Consistency properties of AIC, BIC, Cp and their modifications in the growth curve model under a large-(q, n) framework

Rie Enomoto, Tetsuro Sakurai and Yasunori Fujikoshi

Abstract. This paper is concerned with consistency properties of some criteria for selecting row vectors of a $k \times p$ design matrix within individuals in the growth curve model, based on a sample of size n. Recently Enomoto, Sakurai and Fujikoshi (2013) showed that AIC and its modification have a consistency property for selecting hierarchical models of the row vectors under a condition on the order of the noncentrality matrix, assuming a large-(q, n) asymptotic framework such that $q/n \to d \in [0, 1)$. We extend the result to a family of log-likelihood-based information criteria including AIC and BIC, and Cp. Further, their consistency properties are also obtained under a new condition on the order of the noncentrality matrix. Our results are checked numerically by conducting a Mote Carlo simulation.

AMS 2010 *Mathematics Subject Classification.* 62H12, 62H30.

Key words and phrases. AIC, BIC, Cp, Consistency property, Growth curve model, Large-(q, n) asymptotic framework, Simulation study.

§1. Introduction

The growth curve model introduced by Potthoff and Roy (1964) is written as

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

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

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_q} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_p \\ \vdots & \vdots & \vdots & \vdots \\ t_1^{k-1} & t_2^{k-1} & \cdots & t_p^{k-1} \end{pmatrix}. \tag{1.2}$$

It is important to decide the degree in a polynomial growth curve model. In general, we consider the problem of selecting the row vectors of $\mathbf{X}$. Suppose that j denotes a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements, and $\mathbf{X}_j$ denote the $k_j \times p$ matrix consisting of the rows of $\mathbf{X}$ indexed by the elements of j. Note that $\mathbf{X}_\omega = \mathbf{X}$ and $k_\omega = k$. We will let k_A denote the number of elements of a set A. We then consider the following candidate model M_j with k_j explanatory variables defined by

$$M_j;\ \mathbf{Y} = \mathbf{A}\mathbf{\Theta}_j\mathbf{X}_j + \boldsymbol{\mathcal{E}}, \tag{1.3}$$

where $\mathbf{\Theta}_j$ is a $q \times k_j$ matrix consisting of the columns of $\mathbf{\Theta}$ indexed by the elements of j, and $\boldsymbol{\mathcal{E}}$ has the same distribution as in (1.1). Here we note that the design matrix $\mathbf{A}$ may be also an observation matrix of several explanatory variables. For such an application, see Satoh and Yanagihara (2010). Let $\hat{\mathbf{\Theta}}_j$ and $\hat{\mathbf{\Sigma}}_j$ be the MLE's of $\mathbf{\Theta}_j$ and $\mathbf{\Sigma}$ under M_j, which are given by

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

where $\mathbf{S} = (n-q)^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{A})\mathbf{Y}$, and $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$.

There are several criteria for selecting a "best" model from a family of models M_j. The AIC and the BIC in our problem are given by

$$\mathrm{AIC} = n\log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + 2\left\{qk_j + \frac{1}{2}p(p+1)\right\}, \tag{1.4}$$

$$\mathrm{BIC} = n\log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + (\log n)\left\{qk_j + \frac{1}{2}p(p+1)\right\}. \tag{1.5}$$

Here, the last term $\{qk_j + p(p+1)/2\}$ is the number of independent parameters under M_j. A consistent AIC (CAIC) based on Bozdogan (1987) is given by

$$\mathrm{CAIC} = n\log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + (1 + \log n)\left\{qk_j + \frac{1}{2}p(p+1)\right\}. \tag{1.6}$$

We also consider the other modifications AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$ which are given in Section 2. Further, we consider Cp defined by

$$\mathrm{Cp} = n\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} + 2qk_j, \tag{1.7}$$

and its modification MCp, which is given in Section 2.

In this paper, we assume that the true model is included in the full model M_k. So, without loss of generality, we may assume that the minimum model including the true model is expressed as M_{j_0} for some j_0. Then, the true model is expressed as expressed as

$$M_0 : \mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{A}\boldsymbol{\Theta}_0\mathbf{X}_0, \boldsymbol{\Sigma}_0 \otimes \mathbf{I}_n), \tag{1.8}$$

where $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}_{j_0}$, $\mathbf{X}_0 = \mathbf{X}_{j_0}$, and $\boldsymbol{\Sigma}_0$ is a given positive definite matrix. We write $k_0 = k_{j_0}$. Let a set of candidate models denote by $\mathcal{F}$. The set of all candidate models involves $(2^k - 1)$ candidate models. A candidate model is called an overspecified model or an underspecified model if it includes or does not include the true model M_0. We denote a set of overspecified models and a set of underspecified model by $\mathcal{F}_+$ and $\mathcal{F}_-$, respectively.

In general, it can be seen that the criteria considered in this paper depend through p, n, k_0, k and the characteristic roots of

$$\boldsymbol{\Omega}_j = \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j, \tag{1.9}$$

which is called a noncentrality matrix, where $\boldsymbol{\Gamma}_j = (\mathbf{A}'\mathbf{A})^{1/2}\boldsymbol{\Theta}_0\mathbf{X}_0\boldsymbol{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)}$, $\mathbf{H}_1^{(j)} = (\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1/2})'(\mathbf{X}_j\boldsymbol{\Sigma}_0^{-1}\mathbf{X}_j')^{-1/2}$; $p \times k_j$ and $\left(\mathbf{H}_1^{(j)}, \mathbf{H}_2^{(j)}\right)$ is an orthogonal matrix.

It is known that AIC and Cp have not a consistency, but BIC and CAIC have a consistency property, under a large-sample framework

$$p,\ q \text{ and } k \text{ are fixed},\ n \to \infty, \tag{1.10}$$

and $\boldsymbol{\Omega}_j = \mathrm{O}(n)$. However, it is recently noted that AIC and Cp have a consistency property in a high-dimensional framework. Such results can be found in multivariate regression model, see, Fujikoshi, Sakurai and Yanagihara (2014), Yanagihara, Wakaki and Fujikoshi (2014). Further, Enomoto, Sakurai and Fujikoshi (2013) have noted that AIC and its modification $\mathrm{MAIC_H}$ in our problem have a consistency property for selecting hierarchical models of the row vectors of $\mathbf{X}$ under a large-(q, n) framework such that

$$p \text{ and } k \text{ are fixed},\ q \to \infty,\ n \to \infty,\ q/n \to d \in [0, 1), \tag{1.11}$$

and $\boldsymbol{\Omega}_j = \mathrm{O}(n)$. In this paper we extend such properties to various criteria including AIC, AICc, BIC, CAIC, $\mathrm{MAIC_L}$, $\mathrm{MAIC_H}$, Cp and MCp under $\boldsymbol{\Omega}_j = \mathrm{O}(nq)$ as well as $\boldsymbol{\Omega}_j = \mathrm{O}(n)$. When $\boldsymbol{\Omega}_j = \mathrm{O}(nq)$, it is noted that these criteria have a consistency property, though some condition on the value of d is imposed for AIC. When $\boldsymbol{\Omega}_j = \mathrm{O}(n)$, it is shown that BIC and CAIC have no consistency property, but the other criteria have a consistency property under some additional conditions. More precisely, we note that the probability of selecting the true model by BIC or CAIC tends to zero. Our results are also examined through a simulation experiment.

