Annals of the Institute of Statistical Mathematics (2023) 75:493–509
https://doi.org/10.1007/s10463-022-00852-4

Robust estimation for nonrandomly distributed data

Shaomin Li[1] · Kangning Wang[2] · Yong Xu[3]

Received: 4 May 2021 / Revised: 8 July 2022 / Accepted: 16 August 2022 /
Published online: 12 October 2022

Abstract

In recent years, many methodologies for distributed data have been developed. However, there are two problems. First, most of these methods require the data to be randomly and uniformly distributed across different machines. Second, the methods are mainly not robust. To solve these problems, we propose a distributed pilot modal regression estimator, which achieves robustness and can adapt when the data are stored nonrandomly. First, we collect a random pilot sample from different machines; then, we approximate the global MR objective function by a communication-efficient surrogate that can be efficiently evaluated by the pilot sample and the local gradients. The final estimator is obtained by minimizing the surrogate function in the master machine, while the other machines only need to calculate their gradients. Theoretical results show the new estimator is asymptotically efficient as the global MR estimator. Simulation studies illustrate the utility of the proposed approach.

Keywords Distributed data · Communication-efficient · Modal regression · Robustness

✉ Kangning Wang
wknsuda@126.com

✉ Yong Xu
xuyong0410@126.com

Shaomin Li
lsmjim@bnu.edu.cn

[1] Center for Statistics and Data Science, Beijing Normal University, No.18 Jinfeng Road, Zhuhai 519087, China

[2] School of Statistics, Shandong Technology and Business University, No.191 Binhai Middle Road, Yantai 264005, China

[3] School of Business Administration, Shandong Technology and Business University, No.191 Binhai Middle Road, Yantai 264005, China

1 Introduction

With the rapid development of science and technology, the research and application of massive data have attracted attention from various fields, such as astronomy, economics, and industry (Gopal and Yang, 2013; Battey et al., 2018). Massive data have the characteristics of large volume, high dimensionality, and complex structure. Since it is difficult to process such a large data set on a single machine, the data collection must be distributed on multiple connected machines for processing; one machine is used as the master machine, and the other computers are used as worker machines (Duchi et al., 2014).

In the last few years, a considerable amount of work has been done to develop methodologies for distributed data, and the common methods can be divided into two categories: the "one-shot" approach and the iterative method. The first approach conducts an estimation on each worker machine and transfers the local estimates to the central machine to obtain the final estimator by averaging (Zhang et al., 2013; Lee et al., 2017; Battey et al., 2018; Fan et al., 2019). This method is easy to operate and is highly effective since it requires only one round of "master-and-worker" communication. However, this method requires higher accuracy for the estimates on each working machine, and it might perform poorly when the statistic is nonlinear (Shamir et al., 2014; Jordan et al., 2019; Wang et al., 2022a). Unlike the "one-shot" approach, the iterative method requires multiple rounds of communication between the master and working machines. Through multiple iterations, this method can achieve the same statistical accuracy and convergence speed as the global estimator (Wang et al., 2017; Jordan et al., 2019; Fan et al., 2021).

All these methods have been proven practically useful; however, there are two noteworthy issues. *First*, most of the existing methods must satisfy the assumption of homogeneity, i.e., the data are randomly and uniformly distributed across different machines. In practice, however, this assumption is not common since the data might be recorded by time or location, so the distribution of data is different across machines. *Second*, most of the aforementioned methods are not robust since they are based on either least square or likelihood, i.e., they may be adversely influenced by heavy-tails or outliers. In distributed settings, there are often outliers due to a system breakdown of worker machines, which can lead to a completely wrong final estimates. Thus, robust estimation has recently become an important topic in distributed learning research. Traditional robust methods include Huber's estimation (Huber, 1981) and quantile regression (Koenker and Bassett, 1978). However, these methods lose efficiency when the error distribution is normal or there are no outliers. Yao et al. (2012) proposed a modal regression-based estimator, which can achieve both efficiency and robustness via a tuning parameter.

