The Annals of Statistics
2017, Vol. 45, No. 6, 2400–2426
https://doi.org/10.1214/16-AOS1527

COCOLASSO FOR HIGH-DIMENSIONAL ERROR-IN-VARIABLES REGRESSION

BY ABHIRUP DATTA AND HUI ZOU[1]

Johns Hopkins University and University of Minnesota

Much theoretical and applied work has been devoted to high-dimensional regression with clean data. However, we often face corrupted data in many applications where missing data and measurement errors cannot be ignored. Loh and Wainwright [*Ann. Statist.* **40** (2012) 1637–1664] proposed a nonconvex modification of the Lasso for doing high-dimensional regression with noisy and missing data. It is generally agreed that the virtues of convexity contribute fundamentally the success and popularity of the Lasso. In light of this, we propose a new method named CoCoLasso that is convex and can handle a general class of corrupted datasets. We establish the estimation error bounds of CoCoLasso and its asymptotic sign-consistent selection property. We further elucidate how the standard cross validation techniques can be misleading in presence of measurement error and develop a novel calibrated cross-validation technique by using the basic idea in CoCoLasso. The calibrated cross-validation has its own importance. We demonstrate the superior performance of our method over the nonconvex approach by simulation studies.

1. Introduction. High-dimensional regression has wide applications in various fields such as genomics, finance, medical imaging, climate science, sensor networks, etc. The current inventory of high-dimensional regression methods includes Lasso [24], SCAD [12], elastic net [31], adaptive lasso [30] and Dantzig selector [8] among others. The articles [13] and [14] provide an overview of these existing methods while the book by [6] discusses their statistical properties in finer details. The canonical high-dimensional linear regression model assumes that the number of available predictors (p) is larger than the sample size (n), although the true number of relevant predictors (s) is much less than n. The model is expressed as $y = X\beta^* + w$ where $y = (y_1, \dots, y_n)'$ is the vector of responses, $X = (x_{ij})$ is the $n \times p$ matrix of covariates, β^* is a $p \times 1$ sparse coefficient vector with only s nonzero entries and $w = (w_1, \dots, w_n)'$ is the noise vector.

Much of the existing theoretical and applied work on high-dimensional regression has focused on the clean data case. However, we often face corrupted data

Received November 2015; revised August 2016.

[1] Supported in part by NSF Grant DMS-15-05111.

MSC2010 subject classifications. Primary 62J07; secondary 62F12.

Key words and phrases. Convex optimization, error in variables, high-dimensional regression, LASSO, missing data.

in many applications where the covariates are observed inaccurately or have missing values. Common examples include sensor network data [22], high-throughput sequencing [3] and gene expression data [19]. It is well known that misleading inference results will be obtained if the regression method for clean data is naively applied to the corrupted data. In order to facilitate further discussion, we assume that we observe a corrupted covariate matrix $Z = (z_{ij})_{1 \le i \le n, 1 \le j \le p}$ instead of the true covariate matrix X. Depending on the context, there can be various ways to model the measurement error. In the additive model setup, $z_{ij} = x_{ij} + a_{ij}$ where $A = (a_{ij})$ is the additive error matrix. In the multiplicative error setup, $z_{ij} = x_{ij} m_{ij}$ where m_{ij}s are the multiplicative errors. Missing predictors can be interpreted as a special case of multiplicative measurement errors with $m_{ij} = I$ (x_{ij} is not missing) where $I(\cdot)$ is the indicator function.

Without loss of generality, we take the Lasso as an example to illustrate the impact of measurement errors. We apply the Lasso to the clean data by minimizing

$$1/(2n)\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{1.1}$$

with respect to β. Here, $\lambda > 0$ is the regularization parameter and $\|\cdot\|_p$ denotes the ℓ_p norm for vectors and matrices for $1 \le p \le \infty$. If we ignore the measurement error issue, we would apply the Lasso to the corrupted data by minimizing:

$$1/(2n)\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1.2}$$

However, as pointed out in [20], the resulting estimate of β is often erroneous if the noise is large. We need to find a proper modification of (1.2) such that its solution is comparable/close to the clean Lasso estimate (1.1).

Observe that the clean Lasso objective function can be equivalently formulated as

$$\frac{1}{2}\beta'\Sigma\beta - \rho'\beta + \lambda\|\beta\|_1 \qquad \text{where } \Sigma = \frac{1}{n}X'X, \rho = \frac{1}{n}X'y. \tag{1.3}$$

In [17], Loh and Wainwright use Z and y to construct unbiased surrogates $\widehat{\Sigma}$ for Σ and $\tilde{\rho}$ for ρ. To elucidate, let us consider the classical additive measurement error case. Following [17], assume the additive errors a_{ij} are independent with mean zero and variance τ^2 where τ^2 is a known constant, then

$$E\left[\frac{1}{n}Z'Z\right] = \frac{1}{n}X'X + \tau^2\mathrm{I}, \qquad E\left[\frac{1}{n}Z'y - \frac{1}{n}X'y\right] = 0.$$

Thus, Loh and Wainwright suggested using unbiased surrogates

$$\widehat{\Sigma} = \frac{1}{n}Z'Z - \tau^2\mathrm{I}, \qquad \tilde{\rho} = \frac{1}{n}Z'y \tag{1.4}$$

and then solve the following optimization problem to get an estimate of β:

$$\frac{1}{2}\beta'\widehat{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1. \tag{1.5}$$

Although the above solution is very natural, (1.5) is fundamentally different from the clean Lasso. Notice that $\widehat{\Sigma}$ may not be positive semi-definite. In fact, when $p > n$, $\widehat{\Sigma}$ is guaranteed to have negative eigenvalues. In such instances, the objective function in (1.5) is no longer convex. Moreover, the objective function is unbounded from below when $\widehat{\Sigma}$ has a negative eigenvalue. To overcome these technical difficulties, Loh and Wainwright defined a constrained estimator similar to the constrained (primal) form of the Lasso as

$$\hat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \frac{1}{2}\beta'\widehat{\Sigma}\beta - \tilde{\rho}'\beta \tag{1.6}$$

for some constant R. They also defined a regularized (and constrained) estimator as

$$\hat{\beta} \in \arg\min_{\|\beta\|_1 \le b_o\sqrt{s}} \frac{1}{2}\beta'\widehat{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1 \tag{1.7}$$

for some constants b_0. Note that "∈" not "=" is used in (1.6) and (1.7) because the objective functions may still have multiple local/global minimizers even within the respective regions $\|\beta\|_1 \le R$ and $\|\beta\|_1 \le b_o\sqrt{s}$. Through some careful analysis, Loh and Wainwright showed that, if b_0 and R are properly chosen, a projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of all global minimizers.

In this article, we propose the *Convex Conditioned Lasso* (*CoCoLasso*) that can handle a general class of corrupted datasets including the cases of additive or multiplicative measurement error and random missing data. CoCoLasso automatically enjoys the theoretical and computational benefits of convexity that contribute fundamentally to the success of the Lasso. Theoretically, we derive the desirable statistical error bounds for the CoCoLasso estimate. Additionally, we establish the asymptotic sign-consistent selection property of CoCoLasso. Earlier [23] derived asymptotic selection consistency properties for the estimator in (1.7) only for the restrictive case of additive measurement error. However, our result does not require any specification of the type of measurement error. This is arguably the most general result for sign consistency in presence of measurement error. Loh and Wainwright (2012) did not provide any sign-consistency result for the nonconvex approach.

Our method has another significant advantage over the nonconvex approach by Loh and Wainwright in practice. Loh and Wainwright's method depend on some crucial hidden parameters. First, theoretical results for the constrained estimator (1.6) in [17] assume $R = \|\beta^*\|_1$. As the authors acknowledge, this is very restrictive as $\|\beta^*\|_1$ is unknown. So they prefer the regularized estimator (1.7), where b_0 is critically important because their theory requires that $b_0 \ge \|\beta^*\|_2$ in order to have desirable error bounds and b_0 cannot be too large due to the required lower-RE and upper-RE conditions. See Theorem 1 in [17] for details. Note that both

β^* and s are unknown. One might try to guess b_0 and s based on the naive lasso estimator, but this is not trustworthy as such an initial estimator suffers from the measurement error. The iterative algorithms used to obtain $\hat{\beta}$ in [17] also depend a step-size of $2\alpha_2$ where α_2 is an upper restricted eigenvalue of the design matrix, which is also unknown. Therefore, in practice one has a lot of difficulties in using the nonconvex approach despite the theoretical results offered in [17]. In contrast, CoCoLasso does not have any of these concerns. CoCoLasso uses one tuning parameter λ which can be chosen by cross-validation in practice.

We notice that in the current literature little attention has been paid to the cross validation methods used for corrupted data. Simply replacing Z by X leads to biased version of the cross validation procedure [similar to (1.5) being a biased version of (1.3)]. This leads to inconsistent estimates of β^* obtained through cross-validation. We demonstrate how the ideas used to develop CoCoLasso can be seamlessly adapted to propose new calibrated cross-validation technique tailored for data with measurement error. To our best knowledge, the existing work on high-dimensional regression with measurement error did not touch on this cross-validation issue. The new calibrated cross-validation has its own independent importance.

It is worth pointing out that a Dantzig selector-type estimator named matrix uncertainty (MU) estimator was proposed in [20] for additive measurement error models. An improved version of MU estimator was proposed in [21]. Belloni et al. [1] establishes near-optimal minimax properties of the estimator in [21] and develops a conic-programming based estimator that achieves minimax bounds. Two more conic programming based estimators have been recently proposed in [2] for the same model setup. It has been empirically observed that solving the Lasso problem can be much faster than solving the Dantzig selector [11]. Compared to Dantzig selector-type estimators and the conic programming based estimators, the direct Lasso-modification methods, such as CoCoLasso, would enjoy computational advantages, which is very important for high-dimensional data analysis.

The rest of the article is organized as follows. In Section 2, we define the CoCoLasso estimator. In Section 3, we discuss the main theoretical results. In Section 4, we discuss the consequences of the results in Section 3 for additive and multiplicative measurement error setups. A new cross-validation technique for corrupted data is developed in Section 5. In Section 6, we present simulation results to demonstrate the empirical performance of CoCoLasso.

2. CoCoLasso.

We first introduce some necessary notation. For any matrix $K = (k_{ij})$, we write $K > 0$ (≥ 0) when it is positive (semi-)definite. Let $\|K\|_\infty = \max_i \sum_j |k_{ij}|$ denote the matrix ℓ_∞ norm whereas $\|K\|_{\max} = \max_{i,j} |k_{ij}|$ denote the elementwise maximum norm. Also, let $\Lambda_{\min}(K)$ and $\Lambda_{\max}(K)$ denote the minimum and maximum eigenvalues of K, respectively. We assume that all variables are centered so that the intercept term is not included in the model and the covariance matrix X has normalized columns, that is, $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$ for every $j = 1, \dots, p$. Without loss of generality, assume that $S = \{1, 2, \dots, s\}$ is the true support set of the regression coefficient vector and write $\beta^* = (\beta_S^{*T}, 0')'$ and $X = (X_S, X_{S^c})$. Hence, the true model can be rewritten as $y = X_S\beta_S^* + w$ where the components of β_S^* are nonzero. For any vector v, we can partition it as $v = (v_S', v_{S^c}')'$. Also, we partition Σ as

$$\Sigma = \begin{pmatrix} (1/n)X_S'X_S & (1/n)X_S'X_{S^c} \\ (1/n)X_{S^c}'X_S & (1/n)X_{S^c}'X_{S^c} \end{pmatrix} = \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,S^c} \\ \Sigma_{S^c,S} & \Sigma_{S^c,S^c} \end{pmatrix}.$$

The true design matrix X is fixed. In the theoretical literature on the clean Lasso, it is often assumed that w_i's are independent and identically distributed sub-Gaussian random variables with parameter σ^2. We use the same assumption here.

