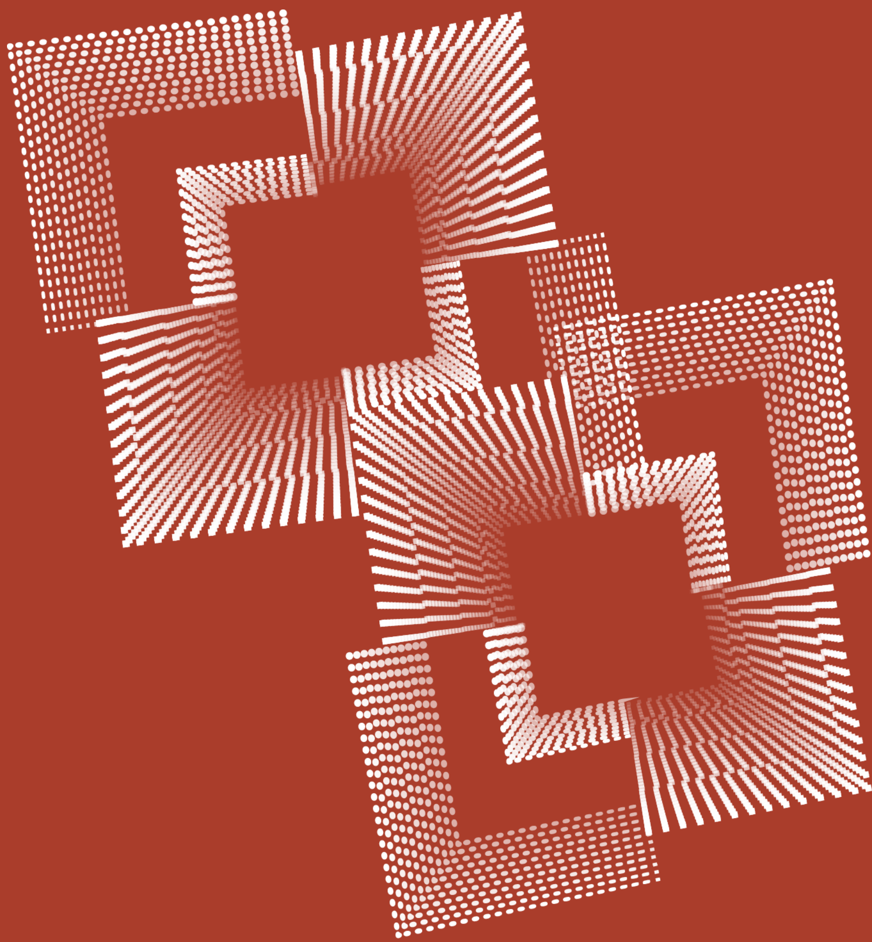
Les Cahiers de la Chaire / N°A-23

Linear and Conic Programming Estimators in High-Dimensional Errors-in-variables Models

Alexandre Belloni, Mathieu Rosenbaum and Alexandre B. Tsybakov



CHAIRE HAVAS-DAUPHINE
ÉCONOMIE ET GESTION DES NOUVELLES DONNÉES
CONNECTING THE DOTS

Linear and Conic Programming Estimators in High-Dimensional Errors-in-variables Models

Alexandre Belloni[1], Mathieu Rosenbaum[2,3] and Alexandre B. Tsybakov[3]

[1] The Fuqua School of Business,
Duke University

[2] Laboratoire de Probabilités et Modèles Aléatoires,
Université Pierre et Marie Curie (Paris 6)

[3] Centre de Recherche en Economie et Statistique,
ENSAE-Paris Tech

August 1, 2014

Abstract

We consider the linear regression model with observation error in the design. In this setting, we allow the number of covariates to be much larger than the sample size. Several new estimation methods have been recently introduced for this model. Indeed, the standard Lasso estimator or Dantzig selector turn out to become unreliable when only noisy regressors are available, which is quite common in practice. We show in this work that under suitable sparsity assumptions, the procedure introduced in [14] is almost optimal in a minimax sense and, despite non-convexities, can be efficiently computed by a single linear programming problem. Furthermore, we provide an estimator attaining the minimax efficiency bound. This estimator is written as a second order cone programming minimisation problem which can be solved numerically in polynomial time.

1 Introduction

We consider the regression model with observation error in the design:

$$y = X\theta^* + \xi, \qquad (1)$$
$$Z = X + W. \qquad (2)$$

Here the random vector $y \in \mathbb{R}^n$ and the random $n \times p$ matrix Z are observed, the $n \times p$ matrix X is unknown, W is an $n \times p$ random noise matrix, $\xi \in \mathbb{R}^n$ is a random noise vector, and θ^* is a vector of unknown parameters to be estimated. For example, the case where the entries of matrix X are missing at random can be reduced to this model. Such linear regressions with errors in both variables have been widely investigated in the literature, see for example [3, 6, 9]. Our work is different in that we consider the setting where the dimension p can be much larger than the sample size n, and θ^* is sparse.

It has been shown in [13] that the presence of observation noise has severe consequences on the usual estimation procedures in the high-dimensional setting. In particular, the

Lasso estimator and Dantzig selector turn out to be inaccurate and fail to identify the sparsity pattern of the vector θ^*. The same paper provides an alternative procedure, called the Matrix Uncertainty selector (MU selector for short), which is robust to the presence of noise. The MU selector $\hat{\theta}^{MU}$ is defined as a solution of the minimisation problem

$$\min\{|\theta|_1 : \ \theta \in \Theta, \ \big|\tfrac{1}{n} Z^T(y - Z\theta)\big|_\infty \leq \mu|\theta|_1 + \tau\}, \tag{3}$$

where $|\cdot|_q$ denotes the ℓ_q-norm for $1 \leq q \leq \infty$, Θ is a given convex subset of $\mathbb{R}^p$ characterising the prior knowledge about θ^*, and the constants μ and τ depend on the level of the noises W and ξ respectively. An extension of the MU selector to generalized linear model is discussed in [17].

In [14], a modification of the MU selector is suggested. It applies when W is a random matrix with independent and zero mean entries W_{ij} such that for any $1 \leq j \leq p$, the sum of expectations

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(W_{ij})^2]$$

is finite and admits a data-driven estimator. This is for example the case in the model with missing data:

$$\widetilde{Z}_{ij} = X_{ij}\eta_{ij}, \quad i = 1, \dots, n, \ j = 1, \dots, p,$$

where for each fixed $j = 1, \dots, p$, the factors $\eta_{ij}, i = 1, \dots, n$, are i.i.d. Bernoulli random variables taking the value 1 with probability $1 - \pi_j$ and 0 with probability π_j, $0 < \pi_j < 1$. Indeed, this model can be rewritten under the form

$$Z_{ij} = X_{ij} + W_{ij},$$

where $Z_{ij} = \widetilde{Z}_{ij}/(1 - \pi_j)$ and $W_{ij} = X_{ij}(\eta_{ij} - (1 - \pi_j))/(1 - \pi_j)$. Therefore, in this model, the σ_j^2 satisfy

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^n X_{ij}^2 \frac{\pi_j}{1 - \pi_j},$$

and it is easily shown that they admit good data-driven estimators $\hat{\sigma}_j^2$, see [14].

The construction of this modified estimator is based on the following idea. We cannot use X in our estimation procedure since only its noisy version Z is available. In particular, the MU selector involves the matrix $Z^T Z/n$ instead of $X^T X/n$. Compared to $X^T X/n$, this matrix contains a bias induced by the diagonal entries of the matrix $W^T W/n$ whose expectations σ_j^2 do not vanish. Therefore, if the σ_j^2 can be estimated, a natural idea is to compensate this bias thanks to these estimates. This leads to a new estimator $\hat{\theta}^C$, called compensated MU selector, and defined as a solution of the minimisation problem

$$\min\{|\theta|_1 : \ \theta \in \Theta, \ \big|\tfrac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta\big|_\infty \leq \mu|\theta|_1 + \tau\}, \tag{4}$$

where $\widehat{D}$ is the diagonal matrix with entries $\hat{\sigma}_j^2$, which are estimators of σ_j^2, and $\mu \geq 0$ and $\tau \geq 0$ are constants chosen according to the level of the noises and the accuracy of the estimators $\hat{\sigma}_j^2$.

Several aspects of the compensated MU selector are studied in [14], in particular the rates of convergence in ℓ_q, the prediction risk and the design of confidence intervals. One of the interest of this modification of the MU selector is that it enables us to obtain bounds for the estimation errors which are decreasing with the sample size n. This is in contrast to the case of the MU selector, where the bounds are small only if the noise W is small. For example, if θ^* is s-sparse, it is shown in [14] that under appropriate assumptions,

$$|\hat{\theta}^C - \theta^*|_q \le C s^{1/q} \sqrt{\frac{\log p}{n}} (|\theta^*|_1 + 1), \quad 1 \le q \le \infty, \tag{5}$$

with probability close to 1, where $C > 0$ is a constant independent of s, p, n, and θ^*. An alternative Lasso type estimator (non-convex program) complemented by an iterative relaxation procedure is introduced in [10]. This method requires the knowledge of the exact value of $|\theta^*|_1$ (or of the property $|\theta^*|_1 \le b\sqrt{s}$ for a constant b), and of an upper bound on $|\theta^*|_2$. Considering the setting where the entries of the regression matrix X are zero-mean subgaussian, it is shown in [10] that if θ^* is s-sparse, under appropriate assumptions, the resulting estimator $\hat{\theta}'$ satisfies

$$|\hat{\theta}' - \theta^*|_2 \le C(\theta^*) s^{1/2} \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1), \tag{6}$$

with probability close to 1, where $C(\theta^*) > 0$ depends on θ^* in a non-specified way. Related covariates selection results are reported in [16]. In [4, 5], the authors propose yet another method of estimation of θ^*, based on orthogonal matching pursuit (OMP). Their procedure needs the parameter s (the exact number of non-zero components of θ^*) as an input. Moreover, they impose the additional assumption that the non-zero components θ_j^* of θ^* are sufficiently large:

$$|\theta_j^*| \ge c \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1), \quad j = 1, \dots, p, \tag{7}$$

where $c > 0$ is a constant. Focusing as in [10] on the case where the entries of the regression matrix X are zero-mean subgaussian, it is shown in [4, 5] that the OMP estimator satisfies a bound analogous to (6) with constant $C(\theta^*) \equiv C > 0$ independent of θ^*, as well as a consistent support recovery property.

