

# Robust Regression via Online Feature Selection

## Abstract

The presence of data corruption in user-generated streaming data, such as social media, motivates a new fundamental problem that learns reliable regression coefficient when features are not accessible entirely at one time. Until now, several important challenges still cannot be handled concurrently: 1) corrupted data estimation when only partial features are accessible; 2) online feature selection when data contains adversarial corruption; and 3) scaling to a massive dataset. This paper proposes a novel Robust regression algorithm via Online Feature Selection (*RoOFS*) that concurrently addresses all the above challenges. Specifically, the algorithm iteratively updates the regression coefficients and the uncorrupted set via a robust online feature substitution method. We also prove that our algorithm has a restricted error bound compared to the optimal solution. Extensive empirical experiments demonstrated that the effectiveness of our new method is superior to that of existing methods in the recovery of both feature selection and regression coefficients, with very competitive efficiency.

## 1 Introduction

In traditional robust regression problem, reliable regression coefficients are learned in the presence of adversarial data corruptions in its response vector. A commonly adopted model from existing methods assumes that the observed response is obtained from the generative model  $\mathbf{y} = X^T \beta^* + \mathbf{u}$ , where  $\beta^*$  is the true regression coefficients we wish to recover and  $\mathbf{u}$  is the corruption vector with adversarial values. In the problem setting, the data matrix  $X$  is assumed to contain all the features that can be accessed at any time and by arbitrarily many times. However, the assumption is no longer suitable to the following scenarios in the applications that contain exponentially increasing user-generated contents: 1) *features are too many to be loaded entirely*. For instance, millions of musics and movies establish an extremely large feature set in a recommending system, which makes it hard to load all the features entirely. 2) *features are generated dynamically*. For example, people create and use new terms and hashtags all the time in Twitter, and the "Likes" (Naylor *et al.* 2012) on newly-generated articles in Facebook can be considered as new features describing the interestingness of the user. Therefore, it is necessary to address online features in traditional robust regression as a new fundamental problem; however, current methods either

focus on robust regression or online feature learning separately.

A large body of literature on robust regression problem has been established over the last few decades. Most of studies focus on handling stochastic noise in small amounts (Loh and Wainwright 2011); however, these methods cannot be applied to data that may exhibit malicious corruption (Chen *et al.* 2013). To recover regression coefficients with adversarial data corruption, Chen *et al.* (Chen *et al.* 2013) proposed a robust algorithm based on trimmed inner product. McWilliams *et al.* (McWilliams *et al.* 2014) proposed a subsampling algorithm for large-scale corrupted linear regression, but their theoretical recovery boundaries are not close to the ground truth (Bhatia *et al.* 2015). Some  $L_1$  penalty based methods (Wright and Ma 2010; Nguyen and Tran 2013) pursue strong recovery results for robust regression problem, but these methods depend on severe restrictions of the data distribution such as row-sampling from an incoherent orthogonal matrix (Nguyen and Tran 2013). Recently, Bhatia *et al.* (Bhatia *et al.* 2015) and Zhang *et al.* (Zhang *et al.* 2017) proposed hard-thresholding based methods with strong guarantees of coefficient recovery under a mild assumption on datasets. However, all these approaches are based on batch feature selection under the assumption that all features can be accessed entirely at any time.

Online feature selection methods (Jiang *et al.* 2006; Wang *et al.* 2014; Yu *et al.* 2016) relaxes the requirement of batch selection and fit the scenarios that feature cannot be accessed entirely at one time. Statistical online feature selection algorithms (Zhou *et al.* 2006; Wu *et al.* 2010; Yu *et al.* 2014) select features via certain statistical quantity such as mutual information, but these methods lack of specific objectives and usually have sub-optimal solutions for some certain tasks. Optimization based approaches (Perkins and Theiler 2003a; Zhu *et al.* 2010) use target oriented objective functions solved by some specific optimization techniques. These methods usually require the regression coefficient  $\beta$  be sparse, i.e.,  $\|\beta\|_0 \leq \mu$ . Grafting (Perkins and Theiler 2003b) and its variation (Zhu *et al.* 2010) relax the hard constraint of feature set into  $L_1$  penalty, which makes it a convex problem. However, the parameter of  $L_1$  norm (Ryali and Menon 2009) is difficult to determine because the usual cross validation strategy is unavailable for the online feature selection scenario (Wang *et al.* 2015). Yang *et al.* (Yang *et al.* 2016) proposed a limited-memory substitution algorithm based on the  $L_0$  norm constraint. Although the hard constraint leads to an NP-hard problem, a theoretical

guarantee for the error bound of their local optimal solution is provided. However, none of these online feature methods can handle the adversarial data corruption.

To the best of our knowledge, our proposed approach is the first robust regression algorithm that can handle the online features with adversarial data corruptions. It is nontrivial to consider online features and data corruption simultaneously in robust regression because 1) robust methods usually estimate data corruption based on the entire data, but online features make the data can only be partially accessible at one time; and 2) online feature selection methods can only select features based on uncorrupted data. Simply using robust regression and online feature selection methods sequentially makes the recovery result of coefficients worse, which is presented in our experiments in Section 5. To address the above challenges, we proposed a new robust regression algorithm via online feature selection (*RoOFS*). The main contributions of our study are summarized as follows: 1) *design of an efficient algorithm to simultaneously address the problem of data corruption and online feature*. The algorithm *RoOFS* is proposed to recover the regression coefficients and uncorrupted set efficiently. Unlike using entire features, our approach alternately estimates the data corruption and selects the feature set via a robust online feature substitution method. 2) *theoretical analysis of the algorithm*. We prove that our method yields a solution with a restricted error bound compared to ground truth coefficients under the Subset Restricted Strong Convexity (*SRSC*) property. 3) *demonstration of empirical effectiveness and efficiency*. Our proposed algorithm was evaluated with 6 competing methods in both robust regression and online feature selection literatures. The results showed that our approach consistently outperforms existing methods in coefficients recovery and uncorrupted set estimation, delivering a competitive running time.

The reminder of this paper is organized as follows. Section 2 gives a formal problem formulation. The proposed *RoOFS* algorithm is presented in Section 3. Section 4 presents the theoretical analysis of proposed algorithm. In Section 5, the experimental results are analyzed and the paper concludes with a summary of our work in Section 6.