Recently, there has been a growing research interest in these two issues. For nonrandomly distributed data, Wang et al. (2020) developed a pseudo-Newton–Raphson algorithm to efficiently estimate generalized linear models; Zhu et al. (2021) developed a distributed least squares approximation algorithm; Wang

et al. (2022b) proposed a communication-efficient estimator for nonrandomly distributed data. In the area of robust estimations for distributed data, Chen et al. (2019) studied the inference problem in quantile regression; Wang and Li (2021) proposed a distributed modal regression method; Tu et al. (2021) proposed a variance reduced median-of-means estimator, based on which they developed a robust distributed inference algorithm. However, to the best of our knowledge, there has been no research considering both issues simultaneously.

In this article, we propose a distributed pilot modal regression (DPMR) estimator, which is robust and can overcome the nonrandom distribution of the data. Actually, the data can be split and stored across the worker machines in any way. Specifically, we start by randomly collecting a total of n data as a pilot sample from each machine and store it in the master machine; then, we approximate the global MR objective function by a surrogate one, which can be efficiently evaluated by using the pilot sample and local gradients from worker machines. We obtain the final estimator by optimizing the surrogate MR objective function on the master machine, while worker machines only need to calculate and transmit the local gradients. Moreover, the communication cost is substantially reduced since only gradient vectors need to be transmitted, instead of Hessian matrices. The asymptotic properties are established under mild conditions, and they confirm that the final estimator is as efficient as the oracle obtained on the full data set as long as $n^2/N \to \infty$, where N is the global sample size.

The materials in the article are organized as follows. Section 2 introduces the new method, the related algorithm and the asymptotic properties. Section 3 reports the simulation studies and the real data example. Proofs of the theorem are presented in Appendix.

2 Distributed pilot modal regression

2.1 Problem setup

Let $\boldsymbol{Z} = (\boldsymbol{X}, Y)$, where $Y \in R$ is a response and $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in R^p$ is the covariate vector. Suppose that $\left\{\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)\right\}_{i=1}^N$ are N independent and identically distributed (i.i.d.) random samples from

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the parameter with true value $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$, the ϵ_is are i.i.d. and independent of $\boldsymbol{X}$ with $E[\epsilon_i|\boldsymbol{X}_i] = 0$.

In the distributed system, suppose that the observations $\{\boldsymbol{Z}_i\}_{i=1}^N$ are stored on K worker machines. Let $S = \{1, \ldots, N\}$ and denote S_k as the set of sample indices stored on the k-th worker machine. Suppose $S_{k_1} \cap S_{k_2} = \emptyset$ for $k_1 \neq k_2$ and $S = \cup_{k=1}^K S_k$. Let $N_k = |S_k|$, then, $N = \sum_{k=1}^K N_k$.

To obtain a robust estimator of $\boldsymbol{\beta}_0$, we propose the modal regression-based estimator (Yao et al., 2012; Yao and Li, 2014) as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{N}} = \arg\max_{\boldsymbol{\beta}\in R^p} \left\{ Q_N^h(\boldsymbol{\beta}) \right\}, \tag{1}$$

where $Q_N^h(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \phi_h\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $\phi_h(\cdot) = h^{-1}\phi(\cdot/h)$, $\phi(\cdot)$ is a kernel density function and $h > 0$ is the bandwidth, which can determine the efficiency and degree of robustness and will be selected by a data adaptive procedure.

