Title

Adaptive Huber Regression.

Permalink

https://escholarship.org/uc/item/4sn8g6n9

Journal

Journal of the American Statistical Association, 115(529)

ISSN

0162-1459

Authors

Sun, Qiang
Zhou, Wen-Xin
Fan, Jianqing

Publication Date

2020

DOI

10.1080/01621459.2018.1543124

Peer reviewed

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Adaptive Huber Regression

Qiang Sun, Wen-Xin Zhou & Jianqing Fan

To cite this article: Qiang Sun, Wen-Xin Zhou & Jianqing Fan (2020) Adaptive Huber Regression, Journal of the American Statistical Association, 115:529, 254-265, DOI: 10.1080/01621459.2018.1543124

To link to this article: https://doi.org/10.1080/01621459.2018.1543124

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 529, 254–265: Theory and Methods
https://doi.org/10.1080/01621459.2018.1543124

Taylor & Francis
Taylor & Francis Group

Adaptive Huber Regression

Qiang Sun[a], Wen-Xin Zhou[b], and Jianqing Fan[c,d]

[a]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [b]Department of Mathematics, University of California, San Diego, La Jolla, CA; [c]School of Data Science, Fudan University, Shanghai, China; [d]Department of Operations Research and Financial Engineering, Princeton University, NJ

ABSTRACT

Big data can easily be contaminated by outliers or contain variables with heavy-tailed distributions, which makes many conventional methods inadequate. To address this challenge, we propose the adaptive Huber regression for robust estimation and inference. The key observation is that the robustification parameter should adapt to the sample size, dimension and moments for optimal tradeoff between bias and robustness. Our theoretical framework deals with heavy-tailed distributions with bounded $(1+\delta)$th moment for any $\delta > 0$. We establish a sharp phase transition for robust estimation of regression parameters in both low and high dimensions: when $\delta \geq 1$, the estimator admits a sub-Gaussian-type deviation bound without sub-Gaussian assumptions on the data, while only a slower rate is available in the regime $0 < \delta < 1$ and the transition is smooth and optimal. In addition, we extend the methodology to allow both heavy-tailed predictors and observation noise. Simulation studies lend further support to the theory. In a genetic study of cancer cell lines that exhibit heavy-tailedness, the proposed methods are shown to be more robust and predictive. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2017
Revised October 2018

KEYWORDS
Adaptive Huber regression;
Bias and robustness tradeoff;
Finite-sample inference;
Heavy-tailed data;
Nonasymptotic optimality;
Phase transition.

1. Introduction

Modern data acquisitions have facilitated the collection of massive and high-dimensional data with complex structures. Along with holding great promises for discovering subtle population patterns that are less achievable with small-scale data, big data have introduced a series of new challenges to data analysis both computationally and statistically (Loh and Wainwright 2015; Fan et al. 2018). During the last two decades, extensive progress has been made toward extracting useful information from massive data with high-dimensional features and sub-Gaussian tails. A random variable Z is said to have sub-Gaussian tails if there exists constants c_1 and c_2 such that $\mathbb{P}(|Z| > t) \leq c_1 \exp(-c_2 t^2)$ for any $t \geq 0$ (Tibshirani 1996; Fan and Li 2001; Efron et al. 2004; Bickel, Ritov, and Tsybakov 2009). We refer to the monographs, Bühlmann and van de Geer (2011) and Hastie, Tibshirani, and Wainwright (2015), for a systematic coverage of contemporary statistical methods for high-dimensional data.

The sub-Gaussian tails requirement, albeit being convenient for theoretical analysis, is not realistic in many practical applications since modern data are often collected with low quality. For example, a recent study on functional magnetic resonance imaging (fMRI) (Eklund, Nichols, and Knutsson 2016) shows that the principal cause of invalid fMRI inferences is that the data do not follow the assumed Gaussian shape, which speaks to the need of validating the statistical methods being used in the field of neuroimaging. In a microarray data example considered in Wang, Peng, and Li (2015), it is observed that some gene expression levels have heavy tails as their kurtosises are much larger than 3, despite of the normalization methods used. In finance, the power-law nature of the distribution of returns has been validated as a stylized fact (Cont 2001). Fan et al. (2016) argued that heavy-tailed distribution is a stylized feature for high-dimensional data and proposed a shrinkage principle to attenuate the influence of outliers. Standard statistical procedures that are based on the method of least squares often behave poorly in the presence of heavy-tailed data. We say a random variable X has heavy tails if $\mathbb{P}(|X| > t)$ decays to zero polynomially in $1/t$ as $t \to \infty$ (Catoni 2012). It is therefore of ever-increasing interest to develop new statistical methods that are robust against heavy-tailed errors and other potential forms of contamination.

In this article, we first revisit the robust regression that was initiated by Peter Huber in his seminal work Huber (1973). Asymptotic properties of the Huber estimator have been well studied in the literature. We refer to Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989), and He and Shao (1996, 2000) for an unavoidably incomplete overview. However, in all of the aforementioned papers, the robustification parameter is suggested to be set as fixed according to the 95% asymptotic efficiency rule. Thus, this procedure cannot estimate the model-generating parameters consistently when the sample distribution is asymmetric.

From a nonasymptotic perspective (rather than an asymptotic efficiency rule), we propose to use the Huber regression with an adaptive robustification parameter, which is referred to as the *adaptive Huber regression*, for robust estimation and

CONTACT Qiang Sun qsun@utstat.toronto.edu Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

inference. Our adaptive procedure achieves the nonasymptotic robustness in the sense that the resulting estimator admits exponential-type concentration bounds when only low-order moments exist. Moreover, the resulting estimator is also an asymptotically unbiased estimate for the parameters of interest. In particular, we do not impose symmetry and homoscedasticity conditions on error distributions, so that our problem is intrinsically different from median/quantile regression models, which are also of independent interest and serve as important robust techniques (Koenker 2005).

We made several major contributions toward robust modeling in this article. First and foremost, we establish nonasymptotic deviation bounds for adaptive Huber regression when the error variables have only finite $(1+\delta)$th moments. By providing a matching lower bound, we observe a sharp phase transition phenomenon, which is in line with that discovered by Devroye et al. (2016) for univariate mean estimation. Second, a similar phase transition for regularized adaptive Huber regression is established in high dimensions. By defining the effective dimension and effective sample size, we present nonasymptotic results under the two different regimes in a unified form. Last, by exploiting the localized analysis developed in Fan et al. (2018), we remove the artificial bounded parameter constraint imposed in previous works; see Loh and Wainwright (2015) and Fan, Li, and Wang (2017). In the supplementary materials, we present a nonasymptotic Bahadur representation for the adaptive Huber estimator when $\delta \geq 1$, which provides a theoretical foundation for robust finite-sample inference.

The rest of the article proceeds as follows. The rest of this section is devoted to related literature. In Section 2, we revisit the Huber loss and robustification parameter, followed by the proposal of adaptive Huber regression in both low and high dimensions. We sharply characterize the nonasymptotic performance of the proposed estimators in Section 3. We describe the algorithm and implementation in Section 5. Section 6 is devoted to simulation studies and a real data application. In Section 4, we extend the methodology to allow possibly heavy-tailed covariates/predictors. All the proofs are collected in the supplementary materials.

1.1. Related Literature

The terminology "robustness" used in this article describes how stable the method performs with respect to the tail-behavior of the data, which can be either sub-Gaussian/subexponential or Pareto-like (Delaigle, Hall, and Jin 2011; Catoni 2012; Devroye et al. 2016). This is different from the conventional perspective of robust statistics under Huber's ϵ-contamination model (Huber 1964), for which a number of depth-based procedures have been developed since the groundbreaking work of Tukey (1975). Significant contributions have also been made in Liu (1990), Liu, Parelius, and Singh (1999), Zuo and Serfling (2000), Mizera (2002), and Mizera and Müller (2004). We refer to Chen, Gao, and Ren (2018) for the most recent result and a literature review concerning this problem.

Our main focus is on the conditional mean regression in the presence of heavy-tailed and asymmetric errors, which automatically distinguishes our method from quantile-based robust regressions (Koenker 2005; Belloni and Chernozhukov 2011; Wang 2013; Fan, Fan, and Barut 2014; Zheng, Peng, and He 2015). In general, quantile regression is biased toward estimating the mean regression coefficient unless the error distributions are symmetric around zero. Another recent work that is related to ours is Alquier, Cottett, and Lecué (2017). They studied a general class of regularized empirical risk minimization procedures with a particular focus on Lipschitz losses, which includes the quantile, hinge, and logistic losses. Different from all these work, our goal is to estimate the mean regression coefficients robustly. The robustness is witnessed by a nonasymptotic analysis: the proposed estimators achieve sub-Gaussian deviation bounds when the regression errors have only finite second moments. Asymptotically, our proposed estimators are fully efficient: they achieve the same efficiency as the ordinary least squares (OLS) estimators.

