High-Dimensional Properties of Information Criteria and Their Efficient Criteria for Multivariate Linear Regression Models with Covariance Structures

Tetsuro Sakurai[*] and Yasunori Fujikoshi[**]

[*] *Center of General Education, Tokyo University of Science, Suwa*
5000-1 Toyohira, Chino, Nagano 391-0292, Japan

[**] *Department of Mathematics, Graduate School of Science*
Hiroshima University
1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

Abstract

In this paper, first we consider high-dimensional consistency properties of information criteria $\mathrm{IC}_{g,d}$ and their efficient criteria $\mathrm{EC}_{g,d}$ for selection of variables in multivariate regression model with covariance structures. The covariance structures considered are (1) independent covariance structure, (2) uniform covariance structure and (3) autoregressive covariance structure. Sufficient conditions for $\mathrm{IC}_{g,d}$ and $\mathrm{EC}_{g,d}$ to be consistent are derived under a high-dimensional asymptotic framework such that the sample size n and the number p of response variables are large as in the way $p/n \to c \in (0, \infty)$. Our results are checked numerically by conducting a Mote Carlo simulation. Next we discuss with high-dimensional properties of AIC and BIC for selecting (4) independence covariance structure with different variances, and (5) no covariance structure, in addition to the covariance structures (1) $\sim$ (3). Some tendancy is pointed through a numerical experiment.

AMS 2000 subject classification: primary 62H12; secondary 62H30

Key Words and Phrases: Consistency property, Information criteria, Covariance structures, Efficient criteria, High-dimensional asymptotic framework, Multivariate linear regression.

1. Introduction

We consider a multivariate linear regression of p response variables $y_1, \ldots, y_p$ on a subset of k explanatory variables $x_1, \ldots, x_k$. Suppose that there are n observations on $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_k)'$, and let $\mathsf{Y} : n \times p$ and $\mathbf{X} : n \times k$ be the observation matrices of $\boldsymbol{y}$ and $\boldsymbol{x}$ with the sample size n, respectively. The multivariate linear regression model including all the explanatory variables is written as

$$\mathsf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n), \tag{1.1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ unknown matrix of regression coefficients and $\boldsymbol{\Sigma}$ is a $p \times p$ unknown covariance matrix. The notation $\mathrm{N}_{n\times p}(\cdot, \cdot)$ means the matrix normal distribution such that the mean of Y is $\mathbf{X}\boldsymbol{\Theta}$ and the covariance matrix of $\mathrm{vec}\,(\mathsf{Y})$ is $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$, or equivalently, the rows of Y are independently normal with the same covariance matrix $\boldsymbol{\Sigma}$. Here, $\mathrm{vec}(\mathsf{Y})$ be the $np \times 1$ column vector obtained by stacking the columns of Y on top of one another. We assume that $\mathrm{rank}(\mathbf{X}) = k$.

We consider the problem of selecting the best model from a collection of candidate models specified by a linear regression of $\boldsymbol{y}$ on subvectors of $\boldsymbol{x}$. Our interest is to examine consistency properties of information criteria and their efficient criteria when $p/n \to c \in (0, \infty)$. When $\boldsymbol{\Sigma}$ is unknown positive definite, it has been pointed (see, e.g., Yanagihara et al. (2015), Fujikoshi et al. (2014), etc.) that AIC and C_p have consistency properties when $p/n \to c \in (0, 1)$, under some conditions, but BIC is not necessarily consistent.

Related to high-dimensional data, it is important to consider selection of regression variables in the case that p is larger than n, and k is also large. When p is large, it will be natural to consider a covariance structure, since a covariance matrix with no covariance structure involves many unknown parameters. One way is to consider a sparse method or a joint regularization of the regression parameters and the inverse covariance matrix, see, e.g., Rothman et al. (2010). As another approach, it may consider to select the regression variables, assuming a simple covariance structure. Related to the later approach, it might occur to select an appropriate simple covariance structure from a set of some simple covariance structures.

In this paper, first we consider the variables selection problems under some simple covariance structures such as (1) independent covariance structure, (2) uniform covariance structure and (3) autoregressive covariance structure, based on model selection approach. We study consistency properties of information criteria including AIC and BIC. These information criteria have a computational problem when k becomes large. In order to avoid such problem, we consider their efficient criteria based on Zho et al. (1986) and Nishii et al. (1988). It is shown that the efficient criteria also consistent in a high dimensional situation. Our results are checked numerically by conducting a Mote Carlo simulation.

Next we discuss with AIC and BIC for selecting (4) independence covariance structure with different variances, and (5) no covariance structure, in addition to covariance structures (1) $\sim$ (3). In a high-dimensional situation, it is noted that the covariance structures except for (2) are identified by AIC and BIC, through a simulation experiment.

The present paper is organized as follows. In section 2, we present notations and preliminaries. In Sections 3, 4 and 5 we show high-dimensional consistencies of information criteria under the covariance structures (1), (2) and (3), respectively. These are also numerically studied. In Section 6, their efficient criteria are shown to be consistent under the same condition as in the case of information criteria. In Section 7, we discuss with selections of covariance structures by AIC and BIC. In Section 8, our conclusions are discussed. The proofs of our results are given in Appendix.

2. Notations and Preliminaries

This paper is concerned with selection of explanatory variables in multivariate regression model (1.1). Suppose that j denotes a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements, and $\mathbf{X}_j$ denotes the $n \times k_j$ matrix consisting the columns of $\mathbf{X}$ indexed by the elements of j. Then, $\mathbf{X}_\omega = \mathbf{X}$. Further, we assume that the covariance matrix $\mathbf{\Sigma}$ have a covariance structure $\mathbf{\Sigma}_g$. Then a generic candidate model can be expressed as

$$M_{g,j} : \ \mathsf{Y} \sim \mathrm{N}_{n \times p}(\mathbf{X}_j \mathbf{\Theta}_j, \mathbf{\Sigma}_{g,j} \otimes \mathbf{I}_n), \tag{2.1}$$

where $\mathbf{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients. We assume that $\mathrm{rank}(\mathbf{X}) = k (< n)$.

As a model selection method, we use a generalized criterion of AIC (Akaike (1973)). When $\mathbf{\Sigma}_{g,j}$ is a $p \times p$ unknown covariance matrix, the AIC (see, e.g., Bedrick and Tsai (1994), Fujikoshi and Satoh (1997)) for $M_{g,j}$ is given by

$$\mathrm{AIC}_{g,j} = n \log |\hat{\mathbf{\Sigma}}_{g,j}| + np(\log 2\pi + 1) + 2 \left\{ k_j p + \frac{1}{2} p(p+1) \right\}, \tag{2.2}$$

where $n\hat{\mathbf{\Sigma}}_{g,j} = \mathsf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathsf{Y}$ and $\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$. The part of "$n \log |\hat{\mathbf{\Sigma}}_{g,j}| + np(\log 2\pi + 1)$" is "$-2 \log \max_{M_{g,j}} f(\mathsf{Y}; \mathbf{\Theta}_j, \mathbf{\Sigma}_{g,j})$", where $f(\mathsf{Y}; \mathbf{\Theta}_j, \mathbf{\Sigma}_{g,j})$ is the density function of Y under $M_{g,j}$. The AIC was introduced as an asymptotic unbiased estimator for the risk function defined as the expected log-predictive-likelihood or equivalently the Kullback-Leibler information, for a candidate model $M_{g,j}$, see, e.g., Fujikoshi and Satoh (1997). When $j = \omega$, the model $M_{g,\omega}$ is called the full model. Note that $\hat{\mathbf{\Sigma}}_{g,\omega}$ and $\mathbf{P}_\omega$ are defined from $\hat{\mathbf{\Sigma}}_{g,j}$ and $\mathbf{P}_j$ as $j = \omega$, $k_\omega = k$ and $\mathbf{X}_\omega = \mathbf{X}$.

In this paper, first we consider the case that the covariance matrix $\mathbf{\Sigma}$ belongs to each of the following three classes;

(1) Independent covariance structure (IND); $\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}_p$,

(2) Uniform covariance structure (UNIF); $\mathbf{\Sigma}_u = \sigma_u^2(\rho_u^{1-\delta_{ij}})_{1 \le i,j \le p}$,

(3) Autoregressive covariance structure (AUTO); $\mathbf{\Sigma}_a = \sigma_a^2(\rho_a^{|i-j|})_{1 \le i,j \le p}$.

Our candidate model can be expressed as (2.1) with $\Sigma_{v,j}$, $\Sigma_{u,j}$ or $\Sigma_{a,j}$ for $\Sigma_{g,j}$. For deriving the maximum likelihood under $M_{g,j}$, we shall use the fact that for any positive definite $\mathbf{\Sigma}_{g,j}$,

$$\begin{aligned}\max_{\mathbf{\Theta}_j} f(\mathbf{Y}; \mathbf{\Theta}_j, \mathbf{\Sigma}_{g,j}) &= np \log |\mathbf{\Sigma}_{g,j}| + np(\log 2\pi + 1) \\ &\quad + \min_{\mathbf{\Theta}_j} \mathrm{tr} \mathbf{\Sigma}_{g,j}^{-1} (\mathbf{Y} - \ X_j \mathbf{\Theta}_j)' (\mathbf{Y} - \ X_j \mathbf{\Theta}_j) \qquad (2.3) \\ &= np \log |\mathbf{\Sigma}_{g,j}| + np \log 2\pi + \mathrm{tr} \mathbf{\Sigma}_{g,j}^{-1} \mathbf{Y} (\mathbf{I}_n - \mathbf{P}_j) \mathbf{Y}.\end{aligned}$$

Let $\hat{\mathbf{\Sigma}}_{g,j}$ be the quantity minimizing the right side of (2.3). Then, in our problem, it satisfies

$$\mathrm{tr} \hat{\mathbf{\Sigma}}_{g,j}^{-1} \mathbf{Y} (\mathbf{I}_n - \mathbf{P}_j) \mathbf{Y} = np.$$

We consider a general information criterion defined by

$$\begin{aligned}\mathrm{IC}_{g,d,j} &= -2 \log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_{g,j}) + dm_{g,j} \\ &= np \log |\mathbf{\Sigma}_{g,j}| + np(\log 2\pi + 1) + dm_{g,j}, \qquad (2.4)\end{aligned}$$

where $m_{g,j}$ is the number of independent unknown parameters under $M_{g,j}$, and d is a positive constant which may depend on n. When $d = 2$ and $d = \log n$,

$$\mathrm{IC}_{g,2,j} = \mathrm{AIC}_{g,j}, \quad \mathrm{IC}_{g,\log n,j} = \mathrm{BIC}_{g,j}.$$

Such general information criterion was considered in a univariate regression model by Nishii (1984).

