# Robust Elastic Net Regression

| Weiyang Liu | Rongmei Lin | Meng Yang |
|:-:|:-:|:-:|
| Peking Univ. | Peking Univ. | Shenzhen Univ. |
| wyliu@pku.edu.cn | scutrmlin@gmail.com | yang.meng@szu.edu.cn |

April 25, 2015

**Abstract**

We propose a robust elastic net (REN) model for high-dimensional sparse regression and give its performance guarantees (both the statistical error bound and the optimization bound). A simple idea of trimming the inner product is applied to the elastic net model. Specifically, we robustify the covariance matrix by trimming the inner product based on the intuition that the trimmed inner product can not be significantly affected by a bounded number of arbitrarily corrupted points (outliers). The REN model can also derive two interesting special cases: robust Lasso and robust soft thresholding. Comprehensive experimental results show that the robustness of the proposed model consistently outperforms the original elastic net and matches the performance guarantees nicely.

## Introduction

Over the past decades, sparse linear regression has been and is still one of the most powerful tools in statistics and machine learning. It seeks to represent a response variable as a sparse linear combination of covariates. Recently, sparse regression has received much attention and found interesting applications in cases where the number of variables $p$ is far greater than the number of observations $n$. The regressor in high-dimensional regime tends to be sparse or near sparse, which guarantees the high-dimensional signal can be efficiently recovered despite the underdetermined nature of the problem [3, 7]. However, data corruption is very common in high-dimensional big data. Research has demonstrated the current sparse linear regression (e.g. Lasso) performs poorly when handling dirty data [6]. Therefore how to robustify the sparse linear regression becomes a major concern that draws increasingly more attentions.

Robust sparse linear regression can be roughly categorized into several lines of researches. One type of representative approaches is to first remove the detected outliers and then perform the regression. However, outlier removal is not suitable for the high-dimensional regime, because outliers might not exhibit any strangeness in the ambient space due to the high-dimensional noise [22]. Another type of approaches include replacing the standard mean squared loss with a more robust loss function such as trimmed loss and median squared loss. Such approaches usually can not give performance guarantees. Methods [12, 15, 13] have been developed to handle arbitrary corruption in the response variable, but fail with corrupted covariates. [14] proposes a robust Lasso that considers the stochastic noise or small bounded noise in the covariates. [5] also considers similar corruption settings and proposes a robust OMP algorithm for Lasso. For the same noise, [17] proposes the matrix uncertainty selector that serves as a robust estimator.

Other than the above noise, the persistent noise [6, 10] which is ill-modeled by stochastic distribution widely exists in many practical applications. We consider this type of noise in the paper. We make no assumptions on the distribution, magnitude or any other such properties of the noise (outliers). The only prior that we use is the bound on the fraction of corrupted samples. Aiming to address such noise problem in sparse linear regression, [6] proposes a robust Lasso that utilizes trimmed statistic. [10] also considers the same noise in the logistic regression using trimmed statistic. Yet robust Lasso still has some inevitable limitations that are inherent from the Lasso [18]:

- Variable Saturation: when $p > n$, the Lasso selects at most $n$ variables after saturation [9]

- Non-Grouping Effect: suppose there is a group of highly correlated variables, the Lasso tends to select only one variable from the group. [24]

- Invalidation: when $p < n$ and there are high correlations between variables, it can be observed that the selection performance of the Lasso is determined by the ridge regression [18].

Elastic net [24] is proposed to address the above shortcomings by combining the Lasso and Ridge constraints. However, similar to the Lasso, its robustness can not be guaranteed. In order to address such shortcomings and simultaneously achieve the robustness, we propose the robust elastic net (REN) for sparse linear regression. Specifically, REN adopts the trimmed inner product to robustly estimate the covariance matrix while adopting a mixed $l_1$ and $l_2$ constraints. REN combines the robust Lasso [6] and the robust soft thresholding (RST) into a unified framework, and provides a unified performance guarantee. Our main theorem provides bounds on the fraction of corrupted samples that REN can tolerate, while still bounding the $l_2/l_1$ error of the support recovery.

## Related Work

Sparse recovery without corruption has been well studied in the past years, even for the high-dimensional settings $p \geq n$. When the covariance matrix satisfies some conditions, e.g. Eigenvalue property [2, 20], $\beta^*$ can be recovered with high probability. It is also learned that some random designs of matrix $\boldsymbol{X}$ also satisfies such property [16]. Various estimators are also proposed to solve the $l_1$ regularized least squares problem, including basis pursuit [4], orthogonal matching pursuit [19], etc. However, these existing methods are not robust to outlier. Lasso and OMP are sensitive to corruptions. Even one entry in $\boldsymbol{X}$ or $\boldsymbol{y}$ may paralyze Lasso and OMP. Some work considers a natural modification of Lasso by adding a corruption term and regularizing it to account for certain types of corruptions. This modification is non-convex due to the bilinearity and therefore the performance guarantees can not be provided. In order to give provable performance guarantees, [14] considers several corruptions including additive noise, multiplicative noise, missing data, and further construct the robust covariance matrix surrogates that could recover the sparse support with high probability. However, it has strong assumption on the corruption type. [6] improves the robust surrogates with trimmed estimators for Lasso. Similarly, [10] applies the trimmed statistics to proposes a robust logistic regression. Such idea of trimming also appears in designing robust PCA [11]. On the other hand, current work on robustifing the elastic net includes reducing the effect of outliers with reweighted procedure [21], which is similar to the idea in [23]. Like most outlier detection based approaches, using reweighted procedure in the elastic net may fail with large-scale high-dimensional data.

## Contributions

We utilize the trimmed statistic to design a robust elastic net for high-dimensional linear regression. The contributions can be summarized as follows.

- The robust elastic net regression is proposed with an efficient solving algorithm. The REN model extends the results in [6].

- We derives the statistical error bound for the support recovery of the REN model. As special cases, we also discuss the statistical error bound for the robust Lasso and the RST.

- To validate the statistical error bound we derived, we further propose an efficient projected gradient descent and gives the provable optimization error bound when using this algorithm. The optimization error bound guarantees the statistical error bound can be achieved in practice.

# Robust Elastic Net Regression

## Problem Setup

We consider the sparse linear regression problem in the high-dimensional settings where the variable number is much larger than the number of observations, i.e. $p \gg n$. Let $\beta^*$ be the $k$-sparse $(k < p)$ groundtruth parameter for the regressor, and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n+n_o} \in \mathbb{R}^p \times \mathbb{R}$ denote the covariate-response observation pairs. Without corruption, we assume these observations follow the linear model:

$$y_i = \langle \boldsymbol{x}_i, \beta^* \rangle + \epsilon_i \tag{1}$$

where $\epsilon_i$ is additive noise that is independent of $\boldsymbol{x}_i$. The actual observation are corrupted by the following model.

**Definition 1** (Fractional Adversarial Corruption). *Up to $n_o$ out of these $n + n_o$ pairs, including both $\boldsymbol{x}_i$ and $y_i$, are arbitrarily corrupted. The outlier fraction is bounded by $\frac{n_o}{n}$.*

Note that we make no assumption on these outliers. The outliers are even not necessarily formed into pairs, namely the corrupted rows in $\boldsymbol{X}$ and $\boldsymbol{y}$ can be different. They can follow any types of distributions. The corruption settings are also the same as [6]. Since the outlier fraction is bounded by $\frac{n_o}{n}$, we assume the outlier fraction is exactly $\frac{n_o}{n}$ without loss of generality. The remaining $n$ non-corrupted samples are called authentic for conciseness, and assumed to obey the standard sub-Gaussian construction [14].

