Variable Selection in Multivariate Linear Regression Models with Fewer Observations than the Dimension

(Last Modified: November 3 2008)

Mariko YAMAMURA[1], Hirokazu YANAGIHARA[2] AND Muni S. SRIVASTAVA[3]

[1]*Division of Biostatistics, School of Pharmaceutical Sciences, Kitasato University*
5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

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

[3]*Department of Statistics, University of Toronto*
100 St. George Street, Toronto, Ontario ON M5S 3G3, Canada

Abstract

This paper deals with selection of variables in multivariate linear regression models with fewer observations than the dimension by using Akaike's information criterion (AIC). It is well known that the AIC cannot be defined when the dimension of an observation is larger than the sample size, because an ordinary estimator of the covariance matrix becomes singular. By replacing the ordinary estimator of the covariance matrix with its ridge-type estimator, we propose a new AIC for selecting variables of multivariate linear regression models even though the dimension of an observation is larger than the sample size. The bias correction term of AIC is evaluated from a remarkable asymptotic theory based on the dimension and the sample size approaching to ∞ simultaneously. By conducting numerical studies, we verify that our new criterion perform well.

AMS 2000 *subject classifications*: Primary 62H12; Secondary 62F07.
Key words: AIC, Empirical Bayes estimator, High dimensional data, Multivariate linear regression model, Selection of variables.

1. Introduction

[2]Corresponding author, E-mail: *yanagi@math.sci.hiroshima-u.ac.jp*

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ matrix of response variables and $\boldsymbol{X}_\omega$ be an $n \times K$ matrix of explanatory variables of full rank K. Suppose that $\boldsymbol{X}$ denotes the $n \times k$ matrix consisting of the columns of $\boldsymbol{X}_\omega$. Then we consider the following multivariate linear regression model with k explanatory variables as the candidate model:

$$M \ : \ \boldsymbol{Y} \ \sim \ N_{n\times p}(\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n), \tag{1.1}$$

(for more detailed information of the multivariate linear regression model, see for examples, Srivastava, 2002, Chapter 9; Timm, 2002, Chapter 4, etc.). We call the model with $\boldsymbol{X} = \boldsymbol{X}_\omega$ the full model. Assume that the true model of $\boldsymbol{Y}$ is expressed as

$$M_* \ : \ \boldsymbol{Y} \ \sim \ N_{n\times p}(\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n), \tag{1.2}$$

where $\boldsymbol{\Gamma}_*$ and $\boldsymbol{\Sigma}_*$ are $n \times p$ and $p \times p$ unknown location and dispersion matrices. In particular, we will assume that a candidate model M is an overspecified model, satisfying $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$, where $\boldsymbol{P}_{\boldsymbol{X}}$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}$, i.e., $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Akaike's information criterion (Akaike, 1973; 1974) is widely used for model selection. The AIC cannot be defined when the dimension p is larger than the sample size n, because an ordinary estimator of $\boldsymbol{\Sigma}$ becomes singular, and thus its inverse matrix does not exist and a logarithm of its determinant becomes $-\infty$. Moreover, there is another trouble in the common AIC. Generally, AIC-type criterion can be defined by adding an estimated bias to the sample Kullback-Leibler (KL) discrepancy (Kullback & Leibler, 1951) function. A bias correction term is evaluated from an ordinary asymptotic theory based on $n \to \infty$ under the fixed p. However, such an ordinary asymptotic theory breaks down when $p > n$. Then the bias correction term does not give a good approximation. Similarly, the Bayesian information criterion (BIC) of Schwarz (1978) is not available because the maximum likelihood estimator (MLE) of $\boldsymbol{\Sigma}$ does not exist when $n < p$ and the Laplace approximation of the integral is no longer available.

An aim of this paper is to propose AIC for selecting variables in multivariate linear model (1.1) when the dimension p is larger than the sample size n. In order to avoid the singularity of an estimated $\boldsymbol{\Sigma}$, we use a ridge-type estimator as in Srivastava and Kubokawa (2007) and Kubokawa and Srivastava (2008). Moreover, we revaluate a bias correction term from a remarkable asymptotic theory based on not only $n \to \infty$ but $p \to \infty$ and $n \to \infty$ simultaneously. By adding the renewal bias correction term to

the sample KL discrepancy with the ridge-type estimator of $\mathbf{\Sigma}$ instead of the ordinary estimator, we propose a new AIC for selecting variables of multivariate linear regression models even when p is larger than n. Through the above framework, Srivastava and Kubokawa (2008) proposed their new AIC for selecting the degree of polynomials of growth curves. Our new AIC can be regarded as an extended versions of their AIC.

This paper is organized in the following ways: In Section 2, we propose the new AIC when $p > n$. In Section 3, we verify performances of proposed criteria by conducting numerical studies. In Section 4, we give discussions and conclusions.