For each of the three covariance structures, consider to select the best model from all the models or a subset of all the models. Let $\mathcal{F}$ be the set of all the candidate models, which is denoted by

$$\{\{1\}, \ldots, \{k\}, \{1, 2\}, \ldots, \{1, \ldots, k\}\},$$

or its subfamily. Then, our model selction criterion is to select the model M_j or the subset j minimizing $\mathrm{IC}_{g,d,j}$, which is written as

$$\hat{j}_{\mathrm{IC}g,d} = \arg \min_{j \in \mathcal{F}} \mathrm{IC}_{g,d,j}. \qquad (2.5)$$

For studying consistency properties of $\hat{j}_{\mathrm{IC}g,d}$, it is assumed that the true model $M_{g,*}$ is included in the full model, i.e.,

$$M_{g,*} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_*\mathbf{\Theta}_*, \mathbf{\Sigma}_{g,*} \otimes \mathbf{I}_n). \tag{2.6}$$

Let us denote the minimum model including the true model by M_{g,j_*}. The true mean of $\mathbf{Y}$ is expresed as

$$\mathbf{X}_*\mathbf{\Theta}_* = \mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}$$

for some $k_{j_*} \times p$ matrix $\mathbf{\Theta}_{j_*}$. So, the notation $\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}$ is also used for the true mean of $\mathbf{Y}$. Let $\mathcal{F}$ separate into two sets, one is a set of overspecified models, i.e., $\mathcal{F}_+ = \{j \in \mathcal{F} \mid j_* \subseteq j\}$ and the other is a set of underspecified models, i.e., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F}$.

Here we list some of our main assumptions:

A1 (The true model): $M_{g,*} \in \mathcal{F}$.

A2 (The asymptotic framework): $p \to \infty,\ n \to \infty,\ p/n \to c \in (0, \infty)$.

It is said that a general model selection criterion $\hat{j}_{\mathrm{IC}g,d}$ is consistent if

$$\lim_{p/n\to c\in(0,\infty)} \Pr(\hat{j}_{\mathrm{IC}g,d} = j_*) = 1.$$

In order to obtain $\hat{j}_{\mathrm{IC}g,d}$, we must calcurate $\mathrm{IC}_{g,d}$ for all the subsets of $j_\omega = \{1, 2, \ldots, k\}$, i.e., $2^k - 1$ $\mathrm{IC}_{g,d}$. This will become extensive computation as k becomes large. As a method of overcoming this weak point, we consider EC criterion based on Zho et al. (1986) and Nishii et al. (1988). Let $j_{(i)}$ be the subset of j_ω omitting the i $(1 \leq i \leq k)$. Then $\mathrm{EC}_{g,d}$ is defined to select

$$\hat{j}_{\mathrm{EC}_{g,d}} = \{i \in j_\omega \mid \mathrm{EC}_{g,d,j_{(i)}} > \mathrm{EC}_{g,d,j_\omega},\ i = 1, \ldots, k\}. \tag{2.7}$$

In Section 6 we shall show that $\hat{j}_{\mathrm{EC}_{g,d}}$ has an consistency property.

3. IC under Independent Covariance Structure

In this section we consider the problem of selecting the regression variables in the multivariate regression model under the assumption that the covariance matrix has an independent covariance structure. A generic candidate model can be expressed as

$$M_{v,j}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_j\mathbf{\Theta}_j, \mathbf{\Sigma}_{v,j} \otimes \mathbf{I}_n), \tag{3.1}$$

where $\mathbf{\Sigma}_{v,j} = \sigma_{v,j}^2\mathbf{I}_p$ and $\sigma_{v,j} > 0$. Then, we have

$$\begin{aligned} -2\log f(\mathbf{Y}; \mathbf{\Theta}_j, \sigma_{v,j}^2) &= np\log(2\pi) + np\log\sigma_{v,j}^2 \\ &\quad + \frac{1}{\sigma_{v,j}^2}\mathrm{tr}(\mathbf{Y} - \mathbf{X}_j\mathbf{\Theta}_j)'(\mathbf{Y} - \mathbf{X}_j\mathbf{\Theta}_j). \end{aligned}$$

Therefore, it is easily seen that the maximum estimators of $\mathbf{\Theta}_j$ and $\sigma_{v,j}^2$ under $M_{v,j}$ are given as

$$\hat{\mathbf{\Theta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{Y}, \quad \hat{\sigma}_{v,j}^2 = \frac{1}{np}\mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}. \tag{3.2}$$

The information criterion (2.4) is given by

$$\mathrm{IC}_{v,d,j} = np\log\hat{\sigma}_{v,j}^2 + np(\log 2\pi + 1) + d \times m_{v,j}, \tag{3.3}$$

where d is a positive constant and $m_{v,j} = k_jp + 1$. Assume that the true model is expressed as

$$M_{v,*}: \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_*\mathbf{\Theta}_*, \sigma_{v,*}^2\mathbf{I}_p \otimes \mathbf{I}_n), \tag{3.4}$$

and denote the minimum model including the true model $M_{v,*}$ by M_{v,j_*}. In general, $\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}$ is distributed as a nocentral Wishart distribution, more precisely

$$\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} \sim \mathrm{W}_p(n - k_j, \mathbf{\Sigma}_{v,*}; (\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}),$$

which implies the following Lemma.

Lemma 3.1. *Under* (3.4), $np\hat{\sigma}_{v,j}^2/\sigma_{v,*}^2$ *is distributed as a noncentral chi-square distribution* $\chi^2_{(n-k_j)p}(\delta^2_{v,j})$, *where*

$$\begin{aligned}\delta^2_{v,j} &= \frac{1}{\sigma^2_{v,*}}\mathrm{tr}(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*} \\ &= \frac{1}{\sigma^2_{v,*}}\mathrm{tr}(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{P}_\omega - \mathbf{P}_j)\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}. \qquad (3.5)\end{aligned}$$

When $j \in \mathcal{F}_+$, *then,* $\delta^2_{v,j} = 0$.

For a sufficient condition for consistency of $\mathrm{IC}_{v,g}$, we assume

$$\text{A3v : For any } j \in \mathcal{F}_-,\ \delta^2_{v,j} = \mathrm{O}(np), \text{and } \lim_{p/n\to c}\frac{1}{np}\delta^2_{v,j} = \eta^2_{v,j} > 0. \qquad (3.6)$$

Theorem 3.1. *Suppose that the assumptions* A1, A2 *and* A3v *are satisfied. Then, the information criteria* $\mathrm{IC}_{v,d}$ *defined by* (3.3) *is consistent if* $d > 1$ *and* $d/n \to 0$.

AIC and BIC satisfy the conditions $d > 1$ and $d/n \to 0$, and we have the following result.

Corollary 3.1. *Under the assumptions* A1, A2 *and* A3v AIC *and* BIC *are consistent.*

In the following we numerically examine the validity of our claims. The true model was assumed as

$$M_{v,*} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_*\mathbf{\Theta}_*, \sigma^2_{v,*}\mathbf{I}_p \otimes \mathbf{I}_n).$$

Here $\mathbf{\Theta}_* : 3 \times p$ was determined by random numbers from the uniform distribution on $(1, 2)$, i.e., i.i.d. from $\mathrm{U}(1, 2)$. The first column of $\mathbf{X}_\omega : n \times 10$ is $\mathbf{1}_n$, and the other elements are i.i.d. from $\mathrm{U}(-1, 1)$. The true variance was set as $\sigma^2_{v,*} = 2$. The five candidate models M_{j_α}, $\alpha = 1, 2, \ldots, 5$ were considered, where $j_\alpha = \{1, \ldots, \alpha\}$, We studied selection percentages of the true model for 10^4 replications under AIC and BIC for

$$(n, p) = (50, 15), (100, 30), (200, 60), (50, 100), (100, 200), (200, 400)$$

The results are given in Tables 4.1 and 4.2. It is seen that the true model has been selected for all the cases except the case $(n, p) = (50, 15)$ of AIC$_v$. In the case $(n, p) = (50, 15)$ of AIC$_v$, the selection percentage is not 100, but it is very high.

Table 3.1. Selection percentages of AIC$_v$ and BIC$_v$ for $p/n = 0.3$

	AIC$_v$			BIC$_v$		
j	(50,15)	(100,30)	(200,60)	(50,15)	(100,30)	(200,60)
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	97.3	99.9	100.0	100.0	100.0	100.0
4	2.4	0.1	0.0	0.0	0.0	0.0
5	0.3	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

Table 3.2. Selection percentages of AIC$_v$ and BIC$_v$ for $p/n = 2$

	AIC$_v$			BIC$_v$		
j	(50,100)	(100,200)	(200,400)	(50,100)	(100,200)	(200,400)
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	100.0	100.0	100.0	100.0	100.0	100.0
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

4. IC under Uniform Covariance Structure

In this section we consider model selection criterion when the covariance matrix has a uniform covariance structure

$$\boldsymbol{\Sigma}_u = \sigma_u^2(\rho_u^{1-\delta_{ij}}) = \sigma_u^2\{(1-\rho_u)\mathbf{I}_p + \rho_u \mathbf{1}_p\mathbf{1}_p'\}. \tag{4.1}$$

The covariance structure is expressed as

$$\boldsymbol{\Sigma}_u = \tau_1\left(\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p\right) + \tau_2\frac{1}{p}\mathbf{G}_p,$$

where

$$\tau_1 = \sigma_u^2(1-\rho_u), \quad \tau_2 = \sigma_u^2\{1+(p-1)\rho_u\}, \quad \mathbf{G}_p = \mathbf{1}_p\mathbf{1}_p',$$

and $\mathbf{1}_p = (1,\ldots,1)'$. Noting that the matrices $\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p$ and $\frac{1}{p}\mathbf{G}_p$ are orthogonal idempotent matrices, we have

$$|\boldsymbol{\Sigma}_u| = \tau_2\tau_1^{p-1}, \quad \boldsymbol{\Sigma}_u^{-1} = \frac{1}{\tau_1}\left(\mathbf{I}_p - \frac{1}{p}\mathbf{G}_p\right) + \frac{1}{\tau_2}\frac{1}{p}\mathbf{G}_p.$$

Now we consider the model $M_{u,j}$ given by

$$M_{u,j}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_{u,j}\otimes\mathbf{I}_n), \tag{4.2}$$

where $\boldsymbol{\Sigma}_{u,j} = \tau_{1j}\left(\mathbf{I}_p - p^{-1}\mathbf{G}_p\right) + \tau_{2j}p^{-1}\mathbf{G}_p$. Let $\mathbf{H} = (\boldsymbol{h}_1, \mathbf{H}_2)$ be an orthognal matrix where $\boldsymbol{h}_1 = p^{-1/2}\mathbf{1}_p$, and let

$$\mathbf{U}_j = \mathbf{H}'\mathbf{W}_j\mathbf{H}, \quad \mathbf{W}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}.$$

