Two Stage Non-penalized Corrected Least Squares for High Dimensional Linear Models with Measurement error or Missing Covariates

Abhishek Kaul[1], Hira L. Koul[2], Akshita Chawla and Soumendra N. Lahiri[3]

NIEHS, Michigan State University, Merck Research Laboratories and North Carolina State University

Abstract

In this paper we propose estimation via bias corrected least squares after model selection for estimation and variable selection in high dimensional linear regression models with measurement error or missing covariates. We show that separating model selection and estimation leads to an improved rate of convergence of the ℓ_2 error compared to the rate $\sqrt{s \log p/n}$ achieved by simultaneous estimation and variable selection methods such as ℓ_1 penalized corrected least squares. If the correct model is selected with high probability then the ℓ_2 rate of convergence for the proposed method is indeed the oracle rate of $\sqrt{s/n}$. Here s, p, n are the number of nonzero parameters, model dimension and sample size respectively. Under general model selection criteria, the proposed method is computationally simpler and statistically at least as efficient as the ℓ_1 penalized corrected least squares method, performs model selection without the availability of the bias correction matrix, and is able to provide estimates with only a small sub-block of the bias correction covariance matrix of order $s \times s$ in comparison to the $p \times p$ correction matrix required for computation of the ℓ_1 penalized version. Furthermore we show that the model selection requirements are met by a correlation screening type method and the ℓ_1 penalized corrected least squares method. Also, the proposed methodology when applied to the estimation of precision matrices with missing observations, is seen to perform at least as well as existing ℓ_1 penalty based methods. All results are supported empirically by a simulation study.

Keywords: High Dimension, Measurement Error, Missing Data.

1 Introduction

Linear regression models with noisy or missing covariates are abound in variety of scientific fields including econometrics, epidemiology and finance. Particular examples of such data include the human microbiome expression data measuring relative abundances of bacteria in the human body, which is often observed only partially, i.e., with several missing observations and gene expression data that are often corrupted with noise or missing values. It is well known that ignoring this measurement error or missing-ness leads to biased parameter estimates, see, e.g., Carroll, Ruppert, Stefansky and Crainiceanu (2006) and Fuller (1987).

[1]Corresponding author. E-mail: abhishek.kaul@nih.gov, Research supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744- 04)

[2]Research supported in part by the NSF DMS grant 1205271

[3]Research supported in part by the NSF DMS grant 1310068.

In the high dimensional setting where the number of parameters may vastly exceed the sample size, several authors including Liang and Li (2009), Loh and Wainwright (2012), Sørensen, Thoresen and Frigessi (2014), and Kaul and Koul (2015), have studied estimators for these models. The common thread of these papers being minimization of an appropriate bias corrected loss function penalized by the ℓ_1 norm of the parameter vector of interest. This approach provides consistent estimates that are also computationally efficient. However, defining the bias corrected loss function in fact requires a bias correction matrix which is typically estimated from data. This matrix being itself high dimensional makes its estimation and thus the implementation of existing methods challenging, if not infeasible.

In this paper we propose a two step estimator for these models and analyse its efficiency in model selection and the rate of ℓ_2 error in estimation. By separating model selection and estimation, it is possible to improve upon the rate of ℓ_2 error in estimation, compared to ℓ_1 penalized methods. Furthermore, our methodology requires only a small sub-block of the bias correction matrix. Thus providing more accurate estimates with lesser information input in comparison to ℓ_1 penalized methods. It is important to note that our error bounds include the uncertainty due to model selection, i.e., although the model selection and estimation are performed in separate steps however the final error bounds remain valid with an associated probability that includes the uncertainty due to model selection.

Loh and Wainwright (2012) show that ℓ_1 penalized corrected least squares method achieves the rate $\sqrt{s \log p/n}$ of the ℓ_2 estimation error, under appropriate conditions. They also empirically show that this rate is optimal. Here p is the dimension of the parameter vector, s represents the number of non zero mean parameters in the model and n is the sample size. In comparison, our two stage methodology enjoys three major advantages. First, the possibility of performing model selection without the availability of the bias correction matrix. Second, being able to provide estimates with only a small sub-block of the bias correction matrix. Lastly, provided one has a reasonable control on the number of incorrectly identified regressors ($\hat{m}$), i.e., provided $\hat{m} = O_P(s)$, the proposed method performs at least as well as ℓ_1- penalized methods. In addition, if the correct model is selected from the first step with probability (w.p.) converging to 1, then the rate of convergence of the ℓ_2-error for the proposed method is shown to be indeed the optimal rate of $\sqrt{s/n}$. We also apply the methodology developed to the problem of precision matrix estimation with observations corrupted with missing values and similarly show that the estimates thus obtained are more efficient in comparison to its ℓ_1 penalized counterpart.

To the best of our knowledge, such two stage refitting procedures were first introduced by Candes and Tao (2007) in the context of Dantzig selector for high dimensional classical linear regression where X is fully observed, and have been investigated by Belloni and Chernuzhokov (2013) with least squares loss again in the linear regression setup without measurement error. In particular, the latter provide a rigorous analysis of the rate of convergence of the ℓ_2 error for the two stage refitting procedure.

Finally, we perform a series of simulated experiments to confirm our theoretical findings. We show empirically that in addition to having higher efficiency in estimation, our methodology provides more accurate model identification compared to the ℓ_1 penalized counterpart and is also computationally faster for larger data sets.

The rest of this paper is organized as follows. Section 2 describes the model under consideration and introduces the notation required for the analysis. Section 3 describes the first step model selection procedure and investigates some theoretical properties of the two possible methods, which can be used to achieve this goal consistently. Section 4 provides some theoretical properties of the second step estimation procedure and describes the associated rates of convergence of estimation error. We then provide an algorithm for precision matrix estimation with observations corrupted with missing data. Section 5 provides a series of simulated experiments. All proofs are relegated to the appendix.

2 Model Setup

We begin by describing the models under consideration. Let $x_i = (x_{i1}, \cdots, x_{ip})', i = 1, \cdots, n$, be vectors of random design variables, where for any vector a, a' denotes its transpose. Let y_i's denote the responses, which are related to x_i's by the relations

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad \text{for some } \beta_0 = (\beta_{01}, ..., \beta_{0p})' \in \mathbb{R}^p, \ 1 \le i \le n. \tag{2.1}$$

Here β_0 is the parameter vector of interest, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ is an $n-$dimensional vector whose components are i.i.d. Gaussian random variables (r.v.'s) with variance σ_ε^2, i.e., $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0, \sigma_\varepsilon^2)$, $1 \le i \le n$. Furthermore, the design variables x_i's are not observed directly. Instead, we observe surrogates z_i, $1 \le i \le n$, obeying one of the following two models.

Additive noise:

$$z_i = x_i + w_i, \quad 1 \le i \le n. \tag{2.2}$$

The covariate noise vectors $w_i = (w_{i1}, .., w_{ip})'$, $1 \le i \le n$, are assumed to be i.i.d. r.v.'s. Furthermore, w_i, x_i, and ε_i, $1 \le i \le n$, are assumed to be mutually independent.

Missing covariates:

$$z_i = x_i \oplus w_i, \quad 1 \le i \le n. \tag{2.3}$$

Here $\oplus$ represents componentwise product and $w_i = (w_{i1}, ..., w_{ip})'$, with the components $\{w_{ij}, 1 \le i \le n\} \sim_{i.i.d.}$ Bernoulli$(1 - \rho_j)$, $1 \le j \le p$.

Let $X = (x_1, ..., x_n)'$ be the unobserved $n \times p$ design matrix and similarly define the $n \times p$ matrices Z, W with the corresponding vectors. For the case of additive noise, the random matrices X and W are assumed to be sub-Gaussian as defined by Loh and Wainwright (2012). This definition is restated below for the convenience of the reader.

Definition 2.1 (sub-Gaussian matrices) We say that a random matrix $X \in \mathbb{R}^{n \times p}$ is sub-Gaussian with parameters (Σ_x, σ_x^2) if the following two conditions hold.

1. Each row $x_i' \in \mathbb{R}^p$ of X is sampled independently from a zero-mean distribution with covariance Σ_x, $1 \le i \le n$.

2. For any unit vector $\delta \in \mathbb{R}^p$ the random variable $\delta' x_i$ is sub-Gaussian in the usual univariate sense with parameter at most σ_x^2, $1 \leq i \leq n$.

Remark 2.1 An elementary property of sub-Gaussianity and X and W being sub-Gaussian imply that in the case of additive noise, Z is also sub-Gaussian. Also, Loh and Wainwright (2012) show as part of the proof of Lemma 4 of their supplement that for the case of missing covariates, the random matrix Z is also sub-Gaussian with parameter σ_x^2, i.e., with the same parameter as for the unobserved sub-Gaussian random matrix X.

3 Notation, Assumptions and Conventions

The parameters p and s are assumed to diverge with the sample size n, however this dependence is suppressed for clarity of the exposition. For the same reason we do not exhibit the dependence of the arrays of x_i's and z_i's on n. For any vector $\delta \in \mathbb{R}^p$, define the support of δ as $\text{supp}(\delta) = \big\{j \in \{1, 2, ..., p\}; \delta_j \neq 0.\big\}$. The ℓ_2 norm of δ is denoted by $\|\cdot\|_2$ and $|\delta|$ shall denote the componentwise absolute value vector. For any two collection of indices S and $\tilde{S}$, we represent $\tilde{S} - S$ as the collection of indices in $\tilde{S}$ but not in S. The cardinality of an index set S will be denoted by either $\text{card}(S)$, and $\|\delta\|_0 := \text{card}\big(\text{supp}(\delta)\big)$. For any two sequences $\{a_n\}$ and $\{b_n\}$ of real numbers, $a_n \preceq b_n$ means that for some constant $0 < c_0 < \infty$, $a_n \leq c_0 b_n$, for n large enough. Similarly, $a_n \preceq_P b_n$ shall denote that $a_n \preceq b_n$ in probability. For matrices M_1 and M_2 we denote $M_1 \oplus M_2$ and $M_1 \ominus M_2$ as the component wise product and division, respectively. For a subset $A \subseteq \{1, 2, \cdots, p\}$, b_A denote the vector of components of b with indices in A. Also, all limits are taken as $n \to \infty$, unless mentioned otherwise. Lastly, $0 < c_0, c_1, c_2 < \infty$, and $0 < c_3 < 1$ shall denote generic constants that may be different in different contexts.

In the above setup we shall consider the model (2.1) in the high dimensional setting where the dimension p of β_0 is allowed to grow exponentially with n. In addition β_0 is assumed to be sparse, i.e., only a small proportion of the parameters are assumed to be non zero. In the sequel,

$$T = \text{supp}(\beta_0), \qquad T^c \text{ denote its compliment set.}$$

By definition, $\text{card}(T) = s$.

Decompose $\beta_0 = \big(\beta_{0T}', \beta_{0T^c}'\big)'$ into its non zero and zero components, and similarly partition $n \times p$ matrices X and Z into columns corresponding to the indices of β_0, i.e., $X = \big(X_T, X_{T^c}\big)$, $Z = \big(Z_T, Z_{T^c}\big)$. Also a $p \times p$ matrix Σ is partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TT^c} \\ \Sigma_{T^cT} & \Sigma_{T^cT^c} \end{pmatrix}. \tag{3.1}$$

Throughout, the parameters s, p and n are assumed to satisfy

$$s \log p / n = o(1).$$

Define

$$\hat{\gamma}^{\text{add}} = n^{-1}Z'y \qquad \text{and} \qquad \hat{\gamma}^{\text{miss}} = n^{-1}Z'y \ominus (\mathbf{1} - \boldsymbol{\rho}), \tag{3.2}$$

where $\mathbf{1}$ is a p-dimensional vector of ones and $\boldsymbol{\rho} = (\rho_1, .., \rho_p)'$. The entities $\hat{\gamma}^{\text{add}}$ and $\hat{\gamma}^{\text{miss}}$ serve as measures of correlation between X and y in the additive error and missing covariates cases, respectively. Also define,

$$\Gamma^{\text{add}} = n^{-1}Z'Z - \Sigma_w, \quad \text{and} \quad \Gamma^{\text{miss}} = n^{-1}Z'Z \ominus M, \tag{3.3}$$

for the additive error and missing covariate cases, respectively. Here, $M = \left[M_{ij}\right]_{i,j=1,\ldots,p}$ is a $p \times p$ matrix with

$$M_{i,j} = \begin{cases} (1-\rho_i)(1-\rho_j)\ ; & i \neq j \\ (1-\rho_i)\ ; & i = j. \end{cases} \tag{3.4}$$

Next, we state the needed assumptions.