As mentioned earlier, in a clean setting where the predictor matrix X is observed accurately, a Lasso estimate is obtained by minimizing (1.3). When the dataset is corrupted by measurement errors, the observed matrix of predictors Z is some function of the true design matrix X and the random error matrix. Based on Z and y, estimates $\widehat{\Sigma}$ and $\tilde{\rho}$ are constructed as surrogates to replace Σ and ρ, respectively, in (1.3). Different pairs of unbiased estimates $(\widehat{\Sigma}, \tilde{\rho})$ are provided in [17] for various types of measurement errors. We will present the actual form of $(\widehat{\Sigma}, \tilde{\rho})$ in Section 4, but for now we only need to assume that $(\widehat{\Sigma}, \tilde{\rho})$ have been constructed.

As discussed earlier, $\widehat{\Sigma}$ is often not positive semi-definite in a high dimensional setup. We now define a nearest positive semi-definite matrix projection operator as follows: for any square matrix K,

$$(K)_+ = \arg\min_{K_1 \geq 0} \|K - K_1\|_{\max}.$$

Then we denote $\widetilde{\Sigma} = (\widehat{\Sigma})_+$ and define our *Convex conditioned Lasso* (*CoCoLasso*) estimate as

$$\hat{\beta} = \arg\min_{\beta} (1/2)\beta'\widetilde{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1. \tag{2.1}$$

We use an alternating direction method of multipliers (ADMM) [5] to obtain $\widetilde{\Sigma}$ from $\widehat{\Sigma}$. The ADMM algorithm is very efficient and details of the algorithm are provided in Appendix A. By definition, $\widetilde{\Sigma}$ is always positive semi-definite. Subsequently, we can reformulate our problem as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|\widetilde{y} - \widetilde{Z}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2.2}$$

where $\widetilde{Z}/\sqrt{n}$ is the Cholesky factor of $\widetilde{\Sigma}$, that is, $\frac{1}{n}\widetilde{Z}'\widetilde{Z} = \widetilde{\Sigma}$ and $\widetilde{y}$ is such that $\widetilde{Z}'\widetilde{y} = \tilde{\rho}$.

Numerically, (2.2) is just like the clean Lasso. One can apply several very fast solvers to solve (2.1), such as the coordinate descent algorithm [15] or the least angle regression algorithm [10]. This is a great advantage for practitioners, as the Lasso solvers are widely used in practice.

Theoretically, (2.1) can be analyzed by the tools for analyzing the clean Lasso. The surrogate $\widehat{\Sigma}$ chosen by [17] is often an unbiased estimate of the true gram matrix Σ, achieving a desired rate of convergence under the max norm. Note that Σ is always positive semi-definite. So by definition, we have

$$\|\widetilde{\Sigma} - \Sigma\|_{\max} \le \|\widetilde{\Sigma} - \widehat{\Sigma}\|_{\max} + \|\widehat{\Sigma} - \Sigma\|_{\max} \le 2\|\widehat{\Sigma} - \Sigma\|_{\max}. \tag{2.3}$$

Equation (2.3) ensures that $\widetilde{\Sigma}$ approximates Σ as well as the initial surrogate $\widehat{\Sigma}$.

Compared with Loh and Wainwright's estimator in [17], CoCoLasso is guaranteed to be convex. This avoids the need of doing any nonconvex analysis of the method. Furthermore, unlike [17] our method does not require any knowledge of $\|\beta\|_1$, and thereby eliminates the need for an initial estimate to obtain a bound for $\|\beta\|_1$. In the next section, we show that CoCoLasso is sign consistent and has desirable ℓ_1, ℓ_2 error bounds.

3. Theoretical analysis.

In this section, we derive the ℓ_1 and ℓ_2 bounds for the statistical error of the CoCoLasso estimate as well as its support recovery probability bounds.

3.1. *Statistical error bounds.*

We assume that $\widehat{\Sigma}$ and $\tilde{\rho}$ are sufficiently "close" to Σ and ρ respectively in the following sense.

DEFINITION 1. Closeness condition: Let us assume that the distribution of $\widehat{\Sigma}$ and $\tilde{\rho}$ are identified by a set of parameters θ. Then there exists universal constants C and c and positive functions ζ and ε_0 depending on θ and σ^2 such that for every $\varepsilon \le \varepsilon_0$, $\widehat{\Sigma}$ and $\tilde{\rho}$ satisfy the following probability statements:

$$\begin{aligned} &\Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \ge \varepsilon) \le C\exp(-cn\varepsilon^2\zeta^{-1}) \qquad \forall i, j = 1, \dots, p, \\ &\Pr(|\tilde{\rho}_j - \rho_j| \ge \varepsilon) \le C\exp(-cns^{-2}\varepsilon^2\zeta^{-1}) \qquad \forall j = 1, \dots, p. \end{aligned} \tag{3.1}$$

The closeness condition requires that the surrogates $\widehat{\Sigma}$ (and hence $\widetilde{\Sigma}$) and $\tilde{\rho}$ are close to Σ and ρ, respectively, in terms of the elementwise maximum norm. We show later in Section 4 that this condition is satisfied by the surrogates defined in [17] for commonly used additive or multiplicative measurement error models.

We also assume the following compatibility or restricted eigenvalue condition:

$$0 < \Omega = \min_{x \ne 0, \|x_{S^c}\|_1 \le 3\|x_S\|_1} \frac{x'\Sigma x}{\|x\|_2^2}. \tag{3.2}$$

Restricted eigenvalue condition similar to this has been used in [26] to obtain bounds of statistical error of the clean Lasso estimate. We show in Lemma 4 that the commonly used version of the restricted eigenvalue condition used to derive the error bounds for the Lasso estimate (see, e.g., [4]) implies condition (3.2). The analogous results for the nonconvex Lasso in [17] are also derived assuming a variant of the restricted eigenvalue condition [4]. Also, the algorithmic results in [17] require an upper restricted eigenvalue condition. We do not need any such assumptions.

We now state the result on the ℓ_1, ℓ_2 and prediction errors of the CoCoLasso estimate. All proofs are provided in Section 8. Note that, for all the theoretical results, C and c denote generic positive constants. Their values may vary from expression to expression.

THEOREM 1. *Under the assumptions stated in equations* (3.1) *and* (3.2), *for* $s\sqrt{(\zeta \log p)/n} < \lambda \leq min(\varepsilon_0, 12\varepsilon_0 \|\beta^*_S\|_\infty)$, *the following results hold with probability at least* $1 - C\exp(-c\log p)$:

$$\text{(3.3)} \quad (\ell_1 \textit{ and } \ell_2 \textit{ error:}) \qquad \|\hat{\beta} - \beta^*\|_2 \leq C\lambda\sqrt{s}/\Omega, \qquad \|\hat{\beta} - \beta^*\|_1 \leq C\lambda s/\Omega,$$

$$\text{(3.4)} \quad (\textit{Prediction error:}) \qquad \|X(\beta^* - \hat{\beta})\|_2/\sqrt{n} \leq C\lambda\sqrt{s}/\sqrt{\Omega}.$$

The finite sample error bounds given in Theorem 1 assume the scaling that $s^2 \log p \ll n$ which is satisfied even when the predictor dimension p varies exponentially with n. For instance, if $p = \mathcal{O}(\exp(n^{c_1}))$ and $s = \mathcal{O}(n^{c_2})$, then $s^2 \log p = o(n)$ as long as $c_1 + 2c_2$ is less than one. The error bounds also depend on the presence of error in the variables through the component ζ. Precise expressions for ζ are derived for the case of additive and multiplicative measurement errors in Section 4.

Theorem 2 of [17] provides error bounds for the estimates obtained by projected gradient descent algorithm for the nonconvex objective function in (1.6). However, owing to the iterative nature of their solutions, the analogous bounds depends on the initial value of β. Specifically, if $\hat{\beta}^{(t)}$ denotes the solution obtained after t iterations starting with an initial value β_0, then the error bounds for $\hat{\beta}^{(t)}$ are inflated by an additional $\mathcal{O}(\alpha^t\|\hat{\beta} - \beta_0\|_2)$ term where $\alpha \in (0, 1)$ and $\hat{\beta}$ denotes the global minimizer of (1.7). Although, this term diminishes at a geometric rate and may seem to be insignificant for large enough t, in practice α depends on the lower and upper restricted eigenvalues of Σ and can be very close to one. Consequently, the rate of decay for this term can be very slow. We have observed in simulations that the error bounds of $\hat{\beta}^{(t)}$ for different choices of initial estimators are drastically different after the same number of iterations. Since, α is not known in practice, the minimum number of iterations required to make this geometric term sufficiently small is also unknown. Hence, the choice of the initial value and number of iterations become very critical to the nonconvex Lasso. CoCoLasso, on the other hand, does not involve any such issues.

3.2. *Sign consistency.* There was no variable selection result for the nonconvex approach in [17]. In this section, we establish the sign consistency of CoCoLasso by assuming the same technical conditions for the sign consistency of the clean lasso. We assume the irrepresentable and minimum eigenvalue conditions on Σ which are sufficient and nearly necessary for sign consistency of the clean Lasso [28–30]:

$$\|\Sigma_{S^c,S}\Sigma_{S,S}^{-1}\|_\infty = 1-\gamma < 1, \qquad \Lambda_{\min}(\Sigma_{S,S}) = C_{\min} > 0. \tag{3.5}$$

The main result on recovery of signed support is stated as follows.

THEOREM 2. *Under the assumptions given in equations* (3.1) *and* (3.5), *for* $\lambda \le \min(\varepsilon_0, 4\varepsilon_0/\gamma)$ *and* $\varepsilon \le \min(\varepsilon_1, \lambda/(\lambda\varepsilon_2 + \varepsilon_3))$ *where* ε_i*'s are bounded positive constants depending of* $\Sigma_{S,S}$, β_S^*, θ *and* σ^2, *the following occurs with probability at least* $1-\delta_1$ *where* $\delta_1 = p^2 C\exp(-cns^{-2}\gamma^2\lambda^2\zeta^{-1}) + p^2 C\exp(-cns^{-2}\varepsilon^2\zeta^{-1})$:

(a) *There exists a unique solution* $\hat{\beta}$ *minimizing* (2.1) *whose support is a subset of the true support.*

(b) $\|\hat{\beta}_S - \beta_S^*\|_\infty \le \kappa\lambda$ *where* $\kappa = (4\|\Sigma_{S,S}^{-1}\|_\infty + C_{\min}^{-1/2})$.

