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.
[1,2]{fanzheyi,qingpeihu}@amss.ac.cn, [3]zhaohui.li@gatech.edu

Abstract

By combining robust regression and prior information, we develop an effective robust regression method that can resist adaptive adversarial attacks. Due to the widespread existence of noise and data corruption, it is necessary to recover the true regression parameters when a certain proportion of the response variables have been corrupted. Methods to overcome this problem often involve robust least-squares regression. However, few methods achieve good performance when dealing with severe adaptive adversarial attacks. Based on the combination of prior information and robust regression via hard thresholding from [1], this paper proposes an algorithm that improves the breakdown point when facing adaptive adversarial attacks. Furthermore, to improve the robustness and reduce the estimation error caused by the inclusion of a prior, the idea of Bayesian reweighting is used to construct a more robust algorithm. We prove the theoretical convergence of proposed algorithms under mild conditions. Extensive experiments show that, under different dataset attacks, our algorithms achieve state-of-the-art results compared with other benchmark algorithms, demonstrating the robustness of the proposed approach.

1 Introduction

Least-squares methods are widely used because of their simplicity and ease of operation. However, due to the inevitable existence of outliers, least-squares methods, such as linear regression, may cause significant bias in practical applications. Therefore, to meet the challenge of learning reliable regression coefficients in the presence of significant corruption in the response vector, this paper focuses on robust least-squares regression (RLSR). RLSR has excellent application value in many fields, such as signal processing [9][13][27], economics [24], industry [22], biology [14], remote sensing [12] and intelligent transportation [26].

Given a data matrix $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, the corresponding response vector $\mathbf{y} \in \mathbb{R}^n$, and a certain number k representing the number of corruptions in the data, the RLSR problem can be described as:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \min_{\substack{\mathbf{w}\in\mathbb{R}^p, S\subset[n] \\ |S|=n-k}} \sum_{i\in S} (y_i - \mathbf{x}_i^T \mathbf{w})^2 \tag{1}$$

That is, we aim to recover the correct point set S and the regression coefficient $\mathbf{w}^*$ simultaneously to achieve the minimum regression error. However, this problem is NP hard, so it is difficult to optimize directly [20].

To solve the above problem, a commonly used data generation model is $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$, where $\mathbf{w}^*$ is the true regression coefficient we wish to recover and $\boldsymbol{\epsilon}$ is a dense white noise vector

†Corresponding authors.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

subject to a specific distribution, that is, $\|\boldsymbol{\epsilon}\|_0 \sim n$. The vector $\mathbf{b}^*$ is k-sparse, which means that there are only k non-zero values, representing k unbounded noise terms in the response vector. After years of development, there are many ways to find a reasonable solution to the problem in Eq. (1). However, these methods typically only achieve good performance under specific conditions. The main challenge is the low breakdown point of conventional methods. The breakdown point k is a measure of robustness, which means the number of corruptions that the RLSR algorithm can tolerate. We can express k as the proportion of all data points: $k = \alpha \cdot n$. Many RLSR algorithms cannot guarantee theoretical convergence as the value of k increases. For example, McWilliams et al. [15] used weighted subsampling for linear regression, but only had a breakdown point of $\alpha = O(1/\sqrt{d})$. Prasad et al. [19] proposed a robust gradient estimator that can be applied to linear regression, but their method only tolerates corruption up to $\alpha = O(1/\log d)$. Other methods may have a higher breakdown point, but tend to assume a specific pattern of data corruption. One representative adversary model for introducing data corruption is the oblivious adversarial attack (OAA), in which the opponent generates k sparse vectors while completely ignoring X, $\mathbf{w}^*$, and $\boldsymbol{\epsilon}$. The work of Bhatia et al. [1] and Suggala et al. [21] reported excellent results against OAAs by using a novel hard thresholding method; indeed, [21] suggested that α may even get close to 1 as $n \to \infty$. The recent online fast robust regression algorithm [17] also has consistent convergence with a mild condition by using Stochastic Gradient Decent (SGD) algorithm. However, these methods cannot resist adaptive adversarial attack (AAA), in which opponents can view X, $\mathbf{w}^*$, and $\boldsymbol{\epsilon}$ before determining $\mathbf{b}^*$. Handling AAA is a challenging task, and so many methods can only guarantee a very low breakdown point, especially when the data distribution is not normal [5][10][16]. Bhatia et al. [2] proposed the thresholding operator-based robust regression method and their breakdown point reach $1/65$ for noiseless model i.e., $\boldsymbol{\epsilon} \equiv 0$. However, their method can give a consistent estimation only in noiseless case. Karmalka et al. [11] had a good result in sparse robust linear regression by applying the L_1 regression and the breakdown point of their method reaches 0.239, but their estimation is consistent only when white noise $\boldsymbol{\epsilon}$ is sparse. Diakonikolas et al. [7] considered the situation that X and $\mathbf{y}$ may have outliers simultaneously, and proposed a filter algorithm in which the error bound is $O(\alpha \log(1/\alpha)\sigma)$. However, their method requires accurate data covariance of the true data distribution or numerous unlabeled correct data to estimate the data covariance, which are often unavailable in practice.

The limitations of the above-mentioned methods can be attributed to a lack of prior knowledge from the real data, making it difficult to distinguish the set of correct points in the case of AAAs. Gülçehre et al. [8] showed that prior information effectively improves the accuracy of machine learning. In many application scenarios in industry, economics, and biology, prior knowledge such as previous experimental data or engineering data are available. The goal of this paper is to propose a new robust regression that can integrate available prior information, even if the prior information is not very accurate.

The typical approach for integrating prior information is the Bayesian method. This provides a way of formalizing the process of learning from data to update beliefs in accordance with recent notions of knowledge synthesis [6]. However, generic Bayesian method is also sensitive to outliers. Thus robust Bayesian method should be considered to produce more reliable estimates in the presence of data corruption. Polson et al. [18] used the local variance to assign each point a local parameter that makes the estimation result robust. Furthermore, Wang et al. [23] proposed a local parameterization method and used empirical Bayesian estimation to determine the global parameters. Bhatia et al. [3] proposed a Bayesian descent method using an unadjusted Langevin algorithm (ULA), which guarantees convergence in a finite number of steps. Wang et al. [25] employed Bayesian reweighting to assign different weights to samples, thus reducing the impact of outliers.

In this paper, we combine the Bayesian method with a hard thresholding method [1] and propose two algorithms, which we call TRIP and BRHT. Through assigning a simple normal prior on the coefficients, TRIP can significantly increase the breakdown point when resisting AAAs. To further improve the accuracy of estimation, we propose BRHT algorithm through applying Bayesian reweighting method [25] to coefficient estimation. Experiments show that BRHT is even more resistant to AAAs and gives lower estimation errors, demonstrating that our method achieves significantly improved robustness.

Our Contributions: The main contribution of this paper is proposing new methods that combining the prior and robust regression to increase the breakdown point when encountering AAAs. We derive the theoretical guarantees given by our proposed algorithms. Compared with the consistent robust

regression (CRR) algorithm proposed in [1], we prove that our algorithms guarantee convergence under a weaker condition, which also shows that our methods improves the breakdown point. We also establish an extended experiment to test the effectiveness of the algorithms. Compared with other basic algorithms, the experimental results show that our methods significantly outperform alternative methods under AAAs. Moreover, BRHT algorithm is also competitive against OAAs.

Paper Organization: We state the problem formulation and present some notation and tools in Section 2. In Section 3, we describe the details of our proposed TRIP and BRHT algorithms. The theoretical properties of these two algorithms are discussed in Section 4. Section 5 presents extensive experimental results that demonstrate the excellent performance of our proposed algorithms. Section 6 concludes this paper.

2 Problem Formulation

In this study, we mainly focus on the problem of RLSR under AAAs. We are given a covariant matrix $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$. The true coefficient of the regression model is denoted by $\mathbf{w}^*$. The response vector $\mathbf{y} \in \mathbb{R}^n$ is generated by:

$$\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon} \tag{2}$$

The perturbations to the response vector consist of two parts: the adversarial corruption vector introduced by $\mathbf{b}^*$, which is a k-sparse vector, and the dense white noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Our goal is to recover the true regression coefficient $\mathbf{w}^*$ while simultaneously determining the corruption set S. To illustrate our problem, we first pay attention to the standard robust regression problem in Eq. (1). From the viewpoint of probability, the problem can be transformed into a log-likelihood version:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \max_{\substack{\mathbf{w}\in\mathbb{R}^p, S\subset[n] \\ |S|=n-k}} \sum_{i\in S} \log \ell(\mathbf{w} \mid y_i, \mathbf{x}_i, \sigma^2) \tag{3}$$

We will try to convert the problem in Eq. (3) into a Bayesian version. From the Bayesian viewpoint, we consider $p_{\mathbf{w}}(\mathbf{w})$ as the prior of the coefficients in the model. In addition, we add the localization parameter $\mathbf{r}$ and its prior $p_{\mathbf{r}}(\mathbf{r})$ to reflect the change introduced by each additional sample. For any subset $S \subseteq [n]$, the distribution of all parameters and data $X_S, \mathbf{y}_S$ is:

$$p(\mathbf{y}_S, \mathbf{w}, \mathbf{r}_S \mid X_S) = p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{r}}(\mathbf{r}_S) \prod_{i\in S} \ell(y_i \mid r_i, \mathbf{w}, \mathbf{x}_i, \sigma^2) \tag{4}$$

The posterior distributions $p(\mathbf{w}, \mathbf{r}_S | X_S, \mathbf{y}_S)$ and $p(\mathbf{y}_S, \mathbf{w}, \mathbf{r}_S | X_S)$ differ by a regularization constant. We ignore this regularization constant in the posterior distribution and only consider the main terms of the parameters. We then formulate the Bayesian RLSR problem of searching for the subset and coefficients by maximizing the log-posterior:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \max_{\substack{\mathbf{w}\in\mathbb{R}^p, \mathbf{r}\in\mathbb{R}^n_+ \\ S\subset[n], |S|=n-k}} \log p_{\mathbf{w}}(\mathbf{w}) + \sum_{i\in S} [\log \ell(y_i \mid r_i, \mathbf{w}, \mathbf{x}_i, \sigma^2) + \log p_{\mathbf{r}}(r_i)] \tag{5}$$

Note that we do not add any prior on σ^2 and only treat this as an adjustable parameter. This is because, in the initial stage of the algorithm described in Section 3, the estimated σ^2 will be large due to the existence of outliers, and this will make the estimation of $\mathbf{w}$ excessively biased to the prior distribution. This bias will be harmful, especially when the prior is not sufficiently accurate. This phenomenon can be observed in Section 3.1. To prove the positive effect of the prior on the RLSR problem, we require the properties of *Subset Strong Convexity (SSC)* and *Subset Strong Smoothness (SSS)*. Given a set $S \subset [n]$, $X_S := [\mathbf{x}_{i\in S}] \in \mathbb{R}^{d\times|S|}$ signifies the matrix with columns in the set S. The smallest and largest eigenvalues of a square symmetric matrix X are denoted by $\lambda_{min}(X)$ and $\lambda_{max}(X)$.

Definition 1 (SSC Property)**.** *A matrix $X \in \mathbb{R}^{d\times n}$ is said to satisfy the SSC property at level m with constant λ_m if the following holds:*

$$\lambda_m \leq \min_{|S|=m} \lambda_{min}(X_S X_S^T) \tag{6}$$

Definition 2 (SSS Property)**.** *A matrix $X \in \mathbb{R}^{d\times n}$ is said to satisfy the SSS property at level m with constant Λ_m if the following holds:*

$$\max_{|S|=m} \lambda_{max}(X_S X_S^T) \leq \Lambda_m \tag{7}$$

Algorithm 1 TRIP: hard **T**hresholding approach to **R**obust regression with s**I**mple **P**rior

Input: Covariates $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, ..., y_n]^T$, prior knowledge $\mathbf{w}_0$, penalty matrix M, corruption index k, tolerance ϵ
Output: solution $\hat{\mathbf{w}}$

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

These two properties are proposed in [2], and are intended to standardize the generation of the data matrix so that it will not be too abnormal. They are used to prove the theorems in Section 4.