Assumptions:

(A1) Additive errors: In the model (2.2), the measurement error matrix $W = (w_1, ..., w_n)'$ is assumed to be sub-Gaussian as defined in (2.1) and w_i, x_i and ε_i are assumed to be mutually independent for all $1 \leq i \leq n$.

(A2) Missing covariates: The components of the vector w_i in the model (2.3) are such that $\{w_{ij}, 1 \leq i \leq n\}$ are i.i.d. Bernoulli$(1-\rho_j), 1 \leq j \leq p$. Also assume that $0 \leq \rho_{\max} := \max\{\rho_j\,;\, 1 \leq j \leq p\} < 1$. Furthermore w_i's are mutually independent of x_i and ε_i, for all $1 \leq i \leq n$.

Unobserved design variables X:

(A3) Assume that the covariance matrix of X satisfies the following conditions, where part (i) is for additive errors and part (ii) is for missing covariates, and where ρ_{max} is as in (A2).

$$\text{(i)} \quad \max|\Sigma^x_{T^cT}\beta^0_T| + 2\frac{\sigma_z}{c_0}(\sigma_\varepsilon + \sigma_x\|\beta_0\|_2)\sqrt{\frac{c_1 \log p}{n}} < \min|\Sigma^x_{TT}\beta^0_T|.$$

$$\text{(ii)} \quad \max|\Sigma^x_{T^cT}\beta^0_T| + 2\frac{\sigma_x}{c_0(1-\rho_{\max})}(\sigma_\varepsilon + \sigma_x\|\beta_0\|_2)\sqrt{\frac{c_1 \log p}{n}} < \min|\Sigma^x_{TT}\beta^0_T|.$$

This assumption is similar to Condition F of Genovese et al. (2012) and is also reminiscent of the 'faithfulness condition' of Bühlmann, Kalisch and Maathuis (2009). In the noiseless setting, it is necessary and sufficient for exact recovery of the support of β_0, (Thm. 2, Genovese et al. 2012).

Random matrices Γ^{add} and Γ^{miss}:

RE: A matrix Γ is said to satisfy the lower restricted eigenvalue condition with curvature $\alpha_1 > 0$ and tolerance $\tau > 0$ if

$$\delta'\Gamma\delta \geq \alpha_1\|\delta\|_2^2 - \tau\|\delta\|_1^2, \qquad \text{for all } \delta \in \mathbb{R}^p. \tag{3.5}$$

RSE(k_n)**:** For any $m \leq k_n$, a matrix Γ is said to satisfy a lower and upper restricted sparse eigenvalue condition with constants $\kappa(m), \phi(m) > 0$, respectively, if

$$\kappa(m) := \inf_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Gamma \delta}{\|\delta\|_2^2} > 0, \quad \phi(m) := \sup_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Gamma \delta}{\|\delta\|_2^2} < \infty. \tag{3.6}$$

Assumption **RE** was introduced by Loh and Wainwright (2012). They prove that this condition holds for Γ^{add} and Γ^{miss}, with asymptotic probability 1 with appropriate choices of α_1 and τ.

Assumption **RSE** controls the minimum and maximum eigenvalues of certain sub-blocks of the matrix Γ. Lemma 3.1 below shows that this condition is satisfied by the matrices Γ^{add} and Γ^{miss} with asymptotic probability 1.

(A4) Parameter vector β_0**:** The minimum magnitude of the components of β_0 satisfies $\min_{j \in T} |\beta_{0j}| \succeq \|\beta_0\|_2 \, s \log p / n$.

The following lemma shows that the random matrices Γ^{add} and Γ^{miss} satisfy the condition **RSE** with asymptotic probability 1, under suitable assumptions. Let

$$\kappa_x(m) := \inf_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Sigma_x \delta}{\|\delta\|_2^2}, \qquad \phi_x(m) := \sup_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Sigma_x \delta}{\|\delta\|_2^2}.$$

Lemma 3.1 (Plausibility of RSE). *Let k_n be any positive sequence satisfying $k_n \log p = o(n)$. Suppose condition* **(A1)** *for the additive error model or condition* **(A2)** *for the missing covariate model hold. Also assume that some constants κ_x and ϕ_x,*

$$0 < \kappa_x \leq \kappa_x(m) \leq \phi_x(m) \leq \phi_x < \infty, \quad \textit{for all } m \leq k_n. \tag{3.7}$$

Then, with $\Gamma = \Gamma^{add}$ or $\Gamma = \Gamma^{miss}$, the following conditions

$$\kappa(m) := \inf_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Gamma \delta}{\|\delta\|_2^2} > 0, \quad \phi(m) := \sup_{\|\delta_{T^c}\|_0 \leq m,\, \delta \neq 0} \frac{\delta' \Gamma \delta}{\|\delta\|_2^2} < \infty,$$

hold uniformly over any $m \leq k_n$ with $\kappa(m) \geq \kappa_x/2$ and $\phi(m) \leq 2\phi_x$, w.p. at least $1 - 2c_3 \exp(-s)/(1 - 1/e)$, for all sufficiently large n.

This lemma shows that for any positive sequence k_n satisfying $k_n \log p = o(n)$, condition **RSE**(k_n) is satisfied by Γ^{add} and Γ^{miss}, with the lower and upper restricted eigenvalues $\kappa(m)$ and $\phi(m)$ being bounded below and above, respectively, for large n, with high probability. This lemma shall play a useful role in the development of the methodology to follow.

4 Step 1: Model Selection

The objective of this first step is to recover the support T of the parameter vector β_0 from the observed variables Z and y. In the sequel $\hat{T}$ denotes the estimate of the support T of β_0

given by the model selection procedure and $\hat{m}$ denotes the number of noise variables selected, i.e.,

$$\hat{m} = \text{card}(\hat{T} - T).$$

We propose the following two possible methods for selecting $\hat{T}$.

CS Screen the corrected absolute correlation vector $\left|\hat{\gamma}^{\text{add}}\right|$ or $\left|\hat{\gamma}^{\text{miss}}\right|$ to select a certain number of indices that are largest in magnitude. The intuition behind this is the same as that of the sure independence screening proposed by Fan and Lv (2008). To see this equivalence for the additive error case notice that $Z'y = X'y + W'y$. Now, by assumption W is independent of y, thus the correlation structure of $n^{-1}Z'y$ will asymptotically be the same as that of $n^{-1}X'y$. These ideas are made rigorous below.

ℓ_1-**CLS** Use ℓ_1 penalized bias corrected least squares as proposed by Loh and Wainwright (2012) to select the indices of the non zero estimates. This estimator is defined as

$$\hat{\beta} = \underset{\|\beta\|_1 \leq b_0\sqrt{s}}{\arg\min} \left\{ \hat{Q}_n(\beta) + \lambda_n \|\beta\|_1 \right\}, \qquad \lambda_n > 0. \tag{4.1}$$

where b_0 is a suitably chosen constant and

$$\hat{Q}_n(\beta) := \frac{1}{2}\beta'\Gamma\beta - \hat{\gamma}'\beta, \tag{4.2}$$

where Γ and $\hat{\gamma}$ are chosen as the corresponding versions in the additive errors or the missing covariates cases. The selected model is $\hat{T} = \text{supp}(\hat{\beta})$.

We begin with the analysis of the **CS** method. Consider the absolute value of the correlation vector $|\hat{\gamma}| := (|\hat{\gamma}_1|, ..., |\hat{\gamma}_p|)'$ defined in (3.2) between the observed variable Z and y, and let $r(\hat{\gamma}) = (r_1(|\hat{\gamma}|), ..., r_p(|\hat{\gamma}|))'$ denote the vector of descending ranks of the components of the vector $\hat{\gamma}$, where rank one signifies the highest magnitude. Then the **CS** method estimates the set of non zero indices by

$$\hat{T}(a_n) = \hat{T}(a_n, Z, y, p) = \{j\,;\, r_j(\hat{\gamma}) \leq a_n,\ 1 \leq j \leq p\}, \tag{4.3}$$

where a_n is a known sequence of positive numbers such that $a_n/s \leq c$, for some constant $c \geq 1$. The following theorem shows that this procedure identifies the support of the parameter vector along with providing a reasonable control on the false positives.

Theorem 4.1 *If either conditions* **(A1)** *and* **(A3i)** *hold for the case of additive errors (2.2) or conditions* **(A2)** *and* **(A3ii)** *hold for the case of missing covariates (2.3), then the estimated set of non zero indices $\hat{T}(a_n)$ of (4.3) satisfies the following.*

$$\textit{(i)}\ \ T \subseteq \hat{T}(a_n), \qquad \textit{(ii)}\ \ \hat{m} \preceq s,$$

w.p. at least $1 - c_1 \exp(-c_2 \log p)$, for all sufficiently large n.

Remark 4.1 A closer look at the proof of Theorem 4.1 shows that if the cardinality of the set T is known and we let $a_n \to s$ in (4.3), then $P\big(\hat{T} = T\big) \to 1$.

In view of Theorem 4.1, choosing a_n appropriately leads to identification of the support of the parameter vector along with a control on the false positives. However, the choice of this thresholding level a_n is determined by the number of non-zero components s, which in practice is unknown. Thus, as is the case with ℓ_1 penalized methods, we shall treat a_n as a tuning parameter and provide a data based strategy to optimally choose this parameter in Section 6.

As stated earlier, the implementation of this method does not require the knowledge of the matrix Σ_w or M. This is especially useful in the case of additive errors where Σ_w is unknown, since we by-pass estimating a p dimensional Σ_w from a very low number or typically available replicates of Z. In addition, this method comes at a cheap computational cost.

Next, we proceed to the $\ell_1 -$ **CLS** method for model selection. Before proceeding, a point of caution here is that this method is not useful for the case of additive errors due to the unavailability of Σ_w. On the other hand, for the case of missing covariates we can estimate ρ_j for all $1 \le j \le p$ by the empirical average of the number of observed entries per column of Z. This in turn enables us to estimate the matrix M and to implement ℓ_1-penalized bias corrected least squares in this case. Thus, the analysis to follow shall focus on model selection by ℓ_1-**CLS** method only for the case of missing covariates.

Another technical reason for not using ℓ_1-**CLS** in the case of additive errors is the non convexity of the loss function $\hat{Q}_n(\beta)$. In comparison, $\hat{Q}_n(\beta)$ is convex in the case of missing covariates, which plays a key role proving the desired model selection property of this methodology.

We begin with the following additional assumption. For some $r > 0$,

$$\|\Gamma^{\text{miss}}\beta_0 - \hat{\gamma}^{\text{miss}}\|_\infty \le r\|\beta_0\|_2\sqrt{\log p}/\sqrt{n}. \tag{4.4}$$

Then we have the following model selection result.

Theorem 4.2 *In addition to (2.1), (2.3), and (A4), suppose the conditions lower-***RE** *and (4.4) hold for Γ^{miss} and $\hat{\gamma}^{miss}$. Then for the method ℓ_1-***CLS** *with $\lambda_n \ge 4r\|\beta_0\|_2\sqrt{\log p}/\sqrt{n}$ in (4.1), where r is as in (4.4), we have, with $\hat{T} = supp(\hat{\beta})$,*

$$\textit{(i)}\ \ T \subseteq \hat{T}, \qquad\qquad \textit{(ii)}\ \ \sqrt{\hat{m}} \le c_0\phi(\hat{m})\sqrt{s}/\alpha_1.$$

Remark 4.2 The result of Theorem 4.2 is not accompanied by a probabilistic statement since this result follows by deterministic arguments on the event where the required assumptions hold. In addition Loh and Wainwright (2012) (Theorem 1 and Corollary 2) show that that the conditions lower-**RE** and (4.4) hold for Γ^{miss} and the pair $(\Gamma^{\text{miss}}, \hat{\gamma}^{\text{miss}})$, respectively, w.p. at least $1 - c_1\exp(-c_2\log p)$, with

$$\alpha_1 = \lambda_{\min}(\Sigma_x)/2 \quad \text{and} \quad r = c_0\frac{\sigma_x}{1-\rho_{\max}}\big(\sigma_\varepsilon + \frac{\sigma_x}{1-\rho_{\max}}\big),$$

where $\lambda_{\min}(\Sigma_x)$ represents the minimum eigenvalue of the matrix Σ_x.