**Definition 2** (Sub-Gaussian Construction). *A random variable $v$ is sub-gaussian with parameter $\sigma$ if $\mathbb{E}\{\exp(tv)\} \leq \exp(\frac{t^2\sigma^2}{2})$ for all real $t$. A random matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is sub-Gaussian with parameter $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$ if: (a) each row $\boldsymbol{x}_i^T \in \mathbb{R}^p$ is sampled independently from a zero-mean distribution with covariance $\frac{1}{n}\Sigma_x$, (b) for any unit vector $\boldsymbol{u} \in \mathbb{R}_p$, the random variable $\boldsymbol{u}^T \boldsymbol{x}_i$ is sub-Gaussian with parameter at most $\sigma_x$.*

We assume the data matrix $\boldsymbol{X}$ is sub-Gaussian with parameter $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$ before adversarial corruption, and the additive noise $\epsilon$ is sub-Gaussian with parameter $\frac{1}{n}\sigma_\epsilon^2$.

## The REN Model

This section presents the details of the proposed robust elastic net model. We use the trimmed statistics to robustify the elastic net. Specifically, we replace the inner product in the original elastic net regression with the trimmed inner product. The simple intuition behind the trimmed statistics is that entries with overly large magnitude in the inner product are likely to be corrupted. If the outliers have too large magnitude, they are not likely to have positive influence on the correlation and thus will be detrimental to the elastic net. Otherwise, they have bounded negative influence on the support recovery. We first consider a simple program

$$\hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \hat{\Gamma} \beta - \langle \hat{\gamma}, \beta \rangle \right\} \tag{2}$$

where $\hat{\Gamma}, \hat{\gamma}$ denote the unbiased estimates of $\Sigma_x$ and $\Sigma_x \beta^*$ respectively. $\beta^*$ is the unique solution of this convex program given $R \geq \|\beta^*\|_1$. Lasso is also a special case of Eq. (2) with $\hat{\Gamma}_{\text{LAS}} = X^T X$ and $\hat{\gamma}_{\text{LAS}} = X^T y$. Under a suitable $\lambda$ and a positive semi-definite $\hat{\Gamma}$, Eq. (2) has an equivalent regularized form as

$$\hat{\beta} \in \arg \min_{\|\beta\|_1 \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^T \hat{\Gamma} \beta - \langle \hat{\gamma}, \beta \rangle + \lambda \|\beta\|_1 \right\} \tag{3}$$

where $\|\beta\|_1$ should be upper bounded by $b\sqrt{k}$ for a suitable constant $b$ if $\hat{\Gamma}$ has negative eigenvalues (in the case of $\hat{\Gamma}_{\text{LAS}}, \hat{\gamma}_{\text{LAS}}$). Since these two forms are equivalent, we merely study the first form. If $\hat{\Gamma}_{\text{EN}} = \alpha X^T X + (1-\alpha) \cdot I, \alpha \in [0,1]$ and $\hat{\gamma}_{\text{EN}} = X^T y$, Eq. (2) becomes the elastic net model. However, both the Lasso and the elastic net are fragile when corruption exists, especially under our noise settings. Thus we use the following trimmed statistic as robust surrogates for $\hat{\Gamma}$ and $\hat{\gamma}$:

$$\{\hat{\Gamma}_{\text{REN}}\}_{ij} = \alpha \sum_{k=1}^{n_o} [X_i, X_j]_{(k)} + (1-\alpha) \cdot I_{ij}$$

$$\{\hat{\gamma}_{\text{REN}}\}_j = \sum_{k=1}^{n_o} [X_j, y]_{(k)} \tag{4}$$

where $\alpha \in [0,1]$, and $[u, v]_{(k)}$ ($u, v \in \mathbb{R}^h$) denotes the $k$th smallest variable in $\{q_i = |u_i \cdot v_i|, \forall i\}$ such that $[u, v]_{(1)} \leq [u, v]_{(2)} \leq \cdots \leq [u, v]_{(h)}$. $I$ is an identity matrix. $X_i$ denotes the $i$th column of the matrix $X$. For example, $\sum_{k=1}^{n_o} [X_i, X_j]_{n_o}$ is to select the $n_o$ smallest element products between $|X_i|$ and $|X_j|$ and sum them. The robust surrogates are equivalent to adopting the summation of top $n_o$ variables in the element-wise vector multiplication. After substituting the robust surrogates, the optimization in REN may become non-convex because $\hat{\Gamma}$ may contain negative eigenvalues. Therefore it is also important to develop an algorithm that can approximate the optimum, otherwise the performance bound may be useless. We use the following polynomial-time projected gradient descent update:

$$\beta^{t+1} = \Pi_{l_1(R)} \left( \beta^t - \frac{1}{\eta} (\hat{\Gamma}_{\text{REN}} \beta^t - \hat{\gamma}_{\text{REN}}) \right) \tag{5}$$

where $\Pi_{l_1(R)}(v) = \arg \min_z \{\|v - z\|_2 \mid \|z\|_1 \leq R\}$ denotes Euclidean projection onto a $l_1$ ball of radius $R$. The optimization error is also bounded (the bound is given in the paper) so that the projected gradient descent can approximate the optimum with satisfactory accuracy. Some remarks for the REN model are in order:

- The REN model contains a family of sparse regression methods. The model parameter $\alpha$ controls the tradeoff between the ridge penalty and the robust trimmed covariance matrix (robust Lasso penalty). The REN model naturally bridges these two robustified model.

- When $\alpha = 0$, the REN model becomes the robust thresholding regression. When $\alpha = 1$, the REN degenerates to the robust Lasso [6].

- The parameter $\alpha$ shrinks the sample covariance matrix towards the identity matrix. Adopting $\hat{\Gamma}_{\mathrm{REN}}$ is mathematically equivalent to replacing the trimmed covariance matrix $\Sigma_x$ with a shrunken version in the robust Lasso.

## Projected Gradient Descent for REN

We present the proposed solving algorithm for the REN model. In order to validate the performance bound of the REN model, it is necessary to consider an algorithm whose solution should be close enough to the global optimum. The gradient of the quadratic loss function takes the form:

$$\nabla \mathcal{L}(\beta) = \hat{\Gamma}_{\mathrm{REN}} \beta - \hat{\gamma}_{\mathrm{REN}}. \tag{6}$$

We use the projected gradient descent that generates a sequence of iterates $\{\beta_t, t = 0, 1, 2, \cdots\}$ by the recursion:

$$\beta^{t+1} = \arg \min_{\|\beta\| \leq R} \left\{ \mathcal{L}(\beta^t) + \langle \nabla \mathcal{L}(\beta^t), \beta - \beta^t \rangle + \frac{\eta}{2} \|\beta - \beta^t\|_2^2 \right\} \tag{7}$$

where $\eta > 0$ is a step-size parameter. This update can be translated to $l_2$ projection onto the $l_1$ ball with radius $R$, which is in fact equivalent to Eq. (5). According to [8], this update can be computed rapidly in $\mathcal{O}(p)$ time. For the regularized version Eq. (3), we just need to include $\lambda \|\beta\|_1$ in the update and perform two projections onto the $l_1$ ball [1].