(c) *If* $|\beta_{\min}^*| \ge \kappa\lambda$, *then* $\operatorname{sign}(\hat{\beta}_S) = \operatorname{sign}(\beta_S^*)$.

If we assume for simplicity that κ is $\mathcal{O}(1)$ and the triplet $\{n, p, s\}$ and β^* satisfy:

$$\begin{gathered} s^2 \log p/n \to 0 \qquad \text{as } n, p \to \infty, \\ |\beta_{\min}^*| \gg s(\zeta \log p/n)^{1/2}, \end{gathered} \tag{3.6}$$

then from the expression of δ_1 in Theorem 2 we can choose λ so that $1-\delta_1$ goes to one, which implies the sign-consistency of the CoCoLasso estimate.

COROLLARY 1. *If* Σ, $\widetilde{\Sigma}$ *and* $\tilde{\rho}$ *satisfy the regularity conditions given in Theorem* 2, *then under the scaling in equation* (3.6), *the CoCoLasso estimate* $\hat{\beta}$ *defined in* (2.1) *is sign-consistent if* $|\beta_{\min}^*| \gg \lambda \gg s(\zeta \log p/n)^{1/2}$ *and we also have the* ℓ_∞ *error bound* $\Pr(\|\hat{\beta}_S - \beta_S^*\|_\infty \le \kappa\lambda) \to 1$.

So far in this section we have derived a general theory for the CoCoLasso where there is no assumption on the type of measurement error and the form of the estimates $\widehat{\Sigma}$ and $\tilde{\rho}$. The only condition that requires a careful check is that the estimates $\widehat{\Sigma}$ and $\tilde{\rho}$ are close enough to Σ and ρ, respectively, in the sense defined in (3.1). In the next section, we consider two specific types of error-in-variables models and use the results of this section to derive the theoretical properties of CoCoLasso estimates for those models.

4. CoCoLasso under two types of measurement errors.

4.1. *Additive error.* We assume that the entries of the observed design matrix Z is contaminated by additive measurement error, that is, $z_{ij} = x_{ij} + a_{ij}$ or in matrix notation, $Z = X + A$ where $A = (a_{ij})$ is the matrix of measurement errors. We also assume that the rows of A are independent and identically distributed with 0 mean, finite covariance Σ_A and sub-Gaussian parameter τ^2. Following [17], we assume that Σ_A is known. The unbiased estimates of Σ and ρ are given by $\widehat{\Sigma}_{\text{add}} = \frac{1}{n}Z'Z - \Sigma_A$ and $\tilde{\rho}_{\text{add}} = \frac{1}{n}Z'y$, respectively. It is easy to observe that $\widehat{\Sigma}_{\text{add}}$ can have negative eigenvalues precluding convex optimization. CoCoLasso estimates for this model will be based on the modified objective function

$$\tilde{f}_{\text{add}}(\beta) = (1/2)\beta'\widetilde{\Sigma}_{\text{add}}\beta - \tilde{\rho}'_{\text{add}}\beta + \lambda\|\beta\|_1 \qquad \text{where } \widetilde{\Sigma}_{\text{add}} = (\widehat{\Sigma}_{\text{add}})_+.$$

The following results show that $\widehat{\Sigma}_{\text{add}}$ and $\tilde{\rho}_{\text{add}}$ satisfy the conditions in equation (3.1).

LEMMA 1. *$\widehat{\Sigma}_{\text{add}}$ and $\tilde{\rho}_{\text{add}}$ satisfy the closeness conditions in* (3.1) *with* $\zeta = \max(\tau^4, \sigma^4, 1)$ *and* $\varepsilon_0 = \tau^2$.

So, even though $\widehat{\Sigma}_{\text{add}}$ may not be positive definite, the surrogates $\widehat{\Sigma}_{\text{add}}$ and $\tilde{\rho}$ satisfy (3.1). The following result is an immediate consequence.

COROLLARY 2. *The results of Theorems* 1 *and* 2 (*and Corollary* 1) *hold for the CoCoLasso estimate for the additive error model under the assumptions* (3.2) *and* (3.5) [*and* (3.6)], *respectively.*

From the expression of ζ in Lemma 1, we observe that for every fixed value of τ^2 and σ^2, the CoCoLasso estimate for additive measurement error achieves statistical consistency for any $\lambda \gg s\sqrt{\zeta \log p/n}$. However, as ζ increases with τ we see that the lower bound for λ required in the Theorem 1 and Corollary 1 also increases with τ. This implies that more penalization is required in presence of larger measurement error to accurately recover the sparse support or equivalently for larger τ we need a larger sample size to achieve the same error bounds.

Note that the additive error covariance Σ_A is assumed to be known in order to compute the CoCoLasso estimate. Similar assumption was used in [17] and [21] as it is unclear how to obtain a data-driven estimate of Σ_A when only one dataset is available. If however, multiple replicates of the data are available, following [17], one can obtain a data-driven estimate $\widehat{\Sigma}_A$ of Σ_A and define $\widehat{\Sigma}_{\text{add}} = \frac{1}{n}Z'Z - \widehat{\Sigma}_A$.

4.2. *Multiplictive error and missing data.* If we assume that the errors are multiplicative, we observe $z_{ij} = x_{ij}m_{ij}$. In matrix notation, we have $Z = X \odot M$ where $M = (m_{ij})$ and $\odot$ denotes the elementwise multiplication operator for vectors and matrices. We assume that the rows of M are independent and identically distributed with mean μ_M, covariance Σ_M and sub-Gaussian parameter τ^2. Under the assumption that the entries of μ_M and $\Sigma_M + \mu_M\mu_M'$ are strictly positive, [17] suggests using the unbiased surrogates $\widehat{\Sigma}_{\text{mult}} = (1/n)ZZ' \oslash (\Sigma_M + \mu_M\mu_M')$

and $\tilde{\rho}_{\text{mult}} = (1/n)Z'y \oslash \mu_M$ where $\oslash$ denotes the elementwise division operator for vectors and matrices. $\widehat{\Sigma}_{\text{mult}}$ once again may not be positive semi-definite. The CoCoLasso estimate $\hat{\beta}$ is obtained as

$$\min_{\beta}(1/2)\beta'(\widetilde{\Sigma}_{\text{mult}})_{+}\beta - \tilde{\rho}'_{\text{mult}}\beta + \lambda\|\beta\|_1 \qquad \text{where } \widetilde{\Sigma}_{\text{mult}} = (\widehat{\Sigma}_{\text{mult}})_{+}.$$

Randomly missing covariates can be formulated as a multiplicative error model. For example, a simple model assumes that x_{ij}'s are missing randomly with probability r and their missing statuses are independent of one another. Then we can defining $z_{ij} = x_{ij}m_{ij}$ where $m_{ij} = I(x_{ij} \text{ is not missing }) \sim \text{Bernoulli}(1-r)$. Other missing data models with different choices of the missing probabilities [e.g., $m_{ij} \sim \text{Bernoulli}(1-r_j)$] will also fall under the same setup. We can obtain estimate of r (or r_j) as the proportion of missing entries in the matrix (or in the jth column). For simplicity, we can assume r is known and then Σ_M and μ_M are known as well.

We now establish analogous results for the CoCoLasso estimate in this multiplicative model setup. Note that as the errors are multiplicative, in order to have all the z_{ij}'s to be close to the respective x_{ij}'s, we need an upper bound for both x_{ij} and m_{ij}. We also need a positive lower bound for the entries of μ_M and $\Sigma_M + \mu_M\mu'_M$ for the expressions of $\widehat{\Sigma}_{\text{mult}}$ and $\tilde{\rho}_{\text{mult}}$ to be meaningful. To ensure these, we impose the following additional set of regularity conditions for the multiplicative setup:

$$\begin{aligned} &\max_{i,j}|X_{ij}| = X_{\max} < \infty, \qquad \min_{i,j}E(m_1m'_1) = M_{\min} > 0, \\ &\min \mu_M = \mu_{\min} > 0, \qquad \max \mu_M = \mu_{\max} < \infty. \end{aligned} \tag{4.1}$$

Under these regularity conditions, the following lemma shows that $\widetilde{\Sigma}_{\text{mult}}$ and $\tilde{\rho}_{\text{mult}}$ satisfies the conditions in (3.1).

LEMMA 2. *$\widehat{\Sigma}_{\text{mult}}$ and $\tilde{\rho}_{\text{mult}}$ satisfy the closeness conditions in* (3.1) *with $\zeta = \max(\tau^4, \sigma^4, 1)$ and $\varepsilon_0 = \tau^2$.*

Having proved Lemma 2, once again we use Theorems 1, 2 and Corollary 1 to have the following results.

COROLLARY 3. *The results of Theorems 1 and 2* (*and Corollary 1*) *hold for the CoCoLasso estimate for the multiplicative error/missing data model under the assumptions* (3.2) *and* (3.5) [*and* (3.6)], *respectively.*

5. Calibrated cross-validation. In applications, cross-validation [16] is a widely used technique for choosing the tuning parameter in penalized methods. However, cross validation for data corrupted with measurement error has received very little attention. In the presence of noisy/corrupted data, naive application of cross-validation is biased and a novel correction is needed. To elucidate, consider the usual K-fold cross validation for selecting the tuning parameter in the clean Lasso. Let (X_k, y_k) denote the true design matrix and response vector for the kth fold of the data for $k = 1, 2, \dots, K$. Likewise, let (X_{-k}, y_{-k}) denote the design matrix and response vector, respectively, after removing the kth fold. In absence of measurement error, the estimate for the prediction error for the kth fold is given by $\frac{1}{n_k}\|y_k - X_k\hat{\beta}_k(\lambda)\|_2^2$ where n_k is the size of the kth fold and $\hat{\beta}_k(\lambda)$ is the Lasso estimate based on X_{-k}, y_{-k} with tuning parameter λ. The optimal λ is obtained by minimizing the total cross-validation error, that is,

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\|y_k - X_k\hat{\beta}_k(\lambda)\|_2^2. \tag{5.1}$$

However, when we face noisy/corrupted data, as X is unknown or partially missing, (5.1) is not directly available. If we naively use the observed data (Z, y), then the cross-validated choice of λ is defined by minimizing

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\|y_k - Z_k\hat{\beta}_k(\lambda)\|_2^2. \tag{5.2}$$

Even when we use the CoCoLasso (or the estimator in 1.7) to compute $\hat{\beta}_k(\lambda)$ based on Z_{-k}, y_{-k}, the above criterion is biased compared to (5.1) in the same way the loss function in (1.5) is a biased version of (1.3).

Using simple algebra, we observe that (5.1) is equivalent to

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k(\lambda)'\Sigma_k\hat{\beta}_k(\lambda) - 2\rho_k'\hat{\beta}_k(\lambda), \tag{5.3}$$

where $\Sigma_k = \frac{1}{n_k}X_k'X_k$ and $\rho_k = \frac{1}{n_k}X_k'y_k$.

