A New Principle for Tuning-Free Huber Regression

Lili Wang[♭], Chao Zheng[♯], Wen Zhou[†], and Wen-Xin Zhou[¶]

[♭] *Zhejiang Gongshang University,* [♯] *University of Southampton*
[†] *Colorado State University,* [¶] *University of California San Diego*

Abstract: The robustification parameter, which balances bias and robustness, has played a critical role in the construction of sub-Gaussian estimators for heavy-tailed and/or skewed data. Although it can be tuned by cross-validation in traditional practice, in large scale statistical problems such as high dimensional covariance matrix estimation and large scale multiple testing, the number of robustification parameters scales with the dimensionality so that cross-validation can be computationally prohibitive. In this paper, we propose a new data-driven principle to choose the robustification parameter for Huber-type sub-Gaussian estimators in three fundamental problems: mean estimation, linear regression, and sparse regression in high dimensions. Our proposal is guided by the non-asymptotic deviation analysis, and is conceptually different from cross-validation which relies on the mean squared error to assess the fit. Extensive numerical experiments and real data analysis further illustrate the efficacy of the proposed methods.

Key words and phrases: Data adaptive, heavy tails, Huber loss, M-estimator, tuning parameters

1. Introduction

Data subject to heavy-tailed and/or skewed distributions are frequently observed across various disciplines (Cont, 2001; Purdom and Holmes, 2005). Rigorously, a random variable X is heavy-tailed if its tail probability $\mathbb{P}(|X| > t)$ decays to zero polynomially in $1/t$ as $t \to \infty$, or equivalently, if X has finite polynomial-order moments. The connection between moment and tail probability is revealed by the property that $\mathbb{E}(|X|^k) = k\int_0^\infty t^{k-1}\mathbb{P}(|X| > t)\,\mathrm{d}t$ for any $k \geq 1$. When the sampling distribution has only a small number of finite moments, with high chance some observations will deviate wildly from the population mean. Such observations are known as outliers caused by a heavy-tailed noise. In contrast, samples generated from a Gaussian or sub-Gaussian distribution (Vershynin, 2012) are strongly concentrated around the expected value, so that the chance of having extreme observations is much smaller.

Heavy-tailed data bring new challenges to conventional statistical methods. For linear models, regression estimators based on the least squares loss are suboptimal, both theoretically and empirically, in the presence of heavy-tailed errors. We refer to Catoni (2012) for a deviation analysis, showing that the deviation of the empirical mean can be much worse for non-Gaussian samples than for Gaussian ones. More broadly, this study exemplifies the pitfalls of asymptotic studies in statistics and inspires new thoughts about the notions of optimality commonly used to assess the performance of estimators. In particular, the minimax optimality under mean squared error does not quite capture the influence of extreme behaviors of estimators. However, these rare events may have catastrophically negative impacts in practice, leading to wrong conclusions or false discoveries. Since Catoni (2012), the non-asymptotic deviation analysis has drawn considerable attention and it is becoming increasingly important to construct sub-Gaussian estimators (see Section S1.2 in the supplementary file) for heavy-tailed data; see, for example, Brownlees, Joly and Lugosi (2015), Minsker (2015, 2018), Hsu and Sabato (2016), Devroye et al. (2016), Lugosi and Mendelson (2016), Fan, Li and Wang (2017), Lugosi and Mendelson (2019), Lecué and Lerasle (2017), and Zhou et al. (2018), among others.

For linear models, Fan, Li and Wang (2017) and Zhou et al. (2018) proposed Huber-type estimators in both low and high dimensional settings and derived non-asymptotic deviation bounds for the estimation error. To implement either Catoni's or Huber-type method, a tuning parameter τ needs to be specified in advance to balance between robustness and bias of the estimation. Deviation analysis suggests that this tuning parameter, which we refer to as the robustification parameter, should adapt to the sample size, dimension, variance of noise and confidence level. Calibration schemes are typically based on cross-validation or Lepski's method, which can be computationally intensive especially for large-scale inference and high dimensional estimation problems where the number of parameters may be exponential in the number of observations. For example, Avella-Medina et al. (2018) proposed adaptive Huber estimators for estimating high dimensional covariance and precision matrices. For a $d \times d$ covariance matrix, although every entry can be robustly estimated by a Huber-type estimator with τ chosen via cross-validation, the overall procedure involves as many as d^2 tuning parameters and therefore the cross-validation method will soon become computationally intractable as d grows. Efficient tuning is important not only for the problem's own interest, but also for applications in a broader context.

This paper develops data-driven Huber-type methods for mean estimation, linear regression, and sparse regression in high dimensions. For each problem, we first provide sub-Gaussian concentration bounds for the Huber-type estimator under minimal moment condition on the errors. These non-asymptotic results guide the choice of key tuning parameters. Some of them are of independent interest and improve the existing results by weakening the sample size scaling. Secondly, we propose a novel data-driven principle to calibrate the robustification parameter $\tau > 0$ in the Huber loss

$$\ell_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \le \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases} \tag{1.1}$$

Huber proposed $\tau = 1.345\sigma$ to retain 95% asymptotic efficiency of the estimator for the normally distributed data, and meanwhile to guarantee the estimator's performance towards arbitrary contamination in a neighborhood of the true model (Huber, 1981; Huber and Ronchetti, 2009). This default setting has found its use in high dimensional statistics even though the asymptotic efficiency is no longer well defined; see, for example, Lambert-Lacroix and Zwald (2011), Elsener and van de Geer (2018), and Loh (2017). Guided by the non-asymptotic deviation analysis, our proposed τ grows with sample size for bias-robustness trade-off. For linear regression under different regimes, the optimal τ depends on the dimension d: $\tau \sim \sigma\sqrt{(n/d)}$ in the low dimensional setting with small d/n and $\tau \sim \sigma\sqrt{n/\log(d)}$ in high dimensions. Lastly, we design simple and fast algorithms to implement our method to calibrate τ.

In this paper, we focus on the notion of tail robustness (Catoni, 2012; Minsker, 2018; Zhou et al., 2018; Fan, Li and Wang, 2017; Avella-Medina et al., 2018), which is characterized by the tight non-asymptotic deviation guarantees of estimators under weak moment assumptions and evidenced by the better finite-sample performance in the presence of heavy-tailed and/or highly skewed noise. It is inherently different from the traditional definition of robustness under Huber's ϵ-contamination model (Huber and Ronchetti, 2009). Following the introduction of the finite sample breakdown point by Donoho and Huber (1983), the traditional robust statistics has focused, in part, on the development of high breakdown point estimators. Informally, the breakdown point of an estimator is defined as the largest proportion of contaminated samples in the data that an estimator can tolerate before produces arbitrarily large estimates (Hampel, 1971; Hampel et al., 1986; Maronna et al., 2018). An estimator with a high breakdown point does not necessarily shed light on its convergence properties, efficiency, and stability. We refer to Portnoy and He (2000) for a review on classical robust statistics. In contrast, a tail robust estimator is resilient to outliers caused by a heavy-tailed noise. Intuitively, the breakdown point describes a form of the worst-case robustness, while our focus corresponds to the average-case robustness.

The remainder of this paper is organized as follows. In Section 2, we revisit Catoni's method on robust mean estimation. Motivated by a careful analysis of the truncated sample mean, we introduce a novel data-driven adaptive Huber estimator. We extend this data-driven tuning scheme to robust regression in Section 3 under both low and high dimensional settings. Extensive numerical experiments are reported in Section 4 to demonstrate the finite sample performance of the proposed procedures. All the proofs, together with technical details and real data examples, are relegated to the supplementary files.

2. Robust data-adaptive mean estimation

2.1 Motivation

To motivate our proposed data-driven scheme for Huber-type estimators, we start with revisiting the mean estimation problem. Let $X_1, \ldots, X_n$ $(n \geq 2)$ be independent and identically distributed $(i.i.d.)$ copies of X with mean μ and finite variance $\sigma^2 > 0$. The sample mean, denoted as $\bar{X}_n$, is the most natural estimator for μ. However, it severely suffers from not being robust to heavy-tailed sampling distributions (Catoni, 2012). In order to cancel, or at least dampen, the erratic fluctuations in $\bar{X}_n$ which are more likely to occur if the distribution of X is heavy-tailed, we consider the truncated sample mean $m_\tau = n^{-1} \sum_{i=1}^n \psi_\tau(X_i)$ for some $\tau > 0$, where

$$\psi_\tau(x) = \mathrm{sign}(x) \min(|x|, \tau) \tag{2.1}$$

is a truncation function on $\mathbb{R}$. Here, the tuning parameter τ controls the bias and tail robustness of m_τ. To see this, note that the bias term $\mathrm{Bias} := \mathbb{E}(m_\tau) - \mu$ satisfies $|\mathrm{Bias}| = |\mathbb{E}\{X - \mathrm{sign}(X)\tau\} I(|X| > \tau)| \leq \tau^{-1}\mathbb{E}(X^2)$. Regarding tail robustness, the following result shows that m_τ with a properly chosen τ is a sub-Gaussian estimator as long as the second moment of X is finite.

Proposition 2.1. Assume that $v_2 := \sqrt{\mathbb{E}(X^2)}$ is finite. For any $z > 0$,

(i) m_τ with $\tau = v\sqrt{n/z}$ for some $v \geq v_2$ satisfies $\mathbb{P}\{|m_\tau - \mu| \geq 2v\sqrt{z/n}\} \leq 2e^{-z}$;

(ii) m_τ with $\tau = cv_2\sqrt{n/z}$ for some $0 < c \leq 1$ satisfies $\mathbb{P}\{|m_\tau - \mu| \geq 2(v_2/c)\sqrt{z/n}\} \leq 2e^{-z/c^2}$.

