HIROSHIMA MATH. J.
30 (2000), 29–53

Modification of *AIC*-type criterion in multivariate normal linear regression with a future experiment

Kenichi SATOH

(Received December 9, 1998)
(Revised March 2, 1999)

ABSTRACT. In this paper we propose a modification of *Predictive AIC* which is an extension of *AIC* to an extrapolation case. This modification reduces bias for both the cases when a candidate model contains the true model and even when it does not contain the true model. Simmulation study shows that our criterion has also a good property in the mean square error.

1. Introduction

We consider multivariate linear regression of response variables $y_1, \ldots, y_p$ on a subset of k_F explanatory variables $x_1, \ldots, x_{k_F}$. Suppose that there are n observations of $y' = (y_1, \ldots, y_p)$ for each fixed explanatory variables $x_F' = (x_1, \ldots, x_{k_F})$. Let Y be an $n \times p$ current observation matrix and X_F an $n \times k_F$ current regression matrix. The multivariate linear regression model including all explanatory variables is written as

$$Y = X_F \Theta_F + \mathscr{E}, \qquad \mathscr{E} \sim N_{n\times p}(O_{n\times p}, \Sigma_F \otimes I_n),$$

where Θ_F is a $k_F \times p$ matrix of unknown parameters, $\mathscr{E}$ is an $n \times p$ error matrix and the rows of $\mathscr{E}$ are assumed to be independently distributed as a p-variate normal distribution with mean zero and covariance matrix Σ_F. We call the model *current full model* or *full model*. The multivariate linear regression has been disscussed in both theoritical and applied statistics, e.g., in a theoritical statistics (Anderson (1958), Rao (1973), Silvey (1970)) and in an applied statistics (Chatterjee and Price (1977), Draper and Smith (1966), Seber (1977)). Mainly our disscussions are based on a multivariate normal distibution. Since our regression model has an normal distributed error matrix, the probability density function of the observation matrix under the full model is given by

$$f_F^Y(Y|\Theta_F, \Sigma_F) = (2\pi)^{-np/2}|\Sigma_F|^{-n/2} \exp\{-\tfrac{1}{2}\mathrm{tr}(Y - X_F\Theta_F)'(Y - X_F\Theta_F)\Sigma_F^{-1}\}.$$

2000 *Mathematics Subject Classification.* 62H12, 62F7.
Key words and phrases. *AIC*, bias reduction, extrapolation, model selection, normal linear regression, small sample.

We consider the problem of selecting a model, from a collection of candidate models specified by linear regression model of y on subvectors of x_F, in order to get better prediction. In general statistical models it is called *model selection* (see Linhart and Zucchini (1986)) and especially, in regression model, it is well-known as a *variable selection* (see Miller (1990)). Here we are concerned with selection methods based on the construction of approximately unbiased estimator of a risk function or an underlying criterion function, in particular the *AIC* (Akaike, 1973). The *AIC* can be motivated by considering the expected log-predictive likelihood or equivalently the Kullback-Leibler information, for a candidate model. It has been pointed by Hurvich and Tsai (1989) that the *AIC* in multivariate normal regression model can drastically underestimate the corresponding risk function when a candidate model is an overspecified model and the sample size is small. Here a candidate model is called an overspecified model or an underspecified model according to whether it does or does not include the true model.

A correction of *AIC* in the multivariate normal linear regression model was proposed by Sugiura (1978), Hurvich and Tsai (1989), providing an exact unbiased estimator for overspecified models. This type of a correction was also proposed in an extended growth curve model by Fujikoshi and Satoh (1996). Further, Fujikoshi and Satoh (1997) pointed that the minimal full model does not always minimize the risk function when the sample size is small. Then they proposed a modification for a class of candidate models including both overspecified and underspecified models. Similarly, Satoh, Kobayashi and Fujikoshi (1997) proposed such a modification of *AIC* in the growth curve model of which the within individual design matrices are balanced type.

The *AIC* has been proposed in order to select a good fitted model for predicting a future observation matrix with the current regression matrix. On the other hand, Satoh (1997) proposed a *Predictive AIC* in order to select a good fitted model for predicting a future observation matrix with a regression matrix different from the current one, in the other words, for an extrapolation case. The criterion is an unbiased estimator of its risk when a candidate model is an overspecified model. It is also an extension of the result of Sugiura (1978) and Hurvich and Tsai (1989). In this paper we propose a modification of the criterion which reduces its bias as much as possible, for both overspecified and underspecified models. In §2 we give some preliminaries for *AIC* and some notes for our frame work of multivariate normal linear regression model with a future experiment. In §3 we disscuss on a Predictive *AIC*, focusing on its difference from *AIC*. In §4 we propose a modification of the Predictive *AIC*, and study its properties. In §5 the proposed criterion is numerically investigated. Proofs for the results in §3 and §4 are presented in §6.

2. Akaike information criterion

2.1. Models

2.1.1. A candidate model

A candidate model can be specified by linear regression model of y on a given subvector of x_F and it is expressed as

$$Y = X\Theta + \mathscr{E}, \qquad \mathscr{E} \sim N_{n\times p}(\mathrm{O}_{n\times p}, \Sigma \otimes I_n),$$

where X is an $n \times k$ $(\leq k_F)$ submatrix of X_F, Θ is a $k \times p$ matrix of unknown parameters, Σ is a $p \times p$ unknown covariance matrix. We call the model *current candidate model.* The probability density function of an observation matrix under the candidate model is given by

$$f^Y(Y|\Theta, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\tfrac{1}{2}\mathrm{tr}(Y - X\Theta)'(Y - X\Theta)\Sigma^{-1}\}.$$

The model may be fitted by maximizing the log-density function with respect to the unknown matrices (Θ, Σ). Let $\hat{\Theta}$ and $\hat{\Sigma}$ be the maximum likelihood estimators of the regression coefficient matrix and the covariance matrix under the candidate model, which are given by

$$\hat{\Theta} = (X'X)^{-1}X'Y, \qquad \hat{\Sigma} = (Y - \hat{Y})'(Y - \hat{Y})/n,$$

respectively. Here $\hat{Y} = X\hat{\Theta}$ is a common predicted matrix of Y under the candidate model. We note that a candidate model is regarded as a set of probability density functions with free (unknown) parameters whose elements have some restrictions. Therefore, a fitted model $f^Y(Y|\hat{\Theta}, \hat{\Sigma})$ is regarded as an element of the candidate model, which may be written as

$$f^Y(Y|\hat{\Theta}, \hat{\Sigma}) \in \{f^Y(Y|\Theta, \Sigma)|(\Theta, \Sigma) \in \mathbf{R}^{kp+p(p+1)/2}\}.$$

2.1.2. The true model

We assume that the true model for a current observation matrix is defined by

$$Y = X_F\Theta_{F_*} + \mathscr{E}, \qquad \mathscr{E} \sim N_{n\times p}(\mathrm{O}_{n\times p}, \Sigma_{F_*} \otimes I_n),$$

where Θ_{F_*} is an $n \times k_F$ true regression coefficient matrix, Σ_{F_*} is a $p \times p$ true covariance matrix. This means that the true model is obtained by specifying a pair of fixed parameter matrices as

$$(\Theta_F, \Sigma_F) = (\Theta_{F_*}, \Sigma_{F_*}),$$

in the current full model. Let f_F^Y be the probability density function of Y

under the current full model. Then the true model is an element of the current full model, i.e.,

$$f_F^Y(Y|\Theta_{F_*}, \Sigma_{F_*}) \in \{f_F^Y(Y|\Theta_F, \Sigma_F)|(\Theta_F, \Sigma_F) \in \mathbf{R}^{k_F p+p(p+1)/2}\}.$$

We may usually discuss on the *minimal full model* which is the minimal candidate model among the candidate models which contains the true model, but we will not refer to the model here. From our assumption on the true model, there exists at least an overspecified model, since the full model is an overspecified model. This implies that the usual estimator of covariance matrix under the full model, $\hat{\Sigma}_F$ is an unbiased estimator for the true covariance matrix Σ_{F_*}. Such a requirement will be reasonable for the case when the number of the available explanatory variables k_F is large but finite. An infinite case is another problem and it is disscussed in Shibata (1981). Our assumption will be too restrictive when k_F is small. The case when the full model does not include the true model should be considered, but it is not treated in this paper.