An important step toward estimation under heavy-tailedness has been made by Catoni (2012), whose focus is on estimating a univariate mean. Let X be a real-valued random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{var}(X) > 0$, and assume that $X_1, \ldots, X_n$ are independent and identically distributed (iid) from X. For any prespecified exception probability $t > 0$, Catoni constructs a robust mean estimator $\widehat{\mu}_{\mathrm{C}}(t)$ that deviates from the true mean μ logarithmically in $1/t$, that is,

$$\mathbb{P}\big[|\hat{\mu}_{\mathrm{C}}(t) - \mu| \leq t\sigma/n^{1/2}\big] \geq 1 - 2\exp(-ct^2), \qquad (1)$$

while the empirical mean deviates from the true mean only polynomially in $1/t^2$, namely sub-Gaussian tails versus Cauchy tail in terms of t. Further, Devroye et al. (2016) developed adaptive sub-Gaussian estimators that are independent of the prespecified exception probability. Beyond mean estimation, Brownlees, Joly, and Lugosi (2015) extended Catoni's idea to study empirical risk minimization problems when the losses are unbounded. Generalizations of the univariate results to those for matrices, such as the covariance matrices, can be found in Catoni (2016), Minsker (2018), Giulini (2017), and Fan, Li, and Wang (2017). Fan, Li, and Wang (2017) modified Huber's procedure (Huber 1973) to obtain a robust estimator, which is concentrated around the true mean with exponentially high probability in the sense of (1), and also proposed a robust procedure for sparse linear regression with asymmetric and heavy-tailed errors.

Notation. We fix some notations that will be used throughout this article. For any vector $\boldsymbol{u} = (u_1, \ldots, u_d)^{\mathrm{T}} \in \mathbb{R}^d$ and $q \geq 1$, $\|\boldsymbol{u}\|_q = (\sum_{j=1}^d |u_j|^q)^{1/q}$ is the ℓ_q norm. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we write $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}$. Moreover, we let $\|\boldsymbol{u}\|_0 = \sum_{j=1}^d 1(u_j \neq 0)$ denote the number of nonzero entries of $\boldsymbol{u}$, and set $\|\boldsymbol{u}\|_\infty = \max_{1\leq j\leq d} |u_j|$. For two sequences of real numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ denotes $a_n \leq Cb_n$ for some constant $C > 0$ independent of n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two scalars, we use $a \wedge b = \min\{a, b\}$ to denote the minimum of a and b. If $\mathbf{A}$ is an $m \times n$ matrix, we use $\|\mathbf{A}\|$ to denote its spectral norm, defined by $\|\mathbf{A}\| = \max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \|\mathbf{A}\boldsymbol{u}\|_2$, where $\mathbb{S}^{n-1} = \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^n$. For an $n \times n$ matrix $\mathbf{A}$, we use $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ to denote the maximum and minimum eigenvalues of $\mathbf{A}$, respectively. For two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semidefinite. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla f \in \mathbb{R}^d$ to denote its gradient vector as long as it exists.

2. Methodology

We consider iid observations $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ that are generated from the following heteroscedastic regression model

$$\begin{aligned} y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad &\text{with} \\ \mathbb{E}(\varepsilon_i | \boldsymbol{x}_i) = 0 \ \text{ and } \ v_{i,\delta} = \mathbb{E}\big(|\varepsilon_i|^{1+\delta}\big) < \infty. \end{aligned} \tag{2}$$

Assuming that the second moments are bounded ($\delta = 1$), the standard OLS estimator, denoted by $\widehat{\boldsymbol{\beta}}^{\text{ols}}$, admits a suboptimal polynomial-type deviation bound, and thus does not concentrate around $\boldsymbol{\beta}^*$ tightly enough for large-scale simultaneous estimation and inference. The key observation that underpins this suboptimality of the OLS estimator is the sensitivity of quadratic loss to outliers (Huber 1973; Catoni 2012), while the Huber regression with a fixed tuning constant may lead to non-negligible estimation bias. To overcome this drawback, we propose to employ the Huber loss with an adaptive robustification parameter to achieve robustness and (asymptotic) unbiasedness simultaneously. We begin with the definitions of the Huber loss and the corresponding robustification parameter.

Definition 1 (Huber loss and robustification parameter). The Huber loss $\ell_\tau(\cdot)$ (Huber 1964) is defined as

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

where $\tau > 0$ is referred to as the robustification parameter that balances bias and robustness.

The loss function $\ell_\tau(x)$ is quadratic for small values of x and becomes linear when x exceeds τ in magnitude. The parameter τ therefore controls the blending of quadratic and ℓ_1 losses, which can be regarded as two extremes of the Huber loss with $\tau = \infty$ and $\tau \to 0$, respectively. Comparing with the least squares, outliers are down weighted in the Huber loss. We will use the name, *adaptive Huber loss*, to emphasize the fact that the parameter τ should adapt to the sample size, dimension and moments for a better tradeoff between bias and robustness. This distinguishes our framework from the classical setting. As $\tau \to \infty$ is needed to reduce the bias when the error distribution is asymmetric, this loss is also called the RA-quadratic (robust approximation to quadratic) loss in Fan, Li, and Wang (2017).

Define the empirical loss function $\mathcal{L}_\tau(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \ell_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)$ for $\boldsymbol{\beta} \in \mathbb{R}^d$. The Huber estimator is defined through the following convex optimization problem

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\beta}). \tag{3}$$

In low dimensions, under the condition that $v_\delta = n^{-1} \sum_{i=1}^n \mathbb{E}(|\varepsilon_i|^{1+\delta}) < \infty$ for some $\delta > 0$, we will prove that $\hat{\boldsymbol{\beta}}_\tau$ with $\tau \asymp \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}\, n^{\max\{1/(1+\delta), 1/2\}}$ (the first factor is kept to show its explicit dependence on the moment) achieves the tight upper bound $d^{1/2} \tau^{-(\delta \wedge 1)} \asymp d^{1/2} n^{-\min\{\delta/(1+\delta), 1/2\}}$. The phase transition at $\delta = 1$ can be easily observed (see Figure 1). When higher moments exist ($\delta \ge 1$), robustification leads to a sub-Gaussian-type deviation inequality in the sense of (1).

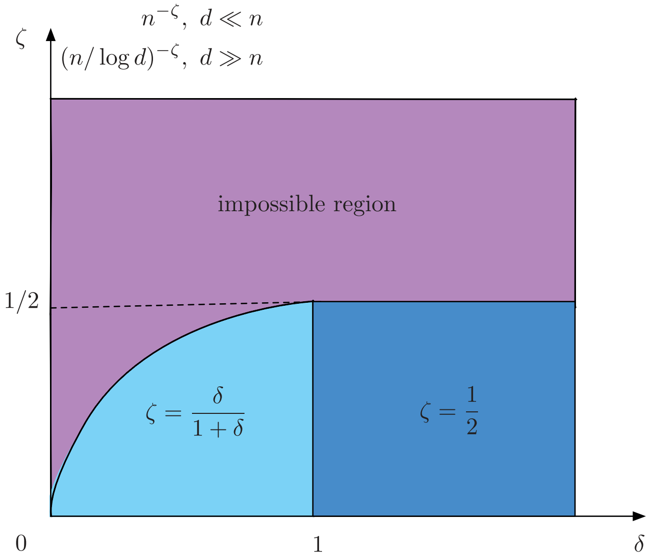

Figure 1. Phase transition in terms of ℓ_2-error for the adaptive Huber estimator. With fixed effective dimension, $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \asymp n_{\text{eff}}^{-\delta/(1+\delta)}$, when $0 < \delta < 1$; $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \asymp n_{\text{eff}}^{-1/2}$, when $\delta \ge 1$. Here n_{eff} is the effective sample size: $n_{\text{eff}} = n$ in low dimensions while $n_{\text{eff}} = n/\log d$ in high dimensions.

In the high-dimensional regime, we consider the following regularized adaptive Huber regression with a different choice of the robustification parameter

$$\widehat{\boldsymbol{\beta}}_{\tau,\lambda} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \big\{\mathcal{L}_\tau(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1\big\}, \tag{4}$$

where $\tau \asymp v_\delta \{n/(\log d)\}^{\max\{1/(1+\delta), 1/2\}}$ and $\lambda \asymp v_\delta \{(\log d)/n\}^{\min\{\delta/(1+\delta), 1/2\}}$ with $v_\delta = \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}$. Let s be the size of the true support $\mathcal{S} = \text{supp}(\boldsymbol{\beta}^*)$. We will show that the regularized Huber estimator achieves an upper bound that is of the order $s^{1/2}\{(\log d)/n\}^{\min\{\delta/(1+\delta), 1/2\}}$ for estimating $\boldsymbol{\beta}^*$ in ℓ_2-error with high probability.

To unify the nonasymptotic upper bounds in the two different regimes, we define the *effective dimension*, d_{eff}, to be d in low dimensions and s in high dimensions. In other words, d_{eff} denotes the number of nonzero parameters of the problem. The *effective sample size*, n_{eff}, is defined as $n_{\text{eff}} = n$ and $n_{\text{eff}} = n/\log d$ in low and high dimensions, respectively. We will establish a phase transition: when $\delta \ge 1$, the proposed estimator enjoys a sub-Gaussian concentration, while it only achieves a slower concentration when $0 < \delta < 1$. Specifically, we show that, for any $\delta \in (0, \infty)$, the proposed estimators with $\tau \asymp \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}\, n_{\text{eff}}^{\max\{1/(1+\delta), 1/2\}}$ achieve the following tight upper bound, up to logarithmic factors

$$\big\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\big\|_2 \lesssim d_{\text{eff}}^{1/2}\, n_{\text{eff}}^{-\min\{\delta/(1+\delta), 1/2\}} \quad \text{with high probability.} \tag{5}$$

This finding is summarized in Figure 1.

3. Nonasymptotic Theory

3.1. Adaptive Huber Regression With Increasing Dimensions

We begin with the adaptive Huber regression in the low-dimensional regime. First, we provide an upper bound for the estimation bias of Huber regression. We then establish the phase transition by establishing matching upper and lower bounds on the ℓ_2-error. The analysis is carried out under both fixed and random designs. The results under random designs are provided in the supplementary materials. We start with the following regularity condition.

