WORKING PAPERS

N° TSE –1032

August 2019

"Inference robust to outliers with L1-norm penalization"

Jad Beyhum

Toulouse School of Economics

Inference robust to outliers with ℓ_1-norm penalization[*]

Jad Beyhum [†]

Toulouse School of Economics, Université Toulouse Capitole

Abstract

This paper considers the problem of inference in a linear regression model with outliers where the number of outliers can grow with sample size but their proportion goes to 0. We apply an estimator penalizing the ℓ_1-norm of a random vector which is non-zero for outliers. We derive rates of convergence and asymptotic normality. Our estimator has the same asymptotic variance as the OLS estimator in the standard linear model. This enables to build tests and confidence sets in the usual and simple manner. The proposed procedure is also computationally advantageous as it amounts to solving a convex optimization program. Overall, the suggested approach constitutes a practical robust alternative to the ordinary least squares estimator.

KEYWORDS: robust regression, ℓ1-norm penalization, unknown variance.

MSC 2010 Subject Classification: Primary 62F35; secondary 62J05, 62J07.

[*] I thank my PhD supervisor Professor Eric Gautier for his availability and great help. I am also grateful to Anne Ruiz-Gazen, Jean-Pierre Florens, Thierry Magnac and Nour Meddahi for useful comments. I acknowledge financial support from the ERC POEMH 337665 grant.

[†] jad.beyhum@gmail.com

1 Introduction

This paper considers a linear regression model with outliers. The statistican observes a dataset of n i.i.d. realizations of an outcome scalar random variable y_i and a random vector of covariates x_i with support in $\mathbb{R}^K$. We assume that the following relationship holds:

$$y_i = x_i^\top \beta + \alpha_i + \epsilon_i \quad \forall i = 1, \dots, n, \ , \tag{1.1}$$

where $\beta \in \mathbb{R}^K$, ϵ_i, the error term, is a scalar random variable such that $\mathbb{E}[x_i \epsilon_i | \alpha_i = 0] = 0$ and α_i is a random variable. It also holds that $\{y_i, x_i, \epsilon_i, \alpha_i\}_i$ are i.i.d. and $\mathbb{E}[x_i x_i^\top | \alpha_i = 0]$ exists and is positive definite. The observation i is called an outlier if $\alpha_i \neq 0$. Let $\mathbb{P}(\alpha_i \neq 0)$, the average proportion of outliers, be denoted p. The goal is to obtain inference results on the parameter β.

This model can represent various situations of practical interest. First, the statistician could be interested in β if it corresponds to the coefficients of the best linear predictor of y_i given x_i conditional on $\alpha_i = 0$. In the presence of outliers, the coefficient of the best linear predictor of y_i given x_i for the whole population may differ greatly from β and hence a statistical analysis based on the whole sample may lead to a poor estimate of β.

Second, if β is given a causal interpretation, then it may represent the causal effect of the regressors for the population of "standard" individuals. That is, for instance, if the aim is to evaluate the effect of treatment which is statistically represented by a binary variable which takes value 1 if the individual is treated, it could be that the effect of treatment is negative for most of the population but strongly positive for a small fraction of the individuals, the outliers. The policy maker may not be willing to implement a policy that has a negative effect on most of the population, giving interest to a statistical procedure that estimates the treatment effect of the large majority of the population.

Finally, β could represent the true coefficient of the best linear predictor of $\tilde{y}_i$ given $\tilde{x}_i$ in a measurement errors model where we do not observe $(\tilde{y}_i, \tilde{x}_i)$ but (y_i, x_i). β may be of interest if one possesses a second sample where x_i is known but not y_i and is interested in predicting y_i from x_i. If the observed variables follow the model $\tilde{y}_i = \tilde{x}_i \beta + \tilde{\epsilon}_i$ with $\mathbb{E}[\tilde{x}_i \tilde{\epsilon}_i] = 0$, this fits our framework with $\epsilon_i = \tilde{\epsilon}_i$ and

$$\alpha_i = y_i - \tilde{y}_i + (\tilde{x}_i - x_i)\beta.$$

Hence, α_i allows for both measurement errors in x_i - called outliers in the x-direction - and in y_i, the outliers in the y-direction, for a small fraction of the population, see Rousseeuw and Leroy (2005) for a precise discussion.

This paper develops results on the estimation of β when the vector $\alpha = (\alpha_1, \dots, \alpha_n)^\top$ is sparse in the sense that p goes to 0 with n. We rely on a variant of the square-root lasso estimator of Belloni et al. (2011) which penalizes the ℓ_1-norm of the vector α. The advantages of our estimator are that the penalty parameter does not depend on the variance of the error term and is computationally tractable. If the vector α is sparse enough, we show that our estimator is $\sqrt{n}$-consistent and asymptotically normal. It has the same asymptotic variance as the OLS estimator in the standard linear model without outliers.

Related literature. This paper is connected to at least two different research fields. First, it draws on the literature on inference in the high-dimensional linear regression model. Our estimator is analogous to the concomitant lasso of Owen (2007). In particular, the computation algorithm outlined in Section 3 is similar to the one proposed for the scaled-lasso estimator introduced in Sun and Zhang (2012). We borrow from this literature by using an ℓ_1-penalized estimator and derive new inference results for the linear regression model with very few outliers.

Next, this paper is related to the literature on robust regression. For detailed accounts of this field, see Rousseeuw and Leroy (2005); Hampel et al. (2011); Maronna et al. (2018). The literature identifies a trade-off between efficiency and robustness, as explicited below. Indeed, M-estimators (such as the Ordinary Least-Squares (OLS) estimator) are often efficient when data is generated by the standard linear model without outliers and Gaussian errors but this comes at the price of robustness, as they can be asymptotically biased in the presence of outliers. By contrast, S-estimators such as the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS) are robust according to several measures of robustness developed in the literature. They are also asymptotically normal in the model with Gaussian errors and without outliers but have a larger asymptotic variance than the OLS estimator in the standard linear model. They also suffer from computational issues because of the non-convexity of their objective functions (see Rousseeuw and Leroy (2005)). The estimator proposed in this paper attains the same asymptotic variance as the OLS estimator in the standard linear model. Unlike in this literature, the computation algorithm outlined in Section 3 relies on a convex program and is computationally tractable. The proposed approach therefore provides a simple efficient alternative to the rest of the literature.

