Robust Lasso with missing and grossly corrupted observations

Nam H. Nguyen
Johns Hopkins University
nam@jhu.edu

Nasser M. Nasrabadi
U.S. Army Research Lab
nasser.m.nasrabadi.civ@mail.mil

Trac D. Tran
Johns Hopkins University
trac@jhu.edu

Abstract

This paper studies the problem of accurately recovering a sparse vector $\beta^\star$ from highly corrupted linear measurements $y = X\beta^\star + e^\star + w$ where $e^\star$ is a sparse error vector whose nonzero entries may be unbounded and w is a bounded noise. We propose a so-called extended Lasso optimization which takes into consideration sparse prior information of both $\beta^\star$ and $e^\star$. Our first result shows that the extended Lasso can faithfully recover both the regression and the corruption vectors. Our analysis is relied on a notion of extended restricted eigenvalue for the design matrix X. Our second set of results applies to a general class of Gaussian design matrix X with i.i.d rows $\mathcal{N}(0, \Sigma)$, for which we provide a surprising phenomenon: the extended Lasso can recover exact signed supports of both $\beta^\star$ and $e^\star$ from only $\Omega(k \log p \log n)$ observations, even the fraction of corruption is arbitrarily close to one. Our analysis also shows that this amount of observations required to achieve exact signed support is optimal.

1 Introduction

One of the central problems in statistics is the linear regression in which the goal is to accurately estimate a regression vector $\beta^\star \in \mathbb{R}^p$ from the noisy observations

$$y = X\beta^\star + w, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is the measurement or design matrix, and $w \in \mathbb{R}^n$ is the stochastic observation vector noise. A particular situation recently attracted much attention from research community concerns with the model in which the number of regression variables p is larger than the number of observations n ($p \geq n$). In such circumstances, without imposing some additional assumptions for this model, it is well known that the problem is ill-posed, and thus the linear regression is not consistent. Accordingly, there have been various lines of work on high dimensional inference based on imposing different types of structure constraints such as sparsity and group sparsity [15] [5] [21]. Among them, the most popular model focused on sparsity assumption of the regression vector. To estimate β, a standard method, namely Lasso [15], was proposed to use l_1-penalty as a surrogate function to enforce sparsity constraint.

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 , \tag{2}$$

where λ is the positive regularization parameter and l_1-norm $\|\beta\|_1$ is defined by $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$.

During the past few years, there has been numerous studies to understand the ℓ_1-regularization for sparse regression models [23] [11] [10] [17] [4] [2] [22]. These works are mainly characterized by

the type of the loss functions considered. For instance, some authors [4] seek to obtain a regression estimate $\widehat{\beta}$ that delivers small prediction error while other authors [2] [11] [22] seek to produce a regressor with minimal parameter estimation error, which is measured by the ℓ_2-norm of $(\widehat{\beta} - \beta^\star)$. Another line of work [23] [17] considers the variable selection in which the goal is to obtain an estimate that correctly identifies the support of the true regression vector. To achieve low prediction or parameter estimation loss, it is now well known that it is both sufficient and necessary to impose certain lower bounds on the smallest singular values of the design matrix [10] [2], while a notion of small mutual incoherence for the design matrix [4] [23] [17] is required to achieve accurate variable selection.

We notice that all the previous work relies on the assumption that the observation noise has bounded energy. Without this assumption, it is very likely that the estimated regressor is either not reliable or unable to identify the correct support. With this observation in mind, in this paper, we extend the linear model (1) by considering the noise with unbounded energy. It is clear that if all the entries of y is corrupted by large error, then it is impossible to faithfully recover the regression vector $\beta^\star$. However, in many practical applications such as face and acoustic recognition, only a portion of the observation vector is contaminated by gross error. Formally, we have the mathematical model

$$y = X\beta^\star + e^\star + w, \tag{3}$$

where $e^\star \in \mathbb{R}^n$ is the sparse error whose locations of nonzero entries are unknown and magnitudes can be arbitrarily large and w is another noise vector with bounded entries. In this paper, we assume that w has a multivariate Gaussian $\mathcal{N}(0, \sigma^2 I_{n\times n})$ distribution. This model also includes as a particular case the missing data problem in which all the entries of y is not fully observed, but some are missing. This problem is particularly important in computer vision and biology applications. If some entries of y are missing, the nonzero entries of $e^\star$ whose locations are associated with the missing entries of the observation vector y have the same values as entries of y but with inverse signs.