The present paper is organized as follows. In Section 2, we summarize modifications of AIC and Cp. Consistency properties of a log-likelihood-based information criterion are given in Section 3. In Section 4 we give consistency properties of Cp and MCp. Numerical experiments are given in Section 5. In Section 6, we summarize our conclusions. The proofs of our results are given in Appendix.

§2. Modifications of AIC and Cp

In this section we summarize modifications of AIC and Cp, and review their bias properties as estimators of the risks. As is well known, the AIC was proposed as an approximately unbiased estimator of the risk defined by the expected $-2\times$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 $-2\times$log-predictive likelihood under M_j is defined by

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

where $\hat{\mathbf{\Sigma}}_j$ and $\hat{\mathbf{\Theta}}_j$ are the maximum likelihood estimators of $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ 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

$$R_A = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_F}\left\{-2\log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j)\right\} + b_A, \tag{2.2}$$

where

$$b_A = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_F}\left\{-2\log f(\mathbf{Y}_F; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j) + 2\log f(\mathbf{Y}; \hat{\mathbf{\Theta}}_j, \hat{\mathbf{\Sigma}}_j)\right\}. \tag{2.3}$$

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

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

and considering an asymptotic approximation of b_A. A bias-corrected AIC is defined by

$$\mathrm{AICc} = n\log|\hat{\mathbf{\Sigma}}_j| + np(\log 2\pi + 1) + b_{A1}, \tag{2.4}$$

where

$$\begin{aligned} b_{A1} =& -np + \frac{n^2(p-k_j)}{n-p+k_j-1} \\ &+ \frac{n(n+q)(n-q-1)k_j}{(n-q-p-1)(n-q-p+k_j-1)}. \end{aligned} \tag{2.5}$$

Note that AICc is an exact unbiased estimator of R_A when M_j is an overspecified model, i.e.

$$\mathrm{E}(\mathrm{AICc}) = R_A, \quad j \in \mathcal{F}_+.$$

The term b_{A1} can be expressed as

$$\begin{aligned} b_{A1} =& 2\left\{qk_j + \frac{1}{2}p(p+1)\right\} + \frac{(p-k_j)(p-k_j+1)^2}{n-p+k_j-1} \\ &+ \frac{k_j(2p+q-k_j+1)(2q+p+1)}{n-q-p-1} \\ &+ \frac{(n+q)k_j(q+p-k_j+1)(p-k_j)}{(n-q-p-1)(n-q-p+k_j-1)}. \end{aligned} \tag{2.6}$$

Therefore, we can easily see that under a large-sample framework

$$\mathrm{AICc} = \mathrm{AIC} + \mathrm{O}(n^{-1}).$$

It is important that a modification has a small bias under underspecified models as well as overspecified models. Let $b_A = b_{A1} + b_{A2}$. It is known (Enomoto, Sakurai and Fujikoshi (2013)) that

$$b_{A2} = -\frac{n(p-k_j)(p-k_j+1)}{n-p+k_j-1} + 2(p-k_j+1)\xi_1 - \xi_2 + \mathrm{O}_g(n^{-1}),$$

where $\mathrm{O}_g(n^i)$ denotes the term of i-th order with respect to n under (1.11),

$$\xi_1 = \mathrm{tr}\left(\mathbf{I}_{p-k_j} + \frac{1}{n}\boldsymbol{\Omega}_j\right)^{-1}, \quad \xi_2 = \xi_1^2 + \mathrm{tr}\left(\mathbf{I}_{p-k_j} + \frac{1}{n}\boldsymbol{\Omega}_j\right)^{-2}. \tag{2.7}$$

A modification under a large-sample framework (1.10) is given by

$$\mathrm{MAIC_L} = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + b_{AL}, \tag{2.8}$$

where

$$b_{AL} = b_{A1} + \tilde{b}_{A2}, \quad \tilde{b}_{A2} = (p-k_j+1)\{2\tilde{\xi}_1 - (p-k_j)\} - \tilde{\xi}_2, \tag{2.9}$$

and

$$\tilde{\xi}_1 = \frac{n}{n-q}\left\{\mathrm{tr}(n\boldsymbol{\Sigma}_j)^{-1}(n-q)\mathbf{S} - k_j\right\},$$
$$\tilde{\xi}_2 = (\tilde{\xi}_1)^2 + \left(\frac{n}{n-q}\right)^2\left[\mathrm{tr}\{(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S}\}^2 - k_j\right].$$

Then, it is known (Satoh, Kobayashi and Fujikoshi (1997)) that under a large-sample framework (1.10)

$$\mathrm{E}\,(b_{AL}) = \begin{cases} b_A + \mathrm{O}(n^{-2}), & j \in \mathcal{F}_+, \\ b_A + \mathrm{O}(n^{-1}), & j \in \mathcal{F}_-. \end{cases}$$

The other modification based on a large-(n,q) framework (1.11) is given by

$$\mathrm{MAIC_H} = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + b_{AH}, \tag{2.10}$$

where

$$b_{AH} = b_{A1} + \hat{b}_{A2}, \quad \hat{b}_{A2} = (p-k_j+1)\{2\hat{\xi}_1 - (p-k_j)\} - \hat{\xi}_2, \tag{2.11}$$

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

$$f = \frac{3(n-q)(p-k_j+1)(n-2p+2k_j-2)}{n(n-p+k_j-1)} \times \left\{\frac{2(n-q+2)(p-k_j+2)}{n+2} + \frac{(n-q-1)(p-k_j-1)}{n-1}\right\}^{-1}.$$

Then, it is known (Enomoto, Sakurai and Fujikoshi (2013)) that under a large-(n,q) framework (1.11)

$$\mathrm{E}\left(\hat{b}_A\right) = \begin{cases} b_A, & j \in \mathcal{F}_+, \\ b_A + \mathrm{O}_g(n^{-1}), & j \in \mathcal{F}_-. \end{cases}$$

