Electronic Journal of Statistics
Vol. 15 (2021) 5933–5994
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1936

Scale calibration for high-dimensional robust regression*

Po-Ling Loh

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
e-mail: pll28@cam.ac.uk

Abstract: We present a new method for robust high-dimensional linear regression when the scale parameter of the additive errors is unknown. The proposed estimator is based on a penalized Huber M-estimator, for which theoretical bounds on the estimation error have recently been proposed in high-dimensional statistics literature. However, the variance of the error term in the linear model is intricately connected to the optimal parameter used to define the shape of the Huber loss. Our main idea is to use an adaptive technique, based on Lepski's method, to overcome the difficulties of solving a joint nonconvex optimization problem with respect to the location and scale parameters. Furthermore, by including a weight term in the definition of the M-estimator, our consistency results hold even when the covariates are heavy-tailed. We then derive asymptotic normality of a one-step estimator constructed from the penalized Huber estimator, which can be used to construct confidence regions for subsets of coordinates. The one-step estimator is shown to be semiparametrically efficient when the covariates are sub-exponential. Our results substantially generalize previous work on high-dimensional inference, derived under sub-Gaussian assumptions on both the covariate and error distributions.

MSC2020 subject classifications: Primary 62F10; secondary 62J07.
Keywords and phrases: Adaptive scale estimation, high-dimensional inference, Lepski's method, robust linear regression, one-step estimation, semiparametric efficiency.

Received April 2020.

1. Introduction

Robust statistics, in its classical form, is a mature and established field [37, 57, 32]. Recently, notions from robust statistics such as ϵ-contamination and influence functions have surfaced in theoretical computer science and machine learning [20, 48]. The use of the Huber loss in place of a squared error loss to encourage robustness has long been adopted in engineering fields, as well [25].

In statistics, a small but growing body of work concerns analyzing high-dimensional analogs of classical robust estimators [47, 78, 56, 10, 53, 23, 70, 71, 27]. The basic premise is that although it is relatively straightforward to devise reasonable high-dimensional estimators, theoretical analysis may become

*Part of this work was completed while the author was visiting the Isaac Newton Institute in Cambridge, UK. The author was supported by NSF grant DMS-1749857.

somewhat trickier in high dimensions [2]. Furthermore, special care must be taken when optimizing such objective functions over a high-dimensional space [1].

Our previous work [53] developed a theory for robust high-dimensional linear regression estimators using penalized M-estimation. The main contribution was to show that global optima of ℓ_1-penalized M-estimators enjoy the same rates of convergence as minimizers of the Lasso program, when the M-estimation loss function is convex and has a bounded derivative—without requiring a Gaussian or sub-Gaussian assumption on the additive errors. In fact, we also established that local optima of penalized M-estimators with a nonconvex, bounded-derivative loss are statistically consistent within a constant-radius region of the global optimum, and such local optima may be obtained via a two-step process initialized using a global optimum of the ℓ_1-penalized Huber loss.

However, a drawback of Loh [53], as well as other related work on penalized M-estimation [23, 71], is that the theoretically optimal choice of the parameter involved in defining the Huber loss depends critically on the scale of the additive errors. This should not be surprising, given that similar complications were recognized in low-dimensional settings for location estimation, when prior knowledge of the scale was unavailable [36]. The "adaptive" methods proposed for low-dimensional robust regression [39, 34] are mostly heuristic suggestions involving, e.g., computing the Huber regression estimate over a grid of values and choosing the parameter that minimizes a surrogate for asymptotic variance. Even in low dimensions, a theoretical gap has remained in terms of how to rigorously calibrate the Huber loss function in a finite-sample setting.

In this paper, we introduce a new solution to the problem of adaptively choosing the scale parameter of a robust M-estimator. The key tool is Lepski's method, and the key observation is that whenever the Huber loss parameter is larger than the true scale parameter of the additive errors, it is possible to derive ℓ_1- and ℓ_2-error bounds on the global optimum that increase linearly with the choice of Huber parameter. This allows us to apply Lepski's method to obtain an estimator that behaves comparably well to the oracle estimator. Importantly, our method bypasses the hard optimization problem of jointly estimating the location and scale. We note that Lepski's method could also be invoked in the low-dimensional, unpenalized setting to rigorously obtain robust regression estimators without needing to optimize a nonconvex problem in an ad hoc manner. In addition to relaxing the usual sub-Gaussian distributional assumptions on the additive errors to a finite variance requirement, we also show how to introduce a weight function to downweight leverage points, thus allowing our theory to be applied to a broader range of heavy-tailed covariate distributions, as well.

We further explain how our estimation results can be used to construct confidence intervals for coordinates of the regression vector when the covariates are sub-exponential. Our approach builds directly on recent literature from high-dimensional inference [75, 43], where confidence regions are derived based on asymptotic normality of one-step corrections of an ℓ_1-penalized M-estimator. However, as the success of these methods relies on suitable nonasymptotic error

bounds on the initial estimator, our results on the ℓ_1-penalized Huber estimator fill a gap by providing an appropriate initial estimator which can be used for a wider range of error distributions. One-step estimators themselves originate from classical robust statistics [7], as a method for improving the efficiency of initial (and more computationally tractable) M-estimators. In the same way, whereas the ℓ_1-penalized Huber estimator may suffer from a loss of efficiency—especially when weight functions are introduced to tame the covariate distribution—we show that our proposed one-step estimators enjoy the property of semiparametric efficiency, thus implying optimality of the resulting confidence regions.

Related work: Other proposals for regression with heavy-tailed errors include work by Hsu and Sabato [35], Minsker [58], and Lugosi and Mendelson [55]. However, many of these methods focus on situations where the covariates are well-behaved, and all of them assume knowledge of an upper bound on the error variance. In contrast, our method produces consistent estimators under much milder assumptions on the covariates, and encompasses situations where preliminary scale estimates are notoriously difficult to obtain. Nonetheless, a benefit of the methods introduced in the aforementioned papers is that they can also be shown to be robust in situations where a constant fraction of the data is *adversarially* contaminated [46, 19, 51, 18, 15, 67, 62, 3].

Another important related work is by Chichignoud et al. [17], who suggest an adaptive method for tuning parameter selection in the Lasso based on Lepski's method. However, the main focus in that paper is in obtaining near-optimal bounds on the ℓ_∞-error. Importantly, the objective function still involves a least-squares loss as in the classical Lasso, whereas our objective functions are designed for robust regression and have the corresponding parameter linked to the regularization parameter involved in the ℓ_1-norm.

On the topic of inference, Belloni et al. [6] introduced a different method for constructing confidence intervals in high-dimensional regression settings based on a one-step correction to ℓ_1-penalized M-estimators. Although this approach is somewhat orthogonal to ours, one benefit of Belloni et al. [6] is that the method can be applied to a broader class of M-estimators than ours, since the smoothness conditions on the loss function are not as stringent. On the other hand, our approach has benefits in terms of semiparametric efficiency for estimation of multiple target parameters (cf. Remark 8 in Section 4.3 below).

Finally, we mention another recent proposal for calibrating the tuning parameter in high-dimensional penalized Huber regression [80]. This is a somewhat heuristic method based on iteratively solving the empirical version of a system of equations which, at the population level, has a unique solution equal to the theoretically optimal parameter. We end by noting that although several alternative tuning-free approaches for high-dimensional regression have been proposed, e.g., based on penalized quantile regression [11, 79, 77, 22], the square root Lasso [5], or the Wilcoxon loss from nonparametric statistics [81], to the best of our knowledge, these alternative approaches also require stronger assumptions on the covariate distributions than we impose in our paper. It is unclear whether the analysis in these papers could be extended to settings where

weights are introduced to dampen the effect of outliers.

Notation: For a vector $v \in \mathbb{R}^p$, we write $\text{supp}(v) \subseteq \{1, \ldots, p\}$ to denote the support of v, and for an arbitrary subset $S \subseteq \{1, \ldots, p\}$, we write $v_S \in \mathbb{R}^S$ to denote the vector v restricted to S. For a matrix M, we write $\|M\|_q$ to denote the ℓ_q-operator norm, and we write $\|M\|_{\max}$ to denote the elementwise ℓ_∞-norm. We write $\text{vec}(M)$ to denote the vectorized version of the matrix. Let $\mathbb{R}_+$ denote the positive reals.

We use the notation c, C', c_0, etc., to denote universal positive constants, where we may use the same notation to refer to different constants as we move between results. We use the abbreviation "w.h.p." to refer to an event occurring with probability tending to 1 as the problem parameters $n, p \to \infty$. We use the standard big-O notation, so that two functions $f(n)$ and $g(n)$ satisfy $f = O(g)$ if there exist a constant C and an integer n_0 such that $f(n) \leq Cg(n)$ for all $n \geq n_0$. We also write $f \precsim g$, and we define $f \succsim g$ (equivalently, $f = \Omega(g)$) analogously. Finally, for sequences of random variables $\{X_n\}$ and $\{Y_n\}$, we write $X_n = \mathcal{O}_P(Y_n)$ to denote boundedness in probability, i.e., for any $\epsilon > 0$, there exist a constant B_ϵ and an integer n_ϵ such that $\mathbb{P}\left(\left|\frac{X_n}{Y_n}\right| > B_\epsilon\right) < \epsilon$ for all $n \geq n_\epsilon$. We write $X_n = o_P(Y_n)$ to mean that $\frac{X_n}{Y_n} \xrightarrow{P} 1$. We write $f(n) = \text{polylog}(n)$ when $f(n) = g(\log n)$, for some polynomial function g.

2. Background and problem setup

We begin by describing the regression model to be studied in our paper. We also discuss several previously existing proposals in the literature.

2.1. Model and assumptions

Consider observations $\{(x_i, y_i)\}_{i=1}^n$ from the linear model

$$y_i = x_i^T \beta^* + \epsilon_i, \tag{2.1}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown regression parameter vector. We will also assume that $\|\beta^*\|_0 \leq k$, where $k < n \ll p$, and denote $S := \text{supp}(\beta^*)$.

We will work in a random design setting, where the x_i's and ϵ_i's are i.i.d. draws from covariate and error distributions, such that $x_i \perp\!\!\!\perp \epsilon_i$ and $\mathbb{E}[\epsilon_i] = \mathbb{E}[x_i] = 0$. Our results could be adapted to the fixed design setting in a fairly straightforward manner; however, we are primarily interested in a setting where the distribution of the covariates is heavy-tailed, leading to high-leverage points. We will denote the covariance matrix of the x_i's by Σ_x. We will also assume that $c_{\min} \leq \lambda_{\min}(\Sigma_x) \leq \lambda_{\max}(\Sigma_x) \leq c_{\max}$, for some constants $c_{\min}, c_{\max} \in (0, \infty)$.

Turning to the error distribution, we will assume throughout the paper that $\sigma^* := \sqrt{\text{Var}(\epsilon_i)}$ is finite. We will assume that the distribution of ϵ_i is symmetric, as is customary in classical robust statistics to ensure consistency of regression M-estimators. Note, however, that this is not a major limitation of our work—we could first postprocess the data to obtain the transformed dataset

$\left\{\left(\frac{y_{2i}-y_{2i-1}}{\sqrt{2}}, \frac{x_{2i}-x_{2i-1}}{\sqrt{2}}\right)\right\}_{i=1}^{\lfloor n/2 \rfloor}$ and then run the regression algorithm on these points.

We will introduce additional assumptions on the distributions of the ϵ_i's and x_i's in Assumptions 1, 2, and 3 later. Recall the following standard definitions of sub-Gaussian and sub-exponential distributions [76], which will be used in the sequel:

Definition 1. *We say that a random variable X is sub-Gaussian with parameter $\sigma > 0$ if*

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(\frac{-t^2}{\sigma^2}\right),$$

for all $t \geq 0$. We say that a random vector $X \in \mathbb{R}^p$ is sub-Gaussian with parameter σ if $v^T X$ is a sub-Gaussian random variable with parameter σ, for any unit vector $v \in \mathbb{R}^p$.

Definition 2. *We say that a random variable X is sub-exponential with parameter $\sigma > 0$ if*

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(\frac{-t}{\sigma}\right),$$

for all $t \geq 0$.

2.2. Previous work

We now briefly describe several previously proposed methods for robust linear regression in high dimensions. We focus on methods that have been devised to handle outliers in the covariates, since our proposed algorithm is provably consistent when the covariate distribution is heavy-tailed, as well. (For additional related work, see the references cited in the introduction.)

The sparse least trimmed squares (LTS) estimator [1] aims to optimize the objective

$$\widehat{\beta} \in \arg\min_{\beta} \left\{\frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2 + \lambda\|\beta\|_1\right\},$$

where the $r_{(i)}^2$'s are the sorted residuals $\{(y_i - x_i^T\beta)^2\}_{i=1}^n$ in ascending order, and $h \leq n$ is a truncation parameter. This is an ℓ_1-penalized version of the least trimmed squares estimator [65]. Although sparse LTS has been shown to perform well in simulations, only a heuristic algorithm has been proposed for optimizing the objective, and statistical guarantees for both global and local optima are absent from the literature.

The S-Bridge estimator [56, 70] is defined via the objective function

$$\widehat{\beta} \in \arg\min_{\beta} \left\{s^2(r(\beta)) + \lambda\|\beta\|_r^r\right\},$$

where $r > 0$, and $s(r(\beta))$ is a robust scale estimator based on the residuals $\{y_i - x_i^T \beta\}_{i=1}^n$. The MM-Bridge estimator is defined by

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{r_i(\beta)}{s(r(\widehat{\beta}_1))} \right) + \lambda \|\beta\|_r^r \right\},$$

where ρ is a robust loss function and $\widehat{\beta}_1$ is an initial estimate of β^*; for $r = 1$, this method is also known as the MM-Lasso. Smucler and Yohai [70] derived the asymptotic consistency of global optima when the loss function ρ is of a redescending type, meaning that ρ' is eventually equal to 0. However, the results are asymptotic, and again, no guarantees are provided for the performance of local optima, which may result from the optimization algorithm proposed by the authors. Penalized S-estimators are further analyzed in Freue et al. [27].

Our work builds upon Loh [53], which studied local and global optima of penalized M-estimators. The main contribution in that work is a rigorous non-asymptotic analysis of global optima in the convex case, as well as an analysis of certain consistent local optima when the objective function is nonconvex. However, the success of the methods proposed in that paper require the parameter of the Huber loss to be chosen correctly, i.e., upper-bounding an expression involving moments and tails of the error distribution. Since this information would generally be unknown a priori, the question of how to choose the Huber parameter in an adaptive manner remained unanswered.

Finally, we mention methods based on joint estimation of location and scale. One natural approach is to jointly minimize the objective function

$$(\widehat{\beta}, \widehat{\sigma}) \in \arg\min_{\beta, \sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left(\frac{y_i - x_i^T \beta}{\sigma} \right) \right\}$$

(or a high-dimensional analog thereof). However, even when the loss function is convex, this leads to a highly nonconvex objective. Iteratively optimizing with respect to β and σ motivates the MM-estimator [83], but theoretical guarantees in terms of both statistical consistency and convergence of the optimization algorithm are largely absent from the literature. Huber [37] also proposed the concomitant estimator:

$$(\widehat{\beta}, \widehat{\sigma}) \in \arg\min_{\beta, \sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left(\frac{y_i - x_i^T \beta}{\sigma} \right) \sigma + a\sigma \right\}, \tag{2.2}$$

where a is an appropriate constant to ensure Fisher consistency. The key insight is that if ℓ is a convex function, the loss function $\mathcal{L}_n(\beta, \sigma)$ appearing in the objective (2.2) is also jointly convex in (β, σ). However, the choice of the correct constant a to provide consistency is somewhat intricate. A small calculation shows that if we denote $\mathcal{L}(\beta, \sigma) = \mathbb{E}[\mathcal{L}_n(\beta, \sigma)]$, we have $\nabla \mathcal{L}(\beta^*, \sigma^*) = 0$ provided

$$a = \mathbb{E}\left[\ell' \left(\frac{\epsilon_i}{\sigma^*} \right) \frac{\epsilon_i}{\sigma^*} - \ell \left(\frac{\epsilon_i}{\sigma^*} \right) \right]$$

holds. Thus, some prior knowledge of the distribution of ϵ_i is required to choose a appropriately. In contrast, our method results in a consistent estimate of β^* whenever ϵ_i has a symmetric distribution. Another important issue is that if ℓ is nonconvex—as is recommended to deal with high-leverage points in the covariates—Huber's estimator (2.2) would no longer be jointly convex, leading to a more tricky analysis of local optima in the (β, σ) parameter space.

3. Adaptive scale estimation

Consider the Huber loss function

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

defined with respect to a parameter $\tau > 0$. Importantly, the Huber loss is differentiable, and $\|\ell'_\tau\|_\infty \le \tau$. We also define a weight function $w : \mathbb{R}^p \to \mathbb{R}_+$, with characteristics which will be described later. We will study the behavior of the ℓ_1-regularized Huber estimator

$$\widehat{\beta}_\tau \in \arg\min_\beta \left\{ \frac{1}{n} \sum_{i=1}^n \ell_\tau \left((x_i^T \beta - y_i) w(x_i) \right) + \lambda \tau \|\beta\|_1 \right\}. \tag{3.1}$$

The idea of downweighting individual terms as a function of the covariates is a classical idea from robust linear regression, where various authors studied M-estimators of the form $\frac{1}{n}\sum_{i=1}^n v(x_i)\ell\left((x_i^T\beta - y_i)w(x_i)\right)$, or more generally, $\frac{1}{n}\sum_{i=1}^n \eta(x_i, x_i^T\beta - y_i)$ (see Hampel [32, Chapter 6.3] and the references cited therein). The motivation for introducing weights in classical settings was to guarantee infinitesimal robustness of the regression estimator by ensuring that the influence function stayed bounded even when the covariates were contaminated. Although our choice to introduce weights only within the individual arguments of the loss function terms does not exactly coincide with the more popular framework of Mallows or Schweppe weights from classical robust statistics, one should keep in mind that the central study in our analysis is somewhat different (concerning robustness to heavy tails in the covariate distribution, rather than a study of influence or other notions of sensitivity). Nonetheless, the idea of downweighting individual arguments can also be found in the paper by Krasker [49]. See Remark 1 for more connections between suitable choices of weight functions for our theory to hold and classical choices of weight functions from robust statistics.

For our theory, we will assume that the weight function satisfies the following properties:

Assumption 1. *Assume that*

(i) $w(x_i)x_i$ *is sub-Gaussian with parameter* b',
(ii) $\|w\|_\infty \le 1$, *and*
(iii) $\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]\right) \ge c'_{\min}$, *for some constant* $c'_{\min} > 0$.

Note that the conditions of Assumption 1 involve both the weight function and the distribution of the x_i's. As noted in Section 3.1 below, when the x_i's are well-behaved (e.g., sub-Gaussian), we may set $w \equiv 1$, somewhat simplifying the analysis. However, we do *not* in general assume that the x_i's follow a sub-Gaussian distribution—Assumption 1 can be satisfied by arbitrarily heavy-tailed distributions, as long as the weight function is chosen appropriately (cf. Example 2 below).

The proof of the following theorem, based on arguments developed in Loh [53], is contained in Appendix B. Recall that σ^* denotes the standard deviation of the error distribution, and is assumed to be finite.

Theorem 1. *Suppose the weight function satisfies Assumption 1. Suppose the Huber parameter satisfies $\tau \geq c_\tau \sigma^*$, the regularization parameter is chosen to be $\lambda = 2c_0\sqrt{\frac{\log p}{n}}$, and the sample size satisfies $n \gtrsim k \log p$. Then the estimate $\widehat{\beta}_\tau$ from ℓ_1-penalized Huber regression with parameter τ satisfies*

$$\|\widehat{\beta}_\tau - \beta^*\|_2 \leq C\tau\sqrt{\frac{k\log p}{n}}, \qquad \text{and}$$
$$\|\widehat{\beta}_\tau - \beta^*\|_1 \leq 4\sqrt{k}\|\widehat{\beta}_\tau - \beta^*\|_2,$$

with probability at least $1 - c_1 p^{-c_2}$, where the c_i's are universal constants.

Importantly, the choice of $\lambda = 2c_0\sqrt{\frac{\log p}{n}}$ in Theorem 1 depends only on a universal constant c_0. This is in contrast to the usual Lasso, which requires the tuning parameter λ to be proportional to the unknown quantity σ^*.

We also comment on the requirement that $\tau \geq c_\tau \sigma^*$, where c_τ is an appropriately defined constant. We will provide a method in the next subsection for adaptively choosing τ without prior knowledge of σ^*, with a guarantee that the estimator obtained from our procedure is at least as good as the estimator obtained by taking the theoretically optimal choice $\tau = c_\tau \sigma^*$. However, suppose momentarily that we are able to set the Huber parameter τ equal to $c_\tau\sigma^*$, and consider for the sake of illustration that the ϵ_i's are drawn from a mixture distribution $(1-\zeta)F + \zeta G$, where F and G are both zero-mean sub-Gaussian distributions with sub-Gaussian parameters $\sigma_F \leq \sigma_G$, and ζ is the mixing probability. Standard results on sub-Gaussian distributions imply that the mixture distribution is also sub-Gaussian, with parameter bounded by σ_G. Thus, Lasso theory implies that $\|\widehat{\beta}_{\text{Lasso}} - \beta^*\|_2 \precsim \sigma_G\sqrt{\frac{k\log p}{n}}$. On the other hand, the variance of the mixture distribution is a weighted combination of the variances of F and G, hence is bounded by a constant multiple of $(1-\zeta)\sigma_F^2 + \zeta\sigma_G^2$. If ζ is close to 0, the result of Theorem 1 translates into the ℓ_2-error bound $\|\widehat{\beta}_\tau - \beta^*\|_2 \precsim \tau\sigma_F\sqrt{\frac{k\log p}{n}}$ on the ℓ_1-penalized Huber estimator. If $\sigma_F \ll \sigma_G$, this can lead to significant gains in the estimation error in comparison to the Lasso.