Remark 4.3 Recall that for the case of missing covariates, $\hat{Q}_n(\beta)$ is convex, and hence, by standard results via first order optimality conditions, $\hat{m} \leq n$, see, e.g., Lemma 5 of Tibshirani (2013). Thus Theorem 4.2 immediately implies that with high probability, $\hat{m} \preceq \phi(n)s$. However, this bound is not sharp since $\phi(n)$ may diverge with n. From here, following the strategy of Belloni and Chernozhukov (2013), we extend the result to obtain the bound $\hat{m} \preceq s$ under an additional assumption. This will be implied by the following lemma.

Lemma 4.1 *Under the conditions of Theorem 4.2,*

$$\hat{m} \leq \frac{c_0}{\alpha_1} s \Big[\min_{m \in \mathcal{M}} \phi(m \wedge n) \Big], \tag{4.5}$$

where $\mathcal{M} = \Big\{ m \in \mathbf{N} \,;\, m > \frac{2c_0}{\alpha_1} s\phi\big(m \wedge n\big) \Big\}$.

This lemma is a consequence of Theorem 4.2 and Lemma 2 of Belloni and Chernozhukov (2013) and thus the short proof is omitted. For details see, page 14 of Belloni and Chernozhukov (2013). As a consequence of this lemma, under the additional assumption

$$\min_{m \in \mathcal{M}} \phi(m \wedge n) \leq c_0, \tag{4.6}$$

$\hat{m} \preceq s$. This result, together with Theorem 4.2 and Remark 4.2, yields that for the case of missing covariates, the model selected via $\ell_1 - \mathbf{CLS}$ satisfies

$$T \subseteq \hat{T}, \qquad \hat{m} \preceq s, \tag{4.7}$$

w.p. at least $1 - c_1 \exp(-c_2 \log p)$, for all sufficiently large n. This concludes this section on the recovery of the support of β_0. We now proceed to the estimation of β_0.

5 Step 2: Estimation

This section shall investigate the estimation properties of the following estimator. With $\hat{Q}_n(\cdot)$ as in (4.2), define the post selection corrected least squares estimator of β_0 as

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \hat{Q}_n(\beta); \;\; \beta_j = 0, \quad \text{for each } j \in \hat{T}^c. \tag{5.1}$$

Notice that the support $\hat{T}$ of the model selected from Step 1 is itself random. We shall show that the estimator $\tilde{\beta}$ performs at least as well as ℓ_1 penalized corrected least squares in terms of the rate of convergence of ℓ_2 estimation error, under suitable assumptions on model selection. More interestingly, $\tilde{\beta}$ has the potential to outperform ℓ_1 penalized methods, depending on the first step model selection. In fact $\tilde{\beta}$ attains the oracle rate $\sqrt{s/n}$ under perfect model selection ($w.p. \to 1$). Furthermore, the implementation of the proposed estimator requires the knowledge of only a sub-block $\big(O(s)$-dimensional$\big)$ of the bias correction matrices Σ_w or M in the additive error or missing covariate cases, respectively.

For any constant $0 < c_3 \leq 1$ and a universal constant D, let

$$e_n(m, c_3) = \sqrt{\frac{m \log p}{n}} + \sqrt{\frac{(m+s)\log(D)}{n}} + \sqrt{\frac{m + s + \log(1/c_3)}{n}}. \tag{5.2}$$

Consider the following assumption.

$$\sup_{\|\delta_{T^c}\|_0 \leq m, ; \|\delta\|_2 > 0} \frac{1}{\|\delta\|_2} \left| \delta' \Gamma \beta_0 - \delta' \hat{\gamma} \right| \leq c_0 r \|\beta_0\|_2 e_n(m, c_3). \tag{5.3}$$

Here $0 < r < \infty$ is a suitably chosen constant depending on the two sources of noise W and ε. Later in this section we show that this uniform bound holds with asymptotic probability 1 for both pairs $\left(\Gamma^{\text{add}}, \hat{\gamma}^{\text{add}}\right)$ and $\left(\Gamma^{\text{miss}}, \hat{\gamma}^{\text{miss}}\right)$. We now state the main result of this section.

Theorem 5.1 *Suppose model is selected by the* **CS** *method and assumptions of Theorem 4.1 hold. Furthermore, assume that the pairs* $(\Gamma^{add}, \hat{\gamma}^{add})$ *for the additive error case or* $(\Gamma^{miss}, \hat{\gamma}^{miss})$ *for the missing covariate case satisfy the uniform deviation condition in (5.3) and the condition lower-***RSE**(a_n) *with* a_n *as in (4.3). Then there exists a universal positive constant* c_0 *such that*

$$\|\tilde{\beta} - \beta_0\|_2 \leq \frac{1}{\kappa(\hat{m})} c_0 r \|\beta_0\|_2 e_n(\hat{m}, c_3), \tag{5.4}$$

holds, w.p. at least $1 - c_1 \exp(-c_2 \log p) - \left(6c_3 \exp(-s)/(1 - 1/e)\right)$, *for all sufficiently large* n.

Corollary 5.1 *Suppose the conditions of Theorem 5.1 hold, and that* $\|\beta_0\|_2 \leq b_0$, *for some constant* $b_0 < \infty$. *Then*

$$\|\tilde{\beta} - \beta_0\|_2 \preceq_P \begin{cases} \sqrt{\frac{s \log p}{n}} \; ; & \text{in general} \\ \sqrt{\frac{s}{n}} + \sqrt{\frac{o(1) s \log p}{n}} \; ; & \text{if } a_n/s \to 1^+, \\ \sqrt{\frac{s}{n}} \; ; & \text{if } a_n = s. \end{cases} \tag{5.5}$$

The proof of this corollary is a direct consequence of Theorem 5.1 and is thus omitted. An immediate consequence of this corollary is that implementing the two stage corrected least squares with the first stage model selection done via the **CS** method will result in estimates that perform at least as well as ℓ_1 penalized counterparts. More importantly, the two stage method has room for improvement for the rate of convergence. In contrast, ℓ_1 penalized methods have a rate of $\sqrt{s \log p/n}$ which is empirically known to be optimal, see, e.g. Loh and Wainwright (2012). In fact under perfect ($w.p. \to 1$) model selection, $\tilde{\beta}$ achieves the $\sqrt{s/n}$, which is the oracle rate of convergence.

The second useful aspect of this method is that implementing the second step estimation requires only an $O(s)$ dimensional block of the p dimensional bias correction matrix Σ_w or M to be known or estimated. In comparison, the ℓ_1 penalized method for simultaneous model

selection and estimation requires the entire p dimensional matrix. Keeping in mind that the dimension p can be growing exponentially with n, estimating Σ_w from the low number of typically available replicates of the design variables may be infeasible.

Next, we focus on the case of missing covariates where model selection is done via ℓ_1-**CLS** method and estimation via (5.1). This shall again yield estimates that are at least as efficient as the estimates based on $\ell_1 - \mathbf{CLS}$ method and shall allow room for improvement in its efficiency.

Theorem 5.2 *Suppose the model (2.1) and (2.3) holds, and let model selection be done via* ℓ_1-**CLS** *method. Assume conditions of Theorem 4.2 and in addition assume that the pair* $\left(\Gamma^{miss}, \hat{\gamma}^{miss}\right)$ *satisfies the uniform deviation condition (5.3) and the matrix* Γ^{miss} *satisfies condition lower-***RSE**(b_n) *and (4.6) for any* $b_n = O(s)$*. Then there exists a universal positive constant* c_0 *such that*

$$\|\tilde{\beta} - \beta_0\|_2 \leq \frac{1}{\kappa(\hat{m})} c_0 r \|\beta_0\|_2 e_n(\hat{m}, c_3).$$

In particular, if $\|\beta_0\|_2 \preceq 1$*, then the rate of convergence described here is at least*

$$\|\tilde{\beta} - \beta_0\|_2 \preceq \sqrt{\frac{s \log p}{n}}.$$

The results of Theorem 5.2 are not accompanied by a probabilistic statement since this result follows by deterministic arguments on the event where the required assumptions hold. In view of Lemma 3.1 and Lemma 5.1, all assumptions made on random quantities made here hold with asymptotic probability 1. Thus, the conclusions of this theorem hold with asymptotic probability 1.

The proof of Theorem 5.2 essentially uses the property (4.7) from the first step model selection and is similar to that of Theorem 5.1, hence omitted. This theorem may also be extended easily to any model selection procedure satisfying (4.7).

The only remaining part is to now show that the uniform deviation assumption (5.3) holds with high probability. This forms the content of the following lemma.

Lemma 5.1 *Suppose the model (2.1) holds and that the covariate noise* W *satisfies conditions* **(A1)** *and* **(A3)** *for additive error and missing covariate cases, respectively. Let*

$$r = \begin{cases} \sigma_z(\sigma_w + \sigma_\varepsilon) & ; \quad \text{for additive error} \\ \frac{\sigma_x}{1-\rho_{\max}}\left(\sigma_\varepsilon + \frac{\sigma_x}{1-\rho_{\max}}\right) & ; \quad \text{for missing covariates.} \end{cases}$$

Then, with $\Gamma = \Gamma^{add}$, $\hat{\gamma} = \hat{\gamma}^{add}$ *or* $\Gamma = \Gamma^{miss}$, $\hat{\gamma} = \hat{\gamma}^{miss}$,

$$\sup_{\|\delta_{T^c}\|_0 \leq m\,;\, \|\delta\|_2 > 0} \frac{1}{\|\delta\|_2} \left| \delta' \Gamma \beta_0 - \delta' \hat{\gamma} \right| \leq c_0 r \|\beta_0\|_2 e_n(m, c_3),$$

w.p. at least $1 - 6c_3 e^{-s}/(1 - 1/e)$*, for all sufficiently large* n*.*

5.1 Application to estimation of precision matrices with missing observations

Estimation of covariance and precision matrices plays an important role in several statistical analyses including the principal component analysis, linear/quadratic discriminant analysis, and graphical modeling. This problem has been extensively researched in the case where the observations are i.i.d. vectors from a multivariate sub-Gaussian distribution, see, Meinhausen and Bühlmann (2006), Friedman, Hastie and Tibshirani (2007) and Bickel and Levina (2008). On the other hand, when the observed vector is corrupted by missing variables, Loh and Wainwright (2012) propose an algorithm based on the ℓ_1 penalized corrected least squares estimator, which provides consistent estimates.

Suppose observations $X_i \in \mathbb{R}^p$, $1 \le i \le n$ are i.i.d. $\mathcal{N}(0, \Sigma)$, where Σ is a positive definite matrix. Then it is well known, see, e.g., Anderson (2003), that for each $1 \le j \le p$, the conditional distribution of the j^{th} component X_i^j, given the rest X_i^{-j}, is again normal distribution, i.e.,

$$X_i^j \Big| X_i^{-j} \sim \mathcal{N}\Big(\Sigma_{j,-j}\Sigma_{-j,-j}^{-1} X_i^{-j} \ , \ \Sigma_{jj} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} \Big).$$

This result can equivalently be written as the linear relation,

$$X_i^j = X_i^{-jT}\theta^j + \varepsilon^j, \quad 1 \le i \le n, \tag{5.6}$$

where $\theta^j = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ is a $p-1$ dimensional vector and ε^j is a vector of i.i.d. Gaussian r.v.'s, independent of X^{-j}. Here $\Sigma_{-j,-j}$ represents the sub-block of Σ with the j^{th} row and column removed. The precision matrix $\Theta := \Sigma^{-1}$ can then be reconstructed from θ^j, $1 \le j \le p$ as follows,

$$\Theta_{jj} = d_j, \ \text{ and } \ \Theta_{-j,j} = -d_j\theta^j, \quad \text{where} \quad d_j := (\Sigma_{jj} - \Sigma_{j,-j}\theta^j)^{-1}. \tag{5.7}$$

When $Z_i = X_i \oplus W_i$ is observed in place of X_i, $1 \le i \le n$ with missing observations as described in (2.3), we can use the methodology proposed in sections 4 and 5 above to estimate the parameters θ^j of the model (5.6) for every $1 \le j \le p$. Note that the response and predictor variables both have missing observations in this case, however the proofs of our results in the previous sections can be easily seen to hold under this setup.

For any matrix $A = (a_{ij})_{1\le i,j\le p}$, let $\|A\|_2 = \max_{1\le j\le p} \big(\sum_{i=1}^p a_{ij}^2 \big)^{1/2}$. Also define $\lambda_{\min}$, $\lambda_{\max}$ as the minimum and maximum eigenvalues of the covariance matrix of interest Σ. Then we have the following algorithm.