The Cp in regression model was proposed by Mallows (1973) for the univariate case. Sparks, Coutsourides and Troskie (1983) extended Mallows' approach to the multivariate case. Fujikoshi and Satoh (1997) gave a more general approach to Cp in the multivariate case. The criterion in the growth curve model may be essentially considered as an approximately unbiased estimator of the risk of M_j defined by

$$R_{\mathrm{C}} = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_{\mathrm{F}}}\left\{\mathrm{tr}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j)'(\mathbf{Y}_{\mathrm{F}} - \hat{\mathbf{Y}}_j)\right\}, \tag{2.12}$$

where $\hat{\mathbf{Y}}_j$ is a predictor of $\mathbf{Y}$ under M_j given by $\hat{\mathbf{Y}}_j = \mathbf{X}_j\hat{\mathbf{\Theta}}_j = \mathbf{P}_j\mathbf{Y}$, and $\mathbf{Y}_F$ is the same random matrix as in (2.1). The risk is expressed as

$$R_{\mathrm{C}} = \mathrm{E}^*_{\mathbf{Y}}\left\{(n-k_j)\mathrm{tr}\hat{\mathbf{\Sigma}}_\omega^{-1}\hat{\mathbf{\Sigma}}_j\right\} + b_{\mathrm{C}}, \tag{2.13}$$

where

$$b_{\mathrm{C}} = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Y}_{\mathrm{F}}}\left\{\mathrm{tr}\mathbf{\Sigma}_0^{-1}(\mathbf{Y}_{\mathrm{F}}-\hat{\mathbf{Y}}_j)'(\mathbf{Y}_{\mathrm{F}}-\hat{\mathbf{Y}}_j) - (n-k_j)\mathrm{tr}\hat{\mathbf{\Sigma}}_\omega^{-1}\hat{\mathbf{\Sigma}}_j\right\}. \tag{2.14}$$

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

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

and by evaluating the bias term b_{C}. Satoh, Kobayashi and Fujikoshi (1997) proposed the following Cp and its modification MCp:

$$\mathrm{Cp} = n\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} + 2qk_j \tag{2.15}$$

$$\begin{aligned}\mathrm{MCp} &= n\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} + q(p+k_j) \\ &\quad - \frac{q(p-k_j)(n-q-k_j)}{n-q-p+k_j-1} + \left(\frac{2k_j-p-1}{n-q-p+k_j-1}\right) \\ &\quad \times \left\{\frac{n(n-q-p+k_j-1)}{n-q}\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} - (n-p+k_j-1)p + qk_j\right\}.\end{aligned} \tag{2.16}$$

The MCp satisfies

$$\mathrm{E(MCp)} = R_{\mathrm{C}}.$$

Further we can write MCp as

$$\begin{aligned}\mathrm{MCp} &= \left\{1 + \frac{2k_j-p+1}{n-q}\right\}n\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} + 2qk_j + p(p-2k_j+1) \\ &= \mathrm{Cp} + (2k_j-p+1)\frac{n}{n-q}\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} + p(p-2k_j+1).\end{aligned} \tag{2.17}$$

§3. Consistency of a log-likelihood-based information criterion

We treat AIC and its modifications as a unified criterion

$$\mathrm{IC}_j = n\log\det(\hat{\mathbf{\Sigma}}_j) + np(\log 2\pi + 1) + m_j, \tag{3.1}$$

which is called a log-likelihood-based information criterion, where m_j is a positive constant expressing a penalty for the complexity of the model (1.3).

A specific criterion is given by specifying the individual penalty term m_j. It contains AIC, BIC, CAIC, AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$ as a special case, as follows.

$$m_j = \begin{cases} 2\{qk_j + p(p+1)/2\} & \text{(AIC)} \\ \{qk_j + p(p+1)/2\}\log n & \text{(BIC)} \\ \{qk_j + p(p+1)/2\}(1+\log n) & \text{(CAIC)} \\ b_{A1,j} & \text{(AICc)} \\ b_{AL,j} & (\mathrm{MAIC_L}) \\ b_{AH,j} & (\mathrm{MAIC_H}) \end{cases}. \tag{3.2}$$

Here the quantities $b_{A1,j}$, $b_{AL,j}$ and $b_{AH,j}$ are the same ones as in (2.6), (2.9) and (2.11), respectively.

In this section we show that the asymptotic probability of selecting the true model by AIC and its modifications goes to 1 when the number q and the sample size n are approaching to ∞ as in (1.11), under some additional assumptions. We denote the AIC for M_j by AIC_j. The best model chosen by minimizing the AIC is written as

$$\hat{j}_{\mathrm{AIC}} = \arg\min_{j\in\mathcal{F}} \mathrm{AIC}_j.$$

Similar notations are used for the other criteria. The consistency property of IC is examined by using a key result (see, e.g., Fujikoshi, Enomoto and Sakurai (2013))

$$\frac{|(n-q)\mathbf{S}|}{|n\hat{\mathbf{\Sigma}}_j|} = \frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)} + \mathbf{B}_{(j)}|}, \tag{3.3}$$

where $\mathbf{W}_{(j)}$ are independently distributed as a Wishart distribution $\mathrm{W}_{p-k_j}(n-q, \mathbf{I}_{p-k_j})$ and a noncentral distribution $\mathrm{W}_{p-k_j}(q, \mathbf{I}_{p-k_j}; \mathbf{\Omega}_j)$, respectively. The matrix $\mathbf{\Omega}_j$ is defined by (1.9).

Our main assumptions are summarized as follows:

A1 (The true model M_0): $j_0 \in \mathcal{F}$.

A2 (The asymptotic framework): $q \to \infty,\ n \to \infty,\ q/n \to d \in [0, 1)$.

A3 (The order assumption (i) of $\mathbf{\Omega}_j$): For $j \in \mathcal{F}_-$,

$$\mathbf{\Omega}_j = n\mathbf{\Delta}_j = \mathrm{O}_g(n) \text{ and } \lim_{q/n\to d} \mathbf{\Delta}_j = \mathbf{\Delta}_j^*.$$

A4 (The order assumption (ii) of $\mathbf{\Omega}_j$): For $j \in \mathcal{F}_-$,

$$\mathbf{\Omega}_j = nq\mathbf{\Xi}_j = \mathrm{O}_g(nq) \text{ and } \lim_{q/n\to d} \mathbf{\Xi}_j = \mathbf{\Xi}_j^*.$$

Our consistency properties of a log-likelihood-based information criterion are given in two theorems, depending on the assumptions A3 and A4 on the order of the noncentrality matrix $\mathbf{\Omega}_j$ as follows.

Theorem 3.1. *Suppose that the assumptions* A1, A2 *and* A3 *are satisfied.*

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

A5*: For any* $j \in \mathcal{F}_-$,

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

Then, the model selection criterion AIC *is consistent,* i.e., *the asymptotic probability of selecting the true model* j_0 *by the* AIC *tends to* 1, *which may be stated as*

$$\lim_{q/n\to d} P(\hat{j}_{\rm AIC}=j_0)=1.$$

(2) *Suppose that*

A6*: For any* $j \in \mathcal{F}_-$,

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

Then, the model selection criteria AICc, $\mathrm{MAIC_L}$ *and* $\mathrm{MAIC_H}$ *are consistent.*

(3) *The model selection criteria* BIC *and* CAIC *are not consistent. More precisely, the probability of selecting the true model by* BIC *or* CAIC *tends to zero.*

Theorem 3.1 is an extension of Enomoto, Sakurai and Fujikoshi (2013) which proves consistency of AIC and $\mathrm{MAIC_H}$ in the case of selection of hierarichical models on the row vectors of $\mathbf{X}$.