These recent developments shed light on errors-in-variables problems in high dimensional settings. However, they are not fully satisfying. Indeed, the following issues are remaining:

• From a practical viewpoint, the use of the above estimators can be intricate. In particular, the minimisation problem (4) is not always a convex one, and [14] does not provide an algorithm enabling to solve it in general case. Although the methods suggested in [10] and in [4, 5] are computationally feasible, they need the knowledge of the parameters $|\theta^*|_1$, $|\theta^*|_2$ or s, which are not available in practice.

• While the bound (5) is more general than (6) (it holds for all q and not only for zero-mean subgaussian X), it is less accurate than (6) in the case $q = 2$ assuming that (6) is established with $C(\theta^*) \equiv C > 0$ independent of θ^*. Indeed, $|\theta^*|_2$ is always smaller than $|\theta^*|_1$. For example, if all components of θ^* take the same value and θ^* is s-sparse, then $|\theta^*|_2 = |\theta^*|_1/\sqrt{s}$. In fact, the optimal rate of convergence in ℓ_q-norm on the class of s-sparse vectors, as a function of s, p, n and the norms $|\theta^*|_r$, remains unknown. When $q = 2$ and X is zero-mean Gaussian, a minimax lower bound including the factor $|\theta^*|_2$ and not $|\theta^*|_1$ is stated without proof in [5]. This, however, does not answer the question in general situation.

The aim of this paper is to provide answers to the above two questions. It is organized as follows. After giving some definitions and assumptions in Sections 2 and 3, we introduce in Section 4 a new estimator $\hat{\theta}$ which is based on second order cone programming and thus can be computed in polynomial time. We show that, under appropriate conditions, this estimator attains bounds of the form

$$|\hat{\theta} - \theta^*|_q \le C s^{1/q} \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1), \quad 1 \le q \le \infty, \tag{8}$$

with probability close to 1, where the constant C does not depend on s, p, n, and θ^*. Contrary to the procedures of [4, 5] and [10], this new estimator does not require the knowledge of $|\theta^*|_1$, $|\theta^*|_2$ or s to be computed. We also do not need a lower bound condition such as (7) on the components of the target vector θ^*. Another difference from the mentioned papers is that our main results do not focus on zero-mean subgaussian regression matrices X, but rather deal with deterministic matrices X commonly appearing in applications. As an easy consequence, we show that the results extend to random matrices X by using suitable deviation properties for the quantity

$$m_2 = \max_{j=1,\dots,p} \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2,$$

where the X_{ij} are the entries of X, as well as checking a restricted eigenvalue type condition on the matrix $X^T X/n$. This extension is possible under the assumption that X is independent of ξ and W.

While the conic programming estimator solves a tractable convex minimisation problem, the compensated MU selector is a non-convex program. Section 5 is devoted to address this issue. We show that under mild assumptions, the compensated MU selector can be reduced to convex programming. In fact, when $\Theta = \mathbb{R}^p$ or Θ is defined by linear constraints, it can even be written as a single linear programming problem, which is of course a computational advantage compared to the estimator based on conic programming. However, the rate of convergence of the compensated MU selector is suboptimal.

Furthermore, in Section 6 we prove minimax lower bounds showing that no estimator can achieve faster rate than that given in (8), up to a logarithmic in s factor, uniformly on a class of s-sparse vectors. Finally, Section 7 provides some simulation results and the proofs are relegated to the appendices.

2 Assumptions on the model

In this section, we introduce the assumptions that will be used below to study the statistical properties of the estimators. Recall that for $\gamma > 0$, the random variable η is said to be *subgaussian with variance parameter* γ^2 (or shortly γ*-subgaussian*) if, for all $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(t\eta)] \le \exp(\gamma^2 t^2/2).$$

A random vector $\zeta \in \mathbb{R}^p$ is said to be *subgaussian with variance parameter* γ^2 (or shortly γ*-subgaussian*) if the inner products (ζ, v) are γ-subgaussian for any $v \in \mathbb{R}^p$ with $|v|_2 = 1$. We shall consider the following assumptions.

(A1) The matrix X is deterministic.

(A2) The elements of the random vector ξ are independent zero-mean subgaussian random variables with variance parameter σ^2.

(A3) The rows w_i, $i = 1, \dots, n$, of the noise matrix W are independent zero-mean sub-gaussian random vectors with variance parameter σ_*^2. Furthermore, W is independent of ξ.

(A4) There exist statistics $\hat{\sigma}_j^2$ such that for any $\varepsilon > 0$, we have

$$\mathbb{P}\big[\max_{j=1,\dots,p} |\hat{\sigma}_j^2 - \sigma_j^2| \geq b(\varepsilon)\big] \leq \varepsilon, \tag{9}$$

where $b(\varepsilon) = c_b\sqrt{\frac{\log(c_b' p/\varepsilon)}{n}}$ for some constants $c_b > 0$ and $c_b' > 0$.

Assumptions (A1) – (A3) are quite standard. Note that we do not assume independence of the components of each w_i. Examples of sufficient conditions for (A4) in the model with missing data are provided in [14].

3 Sensitivities

It is well known, see for example [2], that the performance of the Lasso or Dantzig selector type estimators in high-dimensional linear models is determined by specific characteristics of the Gram matrix

$$\Psi = \frac{1}{n} X^T X,$$

such as the restricted eigenvalue constants. We shall need similar characteristics here. Following [7], we define them in a more general form, so that the required property is a consequence of the restricted eigenvalue property whenever the latter is satisfied. For a vector θ in $\mathbb{R}^p$, we denote by θ_J the vector in $\mathbb{R}^p$ that has the same coordinates as θ on the set of indices $J \subset \{1, \dots, p\}$ and zero coordinates on its complement J^c. We denote by $|J|$ the cardinality of J.

For any $u > 0$ and any subset J of $\{1, \dots, p\}$, consider the cone

$$C_J(u) = \big\{\Delta \in \mathbb{R}^p : \ |\Delta_{J^c}|_1 \leq u|\Delta_J|_1\big\}.$$

The use of such cones to define the restricted eigenvalue constants and other related characteristics of the Gram matrix is standard in the literature on the Lasso and Dantzig selector starting from [2]. For $q \in [1, \infty]$ and an integer $s \in [1, p]$, the paper [7] defines the ℓ_q-*sensitivity* as follows:

$$\kappa_q(s, u) = \min_{J:\ |J| \leq s} \Big(\min_{\Delta \in C_J(u):\ |\Delta|_q = 1} |\Psi\Delta|_\infty \Big).$$

In [7, 8], it is shown that meaningful bounds for various types of errors in sparse linear regression can be obtained in terms of the sensitivities $\kappa_q(s, u)$. In particular, it is proved in [7] that the approach based on sensitivities is more general than that based on restricted eigenvalues or on the coherence condition. In particular, under those assumptions,

$$\kappa_q(s, u) \geq c s^{-1/q},$$

for some constant $c > 0$, which implies the rate optimal bounds for the errors of Lasso and Dantzig selector estimators as in [2]. For convenience, some properties of $\kappa_q(s, u)$ proved in [7] are summarized in Appendix C.

In addition to $\kappa_q(s,u)$, we introduce a *prediction sensitivity* as follows:

$$\kappa_{\rm pr}(s,u) = \min_{J:|J|\le s}\Big(\min_{\Delta\in C_J(u):|\Psi^{1/2}\Delta|_2=1}|\Psi\Delta|_\infty\Big).$$

The sensitivity $\kappa_{\rm pr}(s,u)$ is useful to establish convergence in the prediction norm with fast rates, see (17) in Theorem 2 below. Such rates can be obtained under more general assumptions than rates of convergence in ℓ_q-norm. A discussion of the case of repeated regressors is given in [1]. Lemma 8 in Appendix C shows that $\kappa_{\rm pr}(s,u) > 0$ quite generally. Also, $\kappa_{\rm pr}(s,u) \ge \sqrt{\kappa_1(s,u)}$ (see Lemma 7 in Appendix C).

4 Estimator based on conic programming

In this section, we introduce our conic programming based estimator $\hat\theta$. This estimator is computationally feasible and we provide upper bounds on its estimation and prediction errors. It will be shown in Section 6 that these bounds cannot be improved in a minimax sense. In what follows, we fix a (small) value $\varepsilon > 0$. The probability, with which the bounds on the estimation and prediction errors hold, will be of the form $1 - c\varepsilon$ for some $c > 0$.

To define the estimator $\hat\theta$, we consider the following minimisation problem:

$$\begin{aligned}&\text{minimise } |\theta|_1 + \lambda t \\ &\text{over } (\theta,t) \text{ such that :}\end{aligned} \tag{10}$$

$$\theta\in\Theta,\ \Big|\tfrac{1}{n}Z^T(y-Z\theta)+\widehat D\theta\Big|_\infty \le \mu t+\tau,\ |\theta|_2\le t.$$

Here, λ, μ, and τ are positive tuning constants, and Θ is a given subset of $\mathbb{R}^p$ characterising the prior knowledge about θ. In the results below, μ and τ are of the form

$$\mu = C\sqrt{\frac{\log(p/\varepsilon)}{n}},\quad \tau = C\sqrt{\frac{\log(p/\varepsilon)}{n}}$$

where we denote by $C > 0$ constants depending only on m_2 and on the constants appearing in Assumptions (A1) – (A4). More specifically, in the theory we take

$$\mu = \delta_1'(\varepsilon)+\delta_4'(\varepsilon)+\delta_5(\varepsilon)+b(\varepsilon),\quad \tau = \delta_2(\varepsilon)+\delta_3(\varepsilon)$$

where $\delta_i(\varepsilon)$ and $\delta_i'(\varepsilon)$ are defined in Lemmas 1 and 2 of Appendix A.