Let the density function of $\mathbf{Y}$ under $M_{u,j}$ denote by $f(\mathbf{Y};\boldsymbol{\Theta}_j,\tau_{1j},\tau_{2j})$. From (2.3) we have

$$\begin{aligned} g(\tau_{1j},\tau_{2j}) &= -2\log\max_{\boldsymbol{\Theta}_j} f(\mathbf{Y};\boldsymbol{\Theta}_j,\tau_{1j},\tau_{2j}) \\ &= np\log(2\pi) + n\log\tau_{2j} + n(p-1)\log\tau_{1j} + \mathrm{tr}\boldsymbol{\Psi}_j^{-1}\mathbf{U}_j, \end{aligned}$$

where $\boldsymbol{\Psi}_j = \mathrm{diag}(\tau_{2j},\tau_{1j},\ldots,\tau_{1j})$. Then, it can be shown that the maximum likelihood estimators of τ_{1j} and τ_{2j} under $M_{u,j}$ are given by

$$\begin{aligned} \hat{\tau}_{1j} &= \frac{1}{n(p-1)}\mathrm{tr}\ \mathbf{D}_1\mathbf{U}_j = \frac{1}{n(p-1)}\mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\mathbf{H}_2, \\ \hat{\tau}_{2j} &= \frac{1}{n}\mathrm{tr}\ \mathbf{D}_2\mathbf{U}_j = \frac{1}{n}\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\boldsymbol{h}_1, \end{aligned}$$

where $\mathbf{D}_1 = \mathrm{diag}(0,1,\ldots,1)$ and $\mathbf{D}_2 = \mathrm{diag}(1,0,\ldots,0)$. The number of independent parameters under $M_{u,j}$ is $m_j = k_jp+2$. Noting that $\boldsymbol{\Psi}_j$ is diagonal, we can get the information criterion (2.4) given as

$$\mathrm{IC}_{u,d,j} = n(p-1)\log\hat{\tau}_{1j} + n\log\hat{\tau}_{2j} + np(\log 2\pi + 1) + d(k_jp+2). \tag{4.3}$$

Assume that the true model is expressed as

$$M_{u,*} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_*\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_{u,*} \otimes \mathbf{I}_n), \tag{4.4}$$

where $\boldsymbol{\Sigma}_{u,*} = \tau_{1*}\left(\mathbf{I}_p - p^{-1}\mathbf{G}_p\right) + \tau_{2*}p^{-1}\mathbf{G}_p$, and denote the minimum model including the true model $M_{u,*}$ by M_{u,j_*}. In general, it holds that $\mathbf{U}_j \sim \mathrm{W}_p(n-k_j, \boldsymbol{\Psi}_*; \boldsymbol{\Delta}_j)$, where

$$\boldsymbol{\Delta}_j = (\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*}\mathbf{H})'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*}\mathbf{H},$$

and $\boldsymbol{\Psi}_* = \mathrm{diag}(\tau_{2*}, \tau_{1*}, \ldots, \tau_{1*})$. Therefore, we have the following Lemma (see, e.g., Fujikoshi et al. (2010)).

Lemma 4.1. *Under the true model* (3.4)*, it holds that*
(1) $n(p-1)\tau_{1*}^{-1}\hat{\tau}_{1j}$ *is distributed as a noncentral distribution* $\chi^2_{(p-1)(n-k_j)}(\delta^2_{1j})$*, where*

$$\delta^2_{1j} = \frac{1}{\tau_{1*}}\mathrm{tr}\mathbf{H}_2'(\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*})\mathbf{H}_2.$$

(2) $n\tau_{2*}^{-1}\hat{\tau}_{2j}$ *is distributed as a noncentral distribution* $\chi^2_{n-k_j}(\delta^2_{2j})$*, where*

$$\delta^2_{2j} = \frac{1}{\tau_{2*}}\mathrm{tr}\boldsymbol{h}_1'(\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{X}_{j_*}\boldsymbol{\Theta}_{j_*})\boldsymbol{h}_1.$$

(3) *If* $j \in \mathcal{F}_+$*, then* $\delta_{1j} = 0$ *and* $\delta_{2j} = 0$*.*

For a sufficient condition for consistency of $\mathrm{IC}_{u,d}$, we assume

A3u: For any $j \in \mathcal{F}_-$, $\delta^2_{1j} = \mathrm{O}(np)$, $\delta^2_{2j} = \mathrm{O}(n)$ and

$$\lim_{p/n\to c}\frac{1}{np}\delta^2_{1j} = \eta^2_{1j} > 0, \quad \lim_{p/n\to c}\frac{1}{n}\delta^2_{2j} = \eta^2_{2j} > 0, \tag{4.5}$$

Theorem 4.1. *Suppose that the assumptions* A1, A2 *and* A3u *are satisfied. Then, the information criteria* $\mathrm{IC}_{u,d}$ *defined by* (4.3) *is consistent if* $d > 1$ *and* $d/n \to 0$*.*

Corollary 4.1. *Under the assumptions* A1, A2 *and* A3u AIC *and* BIC *are consistent.*

We tried a numerical experiment under the same setting as in the independence covariance structure except for covariance structure. The true uniform covariance structure was set as the one with $\sigma^2_{u,*} = 2$, $\rho_{u,*} = 0.2$. The results are given Tables 4.1 and 4.2. In general, it seems that AIC selects the true model even a finite setting with a high probability. However, BIC does not always select the true model when (n, p) is small, though it has a consistency property.

Table 4.1. Selection percentages of AIC and BIC for $p/n = 0.3$

	AIC			BIC		
j	(50,15)	(100,30)	(200,60)	(50,15)	(100,30)	(200,60)
1	0.2	0.0	0.0	72.8	1.3	0.0
2	0.6	0.0	0.0	7.6	3.0	0.0
3	96.5	99.9	100.0	19.6	95.7	100.0
4	2.4	0.1	0.0	0.0	0.0	0.0
5	0.3	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

Table 4.2. Selection percentages of AIC and BIC for $p/n = 2$

	AIC			BIC		
j	(50,100)	(100,200)	(200,400)	(50,100)	(100,200)	(200,400)
1	0.1	0.0	0.0	100.0	99.6	0.0
2	3.1	0.0	0.0	0.0	0.4	0.0
3	96.9	100.0	100.0	0.0	0.0	100.0
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

5. IC **under Autoregressive Covariance Structure**

In this section we consider model selection criterion when the covariance matrix $\mathbf{\Sigma}$ has an autoregressive covariance structure

$$\mathbf{\Sigma}_a = \sigma_a^2(\rho_a^{|i-j|})_{1\leq i,j\leq p}. \tag{5.1}$$

Then, it is well known (see, e.g., Fujikoshi et al. (1990))

$$|\mathbf{\Sigma}_a| = (\sigma_a^2)^p(1-\rho_a^2)^{p-1}, \quad \mathbf{\Sigma}_a^{-1} = \frac{1}{\sigma_a^2(1-\rho_a^2)}(\rho_a^2\mathbf{C}_1 - 2\rho_a\mathbf{C}_2 + \mathbf{C}_0),$$

where $\mathrm{C}_0 = \mathbf{I}_p$,

$$\mathbf{C}_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \quad \mathbf{C}_2 = \frac{1}{2}\begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Now we consider the model $M_{a,j}$ given by

$$M_{a,j}: \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\ X_j\mathbf{\Theta}_j, \mathbf{\Sigma}_{a,j}\otimes \mathbf{I}_n), \tag{5.2}$$

where $\mathbf{\Sigma}_{a,j} = \sigma_{a,j}^2(\rho_{a,j}^{|i-j|})$. Then, from (2.3) the maximum likelihood estimate of $\mathbf{\Theta}_j$ is given by

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

and the maximum likelihood estimators of $\rho_{a,j}$ and $\sigma_{a,j}^2$ can be obtained as the minimization of

$$\begin{aligned} &-2\log f(\mathbf{Y};\hat{\mathbf{\Theta}}_{a,j},\sigma_{a,j}^2,\rho_j) = np\log(2\pi) + np\log\sigma_{a,j}^2 + n(p-1)\log(1-\rho_{a,j}^2) \\ &\quad + \frac{1}{\sigma_{a,j}^2(1-\rho_j^2)} + \mathrm{tr}(\rho_{a,j}^2\mathbf{C}_1 - 2\rho_{a,j}\mathbf{C}_2 + \mathbf{C}_0)\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} \end{aligned}$$

with respect to $\sigma_{a,j}$ and $\rho_{a,j}$. Therefore, the maximum likelihood estimators of $\sigma_{a,j}^2$ and $\rho_{a,j}$ are given (see Fujikoshi et al. (1990)) through the following two equations:

$$(1) \quad \hat{\sigma}^2_{a,j} = \frac{n-k_j}{n} \frac{1}{p(1-\hat{\rho}^2_{a,j})}(a_{1j}\hat{\rho}^2_{a,j} - 2a_{2j}\hat{\rho}_{a,j} + a_{0j}), \tag{5.3}$$

$$(2) \quad (p-1)a_{1j}\hat{\rho}^3_{a,j} - (p-2)a_{2j}\hat{\rho}^2_{a,j} - (pa_{1j}+a_{0j})\hat{\rho}_{a,j} + pa_{2j} = 0, \tag{5.4}$$

where $a_{ij} = \mathrm{tr}\mathbf{C}_i\mathbf{S}_j,\ i = 0,1,2$, and $\mathbf{S}_j = (n-k_j)^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}$. Then, the information criterion $\mathrm{IC}_{a,d,j}$ can be written as

$$\begin{aligned}\mathrm{IC}_{a,d,j} =& np\log\hat{\sigma}^2_{a,j} + n(p-1)\log(1-\hat{\rho}^2_{a,j}) + np(\log 2\pi + 1) \\ &+ d(k_j p + 2).\end{aligned} \tag{5.5}$$

Note that the maximum likelihood estimators $\hat{\rho}$ and σ^2 are expressed in terms of a_{0j}, a_{1j} and a_{2j} or

$$b_{0j} = \mathrm{tr}\mathbf{C}_0\mathbf{W}_j, \quad b_{1j} = \mathrm{tr}\mathbf{C}_1\mathbf{W}_j, \quad b_{2j} = \mathrm{tr}\mathbf{C}_2\mathbf{W}_j, \tag{5.6}$$

where

$$\mathbf{W}_j = (n-k_j)\mathbf{S}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}. \tag{5.7}$$

Assume that the true model is expressed as

$$M_{a,*} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_*\mathbf{\Theta}_*, \mathbf{\Sigma}_{a,*} \otimes \mathbf{I}_n), \tag{5.8}$$

where $\mathbf{\Sigma}_{a,*} = \sigma^2_{a,*}(\rho^{|i-j|}_{a,*})$, and denote the minimum model including the true model M_* by M_{j_*}. Then,

$$\mathbf{W}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} \sim \mathrm{W}_p(n-k_j, \mathbf{\Sigma}_{a,*}; \mathbf{\Omega}_j),$$

where

$$\mathbf{\Omega}_j = (\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}.$$

For relating to the noncentrality matrix $\mathbf{\Omega}_j$, we use the following three quantities:

$$\delta_{ij} = \mathrm{tr}\mathbf{C}_i\mathbf{\Omega}_j, \quad i = 0,1,2. \tag{5.9}$$

As a sufficient condition for consistency, we assume

A3a: For any $j \in \mathcal{F}_-$, the order of each element of $\mathbf{\Omega}_j$ is $\mathrm{O}(n)$, $\delta^2_{ij} = \mathrm{O}(np)$, and

$$\lim_{p/n\to c} \frac{1}{np}\delta^2_{ij} = \eta^2_{ij} > 0,\ i = 0,1,2. \tag{5.10}$$

Theorem 5.1. *Suppose that the assumptions* A1, A2 *and* A3a *are satisfied. Then, the information criteria* $\mathrm{IC}_{a,d}$ *defined by* (5.5) *is consistent if* $d > 1$ *and* $d/n \to 0$.

