ESAIM: PS 24 (2020) 688–702
https://doi.org/10.1051/ps/2020014

ESAIM: Probability and Statistics
www.esaim-ps.org

INFERENCE ROBUST TO OUTLIERS WITH ℓ_1-NORM PENALIZATION*

Jad Beyhum**

Abstract. This paper considers inference in a linear regression model with outliers in which the number of outliers can grow with sample size while their proportion goes to 0. We propose a square-root lasso ℓ_1-norm penalized estimator. We derive rates of convergence and establish asymptotic normality. Our estimator has the same asymptotic variance as the OLS estimator in the standard linear model. This enables us to build tests and confidence sets in the usual and simple manner. The proposed procedure is also computationally advantageous, it amounts to solving a convex optimization program. Overall, the suggested approach offers a practical robust alternative to the ordinary least squares estimator.

Mathematics Subject Classification. 62F35; 62J05, 62J07.

Received August 9, 2019. Accepted April 2, 2020.

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 random variable y_i and a random vector of covariates x_i with support in $\mathbb{R}^p$, where p is fixed. We assume that the following relationship holds:

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

where $\beta \in \mathbb{R}^p$, the error term ξ_i is a random variable such that $\mathbb{E}[x_i \xi_i | \alpha_i = 0] = 0$ and α_i is a random variable. It also holds that $\{y_i, x_i, \xi_i, \alpha_i\}_i$ 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 the average proportion of outliers $\mathbb{P}(\alpha_i \neq 0)$ be denoted ϵ. The goal is to obtain inference results on the parameter β.

This model can represent various situations of practical interest. First, the statistician might 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.

*I thank my PhD supervisor Professor Eric Gautier for his availability and valuable help. I am also grateful to Anne Ruiz-Gazen, Jean-Pierre Florens, Thierry Magnac, Nour Meddahi, two anonymous referees and an associate editor of *ESAIM: Probability & Statistics* for their useful comments. I acknowledge financial support from the ERC POEMH 337665 grant.

Keywords and phrases: Robust regression, ℓ1-norm penalization, unknown variance.

Toulouse School of Economics, Université Toulouse Capitole, Toulouse, France.

** Corresponding author: jad.beyhum@gmail.com

Article published by EDP Sciences

Second, β may represent the causal effect of the regressors for the population of "standard" individuals. If the aim is to evaluate the effect of a binary treatment, the effect could be negative for most of the population but strongly positive for a small fraction of individuals, the outliers. In this case, the policy maker may not be willing to implement a policy that has a negative effect on most 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 error model where we do not observe $(\tilde{y}_i, \tilde{x}_i)$ but (y_i, x_i). If the observed variables follow the model $\tilde{y}_i = \tilde{x}_i\beta + \tilde{\xi}_i$ with $\mathbb{E}[\tilde{x}_i\tilde{\xi}_i] = 0$, this fits in our framework with $\xi_i = \tilde{\xi}_i$ 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 (outliers in the x-direction) and in y_i (outliers in the y-direction) for a small fraction of the population. See [17] for a thorough discussion.

This paper develops results on the estimation of β when the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ is sparse in the sense that ϵ goes to 0 with n. We rely on a variant of the square-root lasso estimator of [3] which penalizes the ℓ_1-norm of the vector α. The penalty parameter does not depend on the variance of the error term and the estimator is computationally tractable. We show that if the vector α is sparse enough, then the 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 draws upon the literature in two different research fields. The first is that of inference in the high-dimensional linear regression model. Our estimator is analogous to the concomitant lasso of [16]. The computational algorithm presented in Section 3 is similar to what is proposed for the scaled-lasso estimator introduced in [19]. 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.

The second field is that of robust regression. For detailed accounts of this field, see [10, 15, 17]. The literature identifies a trade-off between efficiency and robustness, as explained below. M-estimators (such as the Ordinary Least-Squares (OLS) estimator) are often efficient when data are generated by the standard linear model with Gaussian errors and without outliers. However, this comes at the cost of robustness; M-estimators may be asymptotically biased in the presence of outliers, even when the proportion of outliers goes to 0 as in the asymptotic setting considered in this paper (see the discussion for the OLS estimator in Sect. 2.2). By contrast, S-estimators such as the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS) are robust under 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. Moreover, the non-convexity of their objectives functions generates computational issues (see [17]). The estimator proposed in this paper attains the same asymptotic variance as the OLS estimator in the standard linear model. The computational algorithm in Section 3 relies on a convex program and is computationally tractable.