Theorem 3.2. *Suppose that the assumptions* A1, A2 *and* A4 *are satisfied.*

(1) *If* $d \in [0, d_{\rm a})$, *then, the model selection criterion* AIC *is consistent. Here* d_a *is given Theorem* 3.1.

(2) *Suppose that for any* $j \in \mathcal{F}_-$, $|\boldsymbol{\Xi}_j| > 0$. *Then, the model selection criteria* AICc, BIC, CAIC, $\mathrm{MAIC_L}$ *and* $\mathrm{MAIC_H}$ *are consistent.*

§4. Consistency of Cp and MCp

In this section we give consistency properties of Cp and MCp. The derivation is done in a way similar to one for a log-likelihood-based information criterion, with the help of

$$\begin{aligned}\frac{n}{n-q}\mathrm{tr}\hat{\boldsymbol{\Sigma}}_j\mathbf{S}^{-1} &= \mathrm{tr}(n\hat{\boldsymbol{\Sigma}}_j)\{(n-q)\mathbf{S}\}^{-1}\\ &= p+\mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1},\end{aligned} \tag{4.1}$$

where $\mathbf{W}_{(j)}$ and $\mathbf{B}_{(j)}$ are the same random matrices as in (3.3).

Theorem 4.3. *Suppose that the assumptions* A1, A2 *and* A3 *are satisfied. Further, assume that*

A7*: For any* $j \in \mathcal{F}_-$,

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

Then, the model selection criteria Cp *and* MCp *are consistent.*

Theorem 4.4. *Suppose that the assumptions* A1, A2 *and* A4 *are satisfied. Further, suppose that for any* $j \in \mathcal{F}_-$, $\mathrm{tr}\boldsymbol{\Xi}_j^* > 0$. *Then, the model selection criteria* Cp *and* MCp *are consistent.*

These results will be worthy of note, since Cp and MCp are known to be inconsistent under a large-sample framework.

§5. Simulation study

In this section, we numerically examine the validity of our claims and the speed of the convergences of the criteria. Monte Carlo simulations were considered for several different values of n and $q = dn$, where $p = 5$, $n = 50, 100, 200$, $n_1 = \cdots = n_q = n/q$ and $d = 0.1, 0.2$. We constructed a 5×5 matrix $\mathbf{X}$ as in (1.2) of explanatory variables with $t_i = 1 + (i-1)(p-1)^{-1}$. The true covariance matrix $\boldsymbol{\Sigma}_0$ was determined such that its (i, j)th element is $\rho^{|i-j|}$, where $\rho = 0.2, 0.8$. We consider the five candidate models $M_1, \ldots, M_5$, where M_j denotes the model with the first j rows of $\mathbf{X}$. So, in this section a subset j means $j = \{1, \ldots, j\}$. We assume that M_2 is the minimum model including the true model. The true model are included in M_2, M_3, M_4, M_5, but it is not included in M_1. Therefore, $\boldsymbol{\Omega}_j = \mathbf{O}$ when M_2, M_3, M_4, M_5, and $\boldsymbol{\Omega}_j \neq \mathbf{O}$ when M_1.

5.1. The case of order assumption (i)

As a realization of $\mathbf{\Omega}_j = \mathrm{O}_g(n)$ we assume that $\mathbf{\Theta}_0 = \mathbf{1}_q\mathbf{1}_2'$. Then, the noncentrality matrix $\mathbf{\Omega}_j$ is expressed as

$$\begin{aligned}
\mathbf{\Omega}_j &= \mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0'\mathbf{\Theta}_0'\mathbf{A}'\mathbf{A}\mathbf{\Theta}_0\mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)} \\
&= \mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0'\mathbf{1}_2\mathbf{1}_q' \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_q \end{pmatrix} \mathbf{1}_q\mathbf{1}_2'\mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)} \\
&= \mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0' \begin{pmatrix} n & n \\ n & n \end{pmatrix} \mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)}.
\end{aligned}$$

Further, $\mathbf{X}_0$, $\mathbf{\Sigma}_0^{-1/2}$ and $\mathbf{H}_2^{(j)}$ do not depend on n and q. Therefore, $\mathbf{\Omega}_j = \mathrm{O}_g(n)$. Moreover, the convergent values in A5, A6 and A7 for consistency are calculated as follows:

ρ	d	$\log\lvert\mathbf{I}_{p-k_j} + \mathbf{\Delta}_j^*\rvert$	$2d + \log(1-d)$	$2d/(1-d) + \log(1-d)$	$\mathrm{tr}\mathbf{\Delta}_j^*$
0.2	0.1	0.440	0.095	0.117	0.552
	0.2	0.440	0.177	0.277	0.552
0.8	0.1	0.614	0.095	0.117	0.847
	0.2	0.614	0.177	0.277	0.847

Here, $\mathcal{F}_-$ contains a subset $\{1\}$ only. Therefore, we can see that in our setting the assumptions A5, A6 and A7 are satisfied.

The selection probabilities (%) based on Monte Carlo simulations with 10^4 iterations are summarized in Tables 1 $\sim$ 4. From these tables we can point the following tendencies.

- We can see that CAIC and BIC have no consistency property. In general, they chose M_1 with high probabilities, though they have a tendency of choosing M_2 for $d = 0.1$ and $\rho = 0.2$.
- $\mathrm{MAIC_H}$ chooses M_2 more frequently than $\mathrm{MAIC_L}$. Similarly, MCp chooses M_2 more frequently than Cp.
- As q increases under n being fixed, AIC, Cp and MCp choose M_2 more frequently, but the other criteria choose M_2 more fewer.
- For the speed of convergences to 1, the case $\rho = 0.8$ is faster than the case $\rho = 0.8$.
- AIC, Cp and MCp have a tendency of choosing overspecified models than AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$.

- AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$ have a tendency of choosing $underspecified modes$ than AIC, Cp and MCp.
- When $d = 0.2$ and n and q are small, AIC chooses the true model more frequently than AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$.
- MCp chooses the true model more frequently than Cp in all the cases except the case; $d = 0.2$, $n = 50$ and $q = 10$.