Corollary 5.1. *Under the the assumptions* A1, A2 *and* A3a AIC *and* BIC *are consistent.*

We tried a numerical experiment under the same setting as in the independence covariance structure and in the uniform covariance structure except for covariance structure. Here the true covariance structure was set as the autoregressive covariance structure $\mathbf{\Sigma}_{a,*} = \sigma^2_{a,*}(\rho^{|i-j|}_{a,*})$ with $\sigma^2_{a,*} = 2$, $\rho_{a,*} = 0.2$. The results are given Tables 5.1 and 5.2. It is seen that AIC and BIC are selecting the true model in all the cases.

Table 5.1. Selection percentages of AIC and BIC for $p/n = 0.3$

	AIC			BIC		
j	(50,15)	(100,30)	(200,60)	(50,15)	(100,30)	(200,60)
1	0.0	0.0	0.0	0.5	0.0	0.0
2	0.0	0.0	0.0	0.1	0.0	0.0
3	97.2	99.9	100.0	99.4	100.0	100.0
4	2.4	0.1	0.0	0.0	0.0	0.0
5	0.3	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

Table 5.2. Selection percentages of AIC and BIC for $p/n = 2$

	AIC			BIC		
j	(50,100)	(100,200)	(200,400)	(50,100)	(100,200)	(200,400)
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	100.0	100.0	100.0	100.0	100.0	100.0
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0
≥ 6	0.0	0.0	0.0	0.0	0.0	0.0

6. Consitency Properties of EC

In this section we are interested in consistency properties of three efficient criteria $\mathrm{EC}_{v,d}$, $\mathrm{EC}_{u,d}$ and $\mathrm{EC}_{a,d}$ defined from $\mathrm{IC}_{v,d}$, $\mathrm{IC}_{u,d}$ and $\mathrm{IC}_{a,d}$ through (2.7). Note that consistency properties of $\mathrm{IC}_{v,d}$, $\mathrm{IC}_{u,d}$ and $\mathrm{IC}_{a,d}$ are given in Theorems 3.1, 4.1 and 5.1. Using these consitency properties it is expected that $\mathrm{EC}_{v,d}$, $\mathrm{EC}_{u,d}$ and $\mathrm{EC}_{a,d}$ have similar consistency properties. In fact, the following result holds.

Theorem 6.1. *Suppose that the assumptions* A1 *and* A2 *are satisfied. Then, it holds that*

(1) *the efficient criterion* $\mathrm{EC}_{v,d}$ *is consistent under the assumption* A3v *if* $d > 1$ *and* $d/n \to 0$.

(2) *the efficient criterion* $\mathrm{EC}_{u,d}$ *is consistent under the assumption* A3u *if* $d > 1$ *and* $d/n \to 0$.

(3) *the efficient criterion* $\mathrm{EC}_{a,d}$ *is consistent under the assumption* A3a *if* $d > 1$ *and* $d/n \to 0$.

In order to examine the validity of the results and the speed of convergences we tried a numerical experiment. The simulation settings are similar to the cases of $\mathrm{IC}_{v,d}$, $\mathrm{IC}_{u,d}$ and $\mathrm{IC}_{a,d}$ except for that the following points: The total number of explanatory variables to be selected was changed to 5 from 10. The true covariance structures were set as follows for IND(independent covariance structure), UNIF(uniform covariance structure) and AUTO(autoregressive covariance structure):

$$\begin{aligned}
\mathrm{IND} :& \boldsymbol{\Sigma} = \sigma_{v,*}^2 \mathbf{I}_p, \quad \sigma_{v,*}^2 = 2. \\
\mathrm{UNIF} :& \boldsymbol{\Sigma} = \sigma_{u,*}^2 (\rho_{u,*}^{1-\delta_{ij}}), \quad \sigma_{u,*}^2 = 2, \quad \rho_{u,*} = 0.9. \\
\mathrm{AUTO} :& \boldsymbol{\Sigma} = \sigma_{a,*}^2 (\rho_{a,*}^{|i-j|}), \quad \sigma_{a,*}^2 = 2, \quad \rho_{a,*} = 0.9.
\end{aligned}$$

Let EC_{A} and EC_{B} be the efficient criterion based on AIC and BIC, respectively. Selection rates of these criteria are given in Tables 6.1 $\sim$ 6.4 for each of three covariance structures. In the tables the column of x_i denotes the selection rate for the i-th explanatory variable x_i. The "Under",

”True” and ”Over” denote the underspecified models, the true model and the overspecified models, respectively.

Table 6.1. Selection percentages of $\mathrm{EC_A}$ and $\mathrm{EC_B}$ for $(n, p) = (20, 10)$

$n = 20, p = 10$		Under	True	Over	x_1	x_2	x_3	x_4	x_5
$\mathrm{EC_A}$	IND	0.03	0.74	0.23	1.00	1.00	0.97	0.13	0.12
	UNIF	0.36	0.47	0.17	1.00	0.83	0.73	0.14	0.12
	AUTO	0.40	0.44	0.17	1.00	0.84	0.68	0.14	0.13
$\mathrm{EC_B}$	IND	0.21	0.77	0.02	1.00	1.00	0.79	0.01	0.01
	UNIF	0.78	0.22	0.01	1.00	0.50	0.37	0.01	0.01
	AUTO	0.81	0.18	0.01	1.00	0.49	0.30	0.01	0.01

Table 6.2. Selection percentages of $\mathrm{EC_A}$ and $\mathrm{EC_B}$ for $(n, p) = (200, 100)$

$n = 200, p = 100$		Under	True	Over	x_1	x_2	x_3	x_4	x_5
$\mathrm{EC_A}$	IND	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	UNIF	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	AUTO	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
$\mathrm{EC_B}$	IND	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	UNIF	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	AUTO	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00

Table 6.3. Selection percentages of $\mathrm{EC_A}$ and $\mathrm{EC_B}$ for $(n, p) = (10, 20)$

$n = 10, p = 20$		Under	True	Over	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\mathrm{EC_A}$	IND	0.07	0.41	0.53	1.00	1.00	0.94	0.37	0.34
	UNIF	0.47	0.15	0.38	0.97	0.85	0.62	0.39	0.35
	AUTO	0.47	0.15	0.38	0.95	0.81	0.65	0.37	0.35
$\mathrm{EC_B}$	IND	0.15	0.54	0.31	1.00	0.99	0.86	0.20	0.19
	UNIF	0.68	0.15	0.17	0.93	0.70	0.43	0.20	0.20
	AUTO	0.71	0.13	0.15	0.87	0.63	0.46	0.20	0.20

Table 6.4. Selection percentages of $\mathrm{EC_A}$ and $\mathrm{EC_B}$ for $(n, p) = (100, 200)$

$n = 100, p = 200$		Under	True	Over	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\mathrm{EC_A}$	IND	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	UNIF	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	AUTO	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
$\mathrm{EC_B}$	IND	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	UNIF	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00
	AUTO	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00

7. Selection of Covariance Structures

In this section we consider high-dimensional properties of AIC and BIC for selection of covariance structures. The covariance structures considered are (1) IND(independent covariance structure), (2) UNIF(uniform covariance structure), (3) AUTO(autoregressive covariance structure), (4) NOST(no covariance structure). For the first three covariance structures, we use the same notation as in Sections 3, 4 and 5. The covariance matrix under NOST is denoted by $\boldsymbol{\Sigma}_{m,j}$. When $\mathrm{E}(\mathbf{Y}) = \mathbf{X}_j\boldsymbol{\Theta}_j$, AIC criteria for these four models are expressed as

$$\begin{aligned}
\mathrm{A}_{v,j} &= np\log\hat{\sigma}_{vj}^2 + np(\log 2\pi + 1) + 2(jp + 1),\\
\mathrm{A}_{u,j} &= n\log\hat{\tau}_{2j} + n(p-1)\log\hat{\tau}_{1j} + np(\log 2\pi + 1) + 2(jp + 2),\\
\mathrm{A}_{a,j} &= np\log\hat{\sigma}_{aj}^2 + n(p-1)\log(1 - \hat{\rho}_{aj}^2) + np(\log 2\pi + 1) + 2(jp + 2),\\
\mathrm{A}_{m,j} &= n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + 2\left\{jp + \frac{1}{2}p(p+1)\right\},
\end{aligned}$$

The expressions for $\hat{\sigma}_{vj}^2$, $\hat{\tau}_{1j}$, $\hat{\tau}_{2j}$, $\hat{\sigma}_{aj}^2$ and $\hat{\rho}_{aj}^2$ are given in Sections 3, 4 and 5. The BIC are defined from AIC by replacing "2" in the "2× the number of independent parameters" to "$\log n$".

In order to see high-dimensional behaviors of AIC and BIC, simulation experiments were done. For multivariate regression models, we selected

$$k_* = 3, \quad k_\omega = 5.$$

The multivariate regression model was set by the same way as in Sections 3, 4 and 5. For the first covariance structure, $\sigma^2 = 2$. For the second and third covariance structures,

$$\sigma^2 = 2, \quad \rho = 0.9.$$

For no-strctured case, the true covariance matrix $\mathbf{\Sigma}_* = (\sigma_{ij})$ was set as follows:

$$\sigma_{ii} = \sigma^2 \left(1 + 3\frac{i-1}{p-1}\right), \quad i = 1, \ldots, p,$$

The other elements of $\mathbf{\Sigma}_* = (\sigma_{ij})$ are i.i.d. from U(0.3, 0.9).