It may seem that using the unbiased surrogates $\widehat{\Sigma}_k$ and $\tilde{\rho}_k$ in (5.3) may overcome the bias issue. However, as $\widehat{\Sigma}_k$ possibly has negative eigenvalues, this will lead to a cross validation function unbounded from below.

In the light of the above discussion, we propose a new cross validation method for corrupted data that adapts the same central idea used to construct CoCoLasso, that is, we can use $(\widehat{\Sigma}_k)_+$ and $\tilde{\rho}_k$ in (5.3). With this correction, the cross-validated λ is defined as

$$\tilde{\lambda} = \arg\min_{\lambda} \sum_{k=1}^{K}\hat{\beta}_k(\lambda)'(\widehat{\Sigma}_k)_+\hat{\beta}_k(\lambda) - 2\tilde{\rho}_k'\hat{\beta}_k(\lambda). \tag{5.4}$$

We call the above procedure the calibrated cross-validation.

6. Numerical studies. We use simulated datasets to evaluate the performance of CoCoLasso. For comparison we also included the nonconvex Lasso (NCL) by Loh and Wainwright described in (1.6).

6.1. *Simulation models.* We considered both additive measurement errors and multiplicative measurement errors in the simulation study.

Additive errors case. We generate data from the model $y \sim N(X\beta^*, \sigma^2 I)$ where

$$\beta^* = (3, 1.5, 0, 0, 2, 0, \dots, 0)'.$$

The sample size n is set to be 100 and $p = 250$. The rows of X are independent and identically distributed normal random variables with mean zero and covariance matrix Σ_X. We consider two models for Σ_X—autoregressive ($\Sigma_{X,ij} = 0.5^{|i-j|}$) and compound symmetry [$\Sigma_{X,ij} = 0.5 + I(i = j) * 0.5$]. We set $\sigma = 3$ giving a signal to noise ratio of 2.36 for autoregressive (AR) and 3.20 for compound symmetry (CS). We generate $Z = X + A$ where the rows of A are independent and identically distributed $N(0, \tau^2 I)$ where $\tau = 0.75$, 1 and 1.25.

Multiplicative errors case. We also evaluated the performance of CoCoLasso and NCL in a multiplicative errors setup. The true model is assumed to be same as in the additive error setup. We now generate $Z = X \odot M$ where we assume that the elements of $M = (m_{ij})$ follow log-normal distribution, that is, $\log(m_{ij})$'s are independent and identically distributed $N(0, \tau^2)$ where $\tau = 0.25$, 0.5 and 0.75.

6.2. *Simulation results and conclusions.* We used 5-fold calibrated cross-validation for the CoCoLasso in our numerical examples. The code for NCL was provided by Dr. Po-Ling Loh. NCL requires an initial estimator. Following [23], the initial estimate is a naive Lasso estimate based on y and Z which is tuned by 5-fold cross validation. NCL also requires knowledge of $\|\beta_S^*\|_1$ for choosing the constraint parameter. Since this is impossible to know beforehand, a naive 5-fold cross validation was used to select the optimal R from 100 equally spaced values in $[R_{\max}/500, 2 * R_{\max}]$ where $R_{\max}$ is the ℓ_1 norm of the initial estimate.

The accuracy of estimators is gauged by the Prediction Error (PE) and the Squared Error (SE) where

$$\mathrm{PE}(\hat{\beta}) = (\beta^* - \hat{\beta})' \Sigma_X (\beta^* - \hat{\beta})$$

and

$$\mathrm{SE}(\hat{\beta}) = \|\beta^* - \hat{\beta}\|_2^2.$$

To evaluate variable selection, we record C and IC that denote the number of correct and incorrect predictors identified, respectively.

Tables 1 and Table 2 summarize the simulation results for the additive error case and the multiplicative error case, respectively. For each of the four statistics C, IC, SE and PE we present the median numbers based on $N = 100$ Monte Carlo simulations. The standard errors of the medians are calculated using bootstrap as

TABLE 1

Summary statistics for the additive error simulation study based on 100 *replications. Reported numbers are the medians and standard errors* (*se*) *are computed by bootstrap. "CoCo" stands for CoCoLasso. "NCL" is the method in Loh and Wainwright* [17]. *AR denotes Autoregressive covariance for the predictors whereas CS denotes compound symmetry covariance*

		$\tau = 0.75$		$\tau = 1.0$		$\tau = 1.25$	
		CoCo	**NCL**	**CoCo**	**NCL**	**CoCo**	**NCL**
AR	C	3 (0)	3 (0)	3 (0)	2 (0.07)	3 (0.07)	2 (0.45)
	IC	11 (0.75)	3 (0.69)	11 (1.09)	1 (0.33)	10 (0.84)	0 (0.17)
	PE	3.66 (0.19)	4.13 (0.26)	5.8 (0.26)	6.91 (0.34)	8.49 (0.5)	10.92 (0.46)
	SE	3.81 (0.19)	3.76 (0.18)	5.57 (0.2)	6.07 (0.27)	7.94 (0.24)	8.36 (0.3)
CS	C	2 (0.18)	2 (0)	2 (0)	1.5 (0.48)	2 (0.03)	1 (0)
	IC	14 (0.64)	11.5 (0.58)	18 (0.71)	7 (0.22)	21 (0.48)	5 (0.28)
	PE	4.49 (0.22)	4.57 (0.31)	6.03 (0.22)	6.91 (0.34)	6.99 (0.25)	10.47 (0.58)
	SE	8.05 (0.33)	8.03 (0.48)	11.01 (0.4)	10.31 (1.0)	12.97 (0.34)	15.06 (1.06)

follows. We calculate the prediction error PE 100 times, once from each dataset. We resample from this sample of PEs to create a bootstrapped sample of size N and calculate the median PE. We repeat this process 500 times and the standard error of the 500 medians gives the bootstrapped standard error for median of PE. We use the same procedure for all the other three statistics as well.

We observe that CoCoLasso is more accurate than NCL, and the gap between the two methods widens as the perturbation level increases (measured by τ). NCL

TABLE 2

Summary statistics for the multiplicative error simulation study based on 100 *replications. Reported numbers are the medians and standard errors* (*se*) *are computed by bootstrap. "CoCo" stands for CoCoLasso. "NCL" is the method in Loh and Wainwright* [17]. *AR denotes Autoregressive covariance for the predictors whereas CS denotes compound symmetry covariance*

		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		CoCo	**NCL**	**CoCo**	**NCL**	**CoCo**	**NCL**
AR	C	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)	2 (0)
	IC	14 (1.41)	12 (2.4)	12 (0.87)	6 (0.74)	10 (0.81)	1 (0.46)
	PE	2.02 (0.15)	2.47 (0.18)	3.25 (0.14)	3.58 (0.25)	7.32 (0.2)	8.32 (0.29)
	SE	1.95 (0.09)	2.26 (0.14)	2.93 (0.14)	3.09 (0.18)	6.19 (0.2)	6.58 (0.26)
CS	C	3 (0)	3 (0)	3 (0.18)	3 (0.18)	2 (0)	1 (0.45)
	IC	15 (0.72)	18 (1.49)	13 (0.77)	11 (0.7)	16 (0.88)	4 (0.36)
	PE	2.23 (0.16)	2.37 (0.1)	3.66 (0.15)	3.82 (0.19)	7.93 (0.3)	9.31 (0.41)
	SE	4.21 (0.27)	4.32 (0.21)	6.11 (0.27)	5.75 (0.26)	10.43 (0.25)	9.34 (0.61)

tends to select a sparser model than CoCoLasso, it often misses importance variables as the noise level is high.

7. Summary. In this paper, we have proposed a novel convex approach to modify the classical Lasso with the clean data to handle the noisy data case. Our approach, named CoCoLasso, is easy to understand, easy to use and has solid theoretical foundations. We also have devised a novel cross validation methods for corrupted data. We have demonstrated the superior performance of our method over the nonconvex approach in Loh and Wainwright [17] by simulation studies.

Cross-validation is an integrated part of many modern statistical methods. In the presence of measurement error, the usual cross-validation has a systematic bias issue which has been ignored in the literature. We have proposed a calibrated cross-validation to fix the bias issue. A future research topic is to prove the consistency of calibrated cross-validation.

Finally, we would like to comment on the generality of the CoCoLasso approach. Although we use the Lasso to illustrate the idea of CoCoLasso, the basic approach of CoCoLasso can be directly used in conjunction with other popular convex penalized methods. For example, the fused Lasso [25] is a popular technique for ordered variable selection. Following the development of CoCoLasso, we can readily develop CoCo-FusedLasso. We opt not to discuss these variants in the present paper.

8. Proofs. In this section, we present the proofs of Theorems 1 and 2 as well as Lemmas 1 and 2. A few useful properties and technical results about sub-Gaussian random variables required in the proofs are provided in Appendix B. Throughout this section, we denote C and c to be universal constants whose values may vary across different expressions. We also introduce a few additional notation used subsequently in the proofs:

$$\begin{aligned}
&D = \widetilde{\Sigma} - \Sigma, \qquad G = \Sigma_{S^c,S}\Sigma_{S,S}^{-1},\\
&\widetilde{G} = \widetilde{\Sigma}_{S^c,S}\widetilde{\Sigma}_{S,S}^{-1}, \qquad H = \widetilde{G} - G,\\
&F = \widetilde{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}, \qquad \phi = \|\Sigma_{S,S}^{-1}\|_\infty,\\
&\psi = \|\Sigma_{S,S}\|_\infty, \qquad B = \|\beta_S^*\|_\infty.
\end{aligned} \tag{8.1}$$

8.1. *Proof of Theorem* 1. We first state and prove a simple result which will be later used in the proof.

LEMMA 3. *For any* $\varepsilon > 0$, *we have*

$$\Pr\bigl(\|\widetilde{\Sigma} - \Sigma\|_{\max} \ge \varepsilon\bigr) \le p^2 \max_{i,j} \Pr\bigl(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \ge \varepsilon/2\bigr). \tag{8.2}$$

PROOF. From equation (2.3), we have

$$\Pr\big(\|\widetilde{\Sigma} - \Sigma\|_{\max} \geq \varepsilon\big) \leq \Pr\big(\|\widehat{\Sigma} - \Sigma\|_{\max} \geq \varepsilon/2\big).$$

The proof then follows using union bounds over $\Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \varepsilon/2)$. □

Note that the compatibility condition (3.2) is slightly different from the restricted eigenvalue condition used to derive ℓ_2 error bounds for the traditional Lasso estimate. However, the following lemma shows that the restricted eigenvalue condition defined in [4] as

$$\min_{\substack{A\subseteq\{1,2,\dots,p\}\\ |A|\leq s}} \min_{\substack{x\neq 0\\ \|x_{A^c}\|_1\leq 4\|x_A\|_1}} \min \frac{x'\Sigma x}{\|x_A\|_2^2} = \Omega' > 0 \tag{8.3}$$

is sufficient to ensure the compatibility condition (3.2).

LEMMA 4. *The restricted eigenvalue condition* (8.3) *implies the compatibility condition* (3.2).

