The Annals of Statistics
2020, Vol. 48, No. 3, 1665–1691
https://doi.org/10.1214/19-AOS1863

ROBUST INFERENCE VIA MULTIPLIER BOOTSTRAP

BY XI CHEN[1] AND WEN-XIN ZHOU[2]

[1]*Stern School of Business, New York University, xchen3@stern.nyu.edu*
[2]*Department of Mathematics, University of California, San Diego, wez243@ucsd.edu*

This paper investigates the theoretical underpinnings of two fundamental statistical inference problems, the construction of confidence sets and large-scale simultaneous hypothesis testing, in the presence of heavy-tailed data. With heavy-tailed observation noise, finite sample properties of the least squares-based methods, typified by the sample mean, are suboptimal both theoretically and empirically. In this paper, we demonstrate that the adaptive Huber regression, integrated with the multiplier bootstrap procedure, provides a useful robust alternative to the method of least squares. Our theoretical and empirical results reveal the effectiveness of the proposed method, and highlight the importance of having inference methods that are robust to heavy tailedness.

1. Introduction. In classical statistical analysis, it is typically assumed that data are drawn from a Gaussian distribution. Although the normality assumption has been widely adopted to facilitate methodological development and theoretical analysis, Gaussian models could be an idealization of the complex random world. The non-Gaussian, or even heavy-tailed, character of the distribution of empirical data has been repeatedly observed in various domains, ranging from genomics, medical imaging to economics and finance. New challenges are thus brought to classical 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. The necessity of robust alternatives to least squares was first noticed by Peter Huber in his seminal work "Robust Estimation of a Location Parameter" (Huber (1964)). Due to the growing complexity of modern data, the notion of robustness is becoming increasingly important in statistical analysis and finds its use in a wide range of applications. We refer to Huber and Ronchetti (2009) for an overview of robust statistics.

Although the past a few decades have witnessed the active development of rich statistical theory on robust estimation, robust statistical inference for heavy-tailed data has always been a challenging problem on which the extant literature has been somewhat silent. Fan, Hall and Yao (2007), Delaigle, Hall and Jin (2011) and Liu and Shao (2014) investigated robust inference that is confined to the Student's t-statistic. However, as pointed out by Devroye et al. (2016) (see Section 8 therein), sharp confidence estimation for heavy-tailed data in the finite sample set-up remains an open problem and a general methodology is still lacking. To that end, this paper makes a further step in studying confidence estimation from a robust perspective. In particular, under linear model with heavy-tailed errors, we address two fundamental problems: (1) confidence set construction for regression coefficients, and (2) large-scale multiple testing with the guarantee of false discovery rate control. The developed techniques provide mathematical underpinnings for a class of robust statistical inference problems. Moreover, sharp exponential-type bounds for the coverage probability of bootstrap confidence set are derived under weak moment assumptions.

Received March 2018; revised May 2019.
MSC2010 subject classifications. Primary 62F35, 62F40; secondary 62J05, 62J15.
Key words and phrases. Confidence set, heavy-tailed data, multiple testing, multiplier bootstrap, robust regression, Wilks' theorem.

1.1. *Confidence sets.* Consider the linear model $Y = \boldsymbol{X}^\intercal \boldsymbol{\theta}^* + \varepsilon$, where $Y \in \mathbb{R}$ is the response variable, $\boldsymbol{X} \in \mathbb{R}^d$ is the (random) vector of covariates, $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_d^*)^\intercal \in \mathbb{R}^d$ is the vector of regression coefficients and ε represents the regression error satisfying $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ and $\sigma^2 = \mathbb{E}(\varepsilon^2|\boldsymbol{X}) < \infty$. Assume that we observe a random sample $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ from $(Y, \boldsymbol{X})$:

$$Y_i = \boldsymbol{X}_i^\intercal \boldsymbol{\theta}^* + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

The intercept term is implicitly assumed in model (1.1) by taking the first element of $\boldsymbol{X}_i$ to be one so that the first element of $\boldsymbol{\theta}^*$ becomes the intercept. The least squares estimator and its variations have been widely adopted to estimate $\boldsymbol{\theta}^*$, which on many occasions achieve the minimax rate in terms of the mean squared error (MSE).

Although the MSE plays an important role in estimation, an estimator that is optimal in MSE might be suboptimal in terms of nonasymptotic deviation, which often relates to the construction of confidence intervals. For example, in the mean estimation problem, although the sample mean has an optimal minimax mean squared error among all mean estimators, its deviation is worse for non-Gaussian samples than for Gaussian ones, and the worst-case deviation is suboptimal when the sampling distribution has heavy tails (Catoni (2012)). More specifically, let $X_1, \ldots, X_n$ be independent random variables from X with mean μ and variance $\sigma^2 > 0$. Consider the empirical mean $\widehat{\mu}_n = (1/n)\sum_{i=1}^n X_i$, applying Chebyshev's inequality delivers a polynomial-type deviation bound

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \geq \sigma\sqrt{\frac{1}{\delta n}}\right) \leq \delta \quad \text{for any } \delta \in (0, 1).$$

In addition, if the distribution of X is sub-Gaussian, that is, $\mathbb{E}\exp(\lambda X) \leq \exp(\sigma^2\lambda^2/2)$ for all λ, then following the terminology in Devroye et al. (2016), $\widehat{\mu}_n$ becomes a sub-Gaussian estimator in the sense that

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| \geq \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right\} \leq \delta.$$

Catoni (2012) established a lower bound for the deviations of $\widehat{\mu}_n$ when the sampling distribution is the least favorable in the class of all distributions with bounded variance: for any $\delta \in (0, e^{-1})$, there is some distribution with mean μ and variance σ^2 such that an independent sample of size n drawn from it satisfies

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| \geq \sigma\sqrt{\frac{1}{\delta n}}\left(1 - \frac{e\delta}{n}\right)^{(n-1)/2}\right\} \geq \delta.$$

This shows that the deviation bound obtained from Chebyshev's inequality is essentially sharp under finite variance condition. The limitation of least squares method arises also in the regression setting, which triggers an outpouring of interest in developing sub-Gaussian estimators, from univariate mean estimation to multivariate or even high dimensional problems, for heavy-tailed data to achieve sharp deviation bounds from a nonasymptotic viewpoint. See, for example, Brownlees, Joly and Lugosi (2015), Minsker (2015), Minsker (2018), Hsu and Sabato (2016), Devroye et al. (2016), Catoni and Giulini (2017), Giulini (2017), Fan, Li and Wang (2017), Sun, Zhou and Fan (2019) and Lugosi and Mendelson (2019), among others. In particular, Fan, Li and Wang (2017) and Sun, Zhou and Fan (2019) proposed adaptive (regularized) Huber estimators with diverging robustification parameters (see Definition 2.1 in Section 2.1), and derived exponential-type deviation bounds when the error distribution only has finite variance. The key observation is that the robustification parameter should adapt to the sample size, dimensionality and noise level for optimal tradeoff between bias and robustness.

All the aforementioned work studies robust estimation through concentration properties, that is, the robust estimator is tightly concentrated around the true parameter with high probability even when the sampling distribution has only a small number of finite moments. In general, concentration inequalities loose constant factors and may result in confidence intervals too wide to be informative. Therefore, an interesting and challenging open problem is how to construct tight confidence sets for $\boldsymbol{\theta}^*$ with finite samples of heavy-tailed data (Devroye et al. (2016)).

This paper addresses this open problem by developing a robust inference framework with nonasymptotic guarantees. To illustrate the key idea, we focus on the classical setting where the parameter dimension d is smaller than but is allowed to increase with the number of observations n. Our approach integrates concentration properties of the adaptive Huber estimator (see Theorems 2.1 and 2.2) and the multiplier bootstrap method. The multiplier bootstrap, also known as the weighted bootstrap, is one of the most widely used resampling methods for constructing a confidence interval/set or for measuring the significance of a test. Its theoretical validity is typically guaranteed by the multiplier central limit theorem (van der Vaart and Wellner (1996)). We refer to Chatterjee and Bose (2005), Arlot, Blanchard and Roquain (2010), Chernozhukov, Chetverikov and Kato (2013, 2014), Spokoiny and Zhilova (2015) and Zhilova (2016) for the most recent progress in the theory and applications of the multiplier bootstrap. In particular, Spokoiny and Zhilova (2015) considered a multiplier bootstrap procedure for constructing likelihood-based confidence sets under a possible model misspecification. For a linear model with sub-Gaussian errors, their results validate the bootstrap procedure when applied to the ordinary least squares (OLS). With heavy-tailed errors in the regression model (1.1), we demonstrate how the adaptive Huber regression and the multiplier bootstrap can be integrated to construct robust and sharp confidence sets for the true parameter $\boldsymbol{\theta}^*$ with a given coverage probability. The validity of the bootstrap procedure in situations with a limited sample size, growing dimensionality and heavy-tailed errors is established. In all these theoretical results, we provide nonasymptotic bounds for the errors of bootstrap approximation. See Theorems 2.3 and 2.4 for finite sample properties of the bootstrap adaptive Huber estimator, including the deviation inequality, Bahadur representation and Wilks' expansion.

An alternative robust inference method is based on the asymptotic theory developed in Zhou et al. (2018); see, for example, Theorems 2.2 and 2.3 therein. Since the asymptotic distribution of either the proposed robust estimator itself or the excess risk depends on σ^2, a direct approach is to replace σ^2 by some sub-Gaussian variance estimator using Catoni's method (Catoni (2012)) or the median-of-means technique (Minsker (2015)), with the advantage of being computationally fast. The disadvantage, however, is two-fold: first, constructing sub-Gaussian variance estimator involves another tuning parameter (for the problem of simultaneously testing m regression models as discussed in the next section, variance estimation brings m additional tuning parameters); second, because the squared heavy-tailed data is highly right-skewed, using the method in Catoni (2012) or Fan, Li and Wang (2017) tends to underestimate the variance, and the median-of-means method is numerically unstable for small or moderate samples. Both methods were examined numerically in Zhou et al. (2018), while the multiplier bootstrap procedure, albeit being more computationally intensive, demonstrates the most desirable finite sample performance.

1.2. *Simultaneous inference.* In addition to building confidence sets for an individual parameter vector, multiple hypothesis testing is another important statistical problem with applications to many scientific fields, where thousands of tests are performed simultaneously (Dudoit and van der Laan (2008), Efron (2010)). Gene microarrays comprise a prototypical example; there, each subject is automatically measured on tens of thousands of features.

Together, the large number of tests together with heavy tailedness bring new challenges to conventional statistical methods, which, in this scenario, often suffer from low power to detect important features and poor reproducibility. Robust alternatives are thus needed for conducting large-scale multiple inference for heavy-tailed data.

In this section, we consider the multiple response regression model

$$y_{ik} = \mu_k + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_k + \varepsilon_{ik}, \quad i = 1, \ldots, n, k = 1, \ldots, m, \tag{1.2}$$

where μ_k is the intercept, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{is})^\intercal$, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{ks})^\intercal \in \mathbb{R}^s$ are s-dimensional vectors of random covariates and regression coefficients, respectively, and ε_{ik} is the regression error. Since our main focus here is the inference for intercepts, we decompose the parameter vector $\boldsymbol{\theta}^*$ in (1.1) into two parts: the intercept μ_k and the slope $\boldsymbol{\beta}_k$. Moreover, we use $\boldsymbol{x}_i$ in (1.2) to distinguish from $\boldsymbol{X}_i$ in (1.1). Write $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\intercal \in \mathbb{R}^m$ and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\intercal \in \mathbb{R}^m$ be the vector of intercepts. Based on random samples $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ from model (1.2), our goal is to simultaneously test the hypotheses

$$H_{0k} : \mu_k = 0 \quad \text{versus} \quad H_{1k} : \mu_k \neq 0, \quad \text{for } k = 1, \ldots, m. \tag{1.3}$$

An iconic example of model (1.2) is the linear pricing model, which subsumes the capital asset pricing model (CAPM) (Sharpe (1964), Lintner (1965)) and the Fama-French three-factor model (Fama and French (1993)). The key implication from the multi-factor pricing theory is that for any asset k, the intercept μ_k should be zero. It is then important to investigate if such a pricing theory, also known as the "mean-variance efficiency" pricing, can be validated by empirical data (Fan, Liao and Yao (2015)). According to the Berk and Green equilibrium (Berk and Green (2004)), inefficient pricing by the market may occur to a small proportion of exceptional assets, namely a very small fraction of the μ_k's are nonzero. To identify positive μ_k's by testing a large number of hypotheses simultaneously, Barras, Scaillet and Wermers (2010) and Lan and Du (2019) developed FDR controlling procedures for data coming from model (1.2), which can be applied to mutual fund selection in empirical finance. We refer to Friguet, Kloareg and Causeur (2009), Desai and Storey (2012), Fan, Han and Gu (2012) and Wang et al. (2017) for more examples from gene expression studies, where the goal is to identify features showing a biological signal of interest.

