2022 年 6 月 17 日 統計数理研究所 オープンハウス

周辺構造モデルにおける二重に頑健なモデル選択基準

馬場 崇充　総合研究大学院大学 統計科学専攻 博士課程 (3年次編入) 3年

1　はじめに

因果推論の一般的な設定において，観測されたデータを用いてナイーブに推定をおこなうと，推定値はバイアスをもつことになる．バイアスを生む原因となっている交絡因子を調整するために，傾向スコアを用いた二重頑健推定というセミパラメトリックアプローチがよくとられる．二重頑健推定量は傾向スコアもしくは結果変数の交絡因子の条件付き期待値のどちらか一方を正しくモデリングできれば一致性を持つことになる．この際のモデル選択基準として，Baba et al. (2017) は周辺構造モデルの情報量規準を与えているものの，両者のモデリングが正しいと仮定した下で導かれたものであるにすぎない，そこで，どちらかのモデリングさえ正しければ数理的に妥当になるような情報量規準を開発した．

2　モデルと仮定

$$y = \sum_{h=1}^{H} t^{(h)} y^{(h)}, \qquad y^{(h)} \sim f(\cdot \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$$

- 割り当て変数 $t^{(h)}$ $(\in \{0,1\}$ かつ $\sum_{h=1}^{H} t^{(h)} = 1)$
- $y^{(h)}$ に対する 説明変数 $\boldsymbol{x}^{(h)}$ $(\in \mathbb{R}^r)$ の回帰モデル $f(y^{(h)} \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$
- 推定対象となるパラメータ $\boldsymbol{\theta}$ $(\in \mathbb{R}^p)$
- $y^{(h)}$ と $t^{(h)}$ の $\boldsymbol{z}$ を条件づけた時の独立性 (強く無視できる割り当て条件)
- 傾向スコアの正値条件 $e^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) = \mathrm{P}(t^{(h)} = 1 | \boldsymbol{z}) > 0$

$d^{(k)}$ は因果効果を推定する集団を規定する変数で，$\boldsymbol{\theta}$ は以下を満たす．

$$\sum_{h,k=1}^{H} \mathrm{E}\left\{ d^{(k)} t^{(k)} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y^{(h)} \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta}) \right\} = \mathbf{0}_p$$

3　モデル選択基準導出の準備

重みを $w^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) \equiv \sum_{k=1}^{H} d^{(k)} e^{(k)}(\boldsymbol{z}; \boldsymbol{\alpha}) / e^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha})$ とし，傾向スコアの重み付き Kullback-Leibler ダイバージェンスを定義．

$$-2 \sum_{i=1}^{N} \sum_{h=1}^{H} \mathrm{E}\{ t_i^{(h)\dagger} w^{(h)}(\boldsymbol{z}_i^{\dagger}) \log f(y_i^{(h)\dagger} \mid \boldsymbol{x}_i^{(h)\dagger}; \hat{\boldsymbol{\theta}}) \} \tag{1}$$

$\dagger$ は将来得られるデータを表す．モデル選択基準の第一項として (1) を推定時のデータで経験的に評価し，第2項として推定時のデータで評価することによるバイアスを漸近評価することで (1) の漸近不偏推定量となるモデル選択基準を導出する．

4　二重に頑健なモデル選択基準

$y^{(h)}$ の周辺対数尤度の交絡因子 $\boldsymbol{z}$ の条件付き期待値を $g^{(h)}(\boldsymbol{x}^{(h)}, \boldsymbol{z}; \boldsymbol{\theta}, \boldsymbol{\beta})$ としたとき，二重頑健推定として，以下の推定方程式を考える．

$$\sum_{h=1}^{H} \frac{\partial}{\partial \boldsymbol{\theta}} \Bigg[t^{(h)} w^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) \log f(y^{(h)} \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta}) + \left\{ \sum_{k=1}^{H} d^{(k)} t^{(k)} - t^{(h)} w^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) \right\} g^{(h)}(\boldsymbol{x}^{(h)}, \boldsymbol{z}; \boldsymbol{\theta}, \boldsymbol{\beta}) \Bigg] = \mathbf{0}$$