2.2. Risk

Our risk function for a candidate model is based on a predictive log-density function, or Kullback-Leibler information introduced by Akaike (1973). This type of risk in our frame work is defined by

$$\begin{aligned} R(X) &= -2E_*^{Y, Y_@}[\log f^Y\{Y_@|\hat{\Theta}(Y), \hat{\Sigma}(Y)\}] \\ &= E_*^{Y, Y_@}[n\log|\hat{\Sigma}| + np\log 2\pi + \mathrm{tr}\{(Y_@ - \hat{Y})'(Y_@ - \hat{Y})\hat{\Sigma}^{-1}\}], \end{aligned}$$

where $Y_@$ may be regarded as an $n \times p$ future observation matrix that has the same distribution as Y and is independent of Y, and E_* denotes the expectation under the true model. The loss function $\log f^Y\{Y_@|\hat{\Theta}(Y), \hat{\Sigma}(Y)\}$ is a function of Y and $Y_@$. The $\log f^Y\{\cdot|\hat{\Theta}(Y), \hat{\Sigma}(Y)\}$ is a function of Y, which gives us a goodness of fit of the candidate model for a current observation. On the other hand, $\log f^Y\{Y_@|\cdot\}$ is a function of $Y_@$ which gives an evaluation for predicting a future observation by using the fitted candidate model. Thus the risk function measures a degree of a fitness for both of the current and the future observations. In principle, we should select the candidate model which minimizes the risk function, i.e., the *population best model.*

It may be noted that the population best model is a function of sample size. If the sample size is large, it will be the minimal full model. But, If the sample size is not large, even the minimal full model will not always be the population best model. In general, the population best model may be smaller or more simple than the minimal full model.

2.3. Criteria

The risk function $R(X)$ depends on the parameters under the true model. Therefore it must be estimated. An estimator of the risk function is called a *criterion*. The Akaike information criterion *AIC* (Akaike, 1973) was proposed as an approximately unbiased estimator for $R(X)$ defined by

$$AIC = -2\{\text{maxmum log-likelihood}\} + 2\{\text{number of unknown parameters}\}$$
$$= n\log|\hat{\Sigma}| + np(\log 2\pi + 1) + 2\{kp + \tfrac{1}{2}p(p+1)\}.$$

The population best model may be estimated as the candidate model which minimizes a criterion. We call the estimated model a *sample best model*. The estimation method does not always select the population best model. However, if the criterion is a good estimator of the risk function, we can expect to select the population best model by the criterion with a high probability. Therefore, it is important to develop a good estimator of the risk function. In this paper, we focus on unbiasedness as one of the good properties.

A correction of *AIC* was obtained by Sugiura (1978), Hurvich and Tsai (1989) and Bedrick and Tsai (1994). It is expressed as

$$CAIC = n\log|\hat{\Sigma}| + np\log 2\pi + \frac{(n+k)np}{n-k-p-1},$$

which is an unbiased estimator for an overspecified model. Fujikoshi and Satoh (1997) proposed a modification of *AIC* which is nearly an unbiased estimator in both underspecified and overspecified models. The modofication is given by

$$MAIC = CAIC + 2k\,\mathrm{tr}(\hat{A} - I_p) - \{\mathrm{tr}(\hat{A} - I_p)\}^2 - \mathrm{tr}\{(\hat{A} - I_p)^2\},$$

where

$$\hat{A} = \frac{n-k}{n-k_F}\hat{\Sigma}_F\hat{\Sigma}^{-1}.$$

They also studied its bias theoritically, and showed that the biases of *MPAIC* is almost the same as the ones of *CAIC* for overspecified models, and *MAIC* gives a considerable improvement on biases for underspecified models.

3. Predictive *AIC*

On basis of a current experiment, we often encounter to predict a future observation for a given future regression matrix, which may be different from a current regression matrix. The situation is called the *extrapolation* or *inter-*

polation case, according to whether a future regression matrix is the same as a current regression matrix or not. Here we treat as an extrapolation case without distingushing these two cases. One possible approach is to estimate the parameters based on Bayesian methods (see Keyes and Levy (1996)). But, in this section, we discuss on the prediction problem by using a model selection method which was proposed by Satoh (1997).

3.1. Future models

3.1.1. A future candidate model

Suppose that we want to predict response variables for a future regression matrix of k_F explanatory variables $x_1, \ldots, x_{k_F}$. Let W_F be an $m \times k_F$ future regression matrix and Z be an $m \times p$ future observation matrix which can not be observed, but required to be predicted. For a future experiment, we assume the following multivariate normal linear regression model,

$$Z = W_F \Theta_F^Z + \mathscr{E}^Z, \mathscr{E}^Z \sim N_{m\times p}(\mathrm{O}_{m\times p}, \Sigma_F^Z \otimes I_n), \qquad \mathscr{E}^Z \text{ is independent of } \mathscr{E},$$

where Θ_F^Z is a $k_F \times p$ matrix of unknown parameters, $\mathscr{E}^Z$ is an $m \times p$ error matrix, the rows of $\mathscr{E}^Z$ is assumed to be independently distributed as a p-variate normal distribution with mean zero and unknown covariance matrix Σ_F^Z, and $\mathscr{E}^Z$ is independent with the current error matrix $\mathscr{E}$. The model is called *future full model.* We note that each column of the future regression matrix corresponds to that of the current regression matrix, respectively. A future candidate model corresponding to a current candidate model is expressed as

$$Z = W\Theta^Z + \mathscr{E}^Z, \qquad \mathscr{E}^Z \sim N_{m\times p}(\mathrm{O}_{m\times p}, \Sigma^Z \otimes I_m).$$

Here W is an $m \times k$ $(\leq k_F)$ future regression matrix, Θ^Z is a $k \times p$ matrix of unknown parameters, and Σ^Z is a $p \times p$ unknown covariance matrix. The probability density function of Z under the future candidate model is given by

$$f^Z(Z|\Theta^Z, \Sigma^Z) = (2\pi)^{-mp/2}|\Sigma^Z|^{-m/2} \exp\{-\tfrac{1}{2}\mathrm{tr}(Z - W\Theta^Z)'(Z - W\Theta^Z)(\Sigma^Z)^{-1}\}.$$

Since we have not any future observation for the future regression matrix, we need to estimate the unknown parameters Θ^Z and Σ^Z, based on a current observation matrix. A natural way is to use estimators under the corresponding current candidate model, i.e.,

$$(\hat{\Theta}^Z, \hat{\Sigma}^Z) = (\hat{\Theta}, \hat{\Sigma}).$$

Furthermore, we may predict the distribution of Z by

$$N_{m\times p}(\hat{Z}, \hat{\Sigma}^Z \otimes I_m),$$

where

$$\hat{Z} = W\hat{\Theta}^Z$$
$$= W(X'X)^{-1}X'Y.$$

Such a prediction procedure has been used by Murray (1977), etc. The probability density function of the future fitted model is given by

$$f^Z(Z|\hat{\Theta}^Z, \hat{\Sigma}^Z) = (2\pi)^{-mp/2}|\hat{\Sigma}|^{-m/2}\exp\{-\tfrac{1}{2}\mathrm{tr}(Z-\hat{Z})'(Z-\hat{Z})\hat{\Sigma}^{-1}\}.$$

3.1.2. The future true model

Recall that the true model for a current experiment is given as an element of the current full model obtained by letting $(\Theta_F, \Sigma_F) = (\Theta_{F_*}, \Sigma_{F_*})$. From a close connection between a current experiments and a future experiment, we assume that the true parameters of the future experiment are equal with those of the current one. More precisely, it is assumed that the future true model is obtained from the future full model by letting

$$(\Theta_F^Z, \Sigma_F^Z) = (\Theta_{F_*}, \Sigma_{F_*}).$$

Let $\hat{Z}_F$ be the conditional mean of Z under the fitted future full model, i.e., $\hat{Z}_F = W_F\hat{\Theta}_F^Z$. Then, under the assumption, we have the following property.

PROPERTY 3.1. *The mean of $\hat{Z}_F$ is equal to that of Z under the future true model, i.e.,*

$$E_*^Y(\hat{Z}_F) = E_*^Z(Z).$$

3.2. Risk and criterion

As a natural risk function which measures a goodness of a future candidate model we may use

$$R(W|X) = -2E_*^Y E_*^Z[\log f^Z\{Z|\hat{\Theta}^Z(Y), \hat{\Sigma}^Z(Y)\}]$$
$$= E_*^Y E_*^Z[m\log|\hat{\Sigma}^Z| + mp\log 2\pi + \mathrm{tr}\{(Z-\hat{Z})'(Z-\hat{Z})\hat{\Sigma}^{Z^{-1}}\}].$$

The risk function is an extension of the *AIC*-type risk function introduced in Section 2.2. In fact, for the case when the future regression matrix is the same as the current one, it is equal to the risk function in Section 2.2. As an estimator of the risk function, Satoh (1997) proposed a Predictive *AIC*,

$$PAIC = m\log|\hat{\Sigma}| + mp\log 2\pi + \frac{(m+\phi)np}{n-k-p-1},$$

which is an unbiased estimator in the overspecified case, where

$$\phi = \mathrm{tr}\{W'W(X'X)^{-1}\}.$$

Similarly we can show that the criterion is equal to *CAIC* when $W_F = X_F$. For its asymptotic properties, see Satoh (1997).

4. Modification of *PAIC*

The *PAIC* proposed by Satoh (1997) is an unbiased estimator of the risk function in Section 3.2 in the overspecified case. In general, a class of candidate models includes both overspecified and underspecified models. Furthermore, it is unknown whether a candidate model is an overspecified model or not. In this section we discuss on a modification for a general candidate model. Our goal is to reduce the bias as much as possible. First we expand a risk function introduced in Section 3.2 and next we consider its estimator.

