ON AIC’S CORRECTED FOR ESTIMATING KULLBACK-LEIBLER INFOMATION FOR LINEAR MODEL SELECTION

Rinya Takahashi, Kazuo Noda, Jinglong Wang, Masashi Itoh

1 Kobe Univ. of Mercantile Marine, Fukue-Minami, Higashi-Nada-ku, Kobe 658-0022, Japan
2 Faculty of Sci. and Tech., Meisei University, 2-1-1, Hodokubo, Hino-shi, Tokyo 191 8506, Japan
3 Dept. of Statist., East China Normal Univ., 3663 Zhongshan Rd. N., Shanghai 200062, China
4 Overseas Environmental Cooperation Center, 3-1-8, Shibakoen, Minato-ku, Tokyo, 105-0011, Japan

1. Introduction

We consider a linear model $M_0 : Y = (Y_1, Y_2, \ldots, Y_n)' \sim N_n(X\beta, \sigma^2 I_n)$, where $X$ is an $n \times (k+1)$ design matrix having $1 = (1, 1, \ldots, 1)'$ as the first column with rank $(X) = k + 1$, $I_n$ is the identity matrix of order $n$, and $(\beta', \sigma^2)'$ is the unknown parameter vector in a parameter space $\Theta_0 = \{ \theta = (\beta', \sigma^2)' \in \mathbb{R}^{k+1} \times (0, \infty), \beta' = (\beta_0, \beta_1, \ldots, \beta_k)' \}$. In the model $M_0$, we consider the problem of selecting the optimum one from the submodels $M_i$, $i = 1, 2, \ldots, 7$, specified by subsets $\Theta_i$ of $\Theta$ in the following way. As given constant vectors, we first write

$$P_0 = (\beta_0, 0, 0, \ldots, 0)' \quad \gamma_1 = (\beta_0, \beta_1, \ldots, \beta_k)' \quad \gamma_2 = (\beta_{1+}, \beta_{2+}, \ldots, \beta_k)'$$

Also, letting

$$C_1 = (I_{k+1}, O)_{(k+1) \times (k+1)}, \quad C_2 = (O, I_{k-s})_{(k-s) \times (k+1)}$$

and $\sigma_0^2$ be a given positive constant, we set restricted parameter spaces $\Theta_i$, $i = 1, 2, \ldots, 7$, as

$$\Theta_1 = \{ \theta | \theta \in \mathbb{R}^{k+1} \times (0, \infty), C_1\beta = \gamma_1 \}, \quad \Theta_2 = \{ \theta | \theta \in \mathbb{R}^{k+1} \times (0, \infty), C_2\beta = \gamma_2 \},$$

$$\Theta_3 = \{ \theta | \beta = \beta_0, \sigma^2 \in (0, \infty) \}, \quad \Theta_4 = \{ \theta | \beta \in \mathbb{R}^{k+1}, \sigma^2 = \sigma_0^2 \},$$

$$\Theta_5 = \{ \theta | \beta \in \mathbb{R}^{k+1}, C_1\beta = \gamma_1, \sigma^2 = \sigma_0^2 \}, \quad \Theta_6 = \{ \theta | \beta \in \mathbb{R}^{k+1}, C_2\beta = \gamma_2, \sigma^2 = \sigma_0^2 \},$$

$$\Theta_7 = \{ \theta | \theta = (\beta_0', \sigma_0^2) \}.$$

Let $f_Y(y, \theta) = \prod_{i=1}^n f_{Y_i}(y_i, \theta)$ be the probability density of $Y$. Omitting the constant term, we express the Kullback-Leibler information in favor of the true value $\theta^+ = (\beta^{+'}, \sigma^{+2})' \in \Theta$ against $\theta_i \in \Theta_i$ as $l(\theta^+, \theta_i) = -2 \int_{\mathbb{R}^n} f_Y(y, \theta) \log f_Y(y, \theta_i) dy$.

We can consider each AIC$_i$ in $M_i$, the Akaike information criterion by Akaike (1973), as an estimator of $l(\theta^+, \theta_i)$. Setting a squared loss function of information criterion, IC$_i$,

$$L(\theta^+, IC_i) = |IC_i - E_{\theta^+} [l(\theta^+, \theta_i)]|^2,$$

we here find the risks of AIC$_i$ and the bias corrected AIC$_i$, AIC$_i^{bc}$, under the same situation as in Noda et al. (1996), where $\hat{\theta}_i$ denotes a maximum likelihood estimator (MLE) in $\Theta_i$. We hence compare the risks of AIC$_i$ with those of AIC$_i^{bc}$, considering their variances and squared biases.

