Asymptotics of AIC, BIC and $C_p$ Model Selection Rules in High-Dimensional Regression

ZHIDONG BAI1,*, KWOK PUI CHOI2, YASUNORI FUKIKOSHI3 and JIANG HU 1,†

1 KLASMOE and School of Mathematics & Statistics, Northeast Normal University, China.
E-mail: *baidz@nenu.edu.cn; †huj156@nenu.edu.cn
2 Department of Statistics and Applied Probability, National University of Singapore, Singapore.
E-mail: stackp@nus.edu.sg
3 Department of Mathematics, Graduate School of Science, Hiroshima University, Japan.
E-mail: 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.

MSC2020 subject classifications: Primary 62J05; 62B20; secondary 62H12
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 $(\tilde{x}_i, y_i)$ for $1 \leq i \leq n$ is the multivariate linear regression (MLR) model. Here $y_i = (y_{i1}, \ldots, y_{ip})^T$ with $y_{ij}$ the $j$-th response variable of the $i$-th observation, and $\tilde{x}_i = (x_{i1}, \ldots, x_{ik})^T$ with $x_{ij}$ the $j$-th predictor variable of the $i$-th observation. Specifically, we consider

$$Y = X\Theta + \varepsilon \Sigma^{1/2},$$

where the $n \times p$ response matrix $Y = (y_{ij}) = (y_1, \ldots, y_n)^T$, the $n \times k$ predictor matrix $X = (\tilde{x}_1, \ldots, \tilde{x}_n)^T = (x_1, \ldots, x_k)$, the $k \times p$ regression coefficient matrix $\Theta = (\theta_1, \ldots, \theta_k)^T$, the $n \times p$
random errors matrix \( E = (e_1, \ldots, e_p) = (e_{ij}) \) and the \( p \times p \) covariance matrix \( \Sigma \). 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 \( \Theta \). 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 \( \Theta \) 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 Mallows’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 \( j \) be a subset of \( \omega = \{1, 2, \cdots, k\} \) and \( X_j = (x_j, j \in j) \) and \( \Theta_j = (\theta_j, j \in j)' \). Denote model \( j \) by

\[
M_j : \quad Y = X_j \Theta_j + E \Sigma^{1/2}. \tag{1.2}
\]

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

\[
A_j = n \log(|\hat{\Sigma}_j|) + 2 \left( |j| \rho + \frac{1}{2} p(\rho + 1) \right) + np(\log(2\pi) + 1), \tag{1.3}
\]

where

\[
n\hat{\Sigma}_j = Y'Q_j Y, \quad Q_j = I_n - P_j, \quad P_j = X_j (X_j'X_j)^{-1} X_j'. \tag{1.4}
\]

Here, \( I_n \) is the identity matrix of order \( n \), \( |j| \) the cardinality of set \( j \), \( |\hat{\Sigma}_j| \) the determinant \( \hat{\Sigma}_j \), \( P_j \) an orthogonal projection of rank \( |j| \) onto the subspace spanned by \( X_j \), and \( Q_j \) the orthogonal projection of rank \( n - |j| \) onto the orthogonal complement subspace spanned by \( X_j \).
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.,

$$B_j = n \log(|\hat{\Sigma}_j|) + \log(n) \left( |\hat{\beta}|^2 + \frac{1}{2} \hat{\beta}(\hat{\beta} + 1) \right) + np(\log(2\pi) + 1).$$  \hspace{1cm} (1.5)

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:

$$C_j = (n - k)\text{tr}(\hat{\Sigma}_j^{-1} \hat{\Sigma}_j) + 2p|\hat{\beta}|.$$  \hspace{1cm} (1.6)

Refer to Fujikoshi (1983); Sparks et al. (1983); Nishii et al. (1988) for additional details of formulas (1.3), (1.5) and (1.6). Then, the AIC, BIC, and $C_p$ selection rules are respectively used to select

$$\hat{\beta}_A = \arg \min_{\beta} A_j, \hspace{0.5cm} \hat{\beta}_B = \arg \min_{\beta} B_j \hspace{0.5cm} \text{and} \hspace{0.5cm} \hat{\beta}_C = \arg \min_{\beta} C_j,$$  \hspace{1cm} (1.7)

where $J$ 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 et al. (1988). Very recently, Fujikoshi et al. (2014); Yanagihara et al. (2015); Yanagihara (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 Shao (1997); Anzanello and Fogliatto (2014); Heinze et al. (2018) 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 et al. (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 et al. (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 et al. (2015) and Fujikoshi et al. (2013), multiple discriminant analysis in Fujikoshi (1983) and Fujikoshi and Sakurai (2016a), principal component analysis in Fujikoshi and Sakurai (2016b) and Bai et al. (2018), and canonical correlation analysis in Nishii et al. (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 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 $(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_n, y_n)$ 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)

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

where $\Theta$ is a $k \times p$ unknown matrix of regression coefficients and $\Sigma$ is a $p \times p$ unknown positive definite covariance matrix.

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

$$M_j: \quad \mathbf{Y} = \mathbf{X}_j\Theta_j + \mathbf{E}\Sigma^{1/2}. \quad (2.1)$$

For simplicity, we refer $M_j$ as model $j$. Model $j$ is equivalent to the multivariate linear regression using just the subset of predictors in $j$ and the corresponding regression coefficients. Denote the true model
Asymptotics of AIC, BIC and $C_p$

as $j_\ast$, and

$$M_{j_\ast} : \ Y = X_{j_\ast} \Theta_{j_\ast} + E \Sigma^{1/2}. $$

We partition $J$ into $J_- \cup \{j_\ast\} \cup J_+$ where

$$J_+ = \{j : j \geq j_\ast\}, \quad J_- = \{j : j \ngeq j_\ast\} = \{j : j_\ast \setminus j \neq \emptyset\}. $$

For any $j \in J_-$, we partition $j$ into $j_+ \cup j_-$, where $j_+ = j \cap j_\ast^c$ and $j_- = j \cap j_\ast$. We say that a model $j$ is over-specified if $j \in J_+$ and under-specified if $j \in J_-$. We call variable $x_j$ if $j \in j_\ast$ a true variable, variable $x_j$ if $j \notin j_\ast$ a spurious variable, and variable $x_j$ a missing variable in $j \in J_- \text{ if } j \in j_\ast \setminus j$.

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 $\omega \in J$, the true model $j_\ast \in J$, $j_\ast$ is fixed, and $|J| = O(n^\ell)$ for some $\ell > 0$.

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

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

for any $\eta > 0$.