The simulation results are given in Table 7.1.

Table 7.1. Selection rates of AIC for the four covariance structures

	$n = 20$, $p = 10$				$n = 200$, $p = 100$			
TRUE	IND	UNIF	AUTO	NOST	IND	UNIF	AUTO	NOST
IND	0.45	0.13	0.14	0.29	0.71	0.14	0.14	0.00
UNIF	0.00	0.69	0.00	0.31	0.00	1.00	0.00	0.00
AUTO	0.00	0.00	0.69	0.31	0.00	0.00	1.00	0.00
NOST	0.03	0.21	0.04	0.72	0.00	0.00	0.00	1.00

It is seen that though it is difficult to select IND covariance structure correctively, but the other three model will be correctly selected. IND covariance structure can be seen as a limit of UNIF and AUTO covariance matrices. So, it seems that it is difficult to select IND covariance structure correctively. On the other hand, we consider another independent covariance structure with different variances given by

$$\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

whose structure is expressed as DIAG. The AIC is given by

$$\text{AIC}_{5j} = n \sum_{i=1}^{p} \log \hat{\sigma}_{ii}^2 + np(\log 2\pi + 1) + 2(jp + p),$$

where $\hat{\sigma}_{ii}^2 = \frac{1}{n} \boldsymbol{e}_i' \mathbf{W}_j \boldsymbol{e}_i$, $\mathbf{W}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}$, and $\boldsymbol{e}_i$ is the p component vector with i-th component 1 and other components zero. In our simulation experiment, the variances were defined by

$$\sigma_i^2 = \sigma^2 \left(1 + 3\frac{i-1}{p-1}\right), \quad i = 1, \ldots, p,$$

where $\sigma^2 = 2$. The simulation result is given in Table 7.2.

Table 7.2. Selection rates of AIC for the five covariance structures

	$n = 20$, $p = 10$				
TRUE	IND	UNIF	AUTO	NOST	DIAG
IND	0.40	0.12	0.11	0.28	0.09
UNIF	0.00	0.68	0.00	0.32	0.00
AUTO	0.00	0.00	0.68	0.32	0.00
NOST	0.01	0.15	0.01	0.59	0.23
DIAG	0.06	0.02	0.03	0.39	0.50

	$n = 200$, $p = 100$				
TRUE	IND	UNIF	AUTO	NOST	DIAG
IND	0.71	0.14	0.14	0.00	0.00
UNIF	0.00	1.00	0.00	0.00	0.00
AUTO	0.00	0.00	1.00	0.00	0.00
NOST	0.00	0.00	0.00	1.00	0.00
DIAG	0.00	0.00	0.00	0.00	1.00

It is seen that AIC is consistent for selection of the five covariance structures in high-dimensional situation.

Similar experiments were done for BIC. The result for selection of the five covariance structures is given in Table 7.3. It seems that BIC coverges more firstly to the true model than AIC, except for NOST. BIC chooses UNIF when the true is NOST. This will come from that our setting for NOST will be near UNIF.

Table 7.3 Selection rates of BIC for the four covariance strucures

	$n = 20$, $p = 10$				
TRUE	IND	UNIF	AUTO	NOST	DIAG
IND	0.75	0.12	0.11	0.00	0.02
UNIF	0.00	1.00	0.00	0.00	0.00
AUTO	0.00	0.00	1.00	0.00	0.00
NOST	0.08	0.52	0.06	0.03	0.32
DIAG	0.34	0.07	0.07	0.01	0.51

	$n = 200$, $p = 100$				
TRUE	IND	UNIF	AUTO	NOST	DIAG
IND	0.96	0.02	0.02	0.00	0.00
UNIF	0.00	1.00	0.00	0.00	0.00
AUTO	0.00	0.00	1.00	0.00	0.00
NOST	0.00	1.00	0.00	0.00	0.00
DIAG	0.00	0.00	0.00	0.00	1.00

8. Concluding Remarks

In this paper, firstly we consider to select regression variables in p variate regression model with one of three covariance structures; (1) IND(independent covariance structure), (2) UNIF(uniform covariance structure), (3) AUTO (autoregressive covariance structure). As a selection method, a general information $\mathrm{IC}_{g,d}$ was considered for each of three covariance structures, where d is a positive constant and may depend on the sample size n. When $d = 2$ and $\log n$, $\mathrm{IC}_{g,d}$ becomes to AIC and BIC, respectively. Under a high-dimensional asymptotic framework $p/n \to c \in (0, \infty)$, it was shown that $\mathrm{IC}_{g,d}$ with $g = v, u$ or a is consistent under the assumption of A3g if $d/n \to 0$ and $d > 1$. Further, in order to avoid a computational problem of $\mathrm{IC}_{g,d}$, we study $\mathrm{EC}_{g,d}$. It was pointed that $\mathrm{EC}_{g,d}$ has a consistency property similar to $\mathrm{IC}_{g,d}$. The result was obtained by assuming normality. It is left to extend the result to the case of non-normality.

Next, under multivariate regression model, we examined to select covariance structures. The covariance structures picked up are (4) NOST(no covariance structure) and (5) IND (independent covariance structure), in addition to the three covariance structures (1), (2) and (3). Two criteria AIC and BIC were examined through a simulation experiment in a high-dimensional setting, It was seen that (i) the four covariance structures except IND can be selected correctly by using AIC, (ii) BIC converges to the true model more firstly than AIC, except for IND. The proofs of theoretical results on these properties will be given in a future work.

Appendix: The Proofs of Theorems 3.1, 4.1 and 5.1

A1. Outline of Our Proofs and Preliminary Lemma

First we explain an outline of our proof. In general, let $\mathcal{F}$ be a finite set of candidate models j(or M_j). Assume that j_* is the minimum model including the true model and $j_* \in \mathcal{F}$. Let $\mathrm{T}_j(n)$ be a general criterion for model j, which depends on parameters p and n. The best model chosen by minimizing $\mathrm{T}_j(p, n)$ is written as $\hat{j}_{\mathrm{T}}(p, n) = \arg\min_{j \in \mathcal{F}} \mathrm{T}_j(p, n)$. Suppose that we are interested in asymptotic behavior of $\hat{j}_{\mathrm{T}}(p, n)$ when p/n tends to $c > 0$. In order to show a consistency of $\mathrm{T}_j(p, n)$, we may check a sufficient condition such that for any $j \neq j_* \in \mathcal{F}$, there exists a sequence $\{a_{p,n}\}$ with $a_{p,n} > 0$,

$$a_{p,n} \left\{ T_j(p, n) - T_{j_*}(p, n) \right\} \xrightarrow{p} b_j > 0.$$

In fact, the condition implies that for any $j \neq j_* \in \mathcal{F}$,

$$P(\hat{j}_{\mathrm{T}}(p, n) = j) \leq P(T_j(p, n) < T_{j_*}(p, n)) \to 0,$$

and

$$P(\hat{j}_{\mathrm{T}}(p, n) = j_*) = 1 - \sum_{j \neq j_* \in \mathcal{F}} P(\hat{j}_{\mathrm{T}}(p, n) = j) \to 1.$$

For the proofs of Theorems 3.1, 4.1 and 5.1, we use the following Lemma frequently.

Lemma A1. *Suppose that a $p \times p$ symmetric random matrix $\mathsf{\mathbf{W}}$ is distributed as a noncentaral Wishart distribution $\mathrm{W}_p(n-k, \mathbf{\Sigma}; \mathbf{\Omega})$. Let $\mathbf{A}$ be a given $p \times p$ symmetric matrix. We consider asymptotic behavior of $\mathrm{tr}\mathbf{A}\mathsf{\mathbf{W}}$ when p and n are large as in the way such that $p/n \to c \in (0, \infty)$, where k is fixed. Suppose that*

$$\lim \frac{1}{p}\mathrm{tr}\mathbf{A}\mathbf{\Sigma} = a^2 > 0, \quad \lim \frac{1}{np}\mathrm{tr}\mathbf{A}\mathbf{\Omega} = \eta^2 > 0. \tag{A.1}$$

Then, it holds that

$$T_{p,n} = \frac{1}{np}\mathrm{tr}\mathbf{A}\mathsf{\mathbf{W}} \xrightarrow{p} a^2 + \eta^2. \tag{A.2}$$

Proof. Let $m = n - k$. We may write $\mathsf{\mathbf{W}}$ as

$$\mathsf{\mathbf{W}} = \mathbf{\Sigma}^{1/2}(\boldsymbol{z}_1\boldsymbol{z}_1' + \cdots + \boldsymbol{z}_m\boldsymbol{z}_m')\mathbf{\Sigma}^{1/2},$$

where $\boldsymbol{z}_i \sim \mathrm{N}_p(\boldsymbol{\zeta}_i, \mathbf{I}_p)$, $i = 1, \ldots, m$ and $\boldsymbol{z}_i$'s are independent. Here, $\mathbf{\Omega} = \boldsymbol{\mu}_1\boldsymbol{\mu}_1' + \cdots + \boldsymbol{\mu}_m\boldsymbol{\mu}_m'$ and $\boldsymbol{\zeta}_i = \mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}_i$, $i = 1, \ldots, m$. Note that $\mathrm{tr}\mathbf{A}\mathsf{\mathbf{W}}$ is expressed as a quadratic form of $\boldsymbol{z} = (\boldsymbol{z}_1^{'}, \ldots, \boldsymbol{z}_m^{'})'$ as follows:

$$\mathrm{tr}\mathbf{A}\mathsf{\mathbf{W}} = \boldsymbol{z}'\mathbf{B}\boldsymbol{z},$$

where $\mathbf{B} = \mathbf{I}_m \otimes \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$. Note that $\boldsymbol{z} \sim \mathrm{N}_{mp}(\boldsymbol{\zeta}, \mathbf{I}_{mp})$, where $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_m')'$. Then, it is known (see, e.g., Gupta) that for any symmetric matrix $\mathbf{B}$,

$$\mathrm{E}[\boldsymbol{z}'\mathbf{B}\boldsymbol{z}] = \mathrm{tr}\mathbf{B} + \boldsymbol{\zeta}'\mathbf{B}\boldsymbol{\zeta},$$
$$\mathrm{Var}(\boldsymbol{z}'\mathbf{B}\boldsymbol{z}) = 2\mathrm{tr}\mathbf{B}^2 + 4\boldsymbol{\zeta}'\mathbf{B}^2\boldsymbol{\zeta}.$$

