Abstract

The robustification parameter, which balances bias and robustness, has played a critical role in the construction of sub-Gaussian estimators for heavy-tailed data. Although it can be tuned by cross-validation, in large-scale statistical problems such as multiple testing and high dimensional covariance matrix estimation, the number of robustification parameters scales with the size of the problem so that cross-validation can be computationally intensive. 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. This procedure is guided by non-asymptotic deviation analysis, and is conceptually different from cross-validation which relies on the mean squared error to assess the fit. We illustrate its promising performance with extensive numerical experiments and real data analysis.

Keywords: Data adaptive; Heavy tails; Huber loss; M-estimator; Tuning parameters
1 Introduction

Data subject to heavy-tailed and/or skewed distributions are frequently observed in many areas, ranging from microarray experiments (Purdom and Holmes, 2005), neuroimaging (Eklund, Nichols and Knutsson, 2016) to finance (Cont, 2001). Rigorously, a random variable $X$ is heavy-tailed if its tail probability $P(|X| > t)$ 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 well known equality that $\mathbb{E}(|X|^k) = k \int_0^\infty t^{k-1} P(|X| > t) \, dt$ 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 their mean. We refer to such observations as distributional outliers. In contrast, data generated from a Gaussian or sub-Gaussian distribution (Vershynin, 2012) are strongly concentrated around their expected value, and the occurrence of even a single outlier will be rare.

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 deviations 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, mini-max 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), non-asymptotic deviation analysis has drawn considerable attention and it is becoming increasingly important to construct sub-Gaussian estimators 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 $\tau$ needs to be specified in advance to ultimately balance between resistance to outliers (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 $\tau$ 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 expensive as $d$ grows. Efficient tuning is important for not only the problem’s own interest, but also its 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 are intended primarily to 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 scheme to calibrate the tuning parameter $\tau$ in the Huber loss which is displayed specified in (9). In Huber’s original proposal (Huber, 1981), $\tau$ is chosen as 1.345 so that the asymptotic efficiency of the estimator is 95% for the normal model. Since then, this has become the default setting and also find 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 non-asymptotic deviation analysis, our proposed $\tau$ grows with sample size for bias-robustness trade-off. For
linear regression under different regimes, the optimal tuning parameter $\tau$ depends on the dimension $d$ differently: $\tau \sim \sigma \sqrt{(n/d)}$ in the low dimensional setting (i.e. $d/n$ is small) and $\tau \sim \sigma \sqrt{n/\log(d)}$ in high dimensions. Thirdly, we provide simple and fast algorithms to implement the data-driven procedure under various scenarios.

The remainder of this paper is organized as follows. In Section 2, we revisit the fundamental mean estimation problem. Motivated by a careful analysis of the truncated sample mean, we introduce a novel data-driven adaptive Huber estimator. We then extend this data-driven tuning scheme to robust regression in Section 3 under both low and high dimensional settings. Extensive numerical experiments are given in Section 4 to demonstrate the finite sample performance of the proposed procedure. All the proofs, together with additional technical details, 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.) random variables from $X$ with mean $\mu$ and finite variance $\sigma^2 > 0$. The sample mean, denoted as $\bar{X}_n$, is the most natural estimator for $\mu$. 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 $\bar{X}_n$, we consider the truncated sample mean

$$m_{\tau} = \frac{1}{n} \sum_{i=1}^{n} \psi_{\tau}(X_i)$$

for some $\tau > 0$, where

$$\psi_{\tau}(x) = \text{sign}(x) \min(|x|, \tau), \quad x \in \mathbb{R}$$

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

**Proposition 1.** Assume that \(v_2 := \sqrt{\mathbb{E}(X^2)}\) is finite. For any \(z > 0,\)

(i) the truncated mean \(m_\tau\) 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) the truncated mean \(m_\tau\) 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 1 shows that how the procedure would perform under various idealized scenarios, as such providing guidance on the choice of \(\tau\). Here \(z > 0\) is a user-specified parameter that controls the confidence level; see further discussions before Remark 2. Given a properly tuned \(\tau\), the sub-Gaussian performance is achieved; conversely, if the resulting estimator performs well, it means that the data are truncated at the right level and therefore can be further exploited. An ideal \(\tau\) is such that the sample mean of truncated data \(\psi_\tau(X_1), \ldots, \psi_\tau(X_n)\) serves as a good estimator of \(\mu\). The influence of distributional outliers is weakened due to 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 observation, we propose to choose \(\tau\) by solving the equation
\[\tau = \left\{ \sum_{i=1}^{n} \psi_\tau^2(X_i) \right\}^{\frac{1}{2}} \sqrt{n/z}; \quad \tau > 0,\]

which is equivalent to
\[\frac{1}{n} \sum_{i=1}^{n} \frac{\psi_\tau^2(X_i)}{\tau^2} = \frac{z}{n}, \quad \tau > 0.\]  

(3)

We will show that under mild conditions, equation (3) has a unique solution, denoted as \(\hat{\tau}_z\), which gives rise to a data-driven mean estimator
\[m_\hat{\tau}_z = \frac{1}{n} \sum_{i=1}^{n} \min(|X_i|, \hat{\tau}_z) \text{sign}(X_i).\]  

(4)
To understand the statistical property of $\hat{\tau}_z$, consider the population version of (3):
\[
\frac{\mathbb{E}\{\psi_\tau^2(X)\}}{\tau^2} = \frac{\mathbb{E}\min(X^2, \tau^2)}{\tau^2} = \frac{z}{n}, \quad \tau > 0.
\] (5)
The following result establishes existence and uniqueness of the solution to (5).

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

(i) Provided $0 < z < n \mathbb{P}(|X| > 0)$, equation (5) has a unique solution, denoted by $\tau_z$, which satisfies
\[
\left\{ \mathbb{E}\min(X^2, q_{z/n}^2) \right\}^{1/2} \sqrt{\frac{n}{z}} \leq \tau_z \leq v_2 \sqrt{\frac{n}{z}},
\]
where $q_\alpha := \inf \{ t : \mathbb{P}(|X| > t) \leq \alpha \}$ is the upper $\alpha$-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 $n \to \infty$.

We now move to the sample version. As a direct consequence of Proposition 2, the following result ensures existence and uniqueness of the solution to equation (3).