Within the robust regression literature some authors have considered the application of ℓ_1-norm penalization to robust estimation. In particular, the model studied in this paper nests the Huber's contamination model for location estimation introduced in Huber et al. (1964). Indeed, if there is a single constant regressor, model nests the following framework:

$$y_i = \beta + \alpha_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0,1)$ i.i.d., $\beta \in \mathbb{R}$ is the mean of y_i for non-outlying coefficients while $\mathbb{E}[y_i|\alpha_i \neq 0]$ is left unrestricted. Chen et al. (2018) show that the minimax lower bound for the squared

ℓ_2-norm estimation error is of order greater than $\max(1/n, p^2)$ under gaussian errors, where $||\alpha||_0$ is the number of outliers in the sample. When $p\sqrt{\log(n)} \to 0$, we attain this lower bound up to a factor $\log(n)^2$. Several strategies have been proposed to tackle this location estimation problem. The one which is closest to the approach studied in this paper is soft-thresholding using a lasso estimator, that is use

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{n} (y_i - \beta - \alpha_i)^2 + \lambda \sum_{i=1}^{n} |\alpha_i|, \ \lambda > 0,$$

see for instance Collier and Dalalyan (2017). We substitute this estimator with a square-root lasso that has the advantage to provide guidance on the choice of the penalty level that is independent from the variance of the errors (see Belloni et al. (2011)). We extend the analysis of this type of estimators to the linear regression model and develop inference results new to the literature. Other ℓ_1-norm penalized estimators for robust linear regression have been studied in the literature such as in Lambert-Lacroix et al. (2011); Dalalyan (2012); Li (2012); Alfons et al. (2013), but the authors do not provide inference results. Fan et al. (2017) considers robust estimation in the case where β is a high-dimensional parameter. Its estimator penalizes the Huber loss function by a term proportional to the ℓ_1-norm of β.

Notations. We use the following notations. For a matrix M, $M^\top$ is its transpose, $||M||_2$, $||M||_1$ and $||M||_\infty$ are, respectively, the ℓ_2-norm, ℓ_1-norm and the sup-norm of the vectorization of M. $||M||_{\text{op}}$ is the operator norm of M and $||M||_0$ is the number of non-zero coefficients in M, that is its ℓ_0-norm. For a probabilistic event $\mathcal{E}$, the fact that it happens w.p.a. 1 (with probability approaching 1) signifies that $\mathbb{P}(\mathcal{E}) \xrightarrow[n\to\infty]{} 1$. Then, for $k = 1, \dots, K$, X_k is the vector $((x_1)_k, \dots, (x_n)_k)^\top$ and X is the matrix $(x_1, \dots, x_n)^\top$. P_X is the projector on the vector space spanned by the columns of the matrix X and $M_X = I_n - P_X$, where I_n is the identity matrix of size n. We denote by y and ϵ, the vectors $(y_1, \dots, y_n)^\top$ and $(\epsilon_1, \dots, \epsilon_n)^\top$, respectively. For a real number $x \in \mathbb{R}$, $\text{sign}(x)$ is equal to 1 if $x \geq 0$ and -1 otherwise.

2 Linear regression with outliers

2.1 Framework

The probabilistic framework consists of a sequence of data generating processes (henceforth, DGPs) that depend on the sample size n. The joint distribution of (x_i, ϵ_i) is independent from the sample size. We consider an asymptotic where n goes to ∞ and where p, the contamination level, depends on n while the number of regressors remains fixed.

The proposed estimation strategy is able to handle models where α is sparse, that is $||\alpha||_0 / n = o_P(1)$ or, in other words, $p \to 0$. Potentially, every individual's y_i can be generated by a distribution that does not follow a linear model but the difference between the distribution of y_i and the one yielded by a linear model can only be important for a negligible proportion of individuals. The subsequent theorems will help to quantify these previous statements.

2.2 Estimation procedure

We consider an estimation procedure that estimates both the coefficients α_i and the effects of the regressors β by a square-root lasso that penalizes only the coefficients α_i, that is

$$(\widehat{\beta}, \widehat{\alpha}) \in \underset{\beta \in \mathbb{R}^K,\ \alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{\sqrt{n}}||y - X\beta - \alpha||_2 + \frac{\lambda}{n}||\alpha||_1,$$

where λ is a penalty level which choice is discussed later. The advantage of the square-root lasso over the lasso estimator is that the penalty level does not depend on an estimate of the variance of ϵ_i. Hence, the proposed procedure is simple in that it does not make use of any tuning parameter unlike the least trimmed squares estimator. An important remark is that if β is such that $X\beta = P_X(y - \widehat{\alpha})$, then

$$\frac{1}{\sqrt{n}}||y - X\beta - \widehat{\alpha}||_2 + \frac{\lambda}{n}||\widehat{\alpha}||_1 \leq \frac{1}{\sqrt{n}}||y - Xb - \widehat{\alpha}||_2 + \frac{\lambda}{n}||\widehat{\alpha}||_1,$$

for any $b \in \mathbb{R}^K$. Therefore, if $X^\top X$ is positive definite, $\widehat{\beta}$ is the OLS estimator of the regression of $y - \widehat{\alpha}$ on X, that is

$$\widehat{\beta} = \left(X^\top X\right)^{-1} X^\top (y - \widehat{\alpha}). \tag{2.1}$$

Then, notice also that for all $\alpha \in \mathbb{R}^n$ and $b \in \mathbb{R}^K$, we have

$$\frac{1}{\sqrt{n}}||M_X(y - \alpha)||_2 + \frac{\lambda}{n}||\alpha||_1 \leq \frac{1}{\sqrt{n}}||y - Xb - \alpha||_2 + \frac{\lambda}{n}||\alpha||_1.$$

Hence, because the value of $\frac{1}{\sqrt{n}}||y - Xb - \alpha||_2 + \frac{\lambda}{n}||\alpha||_1$ is $\frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2 + \frac{\lambda}{n}||\alpha||_1$ whenever $Xb = P_X(y - \alpha)$, it holds that

$$\widehat{\alpha} \in \underset{\alpha \in \mathbb{R}^N}{\arg\min} \frac{1}{\sqrt{n}}||M_X(y - \alpha)||_2 + \frac{\lambda}{n}||\alpha||_1. \tag{2.2}$$

Under assumptions developed below, this procedure yields consistent estimation and asymptotic normality for $\widehat{\beta}$. Remark that model (1) can be seen as a standard linear model with α_i corresponding to the parameter of a dummy variable which value is 1 for the individual i and 0 otherwise. Hence, the estimator can be viewed as the square-root lasso estimator of Belloni et al. (2011). However, our approach is met with additional technical difficulties because we penalize only a subset of the variables and there is no hope to estimate α consistently as each of its entries is indirectly observed only once. As a result, we develop new assumptions and theorems that are better suited for the purposes of this paper.

