Huber lossの共変量に対するロバスト性について

笹井 健行　総合研究大学院大学 統計科学専攻 博士課程(3年編入学)3年

1　問題設定と結果のサマリ

次のモデルからのスパースな回帰係数$\boldsymbol{\beta}^* \in \mathbb{R}^d$を推定したい:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \xi_i, \quad i = 1, \cdots, n, \tag{1}$$

分析者はi.i.d.列$\{y_i, \mathbf{x}_i\}_{i=1}^n$を観測できるものとし，$\{\xi_i\}_{i=1}^n$は$\{\mathbf{x}_i\}_{i=1}^n$とは独立なランダムノイズとする．$\boldsymbol{\beta}^*$の非0要素数を$s$とする．多くの研究では，$\{\mathbf{x}_i\}_{i=1}^n$は多変量正規分布に従う場合を扱っている。この時，さらに$\{\xi_i\}_{i=1}^n$が分散σ^2の正規分布に従うと仮定ことで，通常のlassoの手続き

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left\| y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right\|_2^2 + \lambda_s \|\boldsymbol{\beta}\|_1 \right\}, \tag{2}$$

により，(1)から$\boldsymbol{\beta}^*$を

$$\mathbb{P}\left\{ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \left(\sqrt{s \frac{\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \right\} \geq 1 - \delta, \tag{3}$$

と推定できる。いくつかの研究は，$\{\mathbf{x}_i\}_{i=1}^n$として，多変量正規分布を含むクラスであるL-subGaussianを扱っている。また，$\{\xi_i\}_{i=1}^n$が裾が厚い分布に従っている場合は，lassoの二乗損失をHuber lossに変更したℓ_1-罰則付きHuber回帰により，(3)と定数を除いて同様の結果を得られることが，先行研究により判明している[4]。

ところで，$\{\mathbf{x}_i\}_{i=1}^n$が多変量正規分布やL-subGaussianよりも裾が厚い分布に従う場合を考察した論文は少ない[1, 3, 2]。特に，[3, 2]は，L-subGaussianよりも一回り裾が厚い分布のクラスとして，L-subexponentialと呼ばれる分布を仮定し，議論している。先行研究[3, 2]は(3)のような結果を出そうと試みているが，余分な条件として$s(\log(d/s))^2 \lesssim n$が必要となる。本研究では，$\{\mathbf{x}_i\}_{i=1}^n$が$L$-subexponentialに従う場合であっても，これまで$\{\xi_i\}_{i=1}^n$に対してのみロバスト性を発揮すると考えられてきたℓ_1-罰則付きHuber回帰により，実はこの余分な条件なしに(3)と定数を除いて同様な結果が得られることを示す。

2　定義・例

確率変数fに対してOrliczノルムを次のように定義する:

$$\|f\|_{\psi_\alpha} := \inf\left[\eta > 0 \,:\, \mathbb{E}\exp|f/\eta|^\alpha \leq 2\right] < \infty. \tag{4}$$

ここで，任意の固定した$\mathbf{v} \in \mathbb{R}^d$に対して，$d$次元確率変数$\mathbf{x}$が，次の(5)を$\alpha = 1$で満たす場合は$L$-subexponential，$\alpha = 2$で満たす場合は$L$-subGaussianという:

$$\|\langle \mathbf{x}, \mathbf{v} \rangle\|_{\psi_\alpha} \leq L \left(\mathbb{E}|\langle \mathbf{x}, \mathbf{v} \rangle|^2 \right)^{\frac{1}{2}}. \tag{5}$$

定義より，L-subGaussianはL-subexponentialである。L-subGaussianの例としては，多変量正規分布や各成分が独立なRademacher確率変数に従う分布などがあり，L-subexponentialの例としては，各成分が独立なガンマ分布に従う分布などがある。

3　仮定・結果

Huber lossを以下のように定義する:

$$H(t) = \begin{cases} |t| - 1/2 & (|t| > 1) \\ t^2/2 & (|t| \leq 1) \end{cases}. \tag{6}$$

ノイズと共変量$\{\xi_i, \mathbf{x}_i\}_{i=1}^n$に以下の仮定をおく:

(i) $\mathbf{x}_i$はL-subexponential分布に従い，$\mathbb{E}\mathbf{x}_i = 0$, $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top = I$,

(ii) $\mathbb{E}\xi_i^2 \leq \sigma^2$.