**Proposition 3.** Provided $0 < z < \sum_{i=1}^n I(|X_i| > 0)$, equation (3) admits a unique solution.

Throughout, denote $\hat{\tau}_z$ the solution to (3), which is unique and positive whenever $z < \sum_{i=1}^n I(|X_i| > 0)$. For completeness, we set $\hat{\tau}_z = 0$ on the event \{ $z \geq \sum_{i=1}^n I(|X_i| > 0)$ \}. If the distribution of $X$ satisfies $\mathbb{P}(X = 0) = 0$, then $\hat{\tau}_z > 0$ with probability one, provided $0 < z < n$. With both $\tau_z$ and $\hat{\tau}_z$ well defined, we investigate the statistical property of $\hat{\tau}_z$.

**Theorem 1.** Assume $\mathbb{E}(X^2) < \infty$ and $\mathbb{P}(X = 0) = 0$. For any $1 \leq z < n$ and $0 < r < 1$, we have
\[
\mathbb{P}(\left| \frac{\hat{\tau}_z}{\tau_z} - 1 \right| \geq r) \leq e^{-a_1^2 r^2 z^2 / (2z + 2a_1 rz/3)} + e^{-a_2^2 r^2 z/2} + 2e^{-(a_1 \wedge a_2)^2 z/8},
\] (6)
where
\[
a_1 = a_1(z, r) = \frac{P(\tau_z)}{2Q(\tau_z)} \frac{2 + r}{(1 + r)^2} \quad \text{and} \quad a_2 = a_2(z, r) = \frac{P(\tau_z - \tau_z r)}{2Q(\tau_z)} \frac{2 - r}{1 - r},
\] (7)
with $P(t) = \mathbb{E}\{X^2 I(|X| \leq t)\}$ and $Q(t) = \mathbb{E}\{\psi_t^2(X)\}$. 6
Remark 1. Here we give some direct implications of Theorem 1.

(i) Let $z = z_n \geq 1$ satisfy $z = o(n)$ and $z \to \infty$ as $n \to \infty$. By Proposition 2, $\tau_z \to \infty$ and $\tau_z \sim v_2\sqrt{n/z}$, which further 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 $\kappa \geq 1$ in (6), 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 $\hat{\tau}_z$ satisfies that with probability approaching one, $\tau_z/2 \leq \hat{\tau}_z \leq 3\tau_z/2$.

We end this section with a uniform deviation bound for $m_\tau$. 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 $m_{\hat{\tau}_z}$ is tightly concentrated around the mean with high probability.

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

$$\sup_{\tau_z^*/2 \leq \tau \leq 3\tau_z^*/2} |m_\tau - \mu| \leq 4v_2(z/n)^{1/2} + v_2n^{-1/2}. \quad (8)$$

Let $z = 2\log(n)$ and $\hat{\tau}_z$ be the solution to (3). The mean estimator $m_{\hat{\tau}_z}$ given in (4) satisfies $|m_{\hat{\tau}_z} - \mu| \leq 4v_2\sqrt{2\log(n)/n} + v_2n^{-1/2}$ with probability at least $1 - c_1n^{-c_2}$ for all sufficiently large $n$, 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 $\sigma$. 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. Recall the Huber loss

$$\ell_\tau(x) = \begin{cases} 
  x^2/2 & \text{if } |x| \leq \tau, \\
  \tau|x| - \tau^2/2 & \text{if } |x| > \tau,
\end{cases} \quad (9)$$

where $\tau > 0$ is the robustification parameter. The Huber loss is continuously differentiable with $\ell'_\tau(x) = \psi_\tau(x)$ for $\psi_\tau(\cdot)$ given in (2).
Given a sample of observations \(X_1, \ldots, X_n\) from \(X\) with mean \(\mu\) and finite variance \(\sigma^2\), Huber’s estimator is obtained by solving the optimization problem

\[
\hat{\mu}_\tau = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ell_\tau(X_i - \theta),
\]

or equivalently, \(\hat{\mu}_\tau\) is the unique solution to

\[
0 = \sum_{i=1}^{n} \psi_\tau(X_i - \theta) = \sum_{i=1}^{n} \min(|X_i - \theta|, \tau) \text{sign}(X_i - \theta).
\]  

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

**Proposition 4.** Let \(z > 0\) and \(v \geq \sigma\). Provided \(n \geq 8z\), \(\hat{\mu}_\tau\) with \(\tau = v\sqrt{n/z}\) satisfies the bound \(|\hat{\mu}_\tau - \mu| \leq 4v\sqrt{z/n}\) with probability at least \(1 - 2e^{-z}\).

Proposition 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 (5):

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

According to Proposition 2, provided \(0 < z < n\mathbb{P}(X \neq \mu)\), this equation has a unique solution, denoted by \(\tau_{z,\mu}\), which satisfies

\[
\sqrt{\mathbb{E}\min\{(X - \mu)^2, \bar{q}_\alpha/n\}} \sqrt{\frac{n}{z}} \leq \tau_{z,\mu} \leq \sigma \sqrt{\frac{n}{z}},
\]

where \(\bar{q}_\alpha = \inf\{t : \mathbb{P}(|X - \mu| > t) \leq \alpha\}\) is the upper \(\alpha\)-quantile of \(|X - \mu|\). 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 (11) and (12), a clearly motivated data-driven estimate of \(\mu\) can be obtained by solving the following system of equations:

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

\[13\]
Observe that for any given $\tau > 0$, equation $f_1(\cdot, \tau) = 0$ always admits a unique solution, and for any given $\theta$, equation $f_2(\theta, \cdot) = 0$ has a unique solution provided $z < \sum_{i=1}^n I(X_i \neq \theta)$.

With initial values $\theta^{(0)} = \bar{X}_n$ and $\tau^{(0)} = \hat{\sigma}_n \sqrt{n/z}$ where $\hat{\sigma}_n^2$ denotes the sample variance, we solve (13) successively by computing a sequence of solutions $\{(\theta^{(k)}, \tau^{(k)})\}_{k \geq 1}$ that fulfill $f_2(\theta^{(k-1)}, \tau^{(k)}) = 0$ and $f_1(\theta^{(k)}, \tau^{(k)}) = 0$ for $k \geq 1$. For a predetermined tolerance level $\epsilon$, we stop the algorithm within the $\ell$-th iteration step when \[ \max\{|\theta^{(\ell)} - \theta^{(\ell-1)}|, |\tau^{(\ell)} - \tau^{(\ell-1)}|\} \leq \epsilon. \]

We then use $\theta^{(\ell)}$ as our final robust estimator of $\mu$.

