IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 276–290

DOI: 10.1214/12-IMSCOLL920

Improved matrix uncertainty selector

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

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

$$y = X\theta^* + \xi,$$
$$Z = X + \Xi.$$

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, Ξ 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. We consider the setting where the dimension p can be much larger than the sample size n and θ^* is sparse. Because of the presence of the noise matrix Ξ, the commonly used Lasso and Dantzig selector are unstable. An alternative procedure called the Matrix Uncertainty (MU) selector has been proposed in Rosenbaum and Tsybakov [*The Annals of Statistics* **38** (2010) 2620–2651] in order to account for the noise. The properties of the MU selector have been studied in Rosenbaum and Tsybakov [*The Annals of Statistics* **38** (2010) 2620–2651] for sparse θ^* under the assumption that the noise matrix Ξ is deterministic and its values are small. In this paper, we propose a modification of the MU selector when Ξ is a random matrix with zero-mean entries having the variances that can be estimated. This is, for example, the case in the model where the entries of X are missing at random. We show both theoretically and numerically that, under these conditions, the new estimator called the Compensated MU selector achieves better accuracy of estimation than the original MU selector.

1. Introduction

We consider the model

$$y = X\theta^* + \xi, \tag{1}$$
$$Z = X + \Xi, \tag{2}$$

where 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, Ξ is an $n \times p$ random noise matrix, $\xi \in \mathbb{R}^n$ is a random noise vector, $\theta^* = (\theta_1^*, \dots, \theta_p^*) \in \Theta$ is a vector of unknown parameters to be estimated, and Θ is a given subset of $\mathbb{R}^p$. We consider the problem of estimating an s-sparse vector θ^* (i.e., a vector θ^* having only s non zero components), with p possibly much larger than n. If the matrix X in (1)–(2) is observed without error ($\Xi = 0$), this problem has been recently studied in numerous papers. The proposed estimators mainly rely on ℓ_1 minimization techniques. In particular, this is the case

*Supported in part by ANR "Parcimonie" and by PASCAL-2 Network of Excellence.
[1]Université Pierre et Marie Curie, Paris-6, LPMA, case courrier 188, 4 place Jussieu, 75252 Paris Cedex 05, France and CREST, e-mail: mathieu.rosenbaum@upmc.fr
[2]CREST (ENSAE), 3, av. Pierre Larousse, 92240 Malakoff, France,
e-mail: alexandre.tsybakov@ensae.fr
AMS 2000 subject classifications: Primary 62J05; secondary 62F12
Keywords and phrases: Sparsity, MU selector, matrix uncertainty, errors-in-variables model, measurement error, restricted eigenvalue assumption, missing data

for the widely used Lasso and Dantzig selector, see among others Candès and Tao [6], Bunea et al. [4, 5], Bickel et al. [2], Koltchinskii [8], the book by Bühlmann and van de Geer [3], the lecture notes by Koltchinskii [9], Belloni and Chernozhukov [1] and the references cited therein.

However, it is shown in Rosenbaum and Tsybakov [12] that dealing with a noisy observation of the regression matrix X has severe consequences. In particular, the Lasso and Dantzig selector become very unstable in this context. An alternative procedure, called the matrix uncertainty selector (MU selector for short) is proposed in Rosenbaum and Tsybakov [12] in order to account for the presence of noise Ξ. The MU selector $\hat{\theta}^{MU}$ is defined as a solution of the minimization problem

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

where $|\cdot|_p$ denotes the ℓ_p-norm, $1 \le p \le \infty$, Θ is a given subset of $\mathbb{R}^p$ characterizing the prior knowledge about θ^*, and the constants μ and τ depend on the level of the noises Ξ and ξ respectively. If the noise terms ξ and Ξ are deterministic, it is suggested in Rosenbaum and Tsybakov [12] to choose τ such that

$$\left|\frac{1}{n}Z^T\xi\right|_\infty \le \tau,$$

and to take $\mu = \delta(1 + \delta)$ with δ such that

$$|\Xi|_\infty \le \delta,$$

where, for a matrix A, we denote by $|A|_\infty$ its componentwise ℓ_∞-norm.

In this paper, we propose a modification of the MU selector for the model where Ξ is a random matrix with independent and zero mean entries Ξ_{ij} such that the sums of expectations

$$\sigma_j^2 \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big(\Xi_{ij}^2\big), \quad 1 \le j \le p,$$

are finite and admit data-driven estimators. Our main example where such estimators exist is the model with data missing at random (see below). The idea underlying the new estimator is the following. In the ideal setting where there is no noise Ξ, the estimation strategy for θ^* is based on the matrix X. When there is noise this is impossible since X is not observed and so we have no other choice than using Z instead of X. However, it is not hard to see that under the above assumptions on Ξ, the matrix Z^TZ/n appearing in (3) contains a bias induced by the diagonal entries of the matrix $\Xi^T\Xi/n$ whose expectations σ_j^2 do not vanish. If σ_j^2 can be estimated from the data, it is natural to make a bias correction. This leads to a new estimator $\hat{\theta}$ defined as a solution of the minimization problem

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

where $\widehat{D}$ is the diagonal matrix with entries $\hat{\sigma}_j^2$, which are estimators of σ_j^2, and $\mu \ge 0$ and $\tau \ge 0$ are constants that will be specified later. This estimator $\hat{\theta}$ will be called the Compensated MU selector. In this paper, we show both theoretically and numerically that the estimator $\hat{\theta}$ achieves better performance than the original MU

selector $\hat{\theta}^{MU}$. In particular, under natural conditions given below, the bounds on the error of the Compensated MU selector decrease as $O(n^{-1/2})$ up to logarithmic factors as $n \to \infty$, whereas for the original MU selector $\hat{\theta}^{MU}$ the corresponding bounds do not decrease with n and can be only small if the noise Ξ is small.

Remark 1. The problem (4) is equivalent to

$$\min_{(\theta,u)\in W(\mu,\tau)} |\theta|_1, \tag{5}$$

where

$$W(\mu,\tau) = \left\{(\theta,u) \in \Theta \times \mathbf{R}^p : \left|\frac{1}{n} Z^T(y - Z\theta) + \hat{D}\theta + u\right|_\infty \le \tau,\ |u|_\infty \le \mu|\theta|_1\right\}, \tag{6}$$

with the same μ and τ as in (4) (see the proof in Section 7). This simplifies in some cases the computation of the solution.