Despite the extensive research and wide application of this problem, existing least squares-based methods with normal calibration could fail when applied to heavy-tailed data with a small sample size. To address this challenge, we develop a robust bootstrap procedure for large-scale simultaneous inference, which achieves good numerical performance for a small or moderate sample size. Theoretically, we prove its validity on controlling the false discover proportion (FDP) (see Theorem 4.1).

Finally, we briefly comment on the computation issue. Fast computation of Huber regression is critical to our procedure since the multiplier bootstrap requires solving Huber loss minimization for at least hundreds of times. Ideally, a second order approach (e.g., Newton's method) is preferred. However, the second order derivative of Huber loss does not exist everywhere. To address this issue, we adopt the damped semismooth Netwon method (Qi and Sun (1999)), which is a synergic integration of first and second order methods. The details are provided in Appendix D of the supplemental material (Chen and Zhou (2019)).

1.3. *Organization of the paper.* The rest of the paper proceeds as follows. Section 2.1 presents a series of finite sample results for adaptive Huber regression. Sections 2.2 and 2.3 contain, respectively, the description of the bootstrap procedure for building confidence sets and theoretical guarantees. Two data-driven schemes are proposed in Section 3 for choosing the tuning parameter in the Huber loss. In Section 4, we propose a robust bootstrap calibration method for multiple testing and investigate its theoretical property on controlling the FDP.

The conclusions that are drawn in Sections 2 and 4 are illustrated numerically in Section 5. We conclude with a discussion in Section 6. The supplementary material contains all the proofs and additional simulation studies.

1.4. *Notation*. Let us summarize our notation. For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the the k-dimensional Euclidean space. The inner product of any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^\intercal$, $\boldsymbol{v} = (v_1, \ldots, v_k)^\intercal \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^\intercal \boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. We use the notation $\|\cdot\|_p, 1 \leq p \leq \infty$ for the ℓ_p-norms of vectors in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \leq i \leq k} |u_i|$. For $k \geq 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$. Throughout this paper, we use bold capital letters to represent matrices. For $k \geq 2$, $\mathbf{I}_k$ represents the identity/unit matrix of size k. For any $k \times k$ symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\mathbf{A}\|_2$ is the operator norm of $\mathbf{A}$. We use $\overline{\lambda}_{\mathbf{A}}$ and $\underline{\lambda}_{\mathbf{A}}$ to denote the largest and smallest eigenvalues of $\mathbf{A}$, respectively. For any two real numbers u and v, we write $u \vee v = \max(u, v)$ and $u \wedge v = \min(u, v)$. For two sequences of nonnegative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \geq C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two numbers C_1 and C_2, we write $C_2 = C_2(C_1)$ if C_2 depends only on C_1. For any set $\mathcal{S}$, we use $\operatorname{card}(\mathcal{S})$ and $|\mathcal{S}|$ to denote its cardinality, that is, the number of elements in $\mathcal{S}$.

2. Robust bootstrap confidence sets.

2.1. *Preliminaries*. First, we present some finite sample properties of the adaptive Huber estimator, which are of independent interest and also sharpen the results in Sun, Zhou and Fan (2019).

Let us recall the definition of the Huber loss.

DEFINITION 2.1. The Huber loss $\ell_\tau(\cdot)$ (Huber (1964)) is defined as

$$\ell_\tau(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \tau, \\ \tau|u| - \tau^2/2 & \text{if } |u| > \tau, \end{cases} \tag{2.1}$$

where $\tau > 0$ is a tuning parameter and will be referred to as the *robustification parameter* that balances bias and robustness.

The Huber estimator is defined as

$$\widehat{\boldsymbol{\theta}}_\tau \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\theta}) \quad \text{with } \mathcal{L}_\tau(\boldsymbol{\theta}) = \mathcal{L}_{n,\tau}(\boldsymbol{\theta}) := \sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}). \tag{2.2}$$

The following theorem provides a sub-Gaussian-type deviation inequality and a nonasymptotic Bahadur representation for $\widehat{\boldsymbol{\theta}}_\tau$. The proof is given in the supplement. We first impose the moment conditions.

CONDITION 2.1. (i) There exists some constant $A_0 > 0$ such that for any $\boldsymbol{u} \in \mathbb{R}^d$ and $t \in \mathbb{R}$, $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{Z} \rangle| \geq A_0 \|\boldsymbol{u}\|_2 \cdot t) \leq 2\exp(-t^2)$, where $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$ and $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\intercal)$ is positive definite. (ii) The regression error ε satisfies $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$, $\mathbb{E}(\varepsilon^2|\boldsymbol{X}) = \sigma^2$ and $\mathbb{E}(|\varepsilon|^{2+\delta}|\boldsymbol{X}) \leq \upsilon_{2+\delta}$ almost surely for some $\delta \geq 0$.

Part (i) of Condition 2.1 requires $\boldsymbol{X}$ to be a sub-Gaussian vector. Via one-dimensional marginal, this generalizes the concept of sub-Gaussian random variables to higher dimensions. Typical examples include: (i) Gaussian and Bernoulli random vectors, (ii) spherical random vector,[1] (iii) random vector that is uniformly distributed on the Euclidean ball cen-

[1] A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is said to have a spherical distribution if it is uniformly distributed on the Euclidean sphere in $\mathbb{R}^d$ with center at the origin and radius $\sqrt{d}$.

tered at the origin with radius $\sqrt{d}$, and (iv) random vector that is uniformly distributed on the unit cube $[-1, 1]^d$. In all the above cases, the constant A_0 represents a dimension-free constant. We refer to Chapter 3.4 in Vershynin (2018) for detailed discussions of sub-Gaussian distributions in higher dimensions. Technically, this assumption is needed in order to derive an exponential-type concentration inequality for the quadratic form $\|\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{Z}_i\|_2$, where

$$\boldsymbol{Z}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i, \quad i = 1, \ldots, n. \tag{2.3}$$

To avoid notational clutter, we focus on the homoscedastic model (1.1). The finite sample techniques developed for the results in this section and Section 2.3 can be extended to analyze heteroscedastic models of the form

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}^* + \sigma(\boldsymbol{X}_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

provided that the variance function $\sigma : \mathbb{R}^d \to (0, \infty)$ is such that $\mathbb{E}\{\sigma^2(\boldsymbol{X}_i)\}$ is bounded away from zero. The advantage of bootstrapping over the limiting distribution calibration method is more pronounced in the heteroscedastic model than in the homoscedastic model.

THEOREM 2.1. *Assume Condition 2.1 holds. For any $t > 0$ and $v \geq \upsilon_{2+\delta}^{1/(2+\delta)}$, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ given in (2.2) with $\tau = v(\frac{n}{d+t})^{1/(2+\delta)}$ satisfies*

$$\mathbb{P}\left\{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)\|_2 \geq c_1 v\sqrt{\frac{d+t}{n}}\right\} \leq 2e^{-t} \quad \text{and} \tag{2.4}$$

$$\mathbb{P}\left\{\left\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*) - \frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{Z}_i\right\|_2 \geq c_2 v\frac{d+t}{n}\right\} \leq 3e^{-t} \tag{2.5}$$

as long as $n \geq c_3(d+t)$, where c_1–c_3 are constants depending only on A_0.

The nonasymptotic results in Theorem 2.1 reveal a new perspective for Huber's method: to construct sub-Gaussian estimators for linear regression with heavy-tailed errors, the tuning parameter in the Huber loss should adapt to the sample size, dimension and moments for optimal tradeoff between bias and robustness. The resulting estimator is therefore referred to as the *adaptive Huber estimator*. Specifically, Theorem 2.1 provides the concentration property of the adaptive Huber estimator $\widehat{\boldsymbol{\theta}}_\tau$ and the Fisher expansion for the difference $\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*$. It improves Theorem 2.1 in Zhou et al. (2018) by sharpening the sample size scaling. The classical asymptotic results can be easily derived from the obtained nonasymptotic expansions. In the following theorem, we further study the concentration property of the Wilks' expansion for the excess $\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)$. This new result is directly related to the construction of confidence sets. See Theorem 2.3 below for its counterpart in the bootstrap world.

THEOREM 2.2. *Assume Condition 2.1 holds. Then for any $t > 0$ and $v \geq \upsilon_{2+\delta}^{1/(2+\delta)}$, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ with $\tau = v(\frac{n}{d+t})^{1/(2+\delta)}$ satisfies that with probability at least $1 - 3e^{-t}$,*

$$\left|\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \frac{1}{2}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{Z}_i\right\|_2^2\right| \leq c_4 v^2\frac{(d+t)^{3/2}}{\sqrt{n}} \quad \text{and} \tag{2.6}$$

$$\left|\sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}} - \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{Z}_i\right\|_2\right| \leq c_5 v\frac{d+t}{\sqrt{n}} \tag{2.7}$$

as long as $n \geq c_3(d+t)$, where $c_4, c_5 > 0$ are constants depending on A_0.

REMARK 2.1 (On the robustification parameter τ). Going through the proofs of Theorems 2.1 and 2.2, we see that the robustification parameter τ can be chosen as

$$\tau = v\{n/(d+t)\}^{\eta} \quad \text{for any } \eta \in [1/(2+\delta), 1/2] \text{ and } v \geq \upsilon_{2+\delta}^{1/(2+\delta)}, \tag{2.8}$$

such that the conclusions (2.4)–(2.7) hold as long as $n \gtrsim d+t$. This implies that the existence of higher moments of ε increases the flexibility of choosing τ, whose order ranges from $(\frac{n}{d+t})^{1/(2+\delta)}$ to $(\frac{n}{d+t})^{1/2}$. In practice, $\upsilon_{2+\delta}$ is unknown and thus brings difficulty in calibrating τ. Guided by the theoretical results, in Section 3 we propose a data-dependent procedure to choose τ.

REMARK 2.2 (Sample size scaling). The deviation inequalities in Theorems 2.1 and 2.2 hold under the scaling condition $n \gtrsim d+t$, indicating that as many as $d+t$ samples are required to guarantee the finite sample properties of the estimator. Similar conditions are also imposed for Proposition 2.4 in Catoni (2012) and Theorem 3.1 in Audibert and Catoni (2011). In particular if $\mathbb{E}(\varepsilon^2) < \infty$, taking $t = \log n$ and $\tau \asymp (\frac{n}{d+t})^{1/2}$, the corresponding estimator $\widehat{\boldsymbol{\theta}}_\tau$ satisfies

$$\widehat{\boldsymbol{\theta}}_\tau = \boldsymbol{\theta}^* + \frac{1}{n}\sum_{i=1}^{n} \ell_\tau'(\varepsilon_i)\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i + O\{n^{-1}(d+\log n)\}$$

with probability at least $1 - O(n^{-1})$ under the scaling $n \gtrsim d$. From an asymptotic viewpoint, this implies that if the dimension d, as a function of n, satisfies $d = o(n)$ as $n \to \infty$, then for any deterministic vector $\boldsymbol{u} \in \mathbb{R}^d$, the distribution of the linear contrast $\boldsymbol{u}^\top(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)$ coincides with that of $(1/n)\sum_{i=1}^n \ell_\tau'(\varepsilon_i)\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i$ asymptotically.

REMARK 2.3. To achieve sub-Gaussian behavior, the choice of loss function is not unique. An alternative loss function, which is obtained from the influence function proposed by Catoni and Giulini (2017), is

$$\rho_\tau(u) = \begin{cases} u^2/2 - u^4/(24\tau^2), & \text{if } |u| \leq \sqrt{2}\tau, \\ \dfrac{2\sqrt{2}}{3}\tau|u| - \tau^2/2, & \text{if } |u| > \sqrt{2}\tau. \end{cases} \tag{2.9}$$

The function ρ_τ is convex, twice differentiable everywhere and has bounded derivative that $|\rho_\tau'(u)| \leq (2\sqrt{2}/3)\tau$ for all u. By modifying the proofs of Theorems 2.1 and 2.2, it can be shown that the theoretical properties of the adaptive Huber estimator remain valid for the estimator that minimizes the empirical ρ_τ-loss. Computationally, it can be solved via Newton's method.

