample size n, and the vertical axes represent the model size. Black asterisks, red right-pointing triangles, green crosses and magenta plus signs denote the average sizes of $\tilde{\mathbf{j}}_A \in \mathbf{J}_+$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_+$, $\breve{\mathbf{j}}_A \in \mathbf{J}_+$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_+$, respectively. In this figure, we let $0/0 = 0$.*

5.1. *An auxiliary lemma.* Before starting the proof of our main results, we introduce some basic results from RMT. For any $n \times n$ matrix $\mathbf{A}_n$ with only positive eigenvalues, let $F^{\mathbf{A}_n}$ be the empirical spectral distribution function of $\mathbf{A}_n$, that is,

$$F^{\mathbf{A}_n}(x) = \frac{1}{n}\#\{\lambda_i^{\mathbf{A}_n} \le x\},$$

where $\lambda_i^{\mathbf{A}_n}$ denotes the i-th largest eigenvalue of $\mathbf{A}_n$ and $\#\{\cdot\}$ denotes the cardinality of the set $\{\cdot\}$. If $F^{\mathbf{A}_n}$ has a limit distribution F, then we call it the LSD of sequence $\{\mathbf{A}_n\}$. For any function of bounded variation G on the real line, its Stieltjes transform is defined by

$$s_G(z) = \int \frac{1}{\lambda - z} dG(\lambda), \quad z \in \mathbb{C}^+.$$

If matrix $\mathbf{A}$ is invertible and for any $p \times n$ matrix $\mathbf{C}$, the following formulas will be used frequently,

$$(\mathbf{A} - \mathbf{C}\mathbf{C}')^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{I} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1}, \tag{5.2}$$

which immediately implies

$$(\mathbf{A} - \mathbf{C}\mathbf{C}')^{-1}\mathbf{C} = \mathbf{A}^{-1}\mathbf{C}(\mathbf{I} - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1} \tag{5.3}$$

$$\mathbf{C}'(\mathbf{A}-\mathbf{C}\mathbf{C}')^{-1}=(\mathbf{I}-\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1}. \tag{5.4}$$

For any $z\in\mathbb{C}^+$, we also have

$$\mathbf{C}(\mathbf{C}'\mathbf{C}-z\mathbf{I}_n)^{-1}\mathbf{C}'=\mathbf{I}_p+z(\mathbf{C}\mathbf{C}'-z\mathbf{I}_p)^{-1}, \tag{5.5}$$

which is called the in-out-exchange formula in the sequel. The above equations are straightforward to obtain by basic linear algebra theory; thus, we omit the detailed calculations.

A key tool for the proofs of Lemma 3.1 and Lemma 3.4 is the following lemma, whose proof will be postponed to the Appendix.

LEMMA 5.1. *Let* $\mathbf{M}:=\mathbf{M}(z)=p^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}-z\mathbf{I}_p$, $\boldsymbol{\alpha}_1$ *and* $\boldsymbol{\alpha}_2$ *be* $n\times 1$*-vectors, and* $\boldsymbol{\alpha}_3$ *be a* $p\times 1$*-vector and assume that* $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ *are all bounded in Euclidean norm. Then, under assumptions* (A1) through (A4) *and for any integer* t*, we have that for any* $z\in\mathbb{C}$*,*

$$\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2+\frac{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2}{z(1+\underline{s}_t(z)-\frac{1-c-\alpha_{m-t}}{cz})}\overset{a.s.}{\to}0, \tag{5.6}$$

$$\frac{1}{\sqrt{p}}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_3\overset{a.s.}{\to}0, \tag{5.7}$$

and

$$\begin{aligned}&\frac{1}{p}\boldsymbol{\alpha}_1'\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2\\&+\frac{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2}{z(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})}+\frac{\frac{1}{\underline{s}_t(z)+1}\boldsymbol{\alpha}_1'\mathbf{Q}_{\mathbf{j}_t}\boldsymbol{\alpha}_2}{z^2(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})^2}\overset{a.s.}{\to}0,\end{aligned} \tag{5.8}$$

where $\underline{s}_t(z)$ *is the Stieltjes transform of the LSD of* $\frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}$*.*

REMARK 5.2. *From (1.4) in (Silverstein and Choi, 1995), we have that the Stieltjes transform* $\underline{s}_t(z)$ *is the unique solution on the upper complex plane to the equation*

$$z=-\frac{1}{\underline{s}_t(z)}+\frac{1}{c}\int\frac{t}{1+t\underline{s}_t(z)}dH(t),$$

where H *is the LSD of* $\mathbf{Q}_{\mathbf{j}_{-t}}$*. Thus, we obtain that* $H(\{0\})=\alpha_{m-t}$, $H(\{1\})=1-\alpha_{m-t}$ *and*

$$z=-\frac{1}{\underline{s}_t}+\frac{1}{c}\frac{1-\alpha_{m-t}}{1+\underline{s}_t},$$

which implies

$$\frac{z(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})+\frac{1}{\underline{s}_t(z)+1}}{z^2(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})^2}=\frac{1}{1+\underline{s}(z)}-1 \tag{5.9}$$

and

$$\underline{s}_t(z)=\frac{1-\alpha_{m-t}-c-cz\pm\sqrt{(1-\alpha_{m-t}+c-cz)^2-4c(1-\alpha_{m-t})}}{2cz}.$$

On the basis of the fact that any Stieltjes transform tends to zero as $z\to\infty$, we have

$$\underline{s}_t(z)=\frac{1-\alpha_{m-t}-c-cz+\sqrt{(1-\alpha_{m-t}+c-cz)^2-4c(1-\alpha_{m-t})}}{2cz}$$

and

$$1-\frac{1}{1+\underline{s}_t(z)}=\frac{1-\alpha_{m-t}+c-cz+\sqrt{(1-\alpha_{m-t}+c-cz)^2-4c(1-\alpha_{m-t})}}{2(1-\alpha_{m-t}).}$$

Letting $z\downarrow 0+0i$ and together with (5.14) and $1-\alpha_{m-t}-c>0$, we conclude that

$$\underline{s}_t(z)\to\frac{c}{1-\alpha_{m-t}-c} \tag{5.10}$$

and

$$z\left(1+\underline{s}_t(z)-\frac{1-c-\alpha_{m-t}}{cz}\right)\to-\frac{1-\alpha_{m-t}-c}{c}. \tag{5.11}$$

Here, we have used the fact that when the imaginary part of the square root of a complex number is positive, then its real part has the same sign as the imaginary part; thus,

$$\lim_{z\downarrow 0+i0}\sqrt{(1-\alpha_{m-t}+c-cz)^2-4c(1-\alpha_{m-t})}=-|1-\alpha_{m-t}-c|.$$

Now, we are in position to prove Lemma 3.1 and Lemma 3.4.

5.2. *Proof of Lemma 3.1.* In this subsection, we present the proof of Lemma 3.1.