Especially, when $\mathbf{B} = \mathbf{I}_m \otimes \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$, we have

$$\mathrm{E}[\boldsymbol{z}'\mathbf{B}\boldsymbol{z}] = m\mathrm{tr}\mathbf{A}\mathbf{\Sigma} + \mathrm{tr}\mathbf{A}\mathbf{\Omega},$$
$$\mathrm{Var}(\boldsymbol{z}'\mathbf{B}\boldsymbol{z}) = 2m\mathrm{tr}\,(\mathbf{A}\mathbf{\Sigma})^2 + 4\mathrm{tr}\mathbf{A}\mathbf{\Sigma}\mathbf{A}\mathbf{\Omega}.$$

Under the assumption (A.1), we have

$$\begin{aligned}\mathrm{E}(T_{p,n}) &= \frac{m}{np}\mathrm{tr}\mathbf{A}\mathbf{\Sigma} + \frac{1}{np}\mathrm{tr}\mathbf{A}\mathbf{\Omega} \\ &\to a^2 + \eta^2.\end{aligned}$$

Further,

$$\begin{aligned}\mathrm{Var}(T_{p,n}) &= \frac{2m}{(np)^2}\mathrm{tr}(\mathbf{A\Sigma})^2 + \frac{4}{(np)^2}\mathrm{tr}\mathbf{A\Sigma A\Omega} \\ &\le \frac{2m}{(np)^2}(\mathrm{tr}\mathbf{A\Sigma})^2 + \frac{4}{(np)^2}\mathrm{tr}\mathbf{A\Sigma}\mathrm{tr}\mathbf{A\Omega} \\ &\to 0.\end{aligned}$$

These imply our conclution. □

In a special case $\mathbf{A} = \mathbf{I}_p$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, the assumptions in (A.1) become to

$$\begin{aligned}\lim \frac{1}{p}\mathrm{tr}\mathbf{A\Sigma} &= \lim \frac{1}{p}p\sigma^2 = \sigma^2, \\ \lim \frac{1}{np}\mathrm{tr}\mathbf{A\Omega} &= \lim \frac{1}{np}\mathrm{tr}\mathbf{\Omega} \xrightarrow{p} \eta^2.\end{aligned}$$

The conclusion is that

$$T_{p,n} = \frac{1}{np}\mathrm{tr}\mathsf{W} \xrightarrow{p} \sigma^2(1+\delta^2), \tag{A.3}$$

where $\delta^2 = (1/\sigma^2)\eta^2$. Note that $(1/\sigma^2)\mathrm{tr}\mathsf{W} \sim \chi^2_{(n-k)p}(\delta^2)$. This implies that

$$\frac{1}{\sigma^2}\mathrm{tr}\mathsf{W} \sim \chi^2_{(n-k)p}(\delta^2).$$

Therefore, under a high-dimensional asymptotic framework $p/n \to c \in (0,\infty)$,

$$\frac{1}{np}\chi^2_{(n-k)p}(\delta^2) \xrightarrow{p} 1+\eta^2, \tag{A.4}$$

if $(np)^{-1}\delta^2 \to \eta^2$. More generally, we use the following result.

$$\frac{1}{np}\chi^2_{(n-k)(p-h)}(\delta^2) \xrightarrow{p} 1+\eta^2, \tag{A.5}$$

if $(np)^{-1}\delta^2 \to \eta^2$, where k and h are constants or more generally, they may be the constants satisfying $k/n \to 0$ and $h/n \to 0$. Further, we use the following property for noncentral χ^2-square distribution.

$$\frac{1}{n}\chi^2_{n-k}(\delta^2) \xrightarrow{p} 1+\eta^2, \tag{A.6}$$

if $n^{-1}\delta^2 \to \eta^2$.

A2. The Proof of Theorem 3.1

Using (3.3) we can write

$$\mathrm{IC}_{v,d,j} - \mathrm{IC}_{v,d,j_*} = np\log\hat{\sigma}_{v,j}^2 - np\log\hat{\sigma}_{v,j_*}^2 + d(k_j - k_{j_*})p.$$

Lemma 3.1 shows that

$$\frac{np}{\sigma^2}\hat{\sigma}_j^2 \sim \chi^2_{(n-k_j)p}(\delta_j^2), \quad \delta_j^2 = \frac{1}{\sigma^2}\mathrm{tr}(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*}).$$

In particular,

$$\frac{np}{\sigma^2}\hat{\sigma}_{j_*}^2 \sim \chi^2_{(n-k_{j_*})p}.$$

First, consider the case $j \supset j_*$. Under the assumption A3v,

$$\begin{aligned}\frac{1}{np}\frac{np}{\sigma^2}\hat{\sigma}_j^2 &= \frac{(n-k_j)p}{np}\frac{1}{(n-k_j)p}\frac{1}{\sigma^2}\hat{\sigma}_j^2 \xrightarrow{p} 1+\eta_j^2,\\ \frac{1}{np}\frac{np}{\sigma^2}\hat{\sigma}_{j_*}^2 &= \frac{(n-k_{j_*})p}{np}\frac{1}{(n-k_{j_*})p}\frac{1}{\sigma^2}\hat{\sigma}_{j_*}^2 \xrightarrow{p} 1.\end{aligned}$$

Therefor

$$\begin{aligned}\frac{1}{np}\left(\mathrm{IC}_{v,g,j} - \mathrm{IC}_{v,g,j_*}\right) &= \log\hat{\sigma}_j^2 - \log\hat{\sigma}_{j_*}^2 + \frac{d}{n}(k_j - k_{j_*})\\ &= \log\left(\frac{1}{np}\frac{np}{\sigma^2}\hat{\sigma}_j^2\right) - \log\left(\frac{1}{np}\frac{np}{\sigma^2}\hat{\sigma}_{j_*}^2\right) + \frac{d}{n}(k_j - k_{j_*})\\ &\xrightarrow{p} \log(1+\eta_j^2) + \log 1 + 0 = \log(1+\eta_j^2) > 0,\end{aligned}$$

when $g/n \to 0$.

Next, consider the case $j \supset j_*$. Then

$$\mathrm{IC}_{v,g,j} - \mathrm{IC}_{v,g,j_*} = np\log\frac{\hat{\sigma}_{v,j}^2}{\hat{\sigma}_{v,j_*}^2} + d(k_j - k_{j_*})p.$$

Further,

$$\begin{aligned}\log\frac{\hat{\sigma}_j^2}{\hat{\sigma}_{j_*}^2} &= \log\frac{\mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}}{\mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{j_*})\mathbf{Y}}\\ &= -\log\left(1 + \frac{\mathrm{tr}\mathbf{Y}'(\mathbf{P}_j - \mathbf{P}_{j_*})\mathbf{Y}}{\mathrm{tr}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}}\right)\\ &= -\log\left(1 + \frac{\chi^2_{(k_j-k_{j_*})p}}{\chi^2_{(n-k_j)p}}\right).\end{aligned}$$

Using the fact that $\chi^2_m/m \to 1$ as $m \to \infty$, we have

$$n \log \frac{\hat{\sigma}_j^2}{\hat{\sigma}_{j_*}^2} = -n \log \left\{ 1 + \frac{(k_j - k_{j_*})}{(n - k_j)} \frac{\chi^2_{(k_j - k_{j_*})p}/((k_j - k_{j_*})p)}{\chi^2_{(n-k_j)p}/((n-k_j)p)} \right\} \to -1,$$

and hence

$$\begin{aligned} \frac{1}{p}\{\mathrm{IC}_{v,d,j} - \mathrm{IC}_{v,d,j_*}\} &= n \log \frac{\hat{\sigma}_j^2}{\hat{\sigma}_{j_*}} + 2(k_j - k_{j_*}) \\ &\xrightarrow{p} -(k_j - k_{j_*}) + d(k_j - k_{j_*}) = (d-1)(k_j - k_{j_*}) > 0, \end{aligned}$$

if $d > 1$.

A3. The Proof of Theorem 4.1

Using (4.3) we can write

$$\begin{aligned} \mathrm{IC}_{u,d,j} - \mathrm{IC}_{u,d,j_*} &= n(p-1)(\log \hat{\tau}_{1j}^2 - \log \hat{\tau}_{1j_*}^2) \\ &\quad + n(\log \hat{\tau}_{2j}^2 - \log \hat{\tau}_{2j_*}^2) + d(k_j - k_{j_*})p. \end{aligned}$$

Lemma 4.1 shows that for a general $j \subset \omega$,

$$np\tau_{1*}^{-1}\hat{\tau}_{1j} \sim \chi^2_{(p-1)(n-k_j)}(\delta_{1j}^2), \qquad \text{(A.7)}$$

$$n\tau_{2*}^{-1}\hat{\tau}_{2j} \sim \chi^2_{n-k_j}(\delta_{2j}^2), \qquad \text{(A.8)}$$

where

$$\begin{aligned} \delta_{1j}^2 &= \frac{1}{\tau_{1*}} \mathrm{tr} \mathbf{H}_2'(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})\mathbf{H}_2, \\ \delta_{2j}^2 &= \frac{1}{\tau_{2*}} \mathrm{tr} \boldsymbol{h}_1'(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})'(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{X}_{j_*}\mathbf{\Theta}_{j_*})\boldsymbol{h}_1. \end{aligned}$$

Note thta if $j \supset j_*$, $\delta_{1j} = 0$ and $\delta_{2j} = 0$.

First consider the case $j \subset j_*$. Then, using (A.5) and (A.7), we have

$$\frac{\hat{\tau}_{1j}}{\hat{\tau}_{1j_*}} = \frac{\chi^2_{(n-k_j)p}(\delta_{1j}^2)}{\chi^2_{(n-k_{j_*})p}} \xrightarrow{p} 1 + \eta^2.$$

Similarly, using (A.6) and (A.8), we have

$$\frac{\hat{\tau}_{2j}}{\hat{\tau}_{2j_*}} = \frac{\chi^2_{(n-k_j)p}(\delta^2_{2j})}{\chi^2_{(n-k_{j_*})p}} \xrightarrow{p} 1+\eta^2.$$

These results imply the following:

$$\begin{aligned}
\frac{1}{np}(\mathrm{IC}_{u,d,j} - \mathrm{IC}_{u,d,j_*}) &= \frac{n(p-1)}{np}\log(\hat{\tau}_{1j}/\hat{\tau}_{1j_*}) \\
&\quad + \frac{1}{p}\log(\hat{\tau}_{2j}/\hat{\tau}_{2j_*}) + \frac{d}{n}(k_j - k_{j_*}) \\
&\quad \to \log(1+\eta^2_{1j}) > 0
\end{aligned}$$

if $d/n \to 0$.