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

**Remark 2.1.** (i) For $j \in J$, condition (A4) implies $X_j'X_j$ is invertible because it is a principal submatrix of $X'X$. (ii) Requiring the full model, $\omega \in J$ and the true model $j_\ast \in J$ is natural. The condition that $j_\ast$ is fixed can be relaxed further but for the sake of simpler presentation, it is not pursued in this paper. The role of $|J| = O(n^\ell)$ 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 $|j_\ast|$ is known, we can choose $J$ to consist of all subsets of $\omega$ 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 et al. (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 $A_n$ with only real eigenvalues, let $F^{A_n}$ be the empirical spectral distribution function of $A_n$, that is,

$$F^{A_n}(x) = \frac{1}{n} \left| \left\{ i : \lambda^n_i \leq x \right\} \right|,$$
where $\lambda_i^{A_n}$ denotes the $i$-th largest eigenvalue of $A_n$. If $F^{A_n}$ has a limiting distribution $F$, then we call it the limiting special distribution (LSD) of sequence $\{A_n\}$. For any function of bounded variation $G$ on the real line, its Stieltjes transform is defined by

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

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

$$(A - CC')^{-1} = A^{-1} + A^{-1}C(I_n - C'A^{-1}C)^{-1}C'A^{-1}, \quad (3.1)$$

which immediately implies

$$(A - CC')^{-1}C = A^{-1}C(I_n - C'A^{-1}C)^{-1}, \quad (3.2)$$

$$C'(A - CC')^{-1} = (I_n - C'A^{-1}C)^{-1}C'A^{-1}. \quad (3.3)$$

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

$$C(C'C - zI_n)^{-1}C' = I_p + z(CC' - zI_p)^{-1}, \quad (3.4)$$

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 $\frac{1}{p}E'Q_jE$, denoted by $\s(z)$, is the unique solution on the upper complex plane to the equation

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

where $H$ is the LSD of $Q_j$ (see (1.4) of (Silverstein and Choi, 1995) for more details) and $E$ 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 $M := M(z) = pE'Q_jE - zI_p$, $m = |j|$, $\alpha_1$ and $\alpha_2$ be non-random $p$-vectors, $\alpha_3$ and $\alpha_4$ be non-random $n$-vectors and assume that $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\alpha_4$ are all bounded in Euclidean norm. Then, under conditions (A1) – (A4), we have that for any $z \in C^+$, $t > 0$ and $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\alpha_1'M^{-1}\alpha_2 - \s_{\alpha_1}(z)\alpha_1'\alpha_2\right| \geq \varepsilon\right) = o(n^{-t}), \quad (3.6)$$

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\alpha_1'M^{-1}E'\alpha_3\right| \geq \varepsilon\right) = o(n^{-t}), \quad (3.7)$$

and

$$\mathbb{P}\left(\left|\frac{1}{p}\alpha_3EM^{-1}E'\alpha_4 - \s_{\alpha_3}(z)\alpha_3\alpha_4 + \frac{\s_{\alpha_3}(z)}{\s_{\alpha_1}(z)} + 1\alpha_3'Q_j\alpha_4\right| \geq \varepsilon\right) = o(n^{-t}), \quad (3.8)$$

where $\s_{\alpha_j}(z)$ is the Stieltjes transform of the LSD of $\frac{1}{p}E'Q_jE$ with $c$ and $H$ replaced by $c_n$ and $FQ_j$, 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 \( n^{-1} (E'Q_j E)^{-1} \alpha_2 \), \( n^{1/2} (E'Q_j E)^{-1} \alpha_3 \) and \( \alpha_2 \) \( E'Q_j E)^{-1} \alpha_4 \), which are equivalent to deriving \( \lim_{z \to 0+0i} s_{n,j}(z) \) and \( \lim_{z \to 0+0i} z s_{n,j}(z) \) and then letting \( n \to \infty \). By the fact that \( F_{Q_j}(\{0\}) = F_{Q_j}^{\prime}(\{0\}) = 1 - \alpha_m \) and

\[
z = -\frac{1}{s_{n,j}(z)} + \frac{1 - \alpha_m}{c_n + \frac{z}{1 + s_{n,j}(z)}}.
\]

we have

\[
\frac{z(1 + s_{n,j}(z) + \frac{c_n - 1 + \alpha_m}{c_n z}) + \frac{1}{s_{n,j}(z) + 1}}{z^2(1 + s_{n,j}(z) + \frac{c_n - 1 + \alpha_m}{c_n z})^2} = \frac{1}{1 + s_{n,j}(z)} - 1 \tag{3.9}
\]

and

\[
s_{n,j}(z) = \frac{1 - \alpha_m - c_n - c_n z \pm \sqrt{(1 - \alpha_m + c_n - c_n z)^2 - 4c_n(1 - \alpha_m)}}{2c_n z}.
\]

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

\[
s_{n,j}(z) = \frac{1 - \alpha_m - c_n - c_n z + \sqrt{(1 - \alpha_m + c_n - c_n z)^2 - 4c_n(1 - \alpha_m)}}{2c_n z},
\]

and

\[
1 - \frac{1}{1 + s_{n,j}(z)} = \frac{1 - \alpha_m + c_n - c_n z + \sqrt{(1 - \alpha_m + c_n - c_n z)^2 - 4c_n(1 - \alpha_m)}}{2(1 - \alpha_m)}.
\]

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

\[
s_{n,j}(z) \to \frac{c_n}{1 - \alpha_m - c_n} \tag{3.10}
\]

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,

\[
\lim_{z \to 0+0i} \sqrt{(1 - \alpha_m + c_n - c_n z)^2 - 4c_n(1 - \alpha_m)} = -|1 - \alpha_m - c_n|.
\]

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, \( Q_j \) 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 et al. (2015); Fujikoshi et al. (2013), multiple discriminant analysis Fujikoshi (1983); Fujikoshi and Sakurai (2016a), principal component analysis Fujikoshi and Sakurai (2016b); Bai et al. (2018), and canonical correlation analysis Nishii et al. (1988); Bao et al. (2019). We will pursue these applications in future work.
3.2. Asymptotics of AIC, BIC and $C_p$ selection rules

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

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

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

\[
\Phi_j := \frac{1}{n} \Sigma^{-1/2} \Theta_j^T X_j^T Q_j X_j \Theta_j \Sigma^{-1/2},
\]

\[
\tau_{nj} := (1 - \alpha_m)^{d-p}(1 - \alpha_m)I_p + \Phi_j,
\]

\[
\kappa_{nj} := \text{tr}(\Phi_j)
\]

where $Q_j$ is defined as in (1.4). 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 $j \in J_-$ with $s - m > 0$ such that
   
   \[
   \liminf_{n \to \infty} \log(\tau_{nj}) > (s - m) (2c + \log(1 - c)),
   \]
   
   then the AIC selection rule is strongly consistent.
   
   (ii) If there exists an under-specified model $j \in J_-$ with $s - m > 0$ such that
   
   \[
   \limsup_{n \to \infty} \log(\tau_{nj}) < (s - m) (2c + \log(1 - c)),
   \]
   
   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 $j \in J_-$ with $s - m > 0$ such that
   
   \[
   \liminf_{n \to \infty} \left(\log(\tau_{nj}) - c(s - m) \log(n)\right) > (s - m) \log(1 - c),
   \]
   
   then the BIC selection rule is strongly consistent.

2. If there exists an under-specified model $j \in J_-$ with $s - m > 0$ such that
   
   \[
   \limsup_{n \to \infty} \left(\log(\tau_{nj}) - c(s - m) \log(n)\right) < (s - m) \log(1 - c),
   \]
   
   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.
Asymptotics of AIC, BIC and $C_p$

(2) Suppose $\psi(\alpha, c) > 0$.

(i) If for all $j \in J_-$ with $s - m > 0$ such that
\[
\lim_{n \to \infty} \inf \kappa_{nj} > (s - m) \frac{c(1 - \alpha - 2c)}{1 - \alpha},
\]
then the $C_p$ selection rule is strongly consistent.