3.1. Examples

We now explore the applicability of Theorem 1 in some specific examples. In particular, we will discuss combinations of weight functions and covariate distributions under which the conditions of Assumption 1 are satisfied.

Example 1 (Sub-Gaussian distributions)**.** *When the distribution of x_i is sub-Gaussian, we can simply choose $w \equiv 1$; i.e., we do not need to downweight any of the terms in the objective (3.1) in order to obtain the desired error bounds. Indeed, the vanilla form of Huber regression is known to perform well when leverage points are not present. In particular, the distribution of $w(x_i)x_i$ is sub-Gaussian, and provided $\lambda_{\min}(\Sigma_x) \geq c_{\min} > 0$, we have $\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]\right) \geq c_{\min}$, as well.*

Example 2 (Spherically symmetric distributions)**.** *Now assume that the distribution of x_i is spherically symmetric, meaning $x_i \overset{d}{=} RU$, where U is uniformly distributed on the unit sphere and R is a scalar random variable. Then Σ_x is a multiple of the identity; for this illustration, assume $\Sigma_x = I$ for simplicity. We present families of distributions such that*

$$w(x) = \min\left\{1, \frac{b\sqrt{p}}{\|x\|_2}\right\} \tag{3.2}$$

will satisfy the desired properties for sufficiently large p, where $b > 0$ is a constant which does not depend on p.

We first verify Assumption 1(i). Recall that $\sqrt{p}\frac{x_i}{\|x_i\|_2}$, which is uniformly distributed on the surface of the sphere of radius $\sqrt{p}$, is sub-Gaussian with parameter $\sigma = \Theta(1)$ [76, Theorem 3.4.5]. Hence, for a unit vector $v \in \mathbb{R}^p$, we have

$$\begin{aligned}
&\mathbb{P}(|w(x_i)x_i^T v| \geq t) \\
&\quad = \mathbb{P}\left(|v^T x_i| \geq t, \|x_i\|_2 \leq b\sqrt{p}\right) + \mathbb{P}\left(b\sqrt{p}\frac{|v^T x_i|}{\|x_i\|_2} \geq t, \|x_i\|_2 > b\sqrt{p}\right) \\
&\quad \leq \mathbb{P}\left(\frac{|v^T x_i|}{\|x_i\|_2} \geq \frac{t}{b\sqrt{p}}\right) + \mathbb{P}\left(\frac{|v^T x_i|}{\|x_i\|_2} \geq \frac{t}{b\sqrt{p}}\right) \\
&\quad \leq 4\exp\left(\frac{-t^2}{\sigma^2 b^2}\right),
\end{aligned}$$

from which it follows that $w(x_i)x_i$ is sub-Gaussian with parameter $\Theta(b)$ (cf. Definition 1).

Next, note that for any unit vector $v \in \mathbb{R}^p$, we have

$$\begin{aligned}
&v^T\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]v \\
&\quad = v^T\mathbb{E}\left[x_i x_i^T 1\left\{\|x_i\|_2 \leq b\sqrt{p}\right\}\right]v + v^T\mathbb{E}\left[b^2 p\frac{x_i x_i^T}{\|x_i\|_2^2}1\left\{\|x_i\|_2 > b\sqrt{p}\right\}\right]v \\
&\quad \geq v^T\mathbb{E}\left[b^2 p\frac{x_i x_i^T}{\|x_i\|_2^2}1\left\{\|x_i\|_2 > b\sqrt{p}\right\}\right]v
\end{aligned}$$

$$= v^T \mathbb{E}\left[b^2 p \frac{x_i x_i^T}{\|x_i\|_2^2}\right] v - v^T \mathbb{E}\left[b^2 p \frac{x_i x_i^T}{\|x_i\|_2^2} 1\left\{\|x_i\|_2 \le b\sqrt{p}\right\}\right] v$$

$$\ge v^T \mathbb{E}\left[b^2 p \frac{x_i x_i^T}{\|x_i\|_2^2}\right] v - \sqrt{\mathbb{E}\left[\left(\frac{b^2 p (x_i^T v)^2}{\|x_i\|_2^2}\right)^2\right] \mathbb{P}\left(\|x_i\|_2^2 \le b^2 p\right)}, \tag{3.3}$$

using the Cauchy-Schwarz inequality in the last line.

Since the covariance matrix of the uniform spherically distributed vector $\sqrt{p}\frac{x_i}{\|x_i\|_2}$ *is the identity [24, Theorem 2.7], the first term on the right-hand side of inequality* (3.3) *is equal to* b^2*. For the second term, we use additional properties of moments of spherically symmetric distributions. By Lemma 17, we can compute*

$$\lim_{p\to\infty} \mathbb{E}\left[\left(\frac{p(x_i^T v)^2}{\|x_i\|_2^2}\right)^2\right] = 3.$$

We now study conditions for which

$$\mathbb{P}\left(\|x_i\|_2^2 \le b^2 p\right) \le \frac{1}{12}, \tag{3.4}$$

in which case inequality (3.3) *implies that*

$$\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i) x_i x_i^T\right]\right) \ge \frac{b^2}{2},$$

giving Assumption 1(iii). Note that

$$\mathbb{P}\left(\|x_i\|_2^2 \le b^2 p\right) = \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p} x_{ij}^2 \le b^2\right).$$

If $x_i \sim N(0, I)$*, this inequality will certainly hold for any* $b < 1$ *for sufficiently large* p*, since the* x_{ij}*'s are i.i.d. and the empirical average concentrates. More generally, Guédon and Milman [31] established a similar concentration inequality when* x_i *is a log-concave distribution, with later generalizations to distributions with heavier tails [26].*

In the aforementioned cases, the left-hand side of inequality (3.4) *actually tends to 0 as* $p \to \infty$*. However, we only need the expression to be upper-bounded by a constant. Inequality* (3.4) *can be rewritten as*

$$\mathbb{P}\left(R^2 \le b^2 p\right) \le \frac{1}{12}, \tag{3.5}$$

where we recall that $\mathbb{E}[R^2] = p$ *due to the assumed isotropy condition. Theorem 2.9 of Fang et al. [24] provides the density function for* R*, from which the condition* (3.4) *can further be verified for sufficiently small* b *for various classes of distributions. Note that this line of argument allows the random variable* R*, and consequently also* x_i*, to be arbitrarily heavy-tailed, as long as it possesses a finite second moment.*

Finally, note that while Example 2 is stated for spherically symmetric distributions, a similar weight assignment would work if the x_i's were elliptically symmetric, instead, i.e., Bx_i is spherically symmetric for a well-conditioned matrix $B \in \mathbb{R}^{p\times p}$. In this case, we could define the weight function $w(x) = \min\left\{1, \frac{b\sqrt{p}}{\|Bx\|_2}\right\}$ and follow nearly identical derivations as above. Thus, in practice, one might choose to define the weights according to $w(x_i) = \min\left\{1, \frac{\sqrt{p}}{\|\widehat{\Theta}x_i\|_2}\right\}$, where $\widehat{\Theta}$ is an estimate of Σ^{-1}.

Remark 1. *The choice of weight function (3.2) is a special case of the family of weight functions studied in classical robust regression literature [32, Chapter 6.3], where $w_B(x) = \min\left\{1, \frac{1}{\|Bx\|_2}\right\}$ for $B \in \mathbb{R}^{p\times p}$. Optimal choices of B have accordingly been derived to satisfy various criteria, e.g., maximum efficiency subject to bounds on the gross-error sensitivity and/or local-shift sensitivity, in which case the choice of parameters is implicitly derived from the desired upper bounds. Note, however, that since we are only interested in obtaining high-probability error bounds of the correct order, we do not need as fine-grained a characterization of the matrix B as in the classical setting. Thus, our discussion in Example 2 above, which specifies that $B \asymp \frac{1}{\sqrt{p}}$, is sufficient for our purposes. See also Krasker and Welsch [50] and Huber [38].*

3.2. Lepski's method

We now discuss Lepski's method [52, 9, 14, 59]. Consider $\tau_{\min}$ and $\tau_{\max}$ such that $\tau_{\min} \le c_\tau \sigma^* \le \tau_{\max}$. Let $\tau_j = \tau_{\min} 2^j$, and define

$$\mathcal{J} = \{j \ge 1 : \tau_{\min} \le \tau_j < 2\tau_{\max}\}.$$

Note that $|\mathcal{J}| \le \log_2\left(\frac{2\tau_{\max}}{\tau_{\min}}\right)$.

Let $\widehat{\beta}_{(j)}$ denote the output of the regression procedure with $\tau = \tau_j$, and define

$$j_* = \min\left\{j \in \mathcal{J} : \ \forall i > j \text{ s.t. } i \in \mathcal{J}, \ \|\widehat{\beta}_{(i)} - \widehat{\beta}_{(j)}\|_2 \le 2C\tau_i\sqrt{\frac{k\log p}{n}} \right.$$
$$\left. \text{and } \|\widehat{\beta}_{(i)} - \widehat{\beta}_{(j)}\|_1 \le 8C\tau_i k\sqrt{\frac{\log p}{n}}\right\}. \quad (3.6)$$

(We define $j_* = \infty$ if no such indices exist, but we will show that $j_* < \infty$, w.h.p.) Thus, to compute j_*, we perform pairwise comparisons of regression estimates obtained over the gridding of the interval $[\tau_{\min}, 2\tau_{\max}]$.

Note that if our goal were simply to obtain ℓ_2-consistency, we could apply Lepski's method where j_* is defined only with respect to comparisons involving the ℓ_2-error. However, we will need ℓ_1-error bounds for the one-step derivations later, so we include both deviations in the screening process here. We then have the following result:

Theorem 2. *Under the same conditions as Theorem 1, with probability at least*

$$1 - \log_2\left(\frac{4\tau_{\max}}{\tau_{\min}}\right) cp^{-c'},$$

we have

$$\|\widehat{\beta}_{(j_*)} - \beta^*\|_2 \leq 6Cc_\tau\sigma^*\sqrt{\frac{k\log p}{n}}, \quad \text{and} \quad \|\widehat{\beta}_{(j_*)} - \beta^*\|_1 \leq 24Cc_\tau\sigma^* k\sqrt{\frac{\log p}{n}}.$$

The proof follows from straightforward algebraic manipulations and is contained in Appendix C.

Note that Lepski's method does not correspond to a standard grid search over τ, which would be more reminiscent of the adaptive robust estimation procedures described in the introduction. Indeed, for each candidate value of τ, we perform a type of guided comparison between different values of τ, rather than simply choosing the value of τ that gives rise to the smallest value of some objective function. Furthermore, the output of a Lepski-type procedure does not necessarily correspond to the $\widehat{\beta}_\tau$ arising from the "optimal" choice of $\tau \asymp \sigma^*$. Rather, we are guaranteed that the ℓ_1- and ℓ_2-error of our final estimate is comparable to the *error of the estimator* generated using the optimal parameter. In contrast, the adaptive procedures appearing in robust statistics literature suggest a method for choosing the optimal σ by minimizing an approximation of the variance of the estimator thus produced.

Remark 2. *Note that our algorithm based on Lepski's method requires knowledge of the sparsity level k, which is one drawback of the procedure. An upper bound k' would also be sufficient, in which case the comparisons used to determine j_* in equation (3.6) would involve k' rather than k. On the other hand, the error guarantees would then also be looser.*

We would also need to have an explicit value for C in order to apply Lepski's method. As seen from the proof, the constant C appearing in our bounds depends on universal constants; the choice τ of the parameter used for the robust loss function; and distributional properties of x_i (i.e., the eigenvalue bounds $\{c_{\min}, c_{\max}, c'_{\min}\}$). The last point is somewhat unsatisfactory. However, in practical applications, we might imagine having numerous observed values of the x_i's available, from which we might be able to estimate these quantities. Importantly, we emphasize that our proposed method does not *require any information about the distribution of the ϵ_i's, which we would not be accessible without a good initial estimate of β^* in practice.*

Although we do not include the derivations here, a similar procedure based on ℓ_∞-error comparisons could be used to obtain an estimator based on Lepski's method with ℓ_∞-error guarantees on the same order as the ℓ_1-penalized Huber estimator with a theoretically optimal parameter. Furthermore, such a procedure would not involve knowledge of the sparsity, since ℓ_∞-error bounds are typically $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ and do not depend on the sparsity. On the other

hand, one would need to impose slightly stronger assumptions in order to derive ℓ_∞-error bounds [53].

3.3. Rough scale parameter bounds

Our application of Lepski's method requires specifying choices of $\tau_{\min}$ and $\tau_{\max}$. We now describe how to select these values in a reasonable manner. We assume we have prior knowledge of the constant c_τ, which only depends on characteristics of the covariate distribution and *not* the unknown error distribution. Then it suffices to compute rough bounds $[\sigma_{\min}, \sigma_{\max}]$ on σ^*. By independence, we have

$$\operatorname{Var}(y_i) = \operatorname{Var}(x_i^T \beta^*) + \operatorname{Var}(\epsilon_i).$$

Hence, we have $(\sigma^*)^2 \leq \operatorname{Var}(y_i)$, and we may select $\sigma^2_{\max}$ to be a rough estimate of $\operatorname{Var}(y_i)$.

Various estimators for population means exist that only involve weak distributional assumptions. For instance, the "median of means" estimator takes as input n i.i.d. observations $X_1, \ldots, X_n$, and then computes the means $\{\widehat{\mu}_j\}_{j=1}^K$ of the K disjoint subsets of $N = \lfloor \frac{n}{K} \rfloor$ observations, for a parameter K. The overall estimate $\widehat{\mu}_{MoM}$ is the median of the means $\{\widehat{\mu}_j\}_{j=1}^K$.

Accordingly, we propose to take $\sigma^2_{\max} = 2\widehat{\sigma}^2_{MoM}$, where $\widehat{\sigma}^2_{MoM}$ is the median-of-means estimator computed from the dataset $\{y_i^2\}_{i=1}^n$. Assuming the existence of $(2+\epsilon)$-moments of x_i and ϵ_i, and using the concentration inequality provided in Lemma 13 of Appendix H, we have

$$\sigma^2_{\max} = 2\widehat{\sigma}^2_{MoM} \geq \mathbb{E}[y_i^2] \geq \operatorname{Var}(\epsilon_i) = (\sigma^*)^2$$

and

$$\sigma^2_{\max} \leq \frac{3}{2} \operatorname{Var}(y_i) = \frac{3}{2} \left((\beta^*)^T \Sigma_x \beta^* + (\sigma^*)^2 \right),$$

with probability at least $1 - c\exp(-c'n)$.

We now turn to the problem of choosing $\sigma_{\min}$. Consider the choice $\sigma_{\min} = \frac{\sigma_{\max}}{2^M}$, for some integer M. Let $\widehat{\beta}$ be the final output of Lepski's method. We have the following result:

Theorem 3. *Suppose Lepski's method is performed on the ℓ_1-penalized Huber problem with $\sigma^2_{\max}$ equal to the median-of-means estimator of $\operatorname{Var}(y_i)$ and $\sigma_{\min} = \frac{\sigma_{\max}}{2^M}$. Suppose $x_i^T \beta^*$ and ϵ_i have finite $(2+\epsilon)$-moments. If*

$$\frac{3}{2^{2M+1}} \left((\beta^*)^T \Sigma_x \beta^* + (\sigma^*)^2 \right) \leq (\sigma^*)^2, \tag{3.7}$$

we have

$$\|\widehat{\beta} - \beta^*\|_2 \leq 6Cc_\tau \sigma^* \sqrt{\frac{k \log p}{n}}, \qquad \text{and} \tag{3.8}$$

$$\|\widehat{\beta} - \beta^*\|_1 \leq 24Cc_\tau \sigma^* k \sqrt{\frac{\log p}{n}}, \tag{3.9}$$

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

Note that if $M = o(p^{c'})$, the bounds (3.8) and (3.9) in Theorem 3 hold w.h.p. If we define the signal-to-noise ratio $SNR := \frac{(\beta^*)^T \Sigma_x \beta^*}{(\sigma^*)^2}$, then inequality (3.7) can be rewritten as $\log_2(SNR + 1) \precsim M$, which is a fairly mild assumption. In particular, if $\lambda_{\max}(\Sigma_x)$ and $\|\beta^*\|_2$ are bounded, then SNR is also bounded and we can even choose M to be a constant. Finally, note that some knowledge of the curvature of the covariate distribution (i.e., maximum eigenvalue of Σ_x) can be helpful in determining the choice of M necessary for inequality (3.7) to be satisfied. Note also that in practice, we would not want M to be too large, since the computational complexity of the algorithm will increase linearly with M.

4. One-step estimators

Although we have established the consistency of our estimators under rather weak distributional assumptions on the x_i's and ϵ_i's, the presence of the weight function $w(x)$ may lead to poor efficiency. Classical theory for regression M-estimators suggests that efficiency might be improved by using a loss which is governed by the specific form of the error density. The theory of M-estimation from classical robust statistics also recommends one-step estimators for improved efficiency [65, 45, 69, 29]. In this section, we address the problem of improving efficiency by studying one-step modifications of the estimators proposed in the previous section. Note that recent results in high-dimensional inference have led to theoretical derivations based on similar types of one-step estimators to those analyzed here.

We begin by presenting the "one-step" adjustment which may be performed on an initial estimate $\widehat{\beta}$ to obtain a final estimate $\widehat{b}$ with desirable asymptotic normality properties. The statement of our main theorem about asymptotic normality is provided in Section 4.1, where we also discuss conditions on $\widehat{\beta}$ and additional assumptions to be imposed on the covariate and error distributions in order for the results of the theorem to hold. In particular, the theory from Section 3 shows that the ℓ_1-penalized Huber estimator is a suitable choice for $\widehat{\beta}$. In Section 4.2, we expand upon the specific sense in which $\widehat{b}$ is a more efficient estimator than $\widehat{\beta}$ when the score function ψ is chosen appropriately. In Section 4.3, we provide a method for constructing confidence regions for subsets of regression coefficients based on $\widehat{b}$, which is a natural corollary of our result on asymptotic normality.

Consider a differentiable score function $\widetilde{\psi}$, and let $A(\widetilde{\psi}) = \mathbb{E}[\widetilde{\psi}'(\epsilon_i)]$. Based on an initial estimator $\widehat{\beta}$, define the empirical estimate $\widehat{A}(\widetilde{\psi}) = \frac{1}{n}\sum_{i=1}^n \widetilde{\psi}'(y_i - x_i^T\widehat{\beta})$. Following Bickel [7], we then define the one-step estimator

$$\widehat{b} = \widehat{\beta} + \frac{\widehat{\Theta}}{\widehat{A}(\widetilde{\psi})} \cdot \frac{1}{n}\sum_{i=1}^n \widetilde{\psi}(y_i - x_i^T\widehat{\beta})x_i, \tag{4.1}$$

where $\widehat{\Theta}$ is a suitable estimate of $\Theta_x = \Sigma_x^{-1}$, to be described in the sequel.

For the theory in this section, we will change our notation slightly and adopt the language of scale families. Thus, we write $\epsilon_i = \sigma_\xi^* \xi_i$, where the ξ_i's are i.i.d.

random variables from a fixed reference distribution, and σ_ξ^* is an unknown scale parameter (note that σ_ξ^* agrees with σ^*, the standard deviation of ϵ_i defined earlier, up to a constant factor).

As suggested in Bickel [7], we use a score function ψ_σ of the form $\psi_\sigma(t) = \frac{1}{\sigma}\psi\left(\frac{t}{\sigma}\right)$, and plug in an estimate $\widehat{\sigma}$ of the scale parameter σ_ξ^*. Then the one-step estimator (4.1) becomes

$$\begin{aligned}\widehat{b}_\psi &:= \widehat{\beta} + \frac{\widehat{\Theta}}{\widehat{A}(\psi_{\widehat{\sigma}})} \cdot \frac{1}{n}\sum_{i=1}^n \psi_{\widehat{\sigma}}(y_i - x_i^T\widehat{\beta})x_i \\ &= \widehat{\beta} + \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi_{\widehat{\sigma}})} \cdot \frac{1}{n}\sum_{i=1}^n \psi\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right)x_i, \end{aligned} \tag{4.2}$$

where

$$\widehat{A}(\psi_{\widehat{\sigma}}) = \frac{1}{n}\sum_{i=1}^n \psi'_{\widehat{\sigma}}(y_i - x_i^T\widehat{\beta}) = \frac{1}{\widehat{\sigma}^2} \cdot \frac{1}{n}\sum_{i=1}^n \psi'\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right), \tag{4.3}$$

and the scale estimate $\widehat{\sigma}$ is obtained from the consistent regression parameter estimate $\widehat{\beta}$ via $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\widehat{\beta})^2 / \operatorname{Var}(\xi_i)}$. For ease of notation, we will redefine the term $\widehat{A}(\psi)$ to be equal to the expression (4.3), and let $A(\psi) := \mathbb{E}\left[\frac{1}{(\sigma_\xi^*)^2}\psi'(\xi_i)\right]$.