When $\Theta = \mathbb{R}^p$ or Θ is a subset of $\mathbb{R}^p$ defined by linear constraints, (10) is a conic programming problem. Therefore it can be efficiently solved in polynomial time.

Let $(\hat\theta,\hat t)$ be a solution of (10). We take $\hat\theta$ as estimator of θ^*. It follows that, under Assumptions (A1) – (A4), the feasible set of the minimisation problem (10) is not empty with high probability if ε is small enough (see Lemma 3 in Appendix B).

The following theorem is our main result about the statistical properties of the estimator $\hat\theta$ based on conic programming.

Theorem 1. *Assume (A1)-(A4), and that the true parameter θ^* is s-sparse and belongs to Θ. Let $\varepsilon > 0$ and $1 \le q \le \infty$. Assume also that*

$$\kappa_q(s,1+\lambda)\ge cs^{-1/q}, \tag{11}$$

for some constant $c > 0$ and that

$$s\le c_1\sqrt{n/\log(p/\varepsilon)}, \tag{12}$$

for some small enough constant $c_1 > 0$*. Then, with probability at least* $1 - 8\varepsilon$*,*

$$|\hat{\theta} - \theta^*|_q \;\leq\; Cs^{1/q}\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_2 + 1), \tag{13}$$

for some constants $C > 0$ *and* $c' > 0$ *(here and in the sequel we set* $s^{1/\infty} = 1$*).*

Under the same assumptions, the prediction risk admits the following bound, with probability at least $1 - 8\varepsilon$*,*

$$\tfrac{1}{n}\big|X(\hat{\theta} - \theta^*)\big|_2^2 \;\leq\; Cs\frac{\log(c'p/\varepsilon)}{n}(|\theta^*|_2 + 1)^2 \,. \tag{14}$$

The constants $C > 0$ *and* $c' > 0$ *in* (13) *and* (14) *depend only on* m_2 *and on the constants appearing in Assumptions (A1) – (A4).*

The proof of Theorem 1 is given in Appendix B.

Some remarks are in order here. Theorem 1 is established under the condition $\kappa_q(s, 1+\lambda) \geq cs^{-1/q}$, which holds under standard assumptions on the matrix X. For example, it holds simultaneously for all q under the coherence assumption, see (45) in Appendix C. For $1 \leq q \leq 2$ this condition follows from the restricted eigenvalue (RE) assumption, see (43), (44) in Appendix C. It is shown in [15] that the RE assumption is satisfied with high probability for a large class of random matrices with dependent entries, including random matrices with zero-mean subgaussian rows and non-trivial covariance structure, as well as matrices with zero-mean independent rows and uniformly bounded entries. Theorem 1 extends to such random matrices X as follows. Fix positive constants $\varepsilon, c, \lambda, m_2$ and denote by $\mathcal{P}$ the class of all probability distributions $\mathbf{P}_X$ on the set of $n \times p$ matrices X such that

$$\mathbf{P}_X\Big[\kappa_q(s, 1+\lambda) \geq cs^{-1/q}, \;\; \max_{j=1,\ldots,p} \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \leq m_2\Big] \geq 1 - \varepsilon. \tag{15}$$

Corollary 1. *Let the assumptions of Theorem 1 be satisfied except for (A1) and* (11)*. Let* X *be a random matrix independent of* (ξ, W) *such that* $\mathbf{P}_X \in \mathcal{P}$*. Then,* (13) *and* (14) *hold with probability at least* $1 - 9\varepsilon$*.*

Although we do not pursue it here, Theorem 1 implies results on the correct selection of the sparsity pattern via a thresholding procedure, in the same spirit as it is done in [11].

Importantly, the bound (13) shows that the conic programming estimator is optimal in a minimax sense. Indeed, we give in Section 6 lower bounds for estimation errors which are in agreement with the upper bounds in (13). The conic programming estimator $\hat{\theta}$ achieves this rate with a computationally feasible procedure and does not need the knowledge of the parameters $|\theta^*|_1$, $|\theta^*|_2$ or s.

Inspection of the proof reveals that if condition (12) does not hold, the conclusions of Theorem 1 are valid provided $|\theta^*|_2$ is replaced by $|\theta^*|_1$ in the bounds, thus leading to results analogous to those for the compensated MU selector. The next theorem formally states that. Note that the assumptions are different and somewhat weaker than in Theorem 1.

Theorem 2. *Assume (A1)-(A4), and that the true parameter* θ^* *is* s*-sparse and belongs to* Θ*. Let* $\varepsilon > 0$ *and* $1 \leq q \leq \infty$*. Then, with probability at least* $1 - 8\varepsilon$*,*

$$|\hat{\theta} - \theta^*|_q \;\leq\; \frac{C}{\kappa_q(s, 1+\lambda)}\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_1 + 1), \tag{16}$$

for some constants $C > 0$ and $c' > 0$. Under the same assumptions, the prediction risk admits the following bound, with probability at least $1 - 8\varepsilon$:

$$\tfrac{1}{n}\big|X(\hat{\theta} - \theta^*)\big|_2^2 \;\le\; \frac{C}{\kappa_{\mathrm{pr}}^2(s, 1+\lambda)} \frac{\log(c'p/\varepsilon)}{n}(|\theta^*|_1 + 1)^2\,. \tag{17}$$

Furthermore, under no assumption on X, with the same probability, we have the following "slow rate" bound:

$$\tfrac{1}{n}\big|X(\widehat{\theta} - \theta^*)\big|_2^2 \;\le\; C\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_1 + 1)^2. \tag{18}$$

The constants $C > 0$ and $c' > 0$ in (16) – (18) *depend only on m_2 and on the constants appearing in assumptions (A1) – (A4).*

The proof of Theorem 2 is given in Appendix B.

There are three different results in Theorem 2. The bound (16) is based on the ℓ_q-sensitivity measures without the sparsity condition (12) and recovers the rates of the compensated MU selector. The second result (17) presents a prediction rate but the prediction sensitivity allows for more general designs. Finally the last result in Theorem 2 provides a slow rate of convergence that requires no assumptions on the design matrix.

5 Computation of the compensated MU selector

The goal of this section is to show that the minimisation problem (4) defining the compensated MU selector can be solved numerically in an efficient way. This algorithmic issue can be intricate since the problem is, in general, not convex, except in some specific situations. For example, if $\Theta = (\mathbb{R}^+)^p$, it obviously reduces to linear programming. However, we shall see that under an additional mild technical hypothesis, solutions can be obtained using convex or even linear programming. It is therefore computationally simpler than the conic programming estimator $\hat{\theta}$. We focus here only on algorithmic aspects. Therefore, we do not recall the assumptions under which the problem admits a solution and the estimator enjoys relevant properties. We refer to [14] where these issues are addressed in detail.

For brevity, we write

$$S(\theta) = \tfrac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta$$

and denote by $(\mathcal{U}_r)_{r\ge 0}$ the family of sets

$$\mathcal{U}_r = \{\theta \in \Theta :\, |S(\theta)|_\infty \le \mu r + \tau\}\,.$$

We also define the function φ by

$$\varphi(r) = \min_{\theta \in \mathcal{U}_r} |\theta|_1.$$

We assume in the next theorem that the equation $r = \varphi(r)$ has a solution. Note that φ is decreasing on $[0, \infty)$ and $\varphi(r) \ge 0$. Moreover, for $r, r' \ge 0$, $\alpha \in [0, 1]$ we have $\alpha\mathcal{U}_r + (1 - \alpha)\mathcal{U}_{r'} \subseteq \mathcal{U}_{\alpha r + (1-\alpha) r'}$ so that φ is a convex function and therefore continuous in its domain. In particular, a solution exists provided $\varphi(0) < \infty$.[1]

[1] More generally, since $\varphi(r) < \infty \Leftrightarrow \mathcal{U}_r \ne \emptyset$, we can define $\underline{r} := \inf\{r \ge 0 : \varphi(r) < \infty\}$. A solution exists if and only if $\underline{r} \le \varphi(\underline{r})$.

We now present our algorithm. Consider the following minimisation problem:

$$\begin{aligned} &\text{minimise } t \\ &\text{over } (t, \theta^+, \theta^-) \text{ such that :} \end{aligned} \tag{19}$$

$$\theta^+ - \theta^- \in \Theta, \theta_j^+ \geq 0, \theta_j^- \geq 0, \; j = 1, \ldots, p,$$

$$t = \sum_{j=1}^{p} (\theta_j^+ + \theta_j^-),$$

$$\left| \tfrac{1}{n} Z^T (y - Z(\theta^+ - \theta^-)) + \widehat{D}(\theta^+ - \theta^-) \right|_\infty \leq \mu t + \tau.$$

Here, θ_j^+, θ_j^- are the components of θ^+, θ^- respectively. As previously, μ, τ are positive tuning constants, and Θ is a given subset of $\mathbb{R}^p$ characterising the prior knowledge about θ.

Let $(\hat{t}, \hat{\theta}^+, \hat{\theta}^-)$ be a solution of (19). We set $\widehat{\theta}^{C'} = \hat{\theta}^+ - \hat{\theta}^-$. Note that (19) is a convex program if Θ is a convex set, and it reduces to a linear program if $\Theta = \mathbb{R}^p$ or if Θ is defined by linear constraints. The use of this algorithm is justified by the following theorem.

Theorem 3. *Assume that there exists a solution $\bar{r}$ to the equation $r = \varphi(r)$. Then $\widehat{\theta}^{C'}$ is a solution of the minimisation problem (4). Moreover, any solution $\widehat{\theta}^{C}$ of (4) induces a solution $(\bar{r}, \theta^+, \theta^-)$ of the problem (19), where θ^+, θ^- are vectors with components $\theta_j^+ = \max\{\hat{\theta}_j^C, 0\}$, $\theta_j^- = \max\{-\hat{\theta}_j^C, 0\}$.*

The proof of Theorem 3 is given in Appendix D.