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, if $z \geq 1$ is fixed, there is no net improvement in terms of robustness; instead, we should let $z = z_n$ slowly grow with the sample size to gain robustness without introducing extra bias. Specifically, we choose $z = \log(n)$ throughout the numerical experiments carried out in this paper.

**Remark 2.** Our proposed procedure has some similarities to the estimator considered in Bickel (1975), which is obtained as the solution of $\sum_{i=1}^n \psi_{\hat{\sigma}}(X_i - \theta) = 0$, where $\hat{\sigma}$ is chosen independently as the normalized interquartile range

\[ \hat{\sigma}^{(1)} = \{X_{(n-[n/4]+1)} - X_{([n/4])}\}/2\Phi^{-1}(3/4) \]

or the symmetrized interquartile range

\[ \hat{\sigma}^{(2)} = \text{median}\{|X_i - m|\}/\Phi^{-1}(3/4), \]

where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics and $m$ is the sample median. Provided that $X$ has a symmetric distribution, the consistency of $\hat{\sigma}^{(1)}$ or $\hat{\sigma}^{(2)}$ can be established.

Unlike this classical approach, we waive the symmetry requirement by allowing the robustification parameter to diverge to reduce the bias that is 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 $\mu$ and calibrates $\tau$ simultaneously by solving a system of equations. In fact, as a direct extension of the idea in Section 2.1, we may also tune $\tau$ independently from estimation by
solving
\[ \left( \frac{n}{2} \right)^{-1} \sum_{1 \leq i < j \leq n} \min\{\frac{(X_i - X_j)^2}{2}, \tau^2\} = \frac{z}{n}, \quad \tau > 0. \]

Let \( X' \) be an independent copy of \( X \). Then the population version of this equation is
\[ \mathbb{E} \min\{\frac{(X - X')^2}{2}, \tau^2\}/\tau^2 = z/n, \]
whose solution is unique under mild conditions and scales as \( \sigma \sqrt{n/z} \).

### 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
\[ Y_i = \beta_0^* + X_i^\top \beta^* + \varepsilon_i, \quad i = 1, \ldots, n, \tag{14} \]
where \( Y_i \)’s represent response variables, \( X_i \)’s are \( d \)-dimensional vector of covariates, \( \beta_0^* \) and \( \beta^* \in \mathbb{R}^d \) are 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 simplicity, we also introduce \( Z_i = (1, X_i^\top)^\top \) for \( i = 1, \ldots, n \) and use \( \theta^* = (\beta_0^*, \beta^*)^\top \) to denote the total vector of unknown parameters. The goal is to estimate \( \theta^* \) from observed data \( \{(Y_i, X_i)\}_{i=1}^n \).

#### 3.1 Adaptive Huber regression in low dimensions

We start with the low-dimensional setting 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. The necessity of finding robust alternatives to the least squares was discussed in Huber (1973) under Huber’s \( \epsilon \)-contamination model. Under different hypotheses that allow for heavy-tailed distributions, we refer to Audibert and Catoni (2011) and Sun, Zhou and Fan (2017) for non-asymptotic analysis of Huber-type robust regression methods; the former focuses on the excess risk bounds and the latter provides deviation bounds for the estimator along with non-asymptotic Bahadur representations.
Given any \( \tau > 0 \), Huber’s M-estimator is defined as
\[
\hat{\theta}_\tau = (\hat{\beta}_{0,\tau}, \hat{\beta}_\tau^\top)^\top \in \arg\min_{\theta \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \ell_\tau(Y_i - Z_i^\top \theta),
\]
where \( \ell_\tau(\cdot) \) is given in (9). By the convexity of Huber loss, the solution to (15) is uniquely determined via the first-order condition:
\[
\sum_{i=1}^{n} \psi_\tau(Y_i - Z_i^\top \hat{\theta}_\tau) Z_i = 0.
\]

Most of the desirable features of Huber’s method are established under the assumption that the distribution of errors is symmetric around zero. Due to asymmetry, the bias induced by the Huber loss is nonnegligible. To make this statement precise, note that \( \hat{\theta}_\tau = (\hat{\beta}_{0,\tau}, \hat{\beta}_\tau^\top)^\top \) is a natural M-estimator of
\[
\theta^* = (\beta^*_{0,\tau}, \beta^*_\tau)^\top = \arg\min_{(\beta_0, \beta^\top) \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \mathbb{E}\ell_\tau(Y_i - \beta_0 - X_i^\top \beta),
\]
whereas the true parameters \( \beta^*_{0,\tau} \) and \( \beta^*_\tau \) are identified as argmin \( \beta_0, \beta \sum_{i=1}^{n} \mathbb{E}(Y_i - \beta_0 - X_i^\top \beta)^2 \).
For a fixed \( \tau > 0 \), although \( \hat{\beta}_{0,\tau} \) and \( \hat{\beta}_\tau \) are robust estimates of \( \beta^*_{0,\tau} \) and \( \beta^*_\tau \), respectively, in general \( (\beta_{0,\tau}, \beta_\tau) \) differs from \( (\beta^*_{0,\tau}, \beta^*_\tau) \), as unveiled by the following result.

**Proposition 5.** Assume that \( \varepsilon \) and \( X \) are independent, and that the function \( \alpha \mapsto \mathbb{E}\ell_\tau(\varepsilon - \alpha) \) has a unique minimizer, denoted by \( \alpha_\tau = \arg\min_{\alpha \in \mathbb{R}} \mathbb{E}\ell_\tau(\varepsilon - \alpha) \), which satisfies
\[
\mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau) > 0.
\]
Assume further that \( \mathbb{E}(ZZ^\top) \) is positive definite. Then we have
\[
\beta^*_{0,\tau} = \beta^*_0 + \alpha_\tau \quad \text{and} \quad \beta^*_\tau = \beta^*.
\]
Moreover, \( \alpha_\tau \) with \( \tau > \sigma \) satisfies the bound
\[
|\alpha_\tau| \leq \frac{\sigma^2 - \mathbb{E}\psi^2_\tau(\varepsilon)}{1 - \tau^{-2} \sigma^2} \frac{1}{\tau}.
\]

Note also that Huber loss minimization is equivalent to the following penalized least squares problem (She and Owen, 2011):
\[
(\hat{\mu}_\tau, \hat{\theta}_\tau) = \arg\min_{\mu \in \mathbb{R}^n, \theta \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu_i - Z_i^\top \theta)^2 + \tau \sum_{i=1}^{n} |\mu_i| \right\},
\]

where $\mu = (\mu_1, \ldots, \mu_n)^T$ and $\tilde{\theta}_\tau$ here coincides with that in (15). The loss function in (20) can be written as $\sum_{i=1}^n (Y_i - \mu_i - \beta_0 - X_i^T \beta)^2 / 2 + \tau \sum_{i=1}^n |\mu_i|$. This explains from a different perspective that why the bias arises only at the intercept. The larger the value of $\tau$ is, the sparser the $\hat{\mu}_\tau$ is and therefore the smaller the estimation bias is.