Table 1. Selection probabilities (%) for $d = 0.1$ and $\rho = 0.2$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 5)	M_1	0.8	4.8	7.1	5.2	17.2	37.9	0.6	1.2
	M_2	85.4	91.4	90.2	91.4	82.5	62.1	83.5	86.5
	M_3	9.9	3.1	2.3	2.9	0.3	0.0	10.8	8.4
	M_4	2.8	0.5	0.3	0.5	0.0	0.0	3.6	2.6
	M_5	1.1	0.1	0.1	0.1	0.0	0.0	1.5	1.3
(100, 10)	M_1	0.0	0.3	0.3	0.3	24.1	56.6	0.0	0.1
	M_2	94.7	98.5	98.5	98.5	75.9	43.5	94.0	95.3
	M_3	4.6	1.2	1.1	1.2	0.0	0.0	5.0	4.0
	M_4	0.6	0.0	0.0	0.0	0.0	0.0	0.8	0.6
	M_5	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1
(200, 20)	M_1	0.0	0.0	0.0	0.0	43.1	83.7	0.0	0.0
	M_2	98.8	99.9	99.9	99.9	56.9	16.3	98.8	99.0
	M_3	1.2	0.1	0.1	0.1	0.0	0.0	1.2	1.0
	M_4	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2. Selection probabilities (%) for $d = 0.2$ and $\rho = 0.2$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 10)	M_1	5.5	49.4	55.5	52.0	70.1	92.0	4.2	9.6
	M_2	85.6	50.4	44.4	47.9	29.9	8.0	85.0	83.4
	M_3	7.2	0.2	0.1	0.2	0.0	0.0	8.4	5.5
	M_4	1.4	0.0	0.0	0.0	0.0	0.0	1.9	1.2
	M_5	0.4	0.0	0.0	0.0	0.0	0.0	0.6	0.4
(100, 20)	M_1	0.9	22.9	24.9	23.8	95.9	99.9	0.8	1.7
	M_2	96.6	77.1	75.1	76.2	4.1	0.1	96.6	96.5
	M_3	2.4	0.0	0.0	0.0	0.0	0.0	2.5	1.7
	M_4	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(200, 40)	M_1	0.0	7.0	7.4	7.1	100.0	100.0	0.0	0.1
	M_2	99.7	93.0	92.7	92.9	0.0	0.0	99.7	99.7
	M_3	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.2
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3. Selection probabilities (%) for $d = 0.1$ and $\rho = 0.8$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 5)	M_1	0.0	0.3	0.5	0.3	2.2	9.1	0.0	0.0
	M_2	85.9	96.2	96.8	96.4	97.4	90.9	83.8	87.6
	M_3	10.1	3.0	2.3	2.8	0.3	0.0	11.1	8.5
	M_4	3.0	0.5	0.3	0.5	0.0	0.0	3.7	2.7
	M_5	1.0	0.1	0.1	0.1	0.0	0.0	1.4	1.2
(100, 10)	M_1	0.0	0.0	0.0	0.0	1.5	10.3	0.0	0.0
	M_2	94.9	98.6	98.7	98.7	98.5	89.7	94.4	95.5
	M_3	4.4	1.3	1.2	1.3	0.0	0.0	4.8	3.8
	M_4	0.6	0.1	0.0	0.1	0.0	0.0	0.7	0.5
	M_5	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1
(200, 20)	M_1	0.0	0.0	0.0	0.0	1.5	15.7	0.0	0.0
	M_2	98.9	99.9	99.9	99.9	98.6	84.3	98.7	99.0
	M_3	1.1	0.1	0.1	0.1	0.0	0.0	1.2	1.0
	M_4	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4. Selection probabilities (%) for $d = 0.2$ and $\rho = 0.8$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 10)	M_1	0.5	18.3	23.0	20.4	36.7	73.7	0.3	1.1
	M_2	89.9	81.4	76.7	79.3	63.2	26.3	88.4	91.0
	M_3	8.0	0.4	0.3	0.3	0.1	0.0	9.0	6.3
	M_4	1.3	0.0	0.0	0.0	0.0	0.0	1.9	1.2
	M_5	0.3	0.0	0.0	0.0	0.0	0.0	0.5	0.4
(100, 20)	M_1	0.0	2.0	2.4	2.1	70.6	97.0	0.0	0.0
	M_2	97.7	98.0	97.6	97.9	29.4	3.0	97.5	98.2
	M_3	2.2	0.0	0.0	0.0	0.0	0.0	2.3	1.7
	M_4	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(200, 40)	M_1	0.0	0.1	0.1	0.1	98.3	100.0	0.0	0.0
	M_2	99.8	99.9	99.9	99.9	1.7	0.0	99.8	99.9
	M_3	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.1
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5.2. The case of order assumption (ii)

As a realization of $\mathbf{\Omega}_j = \mathrm{O}_g(nq)$ we assume that $\mathbf{\Theta}_0 = \sqrt{q}\mathbf{1}_q\mathbf{1}_2'$. Then, the noncentrality matrix $\mathbf{\Omega}_j$ is expressed as

$$\begin{aligned}
\mathbf{\Omega}_j &= \mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0'\mathbf{\Theta}_0'\mathbf{A}'\mathbf{A}\mathbf{\Theta}_0\mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)} \\
&= q\mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0'\mathbf{1}_2\mathbf{1}_q' \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_q \end{pmatrix} \mathbf{1}_q\mathbf{1}_2'\mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)} \\
&= \mathbf{H}_2^{(j)\prime}\mathbf{\Sigma}_0^{-1/2\prime}\mathbf{X}_0' \begin{pmatrix} nq & nq \\ nq & nq \end{pmatrix} \mathbf{X}_0\mathbf{\Sigma}_0^{-1/2}\mathbf{H}_2^{(j)}.
\end{aligned}$$

Therefore, $\mathbf{\Omega}_j = \mathrm{O}_g(nq)$. The selection probabilities (%) are summarized in Tables 5 $\sim$ 8. From these tables we can point the following tendencies.

- We can see that all the eight criteria have consistency property.
- $\mathrm{MAIC_H}$ and MCp choose the true model more frequently than $\mathrm{MAIC_L}$ and Cp, respectively.
- As q increases under n being fixed, all the criteria choose the true model more frequently.
- In general, AIC, Cp and MCp have a tendency of choosing larger models when n and q are small.
- It seems that all the criteria doe not choose underspecified models.
- For the speed of convergence to the true model, BIC and CAIC are more faster than the other criteria.

Table 5. Selection probabilities (%) for $d = 0.1$ and $\rho = 0.2$.

(n, q)		AIC	AICc	MAIC_L	MAIC_H	CAIC	BIC	Cp	MCp
(50, 5)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	86.1	96.0	96.9	96.3	99.6	99.9	84.2	87.5
	M_3	9.9	3.3	2.6	3.1	0.4	0.1	10.8	8.6
	M_4	3.0	0.5	0.4	0.5	0.0	0.0	3.6	2.6
	M_5	1.0	0.1	0.1	0.1	0.0	0.0	1.5	1.2
(100, 10)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	95.2	98.7	98.8	98.8	100.0	100.0	94.6	95.7
	M_3	4.2	1.2	1.1	1.2	0.0	0.0	4.6	3.6
	M_4	0.6	0.1	0.1	0.1	0.0	0.0	0.8	0.6
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
(200, 20)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	98.8	99.8	99.8	99.8	100.0	100.0	98.8	99.0
	M_3	1.1	0.2	0.2	0.2	0.0	0.0	1.2	1.0
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6. Selection probabilities (%) for $d = 0.2$ and $\rho = 0.2$.

