論文解説：外れ値にロバストなニューラルネットの学習

奥野彰文 [*1,2], 柳下翔太郎 [3]

[1] 統計数理研究所 統計基盤数理研究系　[2] 理化学研究所 AIP センター
[3] 統計数理研究所 リスク解析戦略研究センター

要旨

本稿は外れ値にロバストなニューラルネットの学習に関する我々のプレプリント Okuno and Yagishita (2024) に関する解説です．解説の平易さを優先するため，厳密な記述については当該プレプリントをご参照ください．

キーワード: ニューラルネット，ロバスト推定，トリム損失，高次変動正則化

1　研究背景

1.1　外れ値にロバストな推定

観測値の中に典型的なパターンから大きく外れた値 (外れ値) が含まれる場合，統計的推定の結果が強い悪影響を受けることがあります．そこで，観測値中に含まれる外れ値をうまく取り除くための方法としてロバスト推定が盛んに研究されています．本研究では特に，外れ値にロバストな回帰分析を取り上げます．

いま，共変量と応答変数に対応する確率変数のペア $(X, Y) \in \Omega \times \mathbb{R}$ があるとし，その観測値として $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ があるとしましょう．Ω は適当な領域として，ここでは簡単のために $\Omega = [-1, 1]^d$ などとしましょう．このとき，回帰分析の目標は，観測値から条件付き期待値

$$f_*(x) = \mathbb{E}[Y \mid X = x]$$

をうまく推定することです．

ここで問題になるのは観測値に含まれる外れ値です．例えばデータ成形時のミスであったり，観測の誤りであったり，さまざまな理由により意図しない外れ値が混入します．普通，回帰分析では y_i が $f_*(x_i)$ から大きくは外れないことを仮定して二乗誤差を最小化しますが，$f_*(x_i)$ から大きく外れてしまっている値を以降では外れ値と呼びましょう．観測値に意図せず外れ値が混入してしまうと，その回帰曲線が外れ値に大きな影響を受けてしまい，通常フィットするべき典型的なパターンから離れてしまうことがあります．この外れ値の悪影響を取り除こうというのがロバスト回帰です (図 1).

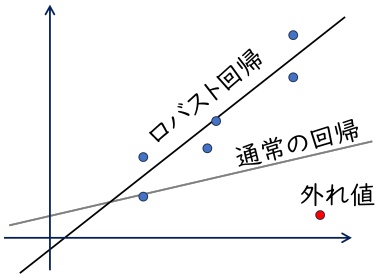


図 1: 線形回帰とロバスト線形回帰

ロバスト回帰として様々な方法が提案されています．例えば二乗誤差を絶対値損失に差し替えたり，二乗関数の裾のペナルティを緩和することで外れ値の影響を軽減する方法がよく用いられます．我々の研究では，裾を緩和するのではなく，典型的なパターンから逸脱した観測値を何割か無視するトリム損失を考えます．

ここでは回帰モデルを $f_\theta(x)$ とし，$r_i(\theta) = y_i - f_\theta(x_i)$ を残差としましょう．絶対残差を小さい順に並び変えた添え字を $(i;\theta)$ とし，つまり $|r_{(1;\theta)}(\theta)| \le |r_{(2;\theta)}(\theta)| \le \cdots \le |r_{(n;\theta)}(\theta)|$ とします．このとき，絶対残差の小さなものから h 個だけを考慮する，つまり上位 $n-h$ 個の大きな残差を無視する

$$T_h(r(\theta)) := \frac{1}{n} \sum_{i=1}^{h} |r_{(i;\theta)}(\theta)|^2 \tag{1}$$

をトリム損失と呼びます．ここで $r(\theta)$ は $r_i(\theta)$ を並べたベクトルとします．h としては例えば典型的には $h \approx 0.9n$ などの値を利用し，理想的には，観測値に混入した $n-h$ 個の外れ値を無視することができます．

1.2　ニューラルネットの学習とロバストネス

ロバスト統計においては典型的に線形回帰か，線形に準ずる方法（たとえばカーネル法など）がよく考えられていますが，ニューラルネットの学習に外れ値が含まれるとどうなるでしょうか．多くの方がご存じの通り，ニューラルネットは非常に表現能力の高い予測モデルで

* 責任著者，okuno@ism.ac.jp

あり[a]，実はトリム損失 (1) を利用しても外れ値に完全にフィットしてしまうことがあります (図 2).