An important example where the values σ_j^2 can be estimated is given by the model with missing data. Assume that the elements X_{ij} of the matrix X are unobservable, and we can only observe

$$\tilde{Z}_{ij} = X_{ij}\eta_{ij}, \quad i = 1,\dots,n,\ j = 1,\dots,p, \tag{7}$$

where for each fixed $j = 1,\dots,p$, the factors $\eta_{ij}, i = 1,\dots,n$, are i.i.d. Bernoulli random variables taking value 1 with probability $1 - \pi_j$ and 0 with probability π_j, $0 < \pi_j < 1$. The data X_{ij} is missing if $\eta_{ij} = 0$, which happens with probability π_j. We can rewrite (7) in the form

$$Z_{ij} = X_{ij} + \Xi_{ij}, \tag{8}$$

where $Z_{ij} = \tilde{Z}_{ij}/(1 - \pi_j)$, $\Xi_{ij} = X_{ij}(\eta_{ij} - (1 - \pi_j))/(1 - \pi_j)$. Thus, we can reduce the model with missing data (7) to the form (2) with a matrix Ξ whose elements Ξ_{ij} have zero mean and variance $X_{ij}^2\pi_j/(1 - \pi_j)$. So,

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \frac{\pi_j}{1 - \pi_j}. \tag{9}$$

In Section 4 below, we show that when the π_j are known, the σ_j^2 admit good data-driven estimators $\hat{\sigma}_j^2$. If the π_j are unknown, they can be readily estimated by the empirical frequencies of 0 that we further denote by $\hat{\pi}_j$. Then the $Z_{ij} = \tilde{Z}_{ij}/(1-\pi_j)$ appearing in (8) are not available and should be replaced by $Z_{ij} = \tilde{Z}_{ij}/(1 - \hat{\pi}_j)$. This slightly changes the model and implies a minor modification of the estimator (cf. Section 4).

2. Definitions and notation

Consider the following random matrices

$$M^{(1)} = \frac{1}{n}X^T\Xi, \qquad M^{(2)} = \frac{1}{n}X^T\xi, \qquad M^{(3)} = \frac{1}{n}\Xi^T\xi,$$
$$M^{(4)} = \frac{1}{n}\big(\Xi^T\Xi - \mathrm{Diag}\{\Xi^T\Xi\}\big), \qquad M^{(5)} = \frac{1}{n}\mathrm{Diag}\{\Xi^T\Xi\} - D,$$

where 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.

Under conditions that will be specified below, the entries of the matrices $M^{(k)}$ are small with probability close to 1. Bounds on the ℓ_∞-norms of the matrices $M^{(k)}$ characterize the stochastic error of the estimation. The accuracy of the estimators is determined by these bounds and by the properties of the Gram matrix

$$\Psi \triangleq \frac{1}{n} X^T X.$$

For a vector θ, we denote by θ_J the vector in $\mathbb{R}^p$ that has the same coordinates as θ on the set of indices $J \subset \{1, \ldots, p\}$ and zero coordinates on its complement J^c. We denote by $|J|$ the cardinality of J.

To state our results in a general form, we follow Gautier and Tsybakov [7] and introduce the sensitivity characteristics related to the action of the matrix Ψ on the cone

$$C_J \triangleq \big\{\Delta \in \mathbb{R}^p : \ |\Delta_{J^c}|_1 \le |\Delta_J|_1\big\},$$

where J is a subset of $\{1, \ldots, p\}$. For $q \in [1, \infty]$ and an integer $s \in [1, p]$, we define the ℓ_q *sensitivity* as follows:

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

We will also consider the *coordinate-wise sensitivities*

$$\kappa_k^*(s) \triangleq \min_{J:\ |J| \le s} \Big(\min_{\Delta \in C_J:\ \Delta_k = 1} |\Psi\Delta|_\infty \Big),$$

where Δ_k is the kth coordinate of Δ, $k = 1, \ldots, p$. To get meaningful bounds for various types of estimation errors, we will need the positivity of $\kappa_q(s)$ or $\kappa_k^*(s)$. As shown in Gautier and Tsybakov [7], this requirement is weaker than the usual assumptions related to the structure of the Gram matrix Ψ, such as the Restricted Eigenvalue assumption and the Coherence assumption. For completeness, we recall these two assumptions.

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

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

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

Assumption C. All the 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 $\rho < 1$.

Note that Assumption C with $\rho < (3s)^{-1}$ implies Assumption RE(s) with $\kappa_{\mathrm{RE}}(s) = \sqrt{1 - 3\rho s}$, see Bickel et al. [2] or Lemma 2 in Lounici [10]. From Proposition 4.2 of Gautier and Tsybakov [7] we get that, under Assumption C with $\rho < (2s)^{-1}$,

$$\kappa_\infty(s) \ge 1 - 2\rho s, \tag{10}$$

which yields the control of the sensitivities $\kappa_q(s)$ for all $1 \le q \le \infty$ since

$$\kappa_q(s) \ge (2s)^{-1/q} \kappa_\infty(s), \qquad \forall\ 1 \le q \le \infty, \tag{11}$$

by Proposition 4.1 of Gautier and Tsybakov [7]. Furthermore, Proposition 9.2 of Gautier and Tsybakov [7] implies that, under Assumption RE(s),

$$\kappa_1(s) \geq (4s)^{-1}\kappa_{\mathrm{RE}}(s), \tag{12}$$

and by Proposition 9.3 of that paper, under Assumption RE($2s$) for any $s \leq p/2$ and any $1 < q \leq 2$, we have

$$\kappa_q(s) \geq C(q)s^{-1/q}\kappa_{\mathrm{RE}}(2s), \tag{13}$$

where $C(q) = 2^{-1/q-1/2}(1 + (q-1)^{-1/q})^{-1}$.

