2020年10月27日　統計数理研究所　オープンハウス

傾向スコア解析における選択的推論

二宮　嘉行　数理・推論研究系　教授

目的

因果推論において標準手法となっている，セミパラメトリックアプローチに基づく傾向スコア解析において，統計学および機械学習の分野で脚光を浴びている手法である選択的推論（選択後推論）[1] を開発する

準備：因果推論モデル

h 番目 $(1 \le h \le H)$ の処置を受けたら $Y^{(h)}$ $(\in \mathbb{R})$ が観測されるモデル

$$Y \equiv \sum_{h=1}^{H} T^{(h)} Y^{(h)} = \sum_{h=1}^{H} T^{(h)} \left\{ \mu^{(h)}(\boldsymbol{X}) + f(\boldsymbol{X}) + \epsilon^{(h)} \right\}$$

- h 番目の処置を受けると 1 で受けないと 0 になるランダムな割り当て変数が $T^{(h)}$，$Y^{(h)}$ と $T^{(h)}$ の交絡変数ベクトルが $\boldsymbol{X}$ $(\in \mathbb{R}^p)$，$f: \mathbb{R}^p \to \mathbb{R}$ が非線形関数，$\epsilon^{(h)}$ が $\mathrm{N}(0, \sigma^2)$ にしたがう非観測変数
- $(c^{(1)}, \ldots, c^{(H)})'$ を $\sum_{h=1}^{H} c^{(h)} = 0$ なるコントラストとし，推定したい因果効果は $\boldsymbol{X} = \boldsymbol{x}$ に対する条件付き平均処置効果 $\sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\boldsymbol{x})$
- $\tilde{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)'$，$\tilde{\boldsymbol{T}}^{(h)} = \mathrm{diag}(T_1^{(h)}, \ldots, T_n^{(h)})$，$\tilde{\boldsymbol{Y}}^{(h)} = (Y_1^{(h)}, \ldots, Y_n^{(h)})'$，$\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$，$\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) = (\mu^{(h)}(\boldsymbol{X}_1), \ldots, \mu^{(h)}(\boldsymbol{X}_n))'$，$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}}) = (f(\boldsymbol{X}_1), \ldots, f(\boldsymbol{X}_n))'$，$\tilde{\boldsymbol{\epsilon}}^{(h)} = (\epsilon_1^{(h)}, \ldots, \epsilon_n^{(h)})'$ とすれば，

$$\tilde{\boldsymbol{Y}} = \sum_{h=1}^{H} \tilde{\boldsymbol{T}}^{(h)} \tilde{\boldsymbol{Y}}^{(h)} = \sum_{h=1}^{H} \tilde{\boldsymbol{T}}^{(h)} \left\{ \tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{\epsilon}}^{(h)} \right\}$$

- 弱く無視できる割り当て条件，正値条件，サンプルの独立性を仮定

準備：傾向スコア解析

傾向スコア $e^{(h)}(\boldsymbol{X}_i) \equiv \mathrm{P}(T_i^{(h)} = 1 \mid \boldsymbol{X}_i)$ を用い

$$\left\| \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{Y}} - \sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) \right\|_2^2$$

の最小化で逆確率重み付け推定を与える

- $\tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)}, \tilde{\boldsymbol{X}}) \equiv \mathrm{diag}\{T_1^{(h)}/e^{(h)}(\boldsymbol{X}_1), \ldots, T_n^{(h)}/e^{(h)}(\boldsymbol{X}_n)\}$ かつ $\tilde{\boldsymbol{T}} = (\tilde{\boldsymbol{T}}^{(1)}, \ldots, \tilde{\boldsymbol{T}}^{(H)})$ として $\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \equiv \sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)}, \tilde{\boldsymbol{X}})$ と定義
- 弱く無視できる割り当て条件のもとで一致性をもつ

準備：因果推論モデルに対する選択的推論

選択されるモデルを $\hat{\boldsymbol{M}}$ とし，以下の被覆を与える $C_j^{\boldsymbol{M}}$ を考える

$$\mathrm{P}\left(\beta_j^{\boldsymbol{M}} \in C_j^{\boldsymbol{M}} \;\middle|\; \hat{\boldsymbol{M}} = \boldsymbol{M} \right) \ge 1 - \alpha$$

- $\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}})$ に対するモデルとして $\boldsymbol{M} \subset \{1, \ldots, p\}$ に属する交絡変数 $\tilde{\boldsymbol{X}}_{\boldsymbol{M}}$ の線形和を考え，推定対象となる回帰係数として次を考えている

$$\boldsymbol{\beta}^{\boldsymbol{M}} \equiv \left(\tilde{\boldsymbol{X}}_{\boldsymbol{M}} \tilde{\boldsymbol{X}}'_{\boldsymbol{M}} \right)^{-1} \tilde{\boldsymbol{X}}_{\boldsymbol{M}} \sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\tilde{\boldsymbol{X}})$$

- 適当な単位ベクトル $\boldsymbol{e}_j$ を用いれば $\beta_j^{\boldsymbol{M}} = \boldsymbol{e}'_j (\tilde{\boldsymbol{X}}_{\boldsymbol{M}} \tilde{\boldsymbol{X}}'_{\boldsymbol{M}})^{-1} \tilde{\boldsymbol{X}}_{\boldsymbol{M}} \sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\tilde{\boldsymbol{X}})$ と書けることを鑑み，$\tilde{\boldsymbol{\eta}}_j \equiv \tilde{\boldsymbol{X}}_{\boldsymbol{M}} (\tilde{\boldsymbol{X}}'_{\boldsymbol{M}} \tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1} \boldsymbol{e}_j$ として $\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{Y}}$ の条件付き分布を考える

引用文献

[1] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44, 907–927.