## 2 Problem Formulation

In this study, we consider the problem of robust regression with adversarial data corruption in the feature selection scenario in which only a few features are accessible at each time. Given data matrix  $X_t \in \mathbb{R}^{p_t \times n}$  where  $p_t$  is the number of features available in the  $t^{\text{th}}$  time interval, and  $n$  are the number of data samples. The data matrix for all the time intervals is represented as  $X = \{X_t\}_{t=1}^T$ . We assume the corresponding response vector  $\mathbf{y} \in \mathbb{R}^{n \times 1}$  is generated using the following model:

$$\mathbf{y} = X^T \beta^* + \mathbf{u} + \varepsilon \quad (1)$$

where  $\beta^*$  represents the  $\mu$ -sparse ground truth coefficients of the regression model i.e.,  $\|\beta^*\|_0 \leq \mu$  and  $\mathbf{u}$  is the unbounded corruption vector introduced by adversarial data attacks.  $\varepsilon \in \mathbb{R}^{n \times 1}$  represents the additive dense noise, where  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . Different from the corruption vector  $\mathbf{u}$  that can be arbitrarily distributed, the dense noise  $\varepsilon_i$  follows normal distribution with zero mean and a relatively small variance  $\sigma$ . The notations used in this paper is summarized in Table 1.

Table 1: Math Notations

Notations	Explanations
$p, n \in \mathbb{R}$	number of entire features and data samples
$p_t \in \mathbb{R}$	number of features in $t^{\text{th}}$ time interval
$\mu \in \mathbb{R}$	ratio of feature sparsity, where $\ \beta\ _0 = \mu$
$X_t \in \mathbb{R}^{p_t \times n}$	data samples containing features in the $t^{\text{th}}$ time interval
$X \in \mathbb{R}^{p \times n}$	data samples containing the entire features
$\beta, \beta^* \in \mathbb{R}^{p \times 1}$	estimated and ground truth regression coefficient
$\mathbf{u} \in \mathbb{R}^{n \times 1}$	corruption vector with adversarial values
$\varepsilon \in \mathbb{R}^{n \times 1}$	dense noise vector, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
$\mathbf{y} \in \mathbb{R}^{n \times 1}$	response vector, where $\mathbf{y} = X^T \beta^* + \mathbf{u} + \varepsilon$
$\mathbf{r} \in \mathbb{R}^{n \times 1}$	residual vector, where $\mathbf{r} = \mathbf{y} - X^T \beta$
$S \subseteq [n]$	estimated uncorrupted set
$S_* \subseteq [n]$	ground truth uncorrupted set, where $S_* = \overline{\text{supp}(\mathbf{u})}$
$\Psi, \Psi_* \subseteq [\mu]$	estimated and ground truth feature set

The goal of our problem is to learn a new robust regression problem with online feature selection, which is to recover the regression coefficients  $\beta^*$  and simultaneously determine the uncorrupted point set  $\hat{S}$  with sequentially accessible features. The problem is formally defined as follows:

$$\begin{aligned} \hat{\beta}, \hat{S} = \arg \min_{\beta, S} \|\mathbf{y}_S - X_S^T \beta\|_2^2 \\ \text{s.t. } S \subseteq [n], |S| \geq \mathcal{G}(\beta), \|\beta\|_0 \leq \mu \end{aligned} \quad (2)$$

Given a subset  $S \subseteq [n]$ ,  $\mathbf{y}_S$  restricts the row of  $\mathbf{y}$  to indices in  $S$  and  $X_S$  signifies that the columns of  $X$  are restricted to indices in  $S$ . Therefore, we have  $\mathbf{y}_S \in \mathbb{R}^{|S| \times 1}$  and  $X_S \in \mathbb{R}^{p \times |S|}$ . We use the notation  $S_* = \overline{\text{supp}(\mathbf{u})}$  to denote the ground truth set of uncorrupted points. Also, for any vector  $\mathbf{v} \in \mathbb{R}^n$ , the notation  $\mathbf{v}_S$  represents the  $|S|$ -dimensional vector containing the components in  $S$ . The notation  $\Psi = \text{supp}(\beta)$  is used to represent the set of selected features, resulting in  $|\Psi| \leq \mu$ . Similarly, we use  $X_\Psi$  to signify the rows of  $X$  are restricted to indices in  $\Psi$  and  $X_{\Psi, S}$  to restrict both the rows and columns in set  $\Psi$  and  $S$ . The function  $\mathcal{G}(\cdot)$  determines the size of uncorrupted data according to the regression coefficients  $\beta$ , which is explained in Section 3. It is worth mentioning that the features of data matrix  $X$  in Equation (2) cannot be loaded entirely, but they can be accessed partially for each time interval. Therefore, the joint optimization of  $\beta$  and  $S$  in our problem are very challenging because it amounts to a non-convex discrete optimization problem under the assumption that data matrix  $X$  cannot be access entirely at one time.

## 3 The Proposed Methodology

To solve the problem in Equation (2) efficiently with the guarantee on the strong recovery of regression coefficients, we propose a novel robust regression algorithm with online feature selection, *RoOFS*. The algorithm is only allowed to access part of features at each time, which are defined as the newly incoming features in our problem. One naive solution to handle the sequentially incoming features is to retain all the features in the memory and then apply traditional robust feature selection methods. However, the solution has two major drawbacks: 1) the feature set can be too large to be retained in the memory, and 2) the algorithm becomes slower and slower when the feature set increases. Therefore, we proposed a new ‘‘robust online substitution’’ method to decide the retained feature set based on an adaptively esti-

rupted corrupted set. The procedure of robust online substitution is defined as follows:

- Update coefficients of retained features  $\Psi$  based on the estimated uncorrupted set  $S$  as follows:  $\beta_\Psi := \beta_\Psi - \eta X_{\Psi,S}^T (X_{\Psi,S}^T \beta_\Psi - \mathbf{y}_S)$ , where  $\eta$  is the step length.
- Retain the top  $\mu$  largest (in magnitude) elements in  $\beta$  and set the rest to zero. Then all the non-zero features will be kept in the retained feature set  $\Psi$ .
- Compute the residual vector  $\mathbf{r} \in \mathbb{R}^{n \times 1}$  with the updated coefficients  $\beta$ , then estimate the uncorrupted feature set  $S$  via a thresholding operator  $\mathcal{H}_\tau(\mathbf{r})$ , where  $\tau$  is the estimated size of uncorrupted set.

