High Dimensional Structured Estimation with Noisy Designs

Amir Asiaee T. * Soumyadeep Chaterjee † Arindam Banerjee *

Abstract

Structured estimation methods, such as LASSO, have received considerable attention in recent years and substantial progress has been made in extending such methods to general norms and non-Gaussian design matrices. In real world problems, however, covariates are usually corrupted with noise and there have been efforts to generalize structured estimation method for noisy covariate setting. In this paper we first show that without any information about the noise in covariates, currently established techniques of bounding statistical error of estimation fail to provide consistency guarantees. However, when information about noise covariance is available or can be estimated, then we prove consistency guarantees for any norm regularizer, which is a more general result than the state of the art. Next, we investigate empirical performance of structured estimation, specifically LASSO, when covariates are noisy and empirically show that LASSO is not consistent or stable in the presence of additive noise. However, prediction performance improves quite substantially when the noise covariance is available for incorporating in the estimator.

1 INTRODUCTION

The study of regression models with error in features predates the twentieth century [13]. In the simplest setting for such models, we assume that instead of observing $(\mathbf{x}_i, y_i)$ from the linear model $y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i \rangle + \epsilon_i$, $(\mathbf{z}_i, y_i)$ is observed, where $\mathbf{z}_i = f(\mathbf{x}_i, \mathbf{w}_i)$ is a noisy version of $\mathbf{x}_i$ corrupted by $\mathbf{w}_i$. The form of function f which we consider in this paper is additive noise. The overall noisy measurement model is:

$$y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i \rangle + \epsilon_i, \quad \boldsymbol{\beta}^* \in \mathbb{R}^p \tag{1.1}$$

$$\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i. \tag{1.2}$$

Given $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$ we want to compute $\hat{\boldsymbol{\beta}}$, which is l_2 consistent, i.e., for the error vector $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\|\Delta\|_2 \leq g(n)$ where $g(n) \to 0$ for $n \to \infty$. Further, we also want to prove non-asymptotic guarantees for statistical recoevery.

Error in features is known with different names in the literature such as measurement error, errors-in-variables, or noisy covariates, and has applications in various areas of science and engineering [5, 13, 19]. The importance of measurement error models is amplified in the era of big data, since large scale and high dimensional data are more prone to noise [14, 19]. In high dimensional setting where $p \gg n$ the classical assumptions required for treatment of measurement error break down [5, 13] and new estimators and methods are required to consistently estimate $\boldsymbol{\beta}^*$. Such challenges have revived measurement error research and several papers have addressed high dimensional issues of those models in recent years [3, 12, 14, 19, 20].

Many recent papers have reported unstable behavior of standard sparse estimators like LASSO [22] and Dantzig selector (DS) [6] under measurement error. These observations, led to suggestion of new estimators [3, 12, 14, 19, 20] for which some knowledge of noise $\mathbf{w}_i$, and/or $\boldsymbol{\beta}^*$ are required for consistent estimation. None of the existing estimators is able to consistently estimate parameters from noisy measurements without noise information, but there is still no theoretical result to show inachievability.

In this paper, we consider regularized (LASSO type) estimators with general norms $R(\cdot)$, when the design matrix X, with $\mathbf{x}_i$ as its rows, is corrupted by additive independent sub-Gaussian noise matrix W (precise definition of sub-Gaussian random variable follows). Therefore, the additive noise model in matrix form becomes:

$$\begin{aligned} Z &= X + W, \quad Z, X, W \in \mathbb{R}^{n \times p} \\ \mathbf{y} &= X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\epsilon} \in \mathbb{R}^p, \end{aligned} \tag{1.3}$$

where matrix Z is the noisy observation (design) matrix with $\mathbf{z}_i$s as its rows which follow additive noise model of (1.2) and $\mathbf{y}$ is generated from linear model of (1.1). Our regularized estimator takes the form:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{C}}{\operatorname{argmin}} \, \mathcal{L}(Z, \mathbf{y}, \boldsymbol{\beta}) + \lambda R(\boldsymbol{\beta}), \tag{1.4}$$

where $\mathcal{L}$ is a loss function, $\mathcal{C} \subseteq \mathbb{R}^p$ and $R(\cdot)$ is a general norm used for regularization and induces some structure (like sparsity) over the unknown parameter $\boldsymbol{\beta}^*$.

To the best of our knowledge none of the previous work in high dimensional measurement error literature (see Section 2 on the related work) has considered structures other than sparsity, i.e. $R(\boldsymbol{\beta}^*) = \|\boldsymbol{\beta}^*\|_1$. However, other structures of $\boldsymbol{\beta}^*$ are of interest in different applications [1, 2, 7, 15]. These structures are formalized as having a small value for $R(\boldsymbol{\beta}^*)$ where R is a suitable norm.

*Department of CSE, University of Minnesota, Twin Cities, MN. {ataheri, banerjee}@cs.umn.edu

†Yahoo! Labs, Sunnyvale, CA. soumyadeep@yahoo-inc.com

In this paper, we first study the properties of the estimator (1.4) where no knowledge of the noise W is available. This is in the sharp contrast to the recent literature [12, 14, 19] where the noise covariance $\Sigma_{\mathbf{w}} = \mathbf{E}[W^T W] \in \mathbb{R}^{p\times p}$ or an estimate of it, is required as a part of estimator. [14] uses a maximum likelihood estimator, which always requires estimation of $\Sigma_{\mathbf{w}}$ in order to establish restricted eigenvalue conditions [18, 24, 23] on the estimated sample covariance $\Sigma_{\mathbf{x}}$. [12] used orthogonal matching pursuit to recover the support of $\boldsymbol{\beta}^*$ without any knowledge of $\Sigma_{\mathbf{w}}$, but it can not establish l_2 consistency without estimating $\Sigma_{\mathbf{w}}$ directly. Our analysis of estimator (1.4) when $\Sigma_{\mathbf{w}}$ is unknown characterizes the upper bound on $\|\Delta\|_2 \leq g(n) + c(\Sigma_{\mathbf{w}})$, where $g(n)$ decays by the rate of $O(1/\sqrt{n})$ but the constant $c(\Sigma_{\mathbf{w}})$, is not vanishing. Thus, the upper bound on the statistical error does not decay to zero, but remains bounded within a norm ball. Second, we prove that when $\Sigma_{\mathbf{w}}$ is available, the regularized estimators like (1.4) are consistent which generalizes the recent work of [14] for the case of $R(.) = \|.\|_1$.

We study the behavior of high dimensional estimators in the presence of the noise and present three key findings. First, we exploit the current bounding techniques [2, 15] and show that the error of regularized estimators in the presence of noise based on current techniques can only be bounded by two terms one of which shrinks as the number of samples increases and the other one is irreducible and depends on the covariance of the noise. Second, when an estimate of the noise covariance is known, we show that existing estimators [15, 22] provide consistent estimates for any norm regularization R. Our analysis generalizes the existing estimators in the noisy setting, which have only considered sparse regression and l_1 norm regularization. Finally, using LASSO as the estimator we empirically show that in the presence of noise in covariates, even estimation followed by significant test fails to detect all important features, whereas our estimator, having knowledge of noise covariance, captures relevant features more accurately.

