Noisy and Missing Data Regression
Supplementary Material

Yudong Chen and Constantine Caramanis
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
ydchen@utexas.edu, caramanis@mail.utexas.edu

Abstract

In this supplementary material, we present the complete details for all the proofs. Also, we provide further numerical simulations.

1 Introduction

In this material, we provide the proofs for the theorems in the submission, as well as further numerical simulations.

As we have noted in the submission, once the support of regressor β^* has been identified, the problem of estimating its non-zero coefficients reduces to one from the classical low-dimensional regime; that is, we need to estimate k values from n linear observations, where $n \gtrsim k$. Therefore, our bounds on the ℓ^2 estimation errors (Theorem 5 and part 2 of Theorem 7 in the main submission) follow from guarantees on support recovery and ℓ^2 error bounds for the low-dimensional problem.

The remainder of this material is organized as follows. In Section 2 we state and prove ℓ^2 error bounds for the low-dimensional problem. In Section 3 we prove our guarantees for support recovery in the high-dimensional regimes, and combine it with the low-dimensional results to obtain ℓ^2 error bounds. We also prove the minimax lower-bounds in this section. Section 4 provides additional numerical simulations.

2 The Low-Dimensional Problem

We first consider the low-dimensional version of the problem where $\beta^* \in \mathbb{R}^k$, with $k \lesssim n$. As noted above, in the high-dimensional sparse-regression setting, once we know the support of β^*, this is precisely the resulting problem. Recall that our basic assumptions in the main submission is that X, W and e obey the Sub-Gaussian Design Model, which is restated below:

Definition 1. Sub-Gaussian Design Model: We assume X, W and $\mathbf{e}$ are sub-Gaussian with parameters $(\frac{1}{n}\Sigma_x, \frac{1}{n})$, $(\frac{1}{n}\Sigma_w, \frac{1}{n}\sigma_w^2)$ and $(\frac{1}{n}\sigma_e^2, \frac{1}{n}\sigma_e^2)$, respectively. We assume they are independent of each other.

We note that in this section our results require no assumptions on the independence of the columns of X, W, or, therefore, of Z; that is, we assume we operate under the sub-Gaussian design model. When $k \lesssim n \log p$, the problem is strongly convex, and in the clean-covariate setting where we know X exactly and completely, the solution is given by the standard least-square estimator:

$$\hat{\beta} \quad = \quad (X^\top X)^{-1} X^\top \mathbf{y} = \arg\min_{\beta} \beta^\top (X^\top X)\beta - 2\mathbf{y}^\top X\beta. \tag{1}$$

In this setting, well-known results establish, among other measures of closeness to β^*, the following (here and in the sequel $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of the matrix A):

Theorem 2 ([6]). *Suppose that (according to the sub-Gaussian design model defined above) X is sub-Gaussian with parameters $(\frac{1}{n}\Sigma_x, \frac{1}{n})$, and the noise vector $\mathbf{e}$ is sub-Gaussian with parameters $(\frac{1}{n}\sigma_e^2, \frac{1}{n}\sigma_e^2)$. Moreover, suppose that $n \gtrsim \frac{k \log p}{\lambda_{\min}(\Sigma_x)}$. Then with high probability, the estimator above satisfies:*

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \frac{\sigma_e}{\lambda_{\min}(\Sigma_x)} \sqrt{\frac{k \log p}{n}}.$$

When we know only Z (a noisy or partially deleted version of X), we consider a modified version of the estimator. Let us generically denote by $\hat{\Sigma}$ the estimator for $X^\top X$, and by $\hat{\gamma}$ our estimate for $X^\top \mathbf{y}$. Thus, in place of $\hat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}$ given in (1), our proposed estimator for $\hat{\beta}$ naturally becomes:

$$\hat{\beta} \quad = \quad (\hat{\Sigma})^{-1}\hat{\gamma} = \arg\min_{\beta} \beta^\top(\hat{\Sigma})\beta - 2\hat{\gamma}^\top\beta, \tag{2}$$

where we require $\hat{\Sigma}$ to be positive semidefinite. For this estimator, we have the following simple but general result.

Theorem 3. *Suppose the following strong convexity condition holds: $\lambda_{\min}\left(\hat{\Sigma}\right) \geq \lambda > 0$. Then the estimation error satisfies: $\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \frac{1}{\lambda}\left\|\hat{\gamma} - \hat{\Sigma}\beta^*\right\|_2$.*

Proof. Let $\Delta = \hat{\beta} - \beta^*$. By optimality of $\hat{\beta}$, we have $(\beta^* + \Delta)^\top(\hat{\Sigma})(\beta^* + \Delta) - 2\hat{\gamma}^\top(\beta^* + \Delta) \leq \beta^{*\top}(\hat{\Sigma})\beta - 2\hat{\gamma}^\top\beta^*$. Rearranging terms gives $\Delta^\top\hat{\Sigma}\Delta \leq 2(\hat{\gamma}^\top - \beta^{*\top}\hat{\Sigma})\Delta$. Under the strong convexity assumption, the l.h.s. is lower-bounded by $\lambda\|\Delta\|_2^2$. The r.h.s. is upper-bounded by $2\left\|\hat{\gamma} - \hat{\Sigma}\beta^*\right\|_2 \|\Delta\|_2$ thanks to Cauchy-Schwarz. The result then follows.[1] □

This result is simple and generic. We specialize to the case of additive noise, and missing variables. In particular, the pair $(\hat{\Sigma}, \hat{\gamma})$ depends on the assumption of what is known.

Additive Noise

For additive noise, the models we use are as follows.

1. Knowledge of Σ_w: we assume we either know or somehow can estimate the noise covariance, $\Sigma_w = \mathbb{E}\left[W^\top W\right]$.
2. Knowledge of Σ_x: in this case, we assume that we either know or somehow can estimate the covariance of the true covariates, $\Sigma_x = \mathbb{E}\left[X^\top X\right]$.
3. Instrumental Variables: in this setting, we assume there are variables $U \in \mathbb{R}^{n\times m}$ with $m \geq k$, whose rows are correlated with the rows of X, but independent of W and $\mathbf{e}$, and that the realization of U is known or can be estimated. Instrumental variables are common in the econometrics literature [2, 1], and are often used when X is not available. In [3] the authors consider instrumental variables for high-dimensional problems when X is observed without noise. To the best of our knowledge, no rigorous finite sample results have been obtained for this approach when one has available a noisy or partially erased version of the covariate matrix X.

We have the following results for the case of additive noise: $Z = X + W$.

[1] Our notation "$\lesssim$" means that we ignore constants that are independent of any scaling variables. We use this throughout.

Corollary 4 (Knowledge of Σ_w)**.** *Suppose $n \gtrsim \frac{(1+\sigma_w^2)^2}{\lambda_{\min}(\Sigma_x)} k \log p$. Then, w.h.p., the estimator built using $\hat{\Sigma} = Z^\top Z - \Sigma_w$ and $\hat{\gamma} = Z^\top y$, satisfies*

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \frac{\left(\sigma_w + \sigma_w^2\right) \|\beta^*\|_2 + \sigma_e\sqrt{1+\sigma_w^2}}{\lambda_{\min}(\Sigma_x)} \sqrt{\frac{k \log p}{n}}.$$

Remark. (1) When $\sigma_w = 0$, the bound reduces to the standard bound for the least-squares estimator, and it implies exact recovery when $\sigma_w = \sigma_e = 0$. (2) If we only have an upper bound, $\overline{\Sigma}_w \succeq \Sigma_w$, then using the same analysis one can show:

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \frac{\left[\left(\sigma_w + \sigma_w^2\right) \|\beta^*\|_2 + \sigma_e\sqrt{1+\sigma_w^2}\right] \sqrt{\frac{k \log p}{n}} + \lambda_{\max}(\overline{\Sigma}_w - \Sigma_w)\|\beta^*\|_2}{\lambda_{\min}(\Sigma_x) - \lambda_{\max}(\overline{\Sigma}_w - \Sigma_w)}.$$

Corollary 5 (Knowledge of Σ_x)**.** *Suppose $n \gtrsim \log p$. Then, w.h.p., the estimator built using $\hat{\Sigma} = \Sigma_x$ and $\hat{\gamma} = Z^\top y$, satisfies*

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \frac{(1+\sigma_w) \|\beta^*\|_2 + \sigma_e\sqrt{1+\sigma_w^2}}{\lambda_{\min}(\Sigma_x)} \sqrt{\frac{k \log p}{n}}.$$

Remark. The bound is linear for σ_w large, but it does *not* vanish when σ_w and σ_e are zero.

Let $\sigma_i(A)$ denote the i-th singular value of A, so, e.g., $\sigma_1(A) = \sigma_{\max}(A)$, the largest singular value of A.

Corollary 6 (Instrumental Variables)**.** *Suppose the Instrumental Variable $U \in \mathbb{R}^{n\times m}$ is zero-mean sub-Gaussian with parameter (Σ_U, σ_u^2), and $\mathbb{E}\left[U^\top X\right] = \Sigma_{UX}$. Let $\sigma_1 = \sigma_1(\Sigma_{UX})$ and $\sigma_k = \sigma_k(\Sigma_{UX})$. If $n \gtrsim \max\left\{1, \frac{\sigma_u^2(1+\sigma_w^2)}{(m/k)\sigma_k^2}\right\} k \log p$, then w.h.p. the estimator built using $\hat{\Sigma} = Z^\top UU^\top Z$, and $\hat{\gamma} = Z^\top UU^\top y$, satisfies*

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \sqrt{\sigma_w^2 \|\beta^*\|^2 + \sigma_e^2} \frac{\sigma_1\sigma_u}{\sigma_k^2\sqrt{k/m}} = \frac{\sqrt{\sigma_w^2 \|\beta^*\|^2 + \sigma_e^2}}{(\sigma_1/\sigma_u)\sqrt{\frac{k}{m}}} \cdot \frac{1}{(\sigma_k/\sigma_1)^2} \sqrt{\frac{k \log p}{n}}.$$

Remark. The first factor can be interpreted as 1/SNR, and the second is a measure of the correlation between X and U (i.e., the strength of the Instrumental Variable).

