High-Dimensional AIC and Consistency Properties of Several Criteria in Multivariate Linear Regression

Yasunori Fujikoshi*, Tetsuro Sakurai** and Hirokazu Yanagihara*

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

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

Abstract

The AIC and C_p and their modifications have been proposed for multivariate linear regression models under a large-sample framework when the sample size n is large, but the dimension p is fixed. In this paper, first we propose a high-dimensional AIC (denoted by HAIC) which is approximately unbiased estimator of the risk under a high-dimensional framework such that $p/n \to c \in (0, 1)$. It is noted that our new criterion do work in a wide range of p and n. Recently Yanagihara, Wakaki and Fujikoshi (2012) noted that AIC has a consistency property under some assumption on a noncentrality matrix when $p/n \to c \in [0, 1)$. In this paper we show that several criteria including HAIC and C_p have also a consistency property under a different assumption from the previous work on the noncentrality matrix when $p/n \to c \in (0, 1)$. Our results are checked numerically by conducting a Mote Carlo simulation.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: AIC, C_p, Consistency property, High-dimensional criteria, Modified criteria, 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 $\mathbf{Y} : n \times p$ 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

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

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 $\mathbf{Y}$ is $\mathbf{X}\boldsymbol{\Theta}$ and the covariance matrix of $\mathrm{vec}\,\mathbf{Y}$ is $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$, i.e., the rows of $\mathbf{Y}$ are independently normal with the same covariance matrix $\boldsymbol{\Sigma}$. We assume that $n - p - k - 1 > 0$, and $\mathrm{rank}(\mathbf{X}) = k$.

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 $\boldsymbol{x}$. A generic candidate model can be expressed in terms of a subset j of the set $\omega = \{1, \ldots, k\}$ of integers and the matrix $\mathbf{X}_j$ consisting of the columns of $\mathbf{X}$ indexed by the k_j integers in j. The candidate model is expressed as

$$M_j : \ \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \mathbf{I}_n), \tag{1.2}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown covariance matrix of the model j.

The AIC (Akaike, 1973) and C_p (Mallows, 1973) for M_j are given by

$$\mathrm{AIC} = n \log |\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + 2\left\{k_j p + \frac{1}{2}p(p+1)\right\}, \tag{1.3}$$

$$\mathrm{C}_p = (n-k)\mathrm{tr}\hat{\boldsymbol{\Sigma}}_\omega^{-1}\hat{\boldsymbol{\Sigma}}_j + 2pk_j, \tag{1.4}$$

where $n\hat{\boldsymbol{\Sigma}}_j = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}$ and $\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$. Note that $\hat{\boldsymbol{\Sigma}}_\omega$ and $\mathbf{P}_\omega$ are defined from $\hat{\boldsymbol{\Sigma}}_j$ and $\mathbf{P}_j$ as $j = \omega$, and $k_\omega = k$, $\mathbf{X}_\omega = \mathbf{X}$. In addition to these criteria, there are several modifications such as CAIC, MAIC, C_p and MC_p (Bedrick and Tsai, 1994; Fujikoshi and Satoh, 1997; see Sections 2 and 3) which were proposed as approximately unbiased estimators of AIC-type and C_p-type risks, based on a large-sample theory. The modifications were studied by assuming that the true model is included into the full model M_ω, and the order of a standardized noncentrality matrix $\boldsymbol{\Omega}_j = \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ is $\mathrm{O}(n^{-1})$, where $\boldsymbol{\Gamma}_j$ is a $r_j \times p$ matrix and $r_j = k - k_j$.

In general, the approximations based on a large-sample framework become inaccurate as the dimension p increases while the sample size n remains fixed. On the other hand, in last year we encounter more and more problems in applications when p is comparable with n or even exceeds it. So, it is important to examine behaviors of these criteria when the dimension is large, for example, a high-dimensional framework such that

$$p/n \to \ c \in (0, 1) \tag{1.5}$$

In this paper we first derive a high-dimensional AIC denoted by HAIC which is an asymptotic unbiased estimator of AIC-type risk under (1.5). It is noted that HAIC includes AIC, CAIC and MAIC since they are obtained from HAIC by considering large-sample asymptotic. Next we show consistency properties of these criteria and C_p, MC_p. Recently Yanagihara, Wakaki and Fujikoshi (2012) pointed out that AIC and CAIC have a consistency property under (1.5) when the order of noncentrality matrix $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is assumed to be $\mathrm{O}(pn)$. In this paper different consistency properties are derived for HAIC, C_p, MC_p and also AIC, CAIC under (1.5), when the order of noncentrality matrix $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is assumed to be $\mathrm{O}(n)$. Our results are also checked numerically by conducting a Mote Carlo simulation.

2. High-Dimensional AIC

As is well known, the AIC was proposed as an approximately unbiased estimator of the expected log-predictive likelihood. Let $f(\mathbf{Y}; \mathbf{\Theta}_j, \mathbf{\Sigma}_j)$ be the density function of $\mathbf{Y}$ under M_j. Then the expected log-predictive likelihood of M_j is defined by

$$R_{\mathrm{A}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_{\mathrm{F}}}[-2\log f(\mathbf{Y}_{\mathrm{F}}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j)], \tag{2.1}$$

where $\hat{\mathbf{\Sigma}}_j$ and $\hat{\mathbf{\Theta}}_j$ are the maximum likelihood estimators of $\mathbf{\Sigma}_j$ and $\mathbf{\Theta}_j$ under M_j, respectively. Here $\mathbf{Y}_{\mathrm{F}} : n \times p$ may be regarded as a future random matrix that has the same distribution as $\mathbf{Y}$ and is independent of $\mathbf{Y}$. Furthermore, E denotes the expectations with respect to the true model. The risk is expressed as

$$R_{\mathrm{A}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_{\mathrm{F}}}[-2\log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j)] + b_{\mathrm{A}}, \tag{2.2}$$

where

$$b_{\mathrm{A}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_{\mathrm{F}}}[-2\log f(\mathbf{Y}_{\mathrm{F}}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j) + 2\log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j)]. \tag{2.3}$$

The AIC and its modifications have been proposed by regarding b_{A} as the bias term when we estimate R_{A} by the $-2\times$(maximum likelihood of the model j) as

$$-2\log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j) = n\log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1),$$

and by evaluating the bias term b_{A}. Although there are many bias-corrected AICs, in this paper we take up two modifications CAIC (Bedrick and Tsai, 1994) and MAIC (Fujikoshi and Satoh, 1997). These modifications are expressed as

$$\mathrm{CAIC} = \mathrm{AIC} + \frac{2(k_j + p + 1)}{n - k_j - p - 1}\left\{k_j p + \frac{1}{2}p(p+1)\right\}, \tag{2.4}$$

$$\mathrm{MAIC} = \mathrm{CAIC} + 2k_j \mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p) - \{\mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p)\}^2 - \mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p)^2, \tag{2.5}$$

where $\mathbf{L}_j$ is defined by

$$\mathbf{L}_j = \frac{n - k_j}{n - k}\hat{\mathbf{\Sigma}}_\omega \hat{\mathbf{\Sigma}}_j^{-1}.$$

