Bernoulli **28**(4), 2022, 2375–2403
https://doi.org/10.3150/21-BEJ1422

Asymptotics of AIC, BIC and C_p model selection rules in high-dimensional regression

ZHIDONG BAI[1,a], KWOK PUI CHOI[2,c], YASUNORI FUJIKOSHI[3,d] and JIANG HU[1,b]

[1] *KLASMOE and School of Mathematics & Statistics, Northeast Normal University, China.* [a] *baizd@nenu.edu.cn,* [b] *huj156@nenu.edu.cn*
[2] *Department of Statistics and Data Science, National University of Singapore, Singapore.* [c] *stackp@nus.edu.sg*
[3] *Department of Mathematics, Graduate School of Science, Hiroshima University, Japan.* [d] *fujikoshi_y@yahoo.co.jp*

Variable selection in multivariate linear regression is essential for the interpretation, subsequent statistical inferences and predictions of the statistical problem at hand. It has a long history of being studied, and many regressor selection criteria have been proposed. Most commonly used criteria include the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Mallow's C_p and their modifications. It is well-known that if the true model is among the candidate models, then BIC is strongly consistent while AIC is not when only the sample size tends to infinity and the numbers of response variables and regressors remain fixed; a setting often described as large-sample. Increasingly, more and more datasets are viewed as high-dimensional in the sense that the number of response variables (p), the number of regressors (k) and the sample size (n) tend to infinity such that $p/n \to c \in (0,1)$ and $k/n \to \alpha \in [0,1)$ with $\alpha + c < 1$. A few recent works reported that, under high dimension, the asymptotic properties of AIC, BIC and C_p selection rules in the large-sample setting do not necessarily carry over in the high-dimensional setting. In this paper, we clarify their asymptotic properties and provide sufficient conditions for which a selection rule is strongly consistent, almost surely under specify and over specify a true model. We do not assume normality in the errors, and we only require finite fourth moment. The main tool employed is random matrix theory techniques. A consequence of this work states that, under certain mild high-dimensional conditions, if the BIC selection rule is strongly consistent then the AIC selection rule is also strongly consistent, but not vice versa. This result is in stark contrast to the large-sample result.

Keywords: AIC; BIC; C_p; strong consistency; high-dimensional criteria; multi-response regression; variable selection; RMT

1. Introduction

In multivariate statistical analysis, the most general and favorable approach to investigate the relationship between a sample of n observations $(\tilde{\mathbf{x}}_i, \mathbf{y}_i)$ for $1 \le i \le n$ is the multivariate linear regression (MLR) model. Here $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ with y_{ij} the j-th response variable of the i-th observation, and $\tilde{\mathbf{x}}_i = (x_{i1}, \ldots, x_{ik})'$ with x_{ij} the j-th predictor variable of the i-th observation. Specifically, we consider

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}\boldsymbol{\Sigma}^{1/2}, \tag{1.1}$$

where the $n \times p$ response matrix $\mathbf{Y} = (y_{ij}) = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$, the $n \times k$ predictor matrix $\mathbf{X} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)' = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, the $k \times p$ regression coefficient matrix $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$, the $n \times p$ random errors matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_p) = (e_{ij})$ and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. MLR has a long history of being studied and applied in too many disciplines to be even listed here. Majority of research and applications confine to the large-sample setting, in which p and k are fixed while n tends to infinity. A main 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 past decades have witnessed breakthroughs in high-throughput biotechnology, telecommunication, surveillance and many other areas which generate huge amount of data. Ever-increasing and faster internet connectivity, and exponential drop in the cost of data-storage over the years have contributed a deluge of data awaiting to be analyzed and made sense of. Therefore, there is a rising need to consider datasets in which p, k and n are large. The model (1.1), hereinafter referred as the full model, is not always a good model for subsequent analyses especially in the high-dimensional context (see Condition (A1)) where predominantly many candidate predictors are erroneously included in the early stage of exploratory data analysis. In other words, rows of Θ that correspond to these candidate predictors are indeed zero. Hence, variable selection in MLR is essential for model interpretation and insight into the statistical problem at hand, and for subsequent statistical inferences and predictions. Many predictor selection criteria have been proposed and studied. Most commonly used criteria include the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Mallow's C_p and their modifications. For close to fifty years since their introduction, these selection rules have been well-studied and their statistical properties are well-understood in the *large-sample* setting: the numbers of predictors and responses are fixed and only the sample size tends to infinity. Much less is known when the dimensionality is large in the sense of Condition (A1). Interestingly, some recent works showed that these selection rules do not converge to the true model in probability in some special cases even under normality assumption of the errors, indicating that statistical properties of these selection rules in the large-sample context may not carry over to the high-dimensional setting. It is therefore desirable to characterize when AIC or BIC or C_p selection rule converges to the true model almost surely without normality assumption in the random errors. This work attempts to fill in this gap. Via random matrix theory techniques, we study the asymptotics of the AIC, BIC and C_p selection rules. A consequence of this work (Corollary 3.10) states that, under mild conditions, if the BIC selection rule is strongly consistent then the AIC selection rule is strongly consistent, but not vice versa. This result is in stark contrast to the well-known large-sample result.

We shall now provide details for the Akaike information criterion (AIC), Bayesian information criterion (BIC) and Mallows's C_p for variable selection problem in MLR. Let $\mathbf{j}$ be a subset of $\omega = \{1, 2, \cdots, k\}$ and $\mathbf{X_j} = (\mathbf{x}_j, j \in \mathbf{j})$ and $\Theta_\mathbf{j} = (\boldsymbol{\theta}_j, j \in \mathbf{j})'$. Denote model $\mathbf{j}$ by

$$\mathbf{Y} = \mathbf{X_j}\Theta_\mathbf{j} + \mathbf{E}\boldsymbol{\Sigma}^{1/2}.$$

Akaike's seminal paper (Akaike, 1973) proposed using Kullback-Leibler divergence as the fundamental basis for model selection, which is defined as follows:

$$A_\mathbf{j} = n\log(|\widehat{\boldsymbol{\Sigma}}_\mathbf{j}|) + 2\left(|\mathbf{j}|p + \frac{1}{2}p(p+1)\right) + np(\log(2\pi) + 1), \tag{1.2}$$

where

$$n\widehat{\boldsymbol{\Sigma}}_\mathbf{j} = \mathbf{Y}'\mathbf{Q_j}\mathbf{Y}, \;\; \mathbf{Q_j} = \mathbf{I}_n - \mathbf{P_j}, \;\; \mathbf{P_j} = \mathbf{X_j}(\mathbf{X_j}'\mathbf{X_j})^{-1}\mathbf{X_j}'. \tag{1.3}$$

Here, $\mathbf{I}_n$ is the identity matrix of order n, $|\mathbf{j}|$ the cardinality of set $\mathbf{j}$, $|\widehat{\boldsymbol{\Sigma}}_\mathbf{j}|$ the determinant $\widehat{\boldsymbol{\Sigma}}_\mathbf{j}$, $\mathbf{P_j}$ an orthogonal projection of rank $|\mathbf{j}|$ onto the subspace spanned by $\mathbf{X_j}$, and $\mathbf{Q_j}$ the orthogonal projection of rank $n - |\mathbf{j}|$ onto the orthogonal complement subspace spanned by $\mathbf{X_j}$.

BIC, 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.,

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

A criterion with behavior related to adjusted R-square and similar to that of the AIC for variable selection in regression models is Mallows's C_p proposed by Mallows (1973). This is defined as follows:

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

Refer to Fujikoshi (1983), Nishii, Bai and Krishnaiah (1988), Sparks, Coutsourides and Troskie (1983) for additional details of formulas (1.2), (1.4) and (1.5). Then, the AIC, BIC, and C_p selection rules are respectively used to select

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

where $\mathbf{J}$ is the set of candidate models.

Suppose the data are generated from a model (hereinafter referred to as the true model) among the candidate models considered. Certain optimality, such as consistency, is desirable for model selection. A model selection rule is said to be weakly consistent if the model it identifies converges to the true model in probability. Strong consistency refers to the model identified by the selection rule converges almost surely to the true model. Clearly, strong consistency implies weak consistency but not vice versa. Moreover, strong consistency provides a deeper understanding of the selection rules. Under a large-sample asymptotic framework, i.e., dimension p is fixed and n tends to infinity, it is well-known that the AIC and C_p selection rules are not strongly consistent (see, for examples, Fujikoshi (1985), Fujikoshi and Veitch (1979)) but the BIC selection rule is strongly consistent, Nishii, Bai and Krishnaiah (1988). Very recently, Fujikoshi, Sakurai and Yanagihara (2014), Yanagihara (2015), Yanagihara, Wakaki and Fujikoshi (2015)) noticed that this asymptotic property is not necessary true in high-dimensional framework. When k and p are large, the large-sample selection criteria admit many variables that are not part of the true model. For example, under large-sample, large-dimensional asymptotic framework (i.e., k is fixed, $p < n$ with $p/n \to c \in [0,1)$) and under normality assumption on the errors, BIC selection rule has been shown to be not consistent, but the AIC and C_p selection rules are weakly consistent.

To clarify these model selection rules, we investigate their asymptotic behavior under a large-model (k), large-sample (n) and large-dimensional response (p); which we coin it as 3L asymptotic framework. Specifically, $\min\{k,p,n\}$ tends to infinity in which $p/n \to c \in (0,1)$, $k/n \to \alpha \in [0,1)$ satisfying $\alpha + c < 1$. Our goal is to provide theoretical understanding of these commonly used selection rules and their modified methods under a 3L framework. Our hope is that this article will stimulate further research in high-dimensional variable selection. We refer readers to three recent reviews by Anzanello and Fogliatto (2014), Heinze, Wallisch and Dunkler (2018), Shao (1997) on comparing the variable selection rules. In this 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, such as Li, Nan and Zhu (2015) and Zou and Hastie (2005). If the model size k is greater than the sample size n, one can use screening methods to reduce the model size to ensure Condition (A1) holds; for examples, the sure independence screening method based on the distance correlation Li, Zhong and Zhu (2012), and interaction pursuit via distance correlation Kong *et al.* (2017). For more details in screening methods, see Fan and Lv (2008, 2010) and references therein. One should note, however, that not all variable screening methods perform well in multiple responses.

We highlight two main contributions of the present paper. First, random matrix theory (RMT) is introduced to study model selection rules in high-dimensional MLR. The new theoretical results and the methods of proofs are applicable to many other model selection rules, such as the modified AIC in Fujikoshi and Satoh (1997) and modified C_p in Bozdogan (1987). The technical tools developed in this paper can be applied to the growth curve model in Enomoto, Sakurai and Fujikoshi (2015)

and Fujikoshi, Enomoto and Sakurai (2013), multiple discriminant analysis in Fujikoshi (1983) and Fujikoshi and Sakurai (2016a), principal component analysis in Fujikoshi and Sakurai (2016b) and Bai, Choi and Fujikoshi (2018), and canonical correlation analysis in Nishii, Bai and Krishnaiah (1988) and Bao *et al.* (2019), just to name some.

Second, we characterize when the selection rule correctly identifies the true model asymptotically under a 3L asymptotic framework without normality assumption in the errors. Moreover, our limited simulation studies suggest that even the finite n results are robust against departure from normal distribution. Specifically, Corollary 3.10 concludes that under a 3L asymptotic framework if the BIC selection rule is strongly consistent so is the AIC selection, but not vise versa. This is in stark contrast to the result in large-sample setting.