The problems of recovering the data under gross error has gained increasing attentions recently with many interesting practical applications [18] [6] [7] as well as theoretical consideration [9] [13] [8]. Another recent line of research on recovering the data from grossly corrupted measurements has been also studied in the context of robust principal component analysis (RPCA) [3] [20] [1]. Let us consider some examples to illustrate:

- *Face recognition.* The model (3) has been originally proposed by Wright et al. [19] in the context of face recognition. In this problem, a face test sample y is assumed to be represented as a linear combination of training faces in the dictionary X, $y = X\beta$ where β is the coefficient vector used for classification. However, it is often the case that the face is occluded by unwanted objects such as glasses, hats etc. These occlusions, which occupy a portion of the test face, can be considered as the sparse error $e^\star$ in the model (3).
- *Subspace clustering.* One of the important problem on high dimensional analysis is to cluster the data points into multiple subspaces. A recent work of Elhamifar and Vidal [6] showed that this problem can be solved by expressing each data point as a sparse linear combination of all other data points. Coefficient vectors recovered from solving the Lasso problems are then employed for clustering. If the data points are represented as a matrix X, then we wish to find a sparse coefficient matrix B such that $X = XB$ and $\mathrm{diag}(B) = 0$. When the data is missing or contaminated with outliers, [6] formulates the problem as $X = XB + E$ and minimize a sum of two ℓ_1-norms with respect to both B and E.
- *Sensor network.* In this model, sensors collect measurements of a signal $\beta^\star$ independently by simply projecting $\beta^\star$ onto row vectors of a sensing matrix X, $y_i = \langle X_i, \beta^\star \rangle$. The measurements y_i are then sent to the center hub for analysis. However, it is highly likely that some sensors might fail to send the measurements correctly and sometimes report totally irrelevant measurements. Therefore, it is more accurate to employ the observation model (3) than model (1).

It is worth noticing that in the aforementioned applications, $e^\star$ plays the role as the sparse (undesired) error. However, in many other applications, $e^\star$ can contain meaningful information, and thus necessary to be recovered. An example of this kind is signal separation, in which $\beta^\star$ and $e^\star$ are two distinct signal components (video or audio). Furthermore, in applications such as classification and

clustering, the assumption that the test sample y is a linear combination of a few training samples in the dictionary (design matrix) X might be violated. This sparse component $e^\star$ can thus be seen as the compensation for linear regression model mismatch.

Given the observation model (1) and the sparsity assumptions on both regression vector $\beta^\star$ and error $e^\star$, we propose the following convex minimization to estimate the unknown parameter $\beta^\star$ as well as the error $e^\star$.

$$\min_{\beta,e} \frac{1}{2} \|y - X\beta - e\|_2^2 + \lambda_\beta \|\beta\|_1 + \lambda_e \|e\|_1 , \tag{4}$$

where λ_β and λ_e are positive regularization parameters. This optimization, we call *extended Lasso*, can be seen as a generalization of the Lasso program. Indeed, by setting $\lambda_e = 0$, (4) returns to the standard Lasso. The additional regularization associated with e encourages sparsity on the error where parameter λ_e controls the sparsity level. In this paper, we focus on the following questions: what are necessary and sufficient conditions for the ambient dimension p, the number of observations n, the sparsity index k of the regression $\beta^\star$ and the fraction of corruption so that (i) the extended Lasso is able (or unable) to recover the exact support sets of both $\beta^\star$ and $e^\star$? (ii) the extended Lasso is able to recover $\beta^\star$ and $e^\star$ with small prediction error and parameter error? We are particularly interested in understanding the asymptotic situation where the the fraction of error is arbitrarily close to 100%.

Previous work. The problem of recovering the estimation vector $\beta^\star$ and error $e^\star$ has originally proposed and analyzed by Wright and Ma [18]. In the absence of the stochastic noise w in the observation model (3), the authors proposed to estimate $(\beta^\star, e^\star)$ by solving the linear program

$$\min_{\beta,e} \|\beta\|_1 + \|e\|_1 \quad \text{s.t.} \quad y = X\beta + e. \tag{5}$$

The result of [18] is asymptotic in nature. They showed that for a class of Gaussian design matrix with i.i.d entries, the optimization (5) can recover $(\beta^\star, e^\star)$ precisely with high probability even when the fraction of corruption is arbitrarily close to one. However, the result holds under rather stringent conditions. In particularly, they require the number of observations n grow proportionally with the ambient dimension p, and the sparsity index k is a very small portion of n. These conditions is of course far from the optimal bound in compressed sensing (CS) and statistics literature (recall $k \leq O(n/\log p)$ is sufficient in conventional analysis [17]).