3. Main results

In this section, we give bounds on the estimation and prediction errors of the Compensated MU selector. For $\varepsilon \geq 0$, we consider the thresholds $b(\varepsilon) \geq 0$ and $\delta_i(\varepsilon) \geq 0$, $i = 1, \ldots, 5$, such that

$$\mathbb{P}\Big(\max_{j=1,\ldots,p} |\hat{\sigma}_j^2 - \sigma_j^2| \geq b(\varepsilon)\Big) \leq \varepsilon, \tag{14}$$

and

$$\mathbb{P}\big(|M^{(i)}|_\infty \geq \delta_i(\varepsilon)\big) \leq \varepsilon, \qquad i = 1, \ldots, 5. \tag{15}$$

Define

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

and $\mathcal{A}(\varepsilon) = \mathcal{A}(\mu(\varepsilon), \tau(\varepsilon))$, where

$$\mathcal{A}(\mu, \tau) \triangleq \left\{\theta \in \Theta : \left|\frac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \leq \mu|\theta|_1 + \tau\right\}, \qquad \forall\ \mu, \tau \geq 0, \tag{16}$$

and Θ is a given subset of $\mathbb{R}^p$. For $\varepsilon \geq 0$, the Compensated MU selector is defined as a solution of the minimization problem

$$\min\big\{|\theta|_1 :\ \theta \in \mathcal{A}(\varepsilon)\big\}. \tag{17}$$

We have the following result.

Theorem 1. *Assume that model (1)–(2) is valid with an s-sparse vector of parameters $\theta^* \in \Theta$, where Θ is a given subset of $\mathbb{R}^p$. For $\varepsilon \geq 0$, set*

$$\nu(\varepsilon) = 2\big(\mu(\varepsilon) + \delta_1(\varepsilon)\big)|\theta^*|_1 + 2\tau(\varepsilon).$$

Then, with probability at least $1 - 6\varepsilon$, the set $\mathcal{A}(\varepsilon)$ is not empty and for any solution $\hat{\theta}$ of (17) we have

$$|\hat{\theta} - \theta^*|_q \leq \frac{\nu(\varepsilon)}{\kappa_q(s)}, \qquad \forall\ 1 \leq q \leq \infty, \tag{18}$$

$$|\hat{\theta}_k - \theta_k^*| \leq \frac{\nu(\varepsilon)}{\kappa_k^*(s)}, \qquad \forall\ 1 \leq k \leq p, \tag{19}$$

$$\frac{1}{n}|X\big(\hat{\theta} - \theta^*\big)|_2^2 \leq \min\left\{\frac{\nu^2(\varepsilon)}{\kappa_1(s)},\ 2\nu(\varepsilon)|\theta^*|_1\right\}. \tag{20}$$

The proof of this theorem is given in Section 7.

Note that (20) contains a bound on the prediction error under no assumption on X:

$$\frac{1}{n}|X(\hat{\theta}-\theta^*)|_2^2 \le 2\nu(\varepsilon)|\theta^*|_1.$$

The other bounds in Theorem 1 depend on the sensitivities. Using (10)–(13) we obtain the following corollary of Theorem 1.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied. Then, with probability at least* $1-6\varepsilon$*, for any solution* $\hat{\theta}$ *of* (17) *we have the following inequalities.*

(i) Under Assumption RE(s):

$$|\hat{\theta}-\theta^*|_1 \le \frac{4\nu(\varepsilon)s}{\kappa_{\mathrm{RE}}(s)}, \tag{21}$$

$$\frac{1}{n}|X(\hat{\theta}-\theta^*)|_2^2 \le \frac{4\nu^2(\varepsilon)s}{\kappa_{\mathrm{RE}}(s)}. \tag{22}$$

(ii) Under Assumption RE(2s), $s \le p/2$*:*

$$|\hat{\theta}-\theta^*|_q \le \frac{4\nu(\varepsilon)s^{1/q}}{\kappa_{\mathrm{RE}}(2s)}, \qquad \forall\ 1<q\le 2. \tag{23}$$

(iii) Under Assumption C with $\rho < \frac{1}{2s}$*:*

$$|\hat{\theta}-\theta^*|_q < \frac{(2s)^{1/q}\nu(\varepsilon)}{1-2\rho s}, \qquad \forall\ 1\le q\le\infty, \tag{24}$$

where we set $1/\infty = 0$*.*

If the components of ξ and Ξ are subgaussian, the values $\delta_i(\varepsilon)$ are of order $O(n^{-1/2})$ up to logarithmic factors, and the value $b(\varepsilon)$ is of the same order in the model with missing data (see Section 4). Then, the bounds for the Compensated MU selector in Theorem 2 are decreasing with rate $n^{-1/2}$ as $n\to\infty$. This is an advantage of the Compensated MU selector as compared to the original MU selector $\hat{\theta}^{MU}$, for which the corresponding bounds do not decrease with n and can be small only if the noise Ξ is small (cf. Rosenbaum and Tsybakov [12]).

If the matrix X is observed without error ($\Xi = 0$), then $\mu(\varepsilon)=0$, $\delta_i(\varepsilon)=0$, $i\ne 2$, and the Compensated MU selector coincides with the Dantzig selector. In this particular case, the results (ii) and (iii) of Theorem 2 improve, in terms of the constants or the range of validity, upon the corresponding bounds in Bickel et al. [2] and Lounici [10].

4. Control of the stochastic error terms

Theorems 1 and 2 are stated with general thresholds $\delta_i(\varepsilon)$ and $b(\varepsilon)$, and can be used both for random or deterministic noises ξ, Ξ and random or deterministic X. In this section, considering $\varepsilon > 0$ we first derive the values $\delta_i(\varepsilon)$ for random ξ and Ξ with subgaussian entries, and then we specify $b(\varepsilon)$ and the matrix $\widehat{D}$ for the model with missing data. Note that, for random ξ and Ξ, the values $\delta_i(\varepsilon)$ and $b(\varepsilon)$ characterize the stochastic error of the estimator.

4.1. Thresholds $\delta_i(\varepsilon)$ under subgaussian noise

Recall that a zero-mean random variable W is said to be γ-subgaussian ($\gamma > 0$) if, for all $t \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(tW)\big] \le \exp\big(\gamma^2 t^2/2\big). \tag{25}$$

In particular, if W is a zero-mean gaussian or bounded random variable, it is subgaussian. A zero-mean random variable W will be called (γ, t_0)-subexponential if there exist $\gamma > 0$ and $t_0 > 0$ such that

$$\mathbb{E}\big[\exp(tW)\big] \le \exp\big(\gamma^2 t^2/2\big), \qquad \forall\ |t| \le t_0. \tag{26}$$

Let the noise terms ξ and Ξ satisfy the following assumption.

