CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-26

Pivotal Estimation Via Self-Normalization for High-Dimensional Linear Models with Errors in Variables

A. BELLONI[1]
V. CHERNOZHUKOV[2]
A. KAUL[3]
M. ROSENBAUM[4]
A. B.TSYBAKOV[5]

Les documents de travail ne reflètent pas la position du CREST et n'engagent que leurs auteurs.
Working papers do not reflect the position of CREST but only the views of the authors.

[1] Duke's Fuqua School of Business. E-mail: abn5@duke.edu

[2] MIT. E-mail : E-mail : vchern@mit.edu

[3] IBM. E-mail : abhishek.kaul@in.ibm.com

[4] CREST ; CEMAP ; Polytechnique. E-mail : mathieu.rosenbaum@polytechnique.edu

[5] CREST; ENSAE; CMC. E-mail: Alexandre.tsybakov@ensae.fr

Submitted to the Annals of Statistics

PIVOTAL ESTIMATION VIA SELF-NORMALIZATION FOR HIGH-DIMENSIONAL LINEAR MODELS WITH ERRORS IN VARIABLES

By Alexandre Belloni, Victor Chernozhukov, Abhishek Kaul,
Mathieu Rosenbaum and Alexandre B. Tsybakov

We propose a new estimator for the high-dimensional linear regression model with measurement error in the design where the number of coefficients is potentially larger than the sample size. The main novelty of our procedure is that the choice of penalty parameters is pivotal. The estimator is based on applying a self-normalization to the constraints that characterize the estimator. Importantly, we show how to cast the computation of the estimator as the solution of a convex program with second order cone constraints. This allows the use of algorithms with theoretical guarantees and enables reliable implementation. Under sparsity assumptions, we derive ℓ_q-rates of convergence and show that consistency can be achieved even if the number of regressors exceeds the sample size. We further provide a simple thresholded estimator that yields a provably sparse estimator with similar ℓ_2 and ℓ_1-rates of convergence.

1. Introduction. In this paper, we consider the high-dimensional linear model with observation error in the design

$$y_i = x_i^T\beta_0 + \xi_i, \quad z_i = x_i + w_i, \quad i = 1, \dots, n, \tag{1.1}$$

where we observe the response variable y_i and the p-dimensional vector z_i, and do not observe the covariates x_i. The scalar errors ξ_i are zero-mean independent random variables and (y_i, z_i) are independent across i. The vector $\beta_0 \in \mathbb{R}^p$ is a vector of unknown parameters to be estimated where the dimension p can be much larger than the sample size n. We assume that β_0 is s-sparse, that is it has at most s non-zero components. The errors in measurements w_i are assumed to be zero-mean and independent of ξ_i. We also assume that the error in measurement covariance matrix $\Gamma = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(w_i w_i^T)$ is diagonal and admits a data-driven estimator $\hat\Gamma$ which is available in several applications as discussed below.

Model (1.1) is motivated by many applications where the covariates may have missing values or are observed with noise. For example, in the field of genomics, the gene expression measurements from microarray data are subject to measurement error. Another example is that of microbiome data where each observation vector

*First Version: 4/10/2016; Current Version: October 11, 2017.

Keywords and phrases: pivotal estimation, self-normalized sums, error in measurement, high-dimensional models

has a significant proportion of missing components. Many other examples arise in empirical economics and finance, see [17] and [25], and consumer surveys in marketing where random subsets of questions are selected for each consumer to reduce the length of the survey. In these settings a data-driven estimator $\hat{\Gamma}$ can be constructed based on auxiliary data without measurement errors [24, 8] or even based on $(y_i, z_i)_{i=1}^n$ alone as in the case of missing at random (as one can estimate the frequencies of missing components, see, e.g. [26, 2]). It has been well-documented that ignoring this measurement error leads to biased parameter estimates even in the fixed p setting, see for example, [17], [13], and [23]. In the high-dimensional framework considered here it is also crucial to account for such measurement errors. In addition to potentially biased estimation, measurement errors can also impact variables selection performance and influence the choice of various penalty parameters, see [29].

High-dimensional linear models with $p \gg n$ and measurement errors have been studied recently by [1], [4], [9], [10], [11],[19], [20], [22], [25], [26], [27] and [28]. The common thread[1] of these papers is to provide estimators along with the corresponding rates of convergence in different norms. Examples of proposed estimators[2] include the orthogonal matching pursuit as defined in [9], the non-convex ℓ_1-penalized regression studied in [22] and the conic programming estimator considered in [1]. In particular, under suitable conditions and appropriate choice of penalty parameters, some of these estimators $\tilde{\beta}$ can attain ℓ_q-rates of convergence of the form

$$\|\tilde{\beta} - \beta_0\|_q \le C(1 + \|\beta_0\|_2) s^{1/q} \sqrt{\frac{\log p}{n}}, \quad 1 \le q \le \infty, \tag{1.2}$$

where $\|\cdot\|_q$ denotes the ℓ_q-norm, and $C > 0$ is a constant independent of s, p and n. It is shown in [1] that these rates are minimax optimal. The rate in (1.2) highlights the impact of the errors in measurements via the ℓ_2-norm term $\|\beta_0\|_2$, which is not present in the case where covariates are observed without error, and the fact that consistency can be achieved in high-dimensional settings even if $p \gg n$. However, estimators suggested in the literature rely on suitable choice of penalty parameters based on some specific knowledge of the model (1.1). To construct these estimators, the variance of the unobserved noise ξ_i and the variance parameters of the measurement noise w_i should typically be known. For some methods, in addition, one needs to have access to the number s of non-zero components or to the ℓ_2-norm of β_0.

In this work, we propose a new estimator of the parameter β_0 in model (1.1) that achieves the optimal rates of convergence in ℓ_q-norm under suitable condi-

[1] We note that all the cited papers assume independent observations except for [27] that allows for the measurement error for each covariate to be a dependent vector across observations.

[2] These estimators were proposed under various conditions on the design matrix, relations between s, p and n, and knowledge of some parameters of the problem.

tions. Moreover, a simple thresholded version of the estimator achieves optimal sparsity, while retaining optimal convergence rates. The main novelty of our procedure is the pivotality of the penalty parameters, which makes the estimator particularly appealing for the practical applications. That is, the penalty parameters do not depend on the number of non-zero components, on the ℓ_2-norm of β_0, the variance parameter of the errors ξ_i, nor on the variance parameters of the errors in the measurements w_i. Furthermore, our estimator is a solution of a convex optimization problem.

Notation. Let $J \subseteq \{1, \dots, p\}$ be a set of integers. We denote by $|J|$ the cardinality of J. For a vector $\theta = (\theta_1, \dots, \theta_p)^T$ 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. We will call θ_J the restriction of θ to J. We denote the ℓ_q-norm of a vector $v \in \mathbb{R}^p$ by $\|v\|_q$. The number of non-zero components of a vector $v \in \mathbb{R}^p$ is denoted by $\|v\|_0$. A centered random variable ξ will be called zero-mean subgaussian with variance parameter σ^2 if $\mathrm{E}[\exp(t\xi)] \leq \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$. A random vector $w \in \mathbb{R}^n$ will be called zero-mean subgaussian with variance parameter σ^2 if all the random variables of the form $v^T w$ where $\|v\|_2 = 1$ are zero-mean subgaussian with variance parameter σ^2. For a matrix A, we denote by $A_{i\cdot}$ and $A_{\cdot j}$ its i-th row and j-th column, respectively. We denote by C, c, C', c' positive constants that can be different on different occurencies.

2. Estimator via self-normalization.

Here we propose a pivotal estimator that does not require knowledge of typically unknown parameters. Our starting point is the moment condition that characterizes the vector of parameters β_0 in (1.1)

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[z_i(y_i - z_i^T\beta_0) + \Gamma\beta_0] = 0, \tag{2.1}$$

where the term $\Gamma\beta_0$ corrects the bias that arises from using the noisy covariates z instead of the unobserved x. The moment condition (2.1) combined with sparsity assumptions on β_0 motivates the use of penalized methods to cope with high-dimensionality.

We now describe the proposed estimation procedure. Let $(\hat\beta, \hat t, \hat u)$ be a solution of the constrained minimization problem

$$\min_{\beta\in\mathbb{R}^p, t\in\mathbb{R}^p, u\in\mathbb{R}^p} \|\beta\|_1 + \lambda_t\|t\|_\infty + \lambda_u\|u\|_\infty : \tag{2.2}$$

$$\begin{array}{l} \left|\frac{1}{n}\sum_{i=1}^n z_{ij}(y_i - z_i^T\beta) + \hat\Gamma_{jj}\beta_j\right| \leq \tau t_j + (1+\tau)b_\epsilon u_j, \\ \left\{\frac{1}{n}\sum_{i=1}^n \{z_{ij}(y_i - z_i^T\beta) + \hat\Gamma_{jj}\beta_j\}^2\right\}^{1/2} \leq t_j, \quad |\beta_j| \leq u_j, \end{array} , \quad \forall j \leq p, \tag{2.3}$$

where $z_i = (z_{ij})_{j=1}^p, \beta = (\beta_j)_{j=1}^p, u = (u_j)_{j=1}^p, t = (t_j)_{j=1}^p$, and λ_t, λ_u, τ are positive tuning parameters set according to Theorem 1 below. As it is standard in the literature, the statistics $(\hat{\Gamma}_{jj})_{j=1}^p$ are given estimators of the diagonal elements of matrix Γ with b_ϵ being a bound on its precision satisfying, for any n,

$$\mathrm{P}(\|\hat{\Gamma} - \Gamma\|_\infty > b_\epsilon) \leq \epsilon \tag{2.4}$$

where $\epsilon \in (0,1)$ is a given number. We use $\hat{\beta}$ as an estimator of β_0 and we call it the *self-normalized conic estimator.*

The proposed method has a self-normalization feature, which is related to the square-root Lasso [3, 6, 30], the STIV estimator [14] and the self-tuned Dantzig estimator [16]. A key point is the direct use of self-normalization in the moment condition (2.1), cf. the second line of constraints in (2.3), instead of working with one scalar noise level as self-normalization quantity. This idea appeared already in the context of instrumental variable regression, cf. (9.22) in [14], as well as in [15] that deals with linear model with no measurement errors and studies a program close to (2.2) - (2.3) with $\hat{\Gamma}_{jj} \equiv 0$, $u_j \equiv 0$.

Importantly, (2.2) - (2.3) is a tractable convex optimization problem with linear and second order cone constraints, for which computationally efficient solvers exist. In particular, we used the R package Rmosek for the computation of the estimator.

In some settings, it is of interest to work with estimators that are also sparse. However, the use of many second order constraints makes unlikely that the estimator $\hat{\beta}$ defined by solving (2.2)-(2.3) is sparse. In such cases, we propose a thresholded version of the self-normalized conic estimator $\hat{\beta}$. Consider the set of components $\hat{T}$ defined as

$$\hat{T} := \left\{ j \in \{1,\ldots,p\} : |\hat{\beta}_j| > \tau \frac{\left\{\frac{1}{n}\sum_{i=1}^n \{z_{ij}(y_i - z_i^T\hat{\beta}) + \hat{\Gamma}_{jj}\hat{\beta}_j\}^2\right\}^{1/2}}{\frac{1}{n}\sum_{i=1}^n z_{ij}^2} \right\} \tag{2.5}$$

where $\hat{\beta}_j$ are the components of $\hat{\beta}$. We define the thresholded self-normalized conic estimator as $\hat{\beta}_{\hat{T}}$ (the restriction of $\hat{\beta}$ to $\hat{T}$).

3. Main results.

In this section we state our assumptions and main theoretical results.

3.1. *Regularity conditions.*

In what follows, we consider a setting where s and p depend on n, and we state the results in the asymptotics as n tends to infinity. Condition A below summarizes the assumptions on the data generating process.

Condition A. *(i) The $n \times p$ matrix $X = [x_1; \ldots; x_n]^T$ is deterministic and the vector β_0 satisfies $\|\beta_0\|_0 \leq s$. (ii) The elements of the random noise vector*

$\xi = (\xi_i)_{i=1}^n$ are independent zero-mean subgaussian random variables with variance parameter $\sigma_\xi^2 \leq C$. (iii) The measurement error vectors $(w_i)_{i=1}^n$ are independent zero-mean subgaussian random vectors with variance parameter $\sigma_w^2 \leq C$, having zero covariances: $\mathrm{E}[w_{ij}w_{ik}] = 0$ for all $1 \leq j < k \leq p$, $i = 1, \ldots, n$. Moreover, $(w_i)_{i=1}^n$ are independent of $\xi = (\xi_i)_{i=1}^n$.

Condition A(i) assumes a deterministic design and the performance of our estimator depends on the Gram matrix $\Psi = \frac{1}{n}X^TX$. Like in problems without measurement errors, some characteristics of Ψ play a key role in the analysis, see [7]. In this paper, we consider ℓ_q-sensitivity characteristics defined for $q \geq 1$ as

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

where $C_J(u) = \{\Delta \in \mathbb{R}^p : \|\Delta_{J^c}\|_1 \leq u\|\Delta_J\|_1\}$, $u > 0$ and $J \subseteq \{1,\ldots,p\}$. These sensitivity characteristics generalize other well known characteristics such as the restricted eigenvalues of [7]. One can find details on their properties in [14] where the notion of sensitivity characteristic was introduced. Sensitivity characteristics have been used previously in several papers including [14, 16, 25, 26] and [1]. For well-behaved designs that are prevalent in the literature, we have $\kappa_q(s,u) \geq cs^{-1/q}$ for $u \geq 1$ and $q \in [1,2]$, where $c > 0$ is a constant, see [14]. Conditions A(ii) and A(iii) are standard in the literature on high-dimensional linear regression with errors in measurements, see [1, 22] among others. Moreover, our analysis relies on the quantity $m_2 := \max_{j=1,\ldots,p}\frac{1}{n}\sum_{i=1}^n x_{ij}^2$ which is typically uniformly bounded for many designs of interest.

For $i = 1,\ldots,n$ and $j = 1,\ldots,p$, we define

$$U_{ij} = z_{ij}(\xi_i - w_i^T\beta_0) + \mathrm{E}[w_{ij}^2]\beta_{0j}$$

and set $\mathcal{U}_{jk} = \{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|U_{ij}|^k]\}^{1/k}$ and $\Delta_j = |\beta_{0j}|\max_{i=1,\ldots,n}|\mathrm{E}[w_{ij}^2] - \Gamma_{jj}|$.

We now state the assumptions on $\mathcal{U}_{jk}$, Δ_j, p and n. Let $\Phi(\cdot)$ denote the standard normal c.d.f.

Condition B. *(i) The estimator $\hat{\Gamma}$ is a diagonal matrix and b_ϵ satisfies (2.4). For some positive sequence ℓ_n tending to infinity, the following conditions hold:*
(ii) $\max_{1\leq j\leq p}\{\mathcal{U}_{j3}/\mathcal{U}_{j2}\}\Phi^{-1}(1-\alpha/(2pn)) \leq n^{1/6}/\ell_n$, *and*
(iii) $\max_{j:\Delta_j>0}\{n^{-1/6}(\Delta_j/\mathcal{U}_{j2}) + n^{-1/2}(\mathcal{U}_{j3}/\mathcal{U}_{j2})^6\}\Phi^{-1}(1-\alpha/(2pn)) \leq 1/\ell_n$.