2.2. *Multiplier bootstrap.* In this section, we go beyond estimation and focus on robust inference. According to (2.7), the distribution of $2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}$ is close to that of $(1/n)\|\sum_{i=1}^n \xi_i \boldsymbol{Z}_i\|_2^2$ provided that d^2/n is small, where $\xi_i = \ell_\tau'(\varepsilon_i)$. As we will see in the proof of Theorem 2.5 that, the truncated random variable ξ_i has mean and variance *approximately* equal to 0 and σ^2, respectively. Heuristically, the multivariate central limit theorem allows us to approximate the distribution of the normalized sum $n^{-1/2}\sum_{i=1}^n \xi_i \boldsymbol{Z}_i$ by $\mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I}_d)$. If this were true, then the distribution of $2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}$ is close to the scaled chi-squared distribution $\sigma^2\chi_d^2$ with d degrees of freedom. This is in line with the asymptotic behavior of the likelihood ratio statistic that was studied in Wilks (1938). With sample size sufficiently large, this result allows to construct confidence sets for $\boldsymbol{\theta}^*$ using quantiles of χ_d^2: for any $\alpha \in (0, 1)$,

$$\mathcal{C}_\alpha^*(\sigma) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) \leq \sigma^2\chi_{d,\alpha}^2/2\}, \tag{2.10}$$

where $\chi^2_{d,\alpha}$ denotes the upper α-quantile of χ^2_d. Estimating the residual variance σ^2 in the construction of $\mathcal{C}^*_\alpha(\sigma)$ is even more challenging when the errors are heavy-tailed. Moreover, as argued in Spokoiny and Zhilova (2015), a possibly low speed of convergence of the likelihood ratio statistic makes the asymptotic Wilks' result hardly applicable to the case of small or moderate samples. Motivated by these two concerns, we have the following goal:

> propose a new method to construct confidence sets for $\boldsymbol{\theta}^*$ that is robust against heavy-tailed error distributions and performs well for a small or moderate sample.

The results in Section 2.1 show that the adaptive Huber estimator provides a robust estimate of $\boldsymbol{\theta}^*$ in the sense that it admits sub-Gaussian-type deviations when the error distribution only has finite variance. To estimate the quantiles of the adaptive Huber estimator and to construct confidence set, we consider the use of multiplier bootstrap. Let $U_1, \ldots, U_n$ be independent and identically distributed (IID) random variables that are independent of the observed data $\mathcal{D}_n := \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ and satisfy

$$\mathbb{E}(U_i) = 0, \qquad \mathbb{E}\big(U_i^2\big) = 1, \quad i = 1, \ldots, n. \tag{2.11}$$

With $W_i := 1 + U_i$ denoting the random weights, the bootstrap Huber loss and bootstrap Huber estimator are defined, respectively, as

$$\mathcal{L}^\flat_\tau(\boldsymbol{\theta}) = \sum_{i=1}^n W_i \ell_\tau\big(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}\big), \qquad \boldsymbol{\theta} \in \mathbb{R}^d \quad \text{and} \tag{2.12}$$

$$\widehat{\boldsymbol{\theta}}^\flat_\tau \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau\|_2 \le R} \mathcal{L}^\flat_\tau(\boldsymbol{\theta}),$$

where $R > 0$ is a prespecified radius parameter. A simple observation is that $\mathbb{E}^*\{\mathcal{L}^\flat_\tau(\boldsymbol{\theta})\} = \mathcal{L}_\tau(\boldsymbol{\theta})$, where $\mathbb{E}^*(\cdot) := \mathbb{E}(\cdot|\mathcal{D}_n)$ is the conditional expectation given the observed data $\mathcal{D}_n$. Therefore, $\widehat{\boldsymbol{\theta}}_\tau \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}^*\{\mathcal{L}^\flat_\tau(\boldsymbol{\theta})\}$ and the difference $\mathcal{L}^\flat_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}^\flat_\tau(\widehat{\boldsymbol{\theta}}^\flat_\tau)$ mimics $\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)$.

Based on this idea, we propose a Huber regression based inference procedure in Algorithm 1, where the bootstrap threshold $z^\flat_\alpha = z^\flat_\alpha(\mathcal{D}_n)$ approximates

$$z_\alpha := \inf\big\{z \ge 0 : \mathbb{P}\big\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > z\big\} \le \alpha\big\}. \tag{2.13}$$

Here $\mathbb{P}$ is the probability measure with respect to the underlying data generating process.

2.3. *Theoretical results*. In this section, we present detailed theoretical results for the bootstrap adaptive Huber estimator, including the deviation inequality, nonasymptotic Bahadur representation (Theorem 2.3), and Wilks' expansions (Theorem 2.4). Moreover, Theorems 2.5 and 2.6 establish the validity of the multiplier bootstrap for estimating quantiles of $\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)$ when the variance σ^2 is unknown. Proofs of the finite sample properties of the bootstrap estimator require new techniques and are more involved than those of Theorems 2.1 and 2.2. We leave them to the supplemental material.

CONDITION 2.2. $U_1, \ldots, U_n$ are IID from a random variable U satisfying $\mathbb{E}(U) = 0$, $\operatorname{var}(U) = 1$ and $\mathbb{P}(|U| \ge t) \le 2\exp(-t^2/A_U^2)$ for all $t \ge 0$.

THEOREM 2.3. *Assume Condition 2.1 with $\delta = 2$ and Condition 2.2 hold. For any $t > 0$ and $v \ge v_4^{1/4}$, the estimator $\widehat{\boldsymbol{\theta}}^\flat_\tau$ with $\tau = v(\frac{n}{d+t})^{1/4}$ and $R \asymp v$ satisfies*:

Algorithm 1 Huber Robust Confidence Set

Input: Data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, number of bootstrap samples B, Huber threshold τ, radius parameter R, confidence level $1-\alpha$

1: Solve the Huber regression in (2.2) and obtain $\widehat{\boldsymbol{\theta}}_\tau$.
2: **for** $b = 1, 2, \ldots, B$ **do**
3: Generate IID random weights $\{W_i\}_{i=1}^n$ satisfying $\mathbb{E}(W_i) = 1$ and $\operatorname{var}(W_i) = 1$.
4: Solve the weighted Huber regression in (2.12) and obtain the "bootstrap" Huber estimator.
5: **end for**
6: Define $\mathbb{P}^*$ be the conditional probability over the random multipliers given the observed data $\mathcal{D}_n = \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, that is, $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot\,|\mathcal{D}_n)$.
7: Compute the upper α-quantile of $\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)$:

$$z_\alpha^\flat = \inf\Big\{z \ge 0 : \mathbb{P}^*\big\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) > z\big\} \le \alpha\Big\}.$$

Output: A confidence set of $\boldsymbol{\theta}^*$ given by $\mathcal{C}_\alpha := \{\boldsymbol{\theta} \in \mathbb{R}^d : \mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) \le z_\alpha^\flat\}$.

1. *with probability (over $\mathcal{D}_n$) at least $1 - 5e^{-t}$,*

$$\mathbb{P}^*\big\{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \boldsymbol{\theta}^*)\|_2 \ge c_1 v (d+t)^{1/2} n^{-1/2}\big\} \le 3e^{-t}, \tag{2.14}$$

2. *with probability (over $\mathcal{D}_n$) at least $1 - 6e^{-t}$,*

$$\mathbb{P}^*\left\{\left\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau) - \frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i) U_i \boldsymbol{Z}_i\right\|_2 \ge c_2 v \frac{d+t}{n}\right\} \le 4e^{-t} \tag{2.15}$$

as long as $n \ge \max\{c_3\kappa_{\boldsymbol{\Sigma}}(d+t), c_4(d+t)^2\}$, where c_1–c_3 are positive constants depending on (A_0, A_U), $c_4 = c_4(A_0) > 0$ and $\kappa_{\boldsymbol{\Sigma}} = \overline{\lambda}_{\boldsymbol{\Sigma}}/\underline{\lambda}_{\boldsymbol{\Sigma}}$ is the condition number of $\boldsymbol{\Sigma}$.

The following theorem is a bootstrap version of Theorem 2.2. Define the random process

$$\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \nabla\mathbb{E}^*\mathcal{L}_\tau^\flat(\boldsymbol{\theta})\big\}, \quad \boldsymbol{\theta} \in \mathbb{R}^d. \tag{2.16}$$

From (2.3) and (2.11) we see that

$$\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta})\big\} = -\sum_{i=1}^n \ell_\tau'(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}) U_i \boldsymbol{Z}_i, \quad \boldsymbol{\theta} \in \mathbb{R}^d.$$

In particular, write $\boldsymbol{\xi}^\flat = \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*) = -\sum_{i=1}^n \ell_\tau'(\varepsilon_i) U_i \boldsymbol{Z}_i$.

THEOREM 2.4. *Assume Condition 2.1 with $\delta = 2$ and Condition 2.2 hold. For any $t > 0$ and $v \ge v_4^{1/4}$, the bootstrap estimator $\widehat{\boldsymbol{\theta}}_\tau^\flat$ with $\tau = v(\frac{n}{d+t})^{1/4}$ and $R \asymp v$ satisfies that, with probability (over $\mathcal{D}_n$) at least $1 - 5e^{-t}$,*

$$\mathbb{P}^*\left\{\left|\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) - \frac{\|\boldsymbol{\xi}^\flat\|_2^2}{2n}\right| \ge c_5 v^2 \frac{(d+t)^{3/2}}{\sqrt{n}}\right\} \le 4e^{-t} \tag{2.17}$$

and

$$\mathbb{P}^*\left\{\left|\sqrt{2(\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat))} - \frac{\|\boldsymbol{\xi}^\flat\|_2}{\sqrt{n}}\right| \ge c_6 v \frac{d+t}{\sqrt{n}}\right\} \le 4e^{-t} \tag{2.18}$$

as long as $n \ge \max\{c_3\kappa_{\boldsymbol{\Sigma}}(d+t), c_4(d+t)^2\}$, where $c_5, c_6 > 0$ are constants depending only on (A_0, A_U).

The results (2.17) and (2.18) are nonasymptotic bootstrap versions of the Wilks' and square-root Wilks' phenomena. In particular, the latter indicates that the square-root excess $\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}}$ is close to $n^{-1/2}\|\boldsymbol{\xi}^\flat\|_2$ with high probability as long as the dimension d of the parameter space satisfies the condition that d^2/n is small.

REMARK 2.4 (Order of robustification parameter). Similar to Remark 2.8, now with finite fourth moment υ_4, the robustification parameter in Theorems 2.3 and 2.4 can be chosen as

$$\tau = \upsilon\{n/(d+t)\}^\eta \quad \text{for any } \eta \in [1/4, 1/2) \text{ and } \upsilon \geq \upsilon_4^{1/4}, \tag{2.19}$$

such that the same conclusions remain valid. Due to Lemma A.2 in the supplemental material, here we require η to be strictly less than $1/2$.

The next result validates the approximation of the distribution of $\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)$ by that of $\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)$ in the Kolmogorov distance. Recall that $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot|\mathcal{D}_n)$ denotes the conditional probability given $\mathcal{D}_n = \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$.

THEOREM 2.5. *Suppose Assumption 2.1 holds with $\delta = 2$ and Condition 2.2 holds with $U \sim \mathcal{N}(0,1)$. For any $t > 0$ and $\upsilon \geq \upsilon_4^{1/4}$, let $\tau = \upsilon(\frac{n}{d+t})^\eta$ for some $\eta \in [1/4, 1/2)$. Then, with probability (over $\mathcal{D}_n$) at least $1 - 6e^{-t}$, it holds for any $z \geq 0$ that*

$$\begin{aligned} &\big|\mathbb{P}\big\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) \leq z\big\} - \mathbb{P}^*\big\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) \leq z\big\}\big| \\ &\quad \leq \Delta_1(n,d,t), \end{aligned} \tag{2.20}$$

where

$$\Delta_1(n,d,t) = C\big\{d^{3/2}n^{-1/2} + d^{1/2}\big\{(d+t)/n\big\}^{1-2\eta} + (d+t)^{3\eta}n^{1/2-3\eta}\big\} + 7e^{-t}$$

with $C = C(A_0, \sigma, \upsilon_4, \upsilon) > 0$.

Theorem 2.5 is in parallel with and can be viewed as a partial extension of Theorem 2.1 in Spokoiny and Zhilova (2015) to the case of heavy-tailed errors. In particular, taking $\eta = 1/4$ in Theorem 2.5 we see that the error term scales as $(d^3/n)^{1/4}$, while in Spokoiny and Zhilova (2015) it is of order $(d^3/n)^{1/8}$. The difference is due to the fact that the latter allows misspecified models as discussed in Remark A.2 therein. In some way, allowing asymmetric and heavy-tailed errors can be regarded as a particular form of misspecification, considering that the OLS is the maximum likelihood estimator at the normal model.

REMARK 2.5 (Asymptotic result). To make asymptotic statements, we assume $n \to \infty$ with an understanding that $d = d(n)$ depends on n and possibly $d \to \infty$ as $n \to \infty$. Theorem 2.5 can be used to show the bootstrap consistency, where the notion of consistency is the one that guarantees asymptotically valid inference. Specifically, it shows that when the dimension d, as a function of n, satisfies $d = o(n^{1/3})$, then with $\tau \asymp (\frac{n}{d+\log n})^\eta$ for some $\eta \in [1/4, 1/2)$, it holds

$$\sup_{z \geq 0}\big|\mathbb{P}\big\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) \leq z\big\} - \mathbb{P}^*\big\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) \leq z\big\}\big| = o_{\mathbb{P}}(1)$$

as $n \to \infty$.