When the objective functions in Eq. (2) and Eq. (3) are convex, or equivalently, $\hat{\Gamma}$ is positive semi-definite, the update in Eq. (7) is guaranteed to converge to the global optimum. Although the $\hat{\Gamma}_{\mathrm{REN}}$ is not positive semi-definite, we still have the optimization bound that guarantees the obtained local optimum is close enough to the global optimum. The algorithm is summarized as follows.

---
**Algorithm 1** Robust Elastic Net
---
**Input:** $X, y, R, n_o, \alpha$
**Output:** $\hat{\beta}$
1: Compute $\hat{\Gamma}_{\mathrm{REN}}$ and $\hat{\gamma}_{\mathrm{REN}}$ via Eq. (4).
2: **while** not reach maximal iteration or satisfied accuracy **do**
3:     Perform the projected gradient descent algorithm by iterating Eq. (7) to solve Eq. (2).
4: **end while**
5: Output the final $\hat{\beta}$.

---

## Performance Guarantees

This section provides the performance guarantees for the REN model and its solving algorithm. The statistical error measures the upper bound of the $l_1/l_2$ difference between estimated $\hat{\beta}$ and the groudtruth $\beta^*$. The optimization error measures the upper bound of the $l_1/l_2$ difference between the $\beta^t$ (after $t-1$ iterations) and the global optimum $\hat{\beta}$. All the detail proofs are provided in the supplementary material.

## Preliminaries

We first introduces a useful type of restricted eigenvalue (RE) conditions [2, 20].

**Definition 3** (Lower-RE Condition). *The matix $\hat{\Gamma}$ satisfies a lower restricted eigenvalue condition with curvature $\mu_1 > 0$ and tolerance $\tau(n,p) > 0$ if for all $\theta \in \mathbb{R}^p$, the following condition holds:*

$$\theta^T \hat{\Gamma} \theta \geq \mu_1 \|\theta\|_2^2 - \tau(n,p)\|\theta\|_1^2 \tag{8}$$

**Definition 4** (Upper-RE Condition). *The matix $\hat{\Gamma}$ satisfies a lower restricted eigenvalue condition with curvature $\mu_2 > 0$ and tolerance $\tau(n,p) > 0$ if for all $\theta \in \mathbb{R}^p$, the following condition holds:*

$$\theta^T \hat{\Gamma} \theta \leq \mu_2 \|\theta\|_2^2 - \tau(n,p)\|\theta\|_1^2 \tag{9}$$

When the Lasso matrix $\hat{\Gamma}_{\text{LAS}}$ satisfies both the lower-RE and upper-RE conditions, the Lasso guarantees low $l_2$ error for $\beta^*$ supported on any subset of size at most $1/\tau_l(n,p)$. In fact, we also prove that the proposed $\hat{\Gamma}_{\text{REN}}$ also satisfies the lower-RE and upper-RE conditions.

## Statistical Error Bound

**Theorem 1.** *Under the sub-Gaussian construction model, If we choose $R = \|\beta^*\|_1$ and the following conditions are satisfied:*

$$n \gtrsim \frac{\sigma_x^4}{\lambda_{min}^2(\Sigma_x)} k \log p, \tag{10}$$

$$\frac{n_o}{n} \lesssim \frac{\lambda_{min}(\Sigma_x)}{\sigma_x^2 k \log p}, \tag{11}$$

$$\max_j \left\{ \|\Sigma_{xj}\|_0 \right\} \leq c_1, \tag{12}$$

*where $c_1$ is an absolute constant and $\Sigma_{xj}$ denotes the jth column of the covariance matrix $\Sigma_x$, then with probability higher than $1 - p^{-2}$, the output of the REN satisfy the following $l_1/l_2$ statistical error bound:*

$$
\begin{aligned}
\frac{1}{2\sqrt{k}}\|\hat{\beta} - \beta^*\|_1 &\leq \|\hat{\beta} - \beta^*\|_2 \\
&\lesssim \frac{32}{\alpha \lambda_{min}(\Sigma_x) + 4(1-\alpha)} \left( \frac{\alpha k n_o \log p}{n} \sigma_x^2 \|\beta^*\|_2 \right. \\
&\left. + \sqrt{\frac{k\sigma_\epsilon^2 \log p}{n}} + \frac{n_o \log p \sqrt{k}\sigma_x}{n} \sqrt{\sigma_\epsilon^2 + \sigma_x^2 \|\beta^*\|_2} \right. \\
&\left. + (1-\alpha)c_2 k \sqrt{\frac{\sigma_x^2 \log p}{n}} \|\beta^*\|_2^2 \right)
\end{aligned}
\tag{13}
$$

*where $\alpha$ is the REN parameters and $c_2$ is an absolute constant.*

**Remark 1.** *The number of non-zero entries in all columns of the covariance matrix $\Sigma_x$ is assumed to be bounded by a constant such that the statistical error bound for the REN model is not trivial. In other words, for every variable, the number of the other correlated variables is bounded.*

**Corollary 1.** *Suppose that, under the sub-Gaussian construction model, Eq. (10), Eq. (11) are both satisfied, $\alpha$ is set to 1, and $R$ is set to $\|\beta^*\|_1$, the output of the REN model is, with probability higher than $1 - p^{-2}$, bounded*

*by*

$$\frac{1}{2\sqrt{k}}\|\hat{\beta} - \beta^*\|_1 \leq \|\hat{\beta} - \beta^*\|_2$$

$$\lesssim \frac{32}{\lambda_{min}(\Sigma_x)}\left(\frac{kn_o \log p}{n}\sigma_x^2\|\beta^*\|_2 + \sqrt{\frac{k\sigma_\epsilon^2 \log p}{n}}. \right. \tag{14}$$

$$\left. + \frac{n_o \log p\sqrt{k}\sigma_x}{n}\sqrt{\sigma_\epsilon^2 + \sigma_x^2\|\beta^*\|_2^2}\right)$$

**Remark 2.** *When $\alpha = 1$, the REN model becomes the robust Lasso. Note that the robust Lasso do not require Eq. (12) to be satisfied. Therefore, the error bound for the REN model ($\alpha = 1$) is non-trivial without any assumptions on the covariance matrix $\Sigma_x$. The error bound for the robust Lasso coincides with [6].*

**Corollary 2.** *Suppose that, under the sub-Gaussian construction model, Eq. (10), Eq. (11) are both satisfied, $\Sigma_x$ is assumed as $\boldsymbol{I}$, $\alpha$ is set to 0, and R is set to $\|\beta^*\|_1$, the output of the REN model is, with probability higher than $1 - p^{-2}$, bounded by*

$$\frac{1}{2\sqrt{k}}\|\hat{\beta} - \beta^*\|_1 \leq \|\hat{\beta} - \beta^*\|_2$$

$$\lesssim 8\left(\sqrt{\frac{k\sigma_\epsilon^2 \log p}{n}} + \frac{n_o \log p\sqrt{k}}{n}\sqrt{1 + \sigma_\epsilon\|\beta^*\|_2^2}. \right. \tag{15}$$

$$\left. + k\sqrt{\frac{\log p}{n}}\|\beta^*\|_2\right)$$

**Remark 3.** *When $\alpha = 0$, the REN model becomes the robust soft thresholding. Note that the statistical error bound for the RST requires $\Sigma_x = \boldsymbol{I}$ to be satisfied. The RST is essentially a $l_1$ relaxation of robust thresholding regression [6].*