Condition 1. The empirical Gram matrix $\mathbf{S}_n := n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ is nonsingular. Moreover, there exist constants c_l and c_u such that $c_l \le \lambda_{\min}(\mathbf{S}_n) \le \lambda_{\max}(\mathbf{S}_n) \le c_u$.

For any $\tau > 0$, $\hat{\boldsymbol{\beta}}_\tau$ given in (3) is natural M-estimator of

$$\boldsymbol{\beta}_\tau^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\{\mathcal{L}_\tau(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\{\ell_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\}, \tag{6}$$

where the expectation is taken over the regression errors. We call $\boldsymbol{\beta}_\tau^*$ the *Huber regression coefficient*, which is possibly different from the vector of true parameters $\boldsymbol{\beta}^*$. The estimation bias, measured by $\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2$, is a direct consequence of robustification and asymmetric error distributions. Heuristically, choosing a sufficiently large τ reduces bias at the cost of losing robustness (the extreme case of $\tau = \infty$ corresponds to the least squares estimator). Our first result shows how the magnitude of τ affects the bias $\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2$. Recall that $v_\delta = n^{-1}\sum_{i=1}^n v_{i,\delta}$ with $v_{i,\delta} = \mathbb{E}(|\varepsilon_i|^{1+\delta})$.

Proposition 1. Assume Condition 1 holds and that v_δ is finite for some $\delta > 0$. Then, the vector $\boldsymbol{\beta}_\tau^*$ of Huber regression coefficients satisfies

$$\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2 \le 2c_l^{-1/2} v_\delta \tau^{-\delta} \tag{7}$$

provided $\tau \ge (4v_\delta \widetilde{M}^2)^{1/(1+\delta)}$ for $0 < \delta < 1$ or $\tau \ge (2v_1)^{1/2}\widetilde{M}$ for $\delta \ge 1$, where $\widetilde{M} = \max_{1\le i\le n}\|\mathbf{S}_n^{-1/2}\boldsymbol{x}_i\|_2$.

The total estimation error $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2$ can therefore be decomposed into two parts

$$\underbrace{\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2}_{\text{Total error}} \le \underbrace{\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^*\|_2}_{\text{estimation error}} + \underbrace{\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2}_{\text{approximation bias}},$$

where the approximation bias is of order $\tau^{-\delta}$. A large τ reduces the bias but compromises the degree of robustness. Thus, an optimal estimator is the one with τ diverging at a certain rate to achieve the optimal tradeoff between estimation error and approximation bias. Our next result presents nonasymptotic upper bounds on the ℓ_2-error with an exponential-type exception probability, when τ is properly tuned. Recall that $\nu_\delta = \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}$ for any $\delta > 0$.

Theorem 1 (Upper bound). Assume Condition 1 holds and $v_\delta < \infty$ for some $\delta > 0$. Let $L = \max_{1\le i\le n}\|\boldsymbol{x}_i\|_\infty$ and assume $n \ge C(L, c_l)d^2 t$ for some $C(L, c_l) > 0$ depending only on L and c_l. Then, for any $t > 0$ and $\tau_0 \ge \nu_\delta$, the estimator $\hat{\boldsymbol{\beta}}_\tau$ with $\tau = \tau_0(n/t)^{\max\{1/(1+\delta),1/2\}}$ satisfies the bound

$$\|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \le 4c_l^{-1}L\tau_0\, d^{1/2}\left(\frac{t}{n}\right)^{\min\{\delta/(1+\delta),1/2\}} \tag{8}$$

with probability at least $1 - (2d+1)e^{-t}$.

Remark 1. It is worth mentioning that the proposed robust estimator depends on the unknown parameter $v_\delta^{1/(1+\delta)}$. Adaptation to the unknown moment is indeed another important problem. In Section 6, we suggest a simple cross-validation scheme for choosing τ with desirable numerical performance. A general adaptive construction of τ can be obtained via Lepski's method (Lepski 1991), which is more challenging due to unspecified constants. In the supplementary materials, we discuss a variant of Lepski's method and establish its theoretical guarantee.

Remark 2. We do not assume $\mathbb{E}(|\varepsilon_i|^{1+\delta}|\boldsymbol{x}_i)$ to be a constant, and hence the proposed method accommodates heteroscedastic regression models. For example, ε_i can take the form of $\sigma(\boldsymbol{x}_i)v_i$, where $\sigma : \mathbb{R}^d \to (0,\infty)$ is a positive function, and v_i are random variables satisfying $\mathbb{E}(v_i) = 0$ and $\mathbb{E}(|v_i|^{1+\delta}) < \infty$.

Remark 3. We need the scaling condition to go roughly as $n \gtrsim d^2 t$ under fixed designs. With random designs, we show that the scaling condition can be relaxed to $n \gtrsim d + t$. Details are given in the supplementary materials.

Theorem 1 indicates that, with only bounded $(1+\delta)$th moment, the adaptive Huber estimator achieves the upper bound $d^{1/2}n^{-\min\{\delta/(1+\delta),1/2\}}$, up to a logarithmic factor, by setting $t = \log(nd)$. A natural question is whether the upper bound in (8) is optimal. To address this, we provide a matching lower bound up to a logarithmic factor. Let $\mathcal{P}_\delta^{v_\delta}$ be the class of all distributions on $\mathbb{R}$ whose $(1+\delta)$th absolute central moment equals v_δ. Let $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}} = (\boldsymbol{x}^1,\ldots,\boldsymbol{x}^d) \in \mathbb{R}^{n\times d}$ be the design matrix and $\mathcal{U}_n = \{\boldsymbol{u} : \boldsymbol{u} \in \{-1,1\}^n\}$.

Theorem 2 (Lower bound). Assume that the regression errors ε_i are iid from a distribution in $\mathcal{P}_\delta^{v_\delta}$ with $\delta > 0$. Suppose there exists a $\boldsymbol{u} \in \mathcal{U}_n$ such that $\|n^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{u}\|_{\min} \ge \alpha$ for some $\alpha > 0$. Then, for any $t \in [0, n/2]$ and any estimator $\widehat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(y_1,\ldots,y_n,t)$ possibly depending on t, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_\delta^{v_\delta}} \mathbb{P}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \ge \alpha c_u^{-1}\nu_\delta\, d^{1/2}\left(\frac{t}{n}\right)^{\min\{\delta/(1+\delta),1/2\}}\right] \ge \frac{e^{-2t}}{2},$$

where $c_u \ge \lambda_{\max}(\mathbf{S}_n)$.

Theorem 2 reveals that root-n consistency with exponential concentration is impossible when $\delta \in (0,1)$. It widens the phenomenon observed in Theorem 3.1 in Devroye et al. (2016) for estimating a mean. In addition to the eigenvalue assumption, we need to assume that there exists a $\boldsymbol{u} \in \mathcal{U}_n \subseteq \mathbb{R}^n$ such that the minimum angle between $n^{-1}\boldsymbol{u}$ and $\boldsymbol{x}^j$ is nonvanishing. This assumption comes from the intuition that the linear subspace spanned by $\boldsymbol{x}^j$ is at most of rank d and thus cannot span the whole space $\mathbb{R}^n$. This assumption naturally holds in the univariate case where $\mathbf{X} = (1, \ldots, 1)^{\mathrm{T}}$ and we can take $\boldsymbol{u} = (1, \ldots, 1)^{\mathrm{T}}$ and $\alpha = 1$. More generally, $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{u}/n\|_{\min} = \min\{|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1|/n, \ldots, |\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^d|/n\}$. Taking $|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1|/n$ for an example, since $\boldsymbol{u} \in \{-1, +1\}^n$, we can assume that each coordinate of $\boldsymbol{x}^1$ is positive. In this case, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1/n = \sum_{i=1}^n |x_i^1|/n \geq \min_i |x_i^1|$, which is strictly positive with probability one, assuming $\boldsymbol{x}^1$ is drawn from a continuous distribution.

Together, the upper and lower bounds show that the adaptive Huber estimator achieves near-optimal deviations. Moreover, it indicates that the Huber estimator with an adaptive τ exhibits a sharp phase transition: when $\delta \geq 1$, $\widehat{\boldsymbol{\beta}}_\tau$ converges to $\boldsymbol{\beta}^*$ at the parametric rate $n^{-1/2}$, while only a slower rate of order $n^{-\delta/(1+\delta)}$ is available when the second moment does not exist.

Remark 4. We provide a parallel analysis under random designs in the supplementary materials. Beyond the nonasymptotic deviation bounds, we also prove a nonasymptotic Bahadur representation, which establishes a linear approximation of the nonlinear robust estimator. This result paves the way for future research on conducting statistical inference and constructing confidence sets under heavy-tailedness. Additionally, the proposed estimator achieves full efficiency: it is as efficient as the OLS estimator asymptotically, while the robustness is characterized via nonasymptotic performance.

3.2. Adaptive Huber Regression in High Dimensions

In this section, we study the regularized adaptive Huber estimator in high dimensions where d is allowed to grow with the sample size n exponentially. The analysis is carried out under fixed designs, and results for random designs are again provided in the supplementary materials. We start with a modified version of the localized restricted eigenvalue (LRE) introduced by Fan et al. (2018). Let $\mathbf{H}_\tau(\boldsymbol{\beta}) = \nabla^2\mathcal{L}_\tau(\boldsymbol{\beta})$ denote the Hessian matrix. Recall that $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*) \subseteq \{1, \ldots, d\}$ is the true support set with $|\mathcal{S}| = s$.

