High-Dimensional AIC in the Growth Curve Model

Yasunori Fujikoshi*, Rie Enomoto** and Tetsuro Sakurai***

**Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, 739-8626, Japan*
***Department of Mathematical Information Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo, 162-8601 Japan*
****Center of General Education, Tokyo University of Science, Suwa, 5000-1 Toyohira, Chino, Nagano 391-0292, Japan*

Abstract

The AIC and its modifications have been proposed for selecting the degree in a polynomial growth curve model under a large-sample framework when the sample size n is large, but the dimension p is fixed. In this paper, first we propose high-dimensional AIC (denoted by HAIC) which is an asymptotic 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 does work in a wide range of p and n. Next we derive asymptotic distributions of AIC and HAIC under the high-dimensional frame work. A sufficient condition is given for that HAIC selects more frequently the true model than AIC. Our results are checked numerically by conducting a Mote Carlo simulation.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: AIC, HAIC, Asymptotic distributions, High-dimensional criteria, Growth curve model.

1. Introduction

We consider the growth curve model introduced by Potthoff and Roy (1964), which is given by

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

where $\mathbf{Y}; n \times p$ is an observation matrix, $\mathbf{A}; n \times q$ is a design matrix across individuals, $\mathbf{X}; q \times p$ is a design matrix within individuals, $\boldsymbol{\Theta}$ is an unknown matrix, and each row of $\mathcal{E}$ is independent and identically distributed as a p-dimensional normal distribution with mean $\mathbf{0}$ and an unknown covariance matrix $\boldsymbol{\Sigma}$. We assume that that $n - p - q - 1 > 0$, and $\mathrm{rank}(\mathbf{X}) = k$. If we consider a polynomial regression of degree $k - 1$ on the time t and with q groups, then

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_q} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_p \\ \vdots & \vdots & \vdots & \vdots \\ t_1^{k-1} & t_2^{k-1} & \cdots & t_p^{k-1} \end{pmatrix}.$$

It is important in a polynomial growth curve model to decide its degree. One way is to treat the problem as the one of selecting models. Related to such problems, consider a set of candidate models $M_1, \ldots, M_k$ where M_j is defined by

$$M_j;\ \mathbf{Y} = \mathbf{A}\boldsymbol{\Theta}_j\mathbf{X}_j + \mathcal{E}, \quad j = 1, \ldots, k, \tag{1.2}$$

where $\boldsymbol{\Theta}_j$ is the $q \times j$ submatrix of $\boldsymbol{\Theta}$, and $\mathbf{X}_j$ is the $j \times p$ submatrix of $\mathbf{X}$ defined by

$$\boldsymbol{\Theta} = (\boldsymbol{\Theta}_j, \boldsymbol{\Theta}_{(j)}), \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_j \\ \mathbf{X}_{(j)} \end{pmatrix}.$$

The AIC (Akaike, 1973) for M_j is given by

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

where $\hat{\boldsymbol{\Sigma}}_j$ is the MLE of $\boldsymbol{\Sigma}$ under M_j, which is given by

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n}(\mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\Theta}}_j\mathbf{X}_j)'(\mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\Theta}}_j\mathbf{X}_j),$$

where $\hat{\boldsymbol{\Theta}}_j = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}\mathbf{S}^{-1}\mathbf{X}_j'(\mathbf{X}_j\mathbf{S}^{-1}\mathbf{X}_j')^{-1}$, $\mathbf{S} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{A})\mathbf{Y}/(n-q)$, and $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. The constant $\{qj + p(p+1)/2\}$ is the number of independent parameters under M_j. In addition to AIC, there are some modifications (see Satoh, Kobayashi and Fujikoshi (1997)) which were proposed as approximately unbiased estimators of AIC-type risk, based on a large-sample theory. The modifications were studied assuming that the true model is included into the largest candidate model M_k.

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

$$n \to \infty, \quad p \to \infty, \quad \frac{p}{n} \to \ c \in [0, 1). \tag{1.4}$$

In this paper we first derive high-dimensional AIC denoted by HAIC which is an asymptotic unbiased estimator of AIC-type risk under (1.4). Note that these criteria are defined for all the subsets by changing the order of the explanatory variables. Next, after we note that these criteria have no consistency property, we obtain asymptotic distributions of AIC and HAIC under (1.4). More precisely, let the values of AIC for model M_j by AIC_j and HAIC_j, and the best subsets chosen by minimizing AIC and HAIC are written as

$$\hat{j}_\mathrm{A} = \arg\min_j \mathrm{AIC}_j, \quad \hat{j}_\mathrm{HA} = \arg\min_j \mathrm{HAIC}_j.$$

Then we shall obtain asymptotic distributions of $\hat{j}_\mathrm{A}$ and $\hat{j}_\mathrm{HA}$. The results include the large-sample asymptotic distributions as their special cases. A sufficient condition is given for that HAIC selects more frequently the true model than AIC. Through simulation experiments, we show that HAIC is better than AIC in the estimation of the risk as well as the probability of selecting the true model.

2. Preliminaries

As is well known, the AIC was proposed as an approximately unbiased estimator of the risk defined by the expected -2 log-predictive likelihood. Let $f(\mathbf{Y};\boldsymbol{\Theta}_j,\boldsymbol{\Sigma}_j)$ be the density function of $\mathbf{Y}$ under M_j. Then the expected log-predictive likelihood of M_j is defined by

$$R_A = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_F}[-2\log f(\mathbf{Y}_F;\hat{\boldsymbol{\Theta}}_j,\hat{\boldsymbol{\Sigma}}_j)] \tag{2.1}$$

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

$$R_A = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_F}[-2\log f(\mathbf{Y};\hat{\boldsymbol{\Theta}}_j,\hat{\boldsymbol{\Sigma}}_j)] + \mathrm{b}_A, \tag{2.2}$$

where

$$\mathrm{b}_A = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_F}[-2\log f(\mathbf{Y}_F;\hat{\boldsymbol{\Theta}}_j,\hat{\boldsymbol{\Sigma}}_j) + 2\log f(\mathbf{Y};\hat{\boldsymbol{\Theta}}_j,\hat{\boldsymbol{\Sigma}}_j)]. \tag{2.3}$$

The AIC and its modifications have been proposed by regarding b_A as the bias term when we estimate R_A by

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

and by evaluating the bias term b_A.