For a justification of these criteria, it was assumed that the true model is included in the full model M_ω. We also assume it in this paper. Let M_{j_0} be the smallest model including the true model, i.e.,

$$M_{j_0} : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}_{j_0}\boldsymbol{\Theta}_{j_0}, \boldsymbol{\Sigma}_{j_0} \otimes \mathbf{I}_n), \tag{2.6}$$

where $\mathbf{X}_{j_0}$ is an $n \times k_{j_0}$ matrix consisting of some columns of $\mathbf{X}$, and $\boldsymbol{\Theta}_{j_0}$ is a $k_{j_0} \times p$ unknown matrix of regression coefficients and $\boldsymbol{\Sigma}_{j_0}$ is a $p \times p$ unknown covariance matrix of the true model. Then, the true model is defined as the model M_{j_0} with given $\boldsymbol{\Theta}_{j_0}$ and $\boldsymbol{\Sigma}_{j_0}$. For simplicity, we write k_{j_0}, M_{j_0}, $\boldsymbol{\Theta}_{j_0}$, $\mathbf{X}_{j_0}$ and $\boldsymbol{\Sigma}_{j_0}$ as k_0, M_0, $\boldsymbol{\Theta}_0$, $\mathbf{X}_0$ and $\boldsymbol{\Sigma}_0$, respectively. Furthermore, we also write the true model as M_0 or j_0, simply.

The bias properties of AIC, CAIC and MAIC have been studied under a large-sample framework,

$$p \text{ and } k \text{ are fixed, } \ n \to \infty, \tag{2.7}$$

and the assumption

$$\boldsymbol{\Omega}_j \equiv \boldsymbol{\Sigma}_0^{-1/2}(\mathbf{X}_0\boldsymbol{\Theta}_0)'(\mathbf{P}_\omega - \mathbf{P}_j)\mathbf{X}_0\boldsymbol{\Theta}_0\boldsymbol{\Sigma}_0^{-1/2} = \mathrm{O}(n). \tag{2.8}$$

More precisely, the bias depends on $\boldsymbol{\Omega}$ through the nonzero roots of $\boldsymbol{\Omega}$ which are the same as the roots of

$$\begin{aligned}\boldsymbol{\Lambda}_j &= n\boldsymbol{\Sigma}_0\{n\boldsymbol{\Sigma}_0 + (\mathbf{X}_0\boldsymbol{\Theta}_0)'(\mathbf{P}_\omega - \mathbf{P}_j)\mathbf{X}_0\boldsymbol{\Theta}_0\}^{-1} \\ &= \{\mathbf{I}_p + (1/n)\boldsymbol{\Sigma}_0^{-1}(\mathbf{X}_0\boldsymbol{\Theta}_0)'(\mathbf{P}_\omega - \mathbf{P}_j)\mathbf{X}_0\boldsymbol{\Theta}_0\}^{-1}\end{aligned} \tag{2.9}$$

The bias b_{A} (see Fujikoshi and Satoh, 1997) was expanded as

$$b_{\mathrm{A}} = b_{\mathrm{AL}} + \mathrm{O}(n^{-1}), \tag{2.10}$$

where

$$\begin{aligned}b_{\mathrm{AL}} = &\frac{2n}{n-k_j-p-1}\left\{k_j p + \frac{1}{2}p(p+1)\right\} \\ &+ 2k_j\mathrm{tr}(\boldsymbol{\Lambda}_j - \mathbf{I}_p) - \{\mathrm{tr}(\boldsymbol{\Lambda}_j - \mathbf{I}_p)\}^2 - \mathrm{tr}(\boldsymbol{\Lambda}_j - \mathbf{I}_p)^2.\end{aligned} \tag{2.11}$$

The results are summarized (see e.g., Fujikoshi and Satoh, 1997) as in the Table 1. In the table, the overspecified model means the model which includes the true model and the underspecified model means the model which is not the overspecified model. The terminologies "overspecified model" and "underspecified model" are the same as in Fujikoshi and Satoh (1997).

Table 1: The orders of biases of AIC, CAIC, MAIC under (2.7)

Candidate model	AIC	CAIC	MAIC
Underspecified	O(1)	O(1)	$O(n^{-1})$
Overspecified	$O(n^{-1})$	0	$O(n^{-2})$

Now we reevaluate the bias b_{A} by the high-dimensional framework (1.5). The standardized noncentrality matrix $\mathbf{\Omega}_j$ can be expressed as $\mathbf{\Omega}_j = \mathbf{\Gamma}_j'\mathbf{\Gamma}_j$, where $\mathbf{\Gamma}_j$ is a $r_j \times p$ matrix and $r_j = k - k_j$. The bias depends on $\mathbf{\Omega}_j$ through its characteristic roots which are the same as the ones of $r_j \times r_j$ matrix $\mathbf{\Gamma}_j\mathbf{\Gamma}_j'$. So, we assume that

$$\mathbf{\Gamma}_j\mathbf{\Gamma}_j' = n\mathbf{\Delta}_j = \mathrm{O}_h(n). \tag{2.12}$$

where $\mathrm{O}_h(n^i)$ denotes the terms of i-th order with respect to n under (1.5).

Theorem 2.1. *Suppose that the true model is included into the full model, and is expressed as in* (2.6). *Then, under* (1.5) *and* (2.12) *the bias term* b_A *in* (2.3) *can be expanded as*

$$b_{\mathrm{A}} = b_{\mathrm{AH}} + \mathrm{O}_h(n^{-1}), \tag{2.13}$$

$$\begin{aligned} b_{\mathrm{AH}} = &\frac{2n}{n-k_j-p-1}\left\{k_jp + \frac{1}{2}p(p+1)\right\} - \frac{nr_j(2k_j+r_j+1)}{n-k_j-p-1} \\ &+ \frac{n}{n-k_j-p-1}\left\{2(r_j+k_j+1)\xi_1 - \xi_2\right\}, \end{aligned} \tag{2.14}$$

where $\xi_1 = \eta_1$, $\xi_2 = \eta_1^2 + \eta_2$, *and*

$$\eta_i = \mathrm{tr}\left(\mathbf{I}_{r_j} + \mathbf{\Delta}_j\right)^{-i} = \mathrm{tr}\mathbf{\Lambda}_j^i - (p - r_j), \ i = 1, 2.$$

Expanding $b_{\rm AH}$ under a large-sample framework, and using that $\eta_1 = {\rm tr}(\mathbf{\Lambda}_j - \mathbf{I}_p) + r_j$ and $\eta_2 = {\rm tr}(\mathbf{\Lambda}_j - \mathbf{I}_p)^2 + 2{\rm tr}(\mathbf{\Lambda}_j - \mathbf{I}_p) + r_j$, we have

$$b_{\rm AH} = b_{\rm AL} + {\rm O}(n^{-1}).$$

From this result it is expected that the high-dimensional approximation $b_{\rm AH}$ will work even in a large-sample situation.