Definition 2 (LRE). The LRE of $\mathbf{H}_\tau$ is defined as

$$\kappa_+(m, \gamma, r) = \sup\left\{\langle \boldsymbol{u}, \mathbf{H}_\tau(\boldsymbol{\beta})\boldsymbol{u}\rangle : (\boldsymbol{u}, \boldsymbol{\beta}) \in \mathcal{C}(m, \gamma, r)\right\},$$

$$\kappa_-(m, \gamma, r) = \inf\left\{\langle \boldsymbol{u}, \mathbf{H}_\tau(\boldsymbol{\beta})\boldsymbol{u}\rangle : (\boldsymbol{u}, \boldsymbol{\beta}) \in \mathcal{C}(m, \gamma, r)\right\},$$

where $\mathcal{C}(m, \gamma, r) := \{(\boldsymbol{u}, \boldsymbol{\beta}) \in \mathbb{S}^{d-1} \times \mathbb{R}^d : \forall J \subseteq \{1, \ldots, d\}$ satisfying $\mathcal{S} \subseteq J, |J| \leq m, \|\boldsymbol{u}_{J^c}\|_1 \leq \gamma\|\boldsymbol{u}_J\|_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq r\}$ is a local ℓ_1-cone.

The LRE is defined in a local neighborhood of $\boldsymbol{\beta}^*$ under ℓ_1-norm. This facilitates our proof, while Fan et al. (2018) use the ℓ_2-norm.

Condition 2. $\mathbf{H}_\tau$ satisfies the LRE condition LRE(k, γ, r), that is, $\kappa_l \leq \kappa_-(k, \gamma, r) \leq \kappa_+(k, \gamma, r) \leq \kappa_u$ for some constants $\kappa_u, \kappa_l > 0$.

The condition above is referred to as the LRE condition (Fan et al. 2018). It is a unified condition for studying generalized loss functions, whose Hessians may possibly depend on $\boldsymbol{\beta}$. For Huber loss, Condition 2 also involves the observation noise. The following definition concerns the restricted eigenvalues (REs) of $\mathbf{S}_n$ instead of $\mathbf{H}_\tau$.

Definition 3 (RE). The restricted maximum and minimum eigenvalues of $\mathbf{S}_n$ are defined, respectively, as

$$\rho_+(m, \gamma) = \sup_{\boldsymbol{u}}\left\{\langle \boldsymbol{u}, \mathbf{S}_n\boldsymbol{u}\rangle : \boldsymbol{u} \in \mathcal{C}(m, \gamma)\right\},$$

$$\rho_-(m, \gamma) = \inf_{\boldsymbol{u}}\left\{\langle \boldsymbol{u}, \mathbf{S}_n\boldsymbol{u}\rangle : \boldsymbol{u} \in \mathcal{C}(m, \gamma)\right\},$$

where $\mathcal{C}(m, \gamma) := \{\boldsymbol{u} \in \mathbb{S}^{d-1} : \forall J \subseteq \{1, \ldots, d\}$ satisfying $\mathcal{S} \subseteq J, |J| \leq m, \|\boldsymbol{u}_{J^c}\|_1 \leq \gamma\|\boldsymbol{u}_J\|_1\}$.

Condition 3. $\mathbf{S}_n$ satisfies the RE condition RE(k, γ), that is, $\kappa_l \leq \rho_-(k, \gamma) \leq \rho_+(k, \gamma) \leq \kappa_u$ for some constants $\kappa_u, \kappa_l > 0$.

To make Condition 2 on $\mathbf{H}_\tau$ practically useful, in what follows, we show that Condition 3 implies Condition 2 with high probability. As before, we write $v_\delta = n^{-1}\sum_{i=1}^n v_{i,\delta}$ and $L = \max_{1 \leq i \leq n}\|\boldsymbol{x}_i\|_\infty$.

Lemma 1. Condition 3 implies Condition 2 with high probability: if $0 < \kappa_l \leq \rho_-(k, \gamma) \leq \rho_+(k, \gamma) \leq \kappa_u < \infty$ for some $k \geq 1$ and $\gamma > 0$, then it holds with probability at least $1 - e^{-t}$ that, $0 < \kappa_l/2 \leq \kappa_-(k, \gamma, r) \leq \kappa_+(k, \gamma, r) \leq \kappa_u < \infty$ provided $\tau \geq \max\{8Lr, c_1(L^2kv_\delta)^{1/(1+\delta)}\}$ and $n \geq c_2L^4k^2t$, where $c_1, c_2 > 0$ are constants depending only on (γ, κ_l).

With the above preparations in place, we are now ready to present the main results on the adaptive Huber estimator in high dimensions.

Theorem 3 (Upper bound in high dimensions). Assume Condition 3 holds with $(k, \gamma) = (2s, 3)$, $v_\delta < \infty$ for some $0 < \delta \leq 1$. For any $t > 0$ and $\tau_0 \geq \nu_\delta$, let $\tau = \tau_0(n/t)^{\max\{1/(1+\delta),1/2\}}$ and $\lambda \geq 4L\tau_0(t/n)^{\min\{\delta/(1+\delta),1/2\}}$. Then with probability at least $1-(2s+1)e^{-t}$, the ℓ_1-regularized Huber estimator $\hat{\boldsymbol{\beta}}_{\tau,\lambda}$ defined in (4) satisfies

$$\left\|\hat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\right\|_2 \leq 3\kappa_l^{-1}s^{1/2}\lambda, \tag{9}$$

as long as $n \geq C(L, \kappa_l)s^2t$ for some $C(L, \kappa_l)$ depending only on (L, κ_l). In particular, with $t = (1 + c)\log d$ for $c > 0$ we have

$$\left\|\hat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\right\|_2 \lesssim \kappa_l^{-1}L\tau_0\, s^{1/2}\left\{\frac{(1+c)\log d}{n}\right\}^{\min\{\delta/(1+\delta),1/2\}} \tag{10}$$

with probability at least $1 - d^{-c}$.

The above result demonstrates that the regularized Huber estimator with an adaptive robustification parameter converges at the rate $s^{1/2}\{(\log d)/n\}^{\min\{\delta/(1+\delta),1/2\}}$ with overwhelming probability. Provided the observation noise has finite variance, the proposed estimator performs as well as the Lasso with sub-Gaussian errors. We advocate the adaptive Huber regression method since sub-Gaussian condition often fails in practice (Wang, Peng, and Li 2015; Eklund, Nichols, and Knutsson 2016).

Remark 5. As pointed out by a reviewer, if one pursues a sparsity-adaptive approach, such as the SLOPE (Bogdan et al.

2015; Bellec, Lecué, and Tsybakov 2018), the upper bound on ℓ_2-error can be improved from $\sqrt{s\log(d)/n}$ to $\sqrt{s\log(ed/s)/n}$. With heavy-tailed observation noise, it is interesting to investigate whether this sharper bound can be achieved by Huber-type regularized estimator. We leave this to future work as a significant amount of additional work is still needed. On the other hand, since $\log(ed/s) = 1 + \log d - \log s$ and $s \le n$, $\log(ed/s)$ scales the same as $\log d$ so long as $\log d > a \log n$ for some $a > 1$.

Remark 6. Analogously to the low-dimensional case, here we impose the sample size scaling $n \gtrsim s^2 \log d$ under fixed designs. In the supplementary materials, we obtain minimax optimal ℓ_1-, ℓ_2- and prediction error bounds for $\hat{\boldsymbol{\beta}}_{\tau,\lambda}$ with random designs under the scaling $n \gtrsim s \log d$.

Finally, we establish a matching lower bound for estimating $\boldsymbol{\beta}^*$. Recall the definition of $\mathcal{U}_n$ in Theorem 2.

Theorem 4 (Lower bound in high dimensions). Assume that ε_i are independent from some distribution in $\mathcal{P}_\delta^{v_\delta}$. Suppose that Condition 3 holds with $k = 2s$ and $\gamma = 0$. Further assume that there exists a set $\mathcal{A}$ with $|\mathcal{A}| = s$ and $\mathbf{u} \in \mathcal{U}_n$ such that $\|\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{u}/n\|_{\min} \ge \alpha$ for some $\alpha > 0$. Then, for any $A > 0$ and s-sparse estimator $\widehat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(y_1, \ldots, y_n, A)$ possibly depending on A, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_\delta^{v_\delta}} \mathbb{P}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \ge v_\delta \frac{\alpha s^{1/2}}{\kappa_u}\left(\frac{A\log d}{2n}\right)^{\min\{\delta/(1+\delta),1/2\}}\right] \ge 2^{-1}d^{-A},$$

as long as $n \ge 2(A\log d + \log 2)$.

Together, Theorems 3 and 4 show that the regularized adaptive Huber estimator achieves the optimal rate of convergence in ℓ_2-error. The proof, which is given in the supplementary materials, involves constructing a subclass of binomial distributions for the regression errors. Unifying the results in low and high dimensions, we arrive at the claim (5) and thus the phase transition in Figure 1.

4. Extension to Heavy-Tailed Designs

In this section, we extend the idea of adaptive Huber regression described in Section 2 to the case where both the covariate vector $\boldsymbol{x}$ and the regression error ε exhibit heavy tails. We focus on the high-dimensional regime $d \gg n$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0 \ll n$. Observe that, for Huber regression, the linear part of the Huber loss penalizes the residuals, and therefore robustifies the quadratic loss in the sense that outliers in the response space (caused by heavy-tailed observation noise) are down weighted or removed. Since no robustification is imposed on the covariates, intuitively, the adaptive Huber estimator may not be robust against heavy-tailed covariates. In what follows, we modify the adaptive Huber regression to robustify both the covariates and regression errors.