Missing Data

For missing data, we assume the erasure probability ρ is known or can be accurately estimated. We use $\hat{\Sigma} = (Z^\top Z) \odot M$ and $\hat{\gamma} = \frac{1}{(1-\rho)} Z^\top y$, where $M_{ij} = \frac{1}{1-\rho}$ if $i = j$ or $\frac{1}{(1-\rho)^2}$ otherwise, and $\odot$ denotes element-wise product. We then have:

Corollary 7 (Missing Data)**.** *If $n \gtrsim \frac{1}{(1-\rho)^4 \lambda_{\min}^2(\Sigma_x)} k \log p$, w.h.p. our estimator satisfies*

$$\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \left(\frac{1}{(1-\rho)^2} \|\beta^*\|_2 + \frac{1}{1-\rho}\sigma_e\right) \frac{1}{\lambda_{\min}(\Sigma_x)} \sqrt{\frac{k \log p}{n}}.$$

Remark. Note that as with the previous results, the dependence on $\|\beta^*\|_2$ is given explicitly.

We now prove Corollary 4-7. The proofs rely on several supporting concentration results in the next subsection; these results will also be used in the proof of the high-dimensional results in the next section.

2.1 Supporting Concentration Results

We state some supporting concentration results, and postpone their proofs to the appendix.

Lemma 8. *[4, Lemma 14] Suppose $Y \in \mathbb{R}^{n\times k}$ is a zero mean sub-Gaussian matrix with parameter $(\frac{1}{n}\Sigma, \frac{1}{n}\sigma^2)$. If $n \gtrsim \log p \geq \log k$, then*

$$\mathbb{P}\left(\left\|X^\top X - \Sigma\right\|_\infty \geq c_0\sigma^2\sqrt{\frac{\log p}{n}}\right) \leq c_1 \exp\left(-c_2 \log p\right).$$

Lemma 9. *Suppose $X \in \mathbb{R}^{n\times k}$, $Y \in \mathbb{R}^{n\times m}$ are zero-mean sub-Gaussian matrices with parameters $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$, $(\frac{1}{n}\Sigma_y, \frac{1}{n}\sigma_y^2)$. Then for any fixed vectors $\mathbf{v}_1$, $\mathbf{v}_2$, we have*

$$\mathbb{P}\left(\left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \geq t\left\|\mathbf{v}_1\right\|\left\|\mathbf{v}_2\right\|\right) \leq 3\exp\left(-cn\min\left\{\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\right\}\right).$$

In particular, if $n \gtrsim \log p \geq \log m \vee \log k$, we have w.h.p.

$$\left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \leq \sigma_x\sigma_y\left\|\mathbf{v}_1\right\|\left\|\mathbf{v}_2\right\|\sqrt{\frac{\log p}{n}}.$$

Setting $\mathbf{v}_1$ to be the i^{th} standard basis vector, and using a union bound over $i = 1, \ldots, m$, we have w.h.p.

$$\left\|\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)v\right\|_\infty \leq \sigma_x\sigma_y\left\|v\right\|\sqrt{\frac{\log p}{n}}.$$

As a simple corollary of this lemma, we get the following.

Corollary 10. *If $X \in \mathbb{R}^{n\times k}$ is a zero-mean sub-Gaussian matrix with parameter $(\frac{1}{n}\sigma_x^2 I, \frac{1}{n}\sigma_x^2)$, and $\mathbf{v}$ is a fixed vector in $\mathbb{R}^n$, then for any $\epsilon \geq 1$, we have*

$$\mathbb{P}\left(\left\|X^\top\mathbf{v}\right\|_2 > \sqrt{\frac{(1+\epsilon)k}{n}}\sigma_x\left\|\mathbf{v}\right\|_2\right) \leq 3\exp\left(-ck\epsilon\right).$$

Lemma 11. *If $X \in \mathbb{R}^{n\times k}$, $Y \in \mathbb{R}^{n\times m}$ are zero mean sub-Gaussian matrices with parameter $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$,$(\frac{1}{n}\Sigma_y, \frac{1}{n}\sigma_y^2)$, then*

$$\mathbb{P}\left(\sup_{\mathbf{v}_1\in\mathbb{R}^m, \mathbf{v}_2\in\mathbb{R}^k, \|\mathbf{v}_1\|=\|\mathbf{v}_2\|=1}\left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \geq t\right) \leq 2\exp\left(-cn\min\left(\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\right) + 6(k+m)\right).$$

In particular, for each $\lambda > 0$, if $n \gtrsim \max\left\{\frac{\sigma_x^2\sigma_y^2}{\lambda^2}, 1\right\}(k+m)\log p$, then w.h.p.

$$\sup_{\mathbf{v}_1\in\mathbb{R}^m, \mathbf{v}_2\in\mathbb{R}^k}\left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \leq \frac{1}{54}\lambda\left\|\mathbf{v}_1\right\|\left\|\mathbf{v}_2\right\|.$$

2.2 Proof of Corollary 4

Substituting $Z = X + W$ into the definition of $\hat{\gamma}$ and $\hat{\Sigma}$, we obtain

$$\begin{aligned}\left\|\hat{\gamma} - \hat{\Sigma}\beta^*\right\|_\infty &= \left\|-X^\top W\beta^* + Z^\top e - (W^\top W - \Sigma_w)\beta^*\right\|_\infty \\ &\leq \left\|X^\top W\beta^*\right\|_\infty + \left\|Z^\top e\right\|_\infty + \left\|(W^\top W - \Sigma_w)\beta^*\right\|_\infty\end{aligned}$$

Using Lemma 9, we have w.h.p.

$$\begin{aligned}\left\|X^\top W\beta^*\right\|_\infty &\leq \sigma_w\left\|\beta\right\|_2\sqrt{\frac{\log p}{n}} \\ \left\|Z^\top e\right\|_\infty &\leq \sigma_e\sqrt{1+\sigma_w^2}\sqrt{\frac{\log p}{n}} \\ \left\|(W^\top W - \Sigma_w)\beta^*\right\|_\infty &\leq \sigma_w^2\left\|\beta\right\|_2\sqrt{\frac{\log p}{n}}.\end{aligned}$$

It follows that

$$\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_2 \leq \sqrt{k}\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_\infty \leq \left[\left(\sigma_w+\sigma_w^2\right)\|\beta^*\|_2+\sigma_e\sqrt{1+\sigma_w^2}\right]\sqrt{\frac{k\log p}{n}}.$$

On the other hand, observe that Z is sub-Gaussian with parameter $(\frac{1}{n}\Sigma_x+\frac{1}{n}\Sigma_w, \frac{1}{n}(1+\sigma_w^2))$. When $n \gtrsim \frac{(1+\sigma_w^2)^2 k\log p}{\lambda_{\min}(\Sigma_x)}$, by Lemma 11 with $\lambda=\lambda_{\min}(\Sigma_x)$, we have $\lambda_1\left(Z^\top Z-(\Sigma_x+\Sigma_w)\right) \leq \frac{1}{54}\lambda_{\min}(\Sigma_x)$ w.h.p. It follows that

$$\begin{aligned}\lambda_{\min}\left(\hat{\Sigma}\right) = \inf_{\|v\|=1} v^\top\hat{\Sigma}v &= \inf_{\|v\|=1} v^\top\left(\Sigma_x+Z^\top Z-(\Sigma_x+\Sigma_w)\right)v \\ &\geq \lambda_{\min}(\Sigma_x)-\lambda_1\left(Z^\top Z-(\Sigma_x+\Sigma_w)\right) \\ &\geq \frac{1}{2}\lambda_{\min}(\Sigma_x).\end{aligned}$$

The corollary then follows by applying Theorem 3.

2.3 Proof of Corollary 5

In this case, we have

$$\begin{aligned}\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_\infty &= \left\|(X^\top X-\Sigma_x)\beta^*+W^\top X\beta^*+Z^\top\mathbf{e}\right\|_\infty \\ &\leq \left\|W^\top X\beta^*\right\|_\infty+\left\|Z^\top\mathbf{e}\right\|_\infty+\left\|(X^\top X-\Sigma_x)\beta^*\right\|_\infty\end{aligned}$$

By Lemma 9, we have w.h.p.

$$\begin{aligned}\left\|W^\top X\beta^*\right\|_\infty &\leq \sigma_w\|\beta\|_2\sqrt{\frac{\log p}{n}} \\ \left\|Z^\top\mathbf{e}\right\|_\infty &\leq \sigma_e\sqrt{1+\sigma_w^2}\sqrt{\frac{\log p}{n}} \\ \left\|(X^\top X-\Sigma_x)\beta^*\right\|_\infty &\leq \|\beta^*\|_2\sqrt{\frac{\log p}{n}}.\end{aligned}$$

So $\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_2 \leq \sqrt{k}\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_\infty \lesssim \left[(1+\sigma_w)\|\beta^*\|_2+\sigma_e\sqrt{1+\sigma_w^2}\right]\sqrt{\frac{k\log p}{n}}$. On the other hand, by assumption $\lambda_{\min}(\hat{\Sigma}) = \lambda_{\min}(\Sigma_x)$. The corollary then follows by applying Theorem 3.

