Complete Bias Modification of Model Risk Based on KL Divergence

Hirofumi Wakaki[1]*, Hirokazu Yanagihara[2,3,4]

[1]*Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan*

[2]*Osaka Central Advanced Mathematical Institute, Osaka Metropolitan University, 3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan*

[3]*Department of Medical Statistics, Research & Development Center, Osaka Medical and Pharmaceutical University, 2-7 Daigaku-machi, Takatsuki, Osaka 569-8686, Japan*

[4]*Mathematical Risk Analysis, Risk Analysis Research Center, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*

December 24, 2025

Abstract

Consider a risk function based on the KL divergence for a variable selection problem in a normal multiple regression model. Fujikoshi and Satoh (1997) proposed Modified AIC (MAIC) as an asymptotically unbiased estimator of this risk function. MAIC corrects for bias in an underspecified model not including the true model. However, it leaves bias in an overspecified model, unlike Corrected AIC, developed by Sugiura (1978). Therefore, the present study proposes an unbiased estimator that can fully correct bias in both overspecified and underspecified models. Numerical experiments are conducted to confirm its performance as an information criterion.

***2020 Mathematics Subject Classification*:** 62J05, 62E15
***Keywords*:** AIC, Information Criterion, MAIC, Normal Multiple Regression Model, Variable Selection, UMVUE

1. Introduction

In this paper, we deal with a normal multiple regression model. In multiple regression models, various combinations of explanatory variables can be considered, and thus, variable selection (i.e., determining which combination yields the optimal model) plays a crucial role. Under the normality assumption, model adequacy is typically evaluated using a risk function based on the Kullback–Leibler (KL) divergence [8], where the model that minimizes this risk function is regarded as optimal. Unfortunately, since the risk function involves unknown parameters, it must be estimated in order to assess model adequacy. The most well-known estimator of this risk function is the Akaike Information Criterion (AIC) [1]. When the model assumes that the n response variables are independently and identically distributed (referred to as an *i.i.d.* model), it is known that AIC serves as an asymptotically unbiased estimator of the risk function. Due to the simplicity of its definition and its versatility, AIC has been widely used across many fields. Although AIC is asymptotically unbiased for an *i.i.d.* model, when the sample size is not sufficiently large, the bias with respect to the risk function can be substantial, which may lead to critical errors in variable selection. In particular, in multiple regression models, AIC tends to underestimate the risk function as the number of explanatory variables in the candidate model increases; that is, AIC deviates below farther from the true risk function as the number of explanatory variables grows. In addition, the variance of AIC as an estimator of the risk function increases with the number of explanatory variables (see e.g., [18]). Consequently, AIC has the drawback of being more likely to select models with a larger number of explanatory variables as optimal. Figure 1 illustrates the values of the risk function and the boxplots of its AIC estimates obtained from 10,000 replications in a polynomial-degree selection problem under the assumptions $n = 30$ and normally distributed errors with standard deviation 0.5. In Figure 1, the left panel shows boxplots of AIC where the horizontal axis is the polynomial degree of the candidate models. As clearly shown, AIC tends to underestimate the risk function, the deviation between AIC and the true risk function grows with the polynomial degree (corresponding to the number of explanatory variables), and the variance of AIC as an estimator also increases with increasing polynomial degree. The underestimation tendency of AIC is particularly pronounced for models that contain the true model, referred to as overspecified models. In Figure 1, since

*Corresponding author (E-mail address: wakaki@hiroshima-u.ac.jp)

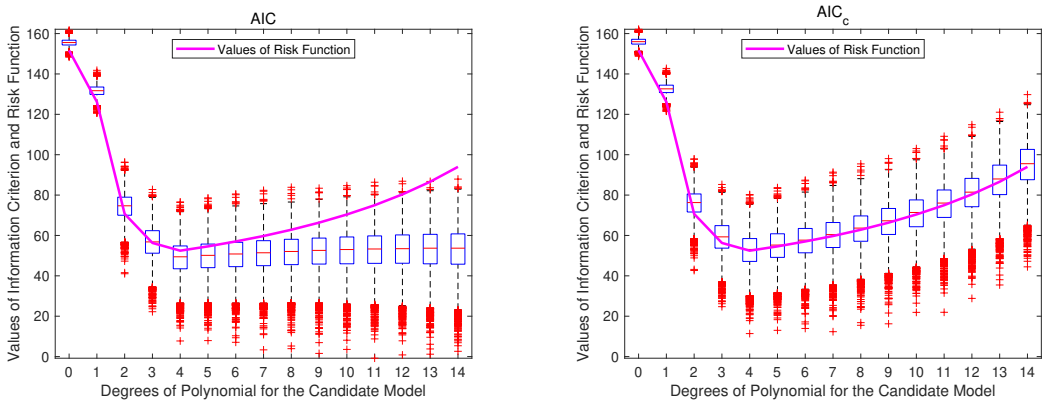

Figure 1: Risk function and boxplots of AIC and AIC$_c$

the true model corresponds to degree 4, models with degree 4 or higher are overspecified. It can be confirmed that for overspecified models, the underestimation of the risk function by AIC is especially severe. For these reasons, much research has been devoted to correcting the bias with respect to the risk function of AIC under the assumption that the candidate model is overspecified, with the aim of improving the shortcomings inherent in AIC.

For the normal multiple regression model considered in the present paper, Sugiura [15] proposed a bias-corrected version of AIC, called the Corrected AIC, which provides an unbiased estimator of the risk function under overspecified models. This criterion later denoted as AIC$_c$ by Hurvich and Tsai [5], and it has become the standard notation used worldwide . Furthermore, Davies *et al.* [2] reported that AIC$_c$ is the uniformly minimum variance unbiased estimator (UMVUE) of the risk function under the assumption that the candidate model is overspecified. The right panel of Figure 1 shows the boxplots of AIC$_c$. The tendency of AIC to underestimate the risk function disappears, and although the variance of AIC$_c$ increases with the number of explanatory variables (i.e., polynomial degree in this case), the corresponding increase in the risk function compensates for this, thereby eliminating the drawback of AIC of selecting models with too many explanatory variables.

However, even if the bias can be fully corrected under overspecified models, there still exists bias with respect to the risk function when the candidate model does not include the true model (underspecified models). As mentioned earlier, for an *i.i.d.* model, AIC remains an asymptotically unbiased estimator of the risk function even under underspecified models. In contrast, for models such as the normal multiple regression model where the response variables are independent but not identically distributed, AIC exhibits a constant bias with respect to the risk function under underspecified models, meaning that it loses its asymptotic unbiasedness. In particular, within the set of all candidate models formed by considering all possible combinations of explanatory variables, most models are underspecified. Moreover, in multiple regression models, the model that minimizes the risk function is either the true model or an underspecified model [19]. Hence, it is evident that information criteria faithful to the risk function are also required under underspecified models.

The first information criterion for the normal multiple regression model to incorporate bias correction under underspecified models was proposed by Sawa [13]. In the paper of Sawa, the criterion was denoted as BIC, with the idea of it being the "next" information criterion after AIC. However, since Schwarz [14] independently proposed the Bayesian Information Criterion using the same acronym, we refer to Sawa's criterion here as Sawa's BIC (SBIC). SBIC modifies AIC by replacing the "2 $\times$ the number of parameters" penalty, and achieves a bias of order $O(n^{-1})$ with respect to the risk function under both underspecified and overspecified models. In comparison, Hurvich and Tsai [6] derived an exact expression for the risk function by using an infinite series expansion in powers of the unknown parameters. Subsequently, Reschenhofer [11] proposed a corrected version of SBIC, referred to here as Corrected SBIC (SBIC$_c$), employing the expansion truncated up to the n^{-1} term. However, since SBIC$_c$ is constructed by simply substituting estimators into the expansion up to the n^{-1} term, the bias of the $O_p(1)$ term propagates into the n^{-1} term. Consequently, although SBIC$_c$ was presented as a correction of SBIC, the order of bias with respect to the risk function remains the same as that of SBIC, that is, $O(n^{-1})$. Other attempts to address bias correction under underspecified models include those by Noda *et al.* [10] and Fujikoshi *et al.* [4]. However, all such information criteria achieve that same order of bias with respect to the risk function, $O(n^{-1})$.

For the normal multiple regression model, the Modified AIC (MAIC) proposed by Fujikoshi and Satoh [3] can be viewed, in a sense, as a combination of AIC$_c$ and SBIC. Specifically, MAIC was constructed by adding to AIC$_c$ an adjusted correction term from SBIC. This additional term is of order $O_p(1)$ under underspecified models and of order $O(n^{-2})$ under overspecified models. As a result, the bias of MAIC itself with respect to the risk function is of order $O(n^{-1})$ under underspecified models and $O(n^{-2})$ under overspecified models. Thus, compared with SBIC and SBIC$_c$, MAIC achieves an improved order of bias under overspecified models. However, in correcting the bias under underspecified models, MAIC sacrifices the unbiasedness of AIC$_c$ under overspecified models, thereby losing the desirable UMVUE property of AIC$_c$ with respect to the risk function.

As noted earlier, the model that minimizes the risk function is either the true model or an underspecified model. Therefore, it is essential to have estimators faithful to the risk function even under underspecified models. On the other hand, under overspecified models, unbiased estimation of the risk function enables the resulting information criterion to have the same natural properties of the risk function, that its value and its growth increase as the number of explanatory variables increases. If one attempts to approximate this behavior with only asymptotic expansions truncated at finite orders, the approximation deteriorates as the number of explanatory variables increases, as illustrated in Figure

1 for the case of AIC. Consequently, the information criterion fails to capture the increasing trend of the risk function with larger models.

From these considerations, it is necessary to develop an information criterion that serves as an unbiased estimator of the risk function under both underspecified and overspecified models. For overspecified models, constructing an unbiased estimator is relatively straightforward, since the bias of "$-2\times$ the maximum log-likelihood function" with respect to the risk function can be expressed in terms of the expectation of the reciprocal of a chi-squared distributed random variable. In contrast, under underspecified models, the bias involves the expectation of the reciprocal of a noncentral chi-squared random variable, which can only be represented as an infinite series in terms of the noncentrality parameter. Furthermore, because the noncentrality parameter depends on unknown parameters, constructing an unbiased estimator in a straightforward manner would require unbiased estimation of all powers of the noncentrality parameter, which is highly challenging.