2.3 Assumptions and results

The first assumption that we make formalizes the hypothesis made on the model that where outlined of the introduction.

Assumption 2.1 *The following holds :*

(i) $\{(x_i, \epsilon_i)\}_i$ *are i.i.d. random variables;*

(ii) $\mathbb{E}[x_i\epsilon_i] = \mathbb{E}[\epsilon_i] = 0$*;*

(iii) $\Sigma = \mathbb{E}[x_i x_i^\top]$ *exists and is positive definite;*

(iv) There exists $\sigma > 0$ *such that* $0 < var[\epsilon_i^2|x_i] = \sigma^2 < \infty$.

This assumption is standard in linear regression models and guarantees that the law of large numbers and the central limit theorem can be applied to certain quantity of interests. The main assumption concerns the choice of the penalty level:

Assumption 2.2 *We have* $\lim\limits_{n\to\infty} \mathbb{P}\left(\lambda \geq 2\sqrt{n}\frac{||M_X\epsilon||_\infty}{||M_X\epsilon||_2}\right) = 1$.

The tuning of λ prescribed by this assumption depends on the distributional assumptions made on ϵ, in particular on the tails. The next lemma provides guidance on how to choose the regularization parameter according to assumptions on ϵ.

Lemma 2.1 *Under Assumption 2.1, it holds that* $2\sqrt{n}\frac{||M_X\epsilon||_\infty}{||M_X\epsilon||_2} = 2\frac{||\epsilon||_\infty}{\sigma} + o_P(||\epsilon||_\infty) + O_P(1)$. *Additionally, if* ψ *is such that* $\lim\limits_{n\to\infty} \mathbb{P}\left(\psi \geq 2\frac{||\epsilon||_\infty}{\sigma}\right) = 1$ *and* $\varphi \to \infty$*, then for any* $c > 1$*,* $\lambda = c\psi + \varphi$ *satisfies Assumption 2.2.*

The proof is given in Appendix. This lemma suppresses the role of the matrix X in the choice of the penalty and simplifies the decision procedure. It leads to the subsequent corollary.

Corollary 2.1 *Under Assumption 2.1, the following holds:*

(i) If ϵ_i *are Gaussian random variables, then* $\lambda = 2c\sqrt{2\log(n)}$ *satisfies Assumption 2.2 for any* $c > 1$*;*

(ii) If ϵ_i *are sub-Gaussian random variables, then there exists a constant* $c > 0$ *such that* $\lambda = c\sqrt{\log(n)}$ *satisfies Assumption 2.2;*

(iii) If ϵ_i are sub-exponential random variables, then then there exists a constant $c > 0$ such that $\lambda = c\log(n)$ satisfies Assumption 2.2.

The proof is given in Appendix. The statistician can use Corollary 2.1 to decide on the penalization parameter given how heavy she expects the tails of the error term to be in her data. In practice, it is advised to choose the smallest penalty verifying Assumption 2.2. This can be done by Monte-Carlo simulations if one is willing to specify the distribution of the errors up to its variance. Notice that heavy-tailed distributions such as sub-exponential random variables are allowed as one can always take $\lambda = \log(n)^{\frac{3}{2}}$ to satisfy Assumption 2.2.

To derive the convergence rate of the estimator, we first bound the estimation error on α and obtain the following result:

Lemma 2.2 *Under assumptions 2.1 and 2.2 and if $\sqrt{p}\max\left(\lambda, ||X||_\infty\right) = o_P(1)$, it holds that*

$$\frac{1}{n}||\widehat{\alpha} - \alpha||_1 = O_P\left(p\lambda\right).$$

The proof is given in Appendix. The rate of convergence of $||\widehat{\alpha} - \alpha||_1/n$ therefore is lower than $p\sqrt{\log(n)}$ if the errors are gaussian or sub-gaussian and we choose the penalty level as in Lemma 2.1. Note that, as standard in works related to the lasso estimator (see Bühlmann and Van De Geer (2011)), in the proof, a condition that states that a compatibility constant is bounded from below with probability approaching one. The condition that $\sqrt{p}||X||_\infty = o_P(1)$ is enough to show that this property holds as shown in Lemma 3.4 in Appendix. It is possible to find other sufficient conditions but it is outside the scope of this paper. Remark that if $\{x_i\}_i$ are i.i.d. sub-Gaussian random vectors then $||X||_\infty = O_P\left(\sqrt{\log(n)}\right)$ allowing for the sparsity level $p = o_P(1/\log(n))$.

Here, we show how to derive the rate of convergence of $\widehat{\beta}$ from Lemma 2.2. Assume that the assumptions of Lemma 2.2 hold. Substituting y by $X\beta + \alpha + \epsilon$ in (2), we obtain

$$\widehat{\beta} - \beta = \left(X^\top X\right)^{-1} X^\top \epsilon + (X^\top X)^{-1} X^\top(\alpha - \widehat{\alpha}). \tag{2.3}$$

Now, notice that $\left(X^\top X\right)^{-1} X^\top(\alpha - \widehat{\alpha}) = (X^\top X/n)^{-1} X^\top(\alpha - \widehat{\alpha})/n$. Because of Assumption 2.1, we can apply the law of large numbers and obtain $\left(X^\top X/n\right)^{-1} = O_P(1)$, which implies that

$$\begin{aligned}\left|\left|\left(X^\top X\right)^{-1} X^\top(\alpha - \widehat{\alpha})\right|\right|_2 &\leq \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{1}{n}\left|\left|X^\top(\alpha - \widehat{\alpha})\right|\right|_2 \\ &= O_P\left(\frac{1}{n}||X||_\infty ||\alpha - \widehat{\alpha}||_1\right) \quad \text{(by Hölder's inequality).} \end{aligned} \tag{2.4}$$

By Lemma 2.2, this implies that

$$||(X^\top X)^{-1}X^\top(\alpha - \widehat{\alpha})||_2 = O_P\left(p\lambda||X||_\infty\right).$$

Finally, we have that $\sqrt{n}(X^\top X)^{-1}X^\top\epsilon \xrightarrow{d} \mathcal{N}(0, \sigma\Sigma^{-1})$. This leads to Theorem 2.2.