Example 3. *The choice $\psi = \frac{-f'}{f}$, where f is the density of ξ_i (assumed to be smooth), will play a prominent role in our analysis. This corresponds to the derivative of the negative log likelihood function. In the case when $\xi_i \sim N(0,1)$, we then have $\psi(t) = t$ and $\psi'(t) = 1$, in which case formula (4.2) reduces to*

$$\widehat{b}_\psi = \widehat{\beta} + \frac{\widehat{\Theta}X^T(y - X\widehat{\beta})}{n},$$

which is the "debiased Lasso" [75, 43, 13, 41, 85]. However, in that line of work, $\widehat{\beta}$ is always taken to be the output of the usual MLE-based objective, whereas we take $\widehat{\beta}$ to be a more general robust high-dimensional estimator with guaranteed statistical consistency properties even when the covariate or error distributions are non-sub-Gaussian.

We now discuss how to obtain a suitable estimate $\widehat{\Theta}$ of Θ_x. Note that Bickel [7] proposes to use $\widehat{\Theta} = \left(\frac{X^TX}{n}\right)^{-1}$; however, when $p > n$, the matrix $\frac{X^TX}{n}$ is not invertible. We instead choose $\widehat{\Theta}$ to be the graphical Lasso estimator [86, 28], obtained by solving the following convex optimization program:

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0}\left\{\operatorname{tr}\left(\Theta^T\widehat{\Sigma}\right) - \log\det(\Theta) + \lambda\sum_{i\neq j}|\Theta_{ij}|\right\}, \tag{4.4}$$

for a suitable choice of $\widehat{\Sigma}$. In particular, we define $\widehat{\Sigma}$ to be the entrywise MoM estimator of the values Σ_x, i.e., $\widehat{\Sigma}_{jk}$ is the MoM estimator based on $\{x_{ij}x_{ik}\}_{i=1}^n$. The reason for choosing this estimator rather than the simpler sample covariance matrix $\widehat{\Sigma} = \frac{X^T X}{n}$ will become clear in the statement of Theorem 4 and proof of Proposition 1 below, which proceed by deriving a high-probability bound of the form $\|\widehat{\Sigma} - \Sigma_x\|_{\max} \precsim \sqrt{\frac{\log p}{n}}$. Although such a bound would hold for the sample covariance matrix if the x_i's were sub-Gaussian, it behooves us to impose such stringent tail assumptions. See also Remark 5 below for an alternative approach involving a reweighted sample covariance matrix and its connection to classical robust regression literature.

4.1. Asymptotic normality

We now derive the limiting distribution of the one-step estimator. Our arguments involve Taylor expansions of the function ψ, so for simplicity, we assume that ψ is thrice-differentiable. We also assume that ψ is an odd function, and suppose the derivatives of ψ are bounded: $\|\psi'\|_\infty, \|\psi^{(2)}\|_\infty, \|\psi^{(3)}\|_\infty < \infty$, where $\psi^{(s)}$ denotes the s^{th} derivative. Extensions to cases where ψ does not satisfy these smoothness criteria (e.g., corresponding to the Huber loss function) may be derived via more careful arguments, but we omit the details here.

We now present the assumptions we will make on the covariate and error distributions in order to guarantee asymptotic normality of the one-step estimator. Our theorem will be stated assuming that $\widehat{\beta}$ satisfies a suitable error bound; thus, if we wish to use the Huber estimator for $\widehat{\beta}$, we will also need the covariates to satisfy the conditions of Assumption 1 in order to guarantee that the results of Section 3 hold, as well.

We make the following assumptions on the distribution of the covariates:

Assumption 2. *Assume that the marginals x_{ij} are sub-exponential with parameter σ_x, for all $1 \le j \le p$. Also suppose $\min_{1\le j\le p} \mathrm{Var}(x_{ij}) \ge c_1$ for some constant $c_1 > 0$.*

Note that the conditions imposed on the covariates in this section are somewhat stronger than the conditions imposed in Section 3 (cf. Assumption 1), since we no longer include a weight function to temper the effect of heavy tails. Thus, unlike the scenario described in Example 2, Assumption 2 does not permit the covariates to have arbitrarily heavy tails. On the other hand, we actually do not require the full power of sub-exponential tails: As our analysis shows, as long as we have a high-probability bound of the form $\|X\|_{\max} \precsim \mathrm{polylog}(p)$ (cf. Lemma 6), the theorems of this section will still continue to hold under a sample size condition of the form $n \succsim k^2\,\mathrm{polylog}(p)$.

We also impose the following assumptions on the additive errors:

Assumption 3. *Assume that*

(i) $\mathbb{E}[\xi_i^4] < \infty$, *and*
(ii) $\psi(\xi_i)$ *is sub-exponential with parameter* $\sigma_\xi = O(1)$.

Note that the conditions appearing in Assumption 3 are fairly mild, e.g., if ψ is bounded, then condition (ii) holds regardless of the tails of ϵ_i. Furthermore, if the ϵ_i's are Gaussian and ψ corresponds to the MLE of ξ_i, then ψ is the identity function and condition (ii) is again satisfied. However, on top of the finite variance bound imposed in Section 3, we now assume that the fourth moments of the ϵ_i's are finite.

Our main result is the following:

Theorem 4. *Suppose Assumptions 2 and 3 hold and* $n \gtrsim k^2 \operatorname{polylog}(p)$*, and* $\widehat{\beta}$ *satisfies the error bounds*

$$\|\widehat{\beta} - \beta^*\|_1 = \mathcal{O}_P\left(\sqrt{\frac{k \log p}{n}}\right), \quad \text{and} \quad \|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}_P\left(k\sqrt{\frac{\log p}{n}}\right). \tag{4.5}$$

Also suppose

$$\left\|\widehat{\Theta} - \Theta_x\right\|_1 = \mathcal{O}_P\left(k\sqrt{\frac{\log p}{n}}\right). \tag{4.6}$$

Let P_J *denote the projection onto any set of* $m = |J|$ *coordinates of fixed dimension. Then the one-step estimator* (4.2) *satisfies*

$$\sqrt{n} P_J(\widehat{b}_\psi - \beta^*) \xrightarrow{d} N\left(0, \frac{\mathbb{E}[\psi^2(\xi_i)]}{(\sigma_\xi^*)^2 A^2(\psi)} \cdot (\Theta_x)_{JJ}\right),$$

as $n, p \to \infty$.

The proof of Theorem 4 is contained in Appendix D. In particular, the error bounds (4.5) follow directly from the guarantees for the Huber estimator derived in Theorem 3 (under the additional distributional conditions stated in Assumptions 2 and 3).

Altogether, we conclude that the limiting distribution of the high-dimensional estimator, restricted to m coordinates, agrees with the result of Bickel [7] for low-dimensional robust M-estimators.

Remark 3. *Note that the assumption* $n \gtrsim k^2 \operatorname{polylog}(p)$ *is somewhat stronger than the sample size condition* $n \gtrsim k \log p$ *usually required for consistency in statistical estimation, in the sense that* $n = \Omega(k^2)$ *as opposed to* $n = \Omega(k)$ *(cf. Theorem 1). However, a similar gap also appears in the analysis of van de Geer et al. [75] and Javanmard and Montanari [43] in the random design setting. As noted by a reviewer, it would be interesting to see whether this sample size requirement could be improved using more refined arguments, but to the best of our knowledge, existing work on the Lasso [44, 4] relies heavily on a Gaussian distributional assumption on the covariates and the specific form of the least-squares objective.*

Remark 4. *We may also compare this result with Section 3 of van de Geer et al. [75], in which Lasso debiasing results are derived for general convex loss functions. Translating to the linear model with i.i.d. (but not necessarily Gaussian)*

additive errors, the proposed one-step estimator takes the form

$$\widehat{b}_\rho = \widehat{\beta} + \widehat{\Theta}_\rho \cdot \frac{1}{n}\sum_{i=1}^n \rho'(y_i - x_i^T\widehat{\beta})x_i, \tag{4.7}$$

where $\widehat{\beta}$ *is the solution to the* ℓ_1*-penalized program*

$$\widehat{\beta} \in \arg\min_\beta \left\{ \frac{1}{n}\sum_{i=1}^n \rho(y_i - x_i^T\beta) + \lambda\|\beta\|_1 \right\}, \tag{4.8}$$

and ρ *is assumed to be a smooth convex function. Furthermore,* $\widehat{\Theta}_\rho$ *is defined to be a sparse approximate inverse of the matrix* $\frac{1}{n}\sum_{i=1}^n \rho''(y_i - x_i^T\widehat{\beta})x_i x_i^T$.

Although clear similarities exist between the one-step estimator (4.7) *and the expression* (4.2)*, with* ρ' *taking the place of* ψ*, the one-step estimator* (4.7) *is only guaranteed to be asymptotically normal when standardized appropriately. Furthermore, note that the M-estimator* (4.8) *is not designed to be robust to contaminated covariates, and in order to obtain appropriate error bounds, much stronger assumptions must be made on the distribution of the* x_i*'s. Importantly, our proposed one-step estimator involves using one loss (the Huber loss) to define the initial estimate* $\widehat{\beta}$*, and then a separate score function* ψ*, which does not necessarily correspond to a derivative of the Huber loss, to obtain both (a) robustness and (b) efficiency.*

Finally, we provide conditions for the inverse covariance matrix estimator $\widehat{\Theta}$ to satisfy the error bound (4.6). Suppose Σ_x satisfies the α-incoherence condition, defined by

$$\max_{e\in S^c} \left\| \Gamma^*_{eS}(\Gamma^*_{SS})^{-1} \right\|_1 \le 1 - \alpha, \tag{4.9}$$

where $\alpha \in (0,1]$, and we denote $\Gamma^* := \Sigma_x \otimes \Sigma_x$ and $S = \mathrm{supp}(\Theta_x)$. We also denote $\kappa_\Sigma := \|\Sigma_x\|_1$ and $\kappa_\Gamma := \left\|(\Gamma^*_{SS})^{-1}\right\|_1$.

Combining a high-probability deviation bound on $\|\widehat{\Sigma} - \Sigma_x\|_{\max}$ (cf. Lemma 11 in Appendix H) with standard derivations for the graphical Lasso [63] yields the following result:

Proposition 1. *Suppose Assumption 2 holds and* $n \gtrsim \mathrm{polylog}(p)$*. Also suppose* Θ_x *satisfies the* α*-incoherence condition* (4.9) *and the regularization parameter satisfies*

$$\frac{c_0\sigma_x^2}{\alpha}\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{1}{6\kappa_\Gamma k}\left(\frac{\alpha}{8}+1\right)^{-1} \min\left\{ \frac{1}{\kappa_\Sigma}, \frac{1}{\kappa_\Sigma^3\kappa_\Gamma}, \frac{\alpha(\alpha/8+1)^{-1}}{8\kappa_\Sigma^3\kappa_\Gamma} \right\}.$$

With probability at least $1 - \exp(-cn)$*, the graphical Lasso estimator* (4.4) *computed with respect to the entrywise MoM estimator* $\widehat{\Sigma}$ *satisfies* $\mathrm{supp}(\widehat{\Theta}) \subseteq \mathrm{supp}(\Theta_x)$*, and*

$$\|\widehat{\Theta} - \Theta_x\|_{\max} \le 2\left\|(\Gamma^*_{SS})^{-1}\right\|_1 \left(1 + \frac{\alpha}{8}\right)\lambda.$$

In particular, if each row of Θ_x is k-sparse, we also have the bound

$$\left|\left|\left|\widehat{\Theta} - \Theta_x\right|\right|\right|_1 \leq 2\left|\left|\left|(\Gamma^*_{SS})^{-1}\right|\right|\right|_1 \left(1 + \frac{\alpha}{8}\right)\lambda k.$$

The proof of Proposition 1 is contained in Appendix E. We see that the final conclusion of the lemma, with $\lambda \asymp \sqrt{\frac{\log p}{n}}$, furnishes the deviation bound (4.6).

Note that simply applying Theorem 1 in Ravikumar et al. [63] would produce a weaker result than we want, since the concentration result in Lemma 11 would fall into the category of "polynomial-type tails," thus yielding a suboptimal sample size requirement. Instead, we derive a statistical error guarantee suitable for our setting, building upon some of the key lemmas in Ravikumar et al. [63].

Remark 5. *Welsch and Ronchetti [82] studied higher-order expansions of various one-step estimators using different proposals for the Hessian term, and pointed out that the critical characteristic for equivalence of first-order terms is a certain bound on the rate of convergence of the Hessian to its expectation. Our estimator (4.2) is most closely related to the "method of scoring" one-step estimator discussed in Welsch and Ronchetti [82]. However, the direct analog of that estimator would involve inverting the matrix $\widehat{\Sigma}^w := \frac{1}{n}\sum_{i=1}^n w^2(x_i)x_i x_i^T$, instead. The main result in Theorem 4 would still hold, since Assumption 1(i) would be enough to guarantee concentration of $\widehat{\Sigma}^w$ to its expectation Σ_x^w, so that the inverse computed with respect to the graphical Lasso would also converge to $\Theta_x^w := (\Sigma_x^w)^{-1}$ (cf. Proposition 1). On the other hand, since Θ_x^w is generally a biased estimator of Θ_x, we would not have the semiparametric efficiency results derived in Section 4.2.*

4.2. Semiparametric efficiency

To make the notions of increased efficiency more precise, we now analyze the one-step estimator $\widehat{b}_\psi$ from the point of view of semiparametric efficiency. In particular, consider the semiparametric regression model

$$y_i = x_i^T\beta_0 + g_0(v_i) + \epsilon_i,$$

where the distribution of the ϵ_i's is unknown and our goal is to estimate the unknown vector β_0 from i.i.d. observations $\{(y_i, x_i, v_i)\}_{i=1}^n$. Recall the notion of semiparametric efficiency:

Definition 3. *An estimate $\widehat{\beta}$ of β_0 is* semiparametrically efficient *if it is regular (i.e., $\sqrt{n}(\widehat{\beta} - \beta_0)$ is asymptotically normal), and the asymptotic variance is minimal among all regular estimates of β_0.*

Additional background material is included in Appendix A. In particular, Theorem 7 states that a lower bound on the variance of any semiparametrically efficient estimator is given by

$$\overline{V} = \left(\mathbb{E}\left[\left(\frac{f'(\epsilon)}{f(\epsilon)}\right)^2\right] \cdot \mathbb{E}\left[(x - \mathbb{E}[x|v])(x - \mathbb{E}[x|v])^T\right]\right)^{-1},$$

where f denotes the density of ϵ_i.

For a fixed set of indices $J \subseteq \{1, \ldots, p\}$, we partition the linear model as

$$y_i = (x_i)_J^T \beta_J^* + (x_i)_{J^c}^T \beta_{J^c}^* + \epsilon_i$$

and consider it as a subclass of the semiparametric regression model

$$y_i = (x_i)_J^T \beta_J^* + g_0\left((x_i)_{J^c}\right) + \epsilon_i. \tag{4.10}$$

We then have the following result, proved in Appendix F:

Theorem 5. *Suppose we have i.i.d. observations from the linear model* (2.1). *Suppose* $\psi = \frac{-f'}{f}$, *where* f *is the pdf of the distribution of* ξ_i, *and the initial estimate* $\widehat{\beta}$ *satisfies the conditions of Theorem 4. Then the one-step estimator* $(\widehat{b}_\psi)_J$ *is semiparametrically efficient for the model* (4.10).

Theorem 5 shows that just as in classical asymptotic theory for M-estimators, a one-step correction with ψ function equal to the (negative) derivative of the log likelihood will yield an estimator with the same asymptotic properties as the maximum likelihood estimator. However, a benefit of using the one-step estimator $\widehat{b}_\psi$ rather than directly using the maximum likelihood estimator is that the latter may be difficult to compute, especially when the negative log likelihood is nonconvex and/or the scale parameter of the error distribution is unknown. Our theory shows that using the Huber estimator $\widehat{\beta}$ for initialization sidesteps both of these potential issues, since the Huber loss is convex and our procedure via Lepski's method adapts to the scale.

Remark 6. *The notions of efficiency we have just described should be contrasted with the discussion of efficiency contained in Loh [53]. Importantly, our present results do not require any conditions for correct support recovery of the regression estimator, which were rather strong requirements imposed in the theory of the aforementioned paper. Furthermore, by using a one-step estimator, we do not require a second subgradient optimization routine performed on a nonconvex objective function in order to achieve efficiency, since a one-step modification of the global optimum of the convex surrogate is sufficient for our purposes.*

Finally, we note that another notion of semiparametric efficiency was recently studied in Jankova and van de Geer [40], involving a more complicated infinite-dimensional model that is allowed to change with n. It was shown that when Θ_x is a sparse matrix, the same bounds may be established for semiparametric efficiency; however, van de Geer [73] showed that without the sparsity condition, the variance of an efficient estimator may in fact be lower. We suspect that these notions could also be adapted to the setting of robust regression estimators discussed in our paper, but such derivations are beyond the scope of our present work.

4.3. Confidence intervals

Our results from Section 4.1 naturally allow us to derive confidence intervals with the correct asymptotic coverage, which we briefly describe here. Further-

more, the semiparametric efficiency result of Section 4.2 provides a type of "optimality" guarantee for the size of the confidence region. We again consider a fixed subset $J \subseteq \{1, \ldots, p\}$, where $|J| = m$.

For an error probability $\alpha \in (0,1)$, we write $\mathcal{B}_{\alpha,J}$ to denote the subset of $\mathbb{R}^J$ corresponding to the direct product of m intervals of the form

$$\left[-\Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right),\ \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right)\right],$$

where Φ is the cdf of a standard normal random variable. In particular, if $Z \sim N(0, I_m)$ is an m-dimensional Gaussian random vector with i.i.d. standard normal components, we have

$$\mathbb{P}(Z \in \mathcal{B}_{\alpha,J}) = \left(1 - 2\left(1 - \frac{1+(1-\alpha)^{1/m}}{2}\right)\right)^m = 1 - \alpha. \tag{4.11}$$

We have the following main result, proved in Appendix G. We impose one additional condition involving the boundedness of $(\psi^2)''$ in order to facilitate our derivations.

Theorem 6. *Let $|J| = m$ be a fixed set of constant cardinality. In addition to the assumptions of Theorem 4, suppose $\|(\psi^2)''\|_\infty < \infty$. An asymptotically valid $(1-\alpha)$-confidence region for the projection β^*_J of the regression vector onto J is given by*

$$P_J \widehat{b}_\psi + \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n \psi^2\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right)}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \left(\widehat{\Theta}_{JJ}\right)^{1/2} \mathcal{B}_{\alpha,J}. \tag{4.12}$$

Note that the region (4.12) is a (pointwise) linear transformation of $\mathcal{B}_{\alpha,J}$.

In the case $m = 1$, the confidence region for a fixed coordinate j reduces to the interval

$$(\widehat{b}_\psi)_j \pm \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n \psi^2\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right)}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \sqrt{\widehat{\Theta}_{jj}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Note that as in Javanmard and Montanari [43], the set $\mathcal{B}_{\alpha,J}$ could be replaced with any other set of measure $1-\alpha$ under an m-dimensional standard normal distribution.

Note that Theorem 6 is a result that holds for *any* choice of score functions ψ, not necessarily corresponding to the score function of the true pdf. Importantly, we can construct valid confidence intervals without needing to know the true distribution of the ϵ_i's. However, in order to construct *optimal* intervals, we would need to use the correct ψ function corresponding to the distribution.

Remark 7. *As mentioned in Remark 4, our recipe for constructing confidence intervals resembles the proposal of van de Geer et al. [75]. However, the key*

difference is that the vanilla Lasso estimator would in general not achieve the correct rates of consistency in order for the confidence intervals to be asymptotically valid for the prescribed sample size scaling. Similarly, Javanmard and Montanari [43] include a section in their paper discussing how to construct confidence intervals in the case of non-Gaussian noise; however, again, they assume that the noise and covariance distributions are sufficiently well-behaved to guarantee fast convergence of the initial Lasso estimator. A way to correct this would be to use the Huber estimator as an initial estimator rather than the Lasso; see the simulations at the end of Section 5.2 for additional discussion and an empirical comparison.

Finally, it is worth discussing the relationship between our proposed method and the robust inference procedures studied in classical robust statistics. These include robust Wald-type and likelihood ratio tests [64, 32], which are more generally applicable to hypothesis testing scenarios involving linear combinations of predictors. Our method resembles Wald-type tests in the sense that they are constructed with respect to a robust M-estimator, and also include robust estimates of the (inverse) covariance—however, our results are primarily designed for hypothesis testing of single coordinates. It is an interesting open question to see if analogs of the robust Wald-type or τ-tests [64] could be derived in the high-dimensional setting. It is plausible that such tests exist using an initial M-estimator such as the regression estimator introduced in this paper (cf. van de Geer and Stucky [74] and Sur et al. [72] for some theory in the non-robust setting).

Remark 8. *As pointed out by a reviewer, an alternative approach proposed by Belloni et al. [6] does not require incoherence assumptions* (4.9) *on the inverse covariance matrix, which are required to ensure the validity of our method. However, since the method of Belloni et al. [6] is a coordinatewise approach, it leads to confidence regions which are direct products of confidence intervals for individual components. This misses out on the optimality property of our confidence regions which is derived from the semiparametric efficiency of our regression estimator; note that in general, the confidence regions constructed in equation* (4.12) *may correspond to affine transformations of cuboids which are not direct products of intervals.*

5. Simulations

We now report the result of experiments that we performed to validate our theoretical predictions.