Algorithm 1:

1. Let $\hat{\Sigma} = n^{-1}Z'Z \ominus M$ with M as defined in (3.4). For each $1 \le j \le p$, define

$$(\Gamma^j, \hat{\gamma}^j) = \Big(\hat{\Sigma}_{-j,-j}, n^{-1}Z^{-jT}Z^j \ominus (\mathbf{1} - \boldsymbol{\rho}^{-j})(1 - \boldsymbol{\rho}_j) \Big),$$

and estimate the support of θ^j by the **CS** method for model selection, i.e., with the thresholding level $a_n = c_0 s$, $c_0 > 1$, let

$$\hat{T}_j(a_n) = \{k \, ; \, r_k(\hat{\gamma}) \le a_n, \ 1 \le k \le (p-1)\}.$$

2. Obtain estimates $\hat{\theta}^j$ by the following optimization,

$$\hat{\theta}^j = \underset{\|\theta\|_1 \le b_0\sqrt{s}}{\arg\min}\, \hat{Q}^j(\theta)\,;\, \theta_k = 0, \qquad \text{for each } k \in \hat{T}_j^c. \tag{5.8}$$

3. Substitute $\hat{\Sigma}$ and $\hat{\theta}^j$, to obtain $\hat{d}_j = (\hat{\Sigma}_{jj} - \hat{\Sigma}_{j,-j}\hat{\theta}^j)^{-1}$ and complete the estimated precision matrix $\tilde{\Theta}$ with $\tilde{\Theta}_{-j,j} = -\hat{d}_j\hat{\theta}^j$ and $\tilde{\Theta}_{j,j} = \hat{d}_j$

4. Set $\hat{\Theta} = \arg\min_{\Theta \in S^p} \|\Theta - \tilde{\Theta}\|_2$, where S^p is the collection of symmetric matrices.

Note that we have placed an additional restriction on the parameter space in step 2 of the algorithm, i.e., $\|\theta^j\|_1 \le b_0\sqrt{s}$, where s represents $\text{card}(T_j)$. This additional restriction does not influence the proofs of Section 4 and 5.

The choice of the thresholding level $a_n = c_0 s$ is required for our proofs. However in practice a_n is a data based tuning parameter as described earlier, see also Remark 6.1.

We now proceed to providing consistency in estimation of the above algorithm. The assumptions required for this purpose are restated below in the present context.

Assumptions:

(G1) The vectors $\theta^j = \Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$ are s-sparse, i.e., for all $1 \le j \le p$, $|T_j| \le s$, where $T_j = \text{Supp}(\theta^j)$. Furthermore, assume that $\|\theta^j\|_1 \le b_0\sqrt{s}$, $1 \le j \le p$ for some constant $b_0 < \infty$.

(G2) The covariance matrix Σ has bounded maximum and minimum eigenvalues, i.e., $0 < \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) < \infty$, and satisfies the following relation for all $1 \le j \le p$,

$$\begin{aligned} &\max |\Sigma^{T_j^c T_j}_{-j,-j}\theta^j_{T_j}| + 2\frac{\sigma}{c_0(1-\rho_{\max})^2}(\sigma_\varepsilon + \sigma_x\|\theta^j\|_2)\sqrt{\frac{c_1 \log p}{n}} \\ &\qquad < \min |\Sigma^{T_j T_j}_{-j,-j}\theta^j_{T_j}|. \end{aligned}$$

Here $\Sigma^{T_j T_j}_{-j,-j}$, and $\Sigma^{T_j^c T_j}_{-j,-j}$ are a partitions of $\Sigma_{-j,-j}$ as described in (3.1).

(G3) The pairs $(\Gamma^j, \hat{\gamma}^j), 1 \le j \le p$, satisfy the following uniform deviation condition. For each $1 \le j \le p$, and for some constant $r < \infty$.

$$\sup_{\|\delta_{T_j^c}\|_0 \le m, ; \|\delta\|_2 > 0} \frac{1}{\|\delta\|_2}\left|\delta'\Gamma^j\theta^j - \delta'\hat{\gamma}^j\right| \le c_0 r\|\theta^j\|_2 e_n(m, c_3). \tag{5.9}$$

Theorem 5.3 *In addition to conditions* **(G1)**, **(G2)** *and* **(G3)**, *assume that the lower-*$\mathbf{RSE}(a_n)$ *condition holds uniformly over the matrices* Γ^j, *and*

$$\left\|\hat{\Sigma} - \Sigma\right\|_\infty \le c_0\sigma_x\sqrt{\frac{\log p}{n}}. \tag{5.10}$$

Then the estimated precision matrix $\hat{\Theta}$ *provided by Algorithm 1 satisfies*

$$\left\|\hat{\Theta} - \Theta\right\|_2 \le c_0\Big(C_2^2 + \frac{C_1^2}{\lambda^2_{\min}} + \frac{\lambda_{\max}}{\lambda_{\min}}C_2\Big)^{1/2} e_n(a_n - s, c_3),$$

w.p. converging to 1, where

$$C_1 = r\lambda_{\max}/\lambda_{min}, \qquad C_2 = \Big(b_0\sigma_x + \big(r\lambda^2_{\max}/\lambda^2_{\min}\big)\Big)\Big/\lambda^2_{\min}.$$

Remark 5.1 Assumption (5.10) is a standard assumption for high dimensional covariance recovery. It can be shown to hold with asymptotic probability 1 with an appropriate choice of constant c_0, see Yuan (2010). The uniform bound of Assumption **(G3)** can be shown to hold using the same arguments as in Lemma 5.1. Lastly, the condition lower-**RSE** can be shown to hold uniformly over $1 \leq j \leq p$ with high probability by applying arguments of Lemma 3.1 along with the observation that in this case uniformly over all $1 \leq j \leq p$,

$$0 < \lambda_{\min} \leq k_{-j,-j}(m) \leq \phi_{-j,-j}(m) \leq \lambda_{\max} < \infty, \tag{5.11}$$

for any $1 \leq m \leq k_n$. Here $\kappa_{-j,-j}$ and $\phi_{-j,-j}$ are as defined in (3.7) with Σ_x replaced by $\Sigma_{-j,-j}$.

6 Simulation Study

In this section we numerically analyse the performance of the methodology developed in this paper. We implement our two step methodology with model selection done via method **(CS)** and refer to these as the post selection estimates. In the following we shall compare the post selection estimates with the ℓ_1 penalized corrected least squares estimates of Loh and Wainwright (L&W) and the ordinary Lasso which disregards covariate noise or missing-ness.

Remark 6.1 *Tuning Parameter:* The model selection step in our post selection estimates involves choosing an appropriate value of the thresholding level a_n. To choose this tuning parameter, we employ standard cross validation logic, i.e., the **CS** method is used to select models for a grid of values of $a_n \in \{1, 2, .., \min\{p, (n/\log p)\}\}$ and corresponding estimates $\tilde{\beta}_{a_n}$ are obtained via (5.1). These estimates are then used on an independent test set to compute the corrected least squares loss

$$L(\tilde{\beta}_{a_n}) = \frac{1}{2}\tilde{\beta}'_{a_n}\Gamma\tilde{\beta}_{a_n} - \hat{\gamma}'\tilde{\beta}_{a_n}.$$

The tuning parameter a_n is chosen as a minimizer of this criteria. Naturally, larger the grid chosen for a_n, more is the computation time necessary. To maintain fairness of comparisons the same cross validation approach is used to choose the tuning parameter λ of ℓ_1 penalized corrected least squares estimator. The tuning parameter for the ordinary Lasso is chosen via similar cross validation with the loss function chosen as ordinary least squares. In both cases λ is allowed to range from zero to one with increments of 0.05.

All simulations are performed in R, the estimates of ordinary Lasso are obtained using the package *glmnet* developed by Friedman et al. (2013). Post selection estimates $\tilde{\beta}$ and L&W estimates are obtained via the projected gradient descent algorithm, see, e.g. Agarwal, Neghban and Wainwright (2012) and Loh and Wainwright (2012).

6.1 Simulation Setup and Results

We begin with the regression setting with additive error or missing covariates. Here, the unobserved design variables $\{x_{ij},\, 1 \leq i \leq n,\, 1 \leq j \leq p,\}$ are chosen as i.i.d. r.v.'s from a standard normal distribution. Components of the parameter vector β_0 are generated independently from a uniform distribution with support $(-4,-1) \cup (1,4)$. The model errors ε_i, $1 \leq i \leq n$ are generated as i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.25$.

To evaluate performance of the estimators, two cases each are presented for additive errors and missing covariates. An independent data set for every combination of the following (n, p, s) settings is generated,

- Sample size n ranging from 100 to 500 with increments of 40. Model dimension p ranging from 100 to 500 with increments of 65 and the number of non zero parameters $s = 4, 8$. Thus leading to $154(11 \times 7 \times 2)$ independent models.
- $p = 750$, $s = 4$ and n ranging from 50 to 500 with increments of 5, thus leading to 91 independent models.

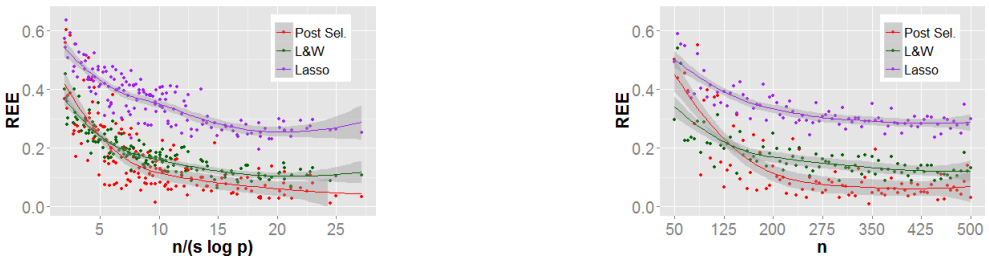
The three estimators to be compared are computed for each generated model and we report the following measures for comparison, (1) relative estimation error, REE:$=\|\beta - \beta_0\|_2/\|\beta_0\|_2$, (2) number of false positives, i.e., number of incorrectly identified zero components, and (3) computation time (in seconds) required to obtain estimates.

Example 1. Regression Setting: Suppose the model (2.1) holds and consider the following two cases.

• *Additive Error*: The covariate noise variables W_i, $1 \leq i \leq n$ are assumed to be i.i.d. Gaussian with mean zero and covariance matrix $\Sigma_w = c_w \left[\sigma_{ij}^w\right]_{i,j=1,..,p}$, where $\sigma_{ij}^w = 0.5^{|i-j|}$ and $c_w = 0.25$. Simulation results are illustrated in Figure 1: **Ad1-Ad4**.

• *Missing Covariates*: The missing covariates are generated as described in assumption **(A2)** where ρ_j, $1 \leq j \leq p$ are chosen independently from a uniform distribution over the support $(0.05, 0.75)$. Simulation results are illustrated in Figure 1: **M1-M4**.