Proposition 2.1 shows that how m_τ would perform under various idealized scenarios, as such providing guidance on the choice of τ. Here $z > 0$ is a user-specified parameter that controls the confidence level; see further discussions before Remark 2.2. Given a properly tuned τ, the sub-Gaussian performance is achieved; conversely, if the resulting estimator performs well, the data have been truncated at the right level and can be further exploited. An ideal τ is such that the sample mean of truncated data $\psi_\tau(X_1), \ldots, \psi_\tau(X_n)$ serves as a good estimator of μ. The influence of outliers caused by a heavy-tailed noise is weakened due to the proper truncation. At the same time, we may expect that the empirical second moment for the same truncated data will provide a reasonable estimate of v_2^2. Motivated by this, we propose to choose $\tau > 0$ by solving $\tau = \{\sum_{i=1}^n \psi_\tau^2(X_i)\}^{1/2}\sqrt{n/z}$, which is equivalent to

$$\frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau^2(X_i)}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{2.2}$$

We will show that under mild conditions, (2.2) has a unique solution $\widehat{\tau}_z$, which gives rise to a data-driven mean estimator

$$m_{\widehat{\tau}_z} = \frac{1}{n}\sum_{i=1}^n \min(|X_i|, \widehat{\tau}_z)\,\mathrm{sign}(X_i). \tag{2.3}$$

To understand the property of $\widehat{\tau}_z$, consider the population version of (2.2):

$$\frac{\mathbb{E}\{\psi_\tau^2(X)\}}{\tau^2} = \frac{\mathbb{E}\{\min(X^2,\tau^2)\}}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{2.4}$$

The following result establishes existence and uniqueness of the solution to (2.4).

Proposition 2.2. Assume that $v_2 = \sqrt{\mathbb{E}(X^2)}$ is finite.

(i) Provided $0 < z < n\mathbb{P}(|X| > 0)$, (2.4) has a unique solution τ_z, which satisfies $[\mathbb{E}\{\min(X^2, q_{z/n}^2)\}]^{1/2}\sqrt{n/z} \le \tau_z \le v_2\sqrt{n/z}$, where $q_\alpha := \inf\{t : \mathbb{P}(|X| > t) \le \alpha\}$ is the upper α-quantile of $|X|$.

(ii) Let $z = z_n > 0$ satisfy $z_n \to \infty$ and $z = o(n)$. Then $\tau_z \to \infty$ and $\tau_z \sim v_2\sqrt{n/z}$ as $n \to \infty$.

As a direct consequence of Proposition 2.2, the following result ensures existence and uniqueness of the solution to (2.2), the empirical counterpart of (2.4).

Proposition 2.3. Provided $0 < z < \sum_{i=1}^n I(|X_i| > 0)$, (2.2) admits a unique solution.

Throughout, denote $\widehat{\tau}_z$ the solution to (2.2), which is unique and positive whenever $z < \sum_{i=1}^n I(|X_i| > 0)$. For completeness, we set $\widehat{\tau}_z = 0$ on $\{z \ge \sum_{i=1}^n I(|X_i| > 0)\}$. If $\mathbb{P}(X = 0) = 0$ and $0 < z < n$, $\widehat{\tau}_z > 0$ with probability one. With both τ_z and $\widehat{\tau}_z$ well defined, we investigate the property of $\widehat{\tau}_z$ below.

Theorem 2.1. Assume $\mathbb{E}(X^2) < \infty$ and $\mathbb{P}(X = 0) = 0$. For any $1 \le z < n$ and $0 < r < 1$, we have

$$\mathbb{P}(|\widehat{\tau}_z/\tau_z - 1| \ge r) \le e^{-a_1^2 r^2 z^2/(2z + 2a_1 rz/3)} + e^{-a_2^2 r^2 z/2} + 2e^{-(a_1 \wedge a_2)^2 z/8}, \tag{2.5}$$

where

$$a_1 = a_1(z,r) = \frac{P(\tau_z)}{2Q(\tau_z)} \frac{2+r}{(1+r)^2} \quad \text{and} \quad a_2 = a_2(z,r) = \frac{P(\tau_z - \tau_z r)}{2Q(\tau_z)} \frac{2-r}{1-r} \tag{2.6}$$

with $P(t) = \mathbb{E}\{X^2 I(|X| \le t)\}$ and $Q(t) = \mathbb{E}\{\psi_t^2(X)\}$.

Remark 2.1. Here we give some direct implications of Theorem 2.1.

(i) Let $z = z_n \ge 1$ satisfy $z = o(n)$ and $z \to \infty$ as $n \to \infty$. By Proposition 2.2, $\tau_z \to \infty$ and $\tau_z \sim v_2\sqrt{n/z}$, which implies $P(\tau_z) \to v_2^2$ and $Q(\tau_z) \to v_2^2$ as $n \to \infty$.

(ii) With $r = 1/2$ and $z = \log^\kappa(n)$ for some $\kappa \ge 1$ in (2.5), the constants $a_1 = a_1(z, 1/2)$ and $a_2 = a_2(z, 1/2)$ satisfy $a_1 \to 5/9$ and $a_2 \to 3/2$ as $n \to \infty$. The resulting $\widehat{\tau}_z$ satisfies that with probability approaching one, $\tau_z/2 \le \widehat{\tau}_z \le 3\tau_z/2$.

We conclude this section with a uniform deviation bound for m_τ. Uniformity of the rate over a neighborhood of the optimal tuning scale requires an additional $\log(n)$-factor. As a result, we show that the data-driven estimator $m_{\widehat{\tau}_z}$ is tightly concentrated around the mean with high probability.

Theorem 2.2. For $z \ge 1$, let $\tau_z^* = v_2\sqrt{n/z}$. Then with probability at least $1 - 2ne^{-z}$,

$$\sup_{\tau_z^*/2 \le \tau \le 3\tau_z^*/2} |m_\tau - \mu| \le 4v_2(z/n)^{1/2} + v_2 n^{-1/2}. \tag{2.7}$$

Therefore, let $z = 2\log(n)$ and $\widehat{\tau}_z$ be the solution to (2.2), we obtain the following concentration inequality for the mean estimator $m_{\widehat{\tau}_z}$ given in (2.3).

Corollary 2.1. With probability at least $1-c_1n^{-c_2}$ for all sufficiently large n, we have

$$|m_{\widehat{\tau}_z}-\mu| \le 4v_2\sqrt{2\log(n)/n}+v_2n^{-1/2}, \tag{2.8}$$

where $c_1, c_2 > 0$ are absolute constants.

2.2 Adaptive Huber estimator

For the truncation method, even with the theoretically desirable tuning parameter $\tau = v_2\sqrt{n/z}$, the deviation of the resulting estimator only scales with v_2 rather than the standard deviation σ. The optimal deviation, which is enjoyed by the sample mean with sub-Gaussian data, is of order $\sigma\sqrt{z/n}$. To achieve such an optimal order, Fan, Li and Wang (2017) modified Huber's method to construct an estimator that exhibits fast (sub-Gaussian type) concentration under finite variance condition.

The Huber loss in (1.1) is continuously differentiable with $\ell'_\tau(x) = \psi_\tau(x)$, where $\psi_\tau(\cdot)$ is defined in (2.1). The Huber's estimator is obtained as $\widehat{\mu}_\tau = \mathrm{argmin}_{\theta\in\mathbb{R}}\sum_{i=1}^n \ell_\tau(X_i-\theta)$, or equivalently, $\widehat{\mu}_\tau$ is the unique solution to

$$0=\sum_{i=1}^n \psi_\tau(X_i-\theta)=\sum_{i=1}^n \min(|X_i-\theta|,\tau)\,\mathrm{sign}(X_i-\theta). \tag{2.9}$$

We refer to Catoni (2012) for a general class of robust mean estimators. The following result from Theorem 5 in Fan, Li and Wang (2017) shows the exponential-type concentration of $\widehat{\mu}_\tau$ when τ is properly calibrated.

Proposition 2.4. Let $z > 0$ and $v \ge \sigma$. Provided $n \ge 8z$, $\widehat{\mu}_\tau$ with $\tau = v\sqrt{n/z}$ satisfies the bound $|\widehat{\mu}_\tau-\mu| \le 4v\sqrt{z/n}$ with probability at least $1-2e^{-z}$.

Proposition 2.4 indicates that a theoretically desirable tuning parameter for the Huber estimator is $\tau \sim \sigma\sqrt{n/z}$. Motivated by the data-driven approach proposed in Section 2.1, we consider the following modification of (2.4):

$$\frac{\mathbb{E}\{\psi_\tau^2(X-\mu)\}}{\tau^2} = \frac{\mathbb{E}[\min\{(X-\mu)^2, \tau^2\}]}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{2.10}$$

According to Proposition 2.2, provided $0 < z < n\mathbb{P}(X \neq \mu)$, (2.10) admits a unique solution $\tau_{z,\mu}$, which satisfies $\sqrt{\mathbb{E}\left[\min\{(X-\mu)^2, \bar{q}_{z/n}\}\right]}\sqrt{n/z} \leq \tau_{z,\mu} \leq \sigma\sqrt{n/z}$, where $\bar{q}_\alpha = \inf\{t : \mathbb{P}(|X-\mu| > t) \leq \alpha\}$. From a large sample perspective, if $z = z_n$ satisfies $z \to \infty$ and $z = o(n)$, then $\tau_{z,\mu} \to \infty$ and $\tau_{z,\mu} \sim \sigma\sqrt{n/z}$ as $n \to \infty$.