5.1. Summary of procedure

We first briefly summarize the steps of the robust regression procedure:

1. Compute rough lower and upper bounds on the scale, using the median of means estimator with tolerance δ and $K = \left\lfloor 8 \log\left(\frac{e^{1/8}}{\delta}\right) \wedge \frac{n}{2} \right\rfloor$ groups: $\sigma^2_{\max} = 2\sigma^2_{MoM}$, and $\sigma_{\min} = \frac{\sigma_{\max}}{2^M}$.
2. Compute the ℓ_1-penalized Huber M-estimator $\widehat{\beta}_\tau$ for all τ in a grid of values within $[c_\tau \sigma_{\min}, c_\tau \sigma_{\max}]$, according to the program (3.1).
3. Use Lepski's method to adaptively choose $\widehat{\beta} = \widehat{\beta}_{(j^*)}$, according to the rule (3.6).
4. Use one-step estimation to improve efficiency, with $\widehat{b}_\psi$ defined according to equation (4.2), with $\widehat{\beta}$ from Lepski's method and $\widehat{\Theta}$ from the graphical Lasso.

Composite gradient descent: In order to obtain the estimators $\widehat{\beta}_\tau$ in the second step above, we employ the composite gradient descent algorithm, which has fast rates of convergence for convex functions [60]. Specifically, the updates are

$$\begin{aligned}\widehat{\beta}^{t+1} &\in \arg\min_\beta \left\{ \mathcal{L}_n(\beta^t) + \langle \nabla \mathcal{L}_n(\beta^t), \beta - \beta^t \rangle + \frac{\eta}{2}\|\beta - \beta^t\|_2^2 + \lambda\tau\|\beta\|_1 \right\} \\ &= \arg\min_\beta \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{1}{\eta}\nabla \mathcal{L}_n(\beta^t) \right) \right\|_2^2 + \frac{\lambda\tau}{\eta}\|\beta\|_1 \right\} \\ &= S_{\lambda\tau/\eta}\left(\beta^t - \frac{1}{\eta}\nabla \mathcal{L}_n(\beta^t) \right),\end{aligned}$$

where $S_{\lambda\tau/\eta}(\beta)$ is the soft-thresholding operator defined componentwise according to

$$S^j_{\lambda\tau/\eta}(\beta) = \operatorname{sign}(\beta_j)\left(|\beta_j| - \frac{\lambda\tau}{\eta} \right)_+.$$

Note also that

$$\nabla \mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^n \ell'_\tau\left((x_i^T\beta - y_i)w(x_i) \right) w(x_i) x_i.$$

5.2. Synthetic data

We first ran experiments involving synthetic data to check the validity of our theory. The simulation results confirm that our estimator is (a) consistent and (b) efficient. For (a), we provide simulation results under two different scenarios:

(i) Additive errors are drawn from a heavy-tailed distribution, but covariates have a sub-Gaussian distribution.
(ii) Both x_i's and ϵ_i's are drawn from heavy-tailed distributions.

In case (i), we generated the x_i's from a standard normal distribution. The ϵ_i's were generated from a t-distribution with five degrees of freedom, to make the fourth moment finite (recall that moments of order five and above do not exist).

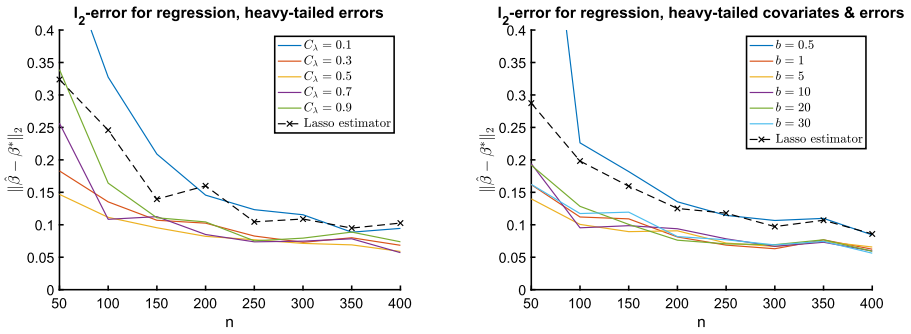
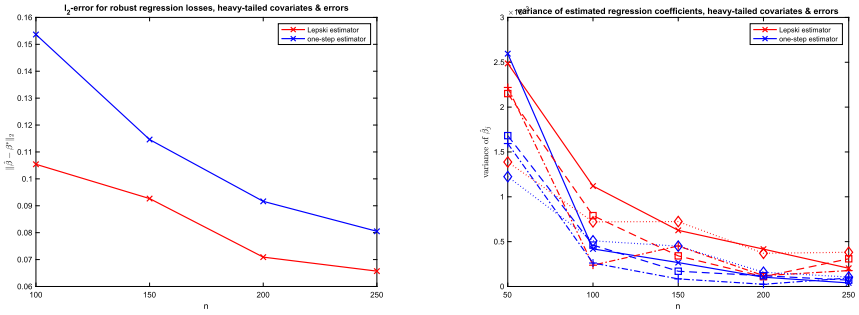


FIG 1. *Plots comparing ℓ_2-error of our Huber regression estimator with tuning parameter chosen adaptively using Lepski's method, for $p = 100$, $k = 4$, and increasing values of n. In plot (a), covariates were generated from a multivariate normal distribution and additive errors were generated from a t-distribution. In plot (b), covariates were generated from a Laplace distribution and additive errors were generated from a t-distribution. Individual points represent an average over 10 independent trials. The performance of the Lasso (dotted line) is also shown in the figure.*

Independently, with probability 0.1, we then multiplied each ϵ_i by 10 to simulate further heavy-tailed contamination. Finally, we scaled the additive errors by 0.1. In case (ii), we generated the ϵ_i's in the same manner as in (i), but generated the coordinates of the x_i's independently from a Laplace distribution with mean 0 and scale parameter 1. Note that in this case, the marginals of the x_i's are sub-exponential.

Further implementation details are as follows: We set the error tolerance $\delta = 0.05$ for the MoM estimator, and took $\sigma_{\min} = \frac{\sigma_{\max}}{2^{2n^{1/3}}}$ and $(c_\tau, C) = (1, 20)$ for the Lepski gridding. We defined the weight function according to the expression (3.2), using $b = 1$ for the simulations in (i) and a range of values for the simulations in (ii). We defined the regularization parameter to be $\lambda = 0.5\sqrt{\frac{\log p}{n}}$ for the simulations in (ii) and $\lambda = C_\lambda\sqrt{\frac{\log p}{n}}$ for a range of values for C_λ for the simulations in (i). We chose the problem dimensions to be $p = 100$ and $k = 4$, and defined β^* to have 1's in the first four components and 0's everywhere else.

Figure 1(a) shows the ℓ_2-error of the adaptively tuned Huber estimator in setting (i), using a range of λ values. Figure 1(b) shows the ℓ_2-error of the adaptively tuned Huber estimator in setting (ii), using a range of λ values. For comparison, we also include error curves for the vanilla Lasso, where the tuning parameter was chosen using 10-fold cross-validation. As expected, the error of both the Huber and Lasso estimators appears to decrease to zero with n. However, the Huber estimator tends to perform better than the Lasso, and the gap becomes more noticeable when both the covariates and errors are heavy-tailed. The precise values of C_λ and b do not seem to affect the performance of the Huber estimator too heavily.



(a) (b)

FIG 2. *Plots comparing ℓ_2-error and variance of estimators obtained via Lepski's method (red) and Lepski's method followed by a one-step correction (blue), when $p = 100$ and $k = 4$. Covariates were generated from a Laplace distribution and additive errors were generated from a t-distribution. Panel (a) shows the error $\|\widehat{\beta} - \beta^*\|_2$, averaged over 10 trials. Panel (b) shows the empirical variance of $\widehat{\beta}_j$, for each of the four nonzero regression coefficients, computed with respect to 10 trials. The coefficients are distinguished in the figure using different line markings.*

In order to explore (b) the relative efficiency of the Huber estimator in comparison to its one-step correction, we borrowed some implementation details from the settings described in (a). We generated the coordinates of the x_i's from a Laplace distribution with mean 0 and scale parameter 1. We generated the ϵ_i's from a t-distribution with five degrees of freedom, scaled by 0.1. We set the error tolerance $\delta = 0.05$ for the MoM estimator, and took $\sigma_{\min} = \frac{\sigma_{\max}}{2^2 n^{1/3}}$ and $(c_\tau, C) = (1, 20)$ for the Lepski gridding. We defined the weight function according to the expression (3.2) with $b = 1$, and we defined the regularization parameter to be $\lambda = 0.5\sqrt{\frac{\log p}{n}}$. We chose the problem dimensions to be $p = 100$ and $k = 4$, and defined β^* to have 1's in the first four components and 0's everywhere else.

For the one-step estimator, we use the formulas in equation (4.2) to define $\widehat{A}$ and $\widehat{\sigma}$. Recall that the pdf of a t-distribution with ν degrees of freedom is equal to

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Then we may compute

$$\psi(t) = \frac{-f'(t)}{f(t)} = \frac{(\nu+1)t}{\nu + t^2}, \quad \text{and} \quad \psi'(t) = \frac{-(\nu+1)t^2 + \nu(\nu+1)}{(\nu + t^2)^2}.$$

Figure 2 shows the results of the simulations. The plot in (a) shows the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$ of the initial Huber estimator in comparison to the ℓ_2-error $\|\widehat{b}_\psi - \beta^*\|_2$ of the one-step estimator. Both curves show vanishing error as n increases—note that our theory does not necessarily imply that the ℓ_2-error of

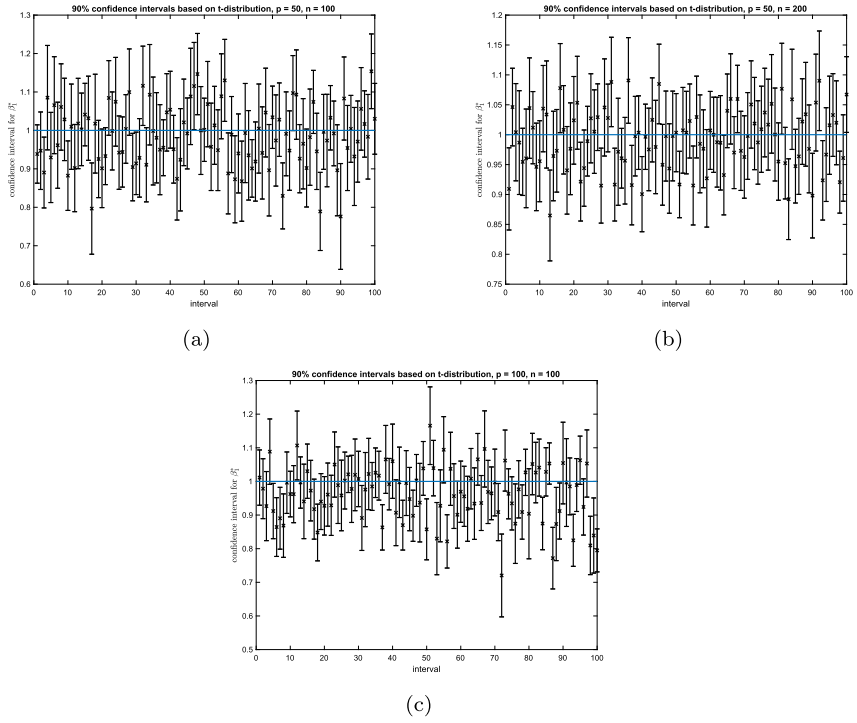


FIG 3. *Plots showing results of confidence interval simulations based on 100 trials. Data were generated with covariates drawn from a Laplace distribution and errors drawn from a t-distribution, and confidence intervals were constructed at the 90% level. Panels (a) and (b) show confidence interval coverage for* $p = 50$ *with* $n = 100$ *and* $n = 200$*. The empirical coverage was 82% and 78%, respectively. Panel (c) shows confidence interval coverage for* $p = 100$ *and* $n = 100$*. The empirical coverage was 74%.*

the one-step estimator will always be smaller than the ℓ_2-error of the initial estimator, and only guarantees that the error will decrease at the same rate, up to constant factors. However, in the plot in (b), we can clearly see that the empirical variance of the estimates of all four of the nonzero coefficients of β^* indeed appears to decrease after the one-step correction, corroborating our theoretical conclusions.

Finally, we provide a set of simulation results illustrating the validity of our method for constructing confidence intervals described in Section 4.3. Figure 3 shows the result of 100 confidence intervals constructed using our procedure when the coordinates of the x_i's are drawn i.i.d. from a Laplace distribution with mean 0 and scale parameter 1, and the ϵ_i's are generated from a t-distribution with five degrees of freedom, scaled by 0.5.

For comparison, we also constructed confidence intervals according to the method suggested by van de Geer et al. [75] and Javanmard and Montanari [43], which essentially corresponds to our one-step procedure with score function

$\psi \equiv 1$ (corresponding to the MLE for Gaussian errors). Furthermore, we set the initial estimator $\widehat{\beta}$ to be equal to the Huber estimator rather than the Lasso, since the Lasso estimator has slower rates of convergence under heavy-tailed covariates and/or error distributions; we take the estimate of variance used in those formulas to be the empirical variance of the residuals computed with respect to $\widehat{\beta}$. We observe that the empirical coverage of the confidence intervals constructed according to our procedure is similar to that of the method using a Gaussian one-step correction: In comparison to the coverage percentages reported in Figure 3, the coverage levels for confidence intervals constructed using a Gaussian correction were (a) 78%, (b) 81%, and (c) 73%. On the other hand, the confidence intervals were on average shorter when using the one-step estimator with score function ψ corresponding to the t-distribution: The average lengths of confidence intervals for t-distribution (normal) corrections were (a) 0.2102 (0.2156), (b) 0.1410 (0.1464), and (c) 0.1823 (0.1921). This supports our theoretical results of semiparametric efficiency.

To check for consistency as $n \to \infty$, we ran the same confidence interval experiment with $p = 10$ and $n = 500$. The results, averaged over 100 trials, are tabulated in Figure 4. Here, t denotes confidence intervals computed with respect to the t-distribution score function, and z denotes confidence intervals computed with respect to the Gaussian one-step correction. We see that the empirical coverage percentages for both methods are roughly equal to 90%, whereas the average lengths of intervals computed using the t-distribution score function are generally slightly smaller. However, the difference between the average lengths of confidence intervals for the two methods vanishes as the number of degrees of freedom ν increases, since the t-distribution tends toward the standard normal.

	$\nu = 5$	$\nu = 6$	$\nu = 7$	$\nu = 8$
empirical coverage (t)	88%	87%	87%	87%
empirical coverage (z)	90%	88%	81%	87%
average length (t)	0.083	0.082	0.082	0.081
average length (z)	0.088	0.084	0.082	0.080

(a)

	$\nu = 5$	$\nu = 6$	$\nu = 7$	$\nu = 8$
empirical coverage (t)	90%	82%	90%	88%
empirical coverage (z)	87%	81%	91%	85%
average length (t)	0.087	0.084	0.084	0.083
average length (z)	0.092	0.086	0.084	0.082

(b)

FIG 4. *Tables of results for confidence interval simulations at the 90% level, based on 100 trials. (a) Data were generated with covariates drawn from a multivariate normal distribution and errors were drawn from a t-distribution with ν degrees of freedom. (b) Data were generated with covariates drawn i.i.d. from a Laplace distribution and errors were drawn from a t-distribution with ν degrees of freedom.*

5.3. Real data experiment

Turning to a real dataset, we analyzed a dataset collected from X-ray microanalysis of archaeological glass vessels [42], which has been analyzed in several other papers on high-dimensional robust linear regression with leverage points [56, 70]. The dataset consists of $n = 180$ observations and $p = 486$ frequencies, which we used as predictors for the contents of compound 13, which is PbO. As discussed in [56], the dataset contains clear outliers.

Following the method of Smucler and Yohai [70] for tuning parameter selection, we chose the parameter λ in our algorithm via 5-fold cross-validation using a τ-scale of the residuals [84, 66]. (Note that our theorems are stated with λ equal to $\sqrt{\frac{\log p}{n}}$ times universal constants, but in practice, choosing λ in a data-driven manner leads to better predictive performance.) Based on this procedure, Lepski's method yielded a sparse vector with six nonzero components. This fit corresponds to the value 0.134 of the τ-scale, which is comparable to the values reported in Smucler and Yohai [70] using alternative methods: MM-Lasso (τ-scale of 0.086, seven selected variables), adaptive MM-Lasso (τ-scale of 0.083, four selected variables), sparse-LTS (τ-scale of 0.329, three selected variables), Lasso (τ-scale of 0.131, seventy selected variables), and adaptive Lasso (τ-scale of 0.138, forty-nine selected variables); note that as is the case for the robust methods advocated in that paper, our method likewise chooses sparser models than the Lasso and adaptive Lasso, making the model easier to interpret, while maintaining good predictive performance.

We also attempted to construct confidence intervals for the selected frequencies. The simulations were inconclusive, due to the fact that various implementations of the graphical Lasso algorithm on the 486×486 matrix of covariates failed to converge. We suspect that this is because the assumption that the population-level inverse covariance matrix is sparse is violated, or the covariate distribution is heavy-tailed and/or possesses extreme outliers, so that the rate of convergence of the estimated covariance matrix $\widehat{\Sigma}$ to Σ_x is too slow. This experiment reveals that the additional assumptions required to construct confidence intervals may be somewhat more stringent than the assumptions needed for consistency in terms of estimation or prediction error.

6. Discussion

Throughout this paper, we have assumed that the variance of ϵ_i is finite. We now describe a small adaptation that applies to the consistency results of Section 3 when the second moment does not exist. Indeed, one can still define σ^* to be a scale parameter of the distribution of ϵ_i (e.g., in the case of a Cauchy distribution). However, the place where we have required existence of second moments in our analysis is in the computation of the rough scale parameter bounds $\sigma_{\min}$ and $\sigma_{\max}$.

Instead of the MoM estimator, we may use the median absolute deviation (MAD) as the scale parameter when the second moments are not finite. Recall

that the population-level MAD is given by

$$\mathrm{MAD}(X) = \mathrm{med}\left(|X - \mathrm{med}(X)|\right),$$

where med denotes the median operator. By Lemma 19 in Appendix H, we know that under the assumption that the distribution of ϵ_i is symmetric and unimodal, we have

$$\mathrm{MAD}(\epsilon_i) \leq \mathrm{MAD}(x_i^T\beta^* + \epsilon_i) = \mathrm{MAD}(y_i),$$

so that the MAD estimate based on the y_i's can indeed be used as an upper bound on the scale of the ϵ_i's, analogous to the case of the variance. Furthermore, concentration inequalities for the empirical version of the MAD estimator can be found, e.g., in [68]. It is an open question whether the analysis of the one-step estimation results in Section 4 can also be adapted to remove the dependence on finiteness of the variance (and/or higher moments).

We also mention an interesting open question of practical relevance: What type of one-step estimator could we use for obtaining a more efficient estimator and/or confidence intervals when the shape of the error distribution is unknown? Some general guidelines for choosing the ψ function in the one-step estimator, or a more principled procedure for flagging outliers and then fitting confidence intervals based on a fitted distribution, would be quite useful in practice.

Finally, an interesting direction to pursue would be whether an approach based on Lepski's method could also be used to adaptively choose the correct parameter for the Huber loss in the case of an ϵ-contaminated model (either in location estimation or linear regression). A related question is how to adaptively choose a trimming parameter for the robust location estimator based on trimmed means. These are both questions of theoretical interest that have largely remained open in the classical robust statistics literature—since they depend on minimizing variance quantities, rather than deriving high-probability error bounds, the machinery developed in this paper does not carry over directly. However, it is plausible that an appropriate modification of the Lepski-based approach may result in theoretically valid conclusions for obtaining a near-optimal estimator from the point of view of variance.

Appendix A: Semiparametric efficiency

In this appendix, we review several concepts in semiparametric estimation. For a more detailed overview, we refer the reader to the textbooks by Bickel et al. [8] or Hansen [33].

Following the treatment of Newey [61], we first define the semiparametric regression model [21]:

Definition 4. *The* semiparametric regression model *characterized by a parameter vector $\beta_0 \in \mathbb{R}^q$ and function g_0 is given by*

$$y_i = x_i^T\beta_0 + g_0(v_i) + \epsilon_i, \qquad \textit{for } 1 \leq i \leq n, \tag{A.1}$$

where the x_i's and v_i's are vectors of exogenous observations, y_i is a scalar response, and ϵ_i is independent additive error.

Semiparametric efficiency is usually established by obtaining lower bounds on the asymptotic variance of an efficient estimator by considering Cramer-Rao bounds for different parametric "submodels," which are models that include the semiparametric model under consideration and are equal to the semiparametric model for a certain value of the parameter. In particular, the Cramer-Rao bound for any parametric subclass must provide a lower bound for the semiparametric estimation problem, as well, and we have the variance lower bound

$$\overline{V} = \sup_{\theta} V_\theta,$$

where V_θ is the Cramer-Rao bound corresponding to a parametric submodel indexed by θ. If one can find a parametric submodel with a Cramer-Rao bound that matches the asymptotic variance of a particular semiparametric estimator, that estimator is guaranteed to be efficient. Note that for multidimensional problems, the supremum is taken with respect to the partial order of positive semidefinite matrices (and the supremum is guaranteed to exist under appropriate regularity conditions, which apply in the setting considered here).