2. AIC for Multivariate Linear Regression Model

2.1. Common Setting $n > p$

In order to consider the case of $p > n$, we first describe for the case $n > p$ in this subsection. Let us consider the following function measuring the fit of the candidate model by the KL discrepancy:

$$d_{\mathrm{KL}}(\mathbf{\Theta}, \mathbf{\Sigma}, \boldsymbol{Y}) = np \log 2\pi + n \log |\mathbf{\Sigma}| + \mathrm{tr}\left\{\mathbf{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta})'(\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta})\right\}. \tag{2.1}$$

The MLEs of $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ are obtained by minimizing (2.1). It is well known that MLEs of $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ are

$$\hat{\mathbf{\Theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \quad \hat{\mathbf{\Sigma}} = \frac{1}{n}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}.$$

Let $\boldsymbol{U}$ be an $n \times p$ matrix of random variables which is independent of $\boldsymbol{Y}$ and whose distribution is the same as that of $\boldsymbol{Y}$. Then, $\boldsymbol{U}$ is regarded as a future observation or imaginary new observation. As a criterion for the goodness of fit of the candidate model, we consider the following risk function based on the prediction:

$$\begin{aligned} R_{\mathrm{KL}} &= E^*_{\boldsymbol{Y}} E^*_{\boldsymbol{U}}[d_{\mathrm{KL}}(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Sigma}}, \boldsymbol{U})] \\ &= np \log 2\pi + nE^*_{\boldsymbol{Y}}[\log |\hat{\mathbf{\Sigma}}|] \\ &\quad + nE^*_{\boldsymbol{Y}}[\mathrm{tr}(\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-1})] + E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\hat{\mathbf{\Sigma}}^{-1}(\mathbf{\Gamma}_* - \boldsymbol{X}\hat{\mathbf{\Theta}})'(\mathbf{\Gamma}_* - \boldsymbol{X}\hat{\mathbf{\Theta}})\right\}\right], \end{aligned} \tag{2.2}$$

where $E^*_{\boldsymbol{Y}}$ and $E^*_{\boldsymbol{U}}$ denote expectations with respect to $\boldsymbol{Y}$ and $\boldsymbol{U}$, respectively. We regard the candidate model with the smallest risk as the best model. However, we cannot use R_{KL} directly because (2.2) includes unknown parameters. Therefore, an estimator of risk

function is used for model selection instead of R_{KL}. It is well known that a rough estimator of R_{KL} is $d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{Y})$. Unfortunately, this rough estimator has a bias given by

$$B_{\mathrm{KL}} = R_{\mathrm{KL}} - E^*_{\boldsymbol{Y}}[d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{Y})]. \tag{2.3}$$

By adding an estimated bias to the rough estimator, the AIC-type criteria can be defined as

$$\mathrm{AIC} = d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{Y}) + \hat{B}_{\mathrm{KL}}. \tag{2.4}$$

Thus, AIC-type criteria is specified by the individual $\hat{B}_{\mathrm{KL}}$.

From an original idea of Akaike (1973; 1974), B_{KL} can be approximated by $2\times$(the number of parameters), i.e., $2pk + p(p+1)$. However, in several situations, the approximated bias has an additional bias. Therefore, many authors have improved the original AIC by evaluating B_{KL} under several assumptions on the true model. When $p = 1$, Sugiura (1978) and Hurvich and Tsai (1989) proposed a bias-corrected AIC by adding an exact B_{KL} under the overspecified model (for more detail on other criteria in the univariate case, see e.g., Konishi & Kitagawa, 2008). By applying their calculations for obtaining the exact B_{KL} to the multivariate case, Bedrick and Tsai (1994) proposed a bias-corrected AIC in the multivariate linear regression model. Fujikoshi and Satoh (1997) extended Bedrick and Tsai's bias-corrected AIC so that the bias can be reduced even if the candidate model is underspecified. The above criteria were proposed under the assumption that a distribution of the true model is the normal distribution. For a violation to normality of the true model, Fujikoshi, Yanagihara and Wakaki (2005) constructed $\hat{B}_{\mathrm{KL}}$ by the bootstrap method. Yanagihara (2006) proposed AIC which is partially constructed by the jackknife method and is adjusted to an exact unbiased estimator of the risk function when the candidate model is correctly specified. Furthermore, Yanagihara, Kamo and Tonda (2006) corrected an additional bias of Yanagihara's bias-corrected AIC.

Although there are many AICs in (1.1) like previous studies above, these results were obtained from an asymptotic theory based on $n \to \infty$ under the fixed p. Hence, these criteria do not work well when p is nearly equal to n. Moreover, if p becomes larger than n, then $\hat{\boldsymbol{\Sigma}}$ becomes singular matrix. This implies that $\hat{\boldsymbol{\Sigma}}^{-1}$ does not exist and $\log|\hat{\boldsymbol{\Sigma}}|$ becomes $-\infty$. Consequently, AIC and its bias-corrected AIC under the ordinary framework cannot be defined when $p > n$. Nevertheless, we can develop AIC-type criteria even when $p > n$.

