High dimensional errors-in-variables models with dependent measurements

Mark Rudelson* and Shuheng Zhou†
*Department of Mathematics,
†Department of Statistics,
University of Michigan, Ann Arbor, MI 48109-1107

February 10, 2015

Abstract

We consider a parsimonious model for fitting observation data $X = X_0 + W$ with two-way dependencies; that is, we use the signal matrix $X_0$ to explain column-wise dependency in $X$, and the measurement error matrix $W$ to explain its row-wise dependency. In the matrix normal setting, we have the following representation where $X$ follows the matrix variate normal distribution with the Kronecker Sum covariance structure:

$$\text{vec} \{X\} \sim \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma = A \oplus B,$$

which is generalized to the subgaussian settings as follows. Suppose that we observe $y \in \mathbb{R}^f$ and $X \in \mathbb{R}^{f \times m}$ in the following model:

$$y = X_0\beta^* + \epsilon$$
$$X = X_0 + W$$

where $X_0$ is a $f \times m$ design matrix with independent subgaussian row vectors, $\epsilon \in \mathbb{R}^m$ is a noise vector and $W$ is a mean zero $f \times m$ random noise matrix with independent subgaussian column vectors, independent of $X_0$ and $\epsilon$. This model is significantly different from those analyzed in the literature. Under sparsity and restricting eigenvalue type of conditions, we show that one is able to recover a sparse vector $\beta^* \in \mathbb{R}^m$ from the following model given a single observation matrix $X$ and the response vector $y$. We establish consistency in estimating $\beta^*$ and obtain the rates of convergence in the $\ell_q$ norm, where $q = 1, 2$ for the Lasso-type estimator, and for $q \in [1, 2]$ for a Dantzig-type conic programming estimator.

1 Introduction

The matrix variate normal model has a long history in psychology and social sciences, and is becoming increasingly popular in biology and genomics, neuroscience, econometric theory, image and signal processing, wireless communication, and machine learning in recent years, see for example [13, 19, 15, 28, 48, 4, 51, 16]. Estimation of the graphical structures corresponding to the matrix normal distribution has been considered in recent work [1, 52, 50, 24, 46, 23, 53].

We call the random matrix $X$ which contains $f$ rows and $m$ columns a single data matrix, or one instance from the matrix variate normal distribution. We say that an $f \times m$ random matrix $X$ follows a matrix normal
distribution with a separable covariance matrix $\Sigma = A \otimes B$, which we write $X_{f \times m} \sim N_{f,m}(M, A_{m \times m} \otimes B_{f \times f})$. This is equivalent to say $\text{vec} \{X\}$ follows a multivariate normal distribution with mean $\text{vec} \{M\}$ and covariance $\Sigma = A \otimes B$. Here, $\text{vec} \{X\}$ is formed by stacking the columns of $X$ into a vector in $\mathbb{R}^{mf}$. Intuitively, $A$ describes the covariance between columns of $X$ while $B$ describes the covariance between rows of $X$. See [13, 19] for more characterization and examples.

In this paper, we introduce the related Kronecker Sum models to encode the covariance structure of a matrix variate distribution. The proposed models and methods incorporate ideas from recent advances in graphical models, high-dimensional regression model with observation errors, and matrix decomposition. Let $A_{m \times m}, B_{f \times f}$ be symmetric positive definite covariance matrices. Denote the Kronecker sum of $A = (a_{ij})$ and $B = (b_{ij})$ by

$$\Sigma = A \oplus B := A \otimes I_f + I_m \otimes B$$

$$= \begin{bmatrix} a_{11}I_f + B & a_{12}I_f & \ldots & a_{1m}I_f \\ a_{21}I_f & a_{22}I_f + B & \ldots & a_{2m}I_f \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}I_f & a_{m2}I_f & \ldots & a_{mm}I_f + B \end{bmatrix}_{(mf) \times (mf)}$$

where $I_f$ is an $f \times f$ identity matrix. This covariance model arises naturally from the context of errors-in-variables regression model which we now introduce. Suppose that we observe $y \in \mathbb{R}^f$ and $X \in \mathbb{R}^{f \times m}$ in the following model:

$$y = X_0 \beta^* + \epsilon \quad \text{(1a)}$$

$$X = X_0 + W \quad \text{(1b)}$$

where $X_0$ is a $f \times m$ design matrix with independent row vectors, $\epsilon \in \mathbb{R}^m$ is a noise vector and $W$ is a mean zero $f \times m$ random noise matrix, independent of $X_0$ and $\epsilon$, with independent column vectors $\omega^1, \ldots, \omega^m$. In particular, we are interested in the additive measurement error model of $X = X_0 + W$ such that

$$\text{vec} \{X\} \sim \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma = A \oplus B := A \otimes I_f + I_m \otimes B$$

where we use one covariance component $A \otimes I_f$ to describe the covariance of matrix $X_0 \in \mathbb{R}^{f \times m}$, which is considered as the signal matrix, and the other component $I_m \otimes B$ to describe that of the noise matrix $W \in \mathbb{R}^{f \times m}$, where $\mathbb{E} \omega^j \otimes \omega^j = B$ for all $j$, where $\omega^j$ denotes the $j^{th}$ column vector of $W$. We will show that our theory and analysis works with a model more general than that in (2). We first state the following assumption.

(A1) We assume $\text{tr}(A) = m$ is a known parameter, where $\text{tr}(A)$ denotes the trace of matrix $A$.

Our focus is on the statistical properties of two estimators for estimating $\beta^*$ in (1a) despite the presence of the additive measurement error $W$ in the observation matrix $X$. We will show the rates of convergence in the $\ell_q$ norm for $q = 1, 2$ for estimating a sparse vector $\beta^* \in \mathbb{R}^m$ in the model (1a) and (1b) using a modified form of the Lasso estimator as studied in [27] in Theorem 2.2, and a modified form of the Dantzig Selector as studied in [2] in Theorem 2.6 for $1 \leq q \leq 2$. We provide a unified analysis of the rates of convergence for both the Lasso-type estimator (4) as well as the Conic Programming estimator (5), which is a Dantzig selector-type, although under slightly different conditions. We first introduce the Lasso-type estimator, adapted from those as considered in Loh and Wainwright [27].
Suppose that \( \hat{\text{tr}}(B) \) is an estimator for \( \text{tr}(B) \). Let
\[
\hat{\Gamma} = \frac{1}{f} X^T X - \frac{\hat{\text{tr}}(B)}{f} I_m \quad \text{and} \quad \hat{\gamma} = \frac{1}{f} X^T y.
\]
(3)

For a chosen penalization parameter \( \lambda \geq 0 \), and parameters \( b_0 \) and \( d \), we consider the following regularized estimation with the \( \ell_1 \)-norm penalty,
\[
\hat{\beta} = \arg \min_{\beta : \|\beta\|_1 \leq b_0 \sqrt{d}} \frac{1}{2} \beta^T \hat{\Gamma} \beta - \langle \hat{\gamma}, \beta \rangle + \lambda \|\beta\|_1,
\]
(4)

which is a variation of the Lasso [43] or the Basis Pursuit [10] estimator.

Recently, Belloni, Rosenbaum and Tsybakov discussed the following conic programming compensated matrix uncertainly (MU) selector [2], which is a variant of the Dantzig selector [5, 32, 33]. Adapted to our setting, it is defined as follows. Let \( \lambda, \mu, \tau > 0 \),
\[
\hat{\beta} = \arg \min \left\{ \|\beta\|_1 + \lambda t : (\beta, t) \in \Upsilon \right\}
\]
where
\[
\Upsilon = \left\{ (\beta, t) : \beta \in \mathbb{R}^m, \left\| \hat{\gamma} - \hat{\Gamma} \beta \right\|_\infty \leq \mu t + \tau, \|\beta\|_2 \leq t \right\}
\]
(5)

where \( \hat{\gamma} \) and \( \hat{\Gamma} \) are as defined in Theorem 2.6 with \( \mu \sim \sqrt{\log m}, \tau \sim \sqrt{\log m} \).

In both Theorems 2.2 and 2.6, we consider the regression model in (1a) and (1b) with subgaussian random design, where \( X_0 = Z_1 A^{1/2} \) is a subgaussian random matrix with independent row vectors, and \( W = B^{1/2} Z_2 \) is a \( f \times m \) random noise matrix with independent column vectors where \( Z_1, Z_2 \) are independent subgaussian random matrices with independent entries (cf. Definition 2.1). This model is significantly different from those analyzed in the literature. For example, unlike the present work, the authors in [27] apply Theorem 3.1 which states a general result on statistical convergence properties of the estimator (4) to cases where \( W \) is composed of independent subgaussian row vectors, when the row vectors of \( X_0 \) are either independent or follow a Gaussian vector auto-regressive model. See also [32, 33, 11, 2] for the corresponding results on the compensated MU selectors, variant on the Orthogonal Matching Pursuit algorithm and the Conic Programming estimator (5).

The other key difference between our framework and the existing work is that we assume that only one observation matrix \( X \) with the single measurement error matrix \( W \) is available. Assuming (A1) allows us to estimate \( \mathbb{E} W^T W \) as required in the estimation procedure (3) directly, given the knowledge that \( W \) is composed of independent column vectors. In contrast, existing work needs to assume that the covariance matrix \( \Sigma_W := \frac{1}{f} \mathbb{E} W^T W \) of the independent row vectors of \( W \) or its functionals are either known a priori, or can be estimated from an dataset independent of \( X \), or from replicated \( X \) measuring the same \( X_0 \); see for example [32, 33, 2, 27, 7]. Such repeated measurements are not always available or are costly to obtain in practice [7].

A noticeable exception is the work of [11], which deals with the scenario when the noise covariance is not assumed to be known. We now elaborate on their result, which is a variant of the orthogonal matching pursuit (OMP) algorithm [45, 44]. Their support recovery result, that is, recovering the support set of \( \beta^* \), applies only to the case when both signal matrix and the measurement error matrix have isotropic subgaussian row vectors; that is, they assume independence among both rows and columns in \( X \) (\( X_0 \) and \( W \)); moreover, their algorithm requires the knowledge of the sparsity parameter \( d \), which is the number of non-zero entries.
in $\beta^*$, as well as a $\beta_{\min}$ condition: $\min_{j \in \text{supp } \beta^*} |\beta^*_j| = \Omega\left(\sqrt{\frac{\log m}{f}} (\|\beta^*\|_2 + 1)\right)$. They recover essentially the same $\ell_2$-error bounds as in [27] and the current work when the covariance $\Sigma_W$ is known. In other words, oblivion in $\Sigma_W$ and a general dependency condition in the data matrix are not simultaneously allowed in existing work.

In contrast, while we assume that $X_0$ is composed of independent subgaussian row vectors, we allow rows of $W$ to be dependent, which brings dependency to the row vectors of the observation matrix $X$. In some sense, we are considering a parsimonious model for fitting observation data with two-way dependencies; that is, we use the signal matrix to explain column-wise dependency in $X$, and the measurement error matrix to explain its row-wise dependency.

We now use an example to motivate (2) and its subgaussian generalization in Definition 2.1. Suppose that there are $f$ patients in a particular study, for which we use $X_0$ to model the ”systolic blood pressure” and $W$ to model the seasonal effects. In this case, $X$ models the fact that among the $f$ patients we measure, each patient has its own row vector of observed set of blood pressures across time, and each column vector in $W$ models the seasonal variation on top of the true signal at a particular day/time. Thus we consider $X$ as measurement of $X_0$ with $W$ being the measurement error. That is, we model the seasonal effects on blood pressures across a set of patients in a particular study with a vector of correlated entries, which allows the reduction to independent case which is commonly assumed in the literature. We refer to [7] for an excellent survey of the classical as well as modern developments in measurement error models. We will continue the discussion of this example in Section 7.

1.1 Assumptions and conditions

We will now define some parameters related to the restricted and sparse eigenvalue conditions that are needed to state our main results. We then state independent isotropic vectors with subgaussian marginals as in Definition 1.5.

Definition 1.1. (Restricted eigenvalue condition $RE(s_0, k_0, A)$). Let $1 \leq s_0 \leq p$, and let $k_0$ be a positive number. We say that a $p \times q$ matrix $A$ satisfies $RE(s_0, k_0, A)$ condition with parameter $K(s_0, k_0, A)$ if for any $\nu \neq 0$,

$$\frac{1}{K(s_0, k_0, A)} := \min_{J \subseteq \{1, \ldots, p\}, \|\nu_J\|_1 \leq k_0 \|\nu_J\|_1} \min_{\|\nu_J\|_2 \leq \|A\nu_J\|_2} > 0.$$  \hspace{1cm} (6)

It is clear that when $s_0$ and $k_0$ become smaller, this condition is easier to satisfy. We also consider the following variation of the baseline RE condition.

Definition 1.2. (Lower-RE condition) [27] The matrix $\Gamma$ satisfies a Lower-RE condition with curvature $\alpha > 0$ and tolerance $\tau > 0$ if

$$\theta^T \Gamma \theta \geq \alpha \|\theta\|_2^2 - \tau \|\theta\|_1^2 \quad \forall \theta \in \mathbb{R}^m.$$  

As $\alpha$ becomes smaller, or as $\tau$ becomes larger, the Lower-RE condition is easier to be satisfied. We show in Lemma 1.3 the the relationships between the two conditions in Definitions 1.1 and 1.2.

Lemma 1.3. Suppose that the Lower-RE condition holds for $\Gamma := A^T A$ with $\alpha, \tau > 0$ such that $\tau (1 + k_0)^2 s_0 \leq \alpha/2$. Then the $RE(s_0, k_0, A)$ condition holds for $A$ with

$$\frac{1}{K(s_0, k_0, A)} \geq \sqrt{\frac{\alpha}{2}} > 0.$$  

4
Assume that $\text{RE}((k_0 + 1)^2, k_0, A)$ holds. Then the Lower-RE condition holds for $\Gamma = A^T A$ with
\[
\alpha = \frac{1}{(k_0 + 1)K^2(s_0, k_0, A)} > 0
\]
where $s_0 = (k_0 + 1)^2$, and $\tau > 0$ which satisfies
\[
\lambda_{\min}(\Gamma) \geq \alpha - \tau s_0 / 4.
\] (7)
The condition above holds for any $\tau \geq \frac{4}{(k_0 + 1)K^2(s_0, k_0, A)} - \frac{4\lambda_{\min}(\Gamma)}{(k_0 + 1)^2}$.

The first part of the Lemma means that, if $k_0$ is fixed, then smaller values of $\tau$ guarantee $\text{RE}(s_0, k_0, A)$ holds with larger $s_0$, that is, a stronger RE condition. The second part of the Lemma implies that a weak RE condition implies that the Lower-RE (LRE) holds with a large $\tau$. On the other hand, if one assumes $\text{RE}((k_0 + 1)^2, k_0, A)$ holds with a large value of $k_0$ (in other words, a strong RE condition), this would imply LRE with a small $\tau$. In short, the two conditions are similar but require tweaking the parameters. Weaker RE condition implies LRE condition holds with a larger $\tau$, and Lower-RE condition with a smaller $\tau$, that is, stronger LRE implies stronger RE. Proof of Lemma 1.3 appears in Section B.

**Definition 1.4. (Upper-RE condition) [27]** The matrix $\Gamma$ satisfies an upper-RE condition with curvature $\bar{\alpha} > 0$ and tolerance $\tau > 0$ if
\[
\theta^T \Gamma \theta \geq \bar{\alpha} \| \theta \|_2^2 + \tau \| \theta \|_1^2 \quad \forall \theta \in \mathbb{R}^m.
\]

**Definition 1.5.** Let $Y$ be a random vector in $\mathbb{R}^p$

1. $Y$ is called isotropic if for every $y \in \mathbb{R}^p$, $\mathbb{E} \left( | \langle Y, y \rangle |^2 \right) = \| y \|_2^2$.

2. $Y$ is $\psi_2$ with a constant $\alpha$ if for every $y \in \mathbb{R}^p$,
\[
\| \langle Y, y \rangle \|_{\psi_2} := \inf \{ t : \mathbb{E} \left( \exp(\langle Y, y \rangle^2 / t^2) \right) \leq 2 \} \leq \alpha \| y \|_2.
\] (8)

The $\psi_2$ condition on a scalar random variable $V$ is equivalent to the subgaussian tail decay of $V$, which means $\mathbb{P}(|V| > t) \leq 2 \exp(-t^2 / \tilde{c}^2)$, for all $t > 0$.

The rest of the paper is organized as follows. In Section 2, we present two main results Theorems 2.2 and 2.6. In Sections 3 and 4, we outline the proofs for Theorems 2.2 and 2.6 respectively. In Section 5, we show a deterministic result as well as its application to the random matrix $\hat{\Gamma} - A$ for $\hat{\Gamma}$ as in (3) with regards to the upper and Lower RE conditions. We note that the bounds corresponding to the Upper RE condition as stated in Lemma 3.3, Corollary 5.1 and Theorem 5.2 are not needed for Theorem 2.2. They are useful to ensure algorithmic convergence and to bound the optimization error for the gradient descent-type of algorithms as considered in [27], when one is interested in approximately solving the non-convex optimization function (4).

In sections 6 and C we show the concentration properties of the gram matrices $XX^T$ and $X^T X$ after we correct them with the corresponding population error terms defined by $\text{tr}(A)I_J$ and $\text{tr}(B)I_m$ respectively. These results might be of independent interests. Technical proofs and additional theoretical results are included in the appendix.

**Notation and definitions.** Let $e_1, \ldots, e_p$ be the canonical basis of $\mathbb{R}^p$. For a set $J \subset \{1, \ldots, p\}$, denote $E_J = \text{span}\{e_j : j \in J\}$. For a matrix $A$, we use $\|A\|_2$ to denote its operator norm. For a set $V \subset \mathbb{R}^p$,
we let $\text{conv } V$ denote the convex hull of $V$. For a finite set $Y$, the cardinality is denoted by $|Y|$. Let $B_1^p$, $B_2^p$ and $S^{p-1}$ be the unit $\ell_1$ ball, the unit Euclidean ball and the unit sphere respectively. For a matrix $A = (a_{ij})_{1 \leq i,j \leq m}$, let $\|A\|_{\text{max}} = \max_{i,j} |a_{ij}|$ denote the entry-wise max norm; Let $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the matrix $\ell_1$ norm. The Frobenius norm is given by $\|A\|_F^2 = \sum_i \sum_j a_{ij}^2$. Let $|A|$ denote the determinant and $\text{tr}(A)$ be the trace of $A$. Let $\varphi_{\text{max}}(A)$ and $\varphi_{\text{min}}(A)$ be the largest and smallest eigenvalues, and $\kappa(A)$ be the condition number for matrix $A$. The operator or $\ell_2$ norm $\|A\|_2^2$ is given by $\varphi_{\text{max}}(AA^T)$.

For a matrix $A$, denote by $r(A)$ the effective rank $\text{tr}(A)/\|A\|_2$. Let $\|A\|_F^2/\|A\|_2^2$ denote the stable rank for matrix $A$. We write $\text{diag}(A)$ for a diagonal matrix with the same diagonal as $A$. For a symmetric matrix $A$, let $\Upsilon(A) = (\upsilon_{ij})$ where $\upsilon_{ij} = \mathbb{I}(a_{ij} \neq 0)$, where $\mathbb{I}(\cdot)$ is the indicator function. Let $I$ be the identity matrix. We let $C$ be a constant which may change from line to line. For two numbers $a, b, a \wedge b := \min(a, b)$, and $a \vee b := \max(a, b)$. We write $a \asymp b$ if $ca \leq b \leq Cb$ for some positive absolute constants $c, C$ which are independent of $n, f, m$ or sparsity parameters. These absolute constants $C, C_1, c, c_1, \ldots$ may change line by line.

2 Main results

In this section, we will introduce a more general model, namely, the subgaussian analog of (2) to model the observational data with measurement error in Definition 2.1. We then state our main results in Theorems 2.2 and 2.6 where we consider the regression model in (1a) and (1b) with random matrices $X_0, W \in \mathbb{R}^{f \times m}$ as defined in Definition 2.1.