The rest of the paper is organized as follows. First we introduce the notation and preliminary definitions. Next, we briefly review the related work in Section 2. In Section 3 we formulate the structured estimation problem under noisy designs assumption using regularized optimization and establish non-asymptotic bounds on the error for sub-Gaussian designs and sub-Gaussian noise. In Section 4, we prove consistency of estimators when an estimate $\hat{\Sigma}_{\mathbf{w}}$ of noise covariance is known. We present supportive numerical simulation results in Section 5 and conclude in Section 6.

NOTATION AND PRELIMINARIES :

We denote matrices by capital letters V, random variables by small letters v and random vectors with bold symbols $\mathbf{v}$. Throughout the paper c_is and C are positive constants. Consider following norm of random variable v: $\|v\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}(|v|^p))^{1/p}$. Then v is sub-Gaussian if $\|v\|_{\psi_2} \leq K_2$ for a constant K_2 [24]. Random vector $\mathbf{v} \in \mathbb{R}^p$ is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{v}, v\rangle$ are sub-Gaussian for all $v \in \mathbb{R}$. The sub-Gaussian norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_{\psi_2} = \sup_{v\in S^{p-1}} \|\langle \mathbf{v}, v\rangle\|_{\psi_2}$. We abuse notation and use shorthand $\mathbf{v} \sim \text{Subg}(0, \Sigma_{\mathbf{v}}, K_{\mathbf{v}})$ for zero mean sub-Gaussian random vector with covariance $\Sigma_{\mathbf{v}}$ and parameter $K_{\mathbf{v}}$, although keeping in mind that no other moments, nor the exact form of the distribution function is known. For any set $A \in \mathbb{R}^p$, the Gaussian width [25] of the set is defined as: $\omega(A) = \mathbb{E}(\sup_{u\in A}\langle g, u\rangle)$, where the expectation is over $g \sim N(0, I_{p\times p})$, a vector of independent zero-mean unit-variance Gaussians. We define the minimum and maximum eigenvalues of a matrix M restricted to set $A \subseteq S^{p-1}$ as $\lambda_{\min}(M|A) = \inf_{\mathbf{u}\in A} \mathbf{u}^T M\mathbf{u}$, and $\lambda_{\max}(M|A) = \sup_{\mathbf{u}\in A} \mathbf{u}^T M\mathbf{u}$ respectively.

2 RELATED WORK

Over the past decade considerable progress has been made on the sparse and structured estimation problems for linear models. Such models assume that the observed pair $(\mathbf{x}_i, y_i)$ follows $y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i\rangle + \epsilon_i$, where $\boldsymbol{\beta}^*$ is sparse or suitably structured according to a norm R [1, 2, 7, 15]. In real world settings, often covariates are noisy, and one observes "noisy" versions $\mathbf{z}_i$ of covariates $\mathbf{x}_i$ corrupted by noise $\mathbf{w}_i$, where $\mathbf{z}_i = f(\mathbf{x}_i, \mathbf{w}_i)$. Two popular model for f are additive, $\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, and multiplicative noise $\mathbf{z}_i = \mathbf{x}_i \circ \mathbf{w}_i$ [12, 14, 19] where $\circ$ is the Hadamard product. Two common choices of $\mathbf{w}_i$ for additive noise case are uniformly bounded [3, 19] and centered subgaussian [12, 14]. In noisy models, a key challenge is to develop estimation methods that are robust to corrupted data, particularly in the high-dimensional regime. Recent work [12, 19] has illustrated empirically that standard estimators like LASSO and DS perform poorly in the presence of measurement errors. Thus, many recent papers proposed modifications of LASSO, DS or Orthogonal Matching Pursuit (OMP) [3, 12, 14, 19, 20] for handling noisy covariates. However, such estimators may become non-convex [14], or require extra information about optimal $\boldsymbol{\beta}^*$ [12, 14]. Further, most of proposed estimators for sub-Gaussian additive noise require an estimate of the noise covariance $\Sigma_{\mathbf{w}}$ in order to establish statistical consistency [3, 12, 14, 20] or impose more stringent condition, like element-wise boundedness on W [3, 19].

Recent literature on regression with additive measurement error in high dimension has focused on sparsity, Table 1 presents key recent works in this area. The first paper in this line of work [19] introduces matrix uncertainty selector (MU) which belongs to constraint family of estimators. As the first attempt for addressing estimation with measurement error in high dimension, MU imposes restrictive conditions on noise W, namely each element of matrix W needs to be bounded. It worth mentioning that MU does not need any information about noise covariance $\Sigma_{\mathbf{w}}$ and as presented in

Table 1: Comparison of estimators for design corrupted with additive sub-Gaussian noise

Name	Estimator	Conditions	Bound for $\|\Delta\|_2$
MU [19]	$\min \|\boldsymbol{\beta}\|_1$ s.t. $\|\frac{1}{n}Z^T(\mathbf{y}-Z\boldsymbol{\beta})\|_\infty \le (1+\delta)\delta\|\boldsymbol{\beta}\|_1+\tau$	$\|\frac{1}{n}Z^T\boldsymbol{\epsilon}\|_\infty \le \tau$ $\forall W_{ij}, \|W_{ij}\| \le \delta$	$c\sqrt{s}(\delta+\delta^2)\|\boldsymbol{\beta}^*\|_1 + C\sqrt{\frac{s\log p}{n}}$
IMU [20]	$\min \|\boldsymbol{\beta}\|_1$ s.t. $\|\frac{1}{n}Z^T(\mathbf{y}-Z\boldsymbol{\beta})+\hat{\Sigma}_{\mathbf{w}}\boldsymbol{\beta}\|_\infty \le \mu\|\boldsymbol{\beta}\|_1+\tau$	$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(W_{ij}^2)$ $\Sigma_{\mathbf{w}} = \text{diag}(\sigma_1,\ldots,\sigma_p)$ $\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$C\|\boldsymbol{\beta}^*\\|_1\sqrt{\frac{s\log p}{n}}$
NCL [14]	$\min \frac{1}{2}\boldsymbol{\beta}^T(\frac{1}{n}Z^TZ-\Sigma_{\mathbf{w}})\boldsymbol{\beta}-\frac{1}{n}\boldsymbol{\beta}^TZ^T\mathbf{y}$ $+\lambda\|\boldsymbol{\beta}\|_1$ s.t. $\|\boldsymbol{\beta}\|_1 \le b_1$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$\max\{c\sqrt{s}\lambda, C\|\boldsymbol{\beta}^*\|_2\sqrt{\frac{s\log p}{n}}\}$
NCC [14]	$\min \frac{1}{2}\boldsymbol{\beta}^T(\frac{1}{n}Z^TZ-\Sigma_{\mathbf{w}})\boldsymbol{\beta}-\frac{1}{n}\boldsymbol{\beta}^TZ^T\mathbf{y}$ s.t. $\|\boldsymbol{\beta}\|_1 \le b_2$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$C\|\boldsymbol{\beta}^*\|_2\sqrt{\frac{s\log p}{n}}$
OMP [12]	OMP for recovery of support indecies S: $\hat{\boldsymbol{\beta}}_S = (Z_S^TZ_S-\Sigma_{\mathbf{w}}^S)(Z_S^T\mathbf{y})$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$ $\forall\beta_i^* \neq 0$ $\|\beta_i^*\| \ge (c\|\boldsymbol{\beta}\|_2+C)\sqrt{\frac{\log p}{n}}$	$(c+C\|\boldsymbol{\beta}^*\|_2)\sqrt{\frac{s\log p}{n}}$

Table 1, it is not consistent, i.e. $c\sqrt{s}(\delta+\delta^2)\|\boldsymbol{\beta}^*\|_1$ term in the upper bound is independent of the number of samples n. This theme repeats in the literature: when $\Sigma_{\mathbf{w}}$ is available proposed estimators are consistent otherwise there is no l_2 recovery guarantee.