2.4 Proof of Corollary 6

First observe that

$$\begin{aligned}\lambda_{\min}(\hat{\Sigma}) &= \lambda_{\min}((X+W)^\top UU^\top(X+W)) \\ &= \sigma_k^2(U^\top X+U^\top W) \\ &= \sigma_k^2(\mathbb{E}\left[U^\top X\right]+(U^\top X-\mathbb{E}\left[U^\top X\right])+U^\top W) \\ &\geq \left[\sigma_k-\sigma_1(U^\top X-\mathbb{E}\left[U^\top X\right])-\sigma_1(U^\top W)\right]^2.\end{aligned}$$

By Lemma 11 with $\lambda=\sigma_k$, we have $\sigma_1(U^\top W) \leq \frac{1}{4}\sigma_k$ and $\sigma_1(U^\top X-\mathbb{E}\left[U^\top X\right]) \leq \frac{1}{4}\sigma_k$ under our assumption, so $\lambda_{\min}(\hat{\Sigma}) \geq \frac{1}{4}\sigma_k^2$. On the other hand,

$$\begin{aligned}\left\|\hat{\Sigma}\beta^*-\hat{\gamma}\right\|_2 &= \left\|(X+W)^\top UU^\top(W\beta^*-\mathbf{e})\right\|_2 \\ &\leq \left\|X^\top UU^\top(W\beta^*+\mathbf{e})\right\|_2+\left\|W^\top UU^\top(W\beta^*+\mathbf{e})\right\|_2.\end{aligned}$$

We bound each term.

1. By Lemma 11 with $\lambda = \sigma_1$, we have $\sigma_1(U^\top X) \leq \frac{3}{2}\sigma_1$. Each entry of $W\beta^* + e$ is i.i.d. zero-mean sub-Gaussian with variance bounded by $\sigma_w^2 \left\|\beta^*\right\|^2 + \sigma_e^2$. Hence by Lemma 9, $\left\|U^\top(W\beta^* + \mathbf{e})\right\|_2 \leq \sqrt{m}\left\|U^\top(W\beta^* + \mathbf{e})\right\|_\infty \leq \sigma_u\sqrt{\sigma_w^2 \left\|\beta^*\right\|^2 + \sigma_e^2}\sqrt{\frac{m\log p}{n}}$. It follows that $\left\|\left(U^\top X\right)^\top U^\top(W\beta^* + \mathbf{e})\right\|_2 \leq 2\sqrt{\sigma_w^2 \left\|\beta^*\right\|^2 + \sigma_e^2}\sqrt{\frac{\sigma_u^2\sigma_1^2 m\log p}{n}}$.
2. By Lemma 11 with $\lambda = \sigma_1$, we have $\left\|W^\top U\right\|_{op} \leq \sigma_1$ under the assumption, so the second term is bounded by $\sigma_w \left\|\beta^*\right\| \sqrt{\frac{\sigma_u^2\sigma_1^2 m\log p}{n}}$.

The result follows from applying Theorem 3.

2.5 Proof of Corollary 7

Let $\Sigma_z = \mathbb{E}\left[Z^\top Z\right]$; we have $(\Sigma_z)_{ij} = (1-\rho)(\Sigma_x)_{ij}$ for $i = j$ and $(\Sigma_z)_{ij} = (1-\rho)^2(\Sigma_x)_{ij}$ for $i \neq j$. Note that the observed matrix Z is sub-Gaussian with parameter $(\frac{1}{n}\Sigma_z, \frac{1}{n})$, which follows from the sub-Gaussianity of X (c.f. [4]). We set $\Delta_z = Z^\top Z - \Sigma_z$. By Lemma 8, we know $\max_i |(\Delta_z)_{ii}| \leq \frac{1}{4}(1-\rho)^2\lambda_{\min}(\Sigma_x)$ w.h.p. When this happens, for each unit norm v, we have

$$\begin{aligned} v^\top(\Delta_z \odot M)v &= \sum_{i,j} v_i v_j (\Delta_z)_{ij}\frac{1}{(1-\rho)^2} + \sum_i v_i^2(\Delta_z)_{ii}\left(\frac{1}{1-\rho} - \frac{1}{(1-\rho)^2}\right) \\ &\leq \frac{1}{(1-\rho)^2} v^\top\Delta_z v + \left\|v\right\|^2 \frac{\rho}{(1-\rho)^2}\max_i |(\Delta_z)_{ii}| \\ &\leq \frac{1}{(1-\rho)^2} v^\top\Delta_z v + \frac{\rho}{4}\lambda_{\min}(\Sigma_x). \end{aligned}$$

By Lemma 11 with $\lambda = \frac{1}{4}(1-\rho)^2\lambda_{\min}(\Sigma_x)$, we obtain $\max_{v:\|v\|=1} v^\top\Delta_z v \leq \frac{1}{4}(1-\rho)^2\lambda_{\min}(\Sigma_x)$, so $v^\top(\Delta_z \odot M)v \leq \frac{1}{4}(1+\rho)\lambda_{\min}(\Sigma_x)$. Because $\hat{\Sigma} = (\Sigma_z + Z^\top Z - \Sigma_z)\odot M = \Sigma_x + \Delta_z \odot M$, it follows that $\lambda_{\min}(\hat{\Sigma}) \geq \lambda_{\min}(\Sigma_x) - \lambda_1(\Delta_z \odot M) \geq \frac{1}{2}\lambda_{\min}(\Sigma_x)$.

On the other hand, observe that

$$\begin{aligned} \left\|\hat{\gamma} - \hat{\Sigma}\beta^*\right\|_\infty &\leq \left\|\hat{\gamma} - \Sigma_x\beta^*\right\|_\infty + \left\|(\hat{\Sigma} - \Sigma_x)\beta^*\right\|_\infty \\ &\leq \left\|\frac{1}{1-\rho}Z^\top X\beta^* - \Sigma_x\beta^*\right\|_\infty + \left\|\frac{1}{1-\rho}Z^\top\mathbf{e}\right\|_\infty + \left\|(\hat{\Sigma} - \Sigma_x)\beta^*\right\|_\infty. \end{aligned}$$

By Lemma 9, w.h.p. the first term is bounded by $\frac{1}{1-\rho}\left\|\beta^*\right\|\sqrt{\frac{\log p}{n}}$, and the second term is bounded by $\frac{1}{1-\rho}\sigma_e\sqrt{\frac{\log p}{n}}$. The magnitude of the i-th term of $(\hat{\Sigma} - \Sigma_x)\beta^*$ is

$$\begin{aligned} \left|((Z^\top Z - \mathbb{E}\left[Z^\top Z\right])_{i-} \odot M_{i-})\beta^*\right| &= \left|(Z^\top Z - \mathbb{E}\left[Z^\top Z\right])_{i-}(M_{i-}^\top \odot \beta^*)\right| \\ &\leq \left\|(Z^\top Z - \mathbb{E}\left[Z^\top Z\right])(M_{i-}^\top \odot \beta^*)\right\|_\infty. \end{aligned}$$

Note that we use M_{i-} to denote the i^{th} row of the matrix M.

Thus, by Lemma 9 and union bound over i, we have

$$\begin{aligned} \left\|(\hat{\Sigma} - \Sigma_x)\beta^*\right\|_\infty &\leq \max_{i=1,\ldots n}\left\|(Z^\top Z - \mathbb{E}\left[Z^\top Z\right])(M_{i-}^\top \odot \beta^*)\right\|_\infty \\ &\leq \sqrt{\frac{\log p}{n}}\max_i \left\|M_{i-}^\top \odot \beta^*\right\|_2 \\ &\leq \frac{1}{(1-\rho)^2}\left\|\beta^*\right\|_\infty\sqrt{\frac{\log p}{n}}. \end{aligned}$$

Combining pieces, we have

$$\begin{aligned}\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_2 &\leq \sqrt{k}\left\|\hat{\gamma}-\hat{\Sigma}\beta^*\right\|_\infty \\ &\leq \left(\frac{1}{(1-\rho)^2}\left\|\beta^*\right\|_2+\frac{1}{1-\rho}\sigma_e\right)\sqrt{\frac{k\log p}{n}}.\end{aligned}$$

The corollary follows by applying Theorem 3.

3 Proofs for the High-Dimensional Problem

In this subsection, we present the details for the proofs of the results in the paper. First, for convenience, we reproduce the statements of all the results that remain to be proven. Using the supporting concentration results in the last section, we prove all the results.

3.1 The Statement of the Results

The following four theorems correspond to Theorem 3, 5, 7, 4 in the main paper, respectively.

Theorem 12. *Under the Independent sub-Gaussian Design model and Additive Noise model, supp-OMP identifies the correct support of β^* with high probability, provided*

$$\begin{aligned}n &\gtrsim (1+\sigma_w^2)^2 k\log p,\\ |\beta_i^*| &\geq 16\left(\sigma_w\left\|\beta^*\right\|_2+\sigma_e\right)\sqrt{\frac{\log p}{n}},\end{aligned}$$

for all $i\in$ supp(β^*).

Theorem 13. *Under the Independent sub-Gaussian Design model and Additive Noise model, the output of estimator* (2) *satisfies:*

1. *(Knowledge of Σ_w):* $\left\|\hat{\beta}-\beta^*\right\|_2 \lesssim \left[\left(\sigma_w+\sigma_w^2\right)\left\|\beta^*\right\|_2+\sigma_e\sqrt{1+\sigma_w^2}\right]\sqrt{\frac{k\log p}{n}}$.
2. *(Knowledge of Σ_x):* $\left\|\hat{\beta}-\beta^*\right\|_2 \lesssim \left[(1+\sigma_w)\left\|\beta^*\right\|_2+\sigma_e\sqrt{1+\sigma_w^2}\right]\sqrt{\frac{k\log p}{n}}$.