**Definition 2.1.** Let $Z$ be an $f \times m$ random matrix with independent entries $Z_{ij}$ satisfying $\mathbb{E}Z_{ij} = 0$, $1 = \mathbb{E}Z_{ij}^2 \leq \|Z_{ij}\|_{\psi_2} \leq K$. Let $Z_1, Z_2$ be independent copies of $Z$. Let $X = X_0 + W$ such that

1. $X_0 = Z_1A_1^{1/2}$ is the design matrix with independent subgaussian row vectors, and
2. $W = B^{1/2}Z_2$ is a random noise matrix with independent subgaussian column vectors.

Throughout this paper, we use $\psi_2$ vector, a vector with subgaussian marginals and subgaussian vector interchangeably. Assumption (A1) allows the covariance model in (2) and its subgaussian variant in Definition 2.1 to be identifiable. In particular, by knowing $\text{tr}(A)$, we can construct an estimator for $\text{tr}(B)$ as follows:

$$\hat{\text{tr}}(B) = \frac{1}{m}(\|X\|_F^2 - f\text{tr}(A)).$$

We next state Theorems 2.2 and 2.6 and their consequences. For the Lasso-type estimator, we are interested in the case where the smallest eigenvalue of the column-wise covariance matrix $A$ does not approach 0 too quickly and the effective rank of the row-wise covariance matrix $B$ is bounded from below (cf. (12)). For the Conic Programming estimator, we impose a restricted eigenvalue condition as formulated in [3, 35] on $A$ and assume that the sparsity of $\beta^*$ is bounded by $o(\sqrt{f/\log m})$.

Before stating our main result for the Lasso-type estimator in Theorem 2.2, we need to introduce some more notation and assumptions. Let $a_{\text{max}} = \max_i a_{ii}$ and $b_{\text{max}} = \max_i b_{ii}$ be the maximum diagonal entries of $A$ and $B$ respectively. In general, under (A1), one can think of $\lambda_{\text{min}}(A) \leq 1$ and for $s \geq 1$,

$$1 \leq a_{\text{max}} \leq p_{\text{max}}(s, A) \leq \lambda_{\text{max}}(A),$$

where $\lambda_{\text{max}}(A)$ denotes the maximum eigenvalue of $A$. 

6
(A2) The minimal eigenvalue $\lambda_{\text{min}}(A)$ of the covariance matrix $A$ is bounded: $1 \geq \lambda_{\text{min}}(A) > 0$.

(A3) Moreover, we assume that the condition number $\kappa(A)$ is upper bounded by $O\left(\sqrt{\frac{f}{\log m}}\right)$.

Throughout the rest of the paper, $s_0 \geq 1$ is understood to be the largest integer chosen such that the following inequality still holds:

$$\sqrt{s_0 \varpi(s_0)} \leq \frac{\lambda_{\text{min}}(A)}{32C} \sqrt{\frac{f}{\log m}} \quad \text{where} \quad \varpi(s_0) := \frac{\rho_{\text{max}}(s_0, A)}{\lambda_{\text{min}}(A)} + \tau_B$$  \hspace{1cm} (10)

where we denote by $\tau_B = \text{tr}(B)/f$ and $C$ is to be defined. Denote by

$$M_A = \frac{64C \varpi(s_0)}{\lambda_{\text{min}}(A)} \geq 64C.$$  \hspace{1cm} (11)

Throughout this paper, for the Lasso-type estimator, we will use the expression

$$\tau := \frac{\alpha}{s_0}, \quad \text{where} \quad \alpha = \frac{\lambda_{\text{min}}(A)}{2};$$

(A2) thus ensures that the Lower-RE condition as in Definition 1.2 is not vacuous. (A3) ensures that (10) holds for some $s_0 \geq 1$.

**Theorem 2.2.** (Estimation for the Lasso-type estimator) Set $1 \leq f \leq m$, and let $d < f/2$. Suppose $m$ is sufficiently large. Suppose (A1), (A2) and (A3) hold. Consider the regression model in (1a) and (1b) with independent random matrices $X_0, W$ as in Definition 2.1, and an error vector $\epsilon_j \in \mathbb{R}^f$ independent of $X_0, W$, with independent entries $\epsilon_j$ satisfying $\mathbb{E}\epsilon_j = 0$ and $\|\epsilon_j\|_{\psi_2} \leq M_e$. Suppose $\hat{\epsilon}(B)$ is an estimator for $\text{tr}(B)$ as constructed in (9). Let $C_0, c' > 0$ be some absolute constants.

Suppose that $\|B\|_2^2 / \|B\|_2^2 \geq \log m$. Suppose that $c'K^4 \leq 1$ and

$$r(B) := \frac{\text{tr}(B)}{\|B\|_2} \geq 16c'K^4 \sqrt{\frac{f}{\log m}} \log \frac{\sqrt{m \log m}}{f}$$  \hspace{1cm} (12)

where $\mathcal{V}$ is a constant which depends on $\lambda_{\text{min}}(A), \rho_{\text{max}}(s_0, A)$ and $\text{tr}(B)/f$.

Let $b_0, \phi$ be numbers which satisfy

$$\frac{M_e^2}{K^2b_0} \leq \phi \leq 1.$$  \hspace{1cm} (13)

Assume that the sparsity of $\beta^*$ satisfies for some $0 < \phi \leq 1$

$$d := |\text{supp}(\beta^*)| \leq \frac{c'\phi K^4}{128M_A^2 \log m} \cdot f.$$  \hspace{1cm} (14)

Let $\hat{\beta}$ be an optimal solution to the Lasso-type estimator as in (4) with

$$\lambda \geq 4\psi \sqrt{\frac{\log m}{f}} \quad \text{where} \quad \psi := C_0D_2K (K \|\beta^*\|_2 + M_e)$$  \hspace{1cm} (15)

with $D_2 := 2(\|A\|_2 + \|B\|_2)$. Then for any $d$-sparse vectors $\beta^* \in \mathbb{R}^m$, such that $\phi b_0^2 \leq \|\beta^*\|_2 \leq b_0^2$, we have with probability at least $1 - 8/m^3$,

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{20}{\alpha} \sqrt{\lambda d} \quad \text{and} \quad \|\hat{\beta} - \beta^*\|_1 \leq \frac{80}{\alpha} \lambda d.$$
We give an outline of the proof of Theorem 2.2 in Section 3. A large deviation bound for the estimator as in (9) is stated in Lemma E.2. The actual proof of Theorem 2.2 appears in Section F.1.

**Remark 2.3.** Denote the Signal-to-noise ratio by

\[ \frac{S}{N} := K^2 \| \beta^* \|_2^2 / M_\varepsilon^2 \] where \( S := K^2 \| \beta^* \|_2^2 \) and \( N := M_\varepsilon^2 \).

The two conditions on \( b_0, \phi \) imply that \( N \leq \phi S \). Notice that this could be restrictive if \( \phi \) is small; hence we prove a slightly more general condition on \( d \) in (27), where (13) is not required. In case \( N \geq S \), and suppose that we set

\[ d \asymp \frac{f}{M_A^2 \log m}. \]

Then the bounds as shown in the Theorem 2.2 statement still hold. For both cases, we require that \( \lambda \asymp (\| A \|_2 + \| B \|_2) K \sqrt{S} / N \sqrt{\log m / f} \). That is, when either the noise level \( M_\varepsilon \) or the signal strength increases, we need to increase \( \lambda \) correspondingly; moreover, when \( N \) dominates the signal \( K^2 \| \beta^* \|_2^2 \), we have

\[ \left\| \hat{\beta} - \beta^* \right\|_2 / \| \beta^* \|_2 \leq \frac{20}{\alpha} D_2 K^2 \sqrt{\frac{N}{S M_A}} \asymp D_2 K^2 \sqrt{\frac{N}{S}} \left( \frac{1}{\omega(s_0)} \right) \]

which eventually becomes a vacuous bound when \( N \gg S \).

**Remark 2.4.** Throughout this paper, we assume that \( C_0 \) is a large enough constant such that for \( c \) as defined in Theorem D.1,

\[ c \min\{ C_0^2, C_0 \} \geq 4. \] (16)

By definition of \( s_0 \), we have for \( \omega^2(s_0) \geq 1,

\[ s_0 \omega^2(s_0) \leq \frac{c' \lambda^2_{\min}(A)}{1024 C_0^2} \frac{f}{\log m} \quad \text{and hence} \]

\[ s_0 \leq \frac{c' \lambda^2_{\min}(A)}{1024 C_0^2} \frac{f}{\log m} \leq \frac{\lambda^2_{\min}(A)}{1024 C_0^2} \frac{f}{\log m} =: s_0. \]

**Remark 2.5.** The proof shows that one can take \( C = C_0 / \sqrt{c'} \), and take

\[ V = 3 e M_A^3 / 2 = \frac{3 e 64^3 C^3 \omega^3(s_0)}{2 \lambda^3_{\min}(A)} \leq \frac{3 e 64^3 C^3 \omega^3(s_0)}{2 (c')^{3/2} \lambda^3_{\min}(A)} \]

Hence a sufficient condition on \( r(B) \) is:

\[ r(B) \geq 16 c' K^4 \frac{f}{\log m} \left( 3 \log \frac{64 C_0 \omega(s_0)}{\sqrt{c'} \lambda_{\min}(A)} + \log \frac{3 e m \log m}{2 f} \right). \] (17)

**Theorem 2.6.** Suppose (A1) holds. Set \( 0 < \delta < 1 \). Suppose that \( f < m < o(\exp(f)) \) and \( 1 \leq d_0 < f \). Let \( \lambda > 0 \) be as defined in (5). Assume that \( \text{RE}(2d_0, 3(1 + \lambda), A^{1/2}) \) holds. Suppose that the sparsity of \( \beta^* \) is bounded by

\[ d_0 = c_0 \sqrt{f / \log m} \] (18)
for some constant \( c_0 > 0 \); Suppose \( k_0 := 1 + \lambda \)

\[
f \geq \frac{2000dK^4}{\delta^2} \log \left( \frac{60em}{d\delta} \right) \quad \text{where}
\]

\[
d = 2d_0 + 2d_0 a_{\text{max}} - \frac{16K^2(2d_0, 3k_0, A^{1/2})(3k_0)^2(3k_0 + 1)}{\delta^2}.
\]

Consider the regression model in (1a) and (1b) with \( X_0, W \) as in Definition 2.1 and an error vector \( \epsilon \in \mathbb{R}^J \), independent of \( X_0, W \), with independent entries \( \epsilon_j \) satisfying \( E\epsilon_j = 0 \) and \( \|\epsilon_j\|_{\psi_2} \leq M_\epsilon \). Let \( \hat{\beta} \) be an optimal solution to the Conic Programming estimator as in (5) with input \((\hat{\gamma}, \hat{\Gamma})\) as defined in (3), where \( \text{tr}(B) \) is as defined in (9). Then with probability at least \( 1 - \frac{c'}{m} - 2\exp(2f/2000K^4) \), for \( 2 \geq q \geq 1 \)

\[
\| \hat{\beta} - \beta^* \|_q \leq CD_2K^2d_0^{1/q} \sqrt{\frac{\log m}{f}} \left( \|\beta^*\|_2 + \frac{M_\epsilon}{K} \right);
\]

(21)

Under the same assumptions, the predictive risk admits the following bounds with the same probability as above,

\[
\frac{1}{f} \left\| X(\hat{\beta} - \beta^*) \right\|^2_2 \leq C'D_2^2K^4d_0^{\frac{\log m}{f}} (\|\beta^*\|_2 + \frac{M_\epsilon}{K})^2
\]

where \( c', C, C' > 0 \) are some absolute constants.

Similar results have been derived in [27, 2], however, under different assumptions on the distribution of the noise matrix \( W \). We give an outline of the proof of Theorem 2.6 in Section 4 while leaving the detailed proof in Section G. While the rates we obtain for both estimators are at the same order for \( q = 1, 2 \), the conditions under which these rates are obtained are somewhat different. We note that following Theorem 2 as in [2], one can show that without the relatively restrictive sparsity condition (18), a bound similar to that in (21) holds, however with \( \|\beta^*\|_2 \) being replaced by \( \|\beta^*\|_1 \), so long as the sample size satisfies the requirement as in (19).

### 3 Outline of the proof of Theorem 2.2

The main focus of the current paper is to apply Theorem 3.1, which follows from Theorem 1 [27], to show Theorem 2.2, which applies to the general subgaussian model as considered in the present work.

**Theorem 3.1.** Consider the regression model in (1a) and (1b). Let \( d \leq f/2 \). Let \( \gamma, \Gamma \) be as constructed in (3). Suppose that the matrix \( \Gamma \) satisfies the Lower-RE condition with curvature \( \alpha > 0 \) and tolerance \( \tau > 0 \),

\[
\sqrt{d} \tau \leq \min \left\{ \frac{\alpha}{32\sqrt{d}}, \frac{\lambda}{4b_0} \right\}
\]

(22)

where \( d, b_0 \) and \( \lambda \) are as defined in (4). Then for any \( d \)-sparse vectors \( \beta^* \in \mathbb{R}^m \), such that \( \|\beta^*\|_2 \leq b_0 \) and

\[
\left\| \gamma - \Gamma \beta^* \right\|_\infty \leq \frac{1}{2} \lambda,
\]

(23)

we have

\[
\left\| \hat{\beta} - \beta^* \right\|_2 \leq \frac{20}{\alpha} \lambda \sqrt{d}, \quad \text{and} \quad \left\| \hat{\beta} - \beta^* \right\|_1 \leq \frac{80}{\alpha} \lambda d
\]

(24)

where \( \hat{\beta} \) is an optimal solution to the Lasso-type estimator as in (4).
In view of Theorem 3.1, we first consider the following deviation bound as stated in Lemma 3.2. One can then combine with Theorem 3.1, Lemmas 3.3 and 3.4 to prove Theorem 2.2. In more details, Lemma 3.3 checks the Lower and the Upper RE conditions on the modified gram matrix:

\[ \hat{\Gamma}_A := X^T X - \hat{\text{tr}}(B)I_m \]  

(25)

while Lemma 3.4 checks condition (22) as stated in Theorem 3.1 for \( \alpha \) and \( \tau \) as derived in Lemma 3.3. The actual proof of Theorem 2.2 appears in Section F.1.

**Lemma 3.2.** Suppose (A1) holds. Let \( X = X_0 + W \), where \( X_0, W \) are as defined in Theorem 2.2. Let \( \hat{\text{tr}}(B) \) be defined as in (9). Suppose that

\[ \|B\|_F^2 / \|B\|_2^2 \geq \log m. \]

Let \( \hat{\Gamma} \) and \( \hat{\gamma} \) be as in (3). On event \( \mathcal{B}_0 \), we have

\[ \left\| \hat{\gamma} - \hat{\Gamma} \beta^* \right\|_\infty \leq 2\psi \sqrt{\frac{\log m}{f}} \quad \text{where} \quad \psi \asymp K\sqrt{S+N} \]

is as defined in Theorem 2.2. Then \( \mathbb{P}(\mathcal{B}_0) \geq 1 - 8/m^3 \).

Proof of Lemma 3.2 appears in Section F.2. We mention in passing that Lemma 3.2 is essential in proving Theorem 2.6 as well.

### 3.1 Preliminary results for Theorem 2.2

We first state Lemma 3.3, which follows immediately from Corollary 5.1. First, we replace (A3) with (A3’)

(A3’) More precisely, we assume for some large enough constant \( C_K \) and \( D_1 = \|A\|_2 + b_{\max}, \lambda_{\min}(A) > C_K D_1 \sqrt{\frac{\log m}{f}} \) where \( m \geq f = \Omega(\frac{D_1^2 K^4 \log m}{\lambda_{\min}(A)^2}) \).

**Lemma 3.3. (Lower and Upper-RE conditions)** Suppose (A1), (A2) and (A3’) hold. Denote by \( V := 3eM_A^3/2 \), where \( M_A \) is as defined in (11). Suppose that \( m \geq 1024 C_0^3 D_1^2 K^4 \log m/\lambda_{\min}(A)^2 \), where \( D_1 = \|A\|_2 + b_{\max} \). Suppose that for some \( c' > 0 \),

\[ \frac{\text{tr}(B)}{\|B\|_2} \geq c' K^4 s_0 \frac{s_0}{\epsilon^2} \log \left( \frac{3em}{s_0 \epsilon} \right) \quad \text{where} \quad \epsilon = \frac{1}{2M_A}. \]

(26)

Let \( \mathcal{A}_0 \) be the event that the modified gram matrix (25) satisfies the Lower as well as Upper RE conditions with curvature \( \alpha = \frac{1}{2} \lambda_{\min}(A) \), smoothness \( \bar{\alpha} = 3 \lambda_{\max}(A)/2 \) and tolerance \( \tau = \frac{\alpha}{2} \). Then \( \mathbb{P}(\mathcal{A}_0) \geq 1 - 4 \exp \left( -\frac{c_2 f}{M_A^2} \log \left( \frac{\log m}{f} \right) \right) - 2 \exp \left( -\frac{4c_2 f}{M_A^4 K^4} \right) - 6/m^3 \).

**Lemma 3.4.** Suppose all conditions in Lemma 3.3 hold. Suppose that

\[ d := |\supp(\beta^*)| \leq C_A \frac{f}{\log m} \left\{ c' D_\phi \lor 2 \right\} \quad \text{where} \quad C_A := \frac{1}{128 M_A^2}, \]

\[ D_\phi = \left( \frac{K^2 M_A^2}{b_0^2} + K^4 \phi \right) \geq K^4 \phi \geq \phi \]  

(27)
where \( c', \phi, b_0, M, \) and \( K \) are as defined in Theorem 2.2, where we assume that \( \| \beta^* \|^2_2 \geq \phi b_0^2 \) for some \( 0 < \phi \leq 1 \). Then the following condition holds

\[
d \leq \frac{s_0}{32} \wedge \left( \frac{s_0}{\alpha} \right)^2 \log m \left( \frac{\psi}{b_0} \right)^2
\]

(28)

where \( \psi \) is as defined in (15) and \( \alpha = \lambda_{\min}(A)/2 \).

Proofs of Lemmas 3.3 and 3.4 appear in Sections F.3 and F.4 respectively.

4 Outline of the proof for Theorem 2.6

Let \( \Psi = \frac{1}{f} X_0^T X_0 \) where \( X_0 = Z_1 A_1^{1/2} \) as in Definition 2.1. Let \( k_0 = 1 + \lambda \). First we need to define the \( \ell_q \)-sensitivity parameter for \( \Psi \) following [2]

\[
\kappa_q(d_0, k_0) = \min_{J:|J| \leq d_0} \min_{\Delta \in \text{Cone}_J(k_0)} \frac{|\Psi \Delta|_{\infty}}{|\Delta|_q}
\]

where

\[
\text{Cone}_J(k_0) = \{x \in \mathbb{R}^m \mid \text{s.t. } \|x_J\|_1 \leq k_0 \|x_J\|_1\}.
\]

Now by Lemma 6 of [2] and Theorem G.1 [35], we can show that the RE condition and the sample requirement are enough to ensure that the \( \ell_q \)-sensitivity parameter satisfies the following lower bound for all \( 1 \leq q \leq 2 \):

\[
\kappa_q(d_0, k_0) \geq cd_0^{-1/q}
\]

(29)

for some constant \( c \). Combining (29) with Lemmas 4.1 to 4.3 gives us both the lower and upper bounds on \( \left\| \frac{1}{f} X_0^T X_0 v \right\|_{\infty} \), with the lower bound being \( \kappa_q(d_0, k_0) \| v \|_q \) and the upper bound as specified in Lemma 4.3. Following some algebraic manipulation, this yields the bound on the \( \| v \|_q \) for all \( 1 \leq q \leq 2 \). We now state Lemmas 4.1 to 4.3 while leaving the detailed proof for the theorem in Section G. Our first goal is to show that the following holds with high probability for the \( \mu, \tau \) as chosen in (30),

\[
\left\| \frac{1}{f} X_0^T (y - X \beta^*) + \frac{1}{f} \hat{\text{tr}}(B) \beta^* \right\|_{\infty} \leq \mu \| \beta^* \|^2_2 + \tau.
\]

This forms the basis for proving the \( \ell_q \) convergence, where \( q \in [1, 2] \), for the Conic Programming estimator (5).

