The Annals of Statistics
2010, Vol. 38, No. 5, 2620–2651
DOI: 10.1214/10-AOS793

SPARSE RECOVERY UNDER MATRIX UNCERTAINTY

BY MATHIEU ROSENBAUM AND ALEXANDRE B. TSYBAKOV[1]

CMAP-École Polytechnique Paris, CREST and LPMA-University of Paris 6

We consider the model

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

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 noise independent of Ξ, and θ^* is a vector of unknown parameters to be estimated. The matrix uncertainty is in the fact that X is observed with additive error. For dimensions p that can be much larger than the sample size n, we consider the estimation of sparse vectors θ^*. Under matrix uncertainty, the Lasso and Dantzig selector turn out to be extremely unstable in recovering the sparsity pattern (i.e., of the set of nonzero components of θ^*), even if the noise level is very small. We suggest new estimators called *matrix uncertainty selectors* (or, shortly, the *MU-selectors*) which are close to θ^* in different norms and in the prediction risk if the restricted eigenvalue assumption on X is satisfied. We also show that under somewhat stronger assumptions, these estimators recover correctly the sparsity pattern.

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 noise independent of Ξ, and $\theta^* = (\theta_1^*, \dots, \theta_p^*)$ is a vector of unknown parameters to be estimated.

We will typically assume that θ^* is s-sparse, that is, that it has only s nonzero components, where $1 \le s \le p$ is some integer. The dimension p can be much larger than the sample size n, but we will typically have in mind the situation where the effective dimension s is much smaller than p and n. We will also assume that the elements of Ξ are small. In this setting we will suggest estimators

Received December 2008; revised October 2009.

[1]Supported in part by the Grant ANR-06-BLAN-0194 and by the PASCAL Network of Excellence.

AMS 2000 subject classifications. Primary 62J05; secondary 62F12.

Key words and phrases. Sparsity, MU-selector, matrix uncertainty, errors-in-variables model, measurement error, sign consistency, oracle inequalities, restricted eigenvalue assumption, missing data, portfolio selection, portfolio replication.

$\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$ that under some assumptions recover θ^* with high accuracy in different norms, as well as under the prediction risk. We will also show that, under somewhat stronger assumptions, these estimators recover correctly the sparsity pattern, that is, the set of nonzero components of θ^*. Our results follow the spirit of the now extensive literature on sparsity with ℓ_1-minimization (see, e.g., [1, 3–8, 11, 18–21, 23–28]). The main difference is in the presence of matrix uncertainty. The matrix X is not known and is observed with error. This leads us to new estimators, called *matrix uncertainty selectors* (or, shortly, the MU-selectors), which are different from the Lasso and Dantzig selector (or their modifications) studied in those papers.

In what follows, without loss of generality, we mainly assume that ξ and Ξ are deterministic and satisfy the assumptions

$$\left|\frac{1}{n} Z^T \xi\right|_\infty \le \varepsilon, \tag{3}$$

$$|\Xi|_\infty \le \delta \tag{4}$$

for some $\varepsilon \ge 0$, $\delta \ge 0$ [a modification of (4) is also used in some cases]. Here $|\cdot|_\infty$ stands for the maximum of components norm. If ξ and Ξ are random, conditions (3) and (4) can be guaranteed with a probability close to 1 under natural assumptions that we discuss below; we also indicate the corresponding values of ε and δ. So, the results that we prove for deterministic ξ and Ξ are extended in a trivial way to random ξ and Ξ satisfying these assumptions. The difference is only in the fact that the results hold on the random event of high probability where (3) and (4) are satisfied. The setting with random X is covered in a similar way. We only need to consider random X for which the restricted eigenvalue (RE) assumption or the Coherence assumption (see below) hold with high probability. Examples of such random X are discussed in the literature [8, 21].

We introduce two versions of MU-selectors. The first one is designed for the case $\xi = 0$, that is, for the problem of solving a large system of linear equations with deterministic or random noise in the matrix. This MU-selector is defined as a solution of the minimization problem

$$\min\{|\theta|_1 : \theta \in \Theta, |y - Z\theta|_\infty \le \delta |\theta|_1\},$$

where $\Theta \subseteq \mathbb{R}^p$ is a given set characterizing the prior knowledge about θ. Here and below $|x|_q$, $q \ge 1$, denotes the ℓ_q-norm of $x \in \mathbb{R}^d$ whatever is $d \ge 1$.

The second version of the MU-selector is defined as a solution of the minimization problem

$$\min\left\{|\theta|_1 : \theta \in \Theta, \left|\frac{1}{n} Z^T (y - Z\theta)\right|_\infty \le \lambda |\theta|_1 + \varepsilon\right\}, \tag{5}$$

where $\lambda \ge 0$ is a factor responsible for matrix uncertainty. If (4) is assumed, we choose λ depending on δ so that $\lambda > 0$ for $\delta > 0$ and $\lambda = 0$ for $\delta = 0$. Note that if

$\Theta = \mathbb{R}^p$ and there is no matrix uncertainty, that is, $\delta = 0$, this second MU-selector becomes the Dantzig selector of [8] based on the data (y, Z). We mainly discuss the choice $\lambda(\delta) = (1 + \delta)\delta$, which corresponds to a noise satisfying (4). This can be also used for $\xi = 0$ by setting $\varepsilon = 0$ in the definition. Nevertheless, for $\xi = 0$ we consider directly the first version of the MU-selector because it is simpler and achieves better error bounds than for $\lambda = (1 + \delta)\delta$.

Note that using in model (1) and (2) the Lasso or Dantzig selector with Z instead of the true X typically leads to satisfactory results for the prediction loss when the noise Ξ is small enough. However, these methods are less efficient in estimation of θ^* and they are especially unstable in selection of the sparsity pattern (cf. Section 7). In particular, they become quite sensitive to the values of θ^*. This is explained by the fact that the true θ^* is no longer guaranteed to stay, with a probability close to 1, in the feasible set of the Dantzig selector (which is also the set containing all the Lasso solutions).

The second MU-selector differs from the Dantzig selector based on the data (y, Z) in that we "penalize more" by enlarging the feasible band for $|\frac{1}{n}Z^T(y - Z\theta)|_\infty$. Indeed, setting $\Theta = \mathbb{R}^p$, a Lasso type analog of this MU-selector can be defined as a solution of the convex minimization problem

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n}|y - Z\theta|_2^2 + \lambda_1|\theta|_1 + \lambda_2|\theta|_1^2 \right\}$$

with some $\lambda_1, \lambda_2 > 0$. To appreciate why there is a similarity, note that for θ to achieve the minimum of such a convex criterion, it is necessary and sufficient to have

$$\begin{aligned} &\left(\frac{1}{n}Z^T(y - Z\theta)\right)_j = \frac{\lambda_1}{2} + \lambda_2|\theta|_1 \operatorname{sign}(\theta_j) \qquad \text{if } \theta_j \neq 0, \\ &\left|\left(\frac{1}{n}Z^T(y - Z\theta)\right)_j\right| \leq \frac{\lambda_1}{2} + \lambda_2|\theta|_1 \qquad \text{if } \theta_j = 0, \end{aligned} \tag{6}$$

where the index j designates the jth component of the corresponding vector and $\operatorname{sign}(\theta_j)$ is the sign of θ_j. Therefore, the set of possible solutions is "tightly" contained in $\{\theta \in \mathbb{R}^p : |\frac{1}{n}Z^T(y - Z\theta)|_\infty \leq \lambda_2|\theta|_1 + \lambda_1/2\}$, which is the feasible set of the MU-selector (5). The analogy is thus in the same spirit as between the Lasso and the Dantzig selector.

The results of this paper can be viewed in several perspectives. First, we can interpret them as a new approach to the inference in errors-in-variables models. The classical ways of treating these models via some versions of least squares or of the method of moments heavily depend on specific identifiability constraints that are violated when $p \gg n$ [12, 16]. Our approach is free of such constraints and requires only a modest price, which is the sparsity of the unknown vector of parameters. Also, on the difference from the results in the conventional errors-in-variables framework, we provide nonasymptotic bounds for the risks of the estimators and guarantee the finite sample variable selection property.

The second perspective is an extension of the theory of ℓ_1-based sparse recovery beyond the restricted isometry/restricted eigenvalue conditions (cf. [1, 8]) that are known to be too strong. We show that small perturbations of the design matrix X that bring these conditions to failure are in fact not so dangerous, once the method of recovery is chosen in a proper way (cf. Remark 4 below).

Finally, the third perspective is in developing simple and efficient tools of sparse recovery for specific applications. We mention here models with missing data, some financial models (portfolio selection, portfolio replication) and inverse problems with unknown operator. They are presented in the next section.

2. Examples of application. Here we explain how several examples of application can be described by model (1) and (2) with a sparse vector of parameters θ^*.

1. *Models with missing data.* Assume that the elements Z_{ij} of matrix Z satisfy

$$Z_{ij} = X_{ij}\eta_{ij}, \tag{7}$$

where X_{ij} are the elements of X and η_{ij} are i.i.d. Bernoulli random variables taking value 1 with probability $1-\pi$ and 0 with probability π, $0 < \pi < 1$. The data X_{ij} is missing if $\eta_{ij} = 0$, which happens with probability π. We are mainly interested in the case of small π. In practice, it is easy to estimate π by the empirical probability of occurrences of zeros in the sample of Z_{ij}, so it is realistic to assume that π is known. Note that we can rewrite (7) in the form

$$Z'_{ij} = X_{ij} + \xi'_{ij}, \tag{8}$$

where $Z'_{ij} = Z_{ij}/(1-\pi)$, $\xi'_{ij} = X_{ij}(\eta_{ij} - E(\eta_{ij}))/(1-\pi)$ and $E(\cdot)$ denotes the expectation. Thus, we can reduce the model with missing data (7) to the form (2) with matrix Ξ whose elements ξ'_{ij} are zero mean bounded random variables. In this case assumption (4) is fulfilled with δ which is not necessarily small, whereas the theoretical bounds obtained below only make sense if δ is small enough. Nevertheless, the variances of ξ'_{ij} are proportional to π, and we will see in Section 6 that, by modifying assumption (4), we obtain bounds for the MU-selector that are small if π is small.

2. *Portfolio selection.* Brodie et al. [2] recently argued that classical methods of portfolio selection are highly unstable. As a remedy, they proposed an algorithm accounting for the sparsity of portfolio weights and studied its numerical performance. A different approach to sparse portfolio selection can be introduced in our framework. Recall that in the traditional Markowitz portfolio selection, the objective is to find a portfolio having minimal variance return for a given expected return. This is stated as the optimization problem

$$\min\Big\{\theta^T X\theta : \theta^T\mu = \beta, \sum_j \theta_j = 1,\ j = 1,\dots,p\Big\}, \tag{9}$$

where $\theta = (\theta_1, \dots, \theta_p)$ is the vector of weights with θ_j representing the proportion of capital invested in the jth asset, X and μ are the covariance matrix and the vector of expected returns of the different assets and β is the desired return of the portfolio.

