On Error Rates of Statistical Model Selection based on Information Criteria

Masayuki GOTOH [1] Toshiyasu MATSUSHIMA [1] Shigeichi HIRASAWA [1]

Abstract — **In this paper, we shall derive the upper bounds on error rates of the statistical model selection using the information criteria, P_e^*. The similar bounds were derived by J.Suzuki [2]. We shall generalize the results for the general model class.**

I. Introduction

Statistical model selection is one of the most important problems in statistics, and many works have left important results behind this field. The conventional model selection criteria (information criteria), such as the Akaike information criterion(AIC), the Bayesian information criterion(BIC), and the minimum description length(MDL) were derived from the different view points one another. The form of the conventional information criterion is given by

$$IC(m) = \log p(x^n|m, \hat{\theta}^{k_m}) - \frac{k_m}{2}c(n), \tag{1}$$

where $x^n = x_1 x_2 \cdots x_n$ is the data sequence with length n emitted from the source, m is denoting a model with k_m-dimensional parameter θ^{k_m}, $\log p(x^n|m, \theta^{k_m})$ is the logarithm likelihood of the model m, $\hat{\theta}^{k_m}$ is the maximum likelihood estimator and $c(n)$ is a penalty term. Here, $x_i \in \mathcal{X}$ for $\forall i$, $\mathcal{X}$ is the sample space, and X is the random variable on $\mathcal{X}$.

The estimator of m is given by maximization of $IC(m)$.

For the consistency of the model selection, which is the property such as the true model can be asymptotically selected, several results have been reported. We interest here more detail valuation for the selection of the true model. Then we define the error probability for the selection of the true model m^*, P_e^* , which is given by

$$P_e^* = P^*\left\{IC(m^*) \le IC(m) \middle| \exists m \ne m^*\right\}, \tag{2}$$

where $P^*\{\cdot\}$ is the probability mass measured by the true distribution $p^*(\cdot)$ from which the data sequence is emitted.

In this paper, we shall derive the upper bounds on error rates of the statistical model selection using the information criteria, P_e^*. The similar bounds were derived by J.Suzuki [2]. This results are, however, for the model class with tree structure, and which have in part the analytical inadequacies. We shall generalize this result for the general model class.

II. Preliminary Definition and Assumption

Definition 1 (the hierarchical (nested) model class) *Let m be a discrete label (index) of model in the discrete and finite model class $\mathcal{M}$, that is, $m \in \mathcal{M}$ and $|\mathcal{M}|$ is bounded. Each model has k_m-dimensional parameter $\theta^{k_m} = (\theta_1, \theta_2, \cdots, \theta_{k_m})^T$ in parameter set $\Theta^{k_m} \subseteq \mathcal{R}^{k_m}$, hence m specifies a parametric model class. Let $\mathcal{H}^{k_m}$ be this model class of m, which is given by $\mathcal{H}^{k_m} = \{p(\cdot) = p(\cdot|m, \theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}\}$, where $p(\cdot)$ represents the probability or density function. Then the hierarchical (nested) model class $\mathcal{H}$ is defined by $\mathcal{H} = \cup_{m\in\mathcal{M}}\mathcal{H}^{k_m}$, where the nested structure $\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots\cdots$ may be satisfied for $m1, m2, \cdots \in \mathcal{M}$ and $k_{m1} < k_{m2} < \cdots$. The data sequence x^n is derived from the distribution $p^*(x^n) = p(x^n|m^*, \theta^{k_{m^*}}_*)$, where m^* is the true model and $\theta^{k_{m^*}}_*$ is the true parameter.* □

[1] Dept. of Industrial and Management Systems Engineering School of Science and Engineering, Waseda University 3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan Email goto@hirasa.mgmt.waseda.ac.jp

Remark the fact such that if $\mathcal{H}^{k_{m^*}} \ne \mathcal{H}$, then there exist m and θ^{k_m} satisfying $m \ne m^*$ and $p(x^n|m, \theta^{k_m}) = p(x^n|m^*, \theta^{k_{m^*}}_*)$. Then, we usually define the true model m^* as follows:

$$m^* = \arg_m \min\left\{k_m \middle| \exists\theta^{k_m}, \forall x^n, p(x^n|m, \theta^{k_m}) = p^*(x^n)\right\}. \tag{3}$$

Suzuki analyzed the error rate of the model selection for the some tree model class [2]. This result is very interesting. However, this is restricted for the tree model class which has the discrete sample space $\mathcal{X}$ with the multinomial distribution. We shall generalize this result in this paper.

III. Main Results

Let $P_e^*[m] = P^*\left[IC(m^*) \le IC(m)\right]$ be the (error) probability of the event such that the model m is selected. At first, we show the following theorem.

Theorem 1 *On some suitable condition, if $\mathcal{H}^{k_m} \subset \mathcal{H}^{k_{m^*}}$ and $\lim_{n\to\infty}\sup \frac{c(n)}{n} = 0$, then the following inequation is satisfied.*

$$P_e^*[m] \le O\left(\frac{\log\log n}{n}\right) \tag{4}$$

Theorem 2 *On some suitable condition, if $\mathcal{H}^{k_{m^*}} \subset \mathcal{H}^{k_m}$, then the following inequation is satisfied.*

$$P_e^*[m] \le O\left(\frac{1}{(k_{m^*} - k_m)^2\left(1 - c(n)\right)^2}\right) \tag{5}$$

From Theorem 1 and 2, we have the following theorem using union bounds.

Theorem 3 *On some suitable condition, the error rate of selection of true model, P_e^*, is given as follows: If $\lim_{n\to\infty}\sup \frac{c(n)}{n} = 0$ and $\lim_{n\to\infty}\inf\left(\frac{n}{\log\log n}\right)^{1/2} c(n) < \infty$, then*

$$P_e^* \le O\left(\frac{1}{\left(1 - c(n)\right)^2}\right), \tag{6}$$

else if $\lim_{n\to\infty}\sup \frac{c(n)}{n} = 0$ and $\lim_{n\to\infty}\inf\left(\frac{n}{\log\log n}\right)^{1/2} c(n) = \infty$, then

$$P_e^* \le O\left(\frac{\log\log n}{n}\right). \tag{7}$$

IV. Conclusion

We have shown the asymptotic order of error rate of the model selection based on the information criteria. Similar results for the model selection based on Bayes decision theory can be derived [1].

References

[1] M.Gotoh, T.Matsushima, and S.Hirasawa: "On Statistical Model Selection based on Bayes Decision Theory (in Japanese)", IEICE Technical Report IT97-21, (1997)

[2] J.Suzuki: "Evaluations for Estimation of an Information Source Based on State Decomposition", IEICE Trans. Fundamentals, Vol.E76-A, No.7, pp.1240-1251, (1993)