The procedure will be repeatedly executed until the residual vector  $\mathbf{r}$  converges. The thresholding operator  $\mathcal{H}_\tau(\cdot)$  is formally defined as follows:

**Definition 1 (Thresholding Operator).** *Defining  $\varphi_v^{-1}(i)$  as the position of the  $i^{\text{th}}$  element in input vector  $\mathbf{v}$ 's ascending order of magnitude and  $\tau$  as the threshold parameter, the thresholding operator of  $\mathbf{v}$  is defined as*

$$\mathcal{H}_\tau(\mathbf{v}) = \{i \in [n] : \varphi_v^{-1}(i) \leq \tau\} \quad (3)$$

To estimate the uncorrupted set  $S$ , the thresholding operator  $\mathcal{H}_\tau(\cdot)$  generally requires two inputs: residual vector  $\mathbf{r}$  and the size of uncorrupted set  $\tau$ . The residual vector  $\mathbf{r}$  can be computed with coefficients  $\beta$  as follows:

$$\mathbf{r} = \mathbf{y} - X^T \beta \quad (4)$$

For the size of uncorrupted set, two general cases are discussed. The first case is that the size can be estimated by users based on their prior knowledge on the data. For instance, if we know the data corruption happens rarely, then we can estimate the uncorrupted size as 95% of the entire data. However, it is hard to obtain prior knowledge on the data in the real-world. Thus, in the second case, we propose a method to adaptively estimate the uncorrupted size based on the residual vector  $\mathbf{r}$ . The method follows an intuition that when the coefficient  $\beta$  is close to  $\beta^*$ , the residuals of uncorrupted samples are smaller than those of corrupted samples in strong possibility. The intuition can be explained by the generative model in Equation (1), where the corrupted samples have the residual  $\mathbf{r} \approx \mathbf{u} + \varepsilon$ , but the residual of uncorrupted samples only contains the white noise  $\varepsilon$ .

The estimation of uncorrupted size can be formalized to solve the following problem:

$$\hat{\tau} := \arg \max_{\substack{\tau \\ \lceil n/2 \rceil < \tau \leq n}} \tau \quad \text{s.t.} \quad r_{\varphi(\tau)} \leq \frac{2\tau r_{\varphi(\tau_o)}}{\tau_o}, \tau \in \mathbb{Z}^+ \quad (5)$$

where  $r_{\varphi(k)}$  represents the  $k^{\text{th}}$  elements of residual vector  $\mathbf{r}$  in ascending order of magnitude. The variable  $\tau_o$  in the constraint is defined as an intermediate variable whose  $r_{\varphi(\tau_o)}$

has the closest value to  $\frac{\|\mathbf{r}_{\mathcal{H}_{\tau'}(\mathbf{r})}\|_2^2}{\tau'}$ , where  $\tau' = \tau - \lceil n/2 \rceil$  and  $\mathcal{H}_{\tau'}(\mathbf{r})$  represent the position set containing the smallest  $\tau'$  elements in residual  $\mathbf{r}$ . The problem in Equation (5) can be solved by searching from  $n$  to  $\lceil n/2 \rceil + 1$  and return the first value  $\hat{\tau}$  which satisfies the constraint. It is important to note that the estimation method in Equation (5) requires the coefficients  $\beta$  to be close to  $\beta^*$ . Thus, we optimize the uncorrupted set  $S$  along with coefficient  $\beta$  until both of them converge.

---

#### Algorithm 1: ROOFS ALGORITHM

---

**Input:** Corrupted training data  $\{\mathbf{x}_i, \mathbf{y}_i\}$ ,  $i = 1 \dots n$ , feature ratio  $\mu$ , tolerance  $\epsilon$

**Output:** solution  $\hat{\beta}$

```

1  $\beta^0 \leftarrow \mathbf{0}$ ,  $\Psi = \emptyset$ ,  $S_0 = [n]$ ,  $t \leftarrow 0$ ,  $k \leftarrow 0$ 
2 repeat
3   Receive features  $X_{\Psi^k}$  from the pool  $\bar{\Psi}$  with index set  $\Psi^k$ 
4    $\Psi = \Psi \cup \Psi^k$ 
5   repeat
6      $\beta_\Psi^{t+1} \leftarrow \beta_\Psi^t - \eta X_{\Psi,S_t}^T (X_{\Psi,S_t}^T \beta_\Psi^t - \mathbf{y}_{S_t})$ 
7     if  $|\Psi| > \mu$  then
8        $\Omega = \arg \min_{\Omega \in \Psi} \|\beta_\Omega\|_1 \quad \text{s.t.} \quad |\Omega| = |\Psi| - \mu$ 
9        $\beta_\Omega = \mathbf{0}$ 
10       $\Psi = \Psi \setminus \Omega$ 
11       $\mathbf{r} = \mathbf{y} - X^T \beta$ 
12       $S_{t+1} \leftarrow \mathcal{H}_\tau(\mathbf{r}^{t+1})$ , where  $\tau$  is the estimated uncorrupted size.
13       $t \leftarrow t + 1$ 
14    until  $\|\mathbf{r}_{S_{t+1}}^{t+1} - \mathbf{r}_{S_t}^t\|_2 < \epsilon n$ 
15     $k \leftarrow k + 1$ 
16 until No more features;
17 return  $\beta^{t+1}$ ,  $S_{t+1}$ 
```

---

The details of *RoOFS* algorithm are presented in Algorithm 1. In Line 3, the algorithm receives data matrix  $X_{\Psi^k}$  with the incoming feature set  $\Psi^k$  at time  $k$ . The new feature set  $\Psi^k$  is combined into the retained feature set  $\Psi$  in Line 4. For each incoming feature set, the algorithm iteratively optimizes the regression coefficients  $\beta$  and the uncorrupted set  $S$  until the value of residual vector  $\mathbf{r}_{S_t}^t$  is converged in Line 14. Specifically, in Line 6, regression coefficients  $\beta$  are updated to a better fit for the current estimated feature set  $\Psi$  and uncorrupted set  $S_t$ . In Line 8, feature set  $\Omega$  that contains features with  $|\Psi| - \mu$  smallest weights in  $\beta$  is selected. Then features in  $\Omega$  are removed from the retained feature set  $\Psi$  and the weights in  $\beta_\Omega$  are reset to zero in Lines 9 and 10. The residual vector  $\mathbf{r}$  is updated in Line 11, while the uncorrupted set  $S_{t+1}$  is estimated in Line 12 by the thresholding operator. Finally, both coefficients  $\beta$  and uncorrupted set  $S$  are returned in Line 17.

## 4 Theoretical Analysis

In this section, we show that the local optimal solution of our algorithm obtains a restricted error bound compared to ground truth solution. To prove the theoretical properties of our algorithm, we require that the least squares function satisfies the *Subset Restricted Strong Convexity* (SRSC), which is defined as follows:

**Definition 2 (SRSC Property).** *The least squares function  $f_S(\beta) = \|\mathbf{y}_S - X_S^T \beta\|_2^2$  satisfies Subset Restricted Strong Convexity (SRSC) Property if the following holds for  $\forall \beta_1, \beta_2 \in \Omega_\mu$  and  $\forall S \in \mathcal{S}_\gamma$ :*

$$f_S(\beta_1) - f_S(\beta_2) \geq \nabla^T f_S(\beta_2)(\beta_1 - \beta_2) + \frac{\varphi_\mu}{2} \|\beta_1 - \beta_2\|_2^2 \quad (6)$$

To provide the local optimality property of our solution, the following two lemmas are first proved.

