モデル選択結果の漸近分布

大阪大学　大学院基礎工学研究科　　伊森晋平
札幌医科大学　医療人育成センター　　加茂憲一

AIC (Akaike, 1974) や BIC (Schwartz, 1978) で知られる情報量規準によるモデル選択を考える. モデル選択結果は確率変数であり, その不確実性を評価することで, モデル選択結果の信頼性を検証することが可能になる. すでに, Shibata (1976) や Nishii (1984) により, 情報量規準によるモデル選択結果の漸近的な性質が与えられている. しかしながら, 実際のデータは有限であるため, 有限標本での不確実性を評価することは肝要である. 本報告では, 真のモデルの選択確率を近似し, 有限標本でのモデル選択結果の妥当性を検証する手法を提案する.

観測されたデータを $y_1, \ldots, y_n \overset{i.i.d.}{\sim} f(y; \boldsymbol{\theta}_*)$ とし, $\boldsymbol{\theta}_* = (\theta_1^*, \ldots, \theta_p^*)$ を未知パラメータとする. 候補のモデル M_k $(k = 1, \ldots, p)$ の下で $\boldsymbol{\theta}_* = \boldsymbol{\theta}_*^{(k)} = (\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k)$ を仮定する. ただし, $\boldsymbol{\theta}_k^* = (\theta_1^*, \ldots, \theta_k^*)$ であり, $\boldsymbol{\tau}_k = (\tau_{k+1}, \ldots, \tau_p)$ は既知とする. さらに, 真のモデルを M_q とする. このとき, モデル M_k における情報量規準を

$$\mathrm{IC}(M_k; c_n) = -2 \sum_{i=1}^{n} \log f(y_i; \hat{\boldsymbol{\theta}}_k) + k c_n$$

と表現する. ただし, c_n は n に関する数列であり,

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}_k}{\mathrm{argmax}} \prod_{i=1}^{n} f(y; \boldsymbol{\theta}^{(k)}),$$

$\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}_k, \boldsymbol{\tau}_k)$ である. 本報告では, Imori, Katayama & Wakaki (2014) で提案された選択手法の拡張である以下の手法を用いて $Pr\{\hat{k}(c_n) = q\}$ の近似式を導く.

$$\hat{k}(c_n) = \max_{1 \le k < p} \{k I(\mathrm{IC}(M_k; c_n) > \mathrm{IC}(M_p; c_n))\} + 1.$$

参考文献

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Imori, S., Katayama, S. and Wakaki, H. (2014). Screening and selection methods in high-dimensional linear regression model. TR 14-01, Statistical Research Group, Hiroshima University, Hiroshima.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

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

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.