For a practical use we need to find estimators for ξ_1 and ξ_2 under (1.5). Naive estimators are given as

$$\begin{aligned}\tilde{\xi}_1 &= {\rm tr}\hat{\mathbf{\Sigma}}_\omega \hat{\mathbf{\Sigma}}_j^{-1} - (p - r_j),\\ \tilde{\xi}_2 &= \left\{{\rm tr}\hat{\mathbf{\Sigma}}_\omega \hat{\mathbf{\Sigma}}_j^{-1} - (p - r_j)\right\}^2 + {\rm tr}\left(\hat{\mathbf{\Sigma}}_\omega \hat{\mathbf{\Sigma}}_j^{-1}\right)^2 - (p - r_j).\end{aligned} \tag{2.15}$$

As one of the more preferable estimators we propose to use

$$\hat{\xi}_1 = \frac{1}{a_1}\tilde{\xi}_1, \quad \hat{\xi}_2 = \frac{1}{a_2}\tilde{\xi}_2, \tag{2.16}$$

where

$$\begin{aligned}a_1 &= \frac{m-p}{m}, \quad m = n - k,\\ a_2 &= \frac{a_1[\{m^2 - (p-1)m - 2\}(r_j + 1) + p]}{(r_j+1)(m-1)(m+2)}.\end{aligned} \tag{2.17}$$

Let $b_{\rm AH}$ be the one obtained from $b_{\rm AH}$ by substituting $\hat{\xi}_1$ and $\hat{\xi}_2$ to ξ_1 and ξ_2, respectively. Then, we propose HAIC by

$${\rm HAIC} = n \log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + \hat{b}_{\rm AH}, \tag{2.18}$$

which has the following property.

Theorem 2.2. *Under assumption* (1.5) *the high-dimensional* AIC, HAIC *defined by* (2.18) *satisfies the following properties:*
(1) *if M_j is an overspecified model,* HAIC *is an exact unbiased estimator of* $R_{\rm A}$,i.e.

$${\rm E(HAIC)} = R_{\rm A}.$$

(2) *if M_j is an underspecified model,*

$${\rm E(HAIC)} = R_{\rm A} + {\rm O}_h(n^{-1}).$$

3. Modifications of C_p

The C_p criterion was essentially proposed (for the univariate case, see Mallows, 1973; for the multivariate case, see Sparks, Coutsourides and Troskie, 1983) as an approximately unbiased estimator of the mean squares of errors of prediction. The risk of M_j may be defined by

$$R_{\mathrm{C}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_{\mathrm{F}}}[\mathrm{tr}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j)'(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j)], \tag{3.1}$$

where $\hat{\mathbf{Y}}_j$ is a predictor of $\mathbf{Y}$ under M_j given by $\hat{\mathbf{Y}}_j = \mathbf{X}_j\hat{\boldsymbol{\Theta}}_j = \mathbf{P}_j\mathbf{Y}$. The risk is expressed as

$$R_{\mathrm{C}} = \mathrm{E}_{\mathbf{Y}}[(n-k)\mathrm{tr}\hat{\boldsymbol{\Sigma}}_{\omega}^{-1}\hat{\boldsymbol{\Sigma}}_j] + b_{\mathrm{C}}, \tag{3.2}$$

where

$$b_{\mathrm{C}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_{\mathrm{F}}}\left[\mathrm{tr}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j)'(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j) - (n-k)\mathrm{tr}\hat{\boldsymbol{\Sigma}}_{\omega}^{-1}\hat{\boldsymbol{\Sigma}}_j\right]. \tag{3.3}$$

Similarly the C_p and its modification have been proposed by regarding b_{C} as the bias term when we estimate R_{C} by a minimum values of standardized residuals sum of squares as

$$(n-k)\mathrm{tr}\hat{\boldsymbol{\Sigma}}_{\omega}^{-1}\hat{\boldsymbol{\Sigma}}_j,$$

and by evaluating the bias term b_{C}.

Assuming that the true model is included in the full model, and is given by (2.6), Fujikoshi and Satoh (1997) showed that

$$b_{\mathrm{C}} = 2pk_j - \frac{p+1}{n-p-k-1}\left\{(k-k_j)p + \mathrm{tr}\boldsymbol{\Omega}_j\right\}. \tag{3.4}$$

If M_j is an overspecified model, under a large-sample framework we have

$$b_{\mathrm{C}} = 2pk_j + \mathrm{O}(n^{-1}),$$

and this leads to the usual C_p criterion. If M_j is an overspecified model, we have

$$b_{\mathrm{C}} = 2pk_j - \frac{(k-k_j)p(p+1)}{n-p-k-1},$$

and under a high-dimensional framework

$$\frac{b_{\mathrm{C}}}{p} \to 2k_j - (k - k_j)\frac{c}{1-c}.$$

In general, we have an exact estimator for b_C given by

$$\hat{b}_{\mathrm{C}} = 2pk_j - (p+1)\mathrm{tr}\hat{\mathbf{\Sigma}}_{\omega}^{-1}(\hat{\mathbf{\Sigma}}_j - \hat{\mathbf{\Sigma}}_{\omega}), \tag{3.5}$$

which leads to a modified criterion

$$\mathrm{MC}_p = \mathrm{C}_p - (p+1)\mathrm{tr}\hat{\mathbf{\Sigma}}_{\omega}^{-1}(\hat{\mathbf{\Sigma}}_j - \hat{\mathbf{\Sigma}}_{\omega}). \tag{3.6}$$

Changing (3.6) to the same expression as in Yanagihara and Satoh (2010) yields

$$\mathrm{MC}_p = \left(1 - \frac{p+1}{n-k}\right)\mathrm{C}_p + p(p+1)\left(\frac{2k_j}{n-k} + 1\right). \tag{3.7}$$

It is expected that MC_p does work well even in a high-dimensional case, since it is an exact unbiased estimator.

4. Consistency of AIC and Its Modifications

In this section we show that the asymptotic probabilities of selecting the true model by AIC and its modifications go to 1 as the sample size and the dimension of response variables approaching to ∞ as in (1.5), under the several assumptions. Let $\mathcal{F}$ be a set of candidate models, which is denoted by $\mathcal{F} = \{j_1, \ldots, j_m\}$, and separate $\mathcal{F}$ into two sets, one is a set of overspecified models, i.e., $\mathcal{F}_+ = \{j \in \mathcal{F} \mid j_0 \subseteq j\}$ and the other is a set of underspecified models, i.e., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F}$. Thus, the true model j_0 can be regarded as the smallest overspecified model. We denote the value of AIC for model M_j by AIC(j). The same notations as the described above are used for other criteria.

The best subsets of ω chosen by minimizing AIC, CAIC, MAIC and HAIC are written as

$$\hat{j}_{\mathrm{A}} = \arg\min_{j\in\mathcal{F}} \mathrm{AIC}(j), \qquad \hat{j}_{\mathrm{CA}} = \arg\min_{j\in\mathcal{F}} \mathrm{CAIC}(j),$$
$$\hat{j}_{\mathrm{MA}} = \arg\min_{j\in\mathcal{F}} \mathrm{MAIC}(j), \quad \hat{j}_{\mathrm{HA}} = \arg\min_{j\in\mathcal{F}} \mathrm{HAIC}(j).$$

Here we list our main assumptions:

A1 (The true model): $j_0 \in \mathcal{F}$.

A2 (The asymptotic framework): $p \to \infty,\ n \to \infty,\ p/n \to c \in (0, 1)$.

A3 (The noncentrality matrix):

For $j \in \mathcal{F}_-,\ \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = n\boldsymbol{\Delta}_j = \mathrm{O}_h(n)$ and $\lim_{p/n\to c} \boldsymbol{\Delta}_j = \boldsymbol{\Delta}_j^*$.