**Lemma 1.** For a given least squares function  $f(\beta) = \|\mathbf{y} - X^T \beta\|_2^2$ , let residual vector  $\mathbf{r} = \mathbf{y} - X^T \beta$  and  $\delta(k)$  be the  $k$ -th position of the ascending order in vector  $\mathbf{r}$ , i.e.  $r_{\delta(1)} \leq r_{\delta(2)} \leq \dots \leq r_{\delta(n)}$ . For any  $1 \leq \tau_1 < \tau_2 \leq n$  and  $\forall \beta^t \in \Omega_m$ , let  $S_1 = \{\delta(i) | 1 \leq i \leq \tau_1\}$  and  $S_2 = \{\delta(i) | 1 \leq i \leq \tau_2\}$ . We then have  $f_{S_1}(\beta^t) \leq f_{S_2}(\beta^t)$ .

*Proof.* Let  $S_3 = \{\delta(i) : \tau_1 + 1 \leq i \leq \tau_2\}$ . Clearly, we have  $f_{S_2}(\beta^t) = f_{S_1}(\beta^t) + f_{S_3}(\beta^t)$ . Moreover, since each element in  $S_3$  is larger than any of the element in  $S_1$ , we have  $f_{S_1}(\beta^t) \leq f_{S_2}(\beta^t) + \frac{|S_3|}{|S_1|} f_{S_1}(\beta^t) \leq \frac{|S_1|}{|S_1| + |S_3|} f_{S_2}(\beta^t) = \frac{\tau_1}{\tau_2} f_{S_2}(\beta^t) \leq f_{S_2}(\beta^t)$ .  $\square$

**Lemma 2.** Let  $\tau_* = \gamma n$  be the true number of uncorrupted samples and  $\tau_t$  be the estimated uncorrupted threshold at the  $t$ -th iteration. If  $\tau_t \leq \tau_*$ , then  $f_{S_t}(\beta^t) \leq f_{S_*}(\beta^t)$ . If  $\tau_t > \tau_*$ , then  $f_{S_t}(\beta^t) \leq \lambda f_{S_*}(\beta^t)$ , where  $\lambda = \left[1 + \frac{128(1-\gamma)}{2\gamma-1}\right]$ .

*Proof.* To simplify the notation, the subscripts  $t$  that signify the  $t$ -th iteration will be omitted and the residual vector  $\mathbf{r}$  is assumed to be sorted in ascending order of magnitude.

We will discuss the  $\tau_t$  value in two different conditions compared to the value of  $\tau_*$ . In the first condition that  $\tau_t \leq \tau_*$ , let  $S_t = \{\delta(i) | 1 \leq i \leq \tau_t\}$  and  $S_* = \{\delta(i) | 1 \leq i \leq \tau_*\}$ , we have  $f_{S_t}(\beta^t) \leq f_{S_*}(\beta^t)$  according to Lemma 1. When  $\tau_t > \tau_*$ , we have the following properties according to the constraint specified in equation (5).

$$r_\tau^2 \leq \left(2 \cdot \frac{\tau r_{\tau_0}}{\tau_0}\right)^2 \stackrel{(a)}{\leq} \frac{64}{\tau'} \|\mathbf{r}_{S_* \cap S_t}\|_2^2$$

$$|S_t \setminus S_*| r_\tau^2 \stackrel{(b)}{\leq} 64(1-\gamma) \cdot \frac{n}{\tau'} \|\mathbf{r}_{S_* \cap S_t}\|_2^2$$

The inequality (a) follows the definition of  $\tau_0$  and the fact that  $|S_* \cap S_t| \geq \tau'$ . The inequality (b) follows  $|S_t \setminus S_*| \leq (1-\gamma) \cdot n$  and  $\|\mathbf{r}_{S_t \setminus S_*}\|_2^2 \leq |S_t \setminus S_*| r_\tau^2$ . Then we have

$$f_{S_t \setminus S_*}(\beta) \leq \left[64(1-\gamma) \cdot \frac{n}{\tau'} + 1\right] f_{S_* \setminus S_t}(\beta)$$

$$+ \left[64(1-\gamma) \cdot \frac{n}{\tau'}\right] f_{S_* \cap S_t}(\beta)$$

$$f_{S_t \setminus S_*}(\beta) + f_{S_* \cap S_t}(\beta) \stackrel{(c)}{\leq} \left[64(1-\gamma) \cdot \frac{n}{\tau'} + 1\right] f_{S_*}(\beta)$$

$$f_{S_t}(\beta) \stackrel{(d)}{\leq} \left[1 + \frac{128(1-\gamma)}{2\gamma-1}\right] f_{S_*}(\beta)$$

The inequality (c) follows  $f_{S_*}(\beta) = f_{S_* \setminus S_t}(\beta) + f_{S_* \cap S_t}(\beta)$  and the inequality (d) follows  $\tau' = \tau_t - \frac{n}{2}$ .  $\square$

**Theorem 3.** Assume that least squares function  $f_S(\beta) = \|\mathbf{y}_S - X_S^T \beta\|_2^2$  satisfies Subset Restricted Strong Convexity (SRSC) Property for  $\forall \beta_1, \beta_2 \in \Omega_\mu$  and  $\forall S \in \mathcal{S}_\gamma$ , then we have

$$f_{\hat{S}}(\hat{\beta}) - f_{S^*}(\beta^*) \leq \frac{\alpha \lambda}{1+\alpha} f_{S^*}(\mathbf{0}) + \left(\frac{\lambda}{1+\alpha} - 1\right) f_{S^*}(\beta^*) \quad (7)$$

where  $\alpha = \left(\frac{1}{\eta \cdot \varphi_\mu}\right)^2$  and  $\lambda = \left[1 + \frac{128(1-\gamma)}{2\gamma-1}\right]$ . Specifically, when the uncorrupted set size is less than ground truth,  $\lambda = 1$ .