In light of (2.9) and (2.10), a clearly motivated data-driven estimate of μ can be obtained by solving the following system of equations:

$$\begin{cases} f_1(\theta,\tau) := \sum_{i=1}^n \psi_\tau(X_i - \theta) = 0, \\ f_2(\theta,\tau) := n^{-1}\sum_{i=1}^n \min\{(X_i-\theta)^2, \tau^2\}/\tau^2 - n^{-1}z = 0, \end{cases} \quad \theta \in \mathbb{R}, \tau > 0. \tag{2.11}$$

Observe that for any given $\tau > 0$, $f_1(\cdot,\tau) = 0$ always admits a unique solution, and for any given θ, $f_2(\theta,\cdot) = 0$ has a unique solution provided $z < \sum_{i=1}^n I(X_i \neq \theta)$. With initial values $\theta^{(0)} = \bar{X}_n$ and $\tau^{(0)} = \widehat{\sigma}_n\sqrt{n/z}$ where $\widehat{\sigma}_n^2$ denotes the sample variance, we can solve (2.11) successively by computing a sequence of solutions $\{(\theta^{(k)}, \tau^{(k)})\}_{k\geq 1}$ satisfying $f_2(\theta^{(k-1)}, \tau^{(k)}) = 0$ and $f_1(\theta^{(k)}, \tau^{(k)}) = 0$ for $k \geq 1$. For a predetermined tolerance ϵ, the algorithm terminates within the ℓ-th iteration when $\max\{|\theta^{(\ell)} - \theta^{(\ell-1)}|, |\tau^{(\ell)} - \tau^{(\ell-1)}|\} \leq \epsilon$ and uses $\theta^{(\ell)}$ as our robust estimator of μ.

In the case of $z = 1$, we see that the algorithm stops in the first iteration and delivers the solution $\bar{X}_n$. According to the results in Section 2.1, for fixed $z \geq 1$, there is no net improvement in terms of tail robustness; instead, we should let $z = z_n$ slowly grow with the sample size to gain tail robustness without introducing extra bias. Specifically, we choose $z = \log(n)$ throughout the numerical experiments in this paper.

Remark 2.2. The proposed estimator is obtained by iteratively solving (2.11), which mimics (1.6) in Bickel (1975) and can be viewed as a variant of (6.28) and (6.29) in Huber and Ronchetti (2009) for joint location and scale estimation. The estimator in Bickel (1975) solves the equation $\sum_{i=1}^n \psi_{\widehat{\sigma}}(X_i - \theta) = 0$, where $\widehat{\sigma}$ is chosen independently as the normalized interquartile range $\widehat{\sigma}^{(1)} = \{X_{(n-[n/4]+1)} - X_{([n/4])}\}/2\Phi^{-1}(3/4)$ or the symmetrized interquartile range $\widehat{\sigma}^{(2)} = \text{median}\{|X_i - m|\}/\Phi^{-1}(3/4)$, where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics and m is the sample median. The consistency of $\widehat{\sigma}^{(1)}$ or $\widehat{\sigma}^{(2)}$ is established under the symmetry assumption of X, but remains unclear for general distributions. On the other hand, similar to Bickel (1975), our proposed estimators of θ and τ are also location and scale equivariant (see Sections S1.7 and S1.8 in the supplementary files).

Unlike this classical approach, we waive the symmetry requirement by allowing the robustification parameter to diverge to reduce the bias induced by the Huber loss when the distribution is asymmetric. Another difference is that Bickel's proposal is a two-step method that estimates the scale and location separately, whereas our procedure estimates μ and calibrates τ simultaneously by solving a system of equations. In fact, as a direct extension of the idea in Section 2.1, we may also tune τ independently from estimation by solving $\binom{n}{2}^{-1}\sum_{1\le i<j\le n}\tau^{-2}\min\{(X_i-X_j)^2/2,\tau^2\}=zn^{-1}$ for $z>0$. Let X' be an independent copy of X. Then the population version of this equation is $\mathbb{E}\left[\min\{(X-X')^2/2,\tau^2\}\right]\tau^{-2}=z/n$, whose solution is unique under mild conditions and scales as $\sigma\sqrt{n/z}$.

Remark 2.3. In this paper, we assume the finite variance of errors. For more subtle scenarios with finite $(1+\delta)$th moment and $0<\delta<1$, the phase transition phenomenon discovered by Devroye et al. (2016) and Sun, Zhou and Fan (2020) suggests that the Huber's M-estimator no longer admits sub-Gaussian type deviation bounds. Developing the corresponding data-driven principle for tuning Huber's method when $\delta<1$ is nontrivial and left as topic for future investigation.

3. Robust data-adaptive linear regression

In this section, we extend the proposed data-driven method for robust mean estimation to regression problems. Consider the linear regression model

$$Y_i=\beta_0^*+\boldsymbol{X}_i^\intercal\boldsymbol{\beta}^*+\varepsilon_i,\quad i=1,\ldots,n, \tag{3.1}$$

where Y_i's represent response variables, $\boldsymbol{X}_i$'s are d-dimensional vectors of covariates, β_0^* and $\boldsymbol{\beta}^*\in\mathbb{R}^d$ denote the intercept and vector of regression coefficients, respectively, and $\varepsilon_1,\ldots,\varepsilon_n$ are independent regression errors with zero mean and finite variance. For ease of presentation, we write $\boldsymbol{Z}_i=(1,\boldsymbol{X}_i^\intercal)^\intercal$ and $\boldsymbol{\theta}^*=(\beta_0^*,\boldsymbol{\beta}^{*\intercal})^\intercal$. The goal is to estimate $\boldsymbol{\theta}^*$ from observed data $\{(Y_i,\boldsymbol{X}_i)\}_{i=1}^n$.

3.1 Adaptive Huber regression in low dimensions

We start with the low-dimensional regime where $d \ll n$. In the presence of heavy-tailed errors, finite sample properties of the least squares method are suboptimal both theoretically and empirically. Under such heavy-tailed models, we refer to Audibert and Catoni (2011) and Sun, Zhou and Fan (2020) for non-asymptotic analysis of Huber-type robust regressions; the former focused on the excess risk bounds and the latter provided deviation bounds for the estimator along with non-asymptotic Bahadur representations.

Given $\tau > 0$, Huber's M-estimator is defined as

$$\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau^\intercal)^\intercal \in \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}} \sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}), \tag{3.2}$$

where $\ell_\tau(\cdot)$ is given in (1.1). By the convexity of Huber loss, the solution to (3.2) is uniquely determined via the first-order condition: $\sum_{i=1}^n \psi_\tau(Y_i - \boldsymbol{Z}_i^\intercal \widehat{\boldsymbol{\theta}}_\tau)\boldsymbol{Z}_i = \mathbf{0}$. Most desirable features of Huber's method are established under the assumption that the error distribution is symmetric around zero. In the absence of symmetry, the bias induced by the Huber loss becomes non-negligible. To make this statement precise, note that $\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau^\intercal)^\intercal$ is a natural M-estimator of

$$\boldsymbol{\theta}_\tau^* = (\beta_{0,\tau}^*, \boldsymbol{\beta}_\tau^{*\intercal})^\intercal = \mathop{\mathrm{argmin}}_{(\beta_0,\boldsymbol{\beta}^\intercal)^\intercal\in\mathbb{R}^{d+1}} \sum_{i=1}^n \mathbb{E}\{\ell_\tau(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})\}, \tag{3.3}$$

whereas the true parameters β_0^* and $\boldsymbol{\beta}^*$ are identified as $\mathrm{argmin}_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^n \mathbb{E}\{(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2\}$. For any fixed $\tau > 0$, though $\widehat{\beta}_{0,\tau}$ and $\widehat{\boldsymbol{\beta}}_\tau$ are robust estimates of $\beta_{0,\tau}^*$ and $\boldsymbol{\beta}_\tau^*$, $(\beta_{0,\tau}^*, \boldsymbol{\beta}_\tau^*)$ differs from $(\beta_0^*, \boldsymbol{\beta}^*)$ in general. The following proposition provides an explicit bound on the bias, complementing the results in Section 4.9.2 in Maronna et al. (2018).

Proposition 3.1. Assume that ε and $\boldsymbol{X}$ are independent, and that the function $\alpha \mapsto \mathbb{E}\{\ell_\tau(\varepsilon - \alpha)\}$ has a unique minimizer $\alpha_\tau = \operatorname{argmin}_{\alpha \in \mathbb{R}} \mathbb{E}\{\ell_\tau(\varepsilon - \alpha)\}$, which satisfies

$$\mathbb{P}(|\varepsilon - \alpha_\tau| \le \tau) > 0. \tag{3.4}$$

Assume further that $\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$ is positive definite. Then we have $\beta^*_{0,\tau} = \beta^*_0 + \alpha_\tau$ and $\boldsymbol{\beta}^*_\tau = \boldsymbol{\beta}^*$. Moreover, α_τ with $\tau > \sigma$ satisfies the bound

$$|\alpha_\tau| \le \frac{\sigma^2 - \mathbb{E}\{\psi^2_\tau(\varepsilon)\}}{1 - \tau^{-2}\sigma^2} \frac{1}{\tau}. \tag{3.5}$$

Note also that the Huber loss minimization is equivalent to the penalized least squares problem (She and Owen, 2011), $(\widehat{\boldsymbol{\mu}}_\tau, \widehat{\boldsymbol{\theta}}_\tau) = \operatorname{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{d+1}} \{\frac{1}{2} \sum_{i=1}^n (Y_i - \mu_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2 + \tau \sum_{i=1}^n |\mu_i|\}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\intercal$ and $\widehat{\boldsymbol{\theta}}_\tau$ here coincides with that in (3.2). This loss function can be written as $\sum_{i=1}^n (Y_i - \mu_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2/2 + \tau \sum_{i=1}^n |\mu_i|$. This explains from a different perspective that the bias arises only at the intercept. The larger the τ is, the sparser the $\widehat{\boldsymbol{\mu}}_\tau$ is and therefore the smaller the estimation bias is.