3 Methodology

We first ignore the localization parameter $\mathbf{r}$ in Eq. (5) and propose a simple method called TRIP in Section 3.1. TRIP demonstrates the effect of a prior on the hard thresholding method. To improve the robustness and the accuracy of the estimation, the BRHT algorithm is proposed in Section 3.2.

3.1 TRIP: Hard Thresholding Approach to Robust Regression with Simple Prior

We propose a robust regression algorithm called TRIP (Algorithm 1), a hard Thresholding approach to Robust regression with sImple Prior. In this subsection, only the prior $p_{\mathbf{w}}(\mathbf{w})$ is considered and the localization parameter $\mathbf{r}$ is not added to the model. We assume the variance σ^2 of ϵ_i can be set by ourselves and that the prior $p_{\mathbf{w}}(\mathbf{w})$ obeys a normal distribution $\mathcal{N}(\mathbf{w}_0, \Sigma_0)$, where $\mathbf{w}_0$ and Σ_0 are determined in advance. Through the above simple parameter settings, the problem in Eq. (5) is transformed into the problem in Eq. (1) with an additional regularization term:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \min_{\substack{\mathbf{w}\in\mathbb{R}^p, S\subset[n] \\ |S|=n-k}} \sum_{i\in S}(y_i - x_i^T\mathbf{w})^2 + (\mathbf{w} - \mathbf{w}_0)^T M(\mathbf{w} - \mathbf{w}_0) \tag{8}$$

where $M = (\Sigma_0/\sigma^2)^{-1}$. To solve this problem, we are motivated by the hard thresholding method proposed by Bhatia [1], which concentrated on recovering the errors instead of selecting the 'cleanest' set. The problem in Eq. (8) can be formulated as $\min_{\mathbf{w}\in\mathbb{R}^p, \|\mathbf{b}\|_0\le k^*} \frac{1}{2}\|X^T\mathbf{w} - (\mathbf{y} - \mathbf{b})\|_2^2 + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T M(\mathbf{w} - \mathbf{w}_0)$. Thus, if we have an estimation $\hat{\mathbf{b}}$ of the corruption vector $\mathbf{b}^*$, the estimation of $\mathbf{w}^*$ can be easily obtained by $\hat{\mathbf{w}} = (XX^T + M)^{-1}[X(\mathbf{y} - \hat{\mathbf{b}}) + M\mathbf{w}_0]$. By substituting this estimation into the optimization problem, we obtain a new formulation of the problem:

$$\min_{\|\mathbf{b}\|_0\le k^*} f(\mathbf{b}) = \frac{1}{2}\|(P_{MX} - I)(\mathbf{y} - \mathbf{b}) + P_{MM}\mathbf{w}_0\|_2^2 \tag{9}$$

where $P_{MX} = X^T(XX^T + M)^{-1}X$, $P_{MM} = X^T(XX^T + M)^{-1}M$. The hard thresholding step in the TRIP algorithm can be viewed as $\mathbf{b}^{t+1} = HT_k(\mathbf{b}^t - \nabla f(\mathbf{b}^t))$, where k is the selected corruption coefficient. The hard thresholding operator HT_k is defined as follows.

Definition 3 (Hard Thresholding)**.** *For any vector $\mathbf{r} \in \mathbb{R}^n$, let $\delta_{\mathbf{r}}^{-1}(i)$ represent the position of the i^{th} element in $\mathbf{r}$, which are arranged in descending order of magnitude. Then, for any $k < n$, the hard thresholding operator is defined as $\hat{\mathbf{r}} = HT_k(\mathbf{r})$, where $\hat{\mathbf{r}}_i = \mathbf{r}_i$ if $\delta_{\mathbf{r}}^{-1}(i) \le k$ and 0 otherwise.*

The difference between the proposed TRIP algorithm and the original CRR [1] is the form of iteration step. The iteration step in both TRIP and CRR can be expressed uniformly as $HT_k(\mathbf{y} - X^T\mathbf{w}^t)$, but $\mathbf{w}^t = (XX^T + M)^{-1}[X(\mathbf{y} - \mathbf{b}^t) + M\mathbf{w}_0]$ in TRIP and $\mathbf{w}^t = (XX^T)^{-1}X(\mathbf{y} - \mathbf{b}^t)$ in CRR. The $\mathbf{w}^t$ in CRR is just a simple least square estimation, while the prior added in TRIP can be regarded to adding a quadratic regularization in each iteration. This quadratic regularization can avoid the candidate of iteration that is too far from the prior mean, which is also helpful to ensure the

Algorithm 2 BRHT: robust **B**ayesian **R**eweighting regression via **H**ard **T**hresholding

Input: Covariates $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, ..., y_n]^T$, prior distribution $p_{\mathbf{r}}(\mathbf{r})$, $p_{\mathbf{w}}(\mathbf{w})$, corruption index k, tolerance ϵ
Output: solution $\hat{\mathbf{w}}$

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

Algorithm 3 VBEM: **V**ariational **B**ayes **E**xpectation **M**aximization

Input: Covariates $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, ..., y_n]^T$, prior distribution $p_{\mathbf{r}}(\mathbf{r})$, $p_{\mathbf{w}}(\mathbf{w})$
Output: solution $\hat{\mathbf{w}}$

repeat
  update $q(\mathbf{r})$
  update $q(\mathbf{w})$
until convergence
return $\hat{\mathbf{w}} \leftarrow$ MAP($q(\mathbf{w})$)

numerical stability of solution. Thus, as long as the prior is not mis-specified too much, TRIP will be more likely to identify the uncorrupted points, and the final result of TRIP will be more robust than CRR.

Therefore, the prior plays an important role in TRIP. However, the weight of a prior in the solution depends entirely on the matrix $M = (\Sigma_0/\sigma^2)^{-1}$. Therefore, if we use an estimation of $\hat{\sigma}^2$ to replace σ^2, or give σ^2 a prior to calculate its posterior distribution, the overestimation of σ^2 will cause a severe increase in M in the initial iteration steps. This will mislead the iteration and cause some deviation in the final results. To overcome this difficulty we can directly treat M as an adjustable parameter to control by specifying the form such as $M = sI$, where s is a positive number, and the suitable parameter can be chosen through 5-fold or 10-fold cross validation.

3.2 BRHT: Robust Bayesian Reweighting Regression via Hard Thresholding

In this subsection, we describe how the Bayesian reweighting method is combined with hard thresholding to give a more robust algorithm, BRHT (Algorithm 2), a robust Bayesian Reweighting regression via Hard Thresholding. We first introduce the reweighted probabilistic model (RPM) proposed in [25] for traditional linear regression. For the covariates X and the response $\mathbf{y}$, the RPM model can be formulated as follows:

$$p(\mathbf{y}, \mathbf{w}, \mathbf{r}|X) = \frac{1}{Z} p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{r}}(\mathbf{r}) \prod_{i=1}^{n} \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)^{r_i} \tag{10}$$

where $\mathbf{r}$ is the local weight assigned to each sample, Z is the normalizing constant, $\ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)$ represents the likelihood of the normal distribution $\mathcal{N}(\mathbf{x}_i^T\mathbf{w}, \sigma^2)$, and $p_{\mathbf{w}}(\mathbf{w})$, $p_{\mathbf{r}}(\mathbf{r})$ are the priors of $\mathbf{w}$ and $\mathbf{r}$, respectively. By ignoring the normalizing constant, the problem in Eq. (5) can be transformed into the following form under this RPM setting:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \max_{\substack{\mathbf{w} \in \mathbb{R}^p, \mathbf{r} \in \mathbb{R}_+^n \\ S \subset [n], |S| = n-k}} \log p_{\mathbf{w}}(\mathbf{w}) + \sum_{i \in S} [r_i \log \ell(y_i | \mathbf{w}, \mathbf{x}_i, \sigma^2) + \log p_{\mathbf{r}}(r_i)] \tag{11}$$

The specific form of the prior $p_{\mathbf{r}}(r_i)$ can be set to any nonnegative random variable distribution, including (but not limited to) the Gamma distribution, Beta distribution, or log-normal distribution. Here, we still use the normal distribution $\mathcal{N}(\mathbf{w}_0, \Sigma_0)$ as the form of $p_{\mathbf{w}}(\mathbf{w})$.

To solve the optimization problem in Eq. (11), we use the two-step BRHT algorithm. The key iteration step in BRHT is $\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{y} - X^T\mathbf{w}^t)$, where $\mathbf{w}^t$ is calculated by maximizing the

log-posterior of the RPM model:

$$(\mathbf{w}^t, \mathbf{r}^t) = \arg \max_{\mathbf{w}\in\mathbb{R}^d, \mathbf{r}\in\mathbb{R}^n_+} \log p_{\mathbf{w}}(\mathbf{w}) + \log p_{\mathbf{r}}(\mathbf{r}) + \sum_{i=1}^{n} r_i \log \ell(y_i - b_i^t \mid \mathbf{w}, \mathbf{x}_i, \sigma^2) \tag{12}$$

However, the direct inference of Eq. (12) is hard because of the nonconvexity of this problem. In general, the parameters in this RPM model can be divided into two parts: the global variable $\mathbf{w}$ and the local latent variable $\mathbf{r}$. To solve the inference problem of RPM, a feasible method is to use variational Bayesian expectation maximization (VBEM). We set $q(\mathbf{w}, \mathbf{r}) = q(\mathbf{w})q(\mathbf{r})$ to approximate the true posterior after several iterations of VBEM (Algorithm 3), and replace the estimation of $\mathbf{w}$ by the maximum a posteriori (MAP) estimation from the final approximate posterior. Full details of the VBEM method are given in Appendix A. It is reasonable to ask why we are using Bayesian reweighting. Note that the iteration step in the TRIP algorithm is $\mathbf{b}^{t+1} \leftarrow HT_k(P_{MX}\mathbf{b}^t + (I - P_{MX})\mathbf{y} - P_{MM}\mathbf{w}_0) = HT_k(\mathbf{y} - X^T\mathbf{w}^t)$, where $\mathbf{w}^t = (XX^T + M)^{-1}[X(\mathbf{y} - \mathbf{b}^t) + M\mathbf{w}_0]$. Although we can show that TRIP already guarantees theoretical convergence, the estimation of $\mathbf{w}^*$ in every iteration still uses least-squares with a penalty term, which is easily affected by the corrupted points. This disadvantage forces us to assign a higher weight to the prior to resist severe data corruption in the case of AAAs. However, a higher weight on the prior means a larger estimation bias. When applying the Bayesian reweighting method, the estimation in each step is more robust than the least-squares result, and thus the weight on the prior can be reduced to guide the iteration. Therefore, the estimation bias is relatively small and the results are more robust. This is reflected in the experimental results presented in Section 5.

We should also explain why we only use a prior for a few parameters. It is important to ensure that the prior weights are neither too high nor too low. As mentioned earlier, if we treat σ^2 as the parameter to be estimated, this places too much weight on the prior. We also ensure that $p_{\mathbf{w}}(\mathbf{w})$ does not create more uncertainty, such as setting $p_{\mathbf{w}}(\mathbf{w})$ to $p(\mathbf{w}, \alpha) = \mathcal{N}(\mathbf{w}_0, \alpha^{-1}\Sigma_0)Gam(\alpha|a_\alpha, b_\alpha)$. Data corruption means that the subset of the training data may vary greatly from the prior information. Thus, when calculating the posterior, the variance of $\mathbf{w}$ controlled by α will be very large to fit the data, and so the prior information $\mathbf{w}_0$ will have lower weights in the inference step. The above problems also mislead the estimation and the selection of subsets, so we do not consider the uncertainty of these quantities and simply treat them as model parameters to be set in advance. An adjustment method for all the parameters in BRHT is described in Appendix D.