Theorem 2.2 *Under assumptions 2.1 and 2.2 and if $\sqrt{p}\max\left(\lambda, ||X||_\infty\right) = o_P(1)$, it holds that*

$$\frac{\widehat{\beta} - \beta}{\max\left(\frac{1}{\sqrt{n}}, p\lambda||X||_\infty\right)} = O_P\left(1\right).$$

This result allows to derive the rates of convergence under different assumptions on the tails of the distributions of the regressors and the error term. For instance, if $\{x_i\}_i$ and $\{\epsilon_i\}_i$ are i.i.d. sub-Gaussian random variables, then $\widehat{\beta}$ is consistent as long as $p\log(n) \to 0$ for the choice of λ proposed in Lemma 2.1. In this case, this implies that the estimator reaches (up to a logarithmic factor) the minimax lower bound for the Huber's contamination location model under gaussian errors, which is $\max(1/n, p^2)$ in ℓ_2-norm according to Chen et al. (2018). We attain the rate $\max(1/n, p^2\log(n))$. Remark also that equation (5) explains the role of $||X||_\infty$ in the convergence rate of $\widehat{\beta}$. For an individual i, if x_i is large then an error in the estimation of α_i can contribute to an error in the estimation of β via the term $(X^\top X)^{-1}X^\top(\alpha - \widehat{\alpha})$ in (4). $||X||_\infty$ measures the maximum influence that an observation can have.

To show that the estimator is asymptotically normal, it suffices to assume that the term $(X^\top X)^{-1}X^\top(\alpha - \widehat{\alpha})$ in (4) vanishes asymptotically:

Theorem 2.3 *Under assumptions 2.1 and 2.2, assuming that $\sqrt{p}\max\left(\lambda, ||X||_\infty\right) = o_P(1)$, $p\lambda\sqrt{n} = o(1)$ and $p\lambda||X||_\infty\sqrt{n} = o_P(1)$, we have*

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2\Sigma^{-1}).$$

Moreover, $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n(y_i - x_i^\top\hat{\beta} - \widehat{\alpha})^2$ and $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ are consistent estimators of, respectively, σ^2 and Σ.

The proof that $\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$ is given in Appendix. When the entries of X and ϵ are sub-Gaussian, for the choice of the penalty prescribed in Lemma 2.1, the contamination level needs to satisfy $p\log(n)\sqrt{n} \to 0$ to be able to use 2.3 to prove asymptotic normality. Notice that the asymptotic variance of our estimator corresponds to the one of the OLS estimator in the standard linear model under homoscedasticity. Hence, confidence sets and tests can be built in the same manner as in the theory of the OLS estimator.

An important last remark concerns the meaning of confidence intervals developed using Theorem 2.3. Note that they are obtained under an asymptotic with triangular array data under which the number of outliers is allowed to go to infinity while the proportion of outliers goes to 0. The interpretation of a 95% confidence interval I built with Theorem 2.3 is as follows: if the number of outliers in our data is low enough and the sample size is large enough, then there is a probability of approximatively 0.95 that β belongs to I.

3 Computation and simulations

3.1 Iterative algorithm

We propose to use an algorithm already introduced in Section 5 of Owen (2007) to compute our estimator. Because $u = \min_{\sigma>0}\left\{\frac{\sigma}{2} + \frac{1}{2\sigma}u^2\right\}$, as long as $\left|\left|y - X\widehat{\beta} - \widehat{\alpha}\right|\right|_2^2 > 0$, we have that

$$(\widehat{\beta}, \widehat{\alpha}, \widehat{\sigma}) \in \underset{\beta\in\mathbb{R}^K,\alpha\in\mathbb{R}^n,\sigma\in\mathbb{R}^+}{\arg\min} \frac{\sigma}{2} + \frac{1}{2\sigma}\left|\left|y - X\beta - \alpha\right|\right|_2^2 + \frac{\lambda}{\sqrt{n}}\left|\left|\alpha\right|\right|_1. \tag{3.1}$$

This is a convex objective and the proposed approach is to iteratively minimize over β, α and σ. Let us start from $\left(\beta^{(0)}, \alpha^{(0)}, \sigma^{(0)}\right)$ and compute the following sequence for $t \in \mathbb{N}^*$ until convergence:

1. $\beta^{(t+1)} \in \underset{\beta\in\mathbb{R}^K}{\arg\min} \left|\left|y - X\beta - \alpha^{(t)}\right|\right|_2^2$;

2. $\alpha^{(t+1)} \in \underset{\alpha\in\mathbb{R}^n}{\arg\min} \left|\left|y - X\beta^{(t+1)} - \alpha\right|\right|_2^2 + \frac{2\lambda\sigma^{(t)}}{\sqrt{n}}\left|\left|\alpha\right|\right|_1$;

3. $\sigma^{(t+1)} = \left|\left|y - X\beta^{(t+1)} - \alpha^{(t+1)}\right|\right|_2$.

The following lemma is a direct consequence of Section 4.2.2. in Giraud (2014) and explains how to perform step 2:

Lemma 3.1 *For $i = 1, \dots, n$, if $\left|y_i - (X\beta^{(t+1)})_i\right| \le \frac{\lambda\sigma^{(t)}}{\sqrt{n}}$ then $\alpha_i^{(t+1)} = 0$. If $\left|y_i - (X\beta^{(t+1)})_i\right| > \frac{\lambda\sigma^{(t)}}{\sqrt{n}}$ then $\alpha_i^{(t+1)} = y_i - \left(X\beta^{(t+1)}\right)_i - sign\left(y_i - \left(X\beta^{(t+1)}\right)_i\right)\frac{\lambda\sigma^{(t)}}{\sqrt{n}}$.*