For any $\alpha \in (0,1)$, let

$$z_\alpha^\flat := \inf\{z \geq 0 : \mathbb{P}^*\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) > z\} \leq \alpha\} \tag{2.21}$$

be the upper α-quantile of $\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)$ under $\mathbb{P}^*$, which serves as an approximate to the target value z_α given in (2.13). As a direct consequence of Theorem 2.5, the following result formally establishes the validity of the multiplier bootstrap for adaptive Huber regression with heavy-tailed error.

THEOREM 2.6 (Validity of multiplier bootstrap). *Assume the conditions of Theorem 2.5 hold and take $\eta = 1/4$. Then, for any $\alpha \in (0,1)$,*

$$\big|\mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > z_\alpha^\flat\} - \alpha\big| \leq \Delta_2(n,d,t), \tag{2.22}$$

where $\Delta_2(n,d,t) = C\{(d+t)^3/n\}^{1/4} + 16e^{-t}$, where $C = C(A_0, \sigma, \upsilon_4, v) > 0$. In particular, taking $\tau \asymp (\frac{n}{d+\log n})^{1/4}$, it holds

$$\sup_{\alpha \in (0,1)} \big|\mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > z_\alpha^\flat\} - \alpha\big| = o(1)$$

provided that $d = d(n)$ satisfies $d = o(n^{1/3})$ as $n \to \infty$.

3. Data-driven procedures for choosing τ. The theoretical results in Sections 2.1 and 2.3 reveal the performance of Huber-type estimators under various idealized scenarios, as such providing guidance on the choice of the key tuning parameter, which is referred to as the robustification parameter that balances bias and robustness. For estimation purpose, we take $\tau = v(\frac{n}{d+t})^{1/2}$ with $v \geq \sigma$; and for bootstrap inference, we choose $\tau = v(\frac{n}{d+t})^{1/4}$ with $v \geq \upsilon_4^{1/4}$. Since both $\sigma^2 = \operatorname{var}(\varepsilon)$ and $\upsilon_4 \geq \mathbb{E}(\varepsilon^4)$ are typically unknown in practice, an intuitive approach is to replace them by the empirical second and fourth moments of the residuals from the ordinary least squares (OLS) estimator, that is, $\widehat{\sigma}^2 := (n-d)^{-1}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\theta}}_{\mathrm{ols}})^2$ and $\widehat{\upsilon}_4 := (n-d)^{-1}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\theta}}_{\mathrm{ols}})^4$. This simple approach performs reasonably well empirically (see Section 5). However, when heavy tails may be a concern, $\widehat{\sigma}^2$ and $\widehat{\upsilon}_4$ are not good estimates of σ^2 and υ_4. In this section, we discuss two data-dependent methods for choosing the tuning parameter τ: the first one uses an adaptive technique based on Lepski's method (Lepskiĭ (1991)), and the second method is inspired by the censored equation approach in Hahn, Kuelbs and Weiner (1990) which was originally introduced in pursing a more robust weak convergence theory for self-normalized sums.

3.1. *Lepski-type method.* Borrowing an idea from Minsker (2018), we first consider a simple adaptive procedure based on Lepski's method. Let $v_{\min}$ and $v_{\max}$ be some crude preliminary lower and upper bounds for the residual standard deviation, that is, $v_{\min} \leq \sigma \leq v_{\max}$. For some prespecified $a > 1$, let $v_j = v_{\min} a^j$ for $j = 0, 1, \ldots$ and define

$$\mathcal{J} = \{j \in \mathbb{Z} : v_{\min} \leq v_j < a v_{\max}\}.$$

It is easy to see that the set $\mathcal{J}$ has its cardinality bounded by $|\mathcal{J}| \leq 1 + \log_a(v_{\max}/v_{\min})$. Accordingly, we define a sequence of candidate parameters $\{\tau_j = v_j(\frac{n}{d+t})^{1/2}, j \in \mathcal{J}\}$ and let $\widehat{\boldsymbol{\theta}}^{(j)}$ be the Huber estimator with $\tau = \tau_j$. Set

$$\begin{aligned}\widehat{j}_{\mathrm{L}} := \min\Big\{ & j \in \mathcal{J} : \|\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}^{(j)}\|_2 \\ & \leq c_0 \upsilon_k \sqrt{\frac{d+t}{n}} \text{ for all } k \in \mathcal{J} \text{ and } k > j \Big\}\end{aligned} \tag{3.1}$$

for some constant $c_0 > 0$. The resulting adaptive estimator is then defined as $\widehat{\boldsymbol{\theta}}_{\mathrm{L}} = \widehat{\boldsymbol{\theta}}^{(\widehat{j}_{\mathrm{L}})}$.

THEOREM 3.1. *Assume that* $c_0 \geq 2c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}$ *for* $c_1 > 0$ *as in Theorem* 2.1. *Then for any* $t > 0$,

$$\|\widehat{\boldsymbol{\theta}}_{\mathrm{L}} - \boldsymbol{\theta}^*\|_2 \leq \frac{3a}{2} c_0 \sigma \sqrt{\frac{d+t}{n}}$$

with probability at least $1 - 3\log_a(a\upsilon_{\max}/\upsilon_{\min})e^{-t}$, *provided* $n \gtrsim d + t$.

Lepski's adaptation method serves a general technique to select the "best" estimator from a collection of certified candidates. The selected estimator adapts to the unknown noise level and satisfies near-optimal probabilistic bounds, while the associated parameter is not necessarily the theoretically optimal one. When applied with the bootstrap, Theorem 2.6 suggests that the dependence on d/n should be slightly adjusted. Since the reuse of the sample brings a big challenge mathematically, we shall prove a theoretical result for the data-driven multiplier bootstrap procedure with sample splitting. However, to avoid notational clutter, we state a two-step procedure without sample splitting, but with the assumption that the second step is carried out on an independent sample.

A Two-Step Data-Driven Multiplier Bootstrap.

STEP 1. Given independent observations $\{(Y_i^{(1)}, \boldsymbol{X}_i^{(1)})\}_{i=1}^n$ from linear model (1.1), first we produce a robust pilot estimator using Lepski's method. Recall that Lepski's method requires initial crude upper and lower bounds for $\upsilon_4 \geq \mathbb{E}(\varepsilon^4)$. Let $\mu_Y = \mathbb{E}(Y)$ and note that $\upsilon_Y := \mathbb{E}(Y - \mu_Y)^4 > \upsilon_4$. We shall use the median-of-means (MOM) estimator of υ_Y as a proxy, which is tuning-free in the sense that the construction does not depend on the noise level (Minsker (2015)). Specifically, we divide the index set $\{1, \ldots, n\}$ into $m \geq 2$ disjoint, equal-length groups $G_1, \ldots, G_m$, assuming n is divisible by m. For $j = 1, \ldots, m$, compute the empirical 4th moment evaluated over observations in group j: $\widehat{\upsilon}_{Y,j} = (1/|G_j|)\sum_{i\in G_j}\{Y_i^{(1)} - \bar{Y}_{G_j}^{(1)}\}^4$ with $\bar{Y}_{G_j}^{(1)} = (1/|G_j|)\sum_{i\in G_j} Y_i^{(1)}$. The MOM estimator of υ_Y is then defined by $\widehat{\upsilon}_{Y,\mathrm{mom}} = \mathrm{median}\{\widehat{\upsilon}_{Y,1}, \ldots, \widehat{\upsilon}_{Y,m}\}$.

Take $\upsilon_{\max} = (2\widehat{\upsilon}_{Y,\mathrm{mom}})^{1/4}$ and $\upsilon_{\min} = a^{-K}\upsilon_{\max}$ for some integer $K \geq 1$ and $a > 1$. Denote $\upsilon_j = a^j\upsilon_{\min}$ for $j = 0, 1, \ldots,$ so that $\mathcal{J} = \{j \in \mathbb{Z} : \upsilon_{\min} \leq \upsilon_j < a\upsilon_{\max}\} = \{0, 1, \ldots, K\}$. Slightly different from above, now we consider a sequence of parameters $\{\tau_j = \upsilon_j(\frac{n}{d+\log n})^{1/4}\}_{j\in\mathcal{J}}$ and let $\widetilde{\boldsymbol{\theta}}^{(j)}$ be the Huber estimator with $\tau = \tau_j$. Set

$$\begin{aligned}\tilde{j} := \min\Bigg\{ & j \in \mathcal{J} : \|\widetilde{\boldsymbol{\theta}}^{(k)} - \widetilde{\boldsymbol{\theta}}^{(j)}\|_2 \\ & \leq c_0\upsilon_k\sqrt{\frac{d+\log n}{n}} \text{ for all } k \in \mathcal{J} \text{ and } k > j\Bigg\}\end{aligned} \tag{3.2}$$

for some constant $c_0 > 0$. Denote by $\widehat{\boldsymbol{\theta}}^{(1)} = \widetilde{\boldsymbol{\theta}}^{(\tilde{j})}$ the corresponding estimator and put $\widehat{\tau} = \tau_{\tilde{j}}$.

STEP 2. Taking $\widehat{\boldsymbol{\theta}}^{(1)}$ and $\widehat{\tau}$ from Step 1, next we apply the multiplier bootstrap procedure to a new sample $(Y_i^{(2)}, \boldsymbol{X}_i^{(2)})\}_{i=1}^n$ that is independent from the previous one. Similarly to (2.2) and (2.12), define

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\mathcal{L}}(\boldsymbol{\theta}) \quad \text{and} \quad \widehat{\boldsymbol{\theta}}^{\flat} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d:\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^{(1)}\|_2\leq\widehat{R}} \widehat{\mathcal{L}}^{\flat}(\boldsymbol{\theta}), \tag{3.3}$$

where $\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_{\widehat{\tau}}(Y_i^{(2)} - \langle \boldsymbol{X}_i^{(2)}, \boldsymbol{\theta}\rangle)$, $\widehat{\mathcal{L}}^{\flat}(\boldsymbol{\theta}) = \sum_{i=1}^n W_i\ell_{\widehat{\tau}}(Y_i^{(2)} - \langle \boldsymbol{X}_i^{(2)}, \boldsymbol{\theta}\rangle)$ and $\widehat{R} = \widehat{\tau}(\frac{d+\log n}{n})^{1/4}$. With the above preparations, we apply Algorithm 1 to construct the confidence

set $\widehat{\mathcal{C}}_\alpha = \{\boldsymbol{\theta} \in \mathbb{R}^d : \widehat{\mathcal{L}}(\boldsymbol{\theta}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}) \le \widehat{z}_\alpha^\flat\}$, where

$$\widehat{z}_\alpha^\flat = \inf\{z \ge 0 : \mathbb{P}\{\widehat{\mathcal{L}}^\flat(\widehat{\boldsymbol{\theta}}) - \mathcal{L}^\flat(\widehat{\boldsymbol{\theta}}^\flat) > z | \bar{\mathcal{D}}_n\} \le \alpha\}$$

with $\bar{\mathcal{D}}_n = \{(Y_i^{(1)}, \boldsymbol{X}_i^{(1)}), (Y_i^{(2)}, \boldsymbol{X}_i^{(2)})\}_{i=1}^n$.

THEOREM 3.2. *Assume $\bar{\mathcal{D}}_n$ is an independent sample from $(Y, \boldsymbol{X})$ satisfying Condition 2.1 and moreover, $\mathbb{E}(|\varepsilon|^{4+\delta}) \le \upsilon_{4+\delta}$ for some $\delta > 0$. Let $W_1, \ldots, W_n$ be IID $\mathcal{N}(1, 1)$ random variables that are independent of $\bar{\mathcal{D}}_n$. Assume further that $d = d(n)$ satisfies $d = o(n^{1/3})$ as $n \to \infty$. Then, for any $\alpha \in (0, 1)$, the confidence set $\widehat{\mathcal{C}}_\alpha$ obtained by the two-step multiplier bootstrap procedure with $m = \lfloor 8 \log n + 1 \rfloor$ and $K = \lfloor \log_a (3\upsilon_Y/\upsilon_4)^{1/4} \rfloor + 1$ satisfies $\mathbb{P}(\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_\alpha) \to 1 - \alpha$ as $n \to \infty$.*

The proof of Theorem 3.2 will be provided in Section C.2 in the supplementary material.

3.2. *Huber-type method.* In Huber's original proposal, robust location estimation with desirable efficiency also depends on the scale parameter σ. For example, in Huber's Proposal 2 (Huber (1964)), the location μ and scale σ are estimated simultaneously by solving a system of "likelihood equations". Similarly in spirit, we propose a new data-driven tuning scheme to calibrate τ by solving a so-called censored equation (Hahn, Kuelbs and Weiner (1990)) instead of a likelihood equation. We first consider mean estimation to illustrate the main idea, and then move forward to the regression problem. Due to space limitations, we leave some discussions and proofs of the theoretical results to Appendix E in the supplemental material (Chen and Zhou (2019)).