*Proof.* As function  $f_S(\beta)$  satisfies SRSC property, we have

$$f_S(\beta_1) - f_S(\beta_2) \geq \langle \nabla f_S(\beta_2), \beta_1 - \beta_2 \rangle + \frac{\varphi_\mu}{2} \|\beta_1 - \beta_2\|_2^2 \quad (8)$$

for  $\forall \beta_1, \beta_2 \in \Omega_\mu$  and  $\forall S \in \mathcal{S}_\gamma$ . Let  $\text{supp}(\beta_1) = \Omega_1$ , then we have

$$f_S(\beta_1) - f_S(\beta_2)$$

$$\geq \min_{\text{supp}(\beta) \subseteq \Omega_1} \left\{ \langle \nabla f_S(\beta_2), \beta - \beta_2 \rangle + \frac{\varphi_\mu}{2} \|\beta - \beta_2\|_2^2 \right\}$$

$$\stackrel{(a)}{=} -\frac{1}{2\varphi_\mu} \left\| \left[ \nabla f_S(\beta_2) \right]_{\Omega_1} \right\|_2^2 \stackrel{(b)}{\geq} -\frac{|\Omega_1|}{2\varphi_\mu} \left\| \left[ \nabla f_S(\beta_2) \right]_{\Omega_1} \right\|_\infty^2$$

$$\geq -\frac{\mu}{2\varphi_\mu} \left\| \nabla f_S(\beta_2) \right\|_\infty^2 \quad (9)$$

The equation (a) solves the minimum value of  $\min_{\text{supp}(\beta) \subseteq \Omega_1} \{\cdot\}$  by setting its gradient to zero, and the inequality (b) follows  $|\Omega_1| \leq \mu$ . Let  $\beta_1, \beta_2$  be ground truth coefficient  $\beta^*$  and estimated solution  $\hat{\beta}$  respectively, and set  $S$  be ground truth uncorrupted set  $S^*$ , then we have

$$f_{S^*}(\hat{\beta}) - f_{S^*}(\beta^*) \leq \frac{\mu}{2\varphi_\mu} \left\| \nabla f_{S^*}(\hat{\beta}) \right\|_\infty^2 \leq \frac{\mu}{2\varphi_\mu} \left( \frac{1}{\eta} \hat{\beta}_{\min} \right)^2$$

where  $\hat{\beta}_{\min} = \min_i |\hat{\beta}_i|$ . According to SRSC property, we have

$$f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \geq \langle \nabla f_{S^*}(\hat{\beta}), -\hat{\beta} \rangle + \frac{\varphi_\mu}{2} \|\hat{\beta}\|_2^2$$

Because of  $\langle \nabla f_{S^*}(\hat{\beta}), -\hat{\beta} \rangle \geq 0$ , we have

$$\mu \hat{\beta}_{\min}^2 \leq \|\hat{\beta}\|_2^2 \leq \frac{2}{\varphi_\mu} \left[ f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \right]$$

$$2\varphi_\mu \cdot \eta^2 \left[ f_{S^*}(\hat{\beta}) - f_{S^*}(\beta^*) \right] \stackrel{(c)}{\leq} \frac{2}{\varphi_\mu} \left[ f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \right]$$

The inequality (c) follows the Equation (9). Let  $\alpha$  be  $\left(\frac{1}{\eta \varphi_\mu}\right)^2$ , then we have

$$f_{S^*}(\hat{\beta}) - f_{S^*}(\beta^*) \leq \alpha \left[ f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \right]$$

$$f_{S^*}(\hat{\beta}) \leq \frac{\alpha}{1+\alpha} \left[ f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \right] + f_{S^*}(\beta^*)$$

According to Lemma 2, we have

$$\frac{1}{\lambda} f_{\hat{S}}(\hat{\beta}) \leq f_{S^*}(\beta^*) \leq \frac{\alpha}{1+\alpha} \left[ f_{S^*}(\mathbf{0}) - f_{S^*}(\hat{\beta}) \right] + f_{S^*}(\beta^*)$$

$$f_{\hat{S}}(\hat{\beta}) - f_{S^*}(\beta^*) \leq \frac{\alpha \lambda}{1+\alpha} f_{S^*}(\mathbf{0}) + \left( \frac{\lambda}{1+\alpha} - 1 \right) f_{S^*}(\beta^*)$$

$\square$

Since the value of  $f_{S^*}(\mathbf{0})$  and  $f_{S^*}(\beta^*)$  are both constants and  $f_{S^*}(\beta^*)$  is close to 0, the error bound of our solution is depended on the value of  $\frac{\alpha \lambda}{1+\alpha}$ . When the ratio of data corruption  $\gamma$  is close to one,  $\lambda$  is close to one according to its definition. In addition, the value of  $\alpha$  is smaller when the parameter  $\varphi_\mu$  of the SRSC property is smaller. Therefore, the error of our solution can be close to zero when both  $\varphi_\mu$  and  $\gamma$  are large enough.

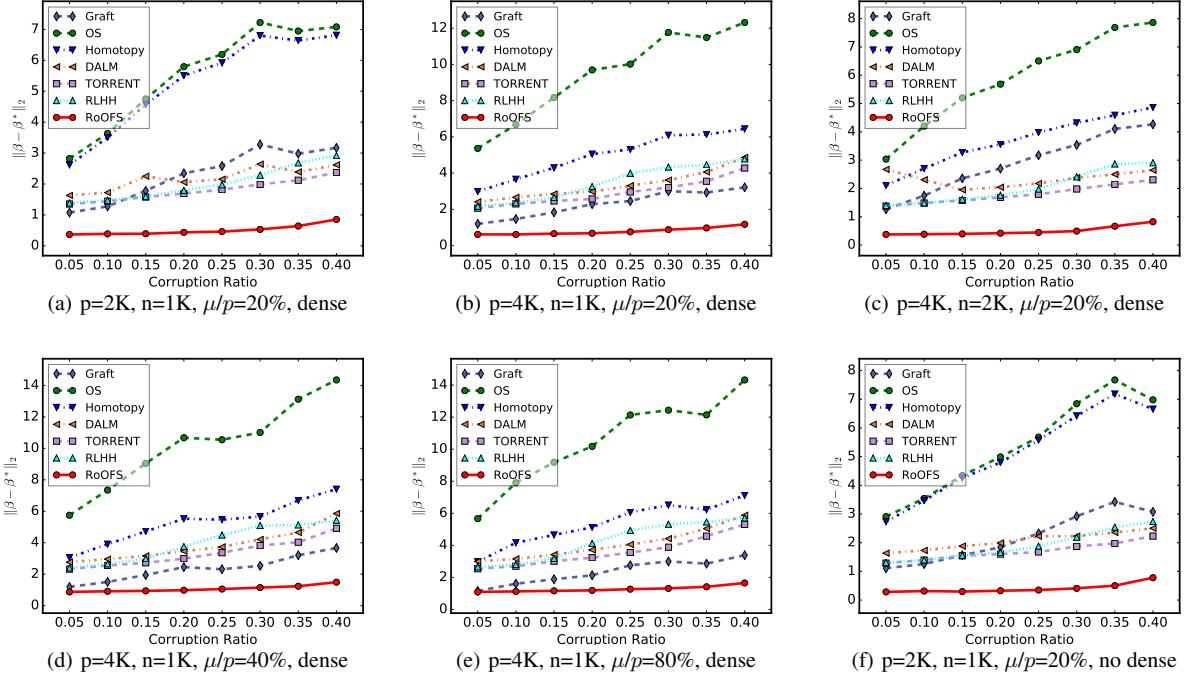


Figure 1: Performance on regression coefficients recovery for different corruption ratios in uniform distribution.