The message delivered by Proposition 5 calls attention to intercept estimation, a problem of independent interest that needs to be treated with greater caution. If the distribution of $\varepsilon$ is asymmetric, $\alpha_\tau$ is typically non-zero for any $\tau > 0$: the smaller the $\tau$ is, the larger the bias becomes and so is the prediction error. To balance bias and robustness, in the following we propose two modifications, one-step and two-step, of Huber’s method that are robust against heavy-tailed and asymmetric error distributions and meanwhile maintain high efficiency in the normal case.

3.1.1 One-step method

As pointed out in Zhou et al. (2018), there is an inherent bias-robustness trade-off in the choice of $\tau$, which should adapt to the sample size, dimension and the variance of noise. Theorem 3 below makes this statement precise. To begin with, we impose the following moment conditions.

**Condition 1.** The covariates $X_1, \ldots, X_n$ are i.i.d. random vectors from $X$. There exists some constant $A_0 > 0$ such that for any $u \in \mathbb{R}^{d+1}$ and $t \in \mathbb{R}$, $\mathbb{P}(|\langle u, z \rangle| \geq A_0 \|u\|_2 \cdot t) \leq 2e^{-t^2}$, where $z = S^{-1/2}Z$ and $S = \mathbb{E}(ZZ^T)$ is positive definite. The regression errors $\varepsilon_i$ are independent and satisfy $\mathbb{E}(\varepsilon_i | X_i) = 0$ and $\mathbb{E}(\varepsilon_i^2 | X_i) = \sigma^2$ almost surely.

**Theorem 3.** Assume Condition 1 holds. For any $z > 0$ and $v \geq \sigma$, the estimator $\hat{\theta}_\tau$ in (15) with $\tau = v\sqrt{n/(d+z)}$ satisfies the bound

$$\|S^{1/2}(\hat{\theta}_\tau - \theta^*)\|_2 \leq c_1 v \sqrt{\frac{d+z}{n}}$$

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 $A_0$.

Theorem 3 establishes a sub-Gaussian concentration bound for $\hat{\theta}_\tau$ under the optimal sampling size scaling, which improves that in Theorem 2.1 in Zhou et al. (2018). To achieve
a sub-Gaussian performance under the finite variance condition, the key observation is that the robustification parameter $\tau$ should adapt to the sample size, dimension, variance of noise and confidence level for optimal trade-off between bias and robustness. Extending our data-driven proposal for mean estimation, we estimate $\theta^*$ and calibrate $\tau$ simultaneously by solving the system of equations

\[
\begin{aligned}
g_1(\theta, \tau) &= \sum_{i=1}^{n} \psi_{\tau}(Y_i - Z_i^T \theta) Z_i = 0, \\
g_2(\theta, \tau) &= (n - d)^{-1} \sum_{i=1}^{n} \min\{(Y_i - Z_i^T \theta)^2, \tau^2\}/\tau^2 - n^{-1}(d + z) = 0,
\end{aligned}
\]

$\theta \in \mathbb{R}^{d+1}, \tau > 0$.

(21)

With initial values $\theta^{(0)} := \hat{\theta}_{\text{ols}} = (\sum_{i=1}^{n} Z_i Z_i^T)^{-1} \sum_{i=1}^{n} Y_i Z_i$ and $\tau^{(0)} = \hat{\sigma}_n \sqrt{n/(d + z)}$ where $\hat{\sigma}_n^2 = (n - d)^{-1} \sum_{i=1}^{n} (Y_i - Z_i^T \hat{\theta}_{\text{ols}})^2$, for $k \geq 1$, compute $\tau^{(k)}$ as the solution to $g_2(\theta^{(k-1)}, \tau^{(k)}) = 0$ and then compute $\theta^{(k)}$ as the solution to $g_1(\theta^{(k)}, \tau^{(k)}) = 0$. Repeat the above steps until convergence or until the maximum number of iterations is reached. Denote by $(\tilde{\theta}, \tilde{\tau})$ the final solution, then set $\tilde{\theta}^1 := \tilde{\theta}_{\tilde{\tau}}$ as our one-step estimator.

The main advantage of the proposed adaptive Huber regression over the classical one with $\tau = 1.345\sigma$ is that the estimation bias with respect to intercept is alleviated. When $\tau$ is of the order $\sigma \sqrt{n/d}$, from (19) in Proposition 5 we see that the order of bias $|\alpha_\tau|$ is roughly $\sigma \sqrt{d/n}$. Examining the proof of Proposition 5, we find that the decrease in bias is linear in $1/\tau$ when the second moment is finite, and is quadratic in $1/\tau$ if the third moment is finite. We contrast this decrease with the stochastic term that dominates the error bound, which is given by the quadratic form of the score function evaluated at $\theta^*$. The order of this term is roughly $\sigma \sqrt{d/n} + \tau d/n$. Balancing out the two terms yields the proposed choice of $\tau$. However, by letting $\tau$ grow with sample size, the bias is reduced at the expense of losing some robustness of the coefficients estimators. When $\tau$ scales as a constant, such as $c\sigma$, the corresponding Huber loss is Lipschitz with bounded score function, and since $\beta^*_\tau = \beta^*$ for any $\tau > 0$, there will be no sacrifice in bias. The tuning constant $c$ is typically chosen 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 recent selective overview. The next result establishes the deviations of the Huber estimator with a fixed $\tau$ from a non-asymptotic viewpoint, representing a
Theorem 4. Suppose Condition 1 and the assumptions in Proposition 5 hold. Assume further that

$$P(|\varepsilon - \alpha \tau| \geq \tau/2) \geq \rho \tau > 0.$$  

(22)

Then for any $z > 0$, the estimator $\hat{\theta}_\tau$ in (15) satisfies

$$\|S^{1/2}(\hat{\theta}_\tau - \theta^*_\tau)\|_2 \lesssim \rho^{-1}A_0 \left( \sigma \sqrt{\frac{d+z}{n}} + \frac{d+z}{n} \right)$$  

(23)

with probability at least $1 - 2e^{-z}$ provided $n \geq c_3(d+z)$, where $c_3 = c_3(\rho \tau, A_0) > 0$.