3.2.1. *Motivation*: *Truncated mean.* Let $X_1, \ldots, X_n$ be IID random variables from X with mean μ and variance $\sigma^2 > 0$. Without loss of generality, we first assume $\mu = 0$. Catoni (2012) proved that the worst case deviations of the sample mean $\bar{X}_n$ are suboptimal for heavy-tailed distributions (see Appendix E.2). To attenuate the erratic fluctuations in $\bar{X}_n$, it is natural to consider the truncated sample mean

$$\widehat{m}_\tau = \frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) \quad \text{for some } \tau > 0, \tag{3.4}$$

where

$$\psi_\tau(u) := \ell_\tau'(u) = \operatorname{sgn}(u)\min\big(|u|, \tau\big), \quad u \in \mathbb{R}, \tag{3.5}$$

and τ is a tuning parameter that balances between bias and robustness. To see this, let $\mu_\tau = \mathbb{E}(\widehat{m}_\tau)$ be the truncated mean. By Markov's inequality, the bias term can be controlled by

$$\begin{aligned} |\mu_\tau| &= \big|\mathbb{E}\big\{X - \operatorname{sgn}(X)\tau\big\} I\big(|X| > \tau\big)\big| \\ &\le \mathbb{E}\big(|X| - \tau\big) I\big(|X| > \tau\big) \\ &\le \frac{\mathbb{E}(X^2 - \tau^2) I\big(|X| > \tau\big)}{\tau} \le \frac{\sigma^2 - \mathbb{E}\psi_\tau^2(X)}{\tau}. \end{aligned} \tag{3.6}$$

The robustness of $\widehat{m}_\tau$, on the other hand, can be characterized via the deviation

$$|\widehat{m}_\tau - \mu_\tau| = \left|\frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) - \mu_\tau\right|.$$

The following result shows that with a properly chosen τ, the truncated sample mean achieves a sub-Gaussian performance under the finite variance condition. Moreover, uniform convergence over a neighborhood of the optimal tuning scale requires an additional $\log(n)$-factor. For every $\tau > 0$, define the truncated second moment

$$\sigma_\tau^2 = \mathbb{E}\{\psi_\tau^2(X)\} = \mathbb{E}\{\min(X^2, \tau^2)\}. \tag{3.7}$$

PROPOSITION 3.1. For any $1 \le t < n\mathbb{P}(|X| > 0)$, let $\tau_t > 0$ be the solution to

$$\frac{\mathbb{E}\{\psi_\tau^2(X)\}}{\tau^2} = \frac{t}{n}, \quad \tau > 0. \tag{3.8}$$

(i) With probability at least $1 - 2e^{-t}$, $\widehat{m}_{\tau_t}$ satisfies

$$|\widehat{m}_{\tau_t} - \mu_{\tau_t}| \le 1.75\sigma_{\tau_t}\sqrt{\frac{t}{n}} \quad \text{and} \quad |\widehat{m}_{\tau_t}| \le \left(0.75\sigma_{\tau_t} + \frac{\sigma^2}{\sigma_{\tau_t}}\right)\sqrt{\frac{t}{n}}. \tag{3.9}$$

(ii) With probability at least $1 - 2e^{\log n - t}$,

$$\max_{\tau_t/2 \le \tau \le 3\tau_t/2} |\widehat{m}_\tau| \le C_t\sqrt{\frac{t}{n}} + \frac{\sigma_{\tau_t}}{\sqrt{n}}, \tag{3.10}$$

where $C_t := \sup_{\sigma_{\tau_t}/2 \le c \le 3\sigma_{\tau_t}/2}\{\sigma_{c(n/t)^{1/2}}\sqrt{2} - c^{-1}\sigma^2_{c(n/t)^{1/2}} + c/3 + c^{-1}\sigma^2\} \le \sqrt{2}\sigma + 2\sigma^2/\sigma_{\tau_t} + \sigma_{\tau_t}/6$.

The next result establishes existence and uniqueness of the solution to equation (3.8).

PROPOSITION 3.2. (i) *Provided* $0 < t < n\mathbb{P}(|X| > 0)$, *equation* (3.8) *has a unique solution, denoted by* τ_t, *which satisfies*

$$\{\mathbb{E}(X^2 \wedge q_{t/n}^2)\}^{1/2}\sqrt{\frac{n}{t}} \le \tau_t \le \sigma\sqrt{\frac{n}{t}},$$

where $q_\alpha := \inf\{z : \mathbb{P}(|X| > z) \le \alpha\}$ *is the upper* α*-quantile of* $|X|$. (ii) *Let* $t = t_n > 0$ *satisfy* $t_n \to \infty$ *and* $t = o(n)$. *Then* $\tau_t \to \infty$, $\sigma_{\tau_t} \to \sigma$ *and* $\tau_t \sim \sigma\sqrt{n/t}$ *as* $n \to \infty$.

According to Proposition 3.1, an ideal τ is such that the sample mean of truncated data $\psi_\tau(X_1), \ldots, \psi_\tau(X_n)$ is tightly concentrated around the true mean. At the same time, it is reasonable to expect that the empirical second moment of $\psi_\tau(X_i)$'s provides an adequate estimate of $\sigma_\tau^2 = \mathbb{E}\{\psi_\tau^2(X)\}$. Motivated by this observation, we propose to choose τ by solving the equation

$$\tau = \left\{\frac{1}{n}\sum_{i=1}^n \psi_\tau^2(X_i)\right\}^{1/2}\sqrt{\frac{n}{t}}, \quad \tau > 0,$$

or equivalently, solving

$$\frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau^2(X_i)}{\tau^2} = \frac{t}{n}, \quad \tau > 0. \tag{3.11}$$

Equation (3.11) is the sample version of (3.8). Provided the solution exists and is unique, denoted by $\widehat{\tau}_t$, we obtain a data-driven estimator

$$\widehat{m}_{\widehat{\tau}_t} = \frac{1}{n}\sum_{i=1}^n \operatorname{sgn}(X_i)\min\bigl(|X_i|, \widehat{\tau}_t\bigr). \tag{3.12}$$

As a direct consequence of Proposition 3.2, the following result ensures existence and uniqueness of the solution to equation (3.11).

PROPOSITION 3.3. *Provided* $0 < t < \sum_{i=1}^n I(|X_i| > 0)$, *equation* (3.11) *has a unique solution*.

Throughout, we use $\widehat{\tau}_t$ to denote the solution to equation (3.11), which is unique and positive when $t < \sum_{i=1}^n I(|X_i| > 0)$. For completeness, we set $\widehat{\tau}_t = 0$ if $t \geq \sum_{i=1}^n I(|X_i| > 0)$. If $\mathbb{P}(X = 0) = 0$, then $\widehat{\tau}_t > 0$ with probability one as long as $0 < z < n$. In the special case of $t = 1$, since $\psi_{\widehat{\tau}_1}(X_i) = X_i$ for all $i = 1, \ldots, n$, equation (3.11) has a unique solution $\widehat{\tau}_1 = (\sum_{i=1}^n X_i^2)^{1/2}$. With both τ_t and $\widehat{\tau}_t$ well defined, next we investigate the statistical property of $\widehat{\tau}_t$.

THEOREM 3.3. *Assume that* $\operatorname{var}(X) < \infty$ *and* $\mathbb{P}(X = 0) = 0$. *For any* $1 \leq t < n$ *and* $0 < r < 1$, *we have*

$$\begin{aligned} &\mathbb{P}\big(|\widehat{\tau}_t/\tau_t - 1| \geq r\big) \\ &\quad \leq e^{-a_1^2 r^2 t^2/(2t + 2a_1 r t/3)} + e^{-a_2^2 r^2 t/2} + 2e^{-(a_1 \wedge a_2)^2 t/8}, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} a_1 = a_1(t, r) &= \frac{P(\tau_t)}{2Q(\tau_t)} \frac{2 + r}{(1 + r)^2}, \\ a_2 = a_2(t, r) &= \frac{P(\tau_t - \tau_t r)}{2Q(\tau_t)} \frac{2 - r}{1 - r}, \end{aligned} \tag{3.14}$$

where $P(z) = \mathbb{E}\{X^2 I(|X| \leq z)\}$ *and* $Q(z) = \mathbb{E}\{\psi_z^2(X)\}$ *for* $z \geq 0$.

More properties of functions $P(z)$ and $Q(z)$ can be found in Appendix E.1 in the supplement.

REMARK 3.1. We discuss some direct implications of Theorem 3.3.

(i) Let $t = t_n \geq 1$ satisfy $t \to \infty$ and $t = o(n)$ as $n \to \infty$. By Proposition 3.2, $\tau_t \to \infty$, $\sigma_{\tau_t} \to \sigma$ and $\tau_t \sim \sigma\sqrt{n/t}$, which further implies $P(\tau_t) \to \sigma$ and $Q(\tau_t) \to \sigma$ as $n \to \infty$.

(ii) With $r = 1/2$ and $t = (\log n)^{1+\kappa}$ for some $\kappa > 0$ in (3.13), the constants $a_1 = a_1(t, 1/2)$ and $a_2 = a_2(t, 1/2)$ satisfy $a_1 \to 5/9$ and $a_2 \to 3/2$ as $n \to \infty$. The resulting $\widehat{\tau}_t$ satisfies that with probability approaching one, $\tau_t/2 \leq \widehat{\tau}_t \leq 3\tau_t/2$.

The following result, which is a direct consequence of (3.10), Theorem 3.3 and Remark 3.1, shows that the data-driven estimator $\widehat{m}_{\widehat{\tau}_t}$ with $t = (\log n)^{1+\kappa}$ ($\kappa > 0$) is tightly concentrated around the mean with high probability.

COROLLARY 3.1. *Assume the conditions of Theorem* 3.3 *hold. Then, the truncated mean* $\widehat{m} = \widehat{m}_{\widehat{\tau}_t}$ *with* $t = (\log n)^{1+\kappa}$ *for some* $\kappa > 0$ *satisfies* $|\widehat{m}| \leq c_1\sqrt{(\log n)^{1+\kappa}/n}$ *with probability greater than* $1 - c_2 n^{-1}$ *as* $n \to \infty$, *where* $c_1, c_2 > 0$ *are constants independent of* n.

3.2.2. *Huber's mean estimator.* For the truncated sample mean, even with the theoretically optimal tuning parameter, the deviation of the estimator only scales with the second moment rather than the ideal scale σ. Indeed, the truncation method described above primarily serves as a heuristic device and paves the way for developing data-driven Huber estimators.

Given IID samples $X_1, \ldots, X_n$ with mean μ and variance σ^2, recall the Huber estimator $\widehat{\mu}_\tau = \operatorname{argmin}_\theta \sum_{i=1}^n \ell_\tau(X_i - \theta)$, which is also the unique solution to

$$\frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i - \theta) = 0, \quad \theta \in \mathbb{R}. \tag{3.15}$$

The nonasymptotic property of $\widehat{\mu}_\tau$ is characterized by a Bahadur-type representation result developed in Zhou et al. (2018): for any $t > 0$, $\widehat{\mu}_\tau$ with $\tau = \sigma\sqrt{n/t}$ satisfies the bound $|\widehat{\mu}_\tau - \mu - (1/n)\sum_{i=1}^n \psi_\tau(\varepsilon_i)| \le c_1\sigma t/\sqrt{n}$ with probability at least $1 - 3e^{-t}$ provided $n \ge c_2 t$, where $c_1, c_2 > 0$ are absolute constants and $\varepsilon_i = X_i - \mu$ are noise variables. In other words, a properly chosen τ is such that the truncated average $(1/n)\sum_{i=1}^n \psi_\tau(\varepsilon_i)$ is resistant to outliers caused by a heavy-tailed 'noise'. Similar to (3.11), now we would like to choose the robustification parameter by solving

$$\frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau^2(\varepsilon_i)}{\tau^2} = \frac{t}{n}, \tag{3.16}$$

which is practically impossible as ε_i's are unobserved realized noise. In light of (3.15) and (3.16), and motivated by Huber's Proposal 2 [page 96 in Huber (1964)] for the simultaneous estimation of location and scale, we propose to estimate μ and calibrate τ by solving the following system of equations

$$\begin{cases} \displaystyle\sum_{i=1}^n \psi_\tau(X_i - \theta) = 0, \\ \displaystyle\frac{1}{n}\sum_{i=1}^n \frac{\psi_\tau^2(X_i - \theta)}{\tau^2} - \frac{t}{n} = 0, \end{cases} \qquad \theta \in \mathbb{R}, \tau > 0.$$

This method of simultaneous estimation can be naturally extended to the regression setting, as discussed in the next section.

A different while comparable proposal is a two-step method, namely M-estimation of μ with auxiliary robustification parameter computed separately by solving

$$\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{\min\{(X_i - X_j)^2/2, \tau^2\}}{\tau^2} = \frac{t}{n}.$$

It is, however, less clear how this method can be generalized to the regression problem. Therefore, our focus will be on the previous approach.