定理. 「結果変数に対する交絡変数のモデル」か「割り当て変数に対する交絡変数のモデル」のどちらか一方が正しい場合の (1) を漸近評価することで二重頑健推定に対する以下のモデル選択基準を得る (Baba and Ninomiya 2021)．

$$\mathrm{DRIC} \equiv -2 \sum_{i=1}^{N} \sum_{h=1}^{H} t_i^{(h)} w^{(h)}(\boldsymbol{z}_i; \hat{\boldsymbol{\alpha}}) \log f(y_i^{(h)} \mid \boldsymbol{x}_i^{(h)}; \hat{\boldsymbol{\theta}}^{\mathrm{DR}}) + 2\mathrm{tr}[\hat{\boldsymbol{A}}^{-1}\{\hat{\boldsymbol{B}} + \hat{\boldsymbol{D}}_1\} + \hat{\boldsymbol{D}}_2 + \hat{\boldsymbol{D}}_3)]$$

$\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2, \hat{\boldsymbol{D}}_3$ は観測データに基づいて経験的に推定される．

5　数値実験

- h $(\in \{1,2,3,4,5\})$ を時刻とし，そのどこかで結果変数 $y^{(h)}$ $(\in \{0,1\})$ は観測される．$(t^{(h)} = 1)$
- 交絡変数 $\boldsymbol{z} = (z_1, z_2)' \sim \mathrm{N}(\mathbf{0}_2, \boldsymbol{I}_2)$
- 割付変数 $t^{(h)} \sim \mathrm{multilogit}(\boldsymbol{z})$
- 推定対象 $(d^{(1)}, d^{(2)}, d^{(3)}, d^{(4)}) = (1,1,1,1)$ とした全群における平均処置効果

$y^{(h)}$ は2値応答であり，ロジット関数の構造として以下の2つを考える．

- linier $0.5 + \theta^*(h-1) + \boldsymbol{\beta}'\boldsymbol{z} + \epsilon \quad (\epsilon \sim \mathrm{N}(0,1))$
- quadratic $0.5 + 0.2(h-1) + \theta^*(h-1)^2 + \boldsymbol{\beta}'\boldsymbol{z} + \epsilon$

真のモデルを linear もしくは quadratic としたとき，DRIC の罰則項が，バイアスを近似できているかどうかを検証する．比較対象は Platt et al. (2013) で提案されている QIC_w で，罰則項は自由パラメータの2倍である．実験の繰り返し回数は $2,000$．

表 1: 離散変数モデルにおけるバイアス評価．MCE の列はモンテカルロ法により評価した真値であり，AE の列はその漸近評価．QIC_w は自由パラメータの2倍

	linear $N = 200$			quadratic $N = 200$		
$(\theta^*$, 傾向スコア, 条件付き分布)	QIC_w	MCE	AE	QIC_w	MCE	AE
(0.05, 正, 正)	4	25.00	21.41	6	40.82	30.99
(0.005, 正, 正)	4	21.40	21.44	6	32.62	32.10
(0.05, 誤, 正)	4	25.78	21.42	6	38.53	31.21
(0.005, 誤, 正)	4	23.67	21.49	6	39.27	32.11
(0.05, 正, 誤)	4	26.06	21.44	6	38.96	31.05
(0.005, 正, 誤)	4	25.07	21.41	6	39.21	32.05

6　結論

2値応答の場合の因果推論における二重に頑健なモデル選択基準を提案し，既存手法よりもバイアスの漸近近似が良いことを示した．さらに連続地応答でも提案手法のほうが既存手法よりもバイアスの漸近近似が良いことを確認している．また，様々なモデル誤特定の条件下で予測の意味で周辺構造のモデル選択がうまくいくことがわかっている (Baba and Ninomiya 2021 参照)．

参考文献

Baba, T., Kanemori, T., and Ninomiya, Y. (2017). A Cp criterion for semiparametric causal inference, *Biometrika*, **104**, 845–861.

Baba, T. and Ninomiya, Y. (2021). Doubly Robust Criterion for Causal Inference, *arXiv*, 2110.14525.

Platt, R. W., Brookhart, M. A., Cole, S. R., Westreich, D., and Schisterman, E. F. (2013). An information criterion for marginal structural models, *Statistics in Medicine*, **32**, 1383–1393.