(ii) If there exists $j \in J_-$ with $s - m > 0$ such that
\[
\lim_{n \to \infty} \sup \kappa_{nj} < (s - m) \frac{c(1 - \alpha - 2c)}{1 - \alpha},
\]
then the $C_p$ selection rule under-specified 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 $j \in J_-$ with $m < s$ since the candidate models $j \in J_-$ 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 $\Phi_j$ is $s$. Thus, $\log(\tau_{nj})$ and $\kappa_{nj}$ are $\sum_{i=1}^{s} \log(1 + \lambda_{ij}^{\Phi_j})$ and $\sum_{i=1}^{s} \lambda_{ij}^{\Phi_j}$, respectively. Here, $\lambda_{ij}^{\Phi_j}$ are the non-zero eigenvalues of $\Phi_j$. If the elements of $X_{j \phi}$ are of $O(1)$, which is a common and easily verified assumption in MLR, then $\|\frac{1}{n} X_{j \phi} Q_j X_j \|$ is bounded. Intuitively, if the elements of $\Theta_j$ are big or the elements of the covariance matrix are small, then the eigenvalues of $\Phi_j$ 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 $\Theta_j$ are small or the elements of the covariance matrix are big, then for $j \notin J$, the eigenvalues of $\Phi_j$ 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 $\Theta$ and $\Sigma$ 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 $\alpha$ 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 et al. (2014); Yanagihara et al. (2013) 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.
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.

![3D and contour plots](image)

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 \( \psi(\alpha, c) > 0 \).

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 \( C_p \) 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 \( C_p \)
Asymptotics of AIC, BIC 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.

![Figure 2. The regions of $R_1$ and $R_2$.](image)

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 $\alpha$ 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 $\Sigma^{-1/2} \Theta$, and does not depend on the choice of nonsingular $p \times p$ matrix $D$ in $YD$. Therefore, in conducting our simulation studies, it suffices to set $\Sigma = I_p$ and vary $\Theta$. We set $d_k = \{1, 2, 3, 4, 5\}$, $\alpha = 0.1$, $c = \{0.2, 0.5, 0.8\}$ and $n = 100, 150$ and $200$, and the values of $k$ and $p$ will follow. We choose $J$ to be the set of all non-empty subsets of $\omega$.

Set $X = UT^{1/2}$, where entries of $U$ are independent and uniformly distributed $U(1, 5)$, and $T = [t_{ij}]_{k \times k}$ 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 $1_5$ is a 5-vector of ones, $t$ and $\theta_s$ to be chosen below. $\Theta_{j_k} = 1_5 \theta_s$ and $\Theta = (\Theta_{j_k})'$. We consider three settings of $t$: (1) $t = 0$; (2) $t = 0.2$ and (3) $t = 0.4$; two settings of $\theta$: (I): $\theta_s = ((-0.5)^0, \ldots, (-0.5)^{p-1})$ and (II): $\theta_s = \sqrt{n}((-0.5)^0, \ldots, (-0.5)^{p-1})$; and three cases for the distribution of $E$: (i) standard normal; (ii) standardized $t$ with five degrees of freedom, i.e., $e_{ij} \sim t_5 / \sqrt{\text{Var}(t_5)}$; and (iii) standardized chi-square distribution with two degrees of freedom, i.e., $e_{ij} \sim (\chi_2^2 - 2)/\sqrt{\text{Var}(\chi_2^2)}$. 
We first explain our choices of $\alpha_k$, $c_n$, 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 $\|\mathbf{\Phi}_1\|$ is bounded whereas Setting (II) ensures $\|\mathbf{\Phi}_1\| \to \infty$. 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 $c_n = 0.5$, $\log(\tau_n(i)) - 4cn log(n) > 4log(1 - c_n)$; and for $c_n = 0.8$, $\log(\tau_n(i)) - 4cn log(n) < 4log(1 - c_n)$. 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 $\alpha$ 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.

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.
Asymptotics of AIC, BIC and $C_p$

| $c_n = 0.2$ | $c_n = 0.5$ | $c_n = 0.8$ |
| $\phi$ | $\psi$ | $\phi$ | $\psi$ | $\phi$ | $\psi$ |
| $\alpha_k = 0.1$ | 0.16 | 0.14 | 0.25 | -0.13 | -0.26 | -5.6 |

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

<table>
<thead>
<tr>
<th>$\alpha = 0.1, c = 0.2, \phi = 0.16$ and $\psi = 0.14$</th>
<th>$\alpha = 0.1, c = 0.5, \phi = 0.25$ and $\psi = -0.13$</th>
<th>$\alpha = 0.1, c = 0.8, \phi = -0.26$ and $\psi = -5.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(k, p, n) = (10, 50, 100)$</td>
<td>$(k, p, n) = (15, 75, 150)$</td>
<td>$(k, p, n) = (20, 100, 200)$</td>
</tr>
<tr>
<td>AIC</td>
<td>BIC</td>
<td>$C_p$</td>
</tr>
<tr>
<td>Under</td>
<td>Over</td>
<td>(Average)</td>
</tr>
<tr>
<td>0</td>
<td>170</td>
<td>0</td>
</tr>
<tr>
<td>130</td>
<td>0</td>
<td>353</td>
</tr>
<tr>
<td>(1.05)</td>
<td>(1.19)</td>
<td>(1.03)</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>673</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>327</td>
<td>0</td>
<td>992</td>
</tr>
<tr>
<td>(1.28)</td>
<td>(3.14)</td>
<td>(1.19)</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 2. Selection times of AIC, BIC and $C_p$ methods under Settings (1) and (1) 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. 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.

5.1. Preliminaries

By (1.3) and (1.7), to prove the strong consistency of the AIC selection rule is equivalent to prove that

$$
P(\hat{J}_A \neq J_*, i.o.) = P(\bigcup_{J \neq J_*} \{A_j < A_{J_*}, i.o.\}) = 0,
$$
\[
\alpha = 0.1, \ c = 0.2, \ \phi = 0.16 \text{ and } \psi = 0.14
\]

\[
(k, p, n) = (10, 20, 100) \quad (k, p, n) = (15, 30, 150) \quad (k, p, n) = (20, 40, 200)
\]

<table>
<thead>
<tr>
<th>(k)</th>
<th>(p)</th>
<th>(n)</th>
<th>(k)</th>
<th>(p)</th>
<th>(n)</th>
<th>(k)</th>
<th>(p)</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>20</td>
<td>100</td>
<td>15</td>
<td>30</td>
<td>150</td>
<td>20</td>
<td>40</td>
<td>200</td>
</tr>
</tbody>
</table>

which implies

\[
1 \over n \log \left( \frac{|n \Sigma_j|}{|n \Sigma_{j-1}|} \right) + 2cn, \quad (5.1)
\]

which implies

\[
1 \over n \log \left( \frac{|n \Sigma_j|}{|n \Sigma_{j-1}|} \right) + 2cn, \quad (5.2)
\]

Table 3. Selection times of AIC, BIC and \(C_p\) methods under Settings (I) 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.