To begin with, suppose we observe independent data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from $(y, \boldsymbol{x})$, which follows the linear model $y = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \varepsilon$. To robustify $\boldsymbol{x}_i$, we define truncated covariates $\boldsymbol{x}_i^{\varpi} = (\psi_\varpi(x_{i1}), \ldots, \psi_\varpi(x_{id}))^{\mathrm{T}}$, where $\psi_\varpi(x) := \min\{\max(-\varpi, x), \varpi\}$ and $\varpi > 0$ is a tuning parameter. Then we consider the modified adaptive Huber estimator (see Fan et al. (2016) for a general robustification principle)

$$\hat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{\mathcal{L}_\tau^\varpi(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\right\}, \qquad (11)$$

where $\mathcal{L}_\tau^\varpi(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \ell_\tau(y_i - \langle \boldsymbol{x}_i^\varpi, \boldsymbol{\beta}\rangle)$ and $\lambda > 0$ is a regularization parameter.

Let $\mathcal{S}$ be the true support of $\boldsymbol{\beta}^*$ with sparsity $|\mathcal{S}| = s$, and denote by $\mathbf{H}_\tau^\varpi(\boldsymbol{\beta}) = \nabla^2\mathcal{L}_\tau^\varpi(\boldsymbol{\beta})$ the Hessian matrix of the modified Huber loss. To investigate the deviation property of $\hat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$, we impose the following mild moment assumptions.

Condition 4. (i) $\mathbb{E}(\varepsilon) = 0$, $\sigma^2 = \mathbb{E}(\varepsilon^2) > 0$ and $v_3 := \mathbb{E}(\varepsilon^4) < \infty$; (ii) the covariate vector $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ is independent of ε and satisfies $M_4 := \max_{1\le j\le d}\mathbb{E}(x_j^4) < \infty$.

We are now in place to state the main result of this section. Theorem 5 demonstrates that the modified adaptive Huber estimator admits exponentially fast concentration when the convariates only have finite fourth moments, although at the cost of stronger scaling conditions.

Theorem 5. Assume Condition 4 holds and let $\mathbf{H}_\tau^\varpi(\cdot)$ satisfy Condition 2 with $k = 2s$, $\gamma = 3$ and $r > 12\kappa_l^{-1}\lambda s$. Then, the modified adaptive Huber estimator $\hat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$ given in (11) satisfies, on the event $\mathcal{E}(\tau, \varpi, \lambda) = \left\{\|(\nabla\mathcal{L}_\tau^\varpi(\boldsymbol{\beta}^*))_{\mathcal{S}}\|_\infty \le \lambda/2\right\}$, that

$$\left\|\hat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda} - \boldsymbol{\beta}^*\right\|_2 \le 3\kappa_l^{-1}s^{1/2}\lambda.$$

For any $t > 0$, let the triplet (τ, ϖ, λ) satisfy

$$\lambda \ge 2M_4\|\boldsymbol{\beta}^*\|_2\, s^{1/2}\varpi^{-2} + 8\left\{v_2 M_2^{1/2} + M_4\|\boldsymbol{\beta}^*\|_2^3\, s^{3/2}\right\}\tau^{-2} + 2\left(2\sigma^2 M_2 + 2M_4\|\boldsymbol{\beta}^*\|_2^2\, s\right)^{1/2}\sqrt{\frac{t}{n}} + \varpi\tau\frac{t}{n}, \qquad (12)$$

where $v_2 = \mathbb{E}(|\varepsilon|^3)$ and $M_2 = \max_{1\le j\le d}\mathbb{E}(x_j^2)$. Then $\mathbb{P}\{\mathcal{E}(\tau, \varpi, \lambda)\} \ge 1 - 2se^{-t}$.

Remark 7. Assume that the quantities v_3, M_4, and $\|\boldsymbol{\beta}^*\|_2$ are all bounded. Taking $t \asymp \log d$ in (12), we see that $\hat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$ achieves a near-optimal convergence rate of order $s\sqrt{(\log d)/n}$ when the parameters (τ, ϖ, λ) scale as

$$\tau \asymp s^{1/2}\left(\frac{n}{\log d}\right)^{1/4}, \quad \varpi \asymp \left(\frac{n}{\log d}\right)^{1/4}, \quad \text{and } \lambda \asymp \sqrt{\frac{s\log d}{n}}.$$

We remark here that the theoretically optimal τ is different from that in the sub-Gaussian design case. See Theorem B.2 in the supplementary materials.

5. Algorithm and Implementation

This section is devoted to computational algorithm and numerical implementation. We focus on the regularized adaptive Huber regression in (4), as (3) can be easily solved via the iteratively reweighted least squares method. To solve the convex optimization problem in (4), standard optimization algorithms, such as the cutting-plane or interior point method, are not scalable to large-scale problems.

In what follows, we describe a fast and easily implementable method using the local adaptive majorize-minimization (LAMM) principle (Fan et al. 2018). We say that a function $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ majorizes $f(\boldsymbol{\beta})$ at the point $\boldsymbol{\beta}^{(k)}$ if

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) \geq f(\boldsymbol{\beta}) \quad \text{and} \quad g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}).$$

To minimize a general function $f(\boldsymbol{\beta})$, a majorize-minimization (MM) algorithm initializes at $\boldsymbol{\beta}^{(0)}$, and then iteratively computes $\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ for $k = 0, 1, \ldots$. The objective value of such an algorithm decreases in each step, since

$$f(\boldsymbol{\beta}^{(k+1)}) \overset{\text{major.}}{\leq} g(\boldsymbol{\beta}^{(k+1)} \mid \boldsymbol{\beta}^{(k)}) \overset{\text{min.}}{\leq} g(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\beta}^{(k)}) \overset{\text{init.}}{=} f(\boldsymbol{\beta}^{(k)}). \tag{13}$$

As pointed out by Fan et al. (2018), the majorization requirement only needs to hold locally at $\boldsymbol{\beta}^{(k+1)}$ when starting from $\boldsymbol{\beta}^{(k)}$. We therefore locally majorize $\mathcal{L}_\tau(\boldsymbol{\beta})$ in (4) at $\boldsymbol{\beta}^{(k)}$ by an isotropic quadratic function

$$g_k(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}) + \langle \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}),\, \boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}\rangle + \frac{\phi_k}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}\|_2^2,$$

where ϕ_k is a quadratic parameter such that $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$. The isotropic form also allows a simple analytic solution to the subsequent majorized optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \langle \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \rangle + \frac{\phi_k}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{14}$$

It can be shown that (14) is minimized at

$$\boldsymbol{\beta}^{(k+1)} = T_{\lambda,\phi_k}(\boldsymbol{\beta}^{(k)}) = S\Big(\boldsymbol{\beta}^{(k)} - \phi_k^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}), \phi_k^{-1}\lambda\Big),$$

where $S(\mathbf{x}, \lambda)$ is the soft-thresholding operator defined by $S(\mathbf{x}, \lambda) = \operatorname{sign}(x_j)\max(|x_j| - \lambda, 0)$. The simplicity of this updating rule is due to the fact that (14) is an unconstrained optimization problem.

To find the smallest ϕ_k such that $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$, the basic idea of LAMM is to start from a relatively small isotropic parameter $\phi_k = \phi_k^0$ and then successfully inflate ϕ_k by a factor $\gamma_u > 1$, say $\gamma_u = 2$. If the solution satisfies $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$, we stop and obtain $\boldsymbol{\beta}^{(k+1)}$, which makes the target value nonincreasing. We then continue with the iteration to produce next solution until the solution sequence $\{\boldsymbol{\beta}^{(k)}\}_{k=1}^\infty$ converges. A simple stopping criterion is $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2 \leq \epsilon$ for a sufficiently small ϵ, say 10^{-4}. We refer to Fan et al. (2018) for a detailed complexity analysis of the LAMM algorithm.

Algorithm 1 LAMM algorithm for regularized adaptive Huber regression.

```
Algorithm: $\{\boldsymbol{\beta}^{(k)}, \phi_k\}_{k=1}^\infty \leftarrow \text{LAMM}(\lambda, \boldsymbol{\beta}^{(0)}, \phi_0, \epsilon)$
Input: $\lambda, \boldsymbol{\beta}^{(0)}, \phi_0, \epsilon$
Initialize: $\phi^{(\ell,k)} \leftarrow \max\{\phi_0, \gamma_u^{-1}\phi^{(\ell,k-1)}\}$
for $k = 0, 1, \ldots$ until $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2 \leq \epsilon$ do
    Repeat
        $\boldsymbol{\beta}^{(k+1)} \leftarrow T_{\lambda,\phi_k}(\boldsymbol{\beta}^{(k)})$
        If $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) < \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$ then $\phi_k \leftarrow \gamma_u\phi_k$
    Until $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$
    Return $\{\boldsymbol{\beta}^{(k+1)}, \phi_k\}$
end for
Output: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(k+1)}$
```

6. Numerical Studies

6.1. Tuning Parameter and Finite Sample Performance

For numerical studies and real data analysis, in the case where the actual order of moments is unspecified, we presume the variance is finite and therefore choose robustification and regularization parameters as follows

$$\tau = c_\tau \times \widehat{\sigma}\left(\frac{n_{\text{eff}}}{t}\right)^{1/2} \quad \text{and} \quad \lambda = c_\lambda \times \widehat{\sigma}\left(\frac{n_{\text{eff}}}{t}\right)^{1/2},$$

where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$ with $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ serves as a crude preliminary estimate of σ^2, and the parameter t controls the confidence level. We set $t = \log n$ for simplicity except for the phase transition plot. The constant c_τ and c_λ are chosen via 3-fold cross-validation from a small set of constants, say $\{0.5, 1, 1.5\}$.