By considering the expectation with respect to the future true model, we can write the risk function of a future candidate model as

$$R(W|X) = E_*^Y[m\log|\hat{\Sigma}| + mp\log 2\pi + \mathrm{tr}\{m\Sigma_{F_*} + (W_F\Theta_{F_*} - W\hat{\Theta})'(W_F\Theta_{F_*} - W\hat{\Theta})\}\hat{\Sigma}^{-1}].$$

Since $\hat{\Theta}$ and $\hat{\Sigma}$ are independent under normality, we have

$$R(W|X) = E_*^Y(m\log|\hat{\Sigma}| + mp\log 2\pi) + r,$$

where

$$\begin{aligned} r &= \mathrm{tr}\{\Gamma E_*^Y(H^{-1})\}, \\ \Gamma &= E_*^Y\{mI_p + \Sigma_{F_*}^{-1/2}(W_F\Theta_{F_*} - W\hat{\Theta})'(W_F\Theta_{F_*} - W\hat{\Theta})\Sigma_{F_*}^{-1/2}\}, \\ H &= \Sigma_{F_*}^{-1/2}\hat{\Sigma}\Sigma_{F_*}^{-1/2}. \end{aligned} \tag{4.1}$$

The term $W_F\Theta_{F_*} - W\hat{\Theta}$ in Γ can be expressed as

$$(W_F\Theta_{F_*} - W\hat{\Theta})\Sigma_{F_*}^{-1/2} = \mathcal{V}_W - P_{W|X}\tilde{\mathcal{E}},$$

where $P_{B|A}$ denotes an $m \times n$ matrix defined by $P_{B|A} = B(A'A)^{-1}A'$ for any $n \times l$ matrix A and $m \times l$ matrix B,

$$\mathcal{V}_W = (W_F - P_{W|X}X_F)\Theta_{F_*}\Sigma_{F_*}^{-1/2}, \qquad \tilde{\mathcal{E}} = \mathcal{E}\Sigma_{F_*}^{-1/2}. \tag{4.2}$$

Note that elements of $\tilde{\mathcal{E}}$ are independently distributed as a standard normal distribution. This implies the following basic Lemma.

LEMMA 4.1. *Let* $\tilde{\mathscr{E}} = (e_1, \ldots e_p)$ *be the random matrix in* (4.2). *Then*

$$E_*^Y(\tilde{\mathscr{E}}' A \tilde{\mathscr{E}}) = I_p \operatorname{tr}(A),$$

where A *is any* $n \times n$ *matrix.*

From (4.1) and Lemma 4.1, we can reduce Γ as follows;

$$\begin{aligned}\Gamma &= mI_p + E_*^Y(\mathscr{V}_W - \Phi\tilde{\mathscr{E}})'(\mathscr{V}_W - \Phi\tilde{\mathscr{E}}) \\ &= mI_p + \Omega_W + E_*^Y(\tilde{\mathscr{E}}' P'_{W|X} P_{W|X} \tilde{\mathscr{E}}) \\ &= mI_p + \Omega_W + I_p \operatorname{tr}(P'_{W|X} P_{W|X}) \\ &= \Omega_W + (m+\phi) I_p,\end{aligned}$$

where

$$\Omega_W = \mathscr{V}_W' \mathscr{V}_W, \qquad \phi = \operatorname{tr}(P'_{W|X} P_{W|X}) = \operatorname{tr}\{W'W(X'X)^{-1}\}.$$

Let

$$\begin{gathered}U_1 = (I_n - P_{X_F})\tilde{\mathscr{E}}, \quad U_2 = (P_{X_F} - P_X)\tilde{\mathscr{E}}, \quad \mathscr{V} = (P_{X_F} - P_X) X_F \Theta_{F_*} \Sigma_{F_*}^{-1/2}, \\ V_1 = U_1'U_1, \quad V_2 = U_2'U_2, \quad \Omega = \mathscr{V}'\mathscr{V}.\end{gathered} \tag{4.3}$$

It is assumed that as in a usual set-up,

$$\Omega = n\Delta = O(n), \qquad \Omega_W = O(m) \tag{4.4}$$

Note that $V_1 = \tilde{\mathscr{E}}'(I_n - P_{X_F})\tilde{\mathscr{E}} \sim W(n - k_F, I_p)$ and hence the distribution of

$$\mathscr{Z} = \frac{1}{(n-k_F)^{1/2}}\{V_1 - (n-k_F)I_P\} \tag{4.5}$$

is asymptotically normal. Using (4.5) we can expand H^{-1} in terms of $\mathscr{Z}$ as in Lemma 4.2.

LEMMA 4.2. *Let* H *be the random matrix in* (4.1). *Then* H^{-1} *can be expanded for large* n *as*

$$H^{-1} = \Lambda - \frac{1}{n^{1/2}} C_1 \Lambda + \frac{1}{n}(C_1^2 \Lambda - C_2 \Lambda) + \frac{1}{n^{3/2}}(C_1 C_2 \Lambda + C_2 C_1 \Lambda - C_3 \Lambda) + O_p(n^{-2}),$$

where Λ *and* C_i $(i = 1, \ldots, 3)$ *are given by*

$$\Lambda = (I_p + \Delta)^{-1}, \qquad C_1 = \Lambda\mathscr{Z} + \frac{1}{n^{1/2}}\Lambda(U_2'\mathscr{V} + \mathscr{V}'U_2),$$

$$C_2 = \Lambda V_2 - k_F \Lambda, \qquad C_3 = -\frac{1}{2} k_F \Lambda \mathscr{Z}.$$

with the notation in (4.3), (4.4) *and* (4.5).

Next we consider an expansion for the expectation of H^{-1}. We use the following basic results; for a proof, see Muirhead (1982) or Siotani, Hayakawa and Fujikoshi (1985, p. 74).

LEMMA 4.3. *Let U_1 and V_1 be the random matrices in* (4.3). *Then*

$$E_*^Y\{\mathrm{tr}(A_1U_1A_2U_1)\} = \mathrm{tr}(A_1A_2'),$$

$$E_*^Y\{\mathrm{tr}(A_3U_1A_4U_1')\} = \mathrm{tr}(A_3)\,\mathrm{tr}(A_4),$$

$$E_*^Y\{\mathrm{tr}(A_4V_1A_5V_1)\} = (n-k_F)\{\mathrm{tr}(A_4)\,\mathrm{tr}(A_5) + (n-k_F+1)\,\mathrm{tr}(A_4A_5)\},$$

where A_1, A_2, A_3, A_4 and A_5 are any constant matrices of $p\times n$, $p\times n$, $n\times n$, $p\times p$ and $p\times p$, respectively.

Using Lemma 4.3 we can evaluate the expectation of each term in an expansion of H^{-1} given in Lemma 4.2.

LEMMA 4.4. *Let C_1, C_2 and C_3 be the random matrices in Lemma* 4.2. *Then*

$$E_*^Y(C_1) = \mathrm{O}_{p\times p},$$

$$E_*^Y(C_2) = -k\Lambda,$$

$$E_*^Y(C_3) = \mathrm{O}_{p\times p},$$

$$E_*^Y(C_1C_2) = \mathrm{O}_{p\times p},$$

$$E_*^Y\{\mathrm{tr}(C_1^2\Lambda\Gamma)\} = 2(p+1)\,\mathrm{tr}(\Lambda^2\Gamma) - \mathrm{tr}(\Lambda)\,\mathrm{tr}(\Lambda^2\Gamma) - \mathrm{tr}(\Lambda^3\Gamma).$$

From Lemmas 4.2 and 4.4, we can write r as

$$\begin{aligned} r &= \mathrm{tr}\{\Gamma E_*^Y(H^{-1})\} \\ &= \mathrm{tr}(\Lambda\Gamma) + \frac{1}{n}(2+2p+k)\,\mathrm{tr}(\Lambda^2\Gamma) - \frac{1}{n}\mathrm{tr}(\Lambda)\,\mathrm{tr}(\Lambda^2\Gamma) - \mathrm{tr}(\Lambda^3\Gamma) + O(mn^{-2}). \end{aligned}$$

It is known (see, Satoh (1997)) that the r under an overspecified model is given by

$$B_* = \frac{np(m+\phi)}{n-k-p-1}. \tag{4.6}$$

This result was obtained by evaluating $E_*^Y(H^{-1})$ exactly. It is a natural consequence that the r in the expanded form contains the terms of $O(n^{-1})$ in an expansion of B_* in (4.6). Here we shall see it directly. Under an overspecified model we have

$$\mathscr{V} = \mathrm{O}_{n\times p}, \qquad \Lambda = I_P, \qquad \mathrm{tr}(\Gamma) = p(m+\phi).$$