Theorem 14. *Under the Independent sub-Gaussian Design model and missing data model, supp-OMP identifies the correct support of β^* provided*

$$\begin{aligned}n &\gtrsim \frac{1}{(1-\rho)^4}k\log p,\\ |\beta_i^*| &\geq \frac{16}{1-\rho}\left(\left\|\beta^*\right\|_2+\sigma_e\right)\sqrt{\frac{\log p}{n}},\end{aligned}$$

for all $i\in$ supp(β^*). Moreover, the output of estimator (2) *with knowledge of ρ satisfies*

$$\left\|\hat{\beta}-\beta^*\right\|_2 \lesssim \left(\frac{1}{(1-\rho)^2}\left\|\beta^*\right\|_2+\frac{1}{1-\rho}\sigma_e\right)\sqrt{\frac{k\log p}{n}}.$$

Theorem 15. *Under the independent Gaussian model, if $n\lesssim\left(\sigma_w^2+\frac{\sigma_z^2\sigma_e^2}{R^2}\right)k\log\left(\frac{p}{k}\right)$ or $b_{\min}\lesssim\sqrt{\left(\sigma_w^2R^2+\sigma_z^2\sigma_e^2\right)\frac{\log(p/k)}{n}}$, then $\mathcal{M}_0\geq 1$.*

3.2 Proof of Theorem 12 and 13

To prove the support recovery guarantees (Theorem 15), we use induction. The inductive assumption is that the previous steps identify a subset I of the true support $I^* = \mathrm{supp}(\beta^*)$. Let $I_r = I^* - I$ be the remaining true support that is yet to be identified. We need to prove that at the current step, supp-OMP picks an index in I_r, i.e., $\|h_{I_r}\|_\infty > |h_i|$ for all $i \in (I^*)^c$.

We use a decoupling argument similar to [8]: consider the oracle which runs supp-OMP over only the true support I^*. Then our supp-OMP identifies I^* if and only if it identifies it in the same order as the oracle. Therefore we can assume I to be independent of X_i and W_i for all $i \in (I^*)^c$. Note that I may still depend on X_{I^*}, W_{I^*}, and $\mathbf{e}$.

Define $\mathcal{P}_I \triangleq Z_I(Z_I^\top Z_I)^{-1}Z_I^\top$. We have

$$\begin{aligned}
\|h_{I_r}\|_\infty &= \left\|Z_{I_r}^\top r\right\|_\infty \\
&= \left\|Z_{I_r}^\top (I-\mathcal{P}_I)(X_{I^*}\beta_{I^*} + \mathbf{e})\right\|_\infty \\
&= \left\|Z_{I_r}^\top (I-\mathcal{P}_I)(Z_{I^*}\beta^*_{I^*} - W_{I^*}\beta^*_{I^*} + \mathbf{e})\right\|_\infty \\
&= \left\|Z_{I_r}^\top (I-\mathcal{P}_I)(Z_{I_r}\beta^*_{I_r} - W_{I^*}\beta^*_{I^*} + \mathbf{e})\right\|_\infty \\
&= \left\|X_{I_r}^\top (I-\mathcal{P}_I)X_{I_r}\beta^*_{I_r} + W_{I_r}^\top (I-\mathcal{P}_I)X_{I_r}\beta^*_{I_r} - Z_{I_r}^\top (I-\mathcal{P}_I)W_I^\top\beta^*_I + Z_{I_r}^\top(I-\mathcal{P}_I)\mathbf{e}\right\|_\infty \\
&\geq \frac{1}{\sqrt{k-i}}\left(\left\|X_{I_r}^\top (I-\mathcal{P}_I)X_{I_r}\beta^*_{I_r}\right\|_2 - \left\|W_{I_r}^\top (I-\mathcal{P}_I)X_{I_r}\beta^*_{I_r}\right\|_2 - \left\|Z_{I_r}^\top (I-\mathcal{P}_I)(W_I\beta^*_I - \mathbf{e})\right\|_2\right); \qquad (3)
\end{aligned}$$

here in the fifth equality we use the relations $Z_{I_r} = W_{I_r} + X_{I_r}$ and $W_I\beta^*_I = W_{I^*}\beta^*_{I^*} - W_{I_r}\beta^*_{I_r}$; the latter is due to $I = I^* - I_r$ by definition of I_r, and the fifth equality follows from expanding these terms. For the first term in (3), we have the following lemma.

Lemma 16. *Under the assumptions of Theorem 12, w.h.p. $\forall I_1 \subseteq I^*$, $I_1^c = I^* - I_1$,*

$$\begin{aligned}
\lambda_{\min}\left(X_{I_1^c}^\top (I-\mathcal{P}_{I_1})X_{I_1^c}\right) &\geq \frac{1}{2}, \\
\lambda_{\max}\left(W_{I_1^c}^\top (I-\mathcal{P}_{I_1})X_{I_1^c}\right) &\leq \frac{1}{8}.
\end{aligned}$$

Proof. By Lemma 11 and a union bound, we have w.h.p. $\forall I_1 \subseteq I^*$, $\lambda_{\min}\left(X_{I_1^c}^\top X_{I_1^c}\right) \geq \frac{1}{2}$. On the other hand, fixing $I_1 \subseteq I^*$, we have

$$\begin{aligned}
\left\|X_{I_1^c}^\top \mathcal{P}_{I_1} X_{I_1^c}\right\|_{op} &= \left\|X_{I_1^c}^\top Z_{I_1}\left(Z_{I_1}^\top Z_{I_1}\right)^{-1} Z_{I_1}^\top X_{I_1^c}\right\|_{op} \\
&\leq \sigma_1^2\left(X_{I_1^c}^\top Z_{I_1}\right)/\sigma_{\min}\left(Z_{I_1}^\top Z_{I_1}\right).
\end{aligned}$$

Again by Lemma 11, $\sigma_{\min}\left(Z_{I_1}^\top Z_{I_1}\right) \geq \frac{1}{2}(1+\sigma_w^2)$ with probability at least $1-\exp\left(cn\frac{1}{(1+\sigma_w^2)^2} + 12k\right)$, and $\sigma_1^2\left(X_{I_1^c}^\top Z_{I_1}\right) \leq \frac{1}{8}$ with probability at least $1-\exp\left(cn\frac{1}{(1+\sigma_w^2)} + 12k\right)$. So a union bound over all I_1 yields w.h.p. $\forall I_1 \subseteq I^*$, $\left\|X_{I_1^c}^\top \mathcal{P}_{I_1} X_{I_1^c}\right\|_{op} \leq \frac{1}{4}$. It follows that

$$\lambda_{\min}\left(X_{I_1^c}^\top (I-\mathcal{P}_{I_1})X_{I_1^c}\right) \geq \lambda_{\min}\left(X_{I_1^c}^\top X_{I_1^c}\right) - \left\|X_{I_1^c}^\top \mathcal{P}_{I_1} X_{I_1^c}\right\|_{op} \geq \frac{1}{4}.$$

Similarly, by Lemma 11 and the union bound, we have w.h.p. $\forall I_1 \subseteq I^*$, $\left\|W_{I_1^c}^\top X_{I_1^c}\right\|_{op} \leq \frac{1}{16}$ and $\left\|W_{I_1^c}^\top \mathcal{P}_{I_1} X_{I_1^c}\right\|_{op} \leq \frac{1}{16}$, hence $\lambda_{\max}\left(W_{I_1^c}^\top (I-\mathcal{P}_{I_1})X_{I_1^c}\right) \leq \left\|W_{I_1^c}^\top X_{I_1^c}\right\|_{op} + \left\|W_{I_1^c}^\top \mathcal{P}_{I_1} X_{I_1^c}\right\|_{op} \leq \frac{1}{8}$. □

Therefore, the first term in (3) is lower bounded by $\frac{1}{4}\left\|\beta^*_{I_r}\right\|_2$, and the second term is upper bounded by $\frac{1}{8}\left\|\beta^*_{I_r}\right\|_2$.

Now consider the third term in (3). By Lemma 11 and a union bound, we have w.h.p. $\sigma_1(W_{I_1}) \leq \frac{3}{2}\sigma_w$ for all I_1. Lemma 9 gives $\|\mathbf{e}\|_2 \leq \frac{3}{2}\sigma_e$. It follows that $\|(I-\mathcal{P}_{I_1})(W_{I_1}\beta_{I_1} - \mathbf{e})\|_2 \leq \sigma_1(I - P_{Z_{I_1}})\left(\sigma_1(W_{I_1})\left\|\beta^*_{I_1}\right\|_2 + \|\mathbf{e}\|_2\right) \leq \frac{3}{2}\left(\sigma_w\left\|\beta^*_{I_1}\right\|_2 + \sigma_e\right)$. Set $v_{I_1} = (I-\mathcal{P}_I)(W_I\beta_I - \mathbf{e})$. Because $Z_{I_1^c}$ and v_{I_1} are independent, Corollary 10 gives $\left\|Z^\top_{I_1^c} v_{I_1}\right\|_2 \leq \sqrt{\frac{(1+\epsilon)(k-i)(1+\sigma_w^2)}{n}}\|v_{I_1}\|_2$ with probability at least $1 - 3\exp\left(ck\epsilon^2\right)$. Using a union bound over all I_1, we conclude that the third term is bounded w.h.p. by $4\sqrt{\frac{(1+\sigma_w^2)(k-i)\log p}{n}}\left(\sigma_w\|\beta^*_I\|_2 + \sigma_e\right)$.

Combining the above bounds, we have

$$\|h_{I_r}\|_\infty \geq \frac{1}{\sqrt{k-i}}\left[\frac{1}{4}\left\|\beta^*_{I_r}\right\|_2 - \frac{1}{8}\left\|\beta^*_{I_r}\right\|_2 - 4\sqrt{\frac{(1+\sigma_w^2)(k-i)\log p}{n}}\left(\sigma_w\|\beta^*_I\|_2 + \sigma_e\right)\right],$$

which is greater than $\frac{1}{8\sqrt{k-i}}\left\|\beta^*_{I_r}\right\|_2$ if all the non-zero entries of β^* are greater than $16\left(\sigma_w\|\beta^*\|_2 + \sigma_e\right)\sqrt{\frac{(1+\sigma_w^2)\log p}{n}}$.