For the justification of AIC, it was assumed that the candidate model M_j include the true model. For the two bias-corrected AICs (see Satoh, Kobayashi and Fujikoshi (1997)), it was assumed that the true model is included in the full model M_k. This assumption is also assumed in this paper. So, without loss of generality, we may assume that the minimum model including the true model is M_{j_0}, and then the true model is expressed as

$$M_{j_0} : \mathbf{Y} \sim N_{n\times p}(\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_{j_0}, \boldsymbol{\Sigma}_0 \otimes \mathbf{I}_n), \tag{2.4}$$

where $\boldsymbol{\Theta}_0$ is a given $q \times j_0$ matrix, and $\boldsymbol{\Sigma}_0$ is a given positive definite matrix. For simplicity, we write $\mathbf{X}_{j_0}$ as $\mathbf{X}_0$. The bias properties of AIC and its modifications have been studied under a large-sample framework,

$$p \text{ and } k \text{ are fixed, } \ n \to \infty. \tag{2.5}$$

In the following we prepare a distributional reduction for the bias b_A. By considering the expectation of $\mathbf{Y}_F$, it is easily seen that

$$\begin{aligned}\mathrm{b}_A = -np + \mathrm{E}^*_{\mathbf{Y}}[\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_0^{-1/2})^{-1}\\ \times \{n\mathbf{I}_p + (\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2} - \hat{\boldsymbol{\Theta}}_j\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2})'\mathbf{A}'\mathbf{A}(\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2} - \hat{\boldsymbol{\Theta}}_j\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2})\}].\end{aligned}$$

Let $\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2); p \times p$ be an orthogonal matrix such that

$$\mathbf{H}_1 = (\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2})'(\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1}\mathbf{X}_j')^{-1/2}, \quad \mathbf{H}_1; p \times j, \quad \mathbf{H}_2; p \times (p-j).$$

Further, define $\mathbf{W}$, $\mathbf{Z}$, $\tilde{\mathbf{Z}}$, $\mathbf{B}$ as follows.

$$\begin{aligned}\mathbf{W} &= \mathbf{H}'\boldsymbol{\Sigma}_0^{-1/2}(n-q)\mathbf{S}\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix},\\ \mathbf{Z} &= (\mathbf{A}'\mathbf{A})^{-1/2}\mathbf{A}'(\mathbf{Y} - \mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0)\boldsymbol{\Sigma}_0^{-1/2},\\ \tilde{\mathbf{Z}} &= \mathbf{Z} + (\mathbf{A}'\mathbf{A})^{-1/2}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2} = (\mathbf{A}'\mathbf{A})^{-1/2}\mathbf{A}'\mathbf{Y}\boldsymbol{\Sigma}_0^{-1/2},\\ \mathbf{B} &= \mathbf{H}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\mathbf{H} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}.\end{aligned}$$

Then, $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as $\mathrm{W}_p(n-q, \mathbf{I}_p)$ and $\mathrm{W}_p(q, \mathbf{I}_p;$ $\mathbf{H}'(\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2})'\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H})$. We can express $\hat{\boldsymbol{\Sigma}}_j$ and its inverse in the terms of W and B as

$$n\boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_0^{-1/2} = \mathbf{H}\begin{pmatrix} \mathbf{I}_j & \mathsf{W}_{12}\mathsf{W}_{22}^{-1} \\ \mathbf{O} & \mathbf{I}_{p-j} \end{pmatrix}\begin{pmatrix} \mathsf{W}_{11\cdot 2} & \mathbf{O} \\ \mathbf{O} & \mathsf{W}_{22}+\mathsf{B}_{22} \end{pmatrix} \times \begin{pmatrix} \mathbf{I}_j & \mathbf{O} \\ \mathsf{W}_{22}^{-1}\mathsf{W}_{21} & \mathbf{I}_{p-j} \end{pmatrix}\mathbf{H}', \quad (2.6)$$

$$(n\boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_0^{-1/2})^{-1} = \mathbf{H}\begin{pmatrix} \mathbf{I}_j & \mathbf{O} \\ -\mathsf{W}_{22}^{-1}\mathsf{W}_{21} & \mathbf{I}_{p-j} \end{pmatrix}\begin{pmatrix} \mathsf{W}_{11\cdot 2}^{-1} & \mathbf{O} \\ \mathbf{O} & (\mathsf{W}_{22}+\mathsf{B}_{22})^{-1} \end{pmatrix} \times \begin{pmatrix} \mathbf{I}_j & -\mathsf{W}_{12}\mathsf{W}_{22}^{-1} \\ \mathbf{O} & \mathbf{I}_{p-j} \end{pmatrix}\mathbf{H}'. \quad (2.7)$$

Lemma 2.1. *Suppose that the true model is given by* (2.4)*. Then, the bias* b_A *for model* M_j *in* (2.2) *or* (2.3) *is expressed as follows:*

$$\mathrm{b}_A = -np + \frac{n(n+q)(n-q-1)j}{(n-q-p-1)(n-q-p+j-1)} + \mathrm{E}\left[n^2\mathrm{tr}(\mathsf{W}_{22}+\mathsf{B}_{22})^{-1}\left(\mathbf{I}+\frac{1}{n}\tilde{\boldsymbol{\Omega}}\right)\right]. \quad (2.8)$$

Here we may assume that W_{22} *and* B_{22} *are independently distributed as* $\mathrm{W}_{p-j}(n-q, \mathbf{I}_{p-j})$ *and* $\mathrm{W}_{p-j}(q, \mathbf{I}_{p-j}; \tilde{\boldsymbol{\Omega}})$*, respectively. The noncentrality matrix is defined by*

$$\tilde{\boldsymbol{\Omega}} = \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \quad (2.9)$$

where $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ *and* $\boldsymbol{\Gamma} = (\mathbf{A}'\mathbf{A})^{1/2}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2$.

Proof. The result is a slight extension of Satoh, Kobayashi and Fujikoshi (1997). From (2.6) and (2.7) we have

$$\mathrm{tr}(n\boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_0^{-1/2})^{-1} = \mathrm{tr}\mathsf{W}_{11\cdot 2}^{-1}(\mathbf{I}_j + \mathsf{W}_{12}\mathsf{W}_{22}^{-2}\mathsf{W}_{21}) + \mathrm{tr}(\mathsf{W}_{22}+\mathsf{B}_{22})^{-1},$$

$$\begin{aligned} &\mathrm{tr}(n\boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_0^{-1/2})^{-1} \\ &\quad\times(\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2} - \hat{\boldsymbol{\Theta}}_j\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2})'\mathbf{A}'\mathbf{A}(\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2} - \hat{\boldsymbol{\Theta}}_j\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2}) \\ &= \mathrm{tr}\mathsf{W}_{11\cdot 2}^{-1}(\mathbf{I}_j, -\mathsf{W}_{12}\mathsf{W}_{22}^{-1})(\mathsf{Z}\mathbf{H})'\mathsf{Z}\mathbf{H}\begin{pmatrix} \mathbf{I}_j \\ -\mathsf{W}_{22}^{-1}\mathsf{W}_{21} \end{pmatrix} \\ &\quad + \mathrm{tr}(\mathsf{W}_{22}+\mathsf{B}_{22})^{-1}(\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2)'\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2. \end{aligned}$$