The message delivered by Proposition 3.1 draws attention to intercept estimation, a problem of independent interest that needs to be treated with greater caution. If the distribution of ε is asymmetric, α_τ is typically non-zero for any $\tau > 0$; the smaller the τ is, the larger the bias becomes and so is the prediction error. To balance bias and tail robustness, in the following we propose two modifications, one-step and two-step, of the Huber's method that are robust against heavy-tailed and asymmetric errors and meanwhile maintain high efficiency for the normal data.

3.1.1 One-step method

As noted in Zhou et al. (2018), there is an inherent bias-robustness trade-off in the choice of τ, which should adapt to the sample size, dimension and the variance of noise. Theorem 3.1 below fine-tunes this statement. To begin with, we impose the following moment conditions.

Condition 3.1. The covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are *i.i.d.* random vectors from $\boldsymbol{X}$. There exists $A_0 > 0$ such that for any $\boldsymbol{u} \in \mathbb{R}^{d+1}$ and $t \in \mathbb{R}$, $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{z} \rangle| \geq A_0 \|\boldsymbol{u}\|_2 \cdot t) \leq e^{-t}$, where $\boldsymbol{z} = \mathbf{S}^{-1/2}\boldsymbol{Z}$ and $\mathbf{S} = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^{\intercal})$ is positive definite. The regression errors ε_i are independent and satisfy $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i) \leq \sigma^2$ almost surely.

Theorem 3.1. Assume Condition 3.1 holds. For any $z > 0$ and $v \geq \sigma$, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ in (3.2) with $\tau = v\sqrt{n/(d+z)}$ satisfies the bound $\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)\|_2 \leq c_1 v \sqrt{(d+z)/n}$ with probability at least $1 - 2e^{-z}$ provided $n \geq c_2(d+z)$, where $c_1, c_2 > 0$ are constants depending only on A_0.

This theorem establishes a sub-Gaussian concentration bound for $\widehat{\boldsymbol{\theta}}_\tau$ under the optimal sampling size scaling. Compared with Theorem 2.1 in Zhou et al. (2018), there are two technical improvements: first, the moment condition on the random predictor is relaxed from sub-Gaussian to sub-exponential; and secondly, the sample size requirement is improved to $n \gtrsim d$, which is in line with the classical asymptotic consistency result that requires $d = o(n)$. To achieve a sub-Gaussian performance under the finite variance condition, the key observation is that the robustification parameter

τ should adapt to the sample size, dimension, variance of noise and confidence level for optimal trade-off between bias and robustness. Extending our proposal for mean estimation, for $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ and $\tau > 0$, we estimate $\boldsymbol{\theta}^*$ and calibrate τ simultaneously by solving the system of equations

$$\begin{cases} g_1(\boldsymbol{\theta}, \tau) := \sum_{i=1}^n \psi_\tau(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}) \boldsymbol{Z}_i = \boldsymbol{0}, \\ g_2(\boldsymbol{\theta}, \tau) := (\tau^2 n)^{-1} \sum_{i=1}^n \min\{(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2, \tau^2\} - n^{-1}(d+z) = 0. \end{cases} \tag{3.6}$$

With initial values $\boldsymbol{\theta}^{(0)} := \widehat{\boldsymbol{\theta}}_{\rm ols} = (\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\intercal)^{-1} \sum_{i=1}^n Y_i \boldsymbol{Z}_i$ and $\tau^{(0)} = \widehat{\sigma}_n \sqrt{n/(d+z)}$ where $\widehat{\sigma}_n^2 = (1/n)\sum_{i=1}^n (Y_i - \boldsymbol{Z}_i^\intercal \widehat{\boldsymbol{\theta}}_{\rm ols})^2$, for $k \geq 1$, solve $g_2(\boldsymbol{\theta}^{(k-1)}, \tau^{(k)}) = 0$ to obtain compute $\tau^{(k)}$ and then compute $\boldsymbol{\theta}^{(k)}$ as the solution to $g_1(\boldsymbol{\theta}^{(k)}, \tau^{(k)}) = 0$. Iterate until convergence and set $\widehat{\boldsymbol{\theta}}^{\rm I} := \widehat{\boldsymbol{\theta}}_{\widehat{\tau}}$ as our one-step estimator, where $(\widehat{\boldsymbol{\theta}}, \widehat{\tau})$ is the final output.

The main advantage of the proposed adaptive Huber regression over the traditional one with $\tau = 1.345\sigma$ is that the estimation bias with respect to intercept is alleviated. Examining the proof of Proposition 3.1, we find that the bias is of order $1/\tau$ when the second moment is finite, and is quadratic in $1/\tau$ if the third moment is finite. The statistical error, on the other hand, is determined by the ℓ_2-norm of the score function evaluated at $\boldsymbol{\theta}^*$ which is of order $\sigma\sqrt{d/n} + \tau d/n$, see Theorem 3.2 below. The overall error is then optimized at $\tau \asymp \sigma\sqrt{n/d}$. For the normal model, since $\max_{1\leq i\leq n} |\varepsilon_i| \sim \sigma\sqrt{2\log(2n)} \lesssim \sigma\sqrt{n/d}$, the adaptive Huber estimator is almost identical to the least squares estimator. Numerical results in Section 4 provide strong support for the tail-adaptivity of our proposed data-driven Huber regression.

When τ scales as a constant, such as $c\sigma$, the corresponding Huber loss is Lipschitz with bounded score function, and since $\boldsymbol{\beta}^*_\tau = \boldsymbol{\beta}^*$ for any $\tau > 0$, no sacrifice in bias will incur for estimating the slope $\boldsymbol{\beta}^*$. Again, constant c is typically tuned to ensure a given level of asymptotic efficiency. Asymptotic properties of general robust M-estimators have been well studied in the literature; see Avella-Medina and Ronchetti (2015) for a selective overview. The next result further complements Theorem 3.1 by establishing the deviations of the Huber estimator with fixed τ from a non-asymptotic viewpoint.

Theorem 3.2. Suppose Condition 3.1 and the assumptions in Proposition 3.1 hold. Assume further that $\rho_\tau := \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau/2) > 0$. Then, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ in (3.2) satisfies $\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*_\tau)\|_2 \lesssim \rho_\tau^{-1} A_0\{\sigma\sqrt{(d+z)/n} + \tau(d+z)/n\}$ for any $z > 0$ with probability at least $1 - 2e^{-z}$ provided $n \geq c_3(d+z)$, where $c_3 > 0$ is a constant depending only on (A_0, ρ_τ).

3.1.2 Two-step method

Motivated by our bias-robustness analysis and the results of finite sample investigation, we further introduce a two-step procedure that estimates the regression coefficients and intercept successively.

In the first step, we compute the Huber estimator $\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}^{*\intercal}_\tau)^\intercal$ by solving (3.2) with $\tau = c\sigma$. We take $c = 1.345$ to retain the 95% efficiency for the normal model. For σ, it can be estimated simultaneously with $\boldsymbol{\theta}^*$ by solving a system of equations as in Huber's "Proposal 2" (Huber, 1964; Huber and Ronchetti, 2009), or we can fix σ at an initial robust estimate and then optimize over $\boldsymbol{\theta}$ (Hampel et al., 1986). We follow the former route and consider an iterative procedure. Start with an initial estimate $\boldsymbol{\theta}^{(0)}$, at iteration $k = 0, 1, 2, \ldots$, we employ a simple procedure to obtain $\widehat{\sigma}^{(k)}$, based on which to produce update $\boldsymbol{\theta}^{(k+1)}$. This step involves two procedures.

Procedure 1: Scale estimation. Using the current estimate $\boldsymbol{\theta}^{(k)}$, we compute the vector of residuals $\boldsymbol{r}^{(k)} = (r_1^{(k)}, \ldots, r_n^{(k)})^\intercal$ and the robustification parameter $\tau^{(k)} = 1.345\widehat{\sigma}^{(k)}$, where $\widehat{\sigma}^{(k)}$ denotes the median absolute deviation (MAD) estimator median$\{|r_i^{(k)} -$ median$(r_i^{(k)})|\}/\Phi^{-1}(3/4)$.

Procedure 2: Weighted least squares. Compute the $n \times n$ diagonal matrix $\mathbf{W}^{(k)} =$ diag$((1 + w_1^{(k)})^{-1}, \ldots, (1 + w_n^{(k)})^{-1})$, where $w_i^{(k)} = |r_i^{(k)}|/\tau^{(k)} - 1$ if $|r_i^{(k)}| > \tau^{(k)}$ and $w_i^{(k)} = 0$ if $|r_i^{(k)}| \leq \tau^{(k)}$. Then we update $\boldsymbol{\theta}^{(k)}$ to produce $\boldsymbol{\theta}^{(k+1)}$ via weighted least squares, that is,

$$\boldsymbol{\theta}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2}{1 + w_i^{(k)}} = (\mathbf{Z}^\intercal \mathbf{W}^{(k)} \mathbf{Z})^{-1} \mathbf{Z}^\intercal \mathbf{W}^{(k)} \boldsymbol{Y},$$

where $\mathbf{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\intercal \in \mathbb{R}^{n \times (d+1)}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$.

Starting with $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{ols}}$, we repeat the above two procedures until convergence. Denote $\widehat{\boldsymbol{\beta}}^{\mathrm{II}} \in \mathbb{R}^d$ the vector of coefficients estimates extracted from the final solution.