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

Step 1: We consider the over-specified case, i.e., \(j \in J_*\). There exist \(m\) and \(j_1 < j_2 < \cdots < j_m\) such that \(J = J_* \cup \{j_1, \ldots, j_m\}\). 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 \(J_*\). Specifically, \(J = J_m \supset J_{m-1} \supset \cdots \supset J_0 = J_*\) where \(J_{t+1} = J_t \cup \{j_{t+1}\}\) 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

\[
1 \over n (A_j - A_{j_*}) = 1 \over n \sum_{t=1}^{m} (A_{j_t} - A_{j_{t-1}}).
\]

Based on the definition of \(A_j\) in (1.3), it follows that

\[
1 \over n (A_{j_t} - A_{j_{t-1}}) = \log \left( \frac{|n \Sigma_j|}{|n \Sigma_{j_{t-1}}|} \right) + 2cn, \quad (5.1)
\]

which implies

\[
1 \over n (A_j - A_{j_*}) = \sum_{t=1}^{m} \left[ \log \left( \frac{|n \Sigma_j|}{|n \Sigma_{j_{t-1}}|} \right) + 2cn \right]. \quad (5.2)
\]
Asymptotics of AIC, BIC and $C_p$

<table>
<thead>
<tr>
<th>$\alpha = 0.1, c = 0.2, \phi = 0.16$ and $\psi = 0.14$</th>
<th>$\alpha = 0.1, c = 0.25, \phi = 0.13$</th>
<th>$\alpha = 0.1, c = 0.8, \phi = -0.26$ and $\psi = -5.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(k, p, n) = (10, 20, 100)$</td>
<td>$(k, p, n) = (15, 30, 150)$</td>
<td>$(k, p, n) = (20, 40, 200)$</td>
</tr>
<tr>
<td>AIC BIC $C_p$ AIC BIC $C_p$ AIC BIC $C_p$</td>
<td>AIC BIC $C_p$ AIC BIC $C_p$ AIC BIC $C_p$</td>
<td>AIC BIC $C_p$ AIC BIC $C_p$ AIC BIC $C_p$</td>
</tr>
<tr>
<td>Under</td>
<td>0 128 0</td>
<td>0 930 0</td>
</tr>
<tr>
<td>True</td>
<td>852 872 656</td>
<td>916 70 632</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>148 (1.06) 0 344 (1.22) 84 (1.04) 0 368 (1.20) 41 (1.02) 0 361 (1.32)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Selection times of AIC, BIC and $C_p$ methods under Settings (2) and (1) 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.

Analogously, we also have

$$\frac{1}{n} \left( B_j - B_{j_s} \right) = \sum_{t=1}^{m} \left[ \log \left( \frac{|n \tilde{\Sigma}_{j_t}|}{|n \tilde{\Sigma}_{j_{t-1}}|} \right) + \log(n) c_n \right]$$

and

$$\frac{1}{n} \left( C_j - C_{j_s} \right) = (1 - \alpha_k) \text{tr} \left( \tilde{\Sigma}_t^{-1} \left( \tilde{\Sigma}_j - \tilde{\Sigma}_{j_s} \right) \right) + 2 mc_n$$

$$= \sum_{t=1}^{m} \left( (1 - \alpha_k) \text{tr} \left( \tilde{\Sigma}_t^{-1} \left( \tilde{\Sigma}_j - \tilde{\Sigma}_{j_{t-1}} \right) \right) + 2 c_n \right).$$

Then, the lemma below follows.

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

$$\mathbb{P} \left( \frac{1}{n} (A_j - A_{j_s}) - \sum_{t=1}^{m} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) - 2mc_n | \geq m \epsilon \right) = o(n^{-t}),$$

(5.5)
\( \alpha = 0.1, c = 0.2, \phi = 0.16 \) and \( \psi = 0.14 \)

\((k, p, n) = (10, 20, 100)\) \((k, p, n) = (15, 30, 150)\) \((k, p, n) = (20, 40, 200)\)

<table>
<thead>
<tr>
<th>( (k, p, n) )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>853</td>
<td>1000</td>
<td>679</td>
<td>911</td>
<td>1000</td>
<td>609</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>147</td>
<td>0</td>
<td>321</td>
<td>89</td>
<td>0</td>
<td>391</td>
</tr>
</tbody>
</table>

\( \alpha = 0.1, c = 0.5, \phi = 0.25 \) and \( \psi = -0.13 \)

<table>
<thead>
<tr>
<th>( (k, p, n) )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>677</td>
<td>1000</td>
<td>9</td>
<td>802</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>323</td>
<td>0</td>
<td>991</td>
<td>198</td>
<td>0</td>
<td>1000</td>
</tr>
</tbody>
</table>

\( \alpha = 0.1, c = 0.8, \phi = -0.26 \) and \( \psi = -5.6 \)

<table>
<thead>
<tr>
<th>( (k, p, n) )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
<th>( \text{AIC} )</th>
<th>( \text{BIC} )</th>
<th>( C_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>1000</td>
<td>0</td>
<td>981</td>
<td>1000</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>489</td>
<td>0</td>
<td>19</td>
<td>489</td>
<td>0</td>
<td>19</td>
</tr>
</tbody>
</table>

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.

\[
P\left( \frac{1}{n} \left( B_j - B_{j^*} \right) - \sum_{t=1}^{m} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) - m c_n \log(n) \right) \geq m \varepsilon \right) = o(n^{-t}), \tag{5.6}\]

\[
P\left( \frac{1}{n} (C_j - C_{j^*}) - m \psi(\alpha_k, c_n) \right) \geq m \varepsilon = o(n^{-t}), \tag{5.7}\]

and

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

**Remark 5.2.** A consequence of this lemma is that \( \frac{1}{nm}(A_j - A_{j^*}) - \phi(\alpha_m, c_n) \rightarrow 0 \) and \( \frac{1}{nm}(B_j - B_{j^*}) - (\log(n) - 2)c_n - \phi(\alpha_m, c_n) \rightarrow 0 \) 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 \( j \in J_{+} \) satisfying \( |j| - k_e = m \), if \( \phi(\alpha_m, c_n) > 0 \), then for sufficiently large \( p \) and \( n \), then a.s. \( j \) 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. \( j \) 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., \( j \in J_{-} \). Let \( j_s \setminus j = \{i_1, \ldots, i_s\} \) and \( j \setminus j_s = \{j_1, \ldots, j_{m} \} \). We first assume \( m \) is positive. Define the model index set \( j_t = j \cup \{i_{t+1}, \ldots, i_s\} \) for
Asymptotics of AIC, BIC and $C_p$

<table>
<thead>
<tr>
<th>$\alpha = 0.1, \ c = 0.2, \ \phi = 0.16$ and $\psi = 0.14$</th>
<th>(k, p, n) = (10, 20, 100)</th>
<th>(k, p, n) = (15, 30, 150)</th>
<th>(k, p, n) = (20, 40, 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>BIC</td>
<td>$C_p$</td>
<td>AIC</td>
</tr>
<tr>
<td>Under</td>
<td>0</td>
<td>983</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>873</td>
<td>17</td>
<td>678</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>127</td>
<td>0</td>
<td>322</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha = 0.1, \ c = 0.5, \ \phi = 0.25$ and $\psi = -0.13$</th>
<th>(k, p, n) = (10, 50, 100)</th>
<th>(k, p, n) = (15, 75, 150)</th>
<th>(k, p, n) = (20, 100, 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>BIC</td>
<td>$C_p$</td>
<td>AIC</td>
</tr>
<tr>
<td>Under</td>
<td>40</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>665</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>295</td>
<td>0</td>
<td>986</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha = 0.1, \ c = 0.8, \ \phi = -0.26$ and $\psi = -5.6$</th>
<th>(k, p, n) = (10, 80, 100)</th>
<th>(k, p, n) = (15, 120, 150)</th>
<th>(k, p, n) = (20, 160, 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>BIC</td>
<td>$C_p$</td>
<td>AIC</td>
</tr>
<tr>
<td>Under</td>
<td>0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>True</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Over (Average)</td>
<td>1000</td>
<td>0</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 6. Selection times of AIC, BIC and $C_p$ methods under Settings (3) and (1) 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.