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.

3.1.2 Two-step method

In the first stage, we compute the Huber estimator $\hat{\theta}_\tau = (\hat{\beta}_{0\tau}, \hat{\beta}_{\tau}^\top)$ by solving the optimization problem in (15) with $\tau = c \sigma$ for some constant $c > 1$. We take $c = 1.345$ so that its efficiency at the normal model is 95%. For $\sigma$, it can be estimated simultaneously with $\theta^*$ by solving a system of equations as in Huber’s proposal 2 (Huber, 1981), or we can fix $\sigma$ at an initial robust estimate and then optimize over $\theta$ (Hampel et al., 1986). We follow the former route and consider an iterative simultaneous estimation procedure, starting at iteration 0 with an initial estimate $\theta^{(0)}$. At iteration $k = 0, 1, 2, \ldots$ it applies a simple procedure to obtain $\hat{\sigma}^{(k)}$, which is then used to update $\theta^{(k)}$, producing $\theta^{(k+1)}$. The procedure involves two steps.

Step 1: Scale estimation. Using the current estimate $\theta^{(k)}$, we compute the vector of residuals $r^{(k)} = (r^{(k)}_1, \ldots, r^{(k)}_n)^\top$ and the robustification parameter $\tau^{(k)} = 1.345\hat{\sigma}^{(k)}$, where $\hat{\sigma}^{(k)}$ denotes the median absolute deviation (MAD) estimator $\text{median}\{|r^{(k)}_i - \text{median}(r^{(k)}_i)|/\Phi^{-1}(3/4)\}$.

Step 2: Weighted least squares. Compute the $n \times n$ diagonal matrix $W^{(k)} = \text{diag}(1 + w^{(k)}_1, \ldots, 1 + w^{(k)}_n)$, where $w^{(k)}_i = |r^{(k)}_i|/\tau^{(k)} - 1$ if $|r^{(k)}_i| > \tau^{(k)}$ and $w^{(k)}_i = 0$ if
Then we update $\theta^{(k)}$ to produce $\theta^{(k+1)}$ via weighted least squares, that is,

$$
\theta^{(k+1)} = \arg\min_{\theta \in \mathbb{R}^{d+1}} \frac{1}{1 + w_i^{(k)}} \left( Y_i - Z_i^\top \theta \right)^2 = (Z^\top W^{(k)} Z)^{-1} Z^\top Y,
$$

where $Z = (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n \times (d+1)}$ and $Y = (Y_1, \ldots, Y_n)^\top$.

Starting with $\theta^{(0)} = \hat{\theta}_{\text{o.ls}}$, we repeat the above two steps for $s = 0, 1, 2, \ldots$ until convergence. We use $\hat{\beta}_{\text{II}} \in \mathbb{R}^d$ to denote the vector of coefficients estimates extracted from the final solution produced by the above procedure.

In the second stage, observe that $\beta_0^* = \mathbb{E}(\delta_i)$, where $\delta_i = Y_i - X_i^\top \beta^*$ are the residuals. To estimate $\beta_0^*$, defining fitted residuals $\hat{\delta}_i = Y_i - X_i^\top \hat{\beta}_{\text{II}}$, we solve the system of equations

$$
\begin{align*}
\left\{ \begin{array}{ll}
f_1(\beta_0, \tau) := & n^{-1} \sum_{i=1}^n \min \{ (\hat{\delta}_i - \beta_0)^2, \tau^2 \} / \tau^2 - n^{-1} \log(n) = 0, \\
f_2(\beta_0, \tau) := & \sum_{i=1}^n \psi_\tau(\hat{\delta}_i - \beta_0) = 0,
\end{array} \right.
\end{align*}
$$

(24)

in the same way as for solving (13) to obtain $\hat{\beta}_0^{\text{II}}$. Stack $\hat{\beta}_0^{\text{II}}$ on top of $\hat{\beta}_{\text{II}}$ we obtain the two-step estimator $\hat{\theta}_{\text{II}} \in \mathbb{R}^{d+1}$ of $\theta^*$.

Both the one-step and two-step methods are computationally efficient. For the former, letting $\tau$ diverge with sample size reduces the estimation bias in intercept at the cost of losing some robustness for estimating coefficients; the latter achieves high degree of robustness for estimating both the intercept and regression coefficients, and therefore takes the biggest advantage when the error distributions are heavy-tailed and skewed. While at the normal model, since $\max_{1 \leq i \leq n} |\varepsilon_i| \sim \sigma \sqrt{2 \log(2n)}$ and the order of $\tau$ is roughly $\sigma \sqrt{(n/d)}$, 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 the proposed data-driven Huber regression.

### 3.2 Adaptive Huber regression in high dimensions

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

Given observations \( \{(Y_i, X_i)\}_{i=1}^n \), the Lasso is the solution to

\[
\hat{\beta}_{\text{lasso}}(\lambda) \in \arg\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \sum_{i=1}^n (Y_i - \beta_0 - X_i^T \beta)^2 + \lambda \| \beta \|_1 \right\},
\]

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, in which the \( \ell_1 \)-norm
penalizes the fitted coefficients to induce sparsity. However, least squares fitting is sensitive
to the tails of error distributions, particularly for ultra-high dimensional covariates as the
maximum spurious correlation between the covariates and the realized noise can be large,
and therefore is not an ideal choice in the presence of heavy-tailed noise.