Ad1 **Ad2**



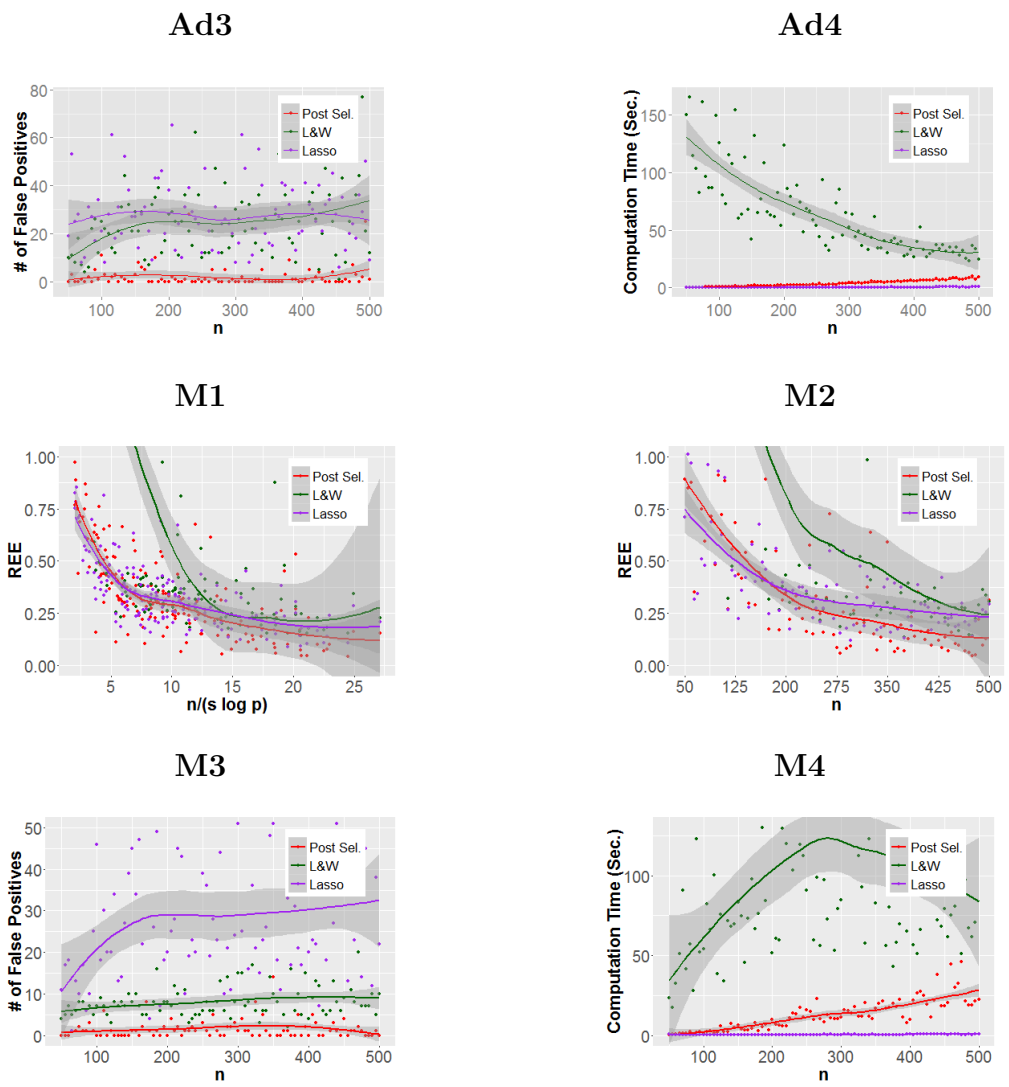


Figure 1: **Ad1 & M1** plots REE against $n/s\log p$, here $n, p \in [100, 500]$, and $s \in \{4, 8\}$ for the additive error and missing covariate cases respectively. **Ad2, Ad3, Ad4 & M2, M3, M4**, plots REE, false positives, and computation time at p=750, against n in the additive and missing cases respectively.

• *Estimation accuracy*: The empirical results support the theoretical findings. Consistency in the ℓ_2 estimation error of the post selection estimator is clearly observed. In addition, the post selection estimates nearly uniformly outperform the two other estimators, see Figure 1: **Ad1, Ad2, M1, M2**. The L&W estimates perform marginally better at lower sample sizes for the additive error case, see Figure 1: **Ad1, Ad2**.

• *False positives*: In both additive and missing covariate cases, the post selection estimates are seen to provide a significant improvement in the control on false positives, see Figure 1: **Ad3, M3**.

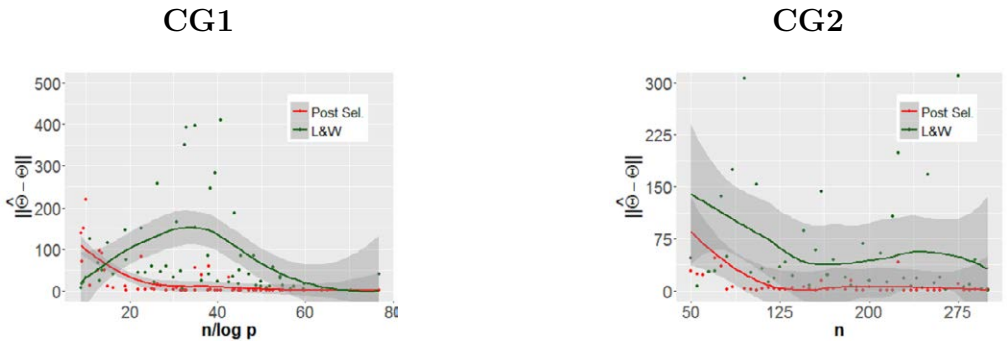
• *Computation time*: The computation time for post selection estimates is significantly quicker in comparison to L&W estimates and comparable to Lasso at larger values of p and a fixed sample size n. However the computation time for post selection estimates increases

with n due to the increase in the grid size of a_n for cross validation. In comparison, the computation time of L&W and Lasso estimates decrease as the grid size for cross validation stays the same and numerical convergence becomes quicker with higher n, see Figure 1: **Ad4, M4**.

Example 2. Graphical Models: In this example we examine the efficacy of the proposed algorithm in estimating the precision matrices for two types of Gaussian graphical models, namely band and cluster structured graphs. These precision matrices are generated by the package "fastclime" developed by Pang, Liu and Vanderbei (2014). For a p-dimensional graph, around $p/20$ band width or clusters are assumed in the two cases, respectively. The adjacency matrices of these graphs with $p = 50$ are illustrated below

Figure 2: Plots of adjacency matrices of banded and cluster precision matrices respectively.

The precision matrices are generated so that the corresponding covariance matrix $\Sigma = \Omega^{-1}$ is normalized to have all diagonal components 1. For further details on the construction of these matrices see, page 5 of Pang, Liu and Vanderbei (2014). Next, the unobserved variables X_i, $1 \leq i \leq n$, are generated as i.i.d. $\mathcal{N}(0, c_x\Sigma)$ for $c_x = 1, 3$. Missing-ness is then induced as $Z_i = X_i * W_i$, $1 \leq i \leq n$ in accordance with (2.3), where ρ_j, $1 \leq j \leq p$, are chosen independently from a uniform distribution over the support $(0.05, 0.75)$. For each model, we compute estimates via the proposed Algorithm 1 and compare it to the estimates based on the ℓ_1 penalized version of L&W. For performance comparison we report $\|\hat{\Theta} - \Theta\|_2$, in addition to false positives identified in the matrix and the computation time required to compute corresponding estimates.



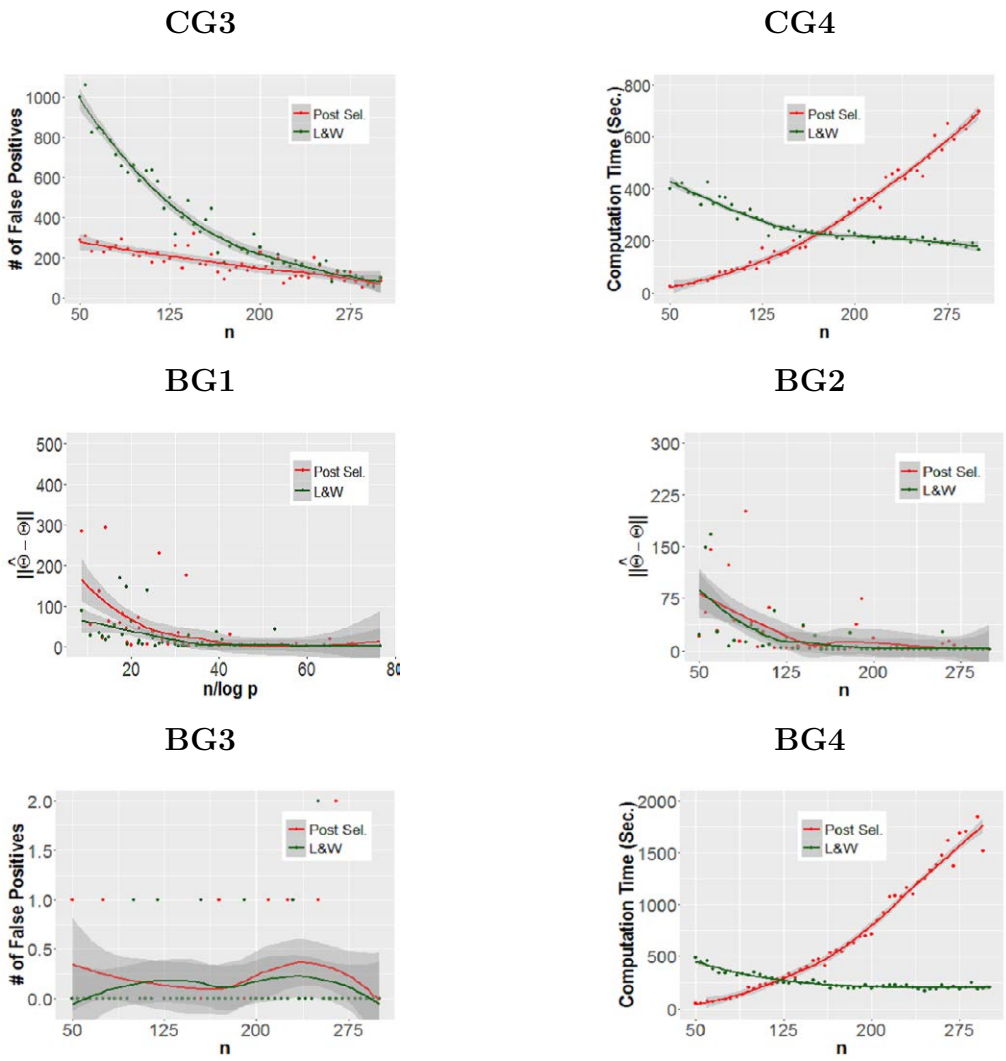


Figure 3: **CG1 & BG1** plots $\|\hat{\Theta} - \Theta\|_2$ against $n/\log p$, here $n, p \in [50, 300]$, for the cluster and banded graph cases respectively. **CG2, CG3, CG4 & BG2, BG3, BG4**, plots $\|\hat{\Theta} - \Theta\|_2$, false positives, and computation time at p=100 against n in the cluster and banded graph cases respectively. Also, $c_x = 3, 1$ in the cluster and banded graph cases respectively.

• *Estimation accuracy*: Post selection estimates provide consistent estimates of Θ. In addition, they are uniformly superior in the case of the cluster graph and perform about as well as L&W estimates in the banded graph case, see Figure 3: **CG1, CG2 & BG1, BG2**. This is due to the constant c_x which is 3 in the latter and 1 in the banded graph case. It is seen that the post selection estimates become uniformly superior as c_x is increased.

• *Computation time*: Although the computation time for L&W estimates in the settings presented here is significantly faster when n increases, however it is also observed that increasing the dimension p significantly favors post selection estimates in terms of computational efficiency.

Note: In Figures 1 and 3, three colors of each dot represent a performance measure corresponding to an independently generated model for the three estimates being compared. To

measure the average performance over the independently simulated models, non parametric regression lines and corresponding confidence bands are drawn, these are made via the Loess method with its smoothing parameter set as 0.75.

7 Appendix

7.1 Proofs for Section 3

The proofs to follow require a probability bound for centered sum of squares of independent sub-Gaussian r.v.'s. This is facilitated by Lemma 14 of Loh and Wainwright (2012) supplement which is restated below for completeness. This lemma in turn is a direct corollary of Lemma 14 of Vershynin (2012).

Lemma 7.1 *If $X \in \mathbb{R}^{n\times p_1}$ is any zero mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2), then for any fixed unit vector $v \in \mathbb{R}^{p_1}$ and $t > 0$,*

$$P\Big(\Big|\|Xv\|_2^2 - E\|Xv\|_2^2\Big| \geq nt\Big) \leq 2\exp\Big(-\frac{1}{c_0}n\min\big\{\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\big\}\Big), \tag{7.1}$$

where $c_0 > 0$ is a universal constant. Moreover, if $Y \in \mathbb{R}^{n\times p_2}$ is a zero mean sub-Gaussian matrix with parameters (Σ_y, σ_y^2) then for every $t > 0$,

$$\begin{aligned} P\Big(\Big\|n^{-1}Y'X - \mathrm{Cov}(y_i, x_i)\Big\|_\infty \geq t\Big) \\ \leq \ 6p_1p_2\exp\Big(-c_0 n\min\big\{\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\big\}\Big). \end{aligned} \tag{7.2}$$

Here y_i and x_i represent the i^{th} rows of Y and X, respectively.