In the present paper, we overcome this difficulty by exploiting properties of the Pochhammer symbol and the Poisson distribution to construct an unbiased estimator of the risk function. The newly proposed information criterion becomes the UMVUE of the risk function, provided that the full model, including all explanatory variables, contains the true model. Through numerical experiments, it confirms that the proposed criterion outperforms other existing criteria as a model selection criterion. Mathematical details are provided in the appendix.

2. Existing Information Criteria

We consider the problem of variable selection for the following normal multiple regression model:

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \quad \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathsf{t}},$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observations, $\boldsymbol{x}_i$ is the $k \times 1$ vector of explanation variables for the i-th individuals, $\boldsymbol{\beta}$ is the unknown vector of regression coefficients, and σ^2 is the unknown variance parameter. We assume that $\mathrm{rank}(\boldsymbol{X}) = k < n + 4$ in order to ensure the existence of the proposed information criterion.

Let j be an arbitrary subset of the index set $\omega = \{1, 2, \ldots, k\}$. Consider fitting the model

$$M_j : \boldsymbol{y} \sim N_n(\boldsymbol{X}_j\boldsymbol{\beta}_j, \sigma_j^2\boldsymbol{I}_n),$$

where $\boldsymbol{X}_j$ consists of the $k_j = \#j$ columns of $\boldsymbol{X}$ corresponding to the indices in j, and $\boldsymbol{\beta}_j$ is the $k_j \times 1$ parameter vector.

Let φ be the true probability density function of $\boldsymbol{y}$ and let $\hat{f}_j$ be the predicted probability density function based on the model M_j, that is, the probability density function of $N_n(\boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2\boldsymbol{I}_n)$, where $(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)$ is the maximum likelihood estimator of $(\boldsymbol{\beta}_j, \sigma_j^2)$ given by

$$\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{y}, \quad \hat{\sigma}_j^2 = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{y}, \tag{2.1}$$

in which $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^{\mathsf{t}}$.

Then the KL divergence [8] of $\hat{f}_j$ from φ is given by

$$\begin{aligned} d_{\mathrm{KL}}(\varphi, \hat{f}_j) &= \mathrm{E}_{\boldsymbol{u}}\left[\log\left\{\frac{\varphi(\boldsymbol{u})}{\hat{f}_j(\boldsymbol{u})}\right\}\right] \\ &= \mathrm{E}_{\boldsymbol{u}}\left[\log\varphi(\boldsymbol{u})\right] + \frac{1}{2}\left\{n\log(2\pi\hat{\sigma}_j^2) + \frac{1}{\hat{\sigma}_j^2}\left\{n\sigma_*^2 + (\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)\right\}\right\}, \end{aligned}$$

where $\mathrm{E}_{\boldsymbol{u}}[\cdot]$ indicates expectation with respect to the distribution of $\boldsymbol{u}$, and we assume that $\boldsymbol{u}$ and $\boldsymbol{y}$ are independent, $\mathrm{E}[\boldsymbol{u}] = \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\eta}_*$, and $\mathrm{Var}[\boldsymbol{u}] = \mathrm{Var}[\boldsymbol{y}] = \sigma_*^2\boldsymbol{I}_n$. The risk function of model M_j is defined by the expectation of the KL divergence. However, since the first term of the KL divergence is independent of model M_j, in practice the risk function is defined as follows, with the first term removed and doubled:

$$R_{\mathrm{KL}}(j) = n\mathrm{E}[\log(2\pi\hat{\sigma}_j^2)] + \mathrm{E}\Big[\frac{1}{\hat{\sigma}_j^2}\left\{n\sigma_*^2 + (\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)\right\}\Big]. \tag{2.2}$$

In this formation of the risk function, the function inside the expectation is called the loss function, that is, $R_{\mathrm{KL}}(j) = E[\mathcal{L}_{\mathrm{KL}}(j)]$. Hence, the loss function $\mathcal{L}_{\mathrm{KL}}(j)$ is given by

$$\mathcal{L}_{\mathrm{KL}}(j) = n\log(2\pi\hat{\sigma}_j^2) + \frac{1}{\hat{\sigma}_j^2}\left\{n\sigma_*^2 + (\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)\right\}. \tag{2.3}$$

Let $\mathcal{J} \subseteq 2^\omega$ be the set of all candidate models. Then we denote the set of overspecified models as $\mathcal{J}_+$, that is,

$$\mathcal{J}_+ = \{j \in \mathcal{J} \mid (\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\eta}_* = \boldsymbol{0}_n\},$$

and the set of underspecified models as $\mathcal{J}_- = 2^\omega \cap \mathcal{J}_+^c$.

Let $\hat{\ell}(j)$ be logarithm of the maximum likelihood of model M_j, that is, $-2\hat{\ell}(j) = n\{\log(2\pi\hat{\sigma}_j^2) + 1\}$. Then the bias of $-2\hat{\ell}(j)$ as an estimator of $R_{\mathrm{KL}}(j)$ is

$$\mathrm{E}[-2\hat{\ell}(j)] - R_{\mathrm{KL}}(j) = n - \mathrm{E}\left[\frac{1}{\hat{\sigma}_j^2}\left\{n\sigma_*^2 + (\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)\right\}\right]. \tag{2.4}$$

We can see that if $j \in \mathcal{J}_+$ and the distribution of $\boldsymbol{y}$ is normal, then the original AIC proposed by Akaike [1],

$$\mathrm{AIC}(j) = -2\hat{\ell}(j) + 2(k_j + 1),$$

has bias of order $O(n^{-1})$, whereas AIC_c (i.e., Corrected AIC) proposed by Sugiura [15],

$$\mathrm{AIC}_c(j) = -2\hat{\ell}(j) - n + \frac{n(n+k_j)}{n-k_j-2} = \mathrm{AIC}(j) + \frac{2(k_j+1)(k_j+2)}{n-k_j-2},$$

is an unbiased estimator of the risk function if $j \in \mathcal{J}_+$. On the other hand, if $j \in \mathcal{J}_-$, then both $\mathrm{AIC}(j)$ and $\mathrm{AIC}_c(j)$ have biases of order $O(1)$.

In comparison, the BIC proposed by Sawa [13] that we denote as SBIC shrinks the bias for the underspecified models. For model j, SBIC is given by

$$\mathrm{SBIC}(j) = \mathrm{AIC}(j) - 2(1-\hat{\gamma}_j)(k_j + 1 - \hat{\gamma}_j),$$

where $\hat{\gamma}_j = \hat{\sigma}_\omega^2/\hat{\sigma}_j^2$. We can see that regardless of whether $j \in \mathcal{J}_-$ or $j \in \mathcal{J}_+$, the order of the bias of SBIC is $O(n^{-1})$; that is, SBIC is always an asymptotic unbiased estimator of the risk function. Let $\gamma_j = n/(n+\theta_j^*)$, where $\theta_j^* = \boldsymbol{\beta}_*^{\mathsf{t}}\boldsymbol{X}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}_*/\sigma_*^2$, in which $\boldsymbol{\beta}_*$ is the vector such that $\boldsymbol{\eta}_* = \boldsymbol{X}\boldsymbol{\beta}_*$, which exists if $\mathcal{J}_+ \neq \emptyset$. Using γ_j, Reschenhofer [11] expanded the bias in (2.4) up to the order n^{-1} as

$$2\left\{(k_j+1) - (1-\gamma_j)(k_j+1-\gamma_j) + \frac{(k_j^2+7k_j+12)\gamma_j^2 - 4(k_j+4)\gamma_j^3 + 6\gamma_j^4}{n-k_j-2}\right\}.$$

Using the equation obtained by replacing γ_j in the above expansion with $\hat{\gamma}_j$ given in the definition of SBIC, Reschenhofer [11] proposed a new information criterion. This information criterion is defined by adding a correction term of order n^{-1} to SBIC, so in the present paper, which is why we refer to this criterion as Corrected SBIC (SBIC_c). For model j, SBIC_c is given by

$$\mathrm{SBIC}_c(j) = \mathrm{SBIC}(j) + \frac{2\{(k_j^2+7k_j+12)\hat{\gamma}_j^2 - 4(k_j+4)\hat{\gamma}_j^3 + 6\hat{\gamma}_j^4\}}{n-k_j-2}.$$

Despite the use of an asymptotic expansion of the bias, unfortunately, the order of bias for SBIC_c remains unchanged from that of SBIC. This is because $2(1-\hat{\gamma}_j)(k_j+1-\hat{\gamma}_j)$, the correction term for SBIC, possesses a bias of order n^{-1}, regardless of whether $j \in \mathcal{J}_-$ or $j \in \mathcal{J}_+$.

Now, recall that Modified AIC (MAIC) proposed by Fujikoshi and Satoh [3], like SBIC and SBIC_c, shrinks the bias for underspecified models. For model j, MAIC is given by

$$\mathrm{MAIC}(j) = \mathrm{AIC}_c(j) - 2\hat{\xi}_j(k_j + \hat{\xi}_j),$$

where $\hat{\xi}_j = 1 - (n-k_j)(n-k)^{-1}\hat{\gamma}_j$ and $\hat{\gamma}_j$ is given in the definition of SBIC. MAIC can be regarded as having been obtained by substituting $\mathrm{AIC}(j)$ with $\mathrm{AIC}_c(j)$ and $1-\hat{\gamma}_j$ with $\hat{\xi}_j$ in the definition formula of SBIC. Note that $\mathrm{AIC}_c(j)$ is an unbiased estimator of the risk function when $j \in \mathcal{J}_+$, and the expectation of $\hat{\xi}_j(k_j+\hat{\xi}_j)$ can be expanded as

$$E\left[\hat{\xi}_j(k_j+\hat{\xi}_j)\right] = \begin{cases} (1-\gamma_j)(k_j+1-\gamma_j) + O(n^{-1}) & (j \in \mathcal{J}_-) \\ O(n^{-2}) & (j \in \mathcal{J}_+) \end{cases}.$$

Hence, we can see that if $j \in \mathcal{J}_-$, then $\mathrm{MAIC}(j)$ has bias of order $O(n^{-1})$, and if $j \in \mathcal{J}_+$, then the order of the bias is $O(n^{-2})$. The order of bias for AIC, AIC_c, SBIC, and SBIC_c for the risk function is summarized in Table 1. In the present paper, we propose a new modified AIC (Completely Modified AIC, MAIC_c) which is an unbiased estimator of $R_{\mathrm{KL}}(j)$ for both $j \in \mathcal{J}_+$ and $j \in \mathcal{J}_-$.