Recently, Fan, Li and Wang (2017) modified Huber’s procedure (Huber, 1973) to ob-
tain an \( \ell_1 \)-regularized robust estimator which fulfills desirable concentration bounds under
only 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

\[
\hat{\theta}_H(\tau, \lambda) \in \arg\min_{\theta = (\beta_0, \beta) \in \mathbb{R}^{d+1}} \{ \mathcal{L}_\tau(\theta) + \lambda \| \beta \|_1 \},
\]

where \( \mathcal{L}_\tau(\theta) := n^{-1} \sum_{i=1}^n \ell_\tau(Y_i - Z_i^T \theta) = n^{-1} \sum_{i=1}^n \ell_\tau(Y_i - \beta_0 - X_i^T \beta) \), \( \tau \) and \( \lambda \) denote the
robustification and regularization parameters, respectively.

Provided finite variance of the distribution of \( \varepsilon_i \), Theorem 5 below reveals that the \( \ell_1 \)-
regularized Huber regression with properly tuned \( (\tau, \lambda) \) gives rise to statistically consistent
estimators with \( \ell_1 \)- and \( \ell_2 \)-errors scaling as \( s \sqrt{\log(d)/n} \) and \( s \sqrt{\log(d)/n} \), respectively,
under the sample size scaling \( n \gtrsim s \log(d) \). These rates are exactly the minimax rates
enjoyed by the standard Lasso with sub-Gaussian errors.

**Theorem 5.** Assume Condition 1 holds and denote by \( \lambda_S \) the minimal eigenvalue of \( S \).
Assume further that the unknown \( \beta^* \) is sparse with \( s = \| \beta^* \|_0 \). Let \( \sigma_{jj} = \mathbb{E}(X_j^2) \) for
\[ j = 1, \ldots, d. \] Then the estimator \( \hat{\theta}_H(\tau, \lambda) \) given in (25) with \( \tau = \sigma \sqrt{n/\log(d)} \) and \( \lambda \) scaling with \( A_0 \max_{1 \leq j \leq d} \sigma_{jj}^{1/2} \sigma \sqrt{\log(d)/n} \) satisfies

\[
\| \hat{\theta}_H(\tau, \lambda) - \theta^* \|_2 \lesssim \frac{\lambda \sqrt{s}}{\lambda_s} \quad \text{and} \quad \| \hat{\theta}_H(\tau, \lambda) - \theta^* \|_1 \lesssim \frac{\lambda s}{\lambda_s} \tag{26}
\]

with probability at least \( 1 - 5d^{-1} \) as long as \( n \geq c_1 s \log(d) \), where \( c_1 > 0 \) is a constant depending only on \( (A_0, \max_{1 \leq j \leq d} \sigma_{jj}, \lambda_S) \).

**Remark 3.** The main purpose of using 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 quadratic loss, such as the nonconvex Tukey and Cauchy losses, the quantile loss and the asymmetric quadratic loss, among others. We refer to Loh (2017), Alquier, Cottet and Lecué (2017) and Fan et al. (2018) for the most recent studies on regularized \( M \)-estimators with general loss functions.

In practice, it is computationally demanding to choose the optimal values of tuning parameters \( \tau \) and \( \lambda \) by a two-dimensional grid search using cross-validation. Combining the data-driven method in Section 3.1 and the refitted cross-validation (RCV) technique (Fan, Guo and Hao, 2012), we consider the following procedure that estimates \( \theta^* \) and tunes \( \tau \) simultaneously. Given a random sample of size \( n \), we first randomly split it to two subsamples \( \{(Y_{1i}, X_{1i})\}_{i=1}^{n_1} \) and \( \{(Y_{2i}, X_{2i})\}_{i=1}^{n_2} \), where \( n_1 = n_2 = n/2 \) if \( n \) is even or \( n_1 = (n + 1)/2 \), \( n_2 = (n - 1)/2 \) otherwise. To begin with, we take cross-validated Lasso estimators computed separately from the two subsamples as initial values. At the \( k \)-th iteration \( (k \geq 1) \), using the estimates \( \hat{\theta}_1^{(k-1)} \) and \( \hat{\theta}_2^{(k-1)} \) from the last iteration, we compute \( \tau^{(k)} \) which is the solution to

\[
\frac{1}{2\{n_1 - \hat{s}_1^{(k-1)}\}} \sum_{i=1}^{n_1} \min\{\frac{(Y_{1i} - Z_{1i}^\top \hat{\theta}_2^{(k-1)})^2}{\tau^2}, \tau^2\} + \frac{1}{2\{n_2 - \hat{s}_2^{(k-1)}\}} \sum_{i=1}^{n_2} \min\{\frac{(Y_{2i} - Z_{2i}^\top \hat{\theta}_1^{(k-1)})^2}{\tau^2}, \tau^2\} = \frac{\log(nd)}{n}, \tag{27}
\]

where \( \hat{s}_1^{(k-1)} = \| \hat{\beta}_1^{(k-1)} \|_0 \) and \( \hat{s}_2^{(k-1)} = \| \hat{\beta}_2^{(k-1)} \|_0 \). Next, take \( \tau = \tau^{(k)} \) and compute \( \hat{\theta}_1^{(k)} \) and
respectively, where $\lambda_1, \lambda_2 > 0$ are chosen via cross-validation. Repeat the above two steps until convergence or until the maximum number of iterations is reached. The resulting $\tau$ is denoted by $\hat{\tau}_{icv}$. The final estimator is then defined as the solution to (25) with $\tau = \hat{\tau}_{icv}$ and $\lambda$ calibrated via cross-validation.

To implement the data-driven Huber regression in high dimensions, again, starting with some initial guess we iteratively solve (27) and (28). For the convex optimization problems in (28), the minimizer satisfies the Karush–Kuhn–Tucker conditions, and therefore can be found by solving the following system of nonsmooth equations:

$$
\begin{align*}
-n^{-1} \sum_i \psi_\tau(Y_i - Z_i^\top \hat{\theta}) &= 0, \\
-n^{-1} \sum_i \psi_\tau(Y_i - Z_i^\top \hat{\theta}) X_{ij} + \lambda \hat{s}_j &= 0, \quad j = 1, \ldots, d \\
\hat{\beta}_j - S(\hat{\beta}_j + \hat{s}_j) &= 0, \quad j = 1, \ldots, d
\end{align*}
$$

(29)

where $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}^\top)^\top \in \mathbb{R}^{d+1}$ with $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\top$, $\hat{s}_j \in \partial |\hat{\beta}_j|$ 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 in solving (29). More specifically, in SNCD we divide (29) 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 reducing the computational cost from $O(nd^2)$ to $O(nd)$ at each iteration. The computational scalability and efficiency are gained especially when $d$ is very large. After obtaining a solution path of (28), we employ the cross-validation method to calibrate $\lambda_1$ and $\lambda_2$ and obtain the associated $\hat{\theta}_1^{(k)}$ and $\hat{\theta}_2^{(k)}$.