Remark 1 The model studied in Yao and Li (2014) is $Mode(Y|x) = x^{\mathrm{T}}\beta$, and they need the bandwidth h in $Q_{\mathrm{N}}^h(\beta)$ to converge to zero to ensure the mode of kernel density function converges to the mode of the distribution of Y. For more studies on modal regression, we refer to Chen et al. (2016) and Feng et al. (2020). As a contrast, the model in our study is $E[Y|x] = x^{\mathrm{T}}\beta$, and we use the similar loss function as modal regression to gain the robustness. If the error distribution is symmetric about zero, the coefficient in modal linear regression will be the same as the coefficients obtained by conventional mean linear regression (Yao and Li, 2014). The idea of using modal regression-based loss is motivated by Yao et al. (2012), and the h in our study plays the same role as h_2 in Yao et al. (2012). So the h in our study is fixed and is selected by (5). The modal regression-based loss is commonly used in robust estimation, such as Zhao et al. (2014) and Wang et al. (2019).

In the distributed system, it is infeasible to solve (1). To realize fast computation, there are two commonly used strategies, i.e., the "one-shot" strategy and iterative algorithms. *However, they are built upon the basis that the data are randomly stored across different machines; otherwise, these methods are not suitable*. Specifically, define the local MR objective functions as $Q_k^h(\boldsymbol{\beta}) = \frac{1}{N_k}\sum_{i\in S_k} \phi_h\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $k = 1, \ldots, K$. The "one-shot" methods first obtain local estimators $\hat{\boldsymbol{\beta}}_k$ by maximizing $Q_k^h(\boldsymbol{\beta}), k = 1, \ldots, K$; then, the resulting estimator $\hat{\boldsymbol{\beta}}^{OS}$ is obtained by simply averaging the local estimators, i.e., $\hat{\boldsymbol{\beta}}^{OS} = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_k$. Notably, when the data are nonrandomly distributed across the worker machines, $\hat{\boldsymbol{\beta}}_k$s can be severely biased, which can lead to inconsistent results. The iterative algorithms, unlike the "one-shot" strategy, use a local Hessian matrix $\nabla^2 Q_k^h(\boldsymbol{\beta}) = \frac{\partial^2 Q_k^h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}$ to replace the global Hessian matrix $\nabla^2 Q_N^h(\boldsymbol{\beta}) = \frac{\partial^2 Q_N^h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}$, which necessitates a critical assumption of homogeneity, i.e.,

$$\left\| \nabla^2 Q_k^h(\boldsymbol{\beta}) - \nabla^2 Q_N^h(\boldsymbol{\beta}) \right\| \leqslant \delta,$$

where δ is a parameter that characterizes the homogeneity. To satisfy the condition, the local data used to construct $\nabla^2 Q_k^h(\boldsymbol{\beta})$ should be identically distributed as the entire data. In the nonrandomly distributed setting, however, the homogeneity assumption does not hold. *Therefore, we aim to solve the optimization problem* (1) *in a nonrandomly distributed manner.*

2.2 Pilot sample surrogate modal regression objective function

The key idea is to replace the global MR objective function $Q_{\mathrm{N}}^h(\boldsymbol{\beta})$ with a surrogate function that is communication-efficient and approximate to $Q_{\mathrm{N}}^h(\boldsymbol{\beta})$ even the data are

nonrandomly distributed. First, we randomly select a pilot sample with size n from all machines and store it in the master machine. The size n may be much smaller than N, and we assume that $n/N \to 0$ with $n \to \infty$ and $N \to \infty$. Denote $\mathcal{P}_k$ as the indices selected from S_k by random sampling without replacement, and $|\mathcal{P}_k| = n_k$. Denote the pilot sample as $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{P}_k$, and $|\mathcal{P}| = n = \sum_{k=1}^{K} n_k$. Then, the pilot estimator could be obtained by

$$\hat{\boldsymbol{\beta}}_{\mathcal{P}} = \arg\min_{\boldsymbol{\beta} \in R^p} \left\{ Q_{\mathcal{P}}^h(\boldsymbol{\beta}) \right\}, \tag{2}$$

where $Q_{\mathcal{P}}^h(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i \in \mathcal{P}} \phi_h\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}\right)$. Since $\mathcal{P}$ is completely randomly selected from all data, $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is consistent regardless of how the data are distributed on each machine. However, the convergence rate of $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is $\sqrt{n}$, which is much smaller than $\sqrt{N}$, the optimal rate.