Assumption N. Let $\gamma_\Xi > 0$, $\gamma_\xi > 0$. The entries Ξ_{ij}, $i = 1, \dots, n,\ j = 1, \dots, p$, of the matrix Ξ are zero-mean γ_Ξ-subgaussian random variables, the n rows of Ξ are independent, and $\mathbb{E}(\Xi_{ij}\Xi_{ik}) = 0$ for $j \ne k$, $i = 1, \dots, n$. The components ξ_i of the vector ξ are independent zero-mean γ_ξ-subgaussian random variables satisfying $\mathbb{E}(\Xi_{ij}\xi_i) = 0$, $i = 1, \dots, n,\ j = 1, \dots, p$.

Assumption N implies that the random variables $\Xi_{ij}\xi_i$, $\Xi_{ij}\Xi_{ik}$ are subexponential. Indeed, if two random variables ζ and η are subgaussian, then for some $c > 0$ we have $\mathbb{E}\exp(c\zeta\eta) < \infty$, which implies that (26) holds for $W = \zeta\eta$ with some γ, t_0 whenever $\mathbb{E}(\zeta\eta) = 0$, cf., e.g., Petrov [11], page 56.

Next, $\zeta_j \triangleq (1/n)\sum_{i=1}^n \Xi_{ij}^2 - \sigma_j^2$ is a zero-mean subexponential random variable with variance $O(1/n)$. It is easy to check that (26) holds for $W = \zeta_j$ with $\gamma = O(1/\sqrt{n})$ and $t_0 = O(n)$.

To simplify the notation, we will use a rougher evaluation valid under Assumption N, namely that all $\Xi_{ij}\xi_i$, $\Xi_{ij}\Xi_{ik}$ are (γ_0, t_0)-subexponential with the same $\gamma_0 > 0$ and $t_0 > 0$, and all ζ_j are $(\gamma_0/\sqrt{n}, t_0 n)$-subexponential. Here the constants γ_0 and t_0 depend only on γ_Ξ and γ_ξ. For $0 < \varepsilon < 1$ and an integer N, set

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

Lemma 1. *Let Assumption N be satisfied, and let X be a deterministic matrix with $\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n X_{ij}^2 \triangleq m_2$. Then for any $0 < \varepsilon < 1$ the bound (15) holds with*

$$\delta_1(\varepsilon) = \gamma_\Xi\sqrt{\frac{2m_2\log(2p^2/\varepsilon)}{n}}, \qquad \delta_2(\varepsilon) = \gamma_\xi\sqrt{\frac{2m_2\log(2p/\varepsilon)}{n}}, \tag{27}$$

$$\delta_3(\varepsilon) = \delta_5(\varepsilon) = \bar\delta(\varepsilon, 2p), \qquad \delta_4(\varepsilon) = \bar\delta\big(\varepsilon, p(p-1)\big). \tag{28}$$

Proof. Use the union bound and the facts that $\mathbb{P}(W > \delta) \le \exp(-\delta^2/(2\gamma^2))$ for a γ-subgaussian W, and $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n W_i > \delta) \le \max(\exp(-n\delta^2/(2\gamma^2)), \exp(-\delta t_0 n/2))$ for a sum of independent (γ, t_0)-subexponential W_i. □

4.2. Data-driven $\widehat{D}$ and $b(\varepsilon)$ for the model with missing data

Consider now the model with missing data (7) and assume that X is non-random. Then we have $\tilde Z_{ij}^2 = X_{ij}^2\eta_{ij}$, which implies:

$$\mathbb{E}\big[\tilde Z_{ij}^2\big] = X_{ij}^2(1 - \pi_j), \qquad j = 1, \dots, p.$$

Hence, $\tilde{Z}_{ij}^2\pi_j/(1-\pi_j)^2$ is an unbiased estimator of $X_{ij}^2\pi_j/(1-\pi_j)$. Then σ_j^2 defined in (9) is naturally estimated by

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_{ij}^2 \frac{\pi_j}{(1-\pi_j)^2}, \tag{29}$$

The matrix $\widehat{D}$ is then defined as a diagonal matrix with diagonal entries $\hat{\sigma}_j^2$. It is not hard to prove that $\hat{\sigma}_j^2$ approximates σ_j^2 in probability with rate $O(n^{-1/2})$ up to a logarithmic factor. For example, let the probability that the data is missing be the same for all j: $\pi_1 = \cdots = \pi_p \triangleq \pi_*$. Then

$$\begin{aligned}\mathbb{P}\big(|\hat{\sigma}_j^2 - \sigma_j^2| \geq b\big) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Z}_{ij}^2\frac{\pi_*}{(1-\pi_*)^2} - X_{ij}^2\frac{\pi_*}{(1-\pi_*)}\right)\right| \geq b\right)\\ &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 - \frac{X_{ij}^2}{(1-\pi_*)}\right| \geq \frac{b}{\pi_*}\right) \leq 2\exp\left(-\frac{2nb^2(1-\pi_*)^4}{\pi_*^2 m_4}\right),\end{aligned}$$

where we have used the fact that $0 \leq Z_{ij}^2 \leq X_{ij}^2(1-\pi_*)^{-2}$, Hoeffding's inequality and the notation $m_4 \triangleq \max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n} X_{ij}^4$. This proves (14) with

$$b(\varepsilon) = \frac{\pi_*}{(1-\pi_*)^2}\sqrt{\frac{m_4\log(2p/\varepsilon)}{2n}}.$$

If π_* is unknown, we replace it by the estimator $\hat{\pi} = \frac{1}{np}\sum_{i,j} 1_{\{\tilde{Z}_{ij}=0\}}$, where $1_{\{\cdot\}}$ denotes the indicator function. Another difference is that $Z_{ij} = \tilde{Z}_{ij}/(1-\pi_j)$ appearing in (8) are not available when π_j's are unknown. Therefore, we slightly modify the estimator using $\tilde{Z}_{ij}$ instead of Z_{ij}; we define $\hat{\theta}$ as a solution of $\min\{|\theta|_1 :\ \theta \in \tilde{\mathcal{A}}(\varepsilon)\}$ with

$$\tilde{\mathcal{A}}(\varepsilon) = \left\{\theta\in\Theta : \left|\frac{1}{n}\tilde{Z}^T\big(y(1-\hat{\pi}) - \tilde{Z}\theta\big) + \widehat{D}\theta\right|_\infty \leq \tilde{\mu}(\varepsilon)|\theta|_1 + \tilde{\tau}(\varepsilon)\right\}, \tag{30}$$

where $\tilde{\mu}(\varepsilon)$ and $\tilde{\tau}(\varepsilon)$ are suitably chosen constants, $\tilde{Z}$ is the $n\times p$ matrix with entries $\tilde{Z}_{ij}$, and $\widehat{D}$ is a diagonal matrix with entries $\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_{ij}^2\hat{\pi}/(1-\hat{\pi})^2$. This modification introduces in the bounds an additional term proportional to $\hat{\pi}-\pi_*$, which is of the order $O((np)^{-1/2})$ in probability and hence is negligible as compared to the error bound for the Compensated MU selector.