Next, consider the case $j \supset j_*$. Then

$$\begin{aligned}
\log\frac{\hat{\tau}_{1j}}{\hat{\tau}_{1j_*}} &= \log\frac{\mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\mathbf{H}_2}{\mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{j_*})\mathbf{Y}\mathbf{H}_2'} \\
&= -\log\left(1 + \frac{\mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{P}_j - \mathbf{P}_{j_*})\mathbf{Y}\mathbf{H}_2}{\mathrm{tr}\mathbf{H}_2'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\mathbf{H}_2}\right) \\
&= -\log\left(1 + \frac{\chi^2_{(k_j-k_{j_*})(p-1)}}{\chi^2_{(n-k_j)(p-1)}}\right) \sim -\frac{k_j - k_{j_*}}{n-k_j}.
\end{aligned}$$

Similarly

$$\begin{aligned}
\log\frac{\hat{\tau}_{2j}}{\hat{\tau}_{2j_*}} &= \log\frac{\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\boldsymbol{h}_1}{v\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{j_*})\mathbf{Y}\boldsymbol{h}_1'} \\
&= -\log\left(1 + \frac{\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{P}_j - \mathbf{P}_{j_*})\mathbf{Y}\boldsymbol{h}_1}{\boldsymbol{h}_1'\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}\boldsymbol{h}_1}\right) \\
&= -\log\left(1 + \frac{\chi^2_{k_j-k_{j_*}}}{\chi^2_{n-k_j}}\right) \sim -\frac{k_j - k_{j_*}}{n-k_j}.
\end{aligned}$$

These results imply the following:

$$\begin{aligned}
\frac{1}{p}(\mathrm{IC}_{u,d,j} - \mathrm{IC}_{u,d,j_*}) &\xrightarrow{p} -(k_j - k_{j_*}) + d(k_j - k_{j_*}) \\
&= (d-1)(k_j - k_{j_*}) > 0,
\end{aligned}$$

if $d > 1$. Combining these to the result in the case $j \subset j_*$, we get Therem 4.1.

A4. The Proof of Theorem 5.1

In this section, for notational simplification, we put $\mathbf{\Sigma}_{a,*}$, $\sigma^2_{a,*}$ and $\rho_{a,*}$ into $\mathbf{\Sigma}$, σ^2 and ρ, respectively. Using (5.5) we can write

$$\begin{aligned}\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*} &= np\log(\hat{\sigma}^2_{a,j}/\hat{\sigma}^2_{a,j_*}) \\ &\quad + n(p-1)\log\left\{(1-\hat{\rho}^2_{a,j})/(1-\hat{\rho}^2_{a,j_*})\right\} + d(k_j - k_{j_*})p.\end{aligned}$$

Further, using (5.3), the above expression can be expressed as

$$\begin{aligned}\mathrm{IC}_{a,d,j}{-}\mathrm{IC}_{a,d,j_*} &= np\log\left\{(n-k_j)/(n-k_{j_*})\right\} - n\log\left\{(1-\hat{\rho}^2_j)/(1-\hat{\rho}^2_{j_*})\right\} \\ &\quad + np\log\left\{(a_{1j}\hat{\rho}^2_j - 2a_{2j}\hat{\rho}_j + a_{0j})/(a_{1j_*}\hat{\rho}^2_{j_*} - 2a_{2j_*}\hat{\rho}_{j_*} + a_{0j_*})\right\} \\ &\quad + d(k_j - k_{j_*})p. \qquad \text{(A.9)}\end{aligned}$$

Here, the maximum likelihood estimators $\hat{\sigma}^2_j$ and $\hat{\rho}_j$ are defined in terms of

$$\begin{aligned}&a_{ij} = \mathrm{tr}\mathbf{C}_i\mathbf{S}_j, \quad i = 0, 1, 2, \\ &(n-k_j)\mathbf{S}_j = \mathbf{W}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} \sim \mathrm{W}_p(n-j, \mathbf{\Sigma}; \mathbf{\Omega}_j).\end{aligned}$$

For the definition of $\mathbf{C}_i$ and $\mathbf{\Omega}_j$, see Section 5. We also use the notation

$$\begin{aligned}&b_{ij} = \mathrm{tr}\mathbf{C}_i\mathbf{W}_j, \quad i = 0, 1, 2, \\ &\mathbf{W}_j = (n-k_j)\mathbf{S}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} \sim \mathrm{W}_p(n-k_j, \mathbf{\Sigma}; \mathbf{\Omega}_j).\end{aligned}$$

First, consider the case $j \subset j_*$. Under the assumption A3a, it is possible to get asymptotic behavior of a_{ij} and b_{ij} which is given in lator. It is shown that

$$\hat{\rho}_j = \frac{a_{2j}}{a_{1j}} + O_p(p^{-1}),$$

and more precisely

$$\hat{\rho}_j = \frac{a_{2j}}{a_{1j}} + \frac{1}{p}\left\{-\frac{a_2(a_0a_1 - a_2^2)}{a_1(a_1-a_2)(a_1+a_2)}\right\} + O_p(p^{-2}).$$

These imply that

$$\begin{aligned}&\log(1-\hat{\rho}^2_j) = \log\left(1 - \frac{a^2_{2j}}{a^2_{1j}}\right) + O_p(p^{-1}), \\ &\log(a_{1j}\hat{\rho}^2_j - 2a_{2j}\hat{\rho}_j + a_{0j}) = \log\left(a_{0j} - \frac{a^2_{2j}}{a_{1j}}\right) + O_p(p^{-2}).\end{aligned}$$

The above expansions hold also for $j = j_*$. Substituting these results to (A.9) and noting $b_{ij} = (n-k_j)a_{ij}$, we have

$$\begin{aligned}
&\frac{1}{np}(\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*}) \\
&= -\frac{1}{p}\left\{\log\left(1-\frac{b_{2j}^2}{b_{1j}^2}\right) - \log\left(1-\frac{b_{2j_*}^2}{b_{1j_*}^2}\right)\right\} \qquad \text{(A.10)} \\
&\quad + \left\{\log\left(b_{0j}-\frac{b_{2j}^2}{b_{1j}}\right) - \log\left(b_{0j_*}-\frac{b_{2j_*}^2}{b_{1j_*}}\right)\right\} + \frac{d}{n}(k_j - k_{j_*}) + O_p(p^{-2}).
\end{aligned}$$

Using Lemma A1, under the assumption A3a we have

$$\begin{aligned}
&\frac{1}{np}b_{0j} = \frac{n-k_j}{n}\frac{1}{(n-k_j)p}b_{0j} \xrightarrow{p} \sigma^2+\eta_{0j}^2, \quad \frac{1}{p}\mathrm{tr}\mathbf{C}_0\mathbf{\Sigma} = \frac{p}{p}\sigma^2 \to \sigma^2, \\
&\frac{1}{np}b_{1j} = \frac{n-k_j}{n}\frac{1}{(n-k_j)p}b_{1j} \xrightarrow{p} \sigma^2+\eta_{1j}^2, \quad \frac{1}{p}\mathrm{tr}\mathbf{C}_1\mathbf{\Sigma} = \frac{p-2}{p}\sigma^2 \to \sigma^2, \\
&\frac{1}{np}b_{2j} = \frac{n-k_j}{n}\frac{1}{(n-k_j)p}b_{2j} \xrightarrow{p} \sigma^2\rho+\eta_{2j}^2, \quad \frac{1}{p}\mathrm{tr}\mathbf{C}_2\mathbf{\Sigma} = \frac{p-1}{p}\sigma^2\rho \to \sigma^2\rho.
\end{aligned}$$

Asymptotic behaviors for b_{ij_*} are obtained from the ones by putting $\eta_{ij_*} = 0$. These imply that

$$\begin{aligned}
&\frac{1}{np}(\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*}) \\
&\quad \xrightarrow{p} \log\left(\sigma^2+\eta_{0j}^2 - \frac{(\sigma^2\rho+\eta_{2j}^2)^2}{\sigma^2+\eta_{1j}^2}\right) - \log\sigma^2(1-\rho^2),
\end{aligned}$$

if $d/n \to 0$. Now we show

$$\log\left(\sigma^2+\eta_{0j}^2 - \frac{(\sigma^2\rho+\eta_{2j}^2)^2}{\sigma^2+\eta_{1j}^2}\right) - \log\sigma^2(1-\rho^2) > 0. \qquad \text{(A.11)}$$

Note that

$$\begin{aligned}
&\frac{1}{np}(\mathrm{tr}\mathbf{C}_0\mathbf{\Omega}_j - \mathrm{tr}\mathbf{C}_1\mathbf{\Omega}_j) \\
&= \frac{1}{np}\mathrm{tr}(\mathbf{C}_0 - \mathbf{C}_1)\mathbf{\Omega}_j \\
&= \frac{1}{np}(\text{the sum of } (1,1) \text{ and } (p,p) \text{ elements of } \mathbf{\Omega}_j) \to 0,
\end{aligned}$$

which implies $\eta_{0j}^2 = \eta_{1j}^2$. Using this equality, we can express

$$\begin{aligned}
&\log\left(\sigma^2 + \eta_{0j}^2 - \frac{(\sigma^2\rho + \eta_{2j}^2)^2}{\sigma^2 + \eta_{1j}^2}\right) - \log\sigma^2(1-\rho^2) \\
&= \log\left(1 + \frac{\eta_{0j}^2}{\sigma^2}\right) + \log\left(1 + \frac{\eta_{0j}^2 - \eta_{2j}^2}{\sigma^2(1-\rho)}\right) + \log\left(1 + \frac{\eta_{0j}^2 + \eta_{2j}^2}{\sigma^2(1+\rho)}\right).
\end{aligned}$$

The first and the third terms are positive. The quantity $\eta_{0j}^2 - \eta_{2j}^2$ in the second term can be expressed as the limit of

$$\frac{1}{np}(\mathrm{tr}\mathbf{C}_0\boldsymbol{\Omega}_j - \mathrm{tr}\mathbf{C}_2\boldsymbol{\Omega}_j) = \frac{1}{np}\mathrm{tr}(\mathbf{C}_0 - \mathbf{C}_2)\boldsymbol{\Omega}_j.$$

Here, the matrix $\mathbf{C}_0 - \mathbf{C}_2$ is expressed as

$$\mathbf{C}_0 - \mathbf{C}_2 = \begin{pmatrix} 1 & -\frac{1}{2} & & 0 & 0 \\ -\frac{1}{2} & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & -\frac{1}{2} \\ 0 & 0 & & -\frac{1}{2} & 1 \end{pmatrix},$$

and

$$\begin{aligned}
&\boldsymbol{x}'(\mathbf{C}_0 - \mathbf{C}_2)\boldsymbol{x} \\
&= x_1^2 - x_1x_2 + x_2^2 - x_2x_3 + x_3^2 + \cdots + x_{p-1}^2 - x_{p-1}x_p + x_p^2 \\
&= \frac{1}{2}\left\{x_1^2 + (x_1 - x_2)^2 + \cdots + (x_{p-1} - x_p)^2 + x_p^2\right\} > 0,
\end{aligned}$$

except for the case $x_1 = \cdots = x_p$. Therefore, the second term is nonnegative, and $(\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*})/(np)$ converges to a positive constant if $d/n \to 0$.