Therefore, the bias can be computed as

$$\begin{aligned}
\mathrm{b}_A &= -np + n^2\mathrm{E}[\mathrm{tr}\mathbf{W}_{11\cdot 2}^{-1}(\mathbf{I}_j + \mathbf{W}_{12}\mathbf{W}_{22}^{-2}\mathbf{W}_{21}) + \mathrm{tr}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}] \\
&\quad + n\mathrm{E}[\mathrm{tr}\mathbf{W}_{11\cdot 2}^{-1}(\mathbf{I}_j, -\mathbf{W}_{12}\mathbf{W}_{22}^{-1})(\mathbf{ZH})'\mathbf{ZH}\begin{pmatrix} \mathbf{I}_j \\ -\mathbf{W}_{22}^{-1}\mathbf{W}_{21}\end{pmatrix} \\
&\quad + \mathrm{tr}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}(\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2)'\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2] \\
&= -np + n^2\left\{\frac{(n-q-1)j}{(n-q-p-1)(n-q-p+j-1)} + \mathrm{E}[\mathrm{tr}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}]\right\} \\
&\quad + n\left\{\frac{q(n-q-1)j}{(n-q-p-1)(n-q-p+j-1)} + \mathrm{tr}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right\},
\end{aligned}$$

which implies

$$\begin{aligned}
\mathrm{b}_A = -np &+ \frac{n(n+q)(n-q-1)j}{(n-q-p-1)(n-q-p+j-1)} \\
&+ \mathrm{E}\left[n^2\mathrm{tr}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}\left(\mathbf{I}+\frac{1}{n}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)\right].
\end{aligned}$$

The final result is obtained by considering an orthogonal transformation $\mathbf{W}_{22} \to \mathbf{F}'\mathbf{W}_{22}\mathbf{F}$, and $\mathbf{B}_{22} \to \mathbf{F}'\mathbf{B}_{22}\mathbf{F}$, where $\mathbf{F} = (\mathbf{F}_1\ \mathbf{F}_2)$ and $\mathbf{F}_1 = \boldsymbol{\Gamma}'(\boldsymbol{\Gamma}\boldsymbol{\Gamma}')^{1/2}$. □

3. High-dimensional AIC

First we consider to evaluate the bias b_A given in Theorem 2.1 under the high-dimensional framework (1.4). Note that $\mathbf{W}_{22}+\mathbf{B}_{22} \sim \mathrm{W}_{p-j}(n, \mathbf{I}_{p-j}; \tilde{\boldsymbol{\Omega}})$. The order of the noncentrality matrix $\tilde{\boldsymbol{\Omega}}$ is $(p-j)\times(p-j)$, and it tends to infinity. However, the matrix is a sparse matrix. In fact, it is possible to reduce evaluating the expectation with respect to a $q \times q$ random matrix. For such a reduction we use the following Lemma due to Sakurai, Nakada and Fujikoshi (2012), based on Kabe (1964).

Lemma 3.1. *Let* $\mathbf{T} \sim \mathrm{W}_p(n, \mathbf{I}_p; \boldsymbol{\Omega})$, *and* $\mathbf{T}$ *and* $\boldsymbol{\Omega}$ *be partitioned as*

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

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*.*

Using Lemma 3.1 we have

$$\mathrm{E}\left[\mathrm{tr}(\mathsf{W}_{22}+\mathsf{B}_{22})^{-1}\left(\mathbf{I}_{p-j}+\frac{1}{n}\tilde{\boldsymbol{\Omega}}\right)\right] = \frac{n-q-1}{n-(p-j)-1}\mathrm{b}_{A1} + \frac{p-j-q}{n-(p-j)-1},$$

where

$$\mathrm{b}_{A1} = \mathrm{E}\left[\mathrm{tr}\mathbf{T}^{-1}\left(\mathbf{I}_q + \frac{1}{n}\boldsymbol{\Omega}\right)\right], \tag{3.1}$$

and $\mathbf{T}$ is the upper $q \times q$ submatrix of $\mathsf{W}_{22}+\mathsf{B}_{22}$. Note that $\mathbf{T}$ is distributed as $\mathrm{W}_q(n, \mathbf{I}_q; \boldsymbol{\Omega})$. When $\boldsymbol{\Omega} = \mathbf{O}$, $\mathbf{T} \sim \mathrm{W}_q(n, \mathbf{I}_q)$, and hence,(see Muirhead (1982), Siotani, Hayakawa and Fujikoshi (1985), etc.)

$$\mathrm{b}_{A1} = \mathrm{E}[\mathrm{tr}\mathbf{T}^{-1}] = \frac{q}{n-q-1}. \tag{3.2}$$

In general, we assume that

$$\boldsymbol{\Omega} = n\boldsymbol{\Delta} = \mathrm{O}_h(n), \tag{3.3}$$

where $\mathrm{O}_h(n^i)$ denotes the terms of i-th order with respect to n under (1.4). Then, the expectation can be expanded, by a result (see, e.g., Fujikoshi (1985)) based a perturbation method, as follows:

$$\mathrm{b}_{A1} = \frac{q}{n} + \frac{1}{n^2}\left\{2(q+1)\xi_1 - \xi_2\right\} + \mathrm{O}_h(n^{-3}), \tag{3.4}$$

where $\xi_1 = \eta_1,\ \xi_2 = \eta_1^2 + \eta_2$ and $\eta_i = \mathrm{tr}\left(\mathbf{I}_q + \boldsymbol{\Omega}/n\right)^{-i},\ i = 1, 2$. Summarizing the result, we have the following theorem.

Theorem 3.1. *Suppose that the true model is expressed as in* (2.4)*. Then, under* (1.4) *and* (3.3) *the bias term* b_A *in* (2.3) *can be expressed as*

$$\begin{aligned}\mathrm{b}_A = &-np + \frac{n(n+q)(n-q-1)j}{(n-q-p-1)(n-q-p+j-1)} \\ &+ n^2 \left[\frac{p-j-q}{n-(p-j)-1} + \frac{n-q-1}{n-(p-j)-1}\mathrm{b}_{A1} \right], \end{aligned} \tag{3.5}$$

where b_{A1} *is given by* (3.4)*. When* $\boldsymbol{\Omega} = \mathbf{O}$*, the term* b_{A1} *is exactly expressed as* (3.2)*.*

For a practical use we derive asymptotic unbiased estimators for ξ_1 and ξ_2 under (1.4). We have seen that

$$\begin{aligned}\mathrm{tr}(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S} &= j + \mathrm{tr}\mathbf{W}_{22}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}, \\ \mathrm{tr}\left\{(n\hat{\boldsymbol{\Sigma}}_j)^{-1}\{(n-q)\mathbf{S}\right\}^2 &= j + \mathrm{tr}\left\{\mathbf{W}_{22}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}\right\}^2.\end{aligned}$$

For a reduction of the right-hand sides, we use the following Lemma.