The remainder of this paper is organized as follows. In Section 2, we introduce the needed notation and state the conditions for the statements of main results. The main results are presented in Section 3, which also includes our recommendation of which selection rule to use for given k, p and n. We present some simulation studies in Section 4 to illustrate and complement our results. Proofs of the main theorems and some preparatory lemmas are given in Section 5 and Section 6, respectively. Section 7 presents the conclusion and discussion. The paper has also an on-line supplementary file (Bai et al., 2022) which includes the proof of Proposition 3.1.

2. Notation and statements of conditions

Throughout this paper, we consider a multivariate linear regression (MLR) of k predictors, response variable is of dimension p, and sample size is n where $k + p < n$. Specifically, let $(\tilde{\mathbf{x}}_1, \mathbf{y}_1), \ldots, (\tilde{\mathbf{x}}_n, \mathbf{y}_n)$ be a random sample drawn from a population. We confine ourselves to the 3L framework: large model (k), large dimension (p) and large sample size (n) which satisfy Condition (A1). For notational simplicity, we do not indicate the dependence of p and k on n. Throughout this paper, we denote the spectral norm for a matrix by $\|\cdot\|$, and we use $o_p(1)$ to denote a scalar negligible in probability. The notation $o(1)$, $o_{a.s}(1)$, $O(1)$, $O_p(1)$ and $O_{a.s.}(1)$ are used in a similar way.

Recall the MLR model (1.1)

$$M: \quad \mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}\mathbf{\Sigma}^{1/2},$$

where Θ is a $k \times p$ unknown matrix of regression coefficients and $\mathbf{\Sigma}$ is a $p \times p$ unknown positive definite covariance matrix.

Let $\mathbf{J}$, which depends on k, be a set of subsets of $\boldsymbol{\omega} := \{1, 2, \cdots, k\}$. For $\mathbf{j} \in \mathbf{J}$, we denote its cardinality by $|\mathbf{j}|$. We also use $|\mathbf{A}|$ to denote the determinant of a matrix $\mathbf{A}$, however, the context will be clear enough that there is no risk of ambiguity. Let $\mathbf{X}_\mathbf{j}$ be the matrix by keeping only all the j-th columns of $\mathbf{X}$ if $j \in \mathbf{j}$. Similarly, $\Theta_\mathbf{j}$ corresponds to the matrix by only keeping all the j-th rows of Θ if $j \in \mathbf{j}$. We define the candidate model corresponding to $\mathbf{j}$ as $M_\mathbf{j}$:

$$M_\mathbf{j}: \quad \mathbf{Y} = \mathbf{X}_\mathbf{j}\Theta_\mathbf{j} + \mathbf{E}\mathbf{\Sigma}^{1/2}. \tag{2.1}$$

For simplicity, we refer $M_\mathbf{j}$ as model $\mathbf{j}$. Model $\mathbf{j}$ is equivalent to the multivariate linear regression using just the subset of predictors in $\mathbf{j}$ and the corresponding regression coefficients. Denote the true model as $\mathbf{j}_*$, and

$$M_{\mathbf{j}_*}: \quad \mathbf{Y} = \mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*} + \mathbf{E}\mathbf{\Sigma}^{1/2}.$$

We partition $\mathbf{J}$ into $\mathbf{J}_- \cup \{\mathbf{j}_*\} \cup \mathbf{J}_+$ where

$$\mathbf{J}_+ = \{\mathbf{j} : \mathbf{j} \supsetneq \mathbf{j}_*\}, \quad \mathbf{J}_- = \{\mathbf{j} : \mathbf{j} \not\supset \mathbf{j}_*\} = \{\mathbf{j} : \mathbf{j}_* \setminus \mathbf{j} \neq \varnothing\}.$$

For any $\mathbf{j} \in \mathbf{J}_-$, we partition $\mathbf{j}$ into $\mathbf{j}_+ \cup \mathbf{j}_-$, where $\mathbf{j}_+ = \mathbf{j} \cap \mathbf{j}_*^c$ and $\mathbf{j}_- = \mathbf{j} \cap \mathbf{j}_*$. We say that a model $\mathbf{j}$ is over-specified if $\mathbf{j} \in \mathbf{J}_+$ and under-specified if $\mathbf{j} \in \mathbf{J}_-$. We call variable $\mathbf{x}_j$ if $j \in \mathbf{j}_*$ a true variable, variable $\mathbf{x}_j$ if $j \notin \mathbf{j}_*$ a spurious variable, and variable $\mathbf{x}_j$ a missing variable in $\mathbf{j} \in \mathbf{J}_-$ if $j \in \mathbf{j}_* \setminus \mathbf{j}$.

The conditions for our results are

(A1): As $n \to \infty$, $c_n \to c \in (0,1)$ and $\alpha_k \to \alpha \in [0,1)$ satisfying $\alpha + c < 1$.

(A2): The full model $\boldsymbol{\omega} \in \mathbf{J}$, the true model $\mathbf{j}_* \in \mathbf{J}$, $\mathbf{j}_*$ is fixed, and $|\mathbf{J}| = O(n^\ell)$ for some $\ell > 0$.

(A3): The entries e_{ij} of $\mathbf{E}$ are independent with zero means, unit variances, uniformly bounded fourth moments and satisfies Lindeberg-type condition:

$$\frac{1}{\eta^4 n^2} \sum_{i,j} \mathbb{E}\left[\left|e_{ij}\right|^4 \mathbb{1}\left\{ \left|e_{ij}\right| \geq \eta\sqrt{n} \right\} \right] = o(1),$$

for any $\eta > 0$.

(A4): Matrix $\mathbf{X}'\mathbf{X}$ is positive definite for all $n > k + p$.

Remark 2.1. *(i)* For $\mathbf{j} \in \mathbf{J}$, condition *(A4)* implies $\mathbf{X}_\mathbf{j}'\mathbf{X}_\mathbf{j}$ is invertible because it is a principal submatrix of $\mathbf{X}'\mathbf{X}$. *(ii)* Requiring the full model, $\boldsymbol{\omega} \in \mathbf{J}$ and the true model $\mathbf{j}_* \in \mathbf{J}$ is natural. The condition that $\mathbf{j}_*$ is fixed can be relaxed further but for the sake of simpler presentation, it is not pursued in this paper. The role of $|\mathbf{J}| = O(n^\ell)$ condition in *(A2)* is to ensure uniform convergence in Theorems 3.4–3.6. This condition is not as restrictive as it appears since majority of commonly used models satisfy this condition. For example, if an upper bound of $|\mathbf{j}_*|$ is known, we can choose $\mathbf{J}$ to consist of all subsets of $\boldsymbol{\omega}$ with cardinality not greater than this upper bound. *(iii)* Condition *(A3)* is commonly assumed in random matrix theory for non-normal distribution.

3. Main results

Before we present our results about the asymptotic properties of AIC, BIC and C_p selection rules, we include some preliminary results from RMT. Proposition 3.1 extends Theorem 1 in Bai, Miao and Pan (2007), who derived the limit of a form of empirical spectral distribution defined with weights depending on the eigenvectors of large-dimensional sample covariance matrix. See Remark 3.3 for other applications. The statements of the asymptotic properties of AIC, BIC and C_p are found in Theorems 3.4, 3.5 and 3.6 respectively in Section 3.2. Based on these theoretical results and simulation studies in Section 4, we come up with recommendations on which selection rule to be preferred in Section 3.3.

3.1. Preliminary results from RMT

We introduce some basic results from RMT and a key proposition, one of the main tools in the paper. For any $n \times n$ matrix $\mathbf{A}_n$ with only real eigenvalues, let $F^{\mathbf{A}_n}$ be the empirical spectral distribution function of $\mathbf{A}_n$, that is,

$$F^{\mathbf{A}_n}(x) = \frac{1}{n}\left|\left\{ i : \lambda_i^{\mathbf{A}_n} \leq x \right\}\right|,$$

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

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

Suppose an $p \times p$ matrix $\mathbf{A}$ is invertible, for any $p \times n$ matrix $\mathbf{C}$, the following identities will be used frequently,

$$(\mathbf{A} - \mathbf{C}\mathbf{C}')^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{I}_n - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1}, \tag{3.1}$$

which immediately implies

$$(\mathbf{A} - \mathbf{C}\mathbf{C}')^{-1}\mathbf{C} = \mathbf{A}^{-1}\mathbf{C}(\mathbf{I}_n - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}, \tag{3.2}$$

$$\mathbf{C}'(\mathbf{A} - \mathbf{C}\mathbf{C}')^{-1} = (\mathbf{I}_n - \mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1}. \tag{3.3}$$

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

$$\mathbf{C}(\mathbf{C}'\mathbf{C} - z\mathbf{I}_n)^{-1}\mathbf{C}' = \mathbf{I}_p + z(\mathbf{C}\mathbf{C}' - z\mathbf{I}_p)^{-1}, \tag{3.4}$$

which is called the in-out-exchange formula herein. Equations (3.1)–(3.4) are straightforward to derive by basic linear algebra, and thus, their proofs are omitted. Note that all vectors in this paper are column vectors, and when the context is clear, we shall not indicate the order of the identity.

It is well known that the Stieltjes transform of the LSD of $\frac{1}{p}\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E}$, denoted by $\underline{s}(z)$, is the unique solution on the upper complex plane to the equation

$$z = -\frac{1}{\underline{s}(z)} + \frac{1}{c}\int \frac{x}{1 + x\underline{s}(z)}dH(x), \tag{3.5}$$

where H is the LSD of $\mathbf{Q}_\mathbf{j}$ (see (1.4) of (Silverstein and Choi, 1995) for more details) and $\mathbf{E}$ is given in (1.1). We state the following proposition, which is a key tool in this paper, and its proof will be given in the supplementary material.

Proposition 3.1. *Let* $\mathbf{M} := \mathbf{M}(z) = p^{-1}\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E} - z\mathbf{I}_p$, $m = |\mathbf{j}|$, $\boldsymbol{\alpha}_1$ *and* $\boldsymbol{\alpha}_2$ *be non-random p-vectors,* $\boldsymbol{\alpha}_3$ *and* $\boldsymbol{\alpha}_4$ *be non-random n-vectors and assume that* $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ *and* $\boldsymbol{\alpha}_4$ *are all bounded in Euclidean norm. Then, under conditions* (A1) – (A4)*, we have that for any* $z \in \mathbb{C}^+$, $t > 0$ *and* $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\boldsymbol{\alpha}_2 - \underline{s}_{n\mathbf{j}}(z)\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2\right| \geq \varepsilon\right) = o(n^{-t}), \tag{3.6}$$

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{\alpha}_1'\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_3\right| \geq \varepsilon\right) = o(n^{-t}), \tag{3.7}$$

and

$$\mathbb{P}\left(\left|\frac{1}{p}\boldsymbol{\alpha}_3'\mathbf{E}\mathbf{M}^{-1}\mathbf{E}'\boldsymbol{\alpha}_4 - \underline{s}_{n\mathbf{j}}(z)\boldsymbol{\alpha}_3'\boldsymbol{\alpha}_4 + \frac{\underline{s}_{n\mathbf{j}}^2(z)}{\underline{s}_{n\mathbf{j}}(z) + 1}\boldsymbol{\alpha}_3'\mathbf{Q}_{\mathbf{j}_t}\boldsymbol{\alpha}_4\right| \geq \varepsilon\right) = o(n^{-t}), \tag{3.8}$$

where $\underline{s}_{n\mathbf{j}}(z)$ *is the Stieltjes transform of the LSD of* $\frac{1}{p}\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E}$ *with* c *and* H *replaced by* c_n *and* $F^{\mathbf{Q}_\mathbf{j}}$*, respectively.*

Remark 3.2. This proposition demonstrates the usefulness of the Stieltjes transform in deriving the limits of the spectra of random matrices. Basically, our aim is to obtain the limits of $n\boldsymbol{\alpha}_1'(\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E})^{-1}\boldsymbol{\alpha}_2$,