Using Lagrange multipliers, problem (9) is reduced to the solution of the linear equation $X\theta = a$ for some vector $a \in \mathbb{R}^p$ depending on μ and X. However, neither the covariance matrix X, nor the mean μ are available. Only their empirical (noisy) versions are observed. Instead of X we have a sample covariance matrix Z, and instead of a a vector of noisy observations y. Thus, we are in the framework of model (1) and (2) with $n = p$ (since X is a square matrix). Direct substitution of noisy values Z and y instead of X and a leads, in general, to instability of the solution of the linear equation because the dimension p can be very high (often 500 assets or more) and X can be either degenerate or with a small minimal eigenvalue. The methods that we suggest below are robust to the variations both of the matrix X and of the right-hand side a.

Another way of looking at sparse portfolio selection is to revise the very problem (9). Note that minimizing $\theta^T X\theta$, where X is the covariance matrix, is motivated by the fact that we would like to get the portfolio with smallest "dispersion." This requirement looks quite natural as long as we remain in the world of the classical second order statistics reasoning. The problem (9) is similar in spirit to "minimal variance unbiased estimation," an old concept which is known to have serious drawbacks. An alternative method would be to look for the *sparsest portfolio with a given daily return* β (we can also consider weakly or monthly returns). The problem can be formalized as follows. Let X_{ij} be the return of the jth asset on day i. The matrix of returns $X = (X_{ij})_{i,j}$ is typically observed with measurement error. This error can be due to an incomplete description of the assets. For example, the only available quantities for the investor are often reduced to the open, high, low and close prices, which leads, in particular, to asynchronous data (especially when one deals with prices from markets belonging to different time zones) and to an underestimation of the order book effects. Indeed, to take into account the liquidity costs, an investor should compute the returns, having in mind the order of magnitude of the number of assets he may have in his portfolio. However, such an accurate computation is only possible for the very few investors having access to order book data. In the case of nonstandard assets such as hedge funds, the measurement error can be also due to uncertainty about the management costs, the rounding approximations used and the way the returns are computed. Thus, instead of X, we in fact observe some other matrix $Z = (Z_{ij})_{i,j}$.

We are looking for the sparsest portfolio, that is, a portfolio that solves the problem

$$\min\Big\{|\theta|_0 : X\theta = \mathbf{b}, \sum_j \theta_j = 1, j = 1, \dots, p\Big\}, \tag{10}$$

where θ is the vector of the proportions of the wealth invested in each asset, $|\theta|_0$ is the number of nonzero components of θ and $\mathbf{b} \in \mathbb{R}^n$ is the vector with all the components equal to β. It is important to note that the sparsest portfolio does not necessarily contain a very small number of assets, in particular, when p is large. The minimization problem (10) is NP-hard, and the standard way to approximate it is to consider its convex relaxation:

$$\min\left\{|\theta|_1 : X\theta = \mathbf{b}, \sum_j \theta_j = 1, j = 1, \ldots, p\right\}. \tag{11}$$

This problem is already numerically solvable, but since X is observed with error, the solution can be unstable. We do not necessarily recover the sparsest solution if we directly plug Z instead of X in (11). A stable alternative that we suggest below is given by solving

$$\min\left\{|\theta|_1 : |\mathbf{b} - Z\theta|_\infty \le \delta|\theta|_1, \sum_j \theta_j = 1, j = 1, \ldots, p\right\}, \tag{12}$$

where δ is an upper bound on the noise level in the matrix X.

3. *Portfolio replication.* Replicating a portfolio, or at least finding the type of assets in a portfolio, has become a very challenging issue in the recent years, especially in the hedge funds context. Indeed, replicating a hedge fund portfolio means obtaining a Profit and Loss profile similar to those of the hedge fund without investing in it (and so avoiding the usual drawbacks of a hedge fund investment such as excessive fees, lack of transparency, lack of liquidity, lack of capacity, etc.). Replicating a portfolio can be done by retrieving the assets belonging to the portfolio. This problem can be formalized through model (1) and (2).

To fix ideas, suppose, for example, that we observe the daily returns y_i, $i = 1, \ldots, T$, of a portfolio. Moreover, assume that the proportion of capital invested in each asset of the portfolio is constant between day 1 and day T. Then, we theoretically have

$$y_i = \sum_{j=1}^{p} \theta_j X_{ij},$$

where p is the total number of different assets in the portfolio, X_{ij} is the return of the jth asset belonging to the portfolio on day i and θ_j the proportion of capital invested in it. As pointed out in the previous example, it is natural to consider that the vector of the portfolio returns $(y_i)_i$ and the matrix of the assets returns $X = (X_{ij})_{i,j}$ are observed with measurement error. Note that in this setup we can also treat the case where y_i and X_{ij} are the absolute returns (differences between the close price and the open price), provided that we define θ_j as the (constant) quantity of the jth asset in the portfolio.

To solve our problem, we could formally consider that any existing asset or derivative can, in principle, belong to the portfolio. This is of course not realistic.

However, it is reasonable to assume that the portfolio is rather sparse and that any asset used in the portfolio has a behavior which is quite close to those of an asset belonging to a restricted, given class of reference assets, especially if this restricted class can still be very large. For example, we will not put all the oil companies in the world in our restricted class. Nevertheless, we suppose that if an oil company is used in the portfolio and is not in the class, its returns profile will look like the returns profile of another oil company which belongs to the restricted class. Consequently, Z will be made from the daily returns of our reference assets. Indeed, for any asset in the portfolio, we will assume that either it belongs to the assets defining Z or it "resembles" one of the assets defining Z. Eventually, Z can be seen as a noisy measurement of X and, thus, the problem is described by model (1) and (2). A numerical illustration is given in Section 7.

4. *Inverse problems with unknown operator.* This setting has been recently discussed by several authors [9, 10, 15, 17, 22]. A typical problem is to recover an unknown function f that belongs to a Hilbert space H based on a noisy observation Y of Af where $A : H \to V$ is a linear operator and V is another Hilbert space. The observation Y with values in V can be written as

$$Y = Af + \zeta, \tag{13}$$

where ζ is a random variable (typically assumed Gaussian) with values in V. Let $\{\phi_j\}_{j=1}^{\infty}$ and $\{\psi_j\}_{j=1}^{\infty}$ be complete orthonormal bases in H and V, respectively. We can write $f = \sum_{j=1}^{\infty} \theta_j^* \phi_j = \sum_{j=1}^{p} \theta_j^* \phi_j + r$ with some coefficients θ_j^*, where the integer p is chosen very large, so that one can consider the remainder term $r \in H$ as negligible. Therefore, we can reduce the problem of estimating f to that of recovering the vector of coefficients $\theta^* = (\theta_1^*, \dots, \theta_p^*)$. Introducing the scalar products $Y_i = (Y, \psi_i)$ and $\xi_i = (\zeta, \psi_i)$, we obtain from (13) the following sequence of real-valued observations:

$$Y_i = \sum_{j=1}^{p} \theta_j^* (A\phi_j, \psi_i) + (Ar, \psi_i) + \xi_i, \qquad i = 1, 2, \dots.$$

If we consider here only the first n observations, assume that $(Ar, \psi_i) = 0$ and define the matrix $X = ((A\phi_j, \psi_i)_{i=1,\dots,n,\, j=1,\dots,p})$, and the vectors $y = (Y_1, \dots, Y_n)$, $\xi = (\xi_1, \dots, \xi_n)$, then we get the linear model (1). As discussed in [9, 10, 15, 17, 22], it is rather frequent in the applications that the operator A is not known, but its action on any given function in H can be observed with some noise. We emphasize that in those papers the noise Ξ is supposed to be small. This is consistent with the strategy of performing many repeated measurements of $(A\phi_j, \psi_i)$ for each pair (i, j). Thus, we have access to observations of the matrix $X = ((A\phi_j, \psi_i)_{i=1,\dots,n,\, j=1,\dots,p})$ with some small noise and, therefore, we are in the framework of model (1) and (2). The results obtained in [9, 10, 15, 17, 22] consider the case $n = p$ and deal with nondegenerate matrices X. This framework is not always convenient, especially if n and p are very large. The approach that

we develop in this paper is more general in the sense that, for example, if $n = p$ we can treat degenerate matrices X that satisfy some regularity assumptions. We also cover the case $p \gg n$, which is a useful extension because by taking a large p we can assure that the residual r is indeed negligible.

3. Sparse solution of linear equations with noisy matrix. In this section we consider the simplest case, $\xi = 0$. Thus, we solve the system of linear equations

$$y = X\theta,$$

where X is an unknown matrix such that we can observe its noisy values

$$Z = X + \Xi,$$

where Ξ satisfies (4).

Let Θ be a given convex subset of $\mathbb{R}^p$. We will assume in this section that there exists an s-sparse solution θ_s of $y = X\theta$ such that $\theta_s \in \Theta$. Consider the estimator $\hat{\theta}$ of θ_s defined as a solution of the following minimization problem:

$$\min\{|\theta|_1 : \theta \in \Theta, |y - Z\theta|_\infty \le \delta|\theta|_1\}. \tag{14}$$

Clearly, (14) is a convex minimization problem. If $\Theta = \mathbb{R}^p$ or if Θ is a linear subspace of $\mathbb{R}^p$ or a simplex (the latter case is interesting, e.g., in the context of portfolio selection), then (14) reduces to a linear programming problem.

Note that under assumption (4) the feasible set of problem (14)

$$\Theta_1 = \{\theta \in \Theta : |y - Z\theta|_\infty \le \delta|\theta|_1\}$$

is nonempty. In fact, $\theta_s \in \Theta_1$ since

$$|y - Z\theta_s|_\infty = |\Xi\theta_s|_\infty \le |\Xi|_\infty|\theta_s|_1 \le \delta|\theta_s|_1. \tag{15}$$

Thus, there always exists a solution $\hat{\theta}$ of (14). But it is not necessarily unique. We will call solutions of (14) the matrix uncertainty selectors (or, shortly, MU-selectors).

To state our assumptions on X, we need some notation. For a vector $\theta \in \mathbb{R}^p$ and a subset J of $\{1, \dots, p\}$, we denote by θ_J the vector in $\mathbb{R}^p$ that has the same coordinates as θ on the set of indices J and zero coordinates on its complement J^c.

We will assume that the matrix X satisfies the following condition (*restricted eigenvalue assumption* [1]):

ASSUMPTION RE(s). There exists $\kappa > 0$ such that

$$\min_{\Delta \ne 0 : |\Delta_{J^c}|_1 \le |\Delta_J|_1} \frac{|X\Delta|_2}{\sqrt{n}|\Delta_J|_2} \ge \kappa$$

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

A detailed discussion of this assumption can be found in [1]. In particular, it is shown in [1] that the restricted eigenvalue assumption is more general than several other similar assumptions used in the sparsity literature [8, 11, 25]. One of such assumptions is the *coherence condition* [11] that has the following form.

ASSUMPTION C. All the diagonal elements of the matrix $\Psi = X^T X/n$ are equal to 1 and all its off-diagonal elements $\Psi_{ij}, i \neq j$, satisfy the coherence condition: $\max_{i\neq j}|\Psi_{ij}| \leq \rho$ with some $\rho < 1$.