$t = 0, 1, \cdots, s$ (with the convention that $j_s = j$), which also indicates that $i_t$ is in $j_{t-1}$ but not in $j_t$. Note also that $j_0 = j \cup j_s$ is an over-specified model. So,

$$
\frac{1}{n}(A_j - A_{j_s}) = \frac{1}{n}(A_{j_s} - A_{j_0}) + \frac{1}{n}(A_{j, j_s - A_j},
$$

$$
\frac{1}{n}(B_j - B_{j_s}) = \frac{1}{n}(B_{j_s} - B_{j_0}) + \frac{1}{n}(B_{j, j_s - B_j},
$$

$$
\frac{1}{n}(C_j - C_{j_s}) = \frac{1}{n}(C_{j_s} - C_{j_0}) + \frac{1}{n}(C_{j, j_s - C_j}.
$$

If $m = 0$, $j \cup j_s = j_s$, 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

$$
\frac{1}{n}(A_{j_s} - A_{j_0}) = \sum_{t=0}^{s-1} \log \left( \frac{|n\sum_{j_{s-t}}|}{|n\sum_{j_{s-t-1}}|} \right) - 2sc_n
$$

(5.9)

$$
\frac{1}{n}(B_{j_s} - B_{j_0}) = \sum_{t=0}^{s-1} \log \left( \frac{|n\sum_{j_{s-t}}|}{|n\sum_{j_{s-t-1}}|} \right) - \log(n)sc_n
\[ \alpha = 0.1, \ c = 0.2, \ \phi = 0.16 \text{ and } \psi = 0.14 \]

<table>
<thead>
<tr>
<th>( (k, p, n) = (10, 20, 100) )</th>
<th>( (k, p, n) = (15, 30, 150) )</th>
<th>( (k, p, n) = (20, 40, 200) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.1, \ c = 0.2, \ \phi = 0.25 \text{ and } \psi = -0.13 )</td>
<td>( \alpha = 0.1, \ c = 0.8, \ \phi = -0.26 \text{ and } \psi = -5.6 )</td>
<td></td>
</tr>
<tr>
<td>( \alpha = 0.1, \ c = 0.8, \ \phi = -0.26 \text{ and } \psi = -5.6 )</td>
<td>( \alpha = 0.1, \ c = 0.8, \ \phi = -0.26 \text{ and } \psi = -5.6 )</td>
<td></td>
</tr>
</tbody>
</table>

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.

and

\[ \frac{1}{n} (C_{\hat{j}} - C_{j_0}) = \sum_{t=0}^{s-1} (1 - \alpha) \mathfrak{tr} \left[ \Sigma_{j_{k-t}}^{-1} \left( \hat{\Sigma}_{j_{k-t}} - \hat{\Sigma}_{j_{k-t-1}} \right) \right] - 2sc_n. \quad (5.10) \]

Then, the lemma below follows.

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

\[ \mathbb{P} \left( \left| \frac{1}{n} (A_{j_{k}} - A_{j_0}) - \log(\tau_{nj}) + s \log(1 - \alpha_m - c_n) + 2sc_n \right| \geq \varepsilon \right) = o(n^{-t}), \quad (5.11) \]

\[ \mathbb{P} \left( \left| \frac{1}{n} (B_{j_{k}} - B_{j_0}) - \log(\tau_{nj}) + s \log(1 - \alpha_m - c_n) + \log(n)sc_n \right| \geq \varepsilon \right) = o(n^{-t}), \quad (5.12) \]

and

\[ \mathbb{P} \left( \left| \frac{1}{n} (C_{j_{k}} - C_{j_0}) - \frac{(1 - \alpha_k)(\kappa_{nj} + sc_n)}{1 - c_n - \alpha_k} \right| \geq \varepsilon \right) = o(n^{-t}). \quad (5.13) \]

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 \( j \) with \( |j_\ast \cap j^-| = s > 0 \) and \( |j_+| = m \geq 0 \), we have for any \( \varepsilon > 0 \) and \( t > 0 \),

\[
P\left( \frac{1}{n} (A_j - A_{j_\ast}) - \sum_{t=1}^{m} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) \right.
- \log(\tau_{nj}) + s \log(1 - \alpha_m - c_n) - 2(m - s) c_n \geq (m + s) \varepsilon \) = o(n^{-t}),
\]

\[
P\left( \frac{1}{n} (B_j - B_{j_\ast}) - \sum_{t=1}^{m} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) \right.
- \log(\tau_{nj}) + s \log(1 - \alpha_m - c_n) - (m - s) c_n \log(n) \geq (m + s) \varepsilon \) = o(n^{-t}),
\]

and

\[
P\left( \frac{1}{n} (C_j - C_{j_\ast}) - m \psi(\alpha_k, c_n) - \frac{(1 - \alpha_k)(\kappa_{nj} + sc_n)}{1 - c - \alpha_k} + 2sc_n \geq (m + s) \varepsilon \right) = o(n^{-t}).
\]

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 \( |J| = O(n^{\ell}) \), for the strong consistency of AIC rule, it is sufficient to prove for all \( j \neq j_\ast \), \( A_j > A_{j_\ast} \) holds with tail probability \( o(n^{-\ell-2}) \). If there exists \( j \in J^+ \) such that for all \( j' \in \{j_\ast\} \cup J^- \), \( A_j < A_{j'} \) holds with tail probability \( o(n^{-\ell-2}) \), then the AIC selection rule almost surely over-specifies the true model. Analogously, if there exists \( j \in J^- \) such that for all \( j' \in \{j_\ast\} \cup J^+ \), \( A_j < A_{j'} \) 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 \( A_j \)'s.