In the second step, observe that $\beta_0^* = \mathbb{E}(\delta_i)$, where $\delta_i = Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* = \beta_0^* + \varepsilon_i$ are the residuals. To estimate β_0^*, defining fitted residuals $\widehat{\delta}_i = Y_i - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}}^{\mathrm{II}}$, we solve the system of equations

$$\begin{cases} f_1(\beta_0, \tau) := (\tau^2 n)^{-1} \sum_{i=1}^n \min\{(\widehat{\delta}_i - \beta_0)^2, \tau^2\} - n^{-1}\log(n) = 0, \\ f_2(\beta_0, \tau) := \sum_{i=1}^n \psi_\tau(\widehat{\delta}_i - \beta_0) = 0, \end{cases} \tag{3.7}$$

in the same way as for solving (2.11) to obtain $\widehat{\beta}_0^{\rm II}$. Then, $\widehat{\boldsymbol{\theta}}^{\rm II} = (\widehat{\beta}_0^{\rm II}, \widehat{\boldsymbol{\beta}}^{\rm II})$ is our two-step estimator of $\boldsymbol{\theta}^*$.

The two-step procedure leverages the fact that, for the asymmetric regression errors with potentially heavy tails, the Huber loss with a fixed τ only introduces bias to the intercept estimation but not to the estimation on the slope coefficients. To alleviate the influence of skewness in the error, in the second step we use the adaptive Huber method with a divergent τ to re-estimate the intercept. The two-step estimator therefore achieves both high degree of tail robustness and unbiasedness.

3.2 Adaptive Huber regression in high dimensions

We now move to the high dimensional setting where $d \gg n$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\intercal \in \mathbb{R}^d$ is sparse with $\|\boldsymbol{\beta}^*\|_0 := \sum_{j=1}^d I(\beta_j^* \neq 0) = s \ll n$. Since the invention of the Lasso (Tibshirani, 1996), a variety of variable selection methods have been developed for finding a small group of response-associated covariates from a large pool. We refer to Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015) for a comprehensive review along this line.

The Lasso estimator is $\widehat{\boldsymbol{\beta}}_{\rm lasso}(\lambda) \in \operatorname{argmin}_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d} \{(2n)^{-1} \sum_{i=1}^n (Y_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1\}$, where $\lambda > 0$ is a regularization parameter. Thinking of the noise variable as being Gaussian, this can be interpreted as a penalized maximum likelihood estimate, where the ℓ_1 penalty encourages sparsity in the estimation. However, least squares fitting is sensitive to the tails of error distributions, particularly for ultra-high dimensional covariates as their spurious correlations with the noise can be large, and therefore is not ideal in the presence of heavy-tailed noise.

Recently, Fan, Li and Wang (2017) modified Huber's procedure to obtain an ℓ_1-regularized robust estimator admitting the desirable concentration bound under finite variance condition on the regression errors. According to the discussions in Section 3.1, the intercept, albeit being often ignored in the literature, plays an important role in the study of robust methods. To take into account the effect of intercept, we consider the regularized Huber estimator of the form

$$\widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau,\lambda) \in \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta}^\intercal)^\intercal\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \big\{\mathcal{L}_\tau(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\beta}\|_1\big\}, \tag{3.8}$$

where $\mathcal{L}_\tau(\boldsymbol{\theta}) := (1/n)\sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n \ell_\tau(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal\boldsymbol{\beta})$, τ and λ are the robustification and regularization parameters, respectively.

Given ε_i with finite variance, Theorem 3.3 below reveals that the ℓ_1-regularized Huber regression with properly tuned (τ,λ) gives rise to consistent estimators with ℓ_1- and ℓ_2-errors scaling as $s\sqrt{\log(d)/n}$ and $\sqrt{s\log(d)/n}$, respectively, under the sample size scaling $n \gtrsim s\log(d)$. These rates are exactly the minimax rates enjoyed by the Lasso with sub-Gaussian errors.

Theorem 3.3. Assume Condition 3.1 holds and denote by $\underline{\lambda}_{\mathbf{S}}$ the minimal eigenvalue of $\mathbf{S}$. Assume further that the unknown $\boldsymbol{\beta}^*$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0$. Let $\sigma_{jj} = \mathbb{E}(X_j^2)$ for $j = 1,\ldots,d$. Then the estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau,\lambda)$ given in (3.8) with $\tau = \sigma\sqrt{n/\log(d)}$ and

λ scaling with $A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2}\sigma\sqrt{\log(d)/n}$ satisfies

$$\|\widehat{\boldsymbol{\theta}}_{\rm H}(\tau,\lambda)-\boldsymbol{\theta}^*\|_2 \lesssim \frac{s^{1/2}\lambda}{\underline{\lambda}_{\mathbf{S}}} \quad \text{and} \quad \|\widehat{\boldsymbol{\theta}}_{\rm H}(\tau,\lambda)-\boldsymbol{\theta}^*\|_1 \lesssim \frac{s\lambda}{\underline{\lambda}_{\mathbf{S}}} \tag{3.9}$$

with probability at least $1-3d^{-1}$ as long as $n \ge c_1 s\log(d)$, where $c_1>0$ is a constant depending only on $(A_0, \max_{1\le j\le d}\sigma_{jj}, \underline{\lambda}_{\mathbf{S}})$.

Theorem 3.3 above complements Theorem 3 in Fan, Li and Wang (2017). The latter provides convergence rates of ℓ_1-penalized Huber's M-estimator under the weakly sparse setting that $\|\boldsymbol{\beta}^*\|_q \le R_q$ for some $0<q\le 1$. Their results, however, do not directly apply to the sparse regime where $q=0$. Moreover, the sub-Gaussian condition imposed in Fan, Li and Wang (2017) is now relaxed to the sub-exponential condition.

Remark 3.1. The main purpose of using the Huber loss for data fitting is to gain robustness against outliers from either contamination models (Huber, 1973) or heavy-tailed models considered in this paper. For other purposes, different loss functions have been proposed to replace the squared loss, such as the nonconvex Tukey and Cauchy losses, the quantile loss and the asymmetric quadratic loss, among others. We refer to Owen (2007), Loh and Wainwright (2015), Loh (2017), Zhou et al. (2018), Mei, Bai and Montanari (2018), Alquier, Cottet and Lecué (2019), and Pan, Sun and Zhou (2019) for more discussions on the regularized M-estimator with different loss functions.

In practice, it is computationally demanding to choose the optimal values of τ and λ by a two-dimensional grid search using cross-validation. We consider the following procedure that estimates $\boldsymbol{\theta}^*$ and tunes τ simultaneously. Given a random sample of size n, we use cross-validated Lasso as an initialization $\widehat{\boldsymbol{\theta}}^{(0)}$. At iteration $k = 1, 2, \ldots$, using the previous estimate $\widehat{\boldsymbol{\theta}}^{(k-1)}$ we compute $\tau^{(k)}$ as the solution to

$$\frac{1}{\{n - \widehat{s}^{(k-1)}\}} \sum_{i=1}^{n} \frac{\min\{(Y_i - \boldsymbol{Z}_i^{\intercal}\widehat{\boldsymbol{\theta}}^{(k-1)})^2, \tau^2\}}{\tau^2} = \frac{\log(nd)}{n}, \tag{3.10}$$

where $\widehat{s}^{(k-1)} = \|\widehat{\boldsymbol{\beta}}^{(k-1)}\|_0$. Next, take $\tau = \tau^{(k)}$ and compute $\widehat{\boldsymbol{\theta}}^{(k)}$ by solving

$$\min_{\boldsymbol{\theta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_\tau(Y_i - \boldsymbol{Z}_i^{\intercal}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{3.11}$$

where λ is chosen via cross-validation. Repeat the above two steps until convergence or until the maximum number of iterations is reached.

To implement the data-driven Huber regression in high dimensions, again, starting with some initial guess we iteratively solve (3.10) and (3.11). For the convex optimization problems in (3.11), the minimizer satisfies the Karush-Kuhn-Tucker conditions, and therefore can be found by solving the following system of nonsmooth equations:

$$\begin{cases} -n^{-1}\sum_i \psi_\tau(Y_i - \boldsymbol{Z}_i^{\intercal}\widehat{\boldsymbol{\theta}}) = 0, \\ -n^{-1}\sum_i \psi_\tau(Y_i - \boldsymbol{Z}_i^{\intercal}\widehat{\boldsymbol{\theta}})X_{ij} + \lambda\widehat{\eta}_j = 0, \quad j = 1, \ldots, d \\ \widehat{\beta}_j - S(\widehat{\beta}_j + \widehat{\eta}_j) = 0, \quad j = 1, \ldots, d \end{cases} \tag{3.12}$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^{\intercal})^{\intercal} \in \mathbb{R}^{d+1}$ with $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)^{\intercal}$, $\widehat{\eta}_j \in \partial|\widehat{\beta}_j|$ and $S(z) = \text{sign}(z)(|z| - 1)_+$ is the soft-thresholding operator. Instead of directly applying the Semismooth Newton Algorithm (SNA) to the entire system of equations, we adapt the Semismooth Newton Coordinate Descent (SNCD) algorithm proposed by Yi and Huang (2017), which combines SNA with cyclic coordinate descent to solve (3.12). More specifically, in

SNCD we divide (3.12) into two parts in order to avoid cumbersome matrix operations as in solving the entire system. In a cyclic fashion, we update the intercept only using the first equation and update the coefficients with its subgradients using the last two equations, therefore reduce the computational cost from $O(nd^2)$ to $O(nd)$ at each iteration. The gain in the computational scalability and efficiency is substantial for large d. After obtaining a solution path of (3.11), we employ the cross-validation method to select λ and then the associated $\widehat{\boldsymbol{\theta}}^{(k)}$.