Lemma 3.2. *Let* $\mathbf{S}_h = \mathbf{X}'\mathbf{X}$ *and* $\mathbf{S}_e$ *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* $\mathbf{X}$ *is a* $q \times p$ *random matrix whose elements are independent normal variates with* $\mathrm{E}(\mathbf{X}) = \mathbf{M}$ *and the common variance* 1*. Put*

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

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 3.2 was essentially obtained in Wakaki, Fujikoshi and Ulyanov (2002) and Fujikoshi, Ulyanov and Shimizu (2010). Lemma 3.2 implies that

$$\mathrm{tr}\mathbf{W}_{22}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}=p-j-q+\mathrm{tr}\mathbf{V}(\mathbf{V}+\mathbf{U})^{-1},$$
$$\mathrm{tr}\left\{\mathbf{W}_{22}(\mathbf{W}_{22}+\mathbf{B}_{22})^{-1}\right\}^2=p-j-q+\mathrm{tr}\left\{\mathbf{V}(\mathbf{V}+\mathbf{U})^{-1}\right\}^2,$$

where $\mathbf{V}\sim \mathrm{W}_q(n-(p-j),\mathbf{I}_q)$ and $\mathbf{U}\sim \mathrm{W}_q(p-j,\mathbf{I}_q;\boldsymbol{\Omega})$. Let $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{U}}$ be defined by

$$\frac{1}{n-(p-j)}\mathbf{V}=\mathbf{I}_q+\frac{1}{\sqrt{n-(p-j)}}\tilde{\mathbf{V}},$$
$$\frac{1}{p-j}\mathbf{U}=\mathbf{I}_q+\frac{n}{p-j}\Delta+\frac{1}{\sqrt{p-j}}\tilde{\mathbf{U}}.$$

Then $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{U}}$ are asymptotically normal, and $\mathrm{E}[\tilde{\mathbf{V}}]=0,\ \ \mathrm{E}[\tilde{\mathbf{U}}]=0$. Therefore

$$\mathbf{V}(\mathbf{V}+\mathbf{U})^{-1}\to\frac{n-(p-j)}{n}(\mathbf{I}_q+\Delta)^{-1}.$$

We define

$$\tilde{\xi}_1=\tilde{\eta}_1,\quad \tilde{\xi}_2=\tilde{\eta}_1^2+\tilde{\eta}_2.$$

where

$$\tilde{\eta}_1=\frac{n}{n-(p-j)}\left\{\mathrm{tr}(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S}-(p-q)\right\},$$
$$\tilde{\eta}_2=\left(\frac{n}{n-(p-j)}\right)^2\left\{\mathrm{tr}\{(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S}\}^2-(p-q)\right\}.$$

Then, $\tilde{\xi}_1$ and $\tilde{\xi}_2$ are asymptotically unbiased estimators of ξ_1 and ξ_2, respectively. Now we modify these estimators so that they are exact estimators when $\mathbf{\Omega} = \mathbf{O}$. Put $\mathbf{Q} = \mathbf{V}(\mathbf{V} + \mathbf{U})^{-1}$. Then, when $\mathbf{\Omega} = \mathbf{O}$, $\mathbf{Q}$ is distributed as a multivariate beta distribution $\beta_q(a/2, b/2)$ with $a = n - (p - j)$ and $b = p - j$. Therefore, we have (see, e.g. Fujikoshi and Satoh (1997))

$$\begin{aligned}
\mathrm{E}[\mathrm{tr}\mathbf{Q}] &= \frac{n - (p - j)}{n} \cdot q, \\
\mathrm{E}[(\mathrm{tr}\mathbf{Q})^2 + \mathrm{tr}\mathbf{Q}^2] &= \frac{n - (p - j)}{6n} \cdot q \cdot \left\{ \frac{4(n - (p - j) + 2)(q + 2)}{n + 2} \right. \\
&\qquad \left. + \frac{2(n - (p - j) - 1)(q - 1)}{n - 1} \right\}.
\end{aligned}$$

Since $\tilde{\xi}_1$ is an unbiased estimator of ξ_1, we modify $\tilde{\xi}_1$ and $\tilde{\xi}_2$ as

$$\hat{\xi}_1 = \tilde{\xi}_1 \quad \text{and} \quad \hat{\xi}_2 = d\tilde{\xi}_2,$$

so that $\hat{\xi}_2$ is an unbiased estimator of ξ_2 when $\mathbf{\Omega} = \mathbf{O}$. For this, from (3.2), (3.4) and (3.5) we may determine d as

$$\frac{1}{n^2} \left\{ 2(q + 1)q - d\mathrm{E}[\tilde{\xi}_2] \right\} = \frac{q(q + 1)}{n(n - q - 1)},$$

since $q/(n - q - 1) = q/n + q(q + 1)/\{n(n - q - 1)\}$. This implies that

$$\begin{aligned}
d = (q + 1) \cdot \frac{n - 2q - 2}{n - q - 1} \cdot \frac{3\{n - (p - j)\}}{n} \\
\times \left[\frac{2(q + 2)\{n - (p - j) + 2\}}{n + 2} + \frac{(q - 1)\{n - (p - j) - 1\}}{n - 1} \right]^{-1}.
\end{aligned} \tag{3.6}$$

Now we define HAIC for M_j as

$$\mathrm{HAIC}_j = n \log |\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + \hat{\mathrm{b}}_j, \tag{3.7}$$

where

$$\begin{aligned}
\hat{\mathrm{b}}_j = -np + \frac{n(n + q)(n - q - 1)j}{(n - q - p - 1)(n - q - p + j - 1)} \\
+ n^2 \left[\frac{p - j - q}{n - (p - j) - 1} + \frac{n - q - 1}{n - (p - j) - 1} \right. \\
\left. \times \left(\frac{q}{n} + \frac{1}{n^2} \left\{ 2(q + 1)\hat{\xi}_1 - \hat{\xi}_2 \right\} \right) \right].
\end{aligned} \tag{3.8}$$

Then we separate a set of candidate models, one of which is a set of overspecified models, candidate models that include the true model, and the other is a set of underspecified models that are not the overspecified models. Note that if M_j is an overspecified model, then $\mathbf{\Omega} = \mathbf{O}$.

Theorem 3.2. *Under assumption* (1.4) *the high-dimensional* AIC*,* HAIC *defined by* (3.7) *satisfies the following properties:*
(1) *if M_j is an overspecified model,* HAIC *is an exact unbiased estimator of R_A,*i.e.

$$\mathrm{E(HAIC)} = R_A.$$

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

$$\mathrm{E(HAIC)} = R_A + \mathrm{O}_h(n^{-1}).$$

Note that these HAIC is defined for all the subsets by changing the order of the explanatory variables.