4 Theoretical Convergence Analysis

In this section, we establish the convergence theory for the TRIP algorithm, and clearly explain how the prior effectively enhances the convergence of the RLSR model. Theorems 1 and 2 summarize the results. We also show the theoretical guarantee of the BRHT algorithm in Theorems 3–5, which further demonstrate the special properties achieved by using Bayesian reweighting. Before presenting the convergence result, we first introduce some notation. Let $\lambda^t := (XX^T + M)^{-1}X(\mathbf{b}^t - \mathbf{b}^*)$, $\mathbf{g} := (I - P_{MX})\epsilon$, and $\mathbf{f} := P_{MM}(\mathbf{w}^* - \mathbf{w}_0)$. Let $S_t := [n]\backslash supp(\mathbf{b}^t)$ be the chosen subset that is considered to be uncorrupted, and $I_t := supp(\mathbf{b}^t) \cup supp(\mathbf{b}^*)$.

Theorem 1. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the given data matrix and $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$ be the corrupted output with sparse corruption of $\|\mathbf{b}^*\|_0 \le k \cdot n$. For a specific positive semi-definite matrix M, X satisfies the SSC and SSS properties such that $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$. Then, if $k > k^*$, it is guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \le \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of TRIP, where $e_0 = O(\sigma\sqrt{(k + k^*)\log\frac{n}{\delta(k+k^*)}})$ under the normal design.*

Theorem 2. *Under the conditions of Theorem 1, and assuming that $\mathbf{x}_i \in \mathbb{R}^d$ are generated from the standard normal distribution, if $k > k^*$, it is guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, the current estimation coefficient $\mathbf{w}_{T_0}$ satisfies $\|\mathbf{w}_{T_0} - \mathbf{w}^*\|_2 \le O(\frac{1}{\sqrt{n}})(\varepsilon + e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ steps.*

For positive semi-definite matrices XX^T and M, $\lambda_{min}(XX^T + M) \ge \lambda_{min}(XX^T) + \lambda_{min}(M)$. Thus, the condition $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$ in Theorem 1 is weaker than the condition $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T)} <$

1 of Lemma 5 of Bhatia [1], which shows that a prior can effectively improve the convergence of the algorithm. Assigning a higher weight to a prior means that M has larger eigenvalues, so that the convergence condition will be more easily satisfied. As a result, the TRIP algorithm can tolerate a higher proportion of outliers than the CRR method of Bhatia [1] and achieves a higher breakdown point. In fact, 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: $k^* \leq k \leq (0.3023 - \sqrt{0.0887 - 0.0040\xi})n$. Details can be found in Appendix C.2. However, the improved convergence comes at the cost of an unavoidable reduction in precision. This can be seen from Theorem 2. If the data corruption is such that k^* is $O(n)$ and the maximum eigenvalue of M is also $O(n)$, then the bias of $\hat{\mathbf{w}}$ cannot be decreased by adding more samples, which shows that there is a trade-off between convergence and accuracy. A reliable prior improves both accuracy and convergence because it has a higher weight. However, an inaccurate prior can also be helpful as long as it is quite different from the distribution of outliers, and the convergence can be improved through a prior with a low weight.

To prove the properties of our BRHT algorithm, we define the following two intermediate variables to simplify the description:

$$U(\mathbf{w}, \mathbf{r}, S) = \log p_{\mathbf{w}}(\mathbf{w}) + \sum_{i \in S} [\log p_{\mathbf{r}}(r_i) + r_i \log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)] \tag{13}$$

$$M(\mathbf{w}, \mathbf{r}, \mathbf{b}) = \log p_{\mathbf{w}}(\mathbf{w}) + \sum_{i} [\log p_{\mathbf{r}}(r_i) + r_i \log \ell(y_i - b_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)] \tag{14}$$

Theorem 5. *Suppose that the prior of r_i is independently and identically distributed (iid). We consider the t^{th} iteration step of the BRHT algorithm, where $\mathbf{w}_t, \mathbf{r}_t = \arg\max_{\mathbf{w}\in\mathbb{R}^d, \mathbf{r}\in\mathbb{R}^n_+} M(\mathbf{w}, \mathbf{r}, \mathbf{b}_t)$ and $\mathbf{b}_t = HT_k(\mathbf{y} - X^T\mathbf{w}_{t-1})$ is obtained from the hard thresholding step. Then, we have that $U(\mathbf{w}_t, \mathbf{r}_t, S_{t+1}) \geq U(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, S_t)$.*

Theorem 6. *Consider a data matrix X and a specific positive semi-definite matrix M satisfying the SSC and SSS properties such that $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$. Then, there exist $\alpha > 0$ and $0 < \gamma \leq 1 + \epsilon$, where ϵ is a small number, such that if $k > k^*$ and Σ in the prior $p_{\mathbf{w}}(\mathbf{w})$ is $\alpha\sigma^2 M^{-1}$, it is guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \leq \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\gamma\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of BRHT, where $e_0 = O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})$ under the normal design.*

Theorem 7. *Under the conditions of Theorem 4 and with $\mathbf{x}_i \in \mathbb{R}^d$ generated from the standard normal distribution, there exist $\alpha > 0$ and $0 < \gamma \leq 1 + \epsilon$, where ϵ is a small number, such that if $k > k^*$ and Σ in the prior $p_{\mathbf{w}}(\mathbf{w})$ is $\alpha\sigma^2 M^{-1}$, it can be guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, the current estimation coefficient $\mathbf{w}_{T_0}$ satisfies $\|\mathbf{w}_{T_0} - \mathbf{w}^*\|_2 \leq O(\frac{1}{\sqrt{n}})(\varepsilon + e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\gamma\|\mathbf{b}^*\|_2}{\varepsilon}))$ steps.*

Theorem 5 shows that our BRHT algorithm is reasonable because it optimizes the problem in Eq. (11) in each step. Theorems 6 and 7 guarantee the convergence of the parameter, which means that if we introduce a prior $p_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \Sigma_0)$ to the TRIP algorithm, it is convergent. There then exists some $\alpha > 0$ such that the prior $p_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \alpha\Sigma_0)$ in the BRHT algorithm guarantees convergent parameters and the bias of the estimation of $\mathbf{w}^*$ will be γ times that of TRIP. Note that α is usually relatively large in practice, which causes a lower prior weight in BRHT. Thus, when the convergence can be guaranteed, the bias of the estimator $\hat{\mathbf{w}}$ can be significantly reduced because BRHT assigns a lower weight to the prior. Even if the data are seriously corrupted, BRHT can ensure good results without significant error. All proofs of these theorems are given in Appendix C.

5 Experiments

In this section, we first consider how to effectively 'corrupt' the dataset using two different attacks: OAA and AAA. We then report an extensive experimental evaluation to verify the robustness of the proposed methods.

Algorithm 4 ADCA: **A**daptive **D**ata **C**orruption **A**lgorithm

Input: Covariates $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, ..., y_n]^T$, true parameter $\mathbf{w}^*$ penalty coefficient δ, corruption index k, tolerance ϵ
Output: solution $\hat{\mathbf{w}}$

$\mathbf{b}^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$,
$P_{\delta X} \leftarrow X^T(XX^T - \delta I)^{-1}X$, $P_\delta \leftarrow X^T(XX^T - \delta I)^{-1}\delta I$
while $\|\mathbf{b}^t - \mathbf{b}^{t-1}\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow HT_k(P_{\delta X}\mathbf{b}^t + (I - P_{\delta X})\mathbf{y} + P_\delta \mathbf{w}^*)$
  $t \leftarrow t + 1$;
end while
$\hat{\mathbf{w}} \leftarrow (XX^T)^{-1}X(\mathbf{y} - \mathbf{b}^t)$
$C \leftarrow supp(\mathbf{b}^t)$
return $\mathbf{y}_C = X_C^T\hat{\mathbf{w}}$

5.1 Data and Metrics

In our experiments, the data generation can be divided into two steps. First, we generate the basic model. The true coefficient $\mathbf{w}^*$ is chosen to be a random unit norm vector. The covariant $\mathbf{x}_i$ are iid in $\mathcal{N}(0, I_d)$. The data are generated by $y_i = \mathbf{x}_i^T w^* + \epsilon_i$, where ϵ_i are iid in $\mathcal{N}(0, \sigma^2)$. We set $\sigma = 1$ in the experiments. The second step is to generate the corrupted data using two kinds of attacks: OAA and AAA, as described in Section 5.2. The aim is to produce k^* corrupted responses in the whole dataset. The prior coefficient $\mathbf{w}_0$ is generated by $\mathbf{w}^* + \nu\mathbf{u}$, where $\mathbf{u}$ is a random unit norm vector and ν is a non-negative number (ν is set to 0.5 unless otherwise stated). Σ_0 takes the form sI, where s takes a different value for each method. All parameters are fixed in each experiment.

Following the setting in [1], we measured the performance of the regression coefficients by the standard L_2 error: $r_{\hat{\mathbf{w}}} = \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$. To judge whether the algorithm had converged, we used the termination criterion $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_2 \leq 10^{-4}$. All results were averaged over 10 runs.

5.2 Corruption Methods

To demonstrate the efficiency of our proposed methods, we apply two different attacks to the dataset: OAA and AAA. The details of these two attacks are shown as follows.

OAA: The set of corrupted points S is selected as a uniformly random k-sized subset of $[n]$, and the corresponding response variables are set as $y_i = \mathbf{x}_i^T\mathbf{w}^* + b_i + \epsilon_i$, where b_i are sampled from the uniform distribution $U[0, 10]$ and the white noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

AAA: We use all information from the true data distribution to corrupt the data, and propose an adaptive data corruption algorithm (ADCA). This algorithm is quite similar to TRIP; full details of ADCA (Algorithm 4) are given in Appendix B. δ is set to $0.1n$ for $n = 1000$, $p = 200$, and to $0.2n$ for $n = 2000$, $p = 100$.

Both TRIP and BRHT employ the prior $p_{\mathbf{w}}(\mathbf{w})$ while the variance Σ_0 of $p_{\mathbf{w}}(\mathbf{w})$ in OAA is set to be four times higher than that in AAA, which means that the prior has a higher weight in AAA. Other parameters for these two algorithms are fixed. The prior distribution $p_{\mathbf{r}}(\mathbf{r})$ is set to the Gamma distribution unless otherwise stated.

We also design another leverage point attack (LPA) to test the robustness of our methods. More results can be seen in Appendix E.

5.3 Methods Comparison

Our methods are compared with three baselines: 1) CRR [1] is an effective robust regression method in cases where there are large numbers of random outliers; 2) Reweighted robust Bayesian regression (RRBR) [25] allows us to judge whether our proposed methods are better than the original method; 3) Rob-ULA [3] is an effective robust Bayesian inference method that approximately converges to the real posterior distribution in a finite number of steps in the presence of outliers. The parameters of RRBR are the same as in the BRHT algorithm, except for the hard thresholding part. For more comparisons of other methods, please see Appendix E.

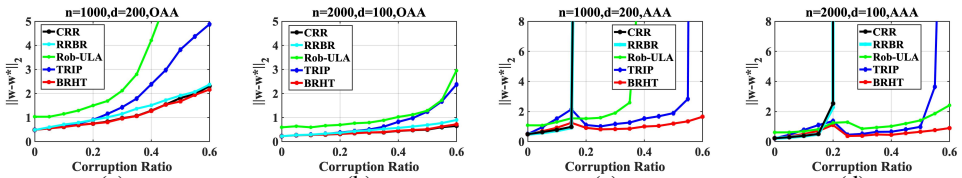

Figure 1: Recovery of parameters with respect to the number of data points n, dimensionality d, and corruption ratio α. TRIP and BRHT are more robust under AAAs than CRR, and BRHT exhibits the best performance in all experiments. RRBR and Rob-ULA show some robustness, but offer slightly worse recovery in some experiments.

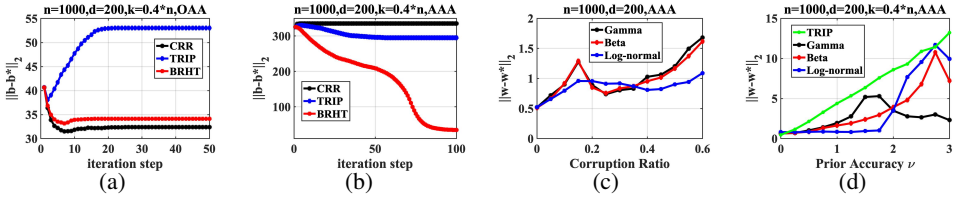

Figure 2: (a), (b) show the convergence characteristics of TRIP and BRHT. Both algorithms exhibit an estimation bias as the price of adding a prior in OAA, but BRHT is more accurate and behaves significantly better in AAA, while TRIP stalls during the iterative process. (c), (d) show the convergence under different weights of the prior $p_{\mathbf{r}}(\mathbf{r})$ and the coefficient prior $p_{\mathbf{w}}(\mathbf{w})$.