## Optimization Error Bound

**Lemma 1.** *Under the sub-Gaussian construction model, $\hat{\Gamma}_{\text{REN}}$ satisfies the lower-RE and upper-RE conditions with the following parameters:*

$$\mu_1 = \alpha\frac{\lambda_{\min}(\Sigma_x)}{2} + (1 - \alpha), \tag{16}$$

$$\mu_2 = \alpha\frac{3\lambda_{\max}(\Sigma_x)}{2} + (1 - \alpha), \tag{17}$$

$$\tau(n, p) \leq \frac{\alpha}{8}\lambda_{\min}(\Sigma), \tag{18}$$

*with high probability.*

**Theorem 2.** *Under the conditions of **Theorem 1** and **Lemma 1**, for the projected gradient descend method, if we choose the step size as $2\mu_2$, then there must exist absolute constants $c_1, c_2 > 0$, $\gamma \in (0, 1)$ such that the following inequalities hold for all $t \in \mathbb{N}$ with high probability,*

$$\|\beta^t - \hat{\beta}\|_2^2 \leq \gamma^t\|\beta^0 - \hat{\beta}\|_2^2 + c_1\frac{\log p}{n}\|\hat{\beta} - \beta^*\|_1^2$$

$$+ \|\hat{\beta} - \beta^*\|_2^2, \tag{19}$$

$$\|\beta^t - \hat{\beta}\|_1 \leq 2\sqrt{k}\|\beta^t - \hat{\beta}\|_2 + 2\sqrt{k}\|\hat{\beta} - \beta^*\|_2$$

$$+ 2\|\hat{\beta} - \beta^*\|_1, \tag{20}$$

*where $\beta^t$ denotes the $t^{\text{th}}$ gradient descent iterate.*
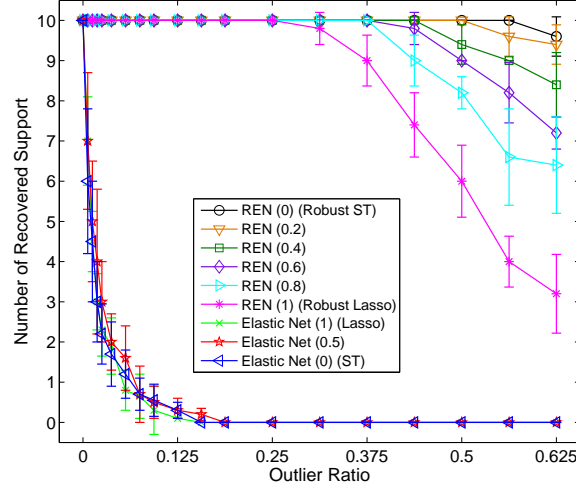
7

Figure 1. Number of recovered support for different methods while $\boldsymbol{X}$ has independent columns. The number beside REN is $\alpha$.
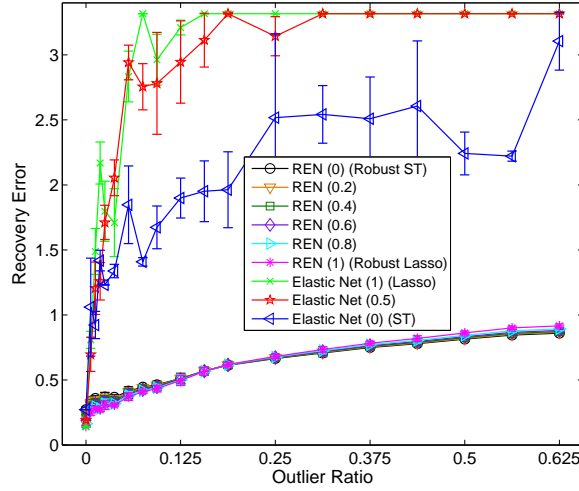


Figure 2. $l_2$ support recovery error for different methods with independent columns of $\boldsymbol{X}$.

**Remark 4.** *The optimization error bound controls the $l_1/l_2$-distance between the iterate $\beta^t$ and the global optimum $\hat{\beta}$ of Eq. (2). The optimization error bound means for a large enough iteration t, the statistical $l_1/l_2$ error bound can be achieved. Note that $\beta^t$ can be computed in polynomial-time while $\hat{\beta}$ is difficult to obtain. Combining the statistical error bound and the optimization error bound gives the conclusion that the $l_1$ and $l_2$ error are bounded as $\mathcal{O}(\frac{k^{3/2}\log p}{n})$ and $\mathcal{O}(\frac{k\log p}{n})$ respectively.*
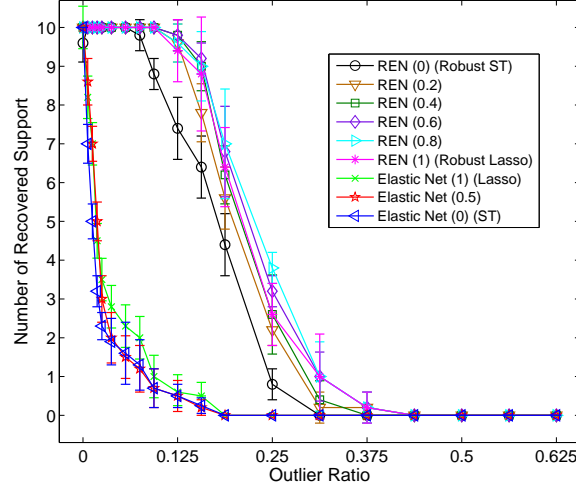
Figure 3. Number of recovered support for different methods while $\boldsymbol{X}$ has dependent columns.
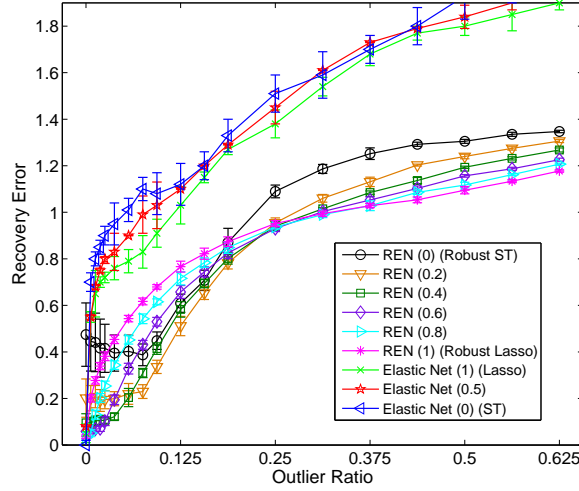


Figure 4. $l_2$ support recovery error for different methods with dependent columns of $\boldsymbol{X}$.