Within the robust regression literature some authors have considered applying ℓ_1-norm penalization to robust estimation. In particular, the model studied in this paper nests the Huber's contamination model for location estimation (see [11])

$$y_i = \beta + \alpha_i + \xi_i, \tag{1.2}$$

where ξ_i are i.i.d. $\mathcal{N}(0,1)$ random variables, $\beta \in \mathbb{R}$ is the mean of y_i for non-outlying coefficients while $\mathbb{E}[y_i|\alpha_i \neq 0]$ is left unrestricted. [4] show that the minimax lower bound for the squared ℓ_2-norm estimation error is of order greater than $\max(1/n, \epsilon^2)$. When $\epsilon \log(n) \to 0$, 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 closest to the approach studied in this paper is soft-thresholding using a lasso estimator

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

see for instance [5]. We use instead a square-root lasso estimator that has the advantage of providing guidance on the choice of penalty level that is independent of the error variance (see [3]). We extend the analysis of this type of estimator to the linear regression model and develop original inference results. Although other ℓ_1-norm penalized estimators for robust linear regression have been studied in the literature (see [1, 6, 8, 12–14, 18]), the authors do not provide inference results. [7] consider robust estimation in the case where β is a high-dimensional parameter. Their estimator penalizes the Huber loss function by a term proportional to the ℓ_1-norm of β.

Notation. We use the following notation. For a matrix M, $M^\top$ is its transpose, $||M||_2$, $||M||_1$ and $||M||_\infty$ are the ℓ_2-norm, ℓ_1-norm and the sup-norm of the vectorization of M, respectively. $||M||_{\text{op}}$ is the operator norm of M and $||M||_0$ is the number of non-zero coefficients in M, that is its ℓ_0-norm. Then, for $k = 1, \dots, p$, 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 projection 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 ξ, the vectors $(y_1, \dots, y_n)^\top$ and $(\xi_1, \dots, \xi_n)^\top$, 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, ξ_i) is independent of the sample size. We consider an asymptotic setting where n goes to ∞ and where ϵ, 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, $\epsilon \to 0$. 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 large for a negligible proportion of individuals. The subsequent theorems will help to quantify these statements.

2.2. Estimation procedure

Note that in this asymptotic setting the OLS estimator is not consistent. Indeed, in the model (1.2) with a single constant regressor equal to 1, the value of the OLS estimator is $\beta + (\sum_{i=1}^n \alpha_i)/n$, which may not converge to β if the nonzero coefficients of α are too large. Therefore, we propose an alternative procedure estimating 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

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

where λ is a penalty level whose 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 ξ_i. Hence, the proposed procedure is simple in that, unlike the least trimmed squares estimator, it does not use any tuning parameter. 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 $b \in \mathbb{R}^p$. 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 $b \in \mathbb{R}^p$, 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 $\frac{1}{\sqrt{n}}||y - Xb - \alpha||_2 + \frac{\lambda}{n}||\alpha||_1 = \frac{1}{\sqrt{n}}||M_X(y-\alpha)||_2 + \frac{\lambda}{n}||\alpha||_1$ when $Xb = P_X(y-\alpha)$, it holds that

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

Under the assumptions developed below, $\widehat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal. Remark that model (1.1) can be seen as a standard linear model with α_i corresponding to the parameter of a dummy variable whose value is 1 for the individual i, and 0 otherwise. Hence, the estimator can be viewed as the square-root lasso estimator of [3]. Our approach encounters additional technical difficulties, given that we penalize only a subset of the variables and there is no way to estimate α consistently as each of its entries is indirectly observed only once. To address this problem, we develop new assumptions and theorems better suited to the purposes of this paper.

2.3. Assumptions and results

Our first assumption formalizes the hypotheses made on the model in the introduction.

