Electronic Journal of Statistics
Vol. 10 (2016) 1729–1750
ISSN: 1935-7524
DOI: 10.1214/15-EJS1095

An $\{\ell_1, \ell_2, \ell_\infty\}$-regularization approach to high-dimensional errors-in-variables models

Alexandre Belloni

The Fuqua School of Business
Duke University
e-mail: abn5@duke.edu

Mathieu Rosenbaum

Laboratoire de Probabilités et Modèles Aléatoires
Université Pierre et Marie Curie (Paris 6)
and
Centre de Recherche en Economie et Statistique
ENSAE-Paris Tech
e-mail: mathieu.rosenbaum@upmc.fr

and

Alexandre B. Tsybakov

Centre de Recherche en Economie et Statistique
ENSAE-Paris Tech
e-mail: alexandre.tsybakov@ensae.fr

Abstract: Several new estimation methods have been recently proposed for the linear regression model with observation errors in the design. Different assumptions on the data generating process have motivated different estimators and analysis. In particular, the literature considered (1) observation errors in the design uniformly bounded by some $\bar{\delta}$, and (2) zero-mean independent observation errors. Under the first assumption, the rates of convergence of the proposed estimators depend explicitly on $\bar{\delta}$, while the second assumption has been essentially applied when an estimator for the second moment of the observational error is available. This work proposes and studies two new estimators which, compared to other procedures for regression models with errors in the design, exploit an additional ℓ_∞-norm regularization. The first estimator is applicable when both (1) and (2) hold but does not require an estimator for the second moment of the observational error. The second estimator is applicable under (2) and requires an estimator for the second moment of the observation error. Importantly, we impose no assumption on the accuracy of this pilot estimator, in contrast to the previously known procedures. As the recent proposals, we allow the number of covariates to be much larger than the sample size. We establish the rates of convergence of the estimators and compare them with the bounds obtained for related estimators in the literature. These comparisons show interesting insights on the interplay of the assumptions and the achievable rates of convergence.

Received December 2014.

1. Introduction

Several new estimation methods have been recently proposed for the linear regression model with observation errors in the design. Such problems arise in a variety of applications, see [7, 8, 10, 11]. In this work we consider the following regression model with observation errors in the design:

$$\begin{aligned} y &= X\theta^* + \xi, \\ Z &= X + W. \end{aligned}$$

Here the random vector $y \in \mathbb{R}^n$ and the random $n \times p$ matrix Z are observed, the $n \times p$ matrix X is unknown, W is an $n \times p$ random noise matrix, and $\xi \in \mathbb{R}^n$ is a random noise vector. The vector of unknown parameters of interest is θ^* which is assumed to belong to a given convex subset Θ of $\mathbb{R}^p$ characterizing some prior knowledge about θ^* (potentially $\Theta = \mathbb{R}^p$). Similarly to the recent literature on this topic, we consider the setting where the dimension p can be much larger than the sample size n and the vector θ^* is s-sparse, which means that it has no more than s non-zero components.

The need for new estimators under errors in the design arises from the fact that standard estimators (e.g. Lasso and Dantzig selector) might become unstable, see [8]. To deal with this framework, various assumptions have been considered, leading to different estimators.

A classical assumption in the literature is a uniform boundedness condition on the errors in the design, namely,

$$|W|_\infty \le \bar{\delta} \;\text{ almost surely,} \tag{1}$$

where $|\cdot|_q$ denotes the ℓ_q-norm for $1 \le q \le \infty$. Note that this assumption allows for various dependences between the errors in the design. In this setting, the Matrix Uncertainty selector (MU selector), which is robust to the presence of errors in the design, is proposed in [8]. The MU selector $\hat{\theta}^{MU}$ is defined as a solution of the minimization problem

$$\min\{|\theta|_1 : \; \theta \in \Theta, \left|\tfrac{1}{n} Z^T(y - Z\theta)\right|_\infty \le \tau_1|\theta|_1 + \tau\}, \tag{2}$$

where the parameters τ_1 and τ depend on the level of the noises of W and ξ respectively. Under appropriate choices of these parameters and suitable assumptions on X, it was shown in [8] that with probability close to 1,

$$|\hat{\theta}^{MU} - \theta^*|_q \le Cs^{1/q}\{\bar{\delta} + \bar{\delta}^2\}|\theta^*|_1 + Cs^{1/q}\sqrt{\frac{\log p}{n}}, \quad 1 \le q \le \infty. \tag{3}$$

Here and in what follows we denote by the same symbol C (or c') different positive constants that do not depend on θ^*, s, n, p, $\bar{\delta}$, but only on the variance parameters σ^2 and σ_*^2 (defined later). The result (3) implies consistency as the sample size n tends to infinity provided that the error in the design goes to zero sufficiently fast to offset $s^{1/q}|\theta^*|_1$, and the number of variables p and the sparsity s of θ^* do not grow too fast relative to the sample size n.

An alternative assumption considered in the literature is that the entries of the random matrix W are independent with zero mean, the values

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(W_{ij}^2),\ j = 1, \dots, p,$$

are finite, and data-driven estimators $\hat{\sigma}_j^2$ of σ_j^2 are available converging with an appropriate rate. This assumption motivated the idea to compensate the bias of using the observable $Z^T Z$ instead of the unobservable $X^T X$ in (2) thanks to the estimates of σ_j^2. This compensated MU selector, introduced in [9] and denoted as $\hat{\theta}^C$, is defined as a solution of the minimization problem

$$\min\{|\theta|_1 :\ \theta \in \Theta,\ \left|\tfrac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \le \tau_1 |\theta|_1 + \tau\},$$

where $\widehat{D}$ is the diagonal matrix with entries $\hat{\sigma}_j^2$ and $\tau_1 > 0$ and $\tau > 0$ are constants chosen according to the level of the noises and the accuracy of the $\hat{\sigma}_j^2$.

Rates of convergence of the compensated MU selector were established in [9]. Importantly, the compensated MU selector can be consistent as the sample size n increases even if the error in the design does not vanish. This is in contrast to the case of the MU selector, where the bounds are small only if the bound on the design error $\bar{\delta}$ is small. In particular, under regularity conditions, when θ^* is s-sparse, it is shown in [9] that with probability close to 1,

$$|\hat{\theta}^C - \theta^*|_q \le C s^{1/q} \sqrt{\frac{\log p}{n}} (|\theta^*|_1 + 1), \quad 1 \le q \le \infty. \tag{4}$$

Under the same alternative assumption, a conic programming based estimator $\hat{\theta}^{Conic}$ has been recently proposed and analyzed in [1]. The estimator $\hat{\theta}^{Conic}$ is defined as the first component of any solution of the optimization problem

$$\min_{(\theta,t) \in \mathbb{R}^p \times \mathbb{R}_+} \{|\theta|_1 + \lambda_2 t :\ \theta \in \Theta,\ \left|\tfrac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \le \tau_2 t + \tau,\ |\theta|_2 \le t\}, \tag{5}$$

where λ_2, τ_2 and τ are some positive tuning constants. Akin to $\hat{\theta}^C$, this estimator compensates for the bias by using the estimators $\hat{\sigma}_j^2$ of σ_j^2. However, it exploits a combination of ℓ_1 and ℓ_2-norm regularization to be more adaptive. It was shown to attain a bound as in (4) and to be computationally feasible since it is cast as a tractable convex optimization problem (a second order cone programming problem). Moreover, under mild additional conditions, with probability close to 1, the estimator (5) achieves improved bounds of the form

$$|\hat{\theta}^{Conic} - \theta^*|_q \le C s^{1/q} \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1), \quad 1 \le q \le \infty, \tag{6}$$

provided that $\widehat{D}$ converges to D in sup-norm with the rate $\sqrt{(\log p)/n}$. It is shown in [1] that the rate of convergence in (6) is minimax optimal in the considered model.

There have been other approaches to the errors-in-variables model, usually exploiting some knowledge about the vector θ^*, see [2, 3, 7, 10]. Assuming $|\theta^*|_1$ is known, [7] proposed an estimator $\hat{\theta}'$ defined as the solution of a non-convex program which can be well approximated by an iterative relaxation procedure. In the case where the entries of the regression matrix X are zero-mean subgaussian and θ^* is s-sparse, under appropriate assumptions, it is proved that for the error in ℓ_2-norm $(q = 2)$,

$$|\hat{\theta}' - \theta^*|_2 \le C(\theta^*) s^{1/2} \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1), \tag{7}$$

with probability close to 1. Here, the value $C(\theta^*)$ depends on θ^*, so that there is no guarantee that the estimator attains the optimal bound as in (6). Assuming that the sparsity s of θ^* is known and the non-zero components of θ^* are separated from zero in so that

$$|\theta_j^*| \ge C \sqrt{\frac{\log p}{n}} (|\theta^*|_2 + 1),$$

an orthogonal matching pursuit algorithm to estimate θ^* is introduced in [2, 3]. Focusing as in [7] on the particular case where the entries of the regression matrix X are zero-mean subgaussian, it is shown in [2, 3] that this last estimator satisfies a bound analogous to (6), as well as a consistent support recovery result, without requiring estimates of σ_j^2.