Table 1: Order of bias of each criterion

	Order	
Information Criterion	Underspecified	Overspecified
AIC	$O(1)$	$O(n^{-1})$
AIC_c	$O(1)$	0
SBIC	$O(n^{-1})$	$O(n^{-1})$
SBIC_c	$O(n^{-1})$	$O(n^{-1})$
MAIC	$O(n^{-1})$	$O(n^{-2})$

It is known that Takeuchi's information criterion (TIC) [16] and the extended information criterion (EIC) proposed by Ishiguro *et al.* [7] both also correct the bias of AIC under model misspecification. Here, the phrase "model misspecification" refers not only to the case of underspecification discussed herein but also to situations in which the probability distribution of the response variable is misspecified. Like AIC, both TIC and EIC are asymptotically unbiased estimators of the risk function under model misspecification for the *i.i.d.* model. However, for the multiple regression model, they are not asymptotically unbiased estimators of the risk function under underspecified models. The orders of their biases are the same as that of AIC (see [17]). On the other hand, Fujikoshi *et al.* [4] reported that, in EIC, resampling the residuals of the full model yields an asymptotically unbiased estimator of the risk function even under underspecified models. In such a case, the order of the bias is the same as that of SBIC (see [4]).

3. Unbiased Estimator of Risk Function

We propose in the following theorem a new modified AIC which is an unbiased estimator of $R_{\mathrm{KL}}(j)$ for both $j \in \mathcal{J}_+$ and $j \in \mathcal{J}_-$. We refer to this new AIC as "Completely Modified AIC" and denote it as MAIC_c.

Theorem 3.1. *Define* MAIC_c *for the model* M_j *as*

$$\mathrm{MAIC}_c(j) = -2\hat{\ell}(j) + \frac{n(n+k_j)}{2}\sum_{r=1}^{\infty}\frac{(r-1)!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r - \frac{n(n-k_j-4)}{2}\sum_{r=1}^{\infty}\frac{r!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r, \tag{3.5}$$

where $\hat{\gamma}_j = \hat{\sigma}_\omega^2/\hat{\sigma}_j^2$ *and* $(a)_r$ *is the Pochhammer symbol defined by*

$$(a)_r = \begin{cases} 1 & (r=0) \\ a(a+1)\cdots(a+r-1) & (r=1,2,\ldots) \end{cases}.$$

Then $\mathrm{MAIC}_c(j)$ *is an unbiased estimator of the risk* $R_{KL}(j)$ *given by* (2.2) *for both* $j \in \mathcal{J}_+$ *and* $j \in \mathcal{J}_-$ *if* $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}_*, \sigma_*^2\boldsymbol{I}_n)$ *for some* $\boldsymbol{\beta}_*$ *and* σ_*^2.

Proof. Note that $\boldsymbol{\eta}_* = \boldsymbol{X}\boldsymbol{\beta}_*$. Since $(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j) = (\boldsymbol{\eta}_* - \boldsymbol{P}_j\boldsymbol{y})^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{P}_j\boldsymbol{y})$ is independent of $\hat{\sigma}_j^2 = n^{-1}\boldsymbol{y}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{y}$, the bias of $-2\hat{\ell}(j)$ as an estimator of $R_{\mathrm{KL}}(j)$ given by (2.4) is

$$\begin{aligned}
&n - \mathrm{E}\left[\frac{1}{\hat{\sigma}_j^2}\left\{n\sigma_*^2 + (\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)^{\mathsf{t}}(\boldsymbol{\eta}_* - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j)\right\}\right] \\
&= n - \mathrm{E}\left[\frac{1}{\hat{\sigma}_j^2}\right]\left\{n\sigma_*^2 + (\boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{P}_j\boldsymbol{X}\boldsymbol{\beta}_*)^{\mathsf{t}}(\boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{P}_j\boldsymbol{X}\boldsymbol{\beta}_*) + \mathrm{tr}\big[\boldsymbol{P}_j\mathrm{Var}(\boldsymbol{y})\big]\right\} \\
&= n - n\mathrm{E}\left[\frac{\sigma_*^2}{n\hat{\sigma}_j^2}\right](n + \theta_j^* + k_j),
\end{aligned} \tag{3.6}$$

where $\theta_j^* = \boldsymbol{\beta}_*^{\mathsf{t}}\boldsymbol{X}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}_*/\sigma_*^2$. The following lemma gives an unbiased estimator of (3.6) as

$$\begin{aligned}
&n - \frac{n(n+k_j)}{2}\sum_{r=1}^{\infty}\frac{(r-1)!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r - n\left\{1 - \frac{n-k_j-4}{2}\sum_{r=1}^{\infty}\frac{r!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r\right\} \\
&= -\frac{n(n+k_j)}{2}\sum_{r=1}^{\infty}\frac{(r-1)!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r + \frac{n(n-k_j-4)}{2}\sum_{r=1}^{\infty}\frac{r!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r.
\end{aligned}$$

□

Lemma 3.1. *Let* $\boldsymbol{y}$ *be an* $n \times 1$ *random vector distributed as* $N_n(\boldsymbol{X}\boldsymbol{\beta}_*, \sigma_*^2\boldsymbol{I}_n)$. *Then if* $n - k > 4$, *the following equalities hold:*

$$E\left[\frac{\sigma_*^2}{n\hat{\sigma}_j^2}\right] = E\left[\frac{1}{2}\sum_{r=1}^{\infty}\frac{(r-1)!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r\right], \quad E\left[\frac{\theta_j\sigma_*^2}{n\hat{\sigma}_j^2}\right] = 1 - \frac{n-k_j-4}{2}E\left[\sum_{r=1}^{\infty}\frac{r!}{(\frac{n-k}{2})_r}\hat{\gamma}_j^r\right],$$

where $\theta_j^* = \boldsymbol{\eta}_*^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\eta}_*/\sigma_*^2$.

The proof of Lemma 3.1 is given in the Appendix.

The two power series in (3.5) can be represented using hypergeometric functions (see Remark A.1), which is useful for actual calculations of MAIC_c because several software libraries for numerical computation include hypergeometric functions. It can be useful to represent power series as integrals (see Lemma A.1).

Remark 3.1. *The* MLE $(\hat{\sigma}_j^2, \hat{\boldsymbol{\beta}}_j)$ *under* M_j *can be expressed as a function of* $(\hat{\sigma}_\omega^2, \hat{\boldsymbol{\beta}}_\omega)$ *as follows:*

$$\begin{aligned}
\hat{\sigma}_j^2 &= \boldsymbol{y}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{y} = \boldsymbol{y}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{y} + \boldsymbol{y}^{\mathsf{t}}\boldsymbol{P}_\omega\boldsymbol{y} - \boldsymbol{y}^{\mathsf{t}}\boldsymbol{P}_\omega\boldsymbol{P}_j\boldsymbol{P}_\omega\boldsymbol{y} \\
&= n\hat{\sigma}_\omega^2 + \hat{\boldsymbol{\beta}}_\omega\boldsymbol{X}^{\mathsf{t}}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}\hat{\boldsymbol{\beta}}_\omega, \\
\hat{\boldsymbol{\beta}}_j &= (\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{y} = (\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{P}_\omega\boldsymbol{y} = (\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^{\mathsf{t}}\boldsymbol{X}_\omega\hat{\boldsymbol{\beta}}_\omega.
\end{aligned}$$

Because $T = (\hat{\sigma}_\omega^2, \hat{\boldsymbol{\beta}}_\omega)$ *is a complete sufficient statistic of a family of normal distributions,* $\{N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n);\ \boldsymbol{\beta} \in \mathbb{R}^k, 0 < \sigma^2\}$, *and* MAIC_c *is a function of* T, *the Lehmann–Scheffé theorem assures that* MAIC_c *is the* UMVUE *of* $R_{\mathrm{KL}}(j)$ *if* $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. *On the other hand, if* $j \subsetneq \omega$ *and* $\boldsymbol{y} \sim N_n(\boldsymbol{X}_j\boldsymbol{\beta}_j, \text{then}\sigma_j^2\boldsymbol{I}_n)$, T *is not complete for the subfamily* $\{N_n(\boldsymbol{X}_j\boldsymbol{\beta}_j, \sigma_j^2\boldsymbol{I}_n);\ \boldsymbol{\beta}_j \in \mathbb{R}^{k_j}, 0 < \sigma_j^2\}$ *and* MAIC_c *is not* UMVUE *of* $R_{\mathrm{KL}}(j)$. *As mentioned in the Introduction,* AIC_c *is the* UMVUE *for this restricted subfamily.*