3.2 Simulations

We apply this algorithm in a small simulation exercise. The data generating process is as follows: there are two regressors x_{1i} and x_{2i}, with $x_{1i} = 1$ for all i and x_{2i} are i.i.d. $\mathcal{N}(0, 1)$ random variables. ϵ_i are i.i.d. $\mathcal{N}(0, 1)$ random variables. The parameter is $\beta = (1, 1)^\top$. Then, we set

$$\alpha_i = \begin{cases} 0 & \text{if } x_{2i} < q_{1-p} \\ \\ 5x_{2i} & \text{if } x_{2i} \geq q_{1-p}, \end{cases}$$

where q_{1-p} is such that $\mathbb{P}(x_{2i} \geq q_{1-p}) = p$. In tables 1, 2 and 3 we present the bias, the variance, the mean squared error (MSE), of $\widehat{\beta}$ for various values of p and n. We use 100 iterations and $\lambda = 2.01\sqrt{2\log(n)}$. This choice corresponds to the one outlined in Corollary 2.1. The bias, the variance and for the naive OLS estimator:

$$\widetilde{\beta}^{OLS} \in \arg\min_{\beta\in\mathbb{R}^K} ||y - X\beta||_2^2$$

are also reported. The coverage of 95% confidence intervals based on the asymptotic variance of Theorem 2.3 are presented in Table 4. The presented results are averages over 8000 replications. We observe that our estimator brings a substantial improvement in estimation precision with respect to the OLS estimator.

Table 1: $p = 0.025$, $n = 100$

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	0.127	0.301	0.278	0.671
var	0.060	0.130	0.097	0.221
MSE	0.076	0.221	0.174	0.671

Table 2: $p = 0.01$, $n = 1000$

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	0.044	0.133	0.120	0.361
var	0.002	0.003	0.004	0.007
MSE	0.004	0.021	0.018	0.152

Table 3: $p = 0.001$, $n = 10000$

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	0.005	0.015	0.017	0.057
var	1.08×10^{-4}	1.28×10^{-4}	2.21×10^{-4}	5.23×10^{-4}
MSE	1.33×10^{-4}	3.53×10^{-4}	5.10×10^{-4}	3.772×10^{-3}

Table 4: Coverage of 95% confidence intervals based on Theorem 2.3

p	n	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
0.025	100	0.82	0.47	0.75	0.20
0.01	1000	0.74	0.16	0.24	0.00
0.001	10000	0.93	0.66	0.68	0.03

Appendix

Proof of Lemma 2.1

We start by proving the next two lemmas:

Lemma 3.2 *Under Assumption 2.1, it holds that* $||P_X\epsilon||_\infty = O_P(1)$.

Proof. We have that $\sqrt{n}(X^\top X)^{-1}X^\top\epsilon \xrightarrow{d} \mathcal{N}(0, \sigma\Sigma^{-1})$, therefore $\sqrt{n}||(X^\top X)^{-1}X^\top\epsilon||_2 = O_P(1)$. Because $X(X^\top X)^{-1}X^\top\epsilon = \frac{X}{\sqrt{n}}\sqrt{n}(X^\top X)^{-1}X^\top\epsilon$, we obtain that

$$||P_X\epsilon||_2 \leq \frac{||X||_2}{\sqrt{n}}\sqrt{n}||(X^\top X)^{-1}X^\top\epsilon||_2 = O_P\left(\frac{||X||_2}{\sqrt{n}}\right) = O_P(1),$$

by the law of large numbers. □

Lemma 3.3 *Under Assumption 2.1, it holds that* $\frac{\sqrt{n}}{||M_X\epsilon||_2} - \frac{1}{\sigma} = o_P(1)$.

Proof. First, remark that, by the Pythagorean theorem,

$$\begin{aligned}||M_X\epsilon||_2^2 &= \left\langle \epsilon - X(X^\top X)^{-1}X^\top\epsilon, \epsilon - X(X^\top X)^{-1}X^\top\epsilon \right\rangle \\ &= ||\epsilon||_2^2 - \epsilon^\top X(X^\top X)^{-1}X^\top\epsilon.\end{aligned}$$

Now, this leads to $\frac{1}{n}||M_X\epsilon||_2^2 = \frac{1}{n}||\epsilon||_2^2 - \frac{1}{n}\epsilon^\top X(X^\top X)^{-1}X^\top\epsilon$. By the law of large numbers and the central limit theorem, we have that $\sqrt{n}(X^\top X)^{-1}X^\top\epsilon = O_P(1)$ and $\frac{1}{\sqrt{n}}X^\top\epsilon = O_P(1)$. This implies that $\epsilon^\top X(X^\top X)^{-1}X^\top\epsilon = O_P(1)$. We also have that $\frac{1}{n}||\epsilon||_2^2 \xrightarrow{\mathbb{P}} \sigma^2$, which leads to $\frac{1}{n}||M_X\epsilon||_2^2 \xrightarrow{\mathbb{P}} \sigma^2$. We conclude by the continuous mapping theorem. □

Now, we proceed with the proof of Lemma 2.1. Notice that

$$\begin{aligned}2\sqrt{n}\frac{||M_X\epsilon||_\infty}{||M_X\epsilon||_2} &\leq \frac{2\sqrt{n}}{||M_X\epsilon||_2}(||\epsilon||_\infty + ||P_X\epsilon||_\infty) \\ &\leq \frac{2}{\sigma}||\epsilon||_\infty + 2\left|\frac{\sqrt{n}}{||M_X\epsilon||_2} - \frac{1}{\sigma}\right| ||\epsilon||_\infty + \frac{2}{\sigma}||P_X\epsilon||_\infty + 2\left|\frac{\sqrt{n}}{||M_X\epsilon||_2} - \frac{1}{\sigma}\right| ||P_X\epsilon||_\infty.\end{aligned}$$

Using lemmas 3.2 and 3.3, we obtain

$$2\sqrt{n}\frac{||M_X\epsilon||_\infty}{||M_X\epsilon||_2} = 2\frac{||\epsilon||_\infty}{\sigma} + o_P(||\epsilon||_\infty) + O_P(1). \tag{3.2}$$

The rest of proof of the lemma is a direct consequence of (7) and the pigeonhole principle.