Remark 2. In this section, we have considered non-random X. Using the same argument, it is easy to derive analogous expressions for $\sigma_i(\varepsilon)$ and $b(\varepsilon)$ when X is a random matrix with independent sub-gaussian entries, and ξ, Ξ (or $\{\eta_{ij}\}$) are independent from X.

5. Confidence intervals

The bounds of Theorems 1 and 2 depend on the unknown matrix X via the sensitivities, and therefore cannot be used to provide confidence intervals. In this section, we show how to address the issue of confidence intervals by deriving other type of bounds based on the empirical sensitivities. Note first that the matrix

$\widehat{\Psi} = \frac{1}{n} Z^T Z - \widehat{D}$ is a natural estimator of the unknown Gram matrix Ψ. It is $\sqrt{n}$-consistent in ℓ_∞-norm under the conditions of the previous section. Therefore, it makes sense to define the empirical counterparts of $\kappa_q(s)$ and $\kappa_k^*(s)$ by the relations:

$$\hat{\kappa}_q(s) \triangleq \min_{J:\ |J|\le s} \Big(\min_{\Delta \in C_J:\ |\Delta|_q = 1} |\widehat{\Psi}\Delta|_\infty \Big),$$

and

$$\hat{\kappa}_k^*(s) \triangleq \min_{J:\ |J|\le s} \Big(\min_{\Delta \in C_J:\ \Delta_k = 1} |\widehat{\Psi}\Delta|_\infty \Big).$$

The values $\hat{\kappa}_q(s)$ and $\hat{\kappa}_k^*(s)$ that we will call the *empirical sensitivities* can be efficiently computed for small s or, alternatively, one can compute data-driven lower bounds on them for any s using linear programming, cf. Gautier and Tsybakov [7].

The following theorem establishes confidence intervals for s-sparse vector θ^* based on the empirical sensitivities.

Theorem 3. *Assume that model (1)–(2) is valid with an s-sparse vector of parameters $\theta^* \in \Theta$, where Θ is a given subset of $\mathbb{R}^p$. Then, with probability at least $1 - 6\varepsilon$, for any solution $\hat{\theta}$ of (17) we have*

$$|\hat{\theta} - \theta^*|_q \le \frac{2(\mu(\varepsilon)|\hat{\theta}|_1 + \tau(\varepsilon))}{\hat{\kappa}_q(s)(1 - \mu(\varepsilon)/\hat{\kappa}_1(s))_+}, \qquad \forall\ 1 \le q \le \infty, \tag{31}$$

$$|\hat{\theta}_k - \theta_k^*| \le \frac{2(\mu(\varepsilon)|\hat{\theta}|_1 + \tau(\varepsilon))}{\hat{\kappa}_k^*(s)(1 - \mu(\varepsilon)/\hat{\kappa}_1(s))_+}, \qquad \forall\ 1 \le k \le p, \tag{32}$$

where $x_+ = \max(0, x)$, and we set $1/0 \triangleq \infty$.

Proof. Set $\Delta = \theta^* - \hat{\theta}$, and write for brevity $S(\theta) = \frac{1}{n} Z^T(y - Z\theta) + \hat{D}\theta$. Using Lemma 2 in Section 7, the fact that $|\Delta_{J^c}|_1 \le |\Delta_J|_1$ where J is the set of non-zero components of θ^* (cf. Lemma 1 in Rosenbaum and Tsybakov [12]) and the definition of the empirical sensitivity $\hat{\kappa}_1(s)$, we find

$$\begin{aligned} |\hat{\Psi}\Delta|_\infty &\le |S(\theta^*)|_\infty + |S(\hat{\theta})|_\infty \\ &\le \mu(\varepsilon)\big(|\theta^*|_1 + |\hat{\theta}|_1\big) + 2\tau(\varepsilon) \\ &\le 2\big(\mu(\varepsilon)|\hat{\theta}|_1 + \tau(\varepsilon)\big) + \mu(\varepsilon)|\Delta|_1 \\ &\le 2\big(\mu(\varepsilon)|\hat{\theta}|_1 + \tau(\varepsilon)\big) + \frac{\mu(\varepsilon)}{\hat{\kappa}_1(s)}|\hat{\Psi}\Delta|_\infty. \end{aligned}$$

This and the definition of $\hat{\kappa}_q(s)$ yield (31). The proof of (32) is analogous, with $\hat{\kappa}_k^*(s)$ used instead of $\hat{\kappa}_q(s)$. □

Remark 3. Note that the bounds (31)–(32) remain valid for $s' \ge s$. Therefore, if one gets an estimator $\hat{s}$ of s such that $\hat{s} \ge s$ with high probability, it can be plugged in into the bounds in order to get completely feasible confidence intervals.

6. Simulations

We consider here the model with missing data (7). Simulations in Rosenbaum and Tsybakov [12] indicate that in this model the MU selector achieves better numerical performance than the Lasso or the Dantzig selector. Here we compare the MU selector with the Compensated MU selector. We design the numerical experiment the following way.