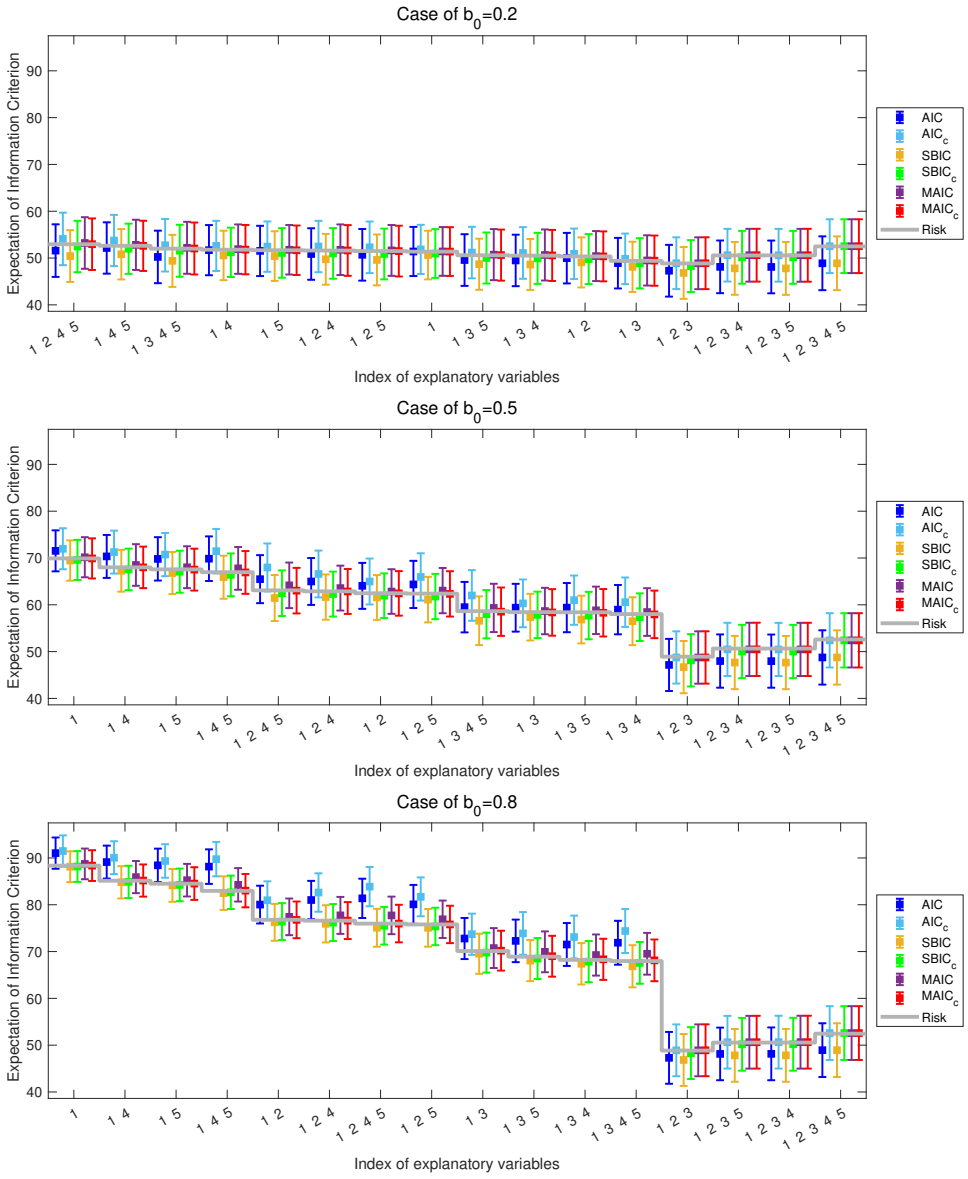

Figure 2: Risk function and expectations of six information criteria

4. Numerical Studies

4..1 Simulation Study

Let $\boldsymbol{Z}$ be an $n \times (k-1)$ matrix whose elements are generated independently from the uniform distribution over $(-1, 1)$ and $\boldsymbol{\Phi}(\rho)$ be the $(k-1) \times (k-1)$ auto-regressive correlation matrix whose (a, b)th element is $\rho^{|a-b|}$. Using $\boldsymbol{Z}$ and $\boldsymbol{\Phi}(\rho)$, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$, as $\boldsymbol{X} = (\mathbf{1}_n, (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Z}\boldsymbol{\Phi}(0.8)^{1/2})$, where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, and $\boldsymbol{J}_n = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{t}}$. Simulation data were generated from $N_n(\boldsymbol{X}\boldsymbol{\beta}_*, \sigma_*^2\boldsymbol{I}_n)$, where $\boldsymbol{\beta}_*$ is the $n \times 1$ vector whose ath element is given by

$$\beta_a^* = \begin{cases} b_0(-1)^{a+1}\{1 + 0.3(a-1)\} & (a = 1, \ldots, k_*) \\ 0 & (a = k_* + 1, \ldots, k) \end{cases}.$$

We required that all candidate models include the intercept. Since the first column of $\boldsymbol{X}$ is the intercept, the set of candidate models is given by $\mathcal{J} = \{j \in 2^\omega \mid j \supseteq \{1\}\}$. It is easy to see that the subset representing the true model was $j_* = \{1, \ldots, k_*\}$. Hence, the sets of overspecified and underspecified models were $\mathcal{J}_+ = \{j \in \mathcal{J} \mid j \supseteq j_*\}$ and $\mathcal{J}_- = \{j \in \mathcal{J} \mid j \not\supseteq j_*\}$, respectively. In all simulation studies conducted here, $n = 30$ and $\sigma^* = 0.5$. Using one simulation data set, we calculated six information criteria — AIC, AIC$_c$, SBIC, SBIC$_c$, MAIC, and MAIC$_c$ — and found model $\hat{j}$ that minimizes each criterion. Furthermore, we also derived model $\hat{j}_\star$ that minimizes the loss function $\mathcal{L}_{\mathrm{KL}}(j)$ in (2.3). In the simulation studies, several expected values were obtained via Monte Carlo integration, with simulation data generated through 10,000 iterations.

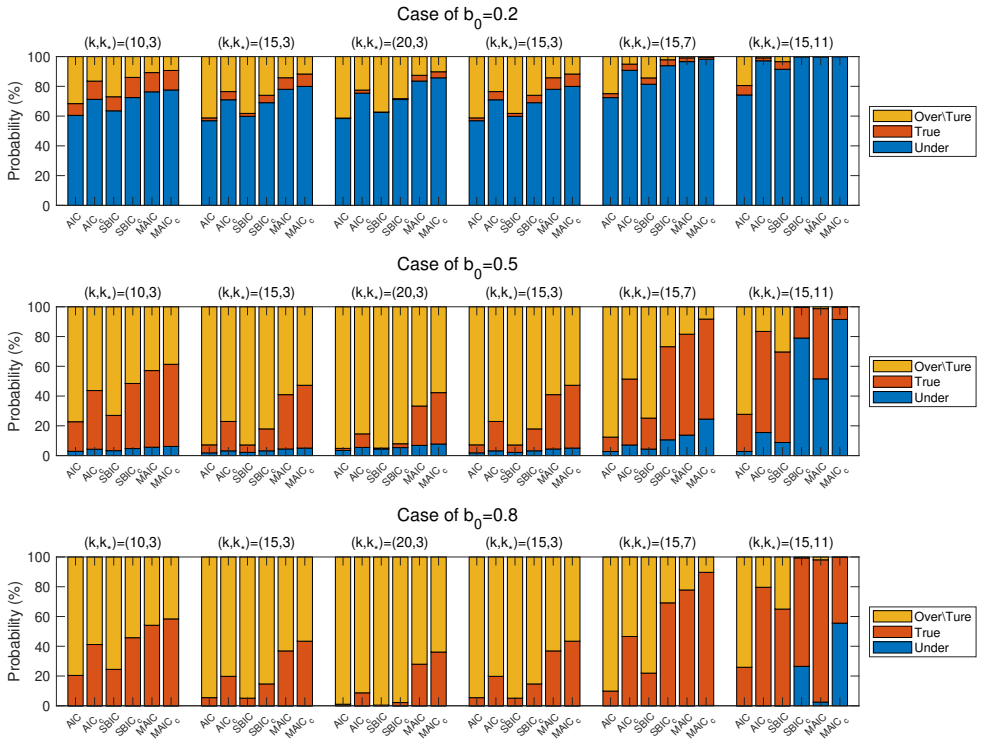

Figure 3: Selection probability of models chosen by each information criterion

First, we investigate the bias of the information criteria by plotting their averages together with the risk function. However, when the number of candidate models is large, such plots are too compressed to distinguish details clearly. Therefore, we conducted the numerical experiment with $(k, k_*) = (5, 3)$. In this setting, the overspecified models are $\{1, 2, 3\}$, $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$, and $\{1, 2, 3, 4, 5\}$, and the other models are all underspecified. Figure 2 presents the expectations of the six information criteria together with the risk function. In the figure, the solid line represents the value of the risk function, the center of each bar denotes the mean of the corresponding information criterion, and the length of each bar is proportional to the standard deviation. The top, middle and bottom panels correspond to the cases $b_0 = 0.2$, 0.5, and 0.8, respectively. In each panel, the candidate models j are listed along the x-axis. The models are arranged so that the order underspecified models are first and then the overspecified models, and within each of these two groups, they are arranged in order of decreasing and increasing values of the risk function, respectively. As b_0 increases, the noncentrality parameter becomes larger, and the difference between the underspecified and overspecified models becomes more pronounced. Note that, in overspecified models, the noncentrality parameter is zero, and thus the values remain the same regardless of b_0. From the figure, it can be observed that AIC and SBIC exhibit large biases under overspecified models. As explained in the Introduction, this is due to deterioration in the accuracy of the asymptotic expansion approximation as the number of explanatory variables increases. Similarly, AIC and AIC_c show large biases under underspecified models, which is expected, since both criteria have a constant bias in such cases. MAIC and SBIC_c exhibit nearly the same extents of bias. This is presumably because SBIC_c, although it has the same order of bias as SBIC, employs the asymptotic expansion of the risk function up to the n^{-1} term under underspecified models. Since SBIC_c relies on an asymptotic expansion, it tends to slightly underestimate the risk function compared with MAIC. MAIC generally exhibits little bias across all candidate models; however, under underspecified models with a large number of explanatory variables, the bias tends to be larger. The proposed MAIC_c, by contrast, is always an unbiased estimator of the risk function, which is visible in the figure. It is often argued that bias correction increases the standard deviation, and in this experiment, the standard deviation does become slightly larger, but this increase is not substantial.