The main purpose of this work is to show that an additional regularization term based on the ℓ_∞-norm leads to improved rates of convergence in several situations. We propose two new estimators for θ^*. The first proposal is applicable under a new combination of the assumptions mentioned above. Namely, we assume that the components of the errors in the design are uniformly bounded by $\bar{\delta}$ as in (1), and that the rows of W are independent and with zero mean. However, we will neither assume that a data-driven estimator $\widehat{D}$ is available, nor that specific features of θ^* are known (e.g. s or $|\theta^*|_1$). The estimator is defined as a solution of a regularized optimization problem which uses simultaneously ℓ_1, ℓ_2 and ℓ_∞ regularization functions. It can be cast as a convex optimization problem and the solution can be easily computed. We study its rates of convergence in various norms in Section 3. One of the conclusions is that for $\bar{\delta} \gg \{(\log p)/n\}^{1/2}$, the new estimator has improved rates of convergence compared to the MU selector. Furthermore, note that the conic estimator $\hat{\theta}^{Conic}$ studied in [1] can also be applied. Indeed, our setting can be embedded into that of [1] with $\widehat{D}$ being the identically zero $p \times p$ matrix, which means that we have an estimator of each σ_j^2 with an error bounded by $\bar{\delta}^2$. Comparing the bounds yields that the conic estimator $\hat{\theta}^{Conic}$ achieves the same rate as our new estimator if $\bar{\delta}$ is smaller than or of the order $\{(\log p)/n\}^{1/4}$. However, there is no bound for $\hat{\theta}^{Conic}$ available when $\bar{\delta} \gg \{(\log p)/n\}^{1/4}$.

The second estimator we propose applies to the same setting as in [1]. The idea of taking advantage of an additional ℓ_∞-norm regularization can be used to improve the conic estimator $\hat{\theta}^{Conic}$ of [1] whenever the rate of convergence of the estimator $\widehat{D}$ for σ_j^2, $j = 1, \ldots, p$, is slower than $\{(\log p)/n\}^{1/2}$. This motivated us to propose and analyze a modification of the conic estimator. We derive new rates of convergence that can lead to improvements. However, we acknowledge that in the case considered in [1], where the rate of convergence of $\widehat{D}$ is $\{(\log p)/n\}^{1/2}$, there is no gain in the rates of convergence when using the additional ℓ_∞-norm regularization. Therefore the proposed estimator is of particular interest in applications where it is costly and disruptive to generate precise measurements of X_i (the i-th row of X: the couple (X_i, Z_i) is observed in that case, and y_i is unobserved), while observations on (y_k, Z_k) are readily available. For example, interventions in experiments can be used to precisely measure a particular covariate of subjects but such interventions can invalidate the use of the outcome y_i for that particular subject. This is relevant in epidemiological research, where accurate measurements are commonly difficult to achieve without expensive interviews and tests which tend to impact the future behavior of the subject invalidating the associated outcome, see [4]. In such settings, we observe n observations of (y_k, Z_k) and n^* observations of the pair (X_i, Z_i). When the (X_i, Z_i) are iid, the later can be typically used to construct estimates $\hat{\sigma}_j^2$ for $\hat{\sigma}_j^2$ so that $|\widehat{D} - D|_\infty = b(\varepsilon) \leq C\sqrt{\{\log(p/\varepsilon)\}/n^*}$ with probability $1 - \varepsilon$. Nonetheless, given the sampling cost structure, it is often the case that $n^* \ll n$ and $b(\varepsilon) \gg \sqrt{(\log p)/n}$.

We emphasize that the new use of the ℓ_∞-norm in the first order condition is precisely what drives our new rates of convergence. In the settings we are concerned with, the rates of convergence are dominated by the "crude" estimators $\widehat{D}$ of D. The impact of using $\widehat{D}$ instead of D is well controlled by $|(\widehat{D} - D)\theta^*|_\infty \leq b(\varepsilon)|\theta^*|_\infty$ since D is diagonal. Previous estimators that rely on ℓ_1 or ℓ_2-regularization terms would not be able to fully exploit this structure and would not be sharp if $b(\varepsilon) \gg \sqrt{(\log p)/n}$. Furthermore, the ℓ_∞-regularization allows us to achieve a convex formulation for the problem. We view the use of the ℓ_∞-regularization as a new way to increase the adaptivity of an estimator that can be of independent interest in other applications.

The paper is organized as follows. Section 2 contains the notation, main assumptions and some preliminary lemmas needed to determine threshold constants in the algorithms. The definition and properties of our first estimator are given in Section 3 whereas those of our second procedure can be found in Section 4. Section 5 contains simulation results. Some auxiliary lemmas are relegated to an appendix.

2. Notation, assumptions, and preliminary lemmas

In this section, we introduce the assumptions which will be required to derive the rates of convergence of the proposed estimators. One set of conditions pertains to the design matrix and the second to the errors in the model. We also state preliminary lemmas related to the stochastic error terms. We start by introducing some notation.

2.1. Notation

Let $J \subset \{1, \dots, p\}$ be a set of integers. We denote by $|J|$ the cardinality of J. For a vector $\theta = (\theta_1, \dots, \theta_p)$ in $\mathbb{R}^p$, we denote by θ_J the vector in $\mathbb{R}^p$ whose j-th component satisfies $(\theta_J)_j = \theta_j$ if $j \in J$, and $(\theta_J)_j = 0$ otherwise. For $\gamma > 0$, the random variable η is said to be *sub-gaussian with variance parameter* γ^2 (or shortly γ*-sub-gaussian*) if, for all $t \in \mathbb{R}$,

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

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

2.2. Design matrix

The performance of the estimators that we consider below is influenced by the properties of the Gram matrix

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

We will assume that:

(A1) The matrix X is deterministic.

In order to characterize the behavior of the design matrix, we set

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

where X_{ij} are the elements of matrix X and we consider the sensitivity characteristics related to the Gram matrix Ψ. For $u > 0$, define the cone

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

where J is a subset of $\{1, \dots, p\}$. For $q \in [1, \infty]$ and an integer $s \in [1, p]$, the ℓ_q*-sensitivity* (cf. [5]) is defined as follows:

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

Like in [5], we use here the sensitivities to derive the rates of convergence of estimators under sparsity. Importantly, as shown in [5], the approach based on sensitivities is more general than that based on the restricted eigenvalue or the coherence condition, see also [1, 6, 9]. In particular, under those conditions, we have $\kappa_q(s, u) \geq c\, s^{-1/q}$ for some constant $c > 0$, which implies the usual optimal bounds for the errors.

2.3. Disturbances

Next we turn to the error W in the design and the error ξ in the regression equation. We will make the following assumptions:

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

(A3) The rows of the noise matrix W are independent zero-mean subgaussian random vectors with variance parameter σ_*^2, and $\mathbb{E}[W_{ij}W_{ik}] = 0$ for all $1 \le j < k \le p$. Furthermore, W is independent of ξ.

2.4. Bounds on the stochastic error terms

We now state some useful lemmas from [1] and [9] that provide bounds to various stochastic error terms that play a role in our analysis. We state them here because they introduce the thresholds δ_i, δ_i' that will be used in the definition of the estimators. In what follows, D is the diagonal matrix with diagonal elements σ_j^2, $j = 1, \dots, p$, and for a square matrix A, we denote by $\mathrm{Diag}\{A\}$ the matrix with the same dimensions as A, the same diagonal elements, and all off-diagonal elements equal to zero.

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

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

where

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

and for an integer N,

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

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

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

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

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

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

where

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

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

The proofs of Lemmas 1 and 2 can be found in [9] and [1] respectively.

3. $\{\ell_1, \ell_2, \ell_\infty\}$-MU selector

In this section, we define and analyze our first estimator. It can be seen as a compromise between the MU selector (2) and the conic estimator (5) achieved thanks to an additional ℓ_∞-norm regularization. In the setting that we consider now, the estimate $\widehat{D}$ is not available but the rows of the design error matrix W are independent with mean 0, and its entries are uniformly bounded. Formally, in this section we make the following assumption:

(A4) Almost surely, $|W|_\infty \le \bar{\delta}$.

Assumptions (A1)–(A4) imply the assumptions in [8]. However, they neither imply or are implied by the assumptions in [9]. That is, it is an intermediary set of conditions relative to the original assumptions for the MU selector in [8] and to those for the compensated MU selector in [9]. Importantly, we do not assume that there are some accurate estimators of σ_j^2.