- We take a matrix X of size 100×500 ($n = 100, p = 500$) which is the normalized version (centered and then normalized so that all the diagonal elements of the associated Gram matrix $X^T X/n$ are equal to 1) of a 100×500 matrix with i.i.d. standard Gaussian entries.
- For a given integer s, we randomly (uniformly) choose s non-zero elements in a vector θ^* of size 500. The associated coefficients θ_j^* are set to 0.5, and all other coefficients are set to 0. We take $s = 1, 2, 3, 5, 10$.
- We set $y = X\theta^* + \xi$, where ξ a vector with i.i.d. zero mean and variance ν^2 normal components, $\nu = 0.05/1.96$.
- We compute the values $Z_{ij} = \tilde{Z}_{ij}/(1-\pi_*)$ with $\tilde{Z}_{ij}$ as in (7)[1], and $\pi_j = 0.1 \triangleq \pi_*$ for all j. (The value π_* rather than its empirical counterpart, which is very close to π_*, is used in the algorithm to simplify the computations).
- We run a linear programming algorithm to compute the solutions of (3) and (17) where we optimize over $\Theta = \mathbb{R}_+^{500}$. To simplify the comparison with Rosenbaum and Tsybakov [12], we write μ in the form $(1+\delta)\delta$ with $\delta = 0, 0.01, 0.05, 0.075, 0.1$. In particular, $\delta = 0$ corresponds to the Dantzig selector based on the noisy matrix Z. In practice, one can use an empirical procedure of the choice of δ described in Rosenbaum and Tsybakov [12]. The choice of τ is not crucial and influences only slightly the output of the algorithm. The results presented below correspond to τ chosen in the same way as in the numerical study in Rosenbaum and Tsybakov [12].
- We compute the error measures

$$\mathrm{Err}_1 = |\hat{\theta} - \theta^*|_2^2 \quad \text{and} \quad \mathrm{Err}_2 = |X(\hat{\theta} - \theta^*)|_2^2.$$

  We also record the retrieved sparsity pattern, which is defined as the set of the non-zero coefficients of $\hat{\theta}$.
- For each value of s we run 100 Monte Carlo simulations.

Tables 1–5 present the empirical averages and standard deviations (in brackets) of Err_1, Err_2, of the number of non-zero coefficients in $\hat{\theta}$ (Nb_1) and of the number of non-zero coefficients in $\hat{\theta}$ belonging to the true sparsity pattern (Nb_2). We also

TABLE 1
Results for the model with missing data, $s = 1$

	Err_1	Err_2	Nb_1	Nb_2	Exact
$\delta = 0$	0.0196 (0.0114)	1.334 (0.5865)	70.13 (10.91)	1 (0)	0
C-$\delta = 0$	0.0225 (0.0145)	1.495 (0.6993)	80.09 (8.343)	1 (0)	0
$\delta = 0.01$	0.0131 (0.0069)	0.9318 (0.3606)	45.45 (9.507)	1 (0)	1
C-$\delta = 0.01$	0.0095 (0.0062)	0.8386 (0.4625)	46.88 (9.737)	1 (0)	0
$\delta = 0.05$	0.0100 (0.0038)	0.8001 (0.2121)	12.45 (5.798)	1 (0)	3
C-$\delta = 0.05$	0.0042 (0.0027)	0.3412 (0.1844)	10.52 (5.764)	1 (0)	6
$\delta = 0.075$	0.0100 (0.0030)	0.8878 (0.1869)	6.28 (4.261)	1 (0)	14
C-$\delta = 0.075$	0.0038 (0.0020)	0.3377 (0.1348)	4.91 (3.674)	1 (0)	21
$\delta = 0.1$	0.0110 (0.0024)	1.038 (0.1582)	3.22 (2.640)	1 (0)	36
C-$\delta = 0.1$	0.0044 (0.0015)	0.4255 (0.1040)	2.37 (2.042)	1 (0)	54

[1]Remark that this experiment slightly differs from those in Rosenbaum and Tsybakov [12] where the matrix taken in (3) has entries $\tilde{Z}_{ij}$.

TABLE 2
Results for the model with missing data, $s = 2$

	Err_1	Err_2	Nb_1	Nb_2	Exact
$\delta = 0$	0.0437 (0.0170)	2.756 (1.060)	80.04 (5.149)	2 (0)	0
C-$\delta = 0$	0.0685 (0.0275)	2.951 (1.129)	92.67 (3.911)	2 (0)	0
$\delta = 0.01$	0.0287 (0.0107)	1.838 (0.5423)	49.29 (6.717)	2 (0)	0
C-$\delta = 0.01$	0.0201 (0.0098)	1.561 (0.6827)	48.18 (6.775)	2 (0)	0
$\delta = 0.05$	0.0264 (0.0093)	2.105 (0.4960)	10.35 (4.631)	2 (0)	1
C-$\delta = 0.05$	0.0125 (0.0066)	0.9796 (0.3849)	7.70 (4.092)	2 (0)	8
$\delta = 0.075$	0.0301 (0.0090)	2.694 (0.5022)	4.77 (2.587)	2 (0)	24
C-$\delta = 0.075$	0.0148 (0.0052)	1.359 (0.3573)	3.41 (1.924)	2 (0)	47
$\delta = 0.1$	0.0371 (0.0086)	3.521 (0.4730)	2.62 (1.046)	2 (0)	65
C-$\delta = 0.1$	0.0218 (0.0059)	2.088 (0.3853)	2.28 (0.617)	2 (0)	77

present the total number of simulations where the sparsity pattern is exactly retrieved (Exact). The lines with "$\delta = v$" for $v = 0, 0.01, 0.05, 0.075, 0.1$ correspond to the MU selector and those with "$C - \delta = v$" to the Compensated MU selector.

The results of the simulations are quite convincing. Indeed, the Compensated MU selector improves upon the MU selector with respect to all the considered criteria, in particular when θ^* is very sparse ($s = 1, 2, 3$). The order of magnitude of the improvement is such that, for the best δ, the errors Err_1 and Err_2 are divided by 2. The improvement is not so significant for larger s, especially for $s = 10$ when the model starts to be not very sparse. For all the values of s, the non-zero coefficients of θ^* are systematically in the sparsity pattern both of the MU selector and of the Compensated MU selector. The total number of non-zero coefficients is always smaller (i.e., closer to the correct one) for the Compensated MU selector. Finally, note that the best results for the error measures Err_1 and Err_2 are obtained with $\delta \leq 0.075$, while the sparsity pattern is better retrieved for $\delta = 0.1$. This reflects a trade-off between estimation and selection.