2. The risks of AIC$_i$ and AIC$_i^{bc}$

The risk function of IC$_i$ corresponding to the loss function $L$ is expressed as

$$R(\theta^+, IC_i) = E_{\theta^+} [L(\theta^+, IC_i)] = Var_{\theta^+} (IC_i) + |B(\theta^+, IC_i)|^2,$$

where $B(\theta^+, IC_i)$ denotes the bias of IC$_i$. 
In case of $i = 1, 2$, we adopt the following $\text{IC}_i$ as $\text{AIC}_i^{bc}$, neglecting the term related to $\hat{\delta}_i$, an unbiased estimator of the noncentrality $\delta_i$, in Noda et al. (1996) for the sake of respecting their variances:

$$\text{AIC}_1^{bc} = n \log(2\pi \hat{\sigma}_1^2) + \frac{n(n + s + 1)}{n - s - 3}, \quad \text{AIC}_2^{bc} = n \log(2\pi \hat{\sigma}_2^2) + \frac{n(n + k - s)}{n - k + s - 2}.$$ 

The risk of $\text{IC}_1$ is expressed through its variance and bias as follows.

$$\text{Var}_{\theta^+}(\text{IC}_1) = \text{Var}_{\theta^+}(\text{IC}_2^{bc}) = n^2 \text{Var}_{\theta^+}(\log \hat{\sigma}_1^2)$$

$$= n^2 \left\{ \sum_{j=0}^{\infty} \frac{(\delta_1/2)^j}{j!} e^{-\delta_1/2} \left[ \psi\left(\frac{n - s - 1 + 2j}{2}\right) + \log 2 - \log(n - s - 1) \right]^2 \right\}$$

$$- 2 \left[ \log \left( \frac{n}{(n - s - 1)\sigma^2} \right) \right] \left[ \sum_{j=0}^{\infty} \frac{(\delta_1/2)^j}{j!} e^{-\delta_1/2} \left\{ \psi\left(\frac{n - s - 1 + 2j}{2}\right) + \log 2 - \log(n - s - 1) \right\} \right]$$

$$- \log \left( \frac{n}{(n - s - 1)\sigma^2} \right)^2$$

$$\left[ \sum_{j=0}^{\infty} \frac{(\delta_1/2)^j}{j!} e^{-\delta_1/2} \left\{ \psi\left(\frac{n - s - 1 + 2j}{2}\right) + \log 2 - \log(n - s - 1) \right\} \right] - \log \left( \frac{n}{(n - s - 1)\sigma^2} \right)^2 \right\},$$

$$B(\theta^+; \text{IC}_1) = - \frac{2(s + 2)(s + 3)}{n - s - 3} + \frac{2n(s + 1)}{(n - s - 3)(n - s - 1)} \delta_1 + o\left(\frac{\delta_1}{n}\right),$$

$$B(\theta^+; \text{IC}_1^{bc}) = - \frac{2n(s + 1)}{(n - s - 3)(n - s - 1)} \delta_1 + o\left(\frac{\delta_1}{n}\right),$$

where $\psi$ denotes the digamma function. As a result, in case of $i = 1$, the risk of $\text{AIC}_1^{bc}$ is smaller than that of $\text{AIC}_i$, if the sample size $n$ is sufficiently large.

The risks of $\text{IC}_i$ in case of $i = 2, 3$ are similarly obtained and evaluated. In case of $i = 4, 5, 6$, the risks of $\text{IC}_i$ depend on $s$ and $k$ so that their evaluations can not be decided without the constraints of the ranges of $s$ and $k$. Roughly speaking, if $s$ and $k$ are sufficiently small to $n$, then each risk of $\text{AIC}_i$ is smaller than that of $\text{AIC}_i^{bc}$. In case of $i = 7$, the risk of $\text{AIC}_i$ is the same as that of $\text{AIC}_i^{bc}$. Hence the risk improvements of $\text{AIC}_i$, $i = 4, 5, 6$, should be considered as those different from $\text{AIC}_i^{bc}$.

REFERENCES


RÉSUMÉ

Des risques de $\text{AIC}$s et $\text{AIC}$s du biais corrigés qui estiment la information de Kullback-Leibler sur la sélection des modèles linéaires sont obtenus et mutuellement comparés sans la condition que la valeur vraie du paramètre est dans le modèle sélectionné. La fonction de risque est ici construite comme la attente de la fonction perte carrée de $\text{IC}$. Ces résultats obtenus sont illustrés par des études de certaines sortes de simulations (qui seront donnés dans la présentation).