Statistica Sinica Preprint No: SS-2022-0375

Title	Sparse and Debiased Adaptive Huber Regression in Distributed Data: Aggregated and Communication-Efficient Approaches
Manuscript ID	SS-2022-0375
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0375
Complete List of Authors	Wei Ma, Junzhuo Gao, Lei Wang and Heng Lian
Corresponding Authors	Lei Wang
E-mails	lwangstat@nankai.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Sparse and debiased adaptive Huber regression in distributed data: aggregated and communication-efficient approaches

Wei Ma[1], Junzhuo Gao[1], Lei Wang[1] and Heng Lian[2]

[1]*Nankai University and* [2]*City University of Hong Kong*

Abstract: Distributed estimation and statistical inference for linear models have drawn much attention recently, but few studies focus on robust learning in the presence of heavy-tailed/asymmetric errors and high-dimensional covariates. Based on adaptive Huber regression to achieve the bias-robustness tradeoff, two classes of sparse and debiased lasso estimators are proposed using aggregated and communication-efficient approaches. To be specific, an aggregated ℓ_1-penalized and a multi-round ℓ_1-penalized communication-efficient adaptive Huber estimators are respectively proposed in the first stage to handle the distributed data with high-dimensional covariates and heavy-tailed/asymmetric errors. To correct the biases caused by the lasso penalty, a unified debiasing framework based on the decorrelated score equations is considered in the second stage. In the third stage, hard-thresholding is used to produce the sparse and debiased lasso estimators. The convergence rates and asymptotic properties of the proposed two estimators are established. The finite-sample performance is studied through simulations and a real data application to Commu-

Address for correspondence: Lei Wang, School of Statistics and Data Science, KLM-DASR, LEBPS and LPMC, Nankai University. E-mail: lwangstat@nankai.edu.cn.

nities and Crime Data Set is also presented to illustrate the validity and feasibility of the proposed estimators.

Key words and phrases: Asymptotic normality; convergence rates; debiased lasso; decorrelated score; thresholding; multi-round.

1. Introduction

With the advancement of science and technology, massive data with large sample size and high-dimensional covariates are stored independently in many different sites, and referred to as *distributed data*. Due to the limitation of storage, computing capability and personal privacy in practice, traditional methods by processing all data simultaneously in one central site are not practical for distributed data. To overcome this problem, distributed estimation and statistical inference have drawn much attention in modern statistical learning recently. The aggregated/divide-and-conquer (Chen and Xie, 2014; Battey et al., 2018; Volgushev et al., 2019) and communication-efficient surrogate likelihood (CSL; Wang et al., 2017; Jordan et al., 2019) are the two well-known methods for dealing with distributed data. The aggregated method conducts local estimators independently and obtains a final estimator via one round communication between the local sites. Unfortunately, a small number sites condition is required to achieve the same convergence rate as using the entire data. On the other hand, the CSL method optimizes a surrogate loss on the central site utilizing the gradient information from all local sites, which is called as communication-efficient since only the gradient information is communicated between the central and local sites at each round. Compared with the aggregated method, the CSL method achieves the optimal convergence rate and relieves the restriction on the number of local sites. However, the majority of existing work focuses on the least squares loss (Lee et al., 2017; Battey et al., 2018; Jordan et al., 2019; Zhao et al., 2020; Fan et al., 2021; Duan et al., 2022), which is not resistant to heavy-tailed/asymmetric errors or outliers, and little knowledge is available about statistical inference for high-dimensional robust regression.

In practice, since distributed data are often collected from different environments/sources with low quality or high level of noise, e.g., misjudgment in functional magnetic resonance imaging studies (Eklund et al., 2016) and large kurtosis values of the gene expression levels (Wang et al., 2015), directly applying the existing distributed methods may lead to large bias and erroneous statistical inference (Chen et al., 2020; Tan et al., 2022), thus it is crucial to analyze the distributed and high-dimensional data robustly and rapidly with theoretical guarantee. In the literature, to overcome both the high dimensionality and heavy-tailed/asymmetric errors, ℓ_1-penalized Huber regression is always considered (Po-Ling Loh, 2018; Han et al., 2022) and then is improved by adaptive Huber regression with a data-driven robustification parameter rather than a fixed one (Fan et al., 2017; Sun et al., 2020; Wang et al., 2021) to balance the tradeoff between bias and robustness. Recently, Luo et al. (2022) studied the ℓ_1-penalized communication-efficient adaptive Huber estimator, but did not obtain a tractable limiting distribution due to the biases caused by the lasso penalty. On the other hand, to produce sparse and asymptotically unbiased estimators for high-dimensional linear and quantile regression models with distributed data, Lee et al. (2017) and Zhao et al. (2020) proposed aggregating the debiased lasso estimators from the local sites and then applying thresholding strategies, which can not be applied to the Huber loss directly.

In this paper, we consider adaptive Huber regression as a robust alternative to the least squares regression, and our goal is to develop two classes of sparse and debiased lasso estimation and statistical inference methods. To the best of our knowledge, these problems have not been investigated due to the following reasons. First, different from the aggregated estimators in Lee et al. (2017) and Zhao et al. (2020), it is difficult to carry out the debiased lasso estimation and statistical inference for adaptive Huber regression, since its loss function is non-smooth and depends on a data-driven robustification parameter. Second, there is no literature studying the debiased lasso estimation and statistical inference for ℓ_1-penalized CSL estimation, which hinders its application in practice. Moreover, since Luo et al. (2022) only considered the first site as the central site for solving the CSL optimization problems and the others just for evaluating gradients, the computing power is not fully utilized and the estimation stability can be improved.

Based on the aggregated and communication-efficient approaches, two classes of sparse and debiased adaptive Huber estimators are respectively proposed based on the following three stages. **(i)** An aggregated ℓ_1-penalized adaptive Huber estimator as well as a multi-round ℓ_1-penalized communication-efficient adaptive Huber estimator are proposed respectively in the first stage. Although the above two ℓ_1-penalized adaptive Huber estimators are sparse, their limiting distributions are untractable due to the biases. **(ii)** A unified debiasing lasso framework based on the decorrelated score equations is proposed in the second stage and then we establish asymptotic normality of estimators with explicit formulas of asymptotic covariance matrices, which can be used to construct confidence intervals or test statistical hypotheses. **(iii)** Due to the debiasing and/or aggregated procedures, the debiased lasso estimators in the second stage are not sparse such that hard-thresholding is necessary to produce the sparse and debiased lasso estimators in the third stage. After these three stages, we show that the proposed two classes of sparse and debiased lasso estimators have the same statistical accuracy as using the entire samples under some regular conditions and have good finite-sample performance in simulation studies.

The rest of the article is organized as follows. In Sections 2 and 3, we introduce the sparse aggregated and communication-efficient debiased adaptive Huber estimators and then investigate their asymptotic properties, respectively. Extensive simulation results are provided in Section 4. An application to the Communities and Crime Data Set is illustrated in Section 5. Some conclusions are given in Section 6. All proofs of Theorems and Corollaries are relegated in the Supplementary Material.

2. Sparse and debiased lasso estimator via aggregation

We adopt the following notations throughout the paper. For a vector $\boldsymbol{u} = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$, denote $\|\cdot\|_q$ $(1 \le q \le \infty)$ as the ℓ_q-norm in $\mathbb{R}^p$: $\|\boldsymbol{u}\|_q = (\sum_{j=1}^p |u_j|^q)^{1/q}$, $\|\boldsymbol{u}\|_\infty = \max_{1\le j\le p} |u_j|$ and $\|\boldsymbol{u}\|_0 = |\operatorname{supp}(\boldsymbol{u})|$, where $\operatorname{supp}(\boldsymbol{u}) = \{j : u_j \neq 0, j = 1, \cdots, p\}$ and $|\cdot|$ denotes the absolute value for a vector or the cardinality for a set. Use u_j and $\boldsymbol{u}_{-j}$ to represent the jth element and the remaining vector when the jth element is removed, respectively. Denote $a_N \lesssim b_N$ $(a_N \gtrsim b_N)$ if a_N is less than (greater than) b_N up to a constant; $a_N \asymp b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$.