Condition B(i) allows for the use of data-driven estimator of Γ. Condition B(ii) is a mild moment condition and allows the application of self-normalized moderate deviation theory, see [12, 18]. In the case of i.i.d. sampling where the covariates x_i are also drawn from a subgaussian distribution with bounded variance parameter, we have $\mathrm{E}[|U_j|^3]^{1/3} \leq C(1 + \|\beta_0\|_2)$ and $\mathrm{E}[|U_j|^2]^{1/2} \geq c$ if ξ is independent of z.

In fact for many designs we have $\max_{j\leq p}\{\mathrm{E}[|U_j|^3]^{1/3}/\mathrm{E}[|U_j|^2]^{1/2}\} \leq C$ so that Condition B(ii) is satisfied provided $\log^3 p = o(n)$. Condition B(iii) provides a mild sufficient condition to handle the non-i.i.d. case where the terms $\mathrm{E}[w_{ij}^2]$ change across i as well. We also view Condition B(iii) a mild moment condition as it is implied by Condition A, B(ii), $\|\beta_0\|_\infty \leq C$ and $\max_{j=1,\ldots,p}\mathcal{U}_{j3}/\mathcal{U}_{j2} \leq C$.

We also introduce a (computable) data-driven quantity

$$H_n = \max_{j=1,\ldots,p}\left\| \frac{1}{n}\sum_{i=1}^n (z_{ij}z_i^T - \hat{\Gamma}_{j\cdot})^T(z_{ij}z_i^T - \hat{\Gamma}_{j\cdot})\right\|_\infty^{1/2}.$$

Let h_ϵ denote its $(1-\epsilon)$-quantile, so that $\mathrm{P}(H_n > h_\epsilon) \leq \epsilon$. Here and in what follows, we assume that $\epsilon \in (0,1)$ is a fixed small number. We will not require the knowledge of h_ϵ to implement the method. We will only need that $h_\epsilon \tau s = o(1)$ as $n \to \infty$. This is implied by mild moment conditions on z_i's and growth conditions on p and s. For many designs, h_ϵ is uniformly bounded as $\epsilon \to 0$ and as the sample size grows (see Lemma 3 in the Appendix). Finally, in order to state our theoretical results below, it will be convenient to define for any $\beta \in \mathbb{R}^p$ the vector $t(\beta) = (t_j(\beta))_{j=1}^p$ where

$$t_j(\beta) := \left\{\frac{1}{n}\sum_{i=1}^n \{z_{ij}(y_i - z_i^T\beta) + \hat{\Gamma}_{j\cdot}\beta\}^2\right\}^{1/2}, \quad j = 1, \ldots, p.$$

3.2. *Properties of self-normalized conic estimator.* The following theorem establishes the rates of convergence of the estimator $\hat{\beta}$.

THEOREM 1. *Let $0 < \alpha < 1$, $0 < \varepsilon < 1$, and $1 \leq q \leq \infty$. Set $\tau = n^{-1/2}\Phi^{-1}(1-\alpha/(2p))$, $\lambda_u = 1/4$ and $\lambda_t = 1/\{4H_n\}$. Assume that*

$$\kappa_q(s,3)s^{1/q} \geq 8s\{(1+\tau)b_\epsilon + \tau h_\epsilon + C'(1+m_2^{1/2})\sqrt{\log(2p^2/\varepsilon)/n}\}$$

where $C' > 0$ is a constant that depends only on σ_w and σ_ξ. Then, under Conditions A and B, for n sufficiently large, with probability at least $1 - \alpha\{1+o(1)\} - 2\epsilon - 9\varepsilon$ we have

$$\|\hat{\beta} - \beta_0\|_q \leq \frac{\tau\|t(\beta_0)\|_\infty}{c'\kappa_q(s,3)} + \frac{(1+\|\beta_0\|_2)(1+m_2^{1/2})}{c'\kappa_q(s,3)}\sqrt{\frac{\log(2p/\varepsilon)}{n}} + \frac{b_\epsilon\|\beta_0\|_\infty}{c'\kappa_q(s,3)},$$

where the constant $c' > 0$ depends only on σ_w and σ_ξ.

Theorem 1 provides a bound on the ℓ_q-rate of convergence that depends on the critical quantities of the data generating process. Indeed, it is characterized via κ_q, m_2, $\|t(\beta_0)\|_\infty$,$\|\beta_0\|_2$, $\|\beta_0\|_\infty$ and b_ϵ which summarizes how good the estimate

$\hat{\Gamma}$ is. The impact of using an estimate $\hat{\Gamma}$ of Γ has a factor of $\|\beta_0\|_\infty$ instead of $\|\beta_0\|_2$.

The next corollary specifies the result of Theorem 1 for the standard configuration of the problem usually considered in the literature. It is described by the following conditions: $b_\epsilon \leq C\sqrt{\log(2p/\epsilon)/n}$, $m_2 \leq C$, $\|t(\beta_0)\|_\infty \leq C(1+\|\beta_0\|_2)$ and $\kappa_q(s,3) \geq cs^{-1/q}$ for $q \in [1,2]$ with high probability. Define $\Omega_X := \{X : \kappa_q(s,3) \geq cs^{-1/q}, \max_{1\leq j\leq p} \frac{1}{n}\sum_{i=1}^n x_{ij}^4 \leq C\}$. We have the following result.

Corollary 1. *Assume that the probability that the design matrix X belongs to Ω_X tends to 1 as $n \to \infty$, and that $b_\epsilon \leq C\sqrt{\log(2p/\epsilon)/n}$. Set $\varepsilon = \epsilon$. Then, under the assumptions of Theorem 1, for n sufficiently large, with probability at least $1-\alpha-11\varepsilon-o(1)$ we have*

$$\|\hat{\beta}-\beta_0\|_q \leq C(1+\|\beta_0\|_2)s^{1/q}\sqrt{\frac{\log(c'p/(\alpha\varepsilon))}{n}}$$

where $C > 0, c' \geq 1$ are constants.

Note that Corollary 1 exhibits the minimax rate of convergence discussed in (1.2). A major point is that the estimator $\hat{\beta}$ achieves the minimax rate without needing to know $\|\beta_0\|_2$, σ_ξ, σ_w or s (or invoking cross-validation) as required for the procedures previously suggested in the literature. Note that cross-validation in the problem that we consider here remains unjustified theoretically.

Remark 1. *If the condition on $\kappa_q(s,3)$ required in Corollary 1 is not satisfied, following the same argument as in [1], we can derive a rate of convergence that depends on $\|\beta_0\|_1$ instead of $\|\beta_0\|_2$, namely*

$$\|\hat{\beta}-\beta_0\|_q \leq \frac{C(1+\|\beta_0\|_1)}{\kappa_q(s,3)}\Big(\sqrt{\frac{\log(c'p)}{n}}+b_\epsilon\Big).$$

Next, we consider the data-driven thresholded estimator $\hat{\beta}_{\widehat{T}}$ based on $\widehat{T}$ defined in (2.5), and we show that it achieves the sparsity $O(s)$, while preserving the optimal ℓ_1 and ℓ_2 rates of convergence.

Theorem 2. *Let $q \in \{1,2\}$. Suppose that $s^{1/q}\sqrt{\log(2pn/\alpha)/n} = o(1)$.Furthermore, assume that there exist constants $0 < c < C < \infty$ such that, for any $1 \leq j \leq p$,*

$$c(1+\|\beta_0\|_2)^2 \leq \frac{1}{n}\sum_{i=1}^n \mathrm{E}[\{z_{ij}(\xi_i - w_i^T\beta_0)+\Gamma_{jj}\beta_{0j}\}^2] \leq C(1+\|\beta_0\|_2)^2.$$

Then, under the assumptions of Corollary 1, for n sufficiently large, with probability at least $1-\alpha-11\varepsilon-o(1)$ we have

$$\|\hat{\beta}_{\widehat{T}}\|_0 \leq Cs \quad \text{and} \quad \|\hat{\beta}_{\widehat{T}}-\beta_0\|_q \leq C(1+\|\beta_0\|_2)s^{1/q}\sqrt{\frac{\log(c'p/\alpha)}{n}}$$

where $C > 0, c' \geq 1$ are constants.

Theorem 2 shows that the estimator $\hat{\beta}_{\widehat{T}}$ inherits the ℓ_1 and ℓ_2 rates of convergence of $\hat{\beta}$ and is also sparse. Estimators with this additional sparsity property have been useful in many settings, see for example [2].

4. Numerical Experiments. In this section, we investigate the performance of the self-normalized conic estimator. In the Supplementary Material we present more detailed simulation results for different designs, which include covariates missing at random, and the thresholded version of the estimator.

We consider the data generating process (1.1) where ξ_i, w_i, x_i are drawn independently satisfying $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$, $w_i \sim \mathcal{N}(0, \sigma_w^2 I_{p\times p})$, $x_i \sim \mathcal{N}(0, \Sigma)$. Here, $I_{p\times p}$ is the identity matrix and Σ is $p \times p$ matrix with elements $\Sigma_{ij} = \rho^{|i-j|}$. We set $\sigma_\xi = 1$, $\sigma_w = 1$ and $\rho = 0.5$. For simplicity, we assume σ_w to be known in all calculations, which means that we set $\hat{\Gamma} = \sigma_w^2 I_{p\times p}$ and $b_\epsilon = 0$. The coefficients of the model are set to $\beta_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)^T$. All results are based on 100 replications.

We compare the performance of the proposed self-normalized estimator (SN-conic) with the performance of the conic estimator (Conic) of [1], the (biased) Lasso estimator of y on z, and the no measurement error Lasso estimator with y on the unobserved x. For numerical comparison, we report the bias, the average accuracy in ℓ_2, and the prediction risk $\|X(\hat{\beta} - \beta_0)\|_2/\sqrt{n}$ for each estimator. The Conic estimator is tuned assuming that $\sigma_\xi = 1$ is known ($\tau = \sigma_\xi\sqrt{\log(p/\varepsilon)/n}$) while the SN-conic has its penalty parameter set to $\tau = \frac{1}{2}n^{-1/2}\Phi^{-1}(1 - \alpha/(2p))$, $\alpha = 0.05$, $\lambda_t = 1$ and $\lambda_u = 0.25$. Computations are performed in R and use the optimization software Mosek, an interior point methods solver, wrapped through the R package Rmosek. All estimates are truncated at 10^{-7}.

The SN-conic estimator provides good results at all three levels of p considered in the simulations, with the performance deteriorating slightly with increase in p. Results, reported in Table 1, support our theoretical results regarding consistency of the proposed estimator. We also observe that SN-conic estimator outperforms the Conic estimator in most of the designs considered. This is attractive as the latter was tuned based on knowing σ_w and σ_ξ whereas the SN-conic estimator does not need the knowledge of these parameters. Another, more obvious observation is the poor performance of (biased) Lasso method that disregards the presence of the measurement error.

Additional simulations are provided in the Supplementary Material, including the case of covariates missing at random. The Supplementary Material also presents results for the proposed thresholded estimator, different choices of parameter, and other estimators in the literature.

			$n = 200$			$n = 250$	
Method	p	**Bias**	**ℓ_2-loss**	**PR**	**Bias**	**ℓ_2-loss**	**PR**
SN-Conic	**10**	0.367	0.875	0.706	0.247	0.716	0.581
	100	0.628	0.891	0.919	0.536	0.822	0.818
	400	0.809	1.064	1.144	0.667	0.878	0.952
Conic	**10**	0.693	1.000	0.992	0.572	0.823	0.836
	100	0.749	1.030	1.785	0.655	0.942	0.971
	400	0.841	1.132	1.184	0.711	0.952	1.014
Lasso (Biased)	**10**	0.900	0.937	1.276	0.883	0.913	1.238
	100	0.941	1.003	1.322	0.909	0.969	1.277
	400	1.002	1.103	1.356	0.966	1.054	1.306
Lasso (No Meas Error)	**10**	0.261	0.331	0.380	0.225	0.282	0.329
	100	0.311	0.370	0.445	0.281	0.332	0.399
	400	0.340	0.389	0.482	0.304	0.356	0.436

			$n = 300$	
Method	p	**Bias**	**ℓ_2-loss**	**PR**
SN-Conic	**10**	0.216	0.717	0.572
	100	0.496	0.716	0.734
	400	0.588	0.782	0.849
Conic	**10**	0.524	0.779	0.787
	100	0.603	0.808	0.878
	400	0.644	0.856	0.927
Lasso (Biased)	**10**	0.860	0.890	1.216
	100	0.907	0.954	1.274
	400	0.929	1.004	1.282
Lasso (No Meas Error)	**10**	0.199	0.262	0.304
	100	0.267	0.307	0.375
	400	0.284	0.327	0.406

TABLE 1

For each estimator we report the bias, average ℓ_2-loss (average of $\|\hat\beta - \beta\|_2$ over replications) and the average prediction risk (PR) given by $\|X(\hat\beta - \beta)\|_2/\sqrt{n}$.

APPENDIX A: MAIN PROOFS

We set $T = \text{supp}(\beta_0)$ with $|T| \leq s$. We begin by stating a technical lemma.

LEMMA 1. *For $a \geq 1$ and $\gamma > 0$, we have*

$$1 - \Phi(a/\{1+\gamma\}) \leq \{1 - \Phi(a)\}\exp(2a^2\gamma).$$

The next lemma deals with $(\hat\beta, \hat t, \hat u)$ defined as the solution to (2.2)-(2.3).

LEMMA 2. *Under Conditions A and B, for $\tau = n^{-1/2}\Phi^{-1}(1 - \alpha/(2p))$, we have that β_0 is feasible in the pivotal conic programming (2.2) with probability at least $1 - \alpha\{1 + o(1)\} - \epsilon$ for n sufficiently large. On that event, we have*

$$\|\hat t\|_\infty - \|t(\beta_0)\|_\infty \leq \frac{1+\lambda_u}{\lambda_t}\|\hat\beta - \beta_0\|_1, \quad \textit{and}$$

$$\|\hat{u}\|_\infty - \|\beta_0\|_\infty \leq \frac{1}{\lambda_u}\|\hat{\beta} - \beta_0\|_1 + \frac{\lambda_t}{\lambda_u} H_n \|\hat{\beta} - \beta_0\|_1.$$

In addition, using $\lambda_u = 1/4$ *and* $\lambda_t = 1/\{4H_n\}$*, we have* $\hat{\beta} - \beta_0 \in C_T(3)$*.*

Proof of Lemma 2. Let $\Gamma_{i,jj} = \mathrm{E}[w_{ij}^2]$, $\Gamma_{jj} = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[w_{ij}^2]$, and $|\beta_0| = (|\beta_{0j}|)_{j=1}^p$. Recall that we defined $t_j(\beta) = \{\frac{1}{n}\sum_{i=1}^n \{z_{ij}(y_i - z_i^T\beta) + \hat{\Gamma}_{j\cdot}\beta\}^2\}^{1/2}$ and $U_{ij} = z_{ij}(y_i - z_i^T\beta_0) + \Gamma_{i,jj}\beta_{0j}$. For $j = 1, \ldots, p$, $i = 1, \ldots, n$, we set $\bar{U}_{ij} = z_{ij}(y_i - z_i^T\beta_0) + \Gamma_{jj}\beta_{0j}$ (note that the optimization problem has $\hat{\Gamma}_{jj}$ instead of $\Gamma_{i,jj}$ or Γ_{jj}). Remark also that by definition $\sum_{i=1}^n U_{ij} = \sum_{i=1}^n \bar{U}_{ij}$. We now show that the triplet $(\beta_0, t(\beta_0), |\beta_0|)$ is feasible with high probability. Indeed, the probability that the triplet $(\beta_0, t(\beta_0), |\beta_0|)$ violates any constraint satisfies

$$\begin{aligned}
&\mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n z_{ij}(y_i - z_i^T\beta_0) + \hat{\Gamma}_{jj}\beta_{0j}\Big| > \tau t_j(\beta_0) + (1+\tau)b_\epsilon|\beta_{0j}|\Big)\\
&= \mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n \bar{U}_{ij} + (\hat{\Gamma}_{jj} - \Gamma_{jj})\beta_{0j}\Big| > \tau t_j(\beta_0) + (1+\tau)b_\epsilon|\beta_{0j}|\Big)\\
&\leq \mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n U_{ij} + (\hat{\Gamma}_{jj} - \Gamma_{jj})\beta_{0j}\Big| > \tau\Big\|\frac{\bar{U}_{\cdot j}}{\sqrt{n}}\Big\|_2 - \tau|(\hat{\Gamma}_{jj} - \Gamma_{jj})\beta_{0j}| + (1+\tau)b_\epsilon|\beta_{0j}|\Big)\\
&\leq \mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n U_{ij}\Big| > \tau\Big\|\frac{\bar{U}_{\cdot j}}{\sqrt{n}}\Big\|_2\Big) + \epsilon\\
&\leq \mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n U_{ij}\Big| > \frac{\tau}{1+n^{-1/3}}\Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Big) + \mathrm{P}\Big(\exists j : (1+n^{-1/3})\Big\|\frac{\bar{U}_{\cdot j}}{\sqrt{n}}\Big\|_2 \leq \Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Big) + \epsilon,
\end{aligned}$$

where we used that $\|\bar{U}_{\cdot j}/\sqrt{n}\|_2 - t_j(\beta_0)| \leq |\hat{\Gamma}_{jj} - \Gamma_{jj}| \cdot |\beta_{0j}|$ by definition and $\max_{j=1,\ldots,p}|\hat{\Gamma}_{jj} - \Gamma_{jj}| \leq b_\epsilon$ with probability $1 - \epsilon$ from Condition B.