Similar to [8], consistent estimates require $\bar{\delta} \to 0$ as the sample size grows. The bound $\bar{\delta}$ does not necessarily scale with p as $\bar{\delta}$ is not derived from (A3). Assumption (A4) is motivated from sampling schemes where the precision with which the covariate X_{ij} is measured can be controlled by the practitioner to some degree and the technology allows $|Z_{ij} - X_{ij}| \le \bar{\delta}$; in principle $\bar{\delta}$ could be made smaller if more expensive measurements were performed. Our finite sample analysis keeps track of the dependence of $\bar{\delta}$ explicitly allowing to capture its impact as a potential function of the sample size. A simple illustrative example is when the $W_{ij}, i = 1, \dots, n, j = 1, \dots, p$ are independent zero-mean random variables that are a.s. bounded by $\bar{\delta}$. Then the vector W_i is sub-gaussian with parameter $\bar{\sigma}_*^2 = \bar{\delta}^2$.

We consider the estimator $\hat{\theta}$ such that $(\hat{\theta}, \hat{t}, \hat{u}) \in \mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}_+$ is a solution of the following minimization problem:

$$\begin{array}{ll} \min_{\theta,t,u} & |\theta|_1 + \lambda_2 t + \lambda_\infty u \\ & (\theta, t, u) \in \Theta, \\ & \left|\frac{1}{n} Z^T(y - Z\theta)\right|_\infty \le \tau_2 t + \bar{\delta}^2 u + \tau, \\ & |\theta|_2 \le t, |\theta|_\infty \le u, \end{array} \tag{8}$$

where $\lambda_2 > 0$ and $\lambda_\infty > 0$ are tuning constants and we allow for $(\theta, t, u) \in \Theta$ where Θ is a pre-specified convex set that contains $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ and characterizes some prior knowledge (a trivial choice is $\Theta = \mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}_+$). This estimator $\hat{\theta}$ will be further referred to as the $\{\ell_1, \ell_2, \ell_\infty\}$-MU selector.

Remark 1 (Safeguard constraints). In order to further bound t and u, we can add constraints that exploit that $|\cdot|_q \leq |\cdot|_1$ for $q \geq 1$. Therefore, the constraints

$$\theta = \theta^+ - \theta^-, \quad \theta^+ \geq 0, \quad \theta^- \geq 0, \quad w = \sum_{j=1}^{p} \{\theta_j^+ + \theta_j^-\}, \quad u \leq t, \quad \text{and} \quad t \leq w$$

preserve the convexity of the optimization problem and can potentially yield additional performance. We note that our theoretical results allow for such safeguard constraints to be included in the estimation.

The estimator above attempts to mimic the conic estimator (5) without estimators $\hat{\sigma}_j^2$ for σ_j^2, $j = 1, \ldots, p$. In order to make θ^* feasible for (8), the contribution of the unknown term $\frac{1}{n}\mathrm{Diag}(W^TW)\theta^*$ needs to be bounded. This is precisely the role of the extra term $\bar{\delta}^2 u$ in the constraint since $|\theta|_\infty \leq u$ and $|\frac{1}{n}\mathrm{Diag}(W^TW)|_\infty \leq \bar{\delta}^2$ almost surely. Note that the use of u and t instead of $|\theta|_\infty$ and $|\theta|_2$ in the constraint makes (8) a convex programming problem.

This new estimator exploits Assumptions (A2)–(A4) to achieve a rate of convergence that is intermediary relative to the rate of the MU selector and to that of the conic estimator.

Set $\tau_2 = \delta_1'(\varepsilon) + \delta_4'(\varepsilon)$ and $\tau = \delta_2(\varepsilon) + \delta_3(\varepsilon)$. Note that τ_2 and τ are of the order $\sqrt{(\log p)/n}$. The next theorem summarizes the performance of the estimator defined by solving (8).

Theorem 1. *Let Assumptions (A1)–(A4) hold. Assume that the true parameter* θ^* *is* s*-sparse and* $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ *belongs to* Θ. *Let* $0 < \varepsilon < 1$, $1 \leq q \leq \infty$ *and* $0 < \lambda_2, \lambda_\infty < \infty$. *Let* $\hat{\theta}$ *be the* $\{\ell_1, \ell_2, \ell_\infty\}$*-MU selector. If* $\kappa_q(s, 1 + \lambda_2 + \lambda_\infty) \geq cs^{-1/q}$ *for some constant* $c > 0$, *then with probability at least* $1 - 7\varepsilon$,

$$|\hat{\theta} - \theta^*|_q \leq Cs^{1/q}\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_1 + 1) + Cs^{1/q}\bar{\delta}^2|\theta^*|_1, \tag{9}$$

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

If in addition, $(1+\lambda_2+\lambda_\infty)\{\bar{\delta}^2\lambda_\infty^{-1} + \lambda_2^{-1}\sqrt{\log(p/\varepsilon)/n}\} \leq c_1\kappa_1(s, 1+\lambda_2+\lambda_\infty)$ *for some small enough constant* c_1, *then, with the same probability, we have*

$$|\hat{\theta} - \theta^*|_q \leq Cs^{1/q}\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_2 + 1) + Cs^{1/q}\bar{\delta}^2|\theta^*|_\infty \tag{10}$$

for some constants $C > 0$ *and* $c' > 0$.

Proof. We proceed in three steps. Step 1 establishes initial relations and the fact that $\Delta = \hat{\theta} - \theta^*$ belongs to $C_J(1 + \lambda_2 + \lambda_\infty)$. Step 2 provides a bound on $|\frac{1}{n}X^TX\Delta|_\infty$. Step 3 establishes the rates of convergence stated in the theorem. We work on the event of probability at least $1 - 7\varepsilon$ where all the inequalities in Lemmas 1 and 2 are realized. Throughout the proof, $J = \{j : \theta_j^* \neq 0\}$. We often make use of the inequalities $|\theta|_\infty \leq |\theta|_2 \leq |\theta|_1, \ \forall\theta \in \mathbb{R}^p$.

Step 1. We first note that

$$\begin{aligned} |\tfrac{1}{n}Z^T(y - Z\theta^*)|_\infty &\leq |\tfrac{1}{n}Z^T\xi|_\infty + |\tfrac{1}{n}Z^TW\theta^*|_\infty \\ &\leq \delta_2(\varepsilon) + \delta_3(\varepsilon) + |\tfrac{1}{n}Z^TW\theta^*|_\infty \end{aligned} \tag{11}$$

with probability at least $1-2\varepsilon$ by Lemma 1. Next, Lemma 2 and the fact that, due to (1), we have $|\frac{1}{n}\mathrm{Diag}(W^TW)|_\infty \le \bar{\delta}^2$ imply

$$\begin{aligned}
|\tfrac{1}{n}Z^TW\theta^*|_\infty &\le |\tfrac{1}{n}X^TW\theta^*|_\infty + |\tfrac{1}{n}W^TW\theta^*|_\infty \\
&\le |\tfrac{1}{n}X^TW\theta^*|_\infty + |\tfrac{1}{n}(W^TW-\mathrm{Diag}(W^TW))\theta^*|_\infty \\
&\quad +|\tfrac{1}{n}\mathrm{Diag}(W^TW)\theta^*|_\infty \\
&\le \delta_1'(\varepsilon)|\theta^*|_2 + \delta_4'(\varepsilon)|\theta^*|_2 + \bar{\delta}^2|\theta^*|_\infty.
\end{aligned} \tag{12}$$

Combining (11) and (12) we get that $(\theta, t, u) = (\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ is feasible for the problem (8), so that

$$|\hat{\theta}|_1 + \lambda_2|\hat{\theta}|_2 + \lambda_\infty|\hat{\theta}|_\infty \le |\hat{\theta}|_1 + \lambda_2\hat{t} + \lambda_\infty\hat{u} \le |\theta^*|_1 + \lambda_2|\theta^*|_2 + \lambda_\infty|\theta^*|_\infty. \tag{13}$$

From (13) we easily obtain

$$|\hat{\theta}_{J^c}|_1 \le (1+\lambda_2+\lambda_\infty)|\hat{\theta}_J - \theta^*|_1.$$

Arguments similar to (13) lead to

$$\begin{aligned}
\hat{t} - |\theta^*|_2 &\le \frac{|\Delta|_1 + \lambda_\infty|\Delta|_\infty}{\lambda_2} \le \frac{(1+\lambda_\infty)}{\lambda_2}|\Delta|_1 \quad \text{and} \\
\hat{u} - |\theta^*|_\infty &\le \frac{|\Delta|_1 + \lambda_2|\Delta|_2}{\lambda_\infty} \le \frac{(1+\lambda_2)}{\lambda_\infty}|\Delta|_1.
\end{aligned}$$

Step 2. We have

$$\begin{aligned}
|\tfrac{1}{n}X^TX\Delta|_\infty &\le |\tfrac{1}{n}Z^TX\Delta|_\infty + |\tfrac{1}{n}W^TX\Delta|_\infty \\
&\le |\tfrac{1}{n}Z^TZ\Delta|_\infty + |\tfrac{1}{n}Z^TW\Delta|_\infty + |\tfrac{1}{n}W^TX\Delta|_\infty \\
&\le |\tfrac{1}{n}Z^T(y-Z\theta^*)|_\infty + |\tfrac{1}{n}Z^T(y-Z\hat{\theta})|_\infty + |\tfrac{1}{n}Z^TW\Delta|_\infty \\
&\quad +|\tfrac{1}{n}W^TX\Delta|_\infty.
\end{aligned}$$

