ダイバージェンスを用いたロバスト推定とモデル選択規準について

大阪大学大学院 基礎工学研究科　倉田 澄人
大阪大学大学院 基礎工学研究科　濱田 悦生

統計的ダイバージェンスに基づいたモデル選択規準には, 広く用いられている AIC や BIC 等がある. これらの規準, 及び最尤推定法は Kullback Leibler divergence (以降, KL-divergence) を基盤に理論が構築されているが, データを発生させる “真の分布” と統計モデルとの “遠さ” を測るダイバージェンスは KL-divergence に限ったものではない.

二つの確率分布 G, F を測るダイバージェンスの一つに, Basu, Harris, Hjort and Jones (1998) [1] により提案されたダイバージェンス (以降, BHHJ-divergence) がある:

$$d_\alpha(G\,;\,F) \;=\; \int \left\{ f(y)^{\alpha+1} - \left(1+\frac{1}{\alpha}\right) f(y)^\alpha g(y) + \frac{1}{\alpha}\, g(y)^{\alpha+1} \right\} dy \quad (\alpha > 0)\,.$$

これに対する最小ダイバージェンス推定 (以降, BHHJ-MDIVE) は, パラメータ α を 0 に近づけると最尤推定と一致し, またある程度大きく取ると, 外れ値の影響を抑えたロバスト推定と捉えることが可能である.

BHHJ-divergence 及び BHHJ-MDIVE に基づいたモデル選択規準の構築を目指した研究としては, 例えば Mattheou *et al.* (2009) [2] による $\mathrm{DIC}^{\mathrm{BHHJ}}$ があるが, これは各データに対して独立同分布性を想定しているという制約がある.

説明変数 $\boldsymbol{x}_i$ によって目的変数 Y_i の説明を試みる一般的な線形回帰モデルであれば,

$$Y_i \;=\; \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i \;\; \left(i \in \{1\,,\,\ldots\,,\,n\}\right), \quad \epsilon_1\,,\,\ldots\,,\,\epsilon_n \overset{i.i.d.}{\sim} N(0\,,\,s)$$

という分布構造が仮定され, $Y_1\,,\,\ldots\,,\,Y_n$ は独立ではあるが, 各分布は $Y_i \;\sim\; N\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\,,\,s\right)$, つまり同分布ではない. このような状況下では, $\mathrm{DIC}^{\mathrm{BHHJ}}$ を利用する為の前提が満足されず, シミュレーションを行っても良い選択精度を示さない場合が多いことが分かった.

そこで, 各データの分布が必ずしも同一でない場合の BHHJ-divergence 及び推定量の漸近分布に関する研究 (Ghosh and Basu (2013) [3]) に基づき, より理論的妥当性を有したモデル選択規準を導出する. そしてシミュレーションを通じてその性能を確認するとともに, 規準の更なる改良や拡張の可能性について考察を行う.

参考文献

[1] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and Efficient Estimation by Minimising a Density Power Divergence, *Biometrika*, **85** (3), 549-559.

[2] Mattheou, K., Lee, S., Karagrigoriou, A. (2009). A Model Selection Criterion Based on the BHHJ Measure of Divergence, *Journal of Statistical Planning and Inference*, **139**, 228-235.

[3] Ghosh, A. and Basu, A. (2013). Robust Estimation for Independent Non-Homogeneous Observations Using Density Power Divergence with Applications to Linear Regression, *Electronic Journal of Statistics*, **7**, 2420-2456.