Note that Assumption C with $\rho < (3\alpha s)^{-1}$ implies Assumption RE(s) with $\kappa = \sqrt{1 - 1/\alpha}$ (cf. [1] or Lemma 2 in [18]).

We now state the main result of this section.

THEOREM 1. *Assume that there exists an s-sparse solution $\theta_s \in \Theta$ of the equation $y = X\theta$. Let* (4) *hold. Then for any solution $\hat{\theta}$ of* (14) *we have the following inequalities*:

(i)

$$\frac{1}{n}|X(\hat{\theta} - \theta_s)|_2^2 \leq 4\delta^2|\hat{\theta}|_1^2. \tag{16}$$

(ii) *If Assumption* RE(s) *holds, then*

$$|\hat{\theta} - \theta_s|_1 \leq \frac{4\sqrt{s}\delta}{\kappa}|\hat{\theta}|_1. \tag{17}$$

(iii) *If Assumption* RE($2s$) *holds, then*

$$|\hat{\theta} - \theta_s|_2 \leq \frac{4\delta}{\kappa}|\hat{\theta}|_1. \tag{18}$$

(iv) *If Assumption* C *holds with $\rho < \frac{1}{3\alpha s}$, $\alpha > 1$, then*

$$|\hat{\theta} - \theta_s|_\infty < 2\biggl(1 + \frac{2}{3\sqrt{s\alpha(\alpha - 1)}}\biggr)\delta|\hat{\theta}|_1. \tag{19}$$

PROOF. Set $\Delta = \hat{\theta} - \theta_s$ and $J = J(\theta_s)$, where $J(\theta)$ denotes the set of nonzero coordinates of θ. Note that

$$\begin{aligned} |X\Delta|_2 &= |Z\hat{\theta} - y - \Xi\hat{\theta}|_2 \\ &\leq \sqrt{n}(|Z\hat{\theta} - y|_\infty + |\Xi\hat{\theta}|_\infty) \\ &\leq \sqrt{n}(\delta|\hat{\theta}|_1 + |\Xi|_\infty|\hat{\theta}|_1) \\ &\leq 2\delta\sqrt{n}|\hat{\theta}|_1, \end{aligned} \tag{20}$$

which proves (16).

Next, by the standard argument (cf. Lemma 1 below or, e.g., [1, 8]) we have

$$|\Delta_{J^c}|_1 \le |\Delta_J|_1.$$

Thus,

$$|\hat{\theta} - \theta_s|_1 \le 2|\Delta_J|_1 \le 2\sqrt{s}|\Delta_J|_2 \le \frac{2\sqrt{s}}{\kappa\sqrt{n}}|X\Delta|_2, \tag{21}$$

where the last inequality follows from Assumption RE(s). Combining (20) and (21), we get (17).

To prove (18), we introduce the set of indices J_1 corresponding to those s coordinates of Δ outside $J = J(\theta_s)$ which are largest in absolute value (we assume without loss of generality that $2s \le p$). Define $J_{01} = J \cup J_1$. By a simple argument that does not use any assumption (cf., e.g., [1, 8] and the papers cited therein), we get

$$|\Delta_{J_{01}^c}|_2 \le \frac{|\Delta_{J^c}|_1}{\sqrt{s}}. \tag{22}$$

Thus,

$$|\Delta_{J_{01}^c}|_2 \le \frac{|\Delta_J|_1}{\sqrt{s}} \le |\Delta_J|_2 \le |\Delta_{J_{01}}|_2,$$

so that

$$|\Delta|_2 \le 2|\Delta_{J_{01}}|_2.$$

Now, by Assumption RE($2s$) and (20),

$$|\Delta_{J_{01}}|_2 \le \frac{1}{\kappa\sqrt{n}}|X\Delta|_2 \le \frac{2\delta}{\kappa}|\hat{\theta}|_1$$

and, hence, (18) follows.

We finally prove (19). Note first that

$$\begin{aligned} |\Psi(\hat{\theta} - \theta_s)|_\infty &\equiv \frac{1}{n}|X^T X(\hat{\theta} - \theta_s)|_\infty \\ &= \frac{1}{n}\max_{1\le j\le p}|\mathbf{x}_{(j)}^T X(\hat{\theta} - \theta_s)| \\ &\le \frac{1}{n}|X(\hat{\theta} - \theta_s)|_2 \max_{1\le j\le p}|\mathbf{x}_{(j)}|_2 = \frac{1}{\sqrt{n}}|X(\hat{\theta} - \theta_s)|_2, \end{aligned} \tag{23}$$

where $\mathbf{x}_{(j)}$ denotes the jth column of X and the last equality uses the fact that $|\mathbf{x}_{(j)}|_2 = \sqrt{n}$ since $|\mathbf{x}_{(j)}|_2^2/n$ are the diagonal elements of $X^T X/n$. Therefore, by (20),

$$|\Psi(\hat{\theta} - \theta_s)|_\infty \le 2\delta|\hat{\theta}|_1. \tag{24}$$

Now, since the jth component of $\Psi(\hat{\theta} - \theta_s)$ is

$$\big(\Psi(\hat{\theta} - \theta_s)\big)_j = (\hat{\theta}_j - \theta_{sj}) + \sum_{i=1, i\neq j}^{p} \Psi_{ij}(\hat{\theta}_i - \theta_{si}),$$

where θ_{si} is the ith component of θ_s, we obtain

$$|\hat{\theta} - \theta_s|_\infty \le 2\delta|\hat{\theta}|_1 + \rho|\hat{\theta} - \theta_s|_1. \tag{25}$$

Recall that Assumption C with $\rho < (3\alpha s)^{-1}$ implies Assumption RE(s) with $\kappa = \sqrt{1 - 1/\alpha}$ (cf. [1] or Lemma 2 in [18]). Thus, we can apply (17) with this value of κ to bound $|\hat{\theta} - \theta_s|_1$ in (25), which finally yields (19). □

REMARK 1. We can replace $|\hat{\theta}|_1$ by $|\theta_s|_1$ in all the inequalities of Theorem 1.

REMARK 2. It is straightforward to deduce a bound for $|\hat{\theta} - \theta_s|_q$ for any $1 \le q \le 2$ from the bounds (17) and (18), as it is done, for example, in [1].

Under the assumptions of part (iv) of Theorem 1, we get

$$|\hat{\theta} - \theta_s|_\infty < C_*(\alpha)\delta|\hat{\theta}|_1, \tag{26}$$

where $C_*(\alpha) = 2(1 + \frac{2}{3\sqrt{\alpha(\alpha-1)}})$ is a constant. Based on this, we can define the thresholded estimator $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)$, where

$$\tilde{\theta}_j = \hat{\theta}_j I\{|\hat{\theta}_j| > \tau\}, \qquad j = 1, \dots, p, \tag{27}$$

with the data-dependent threshold $\tau = C_*(\alpha)\delta|\hat{\theta}|_1$ for some $\alpha > 1$. Here $I\{\cdot\}$ denotes the indicator function. It is useful to note that since the MU-selector $\hat{\theta}$ is, in general, not unique, the thresholded estimator $\tilde{\theta}$ is also not necessarily unique.

Our next result shows that the thresholded estimator $\tilde{\theta}$ recovers the sparsity pattern and, moreover, it recovers the signs of the coordinates of s-sparse solution θ_s (this property is sometimes called the sign consistency; cf. [18, 23, 25, 27]). We define

$$\operatorname{sign}\theta = \begin{cases} -1, & \text{if } \theta < 0, \\ 0, & \text{if } \theta = 0, \\ 1, & \text{if } \theta > 0. \end{cases}$$

THEOREM 2. *Assume that $\theta_s \in \Theta$ is an s-sparse solution of $y = X\theta$, and that $\Theta \subseteq \{\theta \in \mathbb{R}^p : |\theta|_1 \le a\}$ for some $a > 0$. Let (4) and Assumption C hold with $\rho < (3\alpha s)^{-1}$ for some $\alpha > 1$. If*

$$\min_{j \in J(\theta_s)} |\theta_{sj}| > C_*(\alpha)\delta a, \tag{28}$$

then

$$\operatorname{sign}\tilde{\theta}_j = \operatorname{sign}\theta_{sj}, \qquad j = 1, \dots, p, \tag{29}$$

for all $\tilde{\theta}_j$ in (27) such that $\hat{\theta}$ is an MU-selector defined in (14).

PROOF. For $j \notin J(\theta_s)$ we have $\theta_{sj} = 0$ and, thus, by (26), $|\hat{\theta}_j| = |\hat{\theta}_j - \theta_{sj}| < C_*(\alpha)\delta|\hat{\theta}|_1 = \tau$. Therefore, $\tilde{\theta}_j = 0$ for $j \notin J(\theta_s)$. For $j \in J(\theta_s)$ note that (26) implies $|\hat{\theta}_j - \theta_{sj}| < C_*(\alpha)\delta a$. This and assumption (28) yield that $\hat{\theta}_j$ has the same sign as θ_{sj}. $\square$

REMARK 3. Note that, under Assumption C with $\rho < (3\alpha s)^{-1}$ as required in Theorem 2, the s-sparse solution is unique; cf., for example, [18], page 93, so that the right-hand side of (29) is uniquely defined. The estimator $\tilde{\theta}$ is not necessarily unique, nevertheless, Theorem 2 assures that the sign recovery property (29) holds for all versions of $\tilde{\theta}$.

4. Sparse recovery for regression model with unknown design matrix. We consider now the general model (1) and (2) and assume that it holds with an s-sparse vector of unknown parameters $\theta^* = \theta_s \in \Theta$. Because of the presence of noise ξ that is typically not small, we need to change the definition of the MU-selector. We now define the MU-selector $\hat{\theta}$ as a solution of the minimization problem

$$\min\left\{|\theta|_1 : \theta \in \Theta, \left|\frac{1}{n}Z^T(y - Z\theta)\right|_\infty \le (1+\delta)\delta|\theta|_1 + \varepsilon\right\}. \tag{30}$$

Note that if $\delta = 0$ and $\Theta = \mathbb{R}^p$, this MU-selector becomes the Dantzig selector of [8].

Similarly to (14), the problem (30) is a convex minimization problem and it reduces to linear programming if $\Theta = \mathbb{R}^p$, Θ is a linear subspace of $\mathbb{R}^p$ or a simplex.

Throughout this section we will assume for simplicity that the matrix X is normalized, so that all the diagonal elements of the Gram matrix $\Psi = X^T X/n$ are equal to 1. Extensions to general matrices are straightforward, it only modifies the constants in the expression $(1+\delta)\delta|\theta|_1 + \varepsilon$ in (30) and in the theorems.