Thus the r in an expansion form can be reduced as

$$\begin{aligned} r &= \operatorname{tr}(\Gamma) + \frac{1}{n}(2+2p+k)\operatorname{tr}(\Gamma) - \frac{p}{n}\operatorname{tr}(\Gamma) - \operatorname{tr}(\Gamma) + O(mn^{-2}) \\ &= \left\{1 + \frac{1}{n}(k+p+1)\right\}\operatorname{tr}(\Gamma) + O(mn^{-2}) \\ &= B_* + O(mn^{-2}). \end{aligned}$$

Under a general candidate model, it holds that

$$\begin{aligned} r - B_* &= r - \left\{(m+\phi)p + \frac{1}{n}mp(k+p+1) + O(mn^{-2})\right\} \\ &= B_*^* + O(mn^{-2}), \end{aligned}$$

where

$$\begin{aligned} B_*^* &= m\operatorname{tr}(\Lambda\Lambda_W^{-1}) + \phi\operatorname{tr}(\Lambda) - (m+\phi)p \\ &\quad + \frac{m}{n}\{(2+2p+k)\operatorname{tr}(\Lambda^2\Lambda_W^{-1}) - \operatorname{tr}(\Lambda)\operatorname{tr}(\Lambda^2\Lambda_W^{-1}) - \operatorname{tr}(\Lambda^3\Lambda_W^{-1}) - (k+p+1)p\}, \end{aligned}$$

with

$$\Lambda_W = (I_p + m^{-1}\Omega_W)^{-1}.$$

Finally an expansion of the risk function is given in the following theorem.

THEOREM 4.1. *The risk function $R(W|X)$ can be expanded as*

$$R(W|X) = E_*^Y\{mp\log(2\pi) + m\log|\hat{\Sigma}|\} + B_* + B_*^* + O(mn^{-2}),$$

where

$$B_* = \frac{(m+\phi)np}{n-k-p-1},$$

$$\begin{aligned} B_*^* &= m\operatorname{tr}(\Lambda\Lambda_W^{-1}) + \phi\operatorname{tr}(\Lambda) - (m+\phi)p \\ &\quad + \frac{m}{n}\{(2+2p+k)\operatorname{tr}(\Lambda^2\Lambda_W^{-1}) - \operatorname{tr}(\Lambda)\operatorname{tr}(\Lambda^2\Lambda_W^{-1}) - \operatorname{tr}(\Lambda^3\Lambda_W^{-1}) - (k+p+1)p\}, \end{aligned}$$

with

$$\begin{aligned} \phi &= \operatorname{tr}\{W'W(X'X)^{-1}\}, \\ \Lambda_W &= (I_p + m^{-1}\Omega_W)^{-1}, \\ \Lambda &= (I_p + n^{-1}\Omega)^{-1}, \end{aligned}$$

$$\Omega_W = \mathcal{V}_W{}'\mathcal{V}_W,$$

$$\Omega = \mathcal{V}'\mathcal{V},$$

$$\mathcal{V}_W = \{W_F - P_{W|X}X_F\}\Theta_{F_*}\Sigma_{F_*}^{-1/2},$$

$$\mathcal{V} = (I_n - P_X)X_F\Theta_{F_*}\Sigma_{F_*}^{-1/2}.$$

Further, it holds that $B_{\bar{*}} = 0$ *in the overspecified case, and that the error term can be omitted in that case.*

Theorem 4.1 includes the result obtained by Fujikoshi and Satoh (1997) as a special case, which is given by the next corollary.

COROLLARY 4.1. *If the future regression matrix is equal to a current one, then the risk function* $R(X|X)$ *can be expanded as*

$$R(X|X) = E_*^Y\{np\log(2\pi) + n\log|\hat{\Sigma}|\} + B_* + B_{\bar{*}} + O(n^{-1}),$$

where

$$B_* = \frac{(n+k)np}{n-k-p-1},$$

$$B_{\bar{*}} = k\mathrm{tr}(\Lambda) - kp + [(2+2p+k)\,\mathrm{tr}(\Lambda) - \{\mathrm{tr}(\Lambda)\}^2 - \mathrm{tr}(\Lambda^2) - (k+p+1)p]$$

$$= 2k\mathrm{tr}(\Lambda - I_p) - \{\mathrm{tr}(\Lambda - I_p)\}^2 - \mathrm{tr}\{(\Lambda - I_p)^2\},$$

with

$$\Lambda = (I_p + n^{-1}\Omega)^{-1},$$

$$\Omega = \mathcal{V}'\mathcal{V},$$

$$\mathcal{V} = (I_n - P_X)X_F\Theta_{F_*}\Sigma_{F_*}^{-1/2}.$$

Further, it holds that $B_{\bar{*}} = 0$ *in the overspecified case, and that the error term can be omitted in that case.*

Here we note that $B_{\bar{*}}$ contains the term of $O(mn^{-1})$ in Theorem 4.1. The term will be useful when m is not small in comparison with n. On the other hand, the term can be omitted if m is quite smaller than n or $m = O(1)$ as n is large. Even for the case, we still have a valid result with the order $O(n^{-1})$. In this case the following theorem suggests us to select the full model asymptotically.

THEOREM 4.2. *Under the overspecified case the risk function in Theorem* 4.1 *tends to the same value*

$$R(W|X) \to -2E_*^Z\{\log f_F^Z(Z|\Theta_{F_*}, \Sigma_{F_*})\},$$

as $n \to \infty$, *where* f_F^Z *is the probability density function of the future full model. Further its limiting value is the minimum value of the risk function.*

From Theorem 4.2 we can see that the risk function is a function of the sample size. Therefore, the population best model depends on the sample size. Furthermore, it is also a function of the future regression matrix.

EXAMPLE 4.1. *Consider a univariate polynomial regression model given by*

$$y_i \sim N\left(\sum_{j=0}^{k-1} \theta_j s_i^j, \sigma^2\right), \qquad i = 1, \ldots, n.$$

In the model we assume that the current and the future design matrices are expressed as

$$X^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})', \qquad W^{(k)} = (w^{(k)})',$$

where

$$x_i^{(k)} = (s_i^0, \ldots, s_i^{k-1})', \qquad i = 1, \ldots, n,$$

$$w^{(k)} = (t^0, \ldots, t^{k-1})'.$$

Then, the risk function can be expanded with respect to t *as*

$$R(W^{(k)}|X^{(k)}) = \begin{cases} at^{2(k-1)} + O(t^{2(k-2)}), & \text{if } k \geq 2, \\ \text{Constant on } t, & \text{if } k = 1, \end{cases}$$

where $a > 0$. *So, if* $|t|$ *is large, the risk function with order* $O(t^{2(k-1)})$ *come to select a polynomial model with less parameters.*

This result can be seen as follows. The risk function is of W_F and W_F is defined by t. The parameters which contain t are only $\phi = \mathrm{tr}\{W'W(X'X)^{-1}\}$ and Λ_W^{-1}. Since $X'X$ is a positive definite matrix, it holds that $\phi = at^{2(k-1)} + O(t^{2(k-2)})$ where $a > 0$. On the other hand, the fact that $\mathcal{V}_W = O(t^{k-1})$ gives $\Lambda_W^{-1} = O(t^{-2(k-1)})$. □

Next, we consider an estimator of the risk function. The unknown parameters which need to be estimated are only Λ and Λ_W^{-1}. An estimator of Λ_W^{-1} is proposed in the following theorem.

THEOREM 4.3. *Let* Λ_W^{-1} *be the parameter matrix in theorem* 4.1, *and* $\hat{\Lambda}_W^{-1}$ *the estimated matrix defined by*

$$\hat{\Lambda}_W^{-1} = \frac{1}{m}\left\{\frac{1}{c}D + (m - \psi)I_p\right\},$$

where

$$c = \frac{n}{n - k_F - p - 1},$$

$$D = (\hat{Z}_F - \hat{Z})'(\hat{Z}_F - \hat{Z})\hat{\Sigma}_F^{-1},$$

$$\psi = \mathrm{tr}(P_{W_F|X_F} - P_{W|X})'(P_{W_F|X_F} - P_{W|X}).$$

Then it holds that

$$E_*^Y(\hat{\Lambda}_W^{-1}) = \Sigma_{F_*}^{1/2}\Lambda_W^{-1}\Sigma_{F_*}^{-1/2}.$$

We note that the estimator $\hat{\Lambda}_W^{-1}$ is an exact unbiased estimator for Λ_W^{-1} in a general candidate model. Similarly, we also use the estimator for Λ by replacing W_F and W with X_F and X, respectively, i.e., it is given by

$$\hat{\Lambda} = n\left[\frac{1}{c}D + \{n - (k_F - k)\}I_p\right]^{-1},$$

where

$$D = (\hat{Y}_F - \hat{Y})'(\hat{Y}_F - \hat{Y})\hat{\Sigma}_F^{-1}.$$

Therefore, for the case when $W_F = X_F$ it holds that $\hat{\Lambda}_W = \hat{\Lambda}$, or $\hat{\Lambda}^{-1}\hat{\Lambda}_W = I_p$ exactly. Thus we propose a new modification of *PAIC* defined by

$$MPAIC(W|X) = mp\log(2\pi) + m\log|\hat{\Sigma}| + B_* + \hat{B}_*,$$

where $\hat{B}_*$ is defined from B_* by substituting $\hat{\Lambda}_W^{-1}$ and $\hat{\Lambda}$ for Λ_W^{-1} and Λ, respectively. It may be noted that the estimator $\hat{\Lambda}_W$ is not an extension of $\hat{\Lambda}$ which was proposed in *MAIC* (see Fujikoshi and Satoh (1997)).