The results of Step 1 and of Lemmas 1 and 2 imply the following bounds

$$\begin{aligned}
|\tfrac{1}{n}Z^T(y-Z\theta^*)|_\infty &\le \tau_2|\theta^*|_2 + \bar{\delta}^2|\theta^*|_\infty + \tau, \\
|\tfrac{1}{n}Z^T(y-Z\hat{\theta})|_\infty &\le \tau_2\hat{t} + \bar{\delta}^2\hat{u} + \tau \\
&\le \tau_2|\theta^*|_2 + \bar{\delta}^2|\theta^*|_\infty + \tau \\
&\quad +\{\tau_2(1+\lambda_\infty)/\lambda_2 + \bar{\delta}^2(1+\lambda_2)/\lambda_\infty\}|\Delta|_1, \\
|\tfrac{1}{n}W^TX\Delta|_\infty &\le \delta_1|\Delta|_1, \\
|\tfrac{1}{n}Z^TW\Delta|_\infty &\le |\tfrac{1}{n}X^TW\Delta|_\infty + |\tfrac{1}{n}(W^TW-\mathrm{Diag}(W^TW))\Delta|_\infty \\
&\quad +|\tfrac{1}{n}\mathrm{Diag}(W^TW)\Delta|_\infty \\
&\le \delta_1|\Delta|_1 + \delta_4|\Delta|_1 + \bar{\delta}^2|\Delta|_\infty.
\end{aligned}$$

These relations and the inequality $|\Delta|_\infty \le |\Delta|_1$ yield that

$$\begin{aligned}
|\tfrac{1}{n}X^TX\Delta|_\infty &\le 2\tau_2|\theta^*|_2 + 2\bar{\delta}^2|\theta^*|_\infty + 2\tau + (\bar{\delta}^2\{(1+\lambda_2+\lambda_\infty)/\lambda_\infty\} \\
&\quad +\{(1+\lambda_\infty)/\lambda_2\}\tau_2 + 2\delta_1 + \delta_4)|\Delta|_1.
\end{aligned}$$

Step 3. Next note that $|\Delta|_1 \le |\hat{\theta}|_1 + |\theta^*|_1 \le (2+\lambda_2+\lambda_\infty)|\theta^*|_1$. Letting

$$\eta = \bar{\delta}^2\{(1+\lambda_2+\lambda_\infty)/\lambda_\infty\} + \{(1+\lambda_\infty)/\lambda_2\}\tau_2 + 2\delta_1 + \delta_4,$$

we have

$$|\tfrac{1}{n}X^T X\Delta|_\infty \leq 2\tau + (2\tau_2 + 2\bar{\delta}^2 + (2 + \lambda_2 + \lambda_\infty)\eta)|\theta^*|_1.$$

By the definition of the ℓ_q-sensitivity,

$$|\tfrac{1}{n}X^T X\Delta|_\infty \geq \kappa_q(s, 1 + \lambda_2 + \lambda_\infty)|\Delta|_q.$$

Now, (9) follows by combining the last two displays and the assumption on $\kappa_q(s, 1 + \lambda_2 + \lambda_\infty)$. To prove (10), we use that

$$\begin{aligned} |\tfrac{1}{n}X^T X\Delta|_\infty &\leq 2\tau_2|\theta^*|_2 + 2\bar{\delta}^2|\theta^*|_\infty + 2\tau + \eta|\Delta|_1 \\ &\leq 2\tau_2|\theta^*|_2 + 2\bar{\delta}^2|\theta^*|_\infty + 2\tau + \eta|\tfrac{1}{n}X^T X\Delta|_\infty/\kappa_1(s, 1 + \lambda_2 + \lambda_\infty). \end{aligned}$$

Under the condition that $(1+\lambda_2+\lambda_\infty)\{\bar{\delta}^2\lambda_\infty^{-1} + \lambda_2^{-1}\sqrt{\log(p/\varepsilon)/n}\} \leq c_1\kappa_1(s, 1+\lambda_2+\lambda_\infty)$ for c_1 small enough, by definition of η we have $\eta/\kappa_1(s, 1+\lambda_2+\lambda_\infty) \leq c'$ for some $0 < c' < 1$. Thus, we have

$$|\tfrac{1}{n}X^T X\Delta|_\infty \leq c\big(\tau_2|\theta^*|_2 + \bar{\delta}^2|\theta^*|_\infty + \tau\big),$$

which implies (10) in view of the definition of the ℓ_q-sensitivity and the assumption on $\kappa_q(s, 1 + \lambda_2 + \lambda_\infty)$. □

Remark 2 (Relaxation of Assumption (A4)). We have stated Theorem 1 under Assumption (A4) to make the analysis streamlined with the previous literature, see [8]. However, inspection of the proofs shows that a more general condition can be used. The results of Theorem 1 hold with probability at least $1 - 7\varepsilon - \varepsilon'$ if instead of Assumption (A4) we require W to satisfy:

$$\big|\tfrac{1}{n}\mathrm{Diag}(W^T W)\big|_\infty \leq \bar{\delta}^2,$$

with probability at least $1 - \varepsilon'$, for some $\varepsilon' > 0$.

Compared to [8], the results in Theorem 1 exploit the zero-mean condition on the noise matrix W. As in [8], the estimator is consistent as $\bar{\delta}$ goes to zero. In order to compare the rates in Theorem 1 with those for the MU selector, we recall that, by Theorem 3 in [8], the MU selector satisfies

$$|\hat{\theta}^{MU} - \theta^*|_q \leq Cs^{1/q}\sqrt{\frac{\log(c'p/\varepsilon)}{n}} + Cs^{1/q}(\bar{\delta} + \bar{\delta}^2)|\theta^*|_1$$

with probability close to 1. While both rates share some terms, a term of the order $s^{1/q}\bar{\delta}|\theta^*|_1$ appears only in the rate for the MU selector whereas a term of the order $s^{1/q}\sqrt{\log(c'p/\varepsilon)/n}|\theta^*|_1$ appears only for the $\{\ell_1, \ell_2, \ell_\infty\}$-MU selector. Therefore, the improvement upon the original MU selector can be achieved when $\bar{\delta} \gg \sqrt{\log(c'p/\varepsilon)/n}$.

If the additional condition in the second part of Theorem 1 holds, we can use the bound (10) and a better accuracy is achieved by the proposed estimator. In particular, $|\theta^*|_1$ no longer drives the rate of convergence. The impact of $\bar{\delta}$ on this rate is in the term

$$s^{1/q}\bar{\delta}^2|\theta^*|_\infty \quad \text{instead of} \quad s^{1/q}(\bar{\delta}+\bar{\delta}^2)|\theta^*|_1 \tag{14}$$

for the MU selector. Furthermore, the rate of convergence of the new estimator also has a term of the form $|\theta^*|_2 s^{1/q}\sqrt{\log(c'p/\varepsilon)/n}$. Thus the new estimator obtains a better accuracy by exploiting additional assumptions together with the fact that $\bar{\delta}|\theta^*|_1$ is of larger order than $\sqrt{\log(c'p/\varepsilon)/n}|\theta^*|_2$, which holds whenever $\bar{\delta} \gg \sqrt{\log(c'p/\varepsilon)/n}$. Finally, the impact of going from the ℓ_1-norm to the ℓ_2- or ℓ_∞-norms is not negligible neither. For example, if all non-zero components of θ^* are equal to the same constant $a>0$, we have $|\theta^*|_1 = sa$ while $|\theta^*|_2 = a\sqrt{s}$ and $|\theta^*|_\infty = a$. Then, the comparison in (14) is reduces to comparing

$$s^{1/q}\bar{\delta}^2 \quad \text{versus} \quad s^{1+1/q}(\bar{\delta}+\bar{\delta}^2),$$

featuring the maximum contrast between the two rates.

Finally, note that the conic estimator $\hat{\theta}^{Conic}$ studied in [1] can be also applied under the assumptions of this section. Indeed, our setting can be embedded into that of [1] with $\widehat{D}$ being the identically zero $p\times p$ matrix, which means that we have an estimator of each σ_j^2 with an error bounded by $b=\bar{\delta}^2$. The results in [1] assume $b \le C\sqrt{(\log p)/n}$ but they do not apply to designs with b of larger order. Comparing the bound (10) in Theorem 1 to the bound (6) yields that the conic estimator $\hat{\theta}^{Conic}$ achieves the same rate as our new estimator whenever $\bar{\delta}$ is smaller than or of the order $\{(\log p)/n\}^{1/4}$. However, there is no bound for $\hat{\theta}^{Conic}$ available when $\bar{\delta} \gg \{(\log p)/n\}^{1/4}$.