(n, q)		AIC	AICc	MAIC_L	MAIC_H	CAIC	BIC	Cp	MCp
(50, 10)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	90.4	99.5	99.6	99.6	99.9	100.0	88.8	92.2
	M_3	7.7	0.5	0.4	0.4	0.1	0.0	8.7	6.0
	M_4	1.5	0.0	0.0	0.0	0.0	0.0	1.9	1.3
	M_5	0.4	0.0	0.0	0.0	0.0	0.0	0.6	0.5
(100, 20)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	97.4	100.0	100.0	100.0	100.0	100.0	97.3	98.1
	M_3	2.4	0.0	0.0	0.0	0.0	0.0	2.5	1.7
	M_4	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.2
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(200, 40)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	99.9	100.0	100.0	100.0	100.0	100.0	99.9	99.9
	M_3	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 7. Selection probabilities (%) for $d = 0.1$ and $\rho = 0.8$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 5)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	86.9	96.5	97.3	96.7	99.7	100.0	84.6	88.3
	M_3	9.3	2.9	2.3	2.8	0.3	0.0	10.6	8.1
	M_4	2.8	0.4	0.3	0.4	0.0	0.0	3.5	2.6
	M_5	1.0	0.1	0.1	0.1	0.0	0.0	1.3	1.1
(100, 10)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	94.3	98.6	98.9	98.7	100.0	100.0	93.4	94.8
	M_3	4.9	1.3	1.1	1.3	0.0	0.0	5.5	4.4
	M_4	0.8	0.1	0.1	0.1	0.0	0.0	1.0	0.7
	M_5	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1
(200, 20)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	98.9	99.8	99.8	99.8	100.0	100.0	98.9	99.1
	M_3	1.0	0.2	0.2	0.2	0.0	0.0	1.1	0.8
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 8. Selection probabilities (%) for $d = 0.2$ and $\rho = 0.8$.

(n, q)		AIC	AICc	$\mathrm{MAIC_L}$	$\mathrm{MAIC_H}$	CAIC	BIC	Cp	MCp
(50, 10)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	90.6	99.6	99.7	99.6	100.0	100.0	88.9	92.6
	M_3	7.5	0.4	0.3	0.4	0.0	0.0	8.5	5.7
	M_4	1.5	0.0	0.0	0.0	0.0	0.0	2.0	1.3
	M_5	0.4	0.0	0.0	0.0	0.0	0.0	0.6	0.4
(100, 20)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	97.5	99.9	99.9	99.9	100.0	100.0	97.3	98.1
	M_3	2.3	0.1	0.1	0.1	0.0	0.0	2.5	1.8
	M_4	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.1
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
(200, 40)	M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_2	99.9	100.0	100.0	100.0	100.0	100.0	99.9	99.9
	M_3	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1
	M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

§6. Concluding remarks

This paper discusses with consistency properties of a log-likelihood criterion including AIC and its modifications, Cp and its modification MCp for selecting the row vectors of a design matrix $\mathbf{X}$ within individuals in the growth curve model (1.1) under a large-(q, n) framework (1.11). The log-likelihood criterion includes AIC, AICc, BIC, CAIC, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$ as a special case. The consistency properties depend on the order of the noncentrality matrix $\mathbf{\Omega}_j$ of model M_j. When $\mathbf{\Omega}_j = \mathrm{O}(n)$, it is noted that AIC, AICc, $\mathrm{MAIC_L}$ and $\mathrm{MAIC_H}$, Cp and MCp are consistent under some additional assumptions on $\mathbf{\Omega}$. However, BIC and CAIC are not consistent, and more precisely, the probability of selecting the true model by BIC or CAIC tends to zero. When $\mathbf{\Omega}_j = \mathrm{O}(nq)$, it is noted that these criteria have a consistency property, though some condition on the value of d is imposed for AIC.

In a traditional growth curve model it is assumed that the dimension p is not large or moderate. However, it is also important to analysis the data such that p is large. Further, the number k of explanatory variables within individuals will be large. This suggests to study asymptotic properties of these model selection criteria under a high-dimensional framework such that

$$\begin{aligned} &k \to \infty,\ p \to \infty,\ q \to \infty,\ n \to \infty, \\ &k/n \to b \in [0, 1),\ p/n \to c \in [0, 1),\ q/n \to d \in [0, 1), \end{aligned} \tag{6.1}$$

where $1 > c \geq d \geq 0$. Modifications of AIC and Cp and their consistency properties under (6.1) should be also studied. These works are left as a future subject.

§7. Appendix: Proofs of Theorems

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_0 is the minimum model including the true model and $j_0 \in \mathcal{F}$. Let $\mathrm{T}_j(n)$ be a general criterion for model j, which depends on a parameter n. The best model chosen by minimizing $\mathrm{T}_j(n)$ is written as $\hat{j}_{\mathrm{T}}(n) = \arg\min_{j\in\mathcal{F}} \mathrm{T}_j(n)$. Suppose that we are interested in asymptotic behavior of $\hat{j}_{\mathrm{T}}(n)$ when n tends to ∞. In order to show a consistency of $\mathrm{T}_j(n)$, we may check a sufficient condition such that for any $j \neq j_0 \in \mathcal{F}$, there exists a sequence $\{a_n\}$ with $a_n > 0$,

$$a_n \{T_j(n) - T_{j_0}(n)\} \xrightarrow{p} b_j > 0.$$

In fact, the condition implies that for any $j \neq j_0 \in \mathcal{F}$,

$$P(\hat{j}_{\mathrm{T}}(n) = j) \leq P(T_j(n) < T_{j_0}(n)) \to 0,$$

and

$$P(\hat{j}_{\mathrm{T}}(n) = j_0) = 1 - \sum_{j\neq j_0 \in\mathcal{F}} P(\hat{j}_{\mathrm{T}}(n) = j) \to 1.$$

On the other hand, relating to showing an inconsistency of $\hat{j}_{\rm T}(n)$, assume that for some $j \neq j_0 \in \mathcal{F}$ and for a sequence $\{a_n\}$ with $a_n > 0$,

$$a_n \{{\rm T}_j(n) - {\rm T}_{j_0}\} \xrightarrow{p} d_j < 0.$$

Then we have

$$P({\rm T}_j(n) < {\rm T}_{j_0}(n)) \to 1.$$

Further, we have

$$P(\hat{j}_{\rm T}(n) = j_0) \leq P(T_{j_0}(n)) < T_j(n)) = 1 - P({\rm T}_j(n) < {\rm T}_{j_0}(n)) \to 0.$$

This means that $\hat{j}_{\rm T}(n)$ is inconsistent, and further the probability of selecting the true model tends to zero.

