

HSIC に基づく非線形モデルの変数選択

滋賀大学データサイエンス教育研究センター 紅林 亘

滋賀大学データサイエンス学部 松井 秀俊

本研究では、複雑な関係性をもつデータから有効な情報を抽出する方法として、非線形モデルにおける変数選択手法を提案する。HSIC Lasso [1]は非線形回帰問題のための変数選択手法であり、指定した個数の変数を順位とともに選択する。この手法は、HSIC (Hilbert-Schmidt Independence Criterion) によって非加法的な関係性を考慮でき、さらに説明変数が高次元のデータに対して有効であるという 2 つの点を両立する。この点で、スパース加法モデル[2]などの既存手法と比べて好ましい性質を持つ。しかし、この手法の枠組みでは変数間の交互作用を直接的には考慮しないため、あらゆる非加法的な関係性を精度良く検出できるわけではない。我々はこの手法を拡張し、変数間の交互作用を直接的に考慮することで、一般の非加法的な関係性に対してより精度良く変数選択できる手法を提案する。提案手法は、低次の交互作用から逐次的に計算を行うことで計算コストを低減し、HSIC Lasso と同様に高次元データに対しても有効である。

数値実験の結果を図 1 に示す。ただし、説明変数 X_j ($j = 1, 2, \dots, 100$) は標準正規乱数とし、目的変数 Y はモデル $Y = X_1 X_2 + X_3 + 0.1\varepsilon$ (左図)、および $Y = X_1 \sin X_2 + X_3^3 + 0.1\varepsilon$ (ε は標準正規乱数) から生成した。HSIC Lasso は左図のような単純な例でも精度が低い場合がある一方で、提案手法は比較的精度が良いことが分かる。また、右図のように、非線形性が強い場合にも提案手法は有効である。

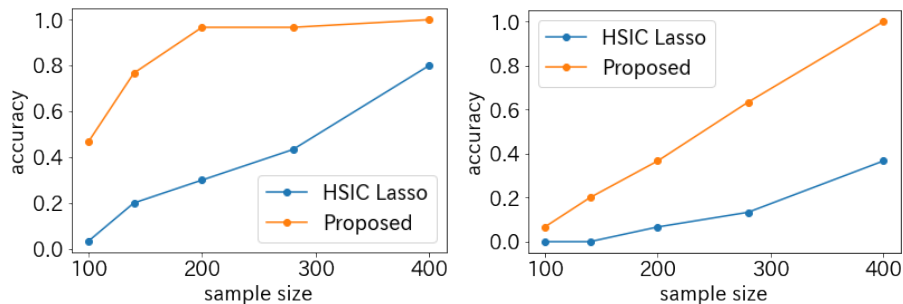


図 1: HSIC Lasso と提案手法の精度の比較。横軸に標本サイズ、縦軸に変数選択の精度 (30回の試行のうち説明変数 X_1, X_2, X_3 が上位3つとして選択された割合) を示す。

参考文献

- [1] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-dimensional feature selection by feature-wise non-linear lasso. *Neural Comput.* 26, 185-207.
- [2] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* 71, 1009-1030.