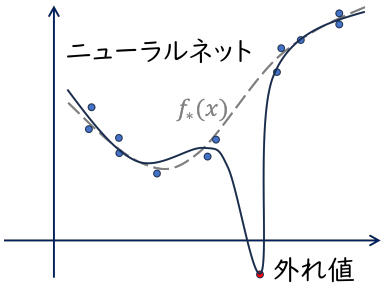


図 2: ニューラルネットを単純なロバスト損失で学習．真の関数 $f_*(x) = \mathbb{E}[Y \mid X = x]$ を学習したいが外れ値に完全に適合している．

これまで考えられてきた典型的な線形回帰などでは，予測モデルの表現能力が乏しいこともあり，外れ値に対応する観測値がトリム損失 (1) で切り捨てる $n - h$ 個の残差に対応しました．一方でニューラルネットは表現能力が高すぎるので，外れ値に完全にフィットすることが可能であり，結果として外れ値がトリム損失 (1) で切り捨てられない h 個の主要項に含まれてしまい，外れ値の影響を強く受けてしまいます．線形回帰やカーネル法などではあまり発生しない現象です．

2 本研究の貢献

2.1 高次変動正則化 (HOVR)

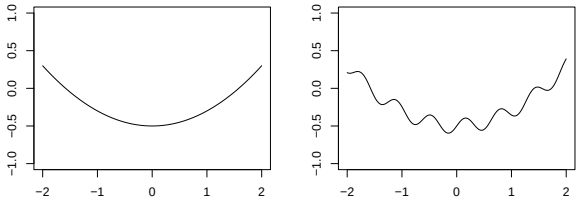
1.2 節で紹介した現象は，ニューラルネットの表現能力が高すぎることに起因しています．そこで本研究では一般の予測モデル f_θ に対して定義できる，高次変動正則化 (Higher-Order Variation Regularization; HOVR):

$$C_{k,q}(f_\theta) := \int_\Omega \left| \frac{\partial^k}{\partial x^k} f_\theta(x) \right|^q \mathrm{d}x \qquad (2)$$

を提案します．一般に予測モデルの微分を計算することは面倒ですが，ニューラルネットの微分は `Pytorch` に実装された `autograd` などを利用して容易に計算可能です．本稿では記号の氾濫を防ぐため，(2) に共変量が $d = 1$ 次元の場合の HOVR を表示していますが，一般の次元に自明に拡張できます．詳細は Okuno and Yagishita (2024) をご覧ください．

実際の曲線について，HOVR の値は図 3 のようになります．(b) は (a) と値そのものは大きく違わないものの，高次の変動が大きい関数であり，HOVR の値で比

[a] 線形の関数しか近似できない線形回帰モデルと異なり，非線形を含めた任意の連続関数を近似できるなど，いわゆる万能近似能力を持ちます．



(a) $C_{2,2}(f) \approx 0.64$.　　(b) $C_{2,2}(f) \approx 197$.

図 3: 高次変動正則化 (HOVR)

較すると約 300 倍の違いがあることがわかります．つまり，HOVR が小さくなるようニューラルネットを学習すると，図 2 のような外れ値への適合を抑えることができます．

なお，HOVR と類似した概念はたくさんあります．例えば全変動 (Total Variation) は (2) の $(k, q) = (0, 1)$ の場合に相当します．また，線形回帰やカーネル法など，パラメータと出力に線形性を持つモデルでは $C_{k,2}(f_\theta)$ がパラメータ正則化 $\|\theta\|_2^2$ と対応します．つまりカーネル法などではパラメータ正則化が変動の正則化と直結するのですが，ニューラルネットなどのより複雑なモデルでは，パラメータを正則化しても変動が正則化できるとは限りません．HOVR はニューラルネットの変動を直接的に制御できる，パラメータ正則化に代わる新しい正則化と思うことができます．

2.2 トリム損失の平易な表現

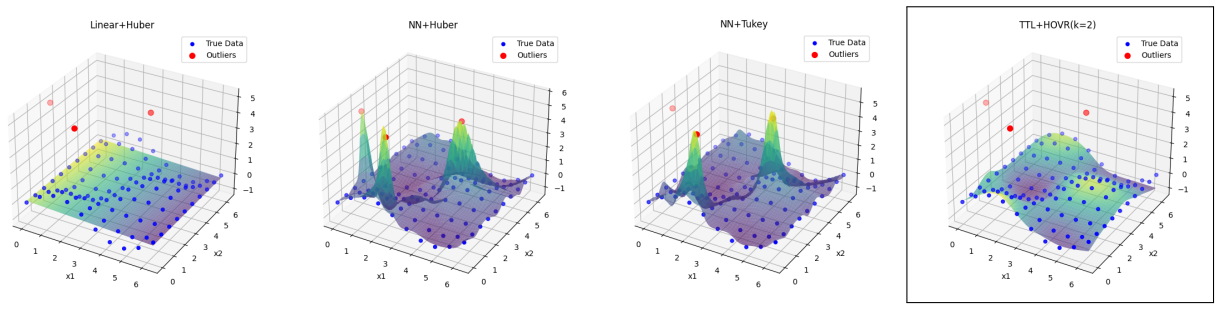
(1) はトリム損失の内部にニューラルネットのパラメータ θ が入っていて，最適化で “扱いにくい” 形なのですが，より扱いやすい形に変形できます．具体的には，余分なパラメータ $\xi \in \mathbb{R}^n$ を追加して