Next, consider the case $j \supset j_*$. From (A.10) we have

$$\begin{aligned}
&\frac{1}{p}\left(\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*}\right) \\
&= -\frac{n}{p}\left\{\log\left(1 - \frac{b_{2j}^2}{b_{1j}^2}\right) - \log\left(1 - \frac{b_{2j_*}^2}{b_{1j_*}^2}\right)\right\} \\
&\quad + n\left\{\log\left(b_{0j} - \frac{b_{2j}^2}{b_{1j}}\right) - \log\left(b_{0j_*} - \frac{b_{2j_*}^2}{b_{1j_*}}\right)\right\} + d(k_j - k_{j_*}) + O_p(p^{-1}).
\end{aligned}$$

Then, it is easely seen that

$$\begin{aligned}&\frac{n}{p}\left\{\log\left(1-\frac{b_{2j}^2}{b_{1j}^2}\right)-\log\left(1-\frac{b_{2j_*}^2}{b_{1j_*}^2}\right)\right\}\\ &\xrightarrow{p}\frac{1}{c}\left\{\log(1-\rho^2)-\log(1-\rho^2)\right\}=0.\end{aligned}$$

Note that

$$\begin{aligned}&n\left\{\log\left(b_{0j}-\frac{b_{2j}^2}{b_{1j}}\right)-\log\left(b_{0j_*}-\frac{b_{2j_*}^2}{b_{1j_*}}\right)\right\}\\ &=n\left\{-(\log b_{1j}-\log b_{1j_*})+\log(b_{0j}b_{1j}-b_{2j}^2)-\log(b_{0j_*}b_{1j_*}-b_{2j_*}^2)\right\}.\end{aligned}$$

Further, we use the following relation between b_{ij} and b_{ij_*}:

$$\begin{aligned}b_{ij}&=\mathrm{tr}\mathbf{C}_i\mathbf{Y}'(\mathbf{I}_n-\mathbf{P}_j)\mathbf{Y}\\ &=\mathrm{tr}\mathbf{C}_i\mathbf{Y}'(\mathbf{I}_n-\mathbf{P}_{j_*})\mathbf{Y}-\mathrm{tr}\mathbf{C}_i\mathbf{Y}'(\mathbf{P}_j-\mathbf{P}_{j_*})\mathbf{Y}\\ &=b_{ij_*}-b_{ij_+},\end{aligned}$$

where $b_{ij_+}=\mathrm{tr}\mathbf{C}_i\mathbf{Y}'(\mathbf{P}_j-\mathbf{P}_{j_*})\mathbf{Y}$. Since $\mathbf{Y}'(\mathbf{P}_j-\mathbf{P}_{j_*})\mathbf{Y}\sim \mathrm{W}_p(k_j-k_{j_*},\mathbf{\Sigma})$, it holds that

$$\frac{1}{(k_j-k_{j_*})p}b_{0j_+}\xrightarrow{p}\sigma^2,\quad \frac{1}{(k_j-k_{j_*})p}b_{1j_+}\xrightarrow{p}\sigma^2,\quad \frac{1}{(k_j-k_{j_*})p}b_{2j_+}\xrightarrow{p}\sigma^2\rho.$$

Therefore, noting that $b_{ij_*}=O_p(np)$ and $b_{ij_+}=O_p(p)$. we have

$$n(\log b_{1j}-\log b_{1j_*})=n\log\left(1-b_{1j_+}b_{1j_*}^{-1}\right)\sim -nb_{1j_+}b_{1j_*}^{-1},$$

and hence

$$n(\log b_{1j}-\log b_{1j_*})\xrightarrow{p}-(k_j-k_{j_*}).$$

Consider

$$f=n\{\log(b_{0j}b_{1j}-b_{2j}^2)-\log(b_{0j_*}b_{1j_*}-b_{2j_*}^2)\}.$$

Sustituting $b_{ij}=b_{ij_*}-b_{ij_+}$ to the above expression, we have

$$f=n\log(1-h)\sim -nh,$$

where

$$h = -\frac{(b_{0j_*} - b_{0j_+})}{(b_{1j_*} - b_{1j_+})}(b_{0j_*}b_{1j_+} + b_{1j_*}b_{0j_+} - 2b_{2j_*}b_{2j_+} - b_{0j_+}b_{0j_+} + b_{2j_+}^2)$$

Considering the limiting values of each terms in h, we have

$$\begin{aligned} f &\xrightarrow{p} -\frac{(k_j - k_{j_*})}{(1-\rho^2)} - \frac{(k_j - k_{j_*})}{(1-\rho^2)} + 2\frac{(k_j - k_{j_*})\rho^2}{(1-\rho^2)} + 0 + 0 \\ &= -2(k_j - k_{j_*}). \end{aligned}$$

Summerizing the above results, for $j \supset j_*$,

$$\begin{aligned} &\frac{1}{p}(\mathrm{IC}_{a,d,j} - \mathrm{IC}_{a,d,j_*}) \\ &\xrightarrow{p} (k_j - k_{j_*}) - 2(k_j - k_{j_*}) + d(k_j - k_{j_*}) = (d-1)(k_j - k_{j_*}) > 0, \end{aligned}$$

if $d > 1$. This completes the proof.

A5. The Proof of Theorem 6.1

For a notational simplicity, we denote $\mathrm{IC}_{g,d,j}$ by $\mathrm{IC}(j)$. Note that $j_\omega = \{1, 2, \ldots, k\}$ and k is finite. Without loss of generality, we may assume that the true model is $j_* = \{1, \ldots, b\}$ and $b = k_{j_*}$. Under the assumption A1, A2 and A3·1∼3, it was shown that our information criterion $\hat{j}_{\mathrm{IC}}$ has a high-dimensiona consistency property. The consitency property was shown by proving that
(1) for $j \in \mathcal{F}_-$,

$$\frac{1}{m_1}\{\mathrm{IC}(j) - \mathrm{IC}(j_*)\} \xrightarrow{p} \gamma_j > 0, \tag{A.12}$$

and
(2) for $j \in \mathcal{F}_+$ and $j \neq j_*$,

$$\frac{1}{m_2}\{\mathrm{IC}(j) - \mathrm{IC}(j_*)\} \xrightarrow{p} (d-1)(k_j - k_{j_*}) > 0, \tag{A.13}$$

where $d > 1$, $m_1 = np$ and $m_2 = p$.

From the definition of our efficient criterion $\hat{j}_{\mathrm{EC}}$, we have

$$\begin{aligned}
&P(\hat{j}_{\mathrm{ED}} = j_*)\\
&= P(\mathrm{IC}(j_{(1)}) - \mathrm{IC}(j_\omega) > 0,\ \ldots,\ \mathrm{IC}(j_{(b)}) - \mathrm{IC}(j_\omega) > 0,\\
&\qquad \mathrm{IC}(j_{(b+1)} - \mathrm{IC}(j_\omega) < 0,\quad \ldots,\quad \mathrm{IC}(j_{(k)}) - \mathrm{IC}(j_\omega) < 0)\\
&= P\left(\bigcap_{i=1}^{b}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega)) > 0) \cap \bigcap_{i=b+1}^{k}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega)) < 0)\right),
\end{aligned}$$

which can be expressed as

$$\begin{aligned}
&1 - P\left(\bigcup_{i=1}^{b}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0) \cup \bigcup_{i=b+1}^{k}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0)\right)\\
&\geq 1 - \sum_{i=1}^{b} P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0) - \sum_{j=b+1}^{k} P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0)\\
&= 1 - \sum_{i=1}^{b}\{1 - P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0)\}\\
&\quad - \sum_{j=b+1}^{k}\{1 - P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0)\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
P(\hat{j}_{\mathrm{ED}} = j_*) \geq & 1 - \sum_{i \in j_*}\{1 - P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0)\}\\
& - \sum_{i \in j_\omega / j_*}\{1 - P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0)\}. \qquad \text{(A.14)}
\end{aligned}$$

Now, consider to evaluate the following probabilities:

$$\begin{aligned}
i \in j_*,&\quad P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0),\\
i \in j_\omega / j_*,&\quad P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0).
\end{aligned}$$

When $i \in j_*$, $j_{(i)} \in \mathcal{F}_-$, and hence, using (A.12), it holds that

$$\begin{aligned}
\frac{1}{m_1}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega)) &= \frac{1}{m_1}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_*)) - \frac{1}{m_1}(\mathrm{IC}(j_\omega) - \mathrm{IC}(j_*))\\
&\xrightarrow{p} \gamma_i + 0 = \gamma_i > 0.
\end{aligned}$$

When $i \in j_\omega/j_*$, $j_{(i)} \in \mathcal{F}_+$ and hence, using (A.13), it holds that

$$\begin{aligned}\frac{1}{m_2}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega)) &= \frac{1}{m_2}(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_*)) - \frac{1}{m_2}(\mathrm{IC}(j_\omega) - \mathrm{IC}(j_*)) \\ &\xrightarrow{p} (d-1)(k-1-b) - (d-1)(k-b) = -(d-1) < 0.\end{aligned}$$

These results imply that

$$\begin{aligned} i \in j_*, &\quad \lim P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) > 0) = 1, \\ i \in j_\omega/j_*, &\quad \lim P(\mathrm{IC}(j_{(i)}) - \mathrm{IC}(j_\omega) < 0) = 1. \end{aligned}$$

Using the above results, we can see that the right-hand side of (A.14) tends to

$$1 - \left[\sum_{j \in j_*}\{1-1\} + \sum_{j \in j_k/j_*}\{1-1\}\right] = 1,$$

and $P(\hat{j}_{\mathrm{ED}} = j_*) \to 1$. This completes the proof of Theorem 6.1.

Acknowledgements

The second author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), 16K00047, 2016-2018.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd. International Symposium on Information Theory* (eds. B. N. Petrov and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Bedrick, E. J. and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

[3] Fujikoshi, Y., Kanda, T. and Tanimura, N. (1990). The growth curve model with an autoregressive covariance structure. *Ann. Inst. Statist. Math.*, **42**, 533–542.

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

[5] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **149**, 199–212.

[6] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley, Hobeken, N.J.

[7] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

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

[9] Rothman, A., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph.. Stat.*, **19**, 947–962.

[10] 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.

[11] Zhao, L. C., Krishnaia, P. R. and Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.