Assumption 2.1. The following holds:

(i) $\{(x_i, \xi_i)\}_i$ are i.i.d. random variables;
(ii) $\mathbb{E}[x_i\xi_i] = \mathbb{E}[\xi_i] = 0$;
(iii) $\Sigma = \mathbb{E}[x_i x_i^\top]$ exists and is positive definite;
(iv) there exists $\sigma > 0$ such that $0 < \mathrm{var}[\xi_i^2|x_i] = \sigma^2 < \infty$. The conditional variance σ^2 does not scale with n.

This assumption is standard in linear regression models. The main assumption concerns the choice of the penalty level:

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

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

Lemma 2.3. *Under Assumption 2.1, it holds that* $2\sqrt{n}\frac{||M_X\xi||_\infty}{||M_X\xi||_2} \leq 2\frac{||\xi||_\infty}{\sigma} + o_P(||\xi||_\infty) + O_P(1)$. *Moreover, if* ψ *is such that* $\lim_{n\to\infty} \mathbb{P}\left(\psi \geq 2\frac{||\xi||_\infty}{\sigma}\right) = 1$ *and* $\varphi \to \infty$, *then for any* $c > 1$, $\lambda = c\psi + \varphi$ *satisfies Assumption 2.2.*

The proof is given in the appendix. This lemma removes the role of the matrix X in the choice of the penalty parameter and simplifies the decision procedure. It leads to the following corollary.

Corollary 2.4. *Under Assumption 2.1, the following holds:*

(i) If ξ_i *are Gaussian random variables, then* $\lambda = 2c\sqrt{2\log(n)}$ *satisfies Assumption 2.2 for any* $c > 1$.
(ii) If ξ_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 ξ_i *are sub-exponential random variables, then there exists a constant* $c > 0$ *such that* $\lambda = c\log(n)$ *satisfies Assumption 2.2.*

The proof is given in the appendix. The statistician can use Corollary 2.4 to decide on the penalization parameter given the expected weight of the tails of the error term in the data. In practice, it is preferable to choose the smallest penalty verifying Assumption 2.2. This can be done using Monte-Carlo simulations if one is willing to specify the distribution of the errors.

When the errors are sub-Gaussian or sub-exponential, the constant c in Corollary 2.4 is unknown. We can circumvent this issue by replacing c by a sequence ϕ such that $\phi \to \infty$. For sub-Gaussian distributions, we can use $\lambda = \log(n)$ and for sub-exponential distributions, take $\lambda = \log(n)^{\frac{3}{2}}$ to satisfy Assumption 2.2. Note that the sequence ϕ also needs to be chosen such that it satisfies the other assumptions made on λ to obtain the asymptotic results below (in Lem. 2.5 and Thms. 2.6 and 2.7).

When the "theoretical" choice of the penalty is unkown, a standard solution in the theory of the lasso estimator is to use an iterative procedure (see [2]), which assumes that we possess an upper bound $\bar{c}$ on c (possibly data-driven). In the first step, we would use $\lambda = \bar{c}\sqrt{\log(n)}$ if the error terms are sub-Gaussian ($\lambda = \bar{c}\log(n)$ in the sub-exponential case). The penalty level of the estimator $(\widehat{\beta}, \widehat{\alpha})$ of this first iteration would satisfy Assumption 2.2 because $\bar{c} \geq c$. Then, one can use the estimated errors $\widehat{\epsilon}_i = y_i - x_i^\top \widehat{\beta} - \widehat{\alpha}$ to estimate c. In the standard literature of the lasso, c corresponds to the variance of the error terms, which is simple to estimate. However, in our case, the proof of Corollary 2.4 shows that c is related to the variance proxy of the errors if they are sub-Gaussian. The variance proxy of sub-Gaussian random variables is complicated to estimate because it is related to the sub-Gaussian norm, which is an infimum. Therefore, this common solution seems hardly feasible in the setting of this paper.

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

Lemma 2.5. *Under Assumptions 2.1 and 2.2 and if* $\sqrt{\epsilon}\max(\lambda, ||X||_\infty) = o_P(1)$*, it holds that*

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

The proof is given in the appendix. The rate of convergence of $||\widehat{\alpha} - \alpha||_1/n$ is therefore lower than $\epsilon\sqrt{\log(n)}$ if the errors are Gaussian or sub-Gaussian and we choose the penalty level as in Lemma 2.4. 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 $\epsilon = o_P(1/\log(n))$.