We now bound each term in the last display separately. By Condition A, note that U_{ij} is a zero-mean random variable. Therefore, applying Lemma 7.4 in [12] together with Condition B that implies

$$\sqrt{n}\tau \max_{1\leq j\leq p}\Big\{\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|U_{ij}|^3]\Big)^{1/3}\Big/\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|U_{ij}|^2]\Big)^{1/2}\Big\} \leq n^{1/6}/\ell_n,$$

where $\ell_n \to \infty$, we have

$$\begin{aligned}
&\mathrm{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n U_{ij}\Big| > \frac{\sqrt{n}\tau}{(1+n^{-1/3})}\Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Big) \leq \{1 - \Phi(\sqrt{n}\tau/\{1+n^{-1/3}\})\}(1 + \frac{A}{\ell_n^3})\\
&\leq \{1 - \Phi(\sqrt{n}\tau)\}\exp\Big(2n^{-1/3}\{\Phi^{-1}(1 - \tfrac{\alpha}{2p})\}^2\Big)\Big(1 + \tfrac{A}{\ell_n^3}\Big)\\
&\leq \{1 - \Phi(\sqrt{n}\tau)\}\exp\big(2/\ell_n^2\big)\Big(1 + \tfrac{A}{\ell_n^3}\Big) = \tfrac{\alpha}{2p}\exp\big(2/\ell_n^2\big)\big(1 + \tfrac{A}{\ell_n^3}\big),
\end{aligned} \tag{A.1}$$

for some universal constant $A > 0$. Here, we have used Lemma 1, and again Condition B that implies $n^{-1/3}\{\Phi^{-1}(1 - \frac{\alpha}{2p})\}^2 \leq 1/\ell_n^2$.

To bound the last term, set $\Sigma_j^2 := \frac{1}{n}\sum_{i=1}^n(\Gamma_{i,jj}-\Gamma_{jj})^2$, $j=1,\dots,p$. Note that

$$\mathrm{P}\Big(\exists j : (1+n^{-1/3})\Big\|\frac{\bar{U}_{\cdot j}}{\sqrt{n}}\Big\|_2 \le \Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Big)$$

is smaller than

$$\mathrm{P}\Big(\exists j : n^{-1/3}\Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2^2 + (1+n^{-1/3})^2\Sigma_j^2\beta_{0j}^2 + \frac{2(1+n^{-1/3})^2\beta_{0j}}{n}\sum_{i=1}^n U_{ij}(\Gamma_{jj}-\Gamma_{i,jj}) < 0\Big).$$

Since $0 \le \Gamma_{i,jj} \le C$ by Condition A, we deduce that $\{U_{ij}(\Gamma_{i,jj}-\Gamma_{jj}) : i = 1,\dots,n\}$ satisfies the moderate deviation condition for self-normalized sums since $\{U_{ij} : i=1,\dots,n\}$ satisfies it by Condition B. Therefore we get

$$\begin{aligned}
&\mathrm{P}\Big(\exists j : \Big|\frac{1}{n}\sum_{i=1}^n U_{ij}(\Gamma_{i,jj}-\Gamma_{jj})\Big| > \frac{\Phi^{-1}(1-\frac{\alpha}{2pn})}{n^{1/2}}\Big\{\frac{1}{n}\sum_{i=1}^n U_{ij}^2(\Gamma_{i,jj}-\Gamma_{jj})^2\Big\}^{1/2}\Big)\\
&\le \frac{\alpha}{n}(1+A/\ell_n^3).
\end{aligned}$$

Note that if $\Big\{\frac{1}{n}\sum_{i=1}^n U_{ij}^2(\Gamma_{i,jj}-\Gamma_{jj})^2\Big\}^{1/2} = 0$ the result is trivial. Furthermore we have

$$\Big\{\frac{1}{n}\sum_{i=1}^n U_{ij}^2(\Gamma_{i,jj}-\Gamma_{jj})^2\Big\}^{1/2} \le \Big\{\frac{1}{n}\sum_{i=1}^n |U_{ij}|^2\Big\}^{1/2}\max_{1\le i\le n}|\Gamma_{i,jj}-\Gamma_{jj}|$$

and by Condition B, for all j such that $\Delta_j := |\beta_{0j}|\max_{1\le i\le n}|\Gamma_{i,jj}-\Gamma_{jj}| > 0$,

$$\Delta_j\Phi^{-1}(1-\alpha/(2pn)) \le \frac{n^{1/6}}{\ell_n}\Big\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]\Big\}^{1/2}.$$

Hence,

$$\begin{aligned}
&\mathrm{P}\Big(\exists j : (1+n^{-1/3})\Big\|\frac{\bar{U}_{\cdot j}}{\sqrt{n}}\Big\|_2 \le \Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Big)\\
&\le \mathrm{P}\Big(\exists j : n^{-1/3}\Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2^2 < 2(1+n^{-1/3})^2\frac{\Phi^{-1}(1-\frac{\alpha}{2pn})}{n^{1/2}}\Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2\Delta_j\Big) + \frac{\alpha}{n}\{1+A/\ell_n^3\}\\
&\le \mathrm{P}\Big(\exists j : \Delta_j > 0 \text{ and } \Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2 < \frac{1}{\ell_n^2}\Big\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]\Big\}^{1/2}\Big) + \frac{\alpha}{n}\{1+A/\ell_n^3\}.
\end{aligned}\tag{A.2}$$

To bound the first term of the RHS of (A.2), set $A_n = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2 1\{|U_{ij}| \le M_n\}]$ for some threshold $M_n > 0$. We have, for $t > 0$,

$$\begin{aligned}
\mathrm{P}(\textstyle\sum_{i=1}^n U_{ij}^2 < A_n n - tn) &\le \mathrm{P}(\textstyle\sum_{i=1}^n U_{ij}^2 1\{|U_{ij}| \le M_n\} < A_n n - tn)\\
&\le \exp(-2t^2 n/M_n^4)
\end{aligned}$$

by Hoeffding's inequality. Moreover,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[U_{ij}^2]-A_n\le\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[U_{ij}^2 1\{|U_{ij}|>M_n\}]\le\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[|U_{ij}|^3]/M_n.$$

Setting $M_n = \frac{2}{n}\sum_{i=1}^n \mathrm{E}[|U_{ij}|^3]/\frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]$, we have $A_n \ge \frac{1}{2n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]$. Then, taking $t = A_n/2$, and using Condition B, we have for any j such that $\Delta_j > 0$:

$$\begin{array}{rl}\mathrm{P}(\frac{1}{n}\sum_{i=1}^n U_{ij}^2 < \frac{1}{4}\frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]) & \le \exp(-2t^2n/(M_n^4)) \\ & \le \exp(-\frac{n}{2^5}\left\{\mathcal{U}_{j2}/\mathcal{U}_{j3}\right\}^{12}) \\ & \le \exp(-\ell_n^2\{\Phi^{-1}(1-\alpha/(2pn))\}^2) \\ & \le \frac{\alpha}{2pn}(1+o(1))\end{array}$$

by Condition B(iii) since $\{\Phi^{-1}(1-\alpha/(2pn))\}^2 \ge c'\log(2pn/\alpha)$ for some universal $c' > 0$ and $\ell_n c' \ge 1$ for n sufficiently large. By the union bound, this implies

$$\mathrm{P}\Big(\exists j:\Delta_j>0 \text{ and } \Big\|\frac{U_{\cdot j}}{\sqrt{n}}\Big\|_2 < o(1)\Big\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[U_{ij}^2]\Big\}^{1/2}\Big)\le\frac{\alpha}{2n}(1+o(1)). \tag{A.3}$$

Combining (A.1), (A.2), (A.3), and using the convergence $\ell_n \to \infty$ and the union bound, we find that the triplet $(\beta_0, t(\beta_0), |\beta_0|)$ is feasible with probability at least $1-\alpha\{1+o(1)\}-\epsilon$.

If the triplet $(\beta_0, t(\beta_0), |\beta_0|)$ is feasible for the problem above, it follows that

$$\|\hat\beta\|_1+\lambda_t\|\hat t\|_\infty+\lambda_u\|\hat u\|_\infty\le\|\beta_0\|_1+\lambda_t\|t(\beta_0)\|_\infty+\lambda_u\|\beta_0\|_\infty. \tag{A.4}$$

By (A.4), and the inequalities $\|t(\hat\beta)\|_\infty \le \|\hat t\|_\infty$ and $\|\hat\beta\|_\infty \le \|\hat u\|_\infty$ from the definition of the estimator, we have that

$$\begin{array}{l}\|\hat t\|_\infty-\|t(\beta_0)\|_\infty\le\frac{1+\lambda_u}{\lambda_t}\|\hat\beta-\beta_0\|_1,\\ \|\hat u\|_\infty-\|\beta_0\|_\infty\le\frac{1}{\lambda_u}\|\beta_0-\hat\beta\|_1+\frac{\lambda_t}{\lambda_u}\{\|t(\beta_0)\|_\infty-\|t(\hat\beta)\|_\infty\}.\end{array}$$

Next, since $|\{\frac{1}{n}\sum_{i=1}^n a_i^2\}^{1/2} - \{\frac{1}{n}\sum_{i=1}^n b_i^2\}^{1/2}| \le \{\frac{1}{n}\sum_{i=1}^n (a_i-b_i)^2\}^{1/2}$, we have

$$\begin{array}{rl}|t_j(\beta_0)-t_j(\hat\beta)|^2 & \le\frac{1}{n}\sum_{i=1}^n\{(z_{ij}z_i^T-\hat\Gamma_{j\cdot})(\hat\beta-\beta_0)\}^2\\ & =(\hat\beta-\beta_0)^T\frac{1}{n}\sum_{i=1}^n(z_{ij}z_i^T-\hat\Gamma_{j\cdot})^T(z_{ij}z_i^T-\hat\Gamma_{j\cdot})(\hat\beta-\beta_0)\ .\\ & \le\|\hat\beta-\beta_0\|_1^2\|\frac{1}{n}\sum_{i=1}^n(z_{ij}z_i^T-\hat\Gamma_{j\cdot})^T(z_{ij}z_i^T-\hat\Gamma_{j\cdot})\|_\infty\end{array}$$

Thus for $H_n = \max_{j=1,\ldots,p}\|\frac{1}{n}\sum_{i=1}^n(z_{ij}z_i^T-\hat\Gamma_{j\cdot})^T(z_{ij}z_i^T-\hat\Gamma_{j\cdot})\|_\infty^{1/2}$, we obtain

$$\|t(\beta_0)-t(\hat\beta)\|_\infty\le H_n\|\hat\beta-\beta_0\|_1 \tag{A.5}$$

and the inequality on $\|\hat u\|_\infty$ stated in the lemma follows.

We now establish the last claim of the lemma. From (A.4), and the inequalities $\|t(\hat\beta)\|_\infty \leq \|\hat t\|_\infty$, $\|\hat\beta\|_\infty \leq \|\hat u\|_\infty$ we get

$$\begin{aligned}\|\hat\beta\|_1 & \leq \|\beta_0\|_1 + \lambda_t\{\|t(\beta_0)\|_\infty - \|t(\hat\beta)\|_\infty\} + \lambda_u\{\|\beta_0\|_\infty - \|\hat\beta\|_\infty\} \\ & \leq \|\beta_0\|_1 + \lambda_t\|t(\beta_0) - t(\hat\beta)\|_\infty + \lambda_u\|\hat\beta - \beta_0\|_1.\end{aligned}$$

Setting $\lambda_t = \frac{1}{4H_n}$, $\lambda_u = 1/4$, and using the fact that $\|\hat\beta\|_1 = \|\hat\beta_T\|_1 + \|\hat\beta_{T^c}\|_1$, we obtain $\frac{1}{2}\|\hat\beta_{T^c}\|_1 \leq \frac{3}{2}\|\beta_0 - \hat\beta_T\|_1$. □

Proof of Theorem 1. Set $Z = [z_1; \ldots; z_n]^T$ and $W = [w_1; \ldots; w_n]^T$. By the triangle inequality,

$$\begin{aligned}\left\|\tfrac{1}{n}X^TX(\hat\beta - \beta_0)\right\|_\infty \leq & \left\|\tfrac{1}{n}Z^T(Y - Z\hat\beta) + \hat\Gamma\hat\beta\right\|_\infty + \left\|(\tfrac{1}{n}Z^TW - \Gamma)\hat\beta\right\|_\infty \\ & + \left\|(\hat\Gamma - \Gamma)\hat\beta\right\|_\infty + \left\|\tfrac{1}{n}Z^T\xi\right\|_\infty + \left\|\tfrac{1}{n}W^TX(\hat\beta - \beta_0)\right\|_\infty.\end{aligned} \tag{A.6}$$

We now bound separately the terms on the RHS in (A.6).