On the other hand, by similar argument as above we have $\|(I-\mathcal{P}_I)(Z_{I_r}\beta_{I_r} - W_{I^*}\beta_{I^*} + \mathbf{e})\|_2 \leq \frac{3}{2}\left(\left\|\beta^*_{I_r}\right\|_2 + \sigma_w\|\beta^*_I\|_2 + \sigma_e\right)$. Note that for each $i \in I^{*c}$, Z_i is independent of Z_I, X_{I^*} and e. Applying Corollary 10 gives w.h.p.

$$\begin{aligned}
|h_i| &= \left|Z_i^\top(I-\mathcal{P}_I)(X_{I^*}\beta_{I^*} + \mathbf{e})\right| \\
&= \left|Z_i^\top(I-\mathcal{P}_I)(Z_{I_r}\beta_{I_r} - W_{I^*}\beta_{I^*} + \mathbf{e})\right| \\
&\leq 4\sqrt{\frac{(1+\sigma_w^2)\log p}{n}}\left(\left\|\beta^*_{I_r}\right\|_2 + \sigma_w\|\beta^*_I\|_2 + \sigma_e\right),
\end{aligned}$$

which is smaller than $\frac{1}{8\sqrt{k-i}}\left\|\beta^*_{I_r}\right\|_2$ provided $n \gtrsim (1+\sigma_w^2)^2 k\log p$, and the nonzeros of β^* are greater than

$$4\left(\sigma_w\|\beta^*\|_2 + \sigma_e\right)\sqrt{\frac{(1+\sigma_w^2)\log p}{n}}$$

. Using a union bound shows this holds for all $i \in I^{*c}$.

We conclude that $\|h_{I_r}\|_\infty > |h_i|$ for all $i \in I^{*c}$ w.h.p. This completes the proof of Theorem 12. Once the correct support of β^* is identified, the problem of estimating the non-zero coefficients of β^* reduces to a low-dimensional problem. Therefore, Theorem 13 follows immediately from Theorem 12, and Corollary 4 and 5.

3.3 Proof of Theorem 14

Note that Z is sub-Gaussian with parameter $\sqrt{\frac{1}{n}}$. Similarly to the proof of Theorem 12, we use induction, the decoupling argument, and the same notation. Therefore, to prove the first part of the theorem, it suffices to show $\|h_{I_r}\|_\infty \geq |h_i|$ for all $i \in (I^*)^c$.

We have

$$\begin{aligned}
\|h_{I_r}\|_\infty &= \left\|Z^\top_{I_r}(I-\mathcal{P}_I)(X_{I^*}\beta_{I^*} + \mathbf{e})\right\|_\infty \\
&\geq \frac{1}{\sqrt{k-i}}\left\|Z^\top_{I_r}(I-\mathcal{P}_I)\left(X_{I_r}\beta^*_{I_r} + (X_I - Z_I)\beta^*_I + \mathbf{e}\right)\right\|_2 \\
&\geq \frac{1}{\sqrt{k-i}}\left(\left\|Z^\top_{I_r}(I-\mathcal{P}_I)X_{I_r}\beta^*_{I_r}\right\|_2 + \left\|Z^\top_{I_r}(I-\mathcal{P}_I)(X_I - Z_I)\beta^*_I\right\|_2 - \left\|Z^\top_{I_r}(I-\mathcal{P}_I)\mathbf{e}\right\|_2\right)
\end{aligned}$$

Consider the first term. We have $\lambda_{\min}(Z^\top_{I_r}X_{I_r}) \geq \frac{1}{2}(1-\rho)$ by Lemma 11. We also have $\lambda_{\min}(Z_I^\top Z_I) \geq \frac{1}{2}(1-\rho)$, $\sigma_1(Z^\top_{I_r}Z_I) \leq \frac{1}{8}(1-\rho)^2$, $\sigma_1(Z_I^\top X_{I_r}) \leq \frac{1}{8}(1-\rho)^2$ by the same lemma. It follows that $\lambda_1(Z^\top_{I_r}\mathcal{P}_I X_{I_r}) = \lambda_1(Z^\top_{I_r}Z_I(Z_I^\top Z_I)^{-1}Z_I^\top X_{I_r}) \leq$

$\sigma_1^2(Z_I^\top Z_{I_r})/\lambda_{\min}(Z_I^\top Z_I) \leq \frac{1}{4}(1-\rho)^3$. We conclude that $\lambda_{\min}(Z_{I_r}^\top(I-\mathcal{P}_I)Z_{I_r}) \geq \lambda_{\min}(Z_I^\top Z_I) - \lambda_1(Z_{I_r}^\top \mathcal{P}_I Z_{I_r}) \geq \frac{1}{4}(1-\rho)$. So the first term is at least $\frac{1-\rho}{4\sqrt{k-i}} \left\|\beta^*_{I_r}\right\|_2$.

For the second term, we apply Lemma 11 to obtain that w.h.p., $\sigma_1(X_I - Z_I) \leq 2$. It follows that

$$\|(I-\mathcal{P}_I)(X_I - Z_I)\beta_I\|_2 \quad \leq \quad 2\|\beta_I\|_2 .$$

By Corollary 10 and a union bound, we obtain

$$\left\|Z_{I_r}^\top (I-\mathcal{P}_I)(X_I - Z_I)\beta_I^*\right\|_2 \leq 2\|\beta_I\|_2 \sqrt{\frac{(k-i)\log p}{n}},$$

which is smaller than $\frac{1-\rho}{8}(1-\rho)\left\|\beta^*_{I_r}\right\|_2$ if the non-zeros are bigger than $\frac{16}{1-\rho}\|\beta_I\|_2\sqrt{\frac{\log p}{n}}$.

Consider the third term. In the proof of Theorem 12 we have shown that $\|\mathbf{e}\|_2 \leq \sigma_e$, so $\|(I-\mathcal{P}_I)\mathbf{e}\|_2 \leq \sigma_e$. w.h.p. By Corollary 10 and a union bound, it follows that $\left\|Z_{I_r}^\top(I-\mathcal{P}_I)\mathbf{e}\right\|_2 \leq \sqrt{\frac{(k-i)\log p}{n}}\sigma_e$, which is smaller than $\frac{1-\rho}{16}\left\|\beta^*_{I_r}\right\|_2$ if non-zeros are bigger than $\frac{16}{1-\rho}\sigma_e\sqrt{\frac{\log p}{n}}$.

Combining the above bounds, we conclude that $\|h_{I_r}\|_\infty \geq \frac{1-\rho}{8\sqrt{k-i}}\|\beta_{I_r}\|_2$ if all the non-zero entries of β^* is greater than $\frac{16}{1-\rho}(\|\beta^*\|_2 + \sigma_e)\sqrt{\frac{\log p}{n}}$.

We now consider $|h_i|$ for $i \in (I^*)^c$. We have

$$\begin{aligned}\|(I-\mathcal{P}_I)(X_{I^*}\beta_{I^*} + \mathbf{e})\|_2 &\leq \|X_{I^*}\beta_{I^*} + \mathbf{e}\|_2 \\ &\leq \frac{3}{2}\|\beta^*\|_2 + \sigma_e.\end{aligned}$$

So by independence of Z_i and X_{I^*} and Corollary 10, we obtain

$$\begin{aligned}|h_i| &= \left|Z_i^\top(I-\mathcal{P}_I)(X_{I^*}\beta_{I^*} + \mathbf{e})\right| \\ &\leq \sqrt{\frac{\log p}{n}}(\frac{3}{2}\|\beta^*\|_2 + \sigma_e),\end{aligned}$$

which is small than $\frac{1-\rho}{8\sqrt{k-i}}\left\|\beta^*_{I_r}\right\|_2$ if all the non-zeros of β^* are bigger than $\frac{16}{1-\rho}(\|\beta^*\|_2 + \sigma_e)\sqrt{\frac{\log p}{n}}$. This completes the proof for the first part of the theorem. The second part of the theorem follows from the first part and Corollary 7.

3.4 Proof of Theorem 15

We use a standard information-theoretical argument. Suppose $P = \{\beta_1, \ldots, \beta_M\}$ be a (δ, p) packing set of the target set T, which means $P \subseteq T$ and for all $\beta_j, \beta_l \in P$, $j \neq l$, we have $\|\beta_j - \beta_l\|_p \geq \delta$. A standard argument [10] converts the problem of bounding the minimax error to a hypothesis testing problem over P. In particular, we have we have

$$\min_{\hat{\beta}} \max_{\beta^* \in T} \mathbb{E}\left\|\hat{\beta} - \beta^*\right\|_p \geq \frac{\delta}{2}\min_{\tilde{\beta}} \mathbb{P}\left(\tilde{\beta} \neq B\right) \tag{4}$$

where $\tilde{\beta}$ is an estimator that takes values in P, and B is uniformly distributed over P. The probability on the R.H.S. can be bounded by Fano's inequality:

$$\min_{\tilde{\beta}} \mathbb{P}\left(\tilde{\beta} \neq B\right) \geq 1 - \frac{I(y, Z; B) + \log 2}{\log M} = 1 - \frac{I(y; B|Z) + \log 2}{\log M};$$

here the equality holds because Z and B are independent and thus if follows from the chain rule that $I(y, Z; B) = I(Z; B) + I(y; B|Z) = I(y; B|Z)$.