Same authors has proposed improved matrix uncertainty selector (IMU) [20] which assumes availability of the diagonal matrix $\hat{\Sigma}_{\mathbf{w}}$ as the covariance of the noise and use it to compensate the effect of the noise. The compensation idea also recurs in the literature where one mitigates Z^TZ by subtracting $\Sigma_{\mathbf{w}}$ and as the result the estimator becomes consistent. Note that both MU and IMU are modification of DS where $\|\boldsymbol{\beta}\|_1$ appears in both constraint and objective of the program. For IMU each row of noise matrix $\mathbf{w}_i$ is sub-Gaussian and independent of $\mathbf{w}_j$, $\mathbf{x}_i$ and ϵ_i and off diagonal of $\Sigma_{\mathbf{w}}$ are zero i.e., W_{ij} are uncorrelated. Following IMU all subsequent work assume sub-Gaussian independent noise and MU and [3] are only estimators that allows general dependence in noise.

Loh and Wainwright [14] proposed a non-convex modification of LASSO (NCL) [22] along with constraint version of it (NCC) which are equivalent by Lagrangian duality (Table 1). In both estimators they substitute the quadratic term X^TX of the LASSO objective with $Z^TZ-\Sigma_{\mathbf{w}}$ which makes the problem non-convex. An interesting aspect of this method is that although a projected gradient algorithm can only reach a local minima, yet any such local minima is guaranteed to have consistency guarantee. Note that for the feasibility of both objectives, [14] requires extra information about the unknown parameter $\boldsymbol{\beta}^*$, particularly b_1 and b_2 should be set to a value greater than $\|\boldsymbol{\beta}^*\|_1$.

In [12], Chen and Caramanis use the OMP [23] for support recovery of a sparse regression problem without knowing the noise covariance. They established non-asymptotic guarantees for support recovery while imposing element-wise lower bound on the absolute value of the support. However, for achieving l_2 consistently, [12] still requires an estimate of the noise covariance $\Sigma_{\mathbf{w}}$, which is in accordance with the requirements of other estimators mentioned above.

Although literature on regression with noisy covariates has only focused on sparsity, the machine learning community recently has made tremendous progress on structured regression that has led to several key publications. [15] provided a general framework for analyzing regularized estimators with decomposable norm of the form $\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta};\mathbf{y},X)+\lambda R(\boldsymbol{\beta})$, and established theoretical guarantees for Gaussian covariates. A number of recent papers [27, 28] have generalized this framework for analyzing estimators with hierarchical structures [9], atomic norms [27] and graphical model structure learning [28]. Recently, [2] established a framework for analyzing regularized estimators with any norm R and sub-Gaussian covariates. On the other hand for constraint estimators [8] has recently generalized the DS for any norm R.

3 STRUCTURED ESTIMATION

We consider the linear model, where covariates are corrupted by additive noise $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle + \epsilon_i$, $\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, where $\mathbf{x}_i \sim \text{Subg}(0,\Sigma_{\mathbf{x}},K_{\mathbf{x}})$, $\epsilon_i \sim \text{Subg}(0,\sigma_\epsilon,K_\epsilon)$ are i.i.d and also independent from one another. Error vector $\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$ is independent from both $\mathbf{x}_i$ and ϵ_i. Since $\mathbf{z}_i$ and $\mathbf{x}_i$ are independent, we have $\Sigma_{\mathbf{z}} = \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}}$ and $\mathbf{z}_i \sim \text{Subg}(0,\Sigma_{\mathbf{z}},K_{\mathbf{z}})$ for $K_{\mathbf{z}} \le c_1K_{\mathbf{x}} + c_2K_{\mathbf{w}}$. In matrix notation, given samples $\{(\mathbf{x}_i,y_i)\}_{i=1}^n$, we obtain

$$\mathbf{y} = X\boldsymbol{\beta}^* + \epsilon, \quad Z = X + W\ . \tag{3.5}$$

The regularized family of estimators in high dimensions is generally characterized as

$$\hat{\boldsymbol{\beta}}_r = \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{2n}\|\mathbf{y}-Z\boldsymbol{\beta}\|_2^2 + \lambda_r R(\boldsymbol{\beta}), \tag{3.6}$$

where $\lambda_r > 0$.

In noiseless scenario, i.e. $Z = X$, (3.6) is called Regularized M-estimators (RME) [2, 15]. R encodes the structure of $\boldsymbol{\beta}^*$. For example, if $\boldsymbol{\beta}^*$ is sparse, i.e. has many zeros, $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and RME (3.6) corresponds to the LASSO problem [22]. When $Z = X$, statistical consistency of RME has been shown for general norms [2]. In the next section, we illustrate that the analysis of [2] can be conducted on RME with noisy design $Z = X + W$, with similar assumptions, but consistency cannot be guaranteed.

3.1 STATISTICAL PROPERTIES For noiseless designs, considerable progress has been made in recent years in the analysis of non-asymptotic estimation error $\|\Delta\|_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ [2, 4, 8, 15, 26]. In this paper, we follow the established analysis techniques, while discussing some of the subtle differences in the results obtained due to presence of the noise in covariates. First we discuss the set of directions which contain the error $\boldsymbol{\Delta}$.

Lemma 1 (Error Set [2]) *Choosing $\lambda_r \geq \alpha R^*(\frac{1}{n} Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*))$ for some $\alpha > 1$, the error vector Δ of RME (3.6) belongs to the restricted error set E_r [2]*

(3.7)
$$E_r = \left\{ \boldsymbol{\Delta} \in \mathbb{R}^p \middle| R(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) \leq R(\boldsymbol{\beta}^*) + \frac{1}{\alpha} R(\boldsymbol{\Delta}) \right\}$$

We name the cone of E_r as $C_r = Cone(E_r)$.

Proof is straightforward and only depend on the optimality of $\hat{\boldsymbol{\beta}}$. Next, we discuss the Restricted Eigenvalue (RE) condition on the design matrix that almost all of the high-dimensional consistency analysis relies on [2, 7, 8, 14, 15, 19, 20].

Definition 1 (Restricted Eigenvalue) The design matrix $Z \in \mathbb{R}^{n \times p}$ satisfies the restricted eigenvalue condition on the spherical cap $A \subset S^{p-1}$, where S^{p-1} is the unit l_2 sphere, if $\frac{1}{\sqrt{n}} \inf_{\mathbf{v} \in A} \|X\mathbf{v}\|_2 \geq \kappa > 0$ or in other words, for $\gamma = \sqrt{n}\kappa$:

$$\inf_{\mathbf{v} \in A} \|X\mathbf{v}\|_2 \geq \gamma > 0 \,. \tag{3.8}$$

Intuitively RE condition means that although for $p \gg n$ the matrix X is not positive definite and the corresponding quadratic form is not strongly convex but in the certain desirable directions represented by A, $||X\mathbf{v}||_2^2$ is strongly convex. In RME these are error vector $\boldsymbol{\Delta}$ directions formulated as $A_r = C_r \cap S^{p-1}$.