2.2. Main Result

In order to guarantee the nonsingularity of an estimator of $\boldsymbol{\Sigma}$, we use the following ridge-type estimator instead of $\hat{\boldsymbol{\Sigma}}$ as in Srivastava and Kubokawa (2008).

$$\hat{\boldsymbol{\Sigma}}_\lambda = \hat{\boldsymbol{\Sigma}} + \frac{\lambda}{n}\boldsymbol{I}_p. \tag{2.5}$$

Let $\alpha_i = \mathrm{tr}(\boldsymbol{\Sigma}_*^i)/p$ $(i = 1, 2)$ and $\boldsymbol{V}$ be a $p \times p$ matrix given by

$$\boldsymbol{V} = n\hat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}. \tag{2.6}$$

Then an unbiased and consistent estimator of α_1 is given by

$$\hat{\alpha}_1 = \frac{1}{p(n-k)}\mathrm{tr}(\boldsymbol{V}). \tag{2.7}$$

From Srivastava and Kubokawa (2007) and Kubokawa and Srivastava (2008), the following λ is chosen by the empirical Bayes method:

$$\lambda = \sqrt{p}\hat{\alpha}_1. \tag{2.8}$$

Note that $\hat{\boldsymbol{\Sigma}}_\lambda$ are independent of $\hat{\boldsymbol{\Theta}}$. By using $\hat{\boldsymbol{\Sigma}}_\lambda$, R_{KL} in (2.2) is redefined as

$$\begin{aligned} R_{\mathrm{KL},\lambda} &= E^*_{\boldsymbol{Y}} E^*_{\boldsymbol{U}}[d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_\lambda, \boldsymbol{U})] \\ &= np\log 2\pi + nE^*_{\boldsymbol{Y}}[\log|\hat{\boldsymbol{\Sigma}}_\lambda|] + (n+k)E^*_{\boldsymbol{Y}}[\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}_*)]. \end{aligned} \tag{2.9}$$

From the simple calculation, we have

$$d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_\lambda, \boldsymbol{Y}) = np(\log 2\pi + 1) + n\log|\hat{\boldsymbol{\Sigma}}_\lambda| - \lambda\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1}).$$

Therefore, the bias of $d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_\lambda, \boldsymbol{Y})$ for $R_{\mathrm{KL},\lambda}$ becomes

$$\begin{aligned} B_{\mathrm{KL},\lambda} &= R_{\mathrm{KL},\lambda} - E^*_{\boldsymbol{Y}}[d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_\lambda, \boldsymbol{Y})] \\ &= -np + E^*_{\boldsymbol{Y}}[\lambda\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1})] + n(n+k)E^*_{\boldsymbol{Y}}\left[\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{V}_\lambda^{-1})\right]. \end{aligned} \tag{2.10}$$

From (2.10), we can see that it is necessary to calculate only $E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{V}_\lambda^{-1})]$ for evaluating the bias of AIC.

For calculating the expectation, there is an useful result in Srivastava and Kubokawa (2008). We shall assume the following conditions:

(C.1) $0 < \lim_{p\to\infty}\alpha_i \equiv \alpha_{i0} < \infty$ $(i = 1, 2)$,

(C.2) $n-k=O(p^{\delta})$ for $0<\delta\leq 1/2$,

(C.3) the maximum eigenvalue of $\boldsymbol{\Sigma}_*$ is bounded in large p.

Then, the following lemma holds (the proof was shown in Srivastava & Kubokawa, 2008).

LEMMA 1. *Let $\boldsymbol{Z}$ be a $p\times p$ matrix of random variables distributed according to $W_p(m,\boldsymbol{\Sigma}_*)$ and $\boldsymbol{Z}_\beta$ denote $\boldsymbol{Z}_\beta=\boldsymbol{Z}+\beta\boldsymbol{I}_p$, where $\beta=\mathrm{tr}(\boldsymbol{Z})/(m\sqrt{p})$. Then, under assumptions that C.1, C.2 and C.3 are satisfied, an expectation of $\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{Z}_\beta^{-1})$ can be expanded as*

$$E[\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{Z}_\beta^{-1})]=b_p(m,\boldsymbol{\alpha})+o(p^{-1/2}).$$

where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2)'$ and $b_p(m,\boldsymbol{\alpha})$ is defined by

$$b_p(m,\boldsymbol{\alpha})=\sqrt{p}\left\{1+\frac{\alpha_2}{p\alpha_1^2}\left(\frac{2}{m}-\frac{m\sqrt{p}}{1+\sqrt{p}}\right)\right\}. \tag{2.11}$$

By applying lemma 1 to (2.10), we obtain the following expansion.

$$B_{\mathrm{KL},\lambda}=-np+E_{\boldsymbol{Y}}^*[\lambda\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1})]+n(n+k)b_p(n-k,\boldsymbol{\alpha})+o(p^{-1/2}). \tag{2.12}$$

