敵対的攻撃に対して頑強な線形回帰

笹井 健行　総合研究大学院大学複合科学研究科統計科学専攻 博士課程（5年一貫制）4年

導入

実データにはノイズが混入することが多く，一般的に分析の際には前処理が要求される．前処理に当たってはデータに対する専門的な知識が要求されることが多いため，詳細が不明なデータに対する分析は一般的には困難である．一方で，データの生成機序に対する詳細な知識がない場合であっても，分析手法自体をノイズに対してロバストにすることで分析が可能となる場合がある．本研究では，ノイズが混入している場合であっても，線形回における回帰係数をロバストに推定する手法を，基礎研究的な観点から提案する．

書誌情報

https://arxiv.org/abs/2102.11120

問題設定

$\{x_i' \in \mathbf{R}^n\}_{i=1}^n$を共変量ベクトルの列，$\{y_i' \in \mathbf{R}\}_{i=1}^n$を出力の列とし，

$$y_i' = x_i'^{\top}\beta^* + \xi_i, \quad i = 1, \cdots, n, \tag{0.1}$$

の関係があるものとする．敵対者は，（正常な）出力と共変量ベクトルの組みの列$\{y_i', x_i'\}_{i=1}^n$からo個を選択し，任意の値に置き換えてよいものとする．置き換えられた後の出力と共変量ベクトルの組みの列を$\{y_i, x_i\}_{i=1}^n$と書くものとする．このとき，$\{y_i, x_i\}_{i=1}^n$からβ^*を精度良く推定する方法を考える．簡単のため，汚染割合$\varepsilon = o/n$は十分小さい定数と仮定する．

Remark 1. 敵対者がサンプルを正常なサンプルに加えてくる場合（*Huber contamination model*）とは異なり，敵対者がサンプルを置き換えるため，正常なサンプルが持つ独立性を破壊できる．この事情から，*Huber's contamination*とは異なる解析手法が要求される．

提案手法

この問題に対して，次のような2段階の推定手法（TWO　STEP WEIGHTED HUBER REGRESSION）を提案する:

i 共変量ベクトル$\{x_i\}_{i=1}^n$から，重み$\{\hat{w}_i\}_{i=1}^n$を計算する

ii $\{y_i, x_i w_i\}_{i=1}^n$に対してHuber 回帰を行う

提案手法のstep 1では，$\{w_i\}_{i=1}^n$が，$\{x_i w_i\}_{i=1}^n$の平均がx_i'の真の平均に近くなるように推定されることが要求される．このような性質を満たす$\{w_i\}_{i=1}^n$を，n, dの多項式の計算量で推定する方法が複数知られている．

定理

提案手法により，次の定理を得た

Theorem 1 (共変量ベクトルがsub-Gaussianの場合). $\{x_i\}_{i=1}^n$ は定数Cに対して$L^4 - L^2$等価性をもつ*sub-Gaussian*分布からの独立なサンプルとし，$\{\xi_i\}_{i=1}^n$ は絶対モーメントが存在する分布からの独立なサンプルとする．また，$n = O(d \log d)$とする．このとき，提案手法により推定された$\hat{\beta}$は次を満たす*:*

$$\|\hat{\beta} - \beta^*\|_2 = O\left(\varepsilon \sqrt{\log \frac{1}{\varepsilon}}\right). \tag{0.2}$$

Theorem 2 (共変量ベクトルがheavy-tailの場合). $\{x_i\}_{i=1}^n$ は定数Cに対して$L^4 - L^2$等価性を満たす分布からの独立なサンプルとし，$\{\xi_i\}_{i=1}^n$は絶対モーメントが存在する分布からの独立なサンプルとする．また，$n = O(d \log d)$とする．このとき，提案手法により推定された$\hat{\beta}$は次を満たす*:*

$$\|\hat{\beta} - \beta^*\|_2 = O\left(\varepsilon \sqrt{\log \frac{1}{\varepsilon}}\right). \tag{0.3}$$