TABLE 3
Results for the model with missing data, $s = 3$

	Err_1	Err_2	Nb_1	Nb_2	Exact
$\delta = 0$	0.0772 (0.0296)	4.361 (1.268)	83.95 (4.177)	3 (0)	0
C-$\delta = 0$	0.1480 (0.0436)	4.258 (1.253)	97.76 (3.262)	3 (0)	0
$\delta = 0.01$	0.0493 (0.0176)	2.929 (0.7907)	49.78 (6.515)	3 (0)	0
C-$\delta = 0.01$	0.0351 (0.0153)	2.328 (0.8442)	48.23 (6.302)	3 (0)	0
$\delta = 0.05$	0.0528 (0.0166)	4.295 (0.7696)	9.82 (3.907)	3 (0)	1
C-$\delta = 0.05$	0.0281 (0.0109)	2.343 (0.6360)	7.02 (3.608)	3 (0)	18
$\delta = 0.075$	0.0643 (0.0161)	5.842 (0.7865)	5.16 (2.086)	3 (0)	29
C-$\delta = 0.075$	0.0384 (0.0106)	3.606 (0.6556)	3.82 (1.177)	3 (0)	57
$\delta = 0.1$	0.0814 (0.0164)	7.792 (0.7434)	3.57 (0.9618)	3 (0)	64
C-$\delta = 0.1$	0.0575 (0.0121)	5.538 (0.6554)	3.13 (0.3912)	3 (0)	89

TABLE 4
Results for the model with missing data, $s = 5$

	Err_1	Err_2	Nb_1	Nb_2	Exact
$\delta = 0$	0.1470 (0.0536)	6.801 (1.686)	87.35 (3.683)	5 (0)	0
C-$\delta = 0$	0.3631 (0.0802)	6.114 (1.490)	104.23 (4.039)	5 (0)	0
$\delta = 0.01$	0.0961 (0.0340)	4.928 (1.180)	49.64 (5.527)	5 (0)	0
C-$\delta = 0.01$	0.0670 (0.0281)	3.627 (1.206)	46.69 (6.298)	5 (0)	0
$\delta = 0.05$	0.1375 (0.0391)	11.100 (1.557)	10.34 (3.347)	5 (0)	6
C-$\delta = 0.05$	0.0864 (0.0307)	7.302 (1.475)	7.42 (2.404)	5 (0)	27
$\delta = 0.075$	0.1769 (0.0427)	15.68 (1.548)	6.85 (1.867)	5 (0)	31
C-$\delta = 0.075$	0.1311 (0.0427)	11.86 (1.737)	5.55 (1.013)	5 (0)	68
$\delta = 0.1$	0.2286 (0.0455)	21.19 (1.385)	5.67 (1.049)	5 (0)	58
C-$\delta = 0.1$	0.1933 (0.0595)	17.71 (2.056)	5.19 (0.6114)	5 (0)	88

7. Proofs

Proof of Remark 1. It is enough to show that $\mathcal{A}(\mu, \tau) = \mathcal{B}(\mu, \tau)$ where

$$\mathcal{B}(\mu, \tau) = \big\{\theta \in \Theta : \ \exists\, u \in \mathbb{R}^p \text{ such that } (\theta, u) \in W(\mu, \tau)\big\}.$$

Let $(\theta, u) \in W(\mu, \tau)$. Using the triangle inequality, we easily get that $\theta \in \mathcal{A}(\mu, \tau)$. Now take $\theta \in \mathcal{A}(\mu, \tau)$. We set

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

and consider $u \in \mathbb{R}^p$ defined by

$$u_i = -N_i 1_{\{|N_i| \le \mu |\theta|_1\}} - \text{sign}(N_i)\mu |\theta|_1 1_{\{|N_i| > \mu |\theta|_1\}},$$

for $i = 1, \ldots, p$, where u_i and N_i are the ith components of u and N respectively. It is easy to check that $(\theta, u) \in W(\mu, \tau)$, which concludes the proof. □

TABLE 5
Results for the model with missing data, $s = 10$

	Err_1	Err_2	Nb_1	Nb_2	Exact
$\delta = 0$	0.4479 (0.1407)	14.56 (3.060)	92.21 (2.881)	10 (0)	0
C-$\delta = 0$	1.208 (0.1705)	11.90 (2.197)	117.23 (6.532)	10 (0)	0
$\delta = 0.01$	0.3512 (0.1263)	13.59 (1.997)	52.76 (5.340)	10 (0)	0
C-$\delta = 0.01$	0.2921 (0.1317)	10.70 (2.049)	48.74 (6.067)	10 (0)	0
$\delta = 0.05$	0.7660 (0.2395)	47.13 (4.389)	20.29 (4.152)	9.96 (0.1959)	0
C-$\delta = 0.05$	0.6919 (0.2696)	41.55 (5.709)	16.99 (4.241)	9.94 (0.2374)	1
$\delta = 0.075$	0.9683 (0.2721)	65.24 (5.496)	16.78 (3.545)	9.85 (0.4092)	0
C-$\delta = 0.075$	0.9443 (0.3067)	61.23 (7.066)	15.00 (3.452)	9.76 (0.5499)	5
$\delta = 0.1$	1.150 (0.2807)	82.86 (6.745)	14.84 (2.948)	9.58 (0.6508)	1
C-$\delta = 0.1$	1.157 (0.3049)	80.43 (8.359)	13.57 (2.804)	9.39 (0.7601)	11

Proof of Theorem 1. The proof is based on two lemmas. For brevity, we will skip the dependence of $b(\varepsilon), \delta_i(\varepsilon)$ and $\nu(\varepsilon)$ on ε.

Lemma 2. *With probability at least* $1-6\varepsilon$, *we have* $\theta^* \in \mathcal{A}(\varepsilon)$.

Proof. We first write that $Z^T(y - Z\theta^*) + n\widehat{D}\theta^*$ is equal to

$$-X^T\Xi\theta^* + X^T\xi + \Xi^T\xi - \big(\Xi^T\Xi - \mathrm{Diag}\{\Xi^T\Xi\}\big)\theta^* - \big(\mathrm{Diag}\{\Xi^T\Xi\} - nD\big)\theta^* + n(\widehat{D} - D)\theta^*.$$