5. Simulation study

We attempt to give an impression of the relative performances of *PAIC* and *MPAIC*. The accuracy of our estimator for the risk function is investigated through Monte Carlo experiments. Bedric and Tsai (1994) report a simulation study indicating that *CAIC* has much smaller biases than those of *AIC* in the overspecified case. As a result it provides better model choices than *AIC* in small sample case. Fujikoshi and Satoh (1997) shows that in terms of bias reduction *MAIC* outperforms *AIC* and *CAIC* in the underspecified case, and performs similarly to *CAIC* in the overspecified case. Satoh (1997) examines a performance of *PAIC* in some extrapolation case. It is reported that *PAIC* has small biases for small sample case and the overspecified cases.

In our experiments $p = 2$ and $n = 20$ as a small sample size or 200 as a large sample size. First, we consider the case when the full model with $k_F = 6$ is the minimal full model, or all the candidate models except the full model are underspecified models. The elements of the matrix X_F and the matrix Θ_{F_*} were defined by realization of independent normal random variables with mean zero and variance 1 and 2, respectively. The case when $k_F = 4$ is also considered and the true matrix Θ_{F_*} is defined by using the submatrix of that for the case $k_F = 6$, which were given by

$$\Theta_{F_*} = \begin{pmatrix} 0.563 & -1.693 \\ -0.687 & -2.698 \\ -1.188 & -0.425 \\ 0.983 & 1.578 \\ 0.676 & 1.554 \\ -1.335 & -0.307 \end{pmatrix} \quad \textit{if} \quad k_F = 6, \qquad = \begin{pmatrix} 0.563 & -1.693 \\ -0.687 & -2.698 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \textit{if} \quad k_F = 4.$$

For the case when $k_F = 4$, the minimal full model is $\{1,2\}$. There exists some overspecified models, which are $\{1,2\}$, $\{1,2,3\}$, $\{1,2,4\}$ and $\{1,2,3,4\}$. The elements of the true covariance matrix Σ_{F_*} were constructed by using a decomposition $\Sigma_{F_*} = \Xi'\Xi$, where Ξ is a $p \times p$ matrix, and by defining elements of Ξ as realizations of independent standard normal random variables. The true covariance matrix was given by

$$\Sigma_{F_*} = \begin{pmatrix} 1.120 & 0.731 \\ 0.731 & 1.256 \end{pmatrix}.$$

Five thousand samples of size $n = 20$, 200 were generated from the true structure. Note that these samples have been generated from the fixed same true model and the same future regression matrix W_F throughout the study. It is also noted that sample sets which are used for evaluating the risk function and for estimating the risk are different each other. In both of cases when $k_F = 4$ and $k_F = 6$, candidate models are considered for all subsets of full explanatory variables. For each candidate model, a risk function, biases and standard deviations of the criteria are presented in Tables 1, 2, 3 and 4. Overspecified models which contain the true model are marked by $*$ and the population best model by $\dagger$.

Tables 1 and 3 present for a small sample case. Biases of *MPAIC* are almost equal with those of *PAIC* for overspecified models. On the other hand, for underspecified models, we can see that biases of *MPAIC* are unformly smaller than those of *PAIC*. Since the standard deviations of *MAPIC* are not so large relative to those of *PAIC*, the mean square errors which is defined by $(Bias)^2 + (S.D.)^2$ are small. It is noted that the population best model for

TABLE 1. Risk, Bias, S.D. and M.S.E. in 5000 repetitions: $n = 20$, $k_F = 4$, $p = 2$, $m = 20$, the true model $= \{1, 2\}$

		Bias		S.D.		M.S.E.	
Model	Risk	PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1}	155.11	4.20	1.33	7.60	7.58	75.40	59.22
{2}	150.18	3.97	1.32	7.71	7.68	75.20	60.72
{3}	172.41	6.57	2.02	6.88	6.69	90.49	48.83
{4}	172.60	6.61	2.01	6.86	6.66	90.75	48.39
{1,2} * †	121.76	0.08	0.16	9.66	9.67	93.32	93.53
{1,3}	156.36	6.92	2.50	7.81	7.79	108.88	66.93
{2,3}	152.01	6.58	2.45	7.92	7.88	106.02	68.09
{1,4}	157.14	6.84	2.38	7.80	7.78	107.62	66.19
{2,4}	152.31	6.49	2.35	7.94	7.89	105.16	67.77
{3,4}	173.41	10.62	3.76	7.04	6.78	162.34	60.10
{1,2,3} *	126.20	0.25	0.32	10.00	10.01	100.06	100.30
{1,2,4} *	126.31	0.06	0.14	9.91	9.94	98.21	98.82
{1,3,4}	158.87	9.98	3.94	8.03	7.99	164.08	79.36
{2,3,4}	154.45	9.49	3.85	8.19	8.12	157.13	80.75
{1,2,3,4} *	131.72	0.18	0.18	10.32	10.32	106.53	106.53

TABLE 2. Risk, Bias, S.D. and M.S.E. in 5000 repetitions: $n = 200$, $k_F = 4$, $p = 2$, $m = 20$, the true model $= \{1, 2\}$

		Bias		S.D.		M.S.E.	
Model	Risk	PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1}	148.98	1.78	0.10	2.26	2.29	8.27	5.25
{2}	145.82	6.61	0.14	2.28	2.27	48.89	5.17
{3}	167.98	5.79	0.10	2.00	1.95	37.52	3.81
{4}	168.34	5.43	0.07	2.00	1.96	33.48	3.84
{1,2} * †	111.41	0.09	0.09	2.86	2.86	8.18	8.18
{1,3}	148.26	2.61	0.12	2.27	2.28	11.96	5.21
{2,3}	145.91	6.82	0.15	2.29	2.28	51.75	5.22
{1,4}	146.05	6.53	0.14	2.29	2.28	47.88	5.21
{2,4}	148.96	2.05	0.10	2.27	2.28	9.35	5.20
{3,4}	167.90	6.09	0.10	2.01	1.93	41.12	3.73
{1,2,3} *	111.62	0.09	0.09	2.87	2.87	8.24	8.24
{1,2,4} *	111.58	0.09	0.09	2.87	2.87	8.24	8.24
{1,3,4}	148.30	2.83	0.13	2.28	2.27	13.20	5.16
{2,3,4}	146.13	6.75	0.16	2.29	2.28	50.80	5.22
{1,2,3,4} *	111.79	0.09	0.09	2.87	2.87	8.24	8.24

TABLE 3. Risk, Bias, S.D. and M.S.E. in 5000 repetitions: $n = 20$, $k_F = 6$, $p = 2$, $m = 20$, the true model $= \{1,2,3,4,5,6\}$

TABLE 3-1.

Model	Risk	Bias		S.D.		M.S.E.	
		PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1}	150.19	5.97	2.28	7.56	7.35	92.79	59.22
{2}	154.95	5.97	2.10	7.35	7.14	89.66	55.38
{3}	167.06	7.46	2.46	6.61	6.42	99.34	47.26
{4}	165.23	7.35	2.42	6.68	6.48	98.64	47.84
{5}	170.29	7.71	2.50	6.44	6.24	100.91	45.18
{6}	161.51	6.13	2.11	7.15	6.99	88.69	53.31
{1,2}	140.91	6.43	2.92	8.67	8.43	116.51	79.59
{1,3}	145.23	8.38	3.72	8.25	7.98	138.28	77.51
{2,3}	154.18	9.63	3.89	7.65	7.37	151.25	69.44
{1,4}	145.65	8.45	3.71	8.30	8.02	140.29	78.08
{2,4}	154.60	9.48	3.81	7.65	7.38	148.39	68.98
{3,4}	164.27	11.64	4.40	6.91	6.67	183.23	63.84
{1,5}	148.64	9.31	4.01	7.96	7.69	150.03	75.21
{2,5}	155.91	9.89	3.95	7.54	7.26	154.66	68.31
{3,5}	167.67	11.94	4.45	6.74	6.49	187.99	61.92
{4,5}	164.45	11.78	4.49	6.86	6.61	185.82	63.85
{1,6}	144.70	7.70	3.43	8.30	8.06	128.18	76.72
{2,6}	146.50	7.85	3.38	8.18	7.94	128.53	74.46
{3,6}	153.88	7.24	2.79	8.03	7.94	116.89	70.82
{4,6}	158.09	9.42	3.67	7.57	7.35	146.04	67.49
{5,6}	161.58	9.65	3.71	7.42	7.22	148.17	65.89
{1,2,3}	138.84	7.66	3.94	9.28	9.03	144.79	97.06
{1,2,4}	140.66	8.91	4.55	9.16	8.86	163.29	99.20
{1,3,4}	143.40	10.57	5.25	8.91	8.58	191.11	101.17
{2,3,4}	154.51	13.71	6.22	7.93	7.60	250.84	96.44
{1,2,5}	140.72	8.03	4.01	9.20	8.96	149.12	96.36
{1,3,5}	146.38	12.31	6.02	8.48	8.16	223.44	102.82
{2,3,5}	155.35	14.04	6.33	7.84	7.50	258.58	96.31
{1,4,5}	147.08	12.73	6.23	8.49	8.14	234.13	105.07
{2,4,5}	154.04	12.74	5.79	8.02	7.74	226.62	93.43
{3,4,5}	160.99	16.16	7.07	7.30	6.99	314.43	98.84

$k_F = 4$ is $\{1,2\}$, which is consistent with the minimal full model, but the population best model for $k_F = 6$ is $\{1,2,3,6\}$, which is not the minimal full model $\{1,2,3,4,5,6\}$. It is an example that the minimal full model is not always the population best model.