PROOF. Let $x \in \mathbb{R}^p$ such that $\|x_{S^c}\|_1 \leq 3\|x_S\|_1$. Let A denote the index set corresponding to the entries of x with s-highest absolute values. Hence, $\|x_{A^c}\|_\infty \leq \|x_A\|_1/s$ and $\|x_S\|_1 \leq \|x_A\|_1$. Also,

$$\begin{aligned}\|x_{A^c}\|_1 &= \|x_{A^c\cap S}\|_1 + \|x_{A^c\cap S^c}\|_1\\ &\leq s\|x_{A^c}\|_\infty + \|x_{S^c}\|_1\\ &\leq \|x_A\|_1 + 3\|x_S\|_1\\ &\leq 4\|x_A\|_1.\end{aligned}$$

Using this, we have $\|x_{A^c}\|_2^2 \leq \|X_{A^c}\|_\infty\|X_{A^c}\|_1 \leq 4\|X_A\|_1^2/s \leq 4\|X_A\|_2^2$. So,

$$\frac{x'\Sigma x}{x'x} = \frac{x'\Sigma x}{\|X_A\|_2^2 + \|X_{A^c}\|_2^2} \geq \frac{x'\Sigma x}{5\|X_A\|_2^2} \geq \frac{\Omega'}{5}.$$ □

PROOF OF THEOREM 1. The general idea of the proof closely resembles the proofs of [6], Lemma 6.3 and Theorem 6.1, for obtaining the error bounds of the traditional Lasso estimate. From the definition of $\hat{\beta}$ in (2.1), we have

$$\frac{1}{2}\hat{\beta}'\widetilde{\Sigma}\hat{\beta} - \tilde{\rho}'\hat{\beta} + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\beta^{*T}\widetilde{\Sigma}\beta^* - \tilde{\rho}'\beta^* + \lambda\|\beta^*\|_1.$$

Expanding $\hat{\beta}$ as $\hat{v} + \beta^*$ where $\hat{v} = \hat{\beta} - \beta^*$, this simplifies to

$$\begin{aligned}\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}\|_1 &\leq \hat{v}'(\tilde{\rho} - \widetilde{\Sigma}\beta^*) + \lambda\|\beta^*\|_1\\ &\leq \|\hat{v}\|_1\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty + \lambda\|\beta^*\|_1.\end{aligned} \tag{8.4}$$

In order to obtain an upper bound for the left-hand side, we first bound the quantity $\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty$. Using triangular inequality, we have

$$\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty \leq \|\tilde{\rho} - \rho\|_\infty + \|\rho - \Sigma\beta^*\|_\infty + \|D\beta^*\|_\infty.$$

Using union bounds on the second equation of (3.1), we see that for $\lambda \leq 6\varepsilon_0$, we have $P(\|\tilde{\rho} - \rho\|_\infty > \lambda/6) \leq pC\exp(-ncs^{-2}\lambda^2\zeta^{-1})$. As $\|D\beta^*\|_\infty \leq sB\|D\|_{\max}$, Lemma 3 along with the first equation of (3.1) implies that for $\lambda \leq 12B\varepsilon_0$, $P(sB\|D\|_{\max} > \lambda/6) \leq p^2C\exp(-ncs^{-2}\lambda^2\zeta^{-1}B^{-2})$. The third component $\rho - \Sigma\beta^* = \frac{1}{n}X'w$ is a linear combination of independent sub-Gaussian errors w. As the columns of X are normalized, invoking property B.2, we have $P(\|\rho - \Sigma\beta^*\|_\infty > \lambda/6) \leq pC\exp(-nc\lambda^2\sigma^{-2})$. Redefining ζ as the maximum of previous ζ and σ^2, we have

$$\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty < \lambda/2 \qquad \text{on } \mathcal{F} \text{ where } P(\mathcal{F}) \geq 1 - p^2C\exp(-ncs^{-2}\lambda^2\zeta^{-1}).$$

For the remainder of the proof, we restrict ourselves to $\mathcal{F}$ adjusting for the probability of $\mathcal{F}^c$. Returning to equation (8.4), we now have on $\mathcal{F}$,

$$\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}\|_1 \leq \frac{\lambda}{2}\|\hat{v}\|_1 + \lambda\|\beta^*\|_1.$$

Since $\beta^*_{S^c} = 0$, we know that $\hat{v}_{S^c} = \hat{\beta}_{S^c}$, $\|\beta^*\|_1 = \|\beta^*_S\|_1$. Also for any vector x, we can write $\|x\|_1 = \|x_S\|_1 + \|x_{S^c}\|_1$. Combining these, we have

$$\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}_S\|_1 + \lambda\|\hat{v}_{S^c}\|_1 \leq \frac{\lambda}{2}\|\hat{v}_S\|_1 + \frac{\lambda}{2}\|\hat{v}_{S^c}\|_1 + \lambda\|\beta^*_S\|_1.$$

Using the fact that $\|\hat{\beta}_S\|_1 \geq \|\beta^*_S\|_1 - \|\hat{v}_S\|_1$, we now have

$$\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{v}_{S^c}\|_1 \leq 3\lambda\|\hat{v}_S\|_1. \tag{8.5}$$

As $\hat{v}'\widetilde{\Sigma}\hat{v} \geq 0$, we have that on $\mathcal{F}$, $\|\hat{v}_{S^c}\|_1 \leq 3\|\hat{v}_S\|_1$. The compatibility condition (3.2) implies that on $\mathcal{F}$, $\|\hat{v}_S\|_1^2 \leq s\|\hat{v}\|_2^2 \leq s\hat{v}'\Sigma\hat{v}/\Omega$. Now

$$\begin{aligned}
\hat{v}'\Sigma\hat{v} + \lambda\|\hat{v}\|_1 &= \hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{v}_S\|_1 + \lambda\|\hat{v}_{S^c}\|_1 + \hat{v}'D\hat{v} \\
&\leq 4\lambda\|\hat{v}_S\|_1 + \hat{v}'D\hat{v} \qquad \text{using equation (8.5)} \\
&\leq 4\lambda\sqrt{s}\sqrt{\frac{\hat{v}'\Sigma\hat{v}}{\Omega}} + \hat{v}'D\hat{v} \qquad \text{using condition (3.2)} \\
&\leq \frac{\hat{v}'\Sigma\hat{v}}{4} + \frac{16\lambda^2 s}{\Omega} + |\hat{v}'D\hat{v}| \qquad \text{using } 4ab \leq a^2/4 + 16b^2.
\end{aligned}$$

The last term on the right-hand side is bounded as follows:

$$\begin{aligned}
|\hat{v}'D\hat{v}| &\leq \|D\|_{\max}\|\hat{v}\|_1^2 = \|D\|_{\max}\left(\|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1\right)^2 \leq 16\|D\|_{\max}\|\hat{v}_S\|_1^2 \qquad \text{on } \mathcal{F} \\
&\leq 16\|D\|_{\max}\frac{s\hat{v}'\Sigma\hat{v}}{\Omega} \qquad \text{(compatibility condition).}
\end{aligned}$$

Using Lemma 3 and the closeness condition (3.1), for ε less than some constant ε_0,

$$P\big(16s\|D\|_{\max} > \Omega/4\big) = P\big(\|D\|_{\max} > \Omega/64s\big) \le p^2 C \exp\big(-ncs^{-2}\varepsilon^2\zeta^{-1}\big).$$

Hence, with probability greater than $1 - p^2C\exp(-ncs^{-2}\varepsilon^2\zeta^{-1}) - p^2C\exp(-ncs^{-2}\lambda^2\zeta^{-1})$ we now have

$$\hat{v}'\Sigma\hat{v} + \lambda\|\hat{v}\|_1 \le \frac{\hat{v}'\Sigma\hat{v}}{4} + \frac{16\lambda^2 s}{\Omega} + \frac{\hat{v}'\Sigma\hat{v}}{4}.$$

We now have the combined inequality:

$$\frac{\hat{v}'\Sigma\hat{v}}{2} + \lambda\|\hat{v}\|_1 \le \frac{16\lambda^2 s}{\Omega}$$

which yields the bounds for both the ℓ_1 error as well as the prediction error in Theorem 1. The ℓ_2 error bound is obtained by one more application of the compatibility condition as $\|\hat{v}\|_2^2 \le \hat{v}'\Sigma\hat{v}/\Omega$. □

8.2. *Proof of Theorem* 2. The proof for the sign consistency result of the CoCoLasso is involved. We first present a series of results required to prove Theorem 2.

LEMMA 5. *Let $\partial\|x\|_1$ denotes the sub-gradient of $\|x\|_1$ for any vector x. Then we have the following results*: (a) *$\hat{\beta}$ is the optimal solution to $\tilde{f}(\beta) = (1/2)\beta'\tilde{\Sigma}\beta - \tilde{\rho}'\beta + \lambda|\beta\|_1$ iff there exists a vector $\tilde{u}$ in $\partial\|\hat{\beta}\|_1$ such that*

$$\tilde{\Sigma}\hat{\beta} - \tilde{\rho} + \lambda\tilde{u} = 0. \tag{8.6}$$

(b) *If $|\tilde{u}_j| < 1$ $\forall j \in S^c$, then any other optimal solution $\tilde{\beta}$ will have support $S(\tilde{\beta}) \subseteq S$.* (c) *If we assume that $\tilde{\Sigma}_{S(\hat{\beta}),S(\hat{\beta})}$ is invertible, then under the conditions of part* (b), *$\tilde{f}(\beta)$ has unique minima.*

PROOF. This lemma is a modified version of [28], Lemma 1. We omit the proof as it is exactly analogous to that in the paper. □

Note that the invertibility assumption of part (c) of Lemma 5 needs to hold to establish the uniqueness of the Lasso solution. We now show that this occurs with probability tending to 1. For notational convenience, we define

$$\delta(\varepsilon,\zeta) = p^2C\exp\big(-cns^{-2}\varepsilon^2\zeta^{-1}\big). \tag{8.7}$$

LEMMA 6. $\Pr(\tilde{\Sigma}_{S,S} > 0) \ge 1 - \delta(\varepsilon,\zeta)$ *for all* $\varepsilon \le \min(\varepsilon_0, C_{\min}/2)$.