4. $\{\ell_1,\ell_2,\ell_\infty\}$-compensated MU selector

In this section, we discuss a modification of the conic estimator proposed in [1]. We introduce an additional ℓ_∞-norm regularization to better adapt to the estimation error in $\widehat{D}$. As discussed in the introduction, this is beneficial when the rate of convergence of $\widehat{D}$ to D is slower than $\sqrt{(\log p)/n}$, which is not covered by [1]. Here we consider the same assumptions as in [1] with the only difference that now we allow for any rate of convergence of $\widehat{D}$ to D. Thus, we replace Assumption (A4) by the following assumption on the availability of estimators for σ_j^2, $j=1,\dots,p$:

(A5) There exist statistics $\hat{\sigma}_j^2$ and positive numbers $b(\varepsilon)$ such that for any $0<\varepsilon<1$, we have

$$\mathbb{P}\left(\max_{j=1,\dots,p}|\hat{\sigma}_j^2-\sigma_j^2| \ge b(\varepsilon)\right) \le \varepsilon.$$

In what follows, we fix ε and set

$$\tau_2 = \delta_1'(\varepsilon)+\delta_4'(\varepsilon), \quad \tau = \delta_2(\varepsilon)+\delta_3(\varepsilon) \quad \text{and} \quad \tau_\infty = b(\varepsilon)+\delta_5(\varepsilon).$$

We are particularly interested in cases where the parameter τ_∞ is of larger order than $\sqrt{(\log p)/n}$. To define the estimator, we consider the following minimization problem:

$$\begin{array}{ll} \min_{\theta,t,u} & |\theta|_1 + \lambda_2 t + \lambda_\infty u \\ & (\theta,t,u) \in \Theta, \\ & \left|\frac{1}{n}Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \leq \tau_2 t + \tau_\infty u + \tau, \\ & |\theta|_2 \leq t, |\theta|_\infty \leq u, \end{array} \tag{15}$$

where $\lambda_2 > 0$ and $\lambda_\infty > 0$ are tuning constants and we allow for $(\theta, t, u) \in \Theta$ where Θ is a pre-specified convex set that contains $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$, see Remark 1.

Let $(\hat\theta, \hat t, \hat u)$ be a solution of (15). We take $\hat\theta$ as estimator of θ^* and we call it the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector. The rates of convergence of this estimator are given in the next theorem.

Theorem 2. *Let Assumptions (A1)–(A3), and (A5) hold. Assume that the true parameter θ^* is s-sparse and $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ belongs to Θ. Let $0 < \varepsilon < 1$ and $1 \leq q \leq \infty$. Suppose also that*

$$\kappa_q(s, 1 + \lambda_2 + \lambda_\infty) \geq cs^{-1/q} \tag{16}$$

for some constant $c > 0$ and that

$$s \leq c_1 \min\left\{\frac{\lambda_2\lambda_\infty}{(1+\lambda_2+\lambda_\infty)^3}\sqrt{\frac{n}{\log(p/\varepsilon)}},\ \frac{\lambda_\infty}{(1+\lambda_2+\lambda_\infty)^2}\frac{1}{b(\varepsilon)}\right\}, \tag{17}$$

for some small enough constant $c_1 > 0$. Let $\hat\theta$ be the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector. Then, with probability at least $1 - 8\varepsilon$,

$$|\hat\theta - \theta^*|_q \quad \leq \quad Cs^{1/q}\sqrt{\frac{\log(c'p/\varepsilon)}{n}}(|\theta^*|_2 + 1) + Cs^{1/q}b(\varepsilon)|\theta^*|_\infty, \tag{18}$$

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

Under the same assumptions with $q = 1$, the prediction error admits the following bound, with the same probability:

$$\tfrac{1}{n}\left|X(\hat\theta - \theta^*)\right|_2^2 \quad \leq \quad Cs\frac{\log(c'p/\varepsilon)}{n}(|\theta^*|_2 + 1)^2 + Csb^2(\varepsilon)|\theta^*|_\infty^2. \tag{19}$$

Proof. Throughout the proof, we assume that we are on the event of probability at least $1 - 8\varepsilon$ where the results of Lemmas 3, 4 and 5 in the Appendix hold. Property (26) in Lemma 4 implies that $\Delta = \hat\theta - \theta^*$ is in the cone $C_J(1+\lambda_2+\lambda_\infty)$, where $J = \{j : \theta_j^* \neq 0\}$. Therefore, by the definition of the ℓ_q-sensitivity and Lemma 5, we have

$$\kappa_q(s, 1+\lambda_2+\lambda_\infty)|\Delta|_q \leq \left|\tfrac{1}{n}X^TX\Delta\right|_\infty \leq \mu_0 + \mu_1|\hat\theta - \theta^*|_1 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty,$$

where μ_0 and μ_2 are of the order $\sqrt{\frac{1}{n}\log(c'p/\varepsilon)}$, and μ_1 and μ_∞ are of the order $\sqrt{\frac{1}{n}\log(c'p/\varepsilon)} + b(\varepsilon)$. Using again (26), we have

$$\begin{aligned} |\Delta|_1 &= |\Delta_{J^c}|_1 + |\Delta_J|_1 \le (2+\lambda_2+\lambda_\infty)|\Delta_J|_1 \\ &\le (2+\lambda_2+\lambda_\infty)s^{1-1/q}|\Delta_J|_q \le (2+\lambda_2+\lambda_\infty)s^{1-1/q}|\Delta|_q. \end{aligned}$$

It follows that

$$(\kappa_q(s, 1+\lambda_2+\lambda_\infty) - (2+\lambda_2+\lambda_\infty)\mu_1 s^{1-1/q})|\Delta|_q \le \mu_0 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty,$$

which implies, by (16),

$$(c - (2+\lambda_2+\lambda_\infty)\mu_1 s)s^{-1/q}|\Delta|_q \le \mu_0 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty,$$

in view of the assumptions of the theorem. Recall that

$$\mu_1 \le a\sqrt{\log(c'p/\varepsilon)/n}\{1+(1+\lambda_2)\lambda_\infty^{-1}+(1+\lambda_\infty)\lambda_2^{-1}\}+ab(\varepsilon)\{1+\lambda_\infty+\lambda_2\}/\lambda_\infty,$$

where $a > 0$ is a constant. Therefore, since we assume (17), and $\{1 + (1 + \lambda_2)\lambda_\infty^{-1} + (1+\lambda_\infty)\lambda_2^{-1}\} \le (1+\lambda_\infty+\lambda_2)^2/(\lambda_\infty\lambda_2)$, relation (18) follows if c_1 is small enough.

To prove (19), write first

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

Next remark that from (18) with $q = 1$, we have

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

Lemma 5 in the Appendix yields

$$\left|\tfrac{1}{n}X^TX\Delta\right|_\infty \le \mu_0 + \mu_1|\hat\theta - \theta^*|_1 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty. \tag{20}$$

Combining the above bound for $|\Delta|_1$ and (20), we get

$$\tfrac{1}{n}\,|X\Delta|_2^2 \le C\frac{s\log(c'p/\varepsilon)}{n}(|\theta^*|_2+1)^2 + Csb^2(\varepsilon)|\theta^*|_\infty^2$$

since $\mu_1 s \le C''$ for some constant $C'' > 0$ under our assumptions. This proves (19). □

Theorem 2 generalizes the results in [1] to estimators $\widehat{D}$ that converge with rate $b(\varepsilon)$ of larger order than $\sqrt{(\log p)/n}$. At the same time, if $b(\varepsilon)$ is smaller than $\sqrt{(\log p)/n}$, both the conic estimator $\hat\theta^{Conic}$ of [1] and the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector achieve the same rate of convergence.

For such designs that condition (17) does not hold, the conclusions of Theorem 2 need to be slightly modified as shown in the next theorem.

Theorem 3. *Let Assumptions (A1)–(A3), and (A5) hold. Assume that the true parameter θ^* is s-sparse and $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ belongs to Θ. Let $0 < \varepsilon < 1$ and $1 \leq q \leq \infty$. Let $\hat{\theta}$ be the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector. Then, with probability at least $1 - 8\varepsilon$,*

$$|\hat{\theta} - \theta^*|_q \quad \leq \quad \frac{C}{\kappa_q(s, 1 + \lambda_2 + \lambda_\infty)} \left\{ \sqrt{\frac{\log(c'p/\varepsilon)}{n}} (|\theta^*|_1 + 1) + b(\varepsilon)|\theta^*|_1 \right\},$$

for some constants $C > 0$ and $c' > 0$.