As shown at the end of this proof, the second term in (A.6) is bounded with probability at least $1 - 6\varepsilon$ as follows:

$$\begin{aligned}\left\|(\tfrac{1}{n}Z^TW - \Gamma)\hat\beta\right\|_\infty & \leq \{\delta_1(\varepsilon) + \delta_4(\varepsilon) + \delta_5(\varepsilon)\}\|\hat\beta - \beta_0\|_1 \\ & + \{\delta_1'(\varepsilon) + \delta_4'(\varepsilon)\}\|\beta_0\|_2 + \delta_5(\varepsilon)\|\beta_0\|_\infty,\end{aligned} \tag{A.7}$$

where the quantities $\delta_i(\varepsilon)$ are defined in Appendix B.1. By Condition B, the third term in (A.6) is bounded with probability at least $1 - \epsilon$ as follows:

$$\begin{aligned}\left\|(\hat\Gamma - \Gamma)\hat\beta\right\|_\infty & \leq \left\|(\hat\Gamma - \Gamma)\beta_0\right\|_\infty + \left\|(\hat\Gamma - \Gamma)(\hat\beta - \beta_0)\right\|_\infty \\ & \leq b_\epsilon\|\beta_0\|_\infty + b_\epsilon\|\hat\beta - \beta_0\|_\infty.\end{aligned}$$

Lemma 4 provides, with probability at least $1 - 2\varepsilon$, the following bound on the fourth term in (A.6) :

$$\left\|\tfrac{1}{n}Z^T\xi\right\|_\infty \leq \left\|\tfrac{1}{n}X^T\xi\right\|_\infty + \left\|\tfrac{1}{n}W^T\xi\right\|_\infty \leq \delta_2(\varepsilon) + \delta_3(\varepsilon).$$

Finally the last term in (A.6) is bounded, with probability at least $1 - \varepsilon$, again via Lemma 4:

$$\left\|\tfrac{1}{n}W^TX(\hat\beta - \beta_0)\right\|_\infty \leq \left\|\tfrac{1}{n}X^TW\right\|_\infty\|\hat\beta - \beta_0\|_1 \leq \delta_1(\varepsilon)\|\hat\beta - \beta_0\|_1.$$

Therefore, with probability at least $1 - 9\varepsilon - \epsilon$ we have

$$\begin{aligned}\left\|\tfrac{1}{n}X^TX(\hat\beta - \beta_0)\right\|_\infty & \leq \left\|\tfrac{1}{n}Z^T(Y - Z\hat\beta) + \hat\Gamma\hat\beta\right\|_\infty \\ & + \tau_0 + \tau_\infty\|\beta_0\|_\infty + \tau_2\|\beta_0\|_2 + \tau_1\|\hat\beta - \beta_0\|_1,\end{aligned} \tag{A.8}$$

where

$$\begin{array}{lll}
\tau_0 & := \delta_2(\varepsilon)+\delta_3(\varepsilon) & \leq \{\sigma_\xi m_2^{1/2}+C_{\xi w}\}\sqrt{\frac{2\log(2p/\varepsilon)}{n}} \\
\tau_\infty & := b_\epsilon+\delta_5(\varepsilon) & \leq b_\epsilon + C_{\xi w}\sqrt{\frac{2\log(2p/\varepsilon)}{n}} \\
\tau_2 & := \delta_1'(\varepsilon)+\delta_4'(\varepsilon) & \leq \{\sigma_w m_2^{1/2}+C_{\xi w}\}\sqrt{\frac{2\log(2p/\varepsilon)}{n}} \\
\tau_1 & := 2\delta_1(\varepsilon)+\delta_4(\varepsilon)+\delta_5(\varepsilon)+b_\epsilon & \leq b_\epsilon+2\{\sigma_w m_2^{1/2}+C_{\xi w}\}\sqrt{\frac{2\log(2p^2/\varepsilon)}{n}}.
\end{array}$$

Here, $C_{\xi w}$ is a positive constant depending only on σ_ξ and σ_w, and the bounds hold for n large enough under the condition $C_{\xi w}\log(p/\varepsilon)=o(n)$.

Next, we bound the first term in (A.6). By the feasibility of $(\hat\beta,\hat t,\hat u)$ in (2.2) we have

$$\left\|\tfrac{1}{n}Z^T(Y-Z\hat\beta)+\hat\Gamma\hat\beta\right\|_\infty \leq \tau\|\hat t\|_\infty+(1+\tau)b_\epsilon\|\hat u\|_\infty.$$

By Lemma 2 and the choices $\lambda_t=1/\{4H_n\}$ and $\lambda_u=1/4$, with probability $1-\alpha\{1+o(1)\}-\epsilon$ we have $\hat\beta-\beta_0\in C_T(3)$ and the bounds on $\|\hat t\|_\infty$ and $\|\hat u\|_\infty$ apply, so that

$$\begin{array}{l}
\left\|\tfrac{1}{n}Z^T(Y-Z\hat\beta)+\hat\Gamma\hat\beta\right\|_\infty \leq \tau\|t(\beta_0)\|_\infty+\tau\frac{1+\lambda_u}{\lambda_t}\|\hat\beta-\beta_0\|_1 \\
+(1+\tau)b_\epsilon\left\{\|\beta_0\|_\infty+\left(\frac{1}{\lambda_u}+\frac{\lambda_t}{\lambda_u}H_n\right)\|\hat\beta-\beta_0\|_1\right\} \\
=\tau\|t(\beta_0)\|_\infty+5\tau H_n\|\hat\beta-\beta_0\|_1+(1+\tau)b_\epsilon\left\{\|\beta_0\|_\infty+5\|\hat\beta-\beta_0\|_1\right\}.
\end{array} \tag{A.9}$$

Next, on the event $\hat\beta-\beta_0\in C_T(3)$ we bound the LHS of (A.8) from below via the ℓ_q-sensitivity. Plugging that lower bound and (A.9) in (A.8) we find

$$\begin{array}{l}
\kappa_q(s,3)\|\hat\beta-\beta_0\|_q\leq\left\|\tfrac{1}{n}X^TX(\hat\beta-\beta_0)\right\|_\infty \\
\leq\tau\|t(\beta_0)\|_\infty+\tau_0+\{(1+\tau)b_\epsilon+\tau_\infty\}\|\beta_0\|_\infty+\tau_2\|\beta_0\|_2+\tilde\mu_1\|\hat\beta-\beta_0\|_1,
\end{array}$$

where $\tilde\mu_1=\tau_1+5\tau H_n+5(1+\tau)b_\epsilon$. Note that

$$\tilde\mu_1\leq\tau_1+5\tau H_\epsilon+5(1+\tau)b_\epsilon\leq(1+\tau)b_\epsilon+\tau h_\epsilon+C'(1+m_2^{1/2})\sqrt{\log(2p^2/\varepsilon)/n}$$

with probability $1-\epsilon$ where $C'=\sigma_w\vee C_{w\xi}$ is bounded by a constant since $\sigma_w\vee\sigma_\xi\leq C$ under Condition A. Moreover, since $\hat\beta-\beta_0\in C_T(3)$ we have

$$\|\hat\beta-\beta_0\|_1\leq 4\|(\hat\beta-\beta_0)_T\|_1\leq 4s^{1-1/q}\|(\hat\beta-\beta_0)_T\|_q\leq 4s^{1-1/q}\|\hat\beta-\beta_0\|_q.$$

Thus under the condition of the theorem on $\kappa_q(s,3)$, we have with probability $1-\alpha\{1+o(1)\}-2\epsilon-9\varepsilon$ that

$$\frac{\kappa_q(s,3)}{2}\|\hat\beta-\beta_0\|_q\leq\tau\|t(\beta_0)\|_\infty+\tau_0+\{(1+\tau)b_\epsilon+\tau_\infty\}\|\beta_0\|_\infty+\tau_2\|\beta_0\|_2.$$

The result follows by noticing that $(1+\tau)b_\epsilon \leq 2b_\epsilon \leq 2\tau_\infty$ for large enough n.

Proof of (A.7). We have

$$\begin{aligned}\left\|(\tfrac{1}{n}Z^TW-\Gamma)\hat\beta\right\|_\infty &\leq \left\|(\tfrac{1}{n}Z^TW-\Gamma)\beta_0\right\|_\infty + \left\|(\tfrac{1}{n}Z^TW-\Gamma)(\hat\beta-\beta_0)\right\|_\infty\\ &\leq \left\|(\tfrac{1}{n}Z^TW-\Gamma)\beta_0\right\|_\infty + \|\tfrac{1}{n}Z^TW-\Gamma\|_\infty\|\hat\beta-\beta_0\|_1\\ &\leq \left\|(\tfrac{1}{n}W^TW-\Gamma)\beta_0\right\|_\infty + \left\|\tfrac{1}{n}X^TW\beta_0\right\|_\infty\\ &+\|\tfrac{1}{n}X^TW\|_\infty\|\hat\beta-\beta_0\|_1 + \|\tfrac{1}{n}W^TW-\Gamma\|_\infty\|\hat\beta-\beta_0\|_1.\end{aligned}$$

By Lemma 4 we get, with probability at least $1-3\varepsilon$,

$$\|\tfrac{1}{n}X^TW\|_\infty \leq \delta_1(\varepsilon)\ ,$$

$$\begin{aligned}\|\tfrac{1}{n}W^TW-\Gamma\|_\infty &\leq \|\tfrac{1}{n}W^TW-\tfrac{1}{n}\mathrm{Diag}(W^TW)\|_\infty + \|\tfrac{1}{n}\mathrm{Diag}(W^TW)-\Gamma\|_\infty\\ &\leq \delta_4(\varepsilon)+\delta_5(\varepsilon).\end{aligned}$$

Finally, Lemma 5 and Lemma 4 yield that, with probability at least $1-3\varepsilon$,

$$\|\tfrac{1}{n}X^TW\beta_0\|_\infty \leq \delta_1'(\varepsilon)\|\beta_0\|_2\ ,$$

$$\begin{aligned}\left\|(\tfrac{1}{n}W^TW-\Gamma)\beta_0\right\|_\infty &\leq \left\|\tfrac{1}{n}(W^TW-\mathrm{Diag}(W^TW))\beta_0\right\|_\infty\\ &+\left\|(\tfrac{1}{n}\mathrm{Diag}(W^TW)-\Gamma)\beta_0\right\|_\infty\\ &\leq \delta_4'(\varepsilon)\|\beta_0\|_2 + \|\tfrac{1}{n}\mathrm{Diag}(W^TW)-\Gamma\|_\infty\|\beta_0\|_\infty\\ &\leq \delta_4'(\varepsilon)\|\beta_0\|_2 + \delta_5(\varepsilon)\|\beta_0\|_\infty.\end{aligned}$$

□

Proof of Corollary 1. By Theorem 1 with probability $1-\alpha\{1+o(1)\}-11\varepsilon$ we have

$$\|\hat\beta-\beta_0\|_q \leq \frac{\tau\|t(\beta_0)\|_\infty}{c'\kappa_q(s,3)} + \frac{(1+\|\beta_0\|_2)(1+m_2^{1/2})}{c'\kappa_q(s,3)}\sqrt{\frac{\log(2p/\varepsilon)}{n}} + \frac{b_\varepsilon\|\beta_0\|_\infty}{c'\kappa_q(s,3)}.$$

Under the additional condition $X\in\Omega_X$, we have by Lemma 3 that $P(H_n\leq C)\geq 1-o(1)$. Therefore we have that with probability $1-\alpha\{1+o(1)\}-11\varepsilon-o(1)$

$$\|\hat\beta-\beta_0\|_q \leq Cs^{1/q}\Big\{\tau\|t(\beta_0)\|_\infty + (1+\|\beta_0\|_2)\sqrt{\frac{\log(2p/\varepsilon)}{n}} + b_\varepsilon\|\beta_0\|_\infty\Big\}$$

since $m_2^{1/2} \leq \{\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n x_{ij}^4\}^{1/4} \leq C^{1/4}$ when $X\in\Omega_X$. Using the triangle inequality, we obtain

$$\begin{aligned}
\|t(\beta_0)\|_\infty & = \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n\{z_{ij}(\xi_i - w_i^T\beta_0)+\hat\Gamma_{jj}\beta_{0j}\}^2\right\}^{1/2} \\
& \le \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n\{z_{ij}(\xi_i - w_i^T\beta_0)\}^2\right\}^{1/2} + |\hat\Gamma_{jj}\beta_{0j}| \\
& \le_{(i)} \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n\{z_{ij}(\xi_i - w_i^T\beta_0)\}^2\right\}^{1/2} \\
& +|\Gamma_{jj}\beta_{0j}| + b_\varepsilon\|\beta_0\|_\infty \\
& \le_{(ii)} \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n z_{ij}^4\right\}^{1/4}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\xi_i - w_i^T\beta_0)^4\right\}^{1/4} \\
& +|\Gamma_{jj}\beta_{0j}| + b_\varepsilon\|\beta_0\|_\infty \\
& \le_{(iii)} \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n x_{ij}^4\right\}^{1/4}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\xi_i - w_i^T\beta_0)^4\right\}^{1/4} \\
& + \max_{j\le p}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n w_{ij}^4\right\}^{1/4}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\xi_i - w_i^T\beta_0)^4\right\}^{1/4} \\
& +|\Gamma_{jj}\beta_{0j}| + b_\varepsilon\|\beta_0\|_\infty,
\end{aligned}$$

where (i) follows from the inequality $\|\hat\Gamma-\Gamma\|_\infty \le b_\varepsilon$ which holds with probability $1-\varepsilon$ by Condition B, (ii) follows from the Cauchy-Schwarz inequality, and (iii) from the triangle inequality. On the event $X \in \Omega_X$, we have $\max_{j\le p}\left\{\frac{1}{n}\sum_{i=1}^n x_{ij}^4\right\}^{1/4} \le C$. Note that w_{ij}, $i = 1,\dots,n$, $j=1,\dots,p$, are σ_w-subgaussian random variables with $\sigma_w \le C$. Therefore, by Lemmas 6 and 7, $\max_{i\le n, j\le p}\frac{1}{n}\sum_{i=1}^n \mathrm{E}[w_{ij}^4] \le C'$, $\mathrm{E}[\max_{i\le n,j\le p}|w_{ij}|^4] \le C'\log^2(pn)$, and

$$\mathrm{E}\left[\max_{j\le p}\left|\frac{1}{n}\sum_{i=1}^n w_{ij}^4 - \mathrm{E}[w_{ij}^4]\right|\right] \le C'\frac{\log(p)}{n}\log^2(pn) + C'\sqrt{\frac{\log(p)}{n}\log^2(pn)} \le o(1)$$

where we have used the relation $\log^3(2p) = o(n)$ following from Condition B(ii). Then using Markov's inequality and the fact that w_i's are subgaussian, with probability $1-o(1)$ we get $\max_{j\le p}\left\{\frac{1}{n}\sum_{i=1}^n w_{ij}^4\right\}^{1/4} \le C''$.

Next note that $\left\{\frac{1}{n}\sum_{i=1}^n(\xi_i - w_i^T\beta_0)^4\right\}^{1/4} \le C'(1+\|\beta_0\|_2)$ with probability $1-o(1)$. Indeed, each of the random variables $\tilde\xi_i := \xi_i - w_i^T\beta_0$ is subgaussian with parameter bounded by $C(1+\|\beta_0\|_2)$. Thus we have

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\tilde\xi_i^4\right) \le \frac{1}{n^2}\sum_{i=1}^n\mathrm{E}[\tilde\xi_i^8] \le C'(1+\|\beta_0\|_2)^8/n.$$

Therefore, using Markov's inequality, we get with probability at least $1-n^{-1/2}$,

$$\begin{aligned}
\tfrac{1}{n}\textstyle\sum_{i=1}^n\tilde\xi_i^4 & \le \left|\tfrac{1}{n}\textstyle\sum_{i=1}^n\tilde\xi_i^4 - \tfrac{1}{n}\sum_{i=1}^n\mathrm{E}[\tilde\xi_i^4]\right| + C(1+\|\beta_0\|_2)^4 \\
& \le n^{1/4}C'(1+\|\beta_0\|_2)^4/\sqrt{n} + C(1+\|\beta_0\|_2)^4 \\
& \le C''(1+\|\beta_0\|_2)^4.
\end{aligned}$$

Thus with probability $1 - \varepsilon - o(1)$ we have

$$\|t(\beta_0)\|_\infty \leq C(1 + \|\beta_0\|_2) + (C + b_\varepsilon)\|\beta_0\|_\infty.$$

Since $\tau = n^{-1/2}\Phi^{-1}(1 - \alpha/(2p)) \leq C\sqrt{\log(2p/\alpha)/n}$, $b_\varepsilon \leq C\sqrt{\log(2p/\varepsilon)/n}$ and $\|\beta_0\|_\infty \leq \|\beta_0\|_2$, the result follows. □

APPENDIX B: AUXILIARY LEMMAS

LEMMA 3. *Under Conditions A and B, if* $\max_{j\leq p} \frac{1}{n}\sum_{i=1}^n x_{ij}^4 \leq C$*, we have that* $P(H_n \leq C') = 1 - o(1)$.