To state the next lemma we need to define

$$\mathcal{B}(m) = \{\delta \in \mathbb{R}^p, \ ; \ \|\delta_{T^c}\|_0 \leq m\}, \quad \mathcal{B}_1(m) = \{\delta \in \mathcal{B}(m), \ ; \ \|\delta\|_2 \leq 1\}.$$

Lemma 7.2 *Let $X \in \mathbb{R}^{n\times p}$ be a sub-Gaussian matrix with parameters (Σ_x, σ_x^2). For a $c_3 \in (0, 1)$, let*

$$r_n = r_n(m, c_3) = \frac{\sigma_x^2}{3c_0}e_n(m, c_3),$$

where $e_n(m, c_3)$ is as in (5.2). Then, for all sufficiently large n,

$$P\left(\sup_{\delta\in\mathcal{B}_1(m)}\Big|\frac{\|X\delta\|_2^2}{n} - E\frac{\|X\delta\|_2^2}{n}\Big| \geq r_n, \ \textit{for all } m \leq n\right) \leq 2c_3e^{-s}/(1 - 1/e).$$

Proof of Lemma 7.2. For every $U \subseteq \{1, ..., p\}$ with $\text{card}(U - T) \leq m$, define $S_U := \{\delta \in \mathbb{R}^p;\ \|\delta\|_2 \leq 1;\ \text{supp}(\delta) \subseteq U\}$ and note that $\mathcal{B}_1(m) = \cup_{card(U-T)\leq m} S_U$. Let $\mathcal{A} = \{u_1, ..., u_k\}$ be a 1/10 cover of a fixed S_U, i.e., for each $\delta \in S_U$ there exists $u_i \in \mathcal{A}$ such that $\|\delta - u_i\|_2 \leq 1/10$. It is known from Ledoux and Talagrand (1991) or Loh and Wainwright (2012) (Supplementary materials, pg.17), that one can construct $\mathcal{A}$ such that $card(\mathcal{A}) \leq 100^{(m+s)}$. Let $\psi(\delta_1, \delta_2) = \delta_1'\big(\frac{X'X}{n} - \Sigma_x\big)\delta_2$. Then by elementary algebra,

$$\psi(\delta, \delta) = \psi(u_i, u_i) + 2\psi(\delta - u_i, u_i) + \psi(\delta - u_i, \delta - u_i) \tag{7.3}$$

Note that by construction, there exists $u_i \in \mathcal{A}$ such that $10(\delta - u_i) \in S_U$ and thus $2\psi(\delta - u_i, u_i) = \frac{2}{10}\psi(\delta_1, u_i)$, for $\delta_1 := 10(\delta - u_i) \in S_U$.

Now expressing the second term on the r.h.s. of (7.3) as,

$$\begin{aligned} \frac{2}{10}\psi(\delta_1, u_i) &= \frac{1}{10}\psi(\delta_1, \delta_1) + \frac{1}{10}\psi(u_i, u_i) - \frac{1}{10}\psi(\delta_1 - u_i, \delta_1 - u_i) \\ &\leq \frac{1}{10}\max_i |\psi(u_i, u_i)| + \frac{5}{10}\sup_{\delta \in S_U} |\psi(\delta, \delta)|, \end{aligned}$$

where the last inequality follows since for any $\delta_1, \delta_2 \in S_U$ we also have $\frac{1}{2}(\delta_1 - \delta_2) \in S_U$. Replacing this inequality in (7.3) we obtain

$$\begin{aligned} &\sup_{\delta \in S_U} |\psi(\delta, \delta)| \leq \frac{11}{10}\max_i |\psi(u_i, u_i)| + \frac{51}{100}\sup_{\delta \in S_U} |\psi(\delta, \delta)|, \text{ or} \\ &\sup_{\delta \in S_U} |\psi(\delta, \delta)| \leq 3\max_i |\psi(u_i, u_i)|. \end{aligned}$$

Applying (7.1) of Lemma 7.1 to each $\psi(u_i, u_i)$ and taking a union bound over the 100^{m+s} such possibilities we obtain,

$$P\left(\sup_{\delta \in S_U} \Big|\frac{\|X\delta\|_2^2}{n} - E\frac{\|X\delta\|_2^2}{n}\Big| \geq 3t\right) \leq 2\,100^{m+s}\,\exp\Big[-c_0 n \min\big(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\big)\Big].$$

Again taking the union bound over all $\binom{p}{m} \leq p^m$ possibilities of U we obtain

$$P\left(\sup_{\delta \in \mathcal{B}_1(m)} \Big|\frac{\|X\delta\|_2^2}{n} - E\frac{\|X\delta\|_2^2}{n}\Big| \geq 3t\right) \leq 2\,100^{m+s}\,p^m \exp\Big[-c_0 n \min\big(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\big)\Big].$$

Choose t such that

$$c_0\sqrt{n}t/\sigma_x^2 = \sqrt{m\log p} + \sqrt{(m+s)\log 100} + \sqrt{(m+s) + \log(1/c_3)}.$$

Then for n large enough $t^2 < t$ and we obtain

$$P\left(\sup_{\delta \in \mathcal{B}_1(m)} \Big|\frac{\|X\delta\|_2^2}{n} - E\frac{\|X\delta\|_2^2}{n}\Big| \geq r_n\right) \leq 2c_3 \exp[-m - s].$$

Summing over all m both sides of this bound in turn yields

$$P\left(\sup_{\delta\in\mathcal{B}_1(m)}\Big|\frac{\|X\delta\|_2^2}{n}-E\frac{\|X\delta\|_2^2}{n}\Big|\geq r_n,\ \text{for all}\, m\leq n\right)$$
$$\leq\ 2c_3\sum_{m=0}^{\infty}\exp[-m-s]=2c_3e^{-s}/(1-1/e).$$

This completes the proof of the lemma □

Proof of Lemma 3.1. Case 1: Additive Error: Apply Lemma 7.2 to the sub-Gaussian matrix Z to obtain w.p. at least $1-2c_3\exp(-s)/(1-1/e)$, for any $\delta\in\mathcal{B}_1(m)$, $m\leq k_n$,

$$\Big|\delta'\big[n^{-1}Z'Z-\Sigma_z\big]\delta\Big|\leq r_n(m,c_3).$$

Now substitute the relation $\Sigma_z=\Sigma_x+\Sigma_w$ in this inequality to obtain

$$\delta'\Sigma_x\delta-r_n(m,c_3)\leq\delta'\Gamma^{\text{add}}\delta\leq\delta'\Sigma_x\delta+r_n(m,c_3). \tag{7.4}$$

Notice that since $k_n\log p=o(n)$ by assumption, $r_n\to 0$. Hence (7.4) and the assumption (3.7) imply that the condition $\mathbf{RSE}(k_n)$ holds for the matrix Γ^{add} w.p. at least $1-2c_3\exp(-s)/(1-1/e)$, for all n large.

Case 2: Missing Covariates: Observe that for any $\delta\in\mathbb{R}^p$ we have

$$\begin{aligned}\Big|\delta'(\Gamma^{\text{miss}}-\Sigma_x)\delta\Big| &= \Big|\delta'\Big(\big(n^{-1}Z'Z-\Sigma_z\big)\ominus M\Big)\delta\Big|\\ &\leq \frac{1}{(1-\rho_{\max})^2}\Big|\delta'\big(n^{-1}Z'Z-\Sigma_z\big)\delta\Big|.\end{aligned}$$

Since Z is sub-Gaussian by Remark 2.1, applying Lemma 7.2 we obtain w.p. at least $1-2c_3\exp(-s)/(1-1/e)$ that

$$\Big|\delta'(\Gamma^{\text{miss}}-\Sigma_x)\delta\Big|\leq\frac{r_n(m,c_3)}{(1-\rho_{\max})^2}\,.$$

The fact $0\leq\rho_{\max}<1$ and the assumption (3.7) imply that the condition $\mathbf{RSE}(k_n)$ holds for the matrix Γ^{miss}, w.p. at least $1-2c_3\exp(-s)/(1-1/e)$, for large enough n. □

7.2 Proofs for Section 4

Proof of Theorem 4.1. Part (ii) of this theorem follows by construction of $\hat{T}(a_n)$. We prove part (i) separately for the two cases of additive errors and missing covariates.

Case 1: Additive Error: It suffices to show that except on a set with asymptotic probability zero,

$$\max|Z'_{T^c}y|<\min|Z'_Ty|. \tag{7.5}$$

Consider

$$\begin{aligned} \big\|\hat{\gamma} - \Sigma_x\beta_0\big\|_\infty &= \big\|n^{-1}Z'y - \Sigma_x\beta_0\big\|_\infty \\ &\leq \big\|n^{-1}Z'X\beta_0 - \Sigma_x\beta_0\big\|_\infty + \big\|n^{-1}Z'\varepsilon\big\|_\infty. \end{aligned} \tag{7.6}$$

Let

$$t_1 = c_0\sigma_z\sigma_\varepsilon\sqrt{(\log p)/n}. \tag{7.7}$$

Use (7.2) of Lemma 7.1 with $t = t_1$, to obtain that

$$\begin{aligned} P\Big(n^{-1}\|Z'\varepsilon\|_\infty \geq t_1\Big) &\leq 6p\,\exp\Big(-cn\min\big\{\frac{t_1^2}{\sigma_z^2\sigma_\varepsilon^2}, \frac{t_1}{\sigma_z\sigma_\varepsilon}\big\}\Big) \\ &\leq c_1\exp\big(-c_2\log p\big). \end{aligned} \tag{7.8}$$

Similarly, upon choosing $t_2 = c_0\sigma_z\sigma_x\|\beta_0\|_2\sqrt{(\log p)/n}$ we obtain

$$P\Big(n^{-1}\|Z'X\beta_0 - \Sigma_x\beta_0\| \geq t_2\Big) \leq c_1\exp\big(-c_2\log p\big). \tag{7.9}$$

Using inequalities (7.8) and (7.9) in (7.6), we obtain w.p. at least $1 - c_1\exp(-c_2\log p)$,

$$\big\|\hat{\gamma} - \Sigma_x\beta_0\big\|_\infty \leq t_1 + t_2,$$

for all sufficiently large n. The proof of (7.5) is now completed upon combining this bound with assumption **(A3i)**. This concludes the proof for additive errors.

Case 2: Missing Covariates: Here, it suffices to show that except on a set with asymptotic probability zero we have,

$$\max|Z'_{T^c}y \ominus (\mathbf{1} - \boldsymbol{\rho})| < \min|Z'_T y \ominus (\mathbf{1} - \boldsymbol{\rho})|. \tag{7.10}$$

For this purpose consider

$$\begin{aligned} &\big\|\hat{\gamma} - \Sigma_x\beta_0\big\|_\infty \\ &= \Big\|n^{-1}Z'y \ominus (\mathbf{1} - \boldsymbol{\rho}) - \Sigma_x\beta_0\|_\infty\Big\|_\infty \\ &\leq \frac{1}{1 - \rho_{\max}}\Big(\big\|n^{-1}Z'X\beta_0 - \mathrm{cov}(z_i, x_i)\beta_0\big\|_\infty + \|n^{-1}Z'\varepsilon\|\Big). \end{aligned} \tag{7.11}$$

Recall as stated in Remark 2.1, Z is a sub-Gaussian matrix with parameter σ_x. Hence, argue as for (7.8) and (7.9), with t_1 as in (7.7) and $t_2 = c_0\sigma_x^2\|\beta_0\|_2\sqrt{(\log p)/n}$, to obtain that

$$\begin{aligned} &P\Big(n^{-1}\|Z'\varepsilon\|_\infty \geq t_1\Big) \leq c_1\exp(-c_2\log p), \\ &P\Big(n^{-1}\|Z'X\beta_0 - \Sigma_x\beta_0\| \geq t_2\Big) \leq c_1\exp(-c_2\log p). \end{aligned}$$

These bounds together with the inequality (7.11) imply that, w.p. at least $1-c_1\exp(-c_2\log p)$,

$$\|\hat{\gamma}-\Sigma_x\beta_0\|_\infty \leq \frac{1}{1-\rho_{\max}}[t_1+t_2],$$

for all sufficiently large n. The claim (7.10) now follows from this bound, assumption **(A3ii)** and the assumption **(A2)** that ensures $0 \leq \rho_{\max} < 1$. □

Proof of Theorem 4.2. Recall that this theorem pertains only to the case of missing covariates. Part (i) of this theorem is a consequence of Theorem 1 of Loh and Wainwright (2012), who show that under the assumed conditions, $\|\hat{\beta}-\beta\|_2 \leq \lambda_n c_0\sqrt{s}/\alpha_1$, w.p. at least $1-c_1\exp(-c_2\log p)$, where α_1 is as in assumption **RE**. This result together with assumption **(A4)** implies that $T \subseteq \hat{T}$, with the same probability, for all sufficiently large n.