Note that under assumptions (3) and (4), the feasible set of the convex problem (30) is nonempty:

$$\Theta_2 \equiv \left\{\theta \in \Theta : \left|\frac{1}{n}Z^T(y - Z\theta)\right|_\infty \le (1+\delta)\delta|\theta|_1 + \varepsilon\right\} \ne \varnothing.$$

To prove this, let us show that the true vector $\theta^* = \theta_s$ belongs to Θ_2. In fact, by (3),

$$\begin{aligned}\left|\frac{1}{n}Z^T(y - Z\theta_s)\right|_\infty &= \left|\frac{1}{n}Z^T(X\theta_s + \xi - Z\theta_s)\right|_\infty \\ &\le \left|\frac{1}{n}Z^T\xi\right|_\infty + \left|\frac{1}{n}Z^T\Xi\theta_s\right|_\infty \\ &\le \varepsilon + \left|\frac{1}{n}Z^T\Xi\theta_s\right|_\infty.\end{aligned} \tag{31}$$

Next, note that, by (4) and by the fact that all the diagonal elements of $X^T X/n$ are equal to 1, the columns $\mathbf{z}_{(j)}$ of matrix Z satisfy $|\mathbf{z}_{(j)}|_2 \le \sqrt{n}(1+\delta)$. Therefore, arguing as in (23), we obtain

$$\left|\frac{1}{n} Z^T \Xi\theta_s\right|_\infty \le \frac{1+\delta}{\sqrt{n}} |\Xi\theta_s|_2 \le (1+\delta)|\Xi\theta_s|_\infty \le (1+\delta)\delta|\theta_s|_1. \tag{32}$$

This and (31) yield

$$\left|\frac{1}{n} Z^T (y - Z\theta_s)\right|_\infty \le (1+\delta)\delta|\theta_s|_1 + \varepsilon.$$

Since we also assume that θ_s belongs to Θ, the fact that $\theta_s \in \Theta_2$ is proved. Thus, there always exists a solution $\hat\theta$ of (30). Of course, it is not necessarily unique.

THEOREM 3. *Assume that model* (1) *and* (2) *holds with an* (*unknown*) s-*sparse parameter vector* $\theta^* = \theta_s \in \Theta$ *and that all the diagonal elements of* $X^T X/n$ *are equal to* 1. *Let* (3) *and* (4) *hold. Set*

$$\nu = 2(2+\delta)\delta|\theta_s|_1 + 2\varepsilon.$$

Then for any solution $\hat\theta$ *of* (30) *we have the following inequalities*:

(i) *Under Assumption* RE(s):

$$|\hat\theta - \theta_s|_1 \le \frac{4\nu s}{\kappa^2}, \tag{33}$$

$$\frac{1}{n}|X(\hat\theta - \theta_s)|_2^2 \le \frac{4\nu^2 s}{\kappa^2}. \tag{34}$$

(ii) *Under Assumption* RE($2s$):

$$|\hat\theta - \theta_s|_q^q \le \left(\frac{4\nu}{\kappa^2}\right)^q s \qquad \forall 1 < q \le 2. \tag{35}$$

(iii) *Under Assumption* C *with* $\rho < \frac{1}{3\alpha s}$, $\alpha > 1$:

$$|\hat\theta - \theta_s|_\infty < \frac{3\alpha+1}{3(\alpha-1)}\nu. \tag{36}$$

PROOF. Set $\Delta = \hat\theta - \theta_s$ and $J = J(\theta_s)$. Note first that (3) and the fact that $\hat\theta$ belongs to the feasible set Θ_2 of (30) imply

$$\begin{aligned}\left|\frac{1}{n} X^T X\Delta\right|_\infty &\le \left|\frac{1}{n} Z^T (y - Z\hat\theta)\right|_\infty + \left|\frac{1}{n}\Xi^T X(\hat\theta - \theta_s)\right|_\infty \\ &\quad + \left|\frac{1}{n} Z^T \xi\right|_\infty + \left|\frac{1}{n} Z^T \Xi\hat\theta\right|_\infty \\ &\le (1+\delta)\delta|\hat\theta|_1 + 2\varepsilon + \left|\frac{1}{n} Z^T \Xi\hat\theta\right|_\infty + \left|\frac{1}{n}\Xi^T X(\hat\theta - \theta_s)\right|_\infty.\end{aligned} \tag{37}$$

Now,

$$|\Xi^T X|_\infty = \max_{1\le j,k\le p} |\xi_{(j)}^T \mathbf{x}_{(k)}| \le \max_{1\le j,k\le p} |\xi_{(j)}|_2 |\mathbf{x}_{(k)}|_2 \le \delta n, \tag{38}$$

where $\xi_{(j)}$ are the columns of Ξ and we used that $|\mathbf{x}_{(k)}|_2 = \sqrt{n}$ by assumption on $X^T X/n$, and $|\xi_{(j)}|_2 \le \delta\sqrt{n}$ by (4). This implies

$$\left|\frac{1}{n}\Xi^T X(\hat\theta - \theta_s)\right|_\infty \le |\hat\theta - \theta_s|_1 \left|\frac{1}{n}\Xi^T X\right|_\infty \le \delta|\hat\theta - \theta_s|_1. \tag{39}$$

Next, as in (32), we obtain

$$\left|\frac{1}{n} Z^T \Xi\hat\theta\right|_\infty \le (1+\delta)\delta|\hat\theta|_1. \tag{40}$$

We now combine (37), (39) and (40) to get

$$\left|\frac{1}{n} X^T X\Delta\right|_\infty \le 2\varepsilon + 2(1+\delta)\delta|\hat\theta|_1 + \delta|\hat\theta - \theta_s|_1 \le \nu. \tag{41}$$

Taking into account (41), the proof of (33), (34) and (35) follows the same lines as the proof of Theorem 7.1 in [1] where we should set $r = \nu/2$, $m = s$.

We now prove (36). We proceed as in the proof of (19) in Theorem 1, with the only difference that now we replace (24) by (41). Thus, instead of (25), we obtain

$$|\hat\theta - \theta_s|_\infty \le \nu + \rho|\hat\theta - \theta_s|_1. \tag{42}$$

Next, recall that Assumption C with $\rho < (3\alpha s)^{-1}$ implies Assumption RE(s) with $\kappa^2 = 1 - 1/\alpha$ (cf. [1] or Lemma 2 in [18]). Using in (42) the bound (35) with $q = 1$ and $\kappa^2 = 1 - 1/\alpha$, we obtain (36). This finishes the proof of the theorem. □

Note that, in contrast to Theorem 1, the bounds of Theorem 3 do not depend on $|\hat\theta|_1$ but on the unknown $|\theta_s|_1$ (cf. definition of ν). This drawback can be corrected for small values of δ, as shown in the next result.

THEOREM 4. *Let the assumptions of Theorem 3 hold and* $\delta < \frac{\kappa^2}{4s}$. *Set*

$$\nu_1 = 2(1+\delta)\delta|\hat\theta|_1 + 2\varepsilon.$$

Then for any solution $\hat\theta$ *of* (30) *we have the following inequalities*:

(i) *Under Assumption* RE(s):

$$|\hat\theta - \theta_s|_1 \le \frac{4\nu_1 s}{\kappa^2}\left(1 - \frac{4\delta s}{\kappa^2}\right)^{-1}, \tag{43}$$

$$\frac{1}{n}|X(\hat\theta - \theta_s)|_2^2 \le \frac{4\nu_1^2 s}{\kappa^2}\left(1 - \frac{4\delta s}{\kappa^2}\right)^{-2}. \tag{44}$$

(ii) *Under Assumption* RE(2s):

$$|\hat{\theta} - \theta_s|_q^q \le \left(\frac{4\nu_1}{\kappa^2}\right)^q \left(1 - \frac{4\delta s}{\kappa^2}\right)^{-q} s \qquad \forall 1 < q \le 2. \tag{45}$$

(iii) *Under Assumption* C *with* $\rho < \frac{1}{3\alpha s}$, $\alpha > 1$, *and* $\delta \le \frac{\kappa^2}{8s}$:

$$|\hat{\theta} - \theta_s|_\infty < \frac{2(3\alpha+1)}{3(\alpha-1)}\nu_1. \tag{46}$$

PROOF. We use the same notation as in the proof of Theorem 3. From (41) and the fact that $|\Delta_{J^c}|_1 \le |\Delta_J|_1$, we obtain

$$\begin{aligned} \frac{1}{n}|X\Delta|_2^2 &\le |\Delta|_1 \left|\frac{1}{n}X^T X\Delta\right|_\infty \\ &\le |\Delta|_1(\nu_1 + \delta|\Delta|_1) \\ &\le 2|\Delta_J|_1(\nu_1 + 2\delta|\Delta_J|_1) \\ &\le 2\sqrt{s}|\Delta_J|_2\big(\nu_1 + 2\sqrt{s}\delta|\Delta_J|_2\big). \end{aligned} \tag{47}$$

Similar arguments as for (21) easily yield the inequality

$$|\Delta_J|_2 \le \frac{2\sqrt{s}\nu_1}{\kappa^2}\left(1 - \frac{4\delta s}{\kappa^2}\right)^{-1} \tag{48}$$

and (43). In the same way, (45) deduces from (48) following the analogous part of the proof of Theorem 7.1 in [1] where we should set $r = (\nu_1/2)(1 - 4\delta s/\kappa^2)^{-1}$ and $m = s$.

Finally, to get the sup-norm inequality (46), we proceed as in the proof of (19) in Theorem 1 or in that of (36) in Theorem 3, with the only difference that instead of (25) we use the bound

$$|\hat{\theta} - \theta_s|_\infty \le \nu_1 + (\rho + \delta)|\hat{\theta} - \theta_s|_1$$

that follows from (41) and the fact that $\delta \le \frac{\kappa^2}{8s}$. This finishes the proof of Theorem 4. □

As in Section 3, we now define a thresholded estimator $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)$ by the formula

$$\tilde{\theta}_j = \hat{\theta}_j I\{|\hat{\theta}_j| > \tau_1\}, \qquad j = 1, \dots, p, \tag{49}$$

where the threshold is given either by

$$\tau_1 = \frac{3\alpha+1}{3(\alpha-1)}\big(2\varepsilon + 2(2+\delta)\delta a\big) \tag{50}$$

for $\alpha > 1$, $a > 0$, or by

$$\tau_1 = \frac{2(3\alpha+1)}{3(\alpha-1)}\big(2\varepsilon + 2(1+\delta)\delta|\hat{\theta}|_1\big) \tag{51}$$

for $\alpha > 1$. Note that the threshold (51) is completely data-driven if ε and δ are known.

The next theorem shows that under some assumptions the thresholded estimator defined in (49) recovers the sparsity pattern and, moreover, it recovers the signs of the coordinates of the s-sparse solution θ_s.

THEOREM 5. *Assume that model* (1) *and* (2) *holds with the s-sparse vector of unknown parameters $\theta^* = \theta_s \in \Theta$ and that* (3), (4) *and Assumption* C *hold with $\rho < (3\alpha s)^{-1}$ for some $\alpha > 1$. Let either $\Theta \subseteq \{\theta \in \mathbb{R}^p : |\theta|_1 \le a\}$ for some $a > 0$ and the threshold τ_1 is given by* (50), *or $\delta \le \frac{\kappa^2}{8s}$ and the threshold τ_1 is given by* (51). *If*

$$\min_{j \in J(\theta_s)} |\theta_{sj}| > \tau_1, \tag{52}$$

then

$$\operatorname{sign} \tilde{\theta}_j = \operatorname{sign} \theta_{sj}, \qquad j = 1, \dots, p, \tag{53}$$

for all $\tilde{\theta}_j$ in (49) *where $\hat{\theta}$ is a MU-selector defined in* (30).