2.1 Aggregated adaptive Huber estimator

Assume N independent and identically distributed (i.i.d.) observations $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ are collected from the following linear regression model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \ldots, N, \tag{2.1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ with $x_{i,1} \equiv 1$ is a p-dimensional vector of covariates, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the true parameter, ε_i is a zero-mean error term independent of $\boldsymbol{x}_i$ with a finite variance σ^2 but can be heavy-tailed and asymmetrically distributed. In this paper, we consider high-dimensional linear models under sparsity, i.e., $\|\boldsymbol{\beta}^*\|_0 = s$, and the global ℓ_1-penalized adaptive Huber estimator can be obtained as follows:

$$\hat{\boldsymbol{\beta}}_{\tau_N} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \{L_{\tau_N}(\boldsymbol{\beta}) + \lambda_N \|\boldsymbol{\beta}\|_1\}, \tag{2.2}$$

where $L_\tau(\boldsymbol{\beta}) = N^{-1} \sum_{i=1}^N \ell_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ with $\ell_\tau(s) = (s^2/2) I(|s| \leq \tau) + (\tau|s| - \tau^2/2) I(|s| > \tau)$, the global robustification parameter $\tau_N > 0$ is allowed to scale with the sample size and parameter dimension, i.e., $\tau_N \asymp \sigma\sqrt{N/\log p}$, and $\lambda_N > 0$ is the global regularization parameter. Under model (2.1), define $\boldsymbol{\beta}_\tau^* \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} E\{L_\tau(\boldsymbol{\beta})\}$ for any τ. Wang et al. (2021) and Han et al. (2022) showed that the slope parts of $\boldsymbol{\beta}_\tau^*$ and $\boldsymbol{\beta}^*$ are the same but the intercept terms have a constant difference depending on τ under some regular conditions, i.e., $\boldsymbol{\beta}_{\tau,-1}^* = \boldsymbol{\beta}_{-1}^*$ and $\beta_{\tau,1}^* = \beta_1^* + \alpha_\tau$. Statistical properties of $\hat{\boldsymbol{\beta}}_{\tau_N}$ and $\boldsymbol{\beta}_{\tau_N}^*$ have been thoroughly studied by Fan et al. (2017) and Sun et al. (2020), and they showed the estimator $\hat{\boldsymbol{\beta}}_{\tau_N}$ with $\tau_N \asymp \sigma\sqrt{N/\log p}$ achieves the optimal tradeoff between estimation error and approximation bias.

While in the distributed setting, it is impractical to store the entire dataset for computing the global estimator based on (2.2) due to the constraint of storage capacity and privacy protocols. In this paper, we assume the entire N observations are stored on M different sites independently and identically, i.e., the mth site has n_m samples such that $N = \sum_{m=1}^{M} n_m$ for $1 \leq m \leq M$. Without loss of generality, we consider $n_1 = \ldots = n_M = n = N/M$ and refer to n as the local sample size. Let $\mathcal{I}_m \subset \{1, \ldots, N\}$ be the index set corresponding to the elements of the mth site, satisfying $\cup_{m=1}^{M} \mathcal{I}_m = \{1, \ldots, N\}$ and $\mathcal{I}_m \cap \mathcal{I}_\ell = \varnothing$ for all $1 \leq m \neq \ell \leq M$. The mth local ℓ_1-penalized adaptive Huber estimator can be obtained by

$$\hat{\boldsymbol{\beta}}_{m,\tau_n} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \{L_{m,\tau_n}(\boldsymbol{\beta}) + \lambda_m \|\boldsymbol{\beta}\|_1\}, \tag{2.3}$$

where $L_{m,\tau}(\boldsymbol{\beta}) = n^{-1} \sum_{i\in\mathcal{I}_m} \ell_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ is the mth local adaptive Huber loss function, $\tau_n > 0$ and $\lambda_m > 0$ are the mth local robustification and regularization parameters, respectively. It should be pointed out the optimal $\tau_n \asymp \sigma\sqrt{N/(M\log p)}$ differs from τ_N, since the local site can only access to $n = N/M$ samples. Subsequently, the aggregated ℓ_1-penalized adaptive Huber estimator of $\boldsymbol{\beta}^*$ is defined as follows:

$$\bar{\boldsymbol{\beta}}_{\tau_n} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}_{m,\tau_n}, \tag{2.4}$$

where $\hat{\boldsymbol{\beta}}_{m,\tau_n}$ is obtained from (2.3). For ease of notations, we omit τ_n and τ_N of the estimators in the rest of this paper, but we should remember that they are τ_n or/and τ_N specifically.

2.2 Thresholding aggregated debiased lasso estimator

Due to the lasso penalty in (2.3) and the aggregation step, $\bar{\boldsymbol{\beta}}$ is non-sparse and generally biased such that its asymptotic distribution is difficult to derive. Our first goal is to propose a sparse aggregated debiased lasso (SADL) adaptive Huber estimator for distributed data.

Without loss of generality, we focus on the estimation and inference of β_j^*, the jth component of $\boldsymbol{\beta}^*$ for $1 \le j \le p$. Motivated by Ning and Liu (2017), the decorrelated score estimating equation for β_j based on the mth site is given as follows:

$$\frac{1}{n}\sum_{i\in\mathcal{I}_m}(-x_{i,j} + \boldsymbol{x}_{i,-j}^{\top}\hat{\boldsymbol{\gamma}}_j^{(m)})\psi_{\tau_n}(y_i - \boldsymbol{x}_{i,-j}^{\top}\hat{\boldsymbol{\beta}}_{m,-j} - x_{i,j}\beta_j) = 0, \tag{2.5}$$

where $\psi_\tau(s) = \nabla_s \ell_\tau(s)$, $\hat{\boldsymbol{\beta}}_{m,-j} \equiv \{\hat{\beta}_{m,k} : k \ne j, 1 \le k \le p\}$, $\boldsymbol{x}_{i,-j} \equiv \{x_{i,k} : k \ne j, 1 \le k \le p\}$, $\hat{\boldsymbol{\gamma}}_j^{(m)}$ is a consistent estimator of $\boldsymbol{\gamma}_j^* \equiv \mathrm{argmin}_{\boldsymbol{\gamma}_j\in\mathbb{R}^{p-1}} E(x_{i,j} -$

$\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j)^2$ and $\hat{\boldsymbol{\gamma}}_j^{(m)}$ can be obtained by

$$\hat{\boldsymbol{\gamma}}_j^{(m)} \in \underset{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \Big\{ \frac{1}{2n} \sum_{i \in \mathcal{I}_m} (x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j)^2 + \omega_{jm} \|\boldsymbol{\gamma}_j\|_1 \Big\}, \tag{2.6}$$

with the regularization parameter ω_{jm}. Actually, (2.5) can be viewed as the residuals of the projection of the score function for β_j onto the closure of the linear span of the score function for the other parameters. The orthogonal property makes sure that the asymptotic normality of the estimator obtained by (2.5) will not be influenced by the slower convergence rate of $\hat{\boldsymbol{\beta}}_{m,-j}$. By replacing $E[(x_{i,j} - \boldsymbol{\gamma}_j^{*\top} \boldsymbol{x}_{i,-j}) I(|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{\tau_n}^*| \le \tau_n)]$ with its empirical counterpart, we get the jth element of the debiased estimator based on the mth site:

$$\hat{\beta}_{m,j}^{\mathbf{d}} = \hat{\beta}_{m,j} - \frac{\sum_{i \in \mathcal{I}_m} (-x_{i,j} + \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)}) \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m)}{\sum_{i \in \mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)}) \times n^{-1} \sum_{i \in \mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m| \le \tau_n)}.$$

Let $\hat{\boldsymbol{\beta}}_m^{\mathbf{d}} = (\hat{\beta}_{m,1}^{\mathbf{d}}, \cdots, \hat{\beta}_{m,p}^{\mathbf{d}})^\top$ and we propose to aggregate the debiased lasso adaptive Huber estimators among the M local sites as

$$\bar{\boldsymbol{\beta}}^{\mathbf{d}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m^{\mathbf{d}}.$$

Although $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ is an asymptotically unbiased estimator, it is not sparse due to the debiasing and averaging procedures. Therefore, hard-thresholding should be applied as a post-processing step to produce a sparse estimator. Given the threshold level ν, we define the hard-thresholding operator $\mathcal{T}_\nu$ such that the jth element of $\mathcal{T}_\nu(\boldsymbol{\beta})$ is $\mathcal{T}_\nu(\beta_j) = \beta_j I\{|\beta_j| \ge \nu\}$ for $1 \le j \le p$. Finally, we get the SADL adaptive Huber estimator

$$\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) = (\mathcal{T}_\nu(\bar{\beta}_1^{\mathbf{d}}), \cdots, \mathcal{T}_\nu(\bar{\beta}_p^{\mathbf{d}}))^\top, \tag{2.7}$$

and we will show that $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ has the same convergence rate as the global adaptive Huber estimator in ℓ_2 error with $\nu \asymp \sqrt{\log p/N}$. Denote $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$ as the debiased lasso adaptive Huber estimator using the entire data.

2.3 Theoretical results

(C1) (i) The error term ε_i's are i.i.d. and independent with $\boldsymbol{x}_i$; (ii) ε_i follows an absolutely continuous random variable with zero-mean and finite variance σ^2; (iii) For any $\tau > 0$, the function $\alpha \mapsto E\{\ell_\tau(\varepsilon - \alpha)\}$ has a unique minimizer $\alpha_\tau = \operatorname{argmin}_{\alpha\in\mathbb{R}} E\{\ell_\tau(\varepsilon - \alpha)\}$ and satisfies $P(|\varepsilon - \alpha_\tau| \le \tau) > 0$.

