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, [2]University of Chinese Academy of Sciences, [3]Georgia Institute of Technology

Regression

Linear regression model:

$$\mathbf{y} = \mathbf{X}^T\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{x} \in [0,1]^{2\times n},$$

Ordinary Least Squares(OLS):

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}Y,$$

The most commonly used method: simple, efficient.
Robustness issue: What if the outliers exists? Only one outlier can destroy the results.

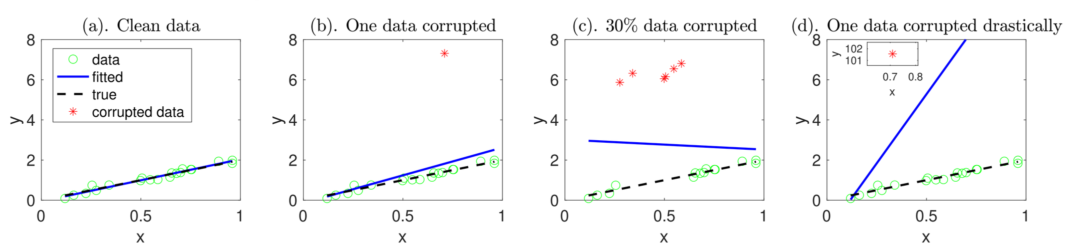

Figure 1:OLS fittings destroyed by the outliers

Research Objectives

Proposing new methods for solving regression problems:

- Robust to outliers.
- Incorporating prior knowledge.

Robust Regression

Model assumption:

$$\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (1)$$

$\mathbf{b}^*$: a k-sparse vector; non-zero elements indicate outliers.
The robust least-squares regression(RLSR) solves:

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

Goal: recover the uncorrupted point set S and the regression coefficient $\mathbf{w}^*$ simultaneously. **NP hard!**
A natural statistical interpretation: maximum likelihood estimation(MLE):

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

Incorporating Prior information, **Bayesian RLSR**: given prior $p_{\mathbf{w}}(\mathbf{w})$; Posterior:

$$p(\mathbf{y}_S, \mathbf{w} \mid X_S) = p_{\mathbf{w}}(\mathbf{w}) \prod_{i\in S} \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2) \quad (3)$$

Maximizing the log-posterior:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \max_{\mathbf{w}\in\mathbb{R}^p, |S|=n-k} \log p_{\mathbf{w}}(\mathbf{w}) + \sum_{i\in S} [\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)] \quad (4)$$

Two types of attacks are considered:
OAA (oblivious adversarial attack): The outliers are independent to the data.
AAA (adaptive adversarial attack): A more severe attack in which outliers are correlated to data.
Difficulties: Algorithms, Reducing bias caused by prior.

Algorithm: TRIP

Hard **T**hresholding approach to **R**obust regression with s**I**mple **P**rior.

- **An elegant posterior**:
$$(\hat{\mathbf{w}}, \hat{S}) = \arg \min_{\substack{\mathbf{w}\in\mathbb{R}^p, S\subset[n]\\ |S|=n-k}} \sum_{i\in S} (y_i - x_i^T\mathbf{w})^2 + (\mathbf{w}-\mathbf{w}_0)^T M(\mathbf{w}-\mathbf{w}_0),$$
given Gaussian prior $p_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \sigma^2 M^{-1})$.
- **Hard thresholding [1] operator**: $\hat{\mathbf{b}} = \mathrm{HT}_k(\mathbf{b})$, where $\hat{\mathbf{b}}_i = \mathbf{b}_i$ if $\delta_{\mathbf{r}}^{-1}(i) \le k$ and 0 otherwise.
- **Intuition**: if only k outliers, then the k elements that have the largest residues are labeled as outliers.
- **TRIP Algorithm**:

$\mathbf{b}^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$,
$P_{MX} \leftarrow X^T(XX^T + M)^{-1}X$,
$P_{MM} \leftarrow X^T(XX^T + M)^{-1}M$
while $\|\mathbf{b}^t - \mathbf{b}^{t-1}\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(P_{MX}\mathbf{b}^t + (I - P_{MX})\mathbf{y} - P_{MM}\mathbf{w}_0)$
  $t \leftarrow t+1$;
end while
return $\hat{\mathbf{w}} \leftarrow (XX^T)^{-1}X(\mathbf{y} - \mathbf{b}^t)$

Reducing Bias :BRHT

Robust **B**ayesian **R**eweighting regression via **H**ard **T**hresholding.

- **Prior are typically imprecise.**
- To reduce its influence, we introduce a localization parameter $\mathbf{r}$ that reflects the influence of each sample [4]:
$$p(\mathbf{y}, \mathbf{w}, \mathbf{r}|X) \propto p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{r}}(\mathbf{r}) \prod_{i=1}^{n} \ell(\mathbf{w}|y_i, \mathbf{x}_i, \sigma^2)^{r_i}$$
- **Intuition**: r_i associated with 'good' sample (small residuals) tends to be large, i.e., contribute more to posterior density.
- **BRHT Algorithm**:

$\mathbf{b}^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$,
while $\|\mathbf{b}^t - \mathbf{b}^{t-1}\|_2 > \epsilon$ **do**
  $\mathbf{w}^t \leftarrow \mathrm{VBEM}(X, \mathbf{y} - \mathbf{b}^t, p_{\mathbf{r}}(\mathbf{r}), p_{\mathbf{w}}(\mathbf{w}))$
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(\mathbf{y} - X^T\mathbf{w}^t)$
  $t \leftarrow t+1$;
end while
return $\hat{\mathbf{w}} \leftarrow (XX^T)^{-1}X(\mathbf{y} - \mathbf{b}^t)$

VBEM: variational Bayesian expectation maximization. Estimating $\mathbf{w}$ with existence of latent variables.

