# 外れ値にロバストな二重頑健推定

藤澤 洋徳　　数理・推論研究系 教授

## 1　はじめに

二重頑健推定は，生起確率モデルと回帰モデルのうち片方が正しければ，応答変数の平均を妥当に推定可能という特徴がある．統計的因果推論でも使われている．本ポスターでは，外れ値にロバストな二重頑健推定を提案する．提案手法には，外れ値の割合が十分に小さくなくても良い，というところにも特徴がある．本研究は原田和治氏(東京医科大)との共同研究です．

## 2　設定

ターゲット： $\mu^* = \mathrm{E}[Y]$

データ：$(X_1, Y_1, T_1), \ldots, (X_n, Y_n, T_n) \sim_{i.i.d.}$
　$Y$は観測される/欠測する: $T_i = 1/0$

仮定： $X$を与えた下で，$Y$と$T$は独立

## 3　二重頑健推定

生起確率モデル： $\pi^*(X) = \Pr(T = 1 \mid X) \leftarrow \pi(X;\alpha)$
　$\hat{\alpha}$: $(T_i, X_i)$sに基づいて推定

回帰モデル: $\mathrm{E}[Y|X] = m^*(X) \leftarrow m(X;\beta)$
　$\hat{\beta}$: 観測された$(Y_i, X_i)$sに基づいて推定

二重頑健推定

$$\sum_{i=1}^{n}\left[\frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\{m(X_i;\hat{\beta})-\mu\}\right]=0$$

$$\hat{\mu}_{\mathrm{DR}}=\sum_{i=1}^{n}\left[\frac{T_i}{\pi(X_i;\hat{\alpha})}Y_i+\left\{1-\frac{T_i}{\pi(X_i;\hat{\alpha})}\right\}m(X_i;\hat{\beta})\right]$$

仮定：$\hat{\alpha}=\alpha^\dagger$かつ$\hat{\beta}=\beta^\dagger$
　$\pi(X;\alpha^\dagger)=\pi^*(X)$または$m(X;\beta^\dagger)=m^*(X)$が成立
　(生起確率モデルと回帰モデルのうち片方が正しい)

結論： $\mathrm{E}[\hat{\mu}_{\mathrm{DR}}]=\mu^*$.

## 4　外れ値の想定

汚染分布

$$g_{Y|X}(y|x) = (1-\varepsilon^*(x))f^*_{Y|X}(y|x)+\varepsilon^*(x)\delta_{Y|X}(y|x)$$
$$g_X(x) = f^*_X(x) \qquad \text{外れ値なし}$$

仮定： $f^*_Y(y)$は$y=\mu^*$で左右対称. $\mathrm{E}_{f^*_Y}[Y]=\mu^*$.
(この想定は中央値＝平均でも使われる．)

重み(密度)関数　$h(z)=h(-z)$ ( e.g., $=\phi(z;0,\tau^2)$ )

外れ値らしさの程度　任意の$x$と固定された$\gamma>0$に対して，

$$\nu_h(x;\mu) = \int h^*_Y(y-\mu)^\gamma(y-\mu)\delta_{Y|X}(y|x)dy \approx 0 \quad \mu\in\mathcal{M}$$

## 5　外れ値にロバストな二重頑健推定

DP-DR (Density Power-Doubly Robust)

$$\sum_{i=1}^{n}\left[\frac{T_i}{\pi(X_i;\hat{\alpha})}h(Y_i-\mu)^\gamma(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}s_f(X_i;\hat{\theta},\mu)\right]=0$$

$$s_f(X;\theta,\mu)=\mathrm{E}_{f_{Y|X;\theta}}[h(Y-\mu)^\gamma(Y-\mu)\,|\,X]$$

工夫: $(Y_i-\mu) \to h(Y_i-\mu)^\gamma(Y_i-\mu)$とバイアス補正.
注意: $f_{Y|X;\theta}$は 適当なモデル; たとえば$\phi(y;x^\top\beta,\sigma^2)$.