PROOF. It goes along the same lines as the proof of Theorem 2.

We can make here the same remarks as in Section 3 about the nonuniqueness of the estimators. Indeed, $\tilde{\theta}$ is not necessarily unique, but Theorem 5 assures the sign recovery property (53) holds for all versions of $\tilde{\theta}$. □

REMARK 4. The argument of this section can be applied with minor modifications to the model

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

This is no longer the errors-in-variables setting, but just the usual regression setting where X is some "nominal" design matrix and Ξ can be viewed as its perturbation. The results of this section suggest that small perturbations of the design matrix X beyond the restricted eigenvalue condition are in fact not so dangerous, once the method of recovery is chosen in a proper way. Indeed, such perturbations lead to the extra terms in the bounds proportional to the ℓ_1-norm of the solution. Roughly speaking, our bounds suggest that the MU-selector is robust with respect to possible violations of the the restricted eigenvalue condition, provided that the perturbations are small enough and the ℓ_1-norm of the true θ is reasonably bounded.

This offers a possible way of relaxing the strong conditions usually imposed in the context of ℓ_1-penalized sparse estimation. Note that another way to do it can be found in [13, 14], suggesting a computationally feasible method of sparse estimation with no assumption on X. However, the oracle inequalities of [13, 14] hold only for the prediction risk.

5. Approximately s-sparse solutions. The results of the previous sections can be easily generalized to the setting where the true θ^* is arbitrary, not necessarily s-sparse. This might be of interest in the context of inverse problems with unknown operator, as discussed in the Introduction. Then the bounds will involve a residual term, which is a difference between θ^* and its s-sparse approximation θ_s. In particular, we can take θ_s as the best s-sparse approximation of θ^*, that is, the vector that coincides with θ^* in the s coordinates with largest absolute values and has other coordinates that vanish.

We will use the following slightly strengthened version of Assumption RE(s), where we only increase a numerical constant in the definition of the set over which the minimum is taken (cf. [1]).

ASSUMPTION RE$(s, 2)$. There exists $\kappa > 0$ such that

$$\min_{\Delta \neq 0 \colon |\Delta_{J^c}|_1 \le 2|\Delta_J|_1} \frac{|X\Delta|_2}{\sqrt{n}|\Delta_J|_2} \ge \kappa$$

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

It is easy to check that Assumption C with $\rho < \frac{1}{5\alpha s}$ for some $\alpha > 1$ implies Assumption RE$(s, 2)$ with $\kappa^2 = 1 - 1/\alpha$ (cf. [1]).

We now state the main result of this section.

THEOREM 6. *Assume that there exists a solution $\theta^* \in \Theta$ of the equation $y = X\theta$. Let (4) hold. Then for any solution $\hat{\theta}$ of (14) we have the following inequalities*:

(i)

$$\frac{1}{n}|X(\hat{\theta} - \theta^*)|_2^2 \le 4\delta^2 |\hat{\theta}|_1^2. \tag{54}$$

(ii) *If Assumption* RE$(s, 2)$ *holds, then*

$$|\hat{\theta} - \theta^*|_1 \le \frac{4\sqrt{s}\delta}{\kappa}|\hat{\theta}|_1 + 6 \min_{J \colon |J| \le s} |\theta^*_{J^c}|_1. \tag{55}$$

(iii) *If Assumption* C *holds with $\rho < \frac{1}{5\alpha s}$, $\alpha > 1$, then*

$$|\hat{\theta} - \theta^*|_\infty < 2\left(1 + \frac{2}{5\sqrt{s\alpha(\alpha - 1)}}\right)\delta|\hat{\theta}|_1 + \frac{6}{5\alpha s} \min_{J \colon |J| \le s} |\theta^*_{J^c}|_1. \tag{56}$$

PROOF. Set $\Delta = \hat{\theta} - \theta^*$ and let $J \subset \{1, \dots, p\}$ be an arbitrary set of indices such that $|J| \le s$. First, note that (54) is already proved in Theorem 1, since (20) is valid with $\Delta = \hat{\theta} - \theta^*$.

We will use the following elementary fact (cf., e.g., [5–8]) that we state for convenience as a lemma.

LEMMA 1. *Let $\hat{\theta}$ be a solution of the problem*

$$\min\{|\theta|_1 : \theta \in \Theta'\},$$

where Θ' is a subset of $\mathbb{R}^p$. Let θ^* be any element of Θ' and J any subset of $\{1, \dots, p\}$. Then for $\Delta = \hat{\theta} - \theta^*$ we have

$$|\Delta_{J^c}|_1 \le |\Delta_J|_1 + 2|\theta^*_{J^c}|_1. \tag{57}$$

PROOF.

$$\begin{aligned} |\theta^*_J|_1 + |\theta^*_{J^c}|_1 = |\theta^*|_1 &\ge |\hat{\theta}|_1 = |\hat{\theta}_J|_1 + |\hat{\theta}_{J^c}|_1 \\ &= |\Delta_J + \theta^*_J|_1 + |\Delta_{J^c} + \theta^*_{J^c}|_1 \\ &\ge |\theta^*_J|_1 - |\Delta_J|_1 + |\Delta_{J^c}|_1 - |\theta^*_{J^c}|_1. \end{aligned}$$

□

To prove (55), consider separately the following two cases: (a) $2|\theta^*_{J^c}|_1 \le |\Delta_J|_1$ and (b) $2|\theta^*_{J^c}|_1 > |\Delta_J|_1$. In case (a) we use (57) to obtain $|\Delta_{J^c}|_1 \le 2|\Delta_J|_1$. Therefore, by Assumption RE(s, 2) and (54),

$$|\Delta_J|_2 \le \frac{1}{\kappa\sqrt{n}}|X\Delta|_2 \le \frac{2\delta}{\kappa}|\hat{\theta}|_1.$$

This and (57) imply that, in case (a),

$$\begin{aligned} |\Delta|_1 &\le 2|\Delta_J|_1 + 2|\theta^*_{J^c}|_1 \le 2\sqrt{s}|\Delta_J|_2 + 2|\theta^*_{J^c}|_1 \\ &\le \frac{4\sqrt{s}\delta}{\kappa}|\hat{\theta}|_1 + 2|\theta^*_{J^c}|_1. \end{aligned} \tag{58}$$

In case (b) we immediately deduce from (57) that $|\Delta|_1 \le 6|\theta^*_{J^c}|_1$. Combining this with (58), we obtain (55).

To prove (56), note that the argument leading to (25) is applicable here with θ^* in place of θ_s. Thus,

$$|\hat{\theta} - \theta^*|_\infty \le 2\delta|\hat{\theta}|_1 + \rho|\hat{\theta} - \theta^*|_1. \tag{59}$$

Now, as mentioned above, Assumption C with $\rho < \frac{1}{5\alpha s}$, $\alpha > 1$, implies Assumption RE(s, 2) with $\kappa^2 = 1 - 1/\alpha$. Using (55) with this value of κ to bound $|\hat{\theta} - \theta_s|_1$ in (59), we arrive at (56). This proves the theorem. □

Note that under Assumption C we can also bound the ℓ_2 norm of the difference $\hat{\theta} - \theta^*$, as well as all its ℓ_r norms with $r > 1$. However, Assumption C is rather

restrictive. For instance, it is not valid for Toeplitz matrices Ψ or for matrices X with independent standard Gaussian entries (for the latter case, Assumption RE is assured with overwhelming probability if s is of a smaller order than $n/\log p$). The next theorem shows that we can bound correctly the ℓ_2 norm $|\hat{\theta} - \theta^*|_2$ under the following condition which is weaker than Assumption C but somewhat stronger than Assumption RE.

ASSUMPTION RE$'(s, 2)$. There exist $\kappa > 0$ and $c_1 > 0$ such that

$$\min_{\Delta \neq 0 \colon |\Delta_{J^c}|_1 \leq 2|\Delta_J|_1 + a} \frac{|X\Delta|_2^2/n + c_1 a |\Delta_J|_2/\sqrt{s}}{|\Delta_J|_2^2} \geq \kappa^2 \tag{60}$$

for all $a \geq 0$ and all subsets J of $\{1, \dots, p\}$ of cardinality $|J| \leq s$.

Note that Assumption RE$(s, 2)$ is a special case of (60) corresponding to $a = 0$. Note also that Assumption RE$'(s, 2)$ is satisfied if the restricted isometry assumption [6–8] holds with the isometry coefficient close enough to 1. This is not hard to show following the lines of [5].

THEOREM 7. *Assume that there exists a solution $\theta^* \in \Theta$ of the equation $y = X\theta$. Let (4) and Assumption* RE$'(2s, 2)$ *hold. Then for any solution $\hat{\theta}$ of (14) we have*

$$|\hat{\theta} - \theta^*|_2 \leq \frac{4\delta}{\kappa} |\hat{\theta}|_1 + \left(4 + \frac{2\sqrt{c_1}}{\kappa}\right) \min_{J \colon |J| \leq s} \frac{|\theta^*_{J^c}|_1}{\sqrt{s}}. \tag{61}$$

PROOF. Set, as before, $\Delta = \hat{\theta} - \theta^*$ and let $J \subset \{1, \dots, p\}$ be an arbitrary set of indices such that $|J| \leq s$. We first note that (57), (22) and the fact that $|\Delta_J|_1 \leq \sqrt{s}|\Delta_J|_2$ imply

$$|\Delta_{J_{01}^c}|_2 \leq |\Delta_J|_2 + \frac{2}{\sqrt{s}} |\theta^*_{J^c}|_1. \tag{62}$$

Consider separately the cases $2|\theta^*_{J^c}|_1/\sqrt{s} \leq |\Delta_{J_{01}}|_2$ and $2|\theta^*_{J^c}|_1/\sqrt{s} > |\Delta_{J_{01}}|_2$.

(a) In the case $2|\theta^*_{J^c}|_1/\sqrt{s} \leq |\Delta_{J_{01}}|_2$ we have $|\Delta_{J_{01}^c}|_1 \leq 2|\Delta_{J_{01}}|_1$. Also, $|J_{01}| \leq 2s$ by the definition of J_{01}. Therefore, using Assumption RE$'(2s, 2)$ with $a = 0$ and (54), we get

$$|\Delta_{J_{01}}|_2 \leq \frac{1}{\kappa\sqrt{n}} |X\Delta|_2 \leq \frac{2\delta}{\kappa} |\hat{\theta}|_1.$$

This and (62) imply

$$|\Delta|_2 \leq |\Delta_{J_{01}}|_2 + |\Delta_J|_2 + \frac{2}{\sqrt{s}} |\theta^*_{J^c}|_1 \leq \frac{4\delta}{\kappa} |\hat{\theta}|_1 + \frac{2}{\sqrt{s}} |\theta^*_{J^c}|_1. \tag{63}$$

Thus, (61) is proved in the case $2|\theta^*_{J^c}|_1/\sqrt{s} \leq |\Delta_{J_{01}}|_2$.