5.4 Recovery Properties of Coefficients and Uncorrupted Sets

CRR, RRBR, and Rob-ULA are excellent robust regression methods or Bayesian inference methods, but they all show their limitations in the face of different kinds of attacks. CRR achieves the best performance in the face of OAAs because it is theoretically unbiased, but it collapses rapidly when facing AAAs, as shown in Figures 1(c) and 1(d). RRBR and Rob-ULA take priors into consideration, but RRBR cannot resist AAAs and Rob-ULA produces poor results under OAAs, as shown in Figures 1(a)–1(d). TRIP produces a good effect against AAAs, while BRHT is not only optimal against AAAs, but also displays a similar effect to CRR in the case of OAAs. This shows that BRHT is the most robust algorithm among those compared in this experiment.

The TRIP and BRHT algorithms are compared in Figures 2(a) and 2(b). The TRIP method incorporates too much prior information in the case of OAAs, resulting in a greater estimation error than those of BRHT and CRR as shown in Figure 2(a). However, under AAAs, the prior information of TRIP and BRHT is enhanced by a factor of four, as described in Section 5.2. BRHT converges under a weaker prior, while TRIP becomes trapped around a local optimum. This shows that BRHT only needs to integrate weak prior information to ensure convergence. Figures 2(c) and 2(d) illustrate the convergence properties under different weights of the prior $p_{\mathbf{r}}(\mathbf{r})$ and coefficient prior $p_{\mathbf{w}}(\mathbf{w})$. Figure 2(c) shows that BRHT is not especially sensitive to the weight prior $p_{\mathbf{r}}(\mathbf{r})$ when this prior is relatively reliable. A log-normal distribution is the best choice when the prior $p_{\mathbf{w}}(\mathbf{w})$ is relatively close to the real parameters and a Gamma distribution is more robust when the prior is imprecise, as shown in Figure 2(d).

6 Conclusion

This paper has described a novel robust regression algorithm named TRIP that achieves strong results in terms of resisting AAAs. By adding a prior to the robust regression via hard thresholding, the recovery of coefficients is significantly improved. Another algorithm, named BRHT, was designed to improve the robustness of TRIP and reduce the estimation error through the use of Bayesian reweighting regression. 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.

There are several interesting future directions to extend current work. Firstly, in this article, we only consider the case when $\mathbf{y}$ is corrupted. One would consider using the prior information to better deal with the problem where both $\mathbf{y}$ and X are corrupted. Secondly, it would be also interesting to further reduce the effect of a prior on the estimation to make it consistent.

Acknowledgments and Disclosure of Funding

This work was partly supported by National Key Research and Development Program of China (2021YFA1000300 and 2021YFA1000301), National Center for Mathematics and Interdisciplinary Sciences, and Key Laboratory of Systems and Control of CAS.

References

[1] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in Neural Information Processing Systems*, 28, 2015.

[3] Kush Bhatia, Yi-An Ma, Anca D Dragan, Peter L Bartlett, and Michael I Jordan. Bayesian robustness: A nonasymptotic viewpoint. *arXiv preprint arXiv:1907.11826*, 2019.

[4] Samprit Chatterjee and Ali S Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical science*, pages 379–393, 1986.

[5] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I. Jordan, Nicolas Flammarion, and Peter L. Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08317*, 2020.

[6] Peter Congdon. *Bayesian statistical modelling*. John Wiley & Sons, 2007.

[7] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[8] Çalar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research*, 17(1):226–257, 2016.

[9] Chao Huang, Dong Wang, and Nitesh V Chawla. Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. *IEEE Transactions on Big Data*, 6(4):702–713, 2017.

[10] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. In *Advances in Neural Information Processing Systems*, volume 34, pages 4475–4488. Curran Associates, Inc., 2021.

[11] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. *arXiv preprint arXiv:1809.08055*, 2018.

[12] Xi Li, Xiaoling Chen, Yousong Zhao, Jia Xu, Fengrui Chen, and Hui Li. Automatic intercalibration of night-time light imagery using robust regression. *Remote sensing letters*, 4(1):45–54, 2013.

[13] Ming Lin, Xiaomin Song, Qi Qian, Hao Li, Liang Sun, Shenghuo Zhu, and Rong Jin. Robust Gaussian process regression for real-time high precision GPS signal enhancement. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2838–2847, 2019.

[14] Vanda M Lourenço, Ana M Pires, and M Kirst. Robust linear regression methods in association studies. *Bioinformatics*, 27(6):815–821, 2011.

[15] Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. *Advances in Neural Information Processing Systems*, 27, 2014.

[16] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[17] Scott Pesme and Nicolas Flammarion. Online robust regression via SGD on the l_1 loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 2540–2552. Curran Associates, Inc., 2020.

[18] Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9(105):501–538, 2010.

[19] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[20] Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bolcskei. Recovery of sparsely corrupted signals. *IEEE Transaction on Information Theory*, 58(5):31153130, may 2012.

[21] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2892–2897. PMLR, 25–28 Jun 2019.

[22] C Ming Wang, Dominic F Vecchia, Matt Young, and Nathan A Brilliant. Robust regression applied to optical-fiber dimensional quality control. *Technometrics*, 39(1):25–33, 1997.

[23] Chong Wang and David M Blei. A general method for robust Bayesian modeling. *Bayesian Analysis*, 13(4):1163–1191, 2018.

[24] Donglin Wang, Don Hong, and Qiang Wu. Prediction of loan rate for mortgage data: Deep learning versus robust regression. *Computational Economics*, pages 1–14, 2022.

[25] Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with Bayesian data reweighting. In *International Conference on Machine Learning*, pages 3646–3655. PMLR, 2017.

[26] He Yan, Yong Qi, Qiaolin Ye, and Dong-Jun Yu. Robust least squares twin support vector regression with adaptive foa and pso for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[27] Abdelhak M Zoubir, Visa Koivunen, Esa Ollila, and Michael Muma. *Robust statistics for signal processing*. Cambridge University Press, 2018.

A Details of Variational Bayesian EM Method

For the RPM model and the given covariates $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, ..., y_n]^T$, and prior distributions $p_{\mathbf{r}}(\mathbf{r})$, $p_{\mathbf{w}}(\mathbf{w})$, the posterior of this RPM model is formulated as

$$\log p_{\mathbf{w}}(\mathbf{w}) + \log p_{\mathbf{r}}(\mathbf{r}) + \sum_{i=1}^{n} r_i \log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)$$

where $\ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)$ represent the likelihood of the normal distribution $\mathcal{N}(x_i^T\mathbf{w}, \sigma^2)$. $p_{\mathbf{w}}(\mathbf{w})$, $p_{\mathbf{r}}(\mathbf{r})$ are the priors of $\mathbf{w}$ and $\mathbf{r}$. $p(\mathbf{w})$ is the density of the normal distribution $\mathcal{N}(\mathbf{w}_0, \Sigma_0)$. We now use variational Bayesian EM to approximate the true posterior.

$$q(\mathbf{w}) \prod_i q(r_i) \approx p(\mathbf{w}, \mathbf{r}|\mathbf{y}, X, \sigma^2)$$

In the following, we derive update equations for these variational parameters.

A.1 Derivation of $q(\mathbf{r})$ (variational E step)

Ignoring terms that do not involve $\mathbf{r}$, we take the expectations of over the remaining terms. We have

$$\begin{aligned}\log q(\mathbf{r}) &= \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{y}, \mathbf{w}, \mathbf{r}|X)] + const \\ &= \log p_{\mathbf{r}}(\mathbf{r}) + \sum_i r_i \mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i|\mathbf{w}, \mathbf{x}_i, \sigma^2)] + const\end{aligned}$$

where $q(\mathbf{w})$ is the form of the normal distribution $\mathcal{N}(\mathbf{w}_N, V_N)$, as will be shown in the derivation of $q(\mathbf{w})$. Using this fact, we have

$$\begin{aligned}\mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i|\mathbf{w}, \mathbf{x}_i, \sigma^2)] &= \mathbb{E}_{q(\mathbf{w})}[-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{1}{2}\log(2\pi\sigma^2)] \\ &= -\frac{1}{2\sigma^2}[(y_i - \mathbf{w}_N^T\mathbf{x}_i)^2 + \mathbf{x}_i^T V_N \mathbf{x}_i] - \frac{1}{2}\log(2\pi\sigma^2)\end{aligned}$$

When the priors $p_{\mathbf{r}}(r_i)$ are independent of each other, then

$$q(r_i) \propto \exp\{\log p_{\mathbf{r}}(r_i) + r_i \mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)]\}$$

If the distribution of $q(r_i)$ is complex, we use a Markov chain Monte Carlo method to simulate the distribution. An easy example for $p_{\mathbf{r}}(r_i)$ is the Gamma distribution $Gam(r_i|a_r, b_r)$. Then,

$$\begin{aligned}q(r_i) &= Gam(r_i|a_N^i, b_N^i) \\ a_N^i &= a_r \\ b_N^i &= b_r - \mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)] \\ \mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)] &= -\frac{1}{2\sigma^2}[(y_i - \mathbf{w}_N^T\mathbf{x}_i)^2 + \mathbf{x}_i^T V_N \mathbf{x}_i] - \frac{1}{2}\log(2\pi\sigma^2)\end{aligned}$$

for which the expectation of r_i under the distribution of $q(r_i)$ can be easily obtained by a_N^i/b_N^i.

A.2 Derivation of $q(\mathbf{w})$ (variational M step)

$$\begin{aligned}\log q(\mathbf{w}) &= \mathbb{E}_{q(\mathbf{r})}[\log p(\mathbf{y}, \mathbf{w}, \mathbf{r}|X)] + const \\ &= \log p_{\mathbf{w}}(\mathbf{w}) + \sum_i \mathbb{E}_{q(\mathbf{r})}(r_i) \log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2) + const \\ &= -\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T\Sigma^{-1}(\mathbf{w} - \mathbf{w}_0) - \frac{1}{2\sigma^2}(\mathbf{y} - X^T\mathbf{w})^T E_r(\mathbf{y} - X^T\mathbf{w}) + const \\ &= -\frac{1}{2}(\mathbf{w} - \mathbf{w}_N)^T V_N^{-1}(\mathbf{w} - \mathbf{w}_N) + const\end{aligned}$$

where E_r is a matrix with diagonal entries of $E_{q(\mathbf{r})}(\mathbf{r})$ and off-diagonal elements of 0. $V_N^{-1} = \frac{1}{\sigma^2}XE_rX^T + \Sigma^{-1}$, $\mathbf{w}_N = V_N(\frac{1}{\sigma^2}XE_r\mathbf{y} + \Sigma^{-1}\mathbf{w}_0)$, and $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}_N, V_N)$.

After several iterations, we use $q(\mathbf{w})\prod_i q(r_i)$ to approximate the true posterior and take $\mathbf{w}_N$ as the MAP estimate of $q(\mathbf{w})$.

