Robust Bayesian Regression via Hard Thresholding

Zheyi Fan[1,2], **Zhaohui Li**[3], **Qingpei Hu**[1,2]
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, China
[3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, USA.

Neurips 2022

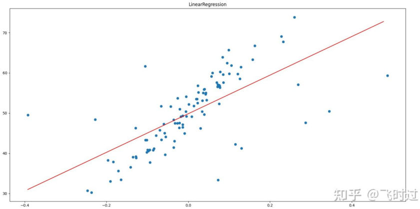
AMSS
Academy of Mathematics and Systems Science,CAS

Georgia Institute of Technology

Introduction

Ordinary Regression:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2$$



- Not stable.
- Easily affected by **outliers**.

Robust least-square regression (RLSR):

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \min_{\substack{\mathbf{w} \in \mathbb{R}^p, S \subset [n] \\ |S| = n-k}} \sum_{i \in S} (y_i - \mathbf{x}_i^T \mathbf{w})^2$$

- RLSR has excellent application value in many fields
- RLSR is hard to solve for this optimization problem is not convex

The weaknesses of previous Methods

Low breakdown point

- (McWilliams et al. 2014): $\alpha = O(1/\sqrt{d})$.
- (Prasad et al. 2018): $\alpha = O(1/\log d)$.

Can only resist OAA (Oblivious adversarial attack)

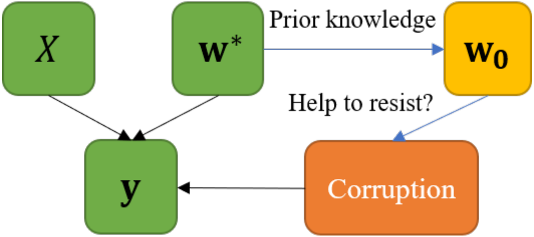
- (Bhatia et al. 2017): proposed the first efficient consistent estimator.
- (Suggala et al. 2019): α is close to 1 as $n \to \infty$.

Can resist AAA (Adaptive adversarial attack)

- Have a low breakdown point
  - (Cherapanamje et al. 2020)(Jambulapat et al.2021)(Pensia et al. 2020)
- (Bhatia et al. 2015): only in noiseless case.
- (Diakonikolas et al. 2019): requires accurate information of data covariance.

How to **increase** the breakdown point of robust regression under AAA?

Motivation



In many application scenarios, **prior knowledge** such as previous experimental data or engineering data are available.

If we do know some prior knowledge such as a rough estimation $\mathbf{w}_0$, can we use this knowledge to **enhance** the effect of robust regression?

Based Work

Consistent Robust Regression (CRR) (Bhatia et al. 2017)

Algorithm 1 CRR: Consistent Robust Regression

Input: Covariates $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, \ldots, y_n]^\top$, corruption index k, tolerance ϵ

$\mathbf{b}^0 \leftarrow \mathbf{0}, t \leftarrow 0,$
$P_X \leftarrow X^\top (XX^\top)^{-1} X$
while $\left\|\mathbf{b}^t - \mathbf{b}^{t-1}\right\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(P_X \mathbf{b}^t + (I - P_X)\mathbf{y})$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}^t \leftarrow (XX^\top)^{-1} X(\mathbf{y} - \mathbf{b}^t)$

The key step in this CRR algorithm is $\mathbf{b}^{t+1} \leftarrow HT_k(P_X \mathbf{b}^t + (I - P_X)\mathbf{y})$. This step can be divided into two sub steps:

$$\mathbf{w}^{t+1} \leftarrow (XX^T)^{-1} X(\mathbf{y} - \mathbf{b}^t)$$
$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T \mathbf{w}^{t+1})$$

This means $\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{b}^t - X^T \mathbf{w}\|^2$

How To Integrate Prior Information Into Algorithm?

CRR

$$\mathbf{w}^t = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{b}^t - X^T\mathbf{w}\|^2$$
$$\mathbf{w}^t \leftarrow (XX^T)^{-1}X(\mathbf{y} - \mathbf{b}^t)$$
$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T\mathbf{w}^t)$$

TRIP:Hard Thresholding Approach to Robust Regression with Simple Prior (Ours)

$$\mathbf{w}^t = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{b}^t - X^T\mathbf{w}\|^2 + (\mathbf{w} - \mathbf{w}_0)^T M(\mathbf{w} - \mathbf{w}_0)$$
$$\mathbf{w}^t \leftarrow (XX^T + M)^{-1}X(\mathbf{y} - \mathbf{b}^t + M\mathbf{w}_0)$$
$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T\mathbf{w}^t)$$

TRIP can be viewed as using MAP to estimate $\mathbf{w}$ by giving $\mathbf{w}$ a prior $\mathcal{N}(\mathbf{w}_0, \Sigma_0)$ and $M = (\Sigma_0/\sigma^2)^{-1}$ in Bayesian view.