(C2) (i) The covariate $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top \in \mathbb{R}^p$ with $x_{i,1} \equiv 1$ is bounded and has bounded kurtosis uniformly, i.e., for some constant $B \ge 1$, $\max_{1\le i\le N} \|\boldsymbol{x}_i\|_\infty \le B$ and $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} E(\boldsymbol{z}_i^\top \boldsymbol{u})^4 < \infty$ with $\boldsymbol{z}_i = \boldsymbol{\Xi}^{-1/2}\boldsymbol{x}_i$, $\boldsymbol{\Xi} = (\Xi_{jk})_{1\le j,k\le p} = E(\boldsymbol{x}_i\boldsymbol{x}_i^\top)$ and $\mathbb{S}^{p-1} = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_2 = 1\}$; (ii) For any $p \times p$ positive semi-definite matrix $\boldsymbol{A} = [A_{jk}]_{1\le j,k\le p}$, denote $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ as the smallest and largest eigenvalues of $\boldsymbol{A}$ respectively. Assume $0 < C_{\min} \le \lambda_{\min}(\boldsymbol{\Xi}) \le \lambda_{\max}(\boldsymbol{\Xi}) \le C_{\max} < \infty$ and $\max_{1\le j\le p} \Xi_{jj} = O(1)$.

(C3) (i) $\boldsymbol{\beta}^*$ is sparse with sparsity s and $s^2 M \log p/N = o(1)$; (ii) $\boldsymbol{\Omega}$ is the inverse matrix of $\boldsymbol{\Xi}$. For any $1 \le j \le p$, $\max_{1\le j\le p} \|\boldsymbol{\Omega}_j\|_0 \le s_1$ for some positive integer s_1, where $\boldsymbol{\Omega}_j$ is the jth row of $\boldsymbol{\Omega}$; (iii) $\max_{i,j} |\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*| \le B$, $s \asymp s_1$ for notational simplicity.

Condition (C1) is often used in robust regression (Han et al., 2022; Luo et al., 2022) and the errors satisfied Condition (C1) include many distributions, such as normal distribution, Chi-square distribution, Student's t-distribution with degrees of freedom greater than 2. (iii) in Condition (C1) ensures that the slope parts of $\boldsymbol{\beta}_\tau^*$ and $\boldsymbol{\beta}^*$ are the same but the intercept terms have a constant difference depending on τ (Proposition 5, Wang et al., 2021). Unlike the Gaussian/sub-Gaussian covariates assumption, Condition (C2) requires a bounded assumption on covariates due to technical barriers, this assumption is widely applied in many literatures, see van de Geer et al. (2014), Zhao et al. (2020), Wang et al. (2021) and Lv and Lian (2022). The compatibility condition is satisfied from the restriction on the eigenvalues (Lee et al., 2017; Battey et al., 2018). Condition (C3) is a common regular condition for the high-dimensional regression models. For example, $s^2 M \log p/N = o(1)$ is a standard sparsity assumption (Han et al., 2022) and $\max_{i,j} |\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*| \le B$ makes sure that the strongly bounded assumption holds.

Theorem 1. *Under Conditions (C1)-(C3), if* $\tau_n \asymp \sigma\sqrt{N/(M\log p)}$, $\lambda_m \asymp$

$\sqrt{M\log p/N}$ *uniformly in* m *and* $\omega_{jm} \asymp \sqrt{M\log p/N}$ *uniformly in* m *and* j, *with* $\log M = O(\log p)$, *then we have*

$$\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{-1} - \boldsymbol{\beta}^*_{-1}\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big), |\bar{\beta}^{\mathbf{d}}_1 - \beta^*_1| = O_p\Big(\sqrt{\frac{M\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big).$$

In addition, if $E(|\varepsilon|^3) < \infty$, *we have*

$$\|\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big).$$

Remark 1. For the intercept term, the convergence rate of $|\bar{\beta}^{\mathbf{d}}_1 - \beta^*_1|$ is slower than that of the slope parts $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{-1} - \boldsymbol{\beta}^*_{-1}\|_\infty$. The reason is that $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ only using τ_n is not large enough to correct the approximation bias $|\alpha_{\tau_n}| \lesssim \sqrt{M\log p/N}$ of the intercept term. Furthermore, given the condition $E(|\varepsilon|^3) < \infty$, we can show that the approximation bias $|\alpha_{\tau_n}| \lesssim M\log p/N$, which is negligible compared with $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{-1} - \boldsymbol{\beta}^*_{-1}\|_\infty$, and thus $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty$ attains the same convergence rate of $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{-1} - \boldsymbol{\beta}^*_{-1}\|_\infty$. For the golden standard estimator $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$ using the entire data, we know that $\|\hat{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty \lesssim \sqrt{\log p/N}$. When $M = O(\sqrt{N/(s^3\log p)})$, it can be seen that $O_p(s^{3/2}M\log p/N)$ becomes $O_p(\sqrt{\log p/N})$, then $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty = O_p(\sqrt{\log p/N})$. Thus, $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ attains the same statistical accuracy as $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$ in term of ℓ_∞ error. The uniform convergence rates of some statistics among M sites can be the same as the rates of the statistics based on the one site as long as M is not too large, e.g., $\log M = O(\log p)$, which is a relatively weak condition and has been used in Lian and Fan (2018).

For any $1 \leq j \leq p$, denote $\boldsymbol{\Theta}_j$ as the jth row of $\boldsymbol{\Theta}$ with $\boldsymbol{\Theta}$ being the inverse matrix of $\boldsymbol{\Sigma} = E\{\boldsymbol{x}_i \boldsymbol{x}_i^\top I(|\varepsilon_{i,\tau_n}| \leq \tau_n)\}$, where $\varepsilon_{i,\tau_n} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*_{\tau_n}$. It can be verified that $\boldsymbol{\Theta}_j = \boldsymbol{\rho}_j / \{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)] E[I(|\varepsilon_{i,\tau_n}| \leq \tau_n)]\}$, where $\boldsymbol{\rho}_j = (-\gamma^*_{j,1}, \ldots, -\gamma^*_{j,(j-1)}, 1, -\gamma^*_{j,j}, \ldots, -\gamma^*_{j,(p-1)})$. Thus, for $1 \leq m \leq M$, an estimator of $\boldsymbol{\Theta}_j$ based on the mth site can be obtained by

$$\hat{\boldsymbol{\Theta}}_j^{(m)} = \hat{\boldsymbol{\rho}}_j^{(m)} / \{n^{-2} \sum_{i \in \mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)}) \sum_{i \in \mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m| \leq \tau_n)\},$$

where $\hat{\boldsymbol{\rho}}_j^{(m)} = (-\hat{\gamma}^{(m)}_{j,1}, \ldots, -\hat{\gamma}^{(m)}_{j,(j-1)}, 1, -\hat{\gamma}^{(m)}_{j,j}, \ldots, -\hat{\gamma}^{(m)}_{j,(p-1)})$.

Theorem 2. *Under the conditions in Theorem 1 and $M = o(\sqrt{N}/(s^{3/2}\log p))$. For any $1 \leq j \leq p$, we have*

$$\bar{\beta}_j^{\mathbf{d}} - \beta_j^* = \frac{1}{N} \sum_{m=1}^{M} \hat{\boldsymbol{\Theta}}_j^{(m)} \sum_{i \in \mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n}) + o_p(N^{-1/2}).$$

Remark 2. Compared with Theorem 1, a stronger condition on M is needed to derive the asymptotic normality since a faster convergence rate is required for the high order term in Taylor expansion of $\bar{\beta}_j^{\mathbf{d}}$ around β_j^*.

Corollary 1. *Under the conditions in Theorem 2, as $N \to \infty$, for $1 \leq j \leq p$, we have*

$$\sqrt{N}(\bar{\beta}_j^{\mathbf{d}} - \beta_j^*)/\sigma_j \xrightarrow{d} N(0,1),$$

where $\sigma_j^2 = E\{\varepsilon_{i,\tau_n}^2 I(|\varepsilon_{i,\tau_n}| \leq \tau_n) + \tau_n^2 I(|\varepsilon_{i,\tau_n}| > \tau_n)\} / \{E(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)^2 [P(|\varepsilon_{i,\tau_n}| \leq \tau_n)]^2\}$.

With Corollary 1, σ_j^2 can be estimated consistently by $\hat{\sigma}_j^2 = M^{-1}\sum_{m=1}^{M}\hat{\sigma}_{jm}^2$ with $\hat{\sigma}_{jm}^2 = n^{-1}\sum_{i\in\mathcal{I}_m}\{(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)^2 I(|y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m| \le \tau_n) + \tau_n{}^2 I(|y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m| > \tau_n)\}\hat{\boldsymbol{\Theta}}_j^{(m)}\hat{\boldsymbol{\Sigma}}^{(m)}\hat{\boldsymbol{\Theta}}_j^{(m)\top}$ and $\hat{\boldsymbol{\Sigma}}^{(m)} = n^{-1}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\boldsymbol{x}_i^\top$. We construct the $100(1-\alpha)\%$ confidence interval for β_j^* as

$$[\bar{\beta}_j^{\mathbf{d}} - N^{-1/2}\hat{\sigma}_j\Phi^{-1}(1-\alpha/2),\ \bar{\beta}_j^{\mathbf{d}} + N^{-1/2}\hat{\sigma}_j\Phi^{-1}(1-\alpha/2)],$$

where $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ upper quantile of standard normal distribution.