Here, we show how to derive the rate of convergence of $\widehat{\beta}$ from Lemma 2.5. Assuming that the conditions of Lemma 2.5 hold and replacing y by $X\beta + \alpha + \xi$ in (2.1), we obtain

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

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

$$\begin{aligned}
\left|\left|\left(X^\top X\right)^{-1} X^\top (\alpha - \widehat{\alpha})\right|\right|_2 &\leq \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{1}{n} \left|\left|X^\top (\alpha - \widehat{\alpha})\right|\right|_2 \\
&= O_P(1)\frac{1}{n}\sqrt{\sum_{k=1}^{K}\left(\sum_{i=1}^{n}(x_i)_k (\alpha_i - \widehat{\alpha}_i)\right)^2} \\
&\leq O_P(1)\frac{1}{n}\sqrt{\sum_{k=1}^{K}||X||_\infty^2 ||\alpha - \widehat{\alpha}||_1^2}
\end{aligned}$$

$$= O_P\left(\frac{1}{n}||X||_\infty||\alpha-\widehat{\alpha}||_1\right). \tag{2.4}$$

By Lemma 2.5, we obtain $||(X^\top X)^{-1}X^\top(\alpha-\widehat{\alpha})||_2 = O_P\left(p\lambda||X||_\infty\right)$. Finally, we have $\sqrt{n}(X^\top X)^{-1}X^\top\xi \xrightarrow{d} \mathcal{N}(0,\sigma\Sigma^{-1})$. This leads to Theorem 2.6.

Theorem 2.6. *Under Assumptions 2.1 and 2.2 and if* $\sqrt{\epsilon}\max(\lambda,||X||_\infty) = o_P(1)$*, it holds that*

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

This result allows us to derive the rate 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 $\{\xi_i\}_i$ are i.i.d. sub-Gaussian random variables, then $\widehat{\beta}$ is consistent as long as $\epsilon\log(n)\to 0$ for the choice of λ proposed in Lemma 2.4. Hence in this case, the estimator reaches (up to a logarithmic factor) the minimax lower bound for the Huber's contamination location model under Gaussian errors, which is $\max(1/n,\epsilon^2)$ in squared ℓ_2-norm according to [4]. We attain the rate $\max(1/n,\epsilon^2\log(n))$. Remark also that equation (2.4) explains the role of $||X||_\infty$ in the rate of convergence 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 (2.3). $||X||_\infty$ 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 (2.3) vanishes asymptotically.

Theorem 2.7. *Under Assumptions 2.1 and 2.2, assuming that* $\sqrt{\epsilon}\max(\lambda,||X||_\infty) = o_P(1)$*,* $\epsilon\lambda\sqrt{n} = o(1)$ *and* $\epsilon\lambda||X||_\infty\sqrt{n} = o_P(1)$*, 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* σ^2 *and* Σ*, respectively.*

The proof that $\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$ is given in appendix. When the entries of X and ξ are sub-Gaussian and the penalty parameter is chosen as in Lemma 2.4, a sufficient condition for the assumptions on λ in Theorem 2.7 is that $\epsilon\log(n)\sqrt{n}\to 0$. Notice that the asymptotic variance of our estimator corresponds to that 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 2.7. They are obtained under an asymptotic setting with triangular array data in which the number of outliers is allowed to go to infinity while the proportion of outliers goes to 0. The 95% confidence interval I built with Theorem 2.7 is interpreted 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 [16] 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

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

This is a convex optimization program 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. $\beta^{(t+1)} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \left|\left|y - X\beta - \alpha^{(t)}\right|\right|_2^2$;
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 [9] and explains how to perform step 2.