PROOF. From equation (8.1), we have

$$\begin{aligned}\Lambda_{\min}(\tilde{\Sigma}_{S,S}) &\ge \Lambda_{\min}(\Sigma_{S,S}) - \big|\Lambda_{\max}(-D_{S,S})\big| \ge C_{\min} - \|D_{S,S}\|_2 \\ &\ge C_{\min} - s\|D_{S,S}\|_{\max} \ge C_{\min} - s\|D\|_{\max} \ge C_{\min}/2,\end{aligned}$$

where the last inequality occurs with probability at least $1 - \delta(\varepsilon, \zeta)$ for $\varepsilon \leq \min(\varepsilon_0, C_{\min}/2)$ □

LEMMA 7. *If $\widehat{\Sigma}$ and $\tilde{\rho}$ satisfy* (3.1), *then there exists positive constants C, c such that for every $\varepsilon \leq min(\varepsilon_0, 1/\phi)$,*

$$\begin{aligned} &\Pr\bigl(\|F\|_\infty \geq \varepsilon\phi^2(1-\phi\varepsilon)^{-1}\bigr) \leq \delta(\varepsilon,\zeta), \\ &\Pr\bigl(\|H\|_\infty \geq \varepsilon\phi(2-\gamma)(1-\phi\varepsilon)^{-1}\bigr) \leq \delta(\varepsilon,\zeta). \end{aligned} \tag{8.8}$$

PROOF. Let $\eta_1 = \|D_{S,S}\|_\infty$ and $\eta_2 = \|D_{S^c,S}\|_\infty$. Now, $\sum_{j=1}^{s}|D_{ij}| \leq s\|D\|_{\max}$ for $(i = 1, \dots, s)$. Consequently, if $\|D\|_{\max} \leq \varepsilon/s$ then both η_1 and η_2 are less than ε. From (3.1) and (8.2), $\Pr(\eta_1 \leq \varepsilon, \eta_2 \leq \varepsilon) \geq 1 - \delta(\varepsilon, \zeta)$ for $\varepsilon \leq \varepsilon_0$. The remainder of the proof follows from [18], Lemma A2. □

PROOF OF THEOREM 2 PART (a). We use a Primal Dual Witness construction technique similar to [28] to prove Theorem 2. Let $\hat{\beta}_S$ be the solution to the restricted modified Lasso program, that is,

$$\hat{\beta}_S = \arg\min_{\beta_S} \tilde{f}_S(\beta_S) \qquad \text{where } \tilde{f}_S(\beta_S) = \frac{1}{2}\beta_S'\widetilde{\Sigma}_{S,S}\beta_S - \tilde{\rho}_S'\beta_S + \lambda\|\beta_S\|_1. \tag{8.9}$$

Let $\hat{\beta} = (\hat{\beta}_S', 0_{(p-s)\times 1}')'$ and $\tilde{u} = (\tilde{u}_S', \tilde{u}_{S^c}')'$ where $\tilde{u}_S \in \partial(\|\hat{\beta}_S\|_1)$ and $\tilde{u}_{S^c}$ is some unspecified $(p-s)\times 1$ vector. From part (a) of Lemma 5, we observe that $\hat{\beta}$ is an optimal solution to (2.1) iff $\{\hat{\beta}, \tilde{u}\}$ satisfies

$$\begin{aligned} \widetilde{\Sigma}_{S,S}\hat{\beta}_S - \tilde{\rho}_S + \lambda\tilde{u}_S &= 0, \\ \widetilde{\Sigma}_{S^c,S}\hat{\beta}_S - \tilde{\rho}_{S^c} + \lambda\tilde{u}_{S^c} &= 0. \end{aligned} \tag{8.10}$$

Solving for $\hat{\beta}_S$ and $\tilde{u}_{S^c}$ from equation (8.10), we have

$$\hat{\beta}_S = \widetilde{\Sigma}_{S,S}^{-1}(\tilde{\rho}_S - \lambda\tilde{u}_S), \qquad \tilde{u}_{S^c} = \widetilde{G}\tilde{u}_S + \frac{1}{\lambda}(\tilde{\rho}_{S^c} - \widetilde{G}\tilde{\rho}_S). \tag{8.11}$$

From parts (b) and (c) of Lemma 5, we see that $\hat{\beta}$ will be the unique solution to (2.1) if $\widetilde{\Sigma}_{S,S}$ is nonsingular and all the entries of $\tilde{u}_{S^c}$ have absolute values less than 1. Lemma 6 provides lower bounds for $\Pr(\widetilde{\Sigma}_{S,S} > 0)$. We now derive the bounds for $\Pr(\|\tilde{u}_{S^c}\|_\infty < 1)$. We expand $\tilde{u}_{S^c}$ as

$$\begin{aligned} \tilde{u}_{S^c} &= G\tilde{u}_S + H\tilde{u}_S + \frac{1}{\lambda}\bigl((\tilde{\rho}_{S^c} - \rho_{S^c}) + (\rho_{S^c} - G\rho_S) + G(\rho_S - \tilde{\rho}_S) - H\tilde{\rho}_S\bigr) \\ &= G\tilde{u}_S + H\biggl(\tilde{u}_s + \frac{1}{\lambda}(\rho_S - \tilde{\rho}_S) - \frac{1}{\lambda}\rho_S\biggr) \\ &\quad + \frac{1}{\lambda}\bigl((\tilde{\rho}_{S^c} - \rho_{S^c}) + (\rho_{S^c} - G\rho_S) + G(\rho_S - \tilde{\rho}_S)\bigr). \end{aligned}$$

Taking the absolute values and using triangular inequalities, we have

$$\|\tilde{u}_{S^c}\|_\infty \le \|G\tilde{u}_S\|_\infty + \|H\|_\infty\left(1 + \frac{1}{\lambda}\|\tilde{\rho}_S - \rho_S\|_\infty + \frac{1}{\lambda}\|\rho_S\|_\infty\right) \\ \times \frac{1}{\lambda}\|\rho_{S^c} - G\rho_S\|_\infty + \left(\frac{1}{\lambda}\|\tilde{\rho}_{S^c} - \rho_{S^c}\|_\infty + \frac{1}{\lambda}\|G(\tilde{\rho}_S - \rho_S)\|_\infty\right).$$

We bound each of the four terms on the right-hand side separately. The irrepresentable condition (3.5) implies that $\|G\tilde{u}_S\|_\infty < (1-\gamma)$. It also implies that for $\lambda \le 4\varepsilon_0/\gamma$ we have

$$\Pr\left(\frac{1}{\lambda}\|\tilde{\rho}_{S^c} - \rho_{S^c}\|_\infty + \frac{1}{\lambda}\|G(\tilde{\rho}_S - \rho_S)\|_\infty < \gamma/2\right) \ge \Pr\left(\frac{1}{\lambda}\|\tilde{\rho} - \rho\|_\infty < \gamma/4\right) \\ \ge 1 - \delta(\lambda\gamma, \zeta),$$

where the last inequality follows from taking union bounds on the second equation in (3.1).

The term $(\rho_{S^c} - G\rho_S) = \frac{1}{n}X'_{S^c}(I - X_S(X'_SX_S)^{-1}X'_S)w$ is a linear combination of sub-Gaussian random variables. A direct application of (B.2) yields that $\Pr((1/\lambda)\|\rho_{S^c} - G\rho_S\|_\infty \ge \gamma/4) \le \delta(\lambda\gamma, \zeta)$ where ζ is redefined as maximum of the previous ζ and σ^2.

Without loss of generality, we assume that $\varepsilon_0 \le 1$. Then with probability greater than $1-\delta(\varepsilon,\zeta)$, we can write $\|\tilde{\rho}_S - \rho_S\|_\infty + \|\rho_S\|_\infty \le \|\tilde{\rho}_S - \rho_S\|_\infty + \|\frac{1}{n}X'_Sw\|_\infty + \|\frac{1}{n}X'_SX_S\beta^*_S\|_\infty \le 2 + B\psi$ for $\varepsilon \le \min(1, \varepsilon_0)$. Combining this with Lemma 7, we have, with probability at least $1 - \delta(\varepsilon, \zeta)$:

$$\|H\|_\infty\left(1 + \frac{1}{\lambda}\|\tilde{\rho}_S - \rho_S\|_\infty + \frac{1}{\lambda}\|\rho_S\|_\infty\right) \le \left(1 + \frac{1}{\lambda}(2 + B\psi)\right)\frac{\varepsilon\phi(2-\gamma)}{(1-\phi\varepsilon)} \le \frac{\gamma}{8}$$

for $\varepsilon \le \varepsilon^*_0$ where $\varepsilon^*_0 = \min(\varepsilon_0, \gamma\lambda\phi^{-1}(8(2-\gamma)(\lambda + 2 + B\psi) + \gamma\lambda)^{-1})$.

Combining all the probabilities and adjusting for the invertibility probability, for $\lambda \le 4\varepsilon_0/\gamma$ and $\varepsilon \le \min(\varepsilon^*_0, C_{\min}/2)$, we have $\Pr(\|\tilde{u}_{S^c}\|_\infty \ge 1 - \gamma/8) \le \delta(\lambda\gamma, \zeta) + \delta(\varepsilon, \zeta)$. □

PROOF OF THEOREM 2 PARTS (B) AND (C). Using the expression of $\hat{\beta}_S$ from equation (8.11), we expand

$$\begin{aligned}\hat{\beta}_S - \beta^*_S &= \tilde{\Sigma}^{-1}_{S,S}\left(\tilde{\rho}_S - \rho_S + \frac{1}{n}X'_SX_S\beta^*_S + \frac{1}{n}X'_Sw - \lambda\tilde{u}_S\right) - \beta^*_S \\ &= F_{S,S}\left(\tilde{\rho}_S - \rho_S + \frac{1}{n}X'_SX_S\beta^*_S + \frac{1}{n}X'_Sw\right) \\ &\quad + \Sigma^{-1}_{S,S}(\tilde{\rho}_S - \rho_S) + \frac{1}{n}\Sigma^{-1}_{S,S}X'_Sw - \lambda\tilde{\Sigma}^{-1}_{S,S}\tilde{u}_S.\end{aligned}$$

We analyze each of the terms above separately. From the definition of sub-Gaussian vectors in (B.2), we observe that $\frac{1}{n}\Sigma_{S,S}^{-1}X_S'w$ is sub-Gaussian with parameter at most $\sigma^2 C_{\min}/n$. This implies that $\|\frac{1}{n}\Sigma_{S,S}^{-1}X_S'w\|_\infty$ is less than $\lambda/\sqrt{C_{\min}}$ with probability at least $1-\delta(\lambda,\zeta)$. Moreover, as $\widetilde{\Sigma}=\Sigma+F$, from Lemma 7 we have with probability at least $1-\delta(\varepsilon,\zeta)$, for $\varepsilon\le\min(\varepsilon_0,(2\phi)^{-1})$:

$$\|\widetilde{\Sigma}_{S,S}\|_\infty\le\phi+\|F\|_\infty\le\phi+\phi^2\varepsilon(1-\phi\varepsilon)^{-1}\le 2\phi.$$

The closeness condition for $\tilde{\rho}$ in equation (3.1) implies that $\|\tilde{\rho}_S-\rho_S\|_\infty\le\lambda$ with probability at least $1-\delta(\lambda,\zeta)$ for $\lambda\le\varepsilon_0$. Following the proof of part (a), we can also conclude that for $\varepsilon\le\varepsilon_0$, we have $\|\tilde{\rho}_S-\rho_S\|_\infty+\|\frac{1}{n}X_S'X_S\beta_S^*\|_\infty+\|\frac{1}{n}X_S'w\|_\infty\le(2+B\psi)$ with probability at least $1-\delta(\varepsilon,\zeta)$. Therefore,

$$\left\|F_{S,S}\left(\tilde{\rho}_S-\rho_S+\frac{1}{n}X_S'X_S\beta_S^*+\frac{1}{n}X_S'w\right)\right\|_\infty<(2+B\psi)\frac{\phi^2\varepsilon}{1-\phi\varepsilon}\le\lambda\phi$$

with probability $1-\delta(\varepsilon,\zeta)$ for $\varepsilon\le\lambda\phi^{-1}(\lambda+2+B\psi)^{-1}$. Combining all the probabilities, we have

$$\begin{aligned}\|\hat{\beta}_S-\beta_S^*\|_\infty&\le\left\|F_{S,S}\left(\tilde{\rho}_S-\rho_S+\frac{1}{n}X_S'X_S\beta_S^*+\frac{1}{n}X_S'w\right)\right\|_\infty\\&\quad+\phi\|\tilde{\rho}_S-\rho_S\|_\infty+\left\|\frac{1}{n}\Sigma_{S,S}^{-1}X_S'w\right\|_\infty+2\lambda\phi\\&\le\lambda\left(4\phi+\frac{1}{\sqrt{C_{\min}}}\right)\end{aligned}$$

with probability $1-\delta(\lambda,\zeta)-\delta(\varepsilon,\zeta)$ for $\varepsilon\le(\varepsilon_0, C_{\min}/2,(2\phi)^{-1},\lambda\phi^{-1}(\lambda+2+B\psi)^{-1})$ and $\lambda\le\varepsilon_0$.