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

\[
\frac{1}{nm} (A_{j_\ast} - A_{j_\ast}) \stackrel{a.s.}{\rightarrow} -\phi(\alpha, c) > 0.
\]

If \( j \in J^- \) with \( |j_+| = m \geq 0 \) and \( |j_\ast \cap j^-| = s > 0 \), by Lemmas 5.1 and 5.4,

\[
\frac{1}{n} (A_j - A_{j_\ast}) = \frac{1}{n} (A_j - A_{j_\ast}) + \frac{1}{n} (A_{j_\ast} - A_{j_\ast}) = \sum_{t=1}^{m} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right) - \sum_{t=1}^{k-k_*} \log \left( \frac{1 - \alpha_t - c_n}{1 - \alpha_t} \right)
\]
\[ \begin{align*}
+ \log(\tau_{nj}) - s \log(1 - \alpha_m - c_n) + 2(m - s - k + k_s)c_n + o_{\text{a.s.}}(k - k_s - m) \\
= \log\left(\frac{\tau_{nj}}{(1 - \alpha_m)^s}\right) - \sum_{t=m-s+1}^{k-k_s} \log\left(\frac{1 - \alpha_t - c_n}{1 - \alpha_t}\right) - 2(k - k_s - m + s)c_n + o_{\text{a.s.}}(k - k_s - m).
\end{align*} \]

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

Next we consider (2). Because when \( j \in J_+ \) with \( |j - k_s| = m \), from the definition of \( \phi(\alpha, c) \) (see Figure 1 for illustration), we know that if \( \phi(\alpha, c) > 0 \), then for any \( \alpha_m \) 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 asymptotically over-specify the true model, i.e., for any \( j \in J_+ \) and for sufficiently large \( p \) and \( n \),

\[ A_j > A_\omega, \quad \text{a.s.} \]

uniformly. For the case of \( j \in J_+ \) with \( |j| = m \geq 0 \) and \( |j| = s > 0 \), note that \( s < k_s \), \( 0 < \Phi_j > 0 \) and \( \log(\tau_{nj}) > s \log(1 - \alpha_m) \) uniformly. If \( m \geq s \), by Lemma 5.4 and \( \phi(\alpha, c) > 0 \),

\[ \frac{1}{n}(A_j - A_{\omega}) = \sum_{t=1}^{m} \log\left(\frac{1 - \alpha_t - c_n}{1 - \alpha_t}\right) + \log(\tau_{nj}) - s \log(1 - \alpha_m - c_n) + 2(m - s)c_n + o_{\text{a.s.}}(m) \\
= \sum_{t=1}^{m-s} \log\left(\frac{1 - \alpha_t - c_n}{1 - \alpha_t}\right) + \log(\tau_{nj}/(1 - \alpha_m)^s) + 2(m - s)c_n + o_{\text{a.s.}}(m), \]

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 \( k_s \) is fixed, \( \alpha_m = m/n \rightarrow 0 \) and \( \lim \inf_{n \rightarrow \infty} \tau_{nj} > 1 \). By the fact

\[ \frac{1}{n}(A_j - A_{\omega}) = \log(\tau_{nj}) - (s - m)(\log(1 - c) + 2c) + o_{\text{a.s.}}(1), \]

and if \( \lim sup_{n \rightarrow \infty} \log(\tau_{nj}) < (s - m)(\log(1 - c) + 2c) \), then from Lemmas 5.1 and 5.4, we know that for sufficiently large \( p \) and \( n \),

\[ A_j < A_\omega, \quad \text{a.s.,} \]

which means that, in this case, the AIC rule asymptotically cannot select the true model \( j_\omega \). 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.

\[ \square \]

5.3. Proof of Theorem 3.5

By (5.6), it is easy to find that BIC rule cannot asymptotically over-specify the true model. In addition, for any \( j \in J_+ \) with \( |j| = m \geq 0 \) and \( |j| = s > 0 \), and \( m < s \), by Lemma 5.4, we obtain that

\[ \frac{1}{n}(B_j - B_{\omega}) = \log(\tau_{nj}) - (s - m)(\log(1 - c) + \log(n)c) + o_{\text{a.s.}}(1), \quad (5.14) \]
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 \( \phi \) by \( \psi \), 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 \( Y \rightarrow Y' \Sigma^{-1/2} \) and \( \Theta \rightarrow \Theta \Sigma^{-1/2} \). Thus, without loss of generality, we assume that \( \Sigma = I_p \) in the sequel. For Lemma 5.1, as \( j_t \) is in \( J_e \) but not in \( J_{t-1} \), we denote \( a_t = Q_{j_{t-1}} x_{j_t}/\|Q_{j_{t-1}} x_{j_t}\| \) for \( t \geq 1 \). Then by Sylvester’s determinant theorem, we have that

\[
|n \hat{\Sigma}_{j_t}| = |Y'Q_{j_t} Y| = |Y'Q_{j_{t-1}} Y - Y'a_t a_t' Y| \tag{6.1}
\]

\[
= |Y'Q_{j_{t-1}} Y|(1 - a_t' Y(Y'Q_{j_{t-1}} Y)^{-1} Y'a_t)
\]

\[
= |n \hat{\Sigma}_{j_{t-1}}|(1 - a_t' Y(Y'Q_{j_{t-1}} Y)^{-1} Y'a_t).
\]

Thus, to prove Lemma 5.1, we need to obtain only the limits of \( a_t' Y(Y'Q_{j_{t-1}} Y)^{-1} Y'a_t \) or similar expressions with different \( j_t \). The proof strategy is that we first define a function

\[
h_n(z) := n^{-1} a_t' Y(n^{-1} Y'Q_{j_{t-1}} Y - z I)^{-1} Y'a_t : \mathbb{C}^+ \rightarrow \mathbb{C}^+,
\]

where \( \mathbb{C}^+ = \{ z \in \mathbb{C}^+ : \Re z > 0 \} \). Next, we prove that outside a null set independent of \( j_{t-1} \), for every \( z \in \mathbb{C}^+ \), \( h_n(z) \) has a limit \( h(z) \in \mathbb{C}^+ \). 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 \( z \in \mathbb{C}^+ \), \( h_n(z) \) \( \overset{a.s.}{\rightarrow} h(z) \). Finally, we let \( z \downarrow 0 + 0i \) and obtain almost surely \( h_n(0) \rightarrow h(0) \) 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 \( Y(Y'Q_{j_t} Y)^{-1} Y' \) is special, and to the best of our knowledge, no known conclusions in RMT can be applied directly to obtain the limit of \( a_t' Y(Y'Q_{j_{t-1}} Y)^{-1} Y'a_t \), 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 \( j \in J_+ \), \( Y'Q_{j} Y = E'Q_{j_t} E \), we have

\[
\log \left( \frac{|n \hat{\Sigma}_{j_t}|}{|n \hat{\Sigma}_{j_{t-1}}|} \right) = \log(1 - a_t' E(E'Q_{j_{t-1}} E)^{-1} E'a_t) \tag{6.2}
\]
and
\[ n\bar{\Sigma}_{jt} - n\bar{\Sigma}_{jt-1} = -E'a_t a'_t E. \quad (6.3) \]

It follows from (5.2) and (6.2) that
\[ \frac{1}{n} (A_j - A_j^*) = \sum_{t=1}^{m} \log(1 - a'_t E(E'Q_{jt-1} E)^{-1} E'a_t + 2c_n). \]

Since \( a_t \) is an eigenvector of \( Q_{jt-1} \), we have \( a'_t Q_{jt-1} a_t = 1 \), which together with Proposition 3.1 and (3.9) implies
\[ \frac{1}{p} a'_t E(\frac{1}{p} E'Q_{jt-1} E - zI_p)^{-1} E'a_t - 1 + \frac{1}{1 + \frac{c_n}{1 - \alpha_{t}}(z)} \rightarrow 0, \quad (6.4) \]

with tail probability \( o(n^{-t}) \) for any fixed \( t > 0 \). Therefore, by (3.10) and as \( n \to \infty \), we have
\[ \frac{1}{nm} (A_j - A_j^*) - \frac{1}{m} \sum_{t=1}^{m} \left( \log(1 - \frac{c_n}{1 - \alpha_{t}}) + 2c_n \right) \rightarrow 0. \quad (6.5) \]

with tail probability \( o(n^{-t}) \) for any fixed \( t > 0 \), which implies (5.5). From integration by parts we have that \( \frac{1}{n^2} (A_j - A_j^*) \) tends to
\[ \int_{0}^{\alpha_m} (\log(1 - \frac{c_n}{1 - t}) + 2c_n) dt = 2c_n \alpha_m + \log \left( \frac{(1 - c_n)(1 - \alpha_m)^{1-\alpha_m}}{(1 - c_n - \alpha_m)^{1-\alpha_m}} \right), \]