Another line of work has also focused on the optimization (5). In both papers of Laska et al. [7] and Li et al. [9], the authors establish that for Gaussian design matrix X, if $n \geq C(k+s)\log p$ where s is the sparsity level of $e^\star$, then the recovery is exact. This follows from the fact that the combination matrix $[X,\ I]$ obeys the restricted isometry property, a well-known property used to guarantee exact recovery of sparse vectors via ℓ_1-minimization. These results, however, do not allow the fraction of corruption close to one.

Among the previous work, the most closely related to the current paper are recent results by Li [8] and Nguyen et al. [13] in which a positive regularization parameter λ is employed to control the sparsity of $e^\star$. Using different methods, both sets of authors show that as λ is deterministically selected to be $1/\sqrt{\log p}$ and X is a sub-orthogonal matrix, then the solution of following optimization is exact even a constant fraction of observation is corrupted. Moreover, [8] establishes a similar result with Gaussian design matrix in which the number of observations is only an order of $k \log p$ - an amount that is known to be optimal in CS and statistics.

$$\min_{\beta,e} \|\beta\|_1 + \lambda \|e\|_1 \quad \text{s.t.} \quad y = X\beta + e. \tag{6}$$

Our contribution. This paper considers a general setting in which the observations are contaminated by both sparse and dense errors. We allow the corruptions to linearly grow with the number of observations and have arbitrarily large magnitudes. We establish a general scaling of the quadruplet (n, p, k, s) such that the extended Lasso stably recovers both the regression and corruption vectors. Of particular interest to us are the following equations:

(a) First, under what scalings of (n, p, k, s) does the extended Lasso obtain the unique solution with small estimation error.

(b) Second, under what scalings of (n, p, k) does the extended Lasso obtain the exact signed support recovery even almost all the observations are corrupted?

(c) Third, under what scalings of (n, p, k, s) does no solution of the extended Lasso specify the correct signed support?

To answer for the first question, we introduce a notion of *extended restricted eigenvalue* for a matrix $[X,\ I]$ where I is an identity matrix. We show that this property satisfies for a general class of random Gaussian design matrix. The answers to the last two questions requires stricter conditions for the design matrix. In particular, for random Gaussian design matrix with i.i.d rows $\mathcal{N}(0, \Sigma)$, we rely on two standard assumptions: invertibility and mutual incoherence.

If we denote $Z = [X,\ I]$ where I is an identity matrix and $\overline{\beta} = [\beta^{\star^T},\ e^{\star^T}]^T$, then the observation vector y is reformulated as $y = Z\overline{\beta} + w$, which is the same as standard Lasso model. However, previous results [2] [17] applying to random Gaussian design matrix are irrelevant to this setting since the Z no longer behave like a Gaussian matrix. To establish theoretical analysis, we need more study on the interaction between the Gaussian and identity matrices. By exploiting the fact that the matrix Z consists of two component where one component has special structure, our analysis reveals an interesting phenomenon: extended Lasso can accurately recover both the regressor $\beta^\star$ and corruption $e^\star$ even when the fraction of corruption is up to 100%. We measure the recoverability of these variables under two criterions: parameter accuracy and feature selection accuracy. Moreover, our analysis can be extended to the situation in which the identity matrix can be replaced by a tight frame D as well as extended to other models such as group Lasso or matrix Lasso with sparse error.

Notation We summarize here some standard notation used throughout the paper. We reserve T and S as the sparse support of $\beta^\star$ and $e^\star$, respectively. Given and design matrix $X \in \mathbb{R}^{n\times p}$ and subsets S and T, we use X_{ST} to denote the $|S| \times |T|$ submatrix obtained by extracting those rows indexed by S and columns indexed by T. We use the notation C_1, C_2, c_1, c_2, etc., to refer to positive constants, whose value may change from line to line. Given two functions f and g, the notation $f(n) = \mathcal{O}(g(n))$ means that there exists a constant $c < +\infty$ such that $f(n) \leq cg(n)$; the notation $f(n) = \Omega(g(n))$ means that $f(n) \geq cg(n)$ and the notation $f(n) = \Theta(g(n))$ means that $f(n) = (g(n))$ and $f(n) = \Omega(g(n))$. The symbol $f(n) = o(g(n))$ means that $f(n)/g(n) \to 0$.

2 Main results

In this section, we provide precise statements of the main results of this paper. In the first sub-section, we establish the parameter estimation and provide a deterministic result which bases on the notion of extended restricted eigenvalue. We further show that the random Gaussian design matrix satisfies this property with high probability. The next sub-section considers the feature estimation. We establish conditions for the design matrix such that the solution of the extended Lasso has the exact signed supports.