$n^{1/2}\boldsymbol{\alpha}_1'(\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_3'\mathbf{E}(\mathbf{E}'\mathbf{Q}_\mathbf{j}\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\alpha}_4$, which are equivalent to deriving $\lim_{z\downarrow 0+0i}\underline{s}_{n\mathbf{j}}(z)$ and $\lim_{z\downarrow 0+0i} z\underline{s}_{n\mathbf{j}}(z)$ and then letting $n\to\infty$. By the fact that $F^{\mathbf{Q}_\mathbf{j}}(\{0\})=\alpha_m$, $F^{\mathbf{Q}_\mathbf{j}}(\{1\})=1-\alpha_m$ and

$$z=-\frac{1}{\underline{s}_{n\mathbf{j}}(z)}+\frac{1}{c_n}\frac{1-\alpha_m}{1+\underline{s}_{n\mathbf{j}}(z)},$$

we have

$$\frac{z(1+\underline{s}_{n\mathbf{j}}(z)+\frac{c_n-1+\alpha_m}{c_nz})+\frac{1}{\underline{s}_{n\mathbf{j}}(z)+1}}{z^2(1+\underline{s}_{n\mathbf{j}}(z)+\frac{c_n-1+\alpha_m}{c_nz})^2}=\frac{1}{1+\underline{s}_{n\mathbf{j}}(z)}-1 \tag{3.9}$$

and

$$\underline{s}_{n\mathbf{j}}(z)=\frac{1-\alpha_m-c_n-c_nz\pm\sqrt{(1-\alpha_m+c_n-c_nz)^2-4c_n(1-\alpha_m)}}{2c_nz}.$$

As any Stieltjes transform tends to zero as $z\to\infty$, we have

$$\underline{s}_{n\mathbf{j}}(z)=\frac{1-\alpha_m-c_n-c_nz+\sqrt{(1-\alpha_m+c_n-c_nz)^2-4c_n(1-\alpha_m)}}{2c_nz},$$

and

$$1-\frac{1}{1+\underline{s}_{n\mathbf{j}}(z)}=\frac{1-\alpha_m+c_n-c_nz+\sqrt{(1-\alpha_m+c_n-c_nz)^2-4c_n(1-\alpha_m)}}{2(1-\alpha_m)}.$$

Letting $z\downarrow 0+0i$ and together with (6.4) and $1-\alpha_m-c_n>0$, we conclude that

$$\underline{s}_{n\mathbf{j}}(z)\to\frac{c_n}{1-\alpha_m-c_n}. \tag{3.10}$$

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. So,

$$\lim_{z\downarrow 0+i0}\sqrt{(1-\alpha_m+c_n-c_nz)^2-4c_n(1-\alpha_m)}=-|1-\alpha_m-c_n|.$$

Finally, letting $n\to\infty$ and applying Proposition 3.1, we obtain the desired limits.

Remark 3.3. We want to point out two other applications in addition to extending the result of Bai et al. (2007) as mentioned at the beginning of Section 3. First, $\mathbf{Q}_\mathbf{j}$ here can be improved to any general non-random projection matrix with rank $m>0$ directly. Second, this kind of random projection matrices appears very often in multivariate statistics analysis, Proposition 3.1 has several potential applications in the growth curve model Enomoto, Sakurai and Fujikoshi (2015), Fujikoshi, Enomoto and Sakurai (2013), multiple discriminant analysis Fujikoshi (1983), Fujikoshi and Sakurai (2016a), principal component analysis Bai, Choi and Fujikoshi (2018), Fujikoshi and Sakurai (2016b), and canonical correlation analysis Bao *et al.* (2019), Nishii, Bai and Krishnaiah (1988). We will pursue these applications in future work.

3.2. Asymptotics of AIC, BIC and C_p selection rules

Define two bivariate functions on $\{(\alpha,c):\ \alpha\in[0,1), c\in(0,1), \alpha+c<1\}$

$$\phi(\alpha,c)=2c+\log\left(\frac{(1-c)^{1-c}(1-\alpha)^{1-\alpha}}{(1-c-\alpha)^{1-c-\alpha}}\right)^{1/\alpha},\ \ \psi(\alpha,c)=\frac{c(1-\alpha-2c)}{1-\alpha-c},$$

which are the limits of $\frac{1}{nk}(A_{\boldsymbol{\omega}}-A_{\mathbf{j}_*})$ and $\frac{1}{nk}(C_{\boldsymbol{\omega}}-C_{\mathbf{j}_*})$, respectively. For $\mathbf{j}\in\mathbf{J}_-$ with $|\mathbf{j}_+|=m\geq 0$ and $|\mathbf{j}_*\cap\mathbf{j}_-^c|=s>0$, we denote

$$\begin{aligned}\Phi_{\mathbf{j}}&:=\frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta_{\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*}'\mathbf{Q}_{\mathbf{j}}\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}\\ \tau_{n\mathbf{j}}&:=(1-\alpha_m)^{s-p}|(1-\alpha_m)\mathbf{I}_p+\Phi_{\mathbf{j}}|\\ \kappa_{n\mathbf{j}}&:=\mathrm{tr}(\Phi_{\mathbf{j}})\end{aligned}$$

where $\mathbf{Q}_{\mathbf{j}}$ is defined as in (1.3). We state below our main theorems for the asymptotics of AIC, BIC and C_p selection rules. Their proofs are presented in Section 5.

Theorem 3.4 (Asymptotics of AIC selection rule). *Suppose conditions* (A1)–(A4) *hold.*

(1) *Suppose* $\phi(\alpha,c)<0$. *The AIC selection rule over-specifies the true model a.s..*
(2) *Suppose* $\phi(\alpha,c)>0$.
  (i) *If for all under-specified models* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\liminf_{n\to\infty}\log(\tau_{n\mathbf{j}})>(s-m)(2c+\log(1-c)),$$

  then the AIC selection rule is strongly consistent.
  (ii) *If there exists an under-specified model* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\limsup_{n\to\infty}\log(\tau_{n\mathbf{j}})<(s-m)(2c+\log(1-c)),$$

  then the AIC selection rule under-specifies the true model a.s..

Theorem 3.5 (Asymptotics of BIC selection rule). *Suppose conditions* (A1)–(A4) *hold.*

(1) *For all under-specified models* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\liminf_{n\to\infty}\Big(\log(\tau_{n\mathbf{j}})-c(s-m)\log(n)\Big)>(s-m)\log(1-c),$$

  then the BIC selection rule is strongly consistent.
(2) *If there exists an under-specified model* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\limsup_{n\to\infty}\Big(\log(\tau_{n\mathbf{j}})-c(s-m)\log(n)\Big)<(s-m)\log(1-c),$$

  then the BIC selection rule under-specifies the true model a.s..

Theorem 3.6 (Asymptotics of C_p selection rule). *Suppose conditions* (A1)–(A4) *hold.*

(1) *Suppose* $\psi(\alpha,c)<0$. *The* C_p *selection rule over-specifies the true model a.s..*

(2) *Suppose* $\psi(\alpha,c)>0$.
(i) *If for all* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\liminf_{n\to\infty}\kappa_{n\mathbf{j}}>(s-m)\frac{c(1-\alpha-2c)}{1-\alpha},$$

then the C_p *selection rule is strongly consistent.*
(ii) *If there exists* $\mathbf{j}\in\mathbf{J}_-$ *with* $s-m>0$ *such that*

$$\limsup_{n\to\infty}\kappa_{n\mathbf{j}}<(s-m)\frac{c(1-\alpha-2c)}{1-\alpha},$$

then the C_p *selection rule under-specifies the true model a.s..*

Remark 3.7. In real datasets, n,p and k are indeed fixed. When it is tenable to assume n is reasonably large, the 3L viewpoint is to regard $\alpha=k/n$ and $c=p/n$ and apply Theorems 3.4–3.6 to study the asymptotics of the AIC, BIC and C_p selection rules.

Remark 3.8. A consequence of these theorems is that over-specified properties of these rules do not need the observed values. And for the under-specified properties, i.e., missing some true variables, one only needs to consider candidate models $\mathbf{j}\in\mathbf{J}_-$ with $m<s$ since the candidate models $\mathbf{j}\in\mathbf{J}_-$ with $m>s$ will not be selected by AIC, BIC and C_p methods asymptotically. Moreover, if $m<s$, then $\alpha_m\to 0$ and the rank of $\Phi_{\mathbf{j}}$ is s. Thus, $\log(\tau_{n\mathbf{j}})$ and $\kappa_{n\mathbf{j}}$ are $\sum_{i=1}^{s}\log(1+\lambda_i^{\Phi_{\mathbf{j}}})$ and $\sum_{i=1}^{s}\lambda_i^{\Phi_{\mathbf{j}}}$, respectively. Here, $\lambda_i^{\Phi_{\mathbf{j}}}$ are the non-zero eigenvalues of $\Phi_{\mathbf{j}}$. If the elements of $\mathbf{X}_{\mathbf{j}_*}$ are of $O(1)$, which is a common and easily verified assumption in MLR, then $\|\frac{1}{n}\mathbf{X}_{\mathbf{j}_*}'\mathbf{Q}_{\mathbf{j}}\mathbf{X}_{\mathbf{j}_*}\|$ is bounded. Intuitively, if the elements of $\Theta_{\mathbf{j}_*}$ are big or the elements of the covariance matrix are small, then the eigenvalues of $\Phi_{\mathbf{j}}$ should be big. According to these three theorems, the under-specified models are unlikely to be selected. On the other hand, if there exists a true variable j, and the elements of θ_j are small or the elements of the covariance matrix are big, then for $j\notin\mathbf{j}$, the eigenvalues of $\Phi_{\mathbf{j}}$ should be small. Thus, in this case, the selection rules are likely to miss the true variable j. How big/small is considered big/small depends on the penalty term. Moreover, in real datasets, both Θ and Σ are unknown, thus we actually cannot know whether some true variables or not.

Remark 3.9. Figure 1 presents 3D and contour plots of $\phi(\alpha,c)>0$ and $\psi(\alpha,c)>0$. It shows that big enough α and c (in the sense that make $\phi(\alpha,c)<0$ and $\psi(\alpha,c)<0$) both result in over-specification of the true model. Moreover, Fujikoshi, Sakurai and Yanagihara (2014), Yanagihara, Wakaki and Fujikoshi (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 when $\alpha=0$.

Combining Theorems 3.4 and 3.5, we have the following corollary.

Corollary 3.10. *Suppose conditions* (A1)–(A4) *hold. Under the condition* $\phi(\alpha,c)>0$*, if the BIC selection rule is strongly consistent, then the AIC selection rule is strongly consistent but not vice versa.*

Remark 3.11. The conclusion in Corollary 3.10 is in stark contrast to the classical result that under large sample framework the BIC selection rule is strongly consistent and the AIC and C_p selection rules are not.