4. Asymptotic distributions of AIC and HAIC

Recently it is known (see Fujikoshi, Sakurai and Yanagihara (2012), Yanagihara, Wakaki and Fujikoshi (2012)) that AIC and its modifications for some multivariate models have consistency property under a high-dimensional asymptotic framework. First we examine whether AIC and HAIC have a consistency property under the high-dimensional framework (1.4). Note that

$$\frac{1}{n}\mathbf{V} = \frac{n-(p-j)}{n} \cdot \frac{1}{n-(p-j)}\mathbf{V} \to (1-c)\mathbf{I}_q,$$
$$\frac{1}{n}\mathbf{U} = \frac{p-j}{n} \cdot \frac{1}{p-j}\mathbf{U} \to c\mathbf{I}_q + \Delta_j.$$

This implies that

$$-\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\mathbf{\Sigma}}_j|} = -\log\frac{|\mathbf{W}_{22}|}{|\mathbf{W}_{22}+\mathbf{B}_{22}|} = -\log\frac{|\mathbf{V}|}{|\mathbf{V}+\mathbf{U}|}$$
$$\to -\log\frac{|(1-c)\mathbf{I}_q|}{|(1-c)\mathbf{I}_q + c\mathbf{I}_q + \Delta_j|} = -\log(1-c)^q + \log|\mathbf{I}_q + \Delta_j|.$$

Therefore we have

$$
\begin{aligned}
&\frac{1}{n}(\mathrm{AIC}_j - \mathrm{AIC}_{j_0}) \\
&\quad = -\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\mathbf{\Sigma}}_j|} - \left(-\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\mathbf{\Sigma}}_0|}\right) + \frac{2q(j-j_0)}{n} \\
&\quad \to \log|\mathbf{I}_q + \Delta_j| = \begin{cases} 0, & j \geq j_0, \\ >0, & j < j_0. \end{cases}
\end{aligned}
$$

Here, AIC_j and AIC_{j_0} is AIC under M_j and M_{j_0}, respectively. This shows that AIC has no consistency property. Similarly HAIC has no consistency property since $(\hat{\mathrm{b}}_j - \hat{\mathrm{b}}_{j_0})/n \to 0$ (see Section 4.2). Further, it holds that for $\ell = 1, \ldots, j_0 - 1$,

$$
\lim P(\hat{j}_{\mathrm{A}} = \ell) = 0, \text{ and } \lim P(\hat{j}_{\mathrm{HA}} = \ell) = 0. \tag{4.1}
$$

4.1 Asymptotic distribution of AIC

We have defined an orthogonal matrix $(\mathbf{H}_1\ \mathbf{H}_2)$ for each $\mathrm{M}_j (j = 1, \ldots, k)$, which is denoted by $(\mathbf{H}_{1j}\ \mathbf{H}_{2j})$, where $\mathbf{H}_{1j} = (\mathbf{X}_0\mathbf{\Sigma}_0^{-1/2})'(\mathbf{X}_0\mathbf{\Sigma}_0^{-1}\mathbf{X}_0')^{-1/2}$; $p \times j$ and $\mathbf{H}_{2j}$ is any matrix such that $(\mathbf{H}_{2j})'\mathbf{H}_{2j} = \mathbf{I}_{p-j}$, and $(\mathbf{H}_{2j})'\mathbf{H}_{1j} = \mathbf{O}$. In order to treat $\mathrm{AIC}_1, \mathrm{AIC}_2, \ldots, \mathrm{AIC}_k$ simultaneously, we define an orthogonal matrix

$$
\mathbf{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k; *),
$$

satisfying $\boldsymbol{h}_1 \in \mathcal{R}[\mathbf{H}_{11}], (\boldsymbol{h}_1, \boldsymbol{h}_2) \in \mathcal{R}[\mathbf{H}_{12}], \cdots, (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k) \in \mathcal{R}[\mathbf{H}_{1k}]$, and the remainder $p - k$ columns are any ones such that $\mathbf{H}$ is an orthogonal matrix. We partition $\mathbf{H}$ as

$$
\mathbf{H} = (\mathbf{H}_1^{(j)}\ \mathbf{H}_2^{(j)}),\ \mathbf{H}_1^{(j)}; p \times j.
$$

Using the new orthogonal matrix $\mathbf{H}$, we consider the random matrices in Section 2 by using the same notations. For example,

$$
\mathbf{W} = \mathbf{H}'\mathbf{\Sigma}_0^{-1/2}(n-q)\mathbf{S}\mathbf{\Sigma}_0^{-1/2}\mathbf{H}, \quad \mathbf{B} = \mathbf{H}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\mathbf{H}.
$$

Let us denote the last $(p-j)\times(p-j)$ submatrices of $\mathbf{W}$ and $\mathbf{B}$ by $\mathbf{W}_{(j)}$ and $\mathbf{B}_{(j)}$, respectively, that is

$$\mathbf{W}=\begin{pmatrix} * & * \\ * & \mathbf{W}_{(j)} \end{pmatrix}, \quad \mathbf{B}=\begin{pmatrix} * & * \\ * & \mathbf{B}_{(j)} \end{pmatrix}.$$

Then

$$\mathbf{W}_{(j)} \sim \mathrm{W}_{p-j}(n-q,\mathbf{I}_{p-j}), \quad \mathbf{B}_{(j)} \sim \mathrm{W}_{p-j}(q,\mathbf{I}_{p-j};\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j),$$

where $\boldsymbol{\Gamma}_j=(\mathbf{A}'\mathbf{A})^{1/2}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)}$. Further, from Lemma 3.2 it is possible to express as

$$\frac{|(n-q)\mathbf{S}|}{|n\hat{\boldsymbol{\Sigma}}_j|}=\frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)}+\mathbf{B}_{(j)}|}=\frac{|\mathbf{V}_{(j)}|}{|\mathbf{V}_{(j)}+\mathbf{U}_{(j)}|},$$

where $\mathbf{V}_{(j)}$ and $\mathbf{U}_{(j)}$ are independently distributed as $\mathrm{W}_q(n-(p-j),\mathbf{I}_q)$ and $\mathrm{W}_q(p-j,\mathbf{I}_q;\boldsymbol{\Omega}_j)$, and $\boldsymbol{\Omega}_j=\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$.

Lemma 4.1. *Suppose that the true model is expressed as* (2.4)*. Then, under* (1.4)*,* $\mathrm{AIC}_j-\mathrm{AIC}_{j_0}$ *is asymptotically distributed as*

$$-\frac{1}{1-c}\left[\{R_1-2(1-c)q\}+\cdots+\{R_{j-j_0}-2(1-c)q\}\right],\ j=j_0+1,\ldots,k,$$

where $R_1,\ldots,R_{k-j_0}$ *are independently distributed as* χ_q^2*.*

Proof. For $j=j_0+1,\ldots,k$, we have

$$\begin{aligned}\mathrm{AIC}_j&-\mathrm{AIC}_{j_0}\\ &=-n\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\boldsymbol{\Sigma}}_j|}-\left(-n\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\boldsymbol{\Sigma}}_{j_0}|}\right)+2(j-j_0)q\\ &=-n\log\Lambda_{(j)}+n\log\Lambda_{(j_0)}+2(j-j_0)q,\end{aligned}$$

where

$$\Lambda_{(j)}=\frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)}+\mathbf{B}_{(j)}|}\sim\Lambda_{p-j,q,n-q}, \quad j=j_0,j_0+1,\ldots,k.$$