Theorem 4.1. *Suppose that the assumptions* A1, A2 *and* A3 *are satisfied.*

(1) *Let* c_{a} (≈ 0.797) *be the constant satisfying* $\log(1-c_{\mathrm{a}})+2c_{\mathrm{a}} = 0$. *Further, assume that* $c \in (0, c_{\mathrm{a}})$, *and*

A4: *For any* $j \in \mathcal{F}_-$ *with* $k_0 - k_j \geq 0$,

$$\log|\mathbf{I}_{r_j} + \boldsymbol{\Delta}_j^*| > (k_0 - k_j)\{2c + \log(1-c)\}.$$

Then, the asymptotic probability of selecting the true model j_0 *by* AIC *tends to* 1, i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{A}} = j_0) = 1.$$

(2) *Suppose that the following assumption* A5 *is satisfied.*

A5: *For any* $j \in \mathcal{F}_-$ *with* $k_0 - k_j \geq 0$,

$$\log|\mathbf{I}_{r_j} + \boldsymbol{\Delta}_j^*| > (k_0 - k_j)\left\{\frac{c}{1-c} + \frac{c}{(1-c)^2} + \log(1-c)\right\}.$$

Then, the asymptotic probability of selecting the true model j_0 *by* CAIC, MAIC *and* HAIC *tends to* 1, i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{TA}} = j_0) = 1,$$

where $\mathrm{TA} = \mathrm{CA}, \mathrm{MA}$ *or* HA.

Yanagihara, Wakaki and Fujikoshi (2012) have shown a consistency of AIC and CAIC. They assumed $\mathbf{\Gamma}_j\mathbf{\Gamma}_j' = \mathrm{O}_h(pn)$ instead of A3, but without A4 and A5. We note that when $\mathbf{\Gamma}_j\mathbf{\Gamma}_j' = \mathrm{O}_h(pn)$, the assumptions 4 and 5 are satisfied.

5. Consistency of C_p and MC_p

In this section we show that the selection-probabilities of selecting the true model by C_p and MC_p go to 1 as the sample size and the dimension of response variables approaching to ∞ as in (1.5), under some assumptions. Similar notations as in Section 3.1 are used.

The best subsets of ω chosen by minimizing C_p and MC_p are written as

$$\hat{j}_{\mathrm{C}} = \arg\min_{j\in\mathcal{F}} \mathrm{C}_p(j), \quad \hat{j}_{\mathrm{MC}} = \arg\min_{j\in\mathcal{F}} \mathrm{MC}_p(j)$$

Theorem 5.1. *Suppose that the assumptions* A1, A2 *and* A3 *are satisfied.*

(1) *Suppose that* $0 < c < 1/2$*, and*

A6: *For any* $j \in \mathcal{F}_-$ *with* $k_0 - k_j > 0$,

$$\mathrm{tr}\mathbf{\Delta}_j^* > (k_0 - k_j)c(1 - 2c).$$

Then, the asymptotic probability of selecting the true model j_0 *by* C_p *tends to* 1*,* i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{C}} = j_0) = 1.$$

(2) *Suppose that*

A7: *For any* $j \in \mathcal{F}_-$ *with* $k_0 - k_j > 0$,

$$\mathrm{tr}\mathbf{\Delta}_j^* > (k_0 - k_j)c.$$

Then, the asymptotic probability of selecting the true model j_0 *by* MC_p *tends to* 1, i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{MC}} = j_0) = 1.$$

For a consistency of C_p we need to assume $0 < c < 1/2$ and A6 for the constant c. On the other hand, for a consistency of MC_p we need to assume A7 only for the constant c.

6. Simulation study

In this section, we numerically examine the validity of our claim. The five candidate models $j_\alpha = \{1, \ldots, \alpha\}$ $(\alpha = 1, \ldots, 5)$, with several different values of n and $p = cn$, were prepared for Monte Carlo simulations, where $n = 20, 50, 100, 500$ and $c = 0.1, 0.2, 0.4$. We generated $z_1, \ldots z_n \sim i.i.d.\ U(-1, 1)$. Using $z_1, \ldots z_n$, we constructed a $n \times 5$ matrix of explanatory variables $\mathbf{X}$ where the (a, b)th element was defined by z_a^{b-1} $(a = 1, \ldots, n; b = 1, \ldots, 5)$. The true model was determined by $\mathbf{\Theta}_0 = \mathbf{1}_3\boldsymbol{\theta}_0'$ and $\mathbf{\Sigma}_0 = \mathbf{\Psi}_0^{1/2}\{(0.2)\mathbf{I}_p + (0.8)\mathbf{1}_p\mathbf{1}_p'\}\mathbf{\Psi}_0^{1/2}$, where $\mathbf{1}_p$ is the p-dimensional vector of ones, and

$$\boldsymbol{\theta}_0 = 2(1, (-0.9), \ldots, (-0.9)^{p-1})', \ \ \mathbf{\Psi}_0 = 2\mathbf{I}_p - \mathrm{diag}(0, 1/p, \ldots, (p-1)/p).$$

Thus, j_1 and j_2 were underspecified models, and j_3, j_4 and j_5 were over-specified models. Moreover, j_3 was the true model. In the above simulation model, convergent values in the conditions for consistency were calculated as

$$\log|\mathbf{I}_4 - \mathbf{\Delta}_{j_1}^*| \approx 3.145, \quad \mathrm{tr}\mathbf{\Delta}_{j_1}^* \approx 22.222,$$
$$\log|\mathbf{I}_3 - \mathbf{\Delta}_{j_2}^*| \approx 1.737, \quad \mathrm{tr}\mathbf{\Delta}_{j_2}^* \approx 4.678.$$

Hence, in the simulated data, all the criteria were consistent in variable selection as $p/n \to c$.

First, we studied performances of AIC, CAIC, MAIC and HAIC as estimators of R_{A}. For each of $j_1, \ldots, j_5$, we computed the average of R_{A}, AIC,

Table 2: Risks and biases of AIC, CAIC, MAIC and HAIC when $n = 20, 50$

k	R_{A}	AIC	CAIC	MAIC	HAIC	R_{A}	AIC	CAIC	MAIC	HAIC
	$(n, p) = (20, 2)$					$(n, p) = (20, 4)$				
1	168.1	-1.97	-4.47	-1.10	-0.80	272.9	6.92	-5.08	-1.74	-0.65
2	141.9	0.18	-4.49	-1.33	-0.58	249.6	13.08	-6.31	-2.69	-0.69
3	130.5	7.58	-0.13	-0.06	-0.13	239.9	28.99	-0.34	-0.15	-0.34
4	135.9	11.70	-0.15	-0.11	-0.15	255.5	42.37	-0.18	-0.09	-0.20
5	142.5	17.13	-0.21	-0.21	-0.21	274.5	59.57	-0.43	-0.43	-0.43
	$(n, p) = (20, 8)$					$(n, p) = (50, 5)$				
1	517.9	80.25	-7.75	-4.49	-1.36	733.7	1.35	-5.16	-1.77	-1.19
2	524.5	116.18	-10.93	-7.44	-1.89	673.2	3.75	-5.77	-1.85	-0.95
3	556.7	178.96	-1.04	-0.44	-1.01	613.8	12.55	-0.62	-0.59	-0.62
4	631.2	251.39	-1.19	-0.87	-1.15	622.4	16.97	-0.53	-0.51	-0.54
5	734.7	354.10	-0.57	-0.57	-0.57	631.5	21.91	-0.65	-0.65	-0.65
	$(n, p) = (50, 10)$					$(n, p) = (50, 20)$				
1	1260.6	35.98	-5.07	-1.84	-0.46	2552.2	355.76	-5.67	-2.64	0.66
2	1209.5	46.33	-6.37	-2.47	-0.42	2557.8	418.50	-7.42	-4.01	0.88
3	1152.1	66.14	0.03	0.14	0.02	2564.9	500.87	2.41	2.82	2.43
4	1175.1	81.44	0.01	0.06	0.00	2658.5	582.97	2.97	3.18	2.98
5	1199.7	98.61	-0.21	-0.21	-0.21	2760.4	674.34	2.68	2.68	2.68