This proves part (b). If $|\beta_{\min}^*|>\lambda(4\phi+\frac{1}{\sqrt{C_{\min}}})$, then the Lasso estimate is sign consistent proving Part(c). □

8.3. *Proofs of Lemmas* 1 *and* 2. We assume sub-Gaussian additive or multiplicative measurement errors in Section 4. The proofs of Lemmas 1 and 2 mainly rely on the properties of sub-Gaussian random variables and vectors which can be found in Appendix B.

PROOF OF LEMMA 1. Let $\Sigma_A=(\sigma_{a,ij})$ and b_j denotes the jth column of any matrix B. Then $\widehat{\Sigma}_{\text{add},jk}-\Sigma_{jk}=\frac{1}{n}a_j'x_k+\frac{1}{n}a_k'x_j+(\frac{1}{n}a_j'a_k-\sigma_{a,jk})$. Since $\frac{1}{n}\|x_j\|_2^2=1$ and the entries of a_j are independent and sub-Gaussian with parameter at most τ^2 for all j, property (B.2) implies that $|(1/n)a_j'x_k|$ and $|(1/n)a_j'x_k|$ are each greater than $\varepsilon/3$ with probability less than $C\exp(-cn\varepsilon^2/\tau^2)$. Let $z_i=(a_{ij},a_{ik})'$. Then z_i's are independent sub-Gaussian vectors with parameter at most τ^2. The tail probability for $\frac{1}{n}a_j'a_k-\sigma_{a,jk}$ can now be made small using Lemma B.1. Hence, $\widehat{\Sigma}_{\text{add}}$ satisfies (3.1) with $\zeta=\max(\tau^4,\tau^2)$ and $\varepsilon_0=c\tau^2$.

We observe that $\tilde{\rho}_{\text{add},j} - \rho_j = \frac{1}{n}a'_j X_S \beta^*_S + \frac{1}{n}a'_j w$. Consequently, $|\tilde{\rho}_{\text{add},j} - \rho_j|$ is less than $B\sum_{i=1}^s |\frac{1}{n}a'_j x_i|$ whose tail probability of exceeding $\varepsilon/2$ is at most $C\exp(-n\varepsilon^2 s^{-2}\tau^{-2}B^{-2})$. Letting $z_i = (a_{ij}, w_i)$, Lemma B.1 can be applied to obtain the tail bound for $\frac{1}{n}a'_j w$. Hence, $\tilde{\rho}_{\text{add}}$ satisfies (3.1) with $\zeta = \max(\sigma^4, \tau^4, 1)$ and $\varepsilon_0 = \tau^2$. □

PROOF OF LEMMA 2. The proof once again relies on Lemma B.1. Let $\Sigma_M = (\sigma_{m,jk})$, then

$$\begin{aligned}\widehat{\Sigma}_{\text{mult},jk} - \Sigma_{jk} &= \frac{1}{n}\sum_{i=1}^n \frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(m_{ij}m_{ik} - \mu_j\mu_k - \sigma_{m,jk}) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk}) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(\mu_j(m_{ik} - \mu_k) + \mu_k(m_{ij} - \mu_j)).\end{aligned}$$

Using the regularity conditions in equation (4.1), we have

$$\begin{aligned}|\widehat{\Sigma}_{\text{mult},jk} - \Sigma_{jk}| \le{}& \frac{1}{M_{\min}}\left|(1/n)\sum_{i=1}^n x_{ij}x_{ik}((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk})\right| \\ &+ \frac{\mu_{\max}}{M_{\min}}\left|(1/n)\sum_{i=1}^n x_{ij}x_{ik}(m_{ik} - \mu_k)\right| \qquad (8.12)\\ &+ \frac{\mu_{\max}}{M_{\min}}\left|(1/n)\sum_{i=1}^n x_{ij}x_{ik}(m_{ij} - \mu_j)\right|.\end{aligned}$$

We denote the three terms on the right-hand side of (8.12) by T_1, T_2 and T_3, respectively. Note that, if $v = (v_1, v_2, \dots, v_n)$ where $v_i = x_{ij}x_{jk}$, then $\|v\|_\infty \le X^2_{\max}$. As, the errors are once again sub-Gaussian, using Lemma B.1, we see that for $\zeta = \max(\tau^4 X^4_{\max}/M^2_{\min}, \tau^2 X^2_{\max}\mu^2_{\max}/M^2_{\min})$ and $\varepsilon \le c\tau^2 X^2_{\max}/M_{\min}$ we have

$$\Pr(T_1 \ge \varepsilon) \le C\exp(-cn\varepsilon^2\zeta^{-1}).$$

The terms T_2 and T_3 can be similarly bounded using property (B.2). This proves that $\widehat{\Sigma}_{\text{mult}}$ satisfies (3.1). We now show that $\tilde{\rho}_{\text{mult}}$ also satisfies (3.1). Recall that $\tilde{\rho}_{\text{mult},j} - \rho_j = (1/n)(z_j - \mu_j x_j)'y/\mu_j$. As $y = X_S\beta^*_S + w$, we have

$$\begin{aligned}|\tilde{\rho}_{\text{mult},j} - \rho_j| &\le \frac{1}{\mu_{\min}}\sum_{k=1}^s \left|\frac{1}{n}(z_j - \mu_j x_j)'x_k\beta^*_k\right| + \frac{1}{\mu_{\min}}\left|\frac{1}{n}(z_j - \mu_j x_j)'w\right| \\ &\le \frac{B}{\mu_{\min}}\sum_{k=1}^s\left|(1/n)\sum_{i=1}^n x_{ij}x_{ik}(m_{ij} - \mu_j)\right|\end{aligned}$$

$$+\frac{1}{\mu_{\min}}\left|(1/n)\sum_{i=1}^{n}x_{ij}w_j(m_{ij}-\mu_j)\right|.$$

Using Lemma B.1, we have for $\zeta = X_{\max}^2 \max(\tau^2 B^2/, \tau^4, \sigma^4)/\mu_{\min}^2$ and $\varepsilon \le cX_{\max}\max(\tau^2, \sigma^2)/\mu_{\min}$:

$$\Pr\left(\frac{1}{\mu_{\min}}\left|(1/n)\sum_{i=1}^{n}x_{ij}w_j(m_{ij}-\mu_j)\right| \ge \varepsilon/2\right) \le C\exp(-cn\varepsilon^2\zeta^{-1}),$$

$$\Pr\left(\frac{B}{\mu_{\min}}\left|(1/n)\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ij}-\mu_j)\right| \ge \varepsilon/2s\right) \le C\exp(-cn\varepsilon^2 s^{-2}\zeta^{-1}),$$

where the last inequality follows from property (B.2). Assuming $X_{\max}$, $\mu_{\min}$, $M_{\min}$ and B to be constants, Lemma 2 is proved with $\varepsilon_0 = \tau^2$ and $\zeta = \max(\tau^4, \sigma^4, 1)$. □

APPENDIX A: ALGORITHM FOR FINDING THE NEAREST POSITIVE SEMI-DEFINITE MATRIX

We use an alternating direction method of multipliers to solve for

$$\hat{A} = \mathop{\arg\min}_{A \ge \varepsilon I} \|A - \widehat{\Sigma}\|_{\max} \tag{A.1}$$

for any $\varepsilon > 0$. We introduce an additional variable B and an equality constraint $B = A - \widehat{\Sigma}$ to rewrite the optimization problem in (A.1) as

$$(\hat{A}, \hat{B}) = \mathop{\arg\min}_{A \ge \varepsilon I, B = A - \widehat{\Sigma}} \|B\|_{\max}. \tag{A.2}$$

To solve (A.2), we will minimize the augmented Lagrangian function:

$$f(A, B, \Lambda) = \frac{1}{2}\|B\|_{\max} - \langle \Lambda, A - B - \widehat{\Sigma}\rangle + \frac{1}{2\mu}\|A - B - \widehat{\Sigma}\|_F^2, \tag{A.3}$$

where μ is some penalty parameter, Λ is the Lagrangian matrix and $\langle\cdot,\cdot\rangle$ denotes the matrix inner product which induces the Frobenius norm $\|\cdot\|_F$. We solve for the minimizer of $f(A, B, \Lambda)$ iteratively using the following three steps at the ith iteration:

$$\begin{aligned} &A \text{ step: } A_{i+1} = \mathop{\arg\min}_{A \ge \varepsilon I} f(A, B_i, \Lambda_i),\\ &B \text{ step: } B_{i+1} = \mathop{\arg\min}_{B} f(A_{i+1}, B, \Lambda_i),\\ &\Lambda \text{ step: } \Lambda_{i+1} = \Lambda_i - \frac{A_{i+1} - B_{i+1} - \widehat{\Sigma}}{\mu}. \end{aligned} \tag{A.4}$$

We now provide the closed-form solutions for the first two steps in equation (A.4). The A step can be simplified as

$$\begin{aligned}\arg\min_{A\geq\varepsilon I} f(A, B_i, \Lambda_i) &= \arg\min_{A\geq\varepsilon I} 1/(2\mu)\|A - B_i - \widehat{\Sigma}\|_F^2 - \langle\Lambda_i, A\rangle\\ &= \arg\min_{A\geq\varepsilon I}\|A - B_i - \widehat{\Sigma} - \mu\Lambda_i\|_F^2.\end{aligned}$$

The unconstrained solution for the A-step is $B_i + \widehat{\Sigma} + \mu\Lambda_i$. Let for any symmetric matrix Z, Z_ε denote the projection of Z into the space of matrices with eigen values greater than ε. If $Z = \sum_j \lambda_j p_j p_j'$ denote the spectral decomposition of Z, then we have $Z_\varepsilon = \sum_j \max(\lambda_j, \varepsilon) p_j p_j'$. Hence, the solution for the A-step is given by

$$A_{i+1} = (B_i + \widehat{\Sigma} + \mu\Lambda_i)_\varepsilon. \tag{A.5}$$

The B-step is equivalent to

$$\begin{aligned}&\arg\min_B \frac{1}{2}\|B\|_{\max} + \frac{1}{2\mu}\|B - (A_{i+1} - \widehat{\Sigma})\|_F^2 - \langle -\Lambda_i, B\rangle\\ &\quad= \arg\min_B \|B - (A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i)\|_F^2 + \mu\|B\|_{\max}.\end{aligned} \tag{A.6}$$

Let for any symmetric matrix M, $\mathrm{vec}_L(M)$ denote the vector containing the lower half elements (including the diagonal) of M. Since vec_l is an injective mapping, we can define an inverse mapping $\mathrm{mat}_l(x)$ such that $\mathrm{mat}_l(\mathrm{vec}_l(M)) = M$ for any symmetric matrix M. The solution to the B-step is given by

$$B_{i+1} = \mathrm{mat}_l\big(\mathrm{vec}_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i) - \ell_1\big(\mathrm{vec}_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i), \mu\big)\big),$$

where for any vector x and $\mu > 0$, $\ell_1(x, \mu)$ is the projection of x into the ℓ_1 ball of radius μ. The algorithm to calculate $\ell_1(x, \mu)$ is provided in [9]. The complete details of the ADMM algorithm are given in Algorithm 1.