Remark 3.2. The above regularized data-adaptive Huber regression method is a direct extension of the one-step method proposed in Section 3.1 to high dimensions. Also, note that Proposition 3.1 holds in high dimensions as long as the population Gram matrix $\mathbf{S}$ is positive definite. Therefore, to further reduce the estimation bias of intercept, we suggest the two-step procedure that estimates the regression coefficients using the standard regularized Huber regression and then estimates the intercept by applying the adaptive-Huber method to fitted residuals as in (3.7). Section 4.3 provides numerical studies of both the one- and two-step regularized adaptive Huber estimators.

4. Empirical analysis

In this section, we examine numerically the finite sample performance of the proposed data-adaptive Huber (DA-Huber) methods for mean estimation and linear regressions. In the supplementary files, using three real data sets, we also demonstrate the desirable performance of the proposed DA-Huber methods in terms of prediction accuracy.

We consider the following four distribution settings to investigate the robustness and efficiency of the proposed method in a wide variety of scenarios.

(1) Normal distribution $\mathcal{N}(0, \sigma^2)$ with mean zero and variance $\sigma^2 > 0$;

(2) Skewed generalized t distribution (Theodossiou, 1998) $\mathsf{sgt}(\mu, \sigma^2, \lambda, p, q)$, where mean $\mu = 0$, variance $\sigma^2 = q/(q-2)$ with $q > 2$, shape $p = 2$ and skewness $\lambda = 0.75$;

(3) Lognormal distribution $\mathsf{LN}(\mu, \sigma)$ with $\mu = 0$ and $\sigma > 0$; and

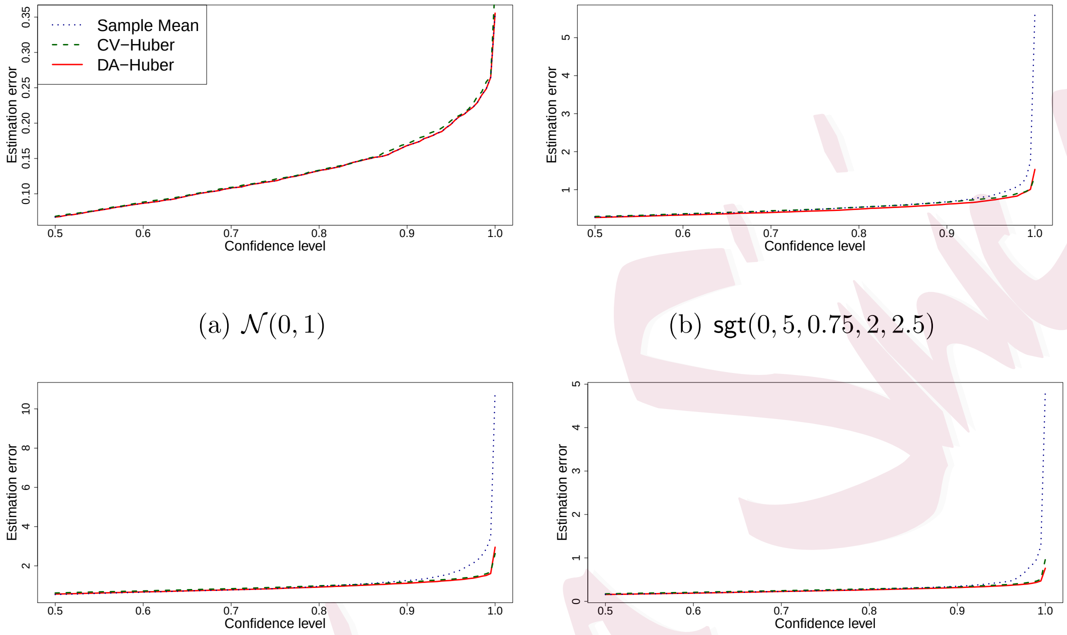
(4) Pareto distribution $\mathsf{Par}(x_m, \alpha)$ with scale $x_m = 1$ and shape $\alpha > 0$.

All above settings but (1) are skewed and might be very heavy-tailed for some choice of the distribution parameters, such as $\alpha < 2$ for the Pareto distribution.

4.1 Mean estimation

For each setting, we generate an independent sample of size $n = 100$ and compute three mean estimators: the sample mean, the Huber estimator with τ chosen via five-fold cross-validation (CV-Huber), and the proposed DA-Huber mean estimator. Figure 1 displays the α-quantile of the estimation error with α ranging from 0.5 to 1 based on 2000 simulations. Figure S1 in the supplementary files reports the boxplots of the estimation error. The DA-Huber estimator and sample mean perform almost identically for the normal data. For the heavy-tailed skewed distributions, the deviation of the sample mean from the population mean grows rapidly with the confidence level, which is in striking contrast to the CV- and DA-Huber estimators.

In Figure 2, we examine the 99%-quantile of the estimation error versus a distribu-



(a) $\mathcal{N}(0,1)$

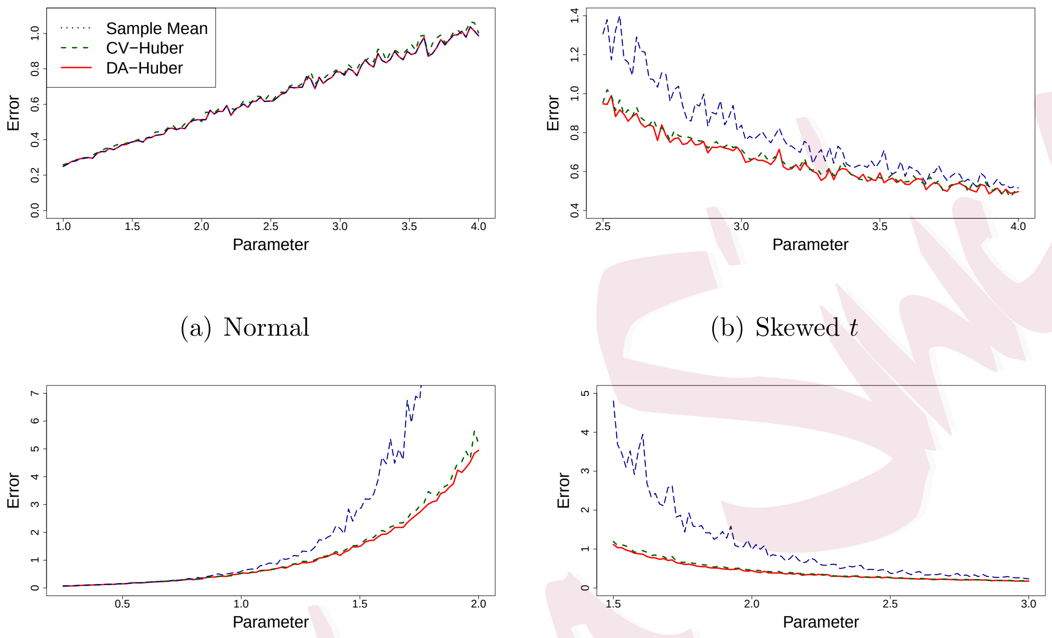
(b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$

(d) $\mathsf{Par}(1,2)$

Figure 1: Estimation error versus confidence level for the sample mean, CV-Huber and DA-Huber estimators based on 2000 simulations.

tion parameter measuring the tail behavior and the skewness. Namely, for normal data we let σ vary between 1 and 4; for skewed generalized t distributions, we increase the shape parameter q from 2.5 to 4; for Lognormal and Pareto distributions, the shape parameters σ and α vary from 0.25 to 2 and 1.5 to 3, respectively. The Huber-type estimators show substantial improvement in deviations from the population mean as the distribution tends to have heavier tails and become more skewed. In summary, the most attractive feature of our method is its adaptivity: (i) it is as efficient as the sample mean for the normal model and is more robust for asymmetric and/or heavy-tailed



(a) Normal

(b) Skewed t

(c) Lognormal

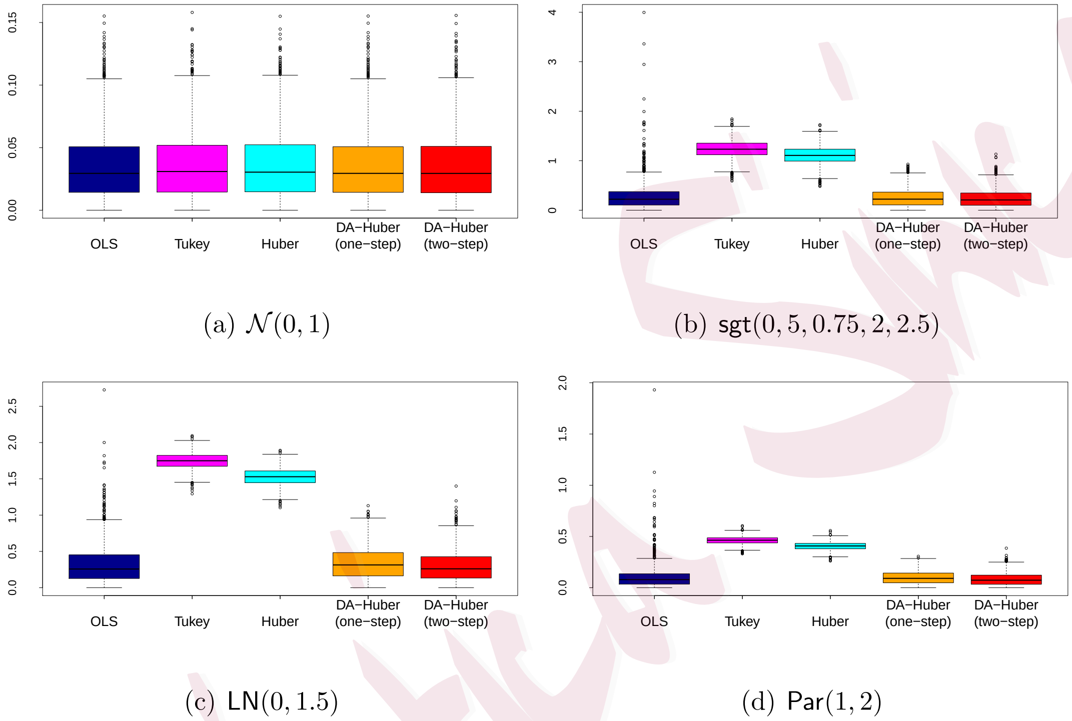
(d) Pareto

Figure 2: Empirical 99%-quantile of the estimation error versus parameter measuring tails and skewness for the sample mean, CV-Huber and DA-Huber estimators.

data; (ii) it performs as good as the cross-validation but with much less computational cost. The latter is particularly important for large-scale inference with a myriad of parameters to be estimated simultaneously.