Remark 2. ランダムベクトルxが定数Cに対して$L^4 - L^2$等価性*($L^4 - L^2$ equivalence)*を満たすとは，定数C及び固定された任意の$v \in \mathbb{R}$に対して

$$\left(\mathbb{E}\left|\langle x, v\rangle\right|^4\right)^{\frac{1}{4}} \leq 2C\left(\mathbb{E}\left|\langle x, v\rangle\right|^2\right)^{\frac{1}{2}} \tag{0.4}$$

が成立すること，つまり尖度が有界であることをいう． 例えば多変量*Gaussian*は*sub-Gaussian*かつ定数Cに対して$L^4 - L^2$等価性を満たす分布であるが，一般的な*sub-Gaussian*は必ずしも定数Cに対して$L^4 - L^2$等価性を満たすとは限らず，Cが次元に依存する場合がある．（詳細が気になる方は，*Mendelson and Zhivotovskiy(2020)*をご参照ください）

先行研究との比較

先行研究との比較は表1及び表2にまとめた．本研究によって得られた2つの結果は，両者とも（定数を除いて）学習限界と一致している．特に，共変量ベクトルがsub-Gaussianに従う場合は，初めて（定数を除いて）学習限界と一致する結果を導出した．

今後の方針

今後の方針としてはβ^*にスパース性がある場合への拡張が考えられる．しかし，提案手法のstep2にL^1罰則を付与するのみだと，推定精度が$\sqrt{o}$に依存してしまい劇的に悪化することは判明している．提案手法は，重みの計算にβの情報を用いていないことから，本研究の方針を採用するのであれば，特にstep1において大きく異なる発想が要求されると考えている．

今回，sub-Gaussianかつ定数Cに対して$L^4 - L^2$等価性を満たす分布を扱ったが，定数Cに対する$L^4 - L^2$等価性の条件を取り除く方針も考えられる．本研究の提案手法で当該条件を用いない場合では，step2の解析でサンプルを$d^2 \log nd$程度を要求することとなってしまうため，step 2で異なる方法または解析手法が要求されると考えている

表 1: 共変量ベクトルがGaussianまたはsub-Gaussianの場合の先行研究との比較

	Diakonikolas et al. (2018)	Baksi and Prasad (2020)	Cherapanamjeri et al.(2020)	本研究
共変量ベクトル	Gaussian	Gaussian	sub-Gaussian with $L^4 - L^2$ equivalence	sub-Gaussian with $L^4 - L^2$ equivalence
ランダムノイズの条件	Gaussian	Gaussian	sub-Gaussian with $L^4 - L^2$ equivalence	absolute moment
計算複雑度	$\mathrm{poly}(n, d)$	$\mathrm{poly}(n, d)$	nd $\mathrm{polylog}(n, d)$	$d^3 + d^2 n$
推定精度	$\varepsilon \log \frac{1}{\varepsilon}$	$\varepsilon \log \frac{1}{\varepsilon}$	$\varepsilon \log \frac{1}{\varepsilon}$	$\varepsilon \sqrt{\log \frac{1}{\varepsilon}}$

表 2: 共変量ベクトルがheavy-tailの場合の先行研究との比較

	Baksi and Prasad (2020)	Cherapanamjeri et al.(2020)	Pensia et al.(2020)	本研究
共変量ベクトル	$L^4 - L^2$ equivalence	$L^4 - L^2$ equivalence	$L^4 - L^2$ equivalence	$L^4 - L^2$ equivalence
ランダムノイズの条件	$L^4 - L^2$ equivalence	$L^4 - L^2$ equivalence	existence of variance	absolute moment
計算複雑度	$\mathrm{poly}(n, d)$	nd $\mathrm{polylog}(n, d)$	$\mathrm{poly}(n, d)$	$d^3 + d^2 n$
推定精度	$\sqrt{\varepsilon}$	$\sqrt{\varepsilon}$	$\sqrt{\varepsilon}$	$\sqrt{\varepsilon}$