(b) It remains to prove (61) in the case $2|\theta^*_{J^c}|_1/\sqrt{s} > |\Delta_{J_{01}}|_2$. This condition and (62) immediately yield

$$|\Delta_{J_{01}^c}|_2 \le \frac{4}{\sqrt{s}}|\theta^*_{J^c}|_1,$$

so that

$$|\Delta|_2 \le |\Delta_{J_{01}}|_2 + \frac{4}{\sqrt{s}}|\theta^*_{J^c}|_1. \tag{64}$$

Next, from (57) we easily get

$$|\Delta_{J_{01}^c}|_1 \le |\Delta_{J_{01}}|_1 + 2|\theta^*_{J^c}|_1.$$

Therefore, using Assumption RE′(2s, 2) with $a = 2|\theta^*_{J^c}|_1$ and (54), we find

$$\begin{aligned}\kappa^2|\Delta_{J_{01}}|_2^2 &\le \frac{1}{n}|X\Delta|_2^2 + 2c_1\frac{|\theta^*_{J^c}|_1|\Delta_{J_{01}}|_2}{\sqrt{2s}}\\ &\le 4\delta^2|\hat{\theta}|_1^2 + 2\sqrt{2}c_1\frac{|\theta^*_{J^c}|_1^2}{s},\end{aligned}$$

where we used that $2|\theta^*_{J^c}|_1/\sqrt{s} > |\Delta_{J_{01}}|_2$. The last display and (64) imply that (61) holds in the case $2|\theta^*_{J^c}|_1/\sqrt{s} > |\Delta_{J_{01}}|_2$. □

6. Random noise. If ξ and Ξ are random and conditions (3) and (4) are satisfied with a probability close to 1, then all the bounds in the above theorems remain valid with the same probability. This holds in different situations under natural assumptions that we briefly discuss in this section.

First, it is not hard to see that if ξ is normal with zero mean and covariance matrix $\sigma^2 I$ where I denotes the identity matrix, and we take

$$\varepsilon = A\sigma\sqrt{\frac{\log p}{n}} \tag{65}$$

for some $A > (1+\delta)\sqrt{2}$, then condition (3) holds with probability at least $1 - p^{1-A^2/2}$. If p is very large, this probability is very close to 1. A similar remark holds for sub-Gaussian ξ.

For more general ξ we can guarantee condition (3) only with a larger value of ε and with a probability that is not as close to 1 as in the Gaussian case. For example, if the components ξ_i of ξ are independent zero mean random variables with uniformly bounded variances, $E(\xi_i^2) \le \sigma^2 < \infty$, $i = 1, \dots, n$, and if the elements $X_{ij}, i = 1, \dots, n$, $j = 1, \dots, p$, of matrix X satisfy

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

for some constant c, then condition (3) holds with probability at least $1 - O(\frac{\log p}{\varepsilon^2 n})$ [18]. In particular, we can take

$$\varepsilon = A\sqrt{\frac{(\log p)^{1+\gamma}}{n}},$$

and then condition (3) holds with probability at least $1 - O((\log p)^{-\gamma})$.

For the choice of δ in condition (4) we can consider the examples related to portfolio selection and to inverse problems with unknown operator; cf. Section 2. In both examples we have repeated measurements. The matrix Z is either the average of several observed matrices with mean X, or the empirical covariance matrix, with X defined as the corresponding population covariance matrix (in the latter case $p = n$). Then the threshold δ in condition (4) can be determined in the same spirit as ε in condition (3). We omit further details.

Finally, consider the model with missing data discussed in Section 2. In this example direct application of condition (4) leads to bounds which are too loose. Indeed, δ can be of the order of $|X|_\infty$. However, we argue that the MU-selector of the form (5) with suitable λ still satisfies good bounds if the probability π that an entry of X is not observed remains small. This needs a refinement of our argument for the particular setting. We sketch it now. Note first that under the assumptions of Theorem 3 for a deterministic matrix X and for $Z_{ij} = X_{ij} + \xi'_{ij}$, where ξ'_{ij} are defined in Section 2, we have, with probability close to 1 when n is large,

$$\left|\frac{1}{n}\Xi^T X\right|_\infty \le \delta_1, \qquad \left|\frac{1}{n}X^T \Xi\right|_\infty \le \delta_1, \tag{66}$$

$$\left|\frac{1}{n}\bigl(\Xi^T \Xi - \operatorname{diag}(\Xi^T \Xi)\bigr)\right|_\infty \le \delta_2, \tag{67}$$

$$\left|\frac{1}{n}\operatorname{diag}(\Xi^T \Xi)\right|_\infty \le C\pi, \tag{68}$$

where Ξ is the matrix with entries ξ'_{ij}, $\operatorname{diag}(\Xi^T \Xi)$ denotes the diagonal matrix having the same diagonal elements as $\Xi^T \Xi$, $C > 0$ is a constant, and $\delta_1, \delta_2 > 0$ are small if n is large. Indeed, (66) and (67) follow from the standard properties of zero mean sub-Gaussian variables, while (68) is due to the fact that the expectations of the diagonal elements of $\frac{1}{n}\Xi^T \Xi$ are proportional to π.

We now observe that under assumptions (66) and (67) the constant $(1 + \delta)\delta$ in (32) can be replaced by $\delta_1 + \delta_2 + C\pi$. This motivates the use of the MU-selector (5) with $\lambda = \delta_1 + \delta_2 + C\pi$. For such an MU-selector we have an analog of Theorem 3 if we replace assumption (4) by assumptions (66) and (67). The only difference is in the form of ν which now becomes a linear combination of δ_1, δ_2 and π. This new value of ν is small for n large enough and small π. In conclusion, the MU-selector (5) with suitable λ achieves good theoretical bounds provided that π is small enough and n is large. This is confirmed by simulations in the next section.

7. Numerical experiments. We present here three illustrative numerical applications. The first two are based on simulated data and the last one on real data.

7.1. *Censored matrix.* We begin with a model where we only observe censored elements of the matrix X. More precisely, for a positive censoring value t, instead of X_{ij}, we observe

$$Z_{ij} = X_{ij} I\{|X_{ij}| \le t\} + t(\operatorname{sign} X_{ij}) I\{|X_{ij}| > t\}. \tag{69}$$

Experiment.

– 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 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 nonzero elements in a vector θ of size 500. The associated values are equal to 0.5. We will take $s = 1, 2, 3, 5, 10$.

– We set $y = X\theta + \xi$, where ξ is a normal random vector with zero mean and covariance matrix $\sigma^2 I$ where $\sigma = 0.05/1.96$ (so that for an element of ξ, the probability of being between -0.05 and 0.05 is 95%).

– We compute the matrix Z following (69) with $t = 0.9$.

– We run a linear programming algorithm to compute the solution of (30) where we optimize over $\Theta = \mathbb{R}_+^{500}$. The value of ε is chosen following (65) with $A = (1+\delta)\sqrt{2}$. We note here that in the simulations below the choice of ε is not crucial because the terms with δ in the definition of the estimator are of a larger order of magnitude. Varying ε within a sufficiently wide range does not essentially modify the simulation results. The choice of parameter δ is done the following way.

Choice of δ. The choice of δ in practice is quite crucial. A very small value of δ means that the matrix uncertainty is not taken into account, whereas a too large value of δ means that we overestimate this uncertainty. In both situations the resulting estimator exhibits poor behavior. Consequently, in practice, it is important to select δ within a reasonable range of values. We suggest to choose the range of candidate δ with the "elbow" rule. We plot the number of retrieved nonzero coefficients as a function of δ. Then we consider that a value of δ can be chosen only if the plot is (or begins to be) flat around it. Usually such a plot is highly decreasing at the beginning and then stabilizes; cf. Figure 1. Following this, we take the values in the flat zone $\delta = 0.05, 0.75, 0.1$ for $s = 1, 2, 3, 5$ and $\delta = 0.01, 0.05, 0.1$ for $s = 10$ (the plot for $s = 10$ suggests to start with smaller values for δ).

– We also compute the Lasso estimator with Mallows' C_p choice of the tuning parameter (we use the Lars R-package of T. Hastie and B. Efron) and the Dantzig selector of [8], with the same value ε. Moreover, we compute the thresholded versions of the estimators (T-Lasso, T-Dantzig, T-δ). More precisely, the retrieved

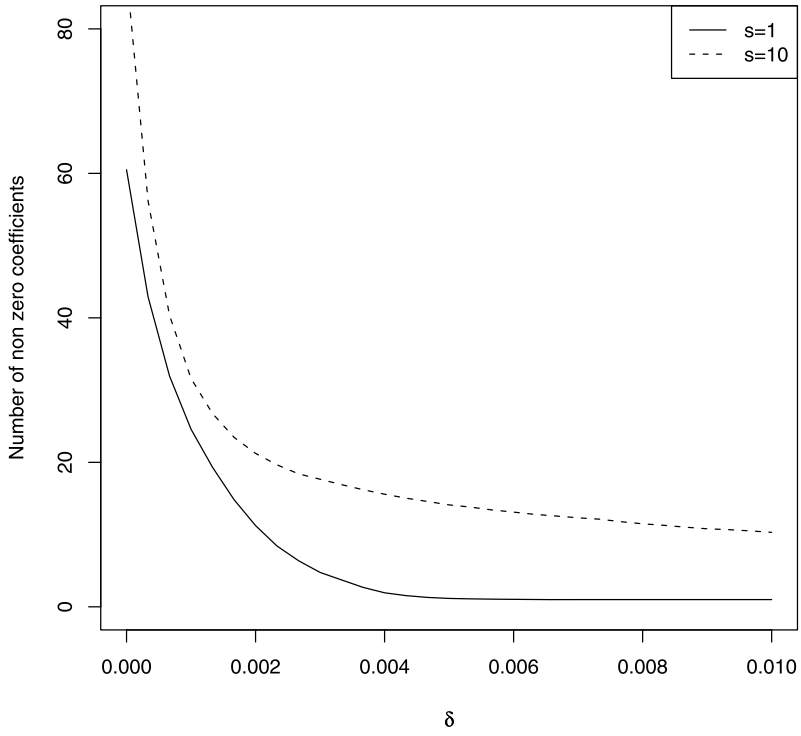

FIG. 1. *Average number of nonzero coefficients in the model with censored matrix for* $s = 1$ *and* $s = 10$.

coefficients whose absolute values are smaller than 20% of the true value of the nonzero coefficients (i.e., smaller than 0.1) are set to zero.

– For all the considered estimators $\hat{\theta}$ of θ 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 nonzero coefficients of $\hat{\theta}$.

– For each value of s we run 100 Monte Carlo simulations.

Results. Tables 1–5 present the empirical averages and standard deviations (in brackets) of Err_1, Err_2, of the number of nonzero coefficients in $\hat{\theta}$ (Nb_1) and of the number of nonzero coefficients in $\hat{\theta}$ belonging to the true sparsity pattern (Nb_2). We also present the total number of simulations where the sparsity pattern is exactly retrieved (Exact). Note that here and in the next numerical examples when a coefficient belonging to the sparsity pattern is retrieved it has systematically the correct sign.