We generate data from the linear model

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where ε_i are iid regression errors and $\boldsymbol{\beta}^* = (5, -2, 0, 0, 3, \underbrace{0, \ldots, 0}_{d-5})^{\mathrm{T}} \in \mathbb{R}^d$. Independent of ε_i, we generate $\boldsymbol{x}_i$ from standard multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. In this section, we set $(n, d) = (100, 5)$, and generate regression errors from three different distributions: the normal distribution $\mathcal{N}(0, 4)$, the t-distribution with degrees of freedom 1.5, and the log-normal distribution $\log\mathcal{N}(0, 4)$. Both t and log-normal distributions are heavy-tailed and produce outliers with high chance.

The results on ℓ_2-error for adaptive Huber regression and the least squares estimator, averaged over 100 simulations, are summarized in Table 1. In the case of normally distributed noise, the adaptive Huber estimator performs as well as the least squares. With heavy-tailed regression errors following Student's t or log-normal distribution, the adaptive Huber regression significantly outperforms the least squares. These empirical results reveal that adaptive Huber regression prevails across various scenarios: not only it provides more reliable estimators in the presence of heavy-tailed and/or asymmetric errors, but also loses almost no efficiency at the normal model.

Table 1. Results for adaptive Huber regression (AHR) and ordinary least squares (OLS) when $n = 100$ and $d = 5$.

Noise	AHR		OLS	
	Mean	SD	Mean	SD
Normal	0.566	0.189	0.567	0.191
Student's *t*	0.806	0.651	1.355	2.306
Log-normal	3.917	3.740	8.529	13.679

NOTE: The mean and SD of ℓ_2-error based on 100 simulations are reported.

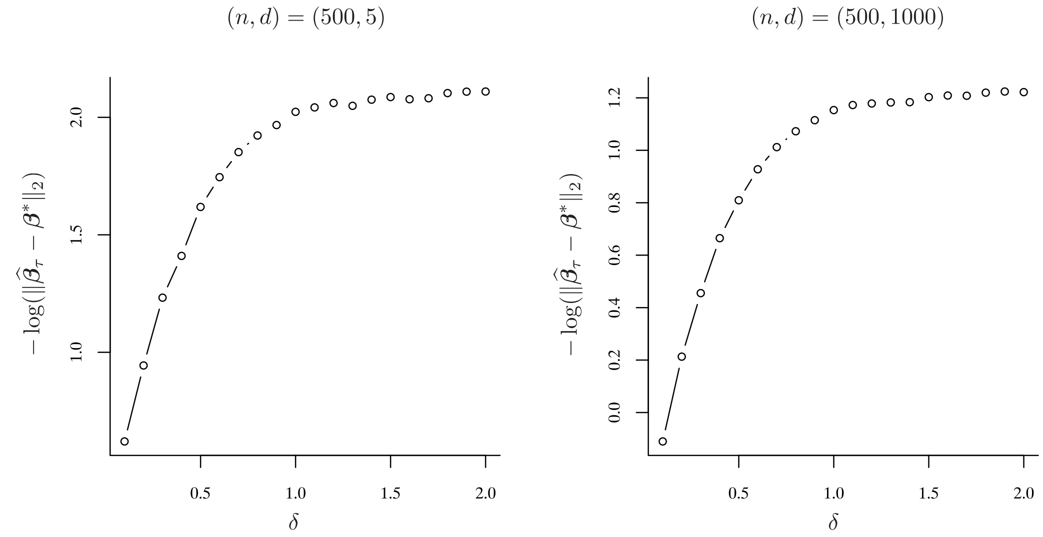

Figure 2. Negative log ℓ_2-error versus δ in low (left panel) and high (right panel) dimensions.

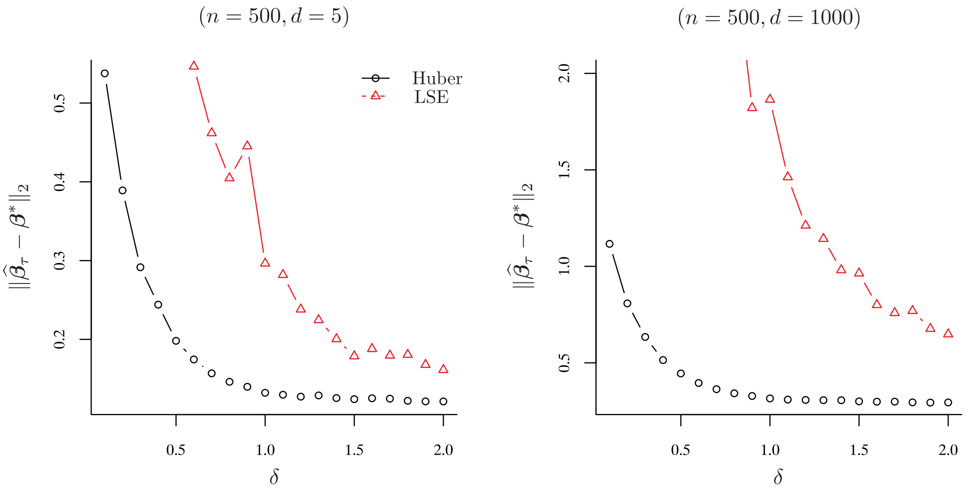

Figure 3. Comparison between the (regularized) adaptive Huber estimator and the (regularized) least squares estimator under ℓ_2-error.

6.2. Phase Transition

In this section, we validate the phase transition behavior of $\|\widehat{\boldsymbol\beta}_\tau - \boldsymbol\beta^*\|_2$ empirically. We generate continuous responses according to (15), where $\boldsymbol\beta^*$ and $\boldsymbol{x}_i$ are set the same way as before. We sample independent errors as $\varepsilon_i \sim t_{\mathrm{df}}$, Student's t-distribution with df degrees of freedom. Note that t_{df} has finite $(1+\delta)$th moments provided $\delta < \mathrm{df} - 1$ and infinite dfth moment. Therefore, we take $\delta = \mathrm{df} - 1 - 0.05$ throughout.

In low dimensions, we take $(n, d) = (500, 5)$ and a sequence of degrees of freedoms (dfs): $\mathrm{df} \in \{1.1, 1.2, \ldots, 3.0\}$; in high dimensions, we take $(n, d) = (500, 1000)$, with the same choice of dfs. Tuning parameters (τ, λ) are calibrated similarly as before. Indicated by the main theorems, it holds

1. (Low dimension):

$$-\log\left(\|\widehat{\boldsymbol\beta}_\tau - \boldsymbol\beta^*\|_2\right) \asymp \frac{\delta}{1+\delta}\log(n) - \frac{1}{1+\delta}\log(v_\delta),$$
$$0 < \delta \le 1,$$

2. (High dimension):

$$-\log\left(\|\widehat{\boldsymbol\beta}_\tau - \boldsymbol\beta^*\|_2\right) \asymp \frac{\delta}{1+\delta}\log\left(\frac{n}{\log d}\right) - \frac{1}{1+\delta}\log(v_\delta),$$
$$0 < \delta \le 1,$$

which are approximately $\log(n) \times \delta/(1+\delta)$ and $\log(n/\log d) \times \delta/(1+\delta)$, respectively, when n is sufficiently large.

Figure 2 displays the negative log ℓ_2-error versus δ in both low and high dimensions over 200 repetitions for each (n, d) combination. The empirically fitted curve closely resembles the theoretical curve displayed in Figure 1. These numerical results are in line with the theoretical findings and empirically validate the phase transition of the adaptive Huber estimator.

We also compared the ℓ_2-error of the adaptive Huber estimator with that of the OLS estimator for t-distributed errors with varying degrees of freedoms. As shown in Figure 3, adaptive Huber exhibits a significant advantage especially when δ is small. The OLS slowly catches up as δ increases.

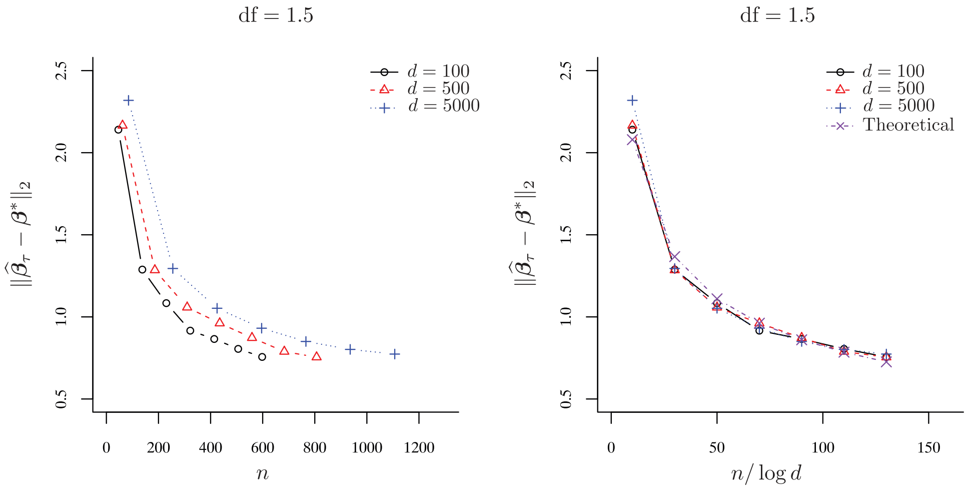

Figure 4. The ℓ_2-error versus sample size n (left panel) and the ℓ_2-error versus effective sample size $n_{\text{eff}} = n/\log d$ (right panel).