CAIC, MAIC and HAIC by Monte Carlo simulations with 10,000 replications. Table 2 shows the risk R_{A} and biases of AIC, CAIC, MAIC and HAIC to R_{A}, defined by $R_{\mathrm{A}} -$(the expectation of the information criterion). In the table, the bold face denotes the true model. Since the tendencies of results were almost the same, we omit the result in the case $n = 100$ and 500 to save the space. From the table, we can see that the biases of the HAIC were the smallest in most cases. Especially, the desirable characteristic of HAIC appeared prominently in the underspecified models. Moreover, in the overspecified, performances of CAIC, MAIC and HAIC were almost the same. In the all criteria, the more increased dimension was, the larger bias appeared.

Next, we studied the probabilities of selecting the model by the AIC, CAIC, MAIC, HAIC, C_p and MC_p, which were evaluated by Monte Carlo simulations with 10,000 iterations. Table 3 shows the probability of selecting the underspecified models, the true model and the overspecified models by each criterion. In the table, columns in $\mathcal{F}_-$, j_0 and $\mathcal{F}_+$ express the probability of selecting the underspecified models, the true model and the overspecified models, respectively. From the table, we can see that all the criteria were consistent in variable selection as $p/n \to c$. However, selection probabilities using C_p were slower to converge to 1 than other criteria. Under finite sample and dimension, when c was small, the performances of CAIC, MAIC and HAIC were better than those of AIC,C_p and MC_p. On the other hand, when c was large, the performances of MC_p were better than those of AIC CAIC, MAIC, HAIC and C_p.

Table 3: Selection probabilities (%) of AIC, CAIC, MAIC, HAIC, C_p and MC_p

			AIC			CAIC			MAIC			HAIC			C_p			MC_p		
c	n	p	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$	$\mathcal{F}_-$	j_0	$\mathcal{F}_+$
0.1	20	2	0.0	69.3	30.7	0.2	94.1	5.7	0.5	97.1	2.5	0.7	97.9	1.5	0.0	74.7	25.3	0.1	82.8	17.1
	50	5	0.0	84.6	15.4	0.0	96.8	3.2	0.0	97.9	2.1	0.0	98.1	1.9	0.0	83.8	16.2	0.0	89.5	10.5
	100	10	0.0	94.2	5.8	0.0	99.0	1.0	0.0	99.3	0.7	0.0	99.3	0.7	0.0	92.0	8.0	0.0	95.8	4.2
	500	50	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	99.9	0.1	0.0	100.0	0.0
0.2	20	4	0.0	68.1	31.9	0.9	98.4	0.6	2.8	97.1	0.2	5.5	94.4	0.1	0.0	68.3	31.7	0.0	86.0	14.0
	50	10	0.0	89.4	10.6	0.0	99.9	0.1	0.0	100.0	0.0	0.0	100.0	0.0	0.0	82.0	18.0	0.0	94.2	5.8
	100	20	0.0	97.2	2.8	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	91.9	8.1	0.0	98.5	1.6
	500	100	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	99.9	0.1	0.0	100.0	0.0
0.4	20	8	0.0	55.3	44.7	99.7	0.3	0.0	99.8	0.2	0.0	99.9	0.1	0.0	0.0	38.9	61.2	0.1	87.8	12.1
	50	20	0.0	85.8	14.2	66.7	33.3	0.0	73.0	27.0	0.0	81.5	18.6	0.0	0.0	52.8	47.2	0.0	96.1	3.9
	100	40	0.0	96.8	3.2	27.5	72.5	0.0	33.0	67.0	0.0	42.3	57.7	0.0	0.0	63.1	36.9	0.0	99.3	0.7
	500	200	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	90.4	9.6	0.0	100.0	0.0

Appendix

A. Preliminary Lemmas

We give three Lemmas which are used in the proofs of Theorems 1 $\sim$ 4. The following Lemma has been essentially used in Sakurai, Nakata and Fujikoshi (2012).

Lemma A.1. *Let* $\mathsf{T} \sim \mathrm{W}_p(n, \mathbf{I}_p; \boldsymbol{\Omega})$*, and* T *and* $\boldsymbol{\Omega}$ *be partitioned as*

$$\mathsf{T} = \begin{pmatrix} \mathsf{T}_{11} & \mathsf{T}_{12} \\ \mathsf{T}_{21} & \mathsf{T}_{22} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

respectively, where $\mathsf{T}_{ij} : p_i \times p_j$ *and* $\boldsymbol{\Omega}_{ij} : p_i \times p_j$*. If* $\boldsymbol{\Omega}_{12}$*,* $\boldsymbol{\Omega}_{21}$ *and* $\boldsymbol{\Omega}_{22}$ *are zero matrices, then*

$$\mathrm{E}[\mathrm{tr}\mathsf{T}^{-1}\mathbf{A}] = \frac{m - p_1 - 1}{m - p - 1}\mathrm{E}[\mathrm{tr}\mathsf{T}_{11}^{-1}\mathbf{A}_{11}] + \frac{1}{m - p - 1}\mathrm{tr}\mathbf{A}_{22},$$

where $\mathbf{A}$ *is a* $p \times p$ *constant matrix partitioned in the same way as the partitions of* T*.*

The following Lemma was given in Fujikoshi (1985).

Lemma A.2. *Suppose that* $\mathsf{T} \sim \mathrm{W}_r(n - k, \mathbf{I}_r; n\boldsymbol{\Delta})$*, and let* $\mathbf{A}; r \times r$ *be a constant matrix. If* r, k *and* $\boldsymbol{\Delta}$ *are fixed and* n *tends to infinity, then*

$$\begin{aligned}\mathrm{E}[\mathrm{tr}\mathsf{T}^{-1}\mathbf{A}] = {} & \frac{1}{n}\mathrm{tr}\mathbf{A}\boldsymbol{\Psi} + \frac{1}{n^2}\left\{(r + k + 2)\mathrm{tr}\mathbf{A}\boldsymbol{\Psi}^2\right. \\ & \left.+\mathrm{tr}\boldsymbol{\Psi}\mathrm{tr}\mathbf{A}\boldsymbol{\Psi} - \mathrm{tr}\mathbf{A}\boldsymbol{\Psi}^3 - \mathrm{tr}\boldsymbol{\Psi}\mathrm{tr}\mathbf{A}\boldsymbol{\Psi}^2\right\} + \mathrm{O}(n^{-3}),\end{aligned}$$

where $\boldsymbol{\Psi} = (\mathbf{I}_r + \boldsymbol{\Delta})^{-1}$*.*

Lemma A.3. *Let* $\mathsf{S}_h = \mathsf{X}'\mathsf{X}$ *and* S_e *be independently distributed as* $\mathrm{W}_p(q, \mathbf{I}_p; \mathbf{M}'\mathbf{M})$ *and* $\mathrm{W}_p(n, \mathbf{I}_p)$*, respectively. Here* X *is a* $q \times p$ *random matrix whose* *elements are independent normal variables with* $\mathrm{E}(\mathbf{X}) = \mathbf{M}$ *and the common variance* 1. *Put*

$$\mathbf{B} = \mathbf{X}\mathbf{X}' \quad \textit{and} \quad \mathbf{W} = \mathbf{B}^{1/2}(\mathbf{X}\mathbf{S}_e^{-1}\mathbf{X}')^{-1}\mathbf{B}^{1/2}.$$