PROOF. We first prove (3.7). By equation (5.1) and the fact that for $\mathbf{j} \in \mathbf{J}_+$,

$$|\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}| = |\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}|,$$

we have

$$\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}}|}\right) = \log(1 - \mathbf{a}'_{t+1}\mathbf{E}(\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_{t+1}) \tag{5.12}$$

and

$$n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t}} - n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-t-1}} = -\mathbf{E}'\mathbf{a}_{t+1}\mathbf{a}'_{t+1}\mathbf{E}. \tag{5.13}$$

It follows from (3.1) and (5.12) that

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \sum_{t=1}^{m}[\log(1 - \mathbf{a}'_t\mathbf{E}(\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t) + 2p/n].$$

Since $\mathbf{a}_t$ is an eigenvector of $\mathbf{Q}_{\mathbf{j}_{-t}}$, we have $\mathbf{a}'_t\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{a}_t = 1$, which together with Lemma 5.1 and (5.9) implies

$$\frac{1}{p}\mathbf{a}'_t\mathbf{E}(\frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E} - z\mathbf{I}_p)^{-1}\mathbf{E}'\mathbf{a}_t \overset{a.s.}{\to} 1 - \frac{1}{1 + \underline{s}_t(z)}. \tag{5.14}$$

Therefore, by (5.10) and as $n \to \infty$, we have

$$\frac{1}{n^2}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = n^{-1}\sum_{t=1}^{m}\left(\log(1 - \frac{c}{1 - \alpha_{m-t}}) + 2c + o_{a.s.}(1)\right). \tag{5.15}$$

If $m/n \to \alpha_m > 0$, then integration by parts indicates that (5.15) tends to

$$\int_0^{\alpha_m}(\log(1 - \frac{c}{1-t}) + 2c)dt = 2c\alpha_m + \log\left(\frac{(1-c)^{1-c}(1-\alpha_m)^{1-\alpha_m}}{(1-c-\alpha_m)^{1-c-\alpha_m}}\right),$$

which implies (3.7).

(3.8) is analogous; thus, we omit the details. Next, we prove (3.9). It follows from (3.3) and (5.13) that

$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = \sum_{t=1}^{m}\left((\frac{k}{n} - 1)\mathbf{a}'_t\mathbf{E}(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t + 2\frac{p}{n}\right). \tag{5.16}$$

By (5.8) and (5.11) and the fact that

$$\mathbf{a}'_t\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{a}_t = 0,$$

we have

$$\mathbf{a}_t'\mathbf{E}(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t = \frac{c}{1-\alpha-c} + o_{a.s.}(1),$$

which together with (5.16) implies

$$\frac{1}{n^2}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) \overset{a.s.}{\to} \frac{c\alpha_m(\alpha-1)}{1-\alpha-c} + 2c\alpha_m.$$

Thus, we complete the proof of Lemma 3.1. □

5.3. *Proof of Lemma 3.4.* We start with $\mathcal{A}_{-t}$. When $t < 0$,

$$\mathcal{A}_{-t} = \log\left(\frac{|\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t-1}}\mathbf{Y}|}\right) - 2p/n.$$

Note that the index set $\mathbf{j}_{-t-1}$ contains one index i_{-t} more than $\mathbf{j}_{-t}$; therefore, from the notation $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{x}_{i_{-t}}/\|\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{x}_{i_{-t}}\|$, we have

$$\mathcal{A}_{-t} = -\log(1 - \mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t) - \frac{2p}{n}. \tag{5.17}$$

To evaluate the limit of $\mathcal{A}_{-t}$, we consider

$$m_{nt} := m_{nt}(z) = -\log(1 - \frac{1}{p}\mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}/p - z\mathbf{I}_p)^{-1}\mathbf{Y}'\mathbf{a}_t) - \frac{2p}{n},$$

where $z \in \mathbb{C}^+$. On the basis of the fact that $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{a}_t$ and the in-out-exchange formula (5.5), we rewrite m_{nt} as

$$m_{nt} = -\log\left(-z\mathbf{a}_t'\left(\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}/p - z\mathbf{I}_n\right)^{-1}\mathbf{a}_t\right) - \frac{2p}{n}. \tag{5.18}$$

Substitute model (2.1) into the above equation and denote

$$\begin{aligned} I_t := I_t(z) &= z\mathbf{a}_t'\left(\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}/p - z\mathbf{I}_n\right)^{-1}\mathbf{a}_t \\ &= z\mathbf{a}_t'\left(\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l} + \mathbf{E})(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}' + \mathbf{E}')\mathbf{Q}_{\mathbf{j}_{-t}}/p - z\mathbf{I}_n\right)^{-1}\mathbf{a}_t \end{aligned}$$

where $\boldsymbol{\ell}_t = \{i_1, \cdots, i_{-t}\}$. Define $\mathbf{B}_1 = \mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l}(\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l})^{-1/2}$ and select $\mathbf{B}_2$ such that $\mathbf{B} = \left(\mathbf{B}_1 \vdots \mathbf{B}_2\right)$ is an $n \times n$ orthogonal matrix. Then, we have

$$I_t = z\mathbf{a}_t'\mathbf{B}\left(\mathbf{B}'\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l} + \mathbf{E})(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}' + \mathbf{E}')\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}/p - z\mathbf{I}_n\right)^{-1}\mathbf{B}'\mathbf{a}_t.$$

With $\mathbf{a}_t'\mathbf{B}_2 = 0$, we obtain

$$\begin{aligned} I_t =& z\tilde{\mathbf{a}}_t'\Big(\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}+\mathbf{E})(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'+\mathbf{E}')\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1/p \\ &- \mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}+\mathbf{E})\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_2(\mathbf{B}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_2/p - z\mathbf{I}_{n+t})^{-1} \\ &\cdot \mathbf{B}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}(\mathbf{E}'+\mathbf{X}_{\boldsymbol{\ell}_l}'\Theta_{\boldsymbol{\ell}_l})\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1/p^2 - z\mathbf{I}_{-t}\Big)^{-1}\tilde{\mathbf{a}}_t \end{aligned}$$

where $\tilde{\mathbf{a}}_t = \mathbf{B}_1'\mathbf{a}_t$. By applying in-out-exchange formula (5.5) to the term

$$\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_2(\mathbf{B}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_2/p - z\mathbf{I}_{n-k_t})^{-1}\mathbf{B}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}/p,$$

we obtain

$$\begin{aligned} I_t = -\tilde{\mathbf{a}}_t'\Big(&\frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}+\mathbf{E})(\frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_2\mathbf{B}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E} - z\mathbf{I}_p)^{-1} \\ &\times(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'+\mathbf{E}')\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1+\mathbf{I}_{-t}\Big)^{-1}\tilde{\mathbf{a}}_t, \end{aligned}$$

which together with $\mathbf{M} = \frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E} - z\mathbf{I}_p$ implies

$$\begin{aligned} I_t = -\tilde{\mathbf{a}}_t'\Big(&\frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}+\mathbf{E})(\mathbf{M} - \frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E})^{-1} \\ &\times(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'+\mathbf{E}')\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1+\mathbf{I}_{-t}\Big)^{-1}\tilde{\mathbf{a}}_t. \end{aligned}$$