To evaluate the performance of the criteria as variable selection methods, we examine which models are selected by the six information criteria and how well the selected models perform in terms of prediction. In this experiment, since the necessity of using information criteria for variable selection decreases as the number of candidate models becomes smaller, we conducted experiments with $k = 10$, 15, and 20. Specifically, we considered two settings: one where $k_* = 3$ and k increase from 10 to 20, and another where $k = 15$ and k_* increases from 3 to 11. This allows us to investigate both the effect of increasing the number of candidate models and the effect of increasing the number of explanatory variables in the true model. Figure 3 classifies the models selected by each information criterion into three categories: underspecified models (Under), the true model (True), and overspecified models excluding the true model (Over\True). As in Figure 2, the top, middle, and bottom panels correspond to the cases $b_0 = 0.2$, 0.5, and 0.8, respectively. To evaluate predictive performance, we compute the expected KL divergence of the selected model as the prediction error, $\mathrm{E}[\mathcal{L}_{\mathrm{KL}}(\hat{j})]$, where $\mathcal{L}_{\mathrm{KL}}(j)$ is the loss function defined in (2.3). Throughout the present paper, model adequacy is assessed using the risk function based on the KL divergence. Therefore, a variable selection method based on an information criterion is regarded as effective if it yields a smaller prediction error. Table 2 reports the prediction errors for each setting, normalized by subtracting the expected minimal loss $\mathrm{E}[\mathcal{L}_{\mathrm{KL}}(\hat{j}_\star)]$. From Figure 3, it is observed that AIC

Table 2: Prediction errors of selected model

b_0	k	k_*	AIC	AIC_c	SBIC	SBIC_c	MAIC	MAIC_c
0.2	10	3	17.36	14.48	16.31	13.86	13.19	12.84
	15	3	35.37	24.11	31.80	24.20	19.38	18.07
	20	3	86.56	42.39	71.49	48.77	27.90	25.01
	15	3	35.37	24.11	31.80	24.20	19.38	18.07
	15	7	41.16	30.48	36.09	28.31	25.99	24.68
	15	11	46.14	36.44	40.49	32.53	32.03	30.76
0.5	10	3	17.33	13.76	15.80	12.50	12.07	11.49
	15	3	35.48	22.95	30.77	22.04	17.23	15.58
	20	3	88.52	41.06	71.20	46.34	25.66	22.15
	15	3	35.48	22.95	30.77	22.04	17.23	15.58
	15	7	40.67	27.01	33.21	21.70	20.87	19.47
	15	11	45.77	35.43	37.59	31.90	33.07	31.42
0.8	10	3	16.96	13.28	15.42	11.91	11.42	10.76
	15	3	35.30	22.66	30.68	21.62	16.77	14.96
	20	3	87.78	39.74	70.14	45.12	24.18	20.62
	15	3	35.30	22.66	30.68	21.62	16.77	14.96
	15	7	40.50	26.47	32.75	20.23	19.09	16.21
	15	11	45.68	33.93	36.83	30.73	28.38	33.14

and SBIC tend to select overspecified models, that is, models with a larger number of explanatory variables. Between the two, AIC exhibits a stronger tendency to select such models. The other four information criteria have been improved with respect to this drawback of AIC, which often favors models with too many explanatory variables. In particular, the proposed MAIC_c more frequently selects either the true model or an underspecified model than do the other three information criteria. Given that the model minimizing the risk function is either the true model or an underspecified model, this property is desirable. Indeed, as shown in Table 2, MAIC_c consistently achieves smaller prediction errors across all settings.

4..2 Example Study

Next, we illustrate model selection using data obtained from Skagerberg *et al*. [12]. They simulated 56 data ($n = 56$) to study the relationship between polymer properties and the process data. The process data consist of 20 different temperature measurements (T_i, $i = 1, \ldots, 20$) taken at equal distances along the reactor, complemented with the wall temperature of the reactor (T_w) and the feed rate of the solvent (S). All temperature values are transformed as $\log_{10}(T - 100)$, where T is temperature.

The corresponding polymer properties are as follows: weight-average molecular weight (M_w), number-average molecular weight (M_n), frequency of long chain branching (LCB), frequency of short chain branching (SCB), and the contents of vinyl groups (VNL) and vinylidene groups (VND) in the polymer chain. For this example study, we used the polymer properties ($M_w^{-1} \times 10^5$, $M_n^{-1} \times 10^6$, LCB, SCB, VNL, VND) as the response variables and the process data ($T_1, \ldots, T_{20}, T_w, S$) as the explanatory variables. The number of explanatory variables in the full model is $k = 23$ because we always add a constant term to a regression. We selected the best model from all 4,194,304 ($= 2^{22}$) candidate models using values of AIC, AIC_c, SBIC, SBIC_c, MAIC and MAIC_c. Figure 4 shows the correlation coefficients between the explanatory variables. Because $T_1, \ldots, T_{20}$ are temperatures at equidistant points, as shown in the picture, temperatures at nearby points are highly correlated. However, as points become more distant, they do not necessarily become less correlated. Surprisingly, there are points that exhibit high negative correlations despite their distance. Variable T_w is highly correlated with the lower index T_i variables, whereas S generally exhibits a negative correlation with other variables. The figure reveals that a considerable number of the explanatory variables have strong intercorrelations, suggesting that this dataset is appropriate for variable selection procedures.

Figure 5 shows the explanatory variables selected by minimizing each information criterion. Panels show the results of variable selection for different response variables: from top to bottom, $M_w^{-1} \times 10^5$, $M_n^{-1} \times 10^6$, LCB, SCB, VNL, and VND. In each panel, all columns except the last two represent explanatory variables; variables that were never selected are shown as white (unfilled). Numerical values in these columns indicate a measure of the relevance of the regression coefficients, defined as $100 \times (1 - p\text{-value})$. Higher value indicates greater relevance of the regression coefficient. Note that these values are computed based on the selected model. The topmost row labeled "Full" shows the results for the full model using all explanatory variables. The penultimate column (UR: use ratio) shows the proportion of selected variables ($100 \times (k_j - 1)/(k - 1)$), and the last column ($R^2$) shows 100 times the coefficient of determination of the selected model. Since the intercept is always included in the model, we subtract one from both the number of selected variables and the total number of variables in UR. Across all response variables, the number of selected variables satisfies the ordering

$$\mathrm{AIC} \geq \mathrm{SBIC} \geq \mathrm{AIC}_c \geq \mathrm{SBIC}_c \geq \mathrm{MAIC} \geq \mathrm{MAIC}_c. \tag{4.7}$$

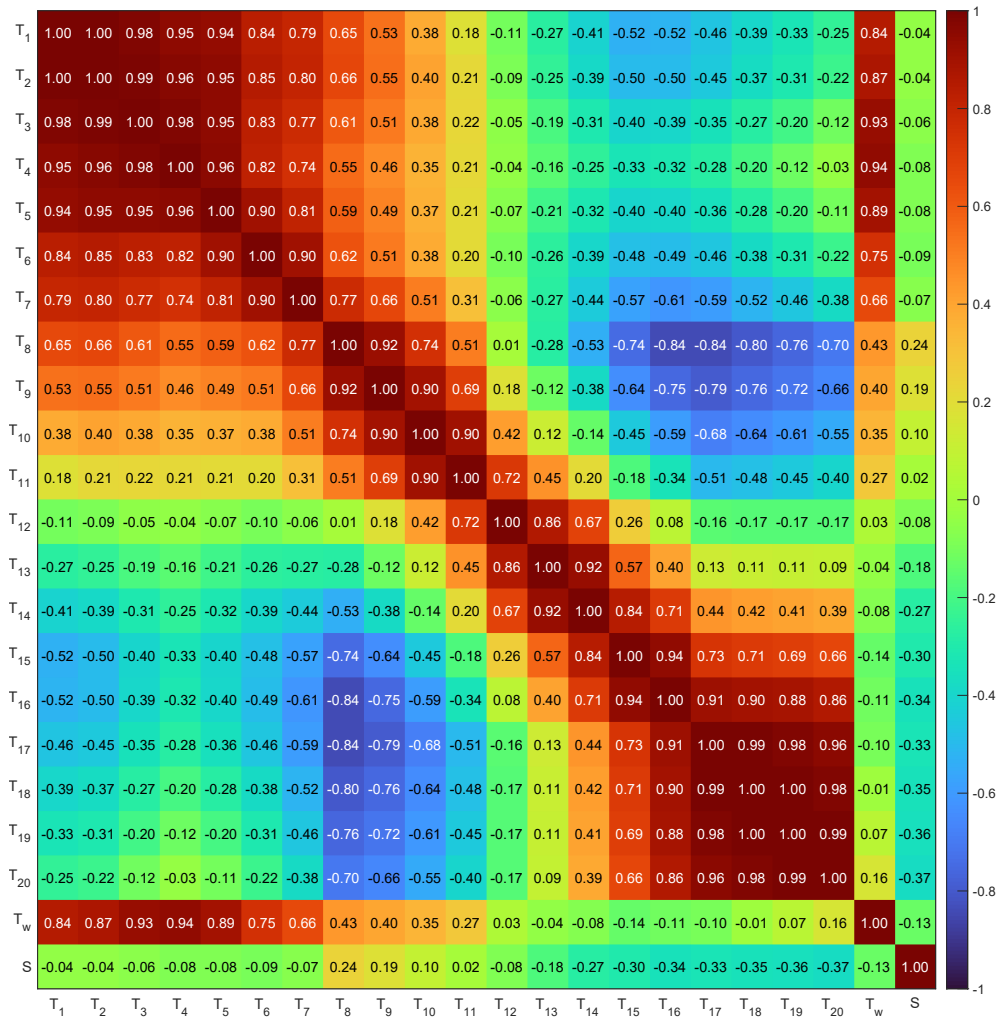

Figure 4: Correlation coefficients between explanatory variables

The results in this figure confirm that the other information criteria correct for AIC's well-known tendency to select models with an excessive number of explanatory variables. Although a model with fewer variables is not necessarily better, from the perspective of interpretability of analysis results, MAIC_c appears to be the most favorable information criterion. In particular, for all response variables, even the models selected by MAIC_c, which use the fewest explanatory variables, still achieve $R^2 > 90\%$. Moreover, variables with low relevance in the full model tend not to be selected, suggesting that the selection by MAIC_c is not unreasonable. The only somewhat puzzling behavior occurs for variable T_9 when the response variable is SCB. Since T_9 has a relevance of only 15.1% in the full model, it is excluded by almost all information criteria; however, it is selected solely by MAICc, where its relevance jumps to 100.0%. As mentioned earlier, this dataset contains many explanatory variables with high mutual correlations. Therefore, it is not surprising that different combinations of variables can be selected depending on the criterion. It is plausible that MAIC_c identifies T_9 as a substitute for variables that other criteria failed to select.