## 5 Experimental Results

In this section, we report the extensive experimental evaluation performed to verify the robustness, effectiveness of feature selection, and efficiency of the proposed method. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@3.6GHz) and 32.0GB memory. Details of both the source code and sample data used in the experiment can be downloaded here<sup>1</sup>.

### 5.1 Datasets and Metrics

To demonstrate the performance of our proposed method, comprehensive experiments are performed in synthetic datasets whose simulation samples were randomly generated according to the model in Equation (1). Specifically, we sample the regression coefficients  $\beta^* \in \mathbb{R}^p$  as a random unit norm vector with feature ratio constraint  $\|\beta\|_0 = \mu$ . The data matrix  $X$  was drawn independently and identically distributed from  $x_i \sim \mathcal{N}(0, I_p)$  and the uncorrupted response variables were generated as  $y_i^* = x_i^T \beta^*$ . The set of uncorrupted samples  $S$  was selected as a uniformly random  $\tau_*$ -sized subset of  $[n]$ . The response vector  $y$  containing corrupted samples was generated as  $y = y^* + u + \varepsilon$ , where the corruption vector  $u$  was sampled from the uniform distribution  $[-5\|y^*\|_\infty, 5\|y^*\|_\infty]$  and the additive dense noise was  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

Following the setting in (Bhatia et al. 2015)(Zhang et al. 2017), we measured the performance of the regression coefficients recovery using the standard  $L_2$  error  $e = \|\hat{\beta} - \beta^*\|_2$ , where  $\hat{\beta}$  represents the recovered coefficients for each method and  $\beta^*$  is the true regression coefficients. To validate

the performance for corrupted set discovery, the F1 score is measured by comparing the discovered corrupted sets with the actual ones. Similarly, the F1 score is also used to measure the effectiveness of feature selection by comparing the selected feature set with actual ones. To compare the scalability of each method, the CPU running time for each of the competing methods was also measured.

### 5.2 Comparison Methods

The following methods are included in the performance comparison presented here: *Grafting* (Perkins and Theiler 2003b). The *Grafting* method is an online version of  $L_1$  regularization approach to select features. *Online Substitution* (OS) (Yang et al. 2016) is a parameter-free online feature selection algorithm with limited-memory. Both *Grafting* and *OS* cannot handle the adversarial data corruption and train models without considering data corruption. We also compared our method to the robust regression methods (Wright and Ma 2010) (Nguyen and Tran 2013). *Homotopy* and *DALM* are two  $L_1$  based solvers that outperform other  $L_1$  methods both in terms of recovery properties and running time (Yang et al. 2010). A hard thresholding method, *TORRENT* (abbreviated "TORR") (Bhatia et al. 2015), developed for robust regression was also compared to our method. As the method requires a parameter for the corruption ratio, which is difficult to estimate in practice, we chose two versions of parameter settings: *TORR\** and *TORR25*. *TORR\** uses the true corruption ratio as its parameter, and *TORR25* applies parameter that is uniformly distributed across the range of  $\pm 25\%$  off the true value. Another recently proposed heuristic hard thresholding method, *RLHH* (Zhang et al. 2017), is also compared in our experiment. The method is a parameter-free approach, where the data corruption is esti-

<sup>1</sup><https://goo.gl/C4HQjo>

Table 2: F1 scores for the performance on uncorrupted set recovery.

	p=2K, n=1K, $\mu/p=20\%$				p=2K, n=2K, $\mu/p=20\%$				p=4K, n=2K, $\mu/p=20\%$			
	10%	20%	30%	40%	10%	20%	30%	40%	10%	20%	30%	40%
<b>Homotopy</b>	0.980	0.912	0.849	0.682	0.977	0.923	0.854	0.834	0.970	0.923	0.845	0.775
<b>DALM</b>	0.976	0.915	0.865	0.825	0.973	0.921	0.885	0.924	0.962	0.946	0.926	0.898
<b>TORR*</b>	0.983	0.950	0.927	0.893	0.983	0.960	0.919	0.934	0.978	0.954	0.934	0.916
<b>TORR25</b>	0.965	0.899	0.842	0.762	0.961	0.909	0.828	0.770	0.958	0.905	0.848	0.752
<b>RLHH</b>	0.979	0.945	0.933	0.901	0.978	0.966	0.936	0.914	0.980	0.959	0.940	0.896
<b>RoOFS</b>	<b>0.991</b>	<b>0.986</b>	<b>0.974</b>	<b>0.933</b>	<b>0.993</b>	<b>0.991</b>	<b>0.976</b>	<b>0.946</b>	<b>0.993</b>	<b>0.988</b>	<b>0.975</b>	<b>0.923</b>
	p=2K, n=1K, $\mu/p=60\%$				p=2K, n=1K, $\mu/p=20\%$ (nd)				p=4K, n=2K, $\mu/p=20\%$ (nd)			
	10%	20%	30%	40%	10%	20%	30%	40%	10%	20%	30%	40%
<b>Homotopy</b>	0.979	0.932	0.829	0.708	0.972	0.923	0.853	0.717	0.985	0.913	0.868	0.789
<b>DALM</b>	0.975	0.939	0.863	0.826	0.965	0.910	0.886	0.842	0.984	0.951	0.937	0.889
<b>TORR*</b>	0.979	0.957	0.937	0.870	0.974	0.950	0.935	0.896	0.988	0.960	0.947	0.904
<b>TORR25</b>	0.952	0.912	0.833	0.690	0.952	0.911	0.859	0.758	0.968	0.908	0.864	0.745
<b>RLHH</b>	0.975	0.959	0.928	0.845	0.973	0.959	0.935	0.907	0.983	0.965	0.940	0.912
<b>RoOFS</b>	<b>0.982</b>	<b>0.984</b>	<b>0.962</b>	<b>0.910</b>	<b>0.989</b>	<b>0.993</b>	<b>0.985</b>	<b>0.947</b>	<b>0.994</b>	<b>0.991</b>	<b>0.988</b>	<b>0.933</b>

