不完全観測における完全データの情報量規準

川崎重工業　技術開発本部　システム技術開発センター　　前田晴義
大阪大学　大学院基礎工学研究科　　伊森晋平
大阪大学　大学院基礎工学研究科　　下平英寿

本報告では, データの一部が観測できない不完全データに対するモデル選択問題を考える. Akaike (1974) により提案された AIC は最もよく用いられるモデル選択手法の一つであるが, AIC は観測部分のみに着目しており, 非観測データへの当てはまりは考慮していない. 一方で, Shimodaira (1994) では, 非観測部分を含めた完全データに対する情報量規準 PDIO が提案された. しかしながら, PDIO の導出には強い仮定が必要であり, 実データへの利用は限定的である. そこで, より弱い仮定の下で利用可能な情報量規準を提案する.

完全データを $X = (x_1, \ldots, x_n)$ とし, 各 $i = 1, \ldots, n$ に対して, $x_i = (y_i, z_i) \overset{i.i.d.}{\sim} p_X(x; \theta)$ を仮定する. ただし, θ は d 次元未知パラメータである. また, 観測データは $Y = (y_1, \ldots, y_n)$, 非観測データは $Z = (z_1, \ldots, z_n)$ とする. このとき, 提案する情報量規準は次式で与えられる:

$$-\sum_{i=1}^{n} \log p_Y(y_i; \hat{\theta}_Y) + \frac{1}{2}\mathrm{tr}(I_X I_Y^{-1}) + \frac{1}{2}d.$$

ただし, $x = (y, z)$ に対し, $p_Y(y; \theta) = \int p_X(x; \theta) dz$ であり, $\hat{\theta}_Y$ は観測データに基づく推定量

$$\hat{\theta}_Y = \underset{\theta}{\mathrm{argmax}} \sum_{i=1}^{n} \log p_Y(y_i; \theta),$$

I_X, I_Y はそれぞれ以下の行列である.

$$I_X = -\int p_X(x; \hat{\theta}_Y) \left. \frac{\partial^2 \log p_X(x; \theta)}{\partial\theta\partial\theta^T} \right|_{\theta=\hat{\theta}_Y} dx,$$
$$I_Y = -\int p_Y(y; \hat{\theta}_Y) \left. \frac{\partial^2 \log p_Y(y; \theta)}{\partial\theta\partial\theta^T} \right|_{\theta=\hat{\theta}_Y} dy.$$

先行研究との比較は当日報告する.

参考文献

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

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. *Selecting Models from Data: AI and Statistics IV* (eds. P. Cheeseman and R. W. Oldford), **89**, 21–30.