The variable selection results in Figure 5 are based on a single application of the variable selection procedure, and therefore do not reveal the potential instability of the results. To investigate the stability of variable selection, the 56 observations were randomly split into 10 validation samples and 46 calibration samples. Variable selection was then performed on each calibration sample, and the variable selection procedure was repeated 1,000 times with different random splits of the calibration samples. Figure 6 presents the results for each response variable. As in Figure 5, from top to bottom, the panels show $M_w^{-1} \times 10^5$, $M_n^{-1} \times 10^6$, LCB, SCB, VNL, and VND. In each panel, all columns except for the last two show percentages of times explanatory variables were selected across the 1000 repetitions.

Let $\hat{j}_a$ denote the model selected in the ath repetition, and define $\boldsymbol{s}(j)$ as the k-dimensional indicator vector for model j: the bth element of $\boldsymbol{s}(j)$ is 1 if b is included in j and 0 otherwise. For example, if $j = \{1, 3, 5\}$ and $k = 10$, then $\boldsymbol{s}(j)$ is the 10-dimensional vector with 1's as the first, third, and fifth elements and 0's as the other elements. The penultimate column (ave.) shows the average number of selected variables, defined as $10^{-3} \sum_{a=1}^{1000} (k_{\hat{j}_a} - 1)/(k-1)$, and the last column shows the similarity (sim.) of the selected models, defined by

$$\text{sim.} = 100 \times \left(1 - \frac{1}{500 \times 999} \sum_{a>b} \frac{1}{k-1} \left\| \boldsymbol{s}(\hat{j}_a) - \boldsymbol{s}(\hat{j}_b) \right\|_1 \right),$$

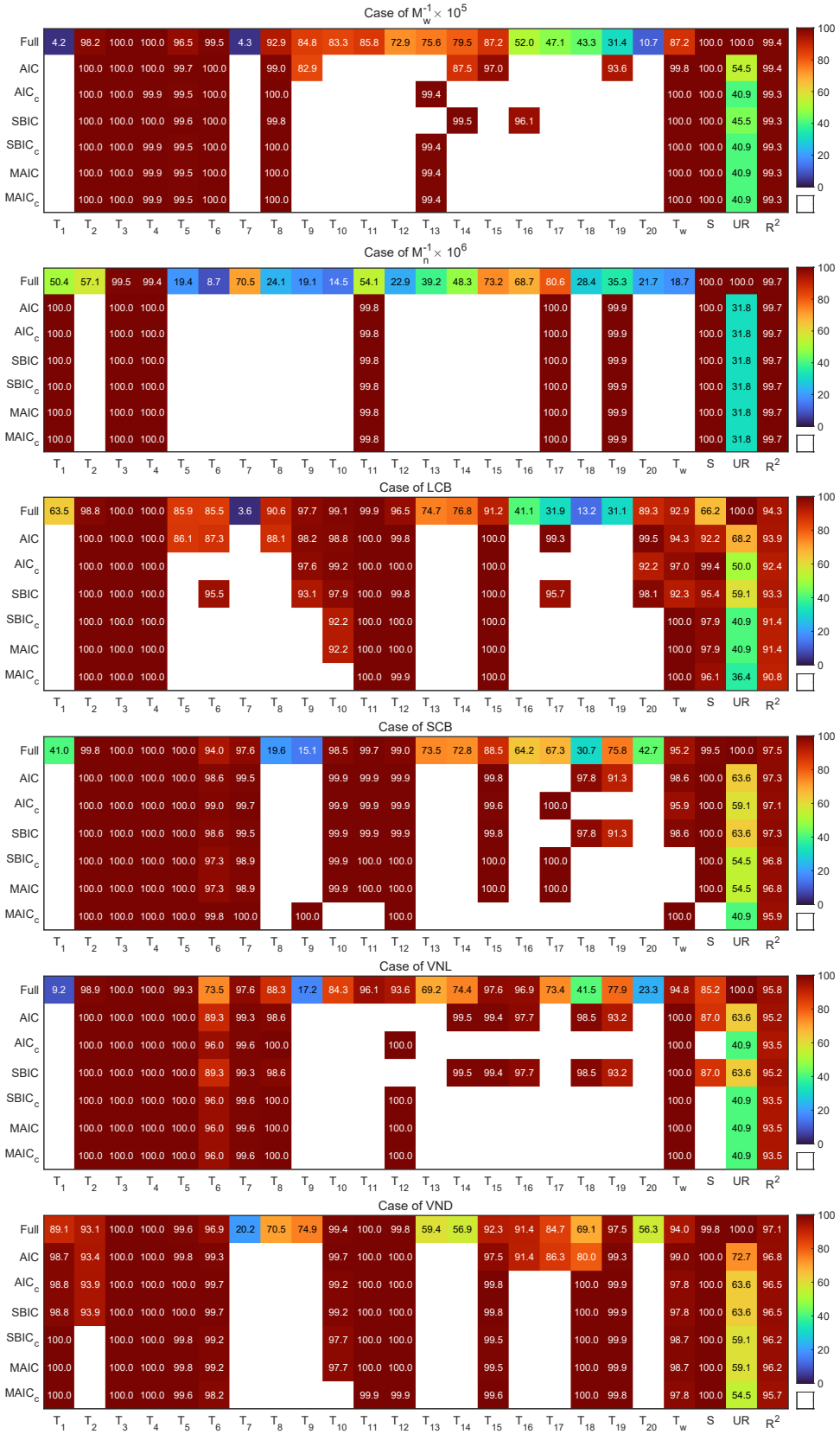

Figure 5: Selected explanatory variables and relevance of regression coefficients

Table 3: Estimated prediction errors of selected model

response variables	AIC	AIC_c	SBIC	SBIC_c	MAIC	MAIC_c
$M_w^{-1} \times 10^5$	380.88	330.80	352.81	323.53	305.27	276.19
$M_n^{-1} \times 10^6$	752.40	652.66	719.63	641.24	616.37	601.27
LCB	81.34	53.26	68.48	52.57	40.06	36.55
SCB	95.87	79.19	88.38	74.35	65.67	62.41
VNL	81.69	55.12	64.18	49.13	45.87	43.55
VND	126.30	101.96	109.47	88.97	78.27	71.38

where $\|\cdot\|_1$ denotes the L_1-norm, that is, the sum of the absolute values of all elements. By definition, similarity equals 100 if the exact same set of variables is selected in every repetition. Smaller similarity values indicate greater instability in the variable selection process. From Figure 6, we observe that, as in the results using the full dataset (Figure 5), the average number of selected variables follows the same ordering as in (4.7). Regarding similarity, for AIC_c, SBIC_c, MAIC, and MAIC_c, the sim. value tends to decrease (i.e., instability increases) as the number of selected variables increases, whereas no such tendency can be observed for AIC and SBIC. The trend with respect can be seen more clearly in the scatterplot of ave. and sim. shown in Figure 7. The scatterplot is across all response variables, numbered in same order as in Figures 5 and 6: 1 – $M_w^{-1} \times 10^5$, 2 – $M_n^{-1} \times 10^6$, 3 – LCB, 4 – SCB, 5 – VNL, and 6 – VND. Consequently, as the number of selected variables becomes relatively large, stability (similarity) becomes lower for AIC_c, SBIC_c, MAIC, and MAIC_c than for AIC and SBIC; however, as the number increases further, the instability trend reverses and AIC and SBIC become less stable. Furthermore, similarity satisfies $\mathrm{MAIC} \leq \mathrm{MAIC}_c$, indicating that MAIC exhibits more instability in variable selection results than MAIC_c. In addition, regarding variable T_9 (discussed above in relation to Figure 5 for the case where the response variable is SCB), it is selected in approximately 58.1% of the repetitions. It is also selected in almost 50% of the repetitions under MAIC, suggesting that the selection of T_9 is not at all surprising.

Unfortunately, although Figure 7 shows the results for each information criterion, it does not reveal how similar the criteria are to one another. Therefore, we attempted to visualize the relationships among the criteria by plotting them two-dimensionally using multidimensional scaling (MDS) based on the dissimilarity measure defined below. For this, we consider the six information criteria AIC, AIC_c, SBIC, SBIC_c, MAIC, and MAIC_c, indexed by $r = 1, \ldots, 6$, and let $\hat{j}_a^{(r)}$ denote the model selected by the rth information criterion in the ath repetition. We define the dissimilarity between the r_1th and r_2th information criteria as

$$\frac{1}{1000} \sum_{a=1}^{1000} \frac{1}{k-1} \left\| \boldsymbol{s}\left(\hat{j}_a^{(r_1)}\right) - \boldsymbol{s}\left(\hat{j}_a^{(r_2)}\right) \right\|_1 .$$

Using the dissimilarity matrix composed of these elements, we performed MDS to plot the six information criteria; the resulting configuration is shown in Figure 8. In each panel, the size of circles is proportional to the corresponding ave. value, with the numerical ave. values also shown. In addition, the stress value is reported above each panel. For all panels, the signs of the axes were adjusted so that AIC appears in the first quadrant. Because the stress values are below 0.05 for all response variables, the configurations can be considered highly faithful representations of the dissimilarities. The MDS results exhibit a U-shaped configuration of the information criteria, although the overall spread varies slightly between response variables. In particular, the horizontal axis roughly orders the criteria according to the magnitude of ave. Regarding the selected models, MAIC and MAIC_c are close to each other, as are AIC_c and SBIC_c. SBIC, on the other hand, is sometimes closer to AIC and sometimes closer to AIC_c, depending on the response variable.