PROOF OF LEMMA 3. We have

$$\begin{aligned} H_n^2 &= \max_{1\leq j,k,\ell\leq p}\left|\frac{1}{n}\sum_{i=1}^n (z_{ij}z_{ik} - \hat\Gamma_{jk})(z_{ij}z_{i\ell} - \hat\Gamma_{j\ell})\right| \\ &\leq \max_{1\leq j,k,\ell\leq p}\left|\frac{1}{n}\sum_{i=1}^n z_{ij}^2 z_{ik}z_{i\ell}\right| + \|\hat\Gamma\|_\infty^2 \\ &+2\|\hat\Gamma\|_\infty \max_{1\leq j,k\leq p}\left|\frac{1}{n}\sum_{i=1}^n z_{ij}z_{ik}\right|. \end{aligned}$$

Note that

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^n z_{ij}^2 z_{ik}z_{i\ell}\right| &\leq \frac{1}{n}\sum_{i=1}^n z_{ij}^2\frac{(z_{ik}^2+z_{i\ell}^2)}{2} \\ &\leq \frac{1}{4}\left|\frac{1}{n}\sum_{i=1}^n z_{ij}^4 + z_{ik}^4\right| + \frac{1}{4}\left|\frac{1}{n}\sum_{i=1}^n z_{ij}^4 + z_{i\ell}^4\right| \\ &\leq \max_{j\leq p}\frac{1}{n}\sum_{i=1}^n z_{ij}^4. \end{aligned}$$

Moreover,

$$\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n z_{ij}^4 \leq 8\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n x_{ij}^4 + 8\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n w_{ij}^4 \leq 8C + 8\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n w_{ij}^4.$$

Since w_{ij} are σ_w-subgaussian random variables with $\sigma_w \leq C$, Lemma 6 yields $M_4 = \mathrm{E}[\max_{i\leq n, j\leq p} w_{ij}^4] \leq C\log^2(pn)$. This and Lemma 7 imply

$$\begin{aligned} \mathrm{E}[\max_{j\leq p}\frac{1}{n}\sum_{i=1}^n w_{ij}^4] &\leq \frac{CM_4\log(2p)}{n} + \sqrt{\frac{CM_4\log(2p)}{n}}\max_{j\leq p}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{E}[w_{ij}^4]\right)^{1/2} \\ &\leq \frac{C'\log^3(pn)}{n} + \sqrt{\frac{C'\log^3(pn)}{n}} = o(1) \end{aligned}$$

where the last equality follows from the relation $\Phi^{-1}(1-\alpha/(2p)) = o(n^{-1/6})$. The result now follows since $\|\hat\Gamma\|_\infty$ is bounded: $\|\hat\Gamma\|_\infty \leq b_\epsilon + \|\Gamma\|_\infty$. □

B.1. Bounds on the stochastic error terms. The following technical lemmas were proved in [1] and [26] and are stated here for completeness. 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 4. *Let $0 < \varepsilon < 1$ and assume Condition A holds. Then, with probability at least $1-\varepsilon$ (for each event),*

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

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

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

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_ξ, σ_w.

LEMMA 5. *Let $0 < \varepsilon < 1$, $\theta^* \in \mathbb{R}^p$ and assume that Condition A holds. Then, with probability at least $1-\varepsilon$, $\left\|\frac{1}{n}X^TW\theta^*\right\|_\infty \le \delta_1'(\varepsilon)\|\theta^*\|_2$, where $\delta_1'(\varepsilon) = \sigma_w\sqrt{\frac{2m_2\log(2p/\varepsilon)}{n}}$. In addition, with probability at least $1-\varepsilon$,*

$$\left\|\tfrac{1}{n}(W^TW - \mathrm{Diag}\{W^TW\})\theta^*\right\|_\infty \le \delta_4'(\varepsilon)\|\theta^*\|_2,$$

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_w.

LEMMA 6. *(1) If X is a centered subgaussian random variable with parameter γ, it follows that for any $k > 0$ $\mathrm{E}[|X|^k] \le k2^{k/2}\gamma^k\Gamma(k/2)$ and for $p \ge 1$ we have $\{\mathrm{E}[|X|^k]\}^{1/k} \le C\gamma\sqrt{k}$. (2) If $X_j, j = 1, \ldots, N$, is a collection of centered subgaussian variables with parameter γ, then for $k \ge 1$ we have $\mathrm{E}\left[\max_{j\le N}|X_j|^k\right] \le \gamma^k\log^{k/2}(NC_k)$ for some constant C_k that depends only on k.*

LEMMA 7. *(e.g.,[1]) Let $X_i, i = 1, \ldots, n$, be independent random vectors in $\mathbb{R}^p$, $p \ge 3$. Define $\bar{m}_k := \max_{j\le p}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|X_{ij}|^k]$ and $M_k \ge \mathbb{E}[\max_{i\le n}\|X_i\|_\infty^k]$. Then*

$$\mathbb{E}\Big[\max_{j\le p}\frac{1}{n}\Big|\sum_{i=1}^n |X_{ij}|^k - \mathbb{E}[|X_{ij}|^k]\Big|\Big] \le 2C^2\frac{\log p}{n}M_k + 2C\sqrt{\frac{\log p}{n}}M_k^{1/2}\bar{m}_k^{1/2}$$

for some universal constant C.

REFERENCES

[1] A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming approaches to high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society, Series B*, 79:939–956, 2017.

[2] A. Belloni, V. Chernozhukov, and A. Kaul. Confidence bands for coefficients in high dimensional linear models with error-in-variables. *Arxiv 1703.00469*, 2017.

[3] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] A. Belloni, M. Rosenbaum, and A. B. Tsybakov. An $\{L_1, L_2, L_\infty\}$-approach to high-dimensional errors-in-variables models. *Electronic Journal of Statistics*, 10(2):1729–1750, 2016.

[5] Alexandre Belloni, Mingli Chen, and Victor Chernozhukov. Quantile graphical models: prediction and conditional independence with applications to financial risk management. *arXiv preprint arXiv:1607.00286*, 2016.

[6] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

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

[8] Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

[9] Y. Chen and C. Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high-dimensional results. *arXiv:1206.0823*, 2012.

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

[11] Abhirup Datta and Hui Zou. Cocolasso for high-dimensional error-in-variables regression. *arXiv preprint arXiv:1510.07123*, 2015.

[12] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.

[13] W.A. Fuller. *Measurement Error Models.* Wiley & Sons, Inc. New York, 1987.

[14] Eric Gautier and Alexandre B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454v4*, 2011.

[15] Eric Gautier and Alexandre B. Tsybakov. Pivotal uniform inference in high-dimensional regression with random design in wide classes of models via linear programming. Unpublished manuscript. 2012.

[16] Eric Gautier and Alexandre B. Tsybakov. Pivotal estimation in high-dimensional regression via linear programming. In *Empirical Inference*, pages 195–204. Springer, 2013.

[17] Zvi Griliches and Jerry A Hausman. Errors in variables in panel data. *Journal of econometrics*, 31(1):93–118, 1986.

[18] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized Cramér-type large deviations for independent random variables. *The Annals of Probability*, 31(4):2167–2215, 2003.

[19] Abhishek Kaul and Hira L Koul. Weighted ℓ_1-penalized corrected quantile regression for high dimensional measurement error models. *Journal of Multivariate Analysis*, 140:72–91, 2015.

[20] Abhishek Kaul, Hira L Koul, Akshita Chawla, and Soumendra N Lahiri. Two stage non-penalized corrected least squares for high dimensional linear models with measurement error or missing covariates. *arXiv preprint arXiv:1605.03154*, 2016.

[21] Roger Koenker, Stephen Portnoy, Pin Tian Ng, Achim Zeileis, Philip Grosjean, and Brian D Ripley. Package quantreg. *Quantile Regression. In*, 5, 2017.

[22] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664,

2012.

[23] L. A. Stefanski R. J. Carroll, D. Ruppert and C. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, New York, 2006.

[24] Marie Reilly and Margaret Sullivan Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

[25] Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[26] Mathieu Rosenbaum and Alexandre B. Tsybakov. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics, 2013.

[27] Mark Rudelson and Shuheng Zhou. High dimensional errors-in-variables models with dependent measurements. *arXiv preprint arXiv:1502.02355*, 2015.

[28] Øystein Sørensen, Arnoldo Frigessi, and Magne Thoresen. Covariate selection in high-dimensional generalized linear models with measurement error. *arXiv preprint arXiv:1407.1070*, 2014.

[29] Øystein Sørensen, Arnoldo Frigessi, and Magne Thoresen. Measurement error in lasso: Impact and correction. *Statistica Sinica*, 25(2):809–829, 2015.

[30] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

B. Supplementary Material: Deferred Proofs.

Proof of Lemma 1. Note that $\log(1-\Phi(t))$ is a concave function so that the supergradient inequality yields

$$\log\{1-\Phi(a/\{1+\gamma\})\} \leq \log\{1-\Phi(a)\} + \frac{\phi(a)}{1-\Phi(a)}\frac{\gamma}{1+\gamma}a,$$

where ϕ denotes the Gaussian density function. The result follows by noting that $\frac{\phi(t)}{1-\Phi(t)} \leq \{t^2+1\}/t \leq 2t$ if $t \geq 1$, and exponentiating both sides of the inequality. $\square$

Proof of Theorem 2. Recall that $t_j(\beta) = \{\frac{1}{n}\sum_{i=1}^n\{z_{ij}(y_i - z_i^T\beta) + \hat{\Gamma}_{jj}\beta_j\}^2\}^{1/2}$ and define $\tilde{t}_j(\beta) = \{\frac{1}{n}\sum_{i=1}^n\{z_{ij}(y_i - z_i^T\beta) + \Gamma_{jj}\beta_j\}^2\}^{1/2}$. Then we can write the threshold in the jth component of the estimator as $\bar{v}_j = \tau t_j(\hat{\beta})/\frac{1}{n}\sum_{i=1}^n z_{ij}^2$, $j=1,\ldots,p$. Further, note that $\mathrm{E}[\tilde{t}_j^2(\beta_0)] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[\{z_{ij}(\xi_i - w_i^T\beta_0) + \Gamma_{jj}\beta_{0j}\}^2]$.

We first derive upper and lower bounds on $t_j(\hat{\beta})$ by controlling the value

$$|t_j(\hat{\beta}) - \{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}| \leq |t_j(\hat{\beta}) - t_j(\beta_0)| + |t_j(\beta_0) - \tilde{t}_j(\beta_0)| + |\tilde{t}_j(\beta_0) - \{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}|$$

via triangle inequality and using the bracketing

$$c(1+\|\beta_0\|_2)^2 \leq \mathrm{E}[\tilde{t}_j^2(\beta_0)] \leq C(1+\|\beta_0\|_2)^2 \tag{B.1}$$

that holds by assumption.

By (A.5) we have with probability $1-o(1)$

$$|t_j(\hat{\beta}) - t_j(\beta_0)| \leq H_n\|\hat{\beta}-\beta_0\|_1 \leq C\|\hat{\beta}-\beta_0\|_1 = o(1+\|\beta_0\|_2)$$

since $H_n \leq C$ with probability $1-o(1)$ by Lemma 3 when $X \in \Omega_X$, Condition A and B hold, and $\|\hat{\beta}-\beta_0\|_1 = o(1+\|\beta_0\|_2)$ by Theorem 1 with $\alpha = \log n$ under the assumed condition $s^{1/q}\sqrt{\log(2pn/\alpha)/n} = o(1)$ for $q \in \{1,2\}$.

Moreover, we have

$$|t_j(\beta_0) - \tilde{t}_j(\beta_0)| \leq |\hat{\Gamma}_{jj} - \Gamma_{jj}|\ |\beta_{0j}| \leq b_\varepsilon\|\beta_0\|_2 = o(\|\beta_0\|_2)$$

under our conditions that imply $b_\varepsilon = o(1)$.

Next, Lemma 7 implies

$$\begin{aligned}\mathrm{E}\left[\max_{j\leq p}|\tilde{t}_j^2(\beta_0) - \mathrm{E}[\tilde{t}_j^2(\beta_0)]|\right] &\leq \tfrac{CM_2\log(3p)}{n} \\ &+\sqrt{\tfrac{CM_2\log(2p)}{n}}\max_{j\leq p}\{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}\end{aligned}$$

where $M_2 := \mathrm{E}[\max_{i\leq n, j\leq p}|z_{ij}(\xi_i - w_i^T\beta_0) + \Gamma_{jj}\beta_{0j}|^2]$. The quantity M_2 satisfies

$$\begin{aligned} M_2 \; & \leq 2\mathrm{E}[\max_{i\leq n, j\leq p}|x_{ij}(\xi_i - w_i^T\beta_0) + \Gamma_{jj}\beta_{0j}|^2] \\ & +2\mathrm{E}[\max_{i\leq n, j\leq p}|w_{ij}(\xi_i - w_i^T\beta_0) + \Gamma_{jj}\beta_{0j}|^2] \\ & \leq 4\max_{i\leq n, j\leq p}|x_{ij}|^2\mathrm{E}[\max_{i\leq n}|\xi_i - w_i^T\beta_0|^2] + 8|\Gamma_{jj}\beta_{0j}|^2 \\ & +4\mathrm{E}[\max_{i\leq n, j\leq p}|w_{ij}(\xi_i - w_i^T\beta_0)|^2] \\ & \leq C\max_{i\leq n, j\leq p}|x_{ij}|^2(1+\|\beta_0\|_2)^2\log(n) + C\|\beta_0\|_2^2 \\ & +4\mathrm{E}[\max_{i\leq n, j\leq p}|w_{ij}|^4]^{1/2}\mathrm{E}[\max_{i\leq n}|\xi_i - w_i^T\beta_0|^4]^{1/2} \\ & \leq C'n^{1/2}(1+\|\beta_0\|_2)^2\log(n) + C\log(pn)(1+\|\beta_0\|_2)^2\log(n) \end{aligned} \tag{B.2}$$

where we used the inequalities $\max_{i\leq n, j\leq p}|x_{ij}| \leq n^{1/4}\max_{i\leq n, j\leq p}\left(\frac{1}{n}\sum_{i=1}^n x_{ij}^4\right)^{1/4} \leq Cn^{1/4}$ and Lemma 6. Finally, note that

$$\begin{aligned} & \max_{j\leq p}|\tilde{t}_j(\beta_0) - \{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}| \leq \max_{j\leq p}\frac{|\tilde{t}_j^2(\beta_0) - \mathrm{E}[\tilde{t}_j^2(\beta_0)]|}{\{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}} \\ & \leq \max_{j\leq p}\frac{|\tilde{t}_j^2(\beta_0) - \mathrm{E}[\tilde{t}_j^2(\beta_0)]|}{c'(1+\|\beta_0\|_2)} \\ & = O_P\Big(\frac{C'n^{1/2} + C\log(pn)}{n}\Big)(1+\|\beta_0\|_2)\log(n)\log(pn) \\ & +O_P\Big(\frac{C'n^{1/2} + C\log(pn)}{n}\Big)^{1/2}(1+\|\beta_0\|_2)\log^{1/2}(n)\log^{1/2}(pn) \\ & = (1+\|\beta_0\|_2)o_P(1) \end{aligned}$$

where we used Markov's inequality, (B.2) and the fact that $\log^2(n)\log^2(pn) = o(n)$, which is due to the relation $\Phi^{-1}(1-\alpha/(2pn)) = o(n^{1/6})$ in Condition B(ii).