因果推論モデルに対する LASSO 推定量の提案

$\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{X}} \boldsymbol{\beta}^{(h)}$ とし，$\tilde{\boldsymbol{X}}$ の次元 p が高いことを想定して次を提案

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ \left\| \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right) \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

- $\hat{\boldsymbol{M}} = \{j : \hat{\beta}_j \neq 0\}$，その 0 でない推定量を集めたものを $\hat{\boldsymbol{\beta}}^{\hat{\boldsymbol{M}}}$，その符号を $\hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \mathrm{sign}(\hat{\boldsymbol{\beta}}^{\hat{\boldsymbol{M}}})$ と記すと，Karuch-Kuhn-Tucker 条件より，任意のモデル $\boldsymbol{M}$ $(\subset \{1, \ldots, p\})$ と任意の符号 $\boldsymbol{s}$ $(\in \{-1, 1\}^{|\boldsymbol{M}|})$ に対して

$$\left\{ \hat{\boldsymbol{M}} = \boldsymbol{M},\ \hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \boldsymbol{s} \right\} = \left\{ \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s}) \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right) \tilde{\boldsymbol{Y}} \le \boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s}) \right\}$$

なる $n \times n$ 行列 $\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})$ と n 次元ベクトル $\boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})$ が存在

主結果

定理 1（因果推論モデル版非漸近多面体補題）: 以下の分布は $\mathrm{Unif}(0, 1)$

$$F_{\gamma_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}), \zeta_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})}^{[\mathcal{V}_{\boldsymbol{s},j}^{-}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}), \mathcal{V}_{\boldsymbol{s},j}^{+}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})]} \left(\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right) \tilde{\boldsymbol{Y}} \right) \mid \hat{\boldsymbol{M}} = \boldsymbol{M},\ \hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \boldsymbol{s}$$

- $F_{\mu, \sigma^2}^{[a,b]}$ は区間 $[a, b]$ に切断した $\mathrm{N}(\mu, \sigma^2)$ の累積分布関数，$\boldsymbol{X}_{\boldsymbol{M}} = (X_j)_{j \in \boldsymbol{M}}$，$\tilde{\boldsymbol{Z}} \equiv \{\boldsymbol{I}_n - \tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{\eta}}'_j\} \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{Y}}$
- $\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\mathcal{V}_{\boldsymbol{s},j}^{-}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\mathcal{V}_{\boldsymbol{s},j}^{+}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\gamma_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\zeta_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\tau_j^{\boldsymbol{M}}(\tilde{\boldsymbol{Y}}^{\dagger}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$，$\rho_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right)$ は適当な関数

定理 2（因果推論モデル版漸近多面体補題）：条件のもとで $\tilde{\boldsymbol{Y}}^{\dagger} - \{\tilde{\boldsymbol{\mu}}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}})\} = \mathrm{o_P}(1)$ ならば，以下の漸近分布は $\mathrm{Unif}(0, 1)$

$$F_{\beta_j^{\boldsymbol{M}} + \tau_j^{\boldsymbol{M}}(\tilde{\boldsymbol{Y}}^{\dagger}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}), \zeta_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})}^{[\mathcal{V}_{\boldsymbol{s},j}^{-}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}), \mathcal{V}_{\boldsymbol{s},j}^{+}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})]} \left(\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right) \tilde{\boldsymbol{Y}} \right) \mid \hat{\boldsymbol{M}} = \boldsymbol{M},\ \hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \boldsymbol{s}$$

- 実際には $\tilde{\boldsymbol{Y}}^{\dagger}$ に代わる自然な推定量を用いる

数値実験：偽被覆率と信頼区間の比較

f: 0 or 線形 or 非線形，μ: 線形，e: ロジスティック，X: 離散 or 連続，n: 1000，p: 25（零は 20，非零は 5）で選択的推論と普通の推論を比較 (TP: 正しく棄却した変数数，FP: 誤って棄却した変数数，FCR: 偽被覆率)

		X: 離散				X: 連続			
f		$\lvert\hat{M}\rvert$	TP	FP	FCR	$\lvert\hat{M}\rvert$	TP	FP	FCR
0	選	13.841	4.979	0.405	0.048	5.287	4.994	0.013	0.053
	普		5.000	8.841	0.874		5.000	0.287	0.491
線	選	20.380	4.938	0.834	0.053	11.844	4.780	0.391	0.054
	普		4.954	15.425	0.968		4.818	7.026	0.904
非	選	17.238	4.963	0.647	0.052	7.050	4.974	0.101	0.054
	普		5.000	12.238	0.933		4.997	2.053	0.724

f: 非線形，X: 連続，で選択的推論（SI）と普通の推論（Naive）を比較