**Remark 4.** 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 5 also holds in high dimensions as long as the population Gram matrix $S$ is positive definite. Therefore, to further reduce the estimation bias of intercept, we suggest a two-step procedure that estimates the regression coefficients using standard regularized Huber regression and then estimates the intercept by applying the adaptive-Huber method to fitted residuals as in (24). Section 4.1.3 provides numerical studies of both the one- and two-step regularized adaptive Huber estimators.

4 Empirical analysis

4.1 Simulated examples

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. 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) $sgt(\mu, \sigma^2, \lambda, p, q)$, where we set mean parameter $\mu = 0$, variance parameter $\sigma^2 = q/(q - 2)$ with $q > 2$, shape parameter $p = 2$ and skewness parameter $\lambda = 0.75$;

(3) Lognormal distribution $LN(\mu, \sigma^2)$ with the log location parameter $\mu = 0$ and log shape parameter $\sigma > 0$;

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

Except the normal distribution, all the other three are skewed and heavy-tailed.

4.1.1 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 $\tau$ chosen via five-fold cross-validation (CV-Huber), and the proposed DA-Huber mean estimator. Figure 1 displays
the boxplots of the estimation error based on 2000 simulations, and Figure 2 illustrates the \( \alpha \)-quantile of the estimation error with \( \alpha \) ranging from 0.5 to 1. The DA-Huber estimator and sample mean perform almost identically for the normal model. 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.

![Boxplots of estimation error for different distributions](image)

(a) \( \mathcal{N}(0,1) \)  
(b) \( \text{sgt}(0,5,0.75,2,2.5) \)  
(c) \( \text{LN}(0,1.5) \)  
(d) \( \text{Par}(1,2) \)

Figure 1: Boxplots of the estimation error for the sample mean, CV-Huber and DA-Huber estimators under different settings based on 2000 simulations.

In Figure 3, we examine the 99%-quantile of the estimation error versus a distribution parameter that measures the tail behavior. Namely, for normal distributions we let \( \sigma \) 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 \( \sigma \) and \( \alpha \) vary from 0.25 to 2 and 1.5 to 3, respectively. The Huber-type estimators show significant improvement in deviations from the population mean as the tails become heavier. 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
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.

![Figure 2](image)

Figure 2: Estimation error versus confidence level for the sample mean, CV-Huber and DA-Huber estimators based on 2000 simulations.

### 4.1.2 Linear regression

We generate data \( \{(Y_i, X_i)\}_{i=1}^{n} \) from linear model (14) with \( n = 500 \) and \( d = 5 \). The intercept and vector of regression coefficients are taken to be \( \beta_0 = 5 \) and \( \beta^* = (1, -1, 1, -1, 1)^T \), respectively. The covariates \( X_i \) are i.i.d. random vectors that consist of independent coordinates from a uniform distribution \( \text{Unif}(-1.5, 1.5) \).

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 (9) and Tukey’s
The biweight loss

\[ \ell^T_T(x) = \begin{cases} 
1 - (1 - x^2/\tau^2)^3 & \text{if } |x| \leq \tau, \\
1 & \text{if } |x| > \tau.
\end{cases} \]

The tuning parameter \( \tau \) in \( \ell^T_T(\cdot) \) and \( \ell_T(\cdot) \) is taken to be 4.685 and 1.345, respectively, according to the asymptotic 95% efficiency rule. We carry out 1000 Monte Carlo simulations to: (1) evaluate the overall performance of the DA-Huber methods comparing with three competing methods, labeled as OLS, Tukey and Huber; see Figures 4 and 5, and (2) demonstrate the robustness of different methods with varying degrees of heavy-tailedness and skewness; see Figures 6 and 7.

Figures 4 and 5 display the boxplots of the estimation error of intercept \(|\hat{\beta}_0 - \beta_0^*|\) and the total \( \ell_2 \)-error \( \|\hat{\theta} - \theta^*\|^2_2 \), respectively, for a fixed distribution parameter as did in Section 4.1.1. 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 nonnegligible bias in classical robust M-estimators, as predicted by theory. In the normal
Figure 4: Boxplots of the estimation error of intercept under different settings.

Figure 5: Boxplots of the total $\ell_2$-error under different settings.
case, the DA-Huber estimator performs almost identically with the OLS and is therefore highly efficient. The $\ell_2$-error of OLS tends to spread out (due to outliers) and thus is not reported in Figure 5. Figures 6 and 7 illustrate, respectively, the average estimation error of intercept and the total $\ell_2$-error versus the distribution parameter that controls the 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 most desirable since it strikes the perfect balance between bias and robustness. Overall, the numerical results indicate that the proposed methods have substantial advantages in the presence of asymmetric and heavy-tailed errors, while maintaining high efficiency for the normal model.

Figure 6: Plots of the average estimation error of intercept $|\hat{\beta}_0 - \beta_0^*|$ versus distribution parameter (that controls tails) for the OLS estimator, standard Tukey’s and Huber’s estimators and data-adaptive Huber estimators (one-step and two-step).
Figure 7: Plots of the average $\ell_2$-error versus distribution parameter (that controls tails) for the OLS estimator, standard Tukey’s and Huber’s estimators and data-adaptive Huber estimators (one-step and two-step).

4.1.3 Sparse linear regression

Now we consider the sparse linear regression model

$$Y_i = \beta_0^* + \mathbf{X}_i^T \mathbf{\beta}^* + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbf{\beta}^* \in \mathbb{R}^d$ is sparse with $s = \|\mathbf{\beta}^*\|_0 \ll n$ and $d \gg n$. In the examples below, we take $n = 250$, $d = 1000$ and $s = 20$. We set $\beta_0^* = 3$ and $\mathbf{\beta}^* = (3, \ldots, 3, 0, \ldots, 0)^T$, where the first $s = 20$ elements of $\mathbf{\beta}^*$ all equal 3 and the rest are zero. As before, the covariates $\mathbf{X}_i$ are i.i.d. random vectors that consist of independent coordinates from $\text{Unif}(-1.5, 1.5)$, and the regression errors are generated from 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 regularization parameters $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ in the optimization programs in (28), producing $\hat{\theta}_1^{(k)}$ and $\hat{\theta}_2^{(k)}$. We evaluate the proposed
regularized DA-Huber estimators by the following measurements.