**Lemma 4.1.** Suppose all conditions in Lemma 3.2 hold. Then on event \( B_0 \) as defined therein, the pair \( (\beta, t) = (\beta^*, \| \beta^* \|_2) \) belongs to the feasible set of the minimization problem (5)

\[
\mu = 2C_0 D_2 K^2 \sqrt{\frac{\log m}{f}} \quad \text{and} \quad \tau = C_0 D_0 K M \sqrt{\frac{\log m}{f}}
\]

(30)

where \( D_0 = (\sqrt{T} + \sqrt{a_{\max}}) \) and \( D_2 = 2(\| A \|^2_2 + \| B \|^2_2) \).
Lemma 4.2. Suppose all conditions in Lemma 4.1 hold. Let $(\hat{\beta}, \hat{\gamma})$ be the optimal solution to (5). Let $S = \text{supp } \beta^*$. On event $B_0$,
\[
\|v_S\|_1 \leq (1 + \lambda)\|v_S\|_1 \quad \text{and} \quad \hat{\gamma} \leq \frac{1}{\lambda} \|v\|_1 + \|\beta^*\|_2.
\]

Lemma 4.3. On event $B_0 \cap B_{10}$, where $B_{10}$ is to be defined in Lemma E.1,
\[
\left\|\frac{1}{f}X_0^T X_0 v\right\|_\infty \leq \mu_1 \|\beta^*\|_2 + \mu_2 \|v\|_1 + \tau
\]
where $\mu_1 = 4D_2K_{m,f}$, $\mu_2 = 2D_2K_{m,f}(\frac{1}{\lambda} + 1)$ and $\tau = 2D_0M_r r_{m,f}$, where $r_{m,f} = C_0K \sqrt{\frac{\log m}{f}}$.

Proofs of Lemmas 4.1 to 4.3 appear in Section G.2.

5 Lower and Upper RE conditions

The goal of this section is to show that for $\Delta$ defined in (34), the presumption in Lemmas H.2 and H.5 as restated in (31) holds with high probability (cf Theorem 5.2). We first state a deterministic result showing that the Lower and Upper RE conditions hold for $\hat{\Gamma}_A$ under condition (31) in Corollary 5.1. This allows us to prove Lemma 3.3 in Sections F.3. See Sections H and I, where we show that Corollary 5.1 follows immediately from the geometric analysis result as stated in Lemma H.5.

Corollary 5.1. Let $1/8 > \delta > 0$. Let $1 \leq \zeta < m/2$. Let $A_{m\times m}$ be a symmetric positive semidefinite covariance matrix. Let $\hat{\Gamma}_A$ be an $m \times m$ symmetric matrix and $\Delta = \hat{\Gamma}_A - A$. Let $E = \bigcup_{|J| \leq \zeta} E_J$, where $E_J = \text{span}\{e_j : j \in J\}$. Suppose that $\forall u, v \in E \cap S^{m-1}$
\[
|u^T \Delta v| \leq \delta \leq \frac{1}{8} \lambda_{\min}(A).
\]
Then the Lower and Upper RE conditions holds: for all $v \in \mathbb{R}^m$,
\[
v^T \hat{\Gamma}_A v \geq \frac{1}{2} \lambda_{\min}(A) \|v\|_2^2 - \frac{\lambda_{\min}(A)}{2 \zeta} \|v\|_1^2
\]
\[
v^T \hat{\Gamma}_A v \leq \frac{3}{2} \lambda_{\max}(A) \|v\|_2^2 + \frac{\lambda_{\min}(A)}{2 \zeta} \|v\|_1^2.
\]

Theorem 5.2. Let $A_{m\times m}$, $B_{x \times f}$ be symmetric positive definite covariance matrices. Let $E = \bigcup_{|J| \leq \zeta} E_J$ for $1 \leq \zeta < m/2$. Let $Z, X$ be $f \times m$ random matrices defined as in Theorem 2.2. Let $\hat{\text{tr}}(B)$ be defined as in (9). Let
\[
\Delta := \hat{\Gamma}_A - A := \frac{1}{f}X^T X - \frac{1}{f} \hat{\text{tr}}(B) I_m - A.
\]
Suppose that for some absolute constant $c' > 0$ and $0 < \varepsilon \leq \frac{1}{c'}$
\[
\frac{\text{tr}(B)}{\|B\|_2} \geq c' K^4 \frac{\zeta}{\varepsilon^2} \log \left(\frac{3em}{c \varepsilon}\right)
\]
where $C = C_0 / \sqrt{c'}$ for $C_0$ as chosen to satisfy (16).
Then with probability at least \( 1 - 4 \exp \left( -c_2 \varepsilon^2 \frac{\text{tr}(B)}{K^4 \|B\|_2} \right) - 2 \exp \left( -c_2 \varepsilon^2 \frac{f}{K^4} \right) - 6/m^3 \), where \( c_2 \geq 2 \), we have for all \( u, v \in E \cap S^{m-1} \) and \( \varpi(\zeta) = \tau_B + \rho_{\text{max}}(\zeta, A) \),

\[
|u^T \Delta v| \leq 8 C \varpi(\zeta) \varepsilon + 2 C_0(\|A\|_2 + b_{\text{max}}) K^2 \sqrt{\frac{\log m}{m}}. \tag{36}
\]

Proof of Theorem 5.2 appears in Section J.

6 Concentration bounds for error-corrected gram matrices

In this section, we show an upper bound on the operator norm convergence as well as an isometry property for estimating \( B \) using the corrected gram matrix \( \tilde{B} := \frac{1}{m} (XX^T - \text{tr}(A) I_f) \). Theorem 6.1 and Corollary 6.2 state that for the matrix \( B \succ 0 \) with the smaller dimension, \( \tilde{B} \) tends to stay positive definite after this error correction step with an overwhelming probability, where we rely on \( f \) being dominated by the effective rank of the positive definite matrix \( A \).

**Theorem 6.1.** Let \( \varepsilon > 0 \). Let \( X \) be defined as in Definition 2.1. Suppose that for some \( c' > 0 \) and \( 0 < \varepsilon < 1/2 \),

\[
\frac{\text{tr}(A)}{\|A\|_2} \geq c' f K^4 \frac{\log(3/\varepsilon)}{\varepsilon^2}. \tag{37}
\]

Then with probability at least \( 1 - 2 \exp \left( -c \varepsilon^2 \frac{m}{K^4} \right) - 4 \exp \left( -c_5 \varepsilon^2 \frac{\text{tr}(A)}{K^4 \|A\|_2} \right) \),

\[
\left\| \frac{1}{m} XX^T - \frac{\text{tr}(A) I_f}{m} - B \right\|_2 \leq C_2 \varepsilon (\tau_A + \|B\|_2)
\]

where \( C_2, c_5 \) are absolute constants depending on \( c', C \), where \( C > 4 \max \left( \frac{1}{c' \varepsilon}, \frac{1}{\sqrt{4 c' c}} \right) \) is a large enough constant.

**Corollary 6.2.** Suppose all conditions in Theorem 6.1 hold. Suppose

\[
\frac{\text{tr}(A)}{\|A\|_2} \geq c' f K^4 \frac{C_2^2}{\delta^2} \log \left( \frac{3C_3}{\delta} \right). \tag{38}
\]

where \( C_3 = C_2 \left( \frac{\tau_A}{\lambda_{\text{min}}(B)} \vee 1 \right) \) for \( C_2 \) as in Theorem 6.1. Then with the probability as stated in Theorem 6.1,

\[
(1 + 2\delta) B \succ \frac{XX^T}{m} - \frac{\text{tr}(A) I_f}{m} \succ (1 - 2\delta) B \succ 0
\]

where for the last inequality to hold, we assume that \( \lambda_{\text{min}}(B) > 0 \).

Proofs of Theorem 6.1 and Corollary 6.2 appear in Section K and Section L. In Appendix C, we show a large deviation bound on the sparse eigenvalues of the error corrected \( \tilde{A} : \tilde{A} := \frac{1}{f} X^T X - \tau_B I_m \).
7 Discussions and future work

The key modeling question is: would each row vector in $W$ for a particular patient across all time points be a correlated normal or subgaussian vector as well? It is our conjecture that combining the newly developed techniques, namely, the concentration of measure inequalities we have derived in the current framework with techniques from existing work, we can handle the case when $W$ follows a matrix normal distribution with a separable covariance matrix $\Sigma_W = C \otimes B$, where $C$ is a $m \times m$ positive semi-definite covariance matrix. Moreover, for this type of "seasonal effects" as the measurement errors, the time varying covariance model would make more sense [54]. We leave the investigation of this more general modeling framework and its estimation procedure to future work. In future work, we will also extend the estimation methods to the settings where the covariates are measured with multiplicative errors which are shown to be reducible to the additive error problem as studied in the present work; see [33, 27]. Moreover, we are interested in applying the analysis and concentration of measure results developed in the current paper to the more general contexts and settings where measurement error models are introduced and investigated; see for example [14, 8, 39, 21, 18, 40, 6, 9, 17, 12, 41, 22, 26, 42, 49, 20, 47, 25, 29, 1, 38, 36, 37].

Acknowledgements

Mark Rudelson is partially supported by NSF grant DMS 1161372 and USAF Grant FA9550-14-1-0009. Shuheng Zhou is supported in part by NSF under Grant DMS-1316731 and Elizabeth Caroline Crosby Funding from the Advance Program at the University of Michigan.

References

[1] G. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics*, 4(2):764–790, 2010.

[2] A. Belloni, M. Rosenbaum, and A. Tsybakov. Linear and conic programming estimators in high-dimensional errors-in-variables models, August 2014. arXiv:0903.2515.

[3] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[4] E. Bonilla, K. Chai, and C. Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems 20 (NIPS 2010)*, 2008.

[5] E. Candès and T. Tao. The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *Annals of Statistics*, 35(6):2313–2351, 2007.

[6] R. Carroll and M. Wand. Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc. B*, 53:573–585, 1991.

[7] R. Carroll, D. Ruppert, L. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models (Second Edition)*. Chapman & Hall, 2006.
[8] R. J. Carroll, P. P. Gallo, and L. J. Gleser. Comparison of least squares and errors-in-variables regression with special reference to randomized analysis of covariance. *Journal of American Statistical Association*, 80:929 – 932, 1985.

[9] R. J. Carroll, M. H. Gail, and J. H. Lubin. Case-control studies with errors in predictors. *Journal of American Statistical Association*, 88:177 – 191, 1993.

[10] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing*, 20:33–61, 1998.

[11] Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *Proceedings of The 30th International Conference on Machine Learning ICML-13*, 2013.

[12] J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1994.

[13] A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, 68:265–274, 1981.

[14] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):138, 1977.

[15] P. Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

[16] B. Efron. Are a set of microarrays independ of each other? *Ann. App. Statist.*, 3(3):922–942, 2009.

[17] J. Fan and Y. K. Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, 21:1900–1925, 1993.

[18] W. A. Fuller. *Measurement error models*. John Wiley and Sons, 1987.

[19] A. Gupta and T. Varga. Characterization of matrix variate normal distributions. *Journal of Multivariate Analysis*, 41:80–88, 1992.

[20] P. Hall and Y. Ma. Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society B*, 69:429–446, 2007.

[21] J. T. Hwang. Multiplicative errors-in-variables models with applications to recent data released by the u.s. department of energy. *Journal of American Statistical Association*, 81:680–688, 1986.

[22] S. J. Iturria, R. J. Carroll, and D. Firth. Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society, Series B, Methodological*, 61:547561, 1999.

[23] A. Kalaitzis, J. Lafferty, N. Lawrence, and S. Zhou. The bigraphical lasso. In *Proceedings of The 30th International Conference on Machine Learning ICML-13*, pages 1229–1237, 2013.

[24] C. Leng and C. Tang. Sparse matrix graphical models. *Journal of American Statistical Association*, 107:1187–1200, 2012.
[25] H. Liang and R. Li. Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104:234–248, 2009.

[26] H. Liang, W. Härdle, and R. J. Carroll. Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.*, 27:1519–1535, 1999.

[27] P. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[28] N. Lu and D. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.*, 73(4):449–457, 2005.

[29] Y. Ma and R. Li. Variable selection in measurement error models. *Bernoulli*, 16:274–300, 2010.

[30] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[31] V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Mathematics 1200*. Springer, 1986.

[32] M. Rosenbaum and A. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[33] M. Rosenbaum and A. Tsybakov. Improved matrix uncertainty selector. *IMS Collections*, 9:276–290, 2013.

[34] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[35] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT’12)*, June 2012.

[36] Ø. Sørensen, A. Frigenssi, and M. Thoresen. Measurement error in Lasso: Impact and likelihood bias correction. *Statistical Sinica Preprint*, 2014.

[37] Ø. Sørensen, A. Frigenssi, and M. Thoresen. Covariate selection in high-dimensional generalized linear models with measurement error, 2014. arXiv:1407.1070.

[38] N. Städler, D. J. Stekhoven, and P. Bühlmann. Pattern alternating maximization algorithm for missing data in high-dimensional problems. *Journal of Machine Learning Research*, 15:1903–1928, 2014.

[39] L. A. Stefanski. The effects of measurement error on parameter estimation. *Biometrika*, 72:583–592, 1985.

[40] L. A. Stefanski. Rates of convergence of some estimators in a class of deconvolution problems. *Statistics and Probability Letters*, 9:229–235, 1990.

[41] L. A. Stefanski and J. R. Cook. Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90:1247–1256, 1995.

[42] K. Strimmer. Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics*, 4(10), 2003.
A Outline

We prove Lemma 1.3 in Section B. As a corollary of Theorem 5.2, we prove Corollary C.2 in Section C. The rest of the appendix contains the technical proofs for the Lasso and the Dantzig-type estimators. In Sections D and E, we present variations of the Hanson-Wright inequality as recently derived in [34] (cf. Lemma D.2), concentration of measure bounds and stochastic error bounds in the error-in-variables models as considered in Theorems 2.2 and 2.6.
B  Proof of Lemma 1.3

We define $\text{Cone}(d_0, k_0)$, where $0 < d_0 < m$ and $k_0$ is a positive number, as the set of vectors in $\mathbb{R}^m$ which satisfy the following cone constraint:

$$\text{Cone}(d_0, k_0) = \{ x \in \mathbb{R}^m | \exists I \in \{1, \ldots, p\}, |I| = d_0 \text{ s.t. } \|x_I\|_1 \leq k_0 \|x_I\|_1 \} .$$

For each vector $x \in \mathbb{R}^p$, let $T_0$ denote the locations of the $s_0$ largest coefficients of $x$ in absolute values. The following elementary estimate \cite{35} will be used in conjunction with the RE condition.

**Lemma B.1.** For each vector $x \in \text{Cone}(s_0, k_0)$, let $T_0$ denote the locations of the $s_0$ largest coefficients of $x$ in absolute values. Then

$$\|x_{T_0}\|_2 \geq \frac{\|x\|_2}{\sqrt{1 + k_0}} . \tag{39}$$

**Proof of Lemma 1.3.** Part I: Suppose that the Lower-RE condition holds for $\Gamma := A^T A$. Let $x \in \text{Cone}(s_0, k_0)$. Then

$$\|x\|_1 \leq (1 + k_0) \|x_{T_0}\|_1 \leq (1 + k_0) s_0 \|x_{T_0}\|_2 .$$

Thus for $x \in \text{Cone}(s_0, k_0) \cap S^{p-1}$ and $\tau (1 + k_0)^2 s_0 \leq \alpha / 2$, we have

$$\|Ax\|_2 = (x^T A^T A x)^{1/2} \geq \left( \alpha \|x\|_2^2 - \tau \|x\|_1^2 \right)^{1/2} \geq \left( \alpha \|x\|_2^2 - \tau (1 + k_0)^2 s_0 \|x_{T_0}\|_2^2 \right)^{1/2} \geq (\alpha - \tau (1 + k_0)^2 s_0)^{1/2} \geq \sqrt{\frac{\alpha}{2}} .$$

Thus the $\text{RE}(s_0, k_0, A)$ condition holds with

$$\frac{1}{K(s_0, k_0, A)} := \min_{x \in \text{Cone}(s_0, k_0)} \frac{\|Ax\|_2}{\|x_{T_0}\|_2} \geq \sqrt{\frac{\alpha}{2}}$$

where we use the fact that for any $J \in \{1, \ldots, p\}$ such that $|J| \leq s_0$, $\|x_J\|_2 \leq \|x_{T_0}\|_2$. We now show the other direction.

Part II. Assume that $\text{RE}(4R^2, 2R - 1, A)$ holds for some integer $R > 1$. Assume that for some $R > 1$

$$\|x\|_1 \leq R \|x\|_2 .$$

Let $(x^*_i)_{i=1}^p$ be non-increasing arrangement of $(\|x_i\|)_{i=1}^p$. Then

$$\|x\|_1 \leq R \left( \sum_{j=1}^s (x^*_j)^2 + \sum_{j=s+1}^{\infty} \left( \frac{\|x\|_1}{j} \right)^2 \right)^{1/2} \leq R \left( \|x_{T_0}^*\|_2 + \|x\|_1 \frac{1}{\sqrt{s}} \right)^{1/2} \leq R \left( \|x\|_2 + \|x\|_1 \frac{1}{\sqrt{s}} \right)$$

18
where \( J := \{1, \ldots, s\} \). Choose \( s = 4R^2 \). Then

\[
\|x\|_1 \leq R \|x^*_J\|_2 + \frac{1}{2} \|x\|_1.
\]

Thus we have

\[
\|x\|_1 \leq 2R \|x^*_J\|_2 \leq 2R \|x^*_J\|_1 \quad \text{and hence} \quad (40)
\]

\[
\|x^*_J\|_1 \leq (2R - 1) \|x^*_J\|_1. \quad (41)
\]

Then \( x \in \text{Cone}(4R^2, 2R - 1) \). Then for all \( x \in \mathcal{S}^{p-1} \) such that \( \|x\|_1 \leq R \|x\|_2 \), we have for \( k_0 = 2R - 1 \) and \( s_0 := 4R^2 \),

\[
x^T \Gamma x \geq \frac{\|x_{T_0}\|_2^2}{K^2(s_0, k_0, A)} \geq \frac{\|x\|_2^2}{\sqrt{s_0}K^2(s_0, k_0, A)} =: \alpha \|x\|_2^2
\]

where we use the fact that \((1 + k_0) \|x_{T_0}\|_2^2 \geq \|x\|_2^2\) by Lemma B.1 with \( x_{T_0} \) as defined therein. Otherwise, suppose that \( \|x\|_1 \geq R \|x\|_2 \). Then for a given \( \tau > 0 \),

\[
\alpha \|x\|_2^2 - \tau \|x\|_1^2 \leq \left( \frac{1}{\sqrt{s_0}K^2(s_0, k_0, A)} - \tau R^2 \right) \|x\|_2^2. \quad (42)
\]

Thus we have by the choice of \( \tau \) as in (7) and (42)

\[
x^T \Gamma x \geq \lambda_{\min}(\Gamma) \|x\|_2^2 \geq \left( \frac{1}{\sqrt{s_0}K^2(s_0, k_0, A)} - \tau R^2 \right) \|x\|_2^2 \geq \alpha \|x\|_2^2 - \tau \|x\|_1^2.
\]

The Lemma thus holds. \( \square \)

### C Sparse eigenvalues

When we subtract a diagonal matrix \( \tau_B I_m \) from the gram matrix \( \frac{1}{f}X^T X \) to form an estimator, we clearly introduce a large number of negative eigenvalues when \( f \ll m \). This in general is a bad idea. However, the sparse eigenvalues for \( \tilde{A} \) can stay pretty close to those of \( A \) as we will show in Corollary C.2 in Section C. We start with a definition.