We would like to emphasize that (19) is not an obvious reformulation because the problem (4) is non-convex. The proof of Theorem 3 exploits the structure of the ℓ_1-norm regularisation. Again, recall that the rates attained by the compensated MU selector are suboptimal. However, it remains attractive compared to the conic programming estimator thanks to the simplicity of its computation.

6 Minimax lower bounds for arbitrary estimators

In this section, we show that the rates of convergence obtained in Theorem 1 are optimal (up to a logarithmic in s term) in a minimax sense for all estimators over the intersection of the class of s-sparse vectors

$$B_0(s) = \{\theta : |\theta|_0 \leq s\}$$

and the ℓ_2-sphere

$$S_2(R) = \{\theta : |\theta|_2 = R\},$$

where $R > 0$. Defining the parameter set as the intersection $\Theta = B_0(s) \cap S_2(R)$ is motivated by the presence of both s and $|\theta^*|_2$ in the upper bounds of Theorem 1. Note that considering a deterministic X means that X is a nuisance parameter of the model. Thus, in the definition of the minimax risk, one should take the maximum not only over Θ but also over a class of possible matrices X. More generally, one can deal with random X and with the maximum over a class of distributions of X. We shall follow this approach with the class of distributions $\mathcal{P}$ introduced in Section 4. The result of Corollary 1 corresponding to (13) can be written as

$$\sup_{\mathbf{P}_X \in \mathcal{P}} \sup_{\theta \in B_0(s) \cap S_2(R)} \mathbb{P}_\theta \Big[|\hat{\theta} - \theta|_q \geq C s^{1/q} \sqrt{\frac{\log(c'p/\varepsilon)}{n}} (R+1) \Big] \leq 9\varepsilon, \tag{20}$$

where, for $\theta^* \in \mathbb{R}^p$, we denote by $\mathbb{P}_{\theta^*}$ the probability measure of the pair (y, Z) satisfying (1)-(2). Our aim now is to prove the reverse inequality to (20) valid for all estimators. For this purpose, instead of the maximum over all $\mathbf{P}_X \in \mathcal{P}$, it suffices to consider one particular distribution $\mathbf{P}_X$. We choose it to be the distribution of Gaussian X with i.i.d. entries. Such matrices satisfy the RE condition with high probability, which implies (15) (see for example [15] for details). Also, we shall assume that ξ and W are Gaussian with i.i.d. entries. In summary, we have the following assumption.

(A5) The elements of the triplet (ξ, X, W) are jointly independent. The components of each of ξ, X, W are i.i.d. Gaussian zero-mean random variables with positive variances σ^2, σ_x^2, and σ_*^2 respectively.

The next theorem provides the desired minimax lower bound.

Theorem 4. *Let $p \geq 2$, $1 \leq q \leq \infty$, $2 \leq s \leq p$, and $R > 0$. Let Assumption (A5) hold, and $s \log(p/s)/n \leq \bar{c}R^2/(R^2+1)$ for some constant $\bar{c} > 0$. Then there exist constants $c > 0$ and $c' > 0$, depending only on $q, \sigma, \sigma_x, \sigma_*, \bar{c}$, such that*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s) \cap S_2(R)} \mathbb{P}_\theta \Big[|\hat{T} - \theta|_q \geq c s^{1/q} \sqrt{\frac{\log(p/s)}{n}} (R+1) \Big] > c', \tag{21}$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators, and we set $s^{1/\infty} = 1$.

The proof of Theorem 4 is given in Appendix D.

7 Monte Carlo study

In this section, we briefly illustrate the empirical performance of the estimators discussed above. We consider the proposed conic programming estimator with $\lambda = 0.5, 0.75,$ and 1 (denoted as Conic (λ)) and the Compensated MU selector (CompMU). To have benchmarks, we also compute the (unfeasible) Dantzig selector which knows X (Dantzig X), and the Dantzig selector that uses only Z (Dantzig Z), ignoring the errors-in-variables issue.

The simulation study uses the following data generating process

$$y_i = x_i^T \theta^* + \xi_i, \qquad z_i = x_i + w_i.$$

Here, ξ_i, w_i, x_i are independent and $\xi_i \sim \mathcal{N}(0, \sigma^2)$, $w_i \sim \mathcal{N}(0, \sigma_*^2 I_{p\times p})$, $x_i \sim \mathcal{N}(0, \Sigma)$ where $I_{p\times p}$ is the identity matrix and Σ is $p \times p$ matrix with elements $\Sigma_{ij} = \rho^{|i-j|}$. We set $\sigma = 0.128$, $\sigma_*^2 = 0.2$, and $\rho = 0.25$. For simplicity, we assume that σ_* and σ are known and we set $\hat{D} = D = \sigma_*^2 I_{p\times p}$. The penalty parameters are set as $\tau = \mu = \sigma\sqrt{\log(p/\varepsilon)/n}$ for $\varepsilon = 0.05$. We consider two choices for the vector of unknown parameters θ^*. The first choice is $\theta^* = 1.25(1,1,1,1,1,0,\dots,0)^T$, which captures the case where the coefficients are well separated from zero. The second choice is $\theta^* = 1.25(1, 1/2, 1/3, 1/4, 1/5, 0, \dots, 0)^T$, which represents the situation where θ^* is sparse with components that are not necessarily well separated from zero.

Table 1 reports the simulation results in the case $\theta^* = 1.25(1,1,1,1,1,0,\dots,0)^T$. As expected, the performance of all the estimators deteriorates as p grows but only slightly. Also, the (unfeasible) estimator based on Dantzig selector that observes X outperforms all feasible options. The estimator that ignores the errors-in-variables issue appears with

First θ^*	$n = 300$ and $p = 10$			$n = 300$ and $p = 50$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
Conic (0.5)	0.0838151	0.1846383	0.1710643	0.0955776	0.2245046	0.2170111
Conic (0.75)	0.0838151	0.1846383	0.1710643	0.0953689	0.2250219	0.2176691
Conic (1)	0.0838151	0.1846383	0.1710643	0.0956858	0.2253614	0.2180705
CompMU	0.1566904	0.2191588	0.2225818	0.1840462	0.2362394	0.2507162
Dantzig X	0.0265486	0.0321528	0.0349530	0.0301636	0.0349420	0.0386731
Dantzig Z	0.2952845	0.3300527	0.3645317	0.3078976	0.4166192	0.4174840

First θ^*	$n = 300$ and $p = 100$			$n = 300$ and $p = 500$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
Conic (0.5)	0.1101178	0.2556778	0.2474407	0.1668239	0.2656095	0.263529
Conic (0.75)	0.0943678	0.2711839	0.2606997	0.1425789	0.2846916	0.2789745
Conic (1)	0.0942906	0.2734750	0.2631424	0.1276741	0.3121221	0.3093194
CompMU	0.1910509	0.2539411	0.2658907	0.2052520	0.2657204	0.2772154
Dantzig X	0.0317776	0.0366155	0.0403419	0.0352309	0.0403134	0.0448000
Dantzig Z	0.3081669	0.4994041	0.4652972	0.3536668	0.6865989	0.5921541

Table 1: Simulation results for the first choice of θ^*. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR).

a higher bias leading to the worse performance in terms of root-squared mean error and empirical risk. The performance of the feasible estimators discussed in this paper is between these two benchmarks. The three conic estimators exhibit a better performance than the compensated MU selector when $p = 10, 50$. For the larger dimensions $p = 100, 500$, their performance becomes similar to that of the compensated MU selector. Nonetheless, the conic estimator with $\lambda = 0.5$ is slightly better than all the other feasible estimators. We also note that for the small dimension $p = 10$, all three conic estimators give the same results. The reason is that the conic constraint was not active for $p = 10$ so that the estimator was the same for the range of λ under consideration. This was not the case for $p = 50, 100, 500$. These findings are very much aligned with the theoretical properties of each estimator and sustain that the impact of errors-in-variables can be substantial.

Table 2 reports the results for the second choice of θ^*, where the coefficients are not well separated from zero. They are qualitatively the same as before. This displays the robustness of the conclusions with respect to possible model selection errors which are unavoidable when coefficients are not well separated from zero.

8 Conclusion

We have studied two estimation methods for high-dimensional linear regression with errors in variables: the compensated MU selector of [14] and the conic programming based estimator. These two procedures are at the same time computationally feasible, realisable in practice (they do not use the knowledge of unaccessible characteristics of the target θ^*) and reasonable in terms of theoretical performances. Namely, the conic programming estimator is rate optimal in a minimax sense, while the compensated MU selector admits somewhat less accurate bounds, with $|\theta^*|_1$ in place of $|\theta^*|_2$. Nevertheless, from an algo-

Second θ^*	$n = 300$ and $p = 10$			$n = 300$ and $p = 50$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
Conic (0.5)	0.0564816	0.1020763	0.0987642	0.0684162	0.1236275	0.1223037
Conic (0.75)	0.0564816	0.1020763	0.0987642	0.0682720	0.1229000	0.1219398
Conic (1)	0.0564816	0.1020763	0.0987642	0.0683291	0.1227749	0.1218898
CompMU	0.0839431	0.1171633	0.1204765	0.1007774	0.1318303	0.1396494
Dantzig X	0.0265486	0.0321528	0.0349530	0.0301636	0.0349420	0.0386731
Dantzig Z	0.1885828	0.2024266	0.2138763	0.1949159	0.2314964	0.2319208

Second θ^*	$n = 300$ and $p = 100$			$n = 300$ and $p = 500$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
Conic (0.5)	0.0714621	0.1374637	0.1349551	0.0945558	0.1472914	0.1477198
Conic (0.75)	0.0713670	0.1378301	0.1353203	0.0824510	0.1589884	0.1565416
Conic (1)	0.0716242	0.1381810	0.1357000	0.0783823	0.1682841	0.1658849
CompMU	0.1063728	0.1405472	0.1479579	0.1131005	0.1477336	0.1545960
Dantzig X	0.0317776	0.0366155	0.0403419	0.0352309	0.0403134	0.0448000
Dantzig Z	0.1978958	0.2536633	0.2432222	0.2152972	0.3145349	0.2766815

Table 2: Simulation results for the second choice of θ^*.

rithmic viewpoint, this last estimator is simpler since, in the cases of major interest, its numerical computation can be reduced to linear programming.