For noiseless case $Z = X$ when $\mathbf{x}_i$ are Gaussian or sub-Gaussian RE condition is satisfied with high probability after a certain sample size $n > n_0$ is reached, where n_0 determines the sample complexity [2, 15]. Interestingly, recent work has shown that the sample complexity is the square of the Gaussian width of A, $n_0 = O(\omega^2(A))$ [2].

Theorem 1 (Deterministic Error Bound [2, 8]) *Assume $\lambda_r \geq \alpha R^*(\frac{1}{n} Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*))$ for some $\alpha > 1$ and sample size $n > n_0$ such that RE condition (3.8) holds over the error directions $A_r = C_r \cap S^{p-1}$, then following deterministic bound holds for RME:*

$$\|\boldsymbol{\Delta}_r\|_2 \leq \frac{\alpha + 1}{\alpha} \frac{\lambda_r}{\kappa} \Psi(C_r) \,, \tag{3.9}$$

where $\Psi(C) = \sup_{\mathbf{u} \in C} \frac{R(\mathbf{u})}{\|\mathbf{u}\|_2}$ is the restricted norm compatibility constant.

Next, we analyze the additive noise case, by (i) obtaining suitable bounds for λ, which sets the scaling of the error bound, and (ii) the sample complexity n_0 for which the RE condition is satisfied with high-probability even with a noisy design Z. Without loss of generality, we will assume $\|\beta^*\|_2 = 1$ for the analysis, noting that the general case follows by a direct scaling of the analysis presented.

3.2 RESTRICTED EIGENVALUE CONDITION For linear models with the square loss function, RE condition is satisfied if (3.8) holds, where $A \subseteq \mathbb{S}^{p-1}$ is a restricted set of directions. Recent literature [2, 7, 15] has proved that the RE condition holds for both Gaussian and sub-Gaussian design matrices. In the following theorem we show that RE condition holds for additive noise in measurement with high probability:

Theorem 2 *For the design matrix of the additive noise in measurement $Z = X + W$ where independent rows of X and W are drawn from $\mathbf{x}_i \sim Subg(0, \Sigma_{\mathbf{x}}, K_{\mathbf{x}})$, and $\mathbf{w}_i \sim Subg(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$, for absolute constants $\eta, c > 0$, with probability at least $(1 - 2\exp(-\eta\omega^2(A)))$, we have:*

$$\inf_{\mathbf{v} \in A} \frac{1}{n} \|Z\mathbf{v}\|_2^2 \geq \lambda_{\min}(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}} | A) \left(1 - c\frac{\omega(A)}{\sqrt{n}}\right) \,, \tag{3.10}$$

where $A \subseteq S^{p-1}$.

Proof. Note that $Z = X + W$ and since rows of X and W are centered independent and sub-Gaussian, as mentioned in section 3 rows of Z are also sub-Gaussian with following distribution $\mathbf{z}_i \sim \text{Subg}(0, \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}}, cK_{\mathbf{x}} + CK_{\mathbf{w}})$. Now we apply Theorem 10 of [2] for RE condition of independent anisotropic sub-Gaussian designs and result follows. ∎

In the noisy design problem, our quantity of interest is the Gaussian width $\omega(A_r)$. For example, L_1 norm in LASSO is a simple special case of this model where β^* is s-sparse and we obtain $\omega(A) \leq \sqrt{s \log p}$ [2, 7]. Further, Group-LASSO is the generalization of LASSO to group-sparse norms, where one considers that the dimensions $1, \ldots, p$ are grouped into n_G disjoint groups each of size at most m_G,

and β^* consists of s_G groups. In this scenario, one obtains $\omega(A) \leq \sqrt{m_G} + \sqrt{s_G \log n_G}$ [9, 21]. The k-support norm was introduced in [1] and [8] provided recovery guarantees for k-support norm for linear models. It was shown in [8] that the Gaussian width of the unit ball of the k-support norm is bounded as $\omega(\Omega_{\|\cdot\|_k^{sp}}) \leq \left(\sqrt{2k \log\left(\frac{pe}{k}\right)} + \sqrt{k}\right)$. For related results we refer the readers to [10]

3.3 REGULARIZATION PARAMETER The statistical analysis of RME requires $\lambda \geq \alpha R^*(\frac{1}{n} Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*))$. For the noiseless case, we note that $\mathbf{y} - Z\boldsymbol{\beta}^* = \mathbf{y} - X\boldsymbol{\beta}^* = \boldsymbol{\epsilon}$, the noise vector, so that $R^*(\frac{1}{n} Z^T(y - Z\beta^*)) = R^*(\frac{1}{n} X^T \boldsymbol{\epsilon})$. Using the fact that X and $\boldsymbol{\epsilon}$ are sub-Gaussian and independent, recent work has shown that $E[R^*(\frac{1}{n} X^T \epsilon)] \leq \frac{c}{\sqrt{n}} \omega(\Omega_R)$, where $\Omega_R = \{\mathbf{u} \in \mathbb{R}^p | R(\mathbf{u}) \leq 1\}$. For l_1 norm, i.e., LASSO, Ω_R is the unit l_1 ball, and $\omega(\Omega_R) \leq c_2 \sqrt{\log p}$. Here we have the following bound on λ:

Theorem 3 *Assume that X and W are matrices with iid rows drawn from zero mean sub-Gaussian distributions. Then,*

(3.11)
$$\mathbf{E}\left[R^*\left(\frac{1}{n} Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*)\right)\right] \leq \nu R(\boldsymbol{\beta}^*) + \frac{C\omega(\Omega_R)}{\sqrt{n}} ,$$

where $\nu = \sup_{\mathbf{u} \in \Omega_R} \|\Sigma_{\mathbf{w}}^{1/2} \mathbf{u}\|_2^2$, and $C > 0$ is a constant dependent on the sub-Gaussian norms of the X and W.

Proof. Noting $Z = X + W$ we can see that

(3.12)
$$Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*) = Z^T(\mathbf{y} - X\boldsymbol{\beta}^* - W\boldsymbol{\beta}^*) = Z^T\boldsymbol{\epsilon} - Z^T W\boldsymbol{\beta}^* .$$

Note that there is an additional term $Z^T W\boldsymbol{\beta}^*$ as a consequence of the noise. Now, by triangle inequality

(3.13)
$$R^*(\frac{1}{n} Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*)) \leq R^*(\frac{1}{n} Z^T\boldsymbol{\epsilon}) + R^*(\frac{1}{n} Z^T W\boldsymbol{\beta}^*) .$$

By existing analysis, we know that $E[R^*(\frac{1}{n} Z^T\boldsymbol{\epsilon})] \leq \frac{c_1}{\sqrt{n}} \omega(\Omega_R)$, along with suitable concentration around the expectation [2]. Therefore, the new component of the analysis focuses on the second term $R^*(\frac{1}{n} Z^T W\boldsymbol{\beta}^*)$, which is a consequence of the noise. For simplicity, we consider the case when X is an isotropic bounded sub-Gaussian vectors such that $\Sigma_{\mathbf{x}} = I_{p \times p}$, with sub-Gaussian norm K_1, and W is composed of independent rows sampled from $\text{Subg}(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$. The following lemma provides a suitable upper bound for the expectation of the second term $R^*(\frac{1}{n} Z^T W\boldsymbol{\beta}^*)$. Note that lemma can be easily extended to anisotropic bounded sub-Gaussian X.