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

An interesting remark is that (3.1) and Theorem 3.1 in [18] imply that

$$\left(\widehat{\beta}, \widehat{\sigma}\right) \in \underset{\beta\in\mathbb{R}^p,\ \sigma\in\mathbb{R}_+}{\arg\min} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^\top\beta}{\sigma}, \frac{\lambda}{\sqrt{n}}\right)\sigma + \frac{\sigma}{2},$$

where $\rho(t, a)$ is Huber's loss function

$$\rho(t, a) = \begin{cases} a\left|t\right| - a^2/2 & \text{if } \left|t\right| > a \\ t^2/2 & \text{otherwise.} \end{cases}$$

3.2. Simulations

We apply this algorithm in a small simulation exercise. Consider a data generating process in which the OLS estimator is asymptotically biased. There are two regressors x_{1i} and x_{2i}, with $x_{1i} = 1$ for all i. The random variables $\{x_{2i}\}_i$ are i.i.d. with a uniform distribution on the interval $[0, 1]$. ξ_i are i.i.d. $\mathcal{N}(0, 1)$ random variables. The parameter is $\beta = (1, 1)^\top$. Then, for $z \in \mathbb{R}$, we set

$$\alpha_i = \begin{cases} 0 & \text{if } x_{2i} < 1 - \epsilon \\ z & \text{if } x_{2i} \geq 1 - \epsilon. \end{cases}$$

In Tables 1–3 we present the bias, the variance and the mean squared error (MSE) of $\widehat{\beta}$ for various values of n, ϵ and z. We use 100 iterations and $\lambda = 2.01\sqrt{2\log(n)}$. This choice corresponds to the one in Corollary sub. The Bias, the variance and the MSE of the naive OLS estimator

$$\widetilde{\beta}^{OLS} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \left|\left|y - X\beta\right|\right|_2^2$$

are also reported. For our estimator, the coverage of 95% confidence intervals based on the asymptotic variance of Theorem 2.7 is presented in Table 4. The coverage reported in the same table for the OLS estimator corresponds

TABLE 1. $n = 1,000$, $\epsilon = 0.005$, $z = 20$.

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	-0.094	-0.197	0.283	0.592
var	0.007	0.012	0.040	0.082
MSE	0.016	0.050	0.120	0.433

TABLE 2. $n = 10,000$, $\epsilon = 0.0005$, $z = 200$.

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	0.009	-0.199	0.027	0.597
var	4×10^{-4}	0.009	0.001	0.074
MSE	5×10^{-4}	0.048	0.002	0.430

TABLE 3. $n = 100,000$, $\epsilon = 0.00005$, $z = 2,000$.

	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
bias	-0.001	-0.192	0.003	0.598
var	4×10^{-5}	0.008	1×10^{-4}	0.072
MSE	4×10^{-5}	0.048	1×10^{-4}	0.429

TABLE 4. Coverage of 95% confidence intervals based on Theorem 2.7.

n	ϵ	z	$\widehat{\beta}_1$	$\widetilde{\beta}_1^{OLS}$	$\widehat{\beta}_2$	$\widetilde{\beta}_2^{OLS}$
1,000	0.005	20	0.77	0.58	0.47	0.17
10,000	0.0005	200	0.93	0.37	0.87	0.05
100,000	0.00005	2000	0.94	0.33	0.94	0.03

to that of the confidence intervals built using the theory of the OLS estimator in the standard linear regression model without outliers. The results are averages over 8000 replications. We observe that our estimator performs better than the OLS estimator.

4. PROOFS

4.1. Proof of Lemma 2.3

We start by proving the next two lemmas:

Lemma 4.1. *Under Assumption 2.1, it holds that* $||P_X\xi||_\infty = O_P(1)$.

Proof. By the law of large numbers, the central limit theorem and Slutsky's theorem, we have $\sqrt{n}(X^\top X)^{-1}X^\top\xi \xrightarrow{d} \mathcal{N}(0, \sigma\Sigma^{-1})$, therefore $\sqrt{n}||(X^\top X)^{-1}X^\top\xi||_2 = O_P(1)$. Because $X(X^\top X)^{-1}X^\top\xi = \frac{X}{\sqrt{n}}\sqrt{n}(X^\top X)^{-1}X^\top\xi$, we obtain that

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

by the law of large numbers. □

Lemma 4.2. *Under Assumption 2.1, it holds that* $\frac{\sqrt{n}}{||M_X\xi||_2} - \frac{1}{\sigma} = o_P(1)$.