Equations (5.2)-(5.5) can be used to separate I_t into the following four parts,

$$I_t = -\tilde{\mathbf{a}}_t'\left(I_{1t} + I_{2t} + I_{2t}' + I_{3t}\right)^{-1}\tilde{\mathbf{a}}_t \tag{5.19}$$

where

$$\begin{aligned} I_{1t} =& \frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\mathbf{M}^{-1}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1 + \frac{1}{p^2}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1 \\ &(\mathbf{I}_{-t} - \frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1)^{-1}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{M}^{-1}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1; \\ I_{2t} =& \frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1(\mathbf{I}_{-t} - \frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1)^{-1}; \\ I_{3t} =& (\mathbf{I}_{-t} - \frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1)^{-1}. \end{aligned}$$

It follows from (5.8) and (5.9) that

$$\frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1 \overset{a.s.}{\to} (1 - \frac{1}{1+\underline{s}(z)})\mathbf{I}_{-t},$$

which implies

$$I_{3t} \overset{a.s.}{\to} (1+\underline{s}(z))\mathbf{I}_{-t}. \tag{5.20}$$

Moreover, from (5.6), (5.7) and Assumption (A4), we have

$$\frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\mathbf{M}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{B}_1 \overset{a.s.}{\to} \mathbf{0}_{-t},$$

and

$$\frac{1}{p}\mathbf{B}_1'\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\mathbf{M}^{-1}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{B}_1 + \frac{p^{-1}\Delta_t}{z(1+\underline{s}(z)-\frac{1-c-\alpha_{m-t}}{cz})} \overset{a.s.}{\to} 0,$$

which together with (5.20) imply

$$I_t + \tilde{\mathbf{a}}_t'\left((1+\underline{s}(z))\mathbf{I}_{-t} - \frac{p^{-1}\Delta_t}{z(1+\underline{s}(z)-\frac{1-c-\alpha_{m-t}}{cz})}\right)^{-1}\tilde{\mathbf{a}}_t \overset{a.s.}{\to} 0.$$

As $z \downarrow 0+0i$ and with (5.10) and the notation

$$\delta_t := \tilde{\mathbf{a}}_t'\left((1-k_{\mathbf{j}_{-t}}/n)\mathbf{I}_{-t} + n^{-1}\Delta_t\right)^{-1}\tilde{\mathbf{a}}_t,$$

we have

$$I_t(0) + (1-\alpha_{m-t}-c)\delta_t \overset{a.s.}{\to} 0.$$

Therefore, we conclude that for $t<0$,

$$\mathcal{A}_{-t} + \log\delta_t + \log(1-\alpha_{m-t}-c) + 2c \overset{a.s.}{\to} 0.$$

Next, we consider $\mathcal{C}_{-t}$ when $t<0$. Recall that

$$\mathcal{C}_{-t} = (1-k/n)\mathbf{a}_t'\mathbf{Y}(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E})^{-1}\mathbf{Y}'\mathbf{a}_t - 2pn^{-1},$$

and let

$$J_t(z) = p^{-1}\mathbf{a}_t'\mathbf{Y}\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\mathbf{Y}'\mathbf{a}_t.$$

Then, by substituting model (2.1) into the above equation, we obtain

$$\begin{aligned}
J_t(z) =& \frac{1}{p}\mathbf{a}_t'(\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l} + \mathbf{E})\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}(\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}' + \mathbf{E}')\mathbf{a}_t \\
=& \frac{1}{p}\mathbf{a}_t'\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{a}_t \\
&+ \frac{1}{p}\mathbf{a}_t'\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\mathbf{E}'\mathbf{a}_t \\
&+ \frac{1}{p}\mathbf{a}_t'\mathbf{E}\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{a}_t \\
&+ \frac{1}{p}\mathbf{a}_t'\mathbf{E}\left(\mathbf{E}'\mathbf{Q}_{\boldsymbol{\omega}}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\mathbf{E}'\mathbf{a}_t \\
:=& J_{1t} + J_{2t} + J_{2t}' + J_{3t},
\end{aligned}$$

where $\boldsymbol{\ell}_t = \{i_{-t+1}, \cdots, i_s\}$. It follows from Lemma 5.1 that

$$J_{1t} + \frac{\frac{1}{p}\mathbf{a}_t'\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{a}_t}{z(1+\underline{s}(z)-\frac{1-c-\alpha}{cz})} \xrightarrow{a.s.} 0,$$

$$J_{2t} \xrightarrow{a.s.} 0 \quad \text{and} \quad J_{3t} + \frac{1}{z(1+\underline{s}(z)-\frac{1-c-\alpha}{cz})} \xrightarrow{a.s.} 0.$$

Note that

$$\begin{aligned}
&\frac{1}{p}\mathbf{a}_t'\mathbf{X}_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}'\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{a}_t \\
=& \frac{1}{p}\tilde{\mathbf{a}}_t'(\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_l})^{1/2}\Theta_{\boldsymbol{\ell}_l}\Theta_{\boldsymbol{\ell}_l}'(\mathbf{X}_{\boldsymbol{\ell}_l}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{X}_{\boldsymbol{\ell}_l})^{1/2}\tilde{\mathbf{a}}_t \\
=& c^{-1}\eta_t.
\end{aligned}$$

Therefore, letting $z \downarrow 0 + 0i$, we obtain

$$J_t(0) \xrightarrow{a.s.} \frac{\eta_t + c}{1-\alpha-c}.$$

Thus, we complete the proof of Lemma 3.4.

6. Conclusion and discussion. In the present paper, we discussed the strong consistency of three fundamental selection criteria, i.e., the AIC, BIC, and C_p, in the linear regression model under an LLL framework. We presented the sufficient and necessary conditions for their strong consistency and determined how the dimension and size of the explanatory variables

and the sample size affect the selection accuracy. Then, we proposed general KOO criteria based on KOO methods and showed the sufficient conditions for their strong consistency. The general KOO criteria have numerous advantages, such as simplicity of expression, ease of computation, limited restrictions, and fast convergence.

The present paper considers only the case in which $\alpha + c < 1$ because $\hat{\mathbf{\Sigma}}_{\mathbf{j}}$ may otherwise be singular. The singularity of $\hat{\mathbf{\Sigma}}_{\mathbf{j}}$ can be avoided by using a ridge-type estimator of the covariance matrix (e.g., (Yamamura et al., 2010; Chen et al., 2011)) or by rewriting the statistics with nonzero eigenvalues (e.g., Zhang et al. (2017)). The consistency properties of the model selection criteria must also be studied when the true model size k_* is large, i.e., k_*/n tends to a constant as $n \to \infty$. All three topics require clarification of the theoretical results of RMT, which is left for future research.

In addition, we intend to obtain the asymptotic distributions of $\breve{A}_j$ and $\breve{C}_j$. In the present paper, we obtained only almost surely the limits of $\breve{A}_j$ and $\breve{C}_j$ and not their convergence rates. If we can determine their asymptotic distributions, such results can be used to construct more reasonable selection criteria. According to the results of (Bai et al., 2007), we guessed that the convergence rate should be $O(n^{-1/2})$.

The main technical tool of the present paper is RMT. In the past two decades, the power of RMT has been partially demonstrated through high-dimensional multivariate analysis. Most subjects in classical multivariate analysis, including the model selection problems considered in this paper, can be (or have been) reexamined by RMT in high-dimensional settings. We hope that RMT will attract more attention in the future research of high-dimensional MLR.