Lemma 2 *Assume that the statistical parameter $\boldsymbol{\beta}^*$ has unit L_2 norm, and the matrices X and W consist of isotropic bounded sub-Gaussian entries with sub-Gaussian norm K_1. Then, the following upper bound holds for the expectation.*

(3.14)
$$\mathbf{E}_{X,W}\left[R^*\left(\frac{1}{n} Z^T W\boldsymbol{\beta}^*\right)\right] \leq R(\boldsymbol{\beta}^*)\nu + K_1 c \frac{\omega(\Omega_R)}{\sqrt{n}} + R(\boldsymbol{\beta}^*)\left[\frac{\eta_0 \Lambda_{\max}(\Sigma_{\mathbf{w}})\omega(\Omega_R)}{\sqrt{n}}\right]$$

where $\nu = \sup_{\mathbf{u} \in \Omega_R} \|\Sigma_{\mathbf{w}}^{1/2} \mathbf{u}\|_2^2$ and $c, c_2 > 0$ are constants.

Proof of Lemma: Note that

(3.15)
$$\mathbf{E}\left[R^*\left(\frac{1}{n} Z^T W\boldsymbol{\beta}^*\right)\right] \leq \mathbf{E}\left[R^*\left(\frac{1}{n} X^T W\boldsymbol{\beta}^*\right)\right] + \mathbf{E}\left[R^*\left(\frac{1}{n} W^T W\boldsymbol{\beta}^*\right)\right] .$$

We upper bound the two terms as follows. First, consider the first term.

(3.16)
$$\mathbf{E}_{X,W}\left[R^*\left(\frac{1}{n} X^T W\boldsymbol{\beta}^*\right)\right] = \mathbf{E}_W\left[\frac{1}{n}\|W\boldsymbol{\beta}^*\|_2\right] \mathbf{E}_X\left[R^*\left(X^T\mathbf{u}\right)\right]$$

where $\mathbf{u} = W\boldsymbol{\beta}^*/\|W\boldsymbol{\beta}^*\|_2 \in \mathbb{S}^{p-1}$ is an unit vector and since X and W are independent the expectation factorizes. Since $W\beta^*$ and $X^T\mathbf{u}$ are sub-Gaussian vectors with i.i.d. rows $(W\beta^*)_i$ and $(X^T\mathbf{u})_i$, each of which is sub-Gaussian with sub-Gaussian norm smaller than K_1, we have:

$$\mathbf{E}_W\left[\frac{1}{n}\|W\boldsymbol{\beta}^*\|_2\right] \leq \frac{1}{n} K_1 \sqrt{n} \tag{3.17}$$

$$\mathbf{E}_X\left[R^*\left(X^T\mathbf{u}\right)\right] \leq c\omega(\Omega_R) , \tag{3.18}$$

so that

$$\mathbf{E}_{X,W}\left[R^*\left(\frac{1}{n} X^T W\boldsymbol{\beta}^*\right)\right] \leq K_1 c \frac{\omega(\Omega_R)}{\sqrt{n}} \tag{3.19}$$

Next, we consider the second term, and note that

(3.20)
$$\mathbf{E}_W\left[R^*\left(\frac{1}{n} W^T W\boldsymbol{\beta}^*\right)\right] = \frac{1}{n}\mathbf{E}_W\left[\sup_{\mathbf{u} \in \Omega_R} \langle W\mathbf{u}, W\boldsymbol{\beta}^*\rangle\right]$$

$$\overset{(a)}{=} \frac{R(\boldsymbol{\beta}^*)}{n}\mathbf{E}_W\left[\sup_{\mathbf{u} \in \Omega_R} \langle W\mathbf{u}, W\mathbf{v}\rangle\right] \tag{3.21}$$

$$\overset{(b)}{\leq} R(\boldsymbol{\beta}^*)\mathbf{E}_W\left[\sup_{\mathbf{u} \in \Omega_R} \frac{1}{n}\|W\mathbf{u}\|_2^2\right] \tag{3.22}$$

(3.23)

where (a) follows from noting that $\mathbf{v} = \boldsymbol{\beta}^*/R(\boldsymbol{\beta}^*) \in \Omega_R$, and (b) follows from the inequality $2\langle W\mathbf{u}, W\mathbf{v}\rangle \leq \|W\mathbf{u}\|_2^2 + \|W\mathbf{v}\|_2^2$, and taking supremum over all $\mathbf{u} \in \Omega_R$.

[24] shows that if W consists of i.i.d. sub-Gaussian rows $\mathbf{w}_i \sim \text{Subg}(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$, then

$$\left| \frac{1}{n}\|W\mathbf{u}\|_2^2 - \|\Sigma_{\mathbf{w}}^{1/2}\mathbf{u}\|_2^2 \right| \leq \max(\delta, \delta^2) \quad \forall \mathbf{u} \in \Omega_R \tag{3.24}$$

with probability at least $1 - 2\exp(-\eta_1\tau^2)$, where $\delta = \frac{\eta_0 \Lambda_{\max}(\Sigma_{\mathbf{w}})\omega(\Omega_R)}{\sqrt{n}} + \frac{\tau}{\sqrt{n}}$, and η_0, η_1 are constants dependent on $K_{\mathbf{w}}$. Therefore, we obtain

$$\sup_{\mathbf{u}\in\Omega_R} \frac{1}{n}\|W\mathbf{u}\|_2^2 \leq \nu + \frac{\eta_0 \Lambda_{\max}(\Sigma_{\mathbf{w}})\omega(\Omega_R)}{\sqrt{n}} + \frac{\tau}{\sqrt{n}}\,, \tag{3.25}$$

with probability at least $1 - 2\exp(-\eta_1\tau^2)$, where $\nu = \sup_{\mathbf{u}\in\Omega_R} \|\Sigma_{\mathbf{w}}^{1/2}\mathbf{u}\|_2^2$.

Therefore,

$$\mathbf{E}\left[R^*\left(\frac{1}{n}W^TW\boldsymbol{\beta}^*\right)\right] \leq R(\boldsymbol{\beta}^*)\left[\nu + \frac{\eta_0 \Lambda_{\max}(\Sigma_{\mathbf{w}})\omega(\Omega_R)}{\sqrt{n}}\right] \tag{3.26}$$

■

Remark 1: For the intuitive interpretation of (3.26), note that when the number of samples n increases sample covariance converges as $\frac{1}{n}W^TW \to \Sigma_{\mathbf{w}} = I$, therefore $\mathbf{E}\left[R^*\left(\frac{1}{n}W^TW\boldsymbol{\beta}^*\right)\right] = R^*\left(\boldsymbol{\beta}^*\right)$ which is not decaying by number of samples. Moreover, $R^*\left(\boldsymbol{\beta}^*\right) = \sup_{\mathbf{u}\neq 0} \frac{\langle \boldsymbol{\beta}^*, \mathbf{u}\rangle}{R(\mathbf{u})} = R(\boldsymbol{\beta}^*) \sup_{\mathbf{u}\neq 0} \frac{\langle \boldsymbol{\beta}^*/R(\boldsymbol{\beta}^*), \mathbf{u}\rangle}{R(\mathbf{u})} = R(\boldsymbol{\beta}^*) \sup_{\mathbf{u}\in\Omega_R} \|\mathbf{u}\|_2^2$ which is exactly RHS when $n \to \infty$.