2.1 Parameter estimation

As in conventional Lasso, to obtain a low parameter estimation bound, it is necessary to impose conditions on the design matrix X. In this paper, we introduce a notion of *extended restricted eigenvalue* (extended RE) condition. Let $\mathbb{C}$ be a restricted set, we say that the matrix X satisfies the extended RE assumption over the set $\mathbb{C}$ if there exists some $\kappa_l > 0$ such that

$$\|Xh + f\|_2 \geq \kappa_l(\|h\|_2 + \|f\|_2) \qquad \text{for all } (h, f) \in \mathbb{C}, \tag{7}$$

where the restricted set $\mathbb{C}$ of interest is defined with $\lambda_n := \lambda_e/\lambda_\beta$ as follow

$$\mathbb{C} := \{(h, f) \in \mathbb{R}^p \times \mathbb{R}^n \mid \|h_{T^c}\|_1 + \lambda_n \|f_{S^c}\|_1 \leq 3\|h_T\|_1 + 3\lambda_n \|f_S\|_1\}. \tag{8}$$

This assumption is a natural extension of the restricted eigenvalue condition and restricted strong convexity considered in [2] [14] and [12]. In the absent of a vector f in the equation (7) and in the set $\mathbb{C}$, this condition returns to the restricted eigenvalue defined in [2]. As explained at more length in [2] and [16], restricted eigenvalue is among the weakest assumption on the design matrix such that the solution of the Lasso is consistent.

With this assumption at hand, we now state the first theorem

Theorem 1. *Consider the optimal solution $(\widehat{\beta}, \widehat{e})$ to the optimization problem (4) with regularization parameters chosen as*

$$\lambda_\beta \geq \frac{2}{\gamma} \left\| X^* w \right\|_\infty \quad \textit{and} \quad \lambda_n := \frac{\lambda_e}{\lambda_\beta} = \gamma \frac{\left\| w \right\|_\infty}{\left\| X^* w \right\|_\infty}, \tag{9}$$

where $\gamma \in (0, 1]$. Assuming that the design matrix X obeys the extended RE, then the error set $(h, f) = (\widehat{\beta} - \beta^\star, \widehat{e} - e^\star)$ is bounded by

$$\left\| h \right\|_2 + \left\| f \right\|_2 \leq 3\kappa_l^{-2} \left(\lambda_\beta \sqrt{k} + \lambda_e \sqrt{s} \right). \tag{10}$$

There are several interesting observations from this theorem

1) The error bound naturally split into two components related to the sparsity indices of $\beta^\star$ and $e^\star$. In addition, the error bound contains three quantity: the sparsity indices, regularization parameters and the extended RE constant. If the terms related to the corruption $e^\star$ are omitted, then we obtain similar parameter estimation bound as the standard Lasso [2] [12].

2) The choice of regularization parameters λ_β and λ_e can make explicitly: assuming w is a Gaussian random vector whose entries are $\mathcal{N}(0, \sigma^2)$ and the design matrix has unit-normed columns, it is clear that with high probability, $\left\| X^* w \right\|_\infty \leq 2\sqrt{\sigma^2 \log p}$ and $\left\| w \right\|_\infty \leq 2\sqrt{\sigma^2 \log n}$. Thus, it is sufficient to select $\lambda_\beta \geq \frac{4}{\gamma}\sqrt{\sigma^2 \log p}$ and $\lambda_e \geq 4\sqrt{\sigma^2 \log n}$.

3) At the first glance, the parameter γ does not seem to have any meaningful interpretation and the $\gamma = 1$ seems to be the best selection due to the smallest estimation error it can produce. However, this parameter actually control the sparsity level of the regression vector with respect to the fraction of corruption. This relation is made via the restricted set $\mathbb{C}$.

In the following lemma, we show that the extended RE condition actually exists for a large class of random Gaussian design matrix whose rows are i.i.d zero mean with covariance Σ. Before stating the lemma, let us define some quantities operating on the covariance matrix Σ: $C_{\min} := \lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ, $C_{\max} := \lambda_{\max}(\Sigma)$ is the biggest eigenvalue of Σ and $\xi(\Sigma) := \max_i \Sigma_{ii}$ is the maximal entry on the diagonal of the matrix Σ.