Algorithm 1 ADMM algorithm for finding the nearest positive semi-definite matrix

```
1: Input $\mu$ and the initial values $B_0$ and $\Lambda_0$
2: At the $i$th step update:
    2.1: (Step $A$) $A_{i+1} = (B_i + \widehat{\Sigma} + \mu\Lambda_i)_\varepsilon$
    2.2: (Step $B$) $B_{i+1} = \mathrm{mat}_l(\mathrm{vec}_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i) - \ell_1(\mathrm{vec}_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i), \mu))$
    2.3: (Step $\Lambda$) $\Lambda_{i+1} = \Lambda_i - \frac{A_{i+1} - B_{i+1} - \widehat{\Sigma}}{\mu}$
3: Repeat Step 2 till convergence
```

APPENDIX B: SUB-GAUSSIAN RANDOM VARIABLES

In our analysis of the CoCoLasso estimate, we have assumed that the errors w are independent and identically distributed sub-Gaussian random variables with parameter τ^2. In this section, we summarize some useful definitions and properties of sub-Gaussian random variables.

DEFINITION B.1 (Sub-Gaussian random variables Vershynin [27]). A random variable Z is said to be sub-Gaussian if there exists a finite $\kappa > 0$ such that $\kappa = \sup_{p\geq 1} p^{-1/2}(E|X|^p)^{\frac{1}{p}}$. κ is referred to as the sub-Gaussian norm of Z denoted by $\|Z\|_\phi$.

Equivalently, a sub-Gaussian random variable Z satisfies the following tail probability bounds:

$$P(|Z| > t) \leq 2\exp(-t^2/2\tau^2) \qquad \text{for all } t > 0. \tag{B.1}$$

To avoid ambiguity, we refer to the sub-Gaussian parameter of Z as the smallest τ^2 satisfying (B.1). Following [27], Lemma 5.5, we observe that there exists universal constants m and M such that $m\|Z\|_\phi^2 \leq \tau^2 \leq M\|Z\|_\phi^2$. We note that if $w = (w_1, w_2, \dots, w_n)'$ that w_i's are independent zero-centered sub-Gaussian random variables, then weighted sums of w_i are also sub-Gaussian and satisfy an useful property [27], Lemma 5.9:

$$\|v'w\|_\phi^2 \leq K\|v\|_2^2 \max_i(\|w_i\|_\phi^2), \tag{B.2}$$

where K is an absolute constant. The tail-probability characterization in (B.1) enables defining sub-Gaussian random vectors in the following sense.

DEFINITION B.2 (Sub-Gaussian random vectors Cai, Zhang and Zhou [7]). A random vector w is said to be sub-Gaussian if there exists $\tau > 0$ such that $\Pr(|v'(w - E(w))| > t) \leq 2\exp(-\frac{t^2}{2\tau^2})$ for all $t > 0$ and $\|v\|_2 = 1$.

From property (B.2), it is clear that if $w = (w_1, w_2, \dots, w_n)'$ is a sub-Gaussian vector with parameter τ^2, then each w_i is also sub-Gaussian with parameter at most τ^2. Conversely, if w_i's are independent and sub-Gaussian random variables with parameter τ_i^2, then $w = (w_1, w_2, \dots, w_n)$ is a sub-Gaussian vector with parameter at most $\tau^2 \leq (KM/m)(\max \tau_i^2)$. We now state and prove another useful result for correlated sub-Gaussian sequences.

LEMMA B.1. *Let $z_i = (x_i, y_i)'$ denote independent and identically distributed vectors with zero mean, covariance $\Sigma = ((\sigma_{ij}))$ and sub-Gaussian parameter τ^2. Then there exists absolute constants C and c such that, for every $\varepsilon \leq c\tau^2\|a\|_\infty$, we have*

$$\Pr\left(\frac{1}{n}\left|\sum_{i=1}^n a_i(x_i y_i - \sigma_{12})\right| \geq \varepsilon\right) \leq C\exp\left(-\frac{nc\varepsilon^2}{\tau^4\|a\|_\infty^2}\right). \tag{B.3}$$

PROOF.

$$
\begin{aligned}
(1/n)\sum_{i=1}^{n} a_i(x_i y_i - \sigma_{12})
&= \frac{1}{4n}\sum_{i=1}^{n} a_i\bigl((x_i + y_i)^2 - (\sigma_{11} + \sigma_{22} + 2\sigma_{12})\bigr) \\
&\quad - \frac{1}{4n}\sum_{i=1}^{n} a_i\bigl((x_i - y_i)^2 - (\sigma_{11} + \sigma_{22} - 2\sigma_{12})\bigr) \\
&= \frac{1}{2n}\sum_{i=1}^{n} a_i\bigl((v_1' z_i)^2 - E\bigl((v_1' z_1)^2\bigr)\bigr) - \frac{1}{2n}\sum_{i=1}^{n} a_i\bigl((v_2' z_i)^2 - E\bigl((v_2' z_1)^2\bigr)\bigr),
\end{aligned}
$$

where $v_1 = (1/\sqrt{2}, 1/\sqrt{2})'$ and $v_1 = (1/\sqrt{2}, -1/\sqrt{2})'$. As $\|v_k\| = 1$, $v_k' z_1$ is sub-Gaussian with parameter at most τ^2 for $k = 1, 2$. Using the relationship between sub-Gaussian and sub-exponential random variables in [27], Lemma 5.14 and Remark 5.18, we see that, for $k = 1, 2$, $(v_k' z_i)^2 - E((v_k' z_1)^2)$ is sub-exponential with parameter at most $c\tau^2$ where c is an absolute constant. As a result $t_i = a_i((v_1' z_i)^2 - E((v_1' z_1)^2)$ is sub-exponential with parameter at most $c\tau^2\|a\|_\infty$. A direct application of [27], Corollary 5.17, now yields for $\varepsilon \le c\tau^2\|a\|_\infty$:

$$
\Pr\biggl(\frac{1}{2n}\biggl|\sum_{i=1}^{n} t_i\biggr| \ge \varepsilon\biggr) \le C\exp\biggl(-\frac{nc\varepsilon^2}{\tau^4\|a\|_\infty^2}\biggr). \qquad \square
$$

Acknowledgment. We thank the Associate Editor and two referees for their helpful and constructive comments that greatly improved this work. The authors thank Dr. Po-Ling Loh for sharing R and Matlab codes for computing the NCL estimator.

REFERENCES

[1] BELLONI, A., ROSENBAUM, M. and TSYBAKOV, A. B. (2014). Linear and conic programming estimators in high-dimensional errors-in-variables models. Preprint. Available at arXiv:1408.0241.

[2] BELLONI, A., ROSENBAUM, M. and TSYBAKOV, A. B. (2016). An $\{\ell_1, \ell_2, \ell_\infty\}$-regularization approach to high-dimensional errors-in-variables models. *Electron. J. Stat.* **10** 1729–1750. MR3522659

[3] BENJAMINI, Y. and SPEED, T. P. (2012). Estimation and correction for GC-content bias in high throughput sequencing. *Nucleic Acids Res.* **40** 72.

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

[6] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

[7] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann*. *Statist*. **38** 2118–2144.

[8] CANDÈS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann*. *Statist*. **35** 2313–2351.

[9] DUCHI, J., SHALEV-SHWARTZ, S., SINGER, Y. and CHANDRA, T. (2008). Efficient projections onto the L1-ball for learning in high dimensions. In *Proceedings of the* 25*th International Conference on Machine Learning* (*ICML* '08) 272–279. ACM, New York.

[10] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann*. *Statist*. **32** 407–499.

[11] EFRON, B., HASTIE, T. and TIBSHIRANI, R. (2007). Discussion: "The Dantzig selector: Statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; MR2382644] by E. Candes and T. Tao. *Ann*. *Statist*. **35** 2358–2364. MR2382646

[12] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J*. *Amer*. *Statist*. *Assoc*. **96** 1348–1360. MR1946581

[13] FAN, J. and LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *International Congress of Mathematicians*. *Vol*. *III* 595–622. Eur. Math. Soc., Zürich. MR2275698

[14] FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist*. *Sinica* **20** 101–148. MR2640659

[15] FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J*. *Stat*. *Softw*. **33** 1–22.

[16] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2011). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York.

[17] LOH, P. L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Ann*. *Statist*. **40** 1637–1664.

[18] MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42.

[19] PURDOM, E. and HOLMES, S. P. (2005). Error distribution for gene expression data. *Stat*. *Appl*. *Genet*. *Mol*. *Biol*. **4** Art. 16, 35. MR2170432

[20] ROSENBAUM, M. and TSYBAKOV, A. B. (2010). Sparse recovery under matrix uncertainty. *Ann*. *Statist*. **38** 2620–2651.

[21] ROSENBAUM, M. and TSYBAKOV, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back*: *High-Dimensional Models and Processes*. *Inst*. *Math*. *Stat*. (*IMS*) *Collect*. **9** 276–290. IMS, Beachwood, OH. MR3202640

[22] SLIJEPCEVIC, S., MEGERIAN, S. and POTKONJAK, M. (2002). Location errors in wireless embedded sensor networks: Sources, models, and effects on applications. *Mob*. *Comput*. *Commun*. *Rev*. **6** 67–78.

[23] SØRENSEN, Ø., FRIGESSI, A. and THORESEN, M. (2013). Measurement error in LASSO: Impact and likelihood bias correction. *Statist*. *Sinica* **23**. To appear.

[24] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J*. *R*. *Stat*. *Soc*., *B* **58** 267–288. MR1379242

[25] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **67** 91–108. MR2136641

[26] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron*. *J*. *Stat*. **3** 1360–1392. MR2576316

[27] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[28] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[29] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[30] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[31] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENT OF BIOSTATISTICS
JOHNS HOPKINS UNIVERSITY
615 N. WOLFE STREET
BALTIMORE, MARYLAND 21205
USA
E-MAIL: abhidatta@jhu.edu

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
224 CHURCH STREET S.E.
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: zouxx019@umn.edu