Acknowledgement We would like to thank Anatoli Juditsky for helpful remarks. This work is supported by GENES, and by the French National Research Agency (ANR) as part of Idex Grant ANR -11- IDEX-0003-02, Labex ECODEC (ANR - 11-LABEX-0047), and IPANEMA grant (ANR-13-BSH1-0004-02).

Appendix A. Bounds on stochastic error terms

In this appendix, we give upper bounds on the stochastic error terms appearing in the main results. In what follows D is the diagonal matrix with diagonal elements σ_j^2, $j = 1, \ldots, p$, and for a square matrix A, we denote by $\mathrm{Diag}\{A\}$ the matrix with the same dimensions as A, the same diagonal elements as A and all off-diagonal elements equal to zero. The following lemma is proved in [14].

Lemma 1. *Let $0 < \varepsilon < 1$ and assume (A1)-(A3). Then, with probability at least $1 - \varepsilon$,*

$$\left|\tfrac{1}{n}X^T W\right|_\infty \le \delta_1(\varepsilon), \quad \left|\tfrac{1}{n}X^T \xi\right|_\infty \le \delta_2(\varepsilon), \quad \left|\tfrac{1}{n}W^T \xi\right|_\infty \le \delta_3(\varepsilon),$$
$$\left|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})\right|_\infty \le \delta_4(\varepsilon), \quad \left|\tfrac{1}{n}\mathrm{Diag}\{W^T W\} - D\right|_\infty \le \delta_5(\varepsilon),$$

where

$$\delta_1(\varepsilon) = \sigma_* \sqrt{\frac{2m_2 \log(2p^2/\varepsilon)}{n}}, \quad \delta_2(\varepsilon) = \sigma \sqrt{\frac{2m_2 \log(2p/\varepsilon)}{n}},$$
$$\delta_3(\varepsilon) = \delta_5(\varepsilon) = \bar{\delta}(\varepsilon, 2p), \quad \delta_4(\varepsilon) = \bar{\delta}(\varepsilon, p(p-1)),$$

and for an integer N,

$$\bar{\delta}(\varepsilon, N) = \max\left(\gamma_0\sqrt{\frac{2\log(N/\varepsilon)}{n}},\ \frac{2\log(N/\varepsilon)}{t_0 n}\right),$$

with γ_0, t_0 are positive constants depending only on σ, σ_*.

We now give the second lemma.

Lemma 2. *Let $0 < \varepsilon < 1$, $\theta^* \in \mathbb{R}^p$ and assume (A1)-(A3). Then, with probability at least $1 - \varepsilon$,*

$$\left|\tfrac{1}{n}X^T W\theta^*\right|_\infty \le \delta_1'(\varepsilon)|\theta^*|_2, \tag{22}$$

where $\delta_1'(\varepsilon) = \sigma_*\sqrt{\frac{2m_2\log(2p/\varepsilon)}{n}}$. In addition, with probability at least $1 - \varepsilon$,

$$\left|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})\theta^*\right|_\infty \le \delta_4'(\varepsilon)|\theta^*|_2, \tag{23}$$

where

$$\delta_4'(\varepsilon) = \max\left(\gamma_2\sqrt{\frac{2\log(2p/\varepsilon)}{n}},\ \frac{2\log(2p/\varepsilon)}{t_2 n}\right),$$

and γ_2, t_2 are positive constants depending only on σ_*.

Proof. If $\theta^* = 0$, the result is obvious. So we assume that $\theta^* \neq 0$. Let $v = \theta^*/|\theta^*|_2$. We can write

$$\left|\tfrac{1}{n}X^T W\theta^*\right|_\infty = |\theta^*|_2 \max_{j=1,\dots,p}\left|\frac{1}{n}\sum_{i=1}^n X_{ij}(w_i, v)\right|, \tag{24}$$

where $(w_i, v) = \sum_{k=1}^p W_{ik}v_k$ and we denote by W_{ik} and v_k the elements of the matrix W and the vector v respectively. By Assumption (A3), the random variable (w_i, v) is subgaussian with variance parameter σ_*^2. Using this together with the independence of the w_i for different i, we get that, for all $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\frac{t}{n}\sum_{i=1}^n X_{ij}(w_i, v)\Big)\Big] &= \prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\frac{t}{n}X_{ij}(w_i, v)\Big)\Big] \\ &\le \prod_{i=1}^n \exp\Big(\frac{\sigma_*^2 t^2 X_{ij}^2}{2n^2}\Big) \le \exp\Big(\frac{\sigma_*^2 m_2 t^2}{2n}\Big).\end{aligned}$$

Thus, the random variable

$$\eta_j = \frac{1}{n}\sum_{i=1}^n X_{ij}(w_i, v)$$

is γ_1-subgaussian with $\gamma_1 = \sigma_*\sqrt{m_2/n}$. This implies the classical tail bound

$$\mathbb{P}[|\eta_j| \ge \delta] \le 2\exp\big(-\delta^2/(2\gamma_1^2)\big),$$

for any $\delta > 0$. This together with (24) and the union bound yields (22).

To prove (23), we shall use the following fact (see for example Lemma 5.14 in [19]): If η is a subgaussian random variable with variance parameter γ, then η^2 is sub-exponential, that is there exist constants $\gamma_0 = \gamma_0(\gamma)$ and $t_0 = t_0(\gamma)$ such that

$$\mathbb{E}[\exp(t\eta^2)] \le \exp(\gamma_0^2 t^2/2), \quad |t| \le t_0. \tag{25}$$

Analogously to (24), we obtain

$$\left|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})\theta^*\right|_\infty = |\theta^*|_2 \max_{j=1,\dots,p} |\eta_j'|, \tag{26}$$

where

$$\eta_j' = \frac{1}{n}\sum_{i=1}^n W_{ij} \sum_{k=1,k\neq j}^p W_{ik} v_k.$$

Now, for all $t \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}[\exp(t\eta_j')] &= \prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\frac{tW_{ij}}{n}\sum_{k=1,k\neq j}^p W_{ik}v_k\Big)\Big] \\ &\le \prod_{i=1}^n \mathbb{E}\Big[\exp\Big\{\frac{t}{2n}\Big(W_{ij}^2 + \big(\sum_{k=1,k\neq j}^p W_{ik}v_k\big)^2\Big)\Big\}\Big].\end{aligned}$$

Then, using Cauchy-Schwarz inequality, we get

$$\mathbb{E}[\exp(t\eta_j')] \le \prod_{i=1}^n \Big\{\mathbb{E}\Big[\exp\Big(\frac{tW_{ij}^2}{n}\Big)\Big]\mathbb{E}\Big[\exp\Big(\frac{t}{n}\big(\sum_{k=1,k\neq j}^p W_{ik}v_k\big)^2\Big)\Big]\Big\}^{1/2}.$$

Recall that Assumption (A3) implies that both W_{ij} and $\sum_{k=1,k\neq j}^p W_{ik}v_k$ are σ_*-subgaussian. Consequently, in view of (25), their squared values are $(\gamma_0(\sigma_*), t_0(\sigma_*))$-sub-exponential, which yields

$$\mathbb{E}[\exp(t\eta_j')] \le \prod_{i=1}^n \exp\Big(\frac{\gamma_0(\sigma_*)^2}{2}\Big(\frac{t}{n}\Big)^2\Big) = \exp\Big(\frac{\gamma_0(\sigma_*)^2 t^2}{2n}\Big), \quad |t| \le t_0(\sigma_*)n.$$

Set $\gamma_2 = \gamma_0(\sigma_*)$ and $t_2 = t_0(\sigma_*)$. The last display states that η_j' is $(\gamma_2/\sqrt{n}, t_2 n)$-sub-exponential. This implies the tail bound

$$\mathbb{P}(|\eta_j'| \ge \delta) \le 2\max\big(\exp(-n\delta^2/(2\gamma_2^2)), \exp(-\delta t_2 n/2)\big),$$

for any $\delta > 0$. This together with (26) and the union bound yields (23). □

Appendix B. Proofs of the upper bounds for the estimation and prediction errors

Set for brevity $\delta_i = \delta_i(\varepsilon)$, $\delta_i' = \delta_i'(\varepsilon)$, $b = b(\varepsilon)$. We first prove some preliminary lemmas.

Lemma 3. *Assume (A1)-(A4). Then with probability at least $1 - 6\varepsilon$, the pair $(\theta, t) = (\theta^*, |\theta^*|_2)$ belongs to the feasible set of the minimisation problem (10).*

Proof. First, note that $Z^T(y - Z\theta^*) + n\widehat{D}\theta^*$ is equal to

$$
\begin{aligned}
&- X^T W\theta^* + X^T\xi + W^T\xi - (W^T W - \text{Diag}\{W^T W\})\theta^* \\
&- (\text{Diag}\{W^T W\} - nD)\theta^* + n(\widehat{D} - D)\theta^*.
\end{aligned}
$$

By definition of δ_i and b, with probability at least $1 - 4\varepsilon$, we have

$$
|\tfrac{1}{n}X^T\xi|_\infty + |\tfrac{1}{n}W^T\xi|_\infty \le \delta_2 + \delta_3 \tag{27}
$$

$$
|(\tfrac{1}{n}\text{Diag}\{W^T W\} - D)\theta^*|_\infty \le |\tfrac{1}{n}\text{Diag}\{W^T W\} - D|_\infty|\theta^*|_\infty \le \delta_5|\theta^*|_2 \tag{28}
$$

$$
|(\widehat{D} - D)\theta^*|_\infty \le b|\theta^*|_\infty \le b|\theta^*|_2, \tag{29}
$$

where in (28) and (29) we have used that the considered matrices are diagonal. Also, by Lemma 2, with probability at least $1 - 2\varepsilon$, we have

$$
|\tfrac{1}{n}X^T W\theta^*|_\infty \le \delta_1'|\theta^*|_2 \tag{30}
$$

$$
|\tfrac{1}{n}(W^T W - \text{Diag}\{W^T W\})\theta^*|_\infty \le \delta_4'|\theta^*|_2. \tag{31}
$$

Combining the decomposition of $Z^T(y - Z\theta^*) + n\widehat{D}\theta^*$ together with (27)-(31), we find that

$$
\left|\tfrac{1}{n}Z^T(y - Z\theta^*) + \widehat{D}\theta^*\right|_\infty \le \mu|\theta^*|_2 + \tau,
$$

with probability at least $1 - 6\varepsilon$, which implies the lemma. □

We now give two lemmas which will be crucial in the proof of our main theorem on the accuracy of the conic programming based estimator (Theorem 1).