3.2.3. *Data-driven Huber regression.* Consider the linear model $Y_i = \boldsymbol{X}_i^\intercal \boldsymbol{\theta}^* + \varepsilon_i$ as in (1.1) and the Huber estimator $\widehat{\boldsymbol{\theta}}_\tau = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\theta})$, where $\mathcal{L}_\tau(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta})$. From the deviation analysis in (2.1) we see that to achieve the sub-Gaussian performance bound, the theoretically desirable tuning parameter for $\widehat{\boldsymbol{\theta}}_\tau$ is $\tau \sim \sigma\sqrt{n/(d+t)}$ with $\sigma^2 = \operatorname{var}(\varepsilon_i)$. Further, by the Bahadur representation (2.5),

$$\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* = \frac{1}{n}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i + \mathcal{R}_\tau,$$

where the remainder $\mathcal{R}_\tau$ is of the order $\sigma(d+t)/n$ with exponentially high probability. This result demonstrates that the robustness is essentially gained from truncating the errors. Motivated by this representation and our discussions in Section 3.2.1, a robust tuning scheme is to find τ such that

$$\tau = \left\{\frac{1}{n}\sum_{i=1}^n \psi_\tau^2(\varepsilon_i)\right\}^{1/2} \sqrt{\frac{n}{d+t}}, \quad \tau > 0. \tag{3.17}$$

Unlike the mean estimation problem, the realized noises ε_i are unobserved. It is therefore natural to calibrate τ using fitted residuals. On the other hand, for a given $\tau > 0$, the Huber

loss minimization is equivalent to the following least squares problem with variable weights:

$$\min_{w_i \geq 0, \boldsymbol{\theta}} \quad \sum_{i=1}^{n} \left\{ \frac{(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta})^2}{w_i + 1} + \tau^2 w_i \right\}, \tag{3.18}$$

where the minimization is over $w_i \geq 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. This equivalence can be derived by writing down the KKT conditions of (3.18). Details will be provided in Remark 3.2 below. By (3.18), this problem can be solved via the iteratively reweighted least squares method.

To summarize, we propose an iteratively reweighted least squares algorithm, which starts at iteration 0 with an initial estimate $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{ols}}$ (the least squares estimator) and involves three steps at each iteration.

Calibration: Using the current estimate $\boldsymbol{\theta}^{(k)}$, we compute the vector of residuals $\boldsymbol{R}^{(k)} = (R_1^{(k)}, \ldots, R_n^{(k)})^\intercal$, where $R_i^{(k)} = Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}^{(k)}$. Then we take $\tau^{(k)}$ as the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\min\{R_i^{(k)2}, \tau^2\}}{\tau^2} = \frac{d+t}{n}, \quad \tau > 0. \tag{3.19}$$

By Proposition 3.3, this equation has a unique positive solution provided $d + t < \sum_{i=1}^{n} I(|R_i^{(k)}| > 0)$.

Weighting: Compute the vector of weights $\boldsymbol{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})^\intercal$, where $w_i^{(k)} = |R_i^{(k)}|/\tau^{(k)} - 1$ if $|R_i^{(k)}| > \tau^{(k)}$ and $w_i^{(k)} = 0$ if $|R_i^{(k)}| \leq \tau^{(k)}$. Then define the diagonal matrix $\mathbf{W}^{(k)} = \mathrm{diag}((1 + w_1^{(k)})^{-1}, \ldots, (1 + w_n^{(k)})^{-1})$.

Weighted least squares: Solve the weighted least squares problem (3.18) with $w_i = w_i^{(k)}$ and $\tau = \tau^{(k)}$ to obtain

$$\boldsymbol{\theta}^{(k+1)} = \left(\mathbf{X}^\intercal \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \boldsymbol{Y},$$

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

Repeat the above three steps until convergence or until the maximum number of iterations is reached.

In addition, from Theorems 2.3–2.5 we see that the validity of the multiplier bootstrap procedure requires a finite fourth moment condition, under which the ideal choice of τ is $\{\upsilon_4 n/(d+t)\}^{1/4}$. To construct data-dependent robust bootstrap confidence set, we adjust equation (3.19) by replacing $R_i^{(k)2}$ and τ^2 therein with $R_i^{(k)4}$ and τ^4, and solve instead

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\min\{R_i^{(k)4}, \tau^4\}}{\tau^4} = \frac{d+t}{n}, \quad \tau > 0. \tag{3.20}$$

Keep the other two steps and repeat until convergence or the maximum number of iterations is reached. Let $\widehat{\boldsymbol{\theta}}_{\widehat{\tau}}$ and $\widehat{\tau}$ be the obtained solutions. Then, we apply Algorithm 1 with $\tau = \widehat{\tau}$ to construct confidence sets.

Finally we discuss the choice of t. Since t appears in both the deviation bound and confidence level, we let $t = t_n$ slowly grow with the sample size to gain robustness without compromising unbiasedness. We take $t = \log n$, a typical slowly growing function of n, in all the numerical experiments carried out in this paper.

REMARK 3.2 (Equivalence between (3.18) and Huber regression). For a given $\boldsymbol{\theta}$ in (3.18), define $R_i = Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}$, $i = 1, \ldots, n$. The KKT condition of the program (3.18) with respect to each w_i under the constraint $w_i \geq 0$ now reads:

$$-\frac{R_i^2}{(w_i+1)^2} + \tau^2 - \lambda_i = 0; \quad w_i \geq 0, \lambda_i \geq 0; \lambda_i w_i = 0,$$

where λ_i is the Lagrangian multiplier. The solution to the KKT condition takes the form:

$$w_i = \frac{|R_i|}{\tau} - 1, \qquad \lambda_i = 0 \quad \text{if } |R_i| \geq \tau,$$
$$w_i = 0, \qquad \lambda_i = \tau^2 - R_i^2 \quad \text{if } |R_i| < \tau.$$

This gives the optimal solution of w_i. By plugging the optimal solution of w_i back into (3.18), we obtain the following optimization with respect to $\boldsymbol{\theta}$:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (2\tau |Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}| - \tau^2) I\big(|Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}| \geq \tau\big)$$
$$+ |Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}|^2 I\big(|Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}| < \tau\big),$$

which is equivalent to Huber regression.

4. Multiple inference with multiplier bootstrap calibration. In this section, we apply the adaptive Huber regression with multiplier bootstrap to simultaneously test the hypotheses in (1.3). Given a random sample $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$ from the multiple response regression model (1.2), we define robust estimators

$$(\widehat{\mu}_k, \widehat{\boldsymbol{\beta}}_k) \in \operatorname*{argmin}_{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^s} \sum_{i=1}^{n} \ell_{\tau_k}(y_{ik} - \mu - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}), \quad k = 1, \ldots, m, \tag{4.1}$$

where τ_k's are robustification parameters.

To conduct simultaneous inference for μ_k's, we use the multiplier bootstrap to approximate the distribution of $\widehat{\mu}_k - \mu_k$. Let W be a random variable with unit mean and variance. Independent of $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$, let $\{W_{ik}, 1 \leq i \leq n, 1 \leq k \leq m\}$ be IID from W. Define the multiplier bootstrap estimators

$$(\widehat{\mu}_k^\flat, \widehat{\boldsymbol{\beta}}_k^\flat) \in \operatorname*{argmin}_{\substack{\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^\intercal)^\intercal: \\ \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_k\|_2 \leq R_k}} \sum_{i=1}^{n} W_{ik} \ell_{\tau_k}(y_{ik} - \mu - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}), \quad k = 1, \ldots, m, \tag{4.2}$$

where $\widehat{\boldsymbol{\theta}}_k = (\widehat{\mu}_k, \widehat{\boldsymbol{\beta}}_k^\intercal)^\intercal$ and R_k's are radius parameters. We will show that the unknown distribution of $\sqrt{n}(\widehat{\mu}_k - \mu_k)$ can be approximated by the conditional distribution of $\sqrt{n}(\widehat{\mu}_k^\flat - \widehat{\mu}_k)$ given $\mathcal{D}_{kn} := \{(y_{ik}, \boldsymbol{x}_i)\}_{i=1}^n$.

The main result of this section establishes validity of the multiplier bootstrap on controlling the FDP in multiple testing. For $k = 1, \ldots, m$, define test statistics $\widehat{T}_k = \sqrt{n}\widehat{\mu}_k$ and the corresponding bootstrap p-values $p_k^\flat = G_k^\flat(|\widehat{T}_k|)$, where $G_k^\flat(z) := \mathbb{P}(\sqrt{n}|\widehat{\mu}_k^\flat - \widehat{\mu}_k| \geq z | \mathcal{D}_{kn})$, $z \geq 0$. For any given threshold $t \in (0, 1)$, the false discovery proportion is defined as

$$\mathrm{FDP}(t) = V(t) / \max\big\{R(t), 1\big\}, \tag{4.3}$$

where $V(t) = \sum_{k \in \mathcal{H}_0} I(p_k^\flat \leq t)$ is the number of false discoveries, $R(t) = \sum_{k=1}^m I(p_k^\flat \leq t)$ is the number of total discoveries and $\mathcal{H}_0 := \{k : 1 \leq k \leq m, \mu_k = 0\}$ is the set of true null hypotheses. For any prespecified $\alpha \in (0, 1)$, applying the the Benjamini and Hochberg (BH) method (Benjamini and Hochberg (1995)) to the bootstrap p-values $p_1^\flat, \ldots, p_m^\flat$ induces a data-dependent threshold

$$t_{\mathrm{BH}}^\flat = p_{(k^\flat)}^\flat \quad \text{with } k^\flat = \max\big\{k : 1 \leq k \leq m, p_{(k)}^\flat \leq \alpha k / m\big\}. \tag{4.4}$$

We reject the null hypotheses for which $p_k^\flat \leq t_{\mathrm{BH}}^\flat$.

CONDITION 4.1. $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$ are IID observations from $(\boldsymbol{y}, \boldsymbol{x})$ that satisfies $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} = (y_1, \ldots, y_m)^\intercal$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\intercal$, $\boldsymbol{\Gamma} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)^\intercal \in \mathbb{R}^{m\times s}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\intercal$. The random vector $\boldsymbol{x} \in \mathbb{R}^s$ satisfies $\mathbb{E}(\boldsymbol{x}) = \boldsymbol{0}$, $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\intercal) = \boldsymbol{\Sigma}$ and $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle| \geq t) \leq 2\exp(-t^2\|\boldsymbol{u}\|_2^2/A_0^2)$ for all $\boldsymbol{u} \in \mathbb{R}^s$, $t \in \mathbb{R}$ and some constant $A_0 > 0$. Independent of $\boldsymbol{x}$, the noise vector $\boldsymbol{\varepsilon}$ has independent elements and satisfies $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $c_l \leq \min_{1\leq k\leq m} \sigma_k \leq \max_{1\leq k\leq m} \upsilon_{k,4}^{1/4} \leq c_u$ for some constants $c_l, c_u > 0$, where $\sigma_k^2 = \mathbb{E}(\varepsilon_k^2)$ and $\upsilon_{k,4} = \mathbb{E}(\varepsilon_k^4)$.

THEOREM 4.1. *Assume Condition 4.1 holds and $m = m(n)$ satisfies $m \to \infty$ and $\log m = o(n^{1/3})$. Moreover, as $(n, m) \to \infty$,*

$$\operatorname{card}\bigl\{k : 1 \leq k \leq m, |\mu_k|/\sigma_k \geq \lambda_0\sqrt{(2\log m)/n}\bigr\} \to \infty \tag{4.5}$$

for some $\lambda_0 > 2$. Then, with

$$\tau_k = \upsilon_k\left\{\frac{n}{s + 2\log(nm)}\right\}^{1/3} \quad \text{and} \quad R_k = \upsilon_k \geq \upsilon_{k,4}^{1/4}, \quad k = 1, \ldots, m,$$

in (4.1) and (4.2), it holds

$$\frac{\mathrm{FDP}(t_{\mathrm{BH}}^\flat)}{(m_0/m)} \to \alpha \quad \text{in probability as } (n, m) \to \infty, \tag{4.6}$$

where $m_0 = \operatorname{card}(\mathcal{H}_0)$.

In practice, conditional quantiles of $\sqrt{n}(\widehat{\mu}_k^\flat - \widehat{\mu}_k)$ can be computed with arbitrary precision by using the Monte Carlo simulations: Independent of the observed data, generate IID random weights $\{W_{ik,b}, 1 \leq i \leq n, 1 \leq k \leq m, 1 \leq b \leq B\}$ from W, where B is the number of bootstrap replications. For each k, the bootstrap samples of $(\widehat{\mu}_k^\flat, \widehat{\boldsymbol{\beta}}_k^\flat)$ are given by

$$(\widehat{\mu}_{k,b}^\flat, \widehat{\boldsymbol{\beta}}_{k,b}^\flat) \in \operatorname*{argmin}_{\mu\in\mathbb{R}, \boldsymbol{\beta}\in\mathbb{R}^s} \sum_{i=1}^n W_{ik,b}\ell_{\tau_k}(y_{ik} - \mu - \boldsymbol{x}_i^\intercal\boldsymbol{\beta}), \quad b = 1, \ldots, B. \tag{4.7}$$

For $k = 1, \ldots, m$, define empirical tail distributions

$$G_{k,B}^\flat(z) = \frac{1}{B+1}\sum_{b=1}^B I\bigl(\sqrt{n}|\widehat{\mu}_{k,b}^\flat - \widehat{\mu}_k| \geq z\bigr), \quad z \geq 0.$$

The bootstrap p-values are thus given by $\{p_{k,B}^\flat = G_{k,B}^\flat(\sqrt{n}|\widehat{\mu}_k|)\}_{k=1}^m$, to which either the BH procedure or Storey's procedure can be applied. For the former, we reject H_{0k} if and only if $p_{k,B}^\flat \leq p_{(k_B^\flat),B}^\flat$, where $k_B^\flat = \max\{k : 1 \leq k \leq m, p_{(k),B}^\flat \leq k\alpha/m\}$ for a predetermined $0 < \alpha < 1$ and $p_{(1),B}^\flat \leq \cdots \leq p_{(m),B}^\flat$ are the ordered bootstrap p-values. See Algorithm 2 for detailed implementations.