Then:

(1) $\mathbf{B}$ *and* $\mathbf{W}$ *are independently distributed as* $\mathrm{W}_q(p, \mathbf{I}_q; \mathbf{M}\mathbf{M}')$ *and* $\mathrm{W}_q(n - p + q, \mathbf{I}_q)$, *respectively.*

(2) *The nonzero characteristic roots of* $\mathbf{S}_h\mathbf{S}_e^{-1}$ *are the same as those of* $\mathbf{B}\mathbf{W}^{-1}$. *In particular*

$$\frac{|\mathbf{S}_e|}{|\mathbf{S}_e + \mathbf{S}_h|} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|}, \quad \mathrm{tr}\mathbf{S}_h\mathbf{S}_e^{-1} = \mathrm{tr}\mathbf{B}\mathbf{W}^{-1},$$

and

$$\mathrm{tr}\mathbf{S}_e(\mathbf{S}_e + \mathbf{S}_h)^{-1} - (p - q) = \mathrm{tr}\mathbf{W}(\mathbf{W} + \mathbf{B})^{-1},$$
$$\mathrm{tr}\left\{\mathbf{S}_e(\mathbf{S}_e + \mathbf{S}_h)^{-1}\right\}^2 - (p - q) = \mathrm{tr}\left\{\mathbf{W}(\mathbf{W} + \mathbf{B})^{-1}\right\}^2.$$

Lemma A.3 was essentially obtained in Wakaki, Fujikoshi and Ulyanov (2002) and Fujikoshi, Ulyanov and Fujikoshi (2010).

B. Proof of Theorem

B.1. Proof of Theorem 2.1

We can write the bias term b_{A} in (2.3) as

$$\begin{aligned} b_{\mathrm{A}} &= \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_{\mathrm{F}}}[\mathrm{tr}\hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{Y}_{\mathrm{F}} - \mathbf{X}_j\hat{\boldsymbol{\Theta}}_j)'(\mathbf{Y}_{\mathrm{F}} - \mathbf{X}_j\hat{\boldsymbol{\Theta}}_j)] - np \\ &= \mathrm{E}^*_{\mathbf{Y}}[\mathrm{tr}\hat{\boldsymbol{\Sigma}}_j^{-1}\{n\boldsymbol{\Sigma}_0 + (\mathbf{X}_0\boldsymbol{\Theta}_0 - \mathbf{X}_j\hat{\boldsymbol{\Theta}}_j)'(\mathbf{X}_0\boldsymbol{\Theta}_0 - \mathbf{X}_j\hat{\boldsymbol{\Theta}}_j)\}] - np. \end{aligned}$$

Noting that $\hat{\boldsymbol{\Sigma}}_j$ and $\mathbf{X}_j\hat{\boldsymbol{\Theta}}_j$ are independent, we can see that

$$b_{\mathrm{A}} = n\mathrm{E}[\mathrm{tr}\mathbf{T}^{-1}\{(n + k_j)\mathbf{I}_p + \boldsymbol{\Omega}_j\}] - np, \tag{B.1}$$

where $\mathsf{T} \sim \mathrm{W}_p(n-k_j, \mathbf{I}_p; \mathbf{\Omega}_j)$. The noncentrality matrix can be expressed as $\mathbf{\Gamma}_j'\mathbf{\Gamma}_j$, where $\mathbf{\Gamma}_j$ is a $r_j \times p$ matrix. We may consider the case $p > r_j$ since p tends to infinity. Then we may regard $\mathbf{\Omega}_j$ as

$$\mathbf{\Omega}_j = \begin{pmatrix} \mathbf{\Omega}_{11,j} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \quad \mathbf{\Omega}_{11,j} = \mathbf{\Gamma}_j\mathbf{\Gamma}_j'.$$

This is shown by considering an orthogonal transformation $\mathsf{T} \to \mathbf{H}'\mathsf{T}\mathbf{H}$. Putting $\mathbf{A} = (n+k_j)\mathbf{I}_p + \mathbf{\Omega}_j$, and using Lemma A.1, we can reduce b_{A} as

$$b_{\mathrm{A}} = n\left[\frac{n-k_j-r_j-1}{n-k_j-p-1}\mathrm{E}[\mathrm{tr}\mathsf{T}_{11}^{-1}\mathbf{A}_{11}] + \frac{(p-r_j)(n+k_j)}{n-k_j-p-1}\right] - np.$$

Further, noting that $\mathbf{\Omega}_{11,j} = n\mathbf{\Delta}_j$, $\mathsf{T}_{11} \sim \mathrm{W}_{r_j}(n-k_j, \mathbf{I}_{r_j}; n\mathbf{\Delta}_j)$, and using Lemma A.2, we obtain

$$\mathrm{E}[\mathrm{tr}\mathsf{T}_{11}^{-1}\mathbf{A}_{11}] = r_j + \frac{1}{n}\{2(r_j+k_j+2)\mathrm{tr}\mathbf{\Psi}_j - (\mathrm{tr}\mathbf{\Psi}_j)^2 - \mathrm{tr}\mathbf{\Psi}_j^2\} + \mathrm{O}(n^{-2}),$$

where $\mathbf{\Psi}_j = (\mathbf{I}_{r_j} + \mathbf{\Delta}_j)^{-1}$. From these we obtain (2.13).

B.2. Proof of Theorem 2.2

When we consider the distribution of the naive estimators in (2.15), we may express as

$$\tilde{\xi}_1 = \mathrm{tr}\mathsf{Q}, \quad \tilde{\xi}_2 = (\mathrm{tr}\mathsf{Q})^2 + \mathrm{tr}\mathsf{Q}^2,$$

where $\mathsf{Q} = \mathsf{W}(\mathsf{W}+\mathsf{B})^{-1}$. Here W and B are independently distributed as $\mathrm{W}_{r_j}(m-p, \mathbf{I}_{r_j})$ and $\mathrm{W}_{r_j}(p, \mathbf{I}_{r_j}; n\mathbf{\Delta}_j)$, respectively.