Lemma 1. *Consider the random Gaussian design matrix whose rows are i.i.d $\mathcal{N}(0, \Sigma)$ and assume $n^2 C_{\max} \xi(\Sigma) = \Theta(1)$. Select*

$$\lambda_n := \frac{\gamma}{\sqrt{\xi(\Sigma) n}} \sqrt{\frac{\log n}{\log p}}, \tag{11}$$

then with probability greater than $1 - c_1 \exp(-c_2 n)$, the matrix X satisfies the extended RE with parameter $\kappa_l = \frac{1}{4\sqrt{2}}$, provided that $n \geq C \frac{\xi(\Sigma)}{C_{\min}} k \log p$ and $s \leq \min\left\{ C_1 \frac{n}{\gamma^2 \log n}, C_2 n \right\}$ for some small constants C_1, C_2.

We would like to make some remarks:

1) The choice of parameter λ_n is nothing special here. When design matrix is Gaussian and independent with the Gaussian stochastic noise w, we can easily show that $\left\| X^* w \right\|_\infty \leq 2\sqrt{\xi(\Sigma) n \delta^2 \log p}$ with probability at least $1 - 2\exp(-\log p)$. Therefore, the selection of λ_n follows from Theorem 1.

2) The proof of this lemma, shown in the Appendix, boils down to control two terms

- *Restricted eigenvalue with X.*

$$\left\| Xh \right\|_2^2 + \left\| f \right\|_2^2 \geq \kappa_r (\left\| h \right\|_2^2 + \left\| f \right\|_2^2) \qquad \text{for all} \quad (h, f) \in \mathbb{C}.$$

- *Mutual incoherence.* Column space of the matrix X is incoherent with the column space of the identity matrix. That is, there exists some $\kappa_m > 0$ such that

$$| \langle Xh, f \rangle | \leq \kappa_m (\left\| h \right\|_2 + \left\| f \right\|_2)^2 \qquad \text{for all} \quad (h, f) \in \mathbb{C}.$$

If the incoherence between these two column spaces is sufficiently small such that $4\kappa_m < \kappa_r$, then we can conclude that $\left\| Xh + f \right\|_2^2 \geq (\kappa_r - 2\kappa_m)(\left\| h \right\|_2 + \left\| f \right\|_2)^2$. The small mutual incoherence

property is especially important since it provides how the regression separates away from the sparse error.

3) To simplify our result, we consider a special case of the uniform Gaussian design, in which $\Sigma = \frac{1}{n} I_{p\times p}$. In this situation, $C_{\min} = C_{\max} = \xi(\Sigma) = 1/n$. We have the following result which is a corollary of Theorem 1 and Lemma 1

Corollary 1 (Standard Gaussian design)**.** *Let X be a standard Gaussian design matrix. Consider the optimal solution $(\widehat{\beta}, \widehat{e})$ to the optimization problem (4) with regularization parameters chosen as*

$$\lambda_\beta \geq \frac{4}{\gamma}\sqrt{\sigma^2 \log p} \quad \textit{and} \quad \lambda_e \geq 4\sqrt{\sigma^2 \log n}, \tag{12}$$

where $\gamma \in (0, 1]$. Also assuming that $n \geq Ck \log p$ and $s \leq \min\{C_1 \frac{n}{\gamma^2 \log n}, C_2 n\}$ for some small constants C_1, C_2. Then with probability greater than $1 - c_1 \exp(-c_2 n)$, the error set $(h, f) = (\widehat{\beta} - \beta^\star, \widehat{e} - e^\star)$ is bounded by

$$\|h\|_2 + \|f\|_2 \leq 384 \left(\frac{1}{\gamma}\sqrt{\sigma^2 k \log p} + \sqrt{\sigma^2 s \log n} \right), \tag{13}$$

Corollary 1 reveals an interesting phenomenon: by setting $\gamma = 1/\sqrt{\log n}$, even when the fraction of corruption is linearly proportional with the number of samples n, the extended Lasso (4) is still capable to recover both coefficient vector $\beta^\star$ and corruption (missing) vector $e^\star$ within a bounded error (13). Without the dense noise w in the observation model (3) ($\sigma = 0$), the extended Lasso recovers the exact solution. This result is impossible to achieve with standard Lasso. Furthermore, if we know in prior that the number of corrupted observations is an order of $\mathcal{O}(n/\log p)$, then selecting $\gamma = 1$ instead of $1/\log n$ will minimize the estimation error (see equation (13)) of Theorem 1.

2.2 Feature selection with random Gaussian design

In many applications, the feature selection criteria is more preferred [17] [23]. Feature selection refers to the property that the recovered parameter has the same signed support as the true regressor. In general, good feature selection implies good parameter estimation but the reverse direction does not usually hold. In this part, we investigate conditions for the design matrix and the scaling of (n, p, k, s) such as both regression and sparse error vectors obtain this criteria.