B Adaptive Data Corruption Method

As the original CRR algorithm can tolerate OAAs to a significant degree, we must use all information contained in the data to fool the estimator. To achieve this goal, we need to find the most suitable subset to corrupt such that the rest of the data can more likely be generated from a completely different distribution. This problem can be formulated as follows:

$$(\hat{\mathbf{w}}, \hat{S}) = \arg \min_{\substack{\mathbf{w}\in\mathbb{R}^p, S\subset[n] \\ |S|=n-k}} \sum_{i\in S} (y_i - x_i^T \mathbf{w})^2 - \delta \|\mathbf{w} - \mathbf{w}^*\|_2^2 \tag{15}$$

where δ is the penalty coefficient that determines the extent to which the parameter leaves the standard value. $\hat{S}$ is the chosen subset that cannot be corrupted. If δ is not very large, then $\sum_{i\in\hat{S}}(y_i - x_i^T\mathbf{w}^*)^2$ will be similar to $\sum_{i\in\hat{S}}(y_i - x_i^T\hat{\mathbf{w}})^2$, and so $\hat{\mathbf{w}}$ can fool the regression model into thinking that $\hat{\mathbf{w}}$ is the true parameter. From this analysis, we find that $\hat{\mathbf{w}}$ can be used to construct the corrupted data. After getting the covariates of the corrupted data, we can define the response of the corrupted data as $y_{c_i} = \mathbf{x}_{c_i}^T \hat{\mathbf{w}}$, where $\mathbf{x}_{c_i}$ is the i^{th} covariate of the corrupted data.

The problem in Eq. (15) is very similar to that in Eq. (8), where M is replaced by $-\delta I$ and $\mathbf{w}_0$ is replaced by $\mathbf{w}^*$. Hence, these two problems can be solved by the same method. Similar to TRIP, we proposed an adaptive data corruption algorithm (ADCA) to solve the corruption problem by replacing some parameters in TRIP. ADCA seriously destroys the data, and when the corruption ratio increases, the solution of Eq. (1) may not be close to the true parameter. However, we will see that, even in this situation, TRIP and BRHT achieve good performance, as shown in Section 5.

C Supplementary Material for Proofs of TRIP and BRHT Algorithms

C.1 SSC/SSS guarantees

In this section, we introduce some theoretical properties of SSC and SSS from [2], which will be used for the convergence analysis of the proposed algorithms.

Definition 4. *A random variable $x \in \mathbb{R}$ is called sub-Gaussian if the following quantity is finite*

$$\sup_{p\geq 1} p^{-1/2} (E[|x|^p])^{1/p}$$

Moreover, the smallest upper bound on this quantity is referred to as the sub-Gaussian norm of x and denoted as $\|x\|_{\psi_2}$

Definition 5. *A vector-valued random variable $\mathbf{x} \in \mathbb{R}^d$ is called sub-Gaussian if its unidimensional marginals $\langle \mathbf{x}, \mathbf{v}\rangle$ are sub-Gaussian for all $\mathbf{v} \in S^{d-1}$. Moreover, its sub-Gaussian norm is defined as follows*

$$\|x\|_{\psi_2} = \sup_{\mathbf{v}\in S^{d-1}} \|\langle \mathbf{x}, \mathbf{v}\rangle\|_{\psi_2}$$

Lemma 8. *Let $X \in \mathbb{R}^{d\times n}$ be a matrix whose columns are sampled i.i.d from a standard Gaussian distribution i.e. $\mathbf{x}_i \sim \mathcal{N}(0, I)$. Then for any $\epsilon > 0$, with probability at least $1 - \delta$, X satisfies*

$$\lambda_{max}(XX^T) \leq n + (1-2\epsilon)^{-1}\sqrt{cnd + c'n\log\frac{2}{\delta}}$$

$$\lambda_{min}(XX^T) \geq n - (1-2\epsilon)^{-1}\sqrt{cnd + c'n\log\frac{2}{\delta}}$$

where $c = 24e^2 log\frac{3}{\epsilon}$ and $c' = 24e^2$.

Theorem 9. *Let $X \in \mathbb{R}^{d\times n}$ be a matrix whose columns are sampled i.i.d from a standard Gaussian distribution i.e. $\mathbf{x}_i \sim \mathcal{N}(0, I)$. Then for any $k > 0$, with probability at least $1 - \delta$, the matrix X satisfies the SSC and SSS properties with constants*

$$\Lambda_k \leq k(1 + 3e\sqrt{6\log\frac{en}{k}}) + O(\sqrt{nd + n\log\frac{1}{\delta}})$$

$$\lambda_k \geq n - (n-k)(1 + 3e\sqrt{6\log\frac{en}{n-k}}) - \Omega(\sqrt{nd + n\log\frac{1}{\delta}})$$

Lemma 10. *Let $X \in \mathbb{R}^{d\times n}$ be a matrix with columns sampled from some sub-Gaussian distribution with sub-Gaussian norm K and convariance Σ. Then for any $\delta > 0$, with probability at least $1-\delta$, each of the following statements holds true:*

$$\lambda_{max}(XX^T) \le \lambda_{max}(\Sigma)\cdot n + C_K \cdot \sqrt{dn} + t\sqrt{n}$$
$$\lambda_{min}(XX^T) \ge \lambda_{min}(\Sigma)\cdot n - C_K \cdot \sqrt{dn} - t\sqrt{n}$$

where $t = \sqrt{\frac{1}{c_K}\log\frac{2}{\delta}}$ and c_K, C_K are absolute constants that depend only on the sub-Gaussian norm K of the distribution.

C.2 Convergence Proof for TRIP

Theorem 1. *Let $X = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the given data matrix and $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$ be the corrupted output with sparse corruptions of $\|\mathbf{b}^*\|_0 \le k\cdot n$. For a specific positive semi-definite matrix M, the data matrix X satisfies the SSC and SSS properties such that $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$. Then, if $k > k^*$, it is guaranteed with a probability of at least $1-\delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \le \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of TRIP, where $e_0 = O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})$ under the normal design.*

Proof. First, we consider the iteration of the TRIP algorithm:

$$\mathbf{b}^{t+1} \leftarrow HT_k(P_{MX}\mathbf{b}^t + (I - P_{MX})y - P_{MM}\mathbf{w}_0)$$

After considering $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$, the iteration step can be rewritten as:

$$\mathbf{b}^{t+1} \leftarrow HT_k(\mathbf{b}^* + X^T\boldsymbol{\lambda}^t + \mathbf{g} + \mathbf{f})$$

where

$$\begin{aligned}\boldsymbol{\lambda}^t &= (XX^T + M)^{-1}X(\mathbf{b}^t - \mathbf{b}^*)\\ \mathbf{g} &= (I - P_{MX})\boldsymbol{\epsilon}\\ \mathbf{f} &= P_{MM}(\mathbf{w}^* - \mathbf{w}_0)\end{aligned}$$

Because $k > k^*$, we use the property of the hard thresholding step:

$$\begin{aligned}\|\mathbf{b}^{t+1}_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^T_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}} + \mathbf{f}_{I^{t+1}})\|_2 &\le \|\mathbf{b}^*_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^T_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}} + \mathbf{f}_{I^{t+1}})\|_2\\ &= \|X^T_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}} + \mathbf{f}_{I^{t+1}}\|_2\end{aligned}$$

Using the trigonometric inequality:

$$\|\mathbf{b}^{t+1}_{I^{t+1}} - \mathbf{b}^*_{I^{t+1}}\|_2 \le 2\|X^T_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}} + \mathbf{f}_{I^{t+1}}\|_2 \le 2\|X^T_{I^{t+1}}\boldsymbol{\lambda}^t\|_2 + 2\|\mathbf{g}_{I^{t+1}}\|_2 + 2\|\mathbf{f}_{I^{t+1}}\|_2$$

Through the SSS and SSC properties of X, we obtain:

$$\begin{aligned}\|X^T_{I^{t+1}}\boldsymbol{\lambda}^t\|_2 &= \|X^T_{I^{t+1}}(XX^T + M)^{-1}X(\mathbf{b}^{t+1} - \mathbf{b}^*)\|_2\\ &= \|X^T_{I^{t+1}}(XX^T + M)^{-1}X_{I^t}(\mathbf{b}^{t+1}_{I^t} - \mathbf{b}^*_{I^t})\|_2\\ &\le \frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T + M)}\|\mathbf{b}^{t+1}_{I^t} - \mathbf{b}^*_{I^t}\|_2\\ &= \frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T + M)}\|\mathbf{b}^{t+1} - \mathbf{b}^*\|_2\end{aligned}$$

According to Bhatia [1], there is a probability of at least $1-\delta$ that, for any set S of size up to $k+k^*$, we can find a uniform bound:

$$\|\boldsymbol{\epsilon}_S\|_2 \le \sigma\sqrt{k+k^*}\sqrt{1 + 2e\sqrt{6\log\frac{en}{\delta(k+k^*)}}} \doteq e_0$$

As for $\|X\boldsymbol{\epsilon}\|_2$, Bhatia [1] gives a consistent bound of $\|X\boldsymbol{\epsilon}\|_2^2 \leq 2\sigma^2\|X\|_F^2 \log(\frac{d}{\delta}) \leq 2\sigma^2 d\Lambda_n \log(\frac{d}{\delta})$, and so:

$$\begin{aligned}\|\mathbf{g}_{I^{t+1}}\|_2 &= \|\boldsymbol{\epsilon}_{I^{t+1}} - X_{I^{t+1}}^T(XX^T+M)^{-1}X\boldsymbol{\epsilon}_{I^{t+1}}\|_2 \leq \|\boldsymbol{\epsilon}_{I^{t+1}}\|_2 + \|X_{I^{t+1}}^T(XX^T+M)^{-1}X\boldsymbol{\epsilon}_{I^{t+1}}\|_2 \\ &\leq e_0 + \sigma\frac{\sqrt{\Lambda_{k+k^*}\Lambda_n}}{\lambda_{min}(XX^T+M)}\sqrt{2d\log(\frac{d}{\delta})} \leq e_0 + \sigma\frac{\sqrt{\Lambda_{k+k^*}\Lambda_n}}{\lambda_n}\sqrt{2d\log(\frac{d}{\delta})} \\ &\leq (1+\sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0\end{aligned}$$

The last inequality holds when n is sufficiently large. Then, we consider $\mathbf{f}_{I^{t+1}}$:

$$\begin{aligned}\|\mathbf{f}_{I^{t+1}}\|_2 &= \|X_{I^{t+1}}^T(XX^T+M)^{-1}M(\mathbf{w}^*-\mathbf{w}_0)\|_2 \\ &\leq \frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)}\|\mathbf{w}^*-\mathbf{w}_0\|_2\end{aligned}$$

We substitute the three calculated terms into the original result to obtain:

$$\begin{aligned}\|\mathbf{b}^{t+1}-\mathbf{b}^*\|_2 \leq 2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)}\|\mathbf{b}^t-\mathbf{b}^*\|_2 + 2(1+\sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0 \\ + 2\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)}\|\mathbf{w}^*-\mathbf{w}_0\|_2\end{aligned}$$

We let $\eta = 2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)}$. Because $\mathbf{b}^0 = 0$:

$$\begin{aligned}\|\mathbf{b}^{t+1}-\mathbf{b}^*\|_2 \leq \eta^t\|\mathbf{b}^*\|_2 + \frac{2}{1-\eta}(1+\sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0 \\ + \frac{2}{1-\eta}\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)}\|\mathbf{w}^*-\mathbf{w}_0\|_2\end{aligned}$$

Suppose that $n > d\log(d)$. Then, $1+\sqrt{\frac{2d}{n}\log(\frac{d}{\delta})} = O(1)$. From the expression of e_0, we have that $e_0 = O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})$. Then, after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$, we obtain:

$$\|\mathbf{b}^{T_0}-\mathbf{b}^*\|_2 \leq \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\|\mathbf{w}^*-\mathbf{w}_0\|_2$$

□

Theorem 2. *Under the conditions of Theorem 1 and assuming that $\mathbf{x}_i \in \mathbb{R}^d$ are generated from the standard normal distribution, for $k > k^*$, it is guaranteed with a probability of at least $1-\delta$ that, for any $\varepsilon, \delta > 0$, the current estimation coefficient $\mathbf{w}_{T_0}$ satisfies $\|\mathbf{w}_{T_0}-\mathbf{w}^*\|_2 \leq O(\frac{1}{\sqrt{n}})(\varepsilon+e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\|\mathbf{w}^*-\mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ steps.*

Proof.