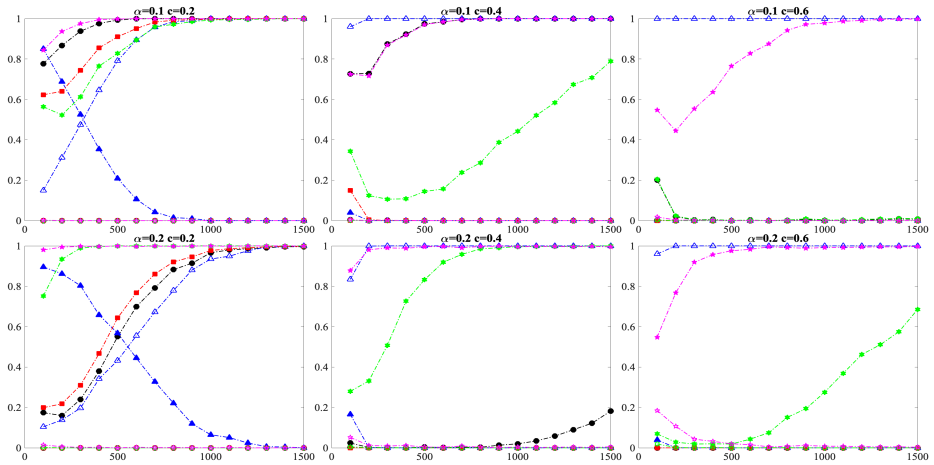
APPENDIX A: SIMULATION RESULTS UNDER NON-NORMAL DISTRIBUTIONS

In this section, we present simulation results under a standardized t distribution with three degrees of freedom and a standardized chi-square distribution with two degrees of freedom. Please see Figures 5-8. These results are similar to those of the normal distribution case. Thus, we guess that the consistency of these estimators depends on only the first two moments.

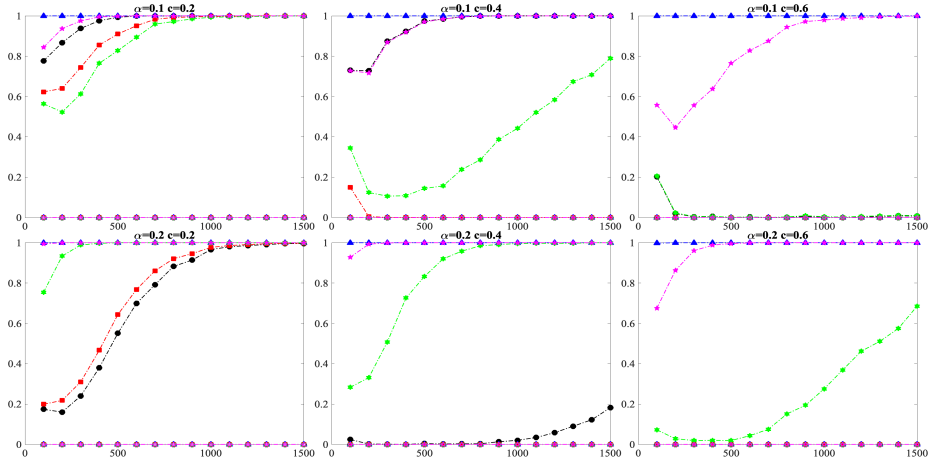
APPENDIX B: PROOF OF LEMMA 5.1

We now present the proof of Lemma 5.1. In the following, C represents a generic constant whose value may vary from line to line.

Proof of Lemma 5.1. According to the truncation approach of Bai

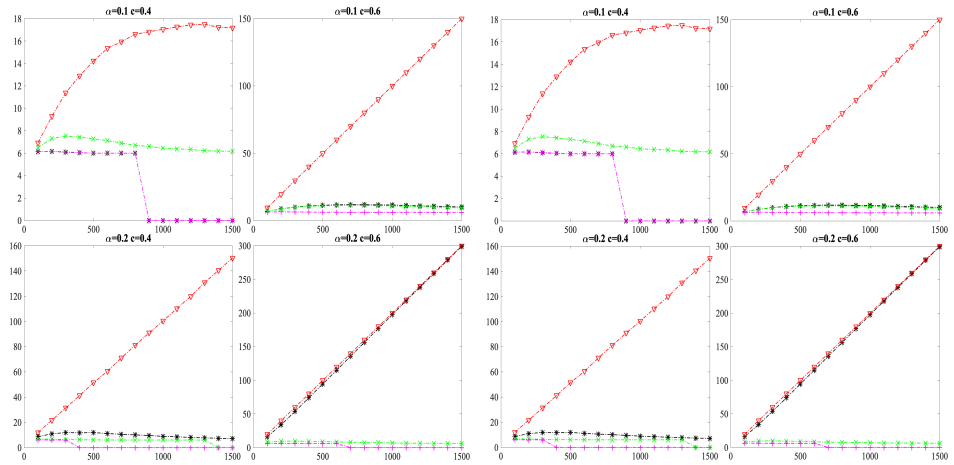


(a) Setting I



(b) Setting II

FIG 5. *Selection percentages under the AIC, BIC and C_p for Settings I and II with a standard t_3 distribution. The horizontal axes represent the sample size n, and the vertical axes represent selection percentage. Black solid circles, blue solid triangles, red solid squares, green solid hexagrams and magenta solid pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A = \mathbf{j}_*$, $\tilde{\mathbf{j}}_B = \mathbf{j}_*$, $\tilde{\mathbf{j}}_C = \mathbf{j}_*$, $\breve{\mathbf{j}}_A = \mathbf{j}_*$ and $\breve{\mathbf{j}}_C = \mathbf{j}_*$, respectively. Correspondingly, black circles, blue triangles, red squares, green hexagrams and magenta pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_B \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_-$, $\breve{\mathbf{j}}_A \in \mathbf{J}_-$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_-$, respectively.*



(a) Setting I (b) Setting II

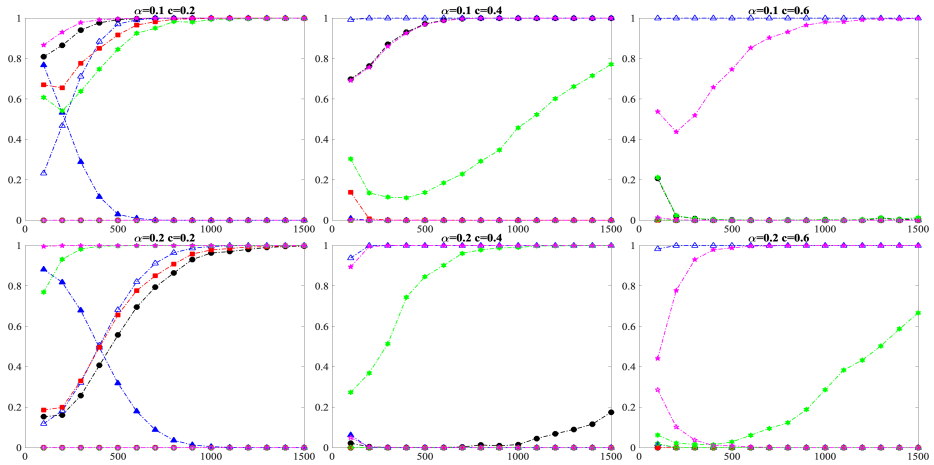
FIG 6. *Overspecified model sizes of the AIC and C_p for Settings I and II with a standard t_3 distribution. The horizontal axes represent the sample size n, and the vertical axes represent the model size. Black asterisks, red right-pointing triangles, green crosses and magenta plus signs denote the average sizes of $\tilde{\mathbf{j}}_A \in \mathbf{J}_+$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_+$, $\breve{\mathbf{j}}_A \in \mathbf{J}_+$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_+$, respectively. In this figure, we let $0/0 = 0$.*

et al. (2007), we can assume that the variables $\{e_{ij}, i = 1 \dots n, j = 1 \dots p\}$ satisfy the following additional condition:

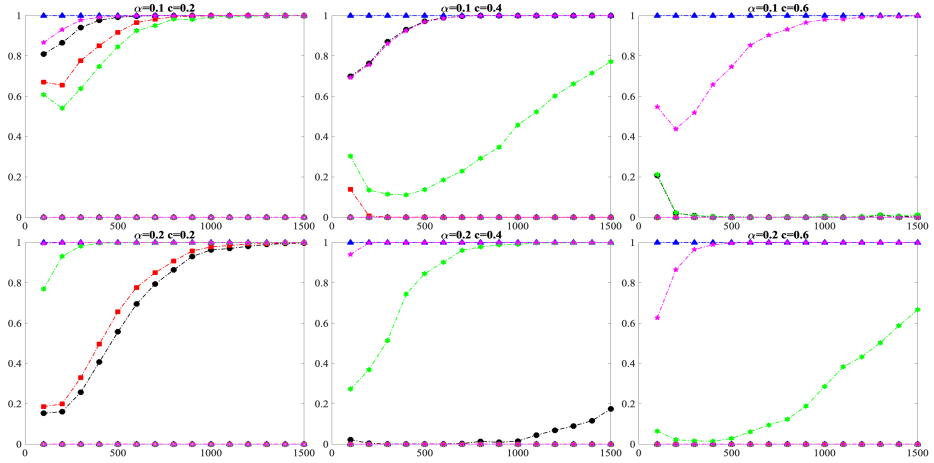
$$|e_{ij}| < C, \quad \text{for all } i, j \text{ and sufficiently large constant } C.$$

As the proof is very similar, we omit the details here. Let $\boldsymbol{\alpha}_{k1}$ be the $p-1$ subvector of $\boldsymbol{\alpha}_1$ with the k-th entry removed, and let α_{k1} be the k-th entry of $\boldsymbol{\alpha}_1$. Analogously, we can define $\boldsymbol{\alpha}_{k2}$ and α_{k2}. Define $\mathbf{M}_k = \frac{1}{p}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k - z\mathbf{I}_{p-1}$, where $\mathbf{E}_k$ is the $n \times (p-1)$ submatrix of $\mathbf{E}$ with the k-th column removed. Denote by $\mathbb{E}_k$ the conditional expectation given $\{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$ and by $\mathbb{E}_0$ the unconditional expectation, where $\mathbf{e}_i$ is the n-vector of the i-th column of $\mathbf{E}$. Then, by inverting the block matrix, we obtain

$$\begin{aligned}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 &= \boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2} + \frac{1}{\beta_k p^2}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2} \\ &\quad -\frac{\alpha_{k2}}{\beta_k p}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k - \frac{\alpha_{k1}}{\beta_k p}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2} + \frac{\alpha_{k1}\alpha_{k2}}{\beta_k}, \end{aligned} \tag{B.1}$$

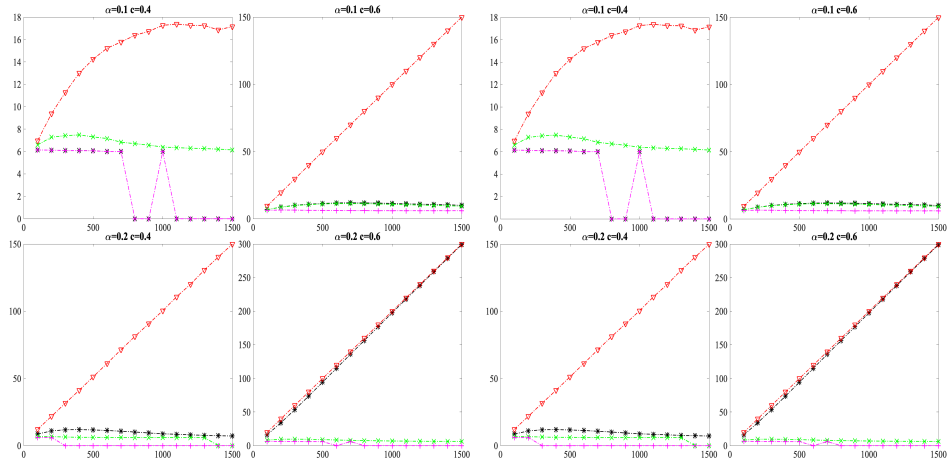


(a) Setting I



(b) Setting II

FIG 7. *Selection percentages under the AIC, BIC and C_p for Settings I and II with standard χ^2_2 distribution. The horizontal axes represent the sample size n, and the vertical axes represent the selection percentages. Black solid circles, blue solid triangles, red solid squares, green solid hexagrams and magenta solid pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A = \mathbf{j}_*$, $\tilde{\mathbf{j}}_B = \mathbf{j}_*$, $\tilde{\mathbf{j}}_C = \mathbf{j}_*$, $\breve{\mathbf{j}}_A = \mathbf{j}_*$ and $\breve{\mathbf{j}}_C = \mathbf{j}_*$, respectively. Correspondingly, black circles, blue triangles, red squares, green hexagrams and magenta pentagrams denote the selection percentages of $\tilde{\mathbf{j}}_A \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_B \in \mathbf{J}_-$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_-$, $\breve{\mathbf{j}}_A \in \mathbf{J}_-$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_-$, respectively.*



(a) Setting I (b) Setting II

FIG 8. *Overspecified model sizes of the AIC and C_p for Settings I and II with a standard chi_2^2 distribution. The horizontal axes represent the sample size n, and the vertical axes represent the model size. Black asterisks, red right-pointing triangles, green crosses and magenta plus signs denote the average sizes of $\tilde{\mathbf{j}}_A \in \mathbf{J}_+$, $\tilde{\mathbf{j}}_C \in \mathbf{J}_+$, $\breve{\mathbf{j}}_A \in \mathbf{J}_+$ and $\breve{\mathbf{j}}_C \in \mathbf{J}_+$, respectively. In this figure we let $0/0 = 0$.*

where

$$\begin{aligned}\beta_k &= \frac{1}{p}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k - z - \frac{1}{p^2}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k \\ &= -z(1 + \frac{1}{p}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_k^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k).\end{aligned}$$

The last equation is from the in-out-exchange formula (5.5) and

$$\widehat{\mathbf{M}}_k := \frac{1}{p}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}} - z\mathbf{I}_n.$$