We now upper-bound the mutual information. For each $j = 1, \ldots, M$, let $\mathbb{P}_j \triangleq \mathbb{P}(y|B = \beta_j, Z)$ be the distribution of y given $B = \beta_j$ when Z is observed. Following [9, 7, 5], we bound $I(y; B|Z)$ using the KL-divergence:

$$I(y; B|Z) = \frac{1}{M} \sum_{ij=1}^{M} \mathbb{E}_Z \left[D\left(\mathbb{P}_j \| \frac{1}{M} \sum_{l=1}^{M} \mathbb{P}_l \right) \right] \leq \frac{1}{M^2} \sum_{i,l=1} \mathbb{E}_Z \left[D\left(\mathbb{P}_j \| \mathbb{P}_l \right) \right]$$

where we use the convexity of KL-divergence in the inequality. Under the independent Gaussian Design model, we have $\Sigma_x = \sigma_x^2 I$ and $\Sigma_w = \sigma_w^2 I$, where $\sigma_x^2 = 1$ denotes the variance of the entries of X. In this case, the KL-divergence $D\left(\mathbb{P}_j \| \mathbb{P}_l \right)$ can be computed explicitly. This is done in [5], which gives

$$D\left(\mathbb{P}_j \| \mathbb{P}_l \right) = \frac{1}{2\sigma^2} \cdot \frac{\sigma_x^4}{\sigma_z^4} \left\| Z(\beta_j - \beta_l) \right\|_2^2,$$

where $\sigma^2 \triangleq \frac{\sigma_x^2 \sigma_w^2}{\sigma_z^2} R^2 + \sigma_e^2$. Taking the expectation over Z, we get

$$\mathbb{E}_Z \left[D\left(\mathbb{P}_j \| \mathbb{P}_l \right) \right] = \frac{n \sigma_x^4}{2\left(\sigma_x^2 \sigma_w^2 R^2 + \sigma_z^2 \sigma_e^2 \right)} \left\| \beta_j - \beta_l \right\|_2^2.$$

Combining pieces, we conclude that

$$\mathcal{M}_0 = \min_{\hat{\beta}} \max_{\beta^* \in T} \mathbb{E} \left\| \hat{\beta} - \beta^* \right\|_0 \geq \frac{\delta}{4}$$

provided

$$1 - \frac{\frac{n\sigma_x^4}{2(\sigma_x^2 \sigma_w^2 R^2 + \sigma_z^2 \sigma_e^2)} \cdot \frac{1}{M^2} \sum_{i,l=1}^{M} \left\| \beta_j - \beta_l \right\|_2^2 + \log 2}{\log M} \geq \frac{1}{2}. \tag{5}$$

It remains to that show Eq.(5) holds by choosing the appropriate P, M and δ.

1) Let $P \subset \mathbb{R}^p$ be the set of vectors which have exactly k entries equal to $\frac{R}{\sqrt{k}}$ and other entries zeros. Clearly P is a $(1, 0)$ packing set of $T_1(R, k)$ with $\log M = \log |P| = \log \binom{p}{k} \geq k \log \frac{p}{k}$. Moreover, for every j, l, $\|\beta_j - \beta_l\|_2 \leq \|\beta_j\|_2 + \|\beta_l\|_2 \leq 2R$. One verifies that Eq. (5) holds when $n \leq c_1 \left(\sigma_w^2 + \frac{\sigma_z^2 \sigma_e^2}{R^2} \right) k \log \left(\frac{p}{k} \right)$ for some absolute constant c_1.

2) Set $M = p - (k - 1)$. Define β^j be such that $\beta_i^j = \frac{R}{\sqrt{k-1}}, i = 1, \ldots, k-1, \beta_j = b_{\min}, \beta_i = 0, i = k+1, \ldots, p, i \neq j$. Let $P = \left\{ \beta^j, j = 1, \ldots, M \right\}$. Clearly P is a $(1, 0)$ packing set of $T_2(R, b_{\min})$. Moreover, for every j, l, $\|\beta_j - \beta_l\|_2 \leq \sqrt{2} b_{\min}$. One verifies that Eq. (5) holds when $b_{\min} \leq c_2 \sqrt{(\sigma_w^2 R^2 + \sigma_z^2 \sigma_e^2) \frac{\log(p/k)}{n}}$ for some absolute constant c_2.

4 Additional Numerical Simulations

In this supplementary material section we provide additional numerical simulations for which space did not permit in the submission. These corroborate the theoretical results presented in the submission, as well as shed further light on the performance of supp-OMP for noisy and missing data. Our results illustrate, in particular, several key points. First, in both the low-dimensional and high-dimensional settings, empirical results demonstrate that the scaling promised in the corollaries to Theorem 3 and Theorem 12 is correct. We demonstrate this by rescaling the error of our experiments, normalizing by the predicted contribution to the error of n, k and p, in order to highlight the dependence on σ_w. Our experiments show a clear alignment of the actual results along the predicted results. The results of this section also show the different regimes of efficacy

of our different estimators for the noisy-covariate setting. Finally, we also compare to the projected gradient method in [4], and demonstrate that in addition to faster running time, we seem to obtain better empirical results at all values of the sparsity parameter, and noise intensity/erasure probability.

We present the low-dimensional results first, and then the high-dimensional results.

4.1 The Low-Dimensional Case

We report some simulation results on our low-dimensional results from Section 2. These results are also relevant to the high-dimension setting, as our OMP algorithm reduces a high-dimensional problem to a low-dimensional one once it identifies the correct support. Note that each of our bounds in Corollary 4 to Corollary 7 scales with $\frac{\log p}{n}$, which is to be expected. Therefore, we focus on verifying the scaling with the other parameters such as k, σ_w, ρ and $\|\beta^*\|$.

We first look at the case with additive noise. We fix $n = 3200$, $\sigma_e = 0$ and $\Sigma_x = I$, and sample all matrices from a Gaussian distribution. k and σ_w take values in $2, 3, \ldots, 7$ and $[0, 2]$, respectively. For each k, we generate the true β^* as a random ± 1 vector; note that $\|\beta^*\| = \sqrt{k}$, which also scales with k. Figure 1 (a) shows the ℓ^2 recovery error under different k and σ_w using the estimator built from knowledge of Σ_w, where one can see the quadratic dependence on σ_w. Corollary 4 predicts that, with fixed n, the error scales proportional to $(\sigma_w + \sigma_w^2)\|\beta^*\|\sqrt{k \log p} = (\sigma_w + \sigma_w^2)k\sqrt{\log p}$; in particular, if we plot the error versus the control parameter $(\sigma_w + \sigma_w^2)k$, all curves should roughly become straight lines through the origin and align with each other. Indeed, this is precisely what we see; the results, representing results averaged over 100 trials, are plotted in Figure 1 (b).

Similarly, we performed simulations for the estimators built from knowledge of Σ_x and from Instrumental Variables. In the latter case, the Instrumental Variable is generated by $U = X\Gamma + E$, where $\Gamma \in \mathbb{R}^{k \times m}$ with $m = 2k$ and the entries of Γ and E being i.i.d. standard Gaussian variables; in this case we have $\sigma_1(\Sigma_{UX}) \approx \sigma_k(\Sigma_{UX}) = \Theta(\sqrt{m})$ and $\sigma_u = \sqrt{k}$. Corollaries 5 and 6 predict that the ℓ^2 errors are proportional to the control parameters $(1 + \sigma_w)k$ and $\sigma_w k$. respectively. These predictions again match well our simulation results shown in Figure 2 (a) and (b).

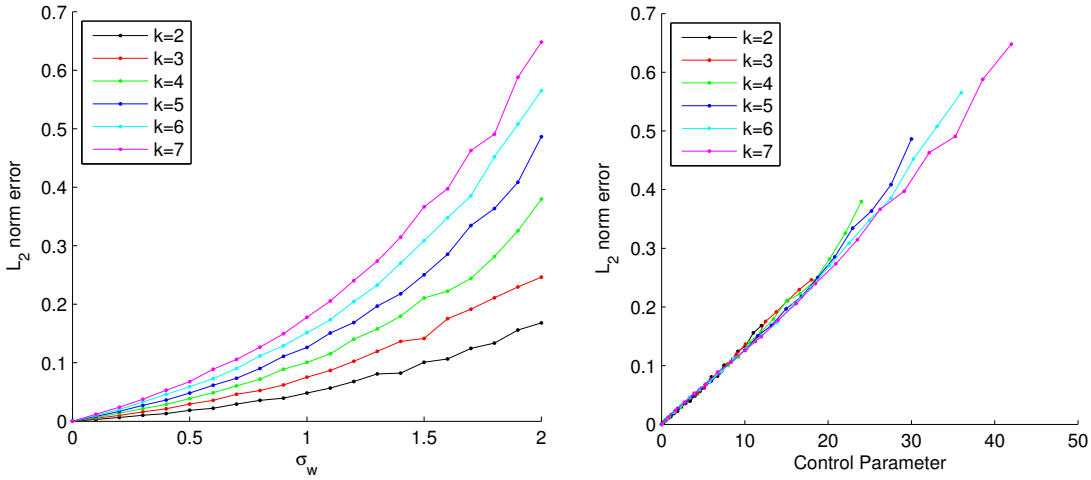
In addition, we compare the performance of the estimators built from Σ_w and Σ_x. Figure 3 shows their recovery error under different σ_w with $k = 7$. The results match the theory, and in particular show that the scaling depends as predicted on σ_w: The Σ_w-estimator performs better for small σ_w, and in particular, delivers exact recovery when $\sigma_w = 0$; the Σ_x-estimator is more favorable for large σ_w due to its linear dependence on σ_w (versus quadratic), but the error does not go to zero when $\sigma_w \to 0$. The crossover occurs roughly at $\sigma_w = 1$.

Finally, we turn to the case with missing data. We perform simulations with parameters $n = 2000$, $k \in \{2, \ldots, 7\}$, $\rho \in [0, 0.8]$, and β^* generated in the same way as above (so that $\|\beta^*\| = k$). With n fixed, Corollary 7 guarantees that the recovery error is bounded by $O(\frac{k}{(1-\rho)^2})$. The simulation results in Figure 4 seem to *outperform* this bound, as the error goes to zero when $\rho \to 0$. If we plot the error versus the control parameter $k\frac{\sqrt{\rho}}{1-\rho}$, then the curves become roughly straight lines and align. It would be interesting in the future to tighten our bound to match this scaling.