Tables 2 and 4 present for a large sample case. Biases of *MPAIC* are almost the same as those of *PAIC* for overspecified models. For underspecified

TABLE 3-2.

Model	Risk	Bias		S.D.		M.S.E.	
		PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1,2,6}	137.36	7.32	3.92	9.30	9.07	140.07	97.63
{1,3,6}	137.69	6.45	3.30	9.46	9.30	131.09	97.38
{2,3,6}	141.71	7.91	3.84	9.10	8.91	145.37	94.13
{1,4,6}	141.29	9.58	4.88	9.05	8.74	173.67	100.20
{2,4,6}	147.75	11.14	5.26	8.49	8.20	196.17	94.90
{3,4,6}	153.69	10.14	4.34	8.36	8.25	172.70	86.89
{1,5,6}	143.54	10.47	5.24	8.84	8.54	187.76	100.38
{2,5,6}	146.85	11.18	5.35	8.55	8.25	198.09	96.68
{3,5,6}	156.21	10.55	4.47	8.23	8.13	179.03	86.07
{4,5,6}	158.78	13.60	5.96	7.77	7.52	245.33	92.07
{1,2,3,4}	140.35	9.83	5.56	9.76	9.46	191.88	120.40
{1,2,3,5}	141.37	9.69	5.34	9.66	9.41	187.21	117.06
{1,2,4,5}	140.92	8.69	4.82	9.60	9.47	167.67	112.91
{1,3,4,5}	144.82	14.31	7.83	9.21	8.83	289.60	139.27
{2,3,4,5}	152.39	14.92	7.51	8.53	8.30	295.36	125.29
{1,2,3,6} †	133.98	3.04	1.95	10.46	10.35	118.65	110.92
{1,2,4,6}	139.42	9.56	5.54	9.77	9.48	186.84	120.56
{1,3,4,6}	138.46	6.85	3.84	10.06	9.90	148.12	112.75
{2,3,4,6}	144.69	10.80	5.65	9.39	9.16	204.81	115.82
{1,2,5,6}	137.61	7.69	4.57	10.07	9.81	160.54	117.12
{1,3,5,6}	140.73	8.82	4.86	9.80	9.60	173.83	115.77
{2,3,5,6}	144.76	11.04	5.79	9.35	9.12	209.30	116.69
{1,4,5,6}	143.58	13.55	7.54	9.32	8.96	270.46	137.13
{2,4,5,6}	146.81	14.42	7.70	9.03	8.68	289.47	134.63
{3,4,5,6}	154.29	13.55	6.49	8.67	8.54	258.77	115.05
{1,2,3,4,5}	141.52	7.98	5.00	10.34	10.19	170.59	128.83
{1,2,3,4,6}	139.72	3.45	2.21	10.88	10.75	130.27	120.44
{1,2,3,5,6}	139.30	2.55	1.69	10.96	10.86	126.62	120.79
{1,2,4,5,6}	141.53	7.90	4.95	10.47	10.28	172.03	130.18
{1,3,4,5,6}	142.89	9.05	5.44	10.38	10.19	189.64	133.42
{2,3,4,5,6}	146.93	12.35	7.00	9.77	9.62	247.97	141.54
{1,2,3,4,5,6} *	146.56	0.02	0.02	11.41	11.41	130.18	130.18

models, although *PAIC* still has large biases, *MPAIC* has only small biases. For both overspecified and underspecified models the standard deviations are not so different and smaller than those of the small sample models. As a consequence, the mean square errors of *MAPC* are quite smaller than those of *PAIC*. On the risk function we can see that those of overspecified models are almost the same value and it agrees with the result of Theorem 4.2.

Table 5 gives frequencies of the sample best model selected by the *PAIC*

TABLE 4. Risk, Bias, S.D. and M.S.E. in 5000 repetitions: $n = 200$, $k_F = 6$, $p = 2$, $m = 20$, the true model = $\{1,2,3,4,5,6\}$

TABLE 4-1.

Model	Risk	Bias		S.D.		M.S.E.	
		PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1}	185.28	7.74	0.05	1.47	1.54	62.06	2.37
{2}	181.38	8.98	0.07	1.48	1.56	82.83	2.43
{3}	182.04	12.83	0.10	1.51	1.51	166.88	2.29
{4}	187.42	10.96	0.08	1.42	1.45	122.13	2.10
{5}	187.66	12.06	0.07	1.37	1.42	147.32	2.02
{6}	184.87	7.99	0.06	1.54	1.59	66.21	2.53
{1,2}	168.60	6.42	0.07	1.76	1.88	44.31	3.53
{1,3}	172.26	11.27	0.11	1.69	1.72	129.86	2.97
{2,3}	172.91	10.52	0.11	1.63	1.66	113.32	2.76
{1,4}	182.09	7.37	0.08	1.54	1.59	56.68	2.53
{2,4}	172.14	12.79	0.11	1.61	1.64	166.17	2.70
{3,4}	177.27	13.33	0.14	1.62	1.59	180.31	2.54
{1,5}	181.13	9.04	0.07	1.51	1.57	84.00	2.46
{2,5}	179.19	6.41	0.08	1.56	1.66	43.52	2.76
{3,5}	178.72	13.05	0.11	1.57	1.58	172.76	2.50
{4,5}	182.99	12.34	0.10	1.47	1.50	154.43	2.26
{1,6}	179.85	−0.69	0.04	1.80	1.85	3.71	3.42
{2,6}	175.66	5.56	0.08	1.68	1.76	33.73	3.10
{3,6}	175.96	9.36	0.12	1.76	1.75	90.70	3.07
{4,6}	181.45	7.18	0.09	1.64	1.67	54.24	2.79
{5,6}	181.88	8.98	0.08	1.57	1.62	83.10	2.63
{1,2,3}	153.98	10.14	0.16	2.00	1.99	106.81	3.98
{1,2,4}	156.28	10.86	0.14	1.95	1.95	121.74	3.82
{1,3,4}	167.29	12.04	0.16	1.78	1.72	148.13	2.98
{2,3,4}	161.78	15.36	0.19	1.79	1.72	239.13	2.99
{1,2,5}	164.75	0.08	0.08	2.00	2.16	4.00	4.67
{1,3,5}	168.52	11.64	0.13	1.74	1.78	138.51	3.18
{2,3,5}	170.33	7.22	0.13	1.74	1.82	55.15	3.32
{1,4,5}	176.61	8.73	0.10	1.60	1.65	78.77	2.73
{2,4,5}	164.69	12.36	0.14	1.78	1.82	155.93	3.33
{3,4,5}	171.32	14.36	0.16	1.73	1.70	209.20	2.91

and *MPAIC*. The limiting distribution of selected order in an autoregression model was obtained by Shibata (1976). Nishii (1984) gave some similar asymptotic results in normal linear regression. They discuss on a probability of selecting the true model, but that is not so important in small sample case. In the study, *PAIC* has more bias under underspecified model than that of *MPAIC*. However, frequencies which the population best model is selected are not so different. Similar numerical result was reported by Satoh (1997).

TABLE 4-2.