Lemma 4. *Assume (A1)-(A4). Let $J = \{j : \theta_j^* \neq 0\}$. Then with probability at least $1 - 6\varepsilon$ (on the same event as in Lemma 3), we have*

$$
|(\hat{\theta} - \theta^*)_{J^c}|_1 \le (1 + \lambda)|(\hat{\theta} - \theta^*)_J|_1, \tag{32}
$$

$$
\hat{t} \le (1/\lambda)|\hat{\theta} - \theta^*|_1 + |\theta^*|_2. \tag{33}
$$

Proof. Set $\Delta = \hat{\theta} - \theta^*$. On the event of Lemma 3, $(\theta^*, |\theta^*|_2)$ belongs to the feasible set of the minimisation problem (10). Consequently,

$$
|\hat{\theta}|_1 + \lambda|\hat{\theta}|_2 \le |\hat{\theta}|_1 + \lambda\hat{t} \le |\theta^*|_1 + \lambda|\theta^*|_2. \tag{34}
$$

This implies

$$
|\Delta_{J^c}|_1 \le |\Delta_J|_1 + \lambda(|\theta^*|_2 - |\hat{\theta}|_2) \le |\Delta_J|_1 + \lambda|\Delta_J|_2 \le (1 + \lambda)|\Delta_J|_1,
$$

and (32) follows. To prove (33), it suffices to note that (34) implies

$$
\lambda\hat{t} \le |\theta^*|_1 - |\hat{\theta}|_1 + \lambda|\theta^*|_2 \le |\hat{\theta} - \theta^*|_1 + \lambda|\theta^*|_2.
$$

□

Lemma 5. *Assume (A1)-(A4). Then, on a subset of the event of Lemma 3 having probability at least $1 - 8\varepsilon$, we have*

$$
\left|\tfrac{1}{n}X^T X(\hat{\theta} - \theta^*)\right|_\infty \le \mu_1|\theta^*|_2 + \mu_2|\hat{\theta} - \theta^*|_1 + \tau_1, \tag{35}
$$

where $\mu_1 = \mu + b + \delta_1' + \delta_4' + \delta_5$, $\mu_2 = \mu/\lambda + b + 2\delta_1 + \delta_4 + \delta_5$ and $\tau_1 = \tau + \delta_2$.

Proof. Throughout the proof, we assume that we are on the event of probability at least $1-6\varepsilon$ where inequalities (27) – (31) hold and $(\theta^*, |\theta^*|_2)$ belongs to the feasible set of the minimisation problem (10). Let $\Delta = \hat{\theta} - \theta^*$. We have that

$$|\tfrac{1}{n}X^T X\Delta|_\infty$$

is smaller than

$$|\tfrac{1}{n}Z^T(Z\hat{\theta} - y) - \widehat{D}\hat{\theta}|_\infty + |(\tfrac{1}{n}Z^T W - D)\hat{\theta}|_\infty$$
$$+|(\widehat{D} - D)\hat{\theta}|_\infty + |\tfrac{1}{n}Z^T\xi|_\infty + |\tfrac{1}{n}W^T X\Delta|_\infty.$$

Using the fact that $(\hat{\theta}, \hat{t})$ belongs to the feasible set of the minimisation problem (10) together with (33), we obtain

$$|\tfrac{1}{n}Z^T(Z\hat{\theta} - y) - \widehat{D}\hat{\theta}|_\infty \le \mu\hat{t} + \tau \le (\mu/\lambda)|\hat{\theta} - \theta^*|_1 + \mu|\theta^*|_2 + \tau.$$

Therefore,

$$|\tfrac{1}{n}X^T X\Delta|_\infty$$

does not exceed

$$(\mu/\lambda)|\hat{\theta} - \theta^*|_1 + \mu|\theta^*|_2 + \tau_1 + |(\tfrac{1}{n}Z^T W - D)\hat{\theta}|_\infty + |(\widehat{D} - D)\hat{\theta}|_\infty + |\tfrac{1}{n}W^T X\Delta|_\infty.$$

We now bound the last expression using that $\hat{\theta} = \theta^* + \Delta$, Assumption (A4), and (29). This gives

$$|\tfrac{1}{n}X^T X\Delta|_\infty \le ((\mu/\lambda) + b)|\Delta|_1 + (\mu + b)|\theta^*|_2 + \tau_1 + |(\tfrac{1}{n}Z^T W - D)\theta^*|_\infty$$
$$+ |(\tfrac{1}{n}Z^T W - D)\Delta|_\infty + |\tfrac{1}{n}W^T X\Delta|_\infty. \tag{36}$$

Remark that

$$|(\tfrac{1}{n}Z^T W - D)\Delta|_\infty \le |\tfrac{1}{n}Z^T W - D|_\infty|\Delta|_1$$
$$\le \big(|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})|_\infty + |\tfrac{1}{n}\mathrm{Diag}\{W^T W\} - D|_\infty + |\tfrac{1}{n}X^T W|_\infty\big)|\Delta|_1.$$

Therefore,

$$|(\tfrac{1}{n}Z^T W - D)\Delta|_\infty \le (\delta_1 + \delta_4 + \delta_5)|\Delta|_1, \tag{37}$$

with probability at least $1-8\varepsilon$ (since we intersect the initial event of probability at least $1-6\varepsilon$ with the event of probability at least $1-2\varepsilon$ where the bounds δ_1 and δ_4 hold for the corresponding terms). Next, on the same event of probability at least $1-8\varepsilon$,

$$|\tfrac{1}{n}W^T X\Delta|_\infty \le |\tfrac{1}{n}X^T W|_\infty|\Delta|_1 \le \delta_1|\Delta|_1. \tag{38}$$

Finally, in view of Lemma 2 and (28), on the initial event of probability at least $1-6\varepsilon$,

$$|(\tfrac{1}{n}Z^T W - D)\theta^*|_\infty$$
$$\le|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})\theta^*|_\infty + |(\tfrac{1}{n}\mathrm{Diag}\{W^T W\} - D)\theta^*|_\infty + |\tfrac{1}{n}X^T W\theta^*|_\infty$$
$$\le(\delta_1' + \delta_4' + \delta_5)|\theta^*|_2. \tag{39}$$

To complete the proof, it suffices to plug (37) – (39) in (36) and to set $\mu_1 = \mu + b + \delta_1' + \delta_4' + \delta_5$ and $\mu_2 = \mu/\lambda + b + 2\delta_1 + \delta_4 + \delta_5$. □

Proof of Theorem 1. Throughout the proof, we assume that we are on the event of probability at least $1-8\varepsilon$ of Lemma 5 where the results of Lemmas 3, 4 and 5 hold. Property (32) in Lemma 4 implies that Δ is in the cone $C_J(1+\lambda)$. Therefore, by definition of ℓ_q-sensitivity and Lemma 5, we have

$$\kappa_q(s,1+\lambda)|\Delta|_q \le \left|\tfrac{1}{n}X^TX\Delta\right|_\infty \le \mu_1|\theta^*|_2+\mu_2|\Delta|_1+\tau_1.$$

Furthermore, using again (32), we have

$$\begin{aligned}|\Delta|_1 &= |\Delta_{J^c}|_1+|\Delta_J|_1 \le (2+\lambda)|\Delta_J|_1\\ &\le (2+\lambda)s^{1-1/q}|\Delta_J|_q \le (2+\lambda)s^{1-1/q}|\Delta|_q.\end{aligned}$$

It follows that

$$(\kappa_q(s,1+\lambda)-(2+\lambda)\mu_2 s^{1-1/q})|\Delta|_q \le \mu_1|\theta^*|_2+\tau_1,$$

which implies

$$(c-(2+\lambda)\mu_2 c_1\sqrt{n/\log(p/\varepsilon)})s^{-1/q}|\Delta|_q \le \mu_1|\theta^*|_2+\tau_1$$

in view of the assumptions of the theorem. Recall that $\mu_2 \le a\sqrt{\log(p/\varepsilon)/n}$, where $a>0$ is a constant. Therefore, (13) follows if $c_1 < c((2+\lambda)a)^{-1}$. To prove (14), write first

$$\tfrac{1}{n}\left|X\Delta\right|_2^2 \le \tfrac{1}{n}\left|X^TX\Delta\right|_\infty|\Delta|_1.$$

Next remark that from (13), we have

$$|\Delta|_1 \le Cs\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_2+1) \tag{40}$$

and recall that from Lemma 5,

$$\left|\tfrac{1}{n}X^TX\Delta\right|_\infty \le C\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(1+|\theta^*|_2+|\Delta|_1). \tag{41}$$

Furthermore, from (40) and (12) we also have

$$|\Delta|_1 \le C(|\theta^*|_2+1),$$

for some constant $C>0$. Then (14) is easily deduced.