Consider the linear model (3) where X is the Gaussian random design matrix whose rows are i.i.d zero mean with covariance matrix Σ. It has been well known in the Lasso that in order to obtain feature selection accuracy, the covariance matrix Σ must obey two properties: invertibility and small mutual coherence restricted on the set T. The first property guarantees that (4) is strictly convex, leading to the unique solution of the convex program, while the second property requires the separation between two components of Σ, one related to the set T and the other to the set T^c must be sufficiently small.

1. **Invertibility.** To guarantee uniqueness, we require Σ_{TT} to be invertible. Particularly, let $C_{\min} = \lambda_{\min}(\Sigma_{TT})$, we require $C_{\min} > 0$.
2. **Mutual incoherence.** For some $\gamma \in (0, 1)$,

$$\left\|\Sigma_{T^cT}^* (\Sigma_{TT})^{-1}\right\|_\infty \leq \frac{1}{2}(1 - \gamma) \tag{14}$$

   where $\|\cdot\|_\infty$ refers to ℓ_∞/ℓ_∞ operator norm. It is worth noting that in the standard Lasso the factor $\frac{1}{2}$ is omitted. Our condition is tighter than condition used to establish feature estimation in the Lasso by a constant factor. In fact, the quantity $1/2$ is nothing special here and we can set any value close to one with a compensation that the number of samples n will increase. Thus, we put $1/2$ for the simplicity of the proof.

Toward the end, we will also elaborate three other quantities operating on the restricted covariance matrix Σ_{TT}: $C_{\max}$, which is defined as the maximum eigenvalue of Σ_{TT}: $C_{\max} := \lambda_{\max}(\Sigma_{TT})$; $D_{\max}^-$ and $D_{\max}^+$, which are denoted as ℓ_∞-norm of matrices Σ_{TT}^{-1} and Σ_{TT} : $D_{\max}^- := \left\|(\Sigma_{TT})^{-1}\right\|_\infty$ and $D_{\max}^+ := \|\Sigma_{TT}\|_\infty$.

Our result also involves in two other quantities operating on the conditional covariance matrix of $(X_{T^c}|X_T)$ defined as

$$\Sigma_{T^c|T} := \Sigma_{T^cT^c} - \Sigma_{T^cT}\Sigma_{TT}^{-1}\Sigma_{TT^c}. \tag{15}$$

We then define $\rho_u(\Sigma_{T^c|T}) = \max_i(\Sigma_{T^c|T})_{ii}$ and $\rho_l(\Sigma_{T^c|T}) = \frac{1}{2}\min_{i\neq j}[(\Sigma_{T^c|T})_{ii} + (\Sigma_{T^c|T})_{jj} - 2(\Sigma_{T^c|T})_{ij}]$. Toward the end, we denote a shorthand ρ_u and ρ_l.

We establish the following result for Gaussian random design whose covariance matrix Σ obeys the two assumptions.

Theorem 2. *(Achievability) Given the linear model (3) with random Gaussian design and the covariance matrix Σ satisfy invertibility and incoherence properties for any $\gamma \in (0, 1)$, suppose we solve the extended Lasso (4) with regularization parameters obeying*

$$\lambda_\beta = \frac{4}{\gamma}\sqrt{\max\{\rho_u, D^+_{\max}\}n\sigma^2\log p} \qquad \textit{and} \qquad \lambda_e = 8\sqrt{\sigma^2\log n}. \tag{16}$$

Also, let $\eta = \frac{1}{32\gamma^2\log n}$, the sequence (n, p, k, s) and regularization parameters λ_β, λ_e satisfying $s \leq \eta n$

$$n \geq \max\left\{C_1\frac{1}{(1-\eta)}\frac{\rho_u}{C_{\min}}k\log(p-k), C_2\frac{\eta}{(1-\eta)^2}\frac{\max\{\rho_u, D^+_{\max}\}}{C_{\min}}k\log(p-k)\log n\right\}, \tag{17}$$

where C_1 and C_2 are numerical constants. In addition, suppose that $\min_{i\in T}|\beta^\star_i| > f_\beta(\lambda_\beta)$ and $\min_{i\in S}|e^\star_i| > f_e(\lambda_\beta, \lambda_e)$ where

$$f_\beta := c_1\frac{\lambda_\beta}{n-s}\sqrt{\frac{k\log(p-k)}{n}}\left\|\Sigma_{TT}^{-1/2}\right\|^2_\infty + 20\sqrt{\frac{\sigma^2\log k}{C_{\min}(n-s)}} \qquad \textit{and} \tag{18}$$

$$f_e := c_2(C_{\max}(k\sqrt{s} + s\sqrt{k}))^{1/2}\frac{\lambda_\beta}{n-s}\sqrt{\frac{k\log(p-k)}{n}}\left\|\Sigma_{TT}^{-1/2}\right\|^2_\infty + c_3\lambda_e. \tag{19}$$