$$\mathbf{w}^t = (XX^T)^{-1}X(\mathbf{y}-\mathbf{b}^t) = (XX^T)^{-1}X(X^T\mathbf{w}^*+\mathbf{b}^*+\boldsymbol{\epsilon}-\mathbf{b}^t) = \mathbf{w}^*+(XX^T)^{-1}X(\boldsymbol{\epsilon}+\mathbf{b}^*-\mathbf{b}^t)$$

$$\begin{aligned}\|\mathbf{w}^t-\mathbf{w}^*\|_2 &= \|(XX^T)^{-1}X(\boldsymbol{\epsilon}+\mathbf{b}^*-\mathbf{b}^t)\|_2 \leq \frac{1}{\lambda_n}(\|X\boldsymbol{\epsilon}\|_2 + \|X(\mathbf{b}^*-\mathbf{b}^t)\|_2) \\ &\leq \frac{\sqrt{\Lambda_n}}{\lambda_n}\sigma\sqrt{2d\log(\frac{d}{\delta})} + \frac{1}{\lambda_n}\|X(\mathbf{b}^*-\mathbf{b}^t)\|_2) \\ &\leq \frac{\sqrt{\Lambda_n}}{\lambda_n}\sigma\sqrt{2d\log(\frac{d}{\delta})} + \frac{\sqrt{\Lambda_n}}{\lambda_n}\Big[\eta^t\|\mathbf{b}^*\|_2 + \frac{2}{1-\eta}(1+\sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0 \\ &+ \frac{2}{1-\eta}\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)}\|\mathbf{w}^*-\mathbf{w}_0\|_2\Big]\end{aligned}$$

when n is sufficiently large. By Lemma 8 and Theorem 9, $\sqrt{\Lambda_n}/\lambda_n$ can then be approximated as $O(1/\sqrt{n})$ and $\sqrt{\Lambda_{k+k^*}}$ can be approximated as $O(\sqrt{k+k^*})$. Then, we have:

$$\|\mathbf{w}_t - \mathbf{w}^*\|_2 \leq O(\frac{1}{\sqrt{n}})(\varepsilon + e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\|\mathbf{w}^* - \mathbf{w}_0\|_2$$

□

Theorem 3. *Let $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the given matrix with each $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Let $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b} + \boldsymbol{\epsilon}$ and $\|\mathbf{b}\|_0 \leq k^*$. Also, let $k^* \leq k$ and suppose $\lim_{n\to\infty} \frac{\lambda_{min}(M)}{n} = \xi$. Then if the following equation holds*

$$2\frac{k+k^*}{n}(1 + 3e\sqrt{6\log\frac{en}{k+k^*}}) < 1 + \xi$$

and $n \geq \Omega(d + \log\frac{1}{\delta})$. Then, with probability at least $1-\delta$, the data satisfies $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$, More specifically, after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$ steps in TRIP algorithm, the estimation coefficient $\mathbf{w}_{T_0}$ satisfies $\|\mathbf{w}_{T_0} - \mathbf{w}^*\|_2 \leq O(\frac{1}{\sqrt{n}})(\varepsilon + e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\|\mathbf{w}^* - \mathbf{w}_0\|_2$.

Proof. We notice that if $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then $\Sigma^{-1/2}\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$. Thus by Theorem 9 and Lemma 8, with the probability at least $1-\delta$, the data matrix $\tilde{X} = \Sigma^{1/2}X$ satisfies SSC and SSS properties with the following constants

$$\Lambda_k \leq k(1 + 3e\sqrt{6\log\frac{en}{k}}) + O(\sqrt{nd + n\log\frac{1}{\delta}}),$$
$$\lambda_{min}(XX^T) \geq n - (1-2\epsilon)^{-1}\sqrt{cnd + c'n\log\frac{2}{\delta}}.$$

As seen in Theorem 1, the convergence of TRIP needs to satisfies $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$. We notice that $\lambda_{min}(XX^T + M) \geq \lambda_{min}(XX^T) + \lambda_{min}(M)$, so the convergence condition can be scaled to $2\Lambda_{k+k^*} \leq \lambda_{min}(XX^T) + \lambda_{min}(M)$. Using the above bounds, the condition is translated into

$$\underbrace{2\frac{k+k^*}{n}(1 + 3e\sqrt{6\log\frac{en}{k+k^*}})}_{(A)} + \underbrace{O(\sqrt{\frac{d}{n} + \frac{1}{n}\log\frac{1}{\delta}})}_{(B)} < 1 + \frac{\lambda_{min}(M)}{n}.$$

For $n = \Omega(d + \frac{1}{\delta})$ and suppose n is large enough, the part (B) goes to 0. Also because $\lim_{n\to\infty}\frac{\lambda_{min}(M)}{n} = \xi$, so the condition becomes

$$2\frac{k+k^*}{n}(1 + 3e\sqrt{6\log\frac{en}{k+k^*}}) < 1 + \xi.$$

□

The condition $2\frac{k+k^*}{n}(1 + 3e\sqrt{6\log\frac{en}{k+k^*}}) < 1 + \xi$ seems quite abstract. By approximating $f(t) = 2t(1+3e\sqrt{6\log\frac{e}{t}})$ using its second order Taylor's expansion at $t = 1/10$, which is shown in Figure 3. We can give an approximated breakdown point of TRIP algorithm when ξ is not too large, i.e.,

$$k^* \leq k \leq (0.3023 - \sqrt{0.0887 - 0.0040\xi})n. \tag{16}$$

C.3 Convergence Proof for BRHT

C.3.1 Proof of Theorem 5

Lemma 4. *For any real function $f(x)$:*

$$\sup_{x\geq 0}[f(x) + ax] \leq \sup_{x\geq 0}[f(x) + bx] \tag{17}$$

for any $b \geq a \geq 0$.

Proof. Suppose the lemma does not hold, that is, $a \geq 0, b \geq a$, but

$$\sup_{x\geq 0}[f(x) + ax] > \sup_{x\geq 0}[f(x) + bx]$$

We select the array $\{x_n\} = \{x_1, x_2, \dots\}$ such that $\lim_{i\to\infty}[f(x_i) + ax_i] = \sup_{x\geq 0}[f(x) + ax]$. Then, we consider the set $S \doteq \{f(x_i) + bx_i | x_i \in \{x_n\}\}$. It is easy to see that $\sup S \geq \sup_{x\geq 0}[f(x) + ax]$ and $\sup S \leq \sup_{x\geq 0}[f(x) + bx]$. As shown above, however, $\sup_{x\geq 0}[f(x) + ax] > \sup_{x\geq 0}[f(x) + bx]$. This is a contradiction, and Lemma 4 is proved. □

Theorem 5. *Suppose the prior of r_i is independently and identically distributed (iid). We consider the t^{th} iteration step of the BRHT algorithm, in which $\mathbf{w}_t, \mathbf{r}_t = \arg\max_{\mathbf{w}\in\mathbb{R}^d, \mathbf{r}\in\mathbb{R}^n_+} M(\mathbf{w}, \mathbf{r}, \mathbf{b}_t)$, where $\mathbf{b}_t = HT_k(\mathbf{y} - X^T\mathbf{w}_{t-1})$ is obtained from the hard thresholding step. Then, we have that $U(\mathbf{w}_t, \mathbf{r}_t, S_{t+1}) \geq U(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, S_t)$.*

Proof. After obtaining $\mathbf{b}_t$ by $\mathbf{b}_t = HT_k(\mathbf{y} - X^T\mathbf{w}_{t-1})$, we consider $M(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, \mathbf{b}_t)$, that is:

$$\begin{aligned} M(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, \mathbf{b}_t) = &\log p_{\mathbf{w}}(\mathbf{w}_{t-1}) + \sum_{i\in S_t}[\log p_{\mathbf{r}}(r_i^{t-1}) + r_i^{t-1}\log \ell(y_i \mid \mathbf{w}_{t-1}, \mathbf{x}_i, \sigma^2)] \\ &+ \sum_{j\in[n]\setminus S_t}[p_{\mathbf{r}}(r_j^{t-1}) + r_j^{t-1}\ell(0)] \end{aligned}$$

where $\ell(0)$ is the value of the likelihood of $\mathcal{N}(0, \sigma^2)$. This is because, after the hard thresholding step and if i is not chosen from the clean set,' $y_i - b_i^t = y_i - (y_i - X^T\mathbf{w}_{t-1}) = X^T\mathbf{w}_{t-1}$. Thus, it can be seen that $\ell(y_i - b_i^t \mid \mathbf{w}_{t-1}, \mathbf{x}_i, \sigma^2) = \ell(\mathbf{x}_i^T\mathbf{w}_{t-1} \mid \mathbf{w}_{t-1}, \mathbf{x}_i, \sigma^2) = \ell(0)$. We consider a pseudo-reweighting process (this is just for the convenience of the proof and does not appear in the algorithm, but does not affect the result of the algorithm). We try to maximize $M(\mathbf{w}_{t-1}, \mathbf{r}, \mathbf{b}_t)$ by varying $\mathbf{r}$. Because of the independence of $p_{\mathbf{r}}(r_i)$ and the definition of $M(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, \mathbf{b}_t)$, the value of $\mathbf{r}$ in S_t is unchanged.

$$\begin{aligned} \tilde{M}_{t-1} &= \max_{\mathbf{r}\in\mathbb{R}^n} M(\mathbf{w}_{t-1}, \mathbf{r}, \mathbf{b}_t) \\ &= \log p_{\mathbf{w}}(\mathbf{w}_{t-1}) + \sum_{i\in S_t}[\log p_{\mathbf{r}}(r_i^{t-1}) + r_i^{t-1}\log \ell(y_i \mid \mathbf{w}_{t-1}, \mathbf{x}_i, \sigma^2)] + kg(0) \\ &= U(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, S_t) + kg(0) \end{aligned}$$

where $g(0)$ is defined as $\max_{r_i}[p_{\mathbf{r}}(r_i) + r_i\ell(0)]$. Next, we consider the update of $\mathbf{w}$. Because:

$$\mathbf{w}_t, \mathbf{r}_t = \arg\max_{\mathbf{w}\in\mathbb{R}^d, \mathbf{r}\in\mathbb{R}^n_+} M(\mathbf{w}, \mathbf{r}, \mathbf{b}_t)$$

it is easy to see that:

$$M(\mathbf{w}_t, \mathbf{r}_t, \mathbf{b}_t) \geq \max_{\mathbf{r}\in\mathbb{R}^n_+} M(\mathbf{w}_{t-1}, \mathbf{r}, \mathbf{b}_t) = \tilde{M}_{t-1}$$

Finally, we examine $\tilde{M}_t = \max_{\mathbf{r}\in\mathbb{R}^n} M(\mathbf{w}_t, \mathbf{r}, \mathbf{b}_{t+1})$. The explicit form of $\tilde{M}_t$ can be given by $\tilde{M}_{t-1}$. We compare the $\tilde{M}_t$ and $M(\mathbf{w}_t, \mathbf{r}_t, \mathbf{b}_t)$:

$$\begin{aligned} \tilde{M}_t - M(\mathbf{w}_t, \mathbf{r}_t, \mathbf{b}_t) = &\log p_{\mathbf{w}}(\mathbf{w}_t) + \sum_{i\in S_{t+1}}[\log p_{\mathbf{r}}(r_i^t) + r_i^t\log \ell(y_i \mid \mathbf{w}_t, \mathbf{x}_i, \sigma^2)] + kg(0) \\ &- \{\log p_{\mathbf{w}}(\mathbf{w}_t) + \sum_{j\in S_t}[\log p_{\mathbf{r}}(r_j^t) + r_j^t\log \ell(y_j \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)] \\ &+ \sum_{j\in[n]\setminus S_t}[p_{\mathbf{r}}(r_j^t) + r_j^t\log \ell(y_j - b_j^t \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)]\} \end{aligned}$$

$$\begin{aligned} = &\sum_{i\in S_{t+1}\setminus S_t}[\log p_{\mathbf{r}}(r_i^t) + r_i^t\log \ell(y_i \mid \mathbf{w}_t, \mathbf{x}_i, \sigma^2)] - \sum_{j\in S_t\setminus S_{t+1}}[\log p_{\mathbf{r}}(r_j^t) + r_j^t\log \ell(y_j \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)] \\ &+ kg(0) - \sum_{j\in[n]\setminus S_t}[p_{\mathbf{r}}(r_j^t) + r_j^t\log \ell(y_j - b_j^t \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)] \end{aligned}$$