Proof of Theorem 3.1

The consistency properties of AIC and $\mathrm{MAIC_H}$ have been essentially proved by Enomoto, Sakurai and Fujikoshi (2013) who proved for the case of selecting hierarchical models of the row vectors of $\mathbf{X}$. The following result was used there:

$$\begin{aligned}\log|\hat{\boldsymbol{\Sigma}}_j| - \log|\hat{\boldsymbol{\Sigma}}_{j_0}| &= \log|n\hat{\boldsymbol{\Sigma}}_j| - \log|n\hat{\boldsymbol{\Sigma}}_{j_0}| \\ &= -\log\frac{|(n-q)\mathbf{S}|}{|n\hat{\boldsymbol{\Sigma}}_j|} + \log\frac{|(n-q)\mathbf{S}|}{|n\hat{\boldsymbol{\Sigma}}_{j_0}|} \\ &= -\log\frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)} + \mathbf{B}_{(j)}|} + \log\frac{|\mathbf{W}_{(j_0)}|}{|\mathbf{W}_{(j_0)} + \mathbf{B}_{(j_0)}|}.\end{aligned}$$

Further, noting that

$$\frac{1}{n}\mathbf{W}_{(j)} \xrightarrow{p} (1-d)\mathbf{I}_{p-k_j}, \text{ and } \frac{1}{n}\mathbf{B}_{(j)} \xrightarrow{p} d\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*,$$

the following result was used:

$$-\log\frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)} + \mathbf{B}_{(j)}|} \xrightarrow{p} \log|\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*| - (p-k_j)\log(1-d).$$

These imply that

$$\begin{aligned}&\frac{1}{n}(\mathrm{IC}_j - \mathrm{IC}_{j_0}) - \frac{1}{n}(m_j - m_{j_0}) \\ &\qquad \xrightarrow{p} \log|\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*| + (k_j - k_0)\log(1-d),\end{aligned} \tag{7.1}$$

since $\boldsymbol{\Delta}_{j_0} = \mathbf{O}$. For the penalty terms, it is easily seen that

$$\frac{1}{n}(m_j - m_{j_0}) \xrightarrow{p} \begin{cases} 2d(k_j - k_0) & \text{(AIC)} \\ 2d(1-d)^{-1}(k_j - k_0) & \text{(AICc)} \end{cases}, \tag{7.2}$$

and

$$\frac{1}{n \log n}(m_j - m_{j_0}) \xrightarrow{p} (k_j - k_0)d, \quad (\mathrm{BIC}, \mathrm{CAIC}) \tag{7.3}$$

Now we shall prove the case of AICc. From (7.1) and (7.2) we have

$$\begin{aligned} \frac{1}{n}(\mathrm{AICc}_j - \mathrm{AICc}_{j_0}) \xrightarrow{p} & \log |\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*| \\ & + (k_j - k_0)\left\{ \frac{2d}{1-d} + \log(1-d) \right\}. \end{aligned} \tag{7.4}$$

Therefore, when $j \in \mathcal{F}_+$, $\log |\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*| = 0$, and hence

$$\mathrm{AICc}_j - \mathrm{AICc}_{j_0} > 0,$$

since $(2d)/(1-d) + \log(1-d)$ is always positive. When $j \in \mathcal{F}_-$, we have $\mathrm{AICc}_j - \mathrm{AICc}_{j_0} > 0$ from (7.4) and the assumption A5. These imply the consistency of AICc. For the case $d = 0$, we need to modify the above proof slightly. For example, we can prove by considering the limit of $(1/q)(\mathrm{AICc}_j - \mathrm{AICc}_{j_0})$ in stead of $(1/n)(\mathrm{AICc}_j - \mathrm{AICc}_{j_0})$. In the following, we give a proof for $0 < d < 1$, and the proof of $d = 0$ is omitted.

For the case of BIC, from (7.1) and (7.3) we have

$$\frac{1}{n \log n}(\mathrm{BIC}_j - \mathrm{BIC}_{j_0}) \xrightarrow{p} (k_j - k_0)d.$$

This implies that for some j such that $j \in \mathcal{F}_-$ and $k_j - k_0 < 0$, $\mathrm{BIC}_j < \mathrm{BIC}_{j_0}$ for large n. These show an inconsistency of BIC and

$$P(\hat{j}_{\mathrm{BIC}} = j_0) \to 0.$$

The case of CAIC is proved similarly as the one of BIC.

In order to prove the case of $\mathrm{MAIC_L}$, it needs to examine asymptotic behavior of $\tilde{b}_{A2}$ as in Enomoto, Sakurai and Fujikoshi (2013), which gives an asymptotic behavior of $\hat{b}_{A2}$. We can express

$$\mathrm{tr}(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S} = j + \mathrm{tr}\mathbf{Q}_j, \quad \mathrm{tr}\big\{(n\hat{\boldsymbol{\Sigma}}_j)^{-1}(n-q)\mathbf{S}\big\}^2 = j + \mathrm{tr}\mathbf{Q}_j^2,$$

where $\mathbf{Q}_j = \mathbf{W}_{(j)}(\mathbf{W}_{(j)} + \mathbf{B}_{(j)})^{-1}$. Using $\mathbf{Q}_j \xrightarrow{p} (1-d)\left(\mathbf{I}_{p-k_j} + \boldsymbol{\Delta}_j^*\right)^{-1}$, we have

$$\begin{aligned} \tilde{\xi}_1 &\xrightarrow{p} \xi_{10} = \mathrm{tr}(\mathbf{I} + \boldsymbol{\Delta}_j^*)^{-1}, \\ \tilde{\xi}_2 &\xrightarrow{p} \xi_{20} = \big\{\mathrm{tr}(\mathbf{I} + \boldsymbol{\Delta}_j^*)^{-1}\big\}^2 + \mathrm{tr}(\mathbf{I} + \boldsymbol{\Delta}_j^*)^{-2}. \end{aligned}$$

From these results we have $(1/n)\tilde{b}_{A2} \to 0$. Therefore, we get that the consistency of $\mathrm{MAIC_L}$ is the same as the one of AICc.

Proof of Theorem 3.2

First we note that when $j \in \mathcal{F}_+$, from (7.1) we have

$$\frac{1}{n}(\mathrm{IC}_j - \mathrm{IC}_{j_0}) - \frac{1}{n}(m_j - m_{j_0}) \xrightarrow{p} (k_j - k_0)\log(1-d), \quad j \in \mathcal{F}_+. \tag{7.5}$$

On the other hand, when $j \in \mathcal{F}_-$,

$$\frac{1}{n}\mathbf{W}_{(j)} \xrightarrow{p} (1-d)\mathbf{I}_{p-k_j}, \quad \frac{1}{nq}\mathbf{B}_{(j)} \xrightarrow{p} \mathbf{\Xi}_j,$$

and hence

$$\begin{aligned}\log|\hat{\mathbf{\Sigma}}_j| - \log|\hat{\mathbf{\Sigma}}_{j_0}| &= -\log\frac{|\mathbf{W}_{(j)}|}{|\mathbf{W}_{(j)} + \mathbf{B}_{(j)}|} + \log\frac{|\mathbf{W}_{(j_0)}|}{|\mathbf{W}_{(j_0)} + \mathbf{B}_{(j_0)}|} \\ &= -\log\frac{|\frac{1}{n}\mathbf{W}_{(j)}|}{|\frac{1}{nq}(\mathbf{W}_{(j)} + \mathbf{B}_{(j)})|q^{p-k_j}} + \log\frac{|\mathbf{W}_{(j_0)}|}{|\mathbf{W}_{(j_0)} + \mathbf{B}_{(j_0)}|}.\end{aligned}$$

These imply that

$$\frac{1}{n\log q}(\mathrm{IC}_j - \mathrm{IC}_{j_0}) - \frac{1}{n\log q}(m_j - m_{j_0}) \xrightarrow{p} p - k_j, \quad j \in \mathcal{F}_-. \tag{7.6}$$

Using (7.5) and (7.6) we can prove Theorem 3.2 by the same line as in Theorem 3.1. Its detail is omitted.