In order to define a new information criterion, we have to estimate the bias, because the bias in (2.12) includes unknown parameters α_1 and α_2. From Srivastava (2005), an unbiased estimator and consistent estimator of α_2 is given by

$$\hat{\alpha}_2=\frac{1}{(n-k-1)(n-k+2)p}\left\{\mathrm{tr}(\boldsymbol{V}^2)-\frac{1}{n-k}(\mathrm{tr}\boldsymbol{V})^2\right\}.$$

Let $\hat{\boldsymbol{\alpha}}=(\hat{\alpha}_1,\hat{\alpha}_2)'$. Then, we can estimate the bias as

$$\hat{B}_{\mathrm{KL},\lambda}=-np+\lambda\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1})+n(n+k)b_p(n-k,\hat{\boldsymbol{\alpha}}).$$

This leads us to the following criterion:

$$\begin{aligned}\mathrm{AIC}_\lambda^{(1)}&=d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\Sigma}}_\lambda,\boldsymbol{Y})+\hat{B}_{\mathrm{KL},\lambda}\\&=np\log 2\pi+n\log|\hat{\boldsymbol{\Sigma}}_\lambda|+n(n+k)b_p(n-k,\hat{\boldsymbol{\alpha}}).\end{aligned} \tag{2.13}$$

On the other hand, we note that an estimator which has a smaller bias evaluates $B_{\mathrm{KL},\lambda}$ more correctly. However, if the candidate model is not overspecified, $\hat{\boldsymbol{\alpha}}$ could have a large bias because an expectation of $\boldsymbol{V}$ becomes $(n-k)\boldsymbol{\Sigma}_*+\boldsymbol{\Gamma}_*'(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\Gamma}_*$. In order

to make such a bias as small as possible, we use the full model to estimate $\boldsymbol{\alpha}$. Let $\boldsymbol{W} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{Y}$. Then we define other estimator of $\boldsymbol{\alpha}$ as $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2)'$, where $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ are given by

$$\tilde{\alpha}_1 = \frac{1}{p(n-K)}\mathrm{tr}(\boldsymbol{W}),\ \tilde{\alpha}_2 = \frac{1}{(n-K-1)(n-K+2)p}\left\{\mathrm{tr}(\boldsymbol{W}^2) - \frac{1}{n-K}(\mathrm{tr}\boldsymbol{W})^2\right\}.$$

By replacing $\hat{\boldsymbol{\alpha}}$ in (2.13) with $\tilde{\boldsymbol{\alpha}}$, we can define the following AIC:

$$\mathrm{AIC}_\lambda^{(2)} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}_\lambda| + n(n+k)b_p(n-k, \tilde{\boldsymbol{\alpha}}). \tag{2.14}$$

It is noted that the bias correction terms of $\mathrm{AIC}_\lambda^{(1)}$ and $\mathrm{AIC}_\lambda^{(2)}$ depend on the data, it may be affected by random fluctuation. Another choice is to use the rough approximations such that $\alpha_1 = \alpha_2 = 1$, and the resulting information criterion is given as follows:

$$\mathrm{AIC}_\lambda^{(3)} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}_\lambda| + n(n+k)b_p(n-k, \mathbf{1}_2), \tag{2.15}$$

where $\mathbf{1}_2 = (1,1)'$.

3. Numerical Studies

3.1. Other Criteria

In previous section, we proposed three AICs based on the asymptotics $p \to \infty$ and $n \to \infty$ simultaneously. However, even if $\hat{\boldsymbol{\Sigma}}_\lambda$ is used as the estimator of $\boldsymbol{\Sigma}$, we can formally define the following AIC which is corresponding to the formula of the crude AIC:

$$\mathrm{AIC}_\lambda^{(0)} = d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_\lambda, \boldsymbol{Y}) + 2\{pk + p(p+1)\}.$$

In order to propose four criteria $\mathrm{AIC}_\lambda^{(j)}$ $(j = 0, 1, 2, 3)$, we avoid the singularity of an estimator of $\boldsymbol{\Sigma}$ by adding $(\lambda/n)\boldsymbol{I}_p$ to $\hat{\boldsymbol{\Sigma}}$. However, we can avoid the singularity by another way if we allow model misspecification. Another choice is to omit correlations between $\boldsymbol{y}_i$, namely, we assume $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_p$ or $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, tentatively. Let $\hat{\boldsymbol{\Sigma}}_{(s)} = \mathrm{tr}(\hat{\boldsymbol{\Sigma}})\boldsymbol{I}_p/p$ and $\hat{\boldsymbol{\Sigma}}_{(d)} = \mathrm{diag}(\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp})$, where $\hat{\sigma}_{ab}$ is the (a,b)th element of $\hat{\boldsymbol{\Sigma}}$. We denote AICs under assumptions $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_p$ and $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ as $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$, respectively. From the original idea of AIC, these two AICs are defined by

$$\mathrm{AIC}_{(s)} = d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_{(s)}, \boldsymbol{Y}) + 2(pk+1)$$

$$
\begin{aligned}
&= np(\log 2\pi + 1) + np\log\{\mathrm{tr}(\hat{\boldsymbol{\Sigma}})/p\} + 2(pk+1), \\
\mathrm{AIC}_{(d)} &= d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_{(d)}, \boldsymbol{Y}) + 2p(k+1) \\
&= np(\log 2\pi + 1) + n\sum_{j=1}^{p} \log \hat{\sigma}_{jj} + 2p(k+1).
\end{aligned}
$$