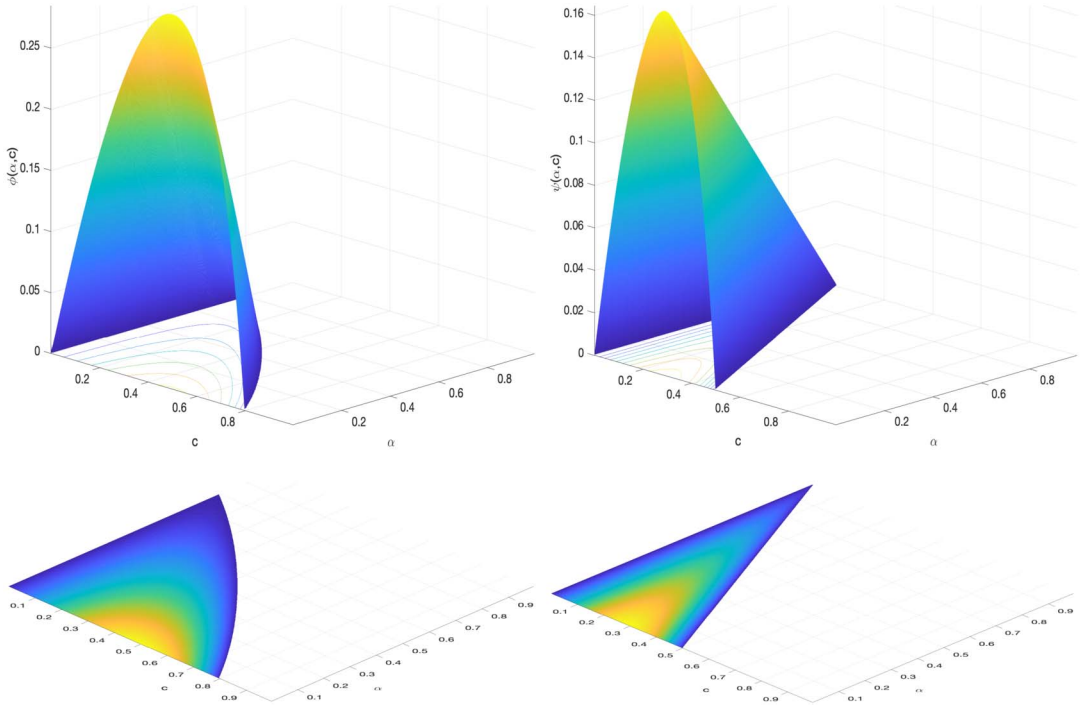

Figure 1. 3D and contour plots for $\phi(\alpha,c) > 0$ and $\psi(\alpha,c) > 0$. The left two figures are a wireframe mesh and a contour plot for $\phi(\alpha,c) > 0$. The right two figures are a wireframe mesh and a contour plot for $\psi(\alpha,c) > 0$.

Remark 3.12. Some recent works (e.g., Fan and Tang (2013)) have shown that both AIC and BIC may not have model selection consistency in high dimensions, while a generalized information criterion (GIC) involving a heavier penalty than BIC can have model selection consistency.

3.3. Recommendations

Based on our main results, we recommend below as to which selection rule, AIC, BIC or C_p, to be preferred in model selection under the 3L framework. Under the 3L high-dimensional framework, we do not recommend the BIC selection rule for model selection as it is prone to miss some true variables.

Denote

$$R_1 = \{(\alpha,c) \in \mathbb{R}^2 : \phi(\alpha,c) < 0,\ 0 < \alpha,c < 1,\ \alpha + c < 1\}$$

and

$$R_2 = \{(\alpha,c) \in \mathbb{R}^2 : \psi(\alpha,c) < 0,\ 0 < \alpha,c < 1,\ \alpha + c < 1\},$$

the regions over which AIC and C_p rules over-specify the true model, respectively. For dataset in which k,p and n may be viewed as large, we can plug in the estimates $\alpha_k = k/n$ and $c_n = p/n$ into Theorems 3.4 and 3.6. For dataset in which k,p and n may be viewed as large, we can plug in the estimates $\alpha_k = k/n$ and $c_n = p/n$ into Theorems 3.4 and 3.6. If $\alpha_k + c_n > 1$, the AIC, BIC and C_p

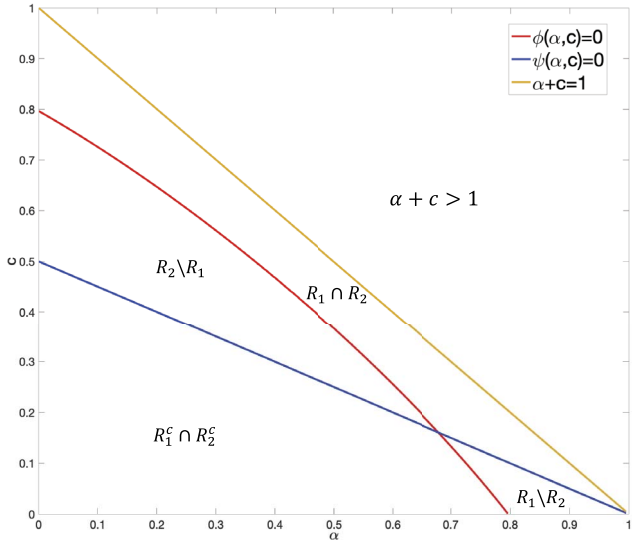

Figure 2. The regions of R_1 and R_2.

selection rules are not applicable. If $(\alpha_k, c_n) \in R_1 \cap R_2$, then both AIC and C_p selection would over-specify the model. If $(\alpha_k, c_n) \in R_1 \setminus R_2$, then AIC rule would over-specify the model, and C_p rule is applicable. If $(\alpha_k, c_n) \in R_2 \setminus R_1$, then C_p rule would over-specify the model, and AIC rule is applicable. If $(\alpha_k, c_n) \in R_1^c \cap R_2^c$, then both AIC and C_p rules would be applicable. For illustration, we present the regions of R_1 and R_2 in Figure 2.

4. Simulation studies

Theorems 3.4–3.6 concern the asymptotic behaviour of the AIC, BIC and C_p variable selection rules when $k, p, n \to \infty$ such that $p/n \to c \in (0,1)$ and $k/n \to \alpha \in [0,1)$ with $\alpha + c < 1$. In practice, k, p and n are fixed, and α and c will be taken as k/n and p/n respectively. Simulation studies are conducted with the following objectives: (i) Explore to what extent their asymptotic properties as delineated in the theorems provide an indication of their performances for finite n; (ii) Provide some empirical observation on their relative rates of convergence in the 3L framework; and (iii) Examine the robustness of our results against departure from normality, finite fourth moment condition on the errors, and collinearity of responses.

It is easy to see that the selection results in these three criteria depend only on the values of $\mathbf{\Sigma}^{-1/2}\Theta$, and does not depend on the choice of nonsingular $p \times p$ matrix $\mathbf{D}$ in $\mathbf{YD}$. Therefore, in conducting our simulation studies, it suffices to set $\mathbf{\Sigma} = \mathbf{I}_p$ and vary Θ. We set $\mathbf{j}_* = \{1,2,3,4,5\}$, $\alpha = 0.1$, $c = \{0.2, 0.5, 0.8\}$ and $n = 100, 150$ and 200, and the values of k and p will follow. We choose $\mathbf{J}$ to be the set of all non-empty subsets of $\boldsymbol{\omega}$.

Set $\mathbf{X} = \mathbf{U}\mathbf{T}^{1/2}$, where entries of $\mathbf{U}$ are independent and uniformly distributed $U(1,5)$, and $\mathbf{T} = [t_{ij}]_{k\times k}$ is a symmetric band matrix with $t_{ii} = 1$, $t_{i,i+1} = t_{i+1,i} = t$, $t_{ij} = 0$ if $|i-j| > 1$. Choose $\mathbf{1}_5$ is a 5-vector of ones, t and $\boldsymbol{\theta}_*$ to be chosen below. $\Theta_{\mathbf{j}_*} = \mathbf{1}_5\boldsymbol{\theta}_*$ and $\Theta = (\Theta'_{\mathbf{j}_*}, \mathbf{0})'$.

We consider three settings of t: (1) $t = 0$; (2) $t = 0.2$ and (3) $t = 0.4$; two settings of $\boldsymbol{\theta}$: (I): $\boldsymbol{\theta}_* = ((-0.5)^0, \ldots, (-0.5)^{p-1})$ and (II): $\boldsymbol{\theta}_* = \sqrt{n}((-0.5)^0, \ldots, (-0.5)^{p-1})$; and three cases for the distribution

of **E**: (i) standard normal; (ii) standardized t with five degrees of freedom, i.e., $e_{ij} \sim t_5/\sqrt{Var(t_5)}$; and (iii) standardized chi-square distribution with two degrees of freedom, i.e., $e_{ij} \sim (\chi_2^2 - 2)/\sqrt{Var(\chi_2^2)}$.

We first explain our choices of α_k, c_n, our settings and the distributions. Table 1 presents the values of $\phi(\alpha_k, c_n)$ and $\psi(\alpha_k, c_n)$. Settings (1), (2) and (3) are for the examination of the correlation of the predictors. Setting I ensures $\|\Phi_{\mathbf{j}}\|$ is bounded whereas Setting (II) ensures $\|\Phi_{\mathbf{j}}\| \to \infty$. Moreover, under Setting (1), $\Phi_{\{1\}}, \ldots, \Phi_{\{5\}}$ are identically distributed. Under Settings (1) and (I), $\log(\tau_{n\{i\}}) > 4(c_n + \log(1 - c_n))$ and $\kappa_{n\{i\}} > 4\psi(\alpha_k, c_n)(1 - \alpha_k - c_n)/(1 - \alpha_k)$. Under Settings (1) and (II), for $c_n = 0.2$ and $c_n = 0.5$, $\log(\tau_{n\{i\}}) - 4c_n \log(n) > 4\log(1 - c_n)$; and for $c_n = 0.8$, $\log(\tau_{n\{i\}}) - 4c_n \log(n) < 4\log(1 - c_n)$. Under Settings (2) and (3), the properties of $\tau_{n\{i\}}$ and $\kappa_{n\{i\}}$ are similar to those under Setting (1). Simulations are also conducted for three distributions so as to have some idea whether the results are distribution dependent for finite n. Based on Theorems 3.4–3.6, we expect that for large enough n, almost surely,

(a) AIC selection rule will over-specify the model in the case where $\{\alpha = 0.1, c = 0.8\}$ but will not in cases where $\{\alpha = 0.1, c = 0.2, 0.5\}$.
(b) BIC will not over-specify the model for our choices of α and c.
(c) C_p selection rule will over-specify the model in the case where $\{\alpha = 0.1, c = 0.5, 0.8\}$ but will not in the case where $\{\alpha = 0.1, c = 0.2\}$;

To explore in greater details the performance of these selection rules, the numbers of times a selection rule under-specifies the true model, exactly identifies it and over-specifies it were computed by Monte Carlo simulations with 1,000 repetitions. We shall call these numbers selection times for short. We first considered the standard normal distribution case and the results are reported in Tables 2–7.

In a repetition in which a selection rule over-specifies the true model, we take note of the number of "spurious" variables. Then we compute the average number of the spurious variables over those repetitions in which the rule over-specifies the true model. The average numbers are reported at the bottom row of each sub-table.

Tables of the simulation results for standardized t-distribution with five degrees of freedom and standardized chi-square distribution with two degrees of freedom are very similar to Tables 2–7. They are not included in this paper due to space consideration. Before we summarize our observations from our simulation studies, we remark that in multivariate linear regression, over-specifying the true model by a reasonable amount is far more desirable than under-specifying it. True variables that are lost when it is under-specified will be lost in any subsequent analysis.

Below are our conclusions based on our simulation studies:

(1) The asymptotic results in Theorems 3.4–3.6 provide very good indication of how the selection rules perform even for moderate values of n, particularly, for over-specification and under-specification. For example, when $\alpha = 0.1$ and $c = 0.2$, the percentage of the AIC selection rule identifying the true model is around 96%, very close to 100% even for $n = 200$.
(2) Under Setting (I), (i) BIC selection rule always under-specifies the true model except when p is small. (ii) In all our repetitions, BIC selection rule does not over-specify the true model in our experiments. (iii) When $\phi(\alpha_k, c_n) > 0$, AIC selection rule is the best among the three selection rules considered. (iv) When $\phi(\alpha, c) > 0$ (resp. $\psi(\alpha_k, c_n) > 0$), even though AIC (resp. C_p) selection rule over-specifies the true model, the average selection sizes are not excessive.
(3) Under Setting (II), (i) When condition (1) in Theorem 3.5 holds, BIC selection rule is the best among the three under consideration especially when $\{p, k\}$ are small. (ii) Almost always, BIC selection rule does not over-specify the true model in our experiments. (iii) The performances of AIC and C_p rules are similar to that under Setting I.