# Experiments

## Experimental Setup

Under the the case where $\boldsymbol{X}$ has independent columns, we generate the authentic rows $\{\boldsymbol{X}^{\mathcal{A}}, \boldsymbol{y}^{\mathcal{A}}\}$ following [6] via sub-Gaussian construction model with $\Sigma_x = \boldsymbol{I}$, $p = 4000$, $n = 1600$, $k = 10$ and $\sigma_\epsilon = 2$. The non-zero entries in $\beta^*$ are randomly selected as $\pm 1$. The corrupted rows $\{\boldsymbol{X}^{\mathcal{O}}, \boldsymbol{y}^{\mathcal{O}}\}$ are generated via the following procedure. We first obtain $\theta^*$ by $\arg\min_{\|\theta\|_1 \leq \|\beta^*\|_1, \theta \in \mathbb{R}^{p-k}} \|, \boldsymbol{y}^{\mathcal{A}} - \boldsymbol{X}^{\mathcal{A}}_{(\Lambda^*)^c}\theta\|_2$. Then we set $\boldsymbol{y}^{\mathcal{O}} = -\boldsymbol{X}^{\mathcal{O}}_{\Lambda^*}\beta^*$ where $\boldsymbol{X}^{\mathcal{O}}_{\Lambda^*} = \frac{3}{\sqrt{n}}\boldsymbol{S}$ and $\boldsymbol{S}$ is a $n_o \times k$ random matrix with each entry being $\pm 1$. Finally we let $\{\boldsymbol{X}^{\mathcal{O}}_{(\Lambda^*)^c}\}_i = (\frac{\boldsymbol{y}_i^{\mathcal{O}}}{\boldsymbol{B}_i^T\theta^*})\boldsymbol{B}_i^T$ where $\boldsymbol{B}_i$ is a $(n-k)$-dimensional vector with i.i.d. standard Gaussian distributed entries. While $\boldsymbol{X}$ has correlated columns, the data is generated using $\sigma_e = 1$, $\Sigma_x$ with diagonal being 1 and the other entries being
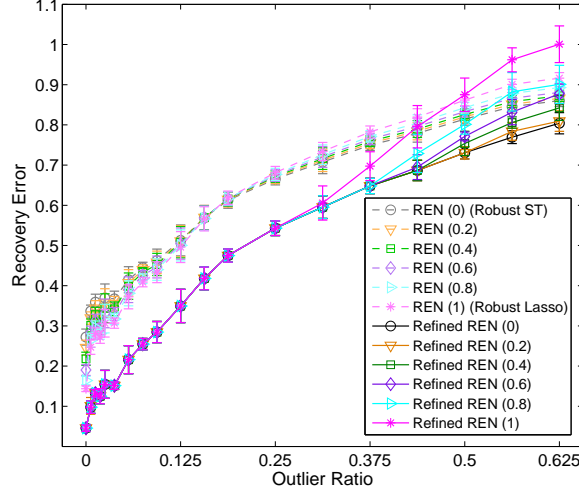
Figure 5. Number of recovered support for different methods while $\boldsymbol{X}$ has dependent columns.

0.4. The other parameters follows the independent case. For more detailed settings refer to the corresponding sections.

## Robustness Evaluation

The section evaluates the robustness of the REN model under different $\alpha$. We use the number of recovered support and the $l_2$ recovery error ($\|\hat{\beta} - \beta^*\|_2 / \|\beta^*\|_2$) for evaluation. We first consider the case where the data matrix $\boldsymbol{X}$ has independent columns. We vary the REN parameter $\alpha$ from 0 to 1 with step 0.2. The number of recovered support is shown in Fig. 1 and the corresponding $l_2$ support recovery error is given in Fig. 2. From Fig. 1 and Fig. 2, one can observe that REN consistently outperforms the original elastic net (including both Lasso and soft thresholding) in terms of both number of recovered support and the $l_2$ recovery error. REN can perfectly recover all the support number when outlier ratio is less than 0.5. Under the independent case, we find that REN with $\alpha = 0$, i.e. RST, performs best and REN with $\alpha = 1$, i.e. robust Lasso, has the worse performance. With $\alpha$ decreasing, the performance of REN gets better. It is partially because the authentic data satisfies $\Sigma_x = \boldsymbol{I}$ that is also the assumption for the statistical error bound of RST. The bound for RST is actually smaller than the bound for robust Lasso, so the performance of RST is the best when $\Sigma_x = \boldsymbol{I}$.

Then we evaluate the performance of REN in the case that $\boldsymbol{X}$ has correlated columns. The results are shown in Fig. 3 and Fig. 4. One can observe that REN still consistently outperforms the original elastic net (including both Lasso and soft thresholding). Although the performance of REN decreases with data containing correlated columns, it can still perfectly recover the locations of support with up to 90 outliers. The corresponding $l_2$ recovery error is also much smaller than the original elastic net. It can be observed that the elastic net with $\alpha = 1$ has the best performance against outliers, which coincides with the statistical error for the robust Lasso, because Robust Lasso do not requires the bounded number of non-zero values in the columns of $\Sigma_x$.

## Refinement for Recovery Error

We observe that even REN recovers most locations of support, but the $l_2$ corresponding recovery error is still very large. (sometime even larger than methods that do not recover the same number of support.) Thus we
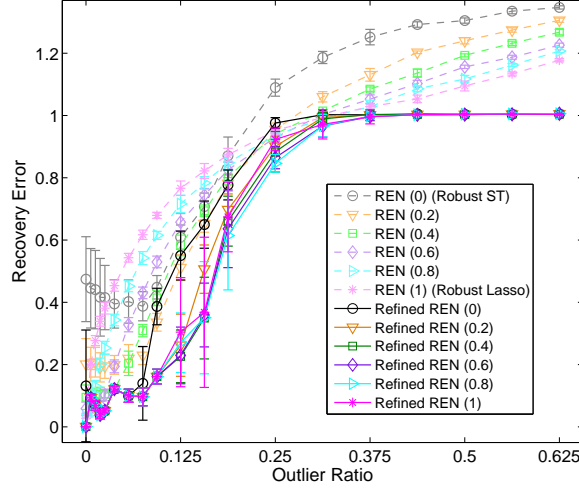
Figure 6. $l_2$ support recovery error for different methods with dependent columns of $\boldsymbol{X}$.

consider a standard refinement strategy for the REN model. When we recover the locations of support, we simply force the other locations in $\hat{\beta}$ to be zero ,and only use these recovered support to perform the least square regression with $\{\hat{\Gamma}_{\text{REN}}\}_{S,S}$ and $\{\hat{\gamma}_{\text{REN}}\}_S$. $S$ is the set of indexes corresponding to the location of the recovered support. Specifically, we use $\hat{\beta}_S = \{\hat{\Gamma}_{\text{REN}}\}_{S,S}^{-1} \cdot \{\hat{\gamma}_{\text{REN}}\}_S$ and set $\hat{\beta}_{S^c}$ as zeros. The refined $l_2$ recovery error for REN is shown in Fig. 5 and Fig. 6. From Fig. 5, one can see that the refined recovery error is much smaller than the original error. However, the refined error gets worse rapidly from a specific outlier ratio. For example, the refined error for the REN ($\alpha = 1$) gets worse rapidly from the outlier ratio 0.3125. This is because the support location recovery starts to get worse from the same point. From Fig. 6, one can still see the advantages of the refinement strategy but with less gain, and the results are not as stable as the independent case. The refined error stays to be 1 from outlier ratio 0.25, which coincides with the results in Fig. 3 that the recovered support locations are close to 0 from outlier ratio 0.25. Being 1 in refined error indicates that nearly all support locations are wrongly estimated.

## Concluding Remarks

We consider the high-dimensional sparse regression problem in this paper. By observing that the elastic net model (including Lasso and soft thresholding) fails with even a very small fraction of outliers, we proposed a robust elastic net model using the trimmed inner product as a robust counterpart to replace the original inner product. The intuition behind is that the trimmed inner product can not be significantly affected by a bounded number of arbitrarily corrupted points, i.e., arbitrary outliers. Such simple idea works extremely well in high-dimensional settings. We further give the performance guarantees (both statistical $l_{1/2}$ error bound and the optimization error bound) for the REN model. The statistical error shows the guarantees for the support recovery, while the optimization bound makes sure that we can use projected gradient descent to approximate the optima and achieve the statistical error bound. Experimental results match the theoretical performance guarantees nicely and also validate the robustness of the REN model.