Proof. Again, throughout the proof, we assume that we are on the event of probability at least $1-8\varepsilon$ where the results of Lemmas 3, 4 and 5 in the Appendix hold. Property (26) in Lemma 4 implies that $\Delta = \hat{\theta} - \theta^*$ is in the cone $C_J(1 + \lambda_2 + \lambda_\infty)$, where $J = \{j : \theta_j^* \neq 0\}$. Since

$$|\Delta|_1 \leq |\hat{\theta}|_1 + |\theta^*|_1 \leq \{|\theta^*|_1 + \lambda_2|\theta^*|_2 + \lambda_\infty|\theta^*|_\infty\} + |\theta^*|_1 \leq (2 + \lambda_2 + \lambda_\infty)|\theta^*|_1,$$

we obtain

$$\begin{aligned} \left|\tfrac{1}{n} X^T X \Delta\right|_\infty &\leq \mu_0 + \mu_1|\Delta|_1 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty \\ &\leq \mu_0 + (\mu_1 + \mu_2 + \mu_\infty)(2 + \lambda_2 + \lambda_\infty)|\theta^*|_1. \end{aligned}$$

Therefore

$$\kappa_q(s, 1 + \lambda_2 + \lambda_\infty)|\Delta|_q \leq \mu_0 + (\mu_1 + \mu_2 + \mu_\infty)(2 + \lambda_2 + \lambda_\infty)|\theta^*|_1,$$

which implies the result. □

The result in Theorem 3 parallels the result for generic designs for the conic estimator [1]. Indeed, this result states that the additional ℓ_∞-regularization does not worsen the guarantees obtained in [1]. For generic designs, our bounds do not achieve the previous dependence on $|\theta^*|_2$ and $|\theta^*|_\infty$ and, instead, the final dependence is on $|\theta^*|_1$.

5. Simulations

This section aims to illustrate the finite sample performance of the proposed estimators. We will focus on the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector only. We consider the following data generating process

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

Here, ξ_i, w_i, x_i are independent and $\xi_i \sim \mathcal{N}(0, \sigma^2)$, $w_i \sim \mathcal{N}(0, \sigma_*^2 I_{p\times p})$, $x_i \sim \mathcal{N}(0, \Sigma)$ where $I_{p\times p}$ is the identity matrix and Σ is $p \times p$ matrix with elements $\Sigma_{ij} = \rho^{|i-j|}$. We consider the vector of unknown parameters $\theta^* = 1.25(1, 1, 1, 1, 1, 0, \ldots, 0)^T$. We set $\sigma = 0.128$, $\sigma_*^2 = 0.5$, and $\rho = 0.25$. We assume that σ is known and we set $\widehat{D} = D = \sigma_*^2 I_{p\times p}$. The penalty parameters are set as $\tau = \sigma\sqrt{\log(p/\varepsilon)/n}$, $b(\varepsilon) = \sigma_*^2\sqrt{\log(p/\varepsilon)/n}$, for $\varepsilon = 0.05$.

In our first set of simulations, we illustrate the finite sample performance of the proposed estimator by setting $\lambda_2 = \lambda_\infty \in \{0.25, 0.5, 0.75, 1\}$. The $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU selector will be denoted by $\{\ell_1, \ell_2, \ell_\infty\}$. We compare its performance with other recent proposals in the literature, namely the conic estimator (denoted as Conic (λ_2) for $\lambda_2 = 0.25, 0.5, 0.75, 1$), and the Compensated MU selector (cMU). We also provide the (infeasible) Dantzig selector which knows X (Dantzig X) and the Dantzig selector that uses only Z (Dantzig Z) as additional benchmark for the performance.

Table 1

Simulation results for 100 replications. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR)

	$n = 300$ and $p = 10$			$n = 300$ and $p = 50$		
Method ($\lambda_2 = \lambda_\infty$)	Bias	RMSE	PR	Bias	RMSE	PR
Dantzig X	0.026	0.032	0.034	0.030	0.034	0.038
Dantzig Z	0.589	0.621	0.711	0.603	0.724	0.752
cMU	0.600	0.652	0.737	0.668	0.707	0.814
Conic (.25)	1.926	1.956	2.316	1.995	2.019	2.408
$\{\ell_1, \ell_2, \ell_\infty\}$(.25)	1.792	1.834	2.145	1.903	1.932	2.287
Conic (.5)	0.318	0.416	0.432	0.366	0.439	0.478
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5)	0.213	0.350	0.338	0.348	0.449	0.460
Conic (.75)	0.317	0.415	0.432	0.366	0.439	0.478
$\{\ell_1, \ell_2, \ell_\infty\}$ (.75)	0.208	0.345	0.333	0.269	0.378	0.389
Conic (1)	0.317	0.415	0.432	0.366	0.439	0.477
$\{\ell_1, \ell_2, \ell_\infty\}$ (1)	0.207	0.345	0.332	0.248	0.369	0.373

Table 2

Simulation results for 100 replications. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR)

	$n = 300$ and $p = 100$			$n = 300$ and $p = 300$		
Method ($\lambda_2 = \lambda_\infty$)	Bias	RMSE	PR	Bias	RMSE	PR
Dantzig X	0.031	0.036	0.040	0.034	0.038	0.043
Dantzig Z	0.603	0.836	0.791	0.633	1.077	0.882
cMU	0.690	0.735	0.847	0.722	0.765	0.884
Conic (.25)	2.019	2.042	2.442	2.083	2.098	2.528
$\{\ell_1, \ell_2, \ell_\infty\}$(.25)	1.936	1.964	2.332	2.001	2.024	2.418
Conic (.5)	0.503	0.647	0.639	0.680	0.888	0.836
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5)	0.417	0.520	0.543	0.469	0.550	0.597
Conic (.75)	0.384	0.469	0.508	0.419	0.496	0.542
$\{\ell_1, \ell_2, \ell_\infty\}$ (.75)	0.325	0.431	0.451	0.386	0.474	0.510
Conic(1)	0.3811186	0.467	0.504	0.404	0.484	0.527
$\{\ell_1, \ell_2, \ell_\infty\}$ (1)	0.290	0.415	0.424	0.357	0.456	0.481

Tables 1 and 2 provide the performance of the proposed estimator when $\lambda_2 = \lambda_\infty$ and the performance of various benchmarks. As discussed in the literature, ignoring the error-in-variables issue can lead to worse performance as seen from the performance of Dantzig Z compared to the (infeasible) Dantzig X. The conic compensated estimator performes better than the compensated MU selector (cMU) when $\lambda_2 \in \{0.5, 0.75, 1\}$. The comparison of the proposed estimator and the conic estimator is easier to establish as we can parametrize them by λ_2 (as we set $\lambda_2 = \lambda_\infty$). In this case the conic estimator penalizes more aggressively the uncertainty of not knowing σ_j^2. In essentially all cases[1] the proposed estimator yields improvements. The introduction of ℓ_∞-norm regularization seems to alleviate regularization bias. Nonetheless, when setting $\lambda_2 = 0.25$ both the conic estimator and the proposed estimator fail in the experiment. This failure occurs by not having enough penalty to control $t - |\theta|_2$ and $u - |\theta|_\infty$ which leads to a large right hand side $\tau_2 t + \tau_\infty u + \tau$ in the constraint

$$\left|\tfrac{1}{n} Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \leq \tau_2 t + \tau_\infty u + \tau$$

in (15) and similarly the right hand side $\tau_2 t + \tau$ in (5). In turn, this leads to substantial regularization bias and therefore underfitting. In fact, detailed inspection of estimators in that case reveals that coefficients are very close to zero for both conic and the proposed estimator.

In the second set of simulations, we explore the performance of the proposed estimator for the case $\lambda_2 \neq \lambda_\infty$. Moreover, we also study a modified estimator that contains safeguard constraints. These constraints aim to mitigate the problem discussed above. The safeguard constraints are described in Remark 1. We denote by $\{\ell_1, \ell_2, \ell_\infty\}^*$ the estimator computed with the safeguards.

We consider the same design as before and we explore some combinations of values $(\lambda_2, \lambda_\infty) \in \{0.25, 0.5, 0.75, 1\} \times \{0.25, 0.5, 0.75, 1\}$ for both proposed estimators (with and without the safeguard constraints).

TABLE 3
Simulation results for 100 replications. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR)

	$n = 300$ and $p = 10$			$n = 300$ and $p = 50$		
Method $(\lambda_2, \lambda_\infty)$	Bias	RMSE	PR	Bias	RMSE	PR
$\{\ell_1, \ell_2, \ell_\infty\}$ (1,1)	0.207	0.345	0.332	0.248	0.369	0.373
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (1,1)	0.207	0.345	0.332	0.248	0.369	0.373
$\{\ell_1, \ell_2, \ell_\infty\}$ (1,.5)	0.253	0.399	0.372	0.521	0.708	0.651
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (1,.5)	0.239	0.362	0.356	0.398	0.472	0.511
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,1)	0.207	0.345	0.332	0.244	0.369	0.372
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.5,1)	0.207	0.345	0.332	0.244	0.369	0.372
$\{\ell_1, \ell_2, \ell_\infty\}$ (.75,.75)	0.208	0.345	0.333	0.269	0.378	0.389
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.75,.75)	0.208	0.345	0.333	0.269	0.378	0.389
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,1)	0.207	0.345	0.332	0.243	0.368	0.371
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,1)	0.207	0.345	0.332	0.243	0.368	0.371
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,.5)	0.213	0.350	0.338	0.348	0.449	0.460
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.5,.5)	0.213	0.350	0.338	0.338	0.422	0.449
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,.5)	0.211	0.350	0.336	0.318	0.408	0.431
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,.5)	0.211	0.350	0.336	0.318	0.408	0.431
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,.25)	1.792	1.834	2.145	1.903	1.932	2.287
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,.25)	0.547	0.605	0.678	0.615	0.657	0.753

Tables 3 and 4 show the performance for different values of λ_2 and λ_∞. We note that these parameters seem to have different impact on the finite sample

[1]The conic compensated estimator performes slightly better only with respect to RMSE in the case of $\lambda_2 = 0.5$. For all other parameters and metrics, the proposed estimator performed slightly better or substantially better.