Model	Risk	Bias		S.D.		M.S.E.	
		PAIC	MPAIC	PAIC	MPAIC	PAIC	MPAIC
{1,2,6}	162.26	−0.88	0.05	2.06	2.15	5.01	4.62
{1,3,6}	164.13	0.49	0.07	2.19	2.37	5.03	5.62
{2,3,6}	166.80	6.86	0.13	1.88	1.92	50.59	3.70
{1,4,6}	175.84	−0.80	0.06	1.88	1.91	4.17	3.65
{2,4,6}	165.40	8.81	0.13	1.85	1.88	81.03	3.55
{3,4,6}	170.64	8.95	0.16	1.91	1.85	83.75	3.44
{1,5,6}	175.41	0.74	0.06	1.84	1.88	3.93	3.53
{2,5,6}	173.97	2.45	0.09	1.75	1.83	9.06	3.35
{3,5,6}	172.92	9.36	0.13	1.82	1.83	90.92	3.36
{4,5,6}	177.16	8.45	0.11	1.69	1.71	74.25	2.93
{1,2,3,4}	141.22	14.92	0.24	2.17	2.07	227.31	4.34
{1,2,3,5}	150.86	3.37	0.15	2.19	2.22	16.15	4.95
{1,2,4,5}	139.10	5.57	0.15	2.39	2.43	36.73	5.92
{1,3,4,5}	161.25	12.80	0.19	1.88	1.81	167.37	3.31
{2,3,4,5}	150.44	15.79	0.23	2.04	1.94	253.48	3.81
{1,2,3,6}	143.14	0.87	0.10	2.42	2.47	6.61	6.11
{1,2,4,6}	149.53	3.59	0.12	2.23	2.23	17.86	4.98
{1,3,4,6}	156.08	2.37	0.15	2.25	2.21	10.67	4.90
{2,3,4,6}	153.57	11.22	0.22	2.10	2.01	130.29	4.08
{1,2,5,6}	157.73	−7.29	0.02	2.32	2.45	58.52	6.00
{1,3,5,6}	160.56	0.84	0.08	2.22	2.41	5.63	5.81
{2,3,5,6}	165.31	2.16	0.13	1.99	2.05	8.62	4.21
{1,4,5,6}	170.32	0.62	0.09	1.92	1.95	4.07	3.81
{2,4,5,6}	158.68	6.86	0.15	2.02	2.06	51.14	4.26
{3,4,5,6}	164.93	9.33	0.18	2.03	1.97	91.16	3.91
{1,2,3,4,5}	123.94	9.39	0.27	2.58	2.53	94.82	6.47
{1,2,3,4,6}	129.18	5.85	0.21	2.55	2.52	40.72	6.39
{1,2,3,5,6}	139.43	−5.82	0.03	2.58	2.65	40.52	7.02
{1,2,4,5,6}	132.09	−1.79	0.06	2.64	2.67	10.17	7.13
{1,3,4,5,6}	150.25	2.83	0.15	2.29	2.28	13.25	5.22
{2,3,4,5,6}	142.44	8.40	0.23	2.30	2.29	75.85	5.29
{1,2,3,4,5,6} * †	112.17	0.11	0.11	2.90	2.90	8.42	8.42

6. Proofs

PROOF OF PROPERTY 3.1.

$$
\begin{aligned}
E_*^Y(\hat{Z}_F) &= E_*^Y(W_F\hat{\Theta}_F) \\
&= W_F(X_F'X_F)^{-1}X_F'E_*^Y(Y) \\
&= W_F(X_F'X_F)^{-1}X_F'X_F\Theta_{F_*} \\
&= W_F\Theta_{F_*}.
\end{aligned}
$$

□

TABLE 5. Frequencies selected by the criterion in 5000 repetitions: $p = 2$, $m = 20$

k_F	n	Model	PAIC	MPAIC.
4	20	$\{1,2\}$ * †	5000	5000
	200	$\{1,2\}$ * †	5000	5000
6	20	$\{1,2,3,6\}$ †	4999	4999
		$\{1,2,3,5\}$		1
		$\{1,2,3,5,6\}$	1	
	200	$\{1,2,3,4,5,6\}$ * †	5000	5000

PROOF OF LEMMA 4.1. Since $\tilde{\mathscr{E}} \sim N_{n\times p}(\mathscr{Z}_{n\times p}, I_p \otimes I_n)$,

$$\begin{aligned} E_*^Y(\tilde{\mathscr{E}}' A \tilde{\mathscr{E}})_{ij} &= E_*^Y(e_i' A e_j) \\ &= \operatorname{tr}\{A E_*^Y(e_j e_i')\} \\ &= \operatorname{tr}(A\delta_{ij} I_n) \\ &= \delta_{ij} \operatorname{tr}(A), \end{aligned}$$

where δ_{ij} is the Kronecker's delta. This completes the proof. □

PROOF OF LEMMA 4.2. Recall that $n\hat{\Sigma} = (Y - X\hat{\Theta})'(Y - X\hat{\Theta})$ and the term $Y - X\hat{\Theta}$ is decomposed to three parts, i.e.,

$$\begin{aligned} Y - X\hat{\Theta} &= Y - X(X'X)^{-1}XY \\ &= (I_n - P_X)(\mathscr{E} + X_F\Theta_{F_*}) \\ &= (I_n - P_{X_F})\mathscr{E} + (P_{X_F} - P_X)(\mathscr{E} + X_F\Theta_{F_*}) \\ &= (U_1 + U_2 + \mathscr{V})\Sigma_{F_*}^{1/2}. \end{aligned}$$

Since $U_1'U_2 = \mathrm{O}_{p\times p}$,

$$\begin{aligned} H &= (U_1 + U_2 + \mathscr{V})'(U_1 + U_2 + \mathscr{V}) \\ &= U_1'U_1 + (U_2 + \mathscr{V})'(U_2 + \mathscr{V}) \\ &= V_1 + V_2 + \Omega + U_2'\mathscr{V} + \mathscr{V}'U_2. \end{aligned}$$

Thus,

$$\begin{aligned} nH &= (n - k_F)^{1/2}\mathscr{Z} + (n - k_F)I_p + V_2 + n\Delta + U_2'\mathscr{V} + \mathscr{V}'U_2 \\ &= n^{1/2}\left\{1 - \frac{k_F}{2n} + O(n^{-2})\right\}\mathscr{Z} + n(I_p + \Delta) - k_F + V_2 + U_2'\mathscr{V} + \mathscr{V}'U_2 \end{aligned}$$

$$= \{n(I_p+\Lambda)\} + (n^{1/2}\mathscr{Z}+U_2'\mathscr{V}+\mathscr{V}'U_2) + (V_2-k_FI_p) - \frac{k_F}{2n^{1/2}}\mathscr{Z} + O(n^{-3/2})$$

$$= n\Lambda^{-1}\left\{I_p + \frac{1}{n^{1/2}}C_1 + \frac{1}{n}C_2 + \frac{1}{n^{3/2}}C_3 + O_p(n^{-5/2})\right\}.$$

By considering a perturbation expansion of H^{-1} when n is large, we obtain

$$\begin{aligned} H^{-1} &= \Lambda - \left(\frac{1}{n^{1/2}}C_1 + \frac{1}{n}C_2 + \frac{1}{n^{3/2}}C_3\right)\Lambda \\ &\quad + \left(\frac{1}{n}C_1^2 + \frac{1}{n^{3/2}}C_1C_2 + \frac{1}{n^{3/2}}C_2C_1\right)\Lambda + O_p(n^{-2}) \\ &= \Lambda - \frac{1}{n^{1/2}}C_1\Lambda + \frac{1}{n}(C_1^2\Lambda - C_2\Lambda) \\ &\quad + \frac{1}{n^{3/2}}(C_1C_2\Lambda + C_2C_1\Lambda - C_3\Lambda) + O_p(n^{-2}), \end{aligned}$$

which completes the proof. □

PROOF OF LEMMA 4.4. Note that

$$E_*^Y(V_2) = (k_F - k)I_p, \quad E_*^Y(\mathscr{Z}) = \mathrm{O}_{p\times p}, \quad E_*^Y(U_1) = \mathrm{O}_{n\times p} \quad E_*^Y(U_2) = \mathrm{O}_{n\times p}.$$

Therefore

$$E_*^Y(C_1) = \mathrm{O}_{p\times p}, \qquad E_*^Y(C_2) = -k\Lambda, \qquad E_*^Y(C_3) = \mathrm{O}_{p\times p}.$$

Since $\mathscr{Z}$ and V_2 are independent, $E_*^Y(C_1C_2) = \mathrm{O}_{p\times p}$.