To prove part (ii) notice that by the first order optimality conditions $\left(\Gamma\hat{\beta}-\hat{\gamma}\right)_{\hat{T}} = \lambda_n$,

$$\begin{aligned}(7.12)\qquad \sqrt{\text{card}(\hat{T})}\lambda_n &\leq \left\|\left(\Gamma\hat{\beta}-\hat{\gamma}\right)_{\hat{T}}\right\|_2 \\ &\leq \sqrt{\text{card}(\hat{T})}\left(\left\|\Gamma\beta_0-\hat{\gamma}\right\|_\infty\right)+\left\|\left(\Gamma(\hat{\beta}-\beta_0)\right)_{\hat{T}}\right\|_2 \\ &= (I)+(II), \qquad \text{(say)}.\end{aligned}$$

Let v be a unit vector such that $\hat{\beta}-\beta_0 = \|\hat{\beta}-\beta_0\|_2 v$ and note that $\|v_{T^c}\|_0 \leq \hat{m}$. Consider the term (II) on the r.h.s of (7.12).

$$\begin{aligned}(7.13)\qquad &\left\|\left(\Gamma(\hat{\beta}-\beta_0)\right)_{\hat{T}}\right\|_2 \\ &\leq \|\hat{\beta}-\beta_0\|_2 \sup_{\|\delta_{T^c}\|_0\leq\hat{m},\|\delta\|_2\leq 1}\left|\delta'\Gamma v\right| \\ &\leq 3\|\hat{\beta}-\beta_0\|_2 \sup_{\|\delta_{T^c}\|_0\leq\hat{m},\|\delta\|_2\leq 1}\left|\delta'\Gamma\delta\right| \leq 3\phi(\hat{m})\|\hat{\beta}-\beta_0\|_2.\end{aligned}$$

Now consider term (I) of (7.12) for the case of missing covariates.

$$\begin{aligned}\left\|\Gamma\beta_0-\hat{\gamma}\right\|_\infty &\leq \|\Gamma\beta_0-\Sigma_x\beta_0\|_\infty+\|\hat{\gamma}-\Sigma_x\beta_0\|_\infty \\ &\leq \left\|\Big((n^{-1}Z'Z-\Sigma_z)\ominus M\Big)\beta_0\right\|_\infty \\ &\quad +\frac{1}{1-\rho_{max}}\Big(\left\|n^{-1}Z'X\beta_0-\Sigma_x\beta_0\right\|_\infty+\left\|n^{-1}Z'\varepsilon\right\|_\infty\Big) \\ &\leq c_0\frac{\sigma_x}{1-\rho_{\max}}\big(\sigma_\varepsilon+\frac{\sigma_x}{1-\rho_{\max}}\big)\|\beta_0\|_2\sqrt{\frac{\log p}{n}} \leq \lambda_n/2.\end{aligned}$$

Here the second inequality follows by basic algebra. The third inequality holds w.p. at least $1-c_1\exp(-c_2\log p)$ which follows by applying (7.2) of Lemma 7.1 separately on each of the three terms. The final inequality follows from the choice of λ_n under the missing covariate case. Combine this result with (7.13) and (7.12) to obtain

$$\sqrt{\text{card}(\hat{T})}\lambda_n[1-2^{-1}] \leq 3\phi(\hat{m})\|\hat{\beta}-\beta_0\|_2\,.$$

On the other hand by part (i), $\|\hat{\beta} - \beta_0\|_2 \leq \lambda_n c_0 \sqrt{s}/\alpha_1$, w.p. at least $1 - c_1 \exp(-c_2 \log p)$, for all sufficiently large n. These facts together with the fact $\hat{m} \leq \text{card}(\hat{T})$ readily imply $\sqrt{\hat{m}} \leq 4\phi(\hat{m}) c_0 \sqrt{s}/\alpha_1$, w.p. at least $1 - c_1 \exp(-c_2 \log p)$, for all sufficiently large n. This completes the proof of the Theorem 4.2. □

7.3 Proofs for Section 5

The proofs for this section shall require the following series of three lemmas. To proceed further we need to define

$$r = \begin{cases} \sigma_z \sigma_\varepsilon \ ; & \text{for additive errors} \\ \sigma_x \sigma_\varepsilon \ ; & \text{for missing covariates.} \end{cases} \tag{7.14}$$

The structure of the proof of the following two lemma's is similar to the proof of Lemma 7.2. All three results provide uniform bounds that hold in probability on different random quantities, for all sufficiently large n.

Lemma 7.3 *Let r be as in (7.14). Then, uniformly over all $m \leq n$ and for any $c_3 \in (0,1)$ and some universal constant D,*

$$\sup_{\delta \in \mathcal{B}(m),\, \|\delta\|_2 > 0} \Big| \frac{\delta' Z' \varepsilon}{n \|\delta\|_2} \Big| \leq c_0 \frac{3r}{2} e_n(m, c_3),$$

w.p. at least $1 - 2c_3 e^{-s}/(1 - 1/e)$, for all sufficiently large n.

Proof of Lemma 7.3 For every $U \subseteq \{1, ..., p\}$ with $\text{card}(U - T) \leq m$, define $S_U := \{\delta \in \mathbb{R}^p;\ \|\delta\|_2 \leq 1;\ \text{supp}(\delta) \subseteq U\}$ and note that $\mathcal{B}_1(m) = \cup_{card(U-T) \leq m} S_U$. Let $\mathcal{A} = \{u_1, ..., u_k\}$ be a $1/3$ cover of a fixed S_U, i.e., $\forall \delta \in S_U \ \exists\ u_i \in \mathcal{A}$ such that $\|\delta - u_i\|_2 \leq 1/3$. It is known from Ledoux and Talagrand (1991) or Loh and Wainwright (2012) (Supplementary materials, pg.17), that we can construct $\mathcal{A}$ such that $card(\mathcal{A}) \leq 9^{(m+s)}$. Then by elementary algebra,

$$n^{-1} \delta' Z' \varepsilon = n^{-1} u_i' Z' \varepsilon + n^{-1} (\delta - u_i)' Z' \varepsilon \tag{7.15}$$

By construction of $\mathcal{A}$, $3(\delta - u_i) \in S_U$ and using (7.15) we obtain,

$$\sup_{\delta \in S_U} \big| n^{-1} \delta' Z' \varepsilon \big| \leq \max_i \big| n^{-1} u_i' Z' \varepsilon \big| + \sup_{\delta \in S_U} \big| \frac{1}{3n} \delta' Z' \varepsilon \big|.$$

Hence $\sup_{\delta \in S_U} \big| n^{-1} \delta' Z' \varepsilon \big| \leq \max_i \big| \frac{3}{2n} u_i' Z' \varepsilon \big|$. Now applying Lemma 7.1, 9^{m+s} times, once for each $n^{-1} u_i' Z' \varepsilon$ and taking a union bound over all such possibilities we obtain,

$$P\left(\sup_{\delta \in S_U} \Big| n^{-1} \delta' Z' \varepsilon \Big| \geq \frac{3t}{2} \right) \leq 2 \cdot 9^{m+s} \cdot \exp\Big[-\frac{1}{c_0} n \min\big(\frac{t^2}{(\sigma_z \sigma_\varepsilon)^2}, \frac{t}{\sigma_z \sigma_\varepsilon} \big) \Big].$$

Again taking the union bound over all $\binom{p}{m} \leq p^m$ possibilities of U we obtain

$$P\left(\sup_{\delta \in \mathcal{B}_1(m)} \Big| n^{-1} \delta' Z' \varepsilon \Big| \geq \frac{3t}{2} \right) \leq 2 \cdot 9^{m+s} \cdot p^m \cdot \exp\Big[-\frac{1}{c_0} n \min\big(\frac{t^2}{(\sigma_z \sigma_\varepsilon)^2}, \frac{t}{\sigma_z \sigma_\varepsilon} \big) \Big].$$

Choose $t = re_n(m, c_3)$ to obtain,

$$P\left(\sup_{\delta\in\mathcal{B}_1(m)}\left|n^{-1}\delta' Z'\varepsilon\right| \geq \frac{3t}{2}\right) \leq 2c_3\exp[-m-s].$$

and thus

$$\begin{aligned} P\left(\sup_{\delta\in\mathcal{B}_1(m)}\left|n^{-1}\delta' Z'\varepsilon\right| \geq \frac{3t}{2}, \text{ for any } m\right) &\leq 2c_3\sum_{m=0}^{\infty}\exp[-m-s] \\ &= 2c_3\exp(-s)/(1-1/e), \end{aligned}$$

thereby completing the proof of the lemma. □

Lemma 7.4 *Let r be as in (7.14). Then uniformly over all $m \leq n$ and for any $c_3 \in (0,1)$ and some universal constant D we have,*

$$\sup_{\delta\in\mathcal{B}(m),\,\|\delta\|_2>0}\frac{1}{\|\delta\|_2}\left|\delta'\big[\frac{Z'W}{n}-\Sigma_x\big]\beta_0\right| \leq c_0\frac{3r}{2}\|\beta_0\|_2 e_n(m,c_3), \tag{7.16}$$

w.p. at least $1-2c_3e^{-s}/(1-1/e)$, for all sufficiently large n.

Proof of Lemma 7.4. Following the same idea as in the proof of Lemma 7.3, construct an 1/3 cover $\mathcal{A}$ of S_U for each U and let $\left|\delta'\big[\frac{Z'W}{n}-\Sigma_w\big]\beta_0\right| = \psi_{zw}(\delta,\beta_0)$. Then

$$\psi_{zw}(\delta,\beta^0) = \psi_{zw}(u_i,\beta_0) + \psi_{zw}(\delta-u_i,\beta_0).$$

This in turn implies that

$$\begin{aligned} \sup_{\delta\in S_U}\left|\psi_{zw}(\delta,\beta_0)\right| &\leq \max_i\left|\psi_{zw}(u_i,\beta_0)\right| + \frac{1}{3}\sup_{\delta\in S_U}\left|\psi_{zw}(\delta,\beta_0)\right|, \text{ or} \\ \sup_{\delta\in S_U}\left|\psi_{zw}(\delta,\beta_0)\right| &\leq \frac{3}{2}\max_i\left|\psi_{zw}(u_i,\beta_0)\right|. \end{aligned}$$

Now, use Lemma 7.1 to obtain

$$P\left(\sup_{\delta\in\mathcal{B}_1}\left|\psi_{zw}(\delta,\beta_0)\right| > \frac{3t}{2}\right) \leq 2\,9^{m+s}\,p^m\,\exp\Big[-c_0 n\min\big(\frac{t^2}{(\sigma_z\sigma_w)^2},\frac{t}{\sigma_z\sigma_w}\big)\Big].$$

Choosing $t = c_0\frac{3r}{2}e_n(m,c_3)$ we obtain,

$$\begin{aligned} P\left(\sup_{\delta\in\mathcal{B}_1(m)}\left|\psi_{zw}(\delta,\beta_0)\right| > \frac{3t}{2}, \text{ for any } m\leq n\right) &\leq 2c_3\sum_{m=0}^{\infty}\exp[-m-s] \\ &= 2c_3e^{-s}/(1-1/e). \end{aligned} \tag{7.17}$$

Thus we obtain uniformly over any $\delta \in \mathcal{B}_1(m)$ and $m \leq n$,

$$\Big|\delta'\big[\frac{Z'W}{n} - \mathrm{cov}(z_i, w_i)\big]\beta_0\Big| \leq c_0\frac{3r}{2}\|\beta_0\|_2 e_n(m, c_3), \tag{7.18}$$

w.p. at least $1 - 2c_3e^{-s}/(1-1/e)$, for all sufficiently large n, thereby completing the proof. □

Proof of Lemma 5.1. Case 1: Additive error: We have

$$\Big|\delta'(\Gamma^{\mathrm{add}}\beta_0 - \hat{\gamma}^{\mathrm{add}})\Big| \leq \Big|\frac{\delta'Z'\varepsilon}{n}\Big| + \Big|\delta'\big[\frac{Z'W}{n} - \Sigma_w\big]\beta_0\Big|. \tag{7.19}$$

Apply Lemmas 7.3 and 7.4 to the two terms on the r.h.s. of this bound and substitute back in (7.19) to obtain the desired result.