Therefore

$$\begin{aligned}
&\mathrm{AIC}_j - \mathrm{AIC}_{j_0}\\
&\quad = n\Big\{(\log\Lambda_{(j_0)} - \log\Lambda_{(j_0+1)}) + (\log\Lambda_{(j_0+1)} - \log\Lambda_{(j_0+2)})\\
&\qquad + \cdots + (\log\Lambda_{(j-1)} - \log\Lambda_{(j)})\Big\} + 2(j-j_0)q\\
&\quad = \sum_{\ell=1}^{j-j_0} n\log\Lambda_{(j_0+\ell-1|j_0+\ell)} + 2(j-j_0)q,
\end{aligned}$$

where $\Lambda_{(j_0+\ell-1|j_0+\ell)} = \Lambda_{(j_0+\ell-1)}/\Lambda_{(j_0+\ell)}$, $\ell = 1,\ldots,j-j_0$. It is shown (see Fujikoshi, Ulyanov and Shimizu (2010), p.61 $\sim$ 63) that $\Lambda_{(j_0-1|j_0)}, \Lambda_{(j_0|j_0+1)}, \ldots, \Lambda_{(j-1|j)}$ are independent, and for $\ell = 1,\ldots,j-j_0$

$$\Lambda_{(j_0+\ell-1|j_0+\ell)} = \frac{\chi^2_{n-q-(p-j_0-\ell)}}{\chi^2_q + \chi^2_{n-q-(p-j_0-\ell)}} = \left(1 + \frac{\chi^2_q}{\chi^2_{n-q-(p-j_j-\ell)}}\right)^{-1}.$$

Therefore, $-(n-p)\log\Lambda_{(j_0+\ell-1|j_0+\ell)}$ is asymptotically distributed as χ^2_q. Using this property we can get the required result. □

Theorem 4.1. *Suppose that the true model is expressed as* (2.4)*. Let* $R_1,\ldots,R_{k-j_0}$ *be independent* χ^2_q *variables, and put*

$$Z_\ell = R_\ell - 2(1-c)q,\ \ \ell = 1,\ldots,k-j_0. \tag{4.2}$$

Then, under (1.4) *it holds that*

$$\begin{aligned}
\lim P(\hat{j}_{\mathrm{A}} = j) &= P(L_{j-j_0} - L_0 > 0,\ldots, L_{j-j_0} - L_{j-j_0-1} > 0)\\
&\quad \times P(L_{j+1-j_0} - L_{j-j_0} < 0,\ldots, L_{k-j_0} - L_{j-j_0} < 0),
\end{aligned} \tag{4.3}$$

where $L_0 = 0$*,* $L_{i-j_0} = \sum_{\ell=1}^{i-j_0} Z_\ell, i = j_0+1,\ldots,k$*. Further, the probability* (4.3) *can be expressed as*

$$\lim P(\hat{j}_{\mathrm{A}} = j) = a_{j-j_0} b_{k-j_0}, \tag{4.4}$$

where $a_0 = b_0 = 1$, $\alpha_i = P(L_i \geq 0)$, *and*

$$a_m = \sum_{[m]} \left\{ \prod_{i=1}^{m} \frac{1}{r_i!} \left(\frac{\alpha_i}{i} \right)^{r_i} \right\}, \quad b_m = \sum_{[m]} \left\{ \prod_{i=1}^{m} \frac{1}{r_i!} \left(\frac{1-\alpha_i}{i} \right)^{r_i} \right\}. \tag{4.5}$$

Here the summation $\sum_{[m]}$ *extends over all* m*-tuples* $(r_1, r_2, \ldots, r_m)$ *of non-negative integers with the property* $r_1 + 2r_2 + \cdots + mr_m = m$.

Proof. Note that

$$\begin{aligned} &\lim P(\hat{j}_{\rm A} = j) = \lim P(\min_{\ell} {\rm AIC}_\ell = {\rm AIC}_j) \\ &= \lim P({\rm AIC}_j - {\rm AIC}_{j_0} \leq {\rm AIC}_\ell - {\rm AIC}_{j_0},\ \ell = j_0, j_0+1, \ldots, k). \end{aligned}$$

Using Lemma 4.1 and noting that $\{R_1, \ldots, R_{j-j_0}\}$ and $\{R_{j+1-j_0}, \ldots, R_{k-j_0}\}$ are independent, we can write the probability as (4.3). Further, it is easy to see that

$$\begin{aligned} &P(L_{j-j_0} - L_0 > 0, \ldots, L_{j-j_0} - L_{j-j_0-1} > 0) \\ &\quad = P\Bigl(\sum_{i=1}^{j-j_0} Z_i > 0, \sum_{i=2}^{j-j_0} Z_i > 0, \ldots, \sum_{i=j-j_0-1}^{j-j_0} Z_i > 0 \Bigr), \\ &P(L_{j+1-j_0} - L_{j-j_0} < 0, \ldots, L_{k-j_0} - L_{j-j_0} < 0) \\ &\quad = P\Bigl(\sum_{i=j-j_0+1}^{j-j_0+1} Z_i < 0, \ldots, \sum_{i=j-j_0+1}^{k-j_0} Z_i < 0 \Bigr). \end{aligned}$$

Applying Spitzer (1956) and Shibata (1976) to the above expressions, we obtain the second result (4.4). □

We note that Theorem 4.1 with $c = 0$ gives asymptotic distribution of AIC under large-sample framework, which was considered by Satoh, Kobayashi and Fujikoshi (1997).

Especially the probability of selecting the true model by AIC is expressed as

$$\lim P(\hat{j}_{\rm A} = j_0) = b_{k-j_0} = \sum_{[k-j_0]} \prod_{i=1}^{k-j_0} \frac{1}{r_i!} \left(\frac{1-\alpha_i}{i} \right)^{r_i},$$

where $\alpha_i = P(L_i \geq 0)$.