The effect of correlations between $\boldsymbol{y}_i$ to model selection can be studied by comparing with $\mathrm{AIC}_{\lambda}^{(j)}$, $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$.

It is noted that $\mathrm{AIC}_{\lambda}^{(0)}$, $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ are the motivated from the conventional AIC but no justification can be guaranteed in the asymptotics of $p \to \infty$.

3.2. Simulation

We evaluate the proposed criteria applied numerically to the polynomial regression model, $\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_p)$ with $n = 20$, $K = 8$ and $p = 20$ or 100, where

$$
\boldsymbol{\Gamma}_* = \boldsymbol{X}_\omega \boldsymbol{\Theta}_*, \quad \boldsymbol{\Sigma}_* = \mathrm{diag}(\psi_1, \ldots, \psi_p)\boldsymbol{\Phi}\mathrm{diag}(\psi_1, \ldots, \psi_p).
$$

In this numerical study, we chose the following $\boldsymbol{X}_\omega$, $\boldsymbol{\Theta}_*$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ as the true model

$$
\begin{aligned}
(\boldsymbol{X}_\omega)_{ij} &= z_i^{j-1} \ (i = 1, \ldots, n; j = 1, \ldots, K), \\
(\boldsymbol{\Theta}_*)_{ij} &= \begin{cases} -\delta\{1 + (p-j+1)/p\} & (i = 1; j = 1, \ldots, p) \\ \delta\{1 + (p-j+1)/p\} & (i = 2, 4; j = 1, \ldots, p) \\ -\delta\{2 + (p-j+1)/p\} & (i = 3; j = 1, \ldots, p) \\ 0 & (i = 5, \ldots, K; j = 1, \ldots, p) \end{cases}, \\
\psi_j &= 2 + (p-j+1)/p \ (j = 1, \ldots, p), \\
(\boldsymbol{\Phi})_{ij} &= \rho^{|i-j|^{1/7}} \ (i = 1, \ldots, p; j = 1, \ldots, p),
\end{aligned}
$$

where an notation $(\boldsymbol{A})_{ij}$ denotes the (i,j)th element of matrix $\boldsymbol{A}$ and $z_1, \ldots, z_n$ are independent random variables from $U(-1, 1)$. Let M_j denote the jth candidate model $(j = 0, \ldots, K-1)$. Note that the candidate models are nested and $\boldsymbol{X}$ in M_j is the submatrix consisting of the first $(j+1)$ columns of $\boldsymbol{X}_\omega$. In a sense, the subindex j express the degree of a polynomial regression in M_j. Moreover, we chose $\delta = 0.0$ or 8.0. It means that there are two types of true model, i.e., a constant model M_0 $(\delta = 0.0)$ and a third degree polynomial model M_3 $(\delta = 8.0)$.

We compared four criteria $\mathrm{AIC}_{\lambda}^{(j)}$ $(j = 0, 1, 2, 3)$ and two criteria $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ on the selection probability of the model chosen by minimizing the criterion. It was evaluated by the Monte Carlo simulation with $10,000$ iterations. The tables 1 and 2 show

the selection probabilities in the case of $\delta = 0.0$ and $\delta = 8.0$, respectively. From tables, we can see that $\mathrm{AIC}_{\lambda}^{(2)}$ and $\mathrm{AIC}_{\lambda}^{(3)}$ selected the true model as the best models with higher probabilities than other criteria. Especially, probabilities to select the true model were over 95 % in any cases. A performance of $\mathrm{AIC}_{\lambda}^{(0)}$ was good when p is large. On the contrary, it became quite bad when p is small. Then, the probability to select the full model became high. It is well known that this is an impermissible character for a variable selector. Moreover, we found that $\mathrm{AIC}_{\lambda}^{(1)}$ tends to choose the model having the smallest number of explanatory variables as the best model. Therefore, when M_3 was the true model, $\mathrm{AIC}_{\lambda}^{(1)}$ worked abnormally. Furthermore, we can see that performances of $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ are not too bad when ρ is low. However, when ρ is high, the performances became worse than that of $\mathrm{AIC}_{0}^{(0)}$. This result means that we should consider correlations to evaluate the goodness of fit of a statistical model correctly if response variables are not independent. We have studied several other settings for simulation, and have obtained similar results.