Table 4
Simulation results for 100 replications. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR)

	$n = 300$ and $p = 100$			$n = 300$ and $p = 300$		
Method $(\lambda_2, \lambda_\infty)$	Bias	RMSE	PR	Bias	RMSE	PR
$\{\ell_1, \ell_2, \ell_\infty\}$ (1,1)	0.290	0.415	0.424	0.357	0.456	0.481
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (1,1)	0.290	0.415	0.424	0.357	0.456	0.481
$\{\ell_1, \ell_2, \ell_\infty\}$ (1,.5)	0.670	0.868	0.826	1.099	1.284	1.322
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (1,.5)	0.471	0.546	0.599	0.581	0.644	0.721
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,1)	0.281	0.412	0.418	0.343	0.450	0.470
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.5,1)	0.281	0.412	0.418	0.343	0.450	0.470
$\{\ell_1, \ell_2, \ell_\infty\}$ (.75,.75)	0.325	0.431	0.451	0.386	0.474	0.510
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.75,.75)	0.325	0.431	0.451	0.386	0.474	0.510
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,1)	0.279	0.411	0.416	0.339	0.448	0.467
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,1)	0.279	0.411	0.416	0.338	0.447	0.466
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,.5)	0.417	0.520	0.543	0.469	0.550	0.597
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.5,.5)	0.397	0.481	0.521	0.459	0.531	0.586
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,.5)	0.372	0.464	0.498	0.432	0.508	0.556
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,.5)	0.371	0.463	0.497	0.435	0.511	0.560
$\{\ell_1, \ell_2, \ell_\infty\}$ (.25,.25)	1.936	1.964	2.332	2.001	2.024	2.418
$\{\ell_1, \ell_2, \ell_\infty\}^*$ (.25,.25)	0.636	0.685	0.785	0.666	0.712	0.819

performance even if $\lambda_2 + \lambda_\infty$ is kept constant. Importantly, we observe that the addition of safeguard constraints virtually always leads to improvements although small (even zero sometimes) for most of the tested parameter values. In the case $\lambda_2 < \lambda_\infty$ using safeguard constraints makes almost no difference and overall performance of both estimators is better. In contrast, the estimators perform worse when $\lambda_2 > \lambda_\infty$ and the safeguard constraints lead to improvements. Finally, as expected, the safeguard constraints improve substantially the performance when $\lambda_2 = \lambda_\infty = 0.25$. In that case, the performance becomes comparable to that of the cMU estimator. Essentially, the safeguard constraints help to avoid severe underfitting. They are very helpful when the performance is below of what can be achieved. Nonetheless, we recommend to keep them in all cases as it does not impact negatively the estimator and the additional computational burden seems minimal.

In our third set of simulations, we consider a situation with a "crude" estimator of D. The design is the same as in our first set of simulations, however, in order to compute the estimator of D, we have independent observations of $(X_i, Z_i), i = 1, \ldots, n^*$ where $n^* \leq n$. We consider $p \in \{50, 100\}$ and $n^* = \{100, 250, 500\}$ and $n = 500$. Table 5 shows how the performance of the estimators change as less precise estimates of D are used. For example, consider the average prediction risk (PR) for $p = 50$ as n^* goes from 500 to 100 so that the quality of the estimator worsens. Table 5 shows that the cMU deteriorates by 0.51 (from 1.23 to 0.72), the conic deteriorates by 0.41 (from 0.82 to 0.41), and the proposed estimator deteriorates by 0.24 (from 0.55 to 0.31). Similarly when $p = 100$, we observe that cMU deteriorates by 0.52, the conic deteriorates by 0.47, and the proposed estimator deteriorates by 0.35. In both cases the findings are aligned with the theoretical results that the $\{\ell_1, \ell_2, \ell_\infty\}$-based estimator is less sensitive to the use of a crude estimator for D.

Table 5
Simulation results for 100 replications. For each estimator we provide average bias (Bias), average root-mean squared error (RMSE), and average prediction risk (PR)

	$n = 500$, $n^* = 100$, $p = 50$			$n = 500$, $n^* = 100$, $p = 100$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
cMU	1.0175038	1.040	1.228	1.064	1.090	1.284
Conic (1)	0.666	0.746	0.815	0.749	0.851	0.915
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,1)	0.427	0.500	0.547	0.545	0.620	0.682
	$n = 500$, $n^* = 250$, $p = 50$			$n = 500$, $n^* = 250$, $p = 100$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
cMU	0.752	0.779	0.915	0.788	0.815	0.956
Conic (1)	0.417	0.471	0.532	0.444	0.496	0.562
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,1)	0.261	0.350	0.369	0.284	0.369	0.393
	$n = 500$, $n^* = 500$, $p = 50$			$n = 500$, $n^* = 500$, $p = 100$		
Method	Bias	RMSE	PR	Bias	RMSE	PR
cMU	0.592	0.622	0.726	0.625	0.656	0.763
Conic (1)	0.313	0.376	0.416	0.339	0.397	0.439
$\{\ell_1, \ell_2, \ell_\infty\}$ (.5,1)	0.196	0.302	0.309	0.221	0.321	0.329

Appendix: Auxiliary lemmas

In what follows, we write for brevity $\delta_i = \delta_i(\varepsilon)$, $\delta_i' = \delta_i'(\varepsilon)$, and we set $\Delta = \hat{\theta} - \theta^*$, $J = \{j : \theta_j^* \neq 0\}$.

Lemma 3. *Assume (A1)–(A3) and (A5). Then with probability at least $1 - 6\varepsilon$, the pair $(\theta, t, u) = (\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ belongs to the feasible set of the minimization problem (15).*

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

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

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

$$|\tfrac{1}{n}X^T\xi|_\infty + |\tfrac{1}{n}W^T\xi|_\infty \le \delta_2 + \delta_3 \tag{21}$$

$$|(\tfrac{1}{n}\mathrm{Diag}\{W^T W\} - D)\theta^*|_\infty \le |\tfrac{1}{n}\mathrm{Diag}\{W^T W\} - D|_\infty |\theta^*|_\infty \le \delta_5 |\theta^*|_\infty \tag{22}$$

$$|(\widehat{D} - D)\theta^*|_\infty \le b(\varepsilon)|\theta^*|_\infty, \tag{23}$$

where in (22) and (23) we have used that the considered matrices are diagonal. Also, by Lemma 2, with probability at least $1 - 2\varepsilon$, we have

$$|\tfrac{1}{n}X^T W\theta^*|_\infty \le \delta_1' |\theta^*|_2 \tag{24}$$

$$|\tfrac{1}{n}(W^T W - \mathrm{Diag}\{W^T W\})\theta^*|_\infty \le \delta_4' |\theta^*|_2. \tag{25}$$

Combining the decomposition of $Z^T(y - Z\theta^*) + n\widehat{D}\theta^*$ together with (21)–(25), we find that

$$\left|\tfrac{1}{n}Z^T(y - Z\theta^*) + \widehat{D}\theta^*\right|_\infty \le \tau_2 |\theta^*|_2 + \tau_\infty |\theta^*|_\infty + \tau,$$

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

Lemma 4. *Let $\hat{\theta}$ be the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU-selector. Assume (A1)–(A3) and (A5). Then with probability at least $1 - 6\varepsilon$ (on the same event as in Lemma 3), we have*

$$|(\hat{\theta} - \theta^*)_{J^c}|_1 \le (1 + \lambda_2 + \lambda_\infty)|(\hat{\theta} - \theta^*)_J|_1, \tag{26}$$

$$\hat{t} - |\theta^*|_2 \le \{(1 + \lambda_\infty)/\lambda_2\}|\hat{\theta} - \theta^*|_1 \quad \textit{and} \quad \hat{u} - |\theta^*|_\infty \le \{(1 + \lambda_2)/\lambda_\infty\}|\hat{\theta} - \theta^*|_1. \tag{27}$$

Proof. Set $\Delta = \hat{\theta} - \theta^*$. On the event of Lemma 3, $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ belongs to the feasible set of the minimization problem (5). Consequently,

$$|\hat{\theta}|_1 + \lambda_2|\hat{\theta}|_2 + \lambda_\infty|\hat{\theta}|_\infty \le |\hat{\theta}|_1 + \lambda_2\hat{t} + \lambda_\infty\hat{u} \le |\theta^*|_1 + \lambda_2|\theta^*|_2 + \lambda_\infty|\theta^*|_\infty. \tag{28}$$