For numerical computations of the asymptotic probabilities, we can use simplified expressions of a_m and b_m. For example, consider

$$a_m = \sum_{[m]} \left\{ \prod_{i=1}^{m} \frac{1}{r_i!} \left(\frac{\alpha_i}{i} \right)^{r_i} \right\}.$$

Then, the a_m's for $m = 1 \sim 5$ are expressed as follows.

$$\begin{aligned}
a_1 &= \alpha_1, \\
a_2 &= \frac{1}{2} \left(\alpha_1^2 + \alpha_2 \right), \\
a_3 &= \frac{1}{6} \left(\alpha_1^3 + 3\alpha_2\alpha_1 + 2\alpha_3 \right), \\
a_4 &= \frac{1}{24} \left(\alpha_1^4 + 6\alpha_2\alpha_1^2 + 8\alpha_3\alpha_1 + 3\alpha_2^2 + 6\alpha_4 \right), \\
a_5 &= \frac{1}{120} \left(\alpha_1^5 + 10\alpha_2\alpha_1^3 + 20\alpha_3\alpha_1^2 + 15\alpha_2^2\alpha_1 \right. \\
&\quad \left. + 30\alpha_4\alpha_1 + 20\alpha_2\alpha_3 + 24\alpha_5 \right).
\end{aligned}$$

4.2 Asymptotic Distribution of HAIC

The main part of HAIC is the same as the one of AIC. So, it is enough to examine asymptotic behaviors of $\hat{\mathrm{b}}_j$ and $\hat{\mathrm{b}}_{j_0}$, where $\hat{\mathrm{b}}_j$ is given by (3.8). Note that $\hat{\xi}_1$ and $\hat{\xi}_2$ converge to ξ_1 and ξ_2, respectively. After much computation, we get

$$\begin{aligned}
\hat{\mathrm{b}}_j - \hat{\mathrm{b}}_{j_0} &= (j - j_0) \left[2q \frac{n^3}{(n-p)^3} + \{2 - (j + j_0)\} \frac{pn^2}{(n-p)^3} \right] \\
&= (j - j_0)(1 - c)^{-3} \left[2q + c\{2 - (j + j_0)\} \right] + \mathrm{o}_h(1).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{HAIC}_j - \mathrm{HAIC}_{j_0} \to & -\frac{1}{1-c} \{R_1 + \cdots + R_{j-j_0}\} \\
& + (j - j_0)(1 - c)^{-3} \left[2q + c\{2 - (j + j_0)\} \right] \\
= & -\frac{1}{1-c} \left[\{R_1 - g_j\} + \cdots + \{R_{j-j_0} - g_j\} \right],
\end{aligned}$$

where

$$g_j = (1-c)^{-2}[2q + c\{2 - (j + j_0)\}]. \tag{4.6}$$

This shows that an asymptotic result of HAIC can be obtained from the one of AIC by changing $Z_\ell = R_\ell - 2(1-c)q$ to $Z_\ell = R_\ell - g_j$. We have the following theorem.

Theorem 4.2. *Suppose that the true model is expressed as* (2.4)*. Let* $R_1, \ldots, R_{k-j_0}$ *be independent* χ_q^2 *variables, and put*

$$Z_\ell = R_\ell - g_j, \ \ \ell = 1, \ldots, k - j_0. \tag{4.7}$$

where g_j *is given by* (4.6)*. Then, under* (1.4) *it holds that* $\lim P(\hat{\jmath}_{\mathrm{HA}} = j)$ *is given by* (4.3) *or* (4.4) *with* Z_ℓ *in* (4.7)*.*

Especially the probability of selecting the true model by HAIC is expressed a

$$\lim P(\hat{\jmath}_{\mathrm{HA}} = j_0) = \tilde{b}_{k-j_0} = \sum_{[k-j_0]} \prod_{i=1}^{k-j_0} \frac{1}{r_i!} \left(\frac{1 - \tilde{\alpha}_i}{i} \right)^{r_i},$$

where $\tilde{\alpha}_i = P(\tilde{L}_i \geq 0)$.

Now we examine which of AIC and HAIC has the more high probability of selecting the true model. For this, let us consider the condition such that

$$\lim P(\hat{\jmath}_{\mathrm{A}} = j_0) \leq \lim P(\hat{\jmath}_{\mathrm{HA}} = j_0).$$

A sufficient condition is given in the following theorem.

Theorem 4.3. *Under the same assumptions as in Theorems* 4.1 *and* 4.2 *it holds that*

(1) *if* $g_j = (1-c)^{-2}[2q + c\{2 - (j + 2j_0)\}] \geq 2(1-c)q$, $j = j_0 + 1, \ldots, k$, *then*

$$\lim P(\hat{\jmath}_{\mathrm{A}} = j_0) \leq \lim P(\hat{\jmath}_{\mathrm{HA}} = j_0).$$

(2) *if* $q + 1 \geq k$, *then*

$$g_j = (1-c)^{-2}[2q + c\{2 - (i + 2j_0)\}] \geq 2(1-c)q, \quad j = 1 + j_0, \ldots, k.$$

Proof. From Theorem 4.1 the asymptotic probability of selecting the true model by AIC is expressed as

$$\begin{aligned}\lim P(\hat{j}_{\mathrm{A}} = j_0) &= P(L_1 < 0, L_2 < 0, \dots, L_{k-j_0} < 0)\\ &= P(S_1 < s_1, S_2 < s_2, \dots, S_{k-j_0} < s_{k-j_0}),\end{aligned}$$

where $L_{j-j_0} = \sum_{i=1}^{j-j_0}\{R_i - 2(1-c)q\}$, $j = j_0+1, \dots, k$, and $R_1, \dots, R_{k-j_0}$ are independent χ_q^2 variables. Here $S_{j-j_0} = \sum_{i=1}^{j-j_0} R_i$ and $s_{j-j_0} = 2(j-j_0)(1-c)q$. Similarly, the asymptotic probability of selecting the true model by HAIC is expressed as

$$\begin{aligned}\lim P(\hat{j}_{\mathrm{HA}} = j_0) &= P(\tilde{L}_1 < 0, \tilde{L}_2 < 0, \dots, \tilde{L}_{k-j_0} < 0)\\ &= P(S_1 < \tilde{s}_1, S_2 < \tilde{s}_2, \dots, S_{k-j_0} < \tilde{s}_{k-j_0}),\end{aligned}$$

where $\tilde{L}_{j-j_0} = \sum_{i=1}^{j-j_0}\{R_i - g_j\}, j = j_0 + 1, \dots, k$. Here $\tilde{s}_{j-j_0} = (j - j_0)g_j$. Suppose that for $j = 1+j_0, \dots, k$, $g_j \geq 2(1-c)q$. Then we have $\tilde{s}_{j-j_0} \geq s_{j-j_0}$. Therefore

$$\begin{aligned}\lim P(\hat{j}_{\mathrm{A}} = j_0) &= P(S_1 < \tilde{s}_1, S_2 < \tilde{s}_2, \dots, S_{k-j_0} < \tilde{s}_{k-j_0})\\ &\geq P(S_1 < s_1, S_2 < s_2, \dots, S_{k-j_0} < s_{k-j_0}) = \lim P(\hat{j}_{\mathrm{A}} = j_0),\end{aligned}$$

which shows the first result. For $j = j_0 + 1, \dots, k$, we have

$$\begin{aligned}g_j - 2(1-c)q &= (1-c)^{-2}[2q + c\{2 - (i + 2j_0)\}] - 2(1-c)q\\ &= (1-c)^{-2}[2q\{1 - (1-c)^3\} + c\{2 - (j + j_0)\}]\\ &\geq (1-c)^{-2}[2q\{1 - (1-c)^3\} + c\{2 - 2k\}]\\ &= 2c(1-c)^{-2}[q(c^2 - 3c + 3) + 1 - k].\end{aligned}$$

Here we used that $1 + j_0 \leq j \leq k$ and $1 \leq j_0 \leq k$. Further, noting that $1 < c^2 - 3c + 3 < 3$ since $0 < c < 1$, $q(c^2 - 3c + 3) + 1 - k > q + 1 - k$. So, if $q + 1 - k \geq 0$, $g_j \geq 2(1-c)q$. This proves the second result. □

Through numerical experiments, we have seen (see Section 5) that HAIC selects the true model more often than AIC. So, it is expected that the sufficient condition is relaxed.