First consider the case when M_j is an overspecified model, then $\mathbf{\Delta}_j = \mathbf{O}$. Then Q is distributed as a multivariate beta distribution $\mathrm{B}_{r_j}(m-p, p)$. Furthermore, we have the following moments from Fujikoshi and Satoh (1997).

$$\mathrm{E}[\mathrm{tr}\mathsf{Q}] = a_1 r_j, \quad \mathrm{E}[(\mathrm{tr}\mathsf{Q})^2 + \mathrm{tr}\mathsf{Q}^2] = a_2 r_j(r_j+1).$$

Therefore

$$\begin{aligned}\mathrm{E}[\hat{b}_{\mathrm{AH}}] &= \frac{2n}{n-k_j-p-1}\left\{k_j p + \frac{1}{2}p(p+1)\right\} - \frac{nr_j(2k_j+r_j+1)}{n-k_j-p-1} \\ &\quad + \frac{n}{n-k_j-p-1}\left\{2(r_j+k_j+1)r_j - r_j^2 - r_j\right\} \\ &= \frac{2n}{n-k_j-p-1}\left\{k_j p + \frac{1}{2}p(p+1)\right\},\end{aligned}$$

which shows Theorem 2.2 (1).

Next consider the proof of Theorem 2.2 (2). Let $\mathbf{U}$ and $\mathbf{V}$ be defined by

$$\frac{1}{p}\mathbf{B} = \mathbf{I}_{r_j} + \frac{n}{p}\boldsymbol{\Delta}_j + \frac{1}{\sqrt{p}}\mathbf{U}, \quad \frac{1}{m-p}\mathbf{W} = \mathbf{I}_{r_j} + \frac{1}{\sqrt{m-p}}\mathbf{V}.$$

Then $\mathbf{W}^{-1}\mathbf{B}$ is expanded as

$$\begin{aligned}\mathbf{W}^{-1}\mathbf{B} &= \frac{p}{m-p}\left(\mathbf{I}_{r_j} + \frac{1}{\sqrt{m-p}}\mathbf{V}\right)^{-1}\left(\mathbf{I}_{r_j} + \frac{n}{p}\boldsymbol{\Delta}_j + \frac{1}{\sqrt{p}}\mathbf{U}\right) \\ &= \frac{p}{m-p}\left[\mathbf{I}_{r_j} + \frac{n}{p}\boldsymbol{\Delta}_j + \left\{\frac{1}{\sqrt{p}}\mathbf{U} - \frac{1}{\sqrt{m-p}}\mathbf{V}(\mathbf{I}_{r_j} + \frac{n}{p}\boldsymbol{\Delta}_j)\right\}\right] \\ &\quad + \mathrm{O}_h(n^{-1}),\end{aligned}$$

by using

$$\left(\mathbf{I}_{r_j} + \frac{1}{\sqrt{m-p}}\mathbf{V}\right)^{-1} = \mathbf{I}_{r_j} - \frac{1}{\sqrt{m-p}}\mathbf{V} + \mathrm{O}_h(n^{-1}).$$

This implies the following:

$$\begin{aligned}\hat{\xi}_1 &= \xi_1 + b_1 + \mathrm{O}_h(n^{-1}), \\ \hat{\xi}_2 &= \xi_2 + b_2 + \mathrm{O}_h(n^{-1}),\end{aligned}$$

where b_1 and b_2 are homogeneous expressions of degree 1 with respect to the elements of $\mathbf{U}$ and $\mathbf{V}$. From these we can see that $\mathrm{E}[\hat{\xi}_1] = \xi_1 + \mathrm{O}_h(n^{-1})$, and $\mathrm{E}[\hat{\xi}_2] = \xi_2 + \mathrm{O}_h(n^{-1})$. This implies Theorem 2.2 (2).

B.3. Proof of Theorem 4.1

First we consider behavior of $|\hat{\boldsymbol{\Sigma}}_\omega|/|\hat{\boldsymbol{\Sigma}}_j|$ for a candidate model $j \in \mathcal{F}$. It is easy to see that

$$\frac{|\hat{\boldsymbol{\Sigma}}_\omega|}{|\hat{\boldsymbol{\Sigma}}_j|} = \frac{|\mathbf{S}_e|}{|\mathbf{S}_e + (\mathbf{S}_j - \mathbf{S}_e)|},$$

where $\mathbf{S}_e$ and $\mathbf{S}_j - \mathbf{S}_e$ are independently distributed as $\mathrm{W}_p(n-k, \mathbf{I}_p)$ and $\mathrm{W}_p(r_j, \mathbf{I}_p; \boldsymbol{\Omega}_j)$, respectively. Using Lemma A.3, we obtain an expression in terms of $r_j \times r_j$ matrices given by

$$\frac{|\hat{\boldsymbol{\Sigma}}_\omega|}{|\hat{\boldsymbol{\Sigma}}_j|} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|}, \tag{B.2}$$

where $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as $\mathrm{W}_{r_j}(n-k-p+r_j, \mathbf{I}_{r_j})$ and $\mathrm{W}_{r_j}(p, \mathbf{I}_{r_j}; \mathbf{\Gamma}_j\mathbf{\Gamma}_j')$, respectively. Using A3 and $r_j = k - k_j$, the distributions of $\mathbf{W}$ and $\mathbf{B}$ are the same as the ones of $\mathrm{W}_{r_j}(n-k_j-p, \mathbf{I}_{r_j})$ and $\mathrm{W}_{r_j}(p, \mathbf{I}_{r_j}; n\mathbf{\Delta}_j)$, respectively. Based on a well-known asymptotic method on Wishart distributions, we can see that under A2

$$\frac{1}{n}\mathbf{B} \xrightarrow{p} c\mathbf{I}_{r_j} + \mathbf{\Delta}_j^*, \quad \frac{1}{n}\mathbf{W} \xrightarrow{p} (1-c)\mathbf{I}_{r_j}, \tag{B.3}$$

which implies

$$\frac{|\hat{\mathbf{\Sigma}}_\omega|}{|\hat{\mathbf{\Sigma}}_j|} \xrightarrow{p} \frac{(1-c)^{r_j}}{|\mathbf{I}_{r_j} + \mathbf{\Delta}_j^*|},$$

where $\xrightarrow{p}$ denotes the convergence in probability. Therefore, we have

$$\begin{aligned}\frac{1}{n}\{\mathrm{AIC}(j) - \mathrm{AIC}(j_0)\} &= -\log\frac{|\hat{\mathbf{\Sigma}}_\omega|}{|\hat{\mathbf{\Sigma}}_j|} + \log\frac{|\hat{\mathbf{\Sigma}}_\omega|}{|\hat{\mathbf{\Sigma}}_0|} + 2(k_j - k_0)\frac{p}{n} \\ &\xrightarrow{p} (k_j - k_0)\{2c + \log(1-c)\} + \log|\mathbf{I}_{r_j} + \mathbf{\Delta}_j^*| \equiv d_{\mathrm{A}}(j). \end{aligned} \tag{B.4}$$

Here we used $\mathbf{\Delta}_j^* = \mathbf{O}$ when $j \in \mathcal{F}_+$. Note that if $0 < c < c_0$, $2c + \log(1-c) > 0$.