Then the following properties holds with probability greater than $1 - c\exp(-c'\max\{\log n, \log pk\})$

1. *The solution pair $(\widehat{\beta}, \widehat{e})$ of the extended Lasso (4) is unique and has exact signed support.*
2. *ℓ_∞-norm bounds: $\left\|\widehat{\beta} - \beta^\star\right\|_\infty \leq f_\beta(\lambda_\beta)$ and $\|\widehat{e} - e^\star\|_\infty \leq f_e(\lambda_\beta)$.*

There are several interesting observations from the theorem

1) The first and important observation is that extended Lasso is robust to arbitrarily large and sparse error observation. In that sense, the extended Lasso can be viewed as a generalization of the Lasso. Under the same invertibility and mutual incoherence assumptions on the covariance matrix Σ as the standard Lasso, the extended Lasso program can recover both the regression vector and error with exact signed supports even when almost all the observations are contaminated by arbitrarily large error with unknown support. What we sacrifice for the corruption robustness is an additional $\log$ factor to the number of samples. We notice that when the error fraction is $\mathcal{O}(n/\log n)$, only $\mathcal{O}(k\log(p-k))$ samples are sufficient to recover the exact signed supports of both regression and sparse error vectors.

2) We consider the special case with Gaussian random design in which the covariance matrix $\Sigma = \frac{1}{n}I_{p\times p}$. In this case, entries of X is i.i.d $\mathcal{N}(0, 1/n)$ and we have quantities $C_{\min} = C_{\max} = D^+_{\max} = D^-_{\max} = \rho_u = \rho_l = 1$. In addition, the invertibility and mutual incoherence properties are automatically satisfied. The theorem implies that when the number of errors s is close to n, the number of samples n needed to recover exact signed supports satisfies $\frac{n}{\log n} = \Omega(k\log(p-k))$. Furthermore, Theorem 2 guarantees consistency in element-wise ℓ_∞-norm of the estimated regression at the rate $\left\|\widehat{\beta} - \beta^\star\right\|_\infty = \mathcal{O}\left(\sqrt{\sigma^2\log p}\sqrt{\frac{k\log(p-k)}{\gamma^2 n}}\right)$.

As γ is chosen to be $1/\sqrt{32\log n}$ (equivalent to establish s close to n), the ℓ_∞ error rate is an order of $\mathcal{O}(\sigma\sqrt{\log p})$, which is known to be the same as that of the standard Lasso.

3) Corollary 1, though interesting, is not able to guarantee stable recovery when the fraction of corruption converges to one. We show in Theorem 2 that this fraction can come arbitrarily close to one by sacrificing a factor of $\log n$ for the number of samples. Theorem 2 also implies that there is a significant difference between recovery to obtain small parameter estimation error versus recovery to obtain correct variable selection. When the amount of corrupted observations is linearly proportional with n, recovering the exact signed supports require an increase from $\Omega(k \log p)$ (in Corollary 1) to $\Omega(k \log p \log n)$ samples (in Theorem 2). This behavior is captured similarly by the standard Lasso, as pointed out in [17], Corollary 2.

Our next theorem show that the number of samples needed to recover accurate signed support is optimal. That is, whenever the rescaled sample size satisfies (20), then for whatever regularization parameters λ_β and λ_e are selected, no solution of the extended Lasso correctly identifies the signed supports with high probability.

Theorem 3. *(Inachievability) Given the linear model (3) with random Gaussian design and the covariance matrix Σ satisfy invertibility and incoherence properties for any $\gamma \in (0, 1)$. Let $\eta = \frac{1}{32\gamma^2 \log(n-s)}$ and the sequence (n, p, k, s) satisfies $s \geq \eta n$ and*

$$n \leq \min \left\{ C_3 \frac{1}{(1-\eta)} \frac{\rho_u}{C_{\min}} k \log(p-k), C_4 \frac{\eta}{(1-\eta)^2} \frac{\min\{\rho_l, D^+_{\max}\}}{C_{\max}} k \log(p-k) \log(1-\eta) n \left(1 + \frac{\sqrt{\sigma^2 \log n}}{\lambda_e}\right)^{-1} \right\}, \tag{20}$$

where C_3 and C_4 are some small universal constants. Then with probability tending to one, no solution pair of the extended Lasso (5) has the correct signed support.