From Lemma (4.3), we have

$$\begin{aligned} \mathrm{tr}(C_1^2\Lambda\Gamma) &= \mathrm{tr}(\Lambda\mathscr{Z}\Lambda\mathscr{Z}\cdot\Lambda\Gamma) \\ &\quad + \frac{1}{n}\mathrm{tr}(\Lambda U_2'\mathscr{V}\Lambda U_2'\mathscr{V}\cdot\Lambda\Gamma) + \frac{1}{n}\mathrm{tr}(\Lambda\mathscr{V}'U_2\Lambda\mathscr{V}'U_2\cdot\Lambda\Gamma) \\ &\quad + \frac{2}{n}\mathrm{tr}(\Lambda U_2'\mathscr{V}\Lambda\mathscr{V}'U_2\cdot\Lambda\Gamma). \end{aligned}$$

Considering the expectation of each term in the above expression, we obtain

$$\begin{aligned} E_*^Y\mathrm{tr}(C_1^2\Lambda\Gamma) &= E_*^Y\mathrm{tr}(\Lambda\mathscr{Z}\Lambda\mathscr{Z}\cdot\Lambda\Gamma) \\ &\quad + \frac{1}{n}E_*^Y\mathrm{tr}(\Lambda U_2'\mathscr{V}\Lambda U_2'\mathscr{V}\cdot\Lambda\Gamma) + \frac{1}{n}E_*^Y\mathrm{tr}(\Lambda\mathscr{V}'U_2\Lambda\mathscr{V}'U_2\cdot\Lambda\Gamma) \\ &\quad + \frac{2}{n}E_*^Y\mathrm{tr}(\Lambda U_2'\mathscr{V}\Lambda\mathscr{V}'U_2\cdot\Lambda\Gamma) \end{aligned}$$

$$= \{\mathrm{tr}(\Lambda^2\Gamma)\,\mathrm{tr}(\Lambda) + (n - k_F + 1)\,\mathrm{tr}(\Lambda^3\Gamma)\} + (n - k_F)\,\mathrm{tr}(\Lambda^3\Gamma)$$
$$- 2(n - k_F)\,\mathrm{tr}(\Lambda^3\Gamma)$$
$$+ \mathrm{tr}\{\Lambda\Gamma(\Lambda - \Lambda^2)\} + \mathrm{tr}\{\Lambda^2\Gamma(I_p - \Lambda)\}$$
$$+ 2\mathrm{tr}(\Lambda^2\Gamma)\,\mathrm{tr}(I_p - \Lambda)$$
$$= 2(p+1)\,\mathrm{tr}(\Lambda^2\Gamma) - \mathrm{tr}(\Lambda\cdot\Lambda^2\Gamma) - \mathrm{tr}(\Lambda^3\Gamma). \qquad \square$$

PROOF OF THEOREM 4.2. From the assumption that $X'X = O(n)$ and $W'W = O(m)$, we have $\phi = O(mn^{-1})$ and $B_* = mp + O(mn^{-1})$. Therefore,

$$B_{\bar{*}} = m\,\mathrm{tr}(\Lambda\Lambda_W^{-1}) - mp + O(mn^{-1}),$$

and hence,

$$B_* + B_{\bar{*}} = m\,\mathrm{tr}(\Lambda\Lambda_W^{-1}) + O(mn^{-1}).$$

On the other hand,

$$m\log|\hat{\Sigma}| = m\log|H| + m\log|\Sigma_{F_*}|.$$

Recall that $H = \Lambda^{-1}\{I_p + n^{-1/2}C_1 + O_p(n^{-1})\}$ and $E_*^Y(C_1) = \mathrm{O}_{p\times p}$. Those imply

$$E(m\log|\hat{\Sigma}|) = m\log|\Sigma_{F_*}| + m\log|\Lambda^{-1}| + O(mn^{-1}).$$

Thus, the risk function can be written as

$$R(W|X) = m\log|\Sigma_{F_*}| + mp\log(2\pi) + m\log|\Lambda^{-1}| + m\,\mathrm{tr}(\Lambda\Lambda_W^{-1}) + O(mn^{-1}).$$

From the fact that Ω and Ω_W are non-negative definite matrix, $m\log|\Lambda^{-1}| + \mathrm{tr}(\Lambda\Lambda_W^{-1})$ attains the minimum value when $\Omega = \Omega_W = O_{p\times p}$, i.e., a candidate model is an overspecified model. Conversely, if a candidate model is an overspecified model, then we have

$$R(W|X) = m\log|\Sigma_{F_*}| + mp\log(2\pi) + mp + O(mn^{-1}). \tag{6.1}$$

Thus we obtain the minimum value which the risk function attains asymptotically. We note that -2 times expected probability density function of the future true model is given by

$$-2E_*^Z\{\log f^Z(Z|\Theta_{F_*}, \Sigma_{F_*})\} = m\log|\Sigma_{F_*}| + mp\log(2\pi)$$
$$+ E_*^Z[\mathrm{tr}\{(Z - W_F\Theta_{F_*})'(Z - W_F\Theta_{F_*})\Sigma_{F_*}^{-1}\}]. \tag{6.2}$$

Since the fomula (6.2) is equal to (6.1) when n is large. This completes the proof. $\square$

PROOF OF THEOREM 4.3. Using the symbols in (4.2), $\hat{Z}_F - \hat{Z}$ can be expressed as follows:

$$\begin{aligned}\hat{Z}_F - \hat{Z} &= W_F\hat{\Theta}_F - W\hat{\Theta} \\ &= (P_{W_F|X_F} - P_{W|X})Y \\ &= (P_{W_F|X_F} - P_{W|X})(X_F\Theta_{F_*} + \mathscr{E}) \\ &= (W_F - P_{W|X}X_F)\Theta_{F_*} + (P_{W_F|X_F} - P_{W|X})\mathscr{E} \\ &= \mathscr{V}_W\Sigma_{F_*}^{1/2} + (P_{W_F|X_F} - P_{W|X})\mathscr{E}. \end{aligned} \tag{6.3}$$

From (6.3) and Lemma 4.1,

$$E_*^Y\{\Sigma_{F_*}^{-1/2}(\hat{Z}_F - \hat{Z})'(\hat{Z}_F - \hat{Z})\Sigma_{F_*}^{-1/2}\} = \Omega_W + \psi I_p.$$

On the other hand, an expectation of the inversed Wishart distribution matrix is obtained (see e.g. Siotani, Hayakawa and Fujikoshi (1985, p. 59)).

$$E_*^Y(n\hat{\Sigma}_F)^{-1} = \frac{1}{n-k-p-1}\Sigma_{F_*}^{-1}.$$

Since $\hat{Z}_F - \hat{Z}$ and $\hat{\Sigma}$ are independent,

$$E_*^Y(D) = c(\Sigma_{F_*}^{1/2}\Omega_W\Sigma_{F_*}^{-1/2} + \psi I_p).$$

Recall that $\Lambda_W = (I_p + m^{-1}\Omega_W)^{-1}$. This completes the proof. □

Acknowledgements

I wish to express my deepest gratitude to Prof. Y. Fujikoshi of Hiroshima University for his comments and advices. Also I would like to thank Prof. M. Ohtaki of Hiroshima University for his support and encouragement. Further, I am grateful to Dr. H. Fujisawa of Tokyo Institute of Technology and Dr. K. Naito of Shimane University for their encouragement.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory, Eds. B. N. Petrov and F. Csáki, pp. 267–281. Budapest: Akadémia Kiado, 1973.

[2] T. W. Anderson, An introduction to multivariate statistical analysis, John Wiley & Sons, New York (1958).

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

[4] S. Chatterjee and B. Price, Regression analysis by example, John Wiley & Sons, New York (1977).
[5] N. Draper and H. Smith, Applied regression analysis, John Wiley & Sons, New York (1966).
[6] Y. Fujikoshi and K. Satoh, Modified *AIC* and C_p criterion in multivariate linear regression, Biometrika **84** (1997), 707–716.
[7] Y. Fujikoshi and K. Satoh, Estimation and model selection in an extended growth curve model, Hiroshima mathematical Journal **26** (1996), 635–647.
T. K. Keyes and M. S. Levy, Goodness of prediction fit for multivariate linear models, J. Amer. Statist. Ass. **91** (1996), 191–197.
H. Linhart and W. Zucchini, Model selection, Wiley (1986).
[8] C. M. Hurvich and C. L. Tsai, Regression and time series model selection in small samples. Biometrika **76** (1989), 297–307.
[9] A. J. Miller, Subset Selection in Regression, Chapman and Hall (1990).
[10] G. D. Murray, A note on the estimation of probability density functions, Biometrika **64** (1977), 150–152.
[11] R. Nishii, Asymptotic properties of criteria for selection of variables in multiple regression. The Annals of Statistics **12** (1984), 758–765.
[12] C. R. Rao, Linear statistical Inference and its application, John Wiley & Sons, New York (1973).
[13] K. Satoh, AIC-type model selection criterion for multivariate linear regression with a future experiment, J. Japan Statist. Soc. **27** (1997), 135–140.
[14] K. Satoh, M. Kobayashi and Y. Fujikoshi, Variable selection for the growth curve model, J. Multivariate Analysis **60** (1997), 277–292.
[15] G. A. F. Seber, Linear regression analysis, John Wiley & Sons, New York (1977).
[16] R. Shibata, Selection of the order of an autoregressive model by Akaike's information criterion, Biometrika **63** (1976), 117–126.
[17] R. Shibata, An optimal selection of regression variables, Biometrika **68** (1981), 45–54.
[18] S. D. Silvey, Statistical Inference, Penguin Books (1970).
[19] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections, Commun. Statist.—Theory Meth. **7** (1978), 13–26.

Department of Environmetrics and Biometrics
Research Institute for Radiation Biology and Medicine
Hiroshima University
1-2-3 Kasumi, Minami-ku, Hiroshima, 734-8553 Japan