Following the hard thresholding step, $\forall i \in S_{t+1} \backslash S_t$ and $\forall j \in S_t \backslash S_{t+1}$, $|y_i - \mathbf{x}_i^T \mathbf{w}_t| \leq |y_j - \mathbf{x}_j^T \mathbf{w}_t|$, and so $\log \ell(y_i \mid \mathbf{w}_t, \mathbf{x}_i, \sigma^2) \geq \log \ell(y_j \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)$. By Lemma 4, we have that:

$$\log p_{\mathbf{r}}(r_i^t) + r_i^t \log \ell(y_i \mid \mathbf{w}_t, \mathbf{x}_i, \sigma^2) \geq \log p_{\mathbf{r}}(r_j^t) + r_j^t \log \ell(y_j \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2)$$

and because $\forall j \in [n] \backslash S_t$, $\log \ell(y_j - b_j^t \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2) \leq \ell(0)$, we have:

$$\log p_{\mathbf{r}}(r_j^t) + r_j^t \log \ell(y_j - b_j^t \mid \mathbf{w}_t, \mathbf{x}_j, \sigma^2) \leq g(0)$$

This proves that:

$$\tilde{M}_t \geq M(\mathbf{w}_t, \mathbf{r}_t, \mathbf{b}_t)$$

Note that $M(\mathbf{w}_t, \mathbf{r}_t, \mathbf{b}_t) \geq \tilde{M}_{t-1}$. Using the expressions for $\tilde{M}_t$ and $\tilde{M}_{t-1}$:

$$\begin{aligned} U(\mathbf{w}_t, \mathbf{r}_t, S_{t+1}) + kg(0) \geq & U(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, S_t) + kg(0) \\ U(\mathbf{w}_t, \mathbf{r}_t, S_{t+1}) \geq & U(\mathbf{w}_{t-1}, \mathbf{r}_{t-1}, S_t) \end{aligned}$$

□

C.3.2 Proof of Theorems 6 and 7

To prove Theorem 4, we require a certain assumption. We will show that this assumption is reasonable through a brief description in Appendix C.2.3.

Assumption 1. *Let X be the given data matrix and $\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$ be the output. For any specific positive semi-definite matrix M, there exist $\alpha > 0$ and $0 < \gamma \leq 1 + \epsilon$, where ϵ is a small positive number, that for any estimation $\hat{\mathbf{b}}$ of $\mathbf{b}^*$, and let $I_{\hat{\mathbf{b}}} = supp(\hat{\mathbf{b}}) \cup supp(\mathbf{b}^*)$, it holds that*

$$u_1 = \|\boldsymbol{\epsilon}_{I_{\hat{\mathbf{b}}}} + X_{I_{\hat{\mathbf{b}}}}^T (\mathbf{w}^* - \mathbf{w}_1)\|_2 \leq \gamma \|\boldsymbol{\epsilon}_{I_{\hat{\mathbf{b}}}} + X_{I_{\hat{\mathbf{b}}}}^T (\mathbf{w}^* - \mathbf{w}_2)\|_2 = \gamma u_2$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are obtained from:

$$\mathbf{w}_1 = VBEM(X, \mathbf{y} - \hat{\mathbf{b}}, p_{\mathbf{r}}(\mathbf{r}), p_{\mathbf{w}}(\mathbf{w}))$$

$$\mathbf{w}_2 = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \|y_i - \hat{b}_i - \mathbf{x}_i^T \mathbf{w}\|^2 + (\mathbf{w} - \mathbf{w}_0)^T M (\mathbf{w} - \mathbf{w}_0)$$

and $p_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \alpha \sigma^2 M^{-1})$

This assumption can be easily understood as making the Bayesian reweighting regression more robust and accurate than simple regression, thus providing a more reliable solution in each iteration step of the BRHT algorithm. This can be explained from the following two aspects: 1) the Bayesian reweighting regression adds smaller weights to points with large deviations, so the regression is less affected by outliers, especially when the estimation $\hat{\mathbf{b}}$ is not very accurate. 2) By considering the robustness of the Bayesian reweighting regression, smaller prior weights are required to recover the true coefficient. Thus, $\mathbf{w}_2$ is closer to the true coefficient $\mathbf{w}^*$ than $\mathbf{w}_1$, which is reflected in the prior shrinkage coefficient α and the error shrinkage coefficient γ.

Theorem 6. *Consider a data matrix X and a specific positive semi-definite matrix M satisfying the SSC and SSS properties such that $2\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)} < 1$. Then, there exist $\alpha > 0$ and $0 < \gamma \leq 1 + \epsilon$, where ϵ is a small number, such that if $k > k^*$ and Σ_0 in the prior $p_{\mathbf{w}}(\mathbf{w})$ is $\alpha\sigma^2 M^{-1}$, it is guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, $\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \leq \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T+M)})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\gamma\|\mathbf{b}^*\|_2}{\varepsilon}))$ iterations of BRHT, where $e_0 = O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})$ under the normal design.*

Proof. The iteration step of the BRHT algorithm is:

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

where $\mathbf{w}^t = VBEM(X, \mathbf{y} - \mathbf{b}^t, p_{\mathbf{r}}(\mathbf{r}), p_{\mathbf{w}}(\mathbf{w}))$ and:

$$\begin{aligned}\|\mathbf{b}^{t+1}_{I^{t+1}} - (\mathbf{y}_{I^{t+1}} - X^T_{I^{t+1}}\mathbf{w}^t)\|_2 &\leq \|\mathbf{b}^*_{I^{t+1}} - (\mathbf{y}_{I^{t+1}} - X^T_{I^{t+1}}\mathbf{w}^t)\|_2 \\ &= \|\mathbf{b}^*_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + \epsilon_{I_{t+1}} + X^T_{I_{t+1}}(\mathbf{w}^* - \mathbf{w}^t))\|_2 \\ &= \|\epsilon_{I_{t+1}} + X^T_{I_{t+1}}(\mathbf{w}^* - \mathbf{w}^t)\|_2\end{aligned}$$

By defining $\hat{\mathbf{w}}^t = (XX^T + M)^{-1}(X(\mathbf{y} - \mathbf{b}^t) + M\mathbf{w}_0)$ and using the trigonometric inequality, we obtain:

$$\begin{aligned}\|\mathbf{b}^{t+1}_{I^{t+1}} - \mathbf{b}^*_{I^{t+1}}\|_2 &\leq 2\|\epsilon_{I_{t+1}} + X^T_{I_{t+1}}(\mathbf{w}^* - \mathbf{w}^t)\|_2 \\ &\leq 2\gamma\|\epsilon_{I_{t+1}} + X^T_{I_{t+1}}(\mathbf{w}^* - \hat{\mathbf{w}}^t)\|_2 \\ &= 2\gamma\|X^T_{I^{t+1}}\lambda^t + \mathbf{g}_{I^{t+1}} + \mathbf{f}_{I^{t+1}}\|_2\end{aligned}$$

The second inequality holds because of assumption 1. λ^t, $\mathbf{g}$, $\mathbf{f}$ have the same meaning as in Theorem 1. Therefore, through the same proof procedure as for Theorem 1, the above inequality can be finally transformed into the following formula:

$$\begin{aligned}\|\mathbf{b}^{t+1} - \mathbf{b}^*\|_2 \leq 2\gamma\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T + M)}\|\mathbf{b}^t - \mathbf{b}^*\|_2 + 2\gamma(1 + \sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0 \\ + 2\gamma\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T + M)}\|\mathbf{w}^* - \mathbf{w}_0\|_2\end{aligned}$$

We let $\eta = 2\gamma\frac{\Lambda_{k+k^*}}{\lambda_{min}(XX^T+M)}$, and because $\mathbf{b}^0 = 0$, we can write:

$$\begin{aligned}\|\mathbf{b}^{t+1} - \mathbf{b}^*\|_2 \leq \eta^t\|\mathbf{b}^*\|_2 + \frac{2\gamma}{1-\eta}(1 + \sqrt{\frac{2d}{n}\log(\frac{d}{\delta})})e_0 \\ + \frac{2\gamma}{1-\eta}\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T + M)}\|\mathbf{w}^* - \mathbf{w}_0\|_2\end{aligned}$$

Suppose that $n > d\log(d)$. Then, $1 + \sqrt{\frac{2d}{n}\log(\frac{d}{\delta})} = O(1)$. From the expression for e_0, we have that $e_0 = O(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}})$. Then, after $T_0 = O(\log(\frac{\|\mathbf{b}^*\|_2}{\varepsilon}))$, we have:

$$\|\mathbf{b}^{T_0} - \mathbf{b}^*\|_2 \leq \varepsilon + O(e_0) + O(\frac{\sqrt{\Lambda_{k+k^*}}\lambda_{max}(M)}{\lambda_{min}(XX^T + M)})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$$

□

Theorem 7. *Under the conditions of Theorem 4 and assuming that $\mathbf{x}_i \in \mathbb{R}^d$ are generated from the standard normal distribution, there exist $\alpha > 0$ and $0 < \gamma \leq 1 + \epsilon$, where ϵ is a small number, such that if $k > k^*$ and Σ_0 in the prior $p_{\mathbf{w}}(\mathbf{w})$ is $\alpha\sigma^2 M^{-1}$, it is guaranteed with a probability of at least $1 - \delta$ that, for any $\varepsilon, \delta > 0$, the current estimation coefficient $\mathbf{w}_{T_0}$ satisfies $\|\mathbf{w}_{T_0} - \mathbf{w}^*\|_2 \leq O(\frac{1}{\sqrt{n}})(\varepsilon + e_0) + O(\frac{\sqrt{k+k^*}\lambda_{\max}(M)}{n^{3/2}})\gamma\|\mathbf{w}^* - \mathbf{w}_0\|_2$ after $T_0 = O(\log(\frac{\gamma\|\mathbf{b}^*\|_2}{\varepsilon}))$ steps.*

The proof of Theorem 7 is the same as that for Theorem 2, so it is omitted here.

C.3.3 Rationality of Assumption 1

In this section, we use some simulations to check whether assumption 1 is true in the iteration of the BRHT algorithm. For this problem, we choose some special M under AAA to make our description more representative. For each corruption ratio, we choose M so as to achieve the minimum fitting error $\|\mathbf{w}_t - \mathbf{w}^*\|_2$ in the TRIP algorithm. We then find a prior shrinkage coefficient α for this M and simulate the BRHT algorithm to show that there exists an error shrinkage coefficient γ such that the following formula holds in all iterative steps:

$$u_{1t} = \|\boldsymbol{\epsilon}_{I_t} + X^T_{I_t}(\mathbf{w}^* - \mathbf{w}_{1t})\|_2 \leq \gamma\|\boldsymbol{\epsilon}_{I_t} + X^T_{I_t}(\mathbf{w}^* - \mathbf{w}_{2t})\|_2 = \gamma u_{2t}$$

Corruption rate	γ
0.2	1.007
0.25	0.937
0.3	0.917
0.35	0.876
0.4	0.879
0.45	0.958
0.5	1.000
0.55	1.008

Table 1: Calculated γ for each corruption rate

Corruption rate	mean(u_{1t}/u_{2t})
0.2	0.989
0.25	0.917
0.3	0.892
0.35	0.857
0.4	0.806
0.45	0.827
0.5	0.772
0.55	0.758

Table 2: Average error ratio for each corruption rate

where $\mathbf{w}_{1t}$ and $\mathbf{w}_{2t}$ are obtained from:

$$\mathbf{w}_{1t} = VBEM(X, \mathbf{y} - \mathbf{b}^t, p_{\mathbf{r}}(\mathbf{r}), p_{\mathbf{w}}(\mathbf{w}))$$

$$\mathbf{w}_{2t} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{n} \|y_i - b_i^t - \mathbf{x}_i^T\mathbf{w}\|^2 + (\mathbf{w} - \mathbf{w}_0)^T M(\mathbf{w} - \mathbf{w}_0)$$

and $p_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \alpha\sigma^2 M^{-1})$. For the chosen corruption ratio, we take eight evenly spaced points from 0.2 to 0.55. This is because the CRR method collapses when the corruption ratio exceeds 0.2, so including the prior become very important at this time. By selecting an appropriate prior shrinkage coefficient α for different M, the overall result is as shown in Figure 4.