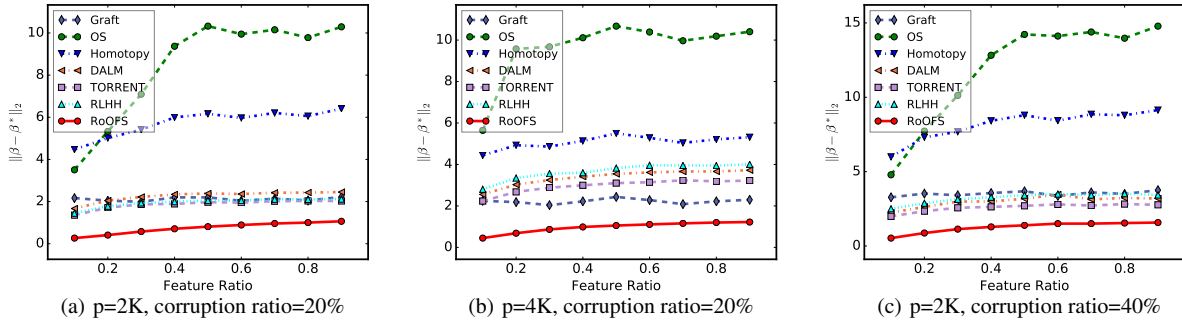


Figure 2: Performance on regression coefficients recovery for different ratios of feature ratio (n=1K, dense noise).

mated by a heuristic hard thresholding method. As all these robust methods are not designed for online feature selection, we run them individually in different feature batches and select features with largest  $\mu$  weights in regression coefficients when  $\|\beta\|_0 = \mu$ . All the results are averaged over 10 runs.

### 5.3 Recovery of regression coefficients

We selected 6 competing methods with which to evaluate the recovery performance of regression coefficients  $\beta$ : *Grafting*, *OS*, *Homotopy*, *DALM*, *TORR*, *RLHH*. Figures 1(a) and 1(b) show the recovery performance for different feature numbers when the data size is fixed. The results show that 1) the proposed method, *RoOFS*, outperforms all the competing methods in all the setting of corruption ratios, and 2) The performance of *RoOFS* is very resistant to the corruption data because the error of *RoOFS* method increases much more slowly than others when corruption ratio increases from 5% to 40%. Figures 1(b) and 1(c) show that when data size increases, we have similar conclusion on the performance except the overall error is decreased since more data is applied. Figures 1(d) and 1(e) show that the result of coefficient recovery remains the same when the number of selected features increase. Figure 1(f) shows that almost all the methods without the dense noise setting perform more than 50% better than that in dense noise settings. Specifically, the error of *RoOFS* is close to zero which means it can almost exactly recover the ground true regression coef-

ficients without the dense noise setting.

Figure 2 shows the recovery performance of regression coefficients in different ratios of feature sparsity. In general, the performance of *RoOFS* method outperforms all the other competing methods in all the data settings. Figure 2(a) and 2(b) show that 1) when the feature ratio increases, the recovery error of *RoOFS* method grows linearly with a small slope, which means our approach can be fitted into different settings of feature sparsity, and 2) the *RoOFS* method performs constantly well when the feature number increases from 2K to 4K. In addition, Figure 2(a) and 2(c) show that the *RoOFS* method is robust to corrupted data, because the error is not significantly impacted when the corruption ratio increases from 20% to 40%.

### 5.4 Recovery of Uncorrupted Set

As the online feature selection methods *Grafting* and *OS* do not explicitly estimate uncorrupted sets, we compared our proposed method with the robust methods: *Homotopy*, *DALM*, *TORR*, and *RLHH*. For the *TORR* algorithm, we use two parameter settings of *TORR\** and *TORR25* for 0% and 25% deviation of true corrupted ratio, respectively. Table 2 shows the following: 1) *RoOFS* outperforms all the other methods up to 14.9% in different settings of data sizes, feature numbers and ratios of feature sparsity. 2) When increasing the corruption ratio, the F1 scores decrease for all the methods. Also, the F1 scores slightly increase 0.5% in aver-



Table 3: F1 score on Performance of Feature Selection (cr=30%).

	p=2K, n=1K, dense						p=2K, n=2K, dense					
	10%	20%	30%	40%	50%	60%	10%	20%	30%	40%	50%	60%
<b>Grafting</b>	0.124	0.219	0.454	0.636	0.759	0.838	0.140	0.223	0.322	0.413	0.546	0.705
<b>OS</b>	0.700	0.646	0.669	0.726	0.810	<b>0.906</b>	0.730	0.671	0.685	0.685	0.704	0.737
<b>Homotopy</b>	0.132	0.223	0.305	0.404	0.498	0.595	0.137	0.238	0.317	0.400	0.496	0.605
<b>DALM</b>	0.221	0.275	0.343	0.433	0.516	0.611	0.347	0.350	0.382	0.461	0.543	0.629
<b>TORR</b>	0.275	0.297	0.368	0.446	0.525	0.615	0.455	0.419	0.446	0.499	0.565	0.639
<b>RLHH</b>	0.193	0.261	0.338	0.416	0.510	0.606	0.440	0.416	0.436	0.489	0.548	0.637
<b>RoOFS</b>	<b>0.890</b>	<b>0.894</b>	<b>0.888</b>	<b>0.884</b>	<b>0.895</b>	0.902	<b>0.916</b>	<b>0.913</b>	<b>0.909</b>	<b>0.905</b>	<b>0.895</b>	<b>0.888</b>
	p=4K, n=2K, dense						p=4K, n=2K, no dense					
	10%	20%	30%	40%	50%	60%	10%	20%	30%	40%	50%	60%
<b>Grafting</b>	0.122	0.214	0.346	0.547	0.704	0.789	0.113	0.205	0.311	0.566	0.705	0.798
<b>OS</b>	0.662	0.638	0.653	0.695	0.827	<b>0.922</b>	0.650	0.627	0.643	0.710	0.823	<b>0.927</b>
<b>Homotopy</b>	0.114	0.205	0.307	0.398	0.500	0.667	0.106	0.206	0.297	0.399	0.500	0.667
<b>DALM</b>	0.206	0.257	0.339	0.424	0.515	0.608	0.215	0.275	0.346	0.433	0.505	0.613
<b>TORR</b>	0.275	0.309	0.359	0.444	0.525	0.617	0.287	0.310	0.381	0.449	0.526	0.616
<b>RLHH</b>	0.180	0.243	0.319	0.424	0.523	0.597	0.215	0.258	0.340	0.425	0.513	0.613
<b>RoOFS</b>	<b>0.872</b>	<b>0.888</b>	<b>0.876</b>	<b>0.888</b>	<b>0.887</b>	0.907	<b>0.880</b>	<b>0.887</b>	<b>0.886</b>	<b>0.884</b>	<b>0.891</b>	0.894