Denote $\beta_k^{tr} = -z[1 + \frac{1}{p}tr(\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_k^{-1})]$. It follows that

$$\frac{1}{\beta_k} = \frac{1}{\beta_k^{tr}} + \frac{\beta_k - \beta_k^{tr}}{p\beta_k\beta_k^{tr}} = \frac{1}{\beta_k^{tr}} + \frac{\xi_k}{\beta_k\beta_k^{tr}}, \tag{B.2}$$

where $\xi_k = p^{-1}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_k^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k - p^{-1}tr\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_k^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}$. It follows from

(B.1) that

$$\begin{aligned}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 - \mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 &= \sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 \\ &= \sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})(\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2}) \\ &= \sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})(\mathcal{M}_1 - \mathcal{M}_2 - \mathcal{M}_3 + \mathcal{M}_4),\end{aligned}$$

where

$$\mathcal{M}_1 = \frac{1}{\beta_k p^2}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2},\ \mathcal{M}_2 = \frac{\alpha_{k2}}{\beta_k p}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k$$
$$\mathcal{M}_3 = \frac{\alpha_{k1}}{\beta_k p}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2},\ \mathcal{M}_4 = \frac{\alpha_{k1}\alpha_{k2}}{\beta_k}.$$

Note that for any fixed $z \in \mathbb{C}^+$, we have

$$\min\{|\beta_k|,\ |\beta_k^{tr}|,\ \|\mathbf{M}_k\|,\ \|\widehat{\mathbf{M}}_k\|\} \geq \Im z = v > 0,$$

and $\|p^{-1/2}\mathbf{E}_k\|$ is almost surely bounded by a constant under Assumption (A2) (see (Bai and Silverstein, 1998, 1999), for example). Thus, together with the condition that $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are both bounded in Euclidean norm, we conclude that for $m \geq 1$

$$\mathbb{E}|\xi_k|^{2m} \leq Cp^{-m}v^{-2m}, \tag{B.3}$$

$$\mathbb{E}|\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}, \mathbf{e}_k\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2}|^m \leq Cp^m v^{-2m} \tag{B.4}$$

(B.5)
$$\max\{\mathbb{E}|\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k|^{2m}, \mathbb{E}|\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\boldsymbol{\alpha}_{k2}|^{2m}\} \leq Cp^m v^{-2m}.$$

Here, we use the quadric form inequality shown in Lemma 2.7 of Bai and Silverstein (1998). Thus, by the Burkholder inequality (see Lemma 2.1 in (Bai and Silverstein, 1998)), we obtain

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})\mathcal{M}_1\right|^4 &\leq C\mathbb{E}\left(\sum_{k=1}^{p}\mathbb{E}_{k-1}|\mathcal{M}_1|^2\right)^2 + C\mathbb{E}\sum_{k=1}^{p}|\mathcal{M}_1|^4 \\ &= O(p^{-2}).\end{aligned}$$

Analogously, we also have

$$\mathbb{E}\left|\sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})\mathcal{M}_2\right|^4 = O(p^{-2}) \text{ and } \mathbb{E}\left|\sum_{k=1}^{p}(\mathbb{E}_k - \mathbb{E}_{k-1})\mathcal{M}_3\right|^4 = O(p^{-2}).$$

For $\mathcal{M}_4$, by (B.3), (B.2) and the Burkholder inequality, we obtain

$$\mathbb{E}\left|\sum_{k=1}^{p}(\mathbb{E}_k-\mathbb{E}_{k-1})\mathcal{M}_4\right|^4=\mathbb{E}\left|\sum_{k=1}^{p}(\mathbb{E}_k-\mathbb{E}_{k-1})\frac{\xi_k}{\beta_k\beta_k^{tr}}\right|^4=O(p^{-2}),$$

which finally implies

$$\mathbb{E}|\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2-\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2|^4=O(p^{-2}).$$

Therefore, by the Borel-Cantelli lemma, we have that

$$\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2-\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2\overset{a.s.}{\to}0. \tag{B.6}$$

Because the entry distributions of $\mathbf{E}$ are identical, we know that the diagonal elements of $\mathbb{E}\mathbf{M}^{-1}$ are the same, denoted as $a_n(z)$ in the following. The off-diagonal elements are also the same, denoted as $b_n(z)$. From (B.1), (B.2) and (B.3), we have

$$a_n(z)=\mathbb{E}\frac{1}{\beta_1}=\mathbb{E}\frac{1}{\beta_1^{tr}}+o(1). \tag{B.7}$$

Let $\lambda_1\geq\cdots\geq\lambda_n\geq 0$ be the eigenvalues of $p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}$ whose corresponding eigenvectors are denoted by $\mathbf{v}_1,\cdots,\mathbf{v}_n$. Note that the rank of $\mathbf{Q}_{\mathbf{j}_{-t}}$ is $n-k_{\mathbf{j}_{-t}}$, the dimension of the null space of $\mathbf{Q}_{\mathbf{j}_{-t}}$ is $k_{\mathbf{j}_{-t}}$, and the null space of $p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}$ is not smaller than that of $\mathbf{Q}_{\mathbf{j}_{-t}}$. Thus, we may select the last $k_{\mathbf{j}_{-t}}$ eigenvectors from the null space of $\mathbf{Q}_{\mathbf{j}_{-t}}$, that is, we may assume that $\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i=0$, $i=n-k_{\mathbf{j}_{-t}}+1,\cdots,n$. Therefore, the matrix $\widehat{\mathbf{M}}_1^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}$ has $k_{\mathbf{j}_{-t}}$-fold eigenvalue 0. By contrast, for all $i=1,\cdots,n-k_{\mathbf{j}_{-t}}$, $\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i\neq 0$ and $\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i$ is also an eigenvector of $p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}$ corresponding to λ_i, that is,

$$p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i=\lambda_i\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i,$$

which is equivalent to

$$(p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}-z\mathbf{I}_n)^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i=(\lambda_i-z)^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{v}_i.$$

Thus, $(\lambda_i-z)^{-1}$ is a nonzero eigenvalue of $\widehat{\mathbf{M}}_1^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}$. Therefore,

$$\frac{1}{p}\mathrm{tr}\widehat{\mathbf{M}}_1^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}=\frac{1}{p}\sum_{i=1}^{n-k_{\mathbf{j}_{-t}}}\frac{1}{\lambda_i-z}=\frac{1}{p}\sum_{i=1}^{n}\frac{1}{\lambda_i-z}+\frac{k_{\mathbf{j}_{-t}}}{pz}=\frac{1}{p}tr\widehat{\mathbf{M}}_1^{-1}+\frac{k_{\mathbf{j}_{-t}}}{pz}.$$

Since the spectra of $p^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}$ and $p^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_1$ differ by $|n-p-1|$ zero eigenvalues, it follows that

$$\mathbb{E}\frac{1}{p}tr\widehat{\mathbf{M}}_1^{-1} = \mathbb{E}\frac{1}{p}tr\mathbf{M}_1^{-1} + \frac{1-n/p}{z} \to \underline{s}_t(z) + \frac{c-1}{cz}.$$

Together with (B.7) and the last equation, we obtain

$$a_n(z) \to -\frac{1}{z(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})}. \tag{B.8}$$

Here, $\alpha_{m-t} = \lim k_{\mathbf{j}_{-t}}/n$.