**Definition C.1.** For \( m < p \), we define the largest and smallest \( m \)-sparse eigenvalue of a \( p \times q \) matrix \( A \) to be

\[
\rho_{\max}(m, A) := \max_{t \neq 0; m-\text{sparse}} \frac{\|At\|_2^2}{\|t\|_2^2}, \quad (43)
\]

\[
\rho_{\min}(m, A) := \min_{t \neq 0; m-\text{sparse}} \frac{\|At\|_2^2}{\|t\|_2^2}. \quad (44)
\]

**Corollary C.2.** Let \( X \) be defined as in Definition 2.1. Let \( \tilde{A} := \frac{1}{f}X^T X - \tau_B I_m \). Suppose

\[
\frac{\text{tr}(B)}{\|B\|_2} \geq c' k^4 \log(3em/k\varepsilon) \frac{\log(3em/k\varepsilon)}{\varepsilon^2}, \quad (45)
\]

19
Then with probability at least $1 - 2 \exp(-c_4 \epsilon^2 K) - 4 \exp(-c_4 \epsilon^2 \frac{\text{tr}(B)}{K \|B\|_2})$,

$$
\rho_{\text{max}}(k, \tilde{A}) \leq \rho_{\text{max}}(k, A)(1 + 10\epsilon) + C_4 \epsilon \tau_B
$$

where $C_4$ is an absolute constant. Moreover, suppose for $C_5 = C_4 \left( \frac{\tau_B}{\rho_{\text{min}}(k, A)} \vee 1 \right)$

$$
\frac{\text{tr}(B)}{\|B\|_2} \geq c' K^4 C_5^2 \frac{\delta^2}{m} \log \left( \frac{12 C_5 \epsilon m}{K \delta} \right).
$$

(46)

Then with the probability as stated immediately above, we have

$$
\rho_{\text{min}}(k, \tilde{A}) \geq \rho_{\text{min}}(k, A)(1 - 2\delta).
$$

D Some auxiliary results

We first need to state the following form of the Hanson-Wright inequality as recently derived in Rudelson and Vershynin [34], and an auxiliary result in Lemma D.2 which may be of independent interests.

**Theorem D.1.** Let $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ be a random vector with independent components $X_i$ which satisfy $\mathbb{E}(X_i) = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let $A$ be an $m \times m$ matrix. Then, for every $t > 0$,

$$
\mathbb{P} (|X^TAX - \mathbb{E}(X^TAX)| > t) \leq 2 \exp \left[ -c \min \left( \frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2} \right) \right].
$$

We note that following the proof of Theorem D.1, it is clear that the following holds: Let $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ be a random vector as defined in Theorem D.1. Let $Y, Y'$ be independent copies of $X$. Let $A$ be an $m \times m$ matrix. Then, for every $t > 0$,

$$
\mathbb{P} (|Y^TAY'| > t) \leq 2 \exp \left[ -c \min \left( \frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2} \right) \right].
$$

(47)

We next need to state Lemma D.2, which we prove in Section M.

**Lemma D.2.** Let $u, w \in S^{f-1}$. Let $A > 0$ be a $m \times m$ symmetric positive definite matrix. Let $Z$ be an $f \times m$ random matrix with independent entries $Z_{ij}$ satisfying $\mathbb{E}Z_{ij} = 0$ and $\|Z_{ij}\|_{\psi_2} \leq K$. Let $Z_1, Z_2$ be independent copies of $Z$. Then for every $t > 0$,

$$
\mathbb{P} \left( |u^T Z_1 A^{1/2} Z_2^T w| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \text{tr}(A)}, \frac{t}{K^2 \|A\|_2^2} \right) \right),
$$

$$
\mathbb{P} \left( |u^T ZAZ^T w - \mathbb{E}u^T ZAZ^T w| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2} \right) \right)
$$

where $c$ is the same constant as defined in Theorem D.1.
E  Stochastic error terms

The following large deviation bounds in Lemmas E.1 and E.2 are the key results in proving Lemmas 3.2 and 4.3. Throughout this section, we denote by:

\[ r_{m,f} = C_0 K \sqrt{\frac{\log m}{f}} \quad \text{and} \quad r_{m,m} = C_0 K \sqrt{\frac{\log m}{m}}. \]

We also define some events \( B_4, B_5, B_6, B_{10} \). Denote by \( B_0 := B_4 \cap B_5 \cap B_6 \), which we use throughout this paper.

**Lemma E.1.** Assume that \( \text{tr}(B) / \|B\|_2 \geq \log m \). Let \( Z, X_0 \) and \( W \) as defined in Theorem 2.2. Let \( Z_0, Z_1 \) and \( Z_2 \) be independent copies of \( Z \). Let \( \epsilon^T \sim Y M_e / K \) where \( Y := e_1^T Z_0^T \). Then on event \( B_4 \), where \( \mathbb{P}(B_4) \geq 1 - \frac{4}{m^3} \), we have

\[ \frac{1}{f} \left\| A^T Z_1 \epsilon \right\|_\infty \leq r_{m,f} \max \{ A \} \sqrt{f} \quad \text{and} \quad \frac{1}{f} \left\| Z_2^T B \epsilon \right\|_\infty \leq r_{m,f} \sqrt{\tau_B} \]

where \( \tau_B = \frac{\text{tr}(B)}{f} \). Moreover, assume that \( \|B\|_F^2 / \|B\|_2^2 \geq \log m \). Then on \( B_5 \), where \( \mathbb{P}(B_5) \geq 1 - 4/m^3 \), we have

\[ \frac{1}{f} \left\| (Z^T B Z - \text{tr}(B) I_m) \beta^* \right\|_\infty \leq r_{m,f} K \| \beta^* \|_2 \sqrt{\frac{\|B\|_F}{\sqrt{f}}} \quad \text{and} \quad \frac{1}{f} \left\| X_0^T W \beta^* \right\|_\infty \leq r_{m,f} K \| \beta^* \|_2 \sqrt{\tau_B} \max \{ A \} \]

Finally, on \( B_{10} \), where \( \mathbb{P}(B_{10}) \geq 1 - 4/m^2 \), we have

\[ \frac{1}{f} \left\| (Z^T B Z - \text{tr}(B) I_m) \right\|_\max \leq r_{m,f} K \frac{\|B\|_F}{\sqrt{f}} \quad \text{and} \quad \frac{1}{f} \left\| X_0^T W \right\|_\max \leq r_{m,f} K \| \beta^* \|_2 \sqrt{\tau_B} \max \{ A \} \]

**Lemma E.2.** Suppose all conditions in Lemma 3.2 hold. Then on event \( B_6 \), which holds with probability \( 1 - \frac{6}{m^2} \), we have for \( D_1 = \|A\|_2 + b_{\max} \)

\[ \frac{1}{f} \left| \text{tr}(B) - \text{tr}(B) \right| \leq C_0 \left( \tau_B + 2 \tau A \frac{1}{2} \max \{ A \} \sqrt{f} + \frac{\|A\|_F}{\sqrt{m}} \right) K^2 \sqrt{\frac{\log m}{m}} \leq 2 D_1 K r_{m,m} \]

where \( C_0 \) satisfies (16) for \( c \) as defined in Theorem D.1.

We prove Lemmas E.1 and E.2 in Section N.

F  Proofs for the Lasso-type estimator

We include a proof for Theorem 3.1 in Section O for the sake of self-containment.
F.1 Proof of Theorem 2.2

First we note that it is sufficient to have (12) in order for (26) to hold. (12) guarantees that for \( \mathcal{V} = 3eM_A^3/2 \)

\[
r(B) := \frac{\text{tr}(B)}{\|B\|_2} \geq 16\epsilon' K^4 \frac{f}{\log m} \sqrt{\mathcal{V} m \log m} \]

\[
\geq 16\epsilon' K^4 \frac{f}{\log m} \log \left( \frac{3emM_A^3 \log m}{2f} \right)
\]

\[
= \epsilon' K^4 \frac{4}{\epsilon^2} \frac{f}{M_A^2 \log m} \log \left( \frac{6emM_A}{s_0} \right)
\]

\[
\geq \epsilon' K^4 \frac{4}{\epsilon^2} s_0 \log \left( \frac{6emM_A}{s_0} \right) = \epsilon' K^4 \frac{4}{\epsilon^2} s_0 \log \left( \frac{3em}{s_0 \epsilon} \right)
\]

where \( \epsilon = \frac{1}{2M_A} \leq \frac{1}{128C} \), and the last inequality holds given that \( k \log(c m/k) \) on the RHS of (48) is a monotonically increasing function of \( k \), and

\[
s_0 \leq \frac{4f}{M_A^2 \log m} \quad \text{and} \quad M_A = \frac{64C(\rho_{\text{max}}(s_0, A) + \tau_B)}{\lambda_{\text{min}}(A)} \geq 64C.
\]

Next we check that the choice of \( d \) as in (14) ensures that (27) holds for Indeed, for \( \epsilon' K^4 \leq 1 \), we have

\[
d \leq C_A (\epsilon' K^4 \lambda \wedge 1) \frac{\phi f}{\log m} \leq C_A (\epsilon' D_f \wedge 1) \frac{f}{\log m}.
\]

By Lemma 3.3, we have on event \( \mathcal{A}_0 \), the modified gram matrix \( \hat{\Gamma}_A := \frac{1}{f} (X^T X - \text{tr}(B) I_m) \) satisfies the Lower RE conditions with

\[
\text{curvature} \quad \alpha = \frac{1}{2} \lambda_{\text{min}}(A) \quad \text{and tolerance} \quad \tau = \frac{\lambda_{\text{min}}(A)}{2s_0} = \frac{\alpha}{s_0}.
\]

Theorem 2.2 follows from Theorem 3.1, so long as we can show that condition (22) holds for \( \lambda \geq 4\psi \sqrt{\log m} \) where the parameter \( \psi \) is as defined (15), and \( \alpha \) and \( \tau = \frac{\alpha}{s_0} \) are as defined immediately above. Combining (49) and (22), we need to show (28) holds. This is precisely the content of Lemma 3.4. This is the end of the proof for Theorem 2.2 \( \square \)

F.2 Proof of Lemma 3.2

First notice that

\[
\hat{\gamma} = \frac{1}{f} \left( X_0^T X_0 \beta^* + W^T X_0 \beta^* + X_0^T \epsilon + W^T \epsilon \right)
\]

\[
\frac{1}{f} (X^T X - \text{tr}(B) I_m) \beta^* = \frac{1}{f} (X_0^T X + W^T X_0 + X_0^T W + W^T W - \text{tr}(B) I_m) \beta^*
\]
Thus
\[
\| \tilde{\gamma} - \tilde{\beta}^*\|_\infty \leq \| \tilde{\gamma} - \frac{1}{f} (X^T X - \hat{u}(B) I_m) \beta^*\|_\infty
\]
\[
= \frac{1}{f} \| X_0^T \epsilon + W^T \epsilon - (W^T W + X_0^T W - \hat{u}(B) I_m) \beta^*\|_\infty
\]
\[
\leq \frac{1}{f} \| X_0^T \epsilon + W^T \epsilon \|_\infty + \frac{1}{f} \| (W^T W - \hat{u}(B) I_m) \beta^*\|_\infty + \frac{1}{f} \| X_0^T W \beta^*\|_\infty
\]
By Lemma E.1 we have on \(B_4\)
\[
\frac{1}{f} \| X_0^T \epsilon + W^T \epsilon \|_\infty = \frac{1}{f} \| A^2 Z_1^T \epsilon + Z_2^T B \epsilon \|_\infty \leq r_{m,f} M \epsilon \left( \frac{1}{2} a_{\max} + \sqrt{T_B} \right),
\]
and on event \(B_5\),
\[
\frac{1}{f} \| (Z^T B Z - \text{tr}(B) I_m) \beta^*\|_\infty + \frac{1}{f} \| X_0^T W \beta^*\|_\infty \leq r_{m,f} K \| \beta^*\|_2 \left( \frac{\| B \|_F}{\sqrt{f}} + \sqrt{T_B} a_{\max} \right)
\]
\[
\leq K \| \beta^*\|_2 r_{m,f} \left( \| B \|_2 + \frac{1}{2} T_B + \frac{1}{2} \| a_{\max} \|\right)
\]
where recall \(\| B \|_F \leq \sqrt{\text{tr}(B)} \| B \|_2^{1/2}\). Denote by \(B_0 := B_4 \cap B_5 \cap B_6\). We have on \(B_0\) and under (A1), by Lemmas E.2 and E.1, for \(f \leq m\)
\[
\frac{1}{f} \left( \| (Z^T B Z - \text{tr}(B) I_m) \beta^*\|_\infty + \| X_0^T W \beta^*\|_\infty \right)
\]
\[
+ \frac{1}{f} \left( \| X_0^T \epsilon + W^T \epsilon \|_\infty + \| \hat{u}(B) - \text{tr}(B) \| \beta^*\|_\infty \right)
\]
\[
\leq K \| \beta^*\|_2 r_{m,f} \left( \| B \|_2 + \frac{1}{2} (a_{\max} + T_B) + 2 \| A \|_2 + 2 \| b_{\max} \| \right)
\]
\[
+ M \epsilon r_{m,f} \left( \sqrt{T_B} + a_{\max}^{1/2} \right)
\]
\[
\leq 2C_0 D_2 K^2 \| \beta^*\|_2 \sqrt{\frac{\log m}{f}} + C_0 D_0 M \epsilon K \sqrt{\frac{\log m}{f}}
\]
where we further bound for \(D_0 := \sqrt{T_B} + a_{\max}^{1/2}\)
\[
D_0 \leq \sqrt{2 \tau_B + a_{\max}} \leq 2 (\tau_B + a_{\max}) \leq 2 (\| A \|_2 + \| B \|_2) = D_2 \text{ and}
\]
\[
\| B \|_2 + \frac{1}{2} (\tau_B + a_{\max}) + 2 \| A \|_2 + 2 \| b_{\max} \| \leq 4 (\| A \|_2 + \| B \|_2) = 2 D_2
\]
given that under (A1) : \(\tau_A = 1, \| A \|_2 \geq a_{\max} \geq a_{\max}^{1/2} \geq 1\). Hence the lemma holds with \(\psi = C_0 D_2 K (K \| \beta^*\|_2 + M \epsilon)\). Finally, we have by the union bound, \(\mathbb{P}(B_0) \geq 1 - 8/m^3\). \(\square\)

**Remark F.1.** Notice that we can have an alternative bound as follows:
\[
LHS \leq \| \beta^*\|_2 C_0 K^2 \sqrt{\frac{\log m}{f}} \sqrt{T_B} \left(2 \| B \|_2^{1/2} + \frac{1}{2} a_{\max}\right) + C_0 K^2 \sqrt{\frac{\log m}{m}} \sqrt{T_B} \left(2 \| B \|_2^{1/2} + \frac{1}{2} a_{\max}\right)
\]
\[
\leq D_2 (K \| \beta^*\|_2 + M \epsilon) r_{m,f}
\]
where $D_2 = \left(2\|B\|_2^{1/2} + \|A\|_2^{1/2}\right)(\sqrt{\tau_B} + 1)$ which can in turn affect the penalty term. This provides a slight improvement in case $\tau_B = O(1)$ while $\|A\|_2 \gg 1$.

F.3 Proof of Lemma 3.3

Condition (26) implies that (35) in Theorem 5.2 holds for $\zeta = s_0$ and $\varepsilon = \frac{1}{2M_A}$. Now, by Theorem 5.2, we have $\forall u, v \in E \cap S^{m-1}$, under (A1) and (A3), condition (31) holds under event $\mathcal{A}_0$, and so long as $m \geq 1024C_0^2D_1^2K^4 \log m/\lambda_{\min}(A)^2$,

$$\frac{|u^T\Delta v|}{s} \leq 8C\varpi(s_0)\varepsilon + 2C_0D_1K^2\sqrt{\log m \over m} =: \delta$$

which holds for all $\varepsilon \leq \frac{1}{2}\lambda_{\min}(A) = \frac{1}{2M_A} \leq \frac{1}{128C}$

with $\mathbb{P}(\mathcal{A}_0) \geq 1 - 4\exp\left(-c_2\varepsilon^2{\rm tr}(B) \over \|B\|_2\|\right) - 2\exp\left(-c_2\varepsilon^2{\hat{f} \over \hat{K}^2}\right) - 6/7^3$. Hence, by Corollary 5.1, $\forall \theta \in \mathbb{R}^m$,

$$\theta^T\hat{A}_{\theta} \geq \alpha \|\theta\|_2^2 - \tau \|\theta\|_1^2$$

and $\theta^T\hat{A}_{\theta} \leq \tilde{\alpha} \|\theta\|_2^2 + \tau \|\theta\|_1^2$

where $\alpha = \frac{1}{2}\lambda_{\min}(A)$ and $\tilde{\alpha} = \frac{3}{2}\lambda_{\max}(A)$ and

$$\frac{512C^2\varpi(s_0)^2\log m}{\lambda_{\min}(A)} \leq \tau = \frac{\alpha}{s_0} \leq \frac{2\alpha}{s_0 + 1} \leq \frac{1024C^2\varpi^2(s_0 + 1)\log m}{\lambda_{\min}(A)}.$$

where we plugged in $s_0$ as defined in (10). The lemma is thus proved in view of Remark F.2. \(\Box\)

Remark F.2. Clearly the condition on $\parallel B \parallel_2$ as stated in Lemma 3.3 ensures that we have for $\varepsilon = \frac{1}{2M_A}$ and $s_0 \asymp \frac{4f}{M_A^2 \log m}$

$$\varepsilon^2{\rm tr}(B) \over K^4 \parallel B \parallel_2 \geq \frac{\varepsilon^2}{K^4}c'K^4s_0\log\left(\frac{3em}{s_0\varepsilon}\right) \geq \frac{1}{4M_A^2K^4}c'K^4M_A^2s_0\log\left(\frac{6emM_A}{s_0}\right) \geq c's_0\log\left(\frac{6emM_A}{s_0}\right)$$

and hence

$$\exp\left(-c_2\varepsilon^2{\rm tr}(B) \over K^4 \parallel B \parallel_2\right) \leq \exp\left(-c_2s_0\log\left(\frac{6emM_A}{s_0}\right)\right) \asymp \exp\left(-c_3{4f \over M_A^2 \log m \log m} \right)$$
F.4 Proof of Lemma 3.4

Let

\[ M_+ = \frac{64C \varpi(s_0 + 1)}{\lambda_{\min}(A)} \]

where \( \varpi(s_0 + 1) = \rho_{\max}(s_0 + 1, A) + \tau_B =: D \)

By definition of \( s_0 \), we have

\[ \sqrt{s_0 + 1} \varpi(s_0 + 1) \geq \frac{\lambda_{\min}(A)}{32C} \sqrt{\frac{f}{\log m}} \]

and hence

\[ s_0 + 1 \geq \frac{\lambda_{\min}^2(A)}{1024C^2 \varpi^2(s_0 + 1) \log m} = \left( \frac{\alpha}{16CD} \right)^2 \frac{f}{\log m} \geq \frac{1}{M_A^2} \frac{f}{\log m} \]

The first inequality in (22) holds given that \( M_+ \leq 2M_A \) and hence

\[ d \leq \frac{1}{64M_A^2} \frac{f}{\log m} \leq \frac{1}{16M_A^2} \frac{f}{\log m} \leq \frac{s_0 + 1}{64} \leq \frac{s_0}{32} \]

Moreover, for \( D = \rho_{\max}(s_0 + 1, A) + \tau_B \leq D_2 \) and \( C = C_0/\sqrt{c} \), we have

\[ d \leq C_{Ac'}D_\phi \frac{f}{\log m} \leq \frac{1}{128M_A^4} \left( \frac{C_0D_2}{CD} \right)^2 D_\phi \frac{f}{\log m} \]

\[ \leq \frac{1}{2} \left( \frac{1}{16CD} \right)^2 \frac{4C_0^2D_2^2D_\phi}{M_A^4} \frac{f}{\log m} \]

\[ \leq \frac{1}{2} \frac{(s_0 + 1)^2 \log m}{\alpha^2} \frac{f}{\log m} \left( \frac{\psi}{b_0} \right)^2 \leq \frac{(s_0)^2 \log m}{\alpha^2} \frac{f}{\log m} \left( \frac{\psi}{b_0} \right)^2 \]

where assuming that \( s_0 \geq 3 \), we have

\[ \frac{2s_0^2}{\alpha^2} \geq \left( \frac{s_0 + 1}{\alpha} \right)^2 \geq \frac{\alpha^2}{(16CD)^4} \left( \frac{f}{\log m} \right)^2 \]

\[ \left( \frac{\psi}{b_0} \right)^2 = 4C_0^2D_2^2K^2 \frac{\beta^*}{b_0^2} (M_\epsilon + K \|\beta^*\|_2)^2 \geq 4C_0^2D_2^2D_\phi \]

We have shown that (28) indeed holds, and the lemma is thus proved. \( \square \)

Remark F.3. Clearly for \( d, b_0, \phi \) as bounded in Theorem 2.2, we have by assumption (13) the following upper and lower bound on \( D_\phi \):

\[ 2K^4 \phi \geq D_\phi := K^4 \left( \frac{M_A^2}{K^2b_0^2} + \phi \right) \geq K^4 \phi. \]

Moreover, in order to obtain the error bounds in Theorem 2.2, (13) is not needed so long as conditions on \( r(B) \) and \( \lambda \) as stated therein hold.
Remark F.4. Examining the proof of Lemma 3.4, we note that the following relaxed condition on $d$ is enough for Theorem 2.2 to hold:

$$d \leq C_A \left\{ C_\phi \sqrt{2} \right\} \frac{f}{\log m} \text{ where } C_\phi = \left( \frac{C_0 D_2}{C D} \right)^2 D_\phi$$

where $C_A = \frac{1}{128M^2}$ and $b_0^2 \geq \|\beta^*\|_2^2 \geq \|b\|_0^2$.