Finally, we evaluate the predictive accuracy using the results from the 1000 repetitions. As an estimate of the prediction error, we quantify as follows. Let $w_a(j)$ denote the estimate of σ_j^2 in the ath repetition, let $\boldsymbol{y}_a^*$ denote the validation sample in the ath repetition, and let $\hat{\boldsymbol{y}}_a^*(j)$ denote the predicted values of $\boldsymbol{y}_a^*$ based on model j in the ath repetition. Then we define our estimate of the prediction error based on the KL divergence as follows:

$$\widehat{\mathrm{PE}} = \frac{1}{1000} \sum_{a=1}^{1000} \left\{ \log\left(\frac{w_a(\hat{j}_a)}{w_a(\omega)}\right) + \frac{1}{10 w_a(\hat{j}_a)} \left\| \boldsymbol{y}_a^* - \hat{\boldsymbol{y}}_a^*(\hat{j}_a) \right\|_2^2 \right\},$$

where $\|\cdot\|_2^2$ denotes the squared L_2-norm, that is, the sum of the squares for all elements. In the expression above, $w_a(j)$ is divided by $w_a(\omega)$ as a form of normalization; note that this operation does not affect the relative ordering of $\widehat{\mathrm{PE}}$. Table 3 reports the values of $\widehat{\mathrm{PE}}$ for each response variable. From this table, we observe that, for all response variables, MAIC_c achieves the smallest values of $\widehat{\mathrm{PE}}$.

5. Discussion

In the present paper, we treated the following linear regression model: $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$. Note that AIC_c by Sugiura [15] and MAIC by Fujikoshi and Satoh [3] were proposed for a multivariate linear model:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n),$$

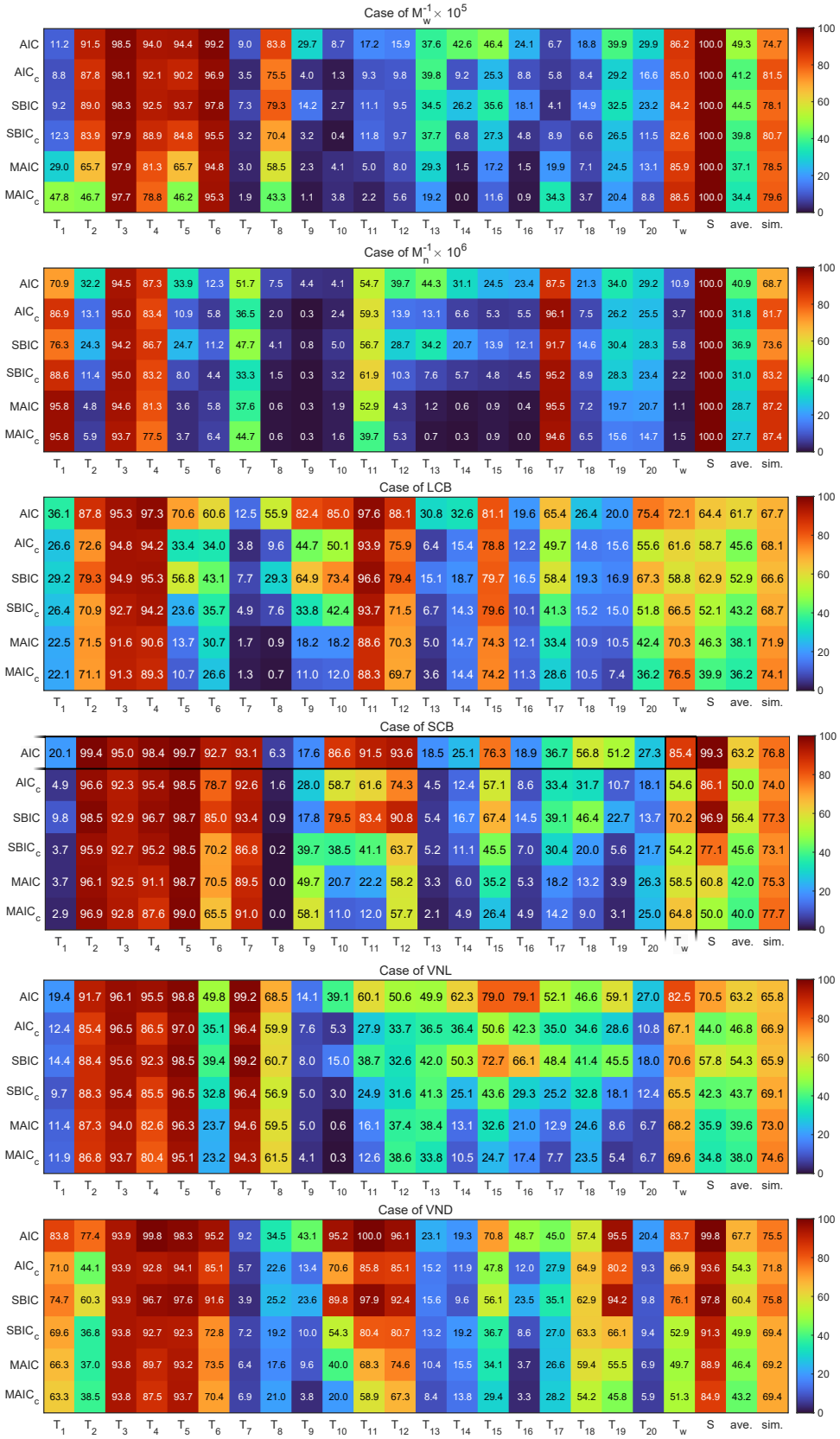

Figure 6: Frequency of selection for each variable

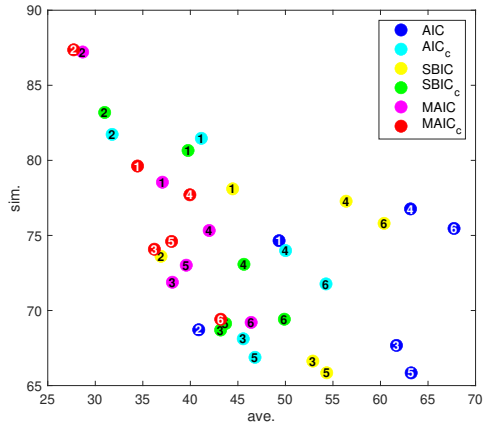

Figure 7: Scatterplot of ave. and sim.

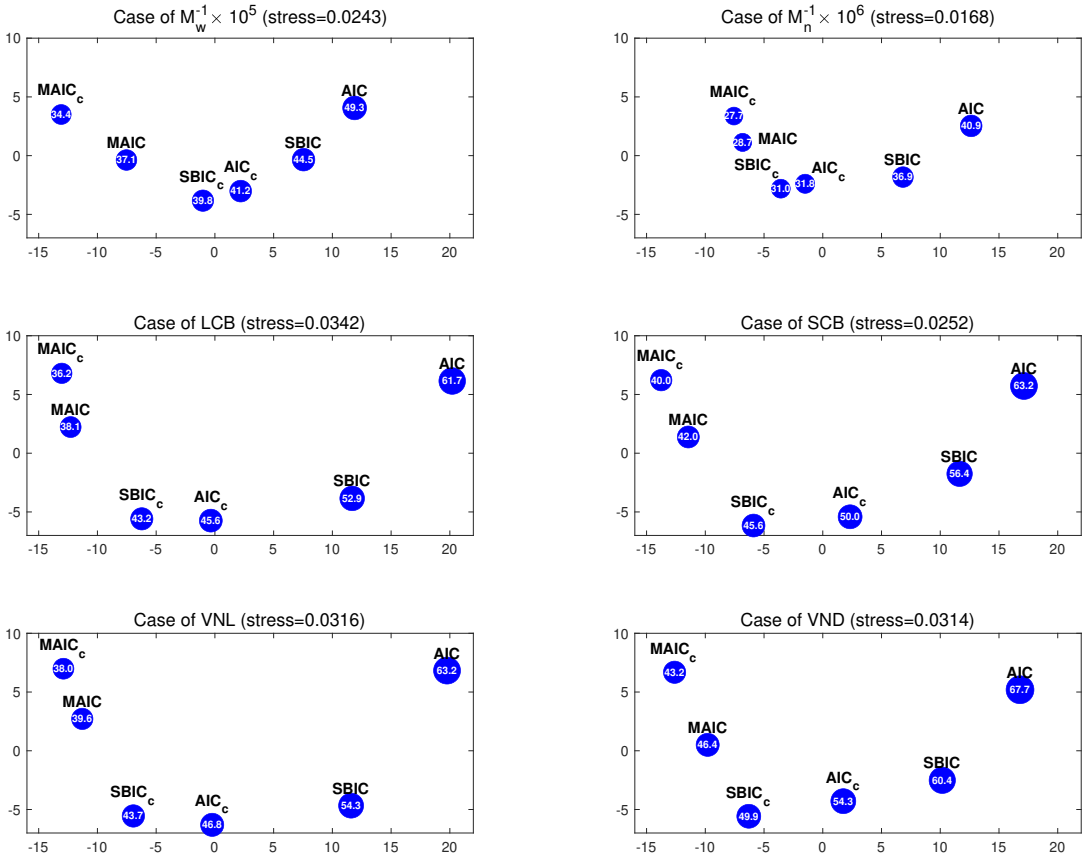

Figure 8: Two-dimensional MDS representation of six information criteria

where $\boldsymbol{Y}$ is an $n \times p$ random matrix of observations, $\boldsymbol{X}$ is an $n \times k$ matrix of explanatory variables, and $\boldsymbol{B}$ is a $k \times p$ matrix of regression coefficients. In order to derive an unbiased estimator of the risk based on the KL divergence, we need to estimate the expectation of the inverse matrix of some noncentral Wishart matrix. If the rank of noncentrality matrix is one, we will be able to use a method similar to the one in the present paper. However, we need to check whether the resulting estimator of the risk still works as well as MAIC does when the rank is greater than one. It is difficult to derive an unbiased estimator in the case that the rank of the noncentrality matrix is greater than one. These problems are left as future work.