Case 2: Missing Covariates: Proceeding as in **Case 2** of the proof of Theorem 4.2, we obtain

$$\begin{aligned}
\big|\delta'\Gamma^{\mathrm{miss}}\beta_0 - \hat{\gamma}^{\mathrm{miss}}\big| &\leq \big|\delta'\big[\Gamma^{\mathrm{miss}} - \Sigma_x\big]\beta_0\big| + \big|\delta'\big(\hat{\gamma}^{\mathrm{miss}} - \Sigma_x\beta_0\big)\big| \\
&\leq \frac{1}{(1-\rho_{max})^2}\big|\delta'\big(n^{-1}Z'Z - \Sigma_z\big)\beta_0\big| \\
&\quad + \frac{1}{1-\rho_{max}}\left(\Big|\delta'\big(\frac{Z'X\beta_0}{n} - \Sigma_x\beta_0\big)\Big| + \big|\delta' n^{-1}Z'\varepsilon\big|\right).
\end{aligned}$$

The claim of the lemma again follows by applying Lemmas 7.3 and 7.4 to the last expression. □

Lemma 7.5 *Let*

$$r = \begin{cases} \sigma_z(\sigma_w + \sigma_\varepsilon) & ; \quad \textit{for additive error} \\ \frac{\sigma_x}{1-\rho_{\max}}\big(\sigma_\varepsilon + \frac{\sigma_x}{1-\rho_{\max}}\big) & ; \quad \textit{for missing covariates,} \end{cases} \tag{7.20}$$

Then uniformly over all $\delta \in \mathcal{B}(m)$, $m \leq n$ and for any $c_3 \in (0,1)$ and some universal constant D we have,

$$\big|\hat{Q}_n(\beta_0 + \delta) - \hat{Q}_n(\beta_0) - \delta'\Gamma\delta/2\big| \leq c_0\|\delta\|_2\|\beta_0\|_2 r e_n(m, c_3),$$

w.p. at least $1 - 6c_3e^{-s}/(1-1/e)$.

Proof of Lemma 7.5. This lemma is a straightforward consequence of Lemma 5.1. Using the definition of $\hat{Q}_n(\cdot)$ we obtain,

$$\big|\hat{Q}(\beta_0 + \delta) - \hat{Q}(\beta_0) - \delta'\Gamma\delta/2\big| \;=\; \big|\delta'(\Gamma\beta_0 - \hat{\gamma})\big|$$

Lemma 5.1 applied to the r.h.s. of this equation yields the desired result. □

Proof of Theorem 5.1. Let $\tilde{\delta} = \tilde{\beta} - \beta_0$, then by Lemma 7.5, w.p. at least $1 - 6c_3\exp(-s)/(1-1/e)$,

$$\left|\hat{Q}(\beta_0+\tilde{\delta}) - \hat{Q}(\beta_0) - \tilde{\delta}'\Gamma\tilde{\delta}/2\right| \leq c_0 r\|\tilde{\delta}\|_2\|\beta_0\|_2 e_n(m,c_3).$$

Also by the model selection step, $T \subseteq \hat{T}$ w.p. at least $1 - c_1\exp(-c_2\log p)$. Hence, on this set, by the definition of the second step estimator, $\hat{Q}(\tilde{\beta}) - \hat{Q}(\beta_0) \leq 0$. This in turn implies that w.p. at least $1 - 6c_3\exp(-s)/(1-1/e) - c_1\exp(-c_2\log p)$,

$$-\tilde{\delta}'\Gamma\tilde{\delta}/2 \geq -c_0 r\|\tilde{\delta}\|_2\|\beta_0\|_2 e_n(m,c_3). \tag{7.21}$$

An application of condition RSE($\hat{m}$) in the inequality (7.21) yields

$$\|\tilde{\delta}\|_2 \leq \frac{c_0 r}{\kappa_x(\hat{m})}\|\beta_0\|_2 e_n(m,c_3), \tag{7.22}$$

w.p. at least $1 - 6c_3\exp(-s)/(1-1/e) - c_1\exp(-c_2\log p)$. This completes the proof of this Theorem. □

The **Proof of Theorem 5.3** shall rely on the following two results. First is Lemma 6 of the supplement of Loh and Wainwright (2012), which is restated below for the convenience of the reader.

Lemma 7.6 *For each* $1 \leq j \leq p$,

$$\frac{1}{\lambda_{\max}(\Sigma)} \leq |d_j| \leq \frac{1}{\lambda_{\min}(\Sigma)} \quad \textit{and} \quad \|\theta^j\|_2 \leq \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$$

Lemma 7.7 *Under the conditions of Theorem 5.3 the following hold.*

(i) $|\hat{d}_j - d_j| \leq c_0 C_2 e_n(a_n - s, c_3)$,

(ii) $\|\tilde{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_2 \leq c_0\Big(C_2^2 + \dfrac{C_1^2}{\lambda_{\min}^2(\Sigma)} + \dfrac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}C_2\Big)^{1/2} e_n(a_n - s, c_3)$,

for all $1 \leq j \leq p$, *w.p. converging to* 1.

Proof of Lemma 7.7 Let $\hat{m} = a_n - s$ and observe that in view of Theorem 4.1 and Theorem 5.1 we have for all $1 \leq j \leq p$,

$$\|\hat{\theta}^j - \theta^j\|_2 \leq \frac{1}{\kappa(\hat{m})} c_0 r\|\theta^j\|_2 e_n(\hat{m},c_3) := c_0 C_1 e_n(\hat{m},c_3), \tag{7.23}$$

w.p. converging to 1. Also, note that by the additional parameter space restriction in the construction of (5.8), $\|\hat{\theta}^j\|_1 \leq b_0\sqrt{s}$. Consider

$$\begin{aligned} |\hat{d}_j^{-1} - d_j^{-1}| &= \left|\left(\hat{\Sigma}_{jj} - \hat{\Sigma}_{j,-j}\hat{\theta}^j\right) - \left(\Sigma_{jj} - \Sigma_{j,-j}\theta^j\right)\right| \\ &\leq \left|\hat{\Sigma}_{jj} - \Sigma_{jj}\right| + \left|\hat{\Sigma}_{j,-j}\hat{\theta}^j - \Sigma_{j,-j}\theta^j\right| := (I) + (II), \text{ (say).} \end{aligned} \tag{7.24}$$

By assumption we have that $(I) \leq c_0\sigma_x\sqrt{\log p/n}$. Now consider the term (II) on the r.h.s of (7.24),

$$\begin{aligned}(II) &\leq \Big|\big(\hat{\Sigma}_{j,-j} - \Sigma_{j,-j}\big)\hat{\theta}^j\Big| + \Big|\Sigma_{j,-j}\big(\hat{\theta}^j - \theta^j\big)\Big| \\ &\leq \|\hat{\Sigma} - \Sigma\|_\infty \|\hat{\theta}^j\|_1 + \|\Sigma_{j,-j}\|_2 \|\hat{\theta}^j - \theta^j\|_2 \\ &\leq c_0\Big(b_0\sigma_x + \frac{\lambda_{\max}}{\lambda_{\min}} r\|\theta^j\|_2\Big) e_n(\hat{m}, c_3). \qquad (7.25)\end{aligned}$$

Combining the bounds for terms (I) and (II) we obtain for all $1 \leq j \leq p$,

$$|\hat{d}_j^{-1} - d_j^{-1}| \leq c_0\Big(b_0\sigma_x + \frac{\lambda_{\max}}{\lambda_{\min}} r\|\theta^j\|_2\Big) e_n(\hat{m}, c_3).$$

Thus applying Lemma 7.6 we obtain,

$$\Big|\frac{d_j}{\hat{d}_j} - 1\Big| \leq |d_j||\hat{d}_j^{-1} - d_j^{-1}| \leq c_0 \frac{1}{\lambda_{\min}}\Big(b_0\sigma_x + \frac{\lambda_{\max}}{\lambda_{\min}} r\|\theta^j\|_2\Big) e_n(\hat{m}, c_3).$$

This in turn implies that $|\hat{d}_j| \leq 2|d_j|$ for n sufficiently large, and hence

$$\begin{aligned}|\hat{d}_j - d_j| \leq |\hat{d}_j|\Big|\frac{d_j}{\hat{d}_j} - 1\Big| &\leq c_0 \frac{1}{\lambda_{\min}^2}\Big(b_0\sigma_x + \frac{\lambda_{\max}}{\lambda_{\min}} r\|\theta^j\|_2\Big) e_n(\hat{m}, c_3), \\ &:= c_0 C_2 e_n(\hat{m}, c_3),\end{aligned}$$

for n sufficiently large. This proves part (i) of this lemma. To prove (ii) consider,

$$\begin{aligned}\|\tilde{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_2^2 &= |\hat{d}_j - d_j|^2 + \|\hat{d}_j\hat{\theta}^j - d_j\theta^j\|_2^2 \\ &\leq |\hat{d}_j - d_j|^2 + 2|d_j|^2\|\hat{\theta}^j - \theta^j\|_2^2 + 2|\hat{d}_j - d_j|^2\|\hat{\theta}^j\|_2^2 \\ &\leq c_0^2\Big(C_2^2 + \frac{C_1^2}{\lambda_{\min}^2(\Sigma)} + \frac{\lambda_{\max}}{\lambda_{\min}} C_2\Big) e_n^2(\hat{m}, c_3)\end{aligned}$$

This completes the proof of the lemma. □

Proof of Theorem 5.3 This proof is a direct consequence of Lemma 7.7 by observing that

$$\|\hat{\Theta} - \Theta\|_2^2 \leq 2\|\hat{\Theta} - \tilde{\Theta}\|_2^2 + 2\|\tilde{\Theta} - \Theta\|_2^2 \leq 4 \max_j \|\tilde{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_2^2. \qquad \square$$

References

1. Agarwal, A., Neghban, S. and Wainwright, M.J. (2012). Fast Global Convergence of gradient methods for High Dimensional Statistical Recovery. *Ann. Statist.* **40**, 2452–2482.

2. Belloni, A., and Chernozhukov, V. (2013). Least Squares After Model Selection in High Dimensional Sparse Models, *Bernoulli.* **19**, 521–547.

3. Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous Analysis of Lasso and Dantzig Selector, *Ann. Statist.*, **37**, 1705–1732.

4. Bickel, P., Levina, E. (2008). Covariance Regularization by Thresholding, *Ann. Statist.*, **36**, 2577–2604.

5. Bühlmann, P. and van de Geer, S. (2011). *Statistics for High Dimensional Data*. Springer-Verlag, Berlin Heidelberg.

6. Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2009). Variable Selection in High Dimensional Linear Models; Partially Faithful Distributions and the PC-Simple Algorithm. *Biometrika*. **97**, 261-278.

7. Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York.

8. Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space . *J.R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911.

9. Friedman, J., Hastie, T., Simon, N., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Statist. Software*, **33**, 1–22.

10. Friedman, J., Hastie, T., Simon, N., Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**, 432–441.

11. Fuller, W.A. (1987). *Measurement Error Models*. Wiley & Sons, Inc. New York.

12. Genovese, C., Jin, J., Wasserman, L., Yao, Z. (2012). A Comparison of the Lasso and Marginal Regression, *J. of Mach. Learn. Res.*, **13**, 2107–2143.

13. Kaul, A. and Koul, H. (2015). Weighted ℓ_1-Penalized Corrected Quantile Regression for High Dimensional Measurement Error Models. *J. Mult. Analysis.*, **140**, 72–91.

14. Ledoux, M. and Talagrand, M. (1991). Probability in Banach Spaces: Isoperimetry and Processes. Springer-Verlag, New York.

15. Liang, H. and Li, R. (2009). Variable Selection for Partially Linear Models with Measurement Errors. *J. Amer. Statist. Assoc.*, **104**, 234–248.

16. Loh, P., and Wainwright, M.J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics* , **40**, 1637–1664.

17. Meinhausen, N. and Búhlmann, P. (2006) High Dimensional graphs and Variable Selection with Lasso. *Annals of Statistics*, **34**, 1436–1462.

18. Pang, H., Liu, H., and Vanderbei, R. (2014). The fastclime Package for Linear Programming and Large-Scale Precision Matrix Estimation in R. *J. Mach. Learn. Res.*, **15** 489-493.

19. Rosenbaum, M. and Tsybakov, A.B. (2010). Sparse recovery under matrix uncertainty. *Annals of Statistics*, **38** 2620–2651.

20. Rosenbaum, M. and Tsybakov, A.B. (2011). Improved matrix uncertainty selector, *Technical Report*. Available at http://arxiv.org/abs/1112.4413.

21. Sørensen, Ø., Frigessi, A., and Thoresen, M. (2014). Covariate Selection in High-Dimensional Generalized Linear Models With Measurement Error. Available at http://arxiv.org/abs

22. Sørensen, Ø., Frigessi, A., and Thoresen, M. (2015). Measurement error in lasso: impact and likelihood bias correction. *Statist. Sinica*, **25(2)**, 809–829.

23. van der Waart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

24. Tibshirani, R.J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.*, **7**, 1456–1490.

25. Vershynin. (2012). Introduction to the Non-Asymptotic Analysis of Random Matrices. *Chapter 5 of Compressed Sensing: Theory and Applications*. Cambridge University Press.

26. Yuan, M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Sparse Linear Programming, *J. of Mach. Learn. Res.*, **11** 2261–2286.