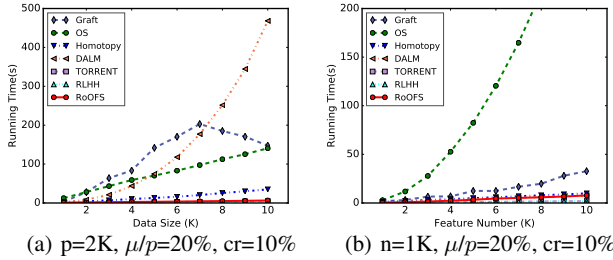


Figure 3: Running time for different data and feature sizes.

age when data size become two times larger, which indicates the number of features has few influence on the estimation of uncorrupted set. 3) The result of *TORR* methods is highly dependent on the corruption ratio parameter: the results of *TORR\** is up to 26.1% better than *TORR25*. It is important to note that the true corruption ratio parameter used in *TORR\** cannot be estimated exactly in practice. 4) Without the dense noise settings, the F1 scores increase less than 1% compared to the F1 score based on the same setting with dense noise, which shows that dense noise has small impact on the performance of uncorrupted set recovery.

### 5.5 Performance of Feature Selection

We selected all the six competing methods to evaluate the performance of feature selection in different settings including data sizes, feature numbers, and dense noises. For each data setting, we chose different ratios of feature sparsity (also known as  $\mu/p$ ) ranging from 10% to 60%. Table 3 shows the following: 1) the F1 scores of *RoOFS* method is up to 69.2% better than other methods, especially when the feature ratio is less than 40%. 2) Although the F1 scores of most methods such as *Grafting* and *TORR* are above 0.6 when the ratio is larger than 50%, the performance degraded significantly when the ratio decreased to 10%. However, the F1 score of *RoOFS* method is constantly higher than 0.85 in all the ratios of features. 3) *OS* method is very competitive in the task of feature selection; however, it still has lower F1

scores in all the settings when the ratio is less than 50%. 4) The setting of dense noise does not have significant impact on the performance of feature selection, since the F1 score without dense noise is only less than 1% larger than that in dense noise setting.

### 5.6 Efficiency

To evaluate the efficiency of our proposed method, we compared the performances of all the competing methods for two difference settings: data sizes and feature numbers. For *Grafting* and *OS* methods, the online features are handled individually due to their design. For the other methods, a hundred features are handled together as a batch. As Figure 3 shows, we found the following: 1) *RoOFS* algorithm has a very competitive efficiency compared to the thresholding based methods, *TORR* and *RLHH*, and significantly outperforms other four methods. 2) The running time of *RoOFS* algorithm increases linearly when both data size and feature number increase, which indicates that our algorithm can be scaled to massive datasets. 3) The running time of *DALM* method increases exponentially when data size increases, however, its efficiency has rarely impacted by increasing the number of features. 4) The efficiency of *Grafting* method fluctuates largely on the different data sizes, which indicates that its running time depends on the data size and content of data.

## 6 Conclusion

In this paper, a novel robust regression algorithm via online feature selection, *RoOFS*, is proposed to recover the regression coefficients and the uncorrupted set under the assumption that features cannot be accessed entirely at one time. To achieve this, we designed a robust online substitution method to alternately estimate the optimal uncorrupted set and substitute the retained feature set with newly updated features. We demonstrate that our algorithm can recover regression coefficients with a restricted error bound compared to ground truth. Extensive experiments on massive simulation data demonstrated that the proposed algorithm outperforms other competing methods in both effectiveness and efficiency.

## References

- Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.
- Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 774–782. JMLR Workshop and Conference Proceedings, May 2013.
- Wei Jiang, Guihua Er, Qionghai Dai, and Jinwei Gu. Similarity-based online feature selection in content-based image retrieval. *IEEE Transactions on Image Processing*, 15(3):702–712, 2006.
- Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.
- Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.
- Rebecca Walker Naylor, Cait Poyner Lamberton, and Patricia M West. Beyond the like button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings. *Journal of Marketing*, 76(6):105–120, 2012.
- Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via  $\ell_1$ -minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.
- Simon Perkins and James Theiler. Online feature selection using grafting. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 592–599, 2003.
- Simon Perkins and James Theiler. Online feature selection using grafting. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML’03, pages 592–599. AAAI Press, 2003.
- S Ryali and V Menon. Feature selection and classification of fmri data using logistic regression with  $\ell_1$  norm regularization. *NeuroImage*, 47:S57, 2009.
- Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):698–710, 2014.
- Jing Wang, Meng Wang, Peipei Li, Luoqi Liu, Zhongqiu Zhao, Xuegang Hu, and Xindong Wu. Online feature selection with group structure analysis. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3029–3041, 2015.
- John Wright and Yi Ma. Dense error correction via  $\ell_1$ -minimization. *IEEE Trans. Inf. Theor.*, 56(7):3540–3560, July 2010.
- Xindong Wu, Kui Yu, Hao Wang, and Wei Ding. Online streaming feature selection. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1159–1166, 2010.
- Allen Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma. Fast  $\ell_1$ -minimization algorithms and an application in robust face recognition: A review. Technical Report UCB/EECS-2010-13, EECS Department, University of California, Berkeley, Feb 2010.
- Haichuan Yang, Ryohei Fujimaki, Yukitaka Kusumura, and Ji Liu. Online feature selection: A limited-memory substitution algorithm and its asynchronous parallel variation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 1945–1954, New York, NY, USA, 2016. ACM.
- Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. Towards scalable and accurate online feature selection for big data. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 660–669. IEEE, 2014.
- Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. Scalable and accurate online feature selection for big data. *ACM Trans. Knowl. Discov. Data*, 11(2):16:1–16:39, December 2016.
- Xuchao Zhang, Liang Zhao, Arnold P. Boedihardjo, and Chang-Tien Lu. Robust regression via heuristic hard thresholding. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3434–3440, 2017.
- Jing Zhou, Dean P Foster, Robert A Stine, and Lyle H Ungar. Streamwise feature selection. *Journal of Machine Learning Research*, 7(Sep):1861–1885, 2006.
- Jun Zhu, Ni Lao, and Eric P Xing. Grafting-light: fast, incremental feature selection and structure learning of markov random fields. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 303–312. ACM, 2010.