An alternative bound for $D_2$ is: $D_2 = a_{\max} + b_{\max} + \frac{\|A\|_F}{\sqrt{m}} \vee \frac{\|B\|_F}{\sqrt{j}}$. We can plug this in the inequality immediately above to relax the condition on $d$.

G Proofs for the Conic Programming estimator

For the set $\text{Cone}_J(k_0)$ as in (39),

$$\kappa_{\text{RE}}(d_0, k_0) := \min_{J: |J| \leq d_0} \min_{\Delta \in \text{Cone}_J(k_0)} \frac{\| \Delta^T \Psi \Delta \|}{\| \Delta J \|_2^2} = \left( \frac{1}{K(d_0, k_0, (1/\sqrt{f}) Z_1 A^{1/2})} \right)^2.$$ 

Recall the following Theorem G.1 from [35].

**Theorem G.1.** Set $0 < \delta < 1$, $k_0 > 0$, and $0 < d_0 < p$. Let $A^{1/2}$ be an $m \times m$ matrix satisfying $\text{RE}(d_0, 3k_0, A^{1/2})$ condition as in Definition 1.1. Let $d$ be as defined in (50)

$$d = d_0 + d_0 \max_j \left\| A^{1/2} \epsilon_j \right\|^2_2 \frac{16 K^2(d_0, 3k_0, A^{1/2})(3k_0)^2(3k_0 + 1)}{\delta^2}. \quad (50)$$

Let $\Psi$ be an $n \times m$ matrix whose rows are independent isotropic $\psi_2$ random vectors in $\mathbb{R}^m$ with constant $\alpha$. Suppose the sample size satisfies

$$n \geq \frac{200d_0 \alpha^4}{\delta^2} \log \left( \frac{60em}{d_0} \right). \quad (51)$$

Then with probability at least $1 - 2 \exp(\delta^2 n/2000 \alpha^2)$, $\text{RE}(d_0, k_0, (1/\sqrt{n}) \Psi A^{1/2})$ condition holds for matrix $(1/\sqrt{n}) \Psi A$ with

$$0 < K(d_0, k_0, (1/\sqrt{n}) \Psi A^{1/2}) \leq \frac{K(d_0, k_0, A^{1/2})}{1 - \delta}. \quad (52)$$

**G.1 Proof of Theorem 2.6**

Suppose $\text{RE}(2d_0, 3k_0, A^{1/2})$ holds. Then for $d$ as defined in (20) and $f = \Omega(dK^4 \log(m/d))$, we have with probability at least $1 - 2 \exp(\delta^2 f/2000K^4)$, the $\text{RE}(2d_0, k_0, (1/\sqrt{f}) Z_1 A^{1/2})$ condition holds with

$$\kappa_{\text{RE}}(2d_0, k_0) = \left( \frac{1}{K(2d_0, k_0, (1/\sqrt{f}) Z_1 A^{1/2})} \right)^2 \geq \left( \frac{1}{2K(2d_0, k_0, A^{1/2})} \right)^2$$

by Theorem G.1.
The rest of the proof follows from [2] Theorem 1 and thus we only provide a sketch. In more details, in view of the lemmas proved in Section 4, we need
\[ \kappa_q(d_0, k_0) \geq cd_0^{-1/q} \]
to hold for some constant \( c \) for \( \Psi := \frac{1}{T}X_0^T X_0 \). It is shown in Appendix C in [2] that under the \( \text{RE}(2d_0, k_0, \frac{1}{\sqrt{T}}Z_1A^{1/2}) \) condition, for any \( d_0 \leq m/2 \) and \( 1 \leq q \leq 2 \), we have
\[ \kappa_1(d_0, k_0) \geq cd_0^{-1} \kappa_{\text{RE}}(d_0, k_0), \]
\[ \kappa_q(d_0, k_0) \geq c(q)d_0^{-1/q} \kappa_{\text{RE}}(2d_0, k_0) \]
where \( c(q) > 0 \) depends on \( k_0 \) and \( q \). The theorem is thus proved following exactly the same line of arguments as in the proof of Theorem 1 in [2] in view of the \( \ell_q \) sensitivity condition derived immediately above, in view of Lemmas 4.1 to 4.3.

For \( v := \hat{\beta} - \beta^* \), we have by Lemmas 4.1 to 4.3
\[ \kappa_q(d_0, k_0) \|v\|_q \leq \left\| \frac{1}{T}X_0^T X_0 v \right\|_\infty \]
\[ \leq \mu_1 \|\beta^*\|_2 + \mu_2 \|v\|_1 + \tau \]
\[ \leq \mu_1 \|\beta^*\|_2 + \mu_2(2 + \lambda) \|v_S\|_1 + \tau \]
\[ \leq \mu_1 \|\beta^*\|_2 + \mu_2(2 + \lambda)d_0^{1-1/q} \|v_S\|_q + \tau \]
\[ \leq \mu_1 \|\beta^*\|_2 + \mu_2(2 + \lambda)d_0^{1-1/q} \|v\|_q + \tau. \]

Thus we have for \( d_0 = c_0 \sqrt{f/\log m} \) sufficiently small,
\[ d_0^{-1/q}(c(q)\kappa_{\text{RE}}(2d_0, k_0) - \mu_2(2 + \lambda)d_0) \|v\|_q \]
\[ \leq (\kappa_q(d_0, k_0) - \mu_2(2 + \lambda)d_0^{1-1/q}) \|v\|_q \]
\[ \leq \mu_1 \|\beta^*\|_2 + \tau \leq 4D_2Kr_mf(\|\beta^*\|_2 + M_\epsilon) \]
where
\[ \mu_2(2 + \lambda)d_0 = 2D_2Kr_mf\left(\frac{1}{\lambda} + 1\right)(2 + \lambda)c_0 \sqrt{f/\log m} = 2c_0C_0D_2K^2(2 + \lambda)(\frac{1}{\lambda} + 1) \]
and thus (21) holds. The prediction error bound follows exactly the same line of arguments as in [2] which we omit here. \( \square \)

G.2 Proof of Lemmas 4.1 to 4.3

We next provide proofs for Lemmas 4.1 to 4.3 in this section.

Remark G.2. The set \( \Upsilon \) in our setting is equivalent to the following:
\[ \Upsilon = \left\{ (\beta, t) : \beta \in \mathbb{R}^m, \left\| \frac{1}{T}X^T(y - X\beta) + \frac{1}{T}G(B)\beta \right\|_\infty \leq \mu t + \tau, \|\beta\|_2 \leq t \right\}. \]
**Proof of Lemma 4.1.** Suppose event $B_0$ holds. Then by the proof of Lemma 3.2,

$$
\|\hat{\beta} - \bar{\beta}\|_{\infty} \leq 2C_0D_2K^2\|\beta^*\|_2\sqrt{\frac{\log m}{f}} + C_0D_0KM\sqrt{\frac{\log m}{f}}.
$$

The lemma follows immediately for the chosen $\mu, \tau$ as in (30) given that $(\beta^*, \|\beta^*\|_2) \in \Upsilon$. □

**Proof of Lemma 4.2.** On event $B_0$, $(\beta, t) = (\beta^*, \|\beta^*\|_2)$ belongs to the feasible set of the minimization problem (5), thus

$$
\left\|\hat{\beta}\right\|_1 + \lambda \left\|\hat{\beta}\right\|_2^2 \leq \left\|\tilde{\beta}\right\|_1 + \lambda \hat{\mu} \leq \|\beta^*\|_1 + \lambda \|\beta^*\|_2
$$

The lemma holds by the triangle inequality. See [2] for details. □

**Proof of Lemma 4.3.** First we rewrite an upper bound for $D = \text{tr}(B)$ and $\hat{D} = \text{tr}(B)$

$$
\|X_0^T X_0^V\|_{\infty} = \left\|(X - W)^T X_0 (\hat{\beta} - \beta^*)\right\|_{\infty} \leq \left\|X^T X_0 (\hat{\beta} - \beta^*)\right\|_{\infty} + \|W^T X_0^V\|_{\infty}
$$

$$
\leq \left\|X^T (X \hat{\beta} - y) - \hat{D}\hat{\beta}\right\|_{\infty} + \|X^T\|_\infty + \left\|(X^T W - D)\hat{\beta}\right\|_{\infty}
$$

$$
+ \left\|\hat{D} - D\right\|_{\infty} + \|W^T X_0^V\|_{\infty} =: f(I + II + III + IV + V)
$$

where

$$
\left\|X^T X_0 (\hat{\beta} - \beta^*)\right\|_{\infty} \leq \left\|X^T (X \hat{\beta} - y + \epsilon)\right\|_{\infty} = \left\|X^T ((X - W)\hat{\beta} - y)\right\|_{\infty} + \|X^T\|_\infty
$$

$$
\leq \left\|X^T (X \hat{\beta} - y) - \hat{D}\hat{\beta}\right\|_{\infty} + \|X^T\|_\infty + \left\|(X^T W - D)\hat{\beta}\right\|_{\infty} + \left\|\hat{D} - D\right\|_{\infty}
$$

where on event $B_0$, we have by Lemma 4.2 and the fact that $\hat{\beta} \in \Upsilon$

$$
I = \left\|\hat{\beta} - \bar{\beta}\right\|_{\infty} = \left\|\frac{1}{\lambda} X^T (y - X \hat{\beta}) + \frac{1}{\lambda} \hat{D}\hat{\beta}\right\|_{\infty}
$$

$$
\leq \mu \hat{\mu} + \tau \leq \mu \left(\frac{1}{\lambda} \|v\|_1 + \|\beta^*\|_2\right) + \tau,
$$

on event $B_4$, $II := \|W^T X_0^V\|_{\infty} \leq r_{m, f} M \epsilon (\alpha_{\text{max}}^{1/2} + \sqrt{T_B})$

and on event $B_6$, $IV := \left\|\hat{D} - D\right\|_{\infty} \leq 2D_1 K r_{m, m} \|\beta^*\|_2$

On event $B_5 \cap B_{10}$, we have

$$
III = \frac{1}{\lambda} \left\|(X^T W - D)\hat{\beta}\right\|_{\infty} \leq \frac{1}{\lambda} \left\|(X^T W - D)\beta^*\right\|_{\infty} + \frac{1}{\lambda} \left\|(X^T W - D)v\right\|_{\infty}
$$

$$
\leq \frac{1}{\lambda} \left\|X^T W \beta^*\right\|_{\infty} + \left\|(W^T W - D)\beta^*\right\|_{\infty}
$$

$$
+ \frac{1}{\lambda} \left\|(Z^T BZ - \text{tr}(B)I_m)\right\|_{\text{max}} + \frac{1}{\lambda} \left\|X_0^T W\right\|_{\text{max}} \|v\|_1
$$

$$
\leq r_{m, f} K (\frac{B\|f\|_{\text{max}}}{\sqrt{T}} + \sqrt{T_B} \alpha_{\text{max}}^{1/2}) (\|v\|_1 + \|\beta^*\|_2),
$$

and $V = \frac{1}{\lambda} \|W^T X_0^V\|_{\text{max}} \|v\|_1 \leq r_{m, f} K \sqrt{T_B} \alpha_{\text{max}}^{1/2} \|v\|_1$. 

28
Thus we have on $B_0 \cap B_{10}$, for $D_0 \leq D_2$,
\[
\left\| \frac{1}{2} X_0^T X_0 v \right\|_\infty \leq \mu \left( \frac{1}{\lambda} \|v\|_1 + \|\beta^*\|_2 \right) + \tau + r_{m,f} M T_{\alpha_{\max}}^{1/2} + III + IV + V
\leq 2D_2 K r_{m,f} \left( \frac{1}{\lambda} + 1 \right) \|v\|_1 + 2 \|\beta^*\|_2 + 2D_0 M r_{m,f}.
\]
The lemma thus holds. □

## H Some geometric analysis results

In order to prove Corollary 5.1, we need to first state some geometric analysis results in this section.

Let us define the following set of vectors in $\mathbb{R}^m$:
\[
\text{Cone}(s_0) := \{v : \|v\|_1 \leq \sqrt{s_0} \|v\|_2\}
\]

For each vector $x \in \mathbb{R}^m$, let $T_0$ denote the locations of the $s_0$ largest coefficients of $x$ in absolute values. Any vector $x \in S^{m-1}$ satisfies:
\[
\|x_{T_0}\|_\infty \leq \|x_{T_0}\|_1 / s_0 \leq \frac{\|x_{T_0}\|_2}{\sqrt{s_0}}
\]

We need to state the following result from [30]. Let $S^{m-1}$ be the unit sphere in $\mathbb{R}^m$, for $1 \leq s \leq m$,
\[
U_s := \{x \in \mathbb{R}^m : |\text{supp}(x)| \leq s\}
\]

The sets $U_s$ is an union of the $s$-sparse vectors. The following three lemmas are well-known and mostly standard; See [30] and [27].

**Lemma H.1.** For every $1 \leq s_0 \leq m$ and every $I \subset \{1, \ldots, m\}$ with $|I| \leq s_0$,
\[
\sqrt{|I|} B_1^m \cap S^{m-1} \subset 2 \text{conv}(U_{s_0} \cap S^{m-1}) =: 2 \text{conv} \left( \bigcup_{|J| \leq s_0} E_J \cap S^{m-1} \right)
\]

and moreover, for $\rho \in (0, 1]$.
\[
\sqrt{|I|} B_1^m \cap \rho B_2^m \subset (1 + \rho) \text{conv}(U_{s_0} \cap B_2^m) =: (1 + \rho) \text{conv} \left( \bigcup_{|J| \leq s_0} E_J \cap S^{m-1} \right)
\]

**Proof.** Fix $x \in \mathbb{R}^m$. Let $x_{T_0}$ denote the subvector of $x$ confined to the locations of its $s_0$ largest coefficients in absolute values; moreover, we use it to represent its 0-extended version $x' \in \mathbb{R}^p$ such that $x'_{T^c} = 0$ and $x'_{T_0} = x_{T_0}$. Throughout this proof, $T_0$ is understood to be the locations of the $s_0$ largest coefficients in absolute values in $x$.  

29
Moreover, let \((x_i^\ast)_{i=1}^m\) be non-increasing rearrangement of \((|x_i|)_{i=1}^m\). Denote by
\[
L = \sqrt{s_0}B_1^m \cap \rho B_2^m
\]
\[
R = 2 \text{conv} \left( \bigcup_{|J| \leq s} E_J \cap B_2^m \right) = 2 \text{conv} \left( E \cap B_2^m \right)
\]
Any vector \(x \in \mathbb{R}^m\) satisfies:
\[
\|x_{T_0}\|_\infty \leq \|x_{T_0}\|_1 / s_0 \leq \frac{\|x_{T_0}\|_2}{\sqrt{s_0}} \tag{55}
\]
It follows that for any \(\rho > 0\), \(s_0 \geq 1\) and for all \(z \in L\), we have the \(i^{th}\) largest coordinate in absolute value in \(z\) is at most \(\sqrt{\frac{s_0}{i}}\).
\[
\sup_{z \in L} \langle x, z \rangle \leq \max_{\|z\|_2 \leq \rho} \langle x_{T_0}, z \rangle + \max_{\|z\|_1 \leq \sqrt{s_0}} \langle x_{T_0}, z \rangle \\
\leq \rho \|x_{T_0}\|_2 + \|x_{T_0}\|_\infty \sqrt{s_0} \\
\leq \|x_{T_0}\|_2 (\rho + 1)
\]
where clearly \(\max_{\|z\|_2 \leq \rho} \langle x_{T_0}, z \rangle = \rho \sum_{i=1}^{s_0} (x_i^2)^{1/2}\). And denote by \(S^J := S^{m-1} \cap E_J\),
\[
\sup_{z \in \mathbb{R}} \langle x, z \rangle = (1 + \rho) \max_{J,|J| \leq s_0} \max_{z \in S^J} \langle x, z \rangle \\
= (1 + \rho) \|x_{T_0}\|_2
\]
given that for a convex function \(\langle x, z \rangle\), the maximum happens at an extreme point, and in this case, it happens for \(z\) such that \(z\) is supported on \(T_0\), such that \(z_{T_0} = \frac{x_{T_0}}{\|x_{T_0}\|_2}\), and \(z_{T_0}^\infty = 0\). \(\square\)

**Lemma H.2.** Let \(1/5 > \delta > 0\). Let \(E = \bigcup_{|J| \leq s_0} E_J\) for \(0 < s_0 < m/2\) and \(k_0 > 0\). Let \(\Delta\) be a \(m \times m\) matrix such that
\[
|u^T \Delta v| \leq \delta \quad \forall u, v \in E \cap S^{m-1} \tag{56}
\]
Then for all \(v \in \left(\sqrt{s_0}B_1^m \cap B_2^m\right)\), we have
\[
|v^T \Delta v| \leq 4\delta. \tag{57}
\]
**Proof.** First notice that
\[
\max_{v \in \left(\sqrt{s_0}B_1^m \cap B_2^m\right)} |v^T \Delta v| \leq \max_{w, u \in \left(\sqrt{s_0}B_1^m \cap B_2^m\right)} |w^T \Delta u| \tag{58}
\]
Now that we have decoupled \(u\) and \(w\) on the RHS of (58), we first fix \(u\). Then for any fixed \(u \in S^{m-1}\) and matrix \(\Delta \in \mathbb{R}^{m \times m}\), \(f(w) = |w^T \Delta u|\) is a convex function of \(w\), and hence for \(w \in \left(\sqrt{s_0}B_1^m \cap B_2^m\right) \subset 2 \text{conv} \left( \bigcup_{|J| \leq s_0} E_J \cap S^{m-1} \right)\),
\[
\max_{w \in \left(\sqrt{s_0}B_1^m \cap B_2^m\right)} |w^T \Delta u| \leq 2 \max_{w \in \text{conv}(E \cap S^{m-1})} |w^T \Delta u| \\
= 2 \max_{w \in E \cap S^{m-1}} |w^T \Delta u|
\]
30
where the maximum occurs at an extreme point of the set \( \text{conv} (E \cap S^{m-1}) \), because of the convexity of the function \( f(w) \).