$$\frac{1}{2} T_h(r(\theta)) = \min_{\xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \|r(\theta) - \xi\|_2^2 + T_h(\xi) \right\} \qquad (3)$$

が成り立ちます (Yagishita, 2024)．パラメータが増えると最適化が困難になりそうに思えますが，右辺ではトリム損失 T_h とニューラルネットのパラメータ θ が分離された形になっているので，むしろ容易に最適化することができます．本研究では (3) を特に Transformed Trimmed Loss (TTL) と呼んでいます．

2.3 Augmented and Regularized Trimmed Loss

本研究では，ロバストな学習を行うためのトリム損失と，ニューラルネットが外れ値に適合しないようにする高次変動正則化 (HOVR) を組みあわせた $T_h(r(\theta)) + \lambda C_{k,q}(f_\theta)$ の最小化により，外れ値にロバストなニューラルネットの学習を行います．最適化の工夫として，特に (3) の関係性を利用すると，Augmented



(a) 線形 +Huber　(b) NN+Huber　(c) NN+Tukey　(d) **提案法**: NN+ARTL

図 4: $f_*(x) = \sin(x_1)\cos(x_2)$ に従い，ノイズを加えて観測値 $(n = 100, d = 2)$ を生成し，3% の外れ値を加えました．ニューラルネット (NN) として，入力が 2 次元，素子数が 100 の中間素子を 3 つ重ね，出力が 1 次元となる多層パーセプトロンを利用します．線形モデルおよび NN にロバスト損失 (Huber 損失，Tukey 損失) を利用した結果，および提案法であるニューラルネット (NN) に ARTL を利用して学習した結果を比較しました．
なお，(a)–(d) は Okuno and Yagishita (2024) Figure 1(a)–(d) に示された $f_*(x) = \sin(2x_1)\cos(2x_2)$ での実験を，$f_*(x) = \sin(x_1)\cos(x_2)$ に変更して実験した結果となっています．

and Regularized Trimmed Loss (ARTL):

$$F_{h,\lambda}(\theta, \xi) := \frac{1}{n}\|r(\theta) - \xi\|_2^2 + T_h(\xi) + \lambda C_{k,q}(f_\theta)$$

のパラメータ (θ, ξ) についての同時最適化が，高次変動を正則化したトリム損失の最適化と等価であることが分かります．ARTL は非常に扱いやすい関数であり，HOVR (2) の積分が Robbins and Monro (1951) などの確率的な方法により巧妙に回避することと併せて効率的に最適化することができます．本研究では特に，ARTL を最適化するための Stochastic Gradient-Supergradient Descent (SGSD) 法を提案し，その収束を理論的に証明しています．

実際に計算をした結果を図 4 に示します．(a) 線形モデルは複雑な外れ値の影響を取り除ける一方で複雑なパターンをとらえることができず，(b), (c) ニューラルネットに単純なロバスト損失を加えても外れ値に適合してしまい，(d) 提案法によりニューラルネットを学習すると非線形のパターンをとらえながら外れ値の影響を取り除くことができています．

3　まとめ

本研究では，表現能力が非常に高いニューラルネットを学習する際，外れ値が含まれるとどうなるのかについて考察し，ロバストな推定法を提案しました．提案法は非常に汎用性が高く，実験で利用したパーセプトロンに限らず，畳み込みニューラルネットや (系列変換タスクであれば) Transformer などといった他の構造のニューラルネットにもそのまま利用できることが強みです．

参考文献

Okuno, A. and Yagishita, S. (2024). Outlier-robust neural network training: Efficient optimization of transformed trimmed loss with variation regularization. *arXiv preprint arXiv:2308.02293*.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

Yagishita, S. (2024). Fast algorithm for sparse least trimmed squares via trimmed-regularized reformulation. *arXiv preprint arXiv:2410.04554*.