6.3. *Effective Sample Size*

In this section, we verify the scaling behavior of $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2$ with respect to the effective sample size. The data are generated in the same way as before except that the errors are drawn from $t_{1.5}$. As discussed in the previous subsection, we take $\delta = 0.45$ and then choose the robustification parameter as $\tau = c_\tau \widehat{v}_\delta (n/\log d)^{1/(1+\delta)}$, where $\widehat{v}_\delta$ is the $(1+\delta)$th sample absolute central moment. For simplicity, we take $c_\tau = 0.5$ here since our goal is to demonstrate the scaling behavior as n grows, instead of to achieve the best finite-sample performance.

The left panel of Figure 4 plots the ℓ_2-error $\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\|_2$ versus sample size over 200 repetitions when the dimension $d \in \{100, 500, 5000\}$. In all three settings, the ℓ_2-error decays as the sample size grows. As expected, the curves shift to the right when the dimension increases. Theorem 3 provides a specific prediction about this scaling behavior: if we plot the ℓ_2-error versus effective sample size $(n/\log d)$, the curves should align roughly with the theoretical curve

$$\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\|_2 \asymp \left(\frac{n}{\log d}\right)^{-\delta/(1+\delta)}$$

for different values of d. This is validated empirically by the right panel of Figure 4. This near-perfect alignment in Figure 4 is also observed by Wainwright (2009) for Lasso with sub-Gaussian errors.

6.4. *A Real Data Example: NCI-60 Cancer Cell Lines*

We apply the proposed methodologies to the NCI-60, a panel of 60 diverse human cancel cell lines. The NCI-60 consists of data on 60 human cancer cell lines and can be downloaded from *http://discover.nci.nih.gov/cellminer/*. More details on data acquisition can be found in Shankavaram et al. (2007). Our aim is to investigate the effects of genes on protein expressions. The gene expression data were obtained with an Affymetrix HG-U133A/B chip, $\log_2$ transformed and normalized with the guanine dytosine robust multi-array analysis. We then combined the same gene expression variables measured by multiple different probes into one by taking their median, resulting in a set of $p = 17{,}924$ predictors. The protein expressions based on 162 antibodies were acquired via reverse-phase protein lysate arrays in their original scale. One observation had to be removed since all values were missing in the gene expression data, reducing the number of observations to $n = 59$.

We first center all the protein and gene expression variables to have mean zero, and then plot the histograms of the kurtosises of all expressions in Figure 5. The left panel in the figure shows that 145 out of 162 protein expressions have kurtosises larger than 3; and 49 larger than 9. In other words, more than 89.5% of the protein expression variables have tails heavier than the normal distribution, and about 30.2% are severely heavy-tailed with tails flatter than t_5, the t-distribution with 5 degrees of freedom. Similarly, about 36.5% of the gene expression variables, even after the $\log_2$-transformation, still exhibit empirical kurtosises larger than that of t_5. This suggests that, regardless of the normalization methods used, genomic data can still exhibit heavy-tailedness, which was also pointed out by Purdom and Holmes (2005).

We order the protein expression variables according to their scales, measured by the SD. We show the results for the protein expressions based on the KRT19 antibody, the protein keratin 19, which constitutes the variable with the largest SD, serving as one dependent variable. KRT19, a type I keratin, also known as Cyfra 21-1, is encoded by the *KRT19* gene. Due to its high sensitivity, the KRT19 antibody is the most used marker for the tumor cells disseminated in lymph nodes, peripheral blood, and bone marrow of breast cancer patients (Nakata et al. 2004). We denote the adaptive Huber regression as AHuber, and that with truncated covariates as TAHuber. We then compare AHuber and TAHuber with Lasso. Both regularization and robustification parameters are chosen by the ten-fold cross-validation.

To measure the predictive performance, we consider a robust prediction loss: the mean absolute error (MAE) defined as

$$\mathrm{MAE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| y_i^{\text{test}} - \langle \boldsymbol{x}_i^{\text{test}}, \widehat{\boldsymbol{\beta}} \rangle \right|,$$

where y_i^{test} and $\boldsymbol{x}_i^{\text{test}}$, $i = 1, \ldots, n_{\text{test}}$, denote the observations of the response and predictor variables in the test data, respectively. We report the MAE via the leave-one-out cross-validation. Table 2 reports the MAE, model size, and selected

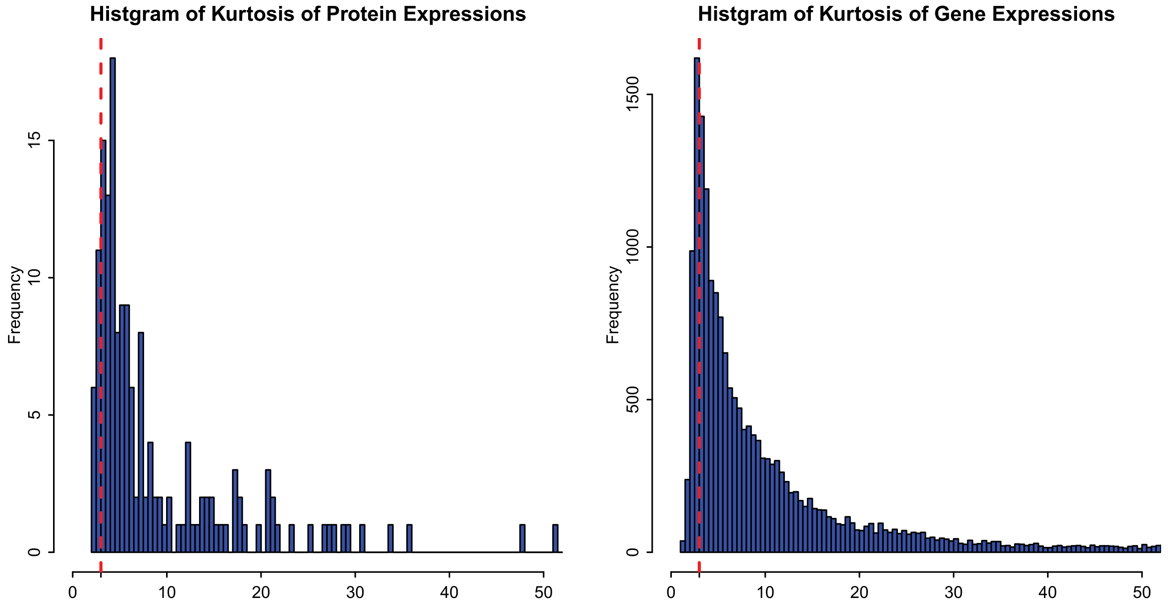

Figure 5. Histogram of kurtosises for the protein and gene expressions. The dashed red line at 3 is the kurtosis of a normal distribution.

Table 2. We report the mean absolute error (MAE) for protein expressions based on the KRT19 antibody from the NCI-60 cancer cell lines, computed from leave-one-out cross-validation.

Method	MAE	Size	Selected genes
Lasso	7.64	42	*FBLIM1, MT1E, EDN2, F3, FAM102B, S100A14, LAMB3, EPCAM, FN1, TM4SF1, UCHL1, NMU, ANXA3, PLAC8, SPP1, TGFBI, CD74, GPX3, EDN1, CPVL, NPTX2, TES, AKR1B10, CA2, TSPYL5, MAL2, GDA, BAMBI, CST6, ADAMTS15, DUSP6, BTG1, LGALS3, IFI27, MEIS2, TOX3, KRT23, BST2, SLPI, PLTP, XIST, NGFRAP1*
AHuber	6.74	11	*MT1E, ARHGAP29, CPCAM, VAMP8, MALL, ANXA3, MAL2, BAMBI, LGALS3, KRT19, TFF3*
TAHuber	5.76	7	*MT1E, ARHGAP29, MALL, ANXA3, MAL2, BAMBI, KRT19*

NOTE: We also report the model size and selected genes for each method.

genes for the considered methods. TAHuber clearly shows the smallest MAE, followed by AHuber and Lasso. The Lasso produces a fairly large model despite the small sample. Now it has been recognized that Lasso tends to select many noise variables along with the significant ones, especially when data exhibit heavy tails.

The Lasso selects a model with 42 genes but excludes the *KRT19* gene, which encodes the protein keratin 19. AHuber finds 11 genes including *KRT19*. TAHuber results in a model with 7 genes: *KRT19*, *MT1E*, *ARHGAP29*, *MALL*, *ANXA3*, *MAL2*, *BAMBI*. First, *KRT19* encodes the keratin 19 protein. It has been reported in Wu et al. (2008) that the *MT1E* expression is positively correlated with cancer cell migration and tumor stage, and the *MT1E* isoform was found to be present in estrogen receptor-negative breast cancer cell lines (Friedline et al. 1998). *ANXA3* is highly expressed in all colon cell lines and all breast-derived cell lines positive for the oestrogen receptor (Ross et al. 2000). A very recent study in Zhou et al. (2017) suggested that silencing the *ANXA3* expression by RNA interference inhibits the proliferation and invasion of breast cancer cells. Moreover, studies in Shangguan et al. (2012) and Kretzschmar (2000) showed that the *BAMBI* transduction significantly inhibited TGF-β/Smad signaling and expression of carcinoma-associated fibroblasts in human bone marrow mesenchymal stem cells (BM-MSCs), and disrupted the cytokine network mediating the interaction between MSCs and breast cancer cells. Consequently, the *BAMBI* transduction abolished protumor effects of BM-MSCs in vitro and in an orthotopic breast cancer xenograft model, and instead significantly inhibited growth and metastasis of coinoculated cancer. *MAL2* expressions were shown to be elevated at both RNA and protein levels in breast cancer (Shehata et al. 2008). It has also been shown that *MALL* is associated with various forms of cancer (Oh et al. 2005; Landi et al. 2014). However, the effect of *ARHGAP29* and *MALL* on breast cancer remains unclear and is worth further investigation.