	$c_n = 0.2$		$c_n = 0.5$		$c_n = 0.8$	
	ϕ	ψ	ϕ	ψ	ϕ	ψ
$\alpha_k = 0.1$	0.16	0.14	0.25	-0.13	-0.26	-5.6

Table 1. Values of $\phi(\alpha_k, c_n)$ and $\psi(\alpha_k, c_n)$.

(4) The simulation results as summarized in Tables 2–7 suggest these selection rules are robust against non-normality of the errors and correlation among the predictors.

5. Proofs of Theorems 3.4–3.6

Before proceeding to the proofs of Theorems 3.4–3.6, we need some preliminary results. They do not just serve the purpose of proofs of the theorems, but have many potential applications in other multivariate analysis problems.

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k, p, n) = (10, 20, 100)$			$(k, p, n) = (15, 30, 150)$			$(k, p, n) = (20, 40, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	170	0	0	247	0	0	737	0
True	870	830	647	899	753	619	961	263	656
Over (Average)	130 (1.05)	0 –	353 (1.19)	101 (1.03)	0 –	381 (1.30)	39 (1.08)	0 –	344 (1.30)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k, p, n) = (10, 50, 100)$			$(k, p, n) = (15, 75, 150)$			$(k, p, n) = (20, 100, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	1000	0	0	1000	0	0	1000	0
True	673	0	8	792	0	0	873	0	0
Over (Average)	327 (1.28)	0 –	992 (3.14)	208 (1.19)	0 –	1000 (6.72)	127 (1.06)	0 –	1000 (10.45)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k, p, n) = (10, 80, 100)$			$(k, p, n) = (15, 120, 150)$			$(k, p, n) = (20, 160, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	1000	0	0	1000	0	0	1000	0
True	0	0	0	0	0	0	0	0	0
Over (Average)	1000 (4.67)	0 –	1000 (5.00)	1000 (9.70)	0 –	1000 (10.00)	1000 (14.82)	0 –	1000 (15.00)

Table 2. Selection times of AIC, BIC and C_p methods under Settings (1) and (I) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k, p, n) = (10, 20, 100)$			$(k, p, n) = (15, 30, 150)$			$(k, p, n) = (20, 40, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	0	0
True	861	1000	641	908	1000	606	963	1000	634
Over	139	0	359	92	0	394	37	0	366
(Average)	(1.06)	–	(1.19)	(1.03)	–	(1.28)	(1.08)	–	(1.28)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k, p, n) = (10, 50, 100)$			$(k, p, n) = (15, 75, 150)$			$(k, p, n) = (20, 100, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	0	0
True	689	1000	170	798	1000	0	861	1000	0
Over	311	0	830	202	0	1000	139	0	1000
(Average)	(1.21)	–	(3.20)	(1.15)	–	(6.65)	(1.12)	–	(10.38)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k, p, n) = (10, 80, 100)$			$(k, p, n) = (15, 120, 150)$			$(k, p, n) = (20, 160, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	513	0	0	982	0	0	1000	0
True	0	487	0	0	18	0	0	0	0
Over	1000	0	1000	1000	0	1000	1000	0	1000
(Average)	(4.70)	–	(5.00)	(9.68)	–	(10.00)	(14.81)	–	(15.00)

Table 3. Selection times of AIC, BIC and C_p methods under Settings (1) and (II) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

5.1. Preliminaries

By (1.2) and (1.6), to prove the strong consistency of the AIC selection rule is equivalent to prove that

$$\mathbb{P}(\hat{\mathbf{j}}_A \neq \mathbf{j}_*, i.o.) = \mathbb{P}(\bigcup_{\mathbf{j} \neq \mathbf{j}_*} \{A_{\mathbf{j}} < A_{\mathbf{j}_*}\}, i.o.) = 0,$$

where $i.o.$ stands for infinitely often. Under condition (A2), i.e., $|\mathbf{J}| = O(n^{\ell})$ for some $\ell > 0$, then, by Borel-Cantelli Lemma, we only need to prove that for any $\mathbf{j} \in \mathbf{J} \backslash \{\mathbf{j}_*\}$, $\mathbb{P}(A_{\mathbf{j}} < A_{\mathbf{j}_*}) = o(n^{-\ell-2})$. The strong consistency of the BIC and C_p selection rules are analogous.

Step 1: We consider the over-specified case, i.e., $\mathbf{j} \in \mathbf{J}_+$. There exist m and $j_1 < j_2 < \cdots < j_m$ such that $\mathbf{j} = \mathbf{j}_* \cup \{j_1, \ldots, j_m\}$. One can construct a sequence of $m+1$ nested models in which we remove one spurious variable at a time until we attain the true model $\mathbf{j}_*$. Specifically, $\mathbf{j} = \mathbf{j}_m \supset \mathbf{j}_{m-1} \supset \cdots \supset \mathbf{j}_0 = \mathbf{j}_*$ where $\mathbf{j}_{t+1} = \mathbf{j}_t \cup \{j_{t+1}\}$ for $t = 0, 1, \ldots, m-1$. We remark that the order of removing which spurious variables makes no difference to our results. We have

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

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k, p, n) = (10, 20, 100)$			$(k, p, n) = (15, 30, 150)$			$(k, p, n) = (20, 40, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	128	0	0	930	0	0	986	0
True	852	872	656	916	70	632	959	14	639
Over (Average)	148 (1.06)	0 –	344 (1.22)	84 (1.04)	0 –	368 (1.20)	41 (1.02)	0 –	361 (1.32)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k, p, n) = (10, 50, 100)$			$(k, p, n) = (15, 75, 150)$			$(k, p, n) = (20, 100, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	1000	0	0	1000	0	0	1000	0
True	700	0	14	785	0	0	877	0	0
Over (Average)	300 (1.26)	0 –	986 (3.19)	215 (1.2)	0 –	1000 (6.57)	123 (1.04)	0 –	1000 (10.48)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k, p, n) = (10, 80, 100)$			$(k, p, n) = (15, 120, 150)$			$(k, p, n) = (20, 160, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	1000	0	0	1000	0	0	1000	0
True	0	0	0	0	0	0	0	0	0
Over (Average)	1000 (4.69)	0 –	1000 (5.00)	1000 (9.71)	0 –	1000 (10.00)	1000 (14.78)	0 –	1000 (15.00)

Table 4. Selection times of AIC, BIC and C_p methods under Settings (2) and (I) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

Based on the definition of $A_{\mathbf{j}}$ in (1.2), it follows that

$$\frac{1}{n}(A_{\mathbf{j}_t} - A_{\mathbf{j}_{t-1}}) = \log\left(\frac{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_t}|}{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{t-1}}|}\right) + 2c_n, \tag{5.1}$$

which implies

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \sum_{t=1}^{m}\left[\log\left(\frac{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_t}|}{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{t-1}}|}\right) + 2c_n\right]. \tag{5.2}$$

Analogously, we also have

$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = \sum_{t=1}^{m}\left[\log\left(\frac{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_t}|}{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{t-1}}|}\right) + \log(n)c_n\right] \tag{5.3}$$

and

$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = (1 - \alpha_k)\,\mathrm{tr}[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_*})] + 2mc_n$$

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k, p, n) = (10, 20, 100)$			$(k, p, n) = (15, 30, 150)$			$(k, p, n) = (20, 40, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	0	0
True	853	1000	679	911	1000	609	959	1000	631
Over	147	0	321	89	0	391	41	0	369
(Average)	(1.07)	–	(1.17)	(1.09)	–	(1.28)	(1.00)	–	(1.28)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k, p, n) = (10, 50, 100)$			$(k, p, n) = (15, 75, 150)$			$(k, p, n) = (20, 100, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	0	0
True	677	1000	9	802	1000	0	883	1000	0
Over	323	0	991	198	0	1000	117	0	1000
(Average)	(1.24)	–	(3.11)	(1.17)	–	(6.63)	(1.09)	–	(10.20)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k, p, n) = (10, 80, 100)$			$(k, p, n) = (15, 120, 150)$			$(k, p, n) = (20, 160, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	511	0	0	981	0	0	1000	0
True	0	489	0	0	19	0	0	0	0
Over	1000	0	1000	1000	0	1000	1000	0	1000
(Average)	(4.69)	–	(5.00)	(9.71)	–	(10.00)	(14.81)	–	(15.00)

Table 5. Selection times of AIC, BIC and C_p methods under Settings (2) and (II) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

$$= \sum_{t=1}^{m} \left((1 - \alpha_k) \operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1} (\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_t} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{t-1}})] + 2c_n \right). \tag{5.4}$$

Then, the lemma below follows.

Lemma 5.1. *Suppose that conditions* (A1) – (A4) *hold. For any over-specified model* $\mathbf{j}$ *with* $|\mathbf{j}| - k_* = m > 0$*, we have for any* $\varepsilon > 0$ *and* $t > 0$*,*

$$\mathbb{P}\left(|\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) - \sum_{t=1}^{m} \log\left(\frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) - 2mc_n| \geq m\varepsilon \right) = o(n^{-t}), \tag{5.5}$$

$$\mathbb{P}\left(|\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) - \sum_{t=1}^{m} \log\left(\frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) - mc_n \log(n)| \geq m\varepsilon \right) = o(n^{-t}), \tag{5.6}$$

$$\mathbb{P}\left(|\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) - m\psi(\alpha_k, c_n)| \geq m\varepsilon \right) = o(n^{-t}), \tag{5.7}$$

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k, p, n) = (10, 20, 100)$			$(k, p, n) = (15, 30, 150)$			$(k, p, n) = (20, 40, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	983	0	0	1000	0	0	1000	0
True	873	17	678	914	0	641	960	0	649
Over	127	0	322	86	0	359	40	0	351
(Average)	(1.08)	–	(1.21)	(1.07)	–	(1.31)	(1.00)	–	(1.31)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k, p, n) = (10, 50, 100)$			$(k, p, n) = (15, 75, 150)$			$(k, p, n) = (20, 100, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	40	1000	0	4	1000	0	1	1000	0
True	665	0	14	802	0	0	898	0	0
Over	295	0	986	194	0	1000	101	0	1000
(Average)	(1.28)	–	(3.20)	(1.15)	–	(5.70)	(1.09)	–	(10.30)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k, p, n) = (10, 80, 100)$			$(k, p, n) = (15, 120, 150)$			$(k, p, n) = (20, 160, 200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	1000	0	0	1000	0	0	1000	0
True	0	0	0	0	0	0	0	0	0
Over	1000	0	1000	1000	0	1000	1000	0	1000
(Average)	(4.68)	–	(5.00)	(10.72)	–	(10.00)	(14.81)	–	(15.00)

Table 6. Selection times of AIC, BIC and C_p methods under Settings (3) and (I) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

and

$$\frac{1}{m}\left(\sum_{t=1}^{m}\log\left(\frac{1-\alpha_t-c_n}{1-\alpha_t}\right)+2c_n\right)\to\phi(\alpha,c). \tag{5.8}$$

Remark 5.2. A consequence of this lemma is that $\frac{1}{nm}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) - \phi(\alpha_m, c_n) \to 0$ and $\frac{1}{nm}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) - (\log(n) - 2)c_n - \phi(\alpha_m, c_n) \to 0$ with tail probability $o(n^{-t})$ for any fixed $t > 0$, respectively. In addition, taking the AIC rule for example, this lemma indicates that for all $\mathbf{j} \in \mathbf{J}_+$ satisfying $|\mathbf{j}| - k_* = m$, if $\phi(\alpha_m, c_n) > 0$, then for sufficiently large p and n, then a.s. $\mathbf{j}$ will not be selected by the AIC rule. On the other hand, if $\phi(\alpha_m, c_n) < 0$, then for sufficiently large p and n, a.s. $\mathbf{j}_*$ will not be selected by the AIC rule, that is, the AIC rule is inconsistent. The arguments for BIC rule and C_p rule are analogous.