5. Simulation study

In this section, we numerically examine the validity of our claim. The five candidate models $M_1, \ldots, M_5$, with several different values of q, $n = \sum_{i=1}^{q} n_i$ and $p = cn$, were prepared for Monte Carlo simulations, where $q = 2$, $n_1 = n_2 = 50, 100$ and $c = 0.1, 0.2, 0.4$. We constructed a $5 \times p$ matrix $\mathbf{X}$ of explanatory variables with $t_i = 1 + i/p$. The true model was determined by $\mathbf{\Theta}_0 = \mathbf{1}_3\mathbf{1}_3'$ and $\mathbf{\Sigma}_0$ whose (i,j)th element was defined by $(0.2)^{|i-j|}$ $(i = 1, \ldots, p; j = 1, \ldots, p)$. Here $\mathbf{1}_p$ was the p-dimensional vector of ones.

First, we studied performances of AIC and HAIC as estimators of R_A. For each of $M_1, \ldots, M_5$, we computed the averages of R_A, AIC and HAIC by Monte Carlo simulations with 10^4 replications. Table 1 shows the risk R_A and the biases of AIC and HAIC to R_A, defined by "$R_A -$ (the expectation of the information criterion)". In the table, j means the model M_j and the bold face denotes the true model. From the table, we can see that the biases of the HAIC were smaller than the ones of AIC. In general, there is a tendency that the biases are large as p increases. But the tendency of HAIC is very weak in the comparison with AIC.

Table 1. Risks and biases of AIC and HAIC

j	R_A	AIC	HAIC	R_A	AIC	HAIC	R_A	AIC	HAIC
	$(n,p) = (100,10)$			$(n,p) = (100,20)$			$(n,p) = (100,40)$		
1	3131.9	13.8	0.5	6218.9	113.8	0.4	13390.0	1148.1	-4.0
2	2881.1	16.8	0.2	5924.4	119.4	-0.1	13056.0	1167.7	-2.1
3	2877.7	18.1	0.2	5919.9	122.9	-0.1	13053.8	1180.7	-2.9
4	2880.6	19.2	0.2	5924.7	126.2	0.0	13067.3	1193.4	-3.1
5	2883.4	20.2	0.2	5929.3	129.2	0.0	13080.1	1205.4	-3.2
	$(n,p) = (200,20)$			$(n,p) = (200,40)$			$(n,p) = (200,80)$		
1	12070.9	50.5	1.7	24305.7	421.1	-3.4	52735.0	4416.4	-6.1
2	11474.5	54.5	2.2	23607.0	427.0	-3.5	51943.4	4434.2	-6.0
3	11458.4	56.1	2.4	23580.1	430.8	-3.5	51902.5	4448.6	-5.9
4	11461.6	57.5	2.5	23585.1	434.1	-3.7	51916.9	4462.3	-5.9
5	11464.6	58.7	2.5	23590.0	437.5	-3.7	51931.0	4475.5	-5.9

Next, we studied the probabilities of selecting the model by AIC and HAIC. Table 2 shows the selection probability by AIC based on Monte Carlo simulations with 10^4 iterations and the asymptotic selection probabilities under a large-sample framework and a high-dimensional framework. In the table, AIC_1 and AIC_2 denote the selection probabilities by Monte Carlo simulations for $n = 100$ and $n = 200$, respectively. The columns of LS and HD express the asymptotic selection probabilities under a large-sample framework and a high-dimensional framework, respectively. Here, we calculated LS and HD as $c = 0.0$ and $c = p/n$, respectively. In the table, the true model is M_3. Similarly, Table 3 shows the selection probabilities of HAIC.

From Tables 2 and 3 we can see that the selection probabilities of AIC and HAIC were closer to HD than LS. Moreover, these selection probabilities approached more to HD as p increases. Both of the probabilities of selecting the true model increase when n increases, but p is fixed. However, the probability of selecting the true model by AIC decreases as p increases, but n is fixed. On the other hand, the probability of selecting the true model by HAIC increases when n and p increases. Especially, the probability approaches to 1. We have shown in Theorem 4.3 that if $q + 1 > k$, the probability of selecting the true model by HAIC is higher than the one by AIC. Moreover, Tables 2 and 3 show that the fact holds in all the cases except for $(n, p) = (100, 10)$. This means that it is expected that a more weak sufficient condition shall be derived.

Table 2. Selection probabilities of AIC (%)

j		$c = 0.1$			$c = 0.2$			$c = 0.4$		
	LS	HD	AIC_1	AIC_2	HD	AIC_1	AIC_2	HD	AIC_1	AIC_2
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	17.9	0.5	0.0	8.1	0.0	0.0	1.9	0.0
3	82.8	78.6	63.0	78.4	73.3	66.7	74.6	59.0	59.2	59.8
4	11.7	13.8	12.5	14.0	16.1	15.3	15.4	21.0	20.4	21.2
5	5.5	7.7	6.6	7.2	10.6	9.8	10.1	20.0	18.5	19.0

Table 3. Selection probabilities of HAIC (%)

j		$c = 0.1$			$c = 0.2$			$c = 0.4$		
	LS	HD	HAIC_1	HAIC_2	HD	HAIC_1	HAIC_2	HD	HAIC_1	HAIC_2
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	28.6	1.3	0.0	28.1	0.1	0.0	49.0	0.2
3	82.8	88.0	61.2	87.6	92.6	66.7	94.5	98.6	50.8	99.5
4	11.7	8.7	7.2	8.4	5.6	4.1	4.5	1.2	0.2	0.3
5	5.5	3.3	3.1	2.8	1.7	1.1	0.9	0.2	0.0	0.0

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

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

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

[5] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2012). High-dimensional AIC and consistency properties of several criteria in multivariate linear regression. TR No. 12-14, *Statistical Research Group, Hiroshima University.*

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

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

[8] Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

[9] 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.*

[10] Satoh, K., Kobayashi, M. and Fujikoshi, Y. (1997). Variable selection for the growth curve model. *J. Multivariate Anal.*, **60**, 277–292.

[11] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.

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

[13] Spitzer, F. (1956). A combinatorial lemma and its application to probability theory. *Trans. Amer. Math. Soc.*, **82**, 323–339.

[14] Wakaki, H., Fujikoshi, Y. and Ulyanov, V. (2002). Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large, TR No. 02-09, *Statistical Research Group*, *Hiroshima University.*

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