Please insert Tables 1 and 2 around here

3.3. An Example

Next, by using a real data in Wille et al. (2004), we show an example of model selection. This data is a microarray experiment. There are 795 genes which may show some association to 39 genes from two biosynthesis pathway in Arabidopsis thaliana. We apply 795 genes to response variables ($p = 795$) and 39 genes to explanatory variables. All variables are logarithmic transformed. Since the sample size is $n = 118$, the dimension p is greater than n in this data. In our analysis, the number of explanatory variables in the full model is $K = 40$, because we always add a constant term to a regression model.

Since the number of candidate models is too large (2^{39}), it is impossible to calculate and compare with the information criteria of all candidate models. Moreover, we cannot use leap-and-bound algorithm (Furnival & Wilson, 1974) or branch-and-bound algorithm (Gatu & Kontoghiorghes, 2006), which are algorithms for searching the best subset of regression models effectively, because $d_{\mathrm{KL}}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}_{\lambda}, \boldsymbol{Y})$ does not decrease when an extra explanatory variable adds to $\boldsymbol{X}$. Therefore, we search the best subset by the forward stepwise selection. Since p is very large, large computational tasks are needed for calculating $|\hat{\boldsymbol{\Sigma}}_{\lambda}|$ and $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\lambda}^{-1})$. In order to save computational tasks, we use the following

formulas to obtain $\log|\hat{\boldsymbol{\Sigma}}_\lambda|$ and $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1})$.

$$\log|\hat{\boldsymbol{\Sigma}}_\lambda| = p\log(\lambda/n) + \log|\boldsymbol{I}_n + \lambda^{-1}\boldsymbol{R}\boldsymbol{R}'|,$$
$$\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_\lambda^{-1}) = np - \lambda^{-1}\mathrm{tr}\{\boldsymbol{R}\boldsymbol{R}'(\boldsymbol{I}_n + \lambda^{-1}\boldsymbol{R}\boldsymbol{R}')^{-1}\},$$

where $\boldsymbol{R} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$. The table 3 shows selected subsets of explanatory variables by each criterion. Names of explanatory variables are corresponding to gene's names described in Lange and Ghassemian (2003). From table 3, we can see that the number of explanatory variables in the best model chosen by $\mathrm{AIC}_\lambda^{(2)}$ is the smallest among all criteria. Criteria $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ chose too many explanatory variables as the best subset.

Please insert Table 3 around here

4. Conclusion and Discussion

In this paper, we proposed new three AICs, $\mathrm{AIC}_\lambda^{(1)}$, $\mathrm{AIC}_\lambda^{(2)}$ and $\mathrm{AIC}_\lambda^{(3)}$, for selecting variables in the multivariate linear model with $p > n$. These are constructed by adding renewal bias correction terms evaluated from a remarkable asymptotic theory, which is based on $p \to \infty$ and $n \to \infty$ simultaneously, to the sample KL discrepancy with $\hat{\boldsymbol{\Sigma}}_\lambda$ instead of $\hat{\boldsymbol{\Sigma}}$. The proposed AICs are specified by how to estimate $\boldsymbol{\alpha}$, i.e., $\hat{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\alpha}}$ and $\mathbf{1}_2$ are used in $\mathrm{AIC}_\lambda^{(1)}$, $\mathrm{AIC}_\lambda^{(2)}$ and $\mathrm{AIC}_\lambda^{(3)}$, respectively.

A simulation shows that performances of $\mathrm{AIC}_\lambda^{(2)}$ and $\mathrm{AIC}_\lambda^{(3)}$ were better than $\mathrm{AIC}_\lambda^{(1)}$, $\mathrm{AIC}_\lambda^{(0)}$, $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$. Especially, in all cases, $\mathrm{AIC}_\lambda^{(2)}$ and $\mathrm{AIC}_\lambda^{(3)}$ selected the true model as the best models with high probabilities. Moreover, in $\mathrm{AIC}_\lambda^{(1)}$, the probability to select the model having the smallest number of explanatory variables as the best model became the highest in all cases. The reason why $\mathrm{AIC}_\lambda^{(1)}$ does not work well is that $\hat{\boldsymbol{\alpha}}$ is used as an estimator of $\boldsymbol{\alpha}$. As stated in Section 2, $\hat{\boldsymbol{\alpha}}$ has a constant bias when the candidate model is not overspecified. Such a bias becomes larger as the number of explanatory variables is decreasing. A large bias causes $\hat{\boldsymbol{\alpha}}$ to increase. An increased $\hat{\boldsymbol{\alpha}}$ leads to a negative $b_p(n-k, \hat{\boldsymbol{\alpha}})$ which is an essential bias correction term. This implies that $\mathrm{AIC}_\lambda^{(1)}$ chooses the model having the smallest number of explanatory variables as the best model. The bias existing in $\hat{\boldsymbol{\alpha}}$ does not appear in $\tilde{\boldsymbol{\alpha}}$. Therefore, $\mathrm{AIC}_\lambda^{(2)}$ works better than $\mathrm{AIC}_\lambda^{(1)}$. On the other hand, $\mathrm{AIC}_\lambda^{(3)}$ tends to have smaller variance than $\mathrm{AIC}_\lambda^{(1)}$, because the $b_p(n-k, \mathbf{1}_2)$ is not random although $b_p(n-k, \hat{\boldsymbol{\alpha}})$ is random. This property