Remark 2: Theorem 3 illustrates that λ does not decay to 0 with increasing sample size, but approaches the operator norm of the covariance matrix $\Sigma_{\mathbf{w}}$. Particularly, when the noise W is i.i.d. with variance $\sigma_{\mathbf{w}}^2$, the error is bounded above by $\sigma_{\mathbf{w}}^2$.

Remark 3: The main consequence of Theorem 3 is to illustrate that the existing technique for proving consistency for the statistical error $\|\Delta\|_2$ of the noiseless estimator fails for RME. We note that in (3.9), when $n > n_0$, κ is a positive quantity that approaches the minimum eigenvalue of $\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}}$ with increasing sample size. Therefore, the scaling of λ determines the error bounds. Theorem 3 proves that the error bound can be as small as the variance of the noise. When $W = 0$, consistency rates are exactly the same as the noiseless case [2].

4 CONSISTENCY WITH NOISE COVARIANCE ESTIMATES

Theorem 3 shows that with no informations about the noise, current analyses can not guarantee statistical consistency for noisy covariates model. At the same time, appearance of $\Sigma_{\mathbf{w}}$ in the upper bound of (3.11), suggests the use of noise covariance estimate to make the estimators consistent. Motivated by this observation and recent line of work [14, 11], we focused on scenarios in which an estimate of the noise covariance matrix $\hat{\Sigma}_{\mathbf{w}}$ is available, e.g., from repeated measurements Z for the same design matrix X, or from independent samples of W. We follow [14] and assume that independent observation from zero mean noise matrix W is possible, from which we estimate the sample covariance as $\hat{\Sigma}_{\mathbf{w}} = \frac{1}{n}W_0^TW_0$. Having $\hat{\Sigma}_{\mathbf{w}}$ in hand we modify RME in the following way. Consider the matrix $\hat{\Gamma} = \frac{1}{n}Z^TZ - \hat{\Sigma}_{\mathbf{w}}$ where $\hat{\Sigma}_{\mathbf{w}}$ compensates the effect of noise W, then:

$$\text{Noisy RME: } \hat{\boldsymbol{\beta}}_r = \underset{R(\boldsymbol{\beta})\leq b}{\operatorname{argmin}} \boldsymbol{\beta}^T\hat{\Gamma}\boldsymbol{\beta} - \boldsymbol{\beta}^T\frac{1}{n}Z^T\mathbf{y} + \lambda R(\boldsymbol{\beta})\,, \tag{4.27}$$

Program (4.27) can be non-convex, because $\hat{\Gamma} = \frac{1}{n}Z^TZ - \hat{\Sigma}_{\mathbf{w}}$ may be indefinite. In such a situation the objective is unbounded below. So we need to impose further constraint of the form $R(\boldsymbol{\beta}) \leq b$ where for the feasibility of $\boldsymbol{\beta}^*$ we set $b = R(\boldsymbol{\beta}^*)$. Our consistency guarantee considers the global solution $\hat{\boldsymbol{\beta}}_r$ of the non-convex problem (4.27). The relation between global and local solutions has been investigated in [14] for the special case of l_1 norm, and for general norms we leave it for the future work. Note that (4.27) "extends" estimator of [14] for any norm, i.e., for $R(\cdot) = \|\cdot\|_1$, (4.27) reduces to the objective of [14].

To show the statistical consistency of $\hat{\boldsymbol{\beta}}$ of noisy RME (NRME), similar to the noiseless case, we need three ingredients, i.e., restricted error set, bound on regularization parameter, and RE condition. The restricted error set of NRME is determined by feasibility of $\hat{\boldsymbol{\beta}}$ as follows:

$$E_w \quad = \quad \left\{\boldsymbol{\Delta} \in \mathbb{R}^p \middle| R(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) \leq R(\boldsymbol{\beta}^*)\right\} \tag{4.28}$$

Note that the restricted error set of the noisy case is a subset of that of noiseless case, i.e., $E_w \subseteq E_r$. Following lemmas shows bounds on λ and RE condition for NRME.

Lemma 3 (Bound on λ for NRME) *With probability* $1 - c_1\exp\{-\min(c_2\tau^2, c_3 n)\}$, $R^*\left(\frac{1}{n}Z^T\mathbf{y} - \hat{\Gamma}\boldsymbol{\beta}^*\right) \leq \frac{c\omega(\Omega_R) + C\tau}{\sqrt{n}}$.

Proof of this lemma follows the same line of proof of Theorem 3, except in this case instead of $R^*\left(\frac{1}{n}W^TW\boldsymbol{\beta}^*\right)$ we end up with $R^*\left(\frac{1}{n}W^TW\boldsymbol{\beta}^* - \frac{1}{n}W_0^TW_0\boldsymbol{\beta}^*\right)$ where W and W_0 have same distributions and cancel out each others effects in expectation. Thus the statement follows.

Lemma 4 (RE condition for NRME) *For matrix* $\hat{\Gamma} = \frac{1}{n}Z^TZ - \hat{\Sigma}_{\mathbf{w}}$ *in the NRME objective with* $Z = X + W$ *where independent rows of* X *and* W *are drawn from* $\mathbf{x}_i \sim$ *Subg*$(0, \Sigma_{\mathbf{x}}, K_{\mathbf{x}})$*, and* $\mathbf{w}_i \sim$ *Subg*$(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$*, and* $\hat{\Sigma}_{\mathbf{w}} = \frac{1}{n}W_0^TW_0$*, for absolute constants* $\eta, c_i > 0$*, with probability*

at least $(1-2\exp(-\eta\omega^2(A_w)))$*, we have:*

$$\begin{aligned}(4.29)\quad &\inf_{\mathbf{v}\in A_w}\mathbf{v}^T\hat{\Gamma}\mathbf{v}\\ &\geq \lambda_{\min}(\Sigma_{\mathbf{x}}|A_w)\left(1-c_1\frac{\omega(A_w)}{\sqrt{n}}\right)\\ &- c_2(\lambda_{\min}(\Sigma_{\mathbf{w}}|A_w)+\lambda_{\max}(\Sigma_{\mathbf{w}}|A_w))\frac{\omega(A_w)}{\sqrt{n}}\,,\end{aligned}$$

where $A_w \subseteq Cone(E_w)\cap S^{p-1}$.