Theorem 3. *Under the conditions in Theorem 1, assume* $\nu = C_0\sqrt{\log p/N}$ *for some sufficiently large constant* C_0 *and* $M = O(\sqrt{N/(s^3\log p)})$*, then we have*

$$\|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}}\Big),\ \|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) - \boldsymbol{\beta}^*\|_2 = O_p\Big(\sqrt{\frac{s\log p}{N}}\Big).$$

3. Sparse and debiased lasso estimator via CSL

3.1 Multi-round communication-efficient adaptive Huber estimator

Although the proposed SADL estimator only needs one round communication between the local and the central sites, evaluating $\hat{\boldsymbol{\Theta}}^{(m)}$ on the mth site still requires to solve p lasso problems, which incurs exorbitant communication or computation costs. Alternatively, it is well-known that the gradient vectors can be easily calculated and communicated between the central and local sites. In this section, we propose another distributed estimator with lower communication cost and higher accuracy.

Inspired by Jordan et al. (2019) and Luo et al. (2022), without loss of generality we regard the first site as the central site, given the total number of rounds T and the estimator $\tilde{\boldsymbol{\beta}}^{[t-1]}$ after the $(t-1)$th iterations for $1 \leq t \leq T$, the tth round ℓ_1-penalized communication-efficient adaptive Huber estimator is given as follows:

$$\tilde{\boldsymbol{\beta}}^{[t]} \equiv \tilde{\boldsymbol{\beta}}^{[t]}_{\tau_n,\tau_N} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\tilde{\mathcal{L}}_1(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{[t-1]}) + \tilde{\lambda}_1^{[t]}\|\boldsymbol{\beta}\|_1\}, \tag{3.8}$$

where

$$\begin{aligned}\tilde{\mathcal{L}}_1(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{[t-1]}) &= L_{1,\tau_n}(\boldsymbol{\beta}) - \langle \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[t-1]}) - \nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[t-1]}), \boldsymbol{\beta}\rangle \\ &= L_{1,\tau_n}(\boldsymbol{\beta}) - \langle \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[t-1]}) - \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}^{[t-1]}), \boldsymbol{\beta}\rangle,\end{aligned}$$

and $\nabla_{\boldsymbol{\beta}} L_{m,\tau}(\boldsymbol{\beta})$ denotes the gradient of the function $L_{m,\tau}(\boldsymbol{\beta})$ and $\langle\cdot,\cdot\rangle$ represents the inner product of two vectors. When $t=1$, we set the initial value $\tilde{\boldsymbol{\beta}}^{[0]} = \hat{\boldsymbol{\beta}}_1$ obtained by (2.3). Note that $\tilde{\mathcal{L}}_1^{[t]}(\boldsymbol{\beta})$ depends on both τ_n and τ_N. For the only nonlocal component $\nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[t-1]})$, each site can calculate $\nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}^{[t-1]})$ locally with τ_N and communicate this gradient to the central site. Hence, it can be seen that this procedure only communicates gradient information and requires one communication round with order $O((M-1)p)$.

To further reduce the impact of the choice of the central site and improve the stability of the estimator, every site can be regarded as a central site and optimize their corresponding optimization problem in parallel. When using the mth site as the central site, the tth round ℓ_1-penalized communication-efficient adaptive Huber estimator is defined as follows:

$$\tilde{\boldsymbol{\beta}}_m^{[t]} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\tilde{\mathcal{L}}_m(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}_m^{[t-1]}) + \tilde{\lambda}_m^{[t]}\|\boldsymbol{\beta}\|_1\},$$

where $\tilde{\mathcal{L}}_m(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}_m^{[t-1]}) = L_{m,\tau_n}(\boldsymbol{\beta}) - \langle \nabla_{\boldsymbol{\beta}} L_{m,\tau_n}(\tilde{\boldsymbol{\beta}}_m^{[t-1]}) - M^{-1}\sum_{m=1}^M \nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}_m^{[t-1]}), \boldsymbol{\beta}\rangle$ and $\tilde{\boldsymbol{\beta}}_m^{[t-1]}$ is the resulting estimator after $(t-1)$th iterations of the mth site. Finally, we derive the tth round aggregated communication-efficient adaptive Huber estimator:

$$\tilde{\boldsymbol{\beta}}_{all}^{[t]} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\boldsymbol{\beta}}_m^{[t]}. \tag{3.9}$$

3.2 Thresholding communication-efficient debiased lasso estimator

Similar with the discussion in Section 2.2, both $\tilde{\boldsymbol{\beta}}^{[t]}$ and $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$ are generally biased and it is hard to obtain their asymptotic distributions. Our second goal is to propose a sparse communication-efficient debiased lasso (SCDL) adaptive Huber estimator. To correct the biases, as long as we get $\tilde{\boldsymbol{\beta}}^{[t]}$ from (3.8), the decorrelated score estimating equation based on $\tilde{\mathcal{L}}_1(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{[t]})$ for β_j is formulated as:

$$\nabla_{\beta_j}\tilde{\mathcal{L}}_1(\beta_j, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]}) - \hat{\boldsymbol{\gamma}}_j^{(1)\top}\nabla_{\boldsymbol{\beta}_{-j}}\tilde{\mathcal{L}}_1(\beta_j, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]}) = 0,$$

where $\nabla_{\beta_j}\tilde{\mathcal{L}}_1(\beta_j, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]})$ and $\nabla_{\boldsymbol{\beta}_{-j}}\tilde{\mathcal{L}}_1(\beta_j, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]})$ denote the gradients of function $\tilde{\mathcal{L}}_1(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}^{[t]})$ with respect to β_j and $\boldsymbol{\beta}_{-j}$ respectively, and $\hat{\boldsymbol{\gamma}}_j^{(1)}$ is obtained by (2.6) based on the central cite. Given the tth round estimator $\tilde{\boldsymbol{\beta}}^{[t]}$ from (3.8), we use the same technique as (2.5) and construct the communication-efficient debiased lasso estimator for β_j^* as follows:

$$\tilde{\beta}_j^{\mathbf{d}[t]} = \tilde{\beta}_j^{[t]} - \frac{\nabla_{\beta_j}\tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[t]}, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]}) - \hat{\boldsymbol{\gamma}}_j^{(1)\top}\nabla_{\boldsymbol{\beta}_{-j}}\tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[t]}, \tilde{\boldsymbol{\beta}}_{-j}^{[t]}|\tilde{\boldsymbol{\beta}}^{[t]})}{n^{-2}\sum_{i\in\mathcal{I}_1}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top\hat{\boldsymbol{\gamma}}_j^{(1)})\sum_{i\in\mathcal{I}_1} I(|y_i - \boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[t]}| \le \tau_n)},$$

and then obtain the multi-round communication-efficient debiased lasso estimator $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]} = (\tilde{\beta}_1^{\mathbf{d}[t]}, \cdots, \tilde{\beta}_p^{\mathbf{d}[t]})^\top$. However, the debiased lasso estimator $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]}$ is no longer sparse such that hard-thresholding is needed to achieve sparsity and reduce the ℓ_2 error. Using the hard-thresholding operator in Section 2.2, finally we get the tth multi-round SCDL adaptive Huber estimator

$$\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]}) = (\mathcal{T}_\nu(\tilde{\beta}_1^{\mathbf{d}[t]}), \cdots, \mathcal{T}_\nu(\tilde{\beta}_p^{\mathbf{d}[t]}))^\top. \tag{3.10}$$

Similarly, the tth multi-round aggregated SCDL estimator is

$$\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]}) = \mathcal{T}_\nu\Big(\frac{1}{M}\sum_{m=1}^{M}\tilde{\boldsymbol{\beta}}_m^{\mathbf{d}[t]}\Big), \tag{3.11}$$

with $\nu \asymp \sqrt{\log p/N}$. We summarize the procedures for computing the SADL and SCDL estimators into two algorithms in the Supplementary Material.

Remark 3. Under some regular conditions, the estimator $\tilde{\boldsymbol{\beta}}^{[T]}$ obtained from (3.8) with $T \asymp \lceil \log M \rceil$ satisfies the bound $\|\tilde{\boldsymbol{\beta}}^{[T]} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{s\log p/N}$, which is the optimal convergence rate of the lasso estimator using the entire data (Luo et al., 2022). Here, denote $\lceil a \rceil$ as the minimum integer bigger than a for $a \in \mathbb{R}$. After the debiasing and hard-thresholding procedure, we will show that the SCDL estimator not only achieves the optimal convergence rate in accuracy of estimation, but also has the asymptotic normality property. To solve (3.8), we apply the local adaptive majorize-minimize (MM) algorithm as in Luo et al. (2022), which is an extended form of the traditional MM algorithm to accommodate the lasso penalty.