4.2 Linear regression

We generate data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from linear model (3.1) with $n = 500$ and $d = 5$. The intercept and vector of regression coefficients are taken to be $\beta_0 = 5$ and $\boldsymbol{\beta}^* = (1, -1, 1, -1, 1)^\intercal$, respectively. The covariates $\boldsymbol{X}_i$ are *i.i.d.* random vectors that consist of independent coordinates from a uniform distribution $\mathsf{Unif}(-1.5, 1.5)$.



(a) $\mathcal{N}(0, 1)$

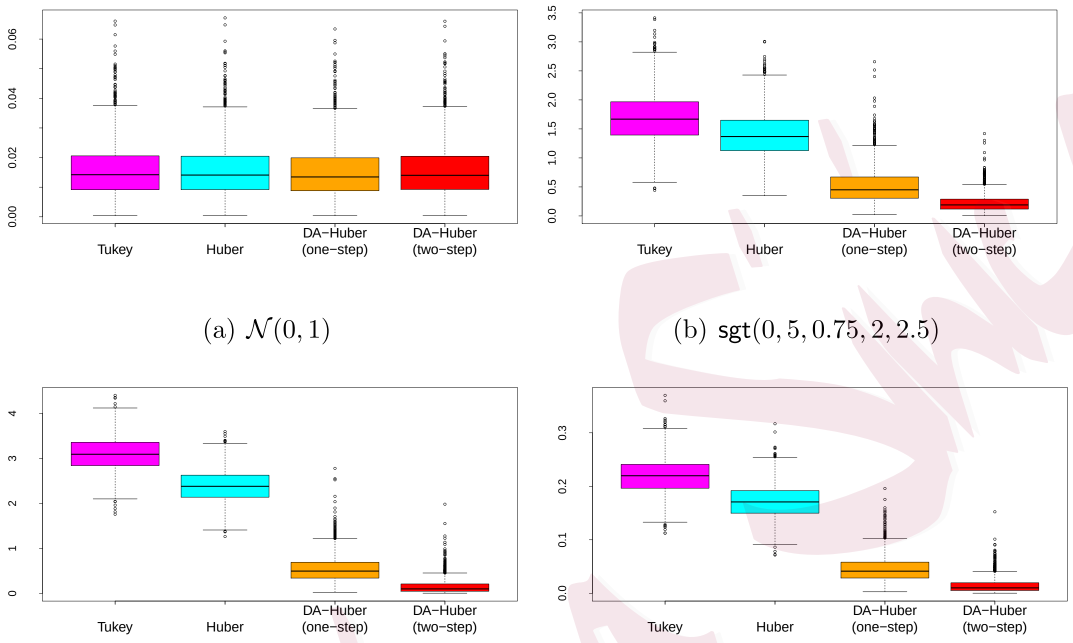
(b) $\mathsf{sgt}(0, 5, 0.75, 2, 2.5)$

(c) $\mathsf{LN}(0, 1.5)$

(d) $\mathsf{Par}(1, 2)$

Figure 3: Estimation errors of intercept under different settings.

We compare the DA-Huber regression estimator with the ordinary least squares (OLS) estimator and classical robust M-estimators with Huber loss $\ell_\tau(\cdot)$ as in (1.1) and Tukey's biweight loss $\ell_\tau^{\mathrm{T}}(x) = \{1 - (1 - x^2/\tau^2)^3\}\mathbb{I}(|x| \le \tau) + \mathbb{I}(|x| > \tau)$. The tuning parameter τ in $\ell_\tau^{\mathrm{T}}(\cdot)$ and $\ell_\tau(\cdot)$ is taken to be 4.685 and 1.345, respectively, according to the 95% efficiency rule. We carry out 1000 Monte Carlo simulations to: (1) evaluate the overall performance of the DA-Huber methods comparing with three competitors, OLS, Tukey, and Huber; see Figures 3 and 4, and (2) explore the robustness of different methods with varying degrees of heavy-tailedness and skewness; see Figures 5 and 6.



(a) $\mathcal{N}(0,1)$

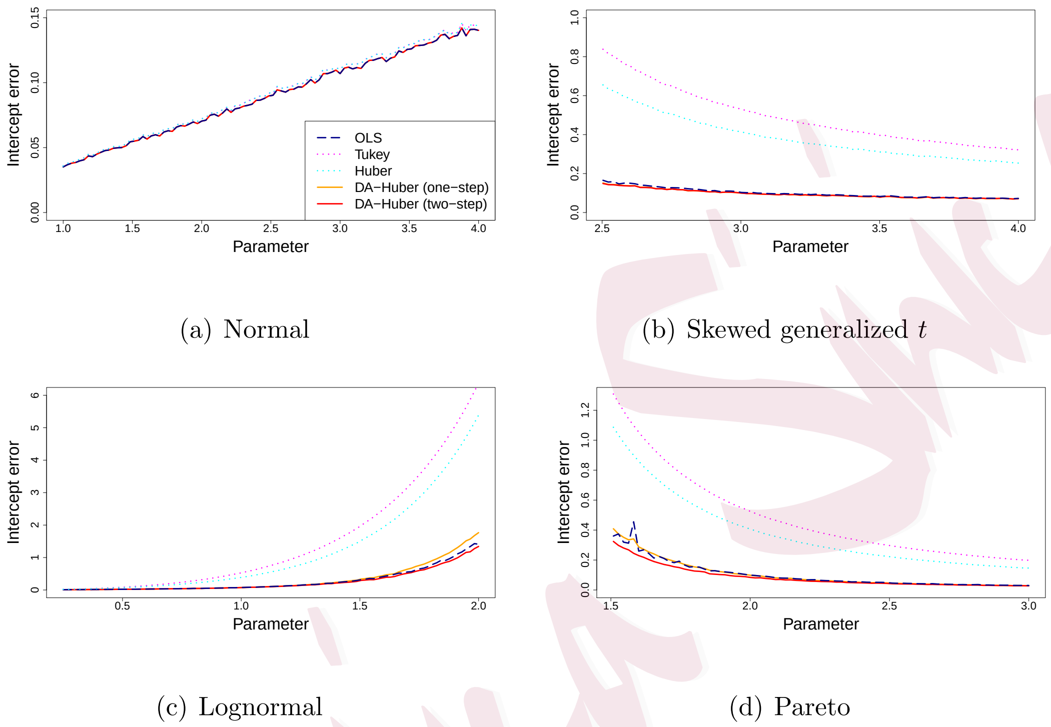
(b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$

(d) $\mathsf{Par}(1,2)$

Figure 4: Total ℓ_2-errors under different settings.

Figures 3 and 4 display the boxplots of the estimation error of intercept $|\widehat{\beta}_0 - \beta_0^*|$ and the total ℓ_2-error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$, respectively, for a fixed distribution parameter as did in Section 4.1. Both the one-step and two-step DA-Huber estimators outperform the other methods across all examples. For estimating the intercept, the DA-Huber rectifies the non-negligible bias in the traditional robust M-estimator, as predicted by theory. In the normal case, the DA-Huber estimator performs almost identically to the OLS and is therefore highly efficient. The ℓ_2-error of OLS tends to spread out (due to outliers) and thus is not reported Figures 5 and 6 illustrate, respectively, the average estimation error of intercept and the total ℓ_2-error versus distribution parameters controlling the



(a) Normal

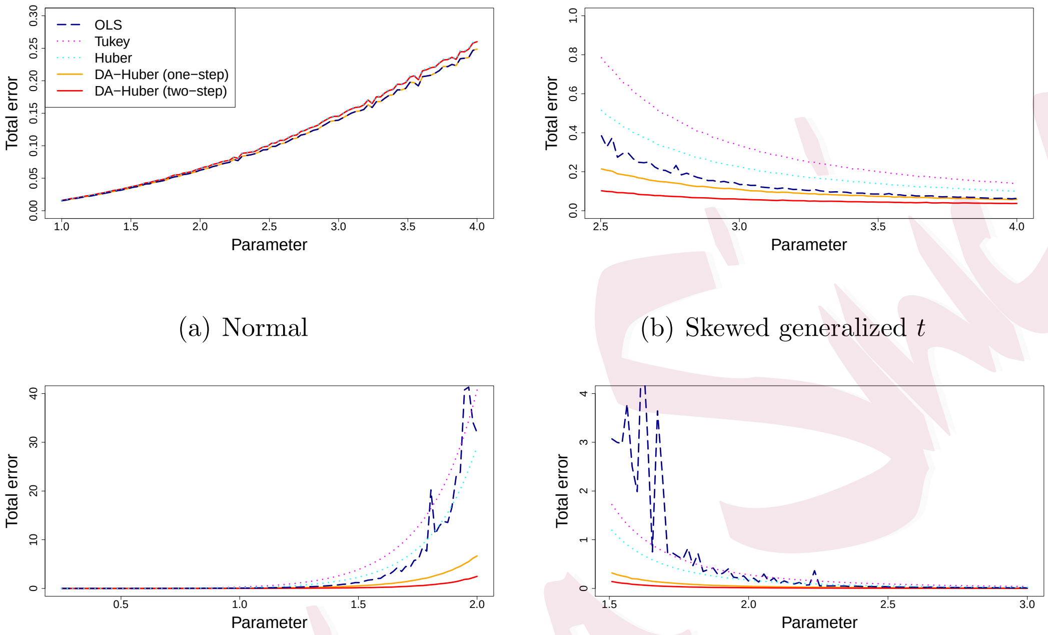
(b) Skewed generalized t

(c) Lognormal

(d) Pareto

Figure 5: Average estimation error of intercept versus distribution parameters controlling tails for the OLS estimator, standard Tukey's and Huber's estimators, and data-adaptive Huber estimators (one-step and two-step).

shape of tails. In the normal case, the one-step DA-Huber and OLS slightly outperform the others; with heavy-tailed and skewed errors, the DA-Huber methods enjoy notable advantage and the two-step approach is the most desirable since it strikes the perfect balance between bias and tail robustness. Overall, the numerical results confirm that the proposed methods have substantial advantages in the presence of asymmetric and heavy-tailed errors, while maintaining high efficiency for the normal model.