Proof of Theorem 2. We place ourselves in the same framework as in the proof of Theorem 1. By the definition of the estimator, $|\Delta|_1 \le |\hat\theta|_1+|\theta^*|_1 \le \{|\theta^*|_1+\lambda|\theta^*|_2\}+|\theta^*|_1 \le (2+\lambda)|\theta^*|_1$ where we have used that $|\theta^*|_2 \le |\theta^*|_1$. This and Lemma 5 yield

$$\left|\tfrac{1}{n}X^TX\Delta\right|_\infty \le \mu_1|\theta^*|_2+\mu_2|\Delta|_1+\tau_1 \le (\mu_1+(2+\lambda)\mu_2)|\theta^*|_1+\tau_1.$$

Therefore, arguing as in the proof of Theorem 1, we find

$$\kappa_q(s,1+\lambda)|\Delta|_q \le (\mu_1+(2+\lambda)\mu_2)|\theta^*|_1+\tau_1,$$

which implies (16). To prove (17), we note that by definition of $\kappa_{\rm pr}$ and the fact that $\Delta \in C_J(1+\lambda)$,

$$\frac{\kappa_{\rm pr}^2(s,1+\lambda)}{n}\left|X\Delta\right|_2^2 \le \tfrac{1}{n}\left|X^TX\Delta\right|_\infty^2 \le \left\{(\mu_1+(2+\lambda)\mu_2)|\theta^*|_1+\tau_1\right\}^2.$$

Finally, (18) follows from

$$\tfrac{1}{n}\left|X\Delta\right|_2^2 \le \tfrac{1}{n}\left|X^TX\Delta\right|_\infty|\Delta|_1 \le \{(\mu_1+(2+\lambda)\mu_2)|\theta^*|_1+\tau_1\}(2+\lambda)|\theta^*|_1$$

since $|\Delta|_1 \le (2+\lambda)|\theta^*|_1$.

Appendix C. Properties of the sensitivities

Here we collect some properties of the sensitivities $\kappa_q(s,u)$ and $\kappa_{\rm pr}(s,u)$. First, following [7], we give a relation between $\kappa_q(s,u)$ and the Restricted Eigenvalue (RE) and Coherence (C) constants. For completeness, we recall the Restricted Eigenvalue and Coherence assumptions.

Assumption RE(s,u). *Let $u>0$, $1\le s\le p$. There exists a constant $\kappa_{\rm RE}(s,u)>0$ such that*

$$\min_{\Delta\in C_J(u)\setminus\{0\}} \frac{|\Delta^T\Psi\Delta|}{|\Delta_J|_2^2} \ge \kappa_{\rm RE}(s,u)$$

for all subsets J of $\{1,\dots,p\}$ of cardinality $|J|\le s$.

Assumption C. *All diagonal elements of Ψ are equal to 1 and all its off-diagonal elements of Ψ_{ij} satisfy the coherence condition: $\max_{i\ne j}|\Psi_{ij}|\le\rho$ for some $\rho<1$.*

Assumption C with $\rho<(cs)^{-1}$ and $c>0$ depending only on u implies Assumption RE(s,u), see [2]. The following lemma due to [7] provides useful relations between the constants $\kappa_{\rm RE}$, ρ and κ_q. In this lemma, we denote by c positive constants that do not depend on s.

Lemma 6. *Let $u>0$, $1\le s\le p$. For any $\alpha\in(0,1)$, there exists $c>0$ such that if Assumption C holds with $\rho<(cs)^{-1}$, then*

$$\kappa_\infty(s,u)\ge\alpha. \tag{42}$$

Next, under Assumption RE(s,u),

$$\kappa_1(s,u)\ge(cs)^{-1}\kappa_{\rm RE}(s,u), \tag{43}$$

and, under Assumption RE$(2s,u)$, for any $s\le p/2$, $1<q\le 2$, we have

$$\kappa_q(s,u)\ge c(q)s^{-1/q}\kappa_{\rm RE}(2s,u), \tag{44}$$

where $c(q)>0$ depends only on u and q. Furthermore,

$$\kappa_q(s,u)\ge(2s)^{-1/q}\kappa_\infty(s,u),\quad \forall\ 1\le q\le\infty. \tag{45}$$

Note that (42) and (45) yield the control of the sensitivities κ_q under the Coherence assumption for all $1\le q\le\infty$. The next lemma relates $\kappa_{\rm pr}$ to κ_1.

Lemma 7. *For any $u>0$, $1\le s\le p$,*

$$\kappa_{\rm pr}(s,u)\ge\sqrt{\kappa_1(s,u)}.$$

Proof. Fix a set J such that $|J|\le s$. Since $\Delta^T\Psi\Delta\le|\Psi\Delta|_\infty|\Delta|_1$, we obtain

$$\begin{aligned}\min_{\Delta\in C_J(u):|\Psi^{1/2}\Delta|_2=1}|\Psi\Delta|_\infty &= \min_{\Delta\in C_J(u):|\Psi^{1/2}\Delta|_2>0}|\Psi\Delta|_\infty/\sqrt{\Delta^T\Psi\Delta}\\ &\ge \min_{\Delta\in C_J(u):|\Psi^{1/2}\Delta|_2>0}\sqrt{|\Psi\Delta|_\infty/|\Delta|_1}\\ &\ge \min_{\Delta\in C_J(u):|\Delta|_1>0}\sqrt{|\Psi\Delta|_\infty/|\Delta|_1}\\ &= \min_{\Delta\in C_J(u):|\Delta|_1=1}\sqrt{|\Psi\Delta|_\infty}\end{aligned}$$

where we used the fact that $\{\Delta:|\Psi^{1/2}\Delta|_2>0\}\subseteq\{\Delta:|\Delta|_1>0,|\Psi\Delta|_\infty>0\}$. Taking the minimum over J such that $|J|\le s$ and using the definitions of $\kappa_{\rm pr}(s,u)$ and $\kappa_1(s,u)$ we obtain the result. □

Lemma 8. *If* $\text{rank}(X) = \min\{n, p\}$, *then for any* $u > 0$, $1 \le s \le p$,

$$\kappa_{\text{pr}}(s, u) > 0.$$

Proof. If $\text{rank}(X) = p$ the result follows trivially, so we assume that $\text{rank}(X) = n < p$. We have

$$\begin{aligned} &\min_{\Delta \in C_J(u): |\Psi^{1/2}\Delta|_2 = 1} |\Psi\Delta|_\infty = \min_{\Delta \in C_J(u): |X\Delta/\sqrt{n}|_2 = 1} |X^T X \Delta/n|_\infty \\ &\ge \min_{\Delta \in \mathbb{R}^p: |X\Delta/\sqrt{n}|_2 = 1} |X^T X\Delta/n|_\infty \ge \min_{\delta \in \mathbb{R}^n: |\delta|_2 = 1} |X^T \delta/\sqrt{n}|_\infty, \end{aligned}$$

where we used the fact that $\delta = X\Delta/\sqrt{n} \in \mathbb{R}^n$. Since $\text{rank}(X^T) = \text{rank}(X) = n$, we have $X^T\delta/\sqrt{n} \ne 0$ for all $\delta \in \mathbb{R}^n \setminus \{0\}$. Since $\{\delta \in \mathbb{R}^n : |\delta|_2 = 1\}$ is compact, the minimum is achieved at some δ^*, with $X^T\delta^*/\sqrt{n} \ne 0$, so that $|X^T\delta^*/\sqrt{n}|_\infty > 0$. Taking the minimum over (the finite collection of) J such that $|J| \le s$ yields the result. □

Appendix D. Proofs of Theorems 3 and 4

Proof of Theorem 3. Let $\bar{r}$ be a solution of the equation $r = \varphi(r)$. We set

$$\mathcal{U}_* = \{\theta \in \Theta : |S(\theta)|_\infty \le \mu|\theta|_1 + \tau\}.$$

The minimisation problem (4) has the form

$$\min_{\theta \in \mathcal{U}_*} |\theta|_1.$$

First remark that

$$\min_{\theta \in \mathcal{U}_*} |\theta|_1 \ge \bar{r}. \tag{46}$$

Indeed, with the convention that the minimum over an empty set is equal to $+\infty$, we get

$$\begin{aligned} \min_{\theta \in \mathcal{U}_*} |\theta|_1 &= \min\big(\min_{\theta \in \mathcal{U}_*: |\theta|_1 \le \bar{r}} |\theta|_1, \min_{\theta \in \mathcal{U}_*: |\theta|_1 > \bar{r}} |\theta|_1 \big) \\ &\ge \min\big(\min_{\theta \in \mathcal{U}_{\bar{r}}: |\theta|_1 \le \bar{r}} |\theta|_1, \bar{r} \big) \\ &\ge \min\big(\min_{\theta \in \mathcal{U}_{\bar{r}}} |\theta|_1, \bar{r} \big) = \min(\varphi(\bar{r}), \bar{r}) = \bar{r}. \end{aligned}$$

Let now $\bar{\theta}$ be any solution of

$$\min_{\theta \in \mathcal{U}_{\bar{r}}} |\theta|_1. \tag{47}$$

Then $\bar{\theta} \in \Theta$, $|\bar{\theta}|_1 = \bar{r}$ and

$$|S(\bar{\theta})|_\infty \le \mu\bar{r} + \tau = \mu|\bar{\theta}|_1 + \tau.$$

Thus, $\bar{\theta} \in \mathcal{U}_*$, which implies

$$\min_{\theta \in \mathcal{U}_*} |\theta|_1 \le |\bar{\theta}|_1 = \bar{r}.$$

This and (46) imply that $\bar{\theta}$ is also a solution of (4) and

$$\min_{\theta \in \mathcal{U}_*} |\theta|_1 = \bar{r}. \tag{48}$$

Hence all solutions of (47) are also solutions of (4). Conversely, if θ' is a solution of (4), then, in view of (48), $|\theta'|_1 = \bar{r}$. This and the fact that $\theta' \in \mathcal{U}_*$ imply that $\theta' \in \Theta$ and

$$|S(\theta')|_\infty \le \mu\bar{r} + \tau.$$

This means that $\theta' \in \mathcal{U}_{\bar{r}}$. Since

$$\min_{\theta\in\mathcal{U}_{\bar{r}}} |\theta|_1 = \bar{r} = |\theta'|_1,$$

we get that θ' is a solution of (47). Consequently, the solutions of (4) and (47) coincide.