Newey [61] presents an approach to compute the variance bound $\overline{V}$ directly by considering the projection of the score function of the semiparametric model onto the tangent set corresponding to the scores of all parametric submodels, where the score of the semiparametric model is the partial derivative of the negative log likelihood with respect to the parameter vector. Formally, consider a parametric submodel parametrized by $\theta = (\beta, \eta)$, where both β and η are vectors, and β corresponds to the q-dimensional parametric part of the original semiparametric model. The overall score function may be partitioned as $S_\theta = (S_\beta, S_\eta)$. By block matrix inversion, we may verify that the Cramer-Rao bound for estimation of β in the parametric submodel is then given by

$$V_\theta = \left(\mathbb{E}[(S_\beta - \widetilde{B}S_\eta)(S_\beta - \widetilde{B}S_\eta)^T]\right)^{-1},$$

where $\widetilde{B} := \mathbb{E}[S_\beta S_\eta^T]\left(E[S_\eta S_\eta^T]\right)^{-1}$. In particular, $\widetilde{B}S_\eta$ is the best linear predictor of S_β as a function of S_η.

We now define the tangent set to be the mean square closure of all q-dimensional linear combinations of scores of parametric submodels:

$$\mathcal{T} = \left\{\mathcal{S} \in \mathbb{R}^q : \mathbb{E}[\|\mathcal{S}\|_2^2] < \infty,\ \exists A_j S_{\theta_j} \text{ s.t. } \mathbb{E}[\|\mathcal{S} - A_j S_{\theta_j}\|_2^2]\right\},$$

where the A_j's are matrices with q rows and the S_{θ_j}'s are the score vectors of various parametric submodels.

We have the following result, which holds generally for semiparametric estimation (not just in the case of the semiparametric regression model):

Lemma 1. *[Newey [61, Theorem 3.2]] Suppose $\mathcal{T}$ is a linear space, and let $S_\beta^{\mathcal{T}}$ denote the projection of S_β on $\mathcal{T}$. Then*

$$\overline{V} = \left(\mathbb{E}\left[(S_\beta - S_\beta^{\mathcal{T}})(S_\beta - S_\beta^{\mathcal{T}})^T\right]\right)^{-1},$$

provided the matrix is nonsingular.

For the model (A.1), we denote a parametrization of $g_0(v)$ as $g(v,\eta)$, where η is a parameter such that $g(v,\eta_0) = g_0(v)$. Then the log likelihood may be written as

$$p_{\beta,\eta}(y|x,v) = \log f\left(y - x^T\beta + g(v,\eta)\right),$$

where f is the density of ϵ_i. Taking partial derivatives and evaluating at the true parameter values (β_0,η_0), we obtain the score functions

$$S_\beta = \frac{f'(\epsilon)}{f(\epsilon)} \cdot x, \qquad S_\eta = \frac{f'(\epsilon)}{f(\epsilon)} \cdot g_\eta,$$

where $\epsilon = y - x^T\beta_0 - g_0(v)$ and $g_\eta := \frac{\partial g(v,\eta)}{\partial \eta}\Big|_{\eta=\eta_0}$. It is not hard to verify that the tangent set is equal to

$$\mathcal{T} = \left\{ \frac{f'(\epsilon)}{f(\epsilon)} \cdot D(v) : \ \mathbb{E}\left[\left(\frac{f'(\epsilon)}{f(\epsilon)}\right)^2 \|D(v)\|_2^2\right] < \infty \right\},$$

using the observation that the parametric submodel with $g(v,\eta) = g_0(v) + \eta^T D(v)$ yields the score $S_\eta = \frac{f'(\epsilon)}{f(\epsilon)} \cdot D(v)$. Furthermore, $\mathcal{T}$ is clearly a linear space.

In order to compute $S_\beta^{\mathcal{T}}$, we use the following result:

Lemma 2. *[Newey [61, Lemma 3.4]] If UW has finite second moment and V and W are functions of some random variable T, such that $\mathbb{E}[UU^T \mid T]$ is constant and positive definite, then the projection of UW on the space*

$$\mathcal{T}_V := \left\{UD(V) : \ \mathbb{E}[\|UD(V)\|_2^2] < \infty\right\}$$

is equal to $U\mathbb{E}[W \mid V]$.

Applying Lemma 2 with $W = x$, $V = v$, and $U = \frac{f'(\epsilon)}{f(\epsilon)}$, we conclude that

$$S_\beta^{\mathcal{T}} = \frac{f'(\epsilon)}{f(\epsilon)} \cdot \mathbb{E}[x|v].$$

Combining this with Lemma 1, we arrive at the following result:

Theorem 7. *Suppose x has finite second moments and*

$$0 < \mathbb{E}\left[\left(\frac{f'(\epsilon)}{f(\epsilon)}\right)^2\right] < \infty.$$

Then

$$\overline{V} = \left(\mathbb{E}\left[\left(\frac{f'(\epsilon)}{f(\epsilon)}\right)^2 \cdot (x - \mathbb{E}[x|v])(x - \mathbb{E}[x|v])^T\right]\right)^{-1},$$

provided the matrix is nonsingular.

Appendix B: Proof of Theorem 1

We begin by analyzing the estimator

$$\widetilde{\beta}_\tau \in \arg \min_{\|\beta-\beta^*\|_2 \leq r} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_\tau \left((x_i^T \beta - y_i) w(x_i) \right) + \lambda\tau \|\beta\|_1 \right\},$$

where we have introduced a side constraint defined in terms of a parameter r to be specified later. We will show that such optima $\widetilde{\beta}_\tau$ lie in the interior of the constraint set, hence agree with the global optima $\widehat{\beta}_\tau$ of the unconstrained problem.

B.1. Main argument

Let

$$\mathcal{L}_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ell_\tau \left((x_i^T \beta - y_i) w(x_i) \right).$$

We first derive a bound on $\|\widetilde{\beta}_\tau - \beta^*\|_2$, assuming the following conditions:

- (Regularization parameter)

$$\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq \frac{\lambda\tau}{2} \tag{B.1}$$

- (RSC condition)

$$\mathcal{L}_n(\beta) - \mathcal{L}_n(\beta^*) - \langle \nabla \mathcal{L}_n(\beta^*), \beta - \beta^* \rangle \geq \alpha \|\beta - \beta^*\|_2^2, \quad \forall \beta \text{ s.t. } \|\Delta\|_2 \leq r \text{ and } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1, \tag{B.2}$$

  where we have denoted $\Delta := \beta - \beta^*$.

In Appendices B.2 and B.3, we will show that the conditions (B.1) and (B.2) hold w.h.p., when $\tau \geq c_\tau \sigma^*$ and $\alpha = \frac{1}{4}\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i) x_i x_i^T\right]\right)$.

We have the basic inequality

$$\mathcal{L}_n(\widetilde{\beta}_\tau) + \lambda\tau\|\widetilde{\beta}_\tau\|_1 \leq \mathcal{L}_n(\beta^*) + \lambda\tau\|\beta^*\|_1. \tag{B.3}$$

Hence,

$$\langle \nabla \mathcal{L}_n(\beta^*), \widetilde{\beta}_\tau - \beta^* \rangle \leq \mathcal{L}_n(\widetilde{\beta}_\tau) - \mathcal{L}_n(\beta^*) \leq \lambda\tau \left(\|\beta^*\|_1 - \|\widetilde{\beta}_\tau\|_1 \right), \tag{B.4}$$

where the first inequality is due to the convexity of $\mathcal{L}_n$. Therefore, we have

$$0 \leq \lambda\tau \left(\|\beta^*\|_1 - \|\widetilde{\beta}_\tau\|_1 \right) + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \|\widetilde{\beta}_\tau - \beta^*\|_1.$$

Denoting $\widehat{\nu} = \widetilde{\beta}_\tau - \beta^*$ and using the bound (B.1), we then have

$$0 \leq \lambda\tau \left(\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1 + \frac{1}{2}\|\widehat{\nu}\|_1 \right),$$

since

$$\|\beta^*\|_1 - \|\widetilde{\beta}_\tau\|_1 = \|\beta^*_S\|_1 - \|\widetilde{\beta}_{\tau,S}\|_1 - \|\widetilde{\beta}_{\tau,S^c}\|_1 \leq \|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1. \tag{B.5}$$

This implies that

$$\|\widehat{\nu}_{S^c}\|_1 \leq 3\|\widehat{\nu}_S\|_1, \tag{B.6}$$

which is the cone condition.

Therefore, the RSC condition (B.2) together with the basic inequality (B.3) implies that

$$\langle \nabla \mathcal{L}_n(\beta^*), \widehat{\nu} \rangle + \alpha \|\widehat{\nu}\|_2^2 \leq \mathcal{L}_n(\widehat{\beta}_\tau) - \mathcal{L}_n(\beta^*) \leq \lambda\tau \left(\|\beta^*\|_1 - \|\widehat{\beta}_\tau\|_1 \right),$$

so combining with inequalities (B.5) and (B.1), we have

$$\begin{aligned}
\alpha\|\widehat{\nu}\|_2^2 &\leq \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \|\widehat{\nu}\|_1 + \lambda\tau \left(\|\beta^*\|_1 - \|\widehat{\beta}_\tau\|_1 \right) \\
&\leq \lambda\tau \left(\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1 + \frac{1}{2}\|\widehat{\nu}\|_1 \right) \\
&\leq \frac{3\lambda\tau}{2}\|\widehat{\nu}_S\|_1 \\
&\leq \frac{3\lambda\tau\sqrt{k}}{2}\|\widehat{\nu}\|_2,
\end{aligned}$$

implying that

$$\|\widehat{\nu}\|_2 \leq \frac{3\lambda\tau\sqrt{k}}{2\alpha}. \tag{B.7}$$

Rewriting the bound (B.7), we conclude that

$$\|\widetilde{\beta}_\tau - \beta^*\|_2 \leq C\tau\sqrt{\frac{k \log p}{n}},$$

with probability at least $1 - c\exp(-c'n)$. Further note that for $n \gtrsim k \log p$, we are guaranteed that

$$C\tau\sqrt{\frac{k \log p}{n}} < r.$$

It follows that $\widetilde{\beta}_\tau$ lies in the interior of the region $\{\beta : \|\beta - \beta^*\|_2 \leq r\}$, so $\widetilde{\beta}_\tau$ must also be a global optimum of the regularized Huber estimator (3.1) that does not include the side constraint. Furthermore, any optima of the unconstrained problem must also lie in the interior of the constraint set.

Finally, note that inequality (B.6) implies

$$\|\widetilde{\beta}_\tau - \beta^*\|_1 = \|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1 \leq 4\|\widehat{\nu}_S\|_1 \leq 4\sqrt{k}\|\widehat{\nu}_S\|_2 \leq 4\sqrt{k}\|\widetilde{\beta}_\tau - \beta^*\|_2,$$

giving the desired ℓ_1-bound. This concludes the proof of the theorem.

B.2. Bound on regularization parameter

We now verify the bound (B.1). Note that

$$\nabla \mathcal{L}_n(\beta^*) = \frac{1}{n}\sum_{i=1}^n \ell_\tau'\left(\epsilon_i w(x_i)\right) w(x_i) x_i.$$

We first condition on the values of the x_i's. For each $1 \le j \le p$, we see that

$$e_j^T \nabla \mathcal{L}_n(\beta^*) = \sum_{i=1}^n \ell_\tau'\left(\epsilon_i w(x_i)\right) \frac{w(x_i) e_j^T x_i}{n}$$

is a sum of independent, zero-mean random variables, where the i^{th} term is bounded by $\frac{\tau w(x_i)|e_j^T x_i|}{n}$. Hence, by Hoeffding's inequality and a union bound, we have

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \ge \frac{\tau t}{n} \cdot \max_{1 \le j \le p} \sqrt{\sum_{i=1}^n w^2(x_i)(e_j^T x_i)^2} \;\middle|\; \{x_i\}_{i=1}^n \right) \le 2p \exp(-2t^2), \tag{B.8}$$

for any $t > 0$. We will take $t \asymp \sqrt{\log p}$.

Furthermore, the random vectors $w(x_i)x_i$ are sub-Gaussian with parameter b' by assumption, so a union bound together with standard concentration inequalities shows that

$$\mathbb{P}\left(\max_{1 \le j \le p} \left| \frac{1}{n}\sum_{i=1}^n w^2(x_i)(e_j^T x_i)^2 - \mathbb{E}\left[w^2(x_i)(e_j^T x_i)^2 \right] \right| \ge (b')^2 s \right) \le 2p \exp\left(\frac{-ns^2}{2} \right), \tag{B.9}$$

for any $s > 0$. In addition,

$$\max_{1 \le j \le p} \mathbb{E}\left[w^2(x_i)(e_j^T x_i)^2 \right] \le \max_{1 \le j \le p} \mathbb{E}\left[(e_j^T x_i)^2 \right] \le \lambda_{\max}(\Sigma_x).$$

Taking $s \asymp \sqrt{\frac{\log p}{n}}$ in the concentration inequality (B.9), we then conclude that

$$\mathbb{P}\left(\max_{1 \le j \le p} \frac{1}{n}\sum_{i=1}^n w^2(x_i)(e_j^T x_i)^2 \le 2\lambda_{\max}(\Sigma_x) \right) \ge 1 - \exp(-c \log p), \tag{B.10}$$

when $n \gtrsim \log p$. Now let E denote the high-probability event appearing on the left-hand side of inequality (B.10), and let

$$F := \left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \le c\tau \lambda_{\max}^{1/2}(\Sigma_x) \sqrt{\frac{\log p}{n}} \right\}.$$

By a conditioning argument, we have

$$\begin{aligned}\mathbb{P}(F^c) &= \int_E \mathbb{P}(F^c \mid \{x_i\}_{i=1}^n) d\mathbb{P}(\{x_i\}_{i=1}^n) + \int_{E^c} \mathbb{P}(F^c \mid \{x_i\}_{i=1}^n) d\mathbb{P}(\{x_i\}_{i=1}^n) \\ &\leq 2\exp(-c\log p),\end{aligned}$$

where the first term is bounded via inequality (B.8) and the second term is bounded by $\mathbb{P}(E^c)$, which is in turn bounded using inequality (B.10).

Hence, we conclude that

$$\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq c_0 \tau \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - cp^{-c'}$, for a universal constant c_0 (note that this constant depends on the bound $c_{\max}$ on $\lambda_{\max}(\Sigma_x)$). In particular, the choice of regularization parameter $\lambda = 2c_0\sqrt{\frac{\log p}{n}}$ ensures that $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq \frac{\lambda\tau}{2}$, w.h.p.

B.3. RSC condition

We now turn to the more challenging task of establishing the RSC condition (B.2). We show that w.h.p., the inequality

$$\mathcal{L}_n(\beta) - \mathcal{L}_n(\beta^*) - \langle \nabla \mathcal{L}_n(\beta^*), \beta - \beta^* \rangle \geq \alpha \|\beta - \beta^*\|_2^2$$

holds uniformly over the set

$$\mathbb{C} := \Big\{ \beta : \ \|\beta - \beta^*\|_2 \leq r, \ \|\beta_{S^c} - \beta^*_{S^c}\|_1 \leq 3\|\beta_S - \beta^*_S\|_1 \Big\}.$$

Defining

$$\mathcal{T}(\beta, \beta^*) := \mathcal{L}_n(\beta) - \mathcal{L}_n(\beta^*) - \langle \nabla \mathcal{L}_n(\beta^*), \beta - \beta^* \rangle,$$

we have

$$\begin{aligned}\mathcal{T}(\beta, \beta^*) = \frac{1}{n}\sum_{i=1}^n \Big(&\ell_\tau\left((x_i^T\beta - y_i)w(x_i)\right) - \ell_\tau\left(\epsilon_i w(x_i)\right) \\ &- \ell'_\tau\left(\epsilon_i w(x_i)\right) w(x_i) x_i^T(\beta - \beta^*) \Big).\end{aligned}$$

Note that for $|u_1|, |u_2| \leq \tau$, we have

$$\ell_\tau(u_1) - \ell_\tau(u_2) - \ell'_\tau(u_2)(u_1 - u_2) = \frac{(u_1 - u_2)^2}{2},$$

whereas the convexity of ℓ_τ implies that

$$\ell_\tau(u_1) - \ell_\tau(u_2) - \ell'_\tau(u_2)(u_1 - u_2) \geq 0, \qquad \forall u_1, u_2 \in \mathbb{R}.$$

Denote $\Delta := \beta - \beta^*$, and define the events

$$A_i^\beta := \left\{|\epsilon_i w(x_i)| \le \frac{\tau}{2}\right\} \bigcap \left\{|(x_i^T(\beta-\beta^*))w(x_i)| \le \frac{\tau}{2}\right\}, \qquad \forall 1 \le i \le n.$$

Note that on the event A_i^β, we have

$$|(x_i^T\beta - y_i)w(x_i)| \le |\epsilon_i w(x_i)| + |(x_i^T(\beta-\beta^*))w(x_i)| \le \tau,$$

so

$$\begin{aligned}\mathcal{T}(\beta,\beta^*) &\ge \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2 1\{A_i^\beta\} \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n \left(w(x_i)(x_i^T(\beta-\beta^*)\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(w(x_i)x_i^T(\beta-\beta^*)\right)^2 1\left\{(A_i^\beta)^c\right\}\right).\end{aligned}$$

We will now prove that the following statements hold, where γ is a sufficiently small constant to be specified later. The proofs of Lemmas 3, 4, and 5 may be found in Appendix B.4.

Lemma 3. *With probability at least* $1 - 2\exp\left(-\frac{c\delta n}{(b')^2} + 2k\log p\right)$, *we have*

$$\frac{1}{n}\sum_{i=1}^n \left(w(x_i)x_i^T(\beta-\beta^*)\right)^2 \ge \mathbb{E}\left[\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2\right] - 459\delta\|\beta-\beta^*\|_2^2, \quad \text{(B.11)}$$

uniformly over $\beta \in \mathbb{C}$.

Lemma 4. *With probability at least* $1-2\exp\left(-\frac{c\delta'\gamma n}{(b')^2} + 2k\log p + \gamma n \log\left(\frac{e}{\gamma}\right)\right)$, *we have*

$$\frac{1}{\gamma n}\sup_{|T|\le\gamma n}\sum_{i\in T}\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2 \le \mathbb{E}\left[\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2\right] + 459\delta'\|\beta-\beta^*\|_2^2, \quad \text{(B.12)}$$

uniformly over $\beta \in \mathbb{C}$. *In particular, taking* $\delta' \asymp \log\left(\frac{e}{\gamma}\right)$ *and assuming* $n \gtrsim k\log p$, *we have*

$$\begin{aligned}\frac{1}{n}\sup_{|T|\le\gamma n}\sum_{i\in T}\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2 \le \gamma\mathbb{E}\left[\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2\right] \\ + c'\gamma\log\left(\frac{e}{\gamma}\right)\|\beta-\beta^*\|_2^2,\end{aligned}$$

with probability at least $1 - 2\exp\left(-c''\gamma\log\left(\frac{e}{\gamma}\right)n\right)$.

Lemma 5. *Let* $r = \frac{\gamma\tau}{8\lambda_{\max}^{1/2}\left(\mathbb{E}\left[w^2(x_i)x_ix_i^T\right]\right)}$ *and suppose* $n \gtrsim k\log p$. *With probability at least* $1 - c\exp(-c'n)$, *we have*

$$\sup_{\beta\in\mathbb{C}}\sum_{i=1}^n 1\left\{(A_i^\beta)^c\right\} \le \gamma n. \quad \text{(B.13)}$$

Combining the results of Lemmas 3, 4, and 5, we see that

$$\begin{aligned}\mathcal{T}(\beta,\beta^*) &\geq \frac{1}{2}\Bigg((1-\gamma)\mathbb{E}\left[\left(w(x_i)x_i^T(\beta-\beta^*)\right)^2\right] \\ &\qquad - \left(459\delta + c'\gamma\log\left(\frac{e}{\gamma}\right)\right)\|\beta-\beta^*\|_2^2\Bigg) \\ &\geq \frac{1}{4}\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i)x_ix_i^T\right]\right)\|\beta-\beta^*\|_2^2,\end{aligned}$$

with probability at least $1-c\exp(-c'n)$, where we choose γ, δ, and δ' such that $\gamma \leq \frac{1}{4}$ and

$$459\delta + c'\gamma\log\left(\frac{e}{\gamma}\right) \leq \frac{1}{4}\lambda_{\min}\left(\mathbb{E}\left[w^2(x_i)x_ix_i^T\right]\right)$$

in order to ensure the second inequality. (Note that $\lim_{\gamma\to 0}\gamma\log\left(\frac{e}{\gamma}\right) = 0$.) This completes the proof.

B.4. Proofs of supporting lemmas

We now provide the proofs of Lemmas 3, 4, and 5.

B.4.1. Proof of Lemma 3

We make use of Lemma 14 in Appendix H. We will apply the lemma to the matrix

$$\Gamma = \frac{1}{n}\sum_{i=1}^n w^2(x_i)x_ix_i^T - \mathbb{E}\left[w^2(x_i)x_ix_i^T\right],$$

with $s = k$. (We will verify the deviation condition (H.1) momentarily.)