# Proof of Theorem 1

## Technical Lemmas

In order to prove Theorem 1, we need two technical lemmas. The first lemma bounds the maximum of independent sub-Gaussian random variables. The second lemma is a standard concentration result for the sum of squares of independent sub-Gaussian random variables.

**Lemma 2.** *Given that $Z_1, \cdots, Z_m$ are $m$ independent sub-Gaussian random variables with parameter $\sigma$, we have $\max_{i=1,\cdots,m} \|Z_i\| \leq 4\sigma\sqrt{\log m + \log p}$ with high probability.*

*Proof.* Define $Z_m = \max_i Z_i$. According to the definition of sub-Gaussianity, we have

$$
\begin{aligned}
\mathbb{E}\{e^{\frac{tZ_m}{\sigma}}\} = \{\max_i e^{\frac{tZ_i}{\sigma}}\} \\
\leq \sum_i \mathbb{E}\{e^{\frac{tZ_i}{\sigma}}\} \\
\leq m e^{\frac{t^2}{2}} \\
= e^{\frac{t^2}{2} + \log m}.
\end{aligned}
\tag{21}
$$

Using Markov inequality, we obatin

$$
\begin{aligned}
P(Z_m \geq \sigma t) = P(e^{\frac{tZ_m}{\sigma}} \geq e^{t^2}) \\
\leq e^{-t^2} \mathbb{E}\{e^{\frac{tZ_m}{\sigma}}\} \\
\leq e^{-t^2 + t^2/2 + \log m} \\
= e^{-\frac{t^2}{2} + \log m}.
\end{aligned}
\tag{22}
$$

By symmetry, we have

$$
P(\min_i Z_i \leq -\sigma t) \leq e^{-\frac{t^2}{2} + \log m}.
\tag{23}
$$

Therefore a union bound gives

$$
\begin{aligned}
P(\max_i |Z_i| \geq \sigma t) \leq P(\max_i Z_i \geq \sigma t) + P(\min_i Z_i \leq -\sigma t) \\
\leq 2e^{-\frac{1}{2}t^2 + \log m},
\end{aligned}
\tag{24}
$$

which yields the result with $t = 4\sqrt{\log m + \log p}$. $\qquad\square$

**Lemma 3.** *Let $Y_1, \cdots, Y_n$ be $n$ i.i.d zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$. Then we have*

$$
|\sum_{i=1}^n Y_i^2 - 1| \leq c_1 \sqrt{\frac{\log p}{n}}
\tag{25}
$$

*with high probability for some absolute constant $c_1$. If $Z_1, \cdots, Z_n$ are also i.i.d. zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$, and independent of $Y_1, \cdots, Y_n$, then we have*

$$
|\sum_{i=1}^n Y_i Z_i| \leq c_2 \sqrt{\frac{\log p}{n}}
\tag{26}
$$

*with high probability for some absolute constant $c_2$.*

*Proof.* The proof follows directly from Lemma 14 in the supplementary material of [14]. $\qquad\square$

## Main Proof

Recall that $\hat{\Gamma}_{\text{REN}} = \alpha \boldsymbol{X}_{\mathcal{T}}^T \boldsymbol{X}_{\mathcal{T}} + (1 - \alpha)\boldsymbol{I}$ and $\hat{\gamma}_{\text{REN}} = \boldsymbol{X}_{\mathcal{T}}^T \boldsymbol{X}_{\mathcal{T}} \beta^*$, where $\mathcal{T}$ is the index set corresponding to the trimmed inner product. Let $\Delta = \hat{\beta} - \beta^*$, $F = \hat{\Gamma}_{\text{REN}} - \alpha \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}} - (1 - \alpha)\boldsymbol{I}$, $f = \hat{\gamma}_{\text{REN}} - \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}} \beta^*$, and $\mathcal{A}$ is the index set of authentic data, $\mathcal{O}$ is the index set of outliers, $S = \text{support}(\beta^*)$. For any vector $b \in \mathbb{R}^p$, we define $b_S$ as the vector with $(b_S)_i = b_i$ for $i \in S$ and $(b_S)_i = 0$ for $i \notin S$.

With $\hat{\beta}$ satisfying the constraint in the optimization problem in Robust Lasso, we have

$$
\begin{aligned}
\|\beta^*\|_1 &\geq \|\beta^* + \Delta\|_1 \\
&= \|\beta^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\
&\geq \|\beta^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1,
\end{aligned}
\tag{27}
$$

which follows that $\|\Delta_{S^c}\| \leq \|\Delta_S\|_1$. Given $\|S\| = k$, we have the following inequality

$$
\begin{aligned}
\|\Delta\|_1 &= \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\
&\leq 2\|\Delta_S\|_1 \\
&\leq 2\sqrt{k}\|\Delta_S\|_2 \\
&\leq 2\sqrt{k}\|\Delta\|_2
\end{aligned}
\tag{28}
$$

Under the assumption for $n$ in this theorem, Lemma 1 in [14] guarantees that the authentic $\boldsymbol{X}_{\mathcal{A}}$ satisfies the Restricted Strong Convexity (RSC) under the assumption of this theorem:

$$
\boldsymbol{u}^T (\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}) \boldsymbol{u} \geq \frac{1}{4} \lambda_{\min}(\Sigma_x) \|\boldsymbol{u}\|_2^2, \quad \forall \boldsymbol{u} : \|\boldsymbol{u}\|_1 \leq 2\sqrt{k}\|\boldsymbol{u}\|_2
\tag{29}
$$

Because $\|\Delta\|_1 \leq 2\sqrt{k}\|\Delta\|_2$, we obtain

$$
\begin{aligned}
\Delta^T \hat{\Gamma}_{\text{REN}}^T \Delta &= \alpha \Delta^T (\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}) \Delta + \Delta^T F \Delta + (1 - \alpha) \Delta^T \Delta \\
&\geq \frac{\alpha}{2} \lambda_{\min}(\Sigma_x) \|\Delta\|_2^2 - \|F\|_\infty \sum_{i,j} |\Delta_i||\Delta_j| + (1 - \alpha) \Delta^T \Delta \\
&\geq \frac{\alpha}{2} \lambda_{\min}(\Sigma_x) \|\Delta\|_2^2 - 4k\|F\|_\infty \|\Delta\|_2^2 + (1 - \alpha)\|\Delta\|_2^2
\end{aligned}
\tag{30}
$$

For $F$, we can bound it with

$$
\begin{aligned}
F_{ij} &= \alpha \left( \langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle_{n_o} - \langle \{\boldsymbol{X}_{\mathcal{A}}\}_i, \{\boldsymbol{X}_{\mathcal{A}}\}_j \rangle \right) \\
&= -\alpha \sum_{k \in \mathcal{T} \cap \mathcal{A}} \boldsymbol{X}_{ki} \boldsymbol{X}_{kj} + \alpha \sum_{k \in \mathcal{T}^c \cap \mathcal{O}} \boldsymbol{X}_{ki} \boldsymbol{X}_{kj} \\
&\leq 2n_o \alpha \left( \max_{k \in \mathcal{A}} |\boldsymbol{X}_{ki}| \right) \left( \max_{k \in \mathcal{A}} |\boldsymbol{X}_{kj}| \right)
\end{aligned}
\tag{31}
$$