This implies

$$\begin{aligned} |\Delta_{J^c}|_1 + \lambda_2|\Delta_{J^c}|_2 + \lambda_\infty|\Delta_{J^c}|_\infty &\le |\Delta_J|_1 + \lambda_2|\Delta_J|_2 + \lambda_\infty|\Delta_J|_\infty \\ &\le (1 + \lambda_2 + \lambda_\infty)|\Delta_J|_1, \end{aligned}$$

and so

$$|\Delta_{J^c}|_1 \le (1 + \lambda_2 + \lambda_\infty)|\Delta_J|_1.$$

and (26) follows. To prove (27), it suffices to note that (28) implies

$$\begin{aligned} \lambda_2\hat{t} &\le |\theta^*|_1 - |\hat{\theta}|_1 + \lambda_2|\theta^*|_2 + \lambda_\infty|\theta^*|_\infty - \lambda_\infty\hat{u} \\ &\le |\hat{\theta} - \theta^*|_1 + \lambda_2|\theta^*|_2 + \lambda_\infty|\hat{\theta} - \theta^*|_\infty \end{aligned}$$

and the result follows since $|\hat{\theta}|_\infty \le \hat{u}$ and $|\hat{\theta} - \theta^*|_\infty \le |\hat{\theta} - \theta^*|_1$. Similar calculations yield the bound for $\hat{u}$. □

Lemma 5. *Let $\hat{\theta}$ be the $\{\ell_1, \ell_2, \ell_\infty\}$-compensated MU-selector. Assume (A1)–(A3) and (A5). Then, on a subset of the event of Lemma 3 having probability at least $1 - 8\varepsilon$, we have*

$$\left|\tfrac{1}{n}X^T X(\hat{\theta} - \theta^*)\right|_\infty \le \mu_0 + \mu_1|\hat{\theta} - \theta^*|_1 + \mu_2|\theta^*|_2 + \mu_\infty|\theta^*|_\infty,$$

where $\mu_0 = \tau + \delta_2 + \delta_3$, $\mu_1 = 2\delta_1 + \delta_4 + \delta_5 + b(\varepsilon) + \{(1 + \lambda_\infty)/\lambda_2\}\tau_2 + \{(1 + \lambda_2)/\lambda_\infty\}\tau_\infty$, $\mu_2 = \tau_2 + \delta_1'$, $\mu_\infty = \tau_\infty + b(\varepsilon) + \delta_4' + \delta_5$.

Note that μ_0 and μ_2 are of the order $\sqrt{\frac{1}{n}\log(c'p/\varepsilon)}$, and μ_1 and μ_∞ are of the order $\sqrt{\frac{1}{n}\log(c'p/\varepsilon)} + b(\varepsilon)$.

Proof. Throughout the proof, we assume that we are on the event of probability at least $1 - 6\varepsilon$ where inequalities (21)–(25) hold and $(\theta^*, |\theta^*|_2, |\theta^*|_\infty)$ belongs to the feasible set of the minimization problem (15). We have

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

Using the fact that $(\hat\theta, \hat t, \hat u)$ belongs to the feasible set of the minimization problem (5) together with (27), we obtain

$$\begin{aligned}|\tfrac{1}{n}Z^T(Z\hat\theta - y) - \widehat{D}\hat\theta|_\infty &\le \tau_2\hat t + \tau_\infty \hat u + \tau\\ &\le \{(1+\lambda_\infty)/\lambda_2\}\tau_2|\Delta|_1 + \tau_2|\theta^*|_2\\ &\quad + \{(1+\lambda_2)/\lambda_\infty\}\tau_\infty|\Delta|_1 + \tau_\infty|\theta^*|_\infty + \tau.\end{aligned}$$

Using that $\hat\theta = \theta^* + \Delta$, Assumption (A5) together with (23) yields that

$$\begin{aligned}|\tfrac{1}{n}X^TX\Delta|_\infty \le&\{(1+\lambda_\infty)/\lambda_2\}\tau_2|\Delta|_1 + \tau_2|\theta^*|_2 + \{(1+\lambda_2)/\lambda_\infty\}\tau_\infty|\Delta|_1\\ &+ \tau_\infty|\theta^*|_\infty + \tau + |(\tfrac{1}{n}Z^TW - D)\hat\theta|_\infty + |(\widehat{D} - D)\hat\theta|_\infty\\ &+ |\tfrac{1}{n}Z^T\xi|_\infty + |\tfrac{1}{n}W^TX\Delta|_\infty\\ \le&\{(1+\lambda_\infty)/\lambda_2\}\tau_2|\Delta|_1 + \tau_2|\theta^*|_2 + \{(1+\lambda_2)/\lambda_\infty\}\tau_\infty|\Delta|_1\\ &+ \tau_\infty|\theta^*|_\infty + \tau + |(\tfrac{1}{n}Z^TW - D)\hat\theta|_\infty + b(\varepsilon)|\theta^*|_\infty + b(\varepsilon)|\Delta|_1\\ &+ \delta_2 + \delta_3 + |\tfrac{1}{n}W^TX\Delta|_\infty.\end{aligned}$$

Now remark that $|(\frac{1}{n}Z^TW - D)\hat\theta|_\infty \le |(\frac{1}{n}Z^TW - D)\Delta|_\infty + |(\frac{1}{n}Z^TW - D)\theta^*|_\infty$. In view of Lemma 2 and (22), on the initial event of probability at least $1-6\varepsilon$,

$$\begin{aligned}&|(\tfrac{1}{n}Z^TW - D)\theta^*|_\infty\\ \le&|\tfrac{1}{n}(W^TW - \mathrm{Diag}\{W^TW\})\theta^*|_\infty + |(\tfrac{1}{n}\mathrm{Diag}\{W^TW\} - D)\theta^*|_\infty \qquad (29)\\ &+\ |\tfrac{1}{n}X^TW\theta^*|_\infty\\ \le&(\delta_4' + \delta_5)|\theta^*|_\infty + \delta_1'|\theta^*|_2. \qquad (30)\end{aligned}$$

Moreover, we have

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

Therefore,

$$|(\tfrac{1}{n}Z^TW - D)\Delta|_\infty \le (\delta_1 + \delta_4 + \delta_5)|\Delta|_1, \qquad (31)$$

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

$$|\tfrac{1}{n}W^TX\Delta|_\infty \le |\tfrac{1}{n}X^TW|_\infty|\Delta|_1 \le \delta_1|\Delta|_1. \qquad (32)$$

To complete the proof, it suffices to plug (30)–(32) in the last inequality for $|\frac{1}{n}X^TX\Delta|_\infty$ and to obtain

$$\begin{aligned}&|\tfrac{1}{n}X^TX\Delta|_\infty\\ &\le [2\delta_1 + \delta_4 + \delta_5 + b(\varepsilon) + \{(1+\lambda_\infty)/\lambda_2\}\tau_2 + \{(1+\lambda_2)/\lambda_\infty\}\tau_\infty]|\Delta|_1\\ &+ \{\tau_2 + \delta_1'\}|\theta^*|_2 + \{\tau_\infty + b(\varepsilon) + \delta_4' + \delta_5\}|\theta^*|_\infty + \tau + \delta_2 + \delta_3. \qquad \square\end{aligned}$$

Acknowledgement

The work of A.B. Tsybakov was supported by GENES, by the French National Research Agency (ANR) under the grants ANR-11-IDEX-0003-02, Labex ECODEC (ANR-11-LABEX-0047), IPANEMA (ANR-13-BSH1-0004-02), and by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

References

[1] Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2014) Linear and conic programming approaches to high-dimensional errors-in-variables models. arxiv:1408.0241

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

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

[4] Freedman, L. S., Midthune, D., Carroll, R. J. and Kipnis, V. (2008) A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, **27**, 5195–5216. MR2516750

[5] Gautier, E. and Tsybakov, A. B. (2011) High-dimensional instrumental variables regression and confidence sets. arxiv:1105.2454 MR2867623

[6] Gautier, E. and Tsybakov, A. B. (2013) Pivotal estimation in high-dimensional regression via linear programming. In: *Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, Schölkopf, B., Luo, Z., Vovk, V. eds., 195–204. Springer, New York e.a. MR3236866

[7] Loh, P.-L. and Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, **40**, 1637–1664. MR3015038

[8] Rosenbaum, M. and Tsybakov, A. B. (2010) Sparse Recovery under Matrix Uncertainty. *Annals of Statistics*, **38**, 2620–2651. MR2722451

[9] Rosenbaum, M. and Tsybakov, A. B. (2013) Improved matrix uncertainty selector. In: *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, Banerjee, M. et al. ed., *IMS Collections*, vol. 9, 276–290, Institute of Mathematical Statistics. MR3202640

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

[11] Sørensen, Ø., Frigessi, A. and Thoresen, M. (2014) Covariate selection in high-dimensional generalized linear models with measurement error. arxiv:1407.1070