5. Numerical studies.

5.1. *Confidence sets.* We first provide simulation studies to illustrate the performance of the robust bootstrap procedure for constructing confidence sets with various heavy-tailed errors. Recall the linear model $Y_i = \boldsymbol{X}_i^\intercal\boldsymbol{\theta}^* + \varepsilon_i$ in (1.1). We simulate $\{\boldsymbol{X}_i\}_{i=1}^n$ from $\mathcal{N}(0, \mathbf{I}_d)$. The true regression coefficient $\boldsymbol{\theta}^*$ is a vector equally spaced between [0, 1]. The errors ε_i are IID from one of the following distributions, standardized to have mean 0 and variance 1.

1. Standard Gaussian distribution $\mathcal{N}(0, 1)$;

Algorithm 2 Huber Robust Multiple Testing

Input: Data $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$, number of bootstrap replications B, thresholding parameters $\{\tau_k\}_{k=1}^m$, nominal level $\alpha \in (0, 1)$.

1: Solve m Huber regressions in (4.1) and obtain $\{(\widehat{\mu}_k, \widehat{\boldsymbol{\beta}}_k)\}_{k=1}^m$.
2: **for** $b = 1, 2, \ldots, B$ **do**
3: Generate IID random weights $\{W_{ik,b}\}_{i\in[n],k\in[m]}$ satisfying $\mathbb{E}(W_{ik,b}) = 1$ and $\operatorname{var}(W_{ik,b}) = 1$.
4: Solve m weighted Huber regressions in (4.7) and obtain $\{(\widehat{\mu}_{k,b}^\flat, \widehat{\boldsymbol{\beta}}_{k,b}^\flat)\}_{k=1}^m$.
5: **end for**
6: **for** $k = 1, \ldots, m$ **do**
7: Compute the bootstrap p-value:

$$p_{k,B}^\flat = \frac{1}{B+1}\sum_{b=1}^{B} I\big(|\widehat{\mu}_{k,b}^\flat - \widehat{\mu}_k| \geq |\widehat{\mu}_k|\big).$$

8: **end for**
9: Sort the bootstrap p-values: $p_{(1),B}^\flat \leq \cdots \leq p_{(m),B}^\flat$.
10: Compute the BH threshold: $k_m^\flat = \max\{1 \leq k \leq m : p_{(k),B}^\flat \leq k\alpha/m\}$.

Output: Set $\{1 \leq k \leq m : p_{k,B}^\flat \leq p_{(k_m^\flat),B}^\flat\}$ of rejections, i.e., reject H_{0k} if

$$p_{k,B}^\flat \leq p_{(k_m^\flat),B}^\flat.$$

2. t_ν-distribution with degrees of freedom $\nu = 3.5$;
3. Gamma distribution with shape parameter 3 and scale parameter 1;
4. t-Weibull mixture (Wbl mix) model: $\varepsilon = 0.5u_t + 0.5u_{\mathrm{W}}$, where u_t follows a standardized t_4-distribution and u_{W} follows a standardized Weibull distribution with shape parameter 0.75 and scale parameter 0.75;
5. Pareto–Gaussian mixture (Par mix) model: $\varepsilon = 0.5u_{\mathrm{P}} + 0.5u_{\mathrm{G}}$, where u_{P} follows a Pareto distribution with shape parameter 4 and scale parameter 1 and $u_{\mathrm{G}} \sim \mathcal{N}(0, 1)$;
6. Lognormal–Gaussian mixture (Logn mix) model: $\varepsilon = 0.5u_{\mathrm{LN}} + 0.5u_{\mathrm{G}}$, where $u_{\mathrm{LN}} = \exp(1.25Z)$ with $Z \sim \mathcal{N}(0, 1)$ and $u_{\mathrm{G}} \sim \mathcal{N}(0, 1)$.

Moreover, we consider three types of random weights as follows:

- Gaussian weights: $W_i \sim \mathcal{N}(0, 1) + 1$;
- Bernoulli weights (rescaled to have mean 1): $W_i \sim 2\,\mathrm{Ber}(0.5)$;
- A mixture of Bernoulli and Gaussian weights considered by Zhilova (2016): $W_i = z_i + u_i + 1$, with $u_i \sim (\mathrm{Ber}(b) - b)\sigma_u$, $b = 0.276$, $\sigma_u = 0.235$, and $z_i \sim \mathcal{N}(0, \sigma_z^2)$, $\sigma_z^2 = 0.038$.

All three weights considered are such that $\mathbb{E}(W_i) = \operatorname{var}(W_i) = 1$. Using nonnegative random weights has the advantage that the weighted objective function is guaranteed to be convex. Numerical results reveal that Gaussian and Bernoulli weights demonstrate almost the same coverage performance.

The number of bootstrap replications is set to be $B = 2000$. Nominal coverage probabilities $1 - \alpha$ are given in the columns, where we consider $1 - \alpha \in \{0.95, 0.90, 0.85, 0.80, 0.75\}$. We report the empirical coverage probabilities from 1000 simulations. We first consider a simple approach for choosing τ, which is set to be $1.2\{\widehat{\upsilon}_4 n/(d + \log n)\}^{1/4}$. Here, $\widehat{\upsilon}_4$ is the empirical fourth moment of the residuals from the OLS and the constant 1.2 (which is slightly larger than 1) is chosen in accordance with Theorem 2.5 which requires $\upsilon \geq \upsilon_4^{1/4}$. This simple ad hoc approach leads to adequate results in most cases. In Section 5.2, we further investigate the empirical performance of the fully data-dependent procedure proposed in Section 3.

TABLE 1
Average coverage probabilities with $n = 100$ and $d = 5$ for different nominal coverage levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$. The weights W_i are generated from $\mathcal{N}(1, 1)$

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.954	0.908	0.842	0.783	0.734
	boot-OLS	0.952	0.908	0.837	0.785	0.735
t_v						
	boot-Huber	0.966	0.904	0.848	0.801	0.748
	boot-OLS	0.954	0.887	0.798	0.710	0.630
Gamma						
	boot-Huber	0.962	0.918	0.860	0.798	0.747
	boot-OLS	0.955	0.910	0.843	0.775	0.700
Wbl mix						
	boot-Huber	0.962	0.907	0.851	0.797	0.758
	boot-OLS	0.944	0.899	0.808	0.775	0.680
Par mix						
	boot-Huber	0.955	0.907	0.856	0.802	0.761
	boot-OLS	0.948	0.900	0.843	0.785	0.738
Logn mix						
	boot-Huber	0.958	0.912	0.860	0.782	0.744
	boot-OLS	0.954	0.912	0.796	0.682	0.616

We compare our method with an OLS-based bootstrap procedure studied in Spokoiny and Zhilova (2015), namely, replacing the weighted Huber loss in (2.12) by the weighted quadratic loss $\mathcal{L}^{\flat}_{\mathrm{ols}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} W_i (Y_i - \boldsymbol{X}_i^{\intercal}\boldsymbol{\theta})^2$.

Consider the sample size $n = 100$ and dimension $d = 5$. Table 1 and Table 2 display the coverage probabilities of the proposed bootstrap Huber method (boot-Huber) and the bootstrap OLS method (boot-OLS). At the normal model, our approach achieves a similar performance as the boot-OLS, which demonstrates the efficiency of adaptive Huber regression. For heavy-tailed errors, our method significantly outperforms the boot-OLS using all three types of random weights. Also, we observe that the Gaussian and Bernoulli weights demonstrate nearly the same desirable performance. For simplicity, we focus on the Gaussian weights throughout the remaining simulation studies.

In Table 3, we increase the sample size to $n = 200$ and retain all the other settings. For most cases of heavy-tailed errors, the coverage probability of the boot-OLS method is lower than the nominal level, sometimes to a large extent. In Table 4, we generate errors from a t-Weilbull mixture distribution and consider different combinations of n ($n \in \{50, 100, 200\}$) and d ($d \in \{2, 5, 10\}$). The robust procedure outperforms the least squares method across most of the settings. Similar phenomena are also observed in other cases of heavy-tailed errors.

We also report the standard deviations of the estimated quantiles of boot-Huber and boot-OLS; see Appendix F.1 in the supplement. The experimental results show that the boot-Huber leads to uniformly smaller standard deviations. Furthermore, we consider more challenging settings with correlated or non-Gaussian designs and nonequally spaced $\boldsymbol{\theta}^*$. The average coverage probabilities of the boot-Huber method are in general close to nominal level, while the boot-OLS leads to severe under-coverage in many heavy-tailed noise settings. More details are presented in Appendix F.2 in the supplementary material (Chen and Zhou (2019)).

TABLE 2
Average coverage probabilities with $n = 100$ and $d = 5$ for different nominal coverage levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Bernoulli weights $W_i \sim 2\,\mathrm{Ber}(0.5)$						
Gaussian						
	boot-Huber	0.947	0.895	0.847	0.792	0.740
	boot-OLS	0.926	0.865	0.824	0.783	0.726
t_ν						
	boot-Huber	0.930	0.897	0.841	0.786	0.755
	boot-OLS	0.884	0.815	0.757	0.703	0.646
Gamma						
	boot-Huber	0.943	0.900	0.861	0.805	0.756
	boot-OLS	0.935	0.882	0.831	0.774	0.720
Wbl mix						
	boot-Huber	0.948	0.894	0.842	0.793	0.739
	boot-OLS	0.931	0.859	0.779	0.721	0.664
Par mix						
	boot-Huber	0.932	0.875	0.832	0.780	0.741
	boot-OLS	0.927	0.871	0.817	0.765	0.716
Logn mix						
	boot-Huber	0.944	0.892	0.846	0.807	0.758
	boot-OLS	0.915	0.838	0.792	0.720	0.674
Mixture weights $W_i = z_i + u_i + 1$, with $u_i \sim (\mathrm{Ber}(b) - b)\sigma_u$, $b = 0.276$, $\sigma_u = 0.235$, and $z_i \sim \mathcal{N}(0, \sigma_z^2)$, $\sigma_z^2 = 0.038$						
Gaussian						
	boot-Huber	0.930	0.864	0.812	0.763	0.698
	boot-OLS	0.920	0.842	0.772	0.695	0.640
t_ν						
	boot-Huber	0.942	0.893	0.844	0.788	0.733
	boot-OLS	0.894	0.792	0.695	0.605	0.528
Gamma						
	boot-Huber	0.930	0.873	0.831	0.782	0.731
	boot-OLS	0.911	0.832	0.754	0.686	0.625
Wbl mix						
	boot-Huber	0.963	0.919	0.874	0.812	0.754
	boot-OLS	0.924	0.759	0.656	0.545	0.459
Par mix						
	boot-Huber	0.942	0.884	0.818	0.750	0.702
	boot-OLS	0.924	0.830	0.736	0.664	0.614
Logn mix						
	boot-Huber	0.940	0.876	0.829	0.796	0.751
	boot-OLS	0.903	0.812	0.736	0.637	0.579

5.2. *Performance of the data-driven tuning approach.* We further investigate the empirical performance of the data-driven procedure proposed in Section 3. We consider lognormal distributions $\mathrm{Logn}(\mu, \sigma)$ with location parameter $\mu = 0$ and varying shape parameters σ. The larger the value of σ is, the heavier the tail is. Moreover, we take $n = 200$, $d = 5$ and $1 - \alpha \in [0.85, 0.99]$ and compare three methods: (1) Huber-based bootstrap procedure with τ calibrated by solving (3.20) (adaptive boot-Huber), (2) Huber-based bootstrap procedure with $\tau = 1.2\{\widehat{\nu}_4 n/(d + \log n)\}^{1/4}$ (boot-Huber), and (3) OLS-based bootstrap method (boot-OLS).