Thus, uniformly over $j \in \{1,\ldots,p\}$, we have

$$|t_j(\hat{\beta}) - \{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}| = (1+\|\beta_0\|_2)o_P(1).$$

This implies that $|t_j(\hat{\beta})|$ satisfies, with high probability, the same bracketing as $|\{\mathrm{E}[\tilde{t}_j^2(\beta_0)]\}^{1/2}|$, cf. (B.1). Since $\frac{1}{n}\sum_{i=1}^n z_{ij}^2$ is bounded away from zero and from above by constants uniformly in j with probability $1-o(1)$, we have $\min_{j\leq p}\bar{v}_j \geq c\tau(1+\|\beta_0\|_2)$ and $\max_{j\leq p}\bar{v}_j \leq C\tau(1+\|\beta_0\|_2)$. Applying (B.4) in Lemma 8 given below with $\nu_{\min} = \min_{j\leq p}\bar{v}_j$ and Corollary 1 with $q=1$ we get

$$\|\hat{\beta}_{\hat{T}}\|_0 \quad \leq s + \tfrac{\|\hat{\beta}-\beta_0\|_1}{c\tau(1+\|\beta_0\|_2)} \leq s + \tfrac{C(1+\|\beta_0\|_2)s\sqrt{\log(c'p/(\alpha\varepsilon))/n}}{c\tau(1+\|\beta_0\|_2)} \leq C's$$

where we have used the fact that $\tau = n^{-1/2}\Phi^{-1}(1-\alpha/(2p)) \geq \sqrt{2\log(2p/\alpha)/n}$.

Similarly, to prove the bounds on the ℓ_1 and ℓ_2 errors of the thresholded estimator, we use inequalities (B.3) and (B.5) in Lemma 8 below with $\nu_{\max} = \max_{j\leq p}\bar{v}_j$, the bounds of Corollary 1 with $q \in \{1,2\}$, and the fact that $\tau \leq C\sqrt{\log(2p/\alpha)/n}$. □

The following lemma provides bounds for general thresholded estimators (see also a related lemma in [5]).

LEMMA 8. *Let $\hat\beta, \beta_0 \in \mathbb{R}^p$ be such that $\|\beta_0\|_0 \le s$. Denote by $\hat\beta^\nu = (\hat\beta^\nu_1, \dots, \hat\beta^\nu_p)$ the vector obtained by thresholding the components $\hat\beta_j$ of $\hat\beta$ as follows: $\hat\beta^\nu_j = \hat\beta_j 1\{|\hat\beta_j| \ge \nu_j\}$ where ν_j are positive numbers. Then,*

$$\|\hat\beta^\nu - \beta_0\|_1 \le \|\hat\beta - \beta_0\|_1 + s\nu_{\max}, \tag{B.3}$$

$$\|\hat\beta^\nu\|_0 \le s + \|\hat\beta - \beta_0\|_1/\nu_{\min}, \tag{B.4}$$

$$\|\hat\beta^\nu - \beta_0\|_2 \le \|\hat\beta - \beta_0\|_2 + 2\sqrt{s}\nu_{\max} + \frac{2\|\hat\beta - \beta_0\|_1}{\sqrt{s}} \tag{B.5}$$

where $\nu_{\max} = \max_{j\le p} \nu_j$ and $\nu_{\min} = \min_{j\le p} \nu_j$.

PROOF OF LEMMA 8. Let $T = \text{supp}(\beta_0)$. The bound (B.3) follows from the chain of inequalities

$$\begin{aligned}\|\hat\beta^\nu - \beta_0\|_1 &= \|(\hat\beta^\nu - \beta_0)_T\|_1 + \|(\hat\beta^\nu)_{T^c}\|_1\\ &\le \|(\hat\beta^\nu - \hat\beta)_T\|_1 + \|(\hat\beta - \beta_0)_T\|_1 + \|(\hat\beta^\nu)_{T^c}\|_1\\ &\le s\nu_{\max} + \|(\hat\beta - \beta_0)_T\|_1 + \|(\hat\beta)_{T^c}\|_1\\ &= s\nu_{\max} + \|\hat\beta - \beta_0\|_1.\end{aligned}$$

To prove the bound (B.4), set $\hat T = \text{supp}(\hat\beta^\nu)$, and note that

$$\|\hat\beta - \beta_0\|_1 \ge \|\hat\beta_{T^c}\|_1 \ge \|\hat\beta_{T^c \cap \hat T}\|_1 \ge \nu_{\min}|T^c \cap \hat T| \ge \nu_{\min}(|\hat T| - |T|) \ge \nu_{\min}(\|\hat\beta^\nu\|_0 - s).$$

We now show (B.5). By the triangle inequality,

$$\|\hat\beta^\nu - \beta_0\|_2 \le \|\hat\beta^\nu - \hat\beta\|_2 + \|\hat\beta - \beta_0\|_2. \tag{B.6}$$

Without loss of generality assume that the order of the components is such that $|\hat\beta^\nu_j - \hat\beta_j|$ is non-increasing in j. Let T_1 be the set of indices j corresponding to the s largest values of $|\hat\beta^\nu_j - \hat\beta_j|$. Similarly, define T_k as the set of indices corresponding to the s largest values of $|\hat\beta^\nu_j - \hat\beta_j|$ outside $\cup_{m=1}^{k-1} T_m$. Therefore, $\hat\beta^\nu - \hat\beta = \sum_{k=1}^{\lceil p/s \rceil} (\hat\beta^\nu - \hat\beta)_{T_k}$. Moreover, $\|(\hat\beta^\nu - \hat\beta)_{T_k}\|_2 \le \|(\hat\beta^\nu - \hat\beta)_{T_{k-1}}\|_1/\sqrt{s}$ in view of the monotonicity of the components. Thus,

$$\begin{aligned}\|\hat\beta^\nu - \hat\beta\|_2 &= \| \textstyle\sum_{k=1}^{\lceil p/s \rceil} (\hat\beta^\nu - \hat\beta)_{T_k}\|_2\\ &\le \|(\hat\beta^\nu - \hat\beta)_{T_1}\|_2 + \textstyle\sum_{k\ge 2} \|(\hat\beta^\nu - \hat\beta)_{T_k}\|_2\\ &\le \|(\hat\beta^\nu - \hat\beta)_{T_1}\|_2 + \textstyle\sum_{k\ge 2} \|(\hat\beta^\nu - \hat\beta)_{T_k}\|_2\\ &\le \nu_{\max}\sqrt{s} + \textstyle\sum_{k\ge 1} \|(\hat\beta^\nu - \hat\beta)_{T_k}\|_1/\sqrt{s}\\ &= \nu_{\max}\sqrt{s} + \|\hat\beta^\nu - \hat\beta\|_1/\sqrt{s}\\ &\le 2\sqrt{s}\nu_{\max} + 2\|\hat\beta - \beta_0\|_1/\sqrt{s}\end{aligned}$$

where we have used the bound $|\hat\beta^\nu_j - \hat\beta_j| \le \nu_j$ valid for all j, and then (B.3). Inequality (B.5) follows by combining the last display with (B.6). □

C. Supplementary Material: Numerical and Optimizational Issues.

The pivotality of the self-normalized estimator is achieved by the introduction of p second order cone constraints which is more computationally demanding than the $2p$ linear constraints (associated with the near zero score condition). In order to solve the optimization problem defined in (2.2), it is convenient to formulate it as

$$\begin{array}{ll} \min\limits_{\theta,\nu} & c^T\theta \\ s.t. & A\theta + \nu = b \\ & (\theta,\nu) \in \mathbb{R}^d \times \mathcal{K}, \end{array}$$

where θ is a vector that contains the positive and negative parts of β and auxiliary variables, and $\mathcal{K}$ is the cartesian product of non-negative cones and second order cones. The introduction of residual variables $\varepsilon_i = y_i - z_i^T(\beta^+ - \beta^-)$ proves to be helpful in the implementation to further exploit sparsity in the design matrix in the p second order constraints. Indeed the additional residual variables allow us to write the second order constraints for $j = 1,\ldots,p$ as

$$\left\{\frac{1}{n}\sum_{i=1}^n \{z_{ij}\varepsilon_i + \hat{\Gamma}_{jj}(\beta_j^+ - \beta_j^-)\}^2\right\}^{1/2} \leq t_j$$

instead of

$$\left\{\frac{1}{n}\sum_{i=1}^n \{z_{ij}(y_i - z_i^T(\beta^+ - \beta^-)) + \hat{\Gamma}_{jj}(\beta_j^+ - \beta_j^-)\}^2\right\}^{1/2} \leq t_j.$$

The difference in these representations come from what multiplies $\beta^+ - \beta^-$. In the first formulation, $\operatorname{diag}(z_{1j}, z_{2j}, \ldots, z_{pj})$ multiplies ε while $e\Gamma_{j\cdot}$ multiplies $\beta^+ - \beta^-$. In the second formulation we have $-z_{\cdot j}z_i^T + e\Gamma_{j\cdot}$ multiplying $\beta^+ - \beta^-$ where e is the n-vector of ones. These three matrices have n rows but they are sparse in the first formulation and typically dense in the second formulation. Since these matrices are formed p times (one for each j), this has non-negligible consequences on the software performance.

D. Supplementary Material: Additional Simulations.

In this section, we provide additional simulations to illustrate the finite sample performance of the self-normalized conic estimator and its thresholded version. We consider both the additive errors in variables (EIV) and the covariates missing at random. The latter is a setting where the bias correction matrix is estimated from the data, while for the additive EIV setting we assume for simplicity that this matrix is known.

Case 1 - Additive EIV: consider the data generating process where ξ_i, w_i, and x_i are independent and Gaussian. More precisely we set $\xi_i \sim N(0, \sigma_\xi^2)$, $w_i \sim N(0, \sigma_w^2 I_{p\times p})$, and $x_i \sim N(0, \Sigma)$ where $I_{p\times p}$ is an identity matrix and Σ is a $p \times p$ matrix with elements $\Sigma_{ij} = \rho^{|i-j|}$. We set $\sigma_\xi = 1$, $\sigma_w = 1$ and $\rho = 0.5$. We assume σ_w to be known in all calculations so we set $\hat{\Gamma} = \sigma_w^2 I_{p\times p}$ so that $b_\epsilon = 0$.

Case 2 - Covariates missing at random: here we consider the case where the error in measurements represents missing data. For this purpose we follow the framework of [26], i.e., we observe $(y_i, \tilde{z}_i, \eta_i,\ i = 1, \dots, n)$ where

$$\tilde{z}_{ij} = x_{ij}\eta_{ij}, \quad \eta_{ij} \text{ i.i.d Bernoulli with parameter } 1 - \pi,$$

the r.v.'s ξ_i, and x_i are generated as for Case 1 and $\eta_{ij} = 0$ indicates that we are missing the observation x_{ij}. For numerical comparisons, the parameter π is chosen uniformly from the interval $(0.1, 0.75)$ for each simulated repetition. The bias correction matrix is estimated as described in [26] and we set $b_\epsilon = c\sqrt{\log(2p/\epsilon)/n}$, $c = 0.25$, $\epsilon = 0.05$.

We consider two types of coefficients β_0: (i) $\beta_0 = (1, \dots, 1, 0, \dots, 0)$ where the first six coefficients are set to one, (ii) $\beta_0 = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{10}, 0, \dots, 0)$. The first vector of coefficients illustrates the case where parameters are well separated from zero, and the second case is without such a separation. The latter typically leads to model selection mistakes with high probability.

The estimators proposed in this paper, namely the self-normalized estimator (SN-conic) and its thresholded version (SN-conic thresholded) are compared with the conic estimator (Conic) proposed in [1]. We also consider the refitted estimator (SN-conic refitted) defined as follows: We first find the selected set of variables for the thresholded SN-conic estimator and then compute the SN-conic estimator for the model restricted to this set. We provide additional benchmarks for the performance, namely the biased and the no measurement error Lasso and Dantzig selector respectively. The biased versions are obtained with the observed design variables z_i's and the no measurement error versions are computed with the unobserved design variables x_i's. We report the following metrics which are computed based on 100 simulated repetitions: bias ($\|\mathrm{E}[\hat{\beta} - \beta_0]\|_2$), root mean squared errorxf (RMSE, $\mathrm{E}[\|\hat{\beta} - \beta_0\|^2]^{1/2}$), prediction bias (PRb, $\|\mathrm{E}(X(\hat{\beta} - \beta_0))\|_2/\sqrt{n}$), ℓ_2-rate (L2, $\mathrm{E}[\|\hat{\beta} - \beta_0\|_2]$), ℓ_1-rate (L1, $\mathrm{E}[\|\hat{\beta} - \beta\|_1]$, prediction risk (PR, $\mathrm{E}[\|X(\hat{\beta} - \beta_0)\|_2]/\sqrt{n}$), false positives (FP, expected number of misidentified zero elements of β_0), false negatives (FN, expected number of misidentified non-zero elements of β_0), and time (average computation time in seconds (CPU: i7@3.1GHz, 16GB RAM)).

The conic estimator is tuned assuming that $\sigma_\xi = 1$ is known and setting $\tau = \sqrt{\log(p/\alpha)/n}$ while the SN-conic has its penalty parameters set to $\tau = cn^{-1/2}\Phi^{-1}(1 - \alpha/2p)$, $c = 1, 0.5$, $\alpha = 0.05$ and $\lambda_t = 1$, $\lambda_u = 0.25$. Similarly to the conic estimator, Lasso and Dantzig selector are also tuned assuming that $\sigma_\xi = 1$ and we set their regularization parameters to $2cn^{-1/2}\Phi^{-1}(1 - \alpha/(2p))$ where

$c = 1.1, \alpha = 0.05$ (as suggested in [3]), and $\sqrt{2(\log p)/n}$, respectively. Computations are performed in R using the optimization software Mosek, an interior point methods solver, wrapped through the R package Rmosek. The Dantzig selector is computed via the R package quantreg [21].

The proposed methods, SN-conic, SN-conic thresholded and SN-conic refitted provide good results at all three levels of p considered in the simulations, with the performance deteriorating slightly with increase in dimension. The results for the additive error case are reported in Tables 2 to 5 and the results for the missing covariates case are reported in Tables 6 to 9. These numerical findings support our theoretical results regarding consistency of the proposed methods. The poor performance of biased Lasso and Dantzig selector highlights the impact of disregarding errors in variables. The self-normalized estimators at $c = 1/2$ outperform the Conic estimator in all designs considered. At $c = 1$ the SN-conic refitted estimator outperforms the Conic estimator at higher dimensions ($p = 100$ and $p = 400$) and they are competitive at $p = 10$. We note that the (non self-tuning) Conic estimator provides a slightly better efficiency in comparison to SN-conic and SN-conic thresholded at $c = 1$.