Proof. First we right the RE condition as follows:

$$\begin{aligned}(4.30)\quad &\inf_{\mathbf{v}\in A_w}\mathbf{v}^T\hat{\Gamma}\mathbf{v}\\ &= \frac{1}{n}X^TX+\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}+\Sigma_{\mathbf{w}}-\hat{\Sigma}_{\mathbf{w}}\\ &= \frac{1}{n}X^TX+\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}+\Sigma_{\mathbf{w}}-\frac{1}{n}W_0^TW_0\end{aligned}$$

Now we lower bound $\frac{1}{n}X^TX$, $\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}$, and upper bound $\frac{1}{n}W_0^TW_0-\Sigma_{\mathbf{w}}$. Note that rows of both W and W_0 are iid sampled from same distribution. Therefore, we need lower and upper RE condition for $\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}$. The result can be instantiated from Theorem 12 of [2] where we have following bounds with probability at least $(1-2\exp(-\eta_i\omega^2(A_w)))$

(4.31)

$$\begin{aligned}\lambda_{\min}(\Sigma_{\mathbf{x}}|A_w)\left(1-c_1\frac{\omega(A_w)}{\sqrt{n}}\right) &\leq \inf_{\mathbf{u}\in A_w}\frac{1}{n}\|X\mathbf{u}\|_2^2\\ -c_2\lambda_{\min}(\Sigma_{\mathbf{x}}|A_w)\frac{\omega(A_w)}{\sqrt{n}} &\leq \inf_{\mathbf{u}\in A_w}\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}\\ c_2\lambda_{\max}(\Sigma_{\mathbf{x}}|A_w)\frac{\omega(A_w)}{\sqrt{n}} &\geq \sup_{\mathbf{u}\in A_w}\frac{1}{n}W^TW-\Sigma_{\mathbf{w}}\end{aligned}$$

Putting together the inequities the lemma follows. ∎

Note that if we set $\Sigma_{\mathbf{w}}=0$ in (4.29) we get the established RE condition of the noiseless case [2].

Corollary 1 *When number of samples* n *passes* $n_0 = O(\omega^2(A_w))$*, the objective of NRME (4.27) becomes strongly convex in the direction of restricted error set* E_w*.*

The following theorem shows that NRME (4.27) consistently estimates $\boldsymbol{\beta}^*$.

Theorem 4 *For the design matrix of the additive noise in measurement* $Z=X+W$ *where independent rows of* X *and* W *are drawn from* $\mathbf{x}_i\sim Subg(0,\Sigma_{\mathbf{x}},K_{\mathbf{x}})$*, and* $\mathbf{w}_i\sim Subg(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$*, and for the noise covariance estimate* $\hat{\Sigma}_{\mathbf{w}}=\frac{1}{n}W_0^TW_0$ *discussed above we have the following error bound for regularized estimator (4.27):*

$$(4.32)\qquad \|\Delta\|_2 \leq \frac{2c\Psi(C_r)}{\kappa}\frac{\omega(\Omega_R)}{\sqrt{n}}\,,$$

with probability greater than $(1-c_3\exp(-c_4\omega^2(A_w)))$*, where* $c_3,c_4>0$ *are constants.*

Proof. We start from the optimality of $\hat{\boldsymbol{\beta}}_r$:

$$\begin{aligned}(4.33)\quad &\hat{\boldsymbol{\beta}}^T\hat{\Gamma}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^T\frac{1}{n}Z^T\mathbf{y}+\lambda R(\hat{\boldsymbol{\beta}})\\ &\leq \boldsymbol{\beta}^{*T}\hat{\Gamma}\boldsymbol{\beta}^*-\boldsymbol{\beta}^{*T}\frac{1}{n}Z^T\mathbf{y}+\lambda R(\boldsymbol{\beta}^*)\\ \Rightarrow\ &\Delta^T\hat{\Gamma}\Delta\leq\Delta^T\Big(\frac{1}{n}Z^T\mathbf{y}-\hat{\Gamma}\boldsymbol{\beta}^*\Big)+\lambda(R(\boldsymbol{\beta}^*)-R(\hat{\boldsymbol{\beta}}))\\ \Rightarrow\ &\Delta^T\hat{\Gamma}\Delta\leq\Delta^T\Big(\frac{1}{n}Z^T\mathbf{y}-\hat{\Gamma}\boldsymbol{\beta}^*\Big)+\lambda R(\Delta)\end{aligned}$$

Equation (4.31) shows that the LHS is lower bounded, with probability at least $(1-2\exp(-\eta_*\omega^2(A_w)))$ where $\eta_*>0$ is a constant, by RE condition as $0\leq\kappa\|\Delta\|_2^2\leq\Delta^T\hat{\Gamma}\Delta$, where $\kappa=\lambda_{\min}(\Sigma_{\mathbf{x}}|A_w)\left(1-c_1\frac{\omega(A_w)}{\sqrt{n}}\right)-c_2(\lambda_{\min}(\Sigma_{\mathbf{w}}|A_w)+\lambda_{\max}(\Sigma_{\mathbf{w}}|A_w))\frac{\omega(A_w)}{\sqrt{n}}$ is a positive constant when $n=O(\omega^2(A_w))$. Next, we bound the first term of the RHS, $\frac{1}{n}\Delta^TZ^T\mathbf{y}$ using Holder's inequality:

$$\begin{aligned}(4.34)\quad \Delta^T\Big(\frac{1}{n}Z^T\mathbf{y}-\hat{\Gamma}\boldsymbol{\beta}^*\Big)&\leq R(\Delta)R^*(\frac{1}{n}Z^T\mathbf{y}-\hat{\Gamma}\boldsymbol{\beta}^*)\\ &\leq R(\Delta)\lambda\end{aligned}$$

where the last inequality is from the definition of λ. Putting the bound back to the original inequality (4.33) we get:

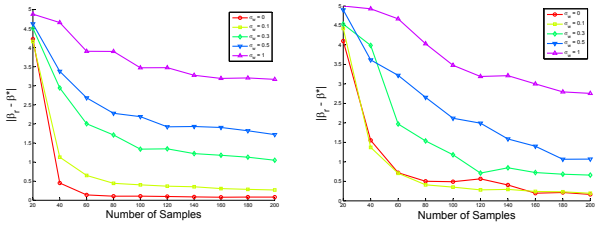
$$(4.35)\quad \|\Delta\|_2^2\leq 2R(\Delta)\frac{\lambda}{\kappa}\leq 2\Psi(C_r)\|\Delta\|_2\frac{\lambda}{\kappa},$$

and using Lemma 4 completes the proof. ∎

Remark: Note that when R is the vector l_1-norm $\omega(\Omega_R)\leq\sqrt{s\log p}$, and we get the rate of $O(\sqrt{\frac{s\log p}{n}})$ for (4.32) which matches the NCL bound of [14]. Note that the NCL [14] bound hinges on the decomposability of the l_1 norm regularizer. Our analysis for (4.32) does not assume decomposability, and follow arguments developed in [8].

5 NUMERICAL SIMULATIONS

In this section we provide numerical simulations to confirm our theoretical results of Section 3. We focus on sparse recovery using noisy RME, i.e., $R(\boldsymbol{\beta})=||\boldsymbol{\beta}||_1$ and investigate l_2-norm consistency.



(a) LASSO (b) Noisy RME

Figure 1: l_2 error vs. number of samples n.