Our Contributions

- We propose new methods that incorporate prior knowledge into robust regression to increase the breakdown point.
- We derive the theoretical properties of the proposed algorithms.
- The simulation results show that our methods significantly outperform alternative methods under AAAs. Moreover, BRHT algorithm is also competitive against OAAs.

Theoretical Convergence

Theorem (Convergence of TRIP)

For break point $\|\mathbf{b}^*\|_0 \le k \cdot n$. Under mild conditions, for $k > k^*$, it is guaranteed with a probability of at least $1-\delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \le \varepsilon + O(O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of TRIP.

Theorem (Convergence of BRHT)

For break point $\|\mathbf{b}^*\|_0 \le k \cdot n$. Under mild conditions, for $k > k^*$, it is guaranteed with a probability of at least $1-\delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \le \varepsilon + O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}}) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\gamma\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of BRHT.

Simulation Studies

Benchmarks: CRR [1]; Reweighted robust Bayesian regression (RRBR) [4]; Rob-ULA [2].

- TRIP and BRHT are more robust under AAAs.
- BRHT Performs the best in all experiments.

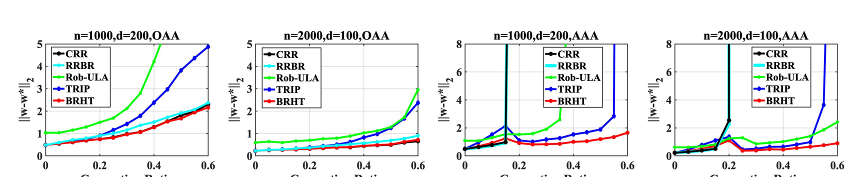

Figure 2:Recovery of parameters

- BRHT is more accurate and behaves significantly better under AAAs, while TRIP stalls during the iterative process.

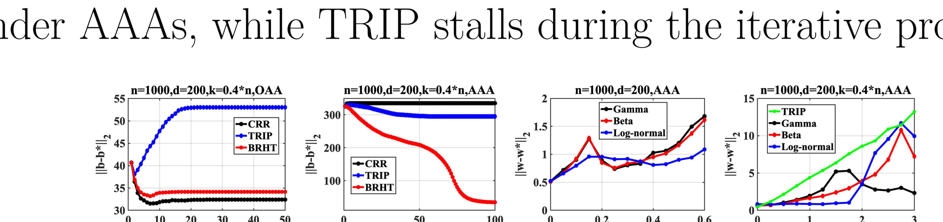

Figure 3:(a),(b), Convergence diagnostic. (c),(d), Impact of prior $p_{\mathbf{r}}(\mathbf{r})$ and $p_{\mathbf{w}}(\mathbf{w})$.

Conceptual Illustration

Recall the toy example. TRIP and BRHT can accurately detect the outliers and estimate coefficients.

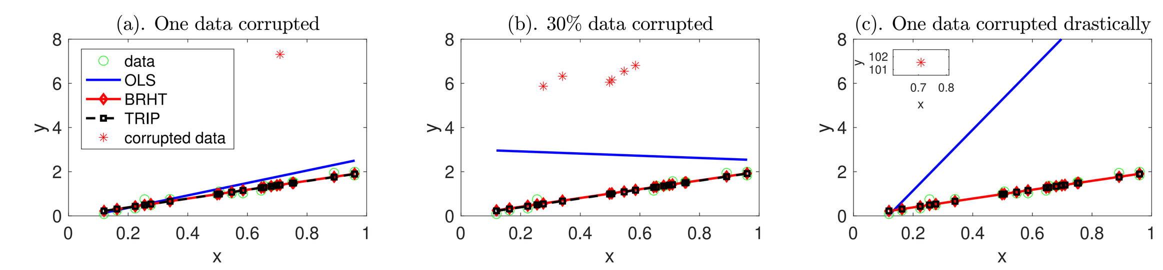

Figure 4:TRIP and BRHT fittings are perfect!

Conclusion

Two algorithms are proposed for robust regression.

- **TRIP**: By incorporating the prior knowledge, we propose robust regression via hard thresholding. The recovery of coefficients is significantly improved.
- **BRHT**: By employing Bayesian reweighting, reduce the estimation bias caused by prior bias.

Both algorithms have strong theoretical guarantees that the algorithms **converge linearly** under a mild condition.
Future research:

- Extend to situations where both $\mathbf{y}$ and X are corrupted.
- Further reduces the effect of a prior on the estimation.

References

[1] Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P.
Consistent robust regression.
Advances in Neural Information Processing Systems 30 (2017).

[2] Bhatia, K., Ma, Y.-A., Dragan, A. D., Bartlett, P. L., and Jordan, M. I.
Bayesian robustness: A nonasymptotic viewpoint.
arXiv preprint arXiv:1907.11826 (2019).

[3] Fan, Z., Li, Z., and Hu, Q.
Robust bayesian regression via hard thresholding.
Conference on Neural Information Processing Systems (2022).

[4] Wang, Y., Kucukelbir, A., and Blei, D. M.
Robust probabilistic modeling with Bayesian data reweighting.
In *International Conference on Machine Learning* (2017), PMLR, pp. 3646–3655.

AMSS Academy of Mathematics and Systems Science,CAS
中国科学院大学 University of Chinese Academy of Sciences
Georgia Institute of Technology