By definition of the $\delta_i(\varepsilon)$ and $b(\varepsilon)$, with probability at least $1-6\varepsilon$ we have

$$\left|\frac{1}{n}X^T\Xi\theta^*\right|_\infty \le \left|\frac{1}{n}X^T\Xi\right|_\infty |\theta^*|_1 \le \delta_1|\theta^*|_1, \tag{33}$$

$$\left|\frac{1}{n}X^T\xi\right|_\infty + \left|\frac{1}{n}\Xi^T\xi\right|_\infty \le \delta_2 + \delta_3, \tag{34}$$

$$\left|\frac{1}{n}\big(\Xi^T\Xi - \mathrm{Diag}\{\Xi^T\Xi\}\big)\theta^*\right|_\infty \le \left|\frac{1}{n}\big(\Xi^T\Xi - \mathrm{Diag}\{\Xi^T\Xi\}\big)\right|_\infty |\theta^*|_1 \le \delta_4|\theta^*|_1, \tag{35}$$

$$\left|\left(\frac{1}{n}\mathrm{Diag}\{\Xi^T\Xi\} - D\right)\theta^*\right|_\infty \le \left|\frac{1}{n}\mathrm{Diag}\{\Xi^T\Xi\} - D\right|_\infty |\theta^*|_1 \le \delta_5|\theta^*|_1, \tag{36}$$

$$|(\widehat{D} - D)\theta^*|_\infty \le b|\theta^*|_1. \tag{37}$$

Therefore $\theta^* \in \mathcal{A}(\varepsilon)$ with probability at least $1-6\varepsilon$. □

Lemma 3. *With probability at least* $1-6\varepsilon$, *for* $\Delta = \hat\theta - \theta^*$ *we have*

$$\left|\frac{1}{n}X^TX\Delta\right|_\infty \le \nu.$$

Proof. Throughout the proof, we assume that we are on event of probability at least $1-6\varepsilon$ where inequalities (33)–(37) hold and $\theta^* \in \mathcal{A}(\varepsilon)$. We have

$$\left|\frac{1}{n}X^TX\Delta\right|_\infty \le \left|\frac{1}{n}Z^T(Z\hat\theta - \Xi\hat\theta - y + \xi)\right|_\infty + \left|\frac{1}{n}\Xi^TX\Delta\right|_\infty.$$

Consequently,

$$\left|\frac{1}{n}X^TX\Delta\right|_\infty \le \left|\frac{1}{n}Z^T(Z\hat\theta - y) - \widehat{D}\hat\theta\right|_\infty + \left|\left(\frac{1}{n}Z^T\Xi - D\right)\hat\theta\right|_\infty + |(\widehat{D} - D)\hat\theta|_\infty + \left|\frac{1}{n}Z^T\xi\right|_\infty + \left|\frac{1}{n}\Xi^TX\Delta\right|_\infty.$$

Using that $\hat\theta \in \mathcal{A}(\varepsilon)$, we easily get that $|\frac{1}{n}X^TX\Delta|_\infty$ is not greater than

$$\mu|\hat\theta|_1 + 2\delta_2 + 2\delta_3 + b|\hat\theta|_1 + \left|\left(\frac{1}{n}Z^T\Xi - D\right)\hat\theta\right|_\infty + \left|\frac{1}{n}\Xi^TX\Delta\right|_\infty.$$

Now remark that

$$\begin{aligned}\left|\left(\frac{1}{n}Z^T\Xi - D\right)\hat\theta\right|_\infty &\le \left|\frac{1}{n}Z^T\Xi - D|_\infty\right|\hat\theta|_1 \\ &\le \left(\left|\frac{1}{n}\big(\Xi^T\Xi - \mathrm{Diag}\{\Xi^T\Xi\}\big)\right|_\infty + \left|\frac{1}{n}\mathrm{Diag}\{\Xi^T\Xi\} - D\right|_\infty + \left|\frac{1}{n}X^T\Xi\right|_\infty\right)|\hat\theta|_1 \\ &\le (\delta_1 + \delta_4 + \delta_5)|\hat\theta|_1.\end{aligned}$$

Finally, using that

$$\left|\frac{1}{n}\Xi^T X\Delta\right|_\infty \le |\hat\theta-\theta^*|_1\left|\frac{1}{n}X^T\Xi\right|_\infty \le \delta_1\big(|\hat\theta|_1+|\theta^*|_1\big)$$

together with the fact that $|\hat\theta|_1 \le |\theta^*|_1$, we obtain the result. □

We now proceed to the proof of Theorem 1. The bounds (18) and (19) follow from Lemma 3, the fact that $|\Delta_{J^c}|_1 \le |\Delta_J|_1$ where J is the set of non-zero components of θ^* (cf. Lemma 1 in Rosenbaum and Tsybakov [12]) and the definition of the sensitivities $\kappa_q(s)$, $\kappa_k^*(s)$. To prove (20), first note that

$$\frac{1}{n}|X\Delta|_2^2 \le \frac{1}{n}|X^TX\Delta|_\infty|\Delta|_1, \tag{38}$$

and use (18) with $q=1$ and Lemma 3. This yields the first term under the minimum on the right hand side of (20). The second term is obtained again from (38), Lemma 3 and the inequality $|\Delta|_1 \le |\hat\theta|_1+|\theta^*|_1 \le 2|\theta^*|_1$. □

Proof of Theorem 2. The bounds (21) and (24) follow by combining (18) with (12) and with (10)–(11) respectively. Next, (22) follows from (20) and (12). Also, as an easy consequence of (18) and (13) with $q=2$ we get

$$|\hat\theta-\theta^*|_2 \le \frac{4\nu(\varepsilon)s^{1/2}}{\kappa_{\mathrm{RE}}(2s)}.$$

Finally, (23) follows from this inequality and (21) using the interpolation formula $|\Delta|_q^q \le |\Delta|_1^{2-q}|\Delta|_2^{2(q-1)}$ for $\Delta=\hat\theta-\theta^*$, and the fact that $\kappa_{\mathrm{RE}}(s) \ge \kappa_{\mathrm{RE}}(2s)$. □

References

[1] Belloni, A. and Chernozhukov, V. (2011). High dimensional sparse econometric models: an introduction. In: *Inverse Problems and High Dimensional Estimation, Stats in the Château 2009*, (Alquier, P., E. Gautier and G. Stoltz, eds.). *Lecture Notes in Statistics* **203** 127–162. Springer, Berlin.

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

[3] Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data*. Springer, New-York.

[4] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *The Annals of Statistics* **35** 1674–1697.

[5] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169–194.

[6] Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics* **35** 2313–2404.

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

[8] Koltchinskii, V. (2009). Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828.

[9] Koltchinskii, V. (2011). Oracle inequalities in empirical risk minimization and sparse recovery problems. *École d'Été de Probabilités de Saint-Flour 2008*. *Lecture Notes in Mathematics* **2033**.

[10] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2** 90–102.
[11] Petrov, V. V. (1995). *Sums of Independent Random Variables.* Oxford University Press.
[12] Rosenbaum, M. and Tsybakov A. B. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics* **38** 2620–2651.