Let now $\hat{\theta}^C = (\theta_1^C, \dots, \theta_p^C)$ be a solution of (4). Then setting $\theta_j^+ = \max\{\hat{\theta}_j^C, 0\}$, $\theta_j^- = \max\{-\hat{\theta}_j^C, 0\}$, $t = |\theta^+|_1 + |\theta^-|_1$ we have that $\hat{\theta}^C = \theta^+ - \theta^-$ and $|\hat{\theta}^C|_1 = t$. Thus, $(|\hat{\theta}^C|_1, \theta^+, \theta^-)$ is feasible for the problem (19). This implies that the minimum in (19) is lower than the minimum in (4), which yields $|\hat{\theta}^{C'}|_1 \le t = |\hat{\theta}^C|_1$. Moreover, for any solution $(\hat{t}, \hat{\theta}^+, \hat{\theta}^-)$ of (19) the difference $\hat{\theta}^{C'} = \hat{\theta}^+ - \hat{\theta}^-$ satisfies

$$\left|\tfrac{1}{n} Z^T(y - Z\hat{\theta}^{C'}) + \widehat{D}\hat{\theta}^{C'}\right|_\infty \le \mu\hat{t} + \tau \le \mu|\hat{\theta}^C|_1 + \tau = \mu\bar{r} + \tau$$

since $\varphi(\bar{r}) = \bar{r}$. Thus, $\hat{\theta}^{C'} \in \mathcal{U}_{\bar{r}}$. Hence, by definition of φ, we have $\varphi(\bar{r}) \le |\hat{\theta}^{C'}|_1$. Therefore, since we have shown before that $|\hat{\theta}^{C'}|_1 \le |\hat{\theta}^C|_1$, we obtain $|\hat{\theta}^{C'}|_1 = |\hat{\theta}^C|_1 = \bar{r}$ and $(\bar{r}, \theta^+, \theta^-)$ is a solution of (19).

Proof of Theorem 4. We denote by $\mathcal{K}(\mathbb{P}, \mathbb{Q})$ the Kullback-Leibler divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$. We shall use the following lemma.

Lemma 9. *Let $\theta \in \mathbb{R}^p$ and $\theta' \in \mathbb{R}^p$ be such that $|\theta|_2 = |\theta'|_2$. Under Assumption (A5),*

$$\mathcal{K}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{n}{2\sigma_\theta^2}|\theta - \theta'|_2^2,$$

where $\sigma_\theta^2 = \big(\sigma_*^2\sigma_x^2|\theta|_2^2 + \sigma^2(\sigma_*^2 + \sigma_x^2)\big)/\sigma_x^4$.

Proof of this lemma is omitted; it is obtained by a direct calculation using the fact that (y, Z) is jointly Gaussian.

We now proceed to the proof of Theorem 4. Throughout, we will denote by c positive constants which may vary from line to line. To derive the lower bounds, we apply Theorem 2.7 in [18]. Thus, we define a finite set of "hypotheses" included in $B_0(s) \cap S_2(R)$. To this end, we first introduce

$$\mathcal{M} = \big\{x \in \{0,1\}^{p-1} : \rho_H(\mathbf{0}, x) = s - 1\big\},$$

where ρ_H denotes the Hamming distance between elements of $\{0,1\}^{p-1}$, and $\mathbf{0}$ is the zero vector. Then, there exists a subset $\mathcal{M}'$ of $\mathcal{M}$ such that for any x, x' in $\mathcal{M}'$ with $x \neq x'$, we have $\rho_H(x, x') > s/16$, and moreover,

$$\log|\mathcal{M}'| \ge c_1' s \log\left(\frac{p}{s}\right)$$

for some absolute constant $c_1' > 0$. Indeed, this follows from the Varshamov-Gilbert bound (see Lemma 2.9 in [18]) if $s-1 > (p-1)/2$ and from Lemma A.3 in [12] if $s-1 \le (p-1)/2$.

We denote by ω_j' the elements of the finite set $\mathcal{M}'$. For $j = 1, \dots, |\mathcal{M}'|$, we define vectors $\omega_j \in \{0,1\}^p$ with components $\omega_{j1} = 0$ and $\omega_{jk} = \omega'_{j(k-1)}$ for $k \ge 2$, where ω_{jk} is the k-th component of ω_j. We also define ω_0 as the vector in $\{0,1\}^p$ with all components equal to 0 except the first one equal to 1.

We now define the set of "hypotheses" $(\bar{\omega}_j, j = 0, \dots, |\mathcal{M}'|+1)$, where $\bar{\omega}_0 = R\omega_0$, and

$$\bar{\omega}_j = \frac{R}{\sqrt{1+\gamma^2(s-1)}}(\omega_0 + \gamma\omega_j), \quad j = 1, \dots, |\mathcal{M}'|+1.$$

Here, γ is a positive parameter to be defined. Note that the sparsity of $\bar{\omega}_j$ is equal to s and that $|\bar{\omega}_j|_2 = R$. Thus all $\bar{\omega}_j$ belong to $B_0(s) \cap S_2(R)$. Moreover, for $j \geq 1$, by Lemma 9, we have

$$\begin{aligned}\mathcal{K}(\mathbb{P}_{\bar{\omega}_j}, \mathbb{P}_{\bar{\omega}_0}) &\leq \frac{cn}{R^2+1}\Big(R^2\Big(\frac{\sqrt{1+\gamma^2(s-1)}-1}{\sqrt{1+\gamma^2(s-1)}}\Big)^2 + \frac{R^2}{1+\gamma^2(s-1)}\gamma^2 s\Big) \\ &\leq cn\frac{R^2\gamma^2 s}{(R^2+1)(1+\gamma^2(s-1))} \leq c_2' n\gamma^2 s \frac{R^2}{R^2+1},\end{aligned}$$

where $c_2' > 0$ is a constant depending only on σ^2, σ_*^2 and σ_x^2. Now, taking

$$\gamma = \Big(\frac{c_1'}{16c_2' n}\log\Big(\frac{p}{s}\Big)\frac{R^2+1}{R^2}\Big)^{1/2}, \tag{49}$$

we obtain, for all j,

$$\mathcal{K}(\mathbb{P}_{\bar{\omega}_j}, \mathbb{P}_{\bar{\omega}_0}) \leq \frac{1}{16}\log|\mathcal{M}'|.$$

Next, for j and j' both different from 0,

$$|\bar{\omega}_j - \bar{\omega}_{j'}|_q = \frac{R\gamma}{\sqrt{1+\gamma^2(s-1)}}\Big(\sum_{k=1}^{p-1}|\omega_{jk} - \omega_{j'k}|^q\Big)^{1/q} \geq cs^{1/q}\frac{R\gamma}{\sqrt{1+\gamma^2(s-1)}}$$

and for $j \neq 0$,

$$|\bar{\omega}_j - \bar{\omega}_0|_q \geq \frac{R\gamma|\omega_j|_q}{\sqrt{1+\gamma^2(s-1)}} \geq cs^{1/q}\frac{R\gamma}{\sqrt{1+\gamma^2(s-1)}}.$$

The definition of γ in (49) and condition (21) imply that, for any j and j',

$$|\bar{\omega}_j - \bar{\omega}_{j'}|_q \geq cs^{1/q}R\gamma \geq cs^{1/q}(R+1)\sqrt{\frac{\log(p/s)}{n}}.$$

We can now apply Theorem 2.7 in [18] to obtain the result.

References

[1] Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *Annals of Statistics,* **42**, 757 –788.

[2] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics,* **37**, 1705 –1732.

[3] G. Casella and R. L. Berger (1990) *Statistical Inference.* Pacific Grove.

[4] Y. Chen and C. Caramanis (2012) Orthogonal matching pursuit with noisy and missing data: Low and high-dimensional results. `arxiv:1206.0823`

[5] Y. Chen and C. Caramanis (2013) Noisy and missing data regression: Distribution-oblivious support recovery. *Proc. of International Conference on Machine Learning (ICML).*

[6] C. L. Cheng and J. W. van Ness (1999) *Statistical Regression with Measurement Error.* John Wiley & Sons.

[7] E. Gautier and A. B. Tsybakov (2011) High-dimensional instrumental variables regression and confidence sets. `arxiv:1105.2454`

[8] E. Gautier and A. B. Tsybakov (2013) Pivotal estimation in high-dimensional regression via linear programming. In: *Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, B.Schölkopf, Z. Luo, V. Vovk eds., 195 - 204. Springer, New York e.a.

[9] M. G. Kendall and A. Stuart (1973) *The Advanced Theory of Statistics Volume Two.* Charles Griffin and Co Ltd, London, Third edition.

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

[11] K. Lounici (2008) Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic J. of Statistics*, **2**, 90 –102.

[12] P. Rigollet and A.B. Tsybakov (2011) Exponential Screening and optimal rates of sparse estimation. *Annals of Statistics*, **39**, 731–771.

[13] M. Rosenbaum and A.B. Tsybakov (2010) Sparse Recovery under Matrix Uncertainty. *Annals of Statistics*, **38**, 2620–2651.

[14] M. Rosenbaum and A.B. Tsybakov (2013) Improved Matrix Uncertainty Selector. In: *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, M.Banerjee et al. eds. *IMS Collections*, vol.9, 276–290, Institute of Mathematical Statistics.

[15] M. Rudelson, S. Zhou (2013) Reconstruction From Anisotropic Random Measurements. *IEEE Trans. on Information Theory*, **56**, 3434 – 3447.

[16] Ø. Sørensen, A. Frigessi and M. Thoresen (2012) Measurement error in Lasso: Impact and likelihood bias correction. `arxiv:1210.5378`

[17] Ø. Sørensen, A. Frigessi and M. Thoresen (2014) Covariate Selection in High-Dimensional Generalized Linear Models with Measurement Error. `arxiv:1407.1070`

[18] A.B. Tsybakov (2009) *Introduction to Nonparametric Estimation.* Springer, New York.

[19] Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In: *Compressed Sensing, Theory and Applications*, ed. Y. Eldar and G. Kutyniok, Chapter 5. Cambridge University Press.