Denoting $\Delta := \beta - \beta^*$, we then have

$$\frac{1}{n}\sum_{i=1}^n\left(w(x_i)x_i^T\Delta\right)^2 \geq \mathbb{E}\left[w^2(x_i)\left(x_i^T\Delta\right)^2\right] - 27\delta\left(\|\Delta\|_2^2 + \frac{\|\Delta\|_1^2}{k}\right),$$

uniformly over all $\Delta\in\mathbb{R}^p$. Now note that for any $\beta\in\mathbb{C}$, we have

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{k}\|\Delta_S\|_2,$$

so

$$\|\Delta\|_2^2 + \frac{\|\Delta\|_1^2}{k} \leq 17\|\Delta\|_2^2,$$

from which inequality (B.11) follows.

Finally, note that the bound (H.1) in the hypothesis of Lemma 14 holds, w.h.p. Indeed, for $\|v\|_2 \leq 1$, the quantity $v^T\Gamma v$ is the recentered average of i.i.d. random variables, each of which is the square of a sub-Gaussian variable with

parameter b'. Thus, a standard ϵ-net argument over $2k$-dimensional subspaces and a union bound over the $\binom{p}{2k}$ choices of the support set implies that

$$\mathbb{P}\left(|v^T\Gamma v| \leq \delta, \quad \forall v \in \mathbb{R}^p \text{ s.t. } \|v\|_0 \leq 2k,\ \|v\|_2 \leq 1\right) \geq 1 - 2\exp\left(-\frac{c\delta n}{(b')^2} + 2k\log p\right) \tag{B.14}$$

(cf. Lemma 15 in Loh and Wainwright [54]). This proves the desired result.

B.4.2. Proof of Lemma 4

The proof is similar to the proof of Lemma 3, except that on top of the arguments used there, we also take a union bound over subsets of size at most γn, leading to an additional factor of $\binom{n}{\gamma n}$ in the error probability. Recalling a standard bound on binomial coefficients, we have

$$\binom{n}{\gamma n} \leq \left(\frac{e}{\gamma}\right)^{\gamma n},$$

and using this expression in the probability bound completes the proof.

B.4.3. Proof of Lemma 5

We write

$$\begin{aligned}\sup_{\beta\in\mathbb{C}} \sum_{i=1}^n 1\{(A_i^\beta)^c\} &\leq \sum_{i=1}^n 1\left\{|\epsilon_i w(x_i)| > \frac{\tau}{2}\right\} \\ &\quad + \sup_{\beta\in\mathbb{C}} \sum_{i=1}^n 1\left\{|(x_i^T(\beta-\beta^*))w(x_i)| > \frac{\tau}{2}\right\} \\ &\leq \sum_{i=1}^n 1\left\{|\epsilon_i| > \frac{\tau}{2}\right\} + \sup_{\beta\in\mathbb{C}} \sum_{i=1}^n 1\left\{|(x_i^T(\beta-\beta^*))w(x_i)| > \frac{\tau}{2}\right\}.\end{aligned} \tag{B.15}$$

For the first term in inequality (B.15), note that by the Chernoff bound in Lemma 16, we have

$$\frac{1}{n}\sum_{i=1}^n 1\left\{|\epsilon_i| > \frac{\tau}{2}\right\} \leq 3\mathbb{P}\left(|\epsilon_i| > \frac{\tau}{2}\right) \leq \frac{3(\sigma^*)^2}{\tau^2/4},$$

with probability at least $1-\exp(-cn)$, where the second inequality comes from Markov's inequality. In particular, we can guarantee that this term is bounded by $\frac{\gamma}{2}$ if we take $\tau \geq c_\tau \sigma^*$, where the constant c_τ depends on γ.

For the second term in inequality (B.15), the bound $1\{x \geq y\} \leq \frac{x}{y}$ for $x \geq 0$ and $y > 0$, together with the Cauchy-Schwarz inequality, implies that

$$\sup_{\beta \in \mathbb{C}} \frac{1}{n} \sum_{i=1}^{n} 1\left\{|(x_i^T(\beta - \beta^*))w(x_i)| > \frac{\tau}{2}\right\} \leq \frac{\sup_{\beta \in \mathbb{C}} \frac{1}{n} \sum_{i=1}^{n} |x_i^T(\beta - \beta^*)w(x_i)|}{\tau/2} \leq \sup_{\beta \in \mathbb{C}} \frac{2}{\tau} \sqrt{\frac{\sum_{i=1}^{n} \left(w(x_i)x_i^T(\beta - \beta^*)\right)^2}{n}}. \tag{B.16}$$

By an analogous argument to the one employed in the proof of Lemma 3, we can derive the bound

$$\frac{1}{n} \sum_{i=1}^{n} \left(w(x_i)x_i^T(\beta - \beta^*)\right)^2 \leq \mathbb{E}\left[\left(w(x_i)x_i^T(\beta - \beta^*)\right)^2\right] + 459\delta\|\beta - \beta^*\|_2^2,$$

with probability at least $1 - 2\exp\left(-\frac{c\delta n}{(b')^2} + 2k\log p\right)$, uniformly over all $\beta \in \mathbb{C}$. Combined with inequality (B.16), this implies that

$$\sup_{\beta \in \mathbb{C}} \frac{1}{n} \sum_{i=1}^{n} 1\left\{|(x_i^T(\beta - \beta^*))w(x_i)| > \frac{\tau}{2}\right\} \leq \frac{2}{\tau}\left(\lambda_{\max}\left(\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]\right) + 459\delta\right)^{1/2} r,$$

w.h.p. If we take $\delta = \frac{3\lambda_{\max}\left(\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]\right)}{459}$, we have

$$\frac{2}{\tau}\left(\lambda_{\max}\left(\mathbb{E}\left[w^2(x_i)x_i x_i^T\right]\right) + 459\delta\right)^{1/2} r = \frac{\gamma}{2}.$$

Thus, both terms in inequality (B.15) are bounded by $\frac{\gamma n}{2}$, leading to the desired result.

Appendix C: Proof of Theorem 2

Let $j' = \min\{j \in \mathcal{J} : \tau_j \geq c_\tau \sigma^*\}$. Then $\tau_{j'} \leq 2c_\tau\sigma^*$. We have

$$\begin{aligned}\mathbb{P}(j_* > j') = \mathbb{P}\Bigg(&\bigcup_{i \in \mathcal{J}: i > j'} \left\{\|\widehat{\beta}_{(i)} - \widehat{\beta}_{(j')}\|_2 > 2C\tau_i\sqrt{\frac{k\log p}{n}}\right\} \\ &\bigcup_{i \in \mathcal{J}: i > j'} \left\{\|\widehat{\beta}_{(i)} - \widehat{\beta}_{(j')}\|_1 > 8C\tau_i k\sqrt{\frac{\log p}{n}}\right\}\Bigg) \\ \leq \mathbb{P}\Bigg(&\|\widehat{\beta}_{(j')} - \beta^*\|_2 > C\tau_{j'}\sqrt{\frac{k\log p}{n}} \quad \text{or}\end{aligned}$$

$$\|\widehat{\beta}_{(j')} - \beta^*\|_1 > 4C\tau_{j'}k\sqrt{\frac{\log p}{n}}\Bigg)$$

$$+ \sum_{i \in \mathcal{J}: i > j'} \mathbb{P}\Bigg(\|\widehat{\beta}_{(i)} - \beta^*\|_2 > C\tau_i\sqrt{\frac{k\log p}{n}} \quad \text{or}$$

$$\|\widehat{\beta}_{(i)} - \beta^*\|_1 > 4C\tau_i k\sqrt{\frac{\log p}{n}}\Bigg)$$

$$\leq cp^{-c'} + \log_2\left(\frac{2\tau_{\max}}{\tau_{\min}}\right) \cdot cp^{-c'},$$

where we have used Theorem 1 and a union bound in the final inequality.

Hence, with probability at least $1 - \log_2\left(\frac{4\tau_{\max}}{\tau_{\min}}\right) \cdot cp^{-c'}$, we have $j' \geq j_*$ and the bounds

$$\|\widehat{\beta}_{(j')} - \beta^*\|_2 \leq C\tau_{j'}\sqrt{\frac{k\log p}{n}},$$
$$\|\widehat{\beta}_{(j')} - \beta^*\|_1 \leq 4C\tau_{j'}k\sqrt{\frac{\log p}{n}}.$$

It follows that

$$\begin{aligned}\|\widehat{\beta}_{(j_*)} - \beta^*\|_2 &\leq \|\widehat{\beta}_{(j_*)} - \widehat{\beta}_{(j')}\|_2 + \|\widehat{\beta}_{(j')} - \beta^*\|_2 \\ &\leq 2C\tau_{j'}\sqrt{\frac{k\log p}{n}} + C\tau_{j'}\sqrt{\frac{k\log p}{n}} \\ &= 3C\tau_{j'}\sqrt{\frac{k\log p}{n}} \\ &\leq 6Cc_\tau\sigma^*\sqrt{\frac{k\log p}{n}},\end{aligned}$$

using the fact that $\tau_{j'} \leq 2c_\tau\sigma^*$ in the final inequality.

Similarly, we have the bound

$$\|\widehat{\beta}_{(j_*)} - \beta^*\|_1 \leq 24Cc_\tau\sigma^* k\sqrt{\frac{\log p}{n}}.$$

Appendix D: Proof of Theorem 4

We first present the main argument, with supporting lemmas in the succeeding subsections.

D.1. Main argument

We write

$$\sqrt{n}(\widehat{b}_\psi - \beta^*) = \sqrt{n}(\widehat{\beta} - \beta^*) + \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right)x_i$$

$$= \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\xi_i) x_i + \sqrt{n} \Bigg\{ (\widehat{\beta} - \beta^*) + \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\psi\left(\frac{y_i - x_i^T \widehat{\beta}}{\widehat{\sigma}} \right) - \psi\left(\frac{y_i - x_i^T \beta^*}{\sigma_\xi^*} \right) \right) x_i \Bigg\}$$
$$:= I + II. \tag{D.1}$$

We first consider the term $I = \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\xi_i) x_i$, which we claim is asymptotically normal. We have

$$\begin{aligned}
&\left\| P_J \left(\frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{\Theta_x}{\sigma_\xi^* A(\psi)} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\xi_i) x_i \right\|_\infty \\
&\quad \le \left\| \left(\frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{\Theta_x}{\sigma_\xi^* A(\psi)} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\xi_i) x_i \right\|_\infty \\
&\quad \le \left\| \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{\Theta_x}{\sigma_\xi^* A(\psi)} \right\|_1 \cdot \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\xi_i) x_i \right\|_\infty .
\end{aligned}$$

By Lemma 7, the second factor is $\mathcal{O}_P(\sqrt{\log p})$. To handle the first factor, we write

$$\begin{aligned}
\left\| \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{\Theta_x}{\sigma_\xi^* A(\psi)} \right\|_1 \le{}& \left\| \frac{1}{\sigma_\xi^* A(\psi)} \left(\widehat{\Theta} - \Theta_x \right) \right\|_1 + \left\| \left(\frac{1}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{1}{\sigma_\xi^* A(\psi)} \right) \Theta_x \right\|_1 \\
&+ \left\| \left(\frac{1}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{1}{\sigma_\xi^* A(\psi)} \right) \left(\widehat{\Theta} - \Theta_x \right) \right\|_1 \\
\le{}& \frac{1}{|\sigma_\xi^* A(\psi)|} \left\| \widehat{\Theta} - \Theta_x \right\|_1 + \left| \frac{1}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{1}{\sigma_\xi^* A(\psi)} \right| \|\Theta_x\|_1 \\
&+ \left| \frac{1}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{1}{\sigma_\xi^* A(\psi)} \right| \left\| \widehat{\Theta} - \Theta_x \right\|_1 \\
={}& \mathcal{O}_P \left(\frac{k \log^{3/2} p}{\sqrt{n}} \right),
\end{aligned} \tag{D.2}$$

where the final inequality leverages Lemma 9 and the condition (4.6).

Together with the convergence statement in Lemma 10, we conclude that I has the desired asymptotic normality property, since

$$\mathcal{O}_P \left(\frac{k \log^{3/2} p}{\sqrt{n}} \right) \cdot \mathcal{O}_P(\sqrt{\log p}) = o_P(1),$$

assuming $n \gtrsim k^2 \operatorname{polylog}(p)$.

We now shift our attention to term II on the right-hand side of equation (D.1). By Taylor's theorem applied to each summand, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right) - \psi\left(\frac{y_i - x_i^T\beta^*}{\sigma_\xi^*}\right)\right)x_i \\ = \frac{1}{n}\sum_{i=1}^{n}\psi'(\xi_i)\widehat{\delta}_i x_i + \frac{1}{n}\sum_{i=1}^{n}\frac{\psi''(\widehat{t}_i)}{2}\widehat{\delta}_i^2 x_i,$$

where $\widehat{\delta}_i := \frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}} - \frac{y_i - x_i^T\beta^*}{\sigma_\xi^*}$ and $\widehat{t}_i$ lies on the segment between $\frac{y_i - x_i^T\beta^*}{\sigma_\xi^*}$ and $\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}$. We have the bound

$$\sqrt{n}\left\|\frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\psi''(\widehat{t}_i)}{2}\widehat{\delta}_i^2 x_i\right\|_\infty \le \frac{\sqrt{n}}{2}\left\|\frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)}\right\|_1 \cdot \left\|\frac{1}{n}\sum_{i=1}^{n}\psi''(\widehat{t}_i)\widehat{\delta}_i^2 x_i\right\|_\infty,$$

and

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\psi''(\widehat{t}_i)\widehat{\delta}_i^2 x_i\right\|_\infty \le \|\psi''\|_\infty\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\delta}_i^2\right|\cdot\|X\|_{\max} \\ = \mathcal{O}_P\left(\frac{k\log p}{n}\cdot\log p\right),$$

using the expansion $\widehat{\delta}_i = \frac{x_i(\beta^* - \widehat{\beta})}{\widehat{\sigma}} + \epsilon_i\left(\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right)$ and the same argument employed to bound the term B_3 in the proof of Lemma 9 and the bound on $\|X\|_{\max}$ from Lemma 6. Altogether, we have the bound

$$\sqrt{n}\left\|\frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\psi''(\widehat{t}_i)}{2}\widehat{\delta}_i^2 x_i\right\|_\infty = \mathcal{O}_P\left(\frac{k\log^2 p}{\sqrt{n}}\right) = o_P(1).$$

Finally, again using the expansion $\widehat{\delta}_i = \frac{x_i(\beta^* - \widehat{\beta})}{\widehat{\sigma}} + \epsilon_i\left(\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right)$, we have

$$\left\|(\widehat{\beta} - \beta^*) + \frac{\widehat{\Theta}}{\widehat{\sigma}\widehat{A}(\psi)}\frac{1}{n}\sum_{i=1}^{n}\psi'(\xi_i)\widehat{\delta}_i x_i\right\|_\infty \\ \le \left\|\left(I - \frac{\widehat{\Theta}}{\widehat{\sigma}^2\widehat{A}(\psi)}\left(\frac{1}{n}\sum_{i=1}^{n}\psi'(\xi_i)x_i x_i^T\right)\right)(\widehat{\beta} - \beta^*)\right\|_\infty \\ + \left\|\frac{\widehat{\Theta}}{\widehat{\sigma}^2\widehat{A}(\psi)}\right\|_1 \cdot \left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right| \cdot \left\|\frac{1}{n}\sum_{i=1}^{n}\psi'(\xi_i)\epsilon_i x_i\right\|_\infty \\ := A_1 + A_2. \tag{D.3}$$

We then bound

$$A_1 \le \left\|\left(I - \frac{\widehat{\Theta}}{\widehat{\sigma}^2\widehat{A}(\psi)}\left(\frac{1}{n}\sum_{i=1}^{n}\psi'(\xi_i)x_i x_i^T\right)\right)\right\|_{\max}\|\widehat{\beta} - \beta^*\|_1$$

$$\leq \left(1 + \left\| \frac{\widehat{\Theta}}{\widehat{\sigma}^2 \widehat{A}(\psi)} \right\|_1 \left\| \frac{1}{n} \sum_{i=1}^n \psi'(\xi_i) x_i x_i^T \right\|_{\max} \right) \|\widehat{\beta} - \beta^*\|_1$$
$$= \mathcal{O}_P\left(1 + \left(1 + \frac{k \log^{3/2} p}{\sqrt{n}}\right) \cdot \log^2 p\right) \cdot \mathcal{O}_P\left(k\sqrt{\frac{\log p}{n}}\right),$$

using inequality (D.2) and Lemma 8, the bound on $\|X\|_{\max}$ from Lemma 6, and the ℓ_1-error bound on $\widehat{\beta}$ in the last inequality. Hence, we conclude that

$$A_1 = \mathcal{O}_P\left(\frac{k \log^{5/2} p}{\sqrt{n}}\right) = o_P(1),$$

under the assumption $n \gtrsim k^2 \operatorname{polylog}(p)$. Next, we bound A_2 by noting that

$$\begin{aligned}\left\| \frac{1}{n} \sum_{i=1}^n \psi'(\xi_i) \epsilon_i x_i \right\|_\infty &\leq \|\psi'\|_\infty \left(\frac{1}{n} \sum_{i=1}^n |\epsilon_i|\right) \cdot \|X\|_{\max} \\ &= \left(\mathbb{E}[|\epsilon_i|] + \mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)\right) \cdot \mathcal{O}_P(\log p) \\ &= \mathcal{O}_P(\log p),\end{aligned}$$

using Lemma 6 and Chebyshev's inequality. Combined with the deviation bound on $\left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma^*}\right|$ from Lemma 8 and the bound on $\left\| \frac{\widehat{\Theta}}{\widehat{\sigma}^2 \widehat{A}(\psi)} \right\|_1$ from inequality (D.2) and Lemma 8, we then have

$$A_2 = \mathcal{O}_P\left(\frac{k \log^{3/2} p}{\sqrt{n}} \cdot \log p\right) = o_P(1),$$

as well. The desired result then follows.

D.2. Supporting lemmas

We begin with a lemma concerning the magnitude of the entries of the design matrix.

Lemma 6. *Suppose Assumption 2 holds. Then*

$$\mathbb{P}\left(\|X\|_{\max} \geq 2\sigma_x \log(np)\right) \leq \frac{1}{np}.$$

Proof. Applying a union bound to the entries of X, we have

$$\mathbb{P}\left(\|X\|_{\max} \geq t\right) \leq np \exp\left(-\frac{t}{\sigma_x}\right).$$

Taking $t = 2\sigma_x \log(np)$ then gives the desired result. □

The next lemma is a concentration inequality derived using Lemma 15:

Lemma 7. *Under Assumptions 2 and 3, and assuming* $n \gtrsim \text{polylog}(p)$, *we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \psi(\xi_i) x_i \right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} \right), \tag{D.4}$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \psi''(\xi_i) x_i \right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} \right), \tag{D.5}$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \psi(\xi_i)\psi'(\xi_i) x_i \right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} \right), \tag{D.6}$$

with probability at least $1 - Cn^{-c} - p^{-c'}$.

Proof. All the inequalities are proved by applying Lemma 15, where the vectors $\{y_i\}_{i=1}^n$ in the lemma are $2p$-dimensional vectors containing the summands in the respective inequalities, together with their additive inverses.

For inequality (D.4), let $y_i = \psi(\xi_i)x_i$. Note that conditioned on the ϵ_i's, the y_i's are independent, zero-mean vectors. Since $\psi(\xi_i)$ is sub-exponential, a union bound gives

$$\mathbb{P}\left(\max_{1 \le i \le n} |\psi(\xi_i)| \le 2\sigma_\xi \log n \right) \ge 1 - \frac{1}{n}$$

(cf. the proof of Lemma 6). Furthermore, by Chebyshev's inequality, we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \psi^2(\xi_i) - \mathbb{E}\left[\psi^2(\xi_i)\right] \right| \le \frac{1}{2}\mathbb{E}\left[\psi^2(\xi_i)\right] \right) \le \frac{4\,\text{Var}[\psi^2(\xi_i)]}{n\mathbb{E}\left[\psi^2(\xi_i)\right]^2}.$$

Hence, defining

$$E := \left\{ \max_{1 \le i \le n} |\psi(\xi_i)| \le 2\sigma_\xi \log n \right\} \bigcap \left\{ \frac{1}{2}\mathbb{E}\left[\psi^2(\xi_i)\right] \le \frac{1}{n}\sum_{i=1}^{n} \psi^2(\xi_i) \le \frac{3}{2}\mathbb{E}\left[\psi^2(\xi_i)\right] \right\},$$

we have $\mathbb{P}(E) \ge 1 - \frac{c}{n}$.

We claim that the conditions (H.2) and (H.3) of Lemma 15 are satisfied with $B_n \asymp \text{polylog}(p)$, conditioned on E. Indeed, we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[y_{ij}^2 \mid \{\epsilon_i\}_{i=1}^n] = \frac{1}{n}\sum_{i=1}^{n} \psi^2(\xi_i)\,\text{Var}(x_{ij}).$$

By assumption, we have $c_1 \le \text{Var}(x_{ij}) \le c_2\sigma_x$ for all j. Furthermore, on the event E, the quantity $\frac{1}{n}\sum_{i=1}^{n}\psi^2(\xi_i)$ is bounded. This establishes inequality (H.2). For condition (H.3), recall that since x_{ij} is sub-exponential, we have

$$\mathbb{E}\left[\exp\left(\frac{|x_{ij}|}{c_3\sigma_x} \right) \right] \le 2$$

for some constant $c_3 > 0$. Then

$$\mathbb{E}\left[|y_{ij}|^{2+\ell} \mid \{\epsilon_i\}_{i=1}^n\right] \leq \max_{1\leq i\leq n} |\psi(\xi_i)|^{2+\ell} \cdot \mathbb{E}\left[|x_{ij}|^{2+\ell}\right] \leq (2\sigma_\xi \log n)^{2+\ell} \cdot c_3 \sigma_x^{2+\ell}$$

and

$$\mathbb{E}\left[\exp\left(\frac{|y_{ij}|}{c_3\sigma_x \cdot 2\sigma_\xi \log n}\right) \Big| \{\epsilon_i\}_{i=1}^n\right] \leq \mathbb{E}\left[\exp\left(\frac{|x_{ij}|}{c_3\sigma_x}\right)\right] \leq 2$$

on E. Hence, by taking $B_n \asymp \mathrm{polylog}(p)$, we can guarantee that condition (H.3) is satisfied.