The Convergence Condition

The Convergence Condition of CRR

$$2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T)} < 1$$

The Convergence Condition of TRIP

$$2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$$

Notice that $\lambda_{min}(XX^T + M) \geq \lambda_{min}(XX^T) + \lambda_{min}(M)$, so TRIP need a weaker condition for convergence than CRR.

Under the condition $\lim_{n\to\infty} \frac{\lambda_{min}(M)}{n} = \xi$, we can give an approximate expression of the breakdown point for TRIP when ξ is not too large

$$k^* \leq k \leq (0.3023 - \sqrt{0.0887 - 0.0040\xi})n$$

How To Decrease The Bias?

In the convergence theory of TRIP, there is an unavoidable bias in the estimation:

$$O\left(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}}\right)\|\mathbf{w}^*-\mathbf{w}_0\|_2$$

If every iteration step is **more robust**, will the estimation bias decrease?

BRHT

TRIP:Hard Thresholding Approach to Robust Regression with Simple Prior

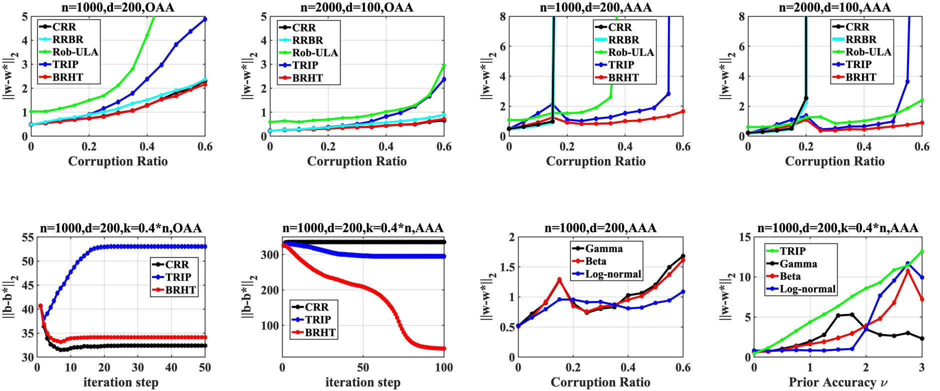
$$\mathbf{w}^t = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{b}^t - X^T\mathbf{w}\|^2 + (\mathbf{w} - \mathbf{w}_0)^T M(\mathbf{w} - \mathbf{w}_0)$$
$$\mathbf{w}^t \leftarrow (XX^T + M)^{-1}X(\mathbf{y} - \mathbf{b}^t + M\mathbf{w}_0)$$
$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T\mathbf{w}^t)$$

BRHT: Robust Bayesian Reweighting Regression via Hard Thresholding

$$(\mathbf{w}^t, \mathbf{r}^t) = \arg\max_{\mathbf{w}\in\mathbb{R}^d, \mathbf{r}\in\mathbb{R}^n_+} \log p_{\mathbf{w}}(\mathbf{w}) + \log p_{\mathbf{r}}(\mathbf{r}) + \sum_{i=1}^{n} r_i \log \ell(y_i - b_i^t \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)$$
$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T\mathbf{w}^t)$$

The estimation of $\mathbf{w}$ is based on Bayesian Reweighting from (Wang et al. 2017), which is a robust Bayesian model for estimating parameters. $\mathbf{r}$ is the weight on each point. The prior $\mathbf{w}$ is also in the form $\mathcal{N}(\mathbf{w}_0, \Sigma_0)$.

Result



Our method is more robust than other methods and shows excellent performance under complex attacks.

Conclusion

- We propose two novel robust regression algorithms TRIP and BRHT, which can tolerate a larger proportion of outliers by incorporating prior information, and BRHT further reduce the estimation error.
- We prove that both algorithms have strong theoretical guarantees and that the algorithms converge linearly under a mild condition.
- Extensive experiments have illustrated that our algorithms outperform benchmark methods in terms of both robustness and efficiency.

Thanks for Listening!