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

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

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

Now, we proceed with the proof of Lemma 2.3. Notice that

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

Using Lemmas 4.1 and 4.2, we obtain

$$2\sqrt{n}\frac{||M_X\xi||_\infty}{||M_X\xi||_2} \leq 2\frac{||\xi||_\infty}{\sigma} + o_P(||\xi||_\infty) + O_P(1). \tag{4.1}$$

The rest of the proof of the lemma is a direct consequence of (4.1) and of the pigeonhole principle.

4.2. Proof of Corollary 2.4

Proof of (i). By Lemma 2.3, it is sufficient to show that for $c > 1$,

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

Let us remember the Gaussian bound (see Lem. B.1 in [9]): for $t \geq 0$, we have $\mathbb{P}\left(\frac{|\xi_i|}{\sigma} \geq t\right) \leq e^{-\frac{t^2}{2}}$. Then, we have

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

Proof of (ii). By Lemma 2.3, it suffices to show that there exists $c > 0$ such that

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

Recall the sub-Gaussian bound (see Prop. 2.5.2 in [20]): for $t \geq 0$, there exists $b > 0$ such that $\mathbb{P}\left(\frac{|\xi_i|}{\sigma} \geq t\right) \leq 2e^{-\frac{t^2}{2b}}$. Then, for $\rho > 1$, we have

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

Proof of (iii). Using Lemma 2.3, we only need to show that there exists $c > 0$ such that

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

Let us state the sub-exponential bound (see Prop. 2.7.1 in [20]): for $t \geq 0$, there exists $b > 0$ such that $\mathbb{P}\left(\frac{|\xi_i|}{\sigma} \geq t\right) \leq 2e^{-\frac{t}{2b}}$. Then, for n large enough and $\rho > 1$, we have

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

4.3. Proof of Lemma 2.5

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 $\delta_{J^c} = \delta - \delta_J$ and introduce the following cone

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

We work with the following compatibility constant

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

and use the following lemma.

Lemma 4.3. *Under Assumption 2.1, if $\sqrt{\epsilon}||X||_\infty = o_P(1)$, there exists $\kappa_* > 0$ such that $\mathbb{P}(\kappa > \kappa_*) \to 1$.*

Proof. Take $\delta \in C$, to show this result, notice that $M_X\delta = \delta - X(X^\top X)^{-1}X^\top\delta$. This implies that

$$||M_X\delta||_2 \geq ||\delta||_2 - ||X(X^\top X)^{-1}X^\top\delta||_2$$
$$= ||\delta||_2 - \left|\left|\sum_{k=1}^{p} X_k\left((X^\top X)^{-1}X^\top\delta\right)_k\right|\right|_2$$
$$\geq ||\delta||_2 - \sum_{k=1}^{p}\left|\left|X_k\left((X^\top X)^{-1}X^\top\delta\right)_k\right|\right|_2$$
$$\geq ||\delta||_2 - \sum_{k=1}^{p}||X_k||_2\left|\left|(X^\top X)^{-1}X^\top\delta\right|\right|_\infty$$

$$\geq ||\delta||_2 - \sum_{k=1}^{p} ||X_k||_2 \left|\left|(X^\top X)^{-1} X^\top \delta\right|\right|_2$$
$$\geq ||\delta||_2 - \sum_{k=1}^{p} ||X_k||_2 \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{1}{n} ||X^\top \delta||_2.$$

Next, as in (2.4), we obtain

$$\begin{aligned}
||M_X \delta||_2 &\geq ||\delta||_2 - \sum_{k=1}^{p} ||X_k||_2 \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{p}}{n} ||X||_\infty ||\delta||_1 \\
&\geq ||\delta||_2 - \sum_{k=1}^{p} ||X_k||_2 \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{p}}{n} ||X||_\infty 4||\delta_J||_1 \quad \text{(because } \delta \in C\text{)} \\
&\geq ||\delta||_2 - \sum_{k=1}^{p} ||X_k||_2 \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} \frac{\sqrt{p}}{n} ||X||_\infty 4\sqrt{||\alpha||_0} ||\delta_J||_2 \quad \text{(because } ||\delta_J||_0 \leq ||\alpha||_0\text{)} \\
&\geq ||\delta||_2 - \sum_{k=1}^{p} \frac{||X_k||_2}{\sqrt{n}} \left|\left|\left(\frac{1}{n} X^\top X\right)^{-1}\right|\right|_{\text{op}} 4\sqrt{p}\sqrt{\frac{||\alpha||_0}{n}} ||X||_\infty ||\delta||_2. \qquad (4.2)
\end{aligned}$$