Step 2: We consider the under-specified case, i.e., $\mathbf{j} \in \mathbf{J}_-$. Let $\mathbf{j}_* \setminus \mathbf{j} = \{i_1, \ldots, i_s\}$ and $\mathbf{j} \setminus \mathbf{j}_* = \{j_1, \ldots, j_m\}$. We first assume m is positive. Define the model index set $\mathbf{j}_t = \mathbf{j} \cup \{i_{t+1}, \cdots, i_s\}$ for $t = 0, 1, \cdots, s$ (with the convention that $\mathbf{j}_s = \mathbf{j}$), which also indicates that i_t is in $\mathbf{j}_{t-1}$ but not in $\mathbf{j}_t$.

$\alpha = 0.1$, $c = 0.2$, $\phi = 0.16$ and $\psi = 0.14$									
	$(k,p,n) = (10,20,100)$			$(k,p,n) = (15,30,150)$			$(k,p,n) = (20,40,200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	0	0
True	860	1000	645	899	1000	620	963	1000	643
Over	140	0	355	101	0	380	37	0	357
(Average)	(1.14)	–	(1.24)	(1.05)	–	(1.32)	(1.03)	–	(1.32)

$\alpha = 0.1$, $c = 0.5$, $\phi = 0.25$ and $\psi = -0.13$									
	$(k,p,n) = (10,50,100)$			$(k,p,n) = (15,75,150)$			$(k,p,n) = (20,100,200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	0	0	0	0	0	0	3	0
True	698	1000	23	801	1000	0	896	997	0
Over	302	0	977	199	0	1000	104	0	1000
(Average)	(1.27)	–	(3.25)	(1.19)	–	(6.69)	(1.18)	–	(10.44)

$\alpha = 0.1$, $c = 0.8$, $\phi = -0.26$ and $\psi = -5.6$									
	$(k,p,n) = (10,80,100)$			$(k,p,n) = (15,120,150)$			$(k,p,n) = (20,160,200)$		
	AIC	BIC	C_p	AIC	BIC	C_p	AIC	BIC	C_p
Under	0	630	0	0	995	0	0	1000	0
True	0	370	0	0	5	0	0	0	0
Over	1000	0	1000	1000	0	1000	1000	0	1000
(Average)	(4.69)	–	(5.00)	(9.68)	–	(10.00)	(14.82)	–	(15.00)

Table 7. Selection times of AIC, BIC and C_p methods under Settings (3) and (II) based on 1,000 replications. When the selection rule over-specifies the true model, we also compute the average number of the spurious variables, simply referred as average.

Note also that $\mathbf{j}_0 = \mathbf{j} \cup \mathbf{j}_*$ is an over-specified model. So,

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \frac{1}{n}(A_{\mathbf{j}_s} - A_{\mathbf{j}_0}) + \frac{1}{n}(A_{\mathbf{j}\cup\mathbf{j}_*} - A_{\mathbf{j}_*}),$$
$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = \frac{1}{n}(B_{\mathbf{j}_s} - B_{\mathbf{j}_0}) + \frac{1}{n}(B_{\mathbf{j}\cup\mathbf{j}_*} - B_{\mathbf{j}_*}),$$
$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = \frac{1}{n}(C_{\mathbf{j}_s} - C_{\mathbf{j}_0}) + \frac{1}{n}(C_{\mathbf{j}\cup\mathbf{j}_*} - C_{\mathbf{j}_*}).$$

If $m = 0$, $\mathbf{j} \cup \mathbf{j}_* = \mathbf{j}_*$ and there will be no second terms in the right hand side of last three equations. Even though $m > 0$, Lemma 5.1 can be directly carried over to estimate these terms. Thus, what we need to consider is the first terms and it follows that

$$\frac{1}{n}(A_{\mathbf{j}_s} - A_{\mathbf{j}_0}) = \sum_{t=0}^{s-1} \log\left(\frac{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t}}|}{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t-1}}|}\right) - 2sc_n \tag{5.9}$$

$$\frac{1}{n}(B_{\mathbf{j}_s} - B_{\mathbf{j}_0}) = \sum_{t=0}^{s-1} \log\left(\frac{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t}}|}{|n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t-1}}|}\right) - \log(n)sc_n$$

and

$$\frac{1}{n}(C_{\mathbf{j}_s} - C_{\mathbf{j}_0}) = \sum_{t=0}^{s-1}(1-\alpha_k)\,\mathrm{tr}[\widehat{\boldsymbol{\Sigma}}_{\omega}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{s-t-1}})] - 2sc_n. \tag{5.10}$$

Then, the lemma below follows.

Lemma 5.3. *Suppose that assumptions* (A1) – (A4) *hold. If* $\mathbf{j} \in \mathbf{J}_-$ *with* $|\mathbf{j}_* \cap \mathbf{j}_-^c| = s > 0$ *and* $|\mathbf{j}_+| = m \geq 0$*, then we have for any* $\varepsilon > 0$ *and* $t > 0$*,*

$$\mathbb{P}\left(|\frac{1}{n}(A_{\mathbf{j}_s} - A_{\mathbf{j}_0}) - \log(\tau_{n\mathbf{j}}) + s\log(1-\alpha_m - c_n) + 2sc_n| \geq s\varepsilon\right) = o(n^{-t}), \tag{5.11}$$

$$\mathbb{P}\left(|\frac{1}{n}(B_{\mathbf{j}_s} - B_{\mathbf{j}_0}) - \log(\tau_{n\mathbf{j}}) + s\log(1-\alpha_m - c_n) + \log(n)sc_n| \geq s\varepsilon\right) = o(n^{-t}), \tag{5.12}$$

and

$$\mathbb{P}\left(|\frac{1}{n}(C_{\mathbf{j}_s} - C_{\mathbf{j}_0}) - \frac{(1-\alpha_k)(\kappa_{n\mathbf{j}} + sc_n)}{1-c_n-\alpha_k} + 2sc_n| \geq s\varepsilon\right) = o(n^{-t}). \tag{5.13}$$

The proofs of Lemmas 5.1 and 5.3 are presented in Section 6. The next lemma is a straightforward consequence of Lemmas 5.1 and 5.3.

Lemma 5.4. *Suppose that assumptions* (A1) – (A4) *hold. For all under-specified models* $\mathbf{j}$ *with* $|\mathbf{j}_* \cap \mathbf{j}_-^c| = s > 0$ *and* $|\mathbf{j}_+| = m \geq 0$*, we have for any* $\varepsilon > 0$ *and* $t > 0$*,*

$$\mathbb{P}\Bigg(\Bigg|\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) - \sum_{t=1}^{m}\log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) - \log(\tau_{n\mathbf{j}}) + s\log(1-\alpha_m - c_n) - 2(m-s)c_n\Bigg| \geq (m+s)\varepsilon\Bigg) = o(n^{-t}),$$

$$\mathbb{P}\Bigg(\Bigg|\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) - \sum_{t=1}^{m}\log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) - \log(\tau_{n\mathbf{j}}) + s\log(1-\alpha_m - c_n) - (m-s)c_n\log(n)\Bigg| \geq (m+s)\varepsilon\Bigg) = o(n^{-t}),$$

and

$$\mathbb{P}\left(\left|\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) - m\psi(\alpha_k, c_n) - \frac{(1-\alpha_k)(\kappa_{n\mathbf{j}} + sc_n)}{1-c-\alpha_k} + 2sc_n\right| \geq (m+s)\varepsilon\right) = o(n^{-t}).$$

Here we let $\sum_{t=1}^{m}\log\left(\frac{1-\alpha_t-c_n}{1-\alpha_t}\right) = 0$ *when* $m = 0$.

Now, using Lemmas 5.1 and 5.4, we prove Theorems 3.4–3.6.

5.2. Proof of Theorem 3.4

As $|\mathbf{J}| = O(n^{\ell})$, for the strong consistency of AIC rule, it is sufficient to prove for all $\mathbf{j} \neq \mathbf{j}_*$, $A_{\mathbf{j}} > A_{\mathbf{j}_*}$ holds with tail probability $o(n^{-\ell-2})$. If there exists $\mathbf{j} \in \mathbf{J}_+$ such that for all $\mathbf{j}' \in \{\mathbf{j}_*\} \cup \mathbf{J}_-$, $A_{\mathbf{j}} < A_{\mathbf{j}'}$ holds with tail probability $o(n^{-\ell-2})$, then the AIC selection rule almost surely over-specifies the true model. Analogously, if there exists $\mathbf{j} \in \mathbf{J}_-$ such that for all $\mathbf{j}' \in \{\mathbf{j}_*\} \cup \mathbf{J}_+$, $A_{\mathbf{j}} < A_{\mathbf{j}'}$ holds with tail probability $o(n^{-\ell-2})$, then the AIC selection rule almost surely under-specifies the true model. Therefore, according to Lemmas 5.1 and 5.4, what remains is to discuss the limits between different $A_{\mathbf{j}}$'s.

We first prove (1) of Theorem 3.4. When $\phi(\alpha, c) < 0$, it follows from Lemma 5.1 that

$$\frac{1}{nm}(A_{\mathbf{j}_*} - A_{\omega}) \xrightarrow{a.s.} -\phi(\alpha, c) > 0.$$

If $\mathbf{j} \in \mathbf{J}_-$ with $|\mathbf{j}_+| = m \geq 0$ and $|\mathbf{j}_* \cap \mathbf{j}_-^c| = s > 0$, by Lemmas 5.1 and 5.4,

$$\begin{aligned}
&\frac{1}{n}(A_{\mathbf{j}} - A_{\omega}) = \frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) + \frac{1}{n}(A_{\mathbf{j}_*} - A_{\omega}) \\
&= \sum_{t=1}^{m} \log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) - \sum_{t=1}^{k-k_*} \log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) \\
&\quad + \log(\tau_{n\mathbf{j}}) - s\log(1-\alpha_m - c_n) + 2(m - s - k + k_*)c_n + o_{a.s.}(k - k_* - m) \\
&= \log\left(\frac{\tau_{n\mathbf{j}}}{(1-\alpha_m)^s}\right) - \sum_{t=m-s+1}^{k-k_*} \log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) - 2(k - k_* - m + s)c_n + o_{a.s.}(k - k_* - m).
\end{aligned}$$

By the definition of $\tau_{n\mathbf{j}}$, it is easy to check $\log\left(\frac{\tau_{n\mathbf{j}}}{(1-\alpha_m)^s}\right) > 0$. When $\phi(\alpha, c) < 0$, we have that $\log\left(\frac{1-\alpha-c}{1-\alpha}\right) + 2c < 0$ and for sufficiently large p and n, $\sum_{t=m-s+1}^{k-k_*} \log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) + 2(k - k_* - m + s)c_n$ is negative with order of $k - k_* - m$ almost surely, which indicate that the limit of $\frac{1}{n}(A_{\mathbf{j}} - A_{\omega})$ is positive. Thus, in this case, the AIC rule asymptotically selects an over-specified model almost surely.