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
\[ \frac{1}{n} (C_j - C_j^*) = \sum_{t=1}^{m} \left( \frac{k}{n} - 1 \right) a'_t E(E'Q_{j\omega} E)^{-1} E'a_t + 2c_n. \quad (6.6) \]

By (3.8) and (3.10) and the fact that
\[ a'_t Q_{j\omega} a_t = 0, \]

we have
\[ a'_t E(E'Q_{j\omega} E)^{-1} E'a_t \rightarrow \frac{c_n}{1 - \alpha_k - c_n}, \]

with tail probability \( o(n^{-t}) \) for any fixed \( t > 0 \), which together with (6.6) implies
\[ \frac{1}{nm} (C_j - C_j^*) - \frac{c_n(\alpha_k - 1)}{1 - \alpha_k - c_n} - 2c_n \rightarrow 0, \]

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

$$n\hat{\Sigma}_{j_t} \sim W_p(n - |j_t|, I_p), \quad n\hat{\Sigma}_{j_{t-1}} \sim W_p(n - |j_{t-1}|, I_p)$$

and

$$\left|\frac{n\hat{\Sigma}_{j_t}}{n\hat{\Sigma}_{j_{t-1}}}\right| \sim \left(1 + \frac{\chi^2_p}{\chi^2_{n-|j_t|-p+1}}\right)^{-1}$$

where $W_p$ and $\chi^2_p$ 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

$$\frac{1}{n}\chi^2_p \rightarrow c_n \quad \text{and} \quad \frac{1}{n}\chi^2_{n-|j_t|-p+1} \rightarrow 1 - \alpha + c_n,$$

with tail probability $o(n^{-1})$ for any $t > 0$, which together with (5.1) imply (5.5) directly.

6.2. Proof of Lemma 5.3

We start with $\frac{1}{n}(A_{j_t} - A_{j_0})$ and $\frac{1}{n}(B_{j_t} - B_{j_0})$. As $i_t$ is in $j_{t-1}$ but not in $j_t$, we denote $a_t = Q_{j_t}x_{i_t}/\|Q_{j_t}x_{i_t}\|$. By the fact that $Q_{j_{t-1}} = Q_{i_t} - a_t a'_t$ and equation (6.1), we have that

$$\log \left(\frac{n\hat{\Sigma}_{j_{t-1}}}{n\hat{\Sigma}_{j_{t-1}}}\right) = -\log(1 - a'_{s-t}Y(Q'_{j_{s-t}}Y)^{-1}Y'a_{s-t}).$$

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

$$m_{nt} := m_{nt}(z) = -\log(1 - \frac{1}{p}a'_{{s-t}}Y(Q'_{j_{s-t}}Y/p - zI_p)^{-1}Y'a_{s-t}),$$

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

$$m_{nt} = -\log \left(-za'_{s-t}Q_{j_{s-t}}YY'Q_{j_{s-t}}/p - zI_n\right)^{-1}a_{s-t}. \quad (6.7)$$

Substitute model (2.1) into the above equation and denote

$$I_t := I_t(z) = za'_{s-t}Q_{j_{s-t}}YY'Q_{j_{s-t}}/p - zI_n)^{-1}a_{s-t}$$

$$= za'_{s-t}Q_{j_{s-t}}(X_{i_t} \Theta \ell_{i_t} + E)(\Theta'_{i_t} X'_{i_t} + E')Q_{j_{s-t}}/p - zI_n)^{-1}a_{s-t},$$

where $\ell_t = \{i_1, \cdots, i_{s-t}\}$. Define $B_1 = Q_{j_{s-t}}X_{i_t} (X'_{i_t} Q_{j_{s-t}} X_{i_t})^{-1/2}$ and select $B_2$ such that $B = (B_1 \oplus B_2)$ is an $n \times n$ orthogonal matrix. Then, we have

$$I_t = za'_{s-t}B (B'Q_{j_{s-t}}X_{i_t} \Theta \ell_{i_t} + E)(\Theta'_{i_t} X'_{i_t} + E')Q_{j_{s-t}}B/p - zI_n)^{-1}B'a_{s-t}. $$
With $a'_{s-t}B_2 = 0$, we obtain

$$I_t = z\tilde{a}'_{s-t}\left( B_1'Q_{j_{s-t}}(X_{\ell t}\Theta_{\ell t} + E)(\Theta'_{\ell t}X'_{\ell t} + E')Q_{j_{s-t}}B_1/p \right. \nonumber$$

$$- B_1'Q_{j_{s-t}}(X_{\ell t}\Theta_{\ell t} + E)E'Q_{j_{s-t}}B_2(2B_2'Q_{j_{s-t}} EE'Q_{j_{s-t}}B_2/p - zI_{n-s+t})^{-1} \nonumber$$

$$\cdot B_2'Q_{j_{s-t}}E(E' + X_{\ell t}\Theta'_{\ell t})Q_{j_{s-t}}B_1/p^2 - zI_{s-t})^{-1}\tilde{a}_{s-t} \nonumber$$

where $\tilde{a}_{s-t} = B_1' a_{s-t}$. By applying in-out-exchange formula (3.4) to the term

$$E'Q_{j_{s-t}}B_2(2B_2'Q_{j_{s-t}} EE'Q_{j_{s-t}}B_2/p - zI_{n-s+t})^{-1}B_2'Q_{j_{s-t}} E/p,$$

we obtain

$$I_t = -\tilde{a}'_{s-t}\left( \frac{1}{p}B_1'Q_{j_{s-t}}(X_{\ell t}\Theta_{\ell t} + E)(\frac{1}{p}E'Q_{j_{s-t}}B_2B_2'Q_{j_{s-t}}E - zI_p)^{-1} \right. \nonumber$$

$$\times (\Theta'_{\ell t}X'_{\ell t} + E')Q_{j_{s-t}}B_1 + I_{s-t})^{-1}\tilde{a}_{s-t}, \nonumber$$

which together with notation $M_{j_{s-t}} := \frac{1}{p}E'Q_{j_{s-t}}E - zI_p$ implies

$$I_t = -\tilde{a}'_{s-t}\left( \frac{1}{p}B_1'Q_{j_{s-t}}(X_{\ell t}\Theta_{\ell t} + E)(M_{j_{s-t}} - \frac{1}{p}E'Q_{j_{s-t}}B_1B_1'Q_{j_{s-t}}E)^{-1} \right. \nonumber$$

$$\times (\Theta'_{\ell t}X'_{\ell t} + E')Q_{j_{s-t}}B_1 + I_{s-t})^{-1}\tilde{a}_{s-t}. \nonumber$$