Supplementary Materials

In the supplementary materials, we provide theoretical analysis under random designs, and proofs of all the theoretical results in this article.

Acknowledgments

The authors thank the editor, associate editor, and two anonymous referees for their valuable comments.

Funding

This work is supported by a Connaught Award, NSERC Grant RGPIN-2018-06484, NSF Grants DMS-1662139, DMS-1712591, and DMS-1811376, NIH Grant 2R01-GM072611-14, and NSFC Grant 11690014.

References

Alquier, P., Cottet, V., and Lecué, G. (2017), "Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures With Lipschitz Loss Functions," arXiv no. 1702.01402. [255]

Belloni, A., and Chernozhukov, V. (2011), "ℓ_1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130. [255]

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2018), "Slope Meets Lasso: Improved Oracle Bounds and Optimality," *The Annals of Statistics*, 46, 3603–3642. [259]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [254]

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015), "SLOPE—Adaptive Variable Selection via Convex Optimization," *The Annals of Applied Statistics*, 9, 1103–1140. [259]

Brownlees, C., Joly, E., and Lugosi, G. (2015), "Empirical Risk Minimization for Heavy-Tailed Losses," *The Annals of Statistics*, 43, 2507–2536. [255]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Heidelberg: Springer. [254]

Catoni, O. (2012), "Challenging the Empirical Mean and Empirical Variance: A Deviation Study," *Annales de I'Institut Henri Poincaré—Probabilités et Statistiques*, 48, 1148–1185. [254,255,256]

——— (2016), "PAC-Bayesian Bounds for the Gram Matrix and Least Squares Regression With a Random Design," arXiv no. 1603.05229. [255]

Chen, M., Gao, C., and Ren, Z. (2018), "Robust Covariance and Scatter Matrix Estimation Under Huber's Contamination Model," *The Annals of Statistics*, 46, 1932–1960. [255]

Cont, R. (2001), "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues," *Quantitative Finance*, 1, 223–236. [254]

Delaigle, A., Hall, P., and Jin, J. (2011), "Robustness and Accuracy of Methods for High Dimensional Data Analysis Based on Student's t-Statistic," *Journal of the Royal Statistical Society*, Series B, 73, 283–301. [255]

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016), "Sub-Gaussian Mean Estimators," *The Annals of Statistics*, 44, 2695–2725. [255,257]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [254]

Eklund, A., Nichols, T., and Knutsson, H. (2016), "Cluster Failure: Why fMRI Inferences for Spatial Extent Have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7900–7905. [254,258]

Fan, J., Fan, Y., and Barut, E. (2014), "Adaptive Robust Variable Selection," *The Annals of Statistics*, 42, 324–351. [255]

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, 247–265. [255,256]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [254]

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018), "I-LAMM for Sparse Learning: Simultaneous Control of Algorithmic Complexity and Statistical Error," *The Annals of Statistics*, 96, 1348–1360. [254,255,258,260]

Fan, J., Wang, W., and Zhu, Z. (2016), "A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery," arXiv no. 1603.08315. [254,259]

Friedline, J. A., Garrett, S. H., Somji, S., Todd, J. H., and Sens, D. A. (1998), "Differential Expression of the MT-1E Gene in Estrogen-Receptor-Positive and -Negative Human Breast Cancer Cell Lines," *The American Journal of Pathology*, 152, 23–27. [263]

Giulini, I. (2017), "Robust PCA and Pairs of Projections in a Hilbert Space," *Electronic Journal of Statistics*, 11, 3903–3926. [255]

Hastie, T., Tibshirani, R., and Wainwright, M. J. (2015), *Statistical Learning With Sparsity: The Lasso and Generalizations*, Boca Raton, FL: CRC Press. [254]

He, X., and Shao, Q.-M. (1996), "A General Bahadur Representation of M-Estimators and Its Application to Linear Regression With Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630. [254]

——— (2000), "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 120–135. [254]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [255,256]

——— (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [254,255,256]

Koenker, R. (2005), *Quantile Regression*, New York: Cambridge University Press. [255]

Kretzschmar, M. (2000), "Transforming Growth Factor-β and Breast Cancer: Transforming Growth Factor-β/Smad Signaling Defects and Cancer," *Breast Cancer Research*, 2, 107–115. [263]

Landi, A., Vermeire, J., Iannucci, V., Vanderstraeten, H., Naessens, E., Bentahir, M., and Verhasselt, B. (2014), "Genome-Wide shRNA Screening Identifies Host Factors Involved in Early Endocytic Events for HIV-1-Induced CD4 Down-Regulation," *Retrovirology*, 11, 118–129. [263]

Lepski, O. V. (1991), "Asymptotically Minimax Adaptive Estimation. I. Upper Bounds. Optimally Adaptive Estimates," *IEEE Transactions on Information Theory*, 36, 682–697. [257]

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414. [255]

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference" (with discussion and a rejoinder by Liu and Singh), *The Annals of Statistics*, 27, 783–858. [255]

Loh, P., and Wainwright, M. J. (2015), "Regularized M-Estimators With Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *Journal of Machine Learning Research*, 16, 559–616. [254,255]

Mammen, E. (1989), "Asymptotics With Increasing Dimension for Robust Regression With Applications to the Bootstrap," *The Annals of Statistics*, 17, 382–400. [254]

Minsker, S. (2018), "Sub-Gaussian Estimators of the Mean of a Random Matrix With Heavy-Tailed Entries," *The Annals of Statistics*, 46, 2871–2903. [255]

Mizera, I. (2002), "On Depth and Deep Points: A Calculus," *The Annals of Statistics*, 30, 1681–1736. [255]

Mizera, I., and Müller, C. H. (2004), "Location-Scale Depth," *Journal of the American Statistical Association*, 99, 949–966. [255]

Nakata, B., Takashima, T., Ogawa, Y., Ishikawa, T., and Hirakawa, K. (2004), "Serum CYFRA 21-1 (Cytokeratin-19 Fragments) Is a Useful Tumour Marker for Detecting Disease Relapse and Assessing Treatment Efficacy in Breast Cancer," *British Journal of Cancer*, 91, 873–878. [262]

Oh, J. H., Yang, J. O., Hahn, Y., Kim, M. R., Byun, S. S., Jeon, Y. J., Kim, J. M., Song, K. S., Noh, S. M., Kim, S., and Yoo, H. S. (2005), "Transcriptome Analysis of Human Gastric Cancer," *Mammalian Genome*, 16, 942–954. [263]

Portnoy, S. (1985), "Asymptotic Behavior of M Estimators of p Regression Parameters When p^2/n Is Large; II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417. [254]

Purdom, E., and Holmes, S. P. (2005), "Error Distribution for Gene Expression Data," *Statistical Applications in Genetics and Molecular Biology*, 4, 16. [262]

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, W., Jeffrey, S. S., Van de Rijn, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, 24, 227–235. [263]

Shangguan, L., Ti, X., Krause, U., Hai, B., Zhao, Y., Yang, Z., and Liu, F. (2012), "Inhibition of TGF-β/Smad Signaling by BAMBI Blocks Differentiation of Human Mesenchymal Stem Cells to Carcinoma-Associated Fibroblasts and Abolishes Their Protumor Effects," *Stem Cells*, 30, 2810–2819. [263]

Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U. Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N. (2007), "Transcript and Protein Expression Profiles of the NCI-60 Cancer Cell Panel: An Integromic Microarray Study," *Molecular Cancer Therapeutics*, 40, 2877–2909. [262]

Shehata, M., Bièche, I., Boutros, R., Weidenhofer, J., Fanayan, S., Spalding, L., Zeps, N., Byth, K., Bright, R. K., Lidereau, R., and Byrne, J. A. (2008), "Nonredundant Functions for Tumor Protein D52-Like Proteins Support Specific Targeting of TPD52," *Clinical Cancer Research*, 14, 5050–5060. [263]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [254]

Tukey, J. W. (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians* (Vol. 2), pp. 523–531. [255]

Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [262]

Wang, L. (2013), "The L_1 Penalized LAD Estimator for High Dimensional Linear Regression," *Journal of Multivariate Analysis*, 120, 135–151. [255]

Wang, L., Peng, B., and Li, R. (2015), "A High-Dimensional Nonparametric Multivariate Test for Mean Vector," *Journal of the American Statistical Association*, 110, 1658–1669. [254,258]

Wu, Y., Siadaty, M. S., Berens, M. E., Hampton, G. M., and Theodorescu, D. (2008), "Overlapping Gene Expression Profiles of Cell Migration and Tumor Invasion in Human Bladder Cancer Identify Metallothionein E1 and Nicotinamide N-Methyltransferase as Novel Regulators of Cell Migration," *Oncogene*, 27, 6679–6689. [263]

Yohai, V. J., and Maronna, R. A. (1979), "Asymptotic Behavior of *M*-Estimators for the Linear Model," *The Annals of Statistics*, 7, 258–268. [254]

Zheng, Q., Peng, L., and He, X. (2015), "Globally Adaptive Quantile Regression With Ultra-High Dimensional Data," *The Annals of Statistics*, 43, 2225–2258. [255]

Zhou, T., Li, Y., Yang, L., Liu, L., Ju, Y., and Li, C. (2017), "Silencing of ANXA3 Expression by RNA Interference Inhibits the Proliferation and Invasion of Breast Cancer Cells," *Oncology Reports*, 37, 388–398. [263]

Zuo, Y., and Serfling, R. (2000), "General Notions of Statistical Depth Function," *The Annals of Statistics*, 28, 461–482. [255]