Clearly the RHS of (58) is bounded by

\[
\max_{u, w \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} |w^T \Delta u| = \max_{u \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} \max_{w \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} |w^T \Delta u| \\
\leq 2 \max_{u \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} \max_{w \in (E \cap S^{m-1})} |w^T \Delta u| \\
= 2 \max_{u \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} g(u)
\]

where the function \( g(u) \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right) \) is defined as

\[
g(u) = \max_{w \in (E \cap S^{m-1})} |w^T \Delta u|
\]

which is convex since it is the maximum of a function \( f_w(u) := |w^T \Delta u| \) which is convex in \( u \) for each \( w \in (E \cap S^{m-1}) \). Thus we have for \( u \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right) \subset 2 \text{conv} \left( \bigcup_{|J| \leq s_0} E_J \cap S^{m-1} \right) =: 2 \text{conv} \left( E \cap S^{m-1} \right) \)

\[
\max_{u \in \left( \sqrt{s_0} B_1^m \cap B_2^m \right)} g(u) \leq 2 \max_{u \in \text{conv} \left( E \cap S^{m-1} \right)} g(u) \\
= 2 \max_{u \in E \cap S^{m-1}} g(u) \\
= 2 \max_{u \in E \cap S^{m-1}} \max_{w \in E \cap S^{m-1}} |w^T \Delta u| \leq 4\delta
\]

where (59) holds given that the maximum occurs at an extreme point of the set \( \text{conv} (E \cap B_2^m) \), because of the convexity of the function \( g(u) \). \( \square \)

**Corollary H.3.** Suppose all conditions in Lemma H.2 hold. Then \( \forall \upsilon \in \text{Cone}(s_0) \),

\[
|\upsilon^T \Delta \upsilon| \leq 4\delta \|\upsilon\|_2^2.
\]

**Proof.** It is sufficient to show that \( \forall \upsilon \in \text{Cone}(s_0) \cap S^{m-1}, \)

\[
|\upsilon^T \Delta \upsilon| \leq 4\delta.
\]

Denote by \( \text{Cone} := \text{Cone}(s_0) \). Clearly this set of vectors satisfy:

\[
\text{Cone} \cap S^{m-1} \subset \left( \sqrt{s_0} B_1^m \cap B_2^m \right)
\]

Thus (61) follows from (57). \( \square \)

**Remark H.4.** Suppose we relax the definition of \( \text{Cone}(s_0) \) to be:

\[
\text{Cone}(s_0) := \{ \upsilon : \|\upsilon\|_1 \leq 2\sqrt{s_0} \|\upsilon\|_2 \}\]

Clearly, \( \text{Cone}(s_0, 1) \subset \text{Cone}(s_0) \). Given that \( \forall \upsilon \in \text{Cone}(s_0, 1) \), we have

\[
\|\upsilon\|_1 \leq 2\|\upsilon_{T_0}\|_1 \leq 2\sqrt{s_0} \|\upsilon_{T_0}\|_2 \leq 2\sqrt{s_0} \|\upsilon\|_2
\]

31
Lemma H.5. Suppose all conditions in Lemma H.2 hold. Then for all $v \in \mathbb{R}^m$,

$$|v^T \Delta v| \leq 4\delta(\|v\|^2_2 + \frac{1}{s_0} \|v\|_1^2) \quad (62)$$

Proof. The lemma follows given that $\forall v \in \mathbb{R}^m$, one of the following must hold:

if $v \in \text{Cone}(s_0)$

$$|v^T \Delta v| \leq 4\delta \|v\|^2_2 \quad (63)$$

otherwise

$$|v^T \Delta v| \leq \frac{4\delta}{s_0} \|v\|_1^2 \quad (64)$$

leading to the same conclusion in (62). We have shown (63) in Lemma H.2. Let $\text{Cone}(s_0)^C$ be the complement set of $\text{Cone}(s_0)$ in $\mathbb{R}^m$. That is, we focus now on the set of vectors such that

$${\text{Cone}(s_0)^C} := \{v : \|v\|_1 \geq \sqrt{s_0} \|v\|_2\}$$

and show that for $u = \sqrt{s_0} \frac{v}{\|v\|_1}$,

$$\frac{|v^T \Delta v|}{\|v\|_1^2} := \frac{1}{s_0} |u^T \Delta u| \leq \frac{1}{s_0} \delta$$

where the last inequality holds by Lemma H.2 given that

$$u \in (\sqrt{s_0} B_1^m \cap B_2^m) \subset 2 \text{ conv } \left( \bigcup_{|J| \leq s_0} E_J \cap B_2^m \right)$$

and thus

$$\frac{|v^T \Delta v|}{\|v\|_1^2} \leq \frac{1}{s_0} \sup_{u \in \sqrt{s_0} B_1^m \cap B_2^m} |u^T \Delta u| \leq \frac{1}{s_0} 4\delta$$

□

I Proof of Corollary 5.1

First we show that for all $v \in \mathbb{R}^m$, (65) holds. It is sufficient to check that the condition (56) in Lemma H.2 holds. Then, (65) follows from Lemma H.5: for $v \in \mathbb{R}^m$,

$$|v^T \Delta v| \leq 4\delta(\|v\|^2_2 + \frac{1}{\zeta} \|v\|_1^2) \leq \frac{1}{2} \lambda_{\text{min}}(A)(\|v\|^2_2 + \frac{1}{\zeta} \|v\|_1^2). \quad (65)$$

The Lower and Upper RE conditions thus immediately follow. The Corollary is thus proved. □
J Proof of Theorem 5.2

To bound the two middle terms, we need the following Lemmas. Proofs for Lemmas J.1 and J.2 appear in Section P. Throughout this section, the choice of $C = C_0/\sqrt{c'}$ satisfies the conditions on $C$ in Lemmas J.1 and J.2, where recall $\min\{C_0, C_0^2\} \geq 4/c$ for $c$ as defined in Theorem D.1. For a set $J \subset \{1, \ldots, m\}$, denote $F_J = A^{1/2} E_J$ where recall $E_J = \text{span}\{e_j : j \in J\}$.

**Lemma J.1.** Suppose all conditions in Theorem 5.2 hold. Let

$$E = \bigcup_{|J|=k} E_J \cap S^{m-1}.$$  

Suppose that for some $c' > 0$ and $\varepsilon \leq \frac{1}{C}$, where $C = C_0/\sqrt{c'}$,

$$r(B) := \frac{\text{tr}(B)}{\|B\|_2} \geq c' k K^4 \frac{\log(3em/k\varepsilon)}{\varepsilon^2}. \quad (66)$$

Then for all vectors $u, v \in E \cap S^{m-1}$, on event $B_1$, where $\mathbb{P}(B_1) \geq 1 - 2 \exp\left(-c_2 \varepsilon^2 \frac{\text{tr}(B)}{K^4 \|B\|_2}\right)$ for $c_2 \geq 2$,

$$|u^T Z^T B Z v - E u^T Z^T B Z v| \leq 4C \varepsilon \text{tr}(B).$$

**Lemma J.2.** Suppose that $\varepsilon \leq 1/C$, where $C$ is as defined in Lemma J.1. Suppose that (66) holds. Let

$$E = \bigcup_{|J|=k} E_J \quad \text{and} \quad F = \bigcup_{|J|=k} F_J. \quad (67)$$

Then on event $B_2$, where $\mathbb{P}(B_2) \geq 1 - 2 \exp\left(-c_2 \varepsilon^2 \frac{\text{tr}(B)}{K^4 \|B\|_2}\right)$ for $c_2 \geq 2$, we have for all vectors $u \in E \cap S^{m-1}$ and $w \in F \cap S^{m-1}$,

$$|w^T Z_1^T B^{1/2} Z_2 w| \leq \frac{C \varepsilon \text{tr}(B)}{(1 - \varepsilon^2) \|B\|_2^{1/2}} \leq 4C \varepsilon \text{tr}(B)/\|B\|_2^{1/2}$$

where $Z_1, Z_2$ are independent copies of $Z$, as defined in Theorem 5.2.

In fact, the same conclusion holds for all $y, w \in F \cap S^{m-1}$; and in particular, for $B = I$, we have the following.

**Corollary J.3.** Suppose all conditions in Lemma J.1 hold. Suppose that $F = A^{1/2} E$ for $E$ as defined in Lemma J.1. Let

$$f \geq c' k K^4 \frac{\log(3em/k\varepsilon)}{\varepsilon^2}. \quad (68)$$

Then on event $B_3$, where $\mathbb{P}(B_3) \geq 1 - 2 \exp\left(-c_2 \varepsilon^2 f \frac{1}{K^4}\right)$, we have for all vectors $w, y \in F \cap S^{m-1}$ and $\varepsilon \leq 1/C$ for $C$ is as defined in Lemma J.1,

$$|y^T (\frac{1}{f} Z^T Z - I) w| \leq 4C \varepsilon. \quad (69)$$
We prove Lemmas J.1 and J.2 and Corollary J.3 in Section P. We are now ready to prove Theorem 5.2.

\textbf{Proof of Theorem 5.2.} Recall the following for $X_0 = Z_1 A^{1/2}$,

\[
\Delta := \tilde{\Gamma}_A - A := \frac{1}{f} X^T X - \hat{\text{tr}}(B) I_m - A \\
= \left( \frac{1}{f} X_0^T X_0 - A \right) + \frac{1}{f} (W^T X_0 + X_0^T W) + \frac{1}{f} (W^T W - \hat{\text{tr}}(B) I_m).
\]

Notice that

\[
\left| u^T (\tilde{\Gamma}_A - A) v \right| = \left| u^T \left( \frac{1}{f} X^T X - \hat{\text{tr}}(B) I_m - A \right) v \right| \\
\leq \left| u^T \left( \frac{1}{f} X_0^T X_0 - A \right) v \right| + \left| u^T \frac{1}{f} (W^T X_0 + X_0^T W) v \right| + \left| u^T \frac{1}{f} (W^T W - \hat{\text{tr}}(B) I_m) v \right| \\
\leq \left| u^T \left( \frac{1}{f} Z_1^T Z_1 A^{1/2} v + u^T A v \right) \right| + \left| u^T \frac{1}{f} (W^T X_0 + X_0^T W) v \right| \\
+ \left| u^T \left( \frac{1}{f} Z_2^T B Z_2 - \tau_B I_m \right) v \right| + \frac{1}{f} \left| \hat{\text{tr}}(B) - \text{tr}(B) \right| \left| u^T v \right| =: I + II + III + IV.
\]

For $u \in E \cap S^{m-1}$, define $h(u) := \frac{A^{1/2} u}{\|A^{1/2} u\|_2}$. The conditions in (66) and (68) hold for $k$. We first bound the middle term as follows. Fix $u, v \in E \cap S^{m-1}$ then on event $B_2$, for $Y = Z_1^T B^{1/2} Z_2$,

\[
\left| u^T (W^T X_0 + X_0^T W) v \right| = \left| u^T Z_2^T B^{1/2} Z_1 A^{1/2} v + u^T A^{1/2} Z_2^T B^{1/2} Z_2 v \right| \\
\leq \left| u^T Y^T h(v) \right| \left\| A^{1/2} v \right\|_2 + \left| h(u)^T Y v \right| \left\| A^{1/2} u \right\|_2 \\
\leq 2 \max_{w \in E \cap S^{m-1}, v \in E \cap S^{m-1}} \left| u^T Y v \right| \| A^{1/2} u \|_2 \rho_{\text{max}}(k, A) \\
\leq 8 \rho_{\text{max}}(k, A) \left( \frac{\rho_{\text{max}}(k, A)}{\|B\|_2} \right)^{1/2}.
\]

We now use Lemma J.1 to bound both $I$ and $III$. We have for $C$ as defined in Lemma J.1, on event $B_1 \cap B_3$,

\[
\left| u^T (Z_2^T B Z_2 - \text{tr}(B) I_m) v \right| \leq 4 C \epsilon \text{tr}(B).
\]

Moreover, by Corollary J.3, we have on event $B_3$, for all $u, v \in E \cap S^{m-1}$,

\[
\left| u^T \left( \frac{1}{f} X_0^T X_0 - A \right) v \right| = \left| u^T A^{1/2} Z_1^T Z_1 A^{1/2} v - u^T A v \right| \\
= \left| h(u)^T \left( \frac{1}{f} Z_1^T Z - I \right) h(v) \right| \left\| A^{1/2} u \right\|_2 \left\| A^{1/2} v \right\|_2 \\
\leq \frac{1}{f} \max_{w, y \in E \cap S^{m-1}} \left| w^T (Z_1^T Z - I) y \right| \rho_{\text{max}}(k, A) \\
\leq 4 C \epsilon \rho_{\text{max}}(k, A).
\]

Thus we have on event $B_1 \cap B_2 \cap B_3$ and for $\tau_B := \text{tr}(B)/f$

\[
I + II + III \leq 4 C \epsilon \left( \rho_{\text{max}}(k, A) + 2 \tau_B \left( \frac{\rho_{\text{max}}(k, A)}{\|B\|_2} \right)^{1/2} + \tau_B \right) \\
\leq 8 C \epsilon \left( \tau_B + \rho_{\text{max}}(k, A) \right).
\]

On event $B_6$, we have for $D_1$ as defined in Lemma E.2,

\[
IV \leq 2 C_0 D_1 K^2 \sqrt{\frac{\log m}{m}}.
\]

The theorem thus holds by the union bound. \qed
\section*{K Proof for Theorem 6.1}

We first state the following bounds in (70) before we prove Theorem 6.1. On event $\mathcal{A}_2$, where $\mathbb{P}(\mathcal{A}_2) \geq 1 - 2\exp\left(-c_3\varepsilon^2 \frac{\text{tr}(A)}{K^4\|A\|_2}\right)$

$$\forall u, w \in S^{f-1} \left| u^TZ_1A^{1/2}Z_2^Tw \right| \leq \frac{4C\varepsilon\text{tr}(A)}{\|A\|_2^{1/2}}. \quad (70)$$

To see this, first note that by Lemma D.2, we have for $t = C\varepsilon\text{tr}(A)/\|A\|_2^{1/2}$ and $\varepsilon \leq 1/2$,

$$\mathbb{P}\left( \left| u^TZ_1A^{1/2}Z_2^Tw \right| > t \right) \leq 2\exp\left(-c\min\left(\frac{C^2\varepsilon^2\text{tr}(A)}{K^4\|A\|_2}, \frac{C\varepsilon\text{tr}(A)}{K^2\|A\|_2}\right)\right)$$

$$\leq 2\exp\left(-c\min\left(C^2, 2C\right)\frac{\varepsilon^2\text{tr}(A)}{K^4\|A\|_2}\right)$$

where recall

$$C' = cC' \min\left(2C, C^2\right) > 4.$$

Before we proceed, we state the following well-known result on volumetric estimate; see e.g. [31].

\textbf{Lemma K.1.} Given $m \geq 1$ and $\varepsilon > 0$. There exists an $\varepsilon$-net $\Pi \subset B_2^m$ of $B_2^m$ with respect to the Euclidean metric such that $B_2^m \subset (1 - \varepsilon)^{-1} \text{conv} \Pi$ and $|\Pi| \leq (1 + 2/\varepsilon)^m$. Similarly, there exists an $\varepsilon$-net of the sphere $S^{m-1}$, $\Pi' \subset S^{m-1}$ such that $|\Pi'| \leq (1 + 2/\varepsilon)^m$.

Choose an $\varepsilon$-net $\Pi \subset S^{f-1}$ such that $|\Pi| \leq (1 + 2/\varepsilon)^f = \exp(f\log(3/\varepsilon))$. The existence of such $\Pi$ is guaranteed by Lemma K.1. By the union bound and Lemma D.2, we have for some $C \geq 2$ and $c' \geq 1$ large enough such that

$$\mathbb{P}\left( \exists u, w \in \Pi \text{s.t.} \left| u^TZ_1A^{1/2}Z_2^Tw \right| \geq C\varepsilon\frac{\text{tr}(A)}{\|A\|_2^{1/2}} \right) \leq 2\exp\left(-c_3\frac{\varepsilon^2\text{tr}(A)}{K^4\|A\|_2}\right).$$

Hence, (70) follows from a standard approximation argument.

\textbf{Lemma K.2.} Let $\varepsilon > 0$. Let $Z$ as defined in Definition 2.1. Assume that

$$\frac{\text{tr}(A)}{\|A\|} \geq c' \frac{\log(3/\varepsilon)}{\varepsilon^2}. $$

Then

$$\mathbb{P}\left( \exists x \in S^{f-1} \left\| A^{1/2}Z^T x \right\|_2 - (\text{tr}(A))^{1/2} > \varepsilon(\text{tr}(A))^{1/2} \right) \leq \exp\left(-c\varepsilon^2 \frac{\text{tr}(A)}{K^4\|A\|}\right).$$

\textbf{Proof of Theorem 6.1.} First we write

$$XX^T - \text{tr}(A)I_f = (Z_1A^{1/2} + B^{1/2}Z_2)(Z_1A^{1/2} + B^{1/2}Z_2)^T - \text{tr}(A)I_f$$

$$= (Z_1A^{1/2} + B^{1/2}Z_2)(Z_2^TB^{1/2} + A^{1/2}Z_1^T) - \text{tr}(A)I_f$$

$$= Z_1A^{1/2}Z_2^TB^{1/2} + B^{1/2}Z_2Z_2^TB^{1/2} + B^{1/2}Z_2A^{1/2}Z_1^T + Z_1AZ_1^T - \text{tr}(A)I_f.$$
Hence,
\[
\begin{align*}
\left| \frac{u^T (X X^T) u}{m} - \frac{u^T \text{tr}(A) I u}{m} - u^T B u \right| & \leq \left| \frac{1}{m} u^T Z_1 A Z_1^T u - \frac{\text{tr}(A)}{m} u^T u \right| + \frac{1}{m} u^T B^{1/2} Z_2 Z_2^T B^{1/2} u - u^T B u \right| + \frac{2}{m} \left| u^T Z_1 A^{1/2} Z_2^T B^{1/2} u \right|.
\end{align*}
\]

where by (70), we have on event $A_2$, for $\tau_A := \frac{\text{tr}(A)}{m}$ and $w := \frac{B^{1/2} u}{\|B^{1/2} u\|_2}$,
\[
\frac{2}{m} \left| u^T Z_1 A^{1/2} Z_2^T B^{1/2} u \right| = \frac{2}{m} \left| u^T Z_1 A^{1/2} Z_2^T w \right| \left\| B^{1/2} u \right\|_2 \leq \frac{8C \varepsilon \text{tr}(A)}{\|A\|^{1/2}} \left\| B^{1/2} u \right\|_2 =: 8C \varepsilon \tau_A \left\| B^{1/2} u \right\|_2 / \|A\|^{1/2}.
\]

Moreover, by the union bound and Lemma K.2, we have on event $A_1$, where $\mathbb{P}(A_1) \geq 1 - \exp(\varepsilon^2 \frac{m}{K} \varepsilon) - \exp(\varepsilon^2 \frac{\text{tr}(A)}{K^2 \|A\|^{1/2}}),$
\[
(1 - \varepsilon) \left\| B^{1/2} u \right\|_2 \leq \frac{1}{\sqrt{m}} \left\| Z_2 B^{1/2} u \right\|_2 \leq (1 + \varepsilon) \left\| B^{1/2} u \right\|_2,
\]
\[
(1 - \varepsilon) \frac{\text{tr}(A)^{1/2}}{\sqrt{m}} \leq \frac{1}{\sqrt{m}} \left\| A^{1/2} Z_1^T u \right\|_2 \leq (1 + \varepsilon) \frac{\text{tr}(A)^{1/2}}{\sqrt{m}}.
\]

Hence on event $A_1$, we have
\[
\frac{1}{m} \left\| A^{1/2} Z_1^T u \right\|_2^2 - \text{tr}(A) \leq \max((1 + \varepsilon)^2 - 1, 1 - (1 - \varepsilon)^2) \frac{\text{tr}(A)}{m},
\]
\[
\frac{1}{m} \left\| Z_2^T B^{1/2} u \right\|_2^2 - u^T B u \leq \max((1 + \varepsilon)^2 - 1, 1 - (1 - \varepsilon)^2) \left\| B^{1/2} u \right\|_2^2.
\]

Thus we have for all $u \in S^{f-1}$, on event $A_1 \cap A_2$, for $C_2 := 4C + 3$
\[
\left| \frac{1}{m} u^T (X X^T) u - u^T \text{tr}(A) I u - u^T B u \right| \leq \left\| Z_2^T B^{1/2} u \right\|_2^2 / m - u^T B u + \frac{1}{m} \left\| A^{1/2} Z_1^T u \right\|_2^2 - \text{tr}(A) + 8C \varepsilon \tau_A \left\| B^{1/2} u \right\|_2 / \|A\|^{1/2} \leq 3\varepsilon \left\| B^{1/2} u \right\|_2^2 + 3\varepsilon \tau_A + 8C \varepsilon \tau_A \left\| B^{1/2} u \right\|_2 / \|A\|^{1/2} \leq C_2 \varepsilon \left\| B^{1/2} u \right\|_2^2 + C_2 \varepsilon \tau_A
\]

where $2\tau_A \left\| B^{1/2} u \right\|_2 \leq \tau_A + \left\| B^{1/2} u \right\|_2^2$. The theorem thus holds. \(\square\)

It remains to prove Lemma K.2.