Then, we have

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

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

End of the proof of Lemma 2.5. Throughout this proof, we work on the event

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

whose probability goes to 1 because of Assumption 2.2, Lemma 4.3 and the condition that $\sqrt{\epsilon}\lambda \to 0$. Let us define $\Delta = \widehat{\alpha} - \alpha$. Now, remark that

$$\begin{aligned} ||\widehat{\alpha}||_1 &= ||\alpha + \Delta||_1 \\ &= ||\alpha + \Delta_J + \Delta_{J^c}||_1 \\ &\geq ||\alpha + \Delta_{J^c}||_1 - ||\Delta_J||_1 . \end{aligned} \tag{4.3}$$

Next, we use the fact that $||\alpha + \Delta_{J^c}||_1 = ||\alpha||_1 + ||\Delta_{J^c}||_1$. Combining this and (4.3), we get

$$||\widehat{\alpha}||_1 \geq ||\alpha||_1 + ||\Delta_{J^c}||_1 - ||\Delta_J||_1 . \tag{4.4}$$

Using (2.2), we have

$$\frac{1}{\sqrt{n}}||M_X(y - \widehat{\alpha})||_2 + \frac{\lambda}{n}||\widehat{\alpha}||_1 \leq \frac{1}{\sqrt{n}}||M_X(y - \alpha)||_2 + \frac{\lambda}{n}||\alpha||_1. \tag{4.5}$$

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

$$\begin{aligned} \frac{1}{\sqrt{n}}||M_X(y - \widehat{\alpha})||_2 - \frac{1}{\sqrt{n}}||M_X(y - \alpha)||_2 &\geq -\frac{1}{\sqrt{n}||M_X\xi||_2} \langle M_X(\xi), \Delta \rangle \\ &\geq -\frac{\lambda}{2n}||\Delta||_1, \end{aligned} \tag{4.6}$$

where (4.6) comes from $\lambda \geq 2\sqrt{n}||M_X\xi||_\infty / ||M_X\xi||_2$. This last inequality is also straightforwardly true when $M_X\xi = 0$. This and (4.5) imply

$$||\hat{\alpha}||_1 \leq \frac{1}{2}||\Delta||_1 + ||\alpha||_1. \tag{4.7}$$

Using (4.4) and (4.7), 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

$$||\Delta_{J^c}||_1 \leq 3\,||\Delta_J||_1 , \tag{4.8}$$

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

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

By Hölder's inequality, this results in

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

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

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

Given that $\Delta \in C$, this implies

$$\frac{1}{n}||M_X(\widehat{\alpha}-\alpha)||_2^2 \leq \frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2 + \frac{4\lambda||M_X\xi||_2}{n^{\frac{3}{2}}}\,||\Delta_J||_1\,. \tag{4.9}$$

By equations (4.4) and (4.5), 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 (4.6), 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

$$\begin{aligned}&\left|\frac{1}{n}||M_X(y-\widehat{\alpha})||_2^2 - \frac{1}{n}||M_X(y-\alpha)||_2^2\right|\\ &\leq \frac{2\lambda}{n}\,||\Delta_J||_1\left(\frac{2}{\sqrt{n}}||M_X(y-\alpha)||_2 + \frac{2\lambda}{n}\,||\Delta_J||_1\right)\\ &\leq \left(\frac{2\lambda}{n}\right)^2||\Delta_J||_1^2 + \frac{4}{\sqrt{n}}||M_X(y-\alpha)||_2\frac{\lambda}{n}\,||\Delta_J||_1\,.\end{aligned} \tag{4.10}$$

Combining (4.9) and (4.10) and noting that $||M_X\xi||_2 = ||M_X(y-\alpha)||_2$, we obtain

