Robust Lasso With Missing and Grossly Corrupted Observations

Nam H. Nguyen and Trac D. Tran, *Senior Member, IEEE*

***Abstract*—This paper studies the problem of accurately recovering a k-sparse vector $\beta^\star \in \mathbb{R}^p$ from highly corrupted linear measurements $y = X\beta^\star + e^\star + w$, where $e^\star \in \mathbb{R}^n$ is a sparse error vector whose nonzero entries may be unbounded and w is a stochastic noise term. 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 as well as the corruption vector. Our analysis relies on the 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 can establish a surprising result: 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 when a linear fraction of observations is grossly corrupted. Our analysis also shows that this amount of observations required to achieve exact signed support is indeed optimal.**

***Index Terms*—Compressed sensing, error correction, high-dimensional inference, ℓ_1-minimization, robust recovery, sparse linear regression.**

I. Introduction

ONE of the central problems in statistics is the problem of linear regression in which the goal is to accurately estimate the 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 noise vector. A particular situation recently attracted much attention from the 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 obvious 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 (see, e.g., [1]–[15]). Among them, the most popular model focused on sparsity assumption of the regression vector. To estimate β, a standard method, namely Lasso [1], was proposed to use ℓ_1-penalty as a surrogate function to enforce sparsity constraint.

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

where λ is the positive regularization parameter, the ℓ_1-norm of the regression vector is $\|\beta\|_1$, defined as $\|\beta\|_1 \triangleq \sum_{i=1}^p |\beta_i|$, and the ℓ_2-norm in the first summand is conventionally defined as $\|a\|_2 \triangleq (\sum_{i=1}^n a_i^2)^{1/2}$ for any $a \in R^n$.

Within the past few years, there has been numerous studies to understand the ℓ_1-regularization aspect of sparse regression models (see, e.g., [5]–[11]). These works are mainly characterized by the type of the loss functions considered. For instance, authors of [9] and [11] seek to obtain a regression estimate $\widehat{\beta}$ that delivers small prediction error, while others [6], [10], [11] 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 (see, e.g., [5], [8], and [16]) 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 (see, e.g., [7] and [10]), while the notion of small mutual coherence for the design matrix (see, e.g., [5], [8], and [9]) is required to achieve accurate variable selection.

We notice that all 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 we fail 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 entries of y are corrupted by large errors, then it is impossible to faithfully recover the regression vector $\beta^\star$. However, in many practical applications such as face recognition, acoustic recognition, and dense sensor network, 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 whose magnitudes can be arbitrarily large, whereas w is the conventional noise vector with bounded energy. 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

Manuscript received December 22, 2011; revised August 10, 2012; accepted September 25, 2012. Date of publication December 11, 2012; date of current version March 13, 2013. This work was supported in part by the National Science Foundation under Grants CCF-1117545 and CCF-0728893; in part by the Army Research Office under Grants 58110-MA-II and 60219-MA; and in part by the Office of Naval Research under Grant N102-183-0208. This paper was presented in part at the Annual Conference on Neural Information Processing Systems, Granada, Spain, December 2011.

The authors are with the Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: nam@jhu.edu; trac@jhu.edu).

Communicated by E. Serpedin, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2232347