統計的因果推論と頑健なモデル選択について

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

本発表では, 有向非巡回グラフ (DAG) によって表現された因果モデルに対するパスの選択, 即ち妥当な因果構造をデータから判断する手法の頑健性について報告する.

m 個の変数 $X_1, \ldots, X_m$ から成る組について, 正規分布を想定した DAG モデルは,

$$X_j \;=\; \mu_j + \sum_{k<j} a_{jk}\, X_k + \epsilon_j\,, \quad \epsilon_j \overset{indep.}{\sim} N(0\,,\, s_j) \qquad (j \;=\; 1\,, \ldots, m)$$

と表現される.

ここで, 現実的には, データの中に平均的挙動からかなり逸脱したものが一部入っていたり, 人為的又は機械的な誤りによって間違ったデータが残されてしまったりと, 様々な異常が観測値の中に存在し得るという問題があり, 母数推定やモデル選択に際しては, そのような異常値に対する頑健性が求められる.

頑健性を持った統計的手法の一つには, Ghosh and Basu (2013) [1] によるダイバージェンスの利用が挙げられる. これは, KL-divergence に基づいた手法と比較して, 外れ値の影響を低減するように調整されたダイバージェンスであり, 統計的仮説検定や多項式回帰モデルの最大次数選択等に応用されている.

本発表では, この頑健性を持ったダイバージェンスに基づいたパス係数の推定及び因果構造の選択手法について, その性能を他の手法と対比しつつ論じる. 特に, 多様な「異常」の入り方を想定して行った幾つかの数値実験の中で, 頑健性を重視したダイバージェンスに基づいた評価規準 (Kurata and Hamada (2017) [2]) は, 条件付独立性をグラフ全体で同時に検定する手法 (Shipley (2013) [3]) が対応出来ない設定や, KL-divergence に基づいた情報量規準が異常値によって選択精度を落としてしまう場面でも, 異常値が混入していない場合に比較的近い精度を発揮する傾向にあることを示す.

参考文献

[1] 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.

[2] Kurata, S. and Hamada, E. (2017). A robust generalization and asymptotic properties of the model selection criterion family, *Communications in Statistics-Theory and Methods* (to be appeared).

[3] Shipley, B. (2013). The AIC model selection method applied to path analytic models compared using a d-separation test, *Ecology*, **94**, 560–564.