Among the self-normalized estimators proposed here, the main difference between SN-conic thresholded and SN conic is in the significantly reduced number of false positives in the thresholded version. This comes at a price of a slightly higher false negative rate especially in the case where the coefficients are not well separated from zero, see Tables 4, 5, 8, 9. Another expected outcome is the reduction in bias obtained via the refitted version. Lastly, in context of computation time it may be of interest to note that it is an average of computation times over 100 replications running parallel on 8 cores and is thus subject to additional computational overhead time per instance, induced by the parallel processing. The computation times for running a dedicated single instance for all of the estimators are three to five times faster.

The Fuqua School of Business
Duke University
100 Fuqua Drive
Durham, NC 27708
E-mail: abn5@duke.edu

Department of Economics
Massachusetts Institute of Technology
52 Memorial Drive
Cambridge, MA 02139
E-mail: vchern@mit.edu

Department of Mathematics and Statistics
Washington State University
Pullman, WA 99164-3113
E-mail: akaul@math.wsu.edu

École Polytechnique
CMAP
91128
Palaiseau, cedex 05
France
E-mail: mathieu.rosenbaum@polytechnique.edu

CREST, ENSAE, Université Paris-Saclay
5, avenue Henry Le Chatelier
91764 Palaiseau Cedex
France
E-mail: alexandre.tsybakov@ensae.fr

$n = 300$	$\boldsymbol{\beta} = (1, 1, 1, 1, 1, 1, 0, \ldots, 0)$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.676	0.917	0.105	0.896	1.799	1.006	0.680	0.020	0.761
	100	0.912	1.116	0.134	1.095	2.267	1.329	2.480	0.040	33.897
	400	1.074	1.233	0.155	1.215	2.551	1.526	387.320	0.000	36.123
SN-conic $(c = 1)$ **(thresholded)**	10	0.678	0.921	0.106	0.898	1.799	1.010	0.040	0.100	0.761
	100	0.922	1.135	0.135	1.110	2.289	1.345	0.010	0.360	33.897
	400	1.097	1.267	0.160	1.245	2.593	1.558	0.000	0.640	36.123
SN-conic $(c = \frac{1}{2})$	10	0.360	0.715	0.066	0.688	1.431	0.629	1.280	0.000	0.762
	100	0.451	0.780	0.075	0.756	1.718	0.750	35.360	0.000	35.233
	400	0.578	0.859	0.090	0.837	1.996	0.865	389.140	0.000	39.797
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.360	0.715	0.066	0.688	1.425	0.629	0.430	0.000	0.762
	100	0.451	0.779	0.075	0.753	1.639	0.751	0.990	0.010	35.233
	400	0.580	0.858	0.091	0.835	1.877	0.868	1.620	0.030	39.797
SN-conic $(c = 1)$ **(refitted)**	10	0.631	0.901	0.099	0.876	1.736	0.953	0.040	0.100	1.370
	100	0.649	0.971	0.101	0.925	1.840	1.001	0.010	0.360	34.644
	400	0.737	1.062	0.112	1.010	1.967	1.065	0.000	0.640	38.541
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.342	0.721	0.064	0.688	1.417	0.616	0.400	0.000	1.342
	100	0.311	0.730	0.066	0.705	1.544	0.651	0.840	0.010	34.399
	400	0.361	0.773	0.074	0.743	1.664	0.678	1.310	0.030	39.610
Conic **(no meas error)**	10	0.568	0.948	0.092	0.916	1.877	0.892	0.430	0.020	0.122
	100	0.652	1.015	0.103	0.982	2.065	1.012	0.320	0.100	0.432
	400	0.729	1.039	0.111	1.015	2.110	1.072	0.720	0.080	1.426
Lasso **(biased)**	10	0.909	0.951	0.134	0.948	2.358	1.314	3.950	0.000	0.047
	100	0.955	1.023	0.139	1.019	3.212	1.371	52.830	0.000	0.388
	400	1.002	1.109	0.141	1.106	4.492	1.403	73.890	0.000	0.786
Lasso **(no meas error)**	10	0.213	0.285	0.033	0.279	0.574	0.321	3.190	0.000	0.040
	100	0.264	0.327	0.040	0.320	0.675	0.393	0.150	0.000	0.302
	400	0.298	0.353	0.043	0.348	0.734	0.433	0.090	0.000	0.666
Dantzig selector **(biased)**	10	0.801	0.861	0.116	0.858	2.274	1.135	3.910	0.000	0.007
	100	0.816	1.354	0.136	1.350	9.823	1.318	88.030	0.000	0.191
	400	0.909	2.375	0.183	2.369	29.694	1.854	267.570	0.000	11.830
Dantzig selector **(no meas error)**	10	0.035	0.216	0.017	0.206	0.514	0.166	3.220	0.000	0.008
	100	0.062	0.598	0.045	0.594	4.061	0.468	70.620	0.000	0.159
	400	0.106	0.963	0.081	0.961	10.813	0.788	206.830	0.000	10.628

TABLE 2

Additive error: numerical comparisons with separated coefficients at $n = 300$

$n = 400$	$\boldsymbol{\beta} = (\mathbf{1, 1, 1, 1, 1, 1, 0, \ldots, 0})$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.596	0.815	0.092	0.794	1.621	0.889	0.590	0.000	0.959
	100	0.774	0.919	0.107	0.907	1.888	1.113	12.050	0.000	40.442
	400	0.882	1.042	0.122	1.027	2.131	1.272	387.790	0.000	48.223
SN-conic $(c = 1)$ **(thresholded)**	10	0.597	0.817	0.092	0.795	1.619	0.890	0.040	0.030	0.959
	100	0.775	0.920	0.108	0.908	1.889	1.114	0.020	0.020	40.442
	400	0.892	1.060	0.124	1.042	2.148	1.287	0.000	0.250	48.223
SN-conic $(c = \frac{1}{2})$	10	0.316	0.640	0.059	0.612	1.286	0.563	1.400	0.000	0.971
	100	0.395	0.659	0.064	0.636	1.482	0.641	48.580	0.000	42.147
	400	0.450	0.712	0.071	0.692	1.670	0.715	389.210	0.000	58.474
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.316	0.640	0.059	0.612	1.282	0.563	0.360	0.000	0.971
	100	0.395	0.657	0.064	0.634	1.425	0.640	1.290	0.000	42.147
	400	0.449	0.707	0.071	0.687	1.542	0.714	1.560	0.000	58.474
SN-conic $(c = 1)$ **(refitted)**	10	0.553	0.786	0.087	0.763	1.549	0.837	0.040	0.030	1.749
	100	0.551	0.743	0.081	0.723	1.480	0.819	0.020	0.020	40.861
	400	0.567	0.840	0.085	0.798	1.583	0.870	0.000	0.250	53.888
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.295	0.633	0.058	0.606	1.268	0.548	0.350	0.000	1.687
	100	0.286	0.624	0.058	0.599	1.342	0.559	1.160	0.000	41.852
	400	0.263	0.662	0.059	0.637	1.460	0.594	1.360	0.000	59.883
Conic **(no meas error)**	10	0.492	0.827	0.081	0.791	1.631	0.783	0.370	0.020	0.149
	100	0.568	0.829	0.085	0.804	1.691	0.859	0.420	0.020	0.428
	400	0.607	0.878	0.090	0.850	1.782	0.925	0.530	0.000	1.662
Lasso **(biased)**	10	0.893	0.923	0.131	0.921	2.299	1.290	3.940	0.000	0.082
	100	0.929	0.981	0.130	0.979	3.068	1.336	54.810	0.000	0.510
	400	0.950	1.036	0.128	1.034	4.285	1.348	106.330	0.000	1.308
Lasso **(no meas error)**	10	0.188	0.240	0.029	0.235	0.486	0.279	3.220	0.000	0.076
	100	0.226	0.282	0.033	0.277	0.572	0.332	0.310	0.000	0.400
	400	0.241	0.290	0.033	0.287	0.608	0.359	0.090	0.000	1.226
Dantzig selector **(biased)**	10	0.798	0.841	0.115	0.838	2.230	1.132	3.870	0.000	0.013
	100	0.814	1.222	0.131	1.220	8.684	1.263	89.340	0.000	0.186
	400	0.843	2.595	0.180	2.590	36.911	1.971	331.630	0.000	15.756
Dantzig selector **(no meas error)**	10	0.022	0.175	0.014	0.169	0.421	0.137	3.340	0.000	0.007
	100	0.050	0.547	0.042	0.545	3.850	0.422	74.870	0.000	0.165
	400	0.103	0.994	0.073	0.992	12.375	0.753	251.020	0.000	13.172

TABLE 3

Additive error: numerical comparisons with separated coefficients at $n = 400$

$n = 300$	$\boldsymbol{\beta = (1, 1/2, 1/3, 1/4, 1/5, 1/10, 0, \ldots, 0)}$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.364	0.469	0.057	0.458	0.952	0.545	0.12	1.15	0.754
	100	0.486	0.55	0.066	0.542	1.128	0.702	5.97	0.84	32.445
	400	0.566	0.618	0.077	0.61	1.274	0.803	379.67	0	34.176
SN-conic $(c = 1)$ **(thresholded)**	10	0.374	0.481	0.059	0.47	0.981	0.563	0	2.34	0.754
	100	0.498	0.563	0.068	0.556	1.16	0.723	0	2.91	32.445
	400	0.58	0.636	0.08	0.629	1.316	0.832	0	3.36	34.176
SN-conic $(c = \frac{1}{2})$	10	0.191	0.383	0.037	0.369	0.789	0.346	0.73	0.65	0.754
	100	0.248	0.4	0.04	0.389	0.891	0.4	27.57	0.21	33.823
	400	0.311	0.453	0.046	0.445	1.096	0.47	382.63	0	39.049
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.194	0.386	0.038	0.372	0.79	0.35	0.21	1.25	0.754
	100	0.25	0.403	0.04	0.39	0.855	0.403	0.96	1.55	33.823
	400	0.313	0.453	0.046	0.445	1.013	0.474	1.42	1.76	39.049
SN-conic $(c = 1)$ **(refitted)**	10	0.33	0.457	0.054	0.447	0.931	0.511	0	2.36	1.306
	100	0.361	0.483	0.053	0.474	1.006	0.554	0	2.91	33.372
	400	0.395	0.539	0.057	0.53	1.141	0.612	0	3.36	37.559
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.179	0.376	0.036	0.362	0.767	0.336	0.2	1.27	1.288
	100	0.197	0.415	0.039	0.403	0.899	0.376	0.78	1.57	34.446
	400	0.215	0.464	0.041	0.451	1.03	0.42	1.03	1.77	40.587
Conic **(no meas error)**	10	0.315	0.453	0.052	0.44	0.925	0.498	0.05	1.31	0.116
	100	0.372	0.483	0.054	0.473	0.991	0.57	0.14	1.81	0.486
	400	0.412	0.525	0.06	0.515	1.083	0.617	0.22	1.99	1.467
Lasso **(biased)**	10	0.577	0.597	0.076	0.594	1.08	0.734	3.34	0.06	0.049
	100	0.608	0.626	0.076	0.623	1.241	0.776	15.4	0.31	0.384
	400	0.634	0.655	0.08	0.653	1.434	0.806	14.29	0.68	0.696
Lasso **(no meas error)**	10	0.187	0.254	0.031	0.249	0.518	0.298	2.63	0.01	0.045
	100	0.241	0.297	0.035	0.292	0.617	0.364	0.19	0.67	0.345
	400	0.264	0.314	0.04	0.311	0.664	0.401	0.07	0.89	0.644
Dantzig selector **(biased)**	10	0.484	0.522	0.062	0.519	1.008	0.588	3.71	0.05	0.01
	100	0.486	0.788	0.076	0.785	5.209	0.721	84.49	0.05	0.197
	400	0.541	1.276	0.104	1.274	15.041	1.027	253.04	0.29	11.757
Dantzig selector **(no meas error)**	10	0.033	0.221	0.018	0.213	0.538	0.168	3.29	0.12	0.006
	100	0.068	0.606	0.049	0.603	4.163	0.472	71.23	0.1	0.179
	400	0.11	0.96	0.082	0.958	10.828	0.788	206.53	0.19	11.037

Table 4

Additive error: numerical comparisons with unseparated coefficients at $n = 300$

$n = 400$	$\boldsymbol{\beta} = (\mathbf{1}, \mathbf{1/2}, \mathbf{1/3}, \mathbf{1/4}, \mathbf{1/5}, \mathbf{1/10}, \mathbf{0}, \ldots, \mathbf{0})$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic ($c = 1$)	10	0.338	0.418	0.05	0.41	0.859	0.501	0.11	1.17	0.988
	100	0.419	0.478	0.061	0.473	1.001	0.615	3.01	0.99	40.715
	400	0.478	0.529	0.072	0.523	1.106	0.688	384.6	0	49.009
SN-conic ($c = 1$) **(thresholded)**	10	0.344	0.425	0.051	0.418	0.874	0.512	0	2.06	0.988
	100	0.429	0.49	0.063	0.484	1.027	0.632	0	2.63	40.715
	400	0.492	0.544	0.075	0.538	1.143	0.713	0	2.92	49.009
SN-conic ($c = \frac{1}{2}$)	10	0.186	0.333	0.034	0.322	0.683	0.319	0.5	0.54	1.032
	100	0.214	0.358	0.039	0.347	0.818	0.358	23.5	0.16	42.653
	400	0.252	0.399	0.043	0.389	0.97	0.406	385.63	0	61.968
SN-conic ($c = \frac{1}{2}$) **(thresholded)**	10	0.188	0.336	0.034	0.326	0.687	0.323	0.18	1.17	1.032
	100	0.217	0.36	0.039	0.349	0.787	0.361	1.11	1.32	42.653
	400	0.255	0.401	0.043	0.391	0.897	0.411	1.67	1.47	61.968
SN-conic ($c = 1$) **(refitted)**	10	0.306	0.402	0.047	0.394	0.822	0.464	0	2.07	1.658
	100	0.318	0.43	0.051	0.423	0.886	0.492	0	2.63	41.148
	400	0.343	0.457	0.056	0.449	0.962	0.524	0	2.92	55.526
SN-conic ($c = \frac{1}{2}$) **(refitted)**	10	0.174	0.329	0.033	0.318	0.671	0.31	0.15	1.17	1.676
	100	0.161	0.362	0.037	0.35	0.789	0.333	1.02	1.32	42.62
	400	0.174	0.42	0.039	0.409	0.955	0.38	1.33	1.51	63.879
Conic **(no meas error)**	10	0.296	0.39	0.046	0.381	0.807	0.455	0.06	1.32	0.149
	100	0.333	0.418	0.051	0.411	0.858	0.505	0.11	1.52	0.47
	400	0.36	0.45	0.057	0.442	0.921	0.546	0.28	1.67	1.559
Lasso **(biased)**	10	0.577	0.589	0.07	0.587	1.072	0.725	3.4	0.02	0.077
	100	0.591	0.605	0.075	0.604	1.187	0.75	18.31	0.23	0.535
	400	0.608	0.626	0.079	0.624	1.345	0.772	16.04	0.45	1.357
Lasso **(no meas error)**	10	0.181	0.233	0.027	0.228	0.487	0.276	3.01	0	0.067
	100	0.216	0.263	0.033	0.259	0.553	0.327	0.19	0.41	0.438
	400	0.228	0.273	0.036	0.269	0.572	0.345	0.04	0.74	1.267
Dantzig selector **(biased)**	10	0.494	0.516	0.057	0.514	0.966	0.594	3.76	0.02	0.011
	100	0.487	0.712	0.065	0.71	4.51	0.681	85.32	0.05	0.171
	400	0.524	1.349	0.112	1.347	17.88	1.052	308.59	0.14	14.493
Dantzig selector **(no meas error)**	10	0.021	0.183	0.014	0.177	0.442	0.143	3.55	0.01	0.01
	100	0.054	0.531	0.043	0.529	3.76	0.417	74.48	0.06	0.16
	400	0.104	1.007	0.078	1.005	12.549	0.759	252.15	0.17	12.881