Next we consider (2). Because when $\mathbf{j} \in \mathbf{J}_+$ with $|\mathbf{j}| - k_* = m$, from the definition of $\phi(\alpha, c)$ (see Figure 1 for illustration), we know that if $\phi(\alpha, c) > 0$, then for any α_m satisfying $\lim \alpha_m \in [0, \alpha]$, we have $\phi(\alpha_m, c) > 0$. Thus, according to Lemma 5.1, if $\phi(\alpha, c) > 0$, AIC rule cannot asymptoticly over-specify the true model, i.e., for any $\mathbf{j} \in \mathbf{J}_+$ and for sufficiently large p and n,

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

uniformly. For the case of $\mathbf{j} \in \mathbf{J}_-$ with $|\mathbf{j}_+| = m \geq 0$ and $|\mathbf{j}_* \cap \mathbf{j}_-^c| = s > 0$, note that $s < k_*$, $\Phi_{\mathbf{j}} > 0$ and $\log(\tau_{n\mathbf{j}}) > s\log(1 - \alpha_m)$ uniformly. If $m \geq s$, by Lemma 5.4 and $\phi(\alpha, c) > 0$,

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

$$= \sum_{t=1}^{m-s} \log\left(\frac{1-\alpha_t - c_n}{1-\alpha_t}\right) + \log(\tau_{n\mathbf{j}}/(1-\alpha_m)^s) + 2(m-s)c_n + o_{a.s.}(m),$$

which is almost surely positive for sufficiently large p and n. Thus, we only need to consider the case in which $m < s$. In this case, since k_* is fixed, $\alpha_m = m/n \to 0$ and $\liminf_{n\to\infty} \tau_{n\mathbf{j}} > 1$. By the fact

$$\frac{1}{n}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) = \log(\tau_{n\mathbf{j}}) - (s-m)(\log(1-c) + 2c) + o_{a.s.}(1),$$

and if $\limsup_{n\to\infty} \log(\tau_{n\mathbf{j}}) < (s-m)(\log(1-c) + 2c)$, then from Lemmas 5.1 and 5.4, 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 rule asymptotically cannot select the true model $\mathbf{j}_*$. On the other hand, the condition $\phi(\alpha, c) > 0$ guarantees that the AIC cannot asymptotically select over-specified models. Then the AIC rule asymptotically selects an under-specified model. Thus, we complete the proof of Theorem 3.4. □

5.3. Proof of Theorem 3.5

By (5.6), it is easy to find that BIC rule cannot asymptoticly over-specify the true model. In addition, for any $\mathbf{j} \in \mathbf{J}_-$ with $|\mathbf{j}_+| = m \geq 0$, $|\mathbf{j}_* \cap \mathbf{j}_-^c| = s > 0$, and $m < s$, by Lemma 5.4, we obtain that

$$\frac{1}{n}(B_{\mathbf{j}} - B_{\mathbf{j}_*}) = \log(\tau_{n\mathbf{j}}) - (s-m)(\log(1-c) + \log(n)c) + o_{a.s.}(1), \tag{5.14}$$

which implies the conclusion (1).

The proof of (2) is analogous with the proof of Theorem 3.4. Thus, the details are not presented here. Then we complete the proof of this theorem. □

5.4. Proof of Theorem 3.6

By the same proof procedure of Theorem 3.4 with replacing ϕ by ψ, we can obtain this theorem. □

6. Proofs of Lemmas 5.1 and 5.3

In this section, we present the technical proofs of Lemma 5.1 and Lemma 5.3. We first briefly describe our proof strategy and the main tools of RMT. Note that the distribution of these statistics in Lemmas 5.1 and 5.3 are invariant under the transformation $\mathbf{Y} \to \mathbf{Y}\boldsymbol{\Sigma}^{-1/2}$ and $\Theta \to \Theta\boldsymbol{\Sigma}^{-1/2}$. Thus, without loss of generality, we assume that $\boldsymbol{\Sigma} = \mathbf{I}_p$ in the sequel. For Lemma 5.1, as j_t is in $\mathbf{j}_t$ but not in $\mathbf{j}_{t-1}$, we denote $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{x}_{j_t}/\|\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{x}_{j_t}\|$ for $t \geq 1$. Then by Sylvester's determinant theorem, we have that

$$\begin{aligned} |n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_t}| =& |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_t}\mathbf{Y}| = |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y} - \mathbf{Y}'\mathbf{a}_t\mathbf{a}_t'\mathbf{Y}| \\ =& |\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y}|(1 - \mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t) \\ =& |n\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}_{t-1}}|(1 - \mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t). \end{aligned} \tag{6.1}$$

Thus, to prove Lemma 5.1, we need to obtain only the limits of $\mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t$ or similar expressions with different $\mathbf{j}_t$. The proof strategy is that we first define a function

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

where $\mathbb{C}^+ = \{z \in \mathbb{C}^+ : \Im z > 0\}$. Next, we prove that outside a null set independent of $\mathbf{j}_{t-1}$, 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, e.g., Lemma 2.14 in (Bai and Silverstein, 2010)), it is sufficient to prove that, for any fixed $z \in \mathbb{C}^+$, $\hbar_n(z) \xrightarrow{a.s.} \hbar(z)$. Finally, we let $z \downarrow 0 + 0i$ and obtain almost surely $\hbar_n(0) \to \hbar(0)$ uniformly. The proof strategy of Lemma 5.3 is analogous.

We remark that this proof approach is common in RMT to obtain the 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 $\mathbf{a}_t'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_t$, we have to derive some new theoretical results for our theorems.

Now, we are in position to prove Lemma 5.1 and Lemma 5.3.

6.1. Proof of Lemma 5.1

We first prove (5.5). By equation (6.1) 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

$$\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_t}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{t-1}}|}\right) = \log(1 - \mathbf{a}_t'\mathbf{E}(\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t) \tag{6.2}$$

and

$$n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_t} - n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{t-1}} = -\mathbf{E}'\mathbf{a}_t\mathbf{a}_t'\mathbf{E}. \tag{6.3}$$

It follows from (5.2) and (6.2) that

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

Since $\mathbf{a}_t$ is an eigenvector of $\mathbf{Q}_{\mathbf{j}_{t-1}}$, we have $\mathbf{a}_t'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{a}_t = 1$, which together with Proposition 3.1 and (3.9) implies

$$\frac{1}{p}\mathbf{a}_t'\mathbf{E}(\frac{1}{p}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{t-1}}\mathbf{E} - z\mathbf{I}_p)^{-1}\mathbf{E}'\mathbf{a}_t - 1 + \frac{1}{1 + \underline{s}_{n\mathbf{j}_{t-1}}(z)} \to 0, \tag{6.4}$$

with tail probability $o(n^{-t})$ for any fixed $t > 0$. Therefore, by (3.10) and as $n \to \infty$, we have

$$\frac{1}{nm}(A_{\mathbf{j}} - A_{\mathbf{j}_*}) - \frac{1}{m}\sum_{t=1}^{m}\left(\log(1 - \frac{c_n}{1-\alpha_t}) + 2c_n\right) \to 0. \tag{6.5}$$

with tail probability $o(n^{-t})$ for any fixed $t > 0$, which implies (5.5). From integration by parts we have that $\frac{1}{n^2}(A_{\mathbf{j}} - A_{\mathbf{j}_*})$ tends to

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

which indicates (5.8).

(5.6) is analogous; thus, we omit the details. Next, we prove (5.7). It follows from (5.4) and (6.3) that

$$\frac{1}{n}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) = \sum_{t=1}^{m} \left((\frac{k}{n} - 1)\mathbf{a}_t' \mathbf{E}(\mathbf{E}'\mathbf{Q}_\omega \mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t + 2c_n \right). \tag{6.6}$$

By (3.8) and (3.10) and the fact that

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

we have

$$\mathbf{a}_t' \mathbf{E}(\mathbf{E}'\mathbf{Q}_\omega \mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_t \to \frac{c_n}{1 - \alpha_k - c_n},$$

with tail probability $o(n^{-t})$ for any fixed $t > 0$, which together with (6.6) implies

$$\frac{1}{nm}(C_{\mathbf{j}} - C_{\mathbf{j}_*}) - \frac{c_n(\alpha_k - 1)}{1 - \alpha_k - c_n} - 2c_n \to 0,$$

with tail probability $o(n^{-t})$ for any fixed $t > 0$. Thus, we complete the proof of Lemma 5.1. □

Remark 6.1. The conclusion (5.5) will be clearer if we assume normality of the errors. Here is a sketch of a more direct understanding. It is well known that if $e_{ij} \sim N(0,1)$, then

$$n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_t} \sim W_p(n - |\mathbf{j}_t|, \mathbf{I}_p), \quad n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{t-1}} \sim W_p(n - |\mathbf{j}_{t-1}|, \mathbf{I}_p)$$

and

$$\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_t}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{t-1}}|} \sim \left(1 + \frac{\chi_p^2}{\chi_{n-|\mathbf{j}_t|-p+1}^2}\right)^{-1}$$

where W_p and χ_p^2 are the Wishart distribution and chi-squared distribution with degrees of freedom p, respectively. From the strong law of large numbers, it is not difficult to check that

$$(1/n)\chi_p^2 \to c_n \quad \text{and} \quad (1/n)\chi_{n-|\mathbf{j}_t|-p+1}^2 \to 1 - \alpha_t - c_n,$$

with tail probability $o(n^{-t})$ for any $t > 0$, which together with (5.1) imply (5.5) directly.

6.2. Proof of Lemma 5.3

We start with $\frac{1}{n}(A_{\mathbf{j}_s} - A_{\mathbf{j}_0})$ and $\frac{1}{n}(B_{\mathbf{j}_s} - B_{\mathbf{j}_0})$. As i_t is in $\mathbf{j}_{t-1}$ but not in $\mathbf{j}_t$, we denote $\mathbf{a}_t = \mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{i_t} / \|\mathbf{Q}_{\mathbf{j}_t}\mathbf{x}_{i_t}\|$. By the fact that $\mathbf{Q}_{\mathbf{j}_{t-1}} = \mathbf{Q}_{\mathbf{j}_t} - \mathbf{a}_t\mathbf{a}_t'$ and equation (6.1), we have that

$$\log\left(\frac{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{s-t}}|}{|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{s-t-1}}|}\right) = -\log(1 - \mathbf{a}_{s-t}'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{s-t}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_{s-t}).$$

To evaluate the limit of right hand side of last equation, we consider

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

where $z \in \mathbb{C}^+$. On the basis of the fact that $\mathbf{a}_{s-t} = \mathbf{Q}_{\mathbf{j}_{s-t}} \mathbf{a}_{s-t}$ and the in-out-exchange formula (3.4), we rewrite m_{nt} as

$$m_{nt} = -\log\left(-z\mathbf{a}'_{s-t}\left(\mathbf{Q}_{\mathbf{j}_{s-t}}\mathbf{Y}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{s-t}}/p - z\mathbf{I}_n\right)^{-1}\mathbf{a}_{s-t}\right). \tag{6.7}$$

Substitute model (2.1) into the above equation and denote

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

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

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

With $\mathbf{a}'_{s-t}\mathbf{B}_2 = 0$, we obtain

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

where $\tilde{\mathbf{a}}_{s-t} = \mathbf{B}'_1\mathbf{a}_{s-t}$. By applying in-out-exchange formula (3.4) to the term