**Proof of Lemma K.2.** Let $x \in S^{f-1}$. Then $Y = Z^T x \in \mathbb{R}^m$ is a random vector with independent coordinates satisfying $\mathbb{E} Y_j = 0$ and $\|Y_j\|_{\psi_2} \leq C K$ for all $j \in 1\ldots m$. The last estimate follows from Hoeffding inequality. By Theorem 2.1 [34],
\[
\mathbb{P} \left( \left\| A^{1/2} Y \right\|_2 - (\text{tr}(A))^{1/2} > \varepsilon (\text{tr}(A))^{1/2} \right) \leq \exp \left( -c \varepsilon^2 \frac{\text{tr}(A)}{K^2 \|A\|} \right).
\]

36
Choose an $\varepsilon$-net $\Pi \subset S^{f-1}$ such that $|\Pi| \leq (3/\varepsilon)^f$. By the union bound and the assumption of the Lemma,

$$\mathbb{P} \left( \exists x \in \Pi \mid \left\| A^{1/2}Z^T x \right\|_2 - (\text{tr}(A))^{1/2} > \varepsilon(\text{tr}(A))^{1/2} \right) \leq |\Pi| \cdot \exp \left( -c\varepsilon^2 \frac{\text{tr}(A)}{K^4 \|A\|} \right) \leq \exp \left( -c'\varepsilon^2 \frac{\text{tr}(A)}{K^4 \|A\|} \right).$$

A standard approximation argument shows that if $\left\| A^{1/2}Z^T x \right\|_2 - (\text{tr}(A))^{1/2} \leq \varepsilon(\text{tr}(A))^{1/2}$ for all $x \in \Pi$, then $\left\| A^{1/2}Z^T x \right\|_2 - (\text{tr}(A))^{1/2} \leq 3\varepsilon(\text{tr}(A))^{1/2}$ for all $x \in S^{f-1}$. This finishes the proof of the Lemma. □

L Proofs of Corollaries 6.2 and C.2

We prove the concentration of measure bounds on error-corrected gram matrices in Corollaries 6.2 and C.2 in this section.

L.1 Proof of Corollary 6.2

**Lower bound:** For all $u \in S^{f-1}$ and

$$\frac{1}{m} u^T (XX^T) u - u^T \frac{\text{tr}(A)I_f}{m} u \geq u^T Bu(1 - 3\varepsilon) - 3\varepsilon \tau_A - 8C \left\| B^{1/2} u \right\|_2 \varepsilon \tau_A / \|A\|^{1/2} \geq u^T Bu(1 - 3\varepsilon - 4C\varepsilon) - 3\varepsilon \tau_A - 4C\varepsilon \tau_A \geq u^T Bu(1 - C_2\varepsilon) - C_2\varepsilon \tau_A \geq u^T Bu(1 - 2\delta)$$

where we bound the term using the fact that $1 \leq \tau_A \leq \lambda_{\max}(B)$ and

$$C_2\tau_A\varepsilon \leq \delta\lambda_{\min}(B) \quad \text{and} \quad \varepsilon \leq \delta\lambda_{\min}(B)/(C_2\tau_A)$$

$$C_2\varepsilon \leq \delta \quad \text{and} \quad C_3\varepsilon \leq \delta \min \left( \frac{\lambda_{\min}(B)}{\tau_A}, 1 \right).$$

By a similar argument, we can prove the upper bound on the isometry property as stated in the corollary. □

L.2 Proof of Corollary C.2

Recall the following

$$\tilde{A} := X^T X - \hat{\text{tr}}(B)I_m = (Z_1A^{1/2} + B^{1/2}Z_2)^T (Z_1A^{1/2} + B^{1/2}Z_2) - \hat{\text{tr}}(B)I_m$$

$$= (Z_2^T B^{1/2} + A^{1/2}Z_1^T) (Z_1A^{1/2} + B^{1/2}Z_2) - \hat{\text{tr}}(B)I_m$$

$$= (Z_2^T B^{1/2}Z_1A^{1/2} + A^{1/2}Z_1^T B^{1/2}Z_2) + A^{1/2}Z_1^T Z_1A^{1/2} + (Z_2^T BZ_2 - \hat{\text{tr}}(B)I_m).$$

37
Hence, for all vectors \( u \in S^{m-1} \cap E \)
\[
\frac{u^T(X^TX)u}{f} - \frac{u^T \text{tr}(B) I u}{f} - u^T A u \leq \frac{1}{f} \left| u^T Z_2 B Z_2^T u - \text{tr}(B) u^T u \right| + \frac{1}{f} \left| u^T A^{1/2} Z_1^T Z_1 A^{1/2} u - u^T A u \right| + \frac{2}{f} \left| u^T A^{1/2} Z_1^T B^{1/2} Z_2 u \right|.
\]

By Lemma J.1, we have on event \( B_1 \),
\[
\forall u \in E \cap S^{m-1} \quad \left| u^T Z^T B Z u - \text{tr}(B) \right| \leq 4C \varepsilon \text{tr}(B);
\]

By Lemma J.2, we have on event \( B_2 \),
\[
\forall u \in E \cap S^{m-1} \quad \left| u^T A^{1/2} Z_1^T B^{1/2} Z_2 u \right| \leq 4C \varepsilon \text{tr}(B) \left\| A^{1/2} u \right\|_2 / \| B \|_2^{1/2}.
\]

For all \( u \in S^{m-1} \cap E \),
\[
8C \varepsilon \tau_B \left\| A^{1/2} u \right\|_2 / \| B \|_2^{1/2} \leq 2(2C \varepsilon^{1/2} \frac{\tau_B}{\| B \|_2^{1/2}}) (2C \varepsilon^{1/2}) \left\| A^{1/2} u \right\|_2
\leq 4C^2 \varepsilon^2 \tau_B + 4 \varepsilon \left\| A^{1/2} u \right\|_2
\leq 4C^2 \varepsilon \tau_B + 4 \varepsilon \left\| A^{1/2} u \right\|_2.
\]

And finally, we have also shown that for all \( u \in E \) on event \( B_9 \),
\[
(1 - \varepsilon) \left\| A^{1/2} u \right\|_2 \leq \frac{1}{\sqrt{f}} \left\| Z_1 A^{1/2} u \right\|_2 \leq (1 + \varepsilon) \left\| A^{1/2} u \right\|_2.
\]

Thus we have for all \( u \in S^{m-1} \cap E \), on event \( B_1 \cap B_2 \cap B_9 \),
\[
\left| \frac{u^T(X^TX)u}{f} - \frac{u^T \text{tr}(B) I u}{f} - u^T A u \right| \leq \frac{1}{f} \left| u^T Z_2 B Z_2^T u - \text{tr}(B) u^T u \right| + \frac{1}{f} \left| Z_1 A^{1/2} u \right|^2_2 - \left\| A^{1/2} u \right\|_2^2 + \frac{2}{f} \left| u^T A^{1/2} Z_1^T B^{1/2} Z_2 u \right|
\leq 4C \varepsilon \tau_B + 6 \varepsilon \left\| A^{1/2} u \right\|_2^2 + 8C \varepsilon \tau_B \left\| A^{1/2} u \right\|_2 / \| B \|_2^{1/2}
\leq 4C \varepsilon \tau_B + 6 \varepsilon \left\| A^{1/2} u \right\|_2^2 + 4C^2 \varepsilon \tau_B + 4 \varepsilon \left\| A^{1/2} u \right\|_2^2
\leq 10 \varepsilon \left\| A^{1/2} u \right\|_2^2 + 4(C^2 + C) \varepsilon \tau_B.
\]

(71)

Upper bound: Thus we have by (71) for the maximum sparse eigenvalue of \( \tilde{A} \) at order \( k \):
\[
\rho_{\text{max}}(k, \tilde{A}) := \max_{u \in E \cap S^{m-1}} \left| u^T \tilde{A} u \right| \leq \max_{u \in E \cap S^{m-1}} \left| u^T \tilde{A} u \right| + \rho_{\text{max}}(k, A)
\leq \rho_{\text{max}}(k, A)(1 + 10 \varepsilon) + C_4 \varepsilon \tau_B
\]

where \( C_4 = 4(C + C^2) \). The upper bound on \( \rho_{\text{max}}(k, \tilde{A} - A) \) in the theorem statement thus holds.
Lower bound: Suppose $C_4 = 4(C + C^2) \lor 10$

$$
\varepsilon \leq \frac{\delta}{C_4} \min \left( \frac{\rho_{\min}(k, A)}{\tau_B}, 1 \right) = \frac{\delta}{C_5} \quad \text{and} \quad C_4 \varepsilon \leq \delta \min \left( \frac{\rho_{\min}(k, A)}{\tau_B}, 1 \right).
$$

We have by (71) for all $u \in S^{m-1} \cap E$, on event $B_1 \cap B_2 \cap B_9$,

$$
\frac{1}{f} u^T (X^T X) u - u^T tr(B) I_m u
\geq u^T A u - \left( 6 \varepsilon u^T A u + 4 C \varepsilon \tau_B + 8 \varepsilon \tau_B \right) \left\| A^{1/2} u \right\|_2 / \left\| B \right\|_2^{1/2}
$$

$$
\geq u^T A u - 6 \varepsilon u^T A u - 4 C \varepsilon \tau_B - 8 \varepsilon \tau_B \left\| A^{1/2} u \right\|_2
\geq u^T A u - 10 \varepsilon u^T A u - 4(C + C^2) \varepsilon \tau_B \geq u^T A u (1 - 10 \varepsilon - \delta)
$$

$$
\geq u^T A u (1 - 2 \delta)
$$

where $4(C + C^2) \varepsilon \tau_B \leq \delta \rho_{\min}(k, A)$ and $10 \varepsilon \leq \delta$. \qed

\section{Proof of Lemma D.2}

Lemma M.1 is a well-known fact.

**Lemma M.1.** Let $A_{uw} := (u \otimes w) \otimes A$ where $u, w \in S^{p-1}$ where $p \geq 2$. Then $\|A_{uw}\|_2 \leq \|A\|_2$ and $\|A_{uw}\|_F \leq \|A\|_F$.

**Proof of Lemma D.2.** Let $z_1, \ldots, z_f, z'_1, \ldots, z'_f \in \mathbb{R}^m$ be the row vectors $Z_1, Z_2$ respectively. Notice that we can write the quadratic form as follows:

$$
u^T Z_1 A^{1/2} Z_2^T w = \sum_{i,j=1}^{m} u_i w_j z_i A^{1/2} z'_j
$$

$$
= \vec \{ Z_1^T \}^T \left( (u \otimes w) \otimes A^{1/2} \right) \vec \{ Z_2^T \} =: \vec \{ Z_1^T \}^T A_{uw} \vec \{ Z_2^T \},
$$

$$
u^T Z A Z^T w = \vec \{ Z^T \}^T \left( (u \otimes w) \otimes A \right) \vec \{ Z \} =: \vec \{ Z^T \}^T A_{uw} \vec \{ Z \}
$$

where clearly by independence of $Z_1, Z_2$.

$$
\mathbb{E} \vec \{ Z_1^T \}^T \left( (u \otimes w) \otimes A^{1/2} \right) \vec \{ Z_2^T \} = 0, \quad \text{and} \quad \mathbb{E} \vec \{ Z \}^T \left( (u \otimes u) \otimes A \right) \vec \{ Z \} = \text{tr}(u \otimes u \otimes A) = \text{tr}(A).
$$

Thus we invoke (47) and Lemma M.1 to show the concentration bounds on event $\{ |u^T Z_1 A^{1/2} Z_2^T w| > t \}$:

$$
\mathbb{P} \left( |u^T Z_1 A^{1/2} Z_2^T w| > t \right) \leq 2 \exp \left( -\min \left( \frac{i^2}{K^4 \left\| A_{uw} \right\|_F^2}, \frac{t}{K^2 \left\| A_{uw} \right\|_2^{1/2}} \right) \right)
$$

$$
\leq 2 \exp \left( -\min \left( \frac{i^2}{K^4 \text{tr}(A)}, \frac{t}{K^2 \left\| A \right\|_2^{1/2}} \right) \right).
$$
Similarly, we have by Theorem D.1 and Lemma M.1,
\[ \mathbb{P} \left( |u^T ZAZ^Tw - \mathbb{E}u^T ZAZ^Tw| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \|A_{uw}\|_F^2}, \frac{t}{K^2 \|A_{uw}\|_2} \right) \right) \]
\[ \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2} \right) \right). \]
The Lemma thus holds. \[ \square \]

### N Stochastic error bounds

Following Lemma D.2, we have for all \( t > 0, B > 0 \) being an \( f \times f \) symmetric positive definite matrix, and \( v, w \in \mathbb{R}^m \)
\[ \mathbb{P} \left( |v^T Z_1^T B^{1/2} Z_2 w| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \text{tr}(B)}, \frac{t}{K^2 \|B\|_{1/2}} \right) \right) \]
\[ \mathbb{P} \left( |v^T Z_1^T B Z w - \mathbb{E}v^T Z_1^T B Z w| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \|B\|_F}, \frac{t}{K^2 \|B\|_2} \right) \right). \]

#### N.1 Proof for Lemma E.1

Let \( e_1, \ldots, e_m \in \mathbb{R}^m \) be the canonical basis spanning \( \mathbb{R}^m \). Let \( x_1, \ldots, x_m, x'_1, \ldots, x'_m \in \mathbb{R}^f \) be the column vectors \( Z_1, Z_2 \) respectively. Let \( Y \sim e_1^T Z_0^T \). Let \( w_i = \frac{A_{1/2}^i}{\|A_{1/2}^i\|_2} \) for all \( i \). By (47), we obtain for \( t' = C_0 M e K \sqrt{\text{tr}(B) \log m} \) and \( t = C_0 K^2 \sqrt{\log \text{tr}(B)}^{1/2} \):
\[ \mathbb{P} \left( \exists j, |e_j^T B^{1/2} Z_2 e_j| > t' \right) = \mathbb{P} \left( \exists j, \frac{M e}{K} |e_j^T Z_0^T B^{1/2} Z_2 e_j| > C_0 M e K \sqrt{\log \text{tr}(B)}^{1/2} \right) \]
\[ \leq \exp(\log m) \mathbb{P} \left( |Y^T B^{1/2} x'_j| > C_0 K^2 \sqrt{\log \text{tr}(B)}^{1/2} \right) \leq 2/m^3 \]
where the last inequality holds by the union bound, given that \( \frac{\text{tr}(B)}{\|B\|_2} \geq \log m \), and for all \( j \)
\[ \mathbb{P} \left( |Y^T B^{1/2} x'_j| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K^4 \text{tr}(B)}, \frac{t}{K^2 \|B\|_{1/2}} \right) \right), \]
\[ \leq 2 \exp \left( -c \min \left( \frac{C_0^2 \log m}{\|B\|_2}, \frac{C_0 \log^{1/2} m \sqrt{\text{tr}(B)}}{\|B\|_2} \right) \right), \]
\[ \leq 2 \exp \left( -c \min(C_0^2, C_0) \log m \right) \leq 2 \exp (-4 \log m). \]
Let $v, w \in S^{m-1}$. Thus we have by Lemma D.2, for $t_0 = C_0 M, K \sqrt{f \log m}$ and $\tau = C_0 K^2 \sqrt{f \log m}$, $w_j = \frac{A^{1/2} e_i}{\|A^{1/2} e_i\|_2}$ and $f \geq \log m$,

\[
P(\exists j, |e^T Z_1 w_j| > t_0) \leq P \left( \exists j, \frac{M_c}{K} |Y^T Z_1 w_j| > C_0 M, K \sqrt{f \log m} \right) \leq mP \left( |Y^T Z_1 w_j| > C_0 K^2 \sqrt{f \log m} \right) = \exp(\log m) P \left( |e^T Z_1^f Z_1 w_j| > \tau \right) \leq 2 \exp \left( -c \min \left( \frac{\tau^2}{K^4 f}, \frac{\tau^2}{K^2} \right) \right),
\]

\[
\leq 2 \exp \left( -c \min \left( \frac{(C_0 K^2 \sqrt{f \log m})^2}{K^4 f}, \frac{C_0 K^2 \sqrt{f \log m}}{K^2} \right) + \log m \right) \leq 2m \exp \left( -c \min \left( C_0^2 \log m, C_0 \log m \right) \right) \leq 2m \exp \left( -c \min \left( C_0^2, C_0 \right) \log m \right) \leq 2 \exp (-3 \log m).
\]

Therefore we have with probability at least $1 - 4/m^3$,

\[
\|Z_1^T B^{1/2} \epsilon\|_\infty := \max_{j=1,\ldots,m} \langle e^T B^{1/2} Z_2, e_j \rangle \leq t' = C_0 M, K \sqrt{\text{tr}(B) \log m}
\]

\[
\|A^{1/2} Z_1^T \epsilon\|_\infty := \max_{j=1,\ldots,m} \langle A^{1/2} e_j, Z_1^T \epsilon \rangle \leq \max_{j=1,\ldots,m} \left\| A^{1/2} e_j \right\|_2 \max_{j=1,\ldots,m} \langle w_j, Z_1^T \epsilon \rangle \leq a_1^{1/2} t_0 = a_1^{1/2} C_0 M, K \sqrt{f \log m}.
\]

The “moreover” part follows exactly the same arguments as above. Denote by $\tilde{\beta}^* := \beta^* / \|\beta^*\|_2 \in E \cap S^{m-1}$ and $w_i := A^{1/2} e_i / \|A^{1/2} e_i\|_2$. By (72)

\[
P \left( \exists i, \langle w_i, Z_1^T B^{1/2} Z_2 \tilde{\beta}^* \rangle \geq C_0 K^2 \sqrt{\log m \text{tr}(B)} \right) \leq \sum_{i=1}^m P \left( \langle w_i, Z_1^T B^{1/2} Z_2 \tilde{\beta}^* \rangle \geq C_0 K^2 \sqrt{\log m \text{tr}(B)} \right) \leq 2 \exp \left( -c \min \left( C_0^2 \log m, C_0 \log m \right) \right) \leq 2/m^3.
\]

Now for $t = C_0 K^2 \sqrt{\log m} \|B\|_F$, and $\|B\|_F / \|B\|_2 \geq \sqrt{\log m}$,

\[
P \left( \exists e_i : \langle e_i, (Z^T BZ - \text{tr}(B) I_m) \tilde{\beta}^* \rangle \geq C_0 K^2 \sqrt{\log m} \|B\|_F \right) \leq 2m \exp \left[ -c \min \left( \frac{\tau^2}{K^4 \|B\|_F^2}, \frac{\tau^2}{K^2 \|B\|_2^2} \right) \right] \leq 2/m^3.
\]

By the two inequalities immediately above, we have with probability at least $1 - 4/m^3$,

\[
\|X_0^T W \beta^*\|_\infty = \|A^{1/2} Z_1^T B^{1/2} Z_2 \tilde{\beta}^*\|_\infty \leq \|\beta^*\|_2 \max_{e_i} \left( \sup_{w_i} \langle w_i, Z_1^T B^{1/2} Z_2 \tilde{\beta}^* \rangle \right) \leq C_0 K^2 \|\beta^*\|_2 \sqrt{\log m \text{max} \sqrt{\text{tr}(B)}}.
\]