Our first observation is that using the Lasso estimator or the Dantzig selector (i.e., ignoring the matrix uncertainty) has severe consequences. These methods ex-

TABLE 1
Results for the model with censored matrix, $s = 1$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.0679	12.33	95.20	1	0
	(0.0128)	(2.016)	(2.245)	(0)	
T-Lasso	0.0271	2.712	1	1	100
	(0.0086)	(0.8615)	(0)	(0)	
Dantzig	0.0399	3.982	56.92	1	0
	(0.0076)	(0.9880)	(5.594)	(0)	
T-Dantzig	0.0260	2.599	1	1	100
	(0.0068)	(0.6860)	(0)	(0)	
$\delta = 0.05$	0.0122	1.231	1.16	1	85
	(0.0027)	(0.2783)	(0.393)	(0)	
T-$\delta = 0.05$	0.0122	1.224	1	1	100
	(0.0028)	(0.2816)	(0)	(0)	
$\delta = 0.075$	0.0064	0.649	1	1	100
	(0.0017)	(0.1715)	(0)	(0)	
T-$\delta = 0.075$	0.0064	0.649	1	1	100
	(0.0017)	(0.1715)	(0)	(0)	
$\delta = 0.1$	0.0023	0.2330	1	1	100
	(0.0008)	(0.0843)	(0)	(0)	
T-$\delta = 0.1$	0.0023	0.2330	1	1	100
	(0.0008)	(0.0843)	(0)	(0)	

TABLE 2
Results for the model with censored matrix, $s = 2$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.1262	23.57	96.47	2	0
	(0.0218)	(3.813)	(1.670)	(0)	
T-Lasso	0.0456	4.688	2.290	2	77
	(0.0194)	(2.157)	(0.5881)	(0)	
Dantzig	0.0792	8.000	68.79	2	0
	(0.0149)	(2.159)	(4.901)	(0)	
T-Dantzig	0.0404	4.0558	2.04	2	97
	(0.0143)	(1.612)	(0.2416)	(0)	
$\delta = 0.05$	0.0064	0.6654	2.15	2	89
	(0.0039)	(0.4247)	(0.4769)	(0)	
T-$\delta = 0.05$	0.0063	0.6535	2	2	100
	(0.0039)	(0.4314)	(0)	(0)	
$\delta = 0.075$	0.0015	0.1535	2	2	100
	(0.0016)	(0.1637)	(0)	(0)	
T-$\delta = 0.075$	0.0015	0.1535	2	2	100
	(0.0016)	(0.1637)	(0)	(0)	
$\delta = 0.1$	0.0059	0.5410	2	2	100
	(0.0045)	(0.3773)	(0)	(0)	
T-$\delta = 0.1$	0.0059	0.5410	2	2	100
	(0.0045)	(0.3773)	(0)	(0)	

TABLE 3
Results for the model with censored matrix, $s = 3$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.1834	34.54	96.91	3	0
	(0.0326)	(6.156)	(1.407)	(0)	
T-Lasso	0.0776	8.832	4.28	3	25
	(0.0306)	(3.907)	(1.068)	(0)	
Dantzig	0.1209	12.27	73.83	3	0
	(0.0259)	(3.556)	(3.945)	(0)	
T-Dantzig	0.0597	6.108	3.40	3	66
	(0.0251)	(2.877)	(0.6164)	(0)	
$\delta = 0.05$	0.0055	0.5287	3.19	3	85
	(0.0059)	(0.4952)	(0.5038)	(0)	
T-$\delta = 0.05$	0.0053	0.5209	3	3	100
	(0.0058)	(0.5064)	(0)	(0)	
$\delta = 0.075$	0.0148	1.296	3.05	3	95
	(0.0110)	(0.7843)	(0.2179)	(0)	
T-$\delta = 0.075$	0.0148	1.302	3	3	100
	(0.0109)	(0.7935)	(0)	(0)	
$\delta = 0.1$	0.0415	3.791	3.02	3	98
	(0.0177)	(1.1552)	(0.1400)	(0)	
T-$\delta = 0.1$	0.0415	3.793	3	3	100
	(0.0177)	(1.159)	(0)	(0)	

TABLE 4
Results for the model with censored matrix, $s = 5$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.3183	57.68	97.57	5	0
	(0.0596)	(10.51)	(1.089)	(0)	
T-Lasso	0.1693	20.22	10.31	5	0
	(0.0551)	(7.408)	(2.331)	(0)	
Dantzig	0.2225	22.68	81.04	5	0
	(0.0429)	(6.275)	(3.967)	(0)	
T-Dantzig	0.1159	12.08	7.87	5	3
	(0.0430)	(5.174)	(1.6891)	(0)	
$\delta = 0.05$	0.0596	4.544	5.52	5	63
	(0.0417)	(2.457)	(0.8423)	(0)	
T-$\delta = 0.05$	0.0592	4.613	5.08	5	92
	(0.0414)	(2.535)	(0.2712)	(0)	
$\delta = 0.075$	0.1327	11.11	5.12	5	91
	(0.0566)	(3.059)	(0.4069)	(0)	
T-$\delta = 0.075$	0.1327	11.14	5.03	5	97
	(0.0565)	(3.097)	(0.1705)	(0)	
$\delta = 0.1$	0.2331	20.29	5.06	5	95
	(0.0698)	(3.154)	(0.2764)	(0)	
T-$\delta = 0.1$	0.2371	20.61	4.97	4.95	98
	(0.0792)	(3.933)	(0.2628)	(0.21)	

TABLE 5
Results for the model with censored matrix, $s = 10$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.7181	100.7	97.98	10	0
	(0.1426)	(19.38)	(0.8364)	(0)	
T-Lasso	0.5560	55.09	27.02	10	0
	(0.1499)	(14.57)	(3.781)	(0)	
Dantzig	0.5625	55.11	87.71	10	0
	(0.1383)	(13.33)	(3.672)	(0)	
T-Dantzig	0.4203	40.41	22.33	9.98	0
	(0.1467)	(12.91)	(3.212)	(0.1400)	
$\delta = 0.01$	0.3142	24.91	31.6	10	0
	(0.1614)	(7.068)	(4.079)	(0)	
T-$\delta = 0.01$	0.2760	20.41	14.13	9.95	0
	(0.1612)	(7.539)	(1.677)	(0.2598)	
$\delta = 0.05$	0.9679	56.18	14.11	9.33	2
	(0.3688)	(14.65)	(2.403)	(0.8724)	
T-$\delta = 0.05$	1.0187	62.38	10.07	8.23	16
	(0.4088)	(17.90)	(1.226)	(1.535)	
$\delta = 0.1$	1.392	98.89	10.31	7.94	14
	(0.2821)	(11.27)	(1.514)	(1.391)	
T-$\delta = 0.1$	1.483	108.1	6.92	5.95	37
	(0.3003)	(12.99)	(1.324)	(1.519)	

hibit erratic behavior already for the minimal sparsity $s = 1$. Though their sets of nonzero components steadily include the relevant set, they are much too large and the results are very far from the correct selection. We also see that the MU-selector strictly improves upon the Lasso estimator and the Dantzig selector for all the considered error criteria and values of s. In particular, for $\delta = 0.1$ and $s = 1, 2, 3, 5$, it almost systematically retrieves the sparsity pattern and the two error measures remain very small. This is obviously no longer the case for the bigger value $s = 10$. However, note that the MU-selector remains quite satisfactory in terms of selecting the sparsity pattern since the average number of retrieved coefficients is about 10 and the average number of retrieved coefficients is about 8. Thresholding the coefficients logically improves the retrieved sparsity patterns of the Lasso estimator and Dantzig selector. Nevertheless, in most of the cases the MU-selector outperforms their thresholded versions as well. This fact is even more significant because we simulate with a threshold which has been well chosen knowing the true value of the nonzero coefficients. In practice, choosing a relevant threshold is a very intricate question since the order of magnitude of the nonzero coefficients is typically unknown. On the other hand, for the MU-selector thresholding can be avoided. Indeed, its effect is not significant, especially when s is small. This is due to the fact that the original (nonthresholded) MU-selector is already very accurate in recovering the sparsity pattern.

Finally, note that the good results for the MU-selector are not due to the fact that we optimize over $\Theta = \mathbb{R}_+^{500}$ instead of $\Theta = \mathbb{R}^{500}$. In particular, taking $\delta = 0$ leads to the same kind of results as those for the Dantzig selector.

7.2. *Model with missing data.* We consider now the model with missing data as defined in Section 2. We design the numerical experiment in the same way as in Section 7.1 except that the observed matrix Z is now given by (7) with $\pi = 0.1$.

Results. The results are given in Tables 6–10. We see that again the Lasso and Dantzig selector are highly unstable in selecting the sparsity pattern, whereas the MU-selector does a good job. The thresholded estimators T-Lasso and T-Dantzig are also quite accurate in retrieving the sparsity pattern, except for $s = 10$. However, in all the cases the MU-selector does it better. The MU-selector with $\delta = 0.05$ (or $\delta = 0.01$ for $s = 10$) has the smallest error measures Err_1 and Err_2, whereas the sparsity pattern is better retrieved for $\delta = 0.1$. This reflects a tradeoff between estimation and selection. Smaller values of δ lead to smaller errors Err_1 and Err_2, whereas larger values of δ lead to a very accurate recovery of the sparsity pattern. The error measures Err_1 and Err_2 of the thresholded estimators T-Lasso and T-Dantzig are somewhat smaller than those of the MU-selector, except for $s = 10$.

TABLE 6
Results for the model with missing data, $s = 1$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.0212	2.606	94.59	1	0
	(0.0105)	(1.232)	(3.256)	(0)	
T-Lasso	0.0011	0.111	1	1	100
	(0.0010)	(0.1019)	(0)	(0)	
Dantzig	0.0109	1.114	64.24	1	0
	(0.0072)	(0.7360)	(11.06)	(0)	
T-Dantzig	0.0011	0.1097	1	1	100
	(0.0010)	(0.1030)	(0)	(0)	
$\delta = 0.05$	0.0041	0.3376	8.55	1	6
	(0.0029)	(0.2218)	(5.087)	(0)	
T-$\delta = 0.05$	0.0022	0.2271	1	1	100
	(0.0012)	(0.1203)	(0)	(0)	
$\delta = 0.075$	0.0039	0.3449	3.99	1	29
	(0.0021)	(0.1625)	(3.090)	(0)	
T-$\delta = 0.075$	0.0031	0.3133	1	1	100
	(0.0011)	(0.1124)	(0)	(0)	
$\delta = 0.1$	0.0047	0.4490	1.94	1	61
	(0.0019)	(0.1356)	(1.605)	(0)	
T-$\delta = 0.1$	0.0044	0.4451	1	1	100
	(0.0012)	(0.1268)	(0)	(0)	