3.3 Theoretical results

Theorem 4. *Under Conditions (C1)-(C3), if* $\tau_N \asymp \sigma\sqrt{N/\log p}$, $\tau_n \asymp \sigma\sqrt{N/(M\log p)}$, $\tilde{\lambda}_m^{[t]} \asymp \sqrt{\log p/N} + (s^2 M\log p/N)^{t/2}\sqrt{\log p/N}$ *uniformly in* m *for* $t = 1, \dots, T$ *and* $\omega_{jm} \asymp \sqrt{M\log p/N}$ *uniformly in* m *and* j*, with* $\log M = O(\log p)$*, then after* $T \asymp \lceil \log M \rceil$ *rounds of communication, we have*

$$\|\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]} - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s\sqrt{M}\log p}{N}\Big),$$
$$\|\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[T]} - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s\sqrt{M}\log p}{N}\Big).$$

Remark 4. Benefiting from the double robustification parameters to adjust bias, the condition $E(|\varepsilon|^3) < \infty$ in Theorem 1 is not needed in Theorem 4 because the approximation error $|\alpha_{\tau_N}| \le \sqrt{\log p/N}$ is comparable with the main term. Moreover, in order to attain the same statistical accuracy in term of ℓ_∞ error, the condition on the number of sites M for $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]}$ can be weakened from $M = O(\sqrt{N/(s^3 \log p)})$ in Theorem 1 to $M = o(N/(s^2 \log p))$, due to the communication-efficient method and double data-adaptive robustification parameters.

Theorem 5. *Under the conditions in Theorem 4 and $M = o(N/(s^2 \log^2 p))$. If $E(|\varepsilon|^3) < \infty$ holds, then for any $1 \le j \le p$, we have*

$$\tilde{\beta}_j^{\mathbf{d}[T]} - \beta_j^* = \frac{1}{N}\tilde{\boldsymbol{\Theta}}_j^{(1)} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N}) + o_p(N^{-1/2}),$$

$$\tilde{\beta}_{all,j}^{\mathbf{d}[T]} - \beta_j^* = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{N}\tilde{\boldsymbol{\Theta}}_j^{(m)} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N}) + o_p(N^{-1/2}),$$

where $\tilde{\boldsymbol{\Theta}}_j^{(m)} = \hat{\boldsymbol{\rho}}_j^{(m)}/\{n^{-2}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)}) \sum_{i\in\mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_m^{[T]}| \le \tau_n)\}$ *and* $\hat{\boldsymbol{\rho}}_j^{(m)} = (-\hat{\gamma}_{j,1}^{(m)}, \ldots, -\hat{\gamma}_{j,(j-1)}^{(m)}, 1, -\hat{\gamma}_{j,j}^{(m)}, \ldots, -\hat{\gamma}_{j,(p-1)}^{(m)})$.

Remark 5. Compared with Theorem 2, the condition on M that guarantees the asymptotic normality in Theorem 5 is weaker. In addition, it should be pointed out that $\|\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[T]} - \boldsymbol{\beta}^*\|_\infty$ attains the same convergence rate as that of $\|\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]} - \boldsymbol{\beta}^*\|_\infty$ after several iterations. Here, $E(|\varepsilon|^3) < \infty$ is necessary to derive the asymptotic normality in Theorem 5 such that the approximation bias $\boldsymbol{\beta}_{\tau_N}^* - \boldsymbol{\beta}^*$ is asymptotically negligible.

Corollary 2. *Under the conditions in Theorem 5, as $N \to \infty$, for $1 \le j \le p$, we have*

$$\sqrt{N}(\tilde{\beta}_j^{\mathbf{d}[T]} - \beta_j^*)/\varrho_j \xrightarrow{d} N(0,1),$$

where $\varrho_j^2 = E\{\varepsilon_{i,\tau_N}^2 I(|\varepsilon_{i,\tau_N}| \le \tau_N) + \tau_N^2 I(|\varepsilon_{i,\tau_N}| > \tau_N)\}/\{E(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)^2 [P(|\varepsilon_{i,\tau_N}| \le \tau_n)]^2\}$.

With Corollary 2, ϱ_j^2 can be estimated by $\tilde{\varrho}_j^2 = M^{-1} \sum_{m=1}^M \tilde{\varrho}_{jm}^2$ consistently with $\tilde{\varrho}_{jm}^2 = n^{-1} \sum_{i \in \mathcal{I}_m} \{(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]})^2 I(|\tilde{y}_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| \le \tau_N) + \tau_N^2 I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| > \tau_N)\} \tilde{\boldsymbol{\Theta}}_j^{(m)} \hat{\boldsymbol{\Sigma}}^{(m)} \tilde{\boldsymbol{\Theta}}_j^{(m)\top}$. Therefore, we can construct the $100(1-\alpha)\%$ confidence interval for β_j^* as

$$[\tilde{\beta}_j^{\mathbf{d}[T]} - N^{-1/2} \tilde{\sigma}_j \Phi^{-1}(1-\alpha/2), \tilde{\beta}_j^{\mathbf{d}[T]} + N^{-1/2} \tilde{\sigma}_j \Phi^{-1}(1-\alpha/2)].$$

Theorem 6. *Under the conditions in Theorem 4, assume* $\nu = C_0 \sqrt{\log p / N}$ *for some sufficiently large constant* C_0*, then we have*

$$\|\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]}) - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}}\Big), \ \|\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]}) - \boldsymbol{\beta}^*\|_2 = O_p\Big(\sqrt{\frac{s \log p}{N}}\Big),$$
$$\|\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[T]}) - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}}\Big), \ \|\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[T]}) - \boldsymbol{\beta}^*\|_2 = O_p\Big(\sqrt{\frac{s \log p}{N}}\Big).$$

Remark 6. Compared with the condition $M = O(\sqrt{N/(s^3 \log p)})$ of the SADL estimator in Theorem 3, the SCDL estimator allows a weaker condition $M = o(N/(s^2 \log p))$ to attain the optimal convergence rate in Theorem 6, which also coincides with our simulation results in Section 4.

4. Simulation studies

In this section, we evaluate the performance of two proposed sparse and debiased adaptive Huber estimators through extensive simulation studies. Consider the following model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i,\ i = 1, \ldots, N,$$

where $\boldsymbol{\beta}^* = (5, 5, 5, 5, 5, 5, 0, \cdots)^\top$, $s = 6$, $x_{i,1} \equiv 1$ and $x_{i,j} \sim N(0, 1)$ are independently and identically distributed for $j = 2, \ldots, p$. Five different errors ε_i are considered: **(1)** $N(0, 1)$: standard normal; **(2)** t_3: t-distribution with 3 degrees of freedom; **(3)** $Pareto(2, 4)$: Pareto distribution with scale parameter 2 and shape parameter 4; **(4)** χ^2_3: Chi-square distribution with degrees of freedom 3; **(5)** $LogN(0, 1)$: Log-normal distribution with local parameter 0 and scale parameter 1. It can be seen that the first two errors are symmetric and the last three errors are skewed. Moreover, t_3, $Pareto(2, 4)$ and χ^2_3 errors are heavy-tailed distributions. In addition, we center the skewed χ^2_3 and $LogN(0, 1)$ errors to identify the intercept term.

All simulations are repeated 200 times and we compare the ℓ_∞ and ℓ_2 errors, i.e., $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$, of the following eight estimators.

(a) the global ℓ_1-penalized adaptive Huber estimator $\hat{\boldsymbol{\beta}}$ using $N = nM$ samples in (2.2);

(b) the sparse and debiased global estimator $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ based on the estimator **(a)**;

(c) the aggregated ℓ_1-penalized adaptive Huber estimator $\bar{\boldsymbol{\beta}}$ in (2.4);

(d) the SADL adaptive Huber estimator $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ in (2.7);

(e) the ℓ_1-penalized communication-efficient adaptive Huber estimator $\tilde{\boldsymbol{\beta}}^{[t]}$ in (3.8);

(f) the SCDL adaptive Huber estimator $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ in (3.10);

(g) the aggregated ℓ_1-penalized communication-efficient adaptive Huber estimator $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$ in (3.9);

(h) the aggregated SCDL adaptive Huber estimator $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ in (3.11).

In practice, the regularization parameters λ_m in (2.3) and ω_{jm} in (2.6) are selected by cross-validation using R packages `adaHuber` and `glmnet`, respectively. The robustification parameter τ_n is determined by a tuning-free principle (Wang et al., 2021; Sun et al., 2020) and we choose $\tau_N = \eta M^{1/2}\tau_n$ according to Theorem 4, where η is a constant determined by the validation set approach. The hard-thresholding parameter ν is determined by five-fold cross-validation according to Theorems 3 and 6.

4.1 Effect of number of rounds and aggregation

In the first experiment, we investigate the performance of the multi-round SCDL and aggregated SCDL estimators, i.e., $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$, by varying the number of rounds from $t = 1, \ldots, 5$. To be specific, we consider the t_3 error with $n = 100$, $p = 200$ and $M = 20$. Based on $\tilde{\boldsymbol{\beta}}^{[t]}$ and $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$ as well as $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$, the simulated ℓ_∞ and ℓ_2 results versus the number of rounds are plotted in **Figure 1**. In addition, the simulated results of the global estimators $\hat{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ are also provided for comparison.