Table 5

Additive error: numerical comparisons with unseparated coefficients at $n = 400$

$n = 300$	$\boldsymbol{\beta} = (\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{0},\ldots,\mathbf{0})$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.744	0.991	0.121	0.93	1.939	1.096	1.14	0.03	0.795
	100	1.049	1.266	0.17	1.194	2.623	1.542	46.61	0.01	17.773
	400	1.091	1.282	0.176	1.225	2.724	1.608	329.8	0	56.727
SN-conic $(c = 1)$ **(thresholded)**	10	0.751	1.004	0.122	0.94	1.946	1.108	0.09	0.29	0.795
	100	1.074	1.303	0.176	1.223	2.632	1.58	0.06	0.85	17.773
	400	1.117	1.322	0.18	1.258	2.723	1.651	0.08	0.82	56.727
SN-conic $(c = \frac{1}{2})$	10	0.466	0.779	0.081	0.715	1.478	0.73	1.88	0	0.772
	100	0.593	0.876	0.096	0.804	1.85	0.904	55.9	0	17.801
	400	0.621	0.878	0.106	0.816	1.87	0.931	342.63	0	57.963
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.466	0.781	0.081	0.716	1.472	0.731	0.27	0.07	0.772
	100	0.593	0.876	0.097	0.803	1.801	0.906	1.09	0.1	17.801
	400	0.622	0.878	0.107	0.815	1.786	0.937	0.77	0.08	57.963
SN-conic $(c = 1)$ **(refitted)**	10	0.697	0.976	0.115	0.904	1.864	1.039	0.09	0.29	1.34
	100	0.8	1.154	0.136	1.044	2.09	1.199	0.05	0.85	18.582
	400	0.741	1.113	0.123	1.013	2	1.13	0.04	0.82	57.818
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.444	0.777	0.079	0.711	1.459	0.711	0.26	0.07	1.383
	100	0.449	0.801	0.081	0.725	1.591	0.75	0.99	0.11	18.693
	400	0.433	0.779	0.085	0.694	1.501	0.706	0.71	0.09	57.093
Conic **(no meas error)**	10	0.589	0.992	0.096	0.922	1.898	0.922	0.43	0.1	0.135
	100	0.68	1.028	0.108	0.963	2.158	1.052	1.27	0.19	0.482
	400	0.737	1.006	0.115	0.968	2.112	1.102	1.19	0.13	1.529
Lasso **(biased)**	10	0.821	0.93	0.13	0.867	2.109	1.212	3.84	0	0.072
	100	0.913	1.035	0.147	0.974	3.127	1.316	40.85	0	0.368
	400	0.858	1.005	0.137	0.944	3.668	1.229	55.22	0	0.506
Lasso **(no meas error)**	10	0.207	0.283	0.032	0.277	0.576	0.319	3.13	0	0.044
	100	0.268	0.323	0.04	0.319	0.671	0.387	0.12	0	0.332
	400	0.296	0.349	0.046	0.346	0.73	0.436	0.05	0	0.512
Dantzig selector **(biased)**	10	0.703	0.851	0.115	0.779	2.033	1.03	3.8	0	0.011
	100	0.767	1.326	0.137	1.297	9.207	1.285	87.15	0	0.171
	400	0.755	2.173	0.176	2.123	26.346	1.69	258.66	0	10.499
Dantzig selector **(no meas error)**	10	0.026	0.219	0.016	0.211	0.524	0.166	3.19	0	0.006
	100	0.056	0.598	0.047	0.595	4.118	0.467	71.05	0	0.153
	400	0.118	0.967	0.077	0.964	10.818	0.786	206.64	0	10.414

TABLE 6

Missing covariates: numerical comparisons with separated coefficients at $n = 300$

$n = 400$	$\boldsymbol{\beta} = (\mathbf{1,1,1,1,1,1,0,\ldots,0})$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.651	0.871	0.107	0.814	1.699	0.963	1.24	0.02	0.852
	100	0.887	1.107	0.136	1.039	2.256	1.298	33.46	0	23.842
	400	0.896	1.1	0.133	1.037	2.259	1.311	297.84	0	228.018
SN-conic $(c = 1)$ **(thresholded)**	10	0.652	0.874	0.107	0.816	1.695	0.966	0.1	0.08	0.852
	100	0.898	1.124	0.138	1.051	2.25	1.314	0.11	0.35	23.842
	400	0.907	1.117	0.135	1.05	2.248	1.326	0.05	0.35	228.018
SN-conic $(c = \frac{1}{2})$	10	0.387	0.646	0.068	0.595	1.237	0.605	2	0	0.841
	100	0.51	0.789	0.084	0.729	1.657	0.792	49.99	0	24.647
	400	0.507	0.767	0.081	0.703	1.618	0.769	329.2	0	224.011
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.387	0.647	0.069	0.595	1.232	0.606	0.4	0.03	0.841
	100	0.51	0.788	0.084	0.728	1.615	0.793	0.88	0.02	24.647
	400	0.507	0.766	0.082	0.702	1.567	0.772	0.97	0.04	224.011
SN-conic $(c = 1)$ **(refitted)**	10	0.605	0.833	0.101	0.774	1.596	0.903	0.1	0.08	1.44
	100	0.659	0.94	0.106	0.853	1.748	0.983	0.09	0.35	24.718
	400	0.586	0.886	0.092	0.799	1.619	0.881	0.04	0.35	223.667
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.369	0.637	0.066	0.584	1.205	0.583	0.4	0.03	1.493
	100	0.387	0.692	0.068	0.625	1.372	0.638	0.79	0.02	25.342
	400	0.34	0.678	0.066	0.61	1.355	0.591	0.91	0.04	217.39
Conic (no meas error)	10	0.512	0.852	0.085	0.797	1.652	0.807	0.61	0.04	0.138
	100	0.607	0.917	0.093	0.868	1.91	0.921	1.22	0.05	0.447
	400	0.631	0.902	0.096	0.858	1.879	0.961	1.11	0.04	2.173
Lasso (biased)	10	0.818	0.918	0.133	0.854	2.111	1.189	3.86	0	0.098
	100	0.866	0.99	0.138	0.922	2.949	1.262	47.07	0	0.539
	400	0.791	0.938	0.118	0.87	3.381	1.126	76.93	0	1.293
Lasso (no meas error)	10	0.186	0.246	0.028	0.24	0.492	0.279	3.16	0	0.047
	100	0.225	0.285	0.032	0.281	0.582	0.335	0.2	0	0.48
	400	0.245	0.292	0.035	0.289	0.614	0.357	0.03	0	1.218
Dantzig selector (biased)	10	0.714	0.847	0.12	0.772	2.048	1.028	3.79	0	0.009
	100	0.74	1.203	0.137	1.169	8.081	1.226	87.39	0	0.157
	400	0.687	2.318	0.175	2.26	31.669	1.703	317.7	0.01	15.053
Dantzig selector (no meas error)	10	0.022	0.188	0.016	0.183	0.457	0.147	3.49	0	0.007
	100	0.053	0.549	0.04	0.546	3.858	0.424	74.92	0	0.162
	400	0.108	0.981	0.074	0.979	12.208	0.743	250.48	0	13.83

TABLE 7

Missing covariates: numerical comparisons with separated coefficients at $n = 400$

$n = 300$	$\boldsymbol{\beta = (1, 1/2, 1/3, 1/4, 1/5, 1/10, 0, \ldots, 0)}$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.412	0.519	0.062	0.492	1.008	0.598	0.25	1.35	0.76
	100	0.52	0.61	0.075	0.59	1.24	0.754	31.12	0.32	18.499
	400	0.635	0.703	0.088	0.68	1.48	0.894	273.15	0.12	54.968
SN-conic $(c = 1)$ **(thresholded)**	10	0.422	0.532	0.063	0.506	1.037	0.618	0.02	2.4	0.76
	100	0.533	0.626	0.078	0.607	1.263	0.78	0.12	3.07	18.499
	400	0.657	0.726	0.091	0.704	1.48	0.937	0.09	3.59	54.968
SN-conic $(c = \frac{1}{2})$	10	0.252	0.404	0.039	0.376	0.77	0.383	0.8	0.74	0.77
	100	0.312	0.447	0.047	0.428	0.941	0.476	40.77	0.19	19.353
	400	0.386	0.505	0.056	0.477	1.125	0.556	296.7	0.07	56.182
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.254	0.407	0.04	0.378	0.774	0.387	0.16	1.42	0.77
	100	0.316	0.451	0.048	0.432	0.927	0.484	0.63	1.79	19.353
	400	0.39	0.508	0.057	0.481	1.07	0.566	1.18	1.88	56.182
SN-conic $(c = 1)$ **(refitted)**	10	0.369	0.499	0.056	0.475	0.973	0.553	0.02	2.42	1.253
	100	0.388	0.537	0.059	0.519	1.094	0.603	0.09	3.07	19.417
	400	0.461	0.609	0.069	0.584	1.255	0.694	0.07	3.59	54.957
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.235	0.402	0.038	0.374	0.771	0.368	0.16	1.44	1.285
	100	0.246	0.433	0.041	0.408	0.888	0.412	0.56	1.79	19.813
	400	0.284	0.476	0.046	0.443	1.005	0.45	1.01	1.93	55.692
Conic **(no meas error)**	10	0.334	0.453	0.05	0.435	0.914	0.502	0.11	1.41	0.136
	100	0.391	0.503	0.059	0.485	1.022	0.587	0.41	1.82	0.491
	400	0.437	0.537	0.063	0.52	1.139	0.637	1.38	1.9	1.594
Lasso **(biased)**	10	0.52	0.569	0.07	0.542	1.011	0.671	3.21	0.07	0.068
	100	0.553	0.597	0.072	0.577	1.186	0.724	8.77	0.42	0.449
	400	0.592	0.641	0.076	0.618	1.464	0.767	16.63	0.74	0.488
Lasso **(no meas error)**	10	0.187	0.252	0.029	0.247	0.521	0.293	2.66	0	0.045
	100	0.248	0.302	0.036	0.298	0.626	0.372	0.09	0.67	0.368
	400	0.266	0.31	0.039	0.305	0.658	0.399	0.07	0.87	0.539
Dantzig selector **(biased)**	10	0.412	0.496	0.057	0.462	0.937	0.518	3.67	0.04	0.011
	100	0.425	0.76	0.067	0.752	4.947	0.69	82.32	0.04	0.185
	400	0.487	1.245	0.101	1.237	14.447	1.013	245.7	0.37	10.478
Dantzig selector **(no meas error)**	10	0.027	0.21	0.017	0.203	0.505	0.163	3.32	0.1	0.006
	100	0.064	0.59	0.046	0.585	4.027	0.464	70.8	0.07	0.169
	400	0.101	0.965	0.08	0.962	10.8	0.79	205.12	0.22	10.499

TABLE 8

Missing covariates: numerical comparisons with unseparated coefficients at $n = 300$

$n = 400$	$\boldsymbol{\beta} = (\mathbf{1}, \mathbf{1/2}, \mathbf{1/3}, \mathbf{1/4}, \mathbf{1/5}, \mathbf{1/10}, \mathbf{0}, \ldots, \mathbf{0})$									
	p	**Bias**	**RMSE**	**PRb**	**L2**	**L1**	**PR**	**FP**	**FN**	**Time**
SN-conic $(c = 1)$	10	0.359	0.451	0.051	0.433	0.884	0.52	0.3	0.94	0.878
	100	0.46	0.532	0.071	0.517	1.103	0.675	21.19	0.41	24.175
	400	0.535	0.601	0.079	0.582	1.282	0.782	222.61	0.12	228.204
SN-conic $(c = 1)$ **(thresholded)**	10	0.367	0.46	0.052	0.444	0.903	0.535	0.01	2.1	0.878
	100	0.475	0.551	0.074	0.537	1.134	0.706	0.07	2.7	24.175
	400	0.549	0.615	0.081	0.596	1.28	0.807	0.07	3.14	228.204
SN-conic $(c = \frac{1}{2})$	10	0.223	0.362	0.033	0.343	0.715	0.343	1.1	0.39	0.836
	100	0.268	0.397	0.045	0.379	0.855	0.416	36.85	0.14	24.105
	400	0.315	0.424	0.05	0.405	0.956	0.48	282.55	0.05	230.558
SN-conic $(c = \frac{1}{2})$ **(thresholded)**	10	0.224	0.364	0.033	0.345	0.718	0.346	0.21	1.25	0.836
	100	0.27	0.399	0.045	0.381	0.838	0.42	0.79	1.3	24.105
	400	0.318	0.426	0.051	0.408	0.919	0.486	1.1	1.67	230.558
SN-conic $(c = 1)$ **(refitted)**	10	0.329	0.442	0.047	0.425	0.873	0.487	0.01	2.12	1.356
	100	0.345	0.478	0.057	0.466	0.985	0.533	0.05	2.7	24.897
	400	0.382	0.498	0.06	0.482	1.024	0.585	0.05	3.14	227.979
SN-conic $(c = \frac{1}{2})$ **(refitted)**	10	0.212	0.359	0.032	0.339	0.708	0.331	0.21	1.26	1.478
	100	0.216	0.394	0.04	0.369	0.823	0.359	0.69	1.32	25.364
	400	0.243	0.417	0.043	0.39	0.885	0.394	0.96	1.73	230.843
Conic **(no meas error)**	10	0.295	0.401	0.044	0.389	0.813	0.453	0.03	1.3	0.144
	100	0.332	0.453	0.053	0.437	0.933	0.513	0.4	1.46	0.505
	400	0.371	0.461	0.057	0.449	0.987	0.567	0.98	1.75	2.446
Lasso **(biased)**	10	0.51	0.551	0.066	0.528	0.974	0.652	3.3	0.04	0.136
	100	0.548	0.588	0.073	0.566	1.144	0.709	14.28	0.26	0.539
	400	0.566	0.609	0.076	0.587	1.371	0.736	19.31	0.53	1.155
Lasso **(no meas error)**	10	0.18	0.233	0.029	0.229	0.48	0.277	2.82	0.01	0.068
	100	0.207	0.252	0.031	0.249	0.529	0.316	0.15	0.46	0.508
	400	0.237	0.281	0.035	0.278	0.587	0.356	0.02	0.7	1.238
Dantzig selector **(biased)**	10	0.418	0.484	0.054	0.455	0.897	0.515	3.61	0.02	0.009
	100	0.44	0.712	0.072	0.703	4.501	0.664	84.48	0.01	0.198
	400	0.467	1.297	0.101	1.29	17.029	1.024	301.62	0.21	14.038
Dantzig selector **(no meas error)**	10	0.028	0.189	0.016	0.183	0.457	0.147	3.35	0.09	0.008
	100	0.056	0.534	0.042	0.531	3.798	0.413	75.69	0.08	0.165
	400	0.111	1.003	0.071	1.001	12.475	0.757	251.71	0.18	13.582

TABLE 9

Missing covariates: numerical comparisons with unseparated coefficients at $n = 400$