Furthermore, from the inverse matrix formula, we obtain

$$b_n(z) \quad = \quad \mathbb{E}\left(\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1}{p\beta_1}\right),$$

where $\mathbf{u}_1 = (1,0,\cdots,0)'$ is a $(p-1)$-dimensional vector, and $\mathbf{e}_1$ is the first column of $\mathbf{E}$. Because $\mathbf{e}_1$ is independent of $\mathbf{E}_1$, by the Cauchy-Schwarz inequality and equation (B.2), we have

$$|b_n(z)| = \left|\mathbb{E}\left(\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\xi_1}{p\beta_1\beta_1^{tr}}\right)\right| = O(p^{-1}). \tag{B.9}$$

Therefore, by combining (B.6)-(B.9) and the C_r inequality, the proof of (5.6) is complete.

For the proof of (5.7), using the inverse of block matrix again, we obtain

$$\begin{aligned}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 &= \boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\boldsymbol{\alpha}_2 + \frac{1}{\beta_k p^2}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\mathbf{E}_k'\boldsymbol{\alpha}_2 \\ &\quad -\frac{\alpha_{k2}}{\beta_k p}\boldsymbol{\alpha}_{k1}'\mathbf{M}_k^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_k - \frac{\alpha_{k1}}{\beta_k p}\mathbf{e}_k'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_k\mathbf{M}_k^{-1}\mathbf{E}_k'\boldsymbol{\alpha}_2 + \frac{\alpha_{k1}\mathbf{e}_k'\boldsymbol{\alpha}_2}{\beta_k}.\end{aligned}$$

By means of the same procedure used in the proof of (5.6), we can obtain that, almost surely,

$$\frac{1}{\sqrt{p}}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 - \frac{1}{\sqrt{p}}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 \to 0.$$

Next, we show that $\frac{1}{\sqrt{p}}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 = o(1)$. Write $\mathbf{M}^{-1} = (M^{ij})$. Then,

we have

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 = \frac{1}{\sqrt{p}}\sum_{ij}\alpha_{i1}\mathbb{E}M^{ij}\mathbf{e}_j'\boldsymbol{\alpha}_2\\
=&\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\alpha_{i1}\left(\mathbb{E}M^{11}\mathbf{e}_1'\boldsymbol{\alpha}_2+(p-1)\mathbb{E}M^{12}\mathbf{e}_2'\boldsymbol{\alpha}_2\right)\\
=&\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\alpha_{i1}\left(\mathbb{E}\frac{\mathbf{e}_1'\boldsymbol{\alpha}_2}{\beta_1}-(p-1)\mathbb{E}\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2}{p\beta_1}\right)\\
=&\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\alpha_{i1}\left(\mathbb{E}\frac{\xi_1\mathbf{e}_1'\boldsymbol{\alpha}_2}{\beta_1\beta_1^{tr}}-(p-1)\mathbb{E}\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{p\beta_1\beta_1^{tr}}\right)\\
=&\frac{1}{p^{3/2}}\sum_{i=1}^{p}\alpha_{i1}\sum_{j=1}^{n}\mathbb{E}\frac{e_{j1}^3(\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_1^{-1}\mathbf{Q}_{\mathbf{j}_{-t}})_{jj}\alpha_{j1}}{(\beta_1^{tr})^2}\\
&-\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\alpha_{i1}\mathbb{E}\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{(\beta_1^{tr})^2}+o(1)\\
=&o(1),
\end{aligned}$$

which completes the proof of (5.7). In the above equation, we use (B.3), the Cauchy-Schwarz inequality, the C_r inequality and the facts that

$$\sum_{i=1}^{p}|\alpha_{i1}|\le\left(p\sum_{i=1}^{p}|\alpha_{i1}^2|\right)^{1/2}=O(\sqrt{p})$$

and

$$\begin{aligned}
&\left|\mathbb{E}\frac{\mathbf{u}_2'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{\sqrt{p}\beta_1\beta_1^{tr}}\right|\\
\le&\left|\mathbb{E}\frac{\mathbf{u}_2'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{\sqrt{p}(\beta_1^{tr})^2}\right|+\mathbb{E}\left|\frac{(\mathbf{u}_2'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2)\xi_1^2}{\sqrt{p}(\beta_1^{tr})^2\beta_1}\right|\\
=&O(p^{-1/2}).
\end{aligned}$$

Next, we give the proof of (5.8). The random part is analogous to the proof of (5.6), which is, almost surely

$$\frac{1}{p}\boldsymbol{\alpha}_1'\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2-\frac{1}{p}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2\to 0,$$

and we omit the details here. Next, we focus on the nonrandom part. By means of the inverse of the block matrix, we have

$$\begin{aligned}
&\frac{1}{p}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_2 = \frac{1}{p}\sum_{ij}\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{e}_i M^{ij}\mathbf{e}_j'\boldsymbol{\alpha}_2 \\
=&\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{e}_1 M^{11}\mathbf{e}_1'\boldsymbol{\alpha}_2 + (p-1)\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{e}_1 M^{12}\mathbf{e}_2'\boldsymbol{\alpha}_2) \\
=&\mathbb{E}\frac{\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{e}_1'\boldsymbol{\alpha}_2}{\beta_1} - (p-1)\mathbb{E}\frac{\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2}{p\beta_1} \\
=&\mathbb{E}\frac{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2}{\beta_1^{tr}} - (p-1)\mathbb{E}\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_1\mathbf{e}_2'\boldsymbol{\alpha}_2}{p\beta_1^{tr}} \\
&+\mathbb{E}\frac{\xi_1\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{e}_1'\boldsymbol{\alpha}_2}{\beta_1\beta_1^{tr}} - (p-1)\mathbb{E}\frac{\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{p\beta_1\beta_1^{tr}}.
\end{aligned}$$

It follows from (B.8) that

$$\mathbb{E}\frac{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2}{\beta_1^{tr}} \to -\frac{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2}{z(1+\underline{s}_t(z)+\frac{c-1+\alpha_{m-t}}{cz})}. \tag{B.10}$$

Using the inverse of block matrix again, we obtain

$$\begin{aligned}
&\mathbb{E}\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_1\mathbf{e}_2'\boldsymbol{\alpha}_2 \\
=&\ \mathbb{E}\frac{\mathbf{e}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_1\mathbf{e}_2'\boldsymbol{\alpha}_2}{\beta_{12}} - \mathbb{E}\frac{1}{\beta_{12}p}\mathbf{e}_2'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_{12}\mathbf{M}_{12}^{-1}\mathbf{E}_{12}'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_1\mathbf{e}_2'\boldsymbol{\alpha}_2,
\end{aligned}$$

which together with (B.2), (B.3) and in-out-exchange foumula (5.5) implies

$$\begin{aligned}
&\mathbb{E}\frac{\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_1\mathbf{e}_2'\boldsymbol{\alpha}_2}{\beta_1^{tr}} \\
=&\mathbb{E}\frac{\boldsymbol{\alpha}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2}{(\beta_{12}^{tr})^2} - \mathbb{E}\frac{\boldsymbol{\alpha}_1\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{E}_{12}\mathbf{M}_{12}^{-1}\mathbf{E}_{12}'\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2}{p(\beta_{12}^{tr})^2} + o(1) \\
=&-\mathbb{E}\frac{z\boldsymbol{\alpha}_1\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_{12}^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2}{(\beta_{12}^{tr})^2} + o(1).
\end{aligned}$$

