HSIC-Lasso を用いたモデル誤特定のもとでの、傾向スコアに対する変数選択法

慶應義塾大学大学院理工学研究科　中村知繁
慶應義塾大学理工学部　南 美穂子

1　はじめに

データ解析におけるモチベーションの 1 つは、処置や介入の因果的効果を推定することである。近年、マーケティングや疫学などの分野で、調査観察データにおける推定バイアスが認知され始め、傾向スコアを用いて、処置群と対照群の分布の偏りを補正した上で処置の効果を推定する事例が増えつつある。調査観察研究において、因果的効果を適切に推定するためには、Imbens and Rubin(2015) でも述べられているが、強く無視できる割り付けが成立するために、十分な数の共変量が得られている必要がある。一方で、議論がそれほど多くなされていないのは、観測された変数のうち、どの変数を傾向スコアの推定する際のモデルに含めるかという変数の選択に関する視点である。

いま、強く無視できる割り付けが成り立つために十分な、処置前の共変量が観測されたとすると、観測された共変量の集合に含まれる変数は、次の 4 ついずれかに分類することができる。

(1) 結果変数と、処置変数の両方に影響を与える変数（交絡変数）
(2) 結果変数のみに影響を与える変数
(3) 処置変数のみに影響を与える変数
(4) 結果変数にも処置変数にも影響を与えない変数。

Brookhart et al.,(2006) におけるシミュレーションでは、(1) と (2) の変数を傾向スコアのモデルに含めて平均処置効果を推定した場合に、その他の組み合わせを傾向スコアのモデルに含めた場合と比較して RMSE が小さくなる結果が得られている。この結果に基づいて、傾向スコアを推定する際には、(1) と (2) に該当する変数を用いるという方法が取られることがある（e.g. Austin, 2011）。

2　本発表の内容

しかし、実際のデータ解析においては、観測された変数は (1)〜(4) のいずれに属するのかが不明瞭な場合が多い。特に、観測された変数が多い場合の因果的効果の推定では、観測された変数を (1)〜(4) のいずれかを判別するのは容易なことではない。また、実際に傾向スコアに対する真のモデルに含まれている変数が観測されずに、変換された変数が観測される場合もある。このような状況における因果的効果の推定は、傾向スコアに対するモデルが誤特定されていると考えるのが適切であるが、このような場合の傾向スコアのモデルに含める変数の選択について議論した論文があまり多くはないのが現状である。

よって、本発表ではまず、モデルが誤特定を含む場合に、(1)〜(4) のどの変数を傾向スコアのモデルに含めれば、RMSE が小さくなるのかについてシミュレーションを通して得られた結果を報告する。次に、近年提案された CBPS（Imai and Ratkovic, 2014）や、Calibration Estimator（Chan et al., 2015）についても、同様のシミュレーションを行い結果を報告する。最後に、観測された変数を (1)〜(4) へと振り分けるためのアルゴリズムとして、HSIC-Lasso（Yamada, et al., 2014）を用いた方法を提案し、シミュレーションを行った結果を報告する。

参考文献 (抜粋)

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006). ”Variable selection for propensity score models”, American Journal of Epidemiology, 163, 1149-1156.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, P. E., and Sugiyama, M. (2014) ”High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso”, Neural Computation, 26, 1, 185-207.

Peter C. Austin (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, Multivariate Behavioral Research, 46, 3, 399-424