Proof of Corollary 2.1

Proof of (i) By Lemma 2.1 it is sufficient to show that for $c > 1$,

$$\lim_{n\to\infty} \mathbb{P}\left(2c\sqrt{2\log(n)} \geq 2\frac{||\epsilon||_\infty}{\sigma}\right) = 1.$$

Let us remember the Gaussian bound (see Lemma B.1 in Giraud (2014)): for $t \geq 0$, we have $\mathbb{P}\left(\frac{|\epsilon_i|}{\sigma} \geq t\right) \leq e^{-\frac{t^2}{2}}$. Then, we have

$$\begin{aligned}\mathbb{P}\left(2c\sqrt{2\log(n)} \leq 2\frac{||\epsilon||_\infty}{\sigma}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(c\sqrt{2\log(n)} \leq \frac{|\epsilon_i|}{\sigma}\right)\\ &\leq ne^{-c\log(n)} = e^{-(c-1)\log(n)} \to 0.\end{aligned}$$

Proof of (ii) By Lemma 2.1 it is sufficient to show that there exists $c > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(c\sqrt{\log(n)} \geq 2\frac{||\epsilon||_\infty}{\sigma}\right) = 1.$$

Recall the sub-Gaussian bound (see Proposition 2.5.2 in Vershynin (2018)): for $t \geq 0$, there exists $b > 0$ such that $\mathbb{P}\left(\frac{|\epsilon_i|}{\sigma} \geq t\right) \leq 2e^{-\frac{t^2}{2b}}$. Then, for $\rho > 1$, we have

$$\begin{aligned}\mathbb{P}\left(2\sqrt{2}\rho\sqrt{b}\sqrt{\log(n)} \leq 2\frac{||\epsilon||_\infty}{\sigma}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(\sqrt{2}\rho\sqrt{b}\sqrt{\log(n)} \leq \frac{|\epsilon_i|}{\sigma}\right)\\ &\leq 2ne^{-\rho^2\log(n)} = 2e^{-(\rho^2-1)\log(n)} \to 0.\end{aligned}$$

Proof of (iii) Using Lemma 2.1, we only need to show that there exists $c > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(c\log(n) \geq 2\frac{||\epsilon||_\infty}{\sigma}\right) = 1.$$

Let us state the sub-exponential bound (see Proposition 2.7.1 in Vershynin (2018)): for $t \geq 0$, there exists $b > 0$ such that $\mathbb{P}\left(\frac{|\epsilon_i|}{\sigma} \geq t\right) \leq 2e^{-\frac{t}{2b}}$. Then, let $\rho > 1$, we have, for n large enough,

$$\begin{aligned}\mathbb{P}\left(4\rho b\log(n) \leq 2\frac{||\epsilon||_\infty}{\sigma}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(2\rho b\log(n) \leq \frac{|\epsilon_i|}{\sigma}\right)\\ &\leq 2ne^{-\rho\log(n)} = 2e^{-(\rho-1)\log(n)} \to 0.\end{aligned}$$

Proof of Lemma 2.2

Compatibility constant. For $\delta \in \mathbb{R}^n$, we denote by $\delta_J \in \mathbb{R}^n$ the vector for which $(\delta_J)_i = \delta_i$ if $\alpha_i \neq 0$ and $(\delta_J)_i = 0$ otherwise. Let us also define $\delta_{J^c} = \delta - \delta_J$. We introduce the following cone:

$$C = \left\{\delta \in \mathbb{R}^n \ s.t. \ ||\delta_{J^c}||_1 \leq 3\, ||\delta_J||_1\right\}.$$

We work with the following compatibility constant defined as

$$\kappa = \min_{\delta \in C, \delta \neq 0} \frac{\sqrt{||\alpha||_0} ||M_X \delta||_2}{||\delta_J||_1}.$$

We use the following lemma:

Lemma 3.4 *Under Assumption 2.1, if* $\sqrt{p}||X||_\infty = o_P(1)$*, there exists* $\kappa_* > 0$ *such that* $\kappa > \kappa_*$ *w.p.a.* 1*.*

Proof. Take $\delta \in C$, to show this result, notice that

$$M_X \delta = \delta - X(X^\top X)^{-1} X^\top \delta.$$

Therefore, we have

$$\begin{aligned}
||M_X \delta||_2 &\geq ||\delta||_2 - ||X(X^\top X)^{-1} X^\top \delta||_2 \\
&= ||\delta||_2 - \left|\left|\sum_{k=1}^K X_k \left((X^\top X)^{-1} X^\top \delta\right)_k\right|\right|_2 \\
&\geq ||\delta||_2 - \sum_{k=1}^K \left|\left|X_k \left((X^\top X)^{-1} X^\top \delta\right)_k\right|\right|_2 \\
&\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|(X^\top X)^{-1} X^\top \delta\right|\right|_\infty \\
&\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|(X^\top X)^{-1} X^\top \delta\right|\right|_2 \\
&\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{1}{n} ||X^\top \delta||_2.
\end{aligned}$$

Next, by Hölder's inequality, we obtain

$$\begin{aligned}
||M_X\delta||_2 &\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{K}}{n}||X||_\infty ||\delta||_1 \\
&\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{K}}{n}||X||_\infty 4||\delta_J||_1 \quad \text{(because } \delta \in C\text{)} \\
&\geq ||\delta||_2 - \sum_{k=1}^K ||X_k||_2 \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{K}}{n}||X||_\infty 4\sqrt{||\alpha||_0}||\delta_J||_2 \quad \text{(because } ||\delta_J||_0 \leq ||\alpha||_0\text{)} \\
&\geq ||\delta||_2 - \sum_{k=1}^K \frac{||X_k||_2}{\sqrt{n}} \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} 4\sqrt{K}\sqrt{\frac{||\alpha||_0}{n}}||X||_\infty ||\delta||_2. \qquad (3.3)
\end{aligned}$$

Next, we have that

$$\begin{aligned}
\kappa &\geq \min_{\delta\in C,\delta\neq 0} \frac{\sqrt{||\alpha||_0}||M_X\delta||_2}{||\delta_J||_1} \\
&\geq \min_{\delta\in C,\delta\neq 0} \frac{\sqrt{||\alpha||_0}||M_X\delta||_2}{\sqrt{||\alpha||_0}\,||\delta_J||_2} \\
&\geq \min_{\delta\in C,\delta\neq 0} \frac{||M_X\delta||_2}{||\delta||_2} \\
&\geq \left(1 - \sum_{k=1}^K \frac{||X_k||_2}{\sqrt{n}} \left|\left|\left(\frac{1}{n}X^\top X\right)^{-1}\right|\right|_{\text{op}} 4\sqrt{K}\sqrt{\frac{||\alpha||_0}{n}}||X||_\infty\right).
\end{aligned}$$