We can find γ for each M by calculating $\max(u_{1t}/u_{2t})$, where $t = 1, ...$ are the iterative steps until convergence. The results are presented in Table 1. For most corruption rates, $\gamma < 1$. However, there are still some cases where γ is greater than 1. This is because, in the iteration process, there are few steps in which u_{1t} and u_{2t} are very close, while in most cases they are well separated. We use another criterion, $mean(u_{1t}/u_{2t})$, to show this phenomenon; the results are presented in Table 2. We can see that, as the corruption ratio increases, $mean(u_{1t}/u_{2t})$ basically exhibits a downward trend and all values are less than 1. This indicates that the performance of BRHT is usually better than that suggested by the theory. This experiment can be used to explain why BRHT usually outperforms TRIP, even when TRIP uses the optimal parameters.

D Choice of Hyperparameters in BRHT

In the BRHT algorithm, the most important parameter for model performance is the parameter in the weight prior $p_{\mathbf{r}}(\mathbf{r})$. According to assumption 1, we must ensure that the Bayesian reweighting regression provides a more robust and accurate solution than traditional least-squares regression. This requires the weight $E_{q(\mathbf{r})}(\mathbf{r})$ in the variational M step of the VBEM algorithm to be relatively insensitive to $\mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)]$, or very few points will have large weights and others will have little impact on the estimates. A relatively sensitive weight will lead to bias, as only a few points of information will be used, and the effects will be even worse when some outliers have not

been detected. Additionally, the weight cannot be too stable, or BRHT will have almost the same performance as TRIP. Here, we present a useful way to determine the hyperparameters so that all uncorrupted points have relatively large weights when the regression result is correct.

Consider the variational E step in the VBEM method. We have:

$$q(r_i) \propto exp\{\log p_{\mathbf{r}}(r_i) + r_i \mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)]\}$$

We use the true likelihood $\log \ell(y_i \mid \mathbf{w}^*, \mathbf{x}_i, \sigma^2)$ to replace $\mathbb{E}_{q(\mathbf{w})}[\log \ell(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2)]$, and we find that:

$$\log \ell(y_i \mid \mathbf{w}^*, \mathbf{x}_i, \sigma^2) = -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w}^*)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

If y_i is not corrupted, then $\frac{1}{\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w}^*)^2 = \frac{1}{\sigma^2}\epsilon_i^2 \leq \chi^2(0.95)$ holds with at least 95% probability, where $\chi^2(0.95)$ is the 95% quantile of the χ^2 distribution with 1 degree of freedom. Under the above condition, with at least 95% probability, it is easy to see that:

$$-\frac{1}{2}\chi^2(0.95) - \frac{1}{2}\log(2\pi\sigma^2) \leq \log \ell(y_i \mid \mathbf{w}^*, \mathbf{x}_i, \sigma^2) \leq -\frac{1}{2}\log(2\pi\sigma^2)$$

Here, we define two posterior distributions of weights in the extreme case where all points fit well or deviate greatly in the true regression model:

$$q_1(r_i) \propto exp[\log p_{\mathbf{r}}(r_i) + r_i(-\frac{1}{2}\log(2\pi\sigma^2))]$$
$$q_2(r_i) \propto exp[\log p_{\mathbf{r}}(r_i) + r_i(-\frac{1}{2}\chi^2(0.95) - \frac{1}{2}\log(2\pi\sigma^2))]$$

Then, the hyperparameter in the weight prior $p_{\mathbf{r}}(\mathbf{r})$ is determined by the following rule:

$$\mathbb{E}_{q_2(r_i)}(r_i) \geq \beta \mathbb{E}_{q_1(r_i)}(r_i)$$

In this paper, $\beta = \frac{1}{2}$. The parameter σ^2 can be replaced by a robust estimate such as the M estimator. After the weight prior $p_{\mathbf{r}}(\mathbf{r})$ has been determined, the hyperparameter Σ in prior $p_{\mathbf{w}}(\mathbf{w})$ can be selected by cross-validation using Σ in the specific form $\Sigma = sI$.

E Additional Experimental Results

In this section, we give more experimental results of TRIP and BRHT in comparison with alternative methods. We also show the robustness of our methods under other attacks. First, We compare TRIP and BRHT with the TORRENT method proposed by Bhatia et al. [2] on both OAA and AAA. TORRENT can resist AAA when the white noise $\boldsymbol{\epsilon}$ is not considered in the model. In order to evaluate the influence of white noise on robust regression, the true data are generated in two ways, one with white noise ($y_i = \mathbf{x}_i^T w^* + \epsilon_i$) and the other without white noise ($y_i = \mathbf{x}_i^T w^*$). Other settings are the same as those in Section 5. The experimental results are shown in Figure 5. Under these two attacks, the performance of TORRENT algorithm is very consistent with that of CRR in both noisy and noiseless settings. TORRENT performs slightly better than CRR in the absence of white noise, as shown in Figure 5(e). However, both CRR and TORRENT perform poorly under AAA. It can be seen that the TRIP and BRHT algorithms are very robust in all cases.

We also consider another leverage point attack (LPA) on data sets. For a point $(\mathbf{x}_i, y_i)$, the leverage value is defined as $h_{ii} = \mathbf{x}_i^T (XX^T)^{-1} \mathbf{x}_i$. In the linear regression, the regression result can be strongly affected by high leverage points[4]. Therefore, if we corrupt those high leverage points, the regression result is more likely to be unstable. If we set the covariant $\mathbf{x}_i$ as iid in $\mathcal{N}(0, I_d)$, then the high leverage points are roughly those points with large norms $\|\mathbf{x}_i\|_2$ since $\frac{1}{n}XX^T$ converges to I_d as $n \to \infty$. According to the above analysis, we set the LPA as follows: choose k points with the largest covariant norm $\|\mathbf{x}_i\|_2$ and set their corresponding y_i to 0. In this experiment, the true coefficient $\mathbf{w}^*$ is chosen to be a random unit norm vector and the covariant $\mathbf{x}_i$ are iid in $\mathcal{N}(0, I_d)$. The true data (before attack) are also generated in two ways, one with white noise $y_i = \mathbf{x}_i^T w^* + \epsilon_i$ and the other without white noise $y_i = \mathbf{x}_i^T w^*$, where ϵ_i are iid in $\mathcal{N}(0, \sigma^2)$. We set $\sigma = 1$ in the experiments. The experimental results are shown in Figure 6. Under LPA, CRR performs poorly and usually collapses first among these methods. Rob-ULA has relatively better performance when the

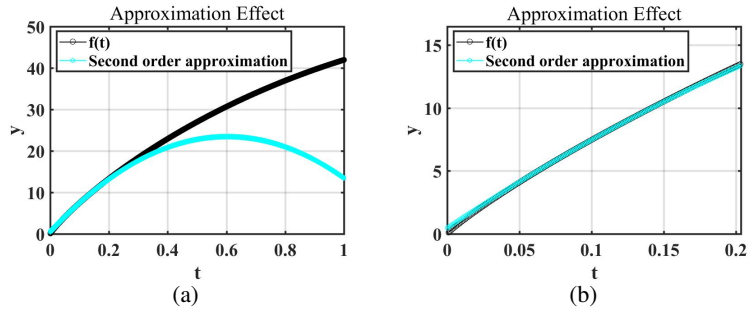

Figure 3: (a) The approximation of the second order Taylor's expansion on $[0, 1]$. (b) Approximation on the interval $[0, 0.2]$.

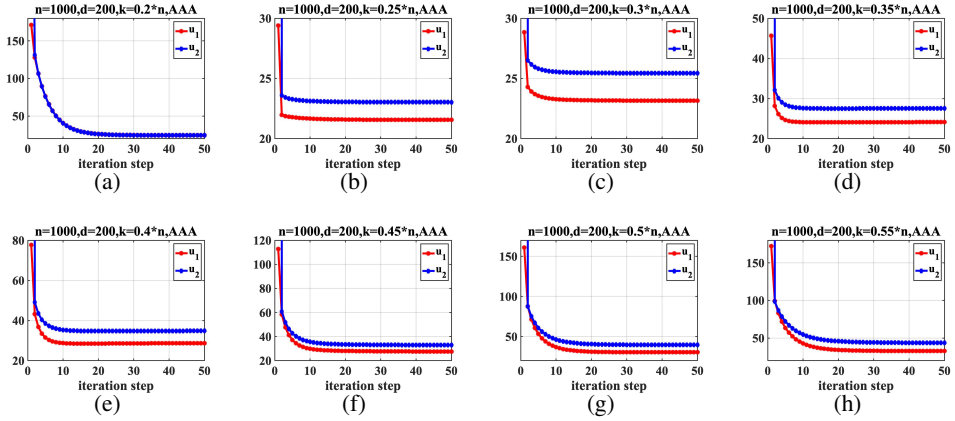

Figure 4: Variation trends of u_{1t} and u_{2t} during the iteration process.

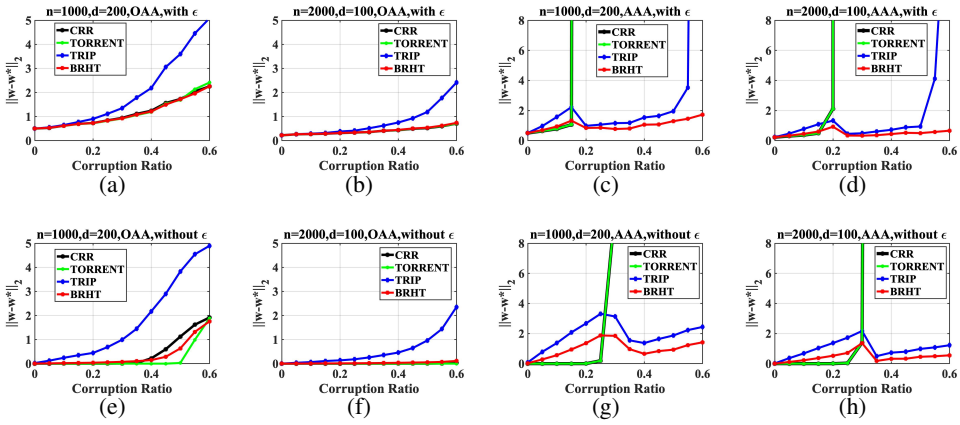

Figure 5: Recovery of parameters with respect to the number of data points n, dimensionality d, and corruption ratio α. (a),(b),(c),(d) consider the case with white noise ϵ, while (e),(f),(g),(h) do not consider white noise. The performance of TORRENT and CRR is similar, and TRIP and BRHT are still more robust under AAAs than CRR and TORRENT.

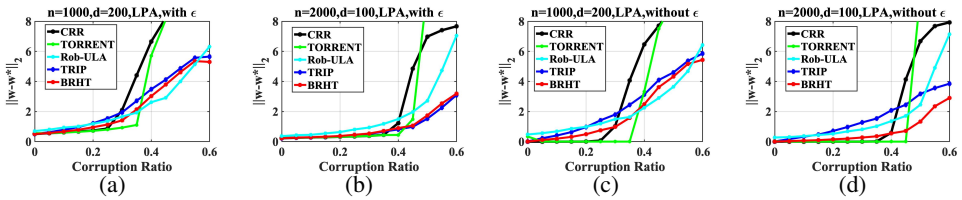

Figure 6: Recovery of parameters with respect to the number of data points n, dimensionality d, and corruption ratio α under LPA. TRIP and BRHT perform significantly better than CRR. BRHT is more robust in all cases. TORRENT and Rob-ULA show robustness in some cases, but still have limitations.

proportion of outliers is high, but there will be relatively large errors in the case of low proportion of outliers. TORRENT is very robust under LPA, especially in the absence of white noise. However, if the data dimension is high and the sample size is small, TORRENT is easier to collapse. The proposed TRIP and BRHT are still better than CRR, and will maintain a robust result even there are lots of outliers. The estimation errors of BRHT are smaller than TRIP, which shows BRHT is the most robust algorithm in this experiment.