Since $\boldsymbol{X}_{ki}, k \in \mathcal{A}$ are independent sub-Gaussian variable with parameter $\frac{1}{n}\sigma_x^2$, Lemma 1 concludes $\max_{k \in \mathcal{A}} |\boldsymbol{X}_{ki}| \lesssim \sigma_x \sqrt{\frac{\log p}{n}}$ with high probability. Thus it follows from a union bound over $(i, j)$ that

$$
\|F\|_\infty \lesssim \alpha \frac{n_o \log p}{n} \sigma_x^2.
\tag{32}
$$

Recall the assumption of the theorem that $\frac{n_o}{n} \lesssim \frac{\lambda_{\min}(\Sigma_x)}{\sigma_x^2 k \log p}$. Thus we have

$$
\|F\|_\infty \leq \alpha \frac{\lambda_{\min}(\Sigma_x)}{16k}
\tag{33}
$$

Combining Eq. (30) and Eq. (33), we have

$$\Delta^T \hat{\Gamma}_{\text{REN}}^T \Delta \geq \frac{\alpha}{4} \lambda_{\min}(\Sigma_x) \|\Delta\|_2^2 + (1-\alpha)\|\Delta\|_2^2 \tag{34}$$

According to Holder's inequality and $\|\Delta\|_1 \leq 2\sqrt{k}\|\Delta\|_2$, we obtain

$$\begin{aligned}
\langle \hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*, \Delta \rangle &\leq \|\hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty \|\Delta\|_1 \\
&\leq 4\sqrt{k}\|\hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty \|\Delta\|_2
\end{aligned} \tag{35}$$

in which we note that

$$\begin{aligned}
\|\hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty &\leq \|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty + \|\hat{\gamma}_{\text{REN}} - \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^*\|_\infty \\
&= \|\alpha \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* + (1-\alpha)\boldsymbol{I}\beta^* - \hat{\Gamma}_{\text{REN}}\beta^* + (1-\alpha)\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - (1-\alpha)\boldsymbol{I}\beta^*\|_\infty \\
&\quad + \|\hat{\gamma}_{\text{REN}} - \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^*\|_\infty \\
&\leq \|\alpha \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* + (1-\alpha)\boldsymbol{I}\beta^* - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty + \|(1-\alpha)\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - (1-\alpha)\boldsymbol{I}\beta^*\|_\infty \\
&\quad + \|\hat{\gamma}_{\text{REN}} - \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^*\|_\infty \\
&= \|F\beta^*\|_\infty + (1-\alpha)\|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - \boldsymbol{I}\beta^*\|_\infty + \|f\|_\infty
\end{aligned} \tag{36}$$

For the first term $\|F\beta^*\|_\infty$ in Eq. (36), we can bound it using $\|F\|_\infty \lesssim \alpha \frac{n_o \log p}{n}\sigma_x^2$ and the $k$-sparsity of $\beta^*$:

$$\|F\beta^*\|_\infty \leq \frac{\alpha\sqrt{k}n_o \log p}{n}\sigma_x^2 \|\beta^*\|_2 \tag{37}$$

For the second term $(1-\alpha)\|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - \boldsymbol{I}\beta^*\|_\infty$ in Eq. (36), we have

$$(1-\alpha)\|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - \boldsymbol{I}\beta^*\|_\infty \leq (1-\alpha)\|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}} - \boldsymbol{I}\|_\infty \cdot \sqrt{k}\|\beta^*\|_2. \tag{38}$$

Using Lemma 2 and the theorem assumption that $\max_j \{\|\Sigma_{xj}\|_0\} \leq c_3$ where $\Sigma_{xj}$ denotes the $j$th column of the covariance matrix $\Sigma_x$, we have

$$(1-\alpha)\|\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}\beta^* - \boldsymbol{I}\beta^*\|_\infty \leq (1-\alpha)c_3\sqrt{\frac{k \log p}{n}}\|\beta^*\|_2. \tag{39}$$

For the third term $\|f\|_\infty$ in Eq. (36), we decompose $f_j$ as

$$\begin{aligned}
f_j &= \langle \boldsymbol{X}_j, \boldsymbol{y} \rangle_{n_o} - \langle \boldsymbol{X}_{\mathcal{A}j}, \boldsymbol{X}_{\mathcal{A}}\beta^* \rangle \\
&= \sum_{i \in \mathcal{T}^c} \boldsymbol{X}_{ij}\boldsymbol{y}_i - \langle \boldsymbol{X}_{\mathcal{A}j}, \boldsymbol{X}_{\mathcal{A}}\beta^* \rangle \\
&= \left( \sum_{i \in \mathcal{A}} \boldsymbol{X}_{ij}\boldsymbol{y}_i - \langle \boldsymbol{X}_{\mathcal{A}j}, \boldsymbol{X}_{\mathcal{A}}\beta^* \rangle \right) - \sum_{i \in \mathcal{T} \cap \mathcal{A}} \boldsymbol{X}_{ij}\boldsymbol{y}_i + \sum_{i \in \mathcal{T}^c \cap \mathcal{O}} \boldsymbol{X}_{ij}\boldsymbol{y}_i \\
&= \langle \boldsymbol{X}_{\mathcal{A}j}, \boldsymbol{\epsilon} \rangle - \sum_{i \in \mathcal{T} \cap \mathcal{A}} \boldsymbol{X}_{ij}\boldsymbol{y}_i + \sum_{i \in \mathcal{T}^c \cap \mathcal{O}} \boldsymbol{X}_{ij}\boldsymbol{y}_i
\end{aligned} \tag{40}$$

From Lemma 2, we obtain $\langle \boldsymbol{X}_{\mathcal{A}j}, \boldsymbol{\epsilon} \rangle \approx \sqrt{\frac{\sigma_\epsilon^2 \log p}{n}}$ with high probability. Under the sub-Gaussian construction model, each $\boldsymbol{y}_i, i \in \mathcal{A}$ is sub-Gaussian with parameter $\frac{\sigma_\epsilon^2 + \sigma_x^2\|\beta^*\|_2^2}{n}$. Based on the Lemma 1, we have

$$\left| -\sum_{i \in \mathcal{T} \cap \mathcal{A}} \boldsymbol{X}_{ij}\boldsymbol{y}_i + \sum_{i \in \mathcal{T}^c \cap \mathcal{O}} \boldsymbol{X}_{ij}\boldsymbol{y}_i \right| \lesssim \frac{n_o \log p}{n}\sigma_x^2 \sqrt{\sigma_\epsilon^2 + \sigma_x^2\|\beta^*\|_2^2} \tag{41}$$

14

which follows from a union bound and gives

$$\|f\|_\infty \lesssim \sqrt{\frac{\sigma_\epsilon^2 \log p}{n}} + \frac{n_o \log p}{n} \sigma_x^2 \sqrt{\sigma_\epsilon^2 + \sigma_x^2 \|\beta^*\|_2^2} \tag{42}$$

Combining Eq. (35), Eq. (36) and the other pieces, we have

$$\begin{aligned}
\langle \hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*, \Delta \rangle \lesssim 4\sqrt{k}\|\Delta\|_2 \bigg( & \frac{\alpha \sqrt{k} n_o \log p}{n} \sigma_x^2 \|\beta\|_2 + \sqrt{\frac{\sigma_\epsilon^2 \log p}{n}} \\
& + \frac{n_o \log p}{n} \sigma_x \sqrt{\sigma_\epsilon^2 + \sigma_x^2 \|\beta^*\|_2^2} + (1-\alpha)c_3 \sqrt{k}\sqrt{\frac{\log p}{n}}\|\beta^*\|_2 \bigg)
\end{aligned} \tag{43}$$