Here, $\mathbf{E}_{12}$ is the $n\times(p-2)$ submatrix of $\mathbf{E}$ with the first and second columns removed. β_{12}, β_{12}^{tr}, $\mathbf{M}_{12}^{-1}$ and $\widehat{\mathbf{M}}_{12}^{-1}$ are denoted analogously. Note that $\mathbb{E}\frac{\xi_1\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{e}_1'\boldsymbol{\alpha}_2}{\beta_1\beta_1^{tr}}$ and $\mathbb{E}\frac{\boldsymbol{\alpha}_1'\mathbf{e}_1\mathbf{u}_1'\mathbf{M}_1^{-1}\mathbf{E}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\alpha}_2\xi_1}{\beta_1\beta_1^{tr}}$ are both $o(1)$ because of (B.2) and (B.3). Similar to (5.9) in Bai et al. (2007), we have

$$\mathbb{E}\boldsymbol{\alpha}_1'\mathbf{Q}_{\mathbf{j}_{-t}}\widehat{\mathbf{M}}_{12}^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1\mathbf{Q}_{\mathbf{j}_{-t}}(-z\underline{s}_t(z)\mathbf{Q}_{\mathbf{j}_{-t}} - z\mathbf{I})^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2 \to 0.$$

As $\boldsymbol{\alpha}_1'\mathbf{Q}_{\mathbf{j}_{-t}}$ and $\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2$ are both eigenvectors of $\mathbf{Q}_{\mathbf{j}_{-t}}$, we obtain

$$\boldsymbol{\alpha}_1\mathbf{Q}_{\mathbf{j}_{-t}}(z\underline{s}_t(z)\mathbf{Q}_{\mathbf{j}_{-t}}+z\mathbf{I})^{-1}\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2=\frac{\boldsymbol{\alpha}_1\mathbf{Q}_{\mathbf{j}_{-t}}\boldsymbol{\alpha}_2}{z\underline{s}_t(z)+z},$$

which together with (B.10) and the fact that

$$\mathbb{E}\beta_1^{-1}-\mathbb{E}\beta_{12}^{-1}\to 0,$$

completes the proof of (5.8). Thus, the proof of this lemma is complete. □

REFERENCES

H. Akaike. Information theory and an extension of the maximum likehood principle. In *Second International Symposium on Information Theory*, pages 267–281. 1973. ISBN 0012-9682. .

M. J. Anzanello and F. S. Fogliatto. A review of recent variable selection methods in industrial and chemometrics applications. *European J. of Industrial Engineering*, 8(5): 619, 2014.

Z. D. Bai, K. P. Choi, and Y. Fujikoshi. Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, 46(3):1050–1076, jun 2018.

Z. D. Bai and J. W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, Jan. 1998.

Z. D. Bai and J. W. Silverstein. Exact separation of eigenvalues of large dimensional sample covariance matrices. *The Annals of Probability*, 27(3):1536–1555, 1999.

Z. D. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices. Second Edition.* Springer Verlag, 2010.

Z. D. Bai, B. Q. Miao, and G. M. Pan. On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572, July 2007.

Z. Bao, J. Hu, G. Pan, and W. Zhou. Canonical correlation coefficients of high-dimensional Gaussian vectors: finite rank case. *The Annals of Statistics (to appear)*, 2018.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017.

H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, sep 1987.

L. Chen, D. Paul, R. Prentice, and P. Wang. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496): 1345–1360, 2011.

R. Enomoto, T. Sakurai, and Y. Fujikoshi. Consistency properties of AIC, BIC, Cp and their modifications in the growth curve model under a large-(q,n) framework. *SUT Journal of Mathematics*, 51(1):59–81, 2015.

Y. Fujikoshi. A criterion for variable selection in multiple discriminant analysis. *Hiroshima Mathematical Journal*, 13(1):203–214, 1983.

Y. Fujikoshi. Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *Journal of Multivariate Analysis*, 17(1):27–37, aug 1985.

Y. Fujikoshi and T. Sakurai. High-dimensional consistency of rank estimation criteria in multivariate linear model. *Journal of Multivariate Analysis*, 149:199–212, 2016a.

Y. Fujikoshi and T. Sakurai. Some Properties of Estimation Criteria for Dimensionality in Principal Component Analysis. *American Journal of Mathematical and Management Sciences*, 35(2):133–142, 2016b.

Y. Fujikoshi and T. Sakurai. Consistency of Test-based Criterion for Selection of Variables in High-dimensional Two Group-Discriminant Analysis. *Japanese Journal of Statistics and Data Science* (to appear), 2018.

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

Y. Fujikoshi and L. G. Veitch. Estimation of dimensionality in canonical correlation analysis. *Biometrika*, 66(2):345–351, 1979.

Y. Fujikoshi, R. Enomoto, and T. Sakurai. High-dimensional in the growth curve model. *Journal of Multivariate Analysis*, 122:239–250, 2013.

Y. Fujikoshi, T. Sakurai, and H. Yanagihara. Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, 123:184–200, 2014.

G. Heinze, C. Wallisch, and D. Dunkler. Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, may 2018.

Y. Li, B. Nan, and J. Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363, jun 2015.

C. L. Mallows. Some Comments on Cp. *Technometrics*, 15(4):661, 1973.

R. Nishii, Z. Bai, and P. R. Krishnaiah. Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Mathematical Journal*, 18(3): 451–462, 1988.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

J. W. Silverstein and S. I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

R. Sparks, D. Coutsourides, and L. Troskie. The multivariate C p. *Communications in Statistics - Theory and Methods*, 12(15):1775–1793, jan 1983.

M. Yamamura, H. Yanagihara, and M. S. Srivastava. Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japanese Journal of Applied Statistics*, 39(1):1–19, 2010.

H. Yanagihara. Conditions for Consistency of a Log-Likelihood-Based Information Criterion in Normal Multivariate Linear Regression Models under the Violation of the Normality Assumption. *Journal of the Japan Statistical Society*, 45(1), 21–56, 2015.

H. Yanagihara, H. Wakaki, and Y. F. Fujikoshi. A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, 9:869–897, 2015.

Q. Zhang, J. Hu, and Z. Bai. Optimal modification of the lrt for the equality of two high-dimensional covariance matrices. *arxiv:1706.06774*, 2017.

L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20(1):1–25, 1986.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.

KLASMOE and School of Mathematics & Statistics
Northeast Normal University, Changchun, China.
E-mail: baizd@nenu.edu.cn; huj156@nenu.edu.cn

Department of Mathematics
Graduate School of Science
Hiroshima University, Hiroshima, Japan.
E-mail: fujikoshi_y@yahoo.co.jp