3 Illustrative simulations

In this section, we provide some simulations to illustrate the possibility of the extended Lasso in recovering the exact regression signed support when a significant fraction of observations is corrupted by large error. Simulations are performed for a range of parameters (n, p, k, s) where the design matrix X is uniform Gaussian random whose rows are i.i.d $\mathcal{N}(0, I_{p\times p})$. For each fixed set of (n, p, k, s), we generate sparse vectors $\beta^\star$ and $e^\star$ where locations of nonzero entries are uniformly random and magnitudes are Gaussian distributed.

In our experiments, we consider varying problem sizes $p = \{128, 256, 512\}$ and three types of regression sparsity indices: sublinear sparsity ($k = 0.2p/\log(0.2p)$), linear sparsity ($k = 0.1p$) and fractional power sparsity ($k = 0.5p^{0.75}$). In all cases, we fixed the error support size $s = n/2$. This means half of the observations is corrupted. By this selection, Theorem 2 suggests that number of samples $n \geq 2Ck \log(p-k) \log n$ to guarantee exact signed support recovery. We choose $\frac{n}{\log n} = 4\theta k \log(p-k)$ where parameter θ is the rescaled sample size. This parameter control the success/failure of the extended Lasso.

In the algorithm, we select $\lambda_\beta = 2\sqrt{\sigma^2 \log p \log n}$ and $\lambda_e = 2\sqrt{\sigma^2 \log n}$ as suggested by Theorem 2, where the noise level $\sigma = 0.1$ is fixed. The algorithm reports a success if the solution pair has the same signed support as $(\beta^\star, e^\star)$. In Fig. 1, each point on the curve represents the average of 100 trials.

As demonstrated by simulations, our extended Lasso is cable to recover the exact signed support of both $\beta^\star$ and $e^\star$ even 50% of the observations are contaminated. Furthermore, up to unknown constants, our theorem 2 and 3 match with simulation results. As the sample size $\frac{n}{\log n} \leq 2k \log(p-k)$, the probability of success starts going to zero, implying the failure of the extended Lasso.

Acknowledgments

We acknowledge support from the Army Research Office (ARO) under Grant 60291-MA and National Science Foundation (NSF) under Grant CCF-1117545.

References

[1] A. Agarwal, S. Negahban, and M. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Proc. 28th Inter. Conf. Mach. Learn. (ICML-11)*, pages 1129–1136, 2011.

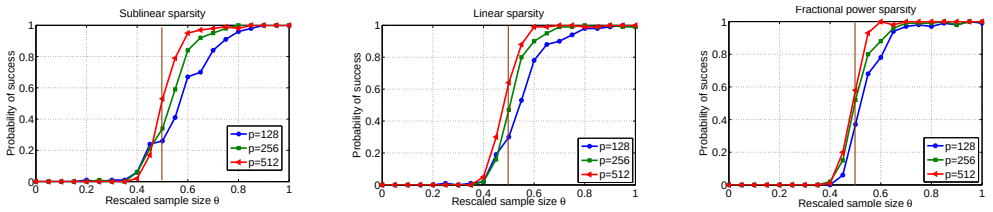

Figure 1: Probability of success in recovering the signed supports

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of statistics*, 37(4):1705–1732, 2009.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Submitted for publication, 2009.

[4] E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *Annals of Statistics*, 37:2145–2177, 2009.

[5] E. J. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of statistics*, 35(6):2313–2351, 2007.

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2797, 2009.

[7] J. N. Laska, M. A. Davenport, and R. G. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *Asilomar conference on Signals, Systems and Computers*, pages 1556–1560, 2009.

[8] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. Preprint, 2011.

[9] Z. Li, F. Wu, and J. Wright. On the systematic measurement matrix for compressed sensing in the presence of gross error. In *Data compression conference (DCC)*, pages 356–365, 2010.

[10] N. Meinshausen and P. Buehlmann. High dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2008.

[11] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of statistics*, 37(1):2246–2270, 2009.

[12] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Preprint, 2010.

[13] N. H. Nguyen and Trac. D. Tran. Exact recoverability from dense corrupted observations via l_1 minimization. preprint, 2010.

[14] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[16] S. A. van de Geer and P. Buehlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3(1360-1392), 2009.

[17] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1 -constrained quadratic programming (lasso). *IEEE Trans. Information Theory*, 55(5):2183–2202, 2009.

[18] J. Wright and Y. Ma. Dense error correction via l1 minimization. *IEEE Transaction on Information Theory*, 56(7):3540–3560, 2010.

[19] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[20] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *Ad. Neural Infor. Proc. Sys. (NIPS)*, pages 2496–2504, 2010.

[21] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

[22] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of statistics*, 37(5):2109–2144, 2009.

[23] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.