41
and
\[
\| (Z^T BZ - \text{tr}(B)I_m) \beta^* \|_\infty = \left\| (Z^T BZ - \text{tr}(B)I_m) \tilde{\beta}^* \right\|_\infty \| \beta^* \|_2
\]
\[
= \| \beta^* \|_2 \left( \sup_{e_i} \langle e_i, (Z^T BZ - \text{tr}(B)I_m) \tilde{\beta}^* \rangle \right) \leq C_0 K^2 \| \beta^* \|_2 \sqrt{\log m \| B \|_F}.
\]

The last two bounds follow exactly the same arguments as above, except that we replace \( \beta^* \) with \( e_j, j = 1, \ldots, m \) and apply the union bounds to \( m^2 \) events instead of \( m \), and thus \( P(B_{10}) \geq 1 - 4/m^2 \). \( \square \)

### N.2 Proof of Lemma E.2

Let \( z_1, \ldots, z_f, z'_1, \ldots, z'_f \in \mathbb{R}^m \) be the row vectors \( Z_1, Z_2 \) respectively. Let \( w_j = B^{1/2} e_j / \| B^{1/2} e_j \|_2 \). First we write
\[
XX^T - \text{tr}(A)I_f = (Z_1 A^{1/2} + B^{1/2} Z_2)(Z_1 A^{1/2} + B^{1/2} Z_2)^T - \text{tr}(A)I_f
\]
\[
= (Z_1 A^{1/2} + B^{1/2} Z_2)(Z_2^T B^{1/2} + A^{1/2} Z_1^T) - \text{tr}(A)I_f
\]
\[
= Z_1 A^{1/2} Z_2^T B^{1/2} + B^{1/2} Z_2 Z_2^T B^{1/2} + B^{1/2} Z_2 A^{1/2} Z_1^T + Z_1 A Z_1^T - \text{tr}(A)I_f.
\]

Hence for \( B = (b_{ij}) \),
\[
\frac{1}{f} \left| \frac{1}{f} \sum_{i=1}^f e_i^T (XX^T) e_i - \frac{\text{tr}(A)}{m} \right| = \left| \frac{1}{f} \sum_{i=1}^f \frac{e_i^T (XX^T) e_i}{m} - \frac{\text{tr}(A)}{m} \right|
\]
\[
\leq \frac{1}{f} \sum_{j=1}^f \frac{1}{m} e_j^T Z_1 A Z_1^T e_j - \frac{\text{tr}(A)}{m} + \frac{1}{f} \sum_{j=1}^f \frac{1}{m} e_j^T B^{1/2} Z_2 Z_2^T B^{1/2} e_j - \frac{\text{tr}(A)}{m}
\]
\[
+ \frac{1}{f} \sum_{j=1}^f \frac{2}{m} e_j^T Z_1 A^{1/2} Z_2^T B^{1/2} e_j.
\]

First we have by Theorem D.1, for \( \mathbb{E} Z_{ij}^2 = 1 \) and \( \tau = C_0 K^2 \sqrt{\log m \| A \|_F} \),
\[
P( \exists j : e_j^T Z_1 A Z_1^T e_j - \text{tr}(A) \geq \tau) = P( \exists j : z_j^T A z_j - \text{tr}(A) \geq \tau)
\]
\[
\leq 2 f \exp \left( -c \min \left( \frac{\tau^2}{K^4 \| A \|_F^2}, \frac{\tau^2}{K^2 \| A \|_2^2} \right) \right)
\]
\[
\leq 2 f \exp \left( -c \min \left( \frac{(C_0 K^2 \sqrt{\log m \| A \|_F})^2}{K^4 \| A \|_F^2}, \frac{(C_0 K^2 \sqrt{\log m \| A \|_F})^2}{K^2 \| A \|_2^2} \right) \right)
\]
\[
\leq 2 \exp (-4 \log m + \log f).
\]

For \( t_0 = C_0 K^2 \sqrt{\log m \cdot \sqrt{m}, \frac{\| A \|_F^2}{\| A \|_2^2}} \geq \log m \), we have by Lemma D.2,
\[
P( \exists j : \left| \frac{1}{m} e_j^T B^{1/2} Z_2 Z_2^T B^{1/2} e_j - b_{jj} \right| \geq b_{jj} t_0 / m) = P( \exists j : |w_j^T Z_2 Z_2^T w_j - m| \geq t_0)
\]
\[
\leq 2 f \exp \left( -c \min \left( \frac{t_0^2}{K^4 m}, \frac{t_0}{K^2} \right) \right) \leq 2 \exp (-4 \log m + \log f)\]
and hence with probability $1 - 4/m^3$

$$\frac{1}{f} \sum_{j=1}^{f} \left| \frac{1}{m} e_j^T Z_1 A Z_2^T e_j - \frac{\tr(A)}{m} \right| \leq \frac{\tau}{m} = C_0 K^2 \sqrt{\frac{\log m \|A\|_F}{m}},$$

and

$$\frac{1}{f} \sum_{j=1}^{f} \left| \frac{1}{m} e_j^T B^{1/2} Z_2 Z_2^T B^{1/2} e_j - b_{jj} \right| \leq \frac{1}{f} \sum_{j=1}^{f} b_{jj} t_0 / m = \frac{\tr(B)}{f} C_0 K^2 \sqrt{\frac{\log m}{m}}$$

which we denote as event $\mathcal{B}_7$. By Lemma D.2, we have for $t = C_0 K^2 \sqrt{\log \tr(A)}$ \(\frac{1}{2}\)

$$\mathbb{P} \left( \exists j : \left| \sum_{j=1}^{f} e_j^T Z_1 A^{1/2} Z_2^T w_j \right| > t \right) \leq 2f \exp \left( -c \min \left( t^2 \frac{K^2}{K^4 \tr(A)}, \frac{t}{K^2 \|A\|_2^2} \right) \right),$$

$$\leq 2f \exp \left( -c \min \left( \left( \frac{C_0 K^2 \sqrt{\log m \tr(A)}}{K^2 \|A\|_2^2} \right)^2, \frac{C_0 K^2 \sqrt{\log m \tr(A)}}{K^2 \|A\|_2^2} \right) \right),$$

$$\leq 2f \exp \left( -c \min \left( C_0^2 \log m, C_0 \log m \right) \right) \leq 2 \exp (-4 \log m + \log f)$$

where recall $\|A\|_F \leq \sqrt{\tr(A)} \|A\|_2^{1/2}$ and hence

$$\frac{\tr(A)}{\|A\|_2} = \frac{\tr(A)}{\|A\|_2^2} \geq \frac{\|A\|_2^2}{\|A\|_2^2} \geq \log m.$$  

Hence with probability $1 - 2/m^3$

$$\frac{1}{f} \sum_{j=1}^{f} \frac{2}{m} \left| e_j^T Z_1 A^{1/2} Z_2^T B^{1/2} e_j \right| \leq \frac{1}{f} C_0 K^2 \sqrt{\log \tr(A)} \frac{1}{2} \sum_{j=1}^{f} \frac{2}{m} \|B^{1/2} e_j\|_2$$

$$\leq 2C_0 K^2 \sqrt{\frac{\log m \tr(A)}{m}} \frac{1}{\sqrt{m}} b_{\max}^{1/2}$$

which we denote as event $\mathcal{B}_8$. Thus on $\mathcal{B}_6 = \mathcal{B}_7 \cap \mathcal{B}_8$,

$$\frac{1}{f} \left| \tr(B) - \tr(B) \right| \leq C_0 K^2 \sqrt{\frac{\log m}{m}} \left( \frac{\tr(B)}{f} + 2 \frac{\tr(A)}{\sqrt{m}} b_{\max}^{1/2} + \frac{\|A\|_F}{\sqrt{m}} \right).$$

\[\square\]

### O Proof of Theorem 3.1

Denote by $\beta = \beta^*$. Let $S := \text{supp } \beta$, $d = |S|$ and

$$v = \hat{\beta} - \beta.$$

where $\hat{\beta}$ is as defined in (4). We first show Lemma O.1, followed by the proof of Theorem 3.1.

**Lemma O.1.** [3, 27] Suppose that (23) holds. Suppose that there exists a parameter $\psi$ such that

$$\sqrt{d} \tau \leq \frac{\psi}{b_0} \sqrt{\frac{\log m}{f}}, \quad \text{and} \quad \lambda \geq 4\psi \sqrt{\frac{\log m}{f}}$$

where $b_0, \lambda$ are as defined in (4). Then $\|v_{S^c}\|_1 \leq 3 \|v_S\|_1.$
Proof. By the optimality of $\hat{\beta}$, we have

$$
\lambda_n \|\beta\|_1 - \lambda_n \|\hat{\beta}\|_1 \geq \frac{1}{2} \hat{\gamma}^T \hat{\beta} - \frac{1}{2} \beta^T \hat{\Gamma} \beta - \langle \hat{\gamma}, v \rangle
$$

$$
= \frac{1}{2} v^T \hat{\Gamma} v + \langle v, \hat{\Gamma} \beta \rangle - \langle v, \hat{\gamma} \rangle
$$

$$
= \frac{1}{2} v^T \hat{\Gamma} v - \langle v, \hat{\gamma} - \hat{\Gamma} \beta \rangle
$$

Hence, we have for $\lambda \geq 4 \psi \sqrt{\frac{\log m}{f}}$,

$$
\frac{1}{2} v^T \hat{\Gamma} v \leq \langle v, \hat{\gamma} - \hat{\Gamma} \beta \rangle + \lambda_n \left( \|\beta\|_1 - \|\hat{\beta}\|_1 \right)
$$

(73)

Hence

$$
v^T \hat{\Gamma} v \leq \lambda_n \left( 2 \|\beta\|_1 - 2 \|\hat{\beta}\|_1 \right) + 2 \psi \sqrt{\frac{\log m}{f}} \|v\|_1
$$

(74)

$$
\leq \lambda_n \left( 2 \|\beta\|_1 - 2 \|\hat{\beta}\|_1 + \frac{1}{2} \|v\|_1 \right)
$$

$$
\leq \lambda_n \frac{1}{2} \left( 5 \|v_S\|_1 - 3 \|v_{sc}\|_1 \right).
$$

(75)

where by the triangle inequality, and $\beta_{sc} = 0$, we have

$$
2 \|\beta\|_1 - 2 \|\hat{\beta}\|_1 + \frac{1}{2} \|v\|_1 = 2 \|\beta_S\|_1 - 2 \|\hat{\beta}_S\|_1 - 2 \|v_{sc}\|_1 + \frac{1}{2} \|v_S\|_1 + \frac{1}{2} \|v_{sc}\|_1
$$

$$
\leq 2 \|v_S\|_1 - 2 \|v_{sc}\|_1 + \frac{1}{2} \|v_S\|_1 + \frac{1}{2} \|v_{sc}\|_1
$$

$$
\leq \frac{1}{2} \left( 5 \|v_S\|_1 - 3 \|v_{sc}\|_1 \right).
$$

(76)

We now give a lower bound on the LHS of (73)

$$
v^T \hat{\Gamma} v \geq \alpha \|v\|^2 - \tau \|v\|_1^2 \geq -\tau \|v\|_1^2
$$

thus

$$
- v^T \hat{\Gamma} v \leq \|v\|_1^2 \leq \|v\|_1 2b_0 \sqrt{d \tau}
$$

$$
\leq \|v\|_1 2b_0 \frac{\psi}{b_0} \sqrt{\frac{\log m}{f}} = \|v\|_1 \frac{2 \psi \sqrt{\log m}}{f}
$$

$$
\leq \frac{1}{2} \lambda \left( \|v_S\|_1 + \|v_{sc}\|_1 \right)
$$

(77)

where we use the assumption that

$$
\sqrt{d \tau} \leq \frac{\psi}{b_0} \sqrt{\frac{\log m}{f}}, \quad \text{and} \quad \|v\|_1 \leq \|\hat{\beta}\|_1 + \|\beta\|_1 \leq 2b_0 \sqrt{d}
$$
which holds by the triangle inequality and the fact that both \( \hat{\beta} \) and \( \beta \) have \( \ell_1 \) norm being bounded by \( b_0 \sqrt{d} \).

Hence by (75) and (77)

\[
0 \leq -v^T \hat{\Gamma} v + \frac{5}{2} \lambda \|v_S\|_1 - \frac{3}{2} \lambda \|v_{SC}\|_1
\]

Hence by (78) and (77)

\[
0 \leq -\nu^T \hat{\Gamma} \nu + 5 \frac{\lambda}{2} \|v_S\|_1 + \frac{1}{2} \lambda \|v_{SC}\|_1 + \frac{5}{2} \lambda \|v_S\|_1 - \frac{3}{2} \lambda \|v_{SC}\|_1
\]

Thus we have

\[
\|v_{SC}\|_1 \leq 3 \|v_S\|_1
\]

Thus Lemma O.1 holds. □

**Proof** of Theorem 3.1. Following the conclusion of Lemma O.1, we have

\[
\|v\|_1 \leq 4 \|v_S\|_1 \leq 4\sqrt{d} \|v\|_2.
\]

Moreover, we have by the lower-RE condition as in Definition 1.2

\[
v^T \hat{\Gamma} v \geq \alpha \|v\|_2^2 - \tau \|v\|_1^2 \geq (\alpha - 16d\tau) \|v\|_2^2 \geq \frac{1}{2} \alpha \|v\|_2^2
\]

where the last inequality follows from the assumption that \( 16d\tau \leq \alpha/2 \).

Combining the bounds in (81), (80) and (74), we have

\[
\frac{1}{2} \alpha \|v\|_2^2 \leq v^T \hat{\Gamma} v \leq \lambda_n \left( 2 \|\beta\|_1 - 2 \|\tilde{\beta}\|_1 \right) + 2\psi \sqrt{\log \frac{m}{f}} \|v\|_1
\]

\[
\leq \frac{5}{2} \lambda_n \|v_S\|_1 \leq 10\lambda_n \sqrt{d} \|v\|_2
\]

And thus we have \( \|v\|_2 \leq 20\lambda_n \sqrt{d} \). The theorem is thus proved. □

**P Proofs of Lemmas J.1 and J.2 and Corollary J.3**

Throughout the following proofs, we denote by \( r(B) = \frac{\text{tr}(B)}{\|B\|_2} \). Let \( \varepsilon \leq \frac{1}{C} \) where \( C \) is large enough so that \( c\varepsilon C^2 \geq 4 \), and hence the choice of \( C = C_0/\sqrt{d} \) satisfies our need.

**Proof** of Lemma J.1. First we prove concentration bounds for all pairs of \( u, v \in \Pi' \), where \( \Pi' \subset S^{m-1} \) is an \( \varepsilon \)-net of \( E \). Let \( t = CK^2 \varepsilon \text{tr}(B) \). We have by Lemma D.2, and the union bound,

\[
\mathbb{P} \left( \exists u, v \in \Pi', \ |u^T Z^T B Z v - \mathbb{E} u^T Z^T B Z v| > t \right)
\]

\[
\leq 2 \|\Pi'\|^2 \exp \left[ -c \min \left( \frac{t^2}{K^4 \|B\|_2^2}, \frac{t}{K^2 \|B\|_2} \right) \right]
\]

\[
\leq 2 \|\Pi'\|^2 \exp \left[ -c \min \left( C^2, \frac{C^2}{\varepsilon} \right) \frac{\varepsilon^2 r(B)}{K^4} \right] \leq 2 \exp \left( -c \varepsilon^2 r(B) / K^4 \right)
\]
where we use the fact that \(|\|B\|_{F}^{2} \leq \|B\|_{2} \text{tr}(B)|

\[ |\Pi'| \leq \left( \frac{m}{k} \right)^{k} (3/\varepsilon)^{k} \leq \exp(k \log(3em/k\varepsilon)) \]

while
\[
c\min \left( \frac{C^2, CK^2}{\varepsilon} \right) \frac{2r(B)}{K^4} = cC^2 \frac{\text{tr}(B)}{\|B\|_{2} K^4} \geq cC^2 k \log \left( \frac{3em}{k\varepsilon} \right) \geq 4 k \log \left( \frac{3em}{k\varepsilon} \right)
\]

Denote by \(B_2\) the event such that for \(\Lambda := \frac{1}{\text{tr}(B)} (Z^T B Z - I),\)
\[
\sup_{u,v \in \Pi'} |v^T \Lambda u| \leq C \varepsilon =: r'_{f,k}
\]
holds. A standard approximation argument shows that under \(B_2\) and for \(\varepsilon \leq 1/2,\)
\[
\sup_{x,y \in S^{m-1} \cap E} |y^T \Lambda x| \leq \frac{r'_{k,f}}{(1 - \varepsilon)^2} \leq 4C \varepsilon.
\]

The lemma is thus proved. □

**Proof of Lemma J.2.** By Lemma D.2, we have for \(t = C\varepsilon \text{tr}(B)/\|B\|_{2}^{1/2}\) for \(C = C_0/\sqrt{c}\)
\[
P \left( \left| w^T Z_1^T B^{1/2} Z_2 u \right| > t \right) \leq \exp \left( -c \min \left( \frac{C^2 \varepsilon^2 r_B}{K^4 \text{tr}(B)}, \frac{C \varepsilon r_B}{K^2} \right) \right)
\]
\[
\leq 2 \exp \left( -c \min \left( C^2 \frac{\varepsilon^2 r_B}{K^4}, \frac{C \varepsilon^2 r_B}{K^2} \right) \right)
\]
Choose an \(\varepsilon\)-net \(\Pi' \subset S^{m-1}\) such that
\[
\Pi' = \bigcup_{|J|=k} \Pi'_J \quad \text{where} \quad \Pi'_J \subset E_J \cap S^{m-1}
\]
is an \(\varepsilon\)-net for \(E_J \cap S^{m-1}\) and
\[
|\Pi'| \leq \left( \frac{m}{k} \right)^{k} (3/\varepsilon)^{k} \leq \exp(k \log(3em/k\varepsilon)).
\]
Similarly, choose \(\varepsilon\)-net \(\Pi\) of \(F \cap S^{m-1}\) of size at most \(\exp(k \log(3em/k\varepsilon))\). By the union bound and Lemma D.2, and for \(K^2 \geq 1,\)
\[
P \left( \exists w \in \Pi, u \in \Pi' \text{ s.t.} \left| w^T Z_1^T B^{1/2} Z_2 u \right| \geq C\varepsilon \text{tr}(B)/\|B\|_{2}^{1/2} \right)
\]
\[
\leq |\Pi'| |\Pi| 2 \exp \left( -c \min \left( CK^2/\varepsilon, C^2 \right) \varepsilon^2 r_B/K^4 \right)
\]
\[
\leq \exp (2k \log(3em/k\varepsilon)) 2 \exp \left( -cC^2 \varepsilon^2 r_B/K^4 \right)
\]
\[
\leq 2 \exp \left( -c_2 \varepsilon^2 r_B/K^4 \right)
\]
where $C$ is large enough such that $c \epsilon' C^2 := C' > 4$ and for $\epsilon \leq \frac{1}{C'}$,

$$c \min \left( CK^2 / \epsilon, C^2 \right) \epsilon^2 \frac{\text{tr}(B)}{\|B\|_2 K^2} \geq C' \epsilon' \log(3em/k\epsilon) \geq 4k \log(3em/k\epsilon).$$

Denote by $\Upsilon := Z^T_1 B^{1/2} Z_2$. A standard approximation argument shows that if

$$\sup_{w \in \Pi, u \in \Pi'} |w^T \Upsilon u| \leq C \epsilon \frac{\text{tr}(B)}{\|B\|_2^{1/2}} =: r_{k,f}$$

an event which we denote by $B_2$, then for all $u \in E$ and $w \in F$,

$$\left| w^T Z^T_1 B^{1/2} Z_2 u \right| \leq \frac{r_{k,f}}{(1 - \epsilon)^2}. \quad (84)$$

The lemma thus holds for $c_2 \geq C'/2 \geq 2$. □

**Proof of Corollary J.3.** Clearly (69) implies that (66) holds for $B = I$. Clearly (68) holds following the analysis of Lemma J.1 by setting $B = I$, while replacing event $B_1$ with $B_3$, which denotes an event such that

$$\sup_{u, v \in \Pi} \frac{1}{T} \left| v^T (Z^T Z - I) u \right| \leq C \epsilon$$

The rest of the proof follows by replacing $E$ with $F$ everywhere. The corollary thus holds. □