TABLE 7
Results for the model with missing data, $s = 2$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.0425	4.675	96.02	2	0
	(0.0162)	(1.786)	(2.074)	(0)	
T-Lasso	0.0047	0.4481	2.02	2	98
	(0.0037)	(0.3318)	(0.1400)	(0)	
Dantzig	0.0269	2.695	74.39	2	0
	(0.0134)	(1.4823)	(5.774)	(0)	
T-Dantzig	0.0046	0.4330	2	2	100
	(0.0035)	(0.3194)	(0)	(0)	
$\delta = 0.05$	0.0131	1.033	6.76	2	13
	(0.0078)	(0.4688)	(3.572)	(0)	
T-$\delta = 0.05$	0.0106	1.018	2	2	100
	(0.0055)	(0.4692)	(0)	(0)	
$\delta = 0.075$	0.0167	1.517	3.20	2	48
	(0.0071)	(0.4557)	(1.489)	(0)	
T-$\delta = 0.075$	0.0160	1.525	2.01	2	99
	(0.0064)	(0.4584)	(0.099)	(0)	
$\delta = 0.1$	0.0247	2.351	2.27	2	77
	(0.0074)	(0.4634)	(0.5264)	(0)	
T-$\delta = 0.1$	0.0245	2.362	2	2	100
	(0.0070)	(0.4731)	(0)	(0)	

TABLE 8
Results for the model with missing data, $s = 3$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.0721	6.828	96.89	3	0
	(0.0251)	(2.116)	(1.449)	(0)	
T-Lasso	0.0134	1.225	3.12	3	88
	(0.0093)	(0.8126)	(0.3250)	(0)	
Dantzig	0.0496	4.844	80.45	3	0
	(0.0204)	(2.075)	(4.693)	(0)	
T-Dantzig	0.0117	1.119	3.05	3	95
	(0.0082)	(0.8438)	(0.2180)	(0)	
$\delta = 0.05$	0.0322	2.591	6.8	3	10
	(0.0138)	(0.7730)	(2.942)	(0)	
T-$\delta = 0.05$	0.0293	2.726	3.04	3	96
	(0.0119)	(0.8735)	(0.1959)	(0)	
$\delta = 0.075$	0.0439	4.0308	3.96	3	50
	(0.0137)	(0.7988)	(1.333)	(0)	
T-$\delta = 0.075$	0.0432	4.1098	3.01	3	99
	(0.0130)	(0.8505)	(0.0994)	(0)	
$\delta = 0.1$	0.0653	6.217	3.21	3	84
	(0.0160)	(0.8355)	(0.5156)	(0)	
T-$\delta = 0.1$	0.0651	6.235	3	3	100
	(0.0158)	(0.8500)	(0)	(0)	

TABLE 9
Results for the model with missing data, $s = 5$

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.1302	9.993	97.24	5	0
	(0.0499)	(2.657)	(1.097)	(0)	
T-Lasso	0.0418	3.331	5.65	5	56
	(0.0326)	(2.056)	(0.899)	(0)	
Dantzig	0.1005	9.371	84.36	5	0
	(0.0443)	(4.113)	(4.009)	(0)	
T-Dantzig	0.0365	3.356	5.38	5	74
	(0.0275)	(2.454)	(0.7454)	(0)	
$\delta = 0.05$	0.1033	8.301	8.19	5	14
	(0.0384)	(1.713)	(2.591)	(0)	
T-$\delta = 0.05$	0.1001	8.900	5.14	5	87
	(0.0362)	(2.146)	(0.3746)	(0)	
$\delta = 0.075$	0.1485	13.25	5.96	5	48
	(0.0415)	(1.716)	(1.272)	(0)	
T-$\delta = 0.075$	0.1477	13.53	5.05	5	95
	(0.0402)	(1.999)	(0.2179)	(0)	
$\delta = 0.1$	0.2133	19.60	5.31	5	79
	(0.0494)	(1.708)	(0.7835)	(0)	
T-$\delta = 0.1$	0.2131	19.70	5.03	5	97
	(0.0488)	(1.904)	(0.1705)	(0)	

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

	$\mathbf{Err_1}$	$\mathbf{Err_2}$	$\mathbf{Nb_1}$	$\mathbf{Nb_2}$	**Exact**
Lasso	0.4746	18.04	98.01	10	0
	(0.1702)	(4.334)	(0.7549)	(0)	
T-Lasso	(0.1710)	(3.357)	(5.317)	(0)	0
	0.4358	13.74	37.44	10	
Dantzig	0.4229	38.13	90.77	10	0
	(0.1684)	(15.95)	(2.853)	(0)	
T-Dantzig	0.3862	34.84	32.77	10	0
	(0.1690)	(16.35)	(5.184)	(0)	
$\delta = 0.01$	0.2891	10.78	47.38	10	0
	(0.1285)	(2.059)	(5.351)	(0)	
T-$\delta = 0.01$	0.2725	12.26	19.93	10	0
	(0.1271)	(2.719)	(3.311)	(0)	
$\delta = 0.05$	0.7710	45.89	18.02	9.91	0
	(0.2755)	(6.212)	(3.720)	(0.2861)	
T-$\delta = 0.05$	0.7719	48.36	13.41	9.73	6
	(0.2807)	(7.134)	(2.015)	(0.6611)	
$\delta = 0.1$	1.182	84.80	13.42	9.37	6
	(0.2983)	(8.477)	(2.324)	(0.8204)	
T-$\delta = 0.1$	1.196	87.42	10.81	8.78	23
	(0.3104)	(9.304)	(1.521)	(1.338)	

Note, however, that we report the results for the performance of T-Lasso and T-Dantzig with a threshold based on the knowledge of the true coefficients.

7.3. *Portfolio replication*. We now present a "toy" application based on financial data. We apply model (1) and (2) and the MU-selector in the context of portfolio replication as described in Section 2. We take the data of the open and close prices of $p = 491$ assets in the Standard and Poors S&P 500 index for the $n = 251$ trading days of 2007. These data are provided by the Yahoo Finance Database. The assets we use are those available for the whole year.

Experiment. Let p_{ij}^o and p_{ij}^c denote the open and close prices of the jth asset for the ith day. Our experiment is the following.

– We consider the matrix $\tilde{X}$ with entries $(\tilde{X})_{ij} = p_{ij}^c - p_{ij}^o$ and define X as the normalized matrix obtained from $\tilde{X}$.

– We pick s assets to build our portfolio. The coordinate of each chosen asset in the vector $\theta \in \mathbb{R}^{491}$ is set to $1/s$ and the other coordinates to 0 [note that, in practice, if the jth asset is in the portfolio, it means that the corresponding coordinate of θ is $1/(s\tilde{\sigma}_j)$, where $\tilde{\sigma}_j$ is the empirical standard deviation of its absolute returns].

– We consider six portfolios (see Table 11).

– We compute $y = X\theta + \xi$ where ξ is the same noise as in Section 7.1. In practice, the noise ξ can reflect an uncertainty about the management costs, a lack of transparency in the definition of the returns of the portfolio or some rounding approximations.

– We consider a matrix uncertainty of the following type: Z is obtained from X by replacing one of its columns by the zero column. The column corresponds to one of the assets in the portfolio. The goal of this manipulation is to mimic the fact that in practice not all the existing assets are in our restricted class. One of the assets in the portfolio does not belong to the restricted class since the corresponding column of X is suppressed. Of course, this asset cannot be retrieved. We suppress the column associated to an asset different from Boeing and Google.

– We solve (30) with such a matrix Z, with $\delta = 0.5$ and ε chosen as in Section 7.1. We also compute the Lasso estimator and the Dantzig selector.

TABLE 11
Initial portfolios

$s = 2$	$s = 3$
Boeing, Goldman Sachs	Boeing, Google, Goldman Sachs
Boeing, Coca Cola	Boeing, Google, Coca Cola
Boeing, Ford	Boeing, Google, Ford

TABLE 12
Retrieved portfolios, MU-selector

Initial portfolio	Retrieved portfolio
B, Goldman Sachs	B, Morgan Stanley, Merrill Lynch
B, Coca Cola	B, Pepsico
B, Ford	B, General Motors
B, G, Goldman Sachs	B, G, Morgan Stanley, Merrill Lynch
B, G, Coca Cola	B, G
B, G, Ford	B, G, General Motors

Results. We write B for Boeing and G for Google. The initial portfolios and the portfolios retrieved by the MU-selector are presented in Table 12.

The results are very satisfying. Indeed, the algorithm almost always finds the correct number of assets in the portfolio and the discarded asset is replaced by one or two assets that are intuitively close to it. Moreover, if one takes $\delta = 0.4$, then for the initial portfolio (Boeing, Google, Coca Cola) the retrieved portfolio becomes (Boeing, Google, Pepsico), whereas the other results remain the same. Finally, note that the Lasso estimator and the Dantzig selector (usual Dantzig selector or MU-selector with $\delta = 0$) systematically output more than 20 assets in the retrieved portfolio.

REFERENCES

[1] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[2] BRODIE, J., DAUBECHIES, I., DE MOLC, C., GIANNONE, D. and LORIS, I. (2009). Sparse and stable Markowitz portfolios. *PNAS* **106** 12267–12272.

[3] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[4] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[5] CANDÈS, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris* **346** 589–592. MR2412803

[6] CANDÈS, E. J., ROMBERG, J. and TAO, T. (2005). Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846

[7] CANDÈS, E. J. and TAO, T. (2006). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. MR2243152

[8] CANDÈS, E. J. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313–2404. MR2382644

[9] CAVALIER, L. and HENGARTNER, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems* **21** 1345–1361. MR2158113

[10] CAVALIER, L. and RAIMONDO, M. (2007). Wavelet deconvolution with noisy eigen-values. *IEEE Trans. Signal Process.* **55** 2414–2424. MR1500172

[11] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[12] DRAPER, N. R. and SMITH, H. (1998). *Applied Regression Analysis*. Wiley, New York. MR1614335
[13] DALALYAN, A. and TSYBAKOV A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.
[14] DALALYAN, A. and TSYBAKOV A. B. (2009). Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of COLT-2009*. Available at arXiv:0903.1223.
[15] EFROMOVICH, S. and KOLTCHINSKII, V. (2001). On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** 2876–2894. MR1872847
[16] FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York. MR0898653
[17] HOFFMANN, M. and REISS, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.* **36** 310–336. MR2387973
[18] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087
[19] LOUNICI, K. (2010). High-dimensional stochastic optimization with the generalized Dantzig estimator. To appear. Available at arXiv:0811.2281v1.
[20] KOLTCHINSKII, V. (2009). Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200
[21] KOLTCHINSKII, V. (2010). Oracle inequalities in empirical risk minimization and sparse recovery problems. St Flour Lecture Notes.
[22] MARTEAU, C. (2007). Regularization of inverse problems with unknown operator. *Math. Methods Statist.* **15** 415–443. MR2301660
[23] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[24] VAN DE GEER, S. A. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809
[25] ZHANG, C. H. and HUANG, J. (2008). The sparsity and biais of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448
[26] ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144. MR2543687
[27] ZHAO, P. and YU, B. (2007). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2567. MR2274449
[28] ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

CMAP-ÉCOLE POLYTECHNIQUE PARIS
UMR CNRS 7641
91128 PALAISEAU CEDEX
FRANCE
E-MAIL: mathieu.rosenbaum@polytechnique.edu

LABORATOIRE DE STATISTIQUE, CREST
3, AV. PIERRE LAROUSSE
92240 MALAKOFF
FRANCE
AND
LPMA (UMR CNRS 7599)
UNIVERSITÉ PARIS 6
4, PLACE JUSSIEU
75252 PARIS, CEDEX 05
FRANCE
E-MAIL: alexandre.tsybakov@upmc.fr