We have the following findings. (i) With t increasing, both the ℓ_∞ and ℓ_2 errors of $\tilde{\boldsymbol{\beta}}^{[t]}$, $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ decrease rapidly and usually attain stable performance after $t = 3$ rounds. Moreover, when $t \geq 4$, $\tilde{\boldsymbol{\beta}}^{[t]}$ and $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$ are close to $\hat{\boldsymbol{\beta}}$ as well as $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ are close to $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$, respectively. In particular, the differences of the ℓ_∞ and ℓ_2 errors between $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ are negligible. (ii) For any fixed t, compared with $\tilde{\boldsymbol{\beta}}^{[t]}$ and $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$, both the ℓ_∞ and ℓ_2 errors of $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ are much smaller, which means the debiasing and thresholding procedures are helpful to improve the accuracy of estimation. Compared with $\tilde{\boldsymbol{\beta}}^{[t]}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})$, $\tilde{\boldsymbol{\beta}}_{all}^{[t]}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})$ perform slightly better when t is small, respectively, by efficiently utilizing statistical structures and similarities among the local losses and benefiting from the averaging step. However, when t increases, the differences become negligible.

Based on the above findings, in the following simulations we fix $T = 5$ and only report the results of $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]})$ for comparison. To simplify notation, we omit "$[T]$" in $\tilde{\boldsymbol{\beta}}^{[T]}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]})$ and use $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$, respectively.

4.2 Effect of heavy-tailed and asymmetric errors

We consider $(n, M) = (100, 5)$ and $p = 200$, 300, 400 under the five different errors **(1)-(5)**. The simulated ℓ_∞ and ℓ_2 results versus different values of p based on the six estimators **(a)-(f)** are shown in **Figure 2**, respectively. To save space, we only show the simulated results under the first three errors and the results under χ^2_3 and $LogN(0,1)$ errors are given in the Supplementary Material. In addition, the computing time and performance of the eight estimators under heteroskedastic error and outliers are also compared in the Supplementary Material.

The three columns in **Figure 2** correspond to the three errors $N(0,1)$, t_3 and $Pareto(2,4)$, respectively. (i) For any fixed p, n and M, compared with the existing distributed estimators $\bar{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ without bias correction, it can be seen that $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ perform much better, respectively, in terms of the ℓ_∞ and ℓ_2 errors, which implies that the debiasing and thresholding procedures are not sensitive to the errors and can efficiently reduce estimation errors for the high-dimensional models. In particular, the bias reduction is substantial under the t_3 and $Pareto(2,4)$ errors. On the other hand, compared with the aggregated estimators $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, it can be seen that the communication-efficient estimators $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ always have much smaller ℓ_∞ and ℓ_2 errors, respectively. Moreover, the performance of $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$

is comparable with the golden standard estimator $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$. It can be seen that both the proposed two estimators $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ can improve the ℓ_∞ and ℓ_2 errors, but $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ performs better than $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ due to the following reasons. Compared with $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, both $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ use double robustification parameters to adjust bias and engage the gradient information of the entire data in the central site. Unfortunately, $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ may involve additional variability from computing the nodewise lasso. Moreover, $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ require much weaker conditions on the number of sites M than that of $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ to achieve the optimal convergence rates. (ii) When the dimension p increases, the ℓ_∞ and ℓ_2 errors of all estimators increase slightly, except for $\bar{\boldsymbol{\beta}}$.

4.3 Effect of number of sites

We fix the local sample size $n = 100$ and the dimension $p = 200$, but vary $M = 5$, 20, 50 to see the influence of the number of sites. The simulated results of the ℓ_∞ and ℓ_2 errors with the different M are reported in **Figure 3**. (i) When the number of sites M increases, the performance of all estimators becomes better as expected, since the total sample size N increases. Compared with the estimators $\bar{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, the sparse and debiased lasso estimators $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ have better performance on the ℓ_∞ and ℓ_2 results, respectively. Moreover, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ achieves the similar performance with the

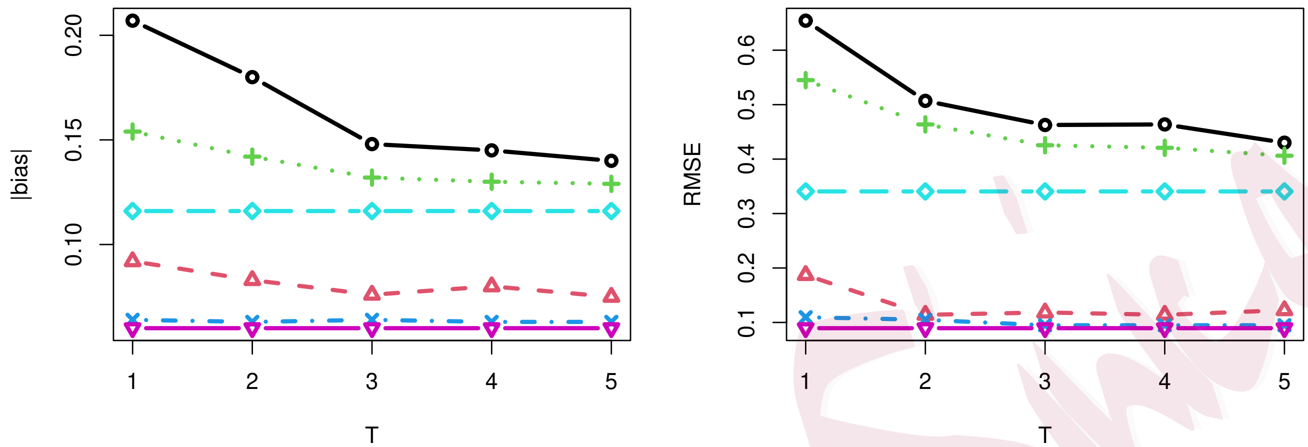

Figure 1: The ℓ_∞ and ℓ_2 errors versus the number of rounds when $(n, M, p) = (100, 20, 200)$ under t_3 error. Here, $\tilde{\boldsymbol{\beta}}^{[t]}(\circ)$, $\tilde{\boldsymbol{\beta}}_{all}^{[t]}(+)$, $\hat{\boldsymbol{\beta}}(\diamond)$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]})(\triangle)$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]})(\times)$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ $(\triangledown)$.

golden standard $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$. (ii) An interesting finding is that the ℓ_∞ and ℓ_2 errors of $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ are decreasing faster than $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, especially when M is small, which shows the advantage of our proposed communication-efficient estimators. Moreover, it can be seen that all errors of $\hat{\boldsymbol{\beta}}$ are even smaller than the errors of $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ when the error follows t_3 distribution and $M \geq 20$, which can be partly explained by the variability of computing nodewise lasso when M is large, i.e., there may exist some unstable estimates in the M local estimators. If one local site returns a bad estimator, the SADL estimator performs worse due to the averaging approach.

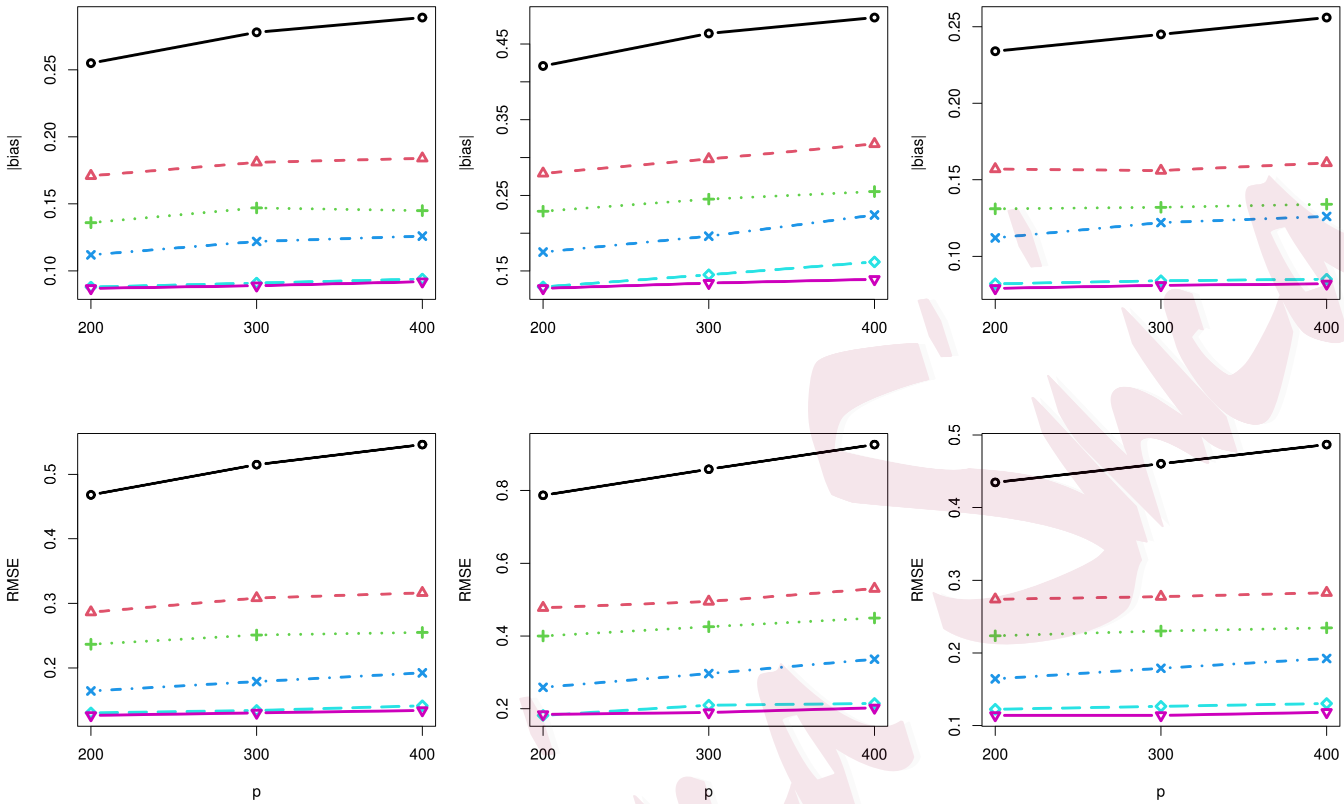

Figure 2: The ℓ_∞ and ℓ_2 errors for $N(0,1)$, t_3 and $Pareto(2,4)$ with varying $p = 200, 300, 400$ when $(n, M) = (100, 5)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