(1) RG, the relative gain of the DA-Huber estimator with respect to the Lasso in terms of $\ell_1$- and $\ell_2$-errors:

$$\text{RG}_1 = \frac{\|\hat{\theta}_H - \theta\|_1}{\|\hat{\theta}_\text{lasso} - \theta\|_1} \quad \text{and} \quad \text{RG}_2 = \frac{\|\hat{\theta}_H - \theta\|_2}{\|\hat{\theta}_\text{lasso} - \theta\|_2};$$

(2) FP, the number of false positives (the number of noise covariates that are selected);

(3) FN, the number of false negatives (the number of signal covariates that are missing).

Table 1 summaries the relative gains of the DA-Huber estimators under $\ell_1$- and $\ell_2$-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 $\ell_1$-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 8, which displays the empirical distributions of the $\ell_2$-errors of the Lasso and DA-Huber estimators.

4.2 Real data examples

In this section, using three real data sets, we demonstrate the desirable performance of the proposed DA-Huber methods in terms of prediction accuracy.

Liu and Rubin (1995) reported a data collected from a clinical trial on endogenous creatinine clearance of 34 male patients where 28 samples are free from missing data. For the four recorded variables, it is known that the level of serum creatinine is closely related to the endogenous creatinine clearance with the body weight and age properly adjusted. Linear model (14) is a natural preliminary fit to the data. In addition, we observe that the empirical kurtosis of the level of serum creatinine is 19.66, which hints potential heavy-tailedness in the data. The second example is the hedonic housing crime data (Harrison and Rubinfeld, 1978), which was originally used to study the association between housing
<table>
<thead>
<tr>
<th>Model</th>
<th>Lasso</th>
<th>DA-Huber (one-step)</th>
<th>DA-Huber (two-step)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>RG₁</td>
<td>1</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>RG₂</td>
<td>1</td>
<td>1.003</td>
</tr>
<tr>
<td>$\mathcal{N}(0, 1)$</td>
<td>FP</td>
<td>87.9</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Skewed generalized $t$</td>
<td>RG₁</td>
<td>1</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>RG₂</td>
<td>1</td>
<td>0.983</td>
</tr>
<tr>
<td>$\text{sgt}(0, 0.75, 2, 2.5)$</td>
<td>FP</td>
<td>86.1</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lognormal</td>
<td>RG₁</td>
<td>1</td>
<td>0.347</td>
</tr>
<tr>
<td></td>
<td>RG₂</td>
<td>1</td>
<td>0.495</td>
</tr>
<tr>
<td>$\text{LN}(0, 1.5)$</td>
<td>FP</td>
<td>80.8</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>0.26</td>
<td>0</td>
</tr>
<tr>
<td>Pareto</td>
<td>RG₁</td>
<td>1</td>
<td>0.653</td>
</tr>
<tr>
<td></td>
<td>RG₂</td>
<td>1</td>
<td>0.845</td>
</tr>
<tr>
<td>$\text{Par}(1, 2)$</td>
<td>FP</td>
<td>85.1</td>
<td>34.5</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: RG, FP and FN of the Lasso and DA-Huber estimators under different models. The results are based on 200 simulations.

market and local air quality. Interestingly, this data also provides some insights on how crime rates vary with respect to house-economics features, such as the proportion of residential land zoned for lots greater than 25,000 square feet, the proportion of non-retail business within a town, proportion of owner units built prior to 1940, proportion of adults without high school education, median value of owner-occupied homes, average number of rooms in owner units, and distance to five employment centers in Boston region. This data set contains 506 locations and the empirical kurtosis of the crime rate is 39.75. The last
data set is the well-known G-Econ data reported by Nordhaus et al. (2006), which was used to show the dependence of gross cell product (GCP) on geographical variables measured on a spatial scale of one degree. The original data contains 27,445 terrestrial grid cells and 47 predictors, and varies abruptly across different latitude and longitude. For example, the sizes of grid cell may change substantially from the equator to the poles. Similar to Nordhaus et al. (2006), we focus on regions from 35 to 50 latitudes (parallel north) that contain a large number of major economic centers, such as Tokyo, New York, Paris and London. Excluding cells with empty inputs, 808 observations remain for studying the relationship between the GCP (in USD) in 1990 and ten explanatory variables as discussed in Nordhaus et al. (2006), including distance to coast, distance to major navigable lakes, distance to major navigable rivers, distance to ice-free ocean, elevation, standard deviation of elevations, elevation from shuttle radar topography mission data, latitude, average precipitation, and average temperature. The empirical kurtosis of the GCP is 256.58, suggesting strong heavy-tailedness.
Figure 9: Comparison of the quantiles of mean absolute prediction errors for the OLS (black diamonds), one-step DA-Huber (blue circles), and two-step DA-Huber (red triangles). The results are based on 100 random splittings.

From the simulation studies in Section 4.1 we see that both the one-step and two-step DA-Huber estimators outperform the OLS in terms of estimation accuracy. For the real data, we focus on the prediction accuracy by investigating the mean absolute prediction errors. Specifically, upon splitting the data into $K = 7$ groups randomly, we predict the responses of one group using the regression coefficients estimated from the other $K - 1$
Various quantile levels of the $K$ mean absolute prediction errors were computed for different estimators. We repeat the random splitting 100 times. Figure 9 displays the empirical medians of $\alpha$-quantiles of the mean absolute prediction errors for the three data sets, where $\alpha$ ranges from 0.1 to 0.9. The two data-adaptive Huber estimators substantially outperform the OLS with smaller prediction errors. When heavy-tailedness prevails and the intercept is nonnegligible, such as the GCP in G-Econ data, the two-step estimator displays the best performance. In general, the one- and two-step methods perform comparably well. For the endogenous creatinine data, the 0.9-quantiles of the mean absolute prediction errors of the three methods are comparable, which is possibly due to the small sample size ($n = 28$). To sum up, the data-adaptive Huber methods provide notably better predictions than the least squares for these three real data examples.

References