4.2 The High-Dimensional Case

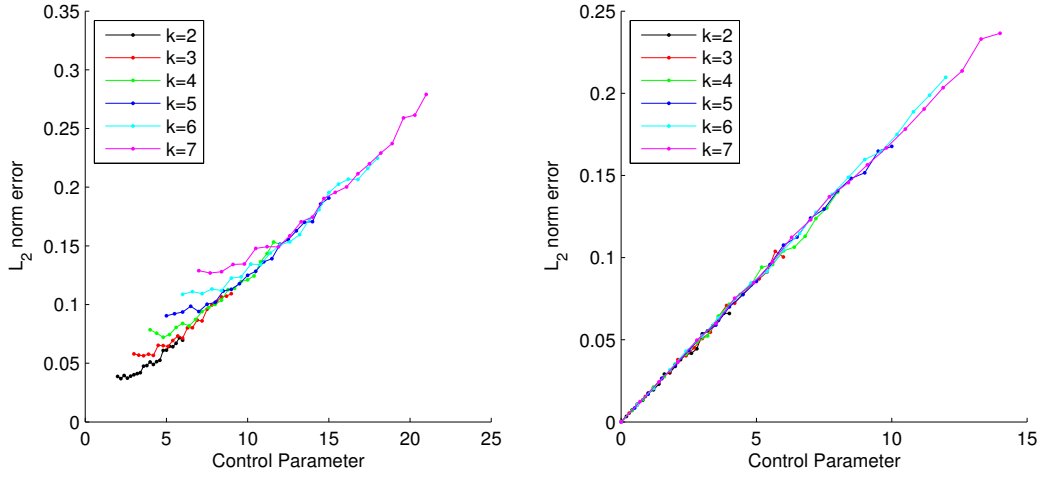
In this subsection, we study the performance of our supp-OMP algorithm for the high-dimensional setting, and compare with the projected gradient method in [4]. We first consider the additive noise case and use the following settings: $p = 450, n = 400, \sigma_e = 0, \Sigma_x = I$, $k \in \{2, \ldots, 7\}\}$, and $\sigma_w \in [0, 1]$. We compare supp-OMP using the Σ_w-estimator and the projected gradient method using the corresponding $\hat{\Sigma}$ and $\hat{\gamma}$. Figure 5 (a) plots the ℓ_2 errors. One observes that OMP outperforms the projected gradient method in all cases.

We also want to point out that supp-OMP enjoys more favorable running time in our experiments, although we do not perform a formal comparison since this depends on the particular implementation of both methods. As is clear from the description of the algorithm, supp-OMP has exactly the same running time as standard OMP.



(a) (b)

Figure 1: ℓ^2 recovery error of the Σ_w-estimator for additive noise versus (a) the noise magnitude σ_w and (b) the control parameter $(\sigma_w + \sigma_w^2)k$. As predicted by Corollary 4, all curves in (b) are roughly straight lines and align. Each point is an average over 200 trials.



(a) (b)

Figure 2: (a) ℓ^2 recovery error of the Σ_x-estimator for additive noise versus the control parameter $(1 + \sigma_w)k$. (b) ℓ^2 recovery error of the IV-estimator versus the control parameter $\sigma_w k$. As predicted by Corollary 5 and 6, all curves are roughly straight lines and align. Each point is an average over 100 trials.

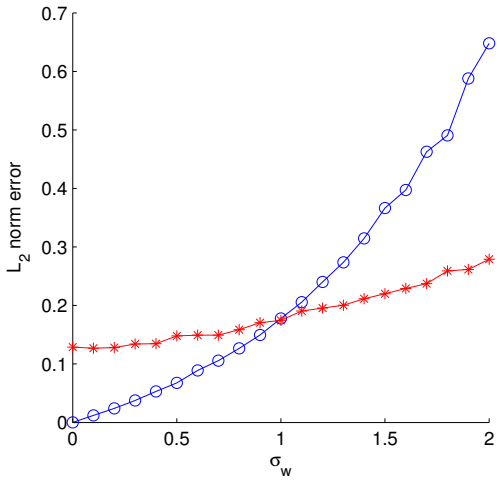
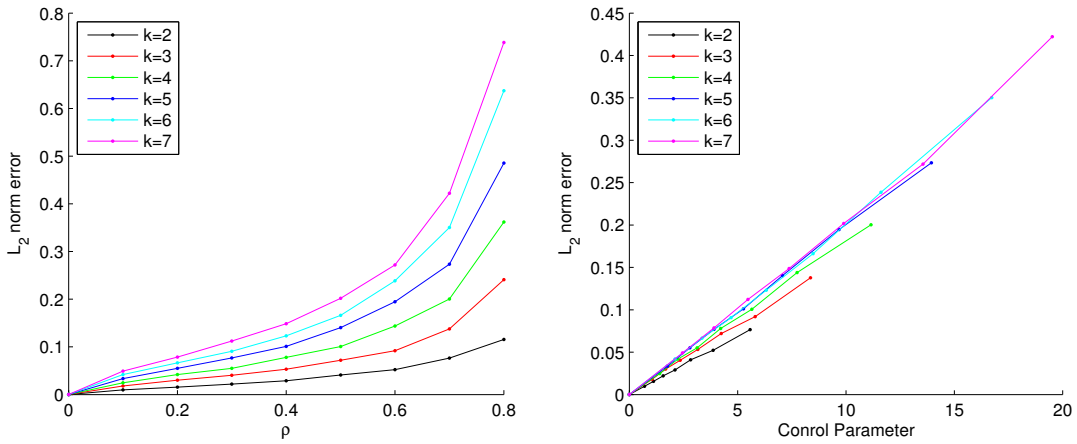
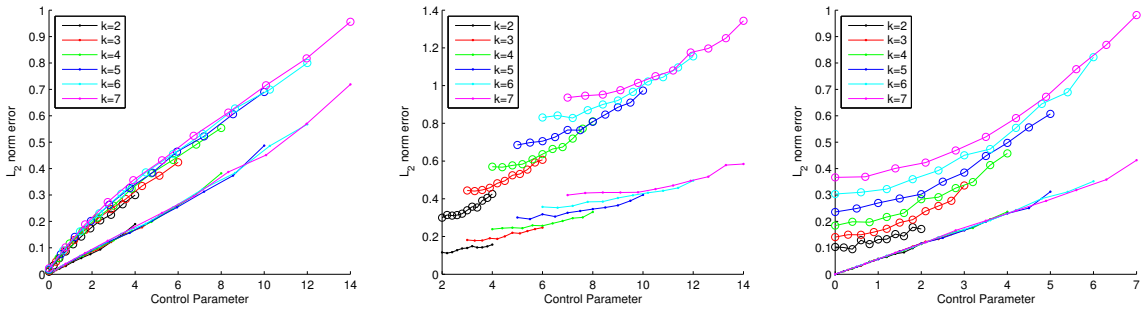


Figure 3: Comparison between recovery errors of the Σ_w- and Σ_x-estimators for additive noise. Each point is an average over 100 trials.



(a) (b)

Figure 4: ℓ^2 recovery error for missing noise versus (a) the erasure probability ρ and (b) the control parameter $\rho\sqrt{k}$. Each point is an average over 200 trials.



(a) (b) (c)

Figure 5: Comparison of the ℓ^2 recovery error of supp-OMP and the projected gradient method under knowledge of (a) Σ_w, (b) Σ_x, and (c) an Instrumental Variable. The error is plotted against the control parameter (a) $(\sigma_w + \sigma_w^2)k$, (b) $(1 + \sigma_w)k$, and (c) $\sigma_w k$. Circles correspond to the projected gradient method and dots to supp-OMP. As claimed, supp-OMP performs better in all cases considered. Each point is an average over 200 trials.

We also consider supp-OMP with the Σ_x- and IV-based estimators. Although not discussed in [4], it is natural to consider the corresponding variants of the projected gradient method which use the $\hat{\Sigma}$ and $\hat{\gamma}$ from knowledge of Σ_x or IVs (c.f. (13) in [4]). We plot the recovery errors for our two estimators in Figure 5 (b) and (c), and again observe better performance of supp-OMP than the projected gradient method.

We next study the case with missing data with the following setting: $p = 750, n = 500, \sigma_e = 0, \Sigma_x = I, k \in \{2, \dots, 7\}\}$, and $\rho \in [0, 0.5]$. The results are displayed in Figure 6, in which supp-OMP shows better performance.

Finally, although we only consider X with independent columns in this paper, and assume the sparsity level k is known, we believe that both these restrictions can be removed. For now, we corroborate this claim via simulation. Figure 6 (a) shows the results under the following choice of covariance matrix of X:

$$(\Sigma_x)_{ij} = \begin{cases} 1 & i = j \\ 0.2 & i \neq j. \end{cases}$$

Again, supp-OMP dominates the projected gradient method in terms of empirical performance. Moreover, the performance degradation due to correlation appears to be less pronounced in supp-OMP (compare Figure 6 (a) and (b)).

A Proof of Supporting Concentration Results

In this section, we provide the proofs to the concentration results for sub-Gaussian random variables that we make extensive use of in Section 2 and 3. We repeat the statements of the results below for convenience.