TABLE 3

Average coverage probabilities with $n = 200$, $d = 5$ for different nominal coverage levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$. The weights W_i are generated from $\mathcal{N}(1, 1)$

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.957	0.910	0.850	0.790	0.736
	boot-OLS	0.955	0.907	0.850	0.789	0.736
t_v						
	boot-Huber	0.958	0.906	0.848	0.798	0.749
	boot-OLS	0.940	0.863	0.772	0.684	0.599
Gamma						
	boot-Huber	0.948	0.899	0.845	0.780	0.726
	boot-OLS	0.944	0.889	0.822	0.751	0.685
Wbl mix						
	boot-Huber	0.954	0.889	0.837	0.775	0.713
	boot-OLS	0.939	0.865	0.784	0.695	0.621
Par mix						
	boot-Huber	0.945	0.898	0.847	0.789	0.738
	boot-OLS	0.941	0.886	0.820	0.757	0.700
Logn mix						
	boot-Huber	0.958	0.916	0.864	0.812	0.748
	boot-OLS	0.938	0.886	0.812	0.718	0.590

TABLE 4

Average coverage probabilities for the Wbl mix error and for different nominal coverage levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$. The weights W_i are generated from $\mathcal{N}(1, 1)$

Approach	d	n	0.95	0.9	0.85	0.8	0.75
boot-Huber							
		50	0.951	0.904	0.848	0.789	0.725
	2	100	0.959	0.914	0.866	0.827	0.771
		200	0.954	0.917	0.856	0.814	0.756
		50	0.982	0.945	0.876	0.826	0.752
	5	100	0.966	0.917	0.855	0.802	0.760
		200	0.950	0.894	0.835	0.777	0.721
		50	0.990	0.972	0.955	0.915	0.881
	10	100	0.980	0.949	0.897	0.850	0.799
		200	0.970	0.922	0.864	0.826	0.777
boot-OLS							
		50	0.942	0.887	0.827	0.758	0.672
	2	100	0.956	0.901	0.849	0.785	0.714
		200	0.947	0.898	0.822	0.763	0.685
		50	0.976	0.911	0.836	0.754	0.688
	5	100	0.954	0.896	0.824	0.751	0.674
		200	0.940	0.868	0.790	0.698	0.622
		50	0.997	0.970	0.919	0.844	0.761
	10	100	0.975	0.921	0.850	0.784	0.719
		200	0.954	0.879	0.816	0.731	0.650

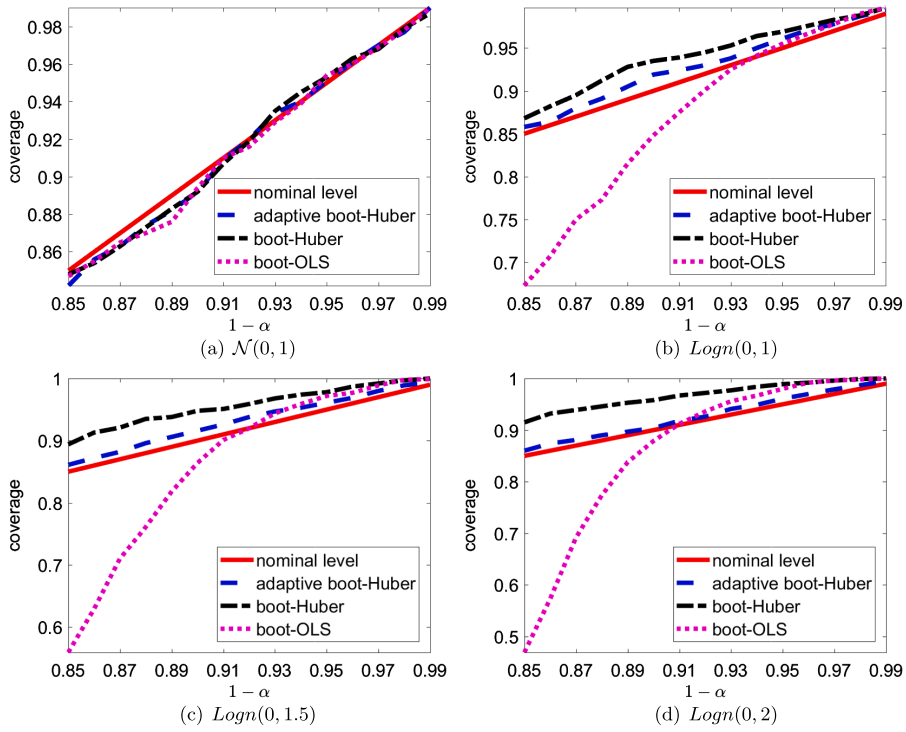

FIG. 1. *Comparison of coverage probabilities for different error distributions when the nominal coverage level* $1-\alpha$ *ranges from* 0.85 *to* 0.99. *Here, x-axis represents* $1-\alpha$ *and y-axis represents the average coverage rates over* 1000 *simulations. The red line represents the nominal coverage probability.*

From Figure 1 and Table 5 we see that, under lognormal models, the coverage probabilities of the adaptive boot-Huber method are closest to nominal levels, while the boot-OLS suffers from distorted empirical coverage: it tends to overestimate the real quantiles at high levels and

TABLE 5

Average coverage probabilities for different nominal coverage levels $1-\alpha \in \{0.99, 0.97, 0.95, 0.9, 0.87\}$. *The weights* W_i *are generated from* $\mathcal{N}(1,1)$

Noise	Approach	0.99	0.97	0.95	0.90	0.87
$\mathcal{N}(0,1)$						
	adaptive boot-Huber	0.993	0.970	0.942	0.896	0.868
	boot-Huber	0.991	0.971	0.946	0.899	0.868
	boot-OLS	0.993	0.970	0.948	0.895	0.868
Logn(0, 1)						
	adaptive boot-Huber	0.994	0.978	0.961	0.919	0.880
	boot-Huber	0.997	0.983	0.969	0.935	0.895
	boot-OLS	0.997	0.978	0.955	0.848	0.750
Logn(0, 1.5)						
	adaptive boot-Huber	0.994	0.980	0.961	0.916	0.882
	boot-Huber	1.000	0.992	0.978	0.948	0.921
	boot-OLS	0.999	0.989	0.972	0.864	0.710
Logn(0, 2)						
	adaptive boot-Huber	0.995	0.979	0.961	0.904	0.881
	boot-Huber	1.000	0.996	0.989	0.958	0.939
	boot-OLS	1.000	0.996	0.980	0.879	0.692

severely underestimate the real quantiles at relatively lower levels. In addition, Figure 1(a) shows that the proposed Huber-based procedure almost loses no efficiency under a normal model.

6. Discussion. In this paper, we have proposed and analyzed robust inference methods for linear models with heavy-tailed errors. Specifically, we use a multiplier bootstrap procedure for constructing sharp confidence sets for adaptive Huber estimators and conducting large-scale simultaneous inference with heavy-tailed panel data. Our theoretical results provide explicit bounds for the bootstrap approximation errors and justify the bootstrap validity; the error of coverage probability is small as long as d^3/n is small. For multiple testing, we show that when the error distributions have finite 4th moments and the dimension m and sample size n satisfy $\log m = o(n^{1/3})$, the bootstrap Huber procedure asymptotically controls the overall false discovery proportion at the nominal level.

The multiplier bootstrap can also be used to construct confidence intervals for the regression coefficients. Let $W_1, \ldots, W_n$ be independent random variables from $2\,\mathrm{Ber}(0.5)$, and define the multiplier bootstrap estimator $\widehat{\boldsymbol{\theta}}_\tau^\flat = (\widehat{\theta}_1^\flat, \ldots, \widehat{\theta}_d^\flat)^\intercal \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}_\tau^\flat(\boldsymbol{\theta})$, where $\mathcal{L}_\tau^\flat(\cdot)$ is given in (2.12). To reduce computational cost (comparing with the nonparametric paired bootstrap), here we recommend using nonnegative weights so that the weighed Huber loss $\mathcal{L}_\tau^\flat(\cdot)$ is still convex. Since each W_i takes the values $\{0, 2\}$ equiprobably, nearly half of the weights are zero, thus reducing the computational complexity of solving weighted Huber regression. Let $\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\theta}_1, \ldots, \widehat{\theta}_d)^\intercal$ be the Huber estimator given in (2.2). For every $1 \le j \le d$ and $q \in (0, 1)$, define the conditional upper q-quantile of $\widehat{\theta}_j^\flat - \widehat{\theta}_j$ given $\mathcal{D}_n$ as

$$c_j^\flat(q) = \inf\{z \in \mathbb{R} : \mathbb{P}^*(\widehat{\theta}_j^\flat - \widehat{\theta}_j > z) \le q\}.$$

At a prescribed confidence level $1 - \alpha \in (0, 1)$, the corresponding multiplier bootstrap confidence intervals for θ_j^*'s are

$$\mathcal{I}_j^\flat = [\widehat{\theta}_j - c_j^\flat(\alpha/2), \widehat{\theta}_j - c_j^\flat(1-\alpha/2)], \quad j = 1, \ldots, d.$$

The Matlab code that implements this procedure and further comparisons are available from https://www.math.ucsd.edu/~wez243/Huber_CI.zip.

Acknowledgements. The authors are grateful to the Editors and reviewers for thoughtful feedback and constructive comments.

The first author was supported in part by NSF Grant IIS-1845444 and the Bloomberg Data Science Research Grant.

The second author was supported in part by NSF Grant DMS-1811376.

SUPPLEMENTARY MATERIAL

Supplement to "Robust inference via multiplier bootstrap" (DOI: 10.1214/19-AOS1863SUPP; .pdf). This supplement material contains (1) the proofs of Theorems 2.2–2.6 and Theorem 3.1 in the main text, (2) implementations of the proposed methods, and (3) additional simulation studies.

REFERENCES

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.* **38** 51–82. MR2589316 https://doi.org/10.1214/08-AOS667

AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *Ann. Statist.* **39** 2766–2794. MR2906886 https://doi.org/10.1214/11-AOS918

BARRAS, L., SCAILLET, O. and WERMERS, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *J. Finance* **65** 179–216.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERK, J. B. and GREEN, R. C. (2004). Mutual fund flows and performance in rational markets. *J. Polit. Econ.* **112** 1269–1295.

BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602 https://doi.org/10.1214/15-AOS1350

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CATONI, O. and GIULINI, L. (2017). Dimension free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. Technical Report.

CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436. MR2157808 https://doi.org/10.1214/009053604000000904

CHEN, X. and ZHOU, W.-X (2020). Supplement to "Robust inference via multiplier bootstrap." https://doi.org/10.1214/19-AOS1863SUPP.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 https://doi.org/10.1214/14-AOS1235

DELAIGLE, A., HALL, P. and JIN, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 283–301. MR2815777 https://doi.org/10.1111/j.1467-9868.2010.00761.x

DESAI, K. H. and STOREY, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *J. Amer. Statist. Assoc.* **107** 135–151. MR2949347 https://doi.org/10.1080/01621459.2011.645777

DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. *Springer Series in Statistics*. Springer, New York. MR2373771 https://doi.org/10.1007/978-0-387-49317-6

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.

FAN, J., HALL, P. and YAO, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.* **102** 1282–1288. MR2372536 https://doi.org/10.1198/016214507000000969

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

FAN, J., LIAO, Y. and YAO, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83** 1497–1541. MR3384226 https://doi.org/10.3982/ECTA12749

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332

GIULINI, I. (2017). Robust PCA and pairs of projections in a Hilbert space. *Electron. J. Stat.* **11** 3903–3926. MR3714302 https://doi.org/10.1214/17-EJS1343

HAHN, M. G., KUELBS, J. and WEINER, D. C. (1990). The asymptotic joint distribution of self-normalized censored sums and sums of squares. *Ann. Probab.* **18** 1284–1341. MR1062070

HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** 18. MR3491112

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

LAN, W. and DU, L. (2019). A factor-adjusted multiple testing procedure with application to mutual fund selection. *J. Bus. Econom. Statist.* **37** 147–157. MR3910232 https://doi.org/10.1080/07350015.2017.1294078

LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. MR1147167 https://doi.org/10.1137/1136085

LINTNER, J. (1965). The valuation of risk assets and the selection of risky investment in stock portfolios and capital budgets. *Rev. Econ. Stat.* **47** 13–37.

LIU, W. and SHAO, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025. MR3262475 https://doi.org/10.1214/14-AOS1249

LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

QI, L. and SUN, D. (1999). A survey of some nonsmooth equations and smoothing Newton methods. In *Progress in Optimization. Appl. Optim.* **30** 121–146. Kluwer Academic, Dordrecht. MR1719516 https://doi.org/10.1007/978-1-4613-3285-5_7

SHARPE, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance* **19** 425–442.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675. MR3405607 https://doi.org/10.1214/15-AOS1355

SUN, Q., ZHOU, W.-X. and FAN, J. (2019). Adaptive Huber regression. Technical Report.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9** 60–62.

ZHILOVA, M. (2016). Non-classical Berry–Esseen inequality and accuracy of the weighted bootstrap. Technical Report.

ZHOU, W.-X., BOSE, K., FAN, J. and LIU, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46** 1904–1931. MR3845005 https://doi.org/10.1214/17-AOS1606