推定方程式のバイアス

$$\pi(X;\hat{\alpha})\to\pi(X;\alpha^\dagger)=\pi(X) \qquad f_{Y|X;\hat{\theta}}\to f_{Y|X;\theta^\dagger}=f_{Y|X}$$

外れ値なしの場合
　$\pi(X)=\pi^*(X)$のとき, バイアスなし.
　$f_{Y|X}=f^*_{Y|X}$のとき, バイアスなし.

外れ値ありの場合
　$\pi(X)=\pi^*(X)$のとき, バイアスはほぼ0.
　$f_{Y|X}=f^*_{Y|X}$のとき, バイアスが存在する.
　　(もしも$\varepsilon(X)$が十分に小さくないならば)

$\varepsilon$DP-DR ($\varepsilon$ Density Power-Doubly Robust)

$$\sum_{i=1}^{n}\left[\frac{T_i}{\pi(X_i;\hat{\alpha})}h(Y_i-\mu)^\gamma(Y_i-\mu)-(1-\hat{\varepsilon})\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}s_f(X_i;\hat{\theta},\mu)\right]=0$$

外れ値ありの場合
　$\pi(X)=\pi^*(X)$のとき, バイアスはほぼ0.
　$f_{Y|X}=f^*_{Y|X}$かつ, $\varepsilon^*(X)=\varepsilon^*$かつ$\hat{\varepsilon}=\varepsilon^*$ならば，バイアスはほぼ0.

## 6　数値実験

NHEFS data
1,566 smokers
Y: weight gain
T: smoking cessation
X: 9 covariates

10% responses were replaced
　by outliers $\sim N(100, 5^2)$
$10^4$ simulations

DR (Doubly Robust)
DR estimate without outliers
　is the target.

| | |
|---|---|
| DR (no outliers) | 5.136 (-) |
| DR | 14.574 (1.57) |
| median (Zhang-DR) | 5.352 (0.20) |
| median (Sued) | 5.353 (0.20) |
| median (TMLE) | 5.363 (0.21) |
| DP-DR ($\gamma=0.05$) | 5.478 (0.27) |
| DP-DR ($\gamma=0.10$) | 5.057 (0.16) |
| DP-DR ($\gamma=0.20$) | 4.983 (0.16) |
| DP-DR ($\gamma=0.50$) | 4.834 (0.16) |
| $\varepsilon$DP-DR ($\gamma=0.05$) | 5.574 (0.29) |
| $\varepsilon$DP-DR ($\gamma=0.10$) | 5.148 (0.15) |
| $\varepsilon$DP-DR ($\gamma=0.20$) | 5.080 (0.15) |
| $\varepsilon$DP-DR ($\gamma=0.50$) | 4.937 (0.15) |

## 0　研究室の研究

キーワード：ロバスト統計．ダイバージェンス．スパース・モデリング．グラフィカル・モデリング．非対称分布．遺伝子発現データ．モデル選択．混合効果モデル．経時データ．欠測データ．多重検定．因果探索．因果推論．多重代入．クラスタリング．高欠測データ．多量の外れ値．異常検知．

企業や異分野との共同研究： 藤澤のHPをご覧ください．具体的な記事や事例も紹介しています．現在も幾つか進行中です．

現学生の研究テーマ
フルタイムA： HSIC Lasso. 歪ノイズ.
社会人A：状態空間モデル＋様々なタイプのデータ処理 (外れ値や欠測)
社会人B： ロバスト推定．非漸近理論．収束レート．
社会人C： 異常検知.

修了生の研究テーマ
フルタイムa： 因果探索．因果推論．ロバスト推定.
社会人a：スパースモデリング（高相関・高欠測・転移学習に関連して）
フルタイムb： ロバスト性とスパース性を併せもつ回帰モデリング
フルタイムc： 多重代入法