If $j \in \mathcal{F}_+$ and $j \neq j_0$, $(k_j - k_0) > 0$, $\mathbf{\Delta}_j^* = \mathbf{O}$, and

$$d_{\mathrm{A}}(j) = (k_j - k_0)\{2c + \log(1-c)\} > 0 \text{ for } 0 < c < c_0.$$

On the other hand, if $j \in \mathcal{F}_-$, $d_{\mathrm{A}}(j) > 0$ when $(k_j - k_0) > 0$. Therefore, if A4 holds, $d_{\mathrm{A}}(j) > 0$ except for $j = j_0$. This implies Theorem 4.1 (1).

Next we consider the case of CAIC. From (2.4),

$$\frac{1}{n}\{\mathrm{CAIC}(j) - \mathrm{CAIC}(j_0)\} = \frac{1}{n}\{\mathrm{AIC}(j) - \mathrm{AIC}(j_0)\} + AD(j),$$

where

$$\begin{aligned} AD(j) = &\frac{2(k_j + p + 1)}{n(n - k_j - p - 1)}\left\{k_j p + \frac{1}{2}p(p+1)\right\} \\ &- \frac{2(k_0 + p + 1)}{n(n - k_0 - p - 1)}\left\{k_0 p + \frac{1}{2}p(p+1)\right\}. \end{aligned}$$

It is easy to see that

$$\lim_{p/n\to c} AD(j) = (k_j - k_0)\left\{\frac{2c^2}{1-c} + \frac{c^2}{(1-c)^2}\right\}.$$

Therefore,

$$\frac{1}{n}\{\mathrm{CAIC}(j) - \mathrm{CAIC}(j_0)\} \xrightarrow{p} d_{\mathrm{CA}}(j), \qquad \text{(B.5)}$$

where

$$d_{\mathrm{CA}}(j) = (k_j - k_0)\left\{\frac{c}{1-c} + \frac{c}{(1-c)^2} + \log(1-c)\right\} + \log|\mathbf{I}_{r_j} + \boldsymbol{\Delta}_j^*|.$$

By the same discussion as in the consistency ot AIC based on $d_{CA}(j)$ we can show Theorem 4.1 (2) in the case of CAIC.

For the case of MAIC and HAIC, we can see that the additional parts to CAIC converge to zero. For example,

$$\frac{1}{n}\{2k_j\mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p) - \{\mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p)\}^2 - \mathrm{tr}(\mathbf{L}_j - \mathbf{I}_p)^2\} \xrightarrow{p} 0.$$

These complete the proof of Theorem 4.1.

B.4. Proof of Theorem 5.1

We use the same notation as in the proof of Theorem 4.1. For a candidate model $j \in \mathcal{F}$, we have

$$\begin{aligned}\mathrm{tr}\hat{\boldsymbol{\Sigma}}_\omega^{-1}\hat{\boldsymbol{\Sigma}}_j &= \mathrm{tr}\mathbf{S}_e^{-1}\{\mathbf{S}_e + (\mathbf{S}_j - \mathbf{S}_e)\}\\ &= p + \mathrm{tr}(\mathbf{S}_j - \mathbf{S}_e)\mathbf{S}_e^{-1} = p + \mathrm{tr}\mathbf{B}\mathbf{W}^{-1}.\end{aligned}$$

Therefore

$$\mathrm{C}_p(j) - \mathrm{C}_p(j_0) = (n-k)\left(\mathrm{tr}\mathbf{B}\mathbf{W}^{-1} - \mathrm{tr}\mathbf{B}_0\mathbf{W}^{-1}\right) + 2p(k_j - k_0),$$

where $\mathbf{B}_0 \sim \mathrm{W}_p(k - k_0, \mathbf{I}_p)$. Using (B.3),

$$\mathrm{tr}\mathbf{B}\mathbf{W}^{-1} \xrightarrow{p} \frac{1}{1-c}\left(cr_j + \mathrm{tr}\boldsymbol{\Delta}_j^*\right), \quad \mathrm{tr}\mathbf{B}_0\mathbf{W}^{-1} \xrightarrow{p} \frac{c(k-k_0)}{1-c},$$

and hence

$$\frac{1}{n}\{\mathrm{C}_p(j) - \mathrm{C}_p(j_0)\} \xrightarrow{p} d_{\mathrm{C}}(j), \qquad \text{(B.6)}$$

where

$$d_{\mathrm{C}}(j) = (k_j - k_0)c\left(\frac{1-2c}{1-c}\right) + \frac{1}{1-c}\mathrm{tr}\boldsymbol{\Delta}_j^*.$$

If $j \in \mathcal{F}_+$ and $j \neq j_0$, $(k_j - k_0) > 0$, $\boldsymbol{\Delta}_j^* = \mathbf{O}$, and

$$d_{\mathrm{C}}(j) = \frac{(k_j - k_0)c(1-2c)}{1-c} > 0 \text{ for } 0 < c < 1/2.$$

On the other hand, if $j \in \mathcal{F}_-$, $d_{\mathrm{C}}(j) > 0$ when $(k_j - k_0) > 0$ and $c > 0$. Therefore, if A6 holds, $d_{\mathrm{C}}(j) > 0$ except for $j = j_0$. This implies Theorem 5.1 (1).

For the proof of (2), it follows from (3.7) that

$$\frac{1}{n}\{\mathrm{MC}_p(j) - \mathrm{MC}_p(j_0)\} = \left(1 - \frac{p+1}{n-k}\right)\frac{1}{n}\{\mathrm{C}_p(j) - \mathrm{C}_p(j_0)\} + \frac{2p(p+1)}{n(n-k)}(k_j - k_0).$$

Note that $\lim_{p/n\to c}(p+1)/(n-k) = 1-c$ and $\lim_{p/n\to c} 2p(p+1)/\{n(n-k)\} = 2c^2$. Therefore we have

$$\frac{1}{n}\{\mathrm{MC}_p(j) - \mathrm{MC}_p(j_0)\} \xrightarrow{p} c(k_j - k_0) + \mathrm{tr}\boldsymbol{\Delta}_j^* \equiv d_{\mathrm{MC}}(j). \tag{B.7}$$

By the same discussion as in the proof of (1), we can show (2).

Acknowledgements

The first author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #22500259, 2010-2012. The third author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Challenging Exploratory Research, #22650058, 2010-2012.

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. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In *Multivariate Analysis-VI* (ed. P. R. Krishnaian), 219–236, Elsevier Science Publishers, B.V.

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

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

[6] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.*, **29**, 295–327.

[7] Kabe, D. G. (1964). A note on the Bartlett decomposition of a Wishart matrix. *J. Roy. Statist. Soc. Ser. B*, **26**, 270–273.

[8] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[9] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, N.Y.

[10] Sakurai, T., Nakada, T. and Fujikoshi, Y. (2012). High-dimensional AICs for selection of redundancy models in discriminant analysis. TR No. 12-13, *Statistical Research Group, Hiroshima University.*

[11] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Ohio.

[12] Sparks, R. S., Coutsourides, D. and Troskie, L. (1983). The multivariate C_p. *Comm. Statist. A Theory Methods*, **12**, 1775–1793.

[13] Wakaki, H., Fujikoshi, Y. and Ulyanov, V. V. (2002). Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large. TR No. 02-10, *Statistical Research Group*, *Hiroshima University.*

[14] Yanagihara, H. and Satoh, K. (2010). An unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.

[15] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2012). A consistency property of AIC for multivariate linear model when the dimension and the sample size are large. TR No. 12-08, *Statistical Research Group*, *Hiroshima University.*