may lead to the good performance of $\mathrm{AIC}_{\lambda}^{(3)}$. In addition, we can see that the higher correlations between $\boldsymbol{y}_i$ were, the worse performances of $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ became. This tendency shows that when there are correlations between $\boldsymbol{y}_i$, the correlation should be evaluated in the discrepancy function for the model selection.

An example of model selection using the real data shows that $\mathrm{AIC}_{\lambda}^{(2)}$ and $\mathrm{AIC}_{\lambda}^{(3)}$ chose the smaller number of explanatory variables than $\mathrm{AIC}_{\lambda}^{(1)}$, $\mathrm{AIC}_{\lambda}^{(0)}$, $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$. This indicates that $\mathrm{AIC}_{\lambda}^{(2)}$ and $\mathrm{AIC}_{\lambda}^{(3)}$ work well. On the other hands, $\mathrm{AIC}_{(s)}$ and $\mathrm{AIC}_{(d)}$ chose too many explanatory variables as the best subset. This indicates that these two criteria do not work well. One reason for this is because correlations between response variables were omitted to calculate the criteria although response variables have clear correlations.

From the viewpoint mentioned above, we recommend the use of $\mathrm{AIC}_{\lambda}^{(2)}$ or $\mathrm{AIC}_{\lambda}^{(3)}$ for selecting variables in multivariate linear regression model with $p > n$.

Acknowledgments

Hirokazu Yanagihara's research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010.

References

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

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

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

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

[5] Fujikoshi, Y., Yanagihara, H. & Wakaki, H. (2005). Bias corrections of some criteria for selection multivariate linear regression models in a general case. *Amer. J. Math. Management Sci.*, **25**, 221–258.

[6] Furnival, G. M. & Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, **16**, 499–511.

[7] Gatu, C. & Kontoghiorghes, E. J. (2006). Branch-and-bound algorithms for computing the best subset regression models. *J. Comput. Graph. Statist.*, **15**, 139–156.

[8] Hurvich, C. M. & Tsai, C.-L. (1989). Regression and times series model selection in small samples. *Biometrika*, **50**, 226–231.

[9] Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Science+Business Media, LLC, New York.

[10] Kubokawa, T. & Srivastava, M. S. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *J. Multivariate Anal.*, **99**, 1906–1928.

[11] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, **22**, 79–86.

[12] Lange, B. & Ghassemian, M. (2003). Genome organization in Arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol. Biol.*, **51**, 925-948.

[13] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[14] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, New York.

[15] Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan. Statist. Soc.*, **35**, 251–272.

[16] Srivastava, M. S. & Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *J. Japan. Statist. Soc.*, **37**, 123–134.

[17] Srivastava, M. S. & Kubokawa, T. (2008). Akaike information criterion for selecting components of the mean vector in high dimensional data with fewer observations. *J. Japan. Statist. Soc.*, **38**, 259–283.

[18] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theory Methods*, **A7**, 13–26.

[19] Timm, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag, New York.

[20] Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. & Bühlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.*, **5**, 1–13.

[21] Yanagihara, H. (2006). Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. *J. Multivariate Anal.*, **97**, 1070–1089.

[22] Yanagihara, H., Kamo, K. & Tonda, T. (2006). Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *TR No. 06-01, Statistical Research Group, Hiroshima University*, Hiroshima, Japan.

TABLE 1. The selection probability of the model chosen by each criterion for 10,000 repetitions ($\delta = 0.0$)