4.4 Coverage probability

We set $n = 200$ and consider the t_3 error to investigate the confidence intervals (CIs) of the proposed two estimators based on the following two cases: (i) varying $p = 200, 400, 600$ with the fixed $M = 5$; (ii) varying $M = 5, 10, 20$

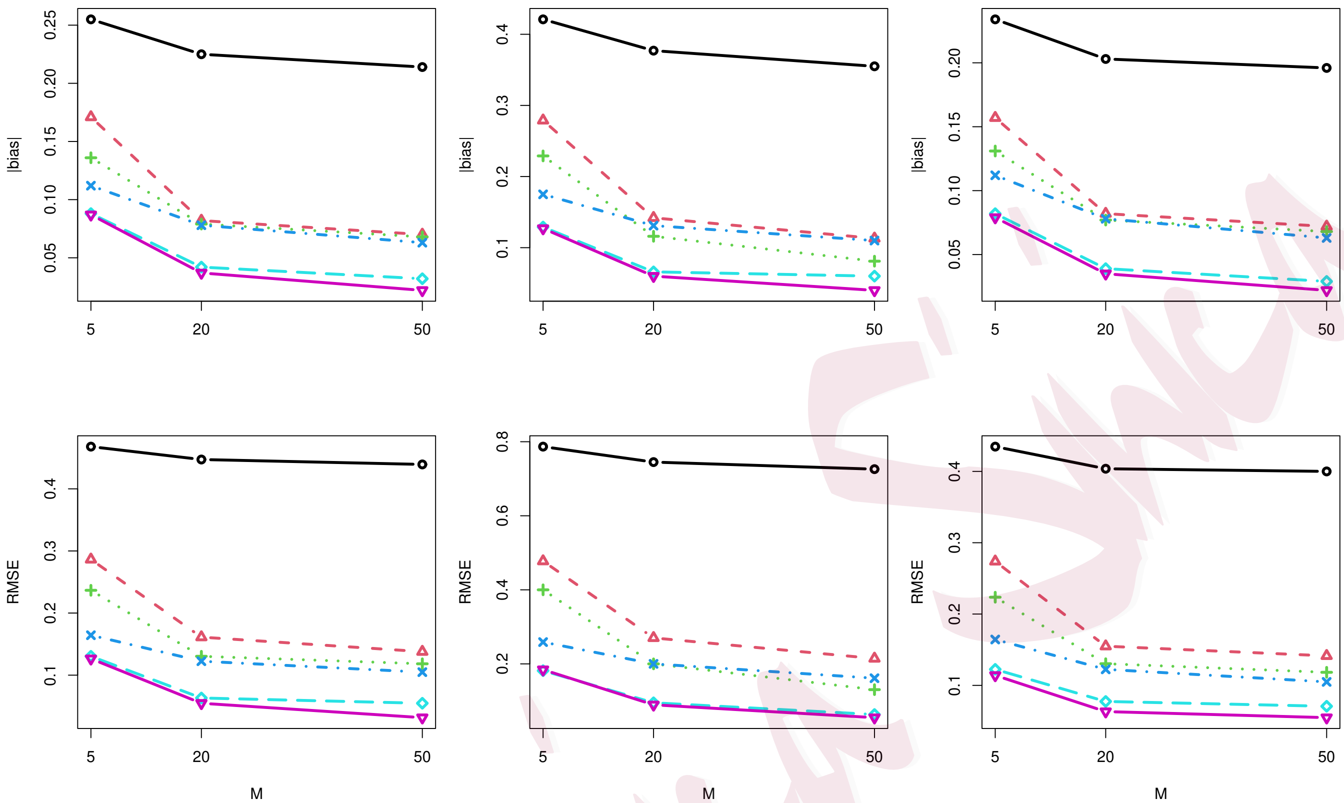

Figure 3: The ℓ_∞ and ℓ_2 errors for $N(0,1)$, t_3 and $Pareto(2,4)$ with varying number of sites $M = 5, 20, 50$ when $(n,p) = (100, 200)$. Here, $\bar{\boldsymbol{\beta}}$ ($\circ$), $\tilde{\boldsymbol{\beta}}$ ($\triangle$), $\hat{\boldsymbol{\beta}}$($+$), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ ($\times$), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ ($\diamond$) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ ($\triangledown$).

with the fixed $p = 200$. Note $\mathcal{S} = \{j | \beta_j^* \neq 0, 1 \leq j \leq p\}$ and $\mathcal{S}^c = \{j | \beta_j^* = 0, 1 \leq j \leq p\}$. For a given set $\mathcal{A} \subset \mathcal{S} \cup \mathcal{S}^c = \{1, \ldots, p\}$, define the average of the coverage probabilities (ACP) of the 95% confidence intervals over the set $\mathcal{A}$ as $\mathrm{ACP}(\mathcal{A}) = \sum_{j \in \mathcal{A}} \mathrm{CP}_j / |\mathcal{A}|$, where CP_j is the empirical coverage probability of the 95% confidence interval for β_j^*. The average lengths (AL) can also be defined similarly. For comparison, we consider the distributed estimators of Battey et al. (2018) by adopting the least squares loss function and denote the resulting estimators as $\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$, $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$ and $\hat{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$, respectively. **Table 1** reports the simulated ACPs with 500 repetitions over the parameter sets $\mathcal{S}$, $\mathcal{S}^c$ and $\mathcal{S}\cup\mathcal{S}^c$, respectively. When N is fixed, from Corollaries 1 and 2, the ALs of proposed estimators depend on the estimation of variance and simulation results show that their values have slight changes across the three different parameter sets, which coincides with the results in Han et al. (2022). Hence, we only report the ALs of the $\mathcal{S}\cup\mathcal{S}^c$ in Table 1. Under the case (i): for any fixed p, the ACPs of the CIs based on Battey et al. (2018) perform badly in all scenarios while the ACPs of the estimators $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ and $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ are close to the nominal level 95% under the three sets $\mathcal{S}$, $\mathcal{S}^c$ and $\mathcal{S}\cup\mathcal{S}^c$. When p increases, the ACPs of the CIs based on the estimators $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ and $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ only decrease slightly in $\mathcal{S}^c$. The main reason is that the model complexity increases when p becomes larger. In addition, the ACPs of $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ are lower than the results of $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ due to the extra variability from the estimation of the inverse covariance matrix. It can be seen that the ALs of $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ and $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ keep stable when p increases. Under the case (ii): as M increases, all the ALs become shorter. According to our theoretical results, the lengths of the ALs are proportional to $M^{-1/2}$ when n is fixed, which is validated by our simulation results in Table 1. These simulation results show the proposed two estimators can make accurate statistical inference.

5. Application

In this section, we apply our method to the Communities and Crime Data Set from the UCI Machine Learning Repository. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. After removing missing values, there are 101 variables with 1993 observations in the 49 states of the United States. We assign each community by the state number to identify its division, which is defined by the Census Bureau-designated regions and divisions, including New England, Mid-Atlantic and so on. Thus there are 9 units, and the number of observations in each unit is 258, 358, 217, 87, 262, 122, 239, 98 and 352. In the real data analysis, we use the total number of violent crimes per $100K$ population (ViolentCrimesPerPop) as the response and the other variables as predictors. After scaling the responses and predictors, we set $M = 9$ by the division and compare the performance of proposed estimators $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ with the global estimator. First, we calculate the estimates of $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$, $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$, $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ and obtain their computation time as 12.753, 4.972, 4.240 seconds, respectively, which also indicates the distributed methods

Table 1: The average of the coverage probabilities (ACPs) and average lengths (ALs) of the 95% confidence intervals over $\mathcal{S}$, $\mathcal{S}^c$ and $\mathcal{S}\cup\mathcal{S}^c$, respectively, with varying p and M.