Lemma 9. *Suppose $X \in \mathbb{R}^{n \times k}$, $Y \in \mathbb{R}^{n \times m}$ are zero-mean sub-Gaussian matrices with parameters $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$, $(\frac{1}{n}\Sigma_y, \frac{1}{n}\sigma_y^2)$. Then for any fixed vector $\mathbf{v}_1, \mathbf{v}_2$, we have*

$$\mathbb{P}\left(\left|\mathbf{v}_1^\top \left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \geq t \left\|\mathbf{v}_1\right\| \left\|\mathbf{v}_2\right\|\right) \leq 3\exp\left(-cn\min\left\{\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\right\}\right).$$

In particular, if $n \gtrsim \log p \geq \log m, \log k$, we have w.h.p.

$$\left|\mathbf{v}_1^\top \left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \leq \sigma_x\sigma_y \left\|\mathbf{v}_1\right\| \left\|\mathbf{v}_2\right\| \sqrt{\frac{\log p}{n}}.$$

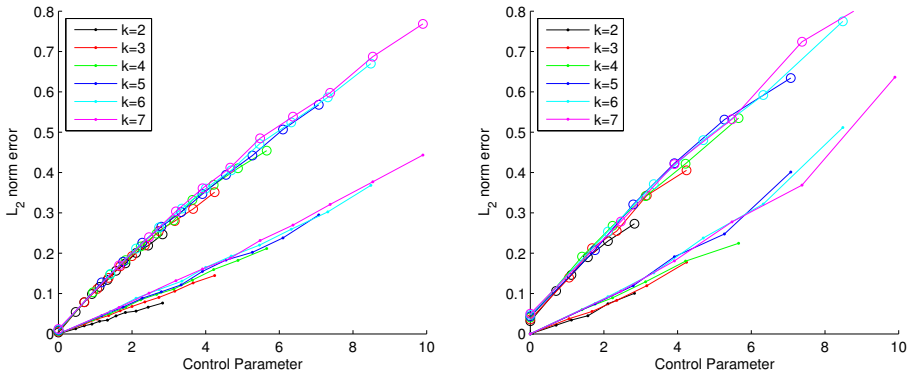


Figure 6: Comparison of the ℓ^2 recovery error of supp-OMP and the projected gradient method for missing data. The error is plotted against the control parameter $k\frac{\sqrt{\rho}}{(1-\rho)}$. (a) Independent columns of X, and (b) Correlated columns. Dots correspond to supp-OMP and circles to the projected gradient method. Our results show that supp-OMP performs better in all cases considered. Each point is an average over 50 trials.

Setting $\mathbf{v}_1$ to be the i^{th} standard basis vector, and using a union bound over $i = 1, \ldots, m$, we have w.h.p.

$$\left\|\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}\right\|_\infty \leq \sigma_x \sigma_y \|\mathbf{v}\| \sqrt{\frac{\log p}{n}}.$$

Proof. Rescaling as necessary, we assume $\sigma_x = \sigma_y = 1$ and $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$. Define $\Phi(x) = \|x\|^2 - \mathbb{E}\left(\|x\|^2\right)$. Then $\left|\mathbf{v}_1^\top \left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| = \frac{1}{2}\left|\Phi(X\mathbf{v}_2 + Y\mathbf{v}_1) - \Phi(X\mathbf{v}_2) - \Phi(Y\mathbf{v}_1)\right|$. Note that $X\mathbf{v}_2 + Y\mathbf{v}_1 = [X, Y][\mathbf{v}_2^\top, \mathbf{v}_1^\top]^\top$, where $X' = [X, Y]$ is zero-mean sub-Gaussian with parameter $(\frac{1}{n}\mathbb{E}\left[X'^\top X'\right], \frac{1}{n})$. Applying (70) in [4] to each of the three terms gives

$$\left|\mathbf{v}_1^\top \left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \geq t,$$

with probability at most $\exp\left(-cn\min\left\{t^2, t\right\}\right)$. □

Corollary 10. *If $X \in \mathbb{R}^{n\times k}$ is a zero-mean sub-Gaussian matrix with parameter $(\frac{1}{n}\sigma_x^2 I, \frac{1}{n}\sigma_x^2)$, and $\mathbf{v}$ is a fixed vector in $\mathbb{R}^n$, then for any $\epsilon \geq 1$, we have*

$$\mathbb{P}\left(\left\|X^\top \mathbf{v}\right\|_2 > \sqrt{\frac{(1+\epsilon)k}{n}}\sigma_x \|\mathbf{v}\|_2\right) \leq 3\exp\left(-ck\epsilon\right)$$

Proof. By assmption, $X^\top$ is zero-mean sub-Gaussian with parameter $(\frac{1}{k}\frac{k}{n}\sigma_x^2 I, \frac{1}{k}\frac{k}{n}\sigma_x^2)$ Note that have $\left\|X^\top \mathbf{v}\right\|_2^2 \leq \left|\mathbf{v}^\top (XX^\top - \frac{k}{n}\sigma_x^2 I)\mathbf{v}\right| + \frac{k}{n}\sigma_x^2 \|\mathbf{v}\|_2^2$. Applying the last lemma with $t = \frac{k}{n}\sigma_x^2\epsilon$, $\epsilon \geq 1$ to the first term, we obtain

$$\mathbb{P}\left(k\left|\mathbf{v}^\top (XX^\top - \frac{k}{n}\sigma_x^2 I)\mathbf{v}\right| > \frac{k}{n}\sigma_x^2\epsilon\|\mathbf{v}\|^2\right) \leq 3\exp\left(-ck\min\left\{\epsilon^2, \epsilon\right\}\right) = 3\exp\left(-ck\epsilon\right).$$

The corollary follows. □

Lemma 11. *If $X \in \mathbb{R}^{n\times k}$, $Y \in \mathbb{R}^{n\times m}$ are zero mean sub-Gaussian matrices with parameter $(\frac{1}{n}\Sigma_x, \frac{1}{n}\sigma_x^2)$,$(\frac{1}{n}\Sigma_y, \frac{1}{n}\sigma_y^2)$, then*

$$\mathbb{P}\left(\sup_{\mathbf{v}_1\in\mathbb{R}^m,\mathbf{v}_2\in\mathbb{R}^k,\|\mathbf{v}_1\|=\|\mathbf{v}_2\|=1} \left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \geq t\right) \leq 2\exp\left(-cn\min(\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}) + 6(k+m)\right).$$

In particular, for each $\lambda > 0$, if $n \gtrsim \max\left\{\frac{\sigma_x^2\sigma_y^2}{\lambda^2}, 1\right\}(k+m)\log p$, then w.h.p.

$$\sup_{\mathbf{v}_1\in\mathbb{R}^m,\mathbf{v}_2\in\mathbb{R}^k} \left|\mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2\right| \qu\leq \quad \frac{1}{54}\lambda\left\|\mathbf{v}_1\right\|\left\|\mathbf{v}_2\right\|$$

Proof. Rescaling as necessary, we assume $\sigma_x = \sigma_y = 1$. Let $\mathcal{A}_1$, be a $1/3$-cover of $\mathbf{v}_1 = \{\mathbf{v}\in\mathbb{R}^m, \|v\| \leq 1\}$; it is known that $|\mathcal{A}_1| \leq 9^{2m}$, and for each $\mathbf{v}$, there is a $u(\mathbf{v}) \in \mathcal{A}_1$ such that $\|\Delta(\mathbf{v})\| \triangleq \|\mathbf{v} - u(\mathbf{v})\| \leq \frac{1}{3}$. Similarly we can find a $1/3$-cover $\mathcal{A}_2$ of $\mathbf{v}_2 = \{\mathbf{v}\in\mathbb{R}^k, \|\mathbf{v}\| \leq 1\}$ with $|\mathcal{A}_2| \leq 9^{2k}$. Defining $\Phi(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{v}_1^\top\left(Y^\top X - \mathbb{E}\left[Y^\top X\right]\right)\mathbf{v}_2$, then

$$\begin{aligned}\sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\mathbf{v}_1,\mathbf{v}_2)| \quad\leq\quad & \max_{u_1\in\mathcal{A}_1,u_2\in\mathcal{A}_2}|\Phi(u_1,u_2)| + \sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\Delta(\mathbf{v}_1),u(\mathbf{v}_2))| \\ & + \sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(u(\mathbf{v}_1),\Delta(\mathbf{v}_2))| + \sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\Delta(\mathbf{v}_1),\Delta(\mathbf{v}_2))|.\end{aligned}$$

Becaue $3\Delta(\mathbf{v}_1), u(\mathbf{v}_1) \in \mathbf{v}_1$, and $3\Delta(\mathbf{v}_2), u(\mathbf{v}_2) \in \mathbf{v}_2$, it follows that

$$\sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\mathbf{v}_1,\mathbf{v}_2)| \quad\leq\quad \max_{u_1,u_2\in\mathcal{A}}|\Phi(u_1,u_2)| + \left(\frac{1}{3}+\frac{1}{3}+\frac{1}{9}\right)\sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\mathbf{v}_1,\mathbf{v}_2)|,$$

hence $\sup_{\mathbf{v}_1\in\mathbf{v}_1,\mathbf{v}_2\in\mathbf{v}_2}|\Phi(\mathbf{v}_1,\mathbf{v}_2)| \leq \frac{9}{2}\max_{u_1,u_2\in\mathcal{A}}|\Phi(u_1,u_2)|$. Using the last lemma and a union bound, we obtain

$$\mathbb{P}\left(\frac{9}{2}\max_{u_1,u_2\in\mathcal{A}}|\Phi(u_1,u_2)| \geq t\right) \leq 9^{2k+2m}\cdot\exp\left(-cn\min\left\{t^2,t\right\}\right) \leq \exp\left(-cn\min\left\{t^2,t\right\} + 6(k+m)\right).$$

□

References

[1] R.J. Carroll. *Measurement error in nonlinear models: a modern perspective*, volume 105. CRC Press, 2006.

[2] W.A. Fuller. Measurement error models. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*, 1, 1987.

[3] E. Gautier and A.B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

[4] P.L. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Arxiv preprint arXiv:1109.3714*, 2011.

[5] P.L. Loh and M.J. Wainwright. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2601–2605. IEEE, 2012.

[6] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[7] G. Raskutti, M.J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Arxiv preprint arXiv:0910.2042*, 2009.

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

[9] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[10] B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.