p	ρ	Criterion	Probability (%)							
			M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7
20	0.2	$\mathrm{AIC}_\lambda^{(0)}$	86.50	7.85	1.86	1.01	0.68	0.65	0.66	0.79
		$\mathrm{AIC}_\lambda^{(1)}$	90.72	7.33	1.33	0.39	0.12	0.05	0.05	0.01
		$\mathrm{AIC}_\lambda^{(2)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	98.43	1.33	0.20	0.03	0.01	0.00	0.00	0.00
		$\mathrm{AIC}_{(s)}$	97.43	2.01	0.25	0.10	0.07	0.01	0.08	0.05
	0.8	$\mathrm{AIC}_\lambda^{(0)}$	91.22	5.36	1.40	0.59	0.35	0.39	0.26	0.43
		$\mathrm{AIC}_\lambda^{(1)}$	49.05	22.74	12.34	6.87	4.41	2.41	1.47	0.71
		$\mathrm{AIC}_\lambda^{(2)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	85.00	7.36	3.05	1.38	0.80	0.78	0.74	0.89
		$\mathrm{AIC}_{(s)}$	82.01	7.83	3.51	1.72	1.05	1.16	1.03	1.69
100	0.2	$\mathrm{AIC}_\lambda^{(0)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(1)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(2)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	99.99	0.01	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(s)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.8	$\mathrm{AIC}_\lambda^{(0)}$	97.88	1.87	0.23	0.01	0.01	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(1)}$	45.90	22.09	13.54	8.83	4.94	2.74	1.39	0.57
		$\mathrm{AIC}_\lambda^{(2)}$	99.41	0.57	0.02	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	98.55	1.32	0.13	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	89.39	6.00	1.80	1.13	0.63	0.29	0.36	0.40
		$\mathrm{AIC}_{(s)}$	86.99	6.53	2.20	1.41	0.78	0.55	0.61	0.93

TABLE 2. The selection probability of the model chosen by each criterion for 10,000 repetitions ($\delta = 8.0$)

p	ρ	Criterion	Probability (%)							
			M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7
20	0.2	$\mathrm{AIC}_\lambda^{(0)}$	0.00	0.00	0.00	68.28	12.50	7.07	5.16	6.99
		$\mathrm{AIC}_\lambda^{(1)}$	99.96	0.00	0.04	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(2)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	0.00	0.00	0.00	91.06	5.47	1.68	0.83	0.96
		$\mathrm{AIC}_{(s)}$	0.00	0.00	0.00	85.50	7.24	2.90	1.70	2.66
	0.8	$\mathrm{AIC}_\lambda^{(0)}$	0.00	0.00	0.00	79.41	9.70	4.21	3.09	3.59
		$\mathrm{AIC}_\lambda^{(1)}$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(2)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	0.00	0.00	0.00	99.99	0.01	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	0.00	0.00	0.00	75.03	10.15	5.34	4.42	5.06
		$\mathrm{AIC}_{(s)}$	0.00	0.00	0.00	70.57	10.63	6.17	5.23	7.40
100	0.2	$\mathrm{AIC}_\lambda^{(0)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(1)}$	53.79	0.00	46.21	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(2)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_{(d)}$	0.00	0.00	0.00	99.75	0.24	0.01	0.00	0.00
		$\mathrm{AIC}_{(s)}$	0.00	0.00	0.00	99.34	0.58	0.04	0.03	0.01
	0.8	$\mathrm{AIC}_\lambda^{(0)}$	0.00	0.00	0.00	96.20	3.11	0.50	0.18	0.01
		$\mathrm{AIC}_\lambda^{(1)}$	99.24	0.75	0.01	0.00	0.00	0.00	0.00	0.00
		$\mathrm{AIC}_\lambda^{(2)}$	0.00	0.00	0.00	99.58	0.37	0.03	0.02	0.00
		$\mathrm{AIC}_\lambda^{(3)}$	0.00	0.00	0.00	97.91	1.83	0.20	0.06	0.00
		$\mathrm{AIC}_{(d)}$	0.00	0.00	0.00	80.00	9.03	4.41	3.25	3.31
		$\mathrm{AIC}_{(s)}$	0.00	0.00	0.00	76.12	9.64	5.00	4.06	5.18

TABLE 3. Selected explanatory variables by six criteria

Name	$\mathrm{AIC}_{\lambda}^{(0)}$	$\mathrm{AIC}_{\lambda}^{(1)}$	$\mathrm{AIC}_{\lambda}^{(2)}$	$\mathrm{AIC}_{\lambda}^{(3)}$	$\mathrm{AIC}_{(d)}$	$\mathrm{AIC}_{(s)}$
AACT1	○				○	○
AACT2					○	○
CMK					○	○
DPPS1						
DPPS2	○				○	○
DPPS3						○
DXPS1		○				
DXPS2	○				○	○
DXPS3	○	○			○	○
DXR	○		○	○	○	○
FPPS1	○	○			○	○
FPPS2						○
GGPPS1						
GGPPS2						
GGPPS3						
GGPPS4		○			○	
GGPPS5		○				
GGPPS6	○				○	○
GGPPS8					○	○
GGPPS9						
GGPPS10					○	
GGPPS11					○	○
GGPPS12	○			○	○	○
GPPS	○	○			○	○
HDR	○	○			○	○
HDS	○			○	○	○
HMGR1	○		○	○	○	○
HMGR2					○	○
HMGS					○	○
IPPI1	○				○	○
IPPI2		○			○	○
MCT					○	○
MECPS	○		○	○	○	○
MK			○	○	○	○
MPDC1						○
MPDC2					○	○
PPDS1			○	○	○	○
PPDS2	○				○	○
UPPS1	○				○	○
Total Number	16	8	5	7	29	30