Now, by Assumption 2.1, we have $\left|\left|\left(X^\top X/n\right)^{-1}\right|\right|_{\text{op}} = O_P(1)$ and that $\sum_{k=1}^K ||X_k||_2/\sqrt{n} = \sum_{k=1}^K \sqrt{(X^\top X/n)_{kk}} = O_P(1)$, both implying that $\frac{1}{\sqrt{n}}\sum_{k=1}^K ||X_k||_2 \left|\left|\left(X^\top X/n\right)^{-1}\right|\right|_{\text{op}} = O_P(1)$. We conclude the proof using that $\sqrt{p}||X||_\infty = o_P(1)$. □

End of the proof of Lemma 2.2

Throughout this proof, we work on the event

$$\left\{\lambda \geq \frac{2\sqrt{n}||M_X\epsilon||_\infty}{||M_X\epsilon||_2}\right\} \cap \{\kappa > \kappa_*\} \cap \left\{\left(\frac{2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{\kappa}\right) < 1\right\},$$

which has probability approaching 1 according to Assumption 2.2, Lemma 3.4 and the condition that $\sqrt{p}\lambda \to 0$. Let us define $\Delta = \widehat{\alpha} - \alpha$. Now, remark that

$$\begin{aligned} ||\widehat{\alpha}||_1 &= ||\alpha + \Delta||_1 \\ &= ||\alpha + \Delta_J + \Delta_{J^c}||_1 \\ &\geq ||\alpha + \Delta_{J^c}||_1 - ||\Delta_J||_1 . \end{aligned} \tag{3.4}$$

Next, we use the fact that $||\alpha + \Delta_{J^c}||_1 = ||\alpha||_1 + ||\Delta_{J^c}||_1$. Combining this and (9), we get

$$||\widehat{\alpha}||_1 \geq ||\alpha||_1 + ||\Delta_{J^c}||_1 - ||\Delta_J||_1 . \tag{3.5}$$

Using (3), we have

$$\frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 + \frac{\lambda}{n}||\widehat{\alpha}||_1 \leq \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2 + \frac{\lambda}{n}||\alpha||_1. \tag{3.6}$$

By convexity, if $M_X\epsilon \neq 0$, it holds that

$$\begin{aligned} \frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 - \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2 &\geq -\frac{1}{\sqrt{n}||M_X\epsilon||_2}\left\langle M_X(\epsilon), \Delta \right\rangle \\ &\geq -\frac{\lambda}{2n}||\Delta||_1, \end{aligned} \tag{3.7}$$

where (12) comes from $\lambda \geq 2\sqrt{n}||M_X\epsilon||_2/||M_X\epsilon||_\infty$. This last inequality is also straightforwardly true when $M_X\epsilon = 0$. This and (11) imply

$$||\hat{\alpha}||_1 \leq \frac{1}{2}||\Delta||_1 + ||\alpha||_1. \tag{3.8}$$

Using (10), we get

$$||\alpha||_1 + ||\Delta_{J^c}||_1 - ||\Delta_J||_1 \leq \frac{1}{2}||\Delta||_1 + ||\alpha||_1.$$

Then, because $||\Delta||_1 = ||\Delta_{J^c}||_1 + ||\Delta_J||_1$, we obtain

$$||\Delta_{J^c}||_1 \leq 3\,||\Delta_J||_1 , \tag{3.9}$$

which implies that $\Delta \in C$. Using $y = X\beta + \alpha + \epsilon$, we get

$$\frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 = \frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 - \frac{2}{n}\left\langle M_X\epsilon, \widehat{\alpha}-\alpha \right\rangle .$$

By Hölder's inequality, this results in

$$\frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 \leq \frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 - \frac{2}{n}||M_X\epsilon||_\infty||\Delta||_1.$$

Because $\lambda \geq 2\sqrt{n}\frac{||M_X\epsilon||_\infty}{||M_X\epsilon||_2}$, we obtain

$$\frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 \leq \frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 + \frac{\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}||\Delta||_1.$$

This implies that

$$\begin{aligned}
&\frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 \\
&\leq \frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 + \frac{\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}||\Delta||_1 \\
&= \frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 + \frac{\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}\left(||\Delta_J||_1 + ||\Delta_{J^c}||_1\right) \\
&\leq \frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 + \frac{4\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}\,||\Delta_J||_1 \quad \text{(because } \Delta \in C\text{)}.
\end{aligned} \tag{3.10}$$

By equations (10) and (11), we have $\frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 - \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2 \leq \frac{\lambda}{n}\left(||\Delta_J||_1 - ||\Delta_{J^c}||_1\right)$. Using the fact that $\Delta \in C$ and (12), this yields

$$\left|\frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 - \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2\right| \leq \frac{2\lambda}{n}\,||\Delta_J||_1\,.$$

Next, notice that

$$\begin{aligned}
&\frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 \\
&= \left(\frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 - \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2\right)\left(\frac{1}{\sqrt{n}}||M_X(y-\widehat{\alpha})||_2 + \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2\right).
\end{aligned}$$

This implies

$$\begin{aligned}
&\left|\frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2\right| \\
&\leq \frac{2\lambda}{n}\,||\Delta_J||_1\left(\frac{2}{\sqrt{n}}||M_X(y-\alpha)||_2 + \frac{2\lambda}{n}\,||\Delta_J||_1\right) \\
&\leq \left(\frac{2\lambda}{n}\right)^2||\Delta_J||_1^2 + \frac{4}{\sqrt{n}}||M_X(y-\alpha)||_2\frac{\lambda}{n}\,||\Delta_J||_1\,.
\end{aligned} \tag{3.11}$$

Combining (15) and (16) and noting that $||M_X\epsilon||_2 = ||M_X(y-\alpha)||_2$, we obtain

$$\frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 \leq \left(\frac{2\lambda}{n}\right)^2||\Delta_J||_1^2 + \frac{4||M_X\epsilon||_2}{\sqrt{n}}\frac{\lambda}{n}\,||\Delta_J||_1 + \frac{4\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}\,||\Delta_J||_1\,.$$