	$\mathcal{S}$	$\mathcal{S}^c$	$\mathcal{S}\cup\mathcal{S}^c$	AL	$\mathcal{S}$	$\mathcal{S}^c$	$\mathcal{S}\cup\mathcal{S}^c$	AL	$\mathcal{S}$	$\mathcal{S}^c$	$\mathcal{S}\cup\mathcal{S}^c$	AL
	$\boldsymbol{p}=200$				$\boldsymbol{p}=400$				$\boldsymbol{p}=600$			
$\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.681	0.750	0.748	0.125	0.644	0.755	0.753	0.125	0.607	0.751	0.750	0.125
$\tilde{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.752	0.746	0.746	0.125	0.743	0.747	0.747	0.125	0.698	0.743	0.743	0.125
$\hat{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.774	0.751	0.751	0.124	0.756	0.755	0.755	0.125	0.740	0.751	0.751	0.124
$\bar{\boldsymbol{\beta}}^{\mathbf{d}}$	0.915	0.950	0.949	0.211	0.881	0.952	0.951	0.211	0.894	0.952	0.951	0.212
$\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$	0.946	0.943	0.943	0.209	0.945	0.943	0.943	0.207	0.925	0.943	0.943	0.207
$\hat{\boldsymbol{\beta}}^{\mathbf{d}}$	0.947	0.950	0.950	0.210	0.946	0.950	0.950	0.208	0.942	0.950	0.950	0.206
	$\boldsymbol{M}=5$				$\boldsymbol{M}=10$				$\boldsymbol{M}=20$			
$\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.681	0.750	0.748	0.125	0.585	0.691	0.688	0.088	0.552	0.736	0.731	0.063
$\tilde{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.752	0.746	0.746	0.125	0.698	0.738	0.737	0.088	0.731	0.738	0.738	0.062
$\hat{\boldsymbol{\beta}}^{\mathbf{d}}_{ols}$	0.774	0.751	0.751	0.124	0.713	0.746	0.745	0.087	0.754	0.749	0.749	0.062
$\bar{\boldsymbol{\beta}}^{\mathbf{d}}$	0.915	0.950	0.949	0.211	0.901	0.939	0.938	0.150	0.919	0.936	0.935	0.107
$\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$	0.946	0.943	0.943	0.209	0.939	0.942	0.942	0.150	0.945	0.941	0.941	0.107
$\hat{\boldsymbol{\beta}}^{\mathbf{d}}$	0.947	0.950	0.950	0.210	0.938	0.950	0.950	0.149	0.960	0.951	0.951	0.107

reduce the computation and storage burden a lot. To further reduce the model complexity, we obtain the sparse estimates $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ with $\nu = 0.06$ and the 95% confidence intervals are also calculated by the normal approximation in Corollaries 1 and 2. It can be seen that 12 predictors are selected by the three estimators and the analysis results are shown in **Table 2**. We find that the point estimates and confidence intervals of the proposed $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ are similar to the results of $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$, which means our proposed estimates and inference results are stable and valid. According to Table 2, we observe that the coefficients of MalePctDivorce and HousVacant are positive, which means higher percentage of males who are divorced and vacant households may lead to increase the number of violent crimes. In addition, the lengths of confidence intervals for $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ are shorter than the lengths for $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$, which also means $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ is better than $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$. Of course, $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ has the shortest confidence interval lengths. If we set $\nu = 0.1$, the selected predictors of the three different estimators have slight differences. For example, different from $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$ and $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$, $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$ selects NumStreet, but tends to not select PctEmploy and MalePctNevMarr; compared with $\hat{\boldsymbol{\beta}}^{\mathbf{d}}$, $\tilde{\boldsymbol{\beta}}^{\mathbf{d}}$ tends to PctUrban.

Table 2: Estimates and 95% confidence intervals for Communities and Crime data.

	$M=9$		$M=1$
Variables	$\mathcal{T}_{0.06}(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$	$\mathcal{T}_{0.06}(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$	$\mathcal{T}_{0.06}(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$
Racepctblack	0.212(0.112,0.312)	0.191(0.094,0.288)	0.196(0.103,0.289)
PctUrban	0.080(0.025,0.135)	0.110(0.054,0.166)	0.086(0.032,0.140)
PctEmploy	0.069(-0.031,0.169)	0.162(0.066,0.259)	0.110(0.014,0.206)
MalePctDivorce	0.206(0.015,0.398)	0.219(0.040,0.398)	0.220(0.059,0.381)
MalePctNevMarr	0.068(-0.019,0.155)	0.109(0.023,0.195)	0.113(0.030,0.196)
PctIlleg	0.183(0.098,0.268)	0.176(0.093,0.259)	0.175(0.096,0.254)
PersPerOccupHous	0.264(0.063,0.465)	0.327(0.129,0.525)	0.271(0.074,0.468)
PctPersDenseHous	0.148(0.033,0.263)	0.171(0.058,0.284)	0.154(0.044,0.264)
HousVacant	0.149(0.052,0.246)	0.184(0.094,0.274)	0.129(0.043,0.215)
PctHousOwnOcc	0.330(0.080,0.580)	0.328(0.087,0.569)	0.275(0.042,0.508)
MedRent	0.191(0.007,0.374)	0.224(0.045,0.403)	0.206(0.032,0.380)
NumStreet	0.111(0.065,0.157)	0.086(0.043,0.129)	0.089(0.051,0.127)

Racepctblack: percentage of population that is African American; PctUrban: percentage of people living in areas classified as urban; PctEmploy: percentage of people 16 and over who are employed; MalePctDivorce: percentage of males who are divorced; MalePctNevMarr: percentage of males who have never married; PctIlleg: percentage of kids born to never married; PersPerOccupHous: mean persons per household; PctPersDenseHous: percent of persons in dense housing (more than 1 person per room); HousVacant: number of vacant households; PctHousOwnOcc: percent of households owner occupied; MedRent: median gross rent; NumStreet: number of homeless people counted in the street.

6. Conclusion

In this paper, we propose two sparse and debiased lasso distributed adaptive Huber regression estimators for distributed data in the presence of the heavy-tailed/asymmetric error and high-dimensional covariates. It should be pointed out that our first proposal is convenient to implement in practice; the second proposal uses double data-adaptive robustification parameters to achieve a balanced tradeoff between statistical optimality and communication efficiency. Compared with the first proposal, the second proposal performs better in simulation studies. In this paper, we consider the covariates are bounded and it is of interest to extend our methods to sub-Gaussian or heavy-tailed predictors in high-dimensional Huber regression models.

Supplementary Material

The Supplementary Material contains the algorithms for computing the proposed two estimators, additional simulation results, and proofs of Theorems and Corollaries.

Acknowledgements

The authors would like to thank the Editor, an Associate Editor, and two anonymous referees for helpful comments and suggestions. Lei Wang's research was supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (12271272). Wei Ma's research was supported by the China Scholarship Council. The authors would like to thank Dongxiao Han for valuable comments. Lei Wang is the corresponding author.

References

Battey, H., Fan, J., Liu, H., Lu, J. and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, **46**, 1352–1382.

Chen, X., Lie, W. and Mao, X. (2020). Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, **21**, 1–43.

Chen, X. and Xie, M. G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, **24**, 1655–1684.

Duan, R., Ning, Y. and Chen, Y. (2022). Heterogeneity-aware and communication efficient distributed statistical inference. *Biometrika*, **109**, 67–83.

Eklund, A., Nichols, T. and Knutsson, H. (2016). Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 7900–7905.

Fan, J., Guo, Y. and Wang, K. (2021). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, **1**, 1–11.

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

Han, D., Huang, J., Lin, Y. and Shen, G. (2022). Robust post-selection inference of high dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors. *Journal of Econometrics*, **230**, 416–431.

Jordan, M. I., Lee, J. D. and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, **114**, 668–681.

Lee, J. D., Liu, Q., Sun, Y. and Taylor, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research*, **18**, 115–144.

Lian, H. and Fan, Z. (2018). Divide-and-Conquer for debiased l_1-norm support vector machine in ultra-high dimensions. *Journal of Machine Learning Research*, **18**, 1–26.

Luo, J., Sun, Q. and Zhou, W. X. (2022). Distributed adaptive Huber regression. *Computational Statistics and Data Analysis*, **169**, 107419.

Lv, S. and Lian, H. (2022). Debiased Distributed Learning for Sparse Partial Linear Models in High Dimensions. *Journal of Machine Learning Research*, **23**, 1–32.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, **45**, 158–195.

Po-Ling Loh. (2021). Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics*, **15**, 5933-5994.

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

Tan, K. M, Battey, H. and Zhou, W. X. (2022). Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research*, **23**, 1-61.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.

Volgushev, S., Chao, S. K. and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics*, **47**, 1634–1662.

Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, **110**, 1658–1669.

Wang, L., Zheng, C. and Zhou, W. X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica*, **31**, 2153-2177.

Wang, J., Kolar, M., Srebro, N. and Zhang, T. (2017). Efficient distributed learning with sparsity. *Journal of Machine Learning Research* , **70**, 3636–3645.

Zhao, W., Zhang, F. and Lian, H. (2020). Debiasing and distributed estimation for high-dimensional quantile regression. *IEEE Transactions on Neural Networks and Learning Systems*, **31**, 2569–2577.

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: maweiha@gmail.com

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: junzhuogao1012@163.com

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: lwangstat@nankai.edu.cn

Department of Mathematics, City University of Hong Kong, China.

E-mail: henglian@cityu.edu.hk