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

$$I_t = -\tilde{a}'_{s-t}\left( I_{1t} + I_{2t} + I_{3t} + I_{4t} \right)^{-1}\tilde{a}_{s-t} \quad (6.8)$$

where

$$I_{1t} = \frac{1}{p}B_1'Q_{j_{s-t}}X_{\ell t}\Theta_{\ell t}M^{-1}_{j_{s-t}}\Theta'_{\ell t}X'_{\ell t}Q_{j_{s-t}}B_1 + \frac{1}{p_2^2}B_1'Q_{j_{s-t}}X_{\ell t}\Theta_{\ell t}M^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1 \nonumber$$

$$\left( I_{s-t} - \frac{1}{p}B_1'Q_{j_{s-t}}EM^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1 \right)^{-1}B_1'Q_{j_{s-t}}EM^{-1}_{j_{s-t}}\Theta'_{\ell t}X'_{\ell t}Q_{j_{s-t}}B_1; \nonumber$$

$$I_{2t} = \frac{1}{p}B_1'Q_{j_{s-t}}X_{\ell t}\Theta_{\ell t}M^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1(I_{s-t} - \frac{1}{p}B_1'Q_{j_{s-t}}EM^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1)^{-1}; \nonumber$$

$$I_{3t} = (I_{s-t} - \frac{1}{p}B_1'Q_{j_{s-t}}EM^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1)^{-1}. \nonumber$$

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

$$\frac{1}{p}B_1'Q_{j_{s-t}}EM^{-1}_{j_{s-t}}E'Q_{j_{s-t}}B_1 - (1 - \frac{1}{1 + s_{j_{s-t}}(z)})I_{s-t} \rightarrow 0, \nonumber$$
Asymptotics of AIC, BIC and $C_p$

which implies

$$I_{3t} - \left(1 + \frac{s_{n.j_s-t}(z)}{j_s-t}\right)I_{s-t} \to 0. \quad (6.9)$$

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

$$\frac{1}{p}B_1'Q_{j_s-t}X_{X_{\ell_t}} \Theta_{\ell_t} M^{-1}_{j_s-t} E'Q_{j_s-t} B_1 \to 0,$$

and

$$\frac{1}{p}B_1'X_{\ell_t} \Theta_{\ell_t} M^{-1}_{j_s-t} \Theta_{\ell_t} X'_{\ell_t} B_1 + \frac{p^{-1}B_1'X_{\ell_t} \Theta_{\ell_t} \Theta_{\ell_t} X'_{\ell_t} B_1}{z(1 + s_{n.j_s-t}(z) - \frac{1-c_n-\alpha_m}{c_n z})} \to 0,$$

which together with (6.9) imply

$$I_t + \tilde{a}_{s-t} - \left(1 + \frac{s_{n.j_s-t}(z)}{j_s-t}\right)I_{s-t} - \frac{p^{-1}B_1'X_{\ell_t} \Theta_{\ell_t} \Theta_{\ell_t} X'_{\ell_t} B_1}{z(1 + s_{n.j_s-t}(z) - \frac{1-c_n-\alpha_m}{c_n z})} \tilde{a}_{s-t} \to 0.$$

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

$$\delta_t := \tilde{a}_{s-t} - \left(1 - \alpha_m\right)I_{s-t} + n^{-1}B_1'X_{\ell_t} \Theta_{\ell_t} \Theta_{\ell_t} X'_{\ell_t} B_1 \to 0,$$

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 $\alpha_{m-t}$ and $\alpha_m$ have the same limit. By basic calculation, we obtain the equation

$$\prod_{t=0}^{s-1} \delta_t = \gamma_{n,j_s},$$

which together with (5.9) implies (5.11).

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

$$\frac{1}{n}(C_j - C_{j_s}) = \sum_{t=0}^{s-1} (1 - \alpha_k) \text{tr}[\Sigma_{\omega}^{-1}(\tilde{\Sigma}_{j_s-t} - \tilde{\Sigma}_{j_s-t-1})] - 2scn$$

$$= \sum_{t=0}^{s-1} (1 - \alpha_k) a_{s-t} Y(E'Q_{\omega} E)^{-1} Y'a_{s-t} - 2scn.$$

Let

$$J_t(z) := J_t(z) = p^{-1}a_{s-t} Y(E'Q_{\omega} E/p - zI_p)^{-1} Y'a_{s-t}.$$
Then, by substituting model (2.1) into the above equation, we obtain
\[
J_t = \frac{1}{p} a'_{s-t} (X_{\ell_t} \Theta_{\ell_t} + E) (E' Q_{\omega} E/p - z I_p)^{-1} (\Theta'_{\ell_t} X'_{\ell_t} + E') a_{s-t}
\]
\[
= \frac{1}{p} a'_{s-t} X_{\ell_t} \Theta_{\ell_t} (E' Q_{\omega} E/p - z I_p)^{-1} \Theta'_{\ell_t} X'_{\ell_t} a_{s-t}
\]
\[
+ \frac{1}{p} a'_{s-t} X_{\ell_t} \Theta_{\ell_t} (E' Q_{\omega} E/p - z I_p)^{-1} E' a_{s-t}
\]
\[
+ \frac{1}{p} a'_{s-t} E (E' Q_{\omega} E/p - z I_p)^{-1} \Theta'_{\ell_t} X'_{\ell_t} a_{s-t}
\]
\[
+ \frac{1}{p} a'_{s-t} E (E' Q_{\omega} E/p - z I_p)^{-1} E' a_{s-t}
\]
\[
:= J_{1t} + J_{2t} + J_{3t} + J_{4t}.
\]
It follows from Proposition 3.1 that with tail probability \(o(n^{-1})\) for any fixed \(t > 0\),
\[
\frac{1}{p} a'_{s-t} X_{\ell_t} \Theta_{\ell_t} \Theta'_{\ell_t} X'_{\ell_t} a_{s-t} \sim \frac{1}{z(1 + s_n \omega(z) - \frac{1-c_n}{c_n^2})} \to 0,
\]
\[
J_{2t} \to 0 \quad \text{and} \quad J_{3t} \to 0 \quad \text{as} \quad \frac{1}{z(1 + s_n \omega(z) - \frac{1-c_n}{c_n^2})} \to 0.
\]
It follows from
\[
\frac{1}{n} \sum_{t=0}^{s-1} a'_{s-t} X_{\ell_t} \Theta_{\ell_t} \Theta'_{\ell_t} X'_{\ell_t} a_{s-t} = \frac{1}{n} \text{tr}[\Theta'_{\ell_t} X'_{\ell_t} \sum_{t=0}^{s-1} a_{s-t} a'_{s-t}] X_{\ell_t} \Theta_{\ell_t} = \kappa_n
\]
and
\[
\lim_{z \to 0^+} \frac{1}{z(1 + s_n \omega(z) - \frac{1-c_n}{c_n^2})} \frac{c_n}{1 - \alpha_k - c_n},
\]
we obtain (5.13). Thus, we complete the proof of Lemma 5.3. \(\square\)

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 \(\Sigma_1\) may be singular. Singularity of \(\Sigma_1\) can be circumvented by using a ridge-type estimator of the covariance matrix (e.g., Yamamura et al. (2010); Chen et al. (2011)). It is of interest to study the asymptotic properties of these model selection criteria when the true model size \(k_s\) is also large, i.e., \(k_s/n\) 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.

Acknowledgement

We thank the AE and the reviewers for careful reading of the manuscript and their helpful comments that improve this paper. 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 and 11971097. 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

This supplementary material gives the proof of Proposition 3.1.

References