According to the optimality of $\hat{\beta}$, we have

$$\frac{1}{2}\hat{\beta}^T \hat{\Gamma}_{\text{REN}}\hat{\beta} - \hat{\gamma}_{\text{REN}}^T \hat{\beta} \le \frac{1}{2}\hat{\beta}^{*T}\hat{\Gamma}_{\text{REN}}\hat{\beta}^* - \hat{\gamma}_{\text{REN}}^T \hat{\beta}^* \tag{44}$$

which can be rearranged as

$$\frac{1}{2}\Delta^T \hat{\Gamma}_{\text{REN}}\Delta \le \langle \hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*, \Delta \rangle \tag{45}$$

Combining all the derived inequalities, we have

$$\begin{aligned}
(\frac{\alpha}{8}\lambda_{\min}(\Sigma_x) + \frac{1-\alpha}{2})\|\Delta\|_2^2 \le \frac{1}{2}\Delta^T \hat{\Gamma}_{\text{REN}}\Delta &\le \langle \hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*, \Delta \rangle \\
&\le 4\sqrt{k}\|\hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*, \Delta\|_\infty \|\Delta\|_2
\end{aligned} \tag{46}$$

which derives

$$\begin{aligned}
\frac{1}{2\sqrt{k}}\|\Delta\|_1 \le \|\Delta\|_2 &\le \frac{4\sqrt{k}}{\frac{\alpha}{8}\lambda_{\min}(\Sigma_x) + \frac{1-\alpha}{2}} \|\hat{\gamma}_{\text{REN}} - \hat{\Gamma}_{\text{REN}}\beta^*\|_\infty \\
&\le \frac{32}{\alpha\lambda_{\min}(\Sigma_x) + 4(1-\alpha)} \bigg( \frac{\alpha k n_o \log p}{n} \sigma_x^2 \|\beta\|_2 + \sqrt{\frac{k\sigma_\epsilon^2 \log p}{n}} \\
&\quad + \frac{n_o \sqrt{k}\log p}{n}\sigma_x \sqrt{\sigma_\epsilon^2 + \sigma_x^2 \|\beta^*\|_2^2} + (1-\alpha)c_3 k\sqrt{\frac{\log p}{n}}\|\beta^*\|_2 \bigg)
\end{aligned} \tag{47}$$

which finishes the proof of the theorem.

# Proof of Corollary 1

In fact, the proof of Corollary 1 follows directly from Theorem 1. The difference is that if $\alpha = 1$ in Eq. (36), then the second term $(1-\alpha)\|X_{\mathcal{A}}^T X_{\mathcal{A}}\beta^* - I\beta^*\|_\infty$ is removed. Therefore, we do not need to use the condition $\max_j \{\|\Sigma_{xj}\|_0\} \le c_3$ to bound this term. So the conclusion of Corollary 1 is the same as Theorem 1 with $\alpha = 1$, but they differs in the required conditions.

# Proof of Corollary 2

The proof of Corollary 2 is essentially the same as Theorem 1. Corollary 2 can be directly concluded with $\alpha = 0$ and $\Sigma_x = I$.

## Proof of Lemma 1

From Eq. (29), Eq. (30) and Eq. (33) in the proof of Theorem 1 and, we can derive the following inequality:

$$
\begin{aligned}
\Delta^T \hat{\Gamma}_{\text{REN}}^T \Delta &= \alpha \Delta^T (\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}}) \Delta + \Delta^T F \Delta + (1-\alpha)\Delta^T \Delta \\
&\geq \frac{\alpha}{2}\lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - \|F\|_\infty \sum_{i,j} |\Delta_i||\Delta_j| + (1-\alpha)\Delta^T \Delta \\
&= \frac{\alpha}{2}\lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - \|F\|_\infty\|\Delta\|_1^2 + (1-\alpha)\|\Delta\|_2^2 \\
&= \left(\frac{\alpha}{2}\lambda_{\min}(\Sigma_x) + (1-\alpha)\right)\|\Delta\|_2^2 - \|F\|_\infty\|\Delta\|_1^2 \\
&\geq \left(\frac{\alpha}{2}\lambda_{\min}(\Sigma_x) + (1-\alpha)\right)\|\Delta\|_2^2 - \frac{\alpha\lambda_{\min}(\Sigma_x)}{8}\|\Delta\|_1^2
\end{aligned}
\tag{48}
$$

which satisfies the lower-RE condition.

Similarly, according to Lemma 1, Lemma 13 in the supplementary material of [14] and the restricted smoothness, we can obtain

$$
\Delta^T \hat{\Gamma}_{\text{REN}}^T \Delta \leq \left(\frac{3\alpha}{2}\lambda_{\max}(\Sigma_x) + (1-\alpha)\right)\|\Delta\|_2^2 + \frac{\alpha\lambda_{\min}(\Sigma_x)}{8}\|\Delta\|_1^2
\tag{49}
$$

which satisfies the upper-RE condition. Therefore, one possible set of values for $\mu_1$, $\mu_2$ and $\tau(n,p)$ is

$$
\mu_1 = \alpha\frac{\lambda_{\min}(\Sigma_x)}{2} + (1-\alpha),
\tag{50}
$$

$$
\mu_2 = \alpha\frac{3\lambda_{\max}(\Sigma_x)}{2} + (1-\alpha),
\tag{51}
$$

$$
\tau(n,p) = \frac{\alpha}{8}\lambda_{\min}(\Sigma),
\tag{52}
$$

Combing pieces, we finish the proof of Lemma 1 in the main paper. It can also be proved with a simple extension using the Appendix 5 in the supplementary material of [6].

## Proof of Theorem 2

Using the Lemma 1 in the main paper, we know that the $\hat{\Gamma}_{\text{REN}}$ satisfies the lower-RE and upper-RE conditions with certain parameters. We can directly follow the Theorem 2 in [14] to prove this theorem.

## References

[1] Alekh Agarwal, Sahand Negahban, Martin J Wainwright, et al. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

[2] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[3] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[4] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[5] Yudong Chen and Constantine Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *Proceedings of The 30th International Conference on Machine Learning*, pages 383–391, 2013.

[6] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 774–782, 2013.

[7] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[8] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[9] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[10] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. In *Advances in Neural Information Processing Systems*, pages 253–261, 2014.

[11] Soren Hauberg, Aasa Feragen, and Michael J Black. Grassmann averages for scalable robust pca. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3810–3817. IEEE, 2014.

[12] Jason N Laska, Mark Davenport, Richard G Baraniuk, et al. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 1556–1560. IEEE, 2009.

[13] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.

[14] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[15] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l1-minimization. *Information Theory, IEEE Transactions on*, 59(4):2017–2035, 2013.

[16] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[17] Mathieu Rosenbaum and Alexandre B Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[19] Joel Tropp et al. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

[20] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[21] Ling Wang, Hong Cheng, Zicheng Liu, and Ce Zhu. A robust elastic net approach for feature learning. *Journal of Visual Communication and Image Representation*, 25(2):313–321, 2014.

[22] Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *Information Theory, IEEE Transactions on*, 59(1):546–572, 2013.

[23] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE, 2011.

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.