上記の仮定の下，推定量

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \lambda_o^2 H\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\lambda_o \sqrt{n}} \right) + \lambda_s \|\boldsymbol{\beta}\|_1 \right\} \tag{7}$$

は，$\lambda_o \asymp \sigma L^2/\sqrt{n}$, $\lambda_s \asymp L^3 \sigma \left(\sqrt{s\frac{\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right)$のもとで以下を満たす:

$$\mathbb{P}\left\{ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim L^3 \sigma \left(\sqrt{s \frac{\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \right\} \geq 1 - \delta. \tag{8}$$

4　なぜHuber lossでうまくいくのか？

今回の設定でℓ_1罰則付きHuber回帰が効果的な理由を標題的に述べると「ノイズ$\{\xi_i\}_{i=1}^n$の挙動を抑えることで$\{\mathbf{x}_i\}_{i=1}^n$が動く余裕を与えられるから」である。以下，詳しく解説する。

Huber lossの微分を次のように定義する:

$$h(t) = \frac{d}{dt}H(t) = \begin{cases} t & (|t| \leq 1) \\ \operatorname{sgn}(t) & (|t| > 1) \end{cases}. \tag{9}$$

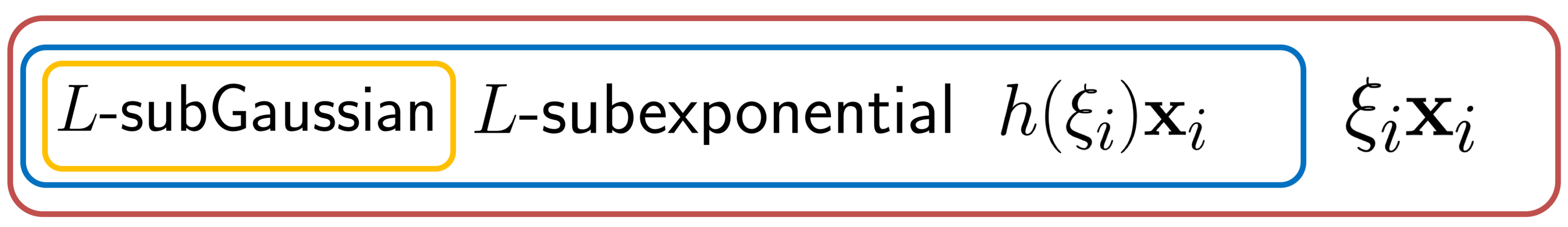
通常のlassoの場合，以下のような量の評価が重要である(r_1, r_2はパラメータ)

$$\sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d} \sum_{i=1}^n \frac{1}{n} \xi_i \mathbf{x}_i^\top \mathbf{v}. \tag{10}$$

一方で，(7)の場合は

$$\sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d} \sum_{i=1}^n \frac{1}{n} h(\xi_i) \mathbf{x}_i^\top \mathbf{v} \tag{11}$$

を評価すればよい。$\mathbf{x}_i$がL-subGaussianかつξ_iがσ-subGaussianの場合，(10)における$\xi_i\mathbf{x}_i$は$L\sigma$-subexponentialとなる。しかし，$\mathbf{x}_i$がL-subexponentialの場合，たとえξ_iがGaussianであっても$\xi_i\mathbf{x}_i$はL-subexponentialとはならない。一方で，(11)における$h(\xi_i)\mathbf{x}_i$は$h(\cdot)$の有界性ゆえにL-subexponentialであり続ける。この現象により，多少の共変量の裾の厚さをHuber lossで吸収できる．



Huber lossにより分布が変化するイメージ

参考文献

[1] J. Fan, W. Wang, and Z. Zhu. *Annals of statistics*, 49(3):1239, 2021.

[2] M. Genzel and C. Kipp. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):15, 2022.

[3] V. Sivakumar, A. Banerjee, and P. K. Ravikumar. *Advances in neural information processing systems*, 28, 2015.

[4] Q. Sun, W. Zhou, and J. Fan. *Journal of the American Statistical Association*, 115(529):254–265, 2020.