Inequalities (D.5) and (D.6) are proved in a similar manner, noting that $\|\psi'\|_\infty, \|\psi''\|_\infty < \infty$ by assumption, so the terms involving $\psi(\xi_i)$ are still sub-exponential. □

The next two lemmas use the preceding concentration results to prove convergence of certain empirical quantities to their population-level counterparts.

Lemma 8. *Under the assumptions of Theorem 4, we have*

$$|\widehat{\sigma} - \sigma_\xi^*| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right),$$

$$\left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right), \quad \textit{and} \quad \left|\frac{1}{\widehat{\sigma}^2} - \frac{1}{(\sigma_\xi^*)^2}\right| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right).$$

Proof. Using the triangle inequality, we write

$$\begin{aligned}\mathrm{Var}(\xi_i) \cdot |\widehat{\sigma}^2 - (\sigma_\xi^*)^2| \leq &\left|\frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\widehat{\beta})^2 - \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta^*)^2\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta^*)^2 - \mathbb{E}[\epsilon_i^2]\right|.\end{aligned}$$

We bound the second term by $\mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)$ via Chebyshev's inequality, using the assumption that $\mathbb{E}[\xi_i^4] < \infty$. Expanding and using the triangle inequality, we bound the first term as

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \left(\left(x_i^T(\beta^* - \widehat{\beta}) + \epsilon_i\right)^2 - \epsilon_i^2\right)\right| \\ &\quad\leq \frac{1}{n}\left|\sum_{i=1}^n \left(x_i^T(\widehat{\beta} - \beta^*)\right)^2\right| + \frac{2}{n}\left|\sum_{i=1}^n \left(x_i^T(\widehat{\beta} - \beta^*)\right)\epsilon_i\right| \\ &\quad\leq (\widehat{\beta} - \beta^*)^T\widehat{\Sigma}(\widehat{\beta} - \beta^*) + 2\left\|\frac{X^T\epsilon}{n}\right\|_\infty \|\widehat{\beta} - \beta^*\|_1. \end{aligned} \tag{D.7}$$

For the first term in inequality (D.7), we show that

$$\left|(\widehat{\beta}-\beta^*)^T\widehat{\Sigma}(\widehat{\beta}-\beta^*)\right| = \mathcal{O}_P\left(\frac{k\log p}{n}\right). \tag{D.8}$$

We use Lemma 14 with $\Gamma = \widehat{\Sigma}$, $\delta = \Theta(1)$, and $s = k$. In particular, we will show that inequality (H.1) holds w.h.p. Then the lemma implies that

$$\left|(\widehat{\beta}-\beta^*)^T\widehat{\Sigma}(\widehat{\beta}-\beta^*)\right| = \mathcal{O}_P\left(\|\widehat{\beta}-\beta^*\|_2^2 + \frac{\|\widehat{\beta}-\beta^*\|_1^2}{k}\right) = \mathcal{O}_P\left(\frac{k\log p}{n}\right).$$

In order to verify the deviation condition (H.1), note that by Lemma 6, we can define $\widetilde{\Sigma} = \frac{\widetilde{X}^T\widetilde{X}}{n}$, where $\widetilde{X}$ is the matrix X with columns truncated according to $\widetilde{x}_i = x_i \cdot 1\{\|x_i\|_\infty > 2\sigma_x \log(np)\}$; then $\widehat{\Sigma} = \widetilde{\Sigma}$, w.h.p. Furthermore, $\widetilde{\Sigma}$ is the sample covariance matrix of bounded i.i.d. random vectors, so we have

$$\sup_{v:\|v\|_0\le 2s, \|v\|_2\le 1} |v^T(\widetilde{\Sigma} - \mathbb{E}(\widetilde{\Sigma}))v| \le \delta',$$

w.h.p., using a standard ϵ-net + union bound argument (cf. inequality (B.14) in the proof of Lemma 3), where $\delta' = O\left(\frac{k\,\mathrm{polylog}\,p}{n})\right)$. Hence, by the triangle inequality, we have

$$\begin{aligned}\sup_{v:\|v\|_0\le 2s, \|v\|_2\le 1} |v^T\widetilde{\Sigma}v| &\le \delta' + \sup_{\|v\|_2\le 1} \mathbb{E}\left[(\widetilde{x}_i^T v)^2\right] \le \delta' + \sup_{\|v\|_2\le 1} \mathbb{E}\left[(x_i^T v)^2\right] \\ &\le \delta' + \lambda_{\max}(\Sigma_x),\end{aligned}$$

giving the desired result. For the second term in inequality (D.7), we have

$$\begin{aligned}\left\|\frac{X^T\epsilon}{n}\right\|_\infty &\le \|X\|_{\max}\cdot\frac{\|\epsilon\|_1}{n} = \mathcal{O}_P(\log p)\cdot\left(\mathbb{E}[|\epsilon_i|] + \mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)\right) \\ &= \mathcal{O}_P(\log p),\end{aligned}$$

using a Chebyshev argument and Lemma 6. Altogether, we conclude that

$$\begin{aligned}|\widehat{\sigma}^2 - (\sigma_\xi^*)^2| &= \mathcal{O}_P\left(\frac{k^2\log^3 p}{n}\right) + \mathcal{O}_P(\log p)\cdot\mathcal{O}_P\left(k\sqrt{\frac{\log p}{n}}\right) + \mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right) \\ &= \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right),\end{aligned} \tag{D.9}$$

using the fact that $n \gtrsim k^2\,\mathrm{polylog}(p)$ by assumption. Finally, note that

$$|\widehat{\sigma} - \sigma_\xi^*| = \frac{|\widehat{\sigma}^2 - (\sigma_\xi^*)^2|}{|\widehat{\sigma} + \sigma_\xi^*|},$$

so when $n \gtrsim k^2 \,\mathrm{polylog}(p)$, we have $|\widehat{\sigma} - \sigma_\xi^*| = \mathcal{O}_P\left(|\widehat{\sigma}^2 - (\sigma_\xi^*)^2|\right)$, from which the desired bound follows.

For the second bound, we simply write

$$\left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right| = \frac{1}{\widehat{\sigma}\sigma_\xi^*}|\widehat{\sigma} - \sigma_\xi^*| \le \frac{1}{\sigma_\xi^*\left(\sigma_\xi^* - \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right)\right)} \cdot \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right)$$
$$= \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right).$$

The third bound may be obtained in a similar manner via inequality (D.9). □

Lemma 9. *Under the assumptions of Theorem 4, we have*

$$|\widehat{A}(\psi) - A(\psi)| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right) \tag{D.10}$$

and

$$\left|\frac{1}{\widehat{\sigma}\widehat{A}(\psi)} - \frac{1}{\sigma_\xi^* A(\psi)}\right| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right). \tag{D.11}$$

Proof. We first prove the bound (D.10). We use the bound on $\left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma^*}\right|$ from Lemma 8.

By the triangle inequality, we have

$$\begin{aligned}|\widehat{A}(\psi) - A(\psi)| \le{}& \left|\frac{1}{\widehat{\sigma}^2} - \frac{1}{(\sigma_\xi^*)^2}\right|\left|\frac{1}{n}\sum_{i=1}^n \psi'\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right)\right| \\ &+ \frac{1}{(\sigma_\xi^*)^2}\left|\frac{1}{n}\sum_{i=1}^n\left(\psi'\left(\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}\right) - \psi'\left(\frac{y_i - x_i^T\beta^*}{\sigma_\xi^*}\right)\right)\right| \\ &+ \frac{1}{(\sigma_\xi^*)^2}\left|\frac{1}{n}\sum_{i=1}^n \psi'(\xi_i) - \mathbb{E}\left[\psi'(\xi_i)\right]\right| \\ :={}& A + B + C.\end{aligned}$$

We now bound each of the terms separately. Note that since $\|\psi'\|_\infty < \infty$ by assumption, term C may be bounded directly by $\mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)$ using Hoeffding's inequality. Furthermore, we have

$$A \le \|\psi'\|_\infty \cdot \left|\frac{1}{\widehat{\sigma}^2} - \frac{1}{(\sigma_\xi^*)^2}\right| = \mathcal{O}_P\left(\frac{k\log^{3/2} p}{\sqrt{n}}\right),$$

using Lemma 8. Finally, defining

$$\widehat{\delta}_i := \frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}} - \frac{y_i - x_i^T\beta^*}{\sigma_\xi^*} = \frac{x_i(\beta^* - \widehat{\beta})}{\widehat{\sigma}} + \epsilon_i\left(\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right),$$

we may use a Taylor series expansion to write

$$B = \frac{1}{(\sigma_\xi^*)^2}\left|\frac{1}{n}\sum_{i=1}^n \psi''(\xi_i)\widehat{\delta}_i + \frac{1}{n}\sum_{i=1}^n \frac{\psi^{(3)}(\widehat{u}_i)}{2}\cdot\widehat{\delta}_i^2\right|,$$

where $\widehat{u}_i$ lies on the segment between $\frac{y_i - x_i^T\beta^*}{\sigma_\xi^*}$ and $\frac{y_i - x_i^T\widehat{\beta}}{\widehat{\sigma}}$, for each i. Applying the triangle inequality and Hölder's inequality then gives

$$\begin{aligned}(\sigma_\xi^*)^2 B &\le \frac{1}{\widehat{\sigma}}\left\|\frac{1}{n}\sum_{i=1}^n \psi''(\xi_i)x_i\right\|_\infty \cdot \|\widehat{\beta}-\beta^*\|_1 + \left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right|\cdot\left|\frac{1}{n}\sum_{i=1}^n \psi''(\xi_i)\epsilon_i\right| \\ &\quad + \|\psi^{(3)}\|_\infty\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{x_i(\widehat{\beta}-\beta^*)}{\widehat{\sigma}}\right)^2 + \frac{1}{n}\sum_{i=1}^n \epsilon_i^2\left(\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right)^2\right) \\ &:= B_1 + B_2 + B_3. \end{aligned} \tag{D.12}$$

We claim that $B = \mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right)$.

Using Lemma 7, together with the assumed ℓ_1-error bound on $\widehat{\beta}$, we have

$$B_1 = \mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)\cdot\mathcal{O}_P\left(k\sqrt{\frac{\log p}{n}}\right).$$

Furthermore, since $\left|\frac{1}{n}\sum_{i=1}^n \psi''(\xi_i)\epsilon_i\right| \le \|\psi''\|_\infty\left(\frac{1}{n}\sum_{i=1}^n|\epsilon_i|\right)$, we have

$$B_2 = \mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right)\cdot\left(\mathbb{E}[|\epsilon_i|] + \mathcal{O}_P\left(\sqrt{\frac{\log p}{n}}\right)\right) = \mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right),$$

using the error bound on $\left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right|$ from Lemma 8, boundedness of $\|\psi''\|_\infty$, and Chebyshev's inequality. Finally, we have

$$\begin{aligned} B_3 &\le \|\psi^{(3)}\|_\infty\left(\frac{1}{\widehat{\sigma}^2}(\widehat{\beta}-\beta^*)^T\widehat{\Sigma}(\widehat{\beta}-\beta^*) + \left|\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma_\xi^*}\right|^2\frac{1}{n}\sum_{i=1}^n\epsilon_i^2\right) \\ &= \mathcal{O}_P\left(\frac{k\log p}{n} + \left(\frac{k\log^{3/2}p}{\sqrt{n}}\right)^2\cdot\sqrt{\frac{\log p}{n}}\right) \\ &= \mathcal{O}_P\left(\frac{k\log p}{n}\right), \end{aligned}$$

via the same argument used in inequality (D.8) of Lemma 8, the assumed ℓ_2-error bound on $\widehat{\beta}$, and the assumption $n \gtrsim k^2\,\mathrm{polylog}(p)$.

Putting the results together, we have

$$|\widehat{A}(\psi) - A(\psi)| = \mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right),$$

as claimed.

To establish inequality (D.11), we write

$$\begin{aligned}\left|\frac{1}{\widehat{A}(\psi)}-\frac{1}{A(\psi)}\right| &= \frac{1}{|\widehat{A}(\psi)A(\psi)|}|\widehat{A}(\psi)-A(\psi)| \\ &\leq \frac{1}{|A(\psi)|\left(|A(\psi)|+\mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right)\right)}\cdot\mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right) \\ &= \mathcal{O}_P\left(\frac{k\log^{3/2}p}{\sqrt{n}}\right). \end{aligned} \tag{D.13}$$

Then

$$\begin{aligned}\left|\frac{1}{\widehat{\sigma}\widehat{A}(\psi)}-\frac{1}{\sigma_\xi^* A(\psi)}\right| \leq \frac{1}{\sigma_\xi^*}\left|\frac{1}{\widehat{A}(\psi)}-\frac{1}{A(\psi)}\right| + \left|\frac{1}{\widehat{\sigma}}-\frac{1}{\sigma_\xi^*}\right|\frac{1}{A(\psi)} \\ + \left|\frac{1}{\widehat{\sigma}}-\frac{1}{\sigma_\xi^*}\right|\cdot\left|\frac{1}{\widehat{A}(\psi)}-\frac{1}{A(\psi)}\right|,\end{aligned}$$

and the bounds from inequality (D.13) and Lemma 8 provide the desired result. □

Finally, we derive an asymptotic normality result, which follows from an application of the multivariate Central Limit Theorem.

Lemma 10. *Under the assumptions of Theorem 4, we have the convergence in distribution*

$$P_J\cdot\frac{\Theta_x}{\sigma_\xi^* A(\psi)}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(\xi_i)x_i \xrightarrow{d} N\left(0,\ \frac{\mathbb{E}[\psi^2(\xi_i)]}{(\sigma_\xi^*)^2A^2(\psi)}\cdot(\Theta_x)_{JJ}\right), \tag{D.14}$$

as $n,p\to\infty$.

Proof. The convergence statement is easy to verify using the assumptions and the multivariate CLT (cf. Lemma 18). Indeed, the mixed third moments of the summands are finite, since they are products of sub-exponential variables, and the variance of the limiting distribution is obtained via the calculation

$$\begin{aligned}\operatorname{Var}\left(P_J\cdot\frac{\Theta_x}{\sigma_\xi^*A(\psi)}\cdot\psi(\xi_i)x_i\right) &= \frac{P_J\Theta_x}{\sigma_\xi^*A(\psi)}\cdot\mathbb{E}\left[\psi^2(\xi_i)\right]\cdot\Sigma_x\frac{\Theta_x P_J^T}{\sigma_\xi^*A(\psi)} \\ &= \frac{\mathbb{E}[\psi^2(\xi_i)]}{(\sigma_\xi^*)^2A^2(\psi)}\cdot P_J\Theta_x P_J^T.\end{aligned}$$

□

Appendix E: Proof of Proposition 1

We adapt an argument from Ravikumar et al. [63], suitable for the present setting. The main technical argument is a primal-dual witness construction,

which shows that the solution of the graphical Lasso restricted to the true support set also yields the unique global optimum when padded with zeros to obtain a $p \times p$ matrix. We only mention the necessary amendments to the arguments used in Ravikumar et al. [63]; for more details, see the paper.

Following the proof of Theorem 1 in Ravikumar et al. [63], we denote $W = \widehat{\Sigma} - \Sigma_x$. We can easily derive the following lemma:

Lemma 11. *The entrywise MoM covariance estimator* $\widehat{\Sigma}$ *satisfies*

$$\left\|\widehat{\Sigma} - \Sigma_x\right\|_{\max} \leq c\sigma_x^2 \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - \exp(-c'n)$.

Proof. We apply Lemma 13 in Appendix H with $\epsilon = 1$, together with a union bound. Note that by assumption, the marginals of the x_i's are sub-exponential random variables, so moments of all orders are finite. Then we have

$$\mathbb{P}\left(\left\|\widehat{\Sigma} - \Sigma_x\right\|_{\max} \geq c\sigma_x^2 t\right) \leq p \exp(-c'nt^2),$$

for all $t > 0$, from which the desired result follows by taking $t \asymp \sqrt{\frac{\log p}{n}}$ and using the sample size assumption $n \gtrsim \log p$. □

By Lemma 11, we can guarantee that

$$\|W\|_{\max} \leq \frac{\alpha\lambda}{8},$$

w.h.p., by choosing the constant c in the bound to be sufficiently small. Next, we define the matrix function

$$R(\Delta) = \widehat{\Theta}^{-1} - \Theta_x^{-1} + \Theta_x^{-1}\Delta\Theta_x^{-1}.$$

By Lemma 5 of Ravikumar et al. [63], we know that $\|\Delta\|_{\max} \leq \frac{1}{3\kappa_\Sigma d}$ implies that

$$\|R(\Delta)\|_{\max} \leq \frac{3k\|\Delta\|_{\max}^2 \kappa_\Sigma^3}{2}.$$

Lemma 6 of Ravikumar et al. [63] then applies directly, as well, stating that if

$$r := 2\kappa_\Gamma(\|W\|_{\max} + \lambda) \leq \min\left\{\frac{1}{3\kappa_\Sigma k}, \frac{1}{3\kappa_\Sigma^3 \kappa_\Gamma k}\right\}, \tag{E.1}$$

we have

$$\|\widehat{\Theta} - \Theta_x\|_{\max} \leq r.$$

Note that the bound (E.1) holds by our assumption on the range of λ. In particular, we have

$$\|R(\widehat{\Theta} - \Theta_x)\|_{\max} \leq \frac{3\kappa_\Sigma^3 k}{2}\|\widehat{\Theta} - \Theta_x\|_{\max}^2$$

$$\leq \frac{3\kappa_\Sigma^3 k}{2} \cdot 4\kappa_\Gamma^2 \left(\frac{\alpha}{8}+1\right)^2 \lambda^2$$
$$\leq \frac{\alpha\lambda}{8},$$

by our assumptions. Lemma 4 of Ravikumar et al. [63] then applies, implying the required strict dual feasibility result and the validity of the primal-dual witness construction argument.

Appendix F: Proof of Theorem 5

By Theorem 4, the asymptotic variance of $\sqrt{n}(\widehat{b}_\psi - \beta^*)_J$ is equal to

$$V_J = \frac{\mathbb{E}[\psi^2(\xi_i)]}{(\sigma_\xi^*)^2 \mathbb{E}\left[\frac{1}{(\sigma_\xi^*)^4}\psi'(\xi_i)\right]^2} \cdot (\Theta_x)_{JJ}.$$

We simply need to note that

$$(\Theta_x)_{JJ} = \left(\mathbb{E}\Big[\Big((x_i)_J - \mathbb{E}\left[(x_i)_J \mid (x_i)_{J^c}\right]\Big)\Big((x_i)_J - \mathbb{E}\left[(x_i)_J \mid (x_i)_{J^c}\right]\Big)\Big]\right)^{-1},$$

so it suffices to prove the equivalence of the terms

$$V_1 := \left(\mathbb{E}\left[\left(\frac{f'_{\sigma_\xi^*}(\epsilon_i)}{f_{\sigma_\xi^*}(\epsilon_i)}\right)^2\right]\right)^{-1}, \qquad \text{and}$$
$$V_2 := \frac{\mathbb{E}\left[\psi^2\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)\right]}{\mathbb{E}\left[\frac{1}{\sigma_\xi^*}\psi'\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)\right]^2},$$

where $f_{\sigma_\xi^*}$ denotes the pdf of ϵ_i. Taking f to be the pdf of $\frac{\epsilon_i}{\sigma_\xi^*}$, we have

$$f_{\sigma_\xi^*}(t) = \frac{1}{\sigma_\xi^*} f\left(\frac{t}{\sigma_\xi^*}\right), \quad \text{and} \quad f'_{\sigma_\xi^*}(t) = \frac{1}{(\sigma_\xi^*)^2} f'\left(\frac{t}{\sigma_\xi^*}\right),$$

so

$$V_1 = \left(\mathbb{E}\left[\frac{1}{(\sigma_\xi^*)^2}\left(\frac{f'\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}{f\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}\right)^2\right]\right)^{-1}.$$

Furthermore, differentiating the equation $\psi(t) = \frac{f'(t)}{f(t)}$, we have

$$\psi'(t) = \frac{f(t)f''(t) - (f'(t))^2}{(f(t))^2},$$

so

$$V_2 = \frac{\mathbb{E}\left[\left(\frac{f'\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}{f\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}\right)^2\right]}{\left(\mathbb{E}\left[\frac{1}{\sigma_\xi^*}\cdot\frac{f\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)f''\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)-\left(f'\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)\right)^2}{\left(f\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)\right)^2}\right]\right)^2}.$$

Furthermore, the square root of the term in the denominator is equal to

$$\begin{aligned}\frac{1}{\sigma_\xi^*}\cdot\mathbb{E}\left[\frac{f(\epsilon)f''(\epsilon)-f'(\epsilon)f'(\epsilon)}{f(\epsilon)^2}\right] &= \int_{-\infty}^{\infty} f''(t)dt - \int_{-\infty}^{\infty}\frac{f'(t)f'(t)}{f(t)}dt \\ &= [f'(t)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\frac{f'(t)f'(t)}{f(t)}dt \\ &= -\mathbb{E}\left[\left(\frac{f'\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}{f\left(\frac{\epsilon_i}{\sigma_\xi^*}\right)}\right)^2\right],\end{aligned}$$

from which we conclude that $V_1 = V_2$. Thus, we have $V_J = \overline{V}$ as well, implying the desired property of asymptotic efficiency.