5.1 l_2 ERROR BOUND Experiments with l_2 norm consistency involves observing the norm of the error $\|\Delta\|_2$ which theory predicts it should decrease with the rate of $\frac{1}{\sqrt{n}}$ and converge to some positive number depending on $\Sigma_{\mathbf{w}}$. We generate synthetic data from the model of Section 3 with $\boldsymbol{\beta}^* = (\overbrace{-2,-2,\ldots,-2}^{s/2},\overbrace{1,1,\ldots,1}^{s/2},\overbrace{0,\ldots,0}^{p-s})$, $\mathbf{x}_i \sim N(0, I_{p\times p})$, $\mathbf{w}_i \sim N(0, \sigma^2_{\mathbf{w}} I_{p\times p})$, and $\epsilon_i \sim N(0, 0.1)$ where $p = 100$, $\sigma^2_{\mathbf{w}} \in \{0, 0.1, 0.3, 0.5, 1\}$ and $s = 10$. Note that setting $\sigma^2_{\mathbf{w}} = 0$ results in the standard noiseless linear model. Figure 1 shows that $\|\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}\|_2$ decreases with increasing number of samples. Each point is an average of 50 runs of the experiment. Clearly, when we increase the noise variance $\sigma^2_{\mathbf{w}}$, LASSO is unable to recover the true parameter vector: with 200 samples in noiseless case error drops to $\|\Delta\|_2 \simeq 0.08$ while with noise of $\sigma_w = 1$ it stays around 3. Next we use the Noisy RME estimator and depict the same diagram in Figure 1(b). In all level of noise, $\|\Delta\|_2$ error drops with the similar rate and with 200 samples converges to smaller value than the original estimator.

5.2 Noisy RME vs. Stable Feature Selection Different level of noise in the covariates will effect the features being picked by LASSO. We perform significant test and show that in the case of noisy covariates it is helpful in recovering the true support of the parameter vector. The major problem with significant testing is that, first, one should solve the estimation problem, e.g., LASSO, several times which is not desirable. Secondly, if LASSO de-selects a feature in first place there is no chance that permutation test can pick it up. We show that Noisy RME can be a suitable replacement for LASSO followed by significant testing.

We pick permutation test [16, 17] as our significant testing method. In permutation test we randomly shuffle the output variables y for $v = 1000$ times and each time perform the estimation using LASSO on $\{(\mathbf{x}_i, \pi(y_i))\}_{i=1}^n$ where π is the permutation function. Name the output of LASSO on each permuted data set as $\tilde{\boldsymbol{\beta}}$ and the output of the LASSO on original samples as $\hat{\boldsymbol{\beta}}$. Then we compute the following probability:

$$p_i = \frac{\text{count}(|\tilde{\beta}_i| \geq |\hat{\beta}_i|)}{v+1} \tag{5.36}$$

For $\hat{\beta}_i$ to be a significant coefficient, p_i should be greater than 0.05. We call those $\hat{\beta}_i$s significant factors. For this experiment we set $\boldsymbol{\beta}^* = (\overbrace{-2,-2,\ldots,-2}^{1-10},0,\ldots,0,\overbrace{1,1,\ldots,1}^{51-60},0,\ldots,0)$.

Figure 2 show the result of stability experiment. Each row of diagrams represent the sparsity pattern (i.e., support) of the estimated vector $\hat{\boldsymbol{\beta}}$ except the lowest row which represent the sparsity pattern of true parameter vector $\boldsymbol{\beta}^*$. Figure 2(a) illustrates the features picked by LASSO. As we expect when the noise level increases LASSO starts selecting incorrect support and missing the correct support. To avoid this we perform permutation test after LASSO and get the 2(b) which clearly conforms more to the support of β^*. Although permutation test removes most of the non-support features, at the same time it discards some support feature for even small amount of noise. In contrast noisy RME of 2(c) consistently selects most part of support even for $\sigma_w = 1$. As we expect number of nonzero elements (selected features) by permutation test (101) is less than features selected by LASSO (127), since significant test only select important subset of picked features. Note that number of features picked by noisy RME (115) is the closest (on average) to actual number of support ($120 = 6 \times 20$).

6 CONCLUSION

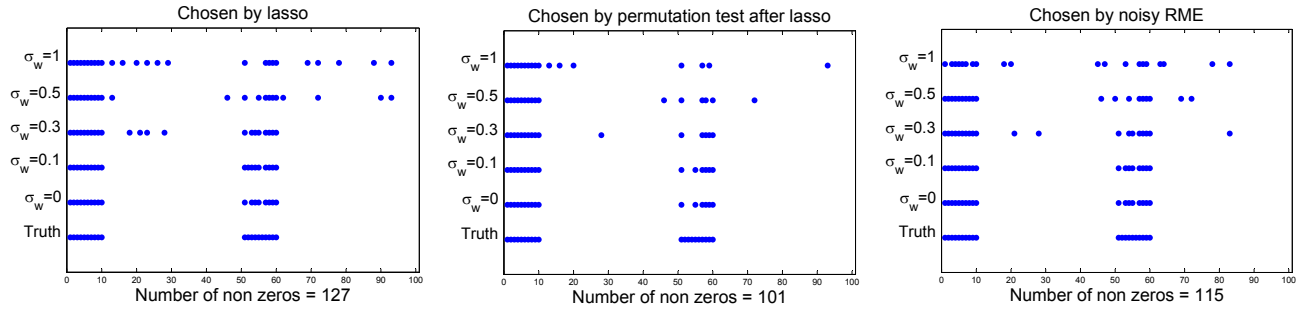
In this paper we investigate consistency of the regularized estimators for structured estimation in high dimensional scaling when covariates are corrupted by additive sub-Gaussian noise. Our analysis holds for any norm R, and shows that when an estimate of the noise covariance is available, our estimators achieve consistent statistical recovery, and recently developed methods for sparse noisy regression are special cases. Finally in the presence of additive noise, our method is stable, i.e., selects the correct support.

Acknowledgment: The research was supported by NSF grants IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, and by NASA grant NNX12AQ39A.

References

[1] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

[2] Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems 27*, pages 1556–1564. 2014.



(a) Features selected by LASSO (b) LASSO followed by permutation test. (c) Features selected by Noisy RME

Figure 2: Comparison between stability of LASSO, LASSO + significant test, and NRME.

[3] Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B. Tsybakov. An $\{l_1, l_2, l_\infty\}$-Regularization Approach to High-Dimensional Errors-in-variables Models. *arXiv*, 2014.

[4] 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.

[5] John P Buonaccorsi. *Measurement error: models, methods, and applications*. CRC Press, 2010.

[6] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313 – 2351, 2007.

[7] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[8] Soumyadeep Chatterjee, Sheng Chen, and Arindam Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1934–1942, 2014.

[9] Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop R Ganguly. Sparse group lasso: Consistency and climate applications. In *SDM*, pages 47–58. SIAM, 2012.

[10] Sheng Chen and Arindam Banerjee. Structured estimation with atomic norms: General bounds and applications. In *Advances in Neural Information Processing Systems*, pages 2890–2898, 2015.

[11] Yudong Chen and Constantine Caramanis. Orthogonal Matching Pursuit with Noisy and Missing Data: Low and High Dimensional Results. 2012.

[12] 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.

[13] W. A. Fuller. *Measurement error models*. J. Wiley & Sons, 1987.

[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] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[16] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

[17] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *The Journal of Machine Learning Research*, 11:1833–1863, 2010.

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

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

[20] Mathieu Rosenbaum and Alexandre B. Tsybakov. Improved Matrix Uncertainty Selector. *arXiv:1112.4413*, 2011.

[21] Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, 2011.

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

[23] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[24] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*. Cambridge University Press, 2012.

[25] Roman Vershynin. Estimation in high dimensions: a geometric perspective. *arXiv:1405.5103*, 2014.

[26] M.J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183 – 2202, 2009.

[27] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 388–396, 2014.

[28] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. Elementary estimators for graphical models. In *Advances in Neural Information Processing Systems*, pages 2159–2167, 2014.