セミパラメトリックな因果推論におけるモデル選択

塩野義製薬株式会社 解析センター 馬場 崇充

1 はじめに

近年，レセプトデータやレジストリ等のデータベースの構築が進んでおり，そのような観察データの活用に対するニーズが高まっている．しかしながら，この場合に集団間の比較に基づく因果推論を考えた時，比較したい集団間における共変量の分布の偏りが問題となり，介入データに対して通常利用される統計手法が適用できない．その解決のためのモデルの 1 つとして周辺構造モデルが重用されている．このモデルは潜在的な結果変数を用いた反実仮想モデルであり，推定は傾向スコア (Rosembaum & Rubin 1983 Biometrika) に基づくことが一般的である．そして，セミパラメトリックな推定法である IPW（inverse probability weighted）推定（Rubin 1985 BayesStat）やその改良版である DR（doubly robust）推定（例えば Bang & Robins 2005 Biometrics）が脚光を浴びている．このような設定においてもモデル選択は不可欠な作業であるが，今まで妥当なモデル選択基準が存在していなかった．そこで Baba et al. (2017 Biometrika) では，因果推論における損失関数を元に，新たなモデル選択基準を提案した．しかしながら，当該論文では因果推論における Estimand として ATE (average treatment effect) のみを対象としており，例えばその他の ATT (average treatment effect on treated) 等を Estimand とした場合には適用できていなかった．そこで，本発表では Fan et al. (2018 JASA) で提案されている傾向スコアを用いた重みのクラスである，Balancing Weights を推定に用いた際のモデル選択基準を導出し，Baba et al. (2017) を因果推論における様々な Estimand に対応できるように拡張する．具体的には以下に示す主結果の導出法を示し，また数値実験により基準の性能を評価した結果を報告する．

2 モデル

Baba et al. (2017) と同様に

$$\boldsymbol{y} = (1-t)\boldsymbol{y}^{(0)} + t\boldsymbol{y}^{(1)} = (1-t)\boldsymbol{X}^{(0)}\boldsymbol{\beta} + t\boldsymbol{X}^{(1)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

というモデルを考える．ここで，$\boldsymbol{y}^{(h)}$ $(\in \mathbb{R}^r)$ は従属変数ベクトル，$\boldsymbol{X}^{(h)}$ $(\in \mathbb{R}^{r\times p})$ はランダムな共変量ベクトル $\boldsymbol{z}$ $(\in \mathbb{R}^q)$ の一部を含んでもよい独立変数行列とする $(h \in \{0,1\})$．そして，$\boldsymbol{\varepsilon}$ は期待値 $\mathbf{0}$ の誤差変数ベクトルであり，通常の回帰と同様に $\boldsymbol{X}^{(h)}$ と $\boldsymbol{\varepsilon}$ は独立であると仮定する．共変量 $\boldsymbol{z}$ を条件付けたもとで，この t が 1 となる確率 $\mathrm{P}(t=1 \mid \boldsymbol{z}) = e(\boldsymbol{z})$ がいわゆる傾向スコア である．ここで，$e^{(0)}(\boldsymbol{z}) = 1 - e(\boldsymbol{z})$，$e^{(1)}(\boldsymbol{z}) = e(\boldsymbol{z})$ とし，$t^{(0)} = 1-t$，$t^{(1)} = t$ と表記する．また，このモデルに対して N 個の独立なサンプルがあるとし，第 i サンプルの変数には添え字 i を付けることにする．

3 主結果

傾向スコア解析において通常仮定される（弱く）無視できる割り当て条件 $y^{(h)} \perp\!\!\!\perp t \mid \boldsymbol{z}$ 等いくつかの条件を仮定する．また，$t=0$ であるときに Balancing Weights を $w(\boldsymbol{z})/e^{(0)}(\boldsymbol{z})$ とし，$t=1$ であるときに $w(\boldsymbol{z})/e^{(1)}(\boldsymbol{z})$ とした $\boldsymbol{\beta}$ の重み付き推定を考える．ここで，$w(\boldsymbol{z}) = 1$ であるときは通常の IPW に相当し，因果推論における Estimand は ATE である．

定理. 傾向スコアが既知であるとし，$\boldsymbol{\Lambda} = \sum_{h=0}^{1} \mathrm{E}[w(\boldsymbol{z})\boldsymbol{X}^{(h)t}\boldsymbol{X}^{(h)}]$ としたとき，Balancing Weights を重み付けした推定量 $\hat{\boldsymbol{\beta}}^{\mathrm{BW}}$ における漸近 C_p 基準は以下で与えられる:

$$\sum_{h=0}^{1}\sum_{i=1}^{N} \frac{t_i^{(h)}}{e^{(h)}(\boldsymbol{z}_i)} \left(\boldsymbol{y}_i - \boldsymbol{X}_i^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{BW}}\right)^t \left(\boldsymbol{y}_i - \boldsymbol{X}_i^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{BW}}\right) + 2\sum_{h=0}^{1} \mathrm{E}\left[\frac{w(\boldsymbol{z})}{e^{(h)}(\boldsymbol{z})}\boldsymbol{\varepsilon}^t\boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^{(h)t}\boldsymbol{\varepsilon}\right]$$

引用文献

1. Baba, T., Kanemori, T., & Ninomiya, Y. (2017). A C_p criterion for semiparametric causal inference. *Biometrika*, **104**, 845-861.