Proofs of Theorems 4.3 and 4.4

Noting that

$$\begin{aligned}\frac{n}{n-q}\mathrm{tr}\hat{\mathbf{\Sigma}}_j\mathbf{S}^{-1} &= \mathrm{tr}(n\hat{\mathbf{\Sigma}}_j)\{(n-q)\mathbf{S}\}^{-1} \\ &= p + \mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1},\end{aligned}$$

we have

$$\frac{1}{n-q}(\mathrm{Cp}_j - \mathrm{Cp}_{j_0}) = \mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1} - \mathrm{tr}\mathbf{B}_{(j_0)}\mathbf{W}_{(j_0)}^{-1} + \frac{2q}{n-q}(k_j - k_0).$$

First consider the case $\mathbf{\Omega}_j = \mathrm{O}_g(n) = n\mathbf{\Delta}_j$. In this case

$$\mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1} \xrightarrow{p} \frac{1}{1-d}\left(d\mathbf{I}_{p-k_j} + \mathbf{\Delta}_j^*\right).$$

When $j \in \mathcal{F}_+$, $\boldsymbol{\Delta}_j = \mathbf{O}$, and hence

$$\begin{aligned}\frac{1}{n-q}\left(\mathrm{Cp}_j - \mathrm{Cp}_{j_0}\right) &\xrightarrow{p} \frac{d}{1-d}(p-k_j) - \frac{d}{1-d}(p-k_0) + \frac{2d}{1-d}(k_j-k_0) \\ &= (k_j-k_0)\cdot\frac{d}{1-d}.\end{aligned}$$

When $j \in \mathcal{F}_-$,

$$\begin{aligned}\frac{1}{n-q}\left(\mathrm{Cp}_j - \mathrm{Cp}_{j_0}\right) &\xrightarrow{p} (k_j-k_0)\cdot\frac{d}{1-d} + \frac{1}{1-d}\mathrm{tr}\boldsymbol{\Delta}_j^* \\ &= \frac{1}{1-d}\{\mathrm{tr}\boldsymbol{\Delta}_j + d(k_j-k_0)\}.\end{aligned}$$

Therefore, if $\mathrm{tr}\boldsymbol{\Delta}_j^* > d(k_0-k_j),\ j \in \mathcal{F}_-$, then Cp is consistent.

Next we consider the case $\boldsymbol{\Omega}_j = \mathrm{O}_g(nq) = nq\boldsymbol{\Xi}_j$. When $j \in \mathcal{F}_+$, the result in the case $\boldsymbol{\Omega}_j = \mathrm{O}_g(n)$, we have that $\{1/(n-q)\}\left(\mathrm{Cp}_j - \mathrm{Cp}_{j_0}\right) > 0$. When $j \in \mathcal{F}_-$, we can see that for $j \in \mathcal{F}_-$

$$\frac{1}{q(n-q)}\{\mathrm{Cp}_j - \mathrm{Cp}_{j_0}\} \xrightarrow{p} \frac{1}{1-d}\mathrm{tr}\boldsymbol{\Xi}_j^*.$$

This implies Theorem 4.4 in the case Cp.

Finally we show that the above consistency properties of Cp hold for MCp. We have seen that when $\boldsymbol{\Omega}_j = \mathrm{O}_g(n)$,

$$\begin{aligned}\frac{n}{n-q}\left(\mathrm{tr}\hat{\boldsymbol{\Sigma}}_j\mathbf{S}^{-1} - \mathrm{tr}\hat{\boldsymbol{\Sigma}}_{j_0}\mathbf{S}^{-1}\right) &= \mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1} - \mathrm{tr}\mathbf{B}_{(j_0)}\mathbf{W}_{(j_0)}^{-1} \\ &\xrightarrow{p} \frac{1}{1-d}\{\mathrm{tr}\boldsymbol{\Delta}_j^* - d(k_j-k_0)\}.\end{aligned}$$

When $\boldsymbol{\Omega}_j = \mathrm{O}_g(nq)$, we have seen that

$$\begin{aligned}\frac{n}{q(n-q)}\left(\mathrm{tr}\hat{\boldsymbol{\Sigma}}_j\mathbf{S}^{-1} - \mathrm{tr}\hat{\boldsymbol{\Sigma}}_{j_0}\mathbf{S}^{-1}\right) &= \frac{1}{q}\left(\mathrm{tr}\mathbf{B}_{(j)}\mathbf{W}_{(j)}^{-1} - \mathrm{tr}\mathbf{B}_{(j_0)}\mathbf{W}_{(j_0)}^{-1}\right) \\ &\xrightarrow{p} \frac{1}{1-d}\mathrm{tr}\boldsymbol{\Xi}_j^*.\end{aligned}$$

Using these results and a relationship between Cp and MCp given in (2.17), we have that MCp has the same consistency properties as the ones of Cp.

Acknowledgement

The first author's research was in part supported by Grant-in-Aid for Research Activity Start-up 25880017. The third author's research was in part supported by Grant-in-Aid for Scientific Research (C) 25330038.

References

[1] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[2] Enomoto, R., Sakurai, T. and Fujikoshi, Y. (2013). Consistency of AIC and its modifications in the growth curve model when the number of groups and the sample size are large. *SUT Journal of Mathematics*. **49**, 93-107.

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

[4] Fujikoshi, Y., Enomoto, R. and Sakurai, T. (2013). High-dimensional AIC in the growth curve model. *J. Multivariate Anal.*, **122**, 239–250.

[5] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **123**, 184–200.

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

[7] 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.

[8] Satoh, K. and Yanagihara, H. (2010). Estimation of varying coefficients for a growth curve model. *Amer. J. Math. Management Sci.*, **30**, 243–256.

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

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

[11] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2014). A consistency property of AIC for multivariate linear model when the dimension and the sample size are large. Submitted for publication.

Rie Enomoto
Department of Mathematical Information Science, Tokyo University of Science
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
E-mail: `j1410701@yahoo.co.jp`

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

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