Appendix G: Proof of Theorem 6

The proof follows in a straightforward manner from Theorem 4, which establishes the weak convergence statement

$$\sqrt{n}P_J(\widehat{b}_\psi - \beta^*) \xrightarrow{d} \frac{\sqrt{\mathbb{E}[\psi^2(\xi_i)]}}{\sigma_\xi^* A(\psi)}\cdot((\Theta_x)_{JJ})^{1/2}\,Z,$$

where $Z \sim N(0, I_m)$. Rearranging, we have

$$\frac{\sqrt{n}\sigma_\xi^* A(\psi)}{\sqrt{\mathbb{E}[\psi^2(\xi_i)]}}\cdot((\Theta_x)_{JJ})^{-1/2}\cdot P_J(\widehat{b}_\psi - \beta^*) \xrightarrow{d} Z.$$

We then use the following lemma:

Lemma 12. *Under the assumptions of the theorem, we have*

$$\frac{\widehat{\sigma}\widehat{A}(\psi)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \psi^2\left((y_i - x_i^T\widehat{\beta})/\widehat{\sigma}\right)}}\cdot\frac{\sqrt{\mathbb{E}[\psi^2(\xi_i)]}}{\sigma_\xi^* A(\psi)}\cdot\left(\widehat{\Theta}_{JJ}\right)^{-1/2}\cdot((\Theta_x)_{JJ})^{1/2} \xrightarrow{P} 1.$$

Proof. Note that it suffices to show the following convergence results:

$$\frac{\widehat{\sigma}\widehat{A}(\psi)}{\sigma_\xi^* A(\psi)} \xrightarrow{P} 1, \tag{G.1}$$

$$\frac{\mathbb{E}[\psi^2(\xi_i)]}{\frac{1}{n}\sum_{i=1}^n \psi^2\left((y_i - x_i^T\widehat{\beta})/\widehat{\sigma}\right)} \xrightarrow{P} 1, \tag{G.2}$$

$$\left(\widehat{\Theta}_{JJ}\right)^{-1}(\Theta_x)_{JJ} \xrightarrow{P} 1, \tag{G.3}$$

since we may combine the statements via Slutsky's theorem to obtain the desired result. Convergence results (G.1) and (G.3) are direct consequences of Lemma 9 and condition (4.6) of Theorem 4, under the assumed sample size scaling. To obtain the convergence result (G.2), we may use a parallel argument to the one employed to bound term B in the proof of Lemma 9. The only difference is that we use a Taylor expansion of ψ^2 rather than ψ'. Note that we have assumed $(\psi^2)''$ to be bounded. Since

$$(\psi^2)' = 2\psi\psi',$$

the terms we need to control replace B_1 and B_2 in inequality (D.12) by the quantities

$$B_1' := \left\| \frac{1}{n}\sum_{i=1}^n \psi(\xi_i)\psi'(\xi_i)x_i \right\|_\infty, \quad B_2' := \left| \frac{1}{n}\sum_{i=1}^n \psi(\xi_i)\psi'(\xi_i)\xi_i \right|.$$

As in the proof of Lemma 9, these terms may be bounded w.h.p. using Lemma 7 and Chebyshev's inequality. □

Hence, by Slutsky's theorem, we also have

$$\frac{\sqrt{n}\widehat{\sigma}\widehat{A}(\psi)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \psi^2\left((y_i - x_i^T\widehat{\beta})/\widehat{\sigma}\right)}} \cdot \left(\widehat{\Theta}_{JJ}\right)^{-1/2} \cdot P_J(\widehat{b}_\psi - \beta^*) \xrightarrow{d} Z.$$

Combined with equation (4.11), we then have

$$\lim_{n,p,k\to\infty} \mathbb{P}\left(\frac{\sqrt{n}\widehat{\sigma}\widehat{A}(\psi)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \psi^2\left((y_i - x_i^T\widehat{\beta})/\widehat{\sigma}\right)}} \cdot \left(\widehat{\Theta}_{JJ}\right)^{-1/2} \cdot P_J(\widehat{b}_\psi - \beta^*) \in \mathcal{B}_{\alpha,J} \right) = 1 - \alpha.$$

Rearranging the argument inside the probability expression yields the desired result.

Appendix H: Additional useful lemmas

We begin with a lemma giving a concentration inequality for the median-of-means estimator:

Lemma 13. *[Bubeck et al. [12, Lemma 2]] Let $0 < \delta < 1$ and $0 < \epsilon \leq 1$, and $n \geq 16 \log\left(\frac{1}{\delta}\right) + 2$. Suppose $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}\left[|X_i - \mu|^{1+\epsilon}\right] = v$. Let $K = \left\lfloor 8 \log\left(\frac{e^{1/8}}{\delta}\right) \wedge \frac{n}{2} \right\rfloor$. Then with probability at least $1 - \delta$,*

$$|\widehat{\mu}_{MoM} - \mu| \leq (12v)^{\frac{1}{1+\epsilon}} \left(\frac{16 \log(e^{1/8}/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

In particular, taking $\delta = ce^{-c'n}$, where c and c' are functions of ϵ, v, and μ, we have $\frac{\mu}{2} \leq \widehat{\mu}_{MoM} \leq \frac{3\mu}{2}$ with probability at least $1 - c\exp(-c'n)$.

The following lemma concerns quadratic forms of a matrix computed with respect to sparse vectors:

Lemma 14. *[Loh and Wainwright [54, Lemma 12]] For a fixed matrix $\Gamma \in \mathbb{R}^{p \times p}$, parameter $s \geq 1$, and tolerance $\delta > 0$, suppose we have the deviation condition*

$$|v^T \Gamma v| \leq \delta, \qquad \forall v \in \mathbb{R}^p \text{ s.t. } \|v\|_0 \leq 2s \text{ and } \|v\|_2 \leq 1. \tag{H.1}$$

Then

$$|v^T \Gamma v| \leq 27\delta \left(\|v\|_2^2 + \frac{\|v\|_1^2}{s}\right), \qquad \forall v \in \mathbb{R}^p.$$

We now have a useful lemma concerning a Gaussian approximation of maxima.

Lemma 15. *[Chernozhukov et al. [16, Corollary 2.1]] Suppose $\{y_i\}_{i=1}^n \subseteq \mathbb{R}^p$ are independent, zero-mean random vectors such that for some constants $c_1, C_1 > 0$ and a sequence of constants $B_n \geq 1$, the following conditions hold uniformly in $1 \leq j \leq p$:*

$$c_1 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[y_{ij}^2] \leq C_1, \qquad \text{and} \tag{H.2}$$

$$\max_{\ell=1,2} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\frac{|y_{ij}|^{2+\ell}}{B_n^\ell}\right) \right\} + \max_{1 \leq i \leq n} \mathbb{E}\left[\exp\left(\frac{|y_{ij}|}{B_n}\right)\right] \leq 4. \tag{H.3}$$

Also suppose $\frac{B_n^2 (\log(pn))^7}{n} \leq C_2 n^{-c_2}$ for some constants $c_2, C_2 > 0$. Then there exist constants $c, C > 0$, depending only on c_1, c_2, C_1, and C_2, such that

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\max_{1 \leq j \leq p} Y_j \leq t\right) - \mathbb{P}\left(\max_{1 \leq j \leq p} Z_j \leq t\right) \right| \leq Cn^{-c},$$

where $Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i$, and $Z \sim \mathcal{N}(0, \mathbb{E}[y_i y_i^T])$ is a multivariate Gaussian vector with components $\{Z_j\}_{j=1}^p$.

Lemma 15 leads to several useful concentration inequalities. In particular, using a union bound together with standard Gaussian tail bounds, we have

$$\mathbb{P}\left(\max_{1\leq j\leq p} Z_j \leq t\right) \leq 2p\exp\left(-\frac{t^2}{2\sigma_Y^2}\right),$$

where $\sigma_Y = \max_{1\leq j\leq p}\mathbb{E}[y_{ij}^2]$, so we have

$$\max_{1\leq j\leq p} Y_j \leq 2\sigma_Y\sqrt{\log p},$$

with probability at least $1 - Cn^{-c} - p^{-c'}$.

Lemma 16. *[Chernoff bound for binomials [76, Theorem 2.3.1]] Let* $X_i \overset{i.i.d.}{\sim}$ *Bernoulli*(p)*, and let* $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$*. For any* $t > p$*, we have*

$$\mathbb{P}(\widehat{\mu} \geq t) \leq \exp\left(-np + nt\log(ep/t)\right).$$

Lemma 17. *[Moments of spherically symmetric distributions [24]] Suppose* $X = RU$*, where* U *is uniformly distributed on a* p*-dimensional unit sphere and* R *is a scalar random variable. For even integers* $2s_i \geq 0$*, where* $1 \leq i \leq p$*, we have*

$$\mathbb{E}\left[\prod_{i=1}^p X_i^{2s_i}\right] = \mathbb{E}[R^{2s}]\pi^{-p/2}\frac{\Gamma(p/2)}{\Gamma(p/2+s)}\prod_{i=1}^p\Gamma\left(\frac{1}{2}+s_i\right),$$

where $s = \sum_{i=1}^p s_i$*.*

Lemma 18. *[Multivariate Lindeberg-Feller CLT [30]] Suppose* $\{x_n\}_{n\geq 1}$ *are independent random vectors such that all mixed third moments are finite. Let* $\mathbb{E}[x_i] = \mu_i$ *and* $\mathrm{Var}[x_i] = Q_i$*, and define*

$$\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \mu_i, \qquad \text{and} \qquad \bar{Q}_n = \frac{1}{n}\sum_{i=1}^n Q_i.$$

Suppose

$$\lim_{n\to\infty}\bar{Q}_n = Q \succeq 0,$$

and for every i*,*

$$\lim_{n\to\infty}\left(\bar{Q}_n\right)^{-1}\frac{Q_i}{n} = 0.$$

Then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i - \bar{\mu}_n\right) \xrightarrow{d} \mathcal{N}(0, Q).$$

Lemma 19. *Suppose* X *and* Y *are independent random variables, where* X *has a symmetric, unimodal density. Then*

$$\mathrm{MAD}(X) \leq \mathrm{MAD}(X+Y).$$

Proof. Without loss of generality, we assume the distribution of X is symmetric around 0. Note that it suffices to show that

$$\mathbb{P}\Big(\,|X+Y-\mathrm{med}(X+Y)|\geq \mathrm{MAD}(X)\Big)\geq \frac{1}{2}.$$

Indeed, we will show this inequality holds for any fixed value $Y=y$:

$$\mathbb{P}\Big(\,|X+y-M|\geq \mathrm{MAD}(X)\Big)\geq \frac{1}{2}, \tag{H.4}$$

where we have denoted $M=\mathrm{med}(X+Y)$. We may then write the left-hand probability as

$$\mathbb{P}\left(X\geq M-y+\mathrm{MAD}(X)\right)+\mathbb{P}\left(X\leq M-y-\mathrm{MAD}(X)\right):=I+II.$$

Note that:

1. If $M-y\geq \mathrm{MAD}(X)$, we have
$$II\geq \mathbb{P}(X\leq 0)\geq \frac{1}{2}.$$
2. If $M-y\leq -\mathrm{MAD}(X)$, we have
$$I\geq \mathbb{P}(X\geq 0)\geq \frac{1}{2}.$$
3. Otherwise, suppose $0\leq M-y<\mathrm{MAD}(X)$ (the case when $M-y$ is negative is analogous). We have the bound
$$\begin{aligned}
&\Big(I+II\Big)-\Big(\mathbb{P}(X\geq \mathrm{MAD}(X))+\mathbb{P}(X\leq -\mathrm{MAD}(X)\Big)\\
&\quad=-\mathbb{P}\Big(\,\mathrm{MAD}(X)\leq X<\mathrm{MAD}(X)+M-y\Big)\\
&\qquad+\mathbb{P}\Big(-\mathrm{MAD}(X)<X\leq -\mathrm{MAD}(X)+M-y\Big)\\
&\quad=-\mathbb{P}\Big(\,\mathrm{MAD}(X)\leq X<\mathrm{MAD}(X)+M-y\Big)\\
&\qquad+\mathbb{P}\Big(\,\mathrm{MAD}(X)-M+y\leq X<\mathrm{MAD}(X)\Big)\\
&\quad\geq 0,
\end{aligned}$$
where the final inequality comes from the assumption that the pdf of X is unimodal, hence is a nonincreasing function on the interval $\mathrm{MAD}(X)\pm(M-y)$. We conclude that
$$I+II\geq \mathbb{P}\Big(|X|\geq \mathrm{MAD}(X)\Big)\geq \frac{1}{2}$$
in this case, as well.

This establishes inequality (H.4). □

Acknowledgments

The author would like to thank Ezequiel Smucler for sharing the archaeological dataset used in the simulations. The author also thanks the AE and anonymous reviewers for thoughtful feedback which greatly improved the manuscript.

References

[1] Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics* **7** 226–248. MR3086417
[2] Avella-Medina, M. (2017). Influence functions for penalized M-estimators. *Bernoulli* **23** 3178–3196. MR3654803
[3] Bakshi, A. and Prasad, A. (2021). Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* 102–115.
[4] Bellec, P. C. and Zhang, C. H. (2019). De-biasing the Lasso with degrees-of-freedom adjustment. *arXiv preprint arXiv:1902.08885*.
[5] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324
[6] Belloni, A., Chernozhukov, V. and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097
[7] Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association* **70** 428–434. MR0386168
[8] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. MR1245941
[9] Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series* 113–133. MR1836557
[10] Bradic, J. (2015). Robustness in sparse linear models: Relative efficiency based on robust approximate message passing. *Electronic Journal of Statistics* **10** 2. MR3581957
[11] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 325–349. MR2815779
[12] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory* **59** 7711–7717. MR3124669
[13] Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* **45** 615–646. MR3650395

[14] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. In *Annales de l'IHP Probabilités et statistiques* **48** 1148–1185. MR3052407
[15] Cherapanamjeri, Y., Aras, E., Tripuraneni, N., Jordan, M. I., Flammarion, N. and Bartlett, P. L. (2020). Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*.
[16] 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
[17] Chichignoud, M., Lederer, J. and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the Lasso with optimality guarantees. *Journal of Machine Learning Research* **17** 1–20. MR3595165
[18] Depersin, J. (2020). A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*.
[19] Diakonikolas, I., Kong, W. and Stewart, A. (2019). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2745–2754. SIAM. MR3909639
[20] Elisseeff, A. and Pontil, M. (2003). Leave-one-out error and stability of learning algorithms with applications. *NATO Science Series Sub Series iii Computer and Systems Sciences* **190** 111–130.
[21] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81** 310–320.
[22] Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics* **42** 324. MR3189488
[23] Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 247–265. MR3597972
[24] Fang, K. T., Kotz, S. and Ng, K. W. (2018). *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC. MR1071174
[25] Fessler, J. A. (2006). Iterative methods for image reconstruction. In *IEEE International Symposium on Biomedical Imaging Tutorial*.
[26] Fradelizi, M., Guédon, O. and Pajor, A. (2014). Thin-shell concentration for convex measures. *Studia Mathematica* **223**. MR3268720
[27] Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M. and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *Annals of Applied Statistics* **13** 2065–2090. MR4037422
[28] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.
[29] Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics* **30** 583–616. MR1902900
[30] Greene, W. H. (2003). *Econometric Analysis*, 5. ed. Prentice Hall, Upper Saddle River, NJ.

[31] Guédon, O. and Milman, E. (2011). Interpolating thin-shell and sharp large-deviation estimates for lsotropic log-concave measures. *Geometric and Functional Analysis* **21** 1043. MR2846382
[32] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Statistics*. Wiley. MR2488795
[33] Hansen, B. E. (2017). *Econometrics*. Available at http://www.ssc.wisc.edu/~bhansen/econometrics.
[34] Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association* **69** 909–923. MR0461779
[35] Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research* **17** 1–40. MR3491112
[36] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35** 73–101. MR0161415
[37] Huber, P. J. (1981). *Robust Statistics*. Wiley New York. MR0606374
[38] Huber, P. J. (1983). Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association* **78** 66–72. MR0696850
[39] Jaeckel, L. A. (1971). Some flexible estimates of location. *The Annals of Mathematical Statistics* 1540–1552. MR0350951
[40] Jankova, J. and van de Geer, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics* **46** 2336–2359. MR3845020
[41] Janková, J. and van de Geer, S. (2021). De-Biased Sparse PCA: Inference for Eigenstructure of Large Covariance Matrices. *IEEE Transactions on Information Theory* **67** 2507–2527. MR4282369
[42] Janssens, K. H., Deraedt, I., Schalm, O. and Veeckman, J. (1998). Composition of 15–17th century archaeological glass vessels excavated in Antwerp, Belgium. In *Modern Developments and Applications in Microbeam Analysis* 253–267. Springer.
[43] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15** 2869–2909. MR3277152
[44] Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for gaussian designs. *The Annals of Statistics* **46** 2593–2622. MR3851749
[45] Jurečková, J. and Portnoy, S. (1987). Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point. *Communications in Statistics–Theory and Methods* **16** 2187–2199. MR0915457
[46] Karmalkar, S. and Price, E. (2019). Compressed sensing with adversarial sparse noise via L1 regression. In *Symposium on Simplicity in Algorithms*. MR3904995
[47] Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American*

Statistical Association **102** 1289–1299. MR2412550
[48] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning* 1885–1894. PMLR.
[49] Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. *Econometrica: Journal of the Econometric Society* 1333–1346. MR0584305
[50] Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American statistical Association* **77** 595–604. MR0675886
[51] Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Annals of Statistics* **48** 906–931. MR4102681
[52] Lepskii, O. V. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications* **35** 454–466. MR1091202
[53] Loh, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics* **45** 866–896. MR3650403
[54] Loh, P. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics* **40** 1637-1664. MR3015038
[55] Lugosi, G. and Mendelson, S. (2019). Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli* **25** 2075–2106. MR3961241
[56] Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics* **53** 44–53. MR2791951
[57] Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. J. Wiley. MR2238141
[58] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468
[59] Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics* **46** 2871–2903. MR3851758
[60] Nesterov, Y. (2007). Gradient methods for minimizing composite objective function CORE Discussion Papers No. 2007076, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).
[61] Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5** 99–135.
[62] Pensia, A., Jog, V. and Loh, P. (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*.
[63] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **4** 935–980. MR2836766
[64] Ronchetti, E. (1982). Robust testing in linear models: The infinitesimal

approach, PhD thesis, ETH Zurich. MR2632390

[65] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79** 871–880. MR0770281

[66] Salibian-Barrera, M., Willems, G. and Zamar, R. (2008). The fast-τ estimator for regression. *Journal of Computational and Graphical Statistics* **17** 659–682. MR2528241

[67] Sasai, T. and Fujisawa, H. (2020). Robust estimation with Lasso when outputs are adversarially contaminated. *arXiv preprint arXiv:2004.05990*.

[68] Serfling, R. and Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics & Probability Letters* **79** 1767–1773. MR2566751

[69] Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* **87** 439–450. MR1173809

[70] Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis* **111** 116–130. MR3630222

[71] Sun, Q., Zhou, W. X. and Fan, J. (2019). Adaptive Huber regression. *Journal of the American Statistical Association* 1–24. MR4078461

[72] Sur, P., Chen, Y. and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields* **175** 487–558. MR4009715

[73] van de Geer, S. (2019). On the asymptotic variance of the debiased Lasso. *Electronic Journal of Statistics* **13** 2970–3008. MR4010589

[74] van de Geer, S. and Stucky, B. (2016). χ^2-Confidence Sets in High-Dimensional Regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014* **11** 279. Springer. MR3616273

[75] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[76] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science* **47**. Cambridge University Press. MR3837109

[77] Wang, L. (2013). The $L1$ penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120** 135–151. MR3072722

[78] Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics* **25** 347–355. MR2380753

[79] Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107** 214–222. MR2949353

[80] Wang, L., Zheng, C., Zhou, W. and Zhou, W. X. (2020a). A new principle for tuning-free Huber regression. *Statistica Sinica*. MR4328856

[81] Wang, L., Peng, B., Bradic, J., Li, R. and Wu, Y. (2020b). A tuning-free robust and efficient approach to high-dimensional regression. *Journal*

of the American Statistical Association 1–44. MR4189748

[82] Welsh, A. H. and Ronchetti, E. (2002). A journey in single steps: Robust one-step M-estimation. *Journal of Statistical Planning and Inference* **103** 287–310. MR1896997

[83] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* **15** 642–656. MR0888431

[84] Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83** 406–413. MR0971366

[85] Yu, Y., Bradic, J. and Samworth, R. J. (2019). Confidence intervals for high-dimensional Cox models. *Statistica Sinica*.

[86] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19-35. MR2367824