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

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

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

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

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

which implies again that

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

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

$$\frac{1}{n}||\Delta||_1 \leq \frac{4}{n}||\Delta_J||_1 \leq \left(1 - \left(\frac{2\sqrt{\frac{||\alpha||_0}{n}}\lambda}{\kappa_*}\right)^2\right)^{-1} \frac{32||M_X\xi||_2\frac{||\alpha||_0}{n}\lambda}{\sqrt{n}\kappa_*^2}. \tag{4.11}$$

Because we work on the event $\kappa > \kappa^*$, we have $\kappa > 0$. Hence, if $M_X\Delta = 0$, then $\Delta_J = 0$ and $\Delta = 0$ using the fact that Δ belongs to C. Therefore, inequality (4.11) also holds when $M_X\Delta = 0$. To conclude the proof, use (4.11), the fact that $||M_X\xi||_2/\sqrt{n} \leq ||\xi||_2/\sqrt{n} = O_P(1)$ by the law of large numbers and the condition $\sqrt{\epsilon}\max(\lambda, ||X||_\infty) = o_P(1)$.

4.4. Proof that $\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$ in Theorem 2.7

We have

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

Then, because of Lemma 2.5, Theorem 2.6, $\epsilon\lambda\sqrt{n} = o(1)$ and $\epsilon\lambda|X|_\infty = o_P(1)$, it holds that

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

Next, we have

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

by the law of large numbers. Then, because the ℓ_2-norm is lower than the ℓ_1-norm, we have

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

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 central limit theorem implies that $||\xi||_2/\sqrt{n}=O_P(1)$. Thus, by the Cauchy-Schwarz inequality, we have

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

and

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

which concludes the proof.

References

[1] A. Alfons, C. Croux and S. Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.* **7** (2013) 226–248.

[2] A. Belloni, V. Chernozhukov, *et al.* Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** (2013) 521–547.

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

[4] M. Chen, C. Gao, Z. Ren, *et al.* Robust covariance and scatter matrix estimation under huber's contamination model. *Ann. Stat.* **46** (2018) 1932–1960.

[5] O. Collier and A.S. Dalalyan. Rate-optimal estimation of p-dimensional linear functionals in a sparse gaussian model. Preprint arXiv:1712.05495 (2017).

[6] A.S. Dalalyan, SOCP based variance free Dantzig selector with application to robust estimation. *C. R. Math.* **350** (2012) 785–788.

[7] J. Fan, Q. Li and Y. Wang, Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc.* **79** (2017) 247–265.

[8] I. Gannaz, Robust estimation and wavelet thresholding in partially linear models. *Stat. Comput.* **17** (2007) 293–310.

[9] C. Giraud, Introduction to high-dimensional statistics. Chapman and Hall/CRC, Boca Raton (2014).

[10] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel. Robust statistics: the approach based on influence functions, Vol. 196. John Wiley & Sons, New Jersey (2011).

[11] P.J. Huber *et al.*, Robust estimation of a location parameter. *Ann. Math. Stat.* **35** (1964) 73–101.

[12] S. Lambert-Lacroix, L. Zwald, *et al.*, Robust regression through the huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5** (2011) 1015–1053.

[13] Y. Lee, S.N. MacEachern, Y. Jung, *et al.*, Regularization of case-specific parameters for robustness and efficiency. *Stat. Sci.* **27** (2012) 350–372.

[14] W. Li. *Simultaneous variable selection and outlier detection using LASSO with applications to aircraft landing data analysis*. Ph.D. thesis, Rutgers University-Graduate School-New Brunswick (2012).

[15] R.A. Maronna, R.D. Martin, V.J. Yohai and M. Salibián-Barrera, Robust statistics: theory and methods (with R). John Wiley & Sons, New Jersey (2018).

[16] A.B. Owen, A robust hybrid of lasso and ridge regression. *Contemp. Math.* **443** (2007) 59–72.

[17] P.J. Rousseeuw and A.M. Leroy. Robust regression and outlier detection, Vol. 589. John Wiley & Sons, New Jersey (2005).

[18] Y. She and A.B. Owen. Outlier detection using nonconvex penalized regression. *J. Am. Stat. Assoc.* **106** (2011) 626–639.

[19] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika* **99** (2012) 879–898.

[20] R. Vershynin, High-dimensional probability: An introduction with applications in data science, Vol. 47. Cambridge University Press, Cambridge (2018).