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

we obtain

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

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

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

Equations (3.1)-(3.4) can be used to separate I_t into the following four parts,

$$I_t = -\tilde{\mathbf{a}}'_{s-t}\left(I_{1t} + I_{2t} + I'_{2t} + I_{3t}\right)^{-1}\tilde{\mathbf{a}}_{s-t} \tag{6.8}$$

where

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

It follows from (3.8) and (3.9) that with tail probability $o(n^{-t})$ for any fixed $t > 0$,

$$\frac{1}{p}\mathbf{B}_1'\mathbf{Q}_{\mathbf{j}_{s-t}}\mathbf{E}\mathbf{M}_{\mathbf{j}_{s-t}}^{-1}\mathbf{E}'\mathbf{Q}_{\mathbf{j}_{s-t}}\mathbf{B}_1 - (1 - \frac{1}{1 + \underline{s}_{n\mathbf{j}_{s-t}}(z)})\mathbf{I}_{s-t} \to 0,$$

which implies

$$I_{3t} - (1 + \underline{s}_{n\mathbf{j}_{s-t}}(z))\mathbf{I}_{s-t} \to 0. \tag{6.9}$$

Moreover, from (3.6), (3.7) and assumption (A1), we have with tail probability $o(n^{-t})$ for any fixed $t > 0$,

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

and

$$\frac{1}{p}\mathbf{B}_1'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\mathbf{M}_{\mathbf{j}_{s-t}}^{-1}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{B}_1 + \frac{p^{-1}\mathbf{B}_1'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{B}_1}{z(1 + \underline{s}_{n\mathbf{j}_{s-t}}(z) - \frac{1-c_n-\alpha_m}{c_n z})} \to 0,$$

which together with (6.9) imply

$$I_t + \tilde{\mathbf{a}}_{s-t}'\left((1 + \underline{s}_{n\mathbf{j}_{s-t}}(z))\mathbf{I}_{s-t} - \frac{p^{-1}\mathbf{B}_1'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{B}_1}{z(1 + \underline{s}_{n\mathbf{j}_{s-t}}(z) - \frac{1-c_n-\alpha_m}{c_n z})}\right)^{-1}\tilde{\mathbf{a}}_{s-t} \to 0.$$

Let $z \downarrow 0 + 0i$, together with (3.10) and the notation

$$\delta_t := \tilde{\mathbf{a}}_{s-t}'\left((1 - \alpha_m)\mathbf{I}_{s-t} + n^{-1}\mathbf{B}_1'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{B}_1\right)^{-1}\tilde{\mathbf{a}}_{s-t},$$

we have

$$I_t(0) + (1 - \alpha_m - c_n)\delta_t \to 0,$$

with tail probability $o(n^{-t})$ for any fixed $t > 0$. Here, we use the fact that α_{m-t} and α_m have the same limit. By basic calculation, we obtain the equation

$$\prod_{t=0}^{s-1}\delta_t = \tau_{n\mathbf{j}}^{-1},$$

which together with (5.9) implies (5.11).

Next, we consider (5.13). Recall (5.10)

$$\begin{aligned}\frac{1}{n}(C_{\mathbf{j}}-C_{\mathbf{j}_*}) &= \sum_{t=0}^{s-1}(1-\alpha_k)\,\mathrm{tr}[\widehat{\mathbf{\Sigma}}_{\omega}^{-1}(\widehat{\mathbf{\Sigma}}_{\mathbf{j}s-t}-\widehat{\mathbf{\Sigma}}_{\mathbf{j}s-t-1})] - 2sc_n \\ &= \sum_{t=0}^{s-1}(1-\alpha_k)\mathbf{a}_{s-t}'\mathbf{Y}(\mathbf{E}'\mathbf{Q}_{\omega}\mathbf{E})^{-1}\mathbf{Y}'\mathbf{a}_{s-t} - 2sc_n.\end{aligned}$$

Let

$$J_t(z) := J_t(z) = p^{-1}\mathbf{a}_{s-t}'\mathbf{Y}\left(\mathbf{E}'\mathbf{Q}_{\omega}\mathbf{E}/p - z\mathbf{I}_p\right)^{-1}\mathbf{Y}'\mathbf{a}_{s-t}.$$

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

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

It follows from Proposition 3.1 that with tail probability $o(n^{-t})$ for any fixed $t>0$,

$$J_{1t} + \frac{\frac{1}{p}\mathbf{a}_{s-t}'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{a}_{s-t}}{z(1+\underline{s}_{n\boldsymbol{\omega}}(z)-\frac{1-c_n-\alpha_k}{c_nz})} \to 0,$$

$$J_{2t}\to 0 \quad \text{and} \quad J_{3t} + \frac{1}{z(1+\underline{s}_{n\boldsymbol{\omega}}(z)-\frac{1-c_n-\alpha_k}{c_nz})} \to 0.$$

It follows from

$$\frac{1}{n}\sum_{t=0}^{t-1}\mathbf{a}_{s-t}'\mathbf{X}_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}\Theta_{\boldsymbol{\ell}_t}'\mathbf{X}_{\boldsymbol{\ell}_t}'\mathbf{a}_{s-t} = \frac{1}{n}\,\mathrm{tr}[\Theta_{\mathbf{j}_*}'\mathbf{X}_{\mathbf{j}_*}'(\sum_{t=0}^{s-1}\mathbf{a}_{s-t}\mathbf{a}_{s-t}')\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}] = \kappa_{n\mathbf{j}}$$

and

$$\lim_{z\downarrow 0+0i}\frac{1}{z(1+\underline{s}_{n\boldsymbol{\omega}}(z)-\frac{1-c_n-\alpha_k}{c_nz})} = \frac{c_n}{1-\alpha_k-c_n},$$

we obtain (5.13). Thus, we complete the proof of Lemma 5.3. □

7. Conclusion and discussion

In this paper, we study strong consistency in variable selection of three commonly used selection criteria, AIC, BIC, and C_p, in multivariate linear regression under a 3L framework. We provide a rather comprehensive description of how the sample size, the number of response variables and the number of predictors will affect the selection accuracy. We confine ourselves to consider the case in which $\alpha + c < 1$ otherwise $\hat{\mathbf{\Sigma}}_{\mathbf{j}}$ may be singular. Singularity of $\hat{\mathbf{\Sigma}}_{\mathbf{j}}$ can be circumvented by using a ridge-type estimator of the covariance matrix (e.g., Chen *et al.* (2011), Yamamura, Yanagihara and Srivastava (2010)). It is of interest to study the asymptotic properties of these model selection criteria when the true model size k_* is also large, i.e., k_*/n tends to a constant as $n \to \infty$.

The key mathematical tool in the present paper is RMT. For the past two decades, RMT techniques have played many significant roles in the study of high-dimensional multivariate statistics analysis. Many results in classical multivariate analysis, including the model selection problems considered in this paper, have been reexamined by RMT in high-dimensional settings. Some of these results show that the classical results in multivariate analysis under the large-sample framework do not carry over to high-dimensional framework. In the big-data era, there is a pressing need to re-examine which classical results in multivariate statistics carry over to and which do not in high-dimensional framework. Our view is that RMT will prove to be a powerful tool in future studies of high-dimensional problems.

Acknowledgements

We thank the AE and the reviewers for careful reading of the manuscript and their helpful comments that improve this paper. The authors in this paper are listed in alphabetical order.

Funding

Bai's research was supported by NSFC No. 12171198 and STDFJ No. 20210101147JC. Choi's research was supported by the Singapore MOE Academic Research Funds R-155-000-188-114. Hu's research was supported by NSFC No. 12171078, 11971097 and National Key R&D Program of China No. 2020YFA0714102. Choi and Fujikoshi acknowledged the funding support of NSFC No.11771073 to visit Northeast Normal University where part of this research project was conducted.

Supplementary Material

Supplement to "Asymptotics of AIC, BIC and C_p model selection rules in high-dimensional regression" (DOI: 10.3150/21-BEJ1422SUPP; .pdf). This supplementary material gives the proof of Proposition 3.1.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. MR0483125

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

Bai, Z.D., Choi, K.P. and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Ann. Statist.* **46** 1050–1076. MR3797996 https://doi.org/10.1214/17-AOS1577

Bai, Z.D., Miao, B. and Pan, G. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35** 1532–1572. MR2330979 https://doi.org/10.1214/009117906000001079

Bai, Z.D. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

Bai, Z.D., Choi, K.P., Fujikoshi, Y. and Hu, J. (2022). Supplement to "Asymptotics of AIC, BIC and C_p model selection rules in high-dimensional regression." https://doi.org/10.3150/21-BEJ1422SUPP

Bao, Z., Hu, J., Pan, G. and Zhou, W. (2019). Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case. *Ann. Statist.* **47** 612–640. MR3909944 https://doi.org/10.1214/18-AOS1704

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. MR0914460 https://doi.org/10.1007/BF02294361

Chen, L.S., Paul, D., Prentice, R.L. and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.* **106** 1345–1360. MR2896840 https://doi.org/10.1198/jasa.2011.ap10599

Enomoto, R., Sakurai, T. and Fujikoshi, Y. (2015). Consistency properties of AIC, BIC, Cp and their modifications in the growth curve model under a large-(q,n) framework. *SUT J. Math.* **51** 59–81. MR3409058

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

Fan, Y. and Tang, C.Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478 https://doi.org/10.1111/rssb.12001

Fujikoshi, Y. (1983). A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.* **13** 203–214. MR0693557

Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate Anal.* **17** 27–37. MR0797518 https://doi.org/10.1016/0047-259X(85)90092-2

Fujikoshi, Y., Enomoto, R. and Sakurai, T. (2013). High-dimensional AIC in the growth curve model. *J. Multivariate Anal.* **122** 239–250. MR3189321 https://doi.org/10.1016/j.jmva.2013.07.006

Fujikoshi, Y. and Sakurai, T. (2016a). High-dimensional consistency of rank estimation criteria in multivariate linear model. *J. Multivariate Anal.* **149** 199–212. MR3507324 https://doi.org/10.1016/j.jmva.2016.04.005

Fujikoshi, Y. and Sakurai, T. (2016b). Some properties of estimation criteria for dimensionality in principal component analysis. *Amer. J. Math. Management Sci.* **35** 133–142.

Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.* **123** 184–200. MR3130429 https://doi.org/10.1016/j.jmva.2013.09.006

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84** 707–716. MR1603952 https://doi.org/10.1093/biomet/84.3.707

Fujikoshi, Y. and Veitch, L.G. (1979). Estimation of dimensionality in canonical correlation analysis. *Biometrika* **66** 345–351. MR0548204 https://doi.org/10.1093/biomet/66.2.345

Heinze, G., Wallisch, C. and Dunkler, D. (2018). Variable selection—a review and recommendations for the practicing statistician. *Biom. J.* **60** 431–449. MR3802914 https://doi.org/10.1002/bimj.201700067

Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Statist.* **45** 897–922. MR3650404 https://doi.org/10.1214/16-AOS1474

Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240 https://doi.org/10.1111/biom.12292

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900 https://doi.org/10.1080/01621459.2012.695654

Mallows, C.L. (1973). Some comments on Cp. *Technometrics* **15** 661.

Nishii, R., Bai, Z.D. and Krishnaiah, P.R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.* **18** 451–462. MR0991240

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. With comments and a rejoinder by the author. MR1466682

Silverstein, J.W. and Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54** 295–309. MR1345541 https://doi.org/10.1006/jmva.1995.1058

Sparks, R.S., Coutsourides, D. and Troskie, L. (1983). The multivariate C_p. *Comm. Statist. Theory Methods* **12** 1775–1793. MR0704853 https://doi.org/10.1080/03610928308828569

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

Yanagihara, H. (2015). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of the normality assumption. *J. Japan Statist. Soc.* **45** 21–56. MR3444402 https://doi.org/10.14490/jjss.45.21

Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.* **9** 869–897. MR3338666 https://doi.org/10.1214/15-EJS1022

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

Received May 2020 and revised September 2021