(a) Normal

(b) Skewed generalized t

(c) Lognormal

(d) Pareto

Figure 6: Average ℓ_2-errors versus distribution parameters controlling tails for the OLS estimator, standard Tukey's and Huber's estimators, and data-adaptive Huber estimators (one-step and two-step).

4.3 Sparse linear regression

Now we consider the sparse linear regression, $Y_i = \beta_0^* + \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i$ with $i = 1, \ldots, n$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0 \ll n$ and $d \gg n$. In simulations, we take $n = 250$, $d = 1000$ and $s = 20$. We set $\beta_0^* = 3$ and $\boldsymbol{\beta}^* = (3, \ldots, 3, 0, \ldots, 0)^\intercal$, where the first $s = 20$ nonzero entries of $\boldsymbol{\beta}^*$ all equal to 3. As before, the covariates $\boldsymbol{X}_i$ are *i.i.d.* random vectors whose independent coordinates are from $\mathsf{Unif}(-1.5, 1.5)$, and ε_i's follow one of the four distributions: normal, skewed generalized t, Lognormal, and Pareto.

To implement the iterative procedure proposed in Section 3.2, at the k-th iteration, we use the five-fold cross-validation to choose $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ in the optimization programs in (3.11), producing $\widehat{\boldsymbol{\theta}}_1^{(k)}$ and $\widehat{\boldsymbol{\theta}}_2^{(k)}$. We evaluate the proposed regularized DA-Huber estimators by the following measurements: RG, the relative gain of the DA-Huber estimator with respect to the Lasso in terms of ℓ_1- and ℓ_2-errors, $\mathrm{RG_q} = \|\widehat{\boldsymbol{\theta}}_{\mathrm{H}} - \boldsymbol{\theta}\|_q / \|\widehat{\boldsymbol{\theta}}_{\mathrm{lasso}} - \boldsymbol{\theta}\|_q$ with $q = 1, 2$; FP, the number of false positives (selected noise covariates); and FN, the number of false negatives (missed signal covariates).

Table 1 summaries the relative gains of the DA-Huber estimators under ℓ_1- and ℓ_2-errors and the numbers of false positive and false negative discoveries. Across all the four models, both one- and two-step DA-Huber estimators outperform the Lasso with smaller ℓ_1-errors and fewer false positive discoveries, therefore are less greedy in model selection. For the normal model, the proposed robust methods and Lasso perform equally well; while in the presence of heavy-tailed skewed errors, the DA-Huber methods lead to remarkably better outputs in regard of both estimation and model selection. Similar phenomenon can also be observed from Figure S2 in the supplementary files, which displays the empirical distributions of the ℓ_2-errors for all estimators.

5. Summary

In this paper, we have proposed a new principle to choose the robustification parameter adaptively from data for a variety of fundamental statistical problems, includ-

Table 1: RG, FP and FN and their standard errors (in brackets) of the Lasso and DA-Huber estimators under different models. The results are based on 200 simulations.

	Lasso	DA-Huber (one-step)	DA-Huber (two-step)	Lasso	DA-Huber (one-step)	DA-Huber (two-step)
	Normal, $\mathcal{N}(0,1)$			$\mathsf{sgt}(0,5,0.75,2,2.5)$		
$\mathrm{RG}_1 \times 100$	100	93.4 (0.6)	91.4 (0.9)	100	87.5 (1.0)	86.2 (0.9)
$\mathrm{RG}_2 \times 100$	100	100.3 (0.2)	102.7 (0.3)	100	98.3 (0.5)	98.1 (0.5)
FP	87.9 (1.7)	77.6 (1.4)	73.5 (2.0)	86.1 (1.8)	63.1 (1.8)	60.7 (1.5)
FN	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Lognormal, $\mathsf{LN}(0,1.5)$			Pareto, $\mathsf{Par}(1,2)$		
$\mathrm{RG}_1 \times 100$	100	34.7 (0.7)	22.7 (0.5)	100	65.3 (1.1)	41.7 (0.8)
$\mathrm{RG}_2 \times 100$	100	49.5 (1.0)	30.5 (0.7)	100	84.5 (0.9)	51.2 (0.9)
FP	80.8 (2.0)	21.9 (0.6)	26.6 (0.7)	85.1 (1.9)	34.5 (1.6)	44.2 (0.9)
FN	0.26 (0.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

ing mean estimations, linear regression and the sparse regression in high dimensions. Inspired by the censored moment equation approach, the proposed principle is genuinely tuning-free and data-adaptive. It is conceptually different from the traditional practice on selecting the robustification parameter based on cross-validation, which is not only computationally demanding but also lacks of the underpinning mathematical guarantees. The proposed principle is guided by non-asymptotic deviation analysis and paves a unified pathway for choosing robustification parameter for tail-robust estimation and inference. Particularly, the analysis guiding our method can be easily extended to a broader class of robust convex loss functions including the pseudo-Huber loss functions. The key is the global Lipschitz and local quadratic geometry of the loss function $\ell_\tau(x) = \tau^2\ell(x/\tau)$. In light of numerical evidences from both synthetic and real data, our proposal outperforms those widely known procedures in terms of estimation, variable selection, and prediction in the presence of heavy-tailed and skewed errors. Finally, an R package that implements the DA-Huber method can be found at https://github.com/XiaoouPan/tfHuber.

Supplementary Materials

The supplementary materials contain the proofs of all the theoretical results in the main text and additional empirical studies.

Acknowledgement

We thank the editor, the associate editor and two referees for insightful comments that have substantially improved this article. Lili Wang is partially supported by NSFC grant 11701509 and the First Class Discipline of Zhejiang-A (Zhejiang Gongshang University-Statistics). Chao Zheng's research is supported by EPSRC Grant

EP/N031938/1. Wen Zhou's research is partially supported by the Department of Energy. Wen-Xin Zhou's research is supported by NSF Grant DMS-1811376.

References

Alquier, P., Cottet, V. and Lecué, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *The Annals of Statistics*, **47**(4), 2117–2144.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, **39**(5), 2766–2794.

Avella-Medina, M., Battey, H. S., Fan, J. and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, **105**(2), 271–284.

Avella-Medina, M. and Ronchetti, E. (2015). Robust statistics: A selective overview and new directions. *WIREs Computational Statistics,* **7**(6), 372–393.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, **70**(350), 428–434.

Brownlees, C., Joly, E. and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, **43**(6), 2507–2536.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, **48**(4), 1148–1185.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, **1**(2), 223–236.

Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P.J. Bickel, K.Doksum, and J.L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, CA.

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, **44**(6), 2695–2725.

Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *The Annals of Statistics,* **46**(6B), 3481–3509.

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

Hampel F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, **42**(6), 1887–1896.

Hampel F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, **17**(18), 1–40.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics,* **35**(1), 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics,* **1**(5), 799–821.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, Second Edition. Wiley, New York.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics,* **5**, 1015–1053.

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306.*

Loh, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, **45**(2), 866–896.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16**(19), 559–616.

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757.*

Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, **47**(2), 783–794.

Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, **46**(6A), 2747–2774.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, **21**(4), 2308–2335.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics,* **46**(6A), 2871–2903.

Maronna, R. A., Marton, R. D., Yohai, V. J., and Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods (with R).* Wiley, New York.

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

Pan, X., Sun, Q. and Zhou, W.-X. (2019). Nonconvex regularized robust regression with oracle properties in polynomial time. *arXiv preprint arXiv:1907.04027.*

Portnoy, S. and He,X. (2000). A robust journey in the new millennium. *Journal of the American Statistical Association*. **95**, 1331–1335.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4, 16.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, **106**(494), 626–639.

Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association*, in press. **115**(529), 254–265.

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science*, **44**(12), 1650–1661.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**(1), 267–288.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 210–268. Cambridge Univ. Press, Cambridge.

Yi, C. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, **26**(3), 547–557.

Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics*, **46**(5), 1904–1931.

Lili Wang, School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China.

E-mail: (liliwang@zjgsu.edu.cn)

Chao Zheng, Southampton Statistical Sciences Research Institute, and Mathematical Sciences, University of Southampton, SO17 1BJ, UK.

E-mail: (c.zheng5@lancaster.ac.uk)

Wen Zhou, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, USA.

E-mail: (riczw@stat.colostate.edu)

Wen-Xin Zhou, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA.

E-mail: (wez243@ucsd.edu)