Appendix

In order to prove Lemma 3.1, we need two lemmas.

Lemma A.1. *Let $a > b > 0$ and $0 \le x \le 1$. Then the following relations hold:*

(1)

$$\sum_{r=1}^{\infty} \frac{(b)_{r-1}}{(a)_r} x^r = x \int_0^1 \frac{(1-t)^{a-1}}{(1-tx)^b} dt$$

(2) *For any positive integer k,*

$$\sum_{r=k+1}^{\infty} \frac{(b)_{r-1}}{(a)_r} x^r = \frac{(b)_k x^{k+1}}{(a)_k} \int_0^1 (1-t)^{a+k-1} (1-tx)^{-(b+k)} dt \le \frac{(b)_k x^{k+1}}{(a)_k (a-b)}$$

(3)

$$\frac{1}{a-b} = \sum_{r=1}^{\infty} \frac{(b)_{r-1}}{(a)_r}$$

Proof. The following formula is obtained by integration by parts:

$$\begin{aligned} & x \int_0^1 (1-t)^{(a+k)-1} (1-tx)^{-(b+k)} dt \\ & \quad = \frac{x}{a+k} + \frac{(b+k)x^2}{a+k} \int_0^1 (1-t)^{(a+k+1)-1} (1-tx)^{-(b+k+1)} dt. \end{aligned}$$

By applying this formula repeatedly for $k = 0, 1, \ldots$, we obtain

$$x \int_0^1 \frac{(1-t)^{a-1}}{(1-tx)^b} dt = \sum_{r=1}^{k} \frac{(b)_{r-1}}{(a)_r} x^r + \frac{(b)_k x^{k+1}}{(a)_k} \int_0^1 (1-t)^{a+k-1} (1-tx)^{-(b+k)} dt$$

Since

$$\int_0^1 (1-t)^{a+k-1} (1-tx)^{-(b+k)} dt \le \int_0^1 (1-t)^{a-b-1} dt = \frac{1}{a-b},$$

to prove (1) and (2), it suffices to prove $\lim_{k\to\infty} (b)_k/(a)_k = 0$, which is shown as follows:

$$\log \frac{(b)_k}{(a)_k} = \sum_{r=0}^{k-1} \log\left(1 - \frac{a-b}{a+r}\right) < -\sum_{r=0}^{k-1} \frac{a-b}{a+r} \to -\infty \ (k \to \infty),$$

since $\log(1-x) < -x$ if $0 < x < 1$. Substituting $x = 1$ into (1), we obtain (3). □

Lemma A.2. *Let K be a random variable distributed as* $\mathrm{Po}(\lambda)$*, the Poisson distribution with mean λ, and g be an arbitrary function such that $E[g(K)]$ exists. Then*

$$E[\lambda g(K)] = E[K g(K-1)].$$

Proof.

$$\begin{aligned} \mathrm{E}[\lambda g(K)] &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} g(k) \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} (k+1) g(k) = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} k\, g(k-1) \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} k\, g(k-1) = \mathrm{E}[K g(K-1)]. \end{aligned}$$

□

Proof of Lemma 3.1. Let $W_j = n\hat{\sigma}_j^2/\sigma_*^2$ for $j \subset \omega$. Additionally, let

$$V_j = W_j - W_\omega = \frac{\boldsymbol{y}^\mathsf{t}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{y} - \boldsymbol{y}^\mathsf{t}(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{y}}{\sigma_*^2} = \frac{\boldsymbol{y}^\mathsf{t}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{y}}{\sigma_*^2}.$$

Then V_j and W_ω are independent, since

$$\mathrm{Cov}[(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{y}, (\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{y}] = \sigma_*^2(\boldsymbol{P}_\omega - \boldsymbol{P}_j)(\boldsymbol{I}_n - \boldsymbol{P}_\omega) = \boldsymbol{O}_n.$$

It can be easily seen that V_j is distributed as $\chi_{k-k_j}(\theta_j^*)$, the noncentral chi-square distribution with degree of freedom $k-k_j$ and noncentrality parameter θ_j^*, since $(\boldsymbol{P}_\omega - \boldsymbol{P}_j)$ is idempotent and its rank is $k - k_j$. The assumption $\mathcal{J}_+ \neq \emptyset$ gives that $\theta_\omega^* = 0$ and W_ω is distributed as χ_{n-k}^2.

Let $(K, \tilde{V}_j)$ be a pair of random variables independent of W_ω. Suppose that K is distributed as $\mathrm{Po}(\theta_j^*/2)$, and the conditional distribution of $\tilde{V}_j$ given $K = \ell$ is $\chi_{k-k_j+2\ell}^2$. Then the distribution of $\tilde{V}_j$ is the same as that of V_j, that is, $\chi_{k-k_j}^2(\theta_j^*)$ (see, e.g., [9]). Then we have

$$\mathrm{E}\left[\frac{\sigma_*^2}{n\hat{\sigma}_j^2}\right] = \mathrm{E}\left[\mathrm{E}\left[\left(W_\omega + \tilde{V}_j\right)^{-1} \,\middle|\, K\right]\right] = \mathrm{E}\left[\frac{1}{n - k_j + 2K - 2}\right],$$

$$\mathrm{E}\left[\hat{\gamma}_j^r\right] = \mathrm{E}\left[\mathrm{E}\left[\left(\frac{W_\omega}{W_\omega + \tilde{V}_j}\right)^r \,\middle|\, K\right]\right] = \mathrm{E}\left[\frac{\left(\frac{n-k}{2}\right)_r}{\left(\frac{n-k_j+2K}{2}\right)_r}\right].$$

Here we used the fact that the conditional distribution of $W_\omega/(W_\omega + \tilde{V}_j)$ given $K = \ell$ is the beta distribution with parameters $(n-k)/2$ and $(n - k_j + 2\ell)/2$.

Hence, the first equation in Lemma 3.1. can be rewritten as

$$\mathrm{E}\left[\frac{1}{2}\sum_{r=1}^{\infty}\frac{(r-1)!}{\left(\frac{n-k}{2}\right)_r}\hat{\gamma}_j^r\right] = \mathrm{E}\left[\frac{1}{2}\sum_{r=1}^{\infty}\frac{(1)_{r-1}}{\left(\frac{n-k_j+2K}{2}\right)_r}\right] = \mathrm{E}\left[\frac{1}{2}\frac{1}{\left(\frac{n-k_j+2K}{2} - 1\right)}\right],$$

where the last equality was obtained using Lemma A.1.

From Lemma A.2, we find

$$\mathrm{E}\left[\frac{\theta_j\sigma_*^2}{n\hat{\sigma}_j^2}\right] = \mathrm{E}\left[\frac{\frac{\theta_j}{2}}{\frac{n-k_j+2K}{2} - 1}\right] = \mathrm{E}\left[\frac{K}{\frac{n-k_j+2(K-1)}{2} - 1}\right] = 1 - \mathrm{E}\left[\frac{\frac{n-k_j-4}{2}}{\frac{n-k_j+2K}{2} - 2}\right].$$

Using Lemma A.1, we obtain

$$\mathrm{E}\left[\sum_{r=1}^{\infty}\frac{r!}{\left(\frac{n-k}{2}\right)_r}\hat{\gamma}_j^r\right] = \mathrm{E}\left[\sum_{r=1}^{\infty}\frac{(2)_{r-1}}{\left(\frac{n-k_j+2K}{2}\right)_r}\right] = \mathrm{E}\left[\frac{1}{\frac{n-k_j+2K}{2} - 2}\right].$$

These two formulas prove the second equation in Lemma 3.1. □

Remark A.1. *The power series in* Lemma A.1 *can be written as*

$$\sum_{r=1}^{\infty}\frac{(b)_{r-1}}{(a)_r}x^r = \frac{x}{a}\sum_{r=1}^{\infty}\frac{(b)_{r-1}}{(a+1)_{r-1}}x^{r-1} = \frac{x}{a}{}_2F_1(b, 1; a+1; x),$$

where ${}_2F_1(a, b; c; x) = \sum_{r=0}^{\infty}\frac{(a)_r(b)_r}{(c)_r}x^r$ *is a hypergeometric function. One can easily calculate* MAIC_c *for actual data using numerical software that handles hypergeometric functions.*

Acknowledgements

The second author's research was partially supported by JSPS KAKENHI Grant Number 23H00809. The authors thank FORTE Science Communications (https://www.forte-science.co.jp/) for English language editing.

References

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

[2] Davies, S. J., Neath, A. A. & Cavanaugh, J. E. (2006). Estimation optimality of corrected AIC and modified C_p in linear regression model. *Int. Stat. Rev.*, **74**, 161–168.

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

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

[5] Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

[6] Hurvich, C. M. & Tsai, C. L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, **78**, 499–509.

[7] Ishiguro, M., Sakamoto, Y. & Kitagawa, G. (1997). Bootstrapping log likelihood AIC. *Ann. Inst. Statist. Math.*, **49**, 411–434.

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

[9] Muirhead, R. J. (1986). *Aspects of Multivariate Statistical Theory*, New York: Wiley.

[10] Noda, K., Miyaoka, E. & Itoh, M. (1996). On bias correction of the Akaike information criterion in linear models. *Commun. Statist. Theory Methods*, **25**, 1845–1857.

[11] Reschenhofer, E. (1999). Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory*, **15**, 377–387.

[12] Skagerberg, B., Macgregor, J. F. & Kiparissides, C. (1992). Multivariate data analysis applied to low-density polyethylene reactors. *Chemom. Intell. Lab. Syst.*, **14**, 341–356.

[13] Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46**, 1273–1291.

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

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

[16] Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Math. Sci.*, **153**, 12–18 (in Japanese).

[17] 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.

[18] Yanagihara, H. & Ohmoto (2005). On distribution of AIC in linear regression models. *J. Stat. Plan. Infer.*, **133**, 417–433.

[19] Yanagihara, H., Kamo, K., Imori, S. & Yamamura, M. (2017). A study on the bias-correction effect of the AIC for selecting variables in normal multivariate linear regression models under model misspecification. *REVSTAT-Stat. J.*, **15**, 299–332.