Now, because $\Delta \in C$, this implies that

$$\frac{1}{n}||M_X\Delta||_2^2 \le \left(\frac{2\lambda}{n}\right)^2 \left(\frac{\sqrt{||\alpha||_0}||M_X\Delta||_2}{\kappa}\right)^2 + \frac{8\lambda||M_X\epsilon||_2}{n^{\frac{3}{2}}}\frac{\sqrt{||\alpha||_0}||M_X\Delta||_2}{\kappa}.$$

From now on assume that $||M_X\Delta||_2 \neq 0$, we get

$$\frac{1}{n}||M_X\Delta||_2 \le \left(1-\left(\frac{2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{\kappa}\right)^2\right)^{-1}\frac{8||M_X\epsilon||_2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{n\kappa},$$

which implies again that

$$\frac{1}{n}\left||\Delta_J\right||_1 \le \left(1-\left(\frac{2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{\kappa}\right)^2\right)^{-1}\frac{8||M_X\epsilon||_2\frac{||\alpha||_0}{n}\lambda}{\sqrt{n}\kappa^2}.$$

Finally, as $\Delta \in C$, we have

$$\begin{aligned}\frac{1}{n}||\Delta||_1 &= \frac{1}{n}\left(||\Delta_J||_1 + ||\Delta_{J^c}||_1\right)\\ &\le \frac{4}{n}\left||\Delta_J\right||_1\\ &\le \left(1-\left(\frac{2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{\kappa_*}\right)^2\right)^{-1}\frac{32||M_X\epsilon||_2\frac{||\alpha||_0}{n}\lambda}{\sqrt{n}\kappa_*}. \qquad (3.12)\end{aligned}$$

The last inequality also holds if $M_X\Delta = 0$ because, as $\kappa > 0$ because we are on the event $\kappa > \kappa^*$, this implies that $\Delta_J = 0$ and hence $\Delta = 0$ using the fact that Δ belongs to C. To conclude the proof, use (17), the fact that $||M_X\epsilon||_2/\sqrt{n} \le ||\epsilon||_2/\sqrt{n} = O_P(1)$ by the law of large numbers and the condition $\sqrt{p}\max\left(\lambda, ||X||_\infty\right) = o_P(1)$.

Proof that $\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$ in Theorem 2.3

We have

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\left\|y - X\widehat{\beta} - \widehat{\alpha}\right\|_2^2\\ &= \frac{1}{n}\left\|X\left(\beta-\widehat{\beta}\right) + (\alpha - \widehat{\alpha}) + \epsilon\right\|_2^2\\ &= \frac{1}{n}\left\|X\left(\beta-\widehat{\beta}\right)\right\|_2^2 + \frac{1}{n}\left||\alpha-\widehat{\alpha}\right||_2^2 + \frac{2}{n}\left\langle X\left(\beta-\widehat{\beta}\right), \alpha-\widehat{\alpha}\right\rangle\\ &\quad + \frac{2}{n}\left\langle X\left(\beta-\widehat{\beta}\right), \epsilon\right\rangle + \frac{2}{n}\left\langle \alpha-\widehat{\alpha}, \epsilon\right\rangle + \frac{1}{n}\left||\epsilon\right||_2^2.\end{aligned}$$

Then, because of Lemma 2.2, Theorem 2.2, $p\lambda\sqrt{n} = o(1)$ and $p\lambda \left|X\right|_\infty = o_P(1)$, it holds that

$$\begin{aligned}
&\left|\left|\widehat{\alpha} - \alpha\right|\right|_1 = o_P\left(\sqrt{n}\right);\\
&\left|\left|\widehat{\beta} - \beta\right|\right|_2 = o_P\left(1\right).
\end{aligned}$$

Next, we have

$$\frac{1}{n}\left|\left|X\left(\beta - \widehat{\beta}\right)\right|\right|_2^2 \leq \frac{1}{n}\left|\left|X\right|\right|_2^2 \left|\left|\widehat{\beta} - \beta\right|\right|_2^2 = o_P(1),$$

by the law of large numbers. Then, by Hölder's inequality, we obtain that

$$\frac{1}{n}\left|\left|\alpha - \widehat{\alpha}\right|\right|_2^2 \leq \frac{1}{n}\left|\left|\alpha - \widehat{\alpha}\right|\right|_1^2 = o_P(1).$$

By the Cauchy-Schwarz inequality, this also leads to $\left\langle X\left(\beta - \widehat{\beta}\right), \alpha - \widehat{\alpha}\right\rangle/n = o_P(1)$. Then, by Assumption 2.1, the law of large numbers implies that $\left|\left|\epsilon\right|\right|_2/\sqrt{n} = O_P(1)$. Thus, by the Cauchy-Schwartz inequality, we have

$$\frac{1}{n}\left\langle X\left(\beta - \widehat{\beta}\right), \epsilon\right\rangle \leq \frac{1}{n}\left|\left|X\left(\beta - \widehat{\beta}\right)\right|\right|_2 \left|\left|\epsilon\right|\right|_2 = o_P(1)$$

and

$$\left\langle \alpha - \widehat{\alpha}, \epsilon\right\rangle \leq \frac{1}{n}\left|\left|\widehat{\alpha} - \alpha\right|\right|_2 \left|\left|\epsilon\right|\right|_2 = o_P(1),$$

which concludes the proof.

References

Andreas Alfons, Christophe Croux, and Sarah Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, pages 226–248, 2013.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Olivier Collier and Arnak S Dalalyan. Rate-optimal estimation of p-dimensional linear functionals in a sparse gaussian model. *arXiv preprint arXiv:1712.05495*, 2017.

Arnak S Dalalyan. SOCP based variance free Dantzig selector with application to robust estimation. *Comptes Rendus Mathematique*, 350(15-16):785–788, 2012.

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

Christophe Giraud. *Introduction to high-dimensional statistics.* Chapman and Hall/CRC, 2014.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Sophie Lambert-Lacroix, Laurent Zwald, et al. Robust regression through the huber?s criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053, 2011.

Wei Li. *Simultaneous variable selection and outlier detection using LASSO with applications to aircraft landing data analysis.* PhD thesis, Rutgers University-Graduate School-New Brunswick, 2012.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R).* Wiley, 2018.

Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443 (7):59–72, 2007.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.