In the second step, we regard the pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ as an initial estimator. By Taylor expansion around $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$, $Q_{\mathrm{N}}^h(\boldsymbol{\beta})$ can be represented as

$$Q_{\mathrm{N}}^h(\boldsymbol{\beta}) = Q_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) + \langle \nabla Q_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathcal{P}} \rangle + R_{\mathrm{N}}(\boldsymbol{\beta}), \tag{3}$$

where

$$\nabla Q_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in S_k} \boldsymbol{X}_i \dot{\phi}_h\left(Y_i - \boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathcal{P}}\right) = \frac{1}{N} \sum_{k=1}^{K} N_k \nabla Q_k^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}),$$

with $\dot{\phi}_h(\cdot)$ being the first derivative of $\phi_h(\cdot)$, and $\langle \cdot, \cdot \rangle$ denotes the inner product. In the distributed system, it requires one communication round to evaluate $\nabla Q_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}})$ and $R_{\mathrm{N}}(\boldsymbol{\beta})$ in (3). However, unlike the p-dim gradient vector, $R_{\mathrm{N}}(\boldsymbol{\beta})$ involves the calculation of global higher-order derivatives which require communicating more than $O(p^2)$ bits from each machine. To reduce the communication cost, we replace $R_{\mathrm{N}}(\boldsymbol{\beta})$ by a pilot sample version on the master machine,

$$R_{\mathcal{P}}(\boldsymbol{\beta}) = Q_{\mathcal{P}}^h(\boldsymbol{\beta}) - Q_{\mathcal{P}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) - \langle \nabla Q_{\mathcal{P}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathcal{P}} \rangle.$$

Then, we omit the additive constant in $Q_{\mathrm{N}}^h(\boldsymbol{\beta})$ and define the pilot sample surrogate MR objective function as

$$\widetilde{L}_N^h(\boldsymbol{\beta}) = Q_{\mathcal{P}}^h(\boldsymbol{\beta}) - \langle \nabla Q_{\mathcal{P}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) - \nabla Q_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathcal{P}}), \boldsymbol{\beta} \rangle.$$

Finally, we obtain the communication-efficient estimator by

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} = \arg\max_{\boldsymbol{\beta} \in R^p} \{ \widetilde{L}_{\mathrm{N}}^h(\boldsymbol{\beta}) \}. \tag{4}$$

Remark 2 The modal regression-based estimator is a widely used robust method, and we have applied this method to longitudinal data (Wang et al., 2019) and randomly distributed data (Wang and Li, 2021) before. However, when the data are nonrandomly stored, the estimator proposed by Wang and Li (2021) will fail. To the best of our knowledge, this is the first to study the robust estimation for nonrandomly distributed data. In this study, to overcome the challenges imposed by

distribution nonrandomness, we adapt the pilot sampling which is also used in Wang et al. (2020) and Pan et al. (2021). In their studies, they first calculated an initial estimate by the pilot sample and then upgraded the estimate by a one-step Newton–Raphson-type algorithm. Unlike them, the pilot sample in this study is used not only to provide an initial estimator, but also, more importantly, to update the Hessian matrix at each iteration. Therefore, our method does not require the transmission of the Hessian matrices and thus has a communication advantage.

2.3 Asymptotic properties and algorithm

The following regularity conditions are required for the theoretical development.

(A1) The parameter space $\mathscr{H}$ is a compact subset of R^p and the true value $\boldsymbol{\beta}_0$ lies in the interior of $\mathscr{H}$.
(A2) There exists a positive definite matrix $\boldsymbol{\Sigma}$ such that $\frac{1}{n}\sum_{i\in\mathcal{P}} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \to_p \boldsymbol{\Sigma}$, as $n \to \infty$, and $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \to_p \boldsymbol{\Sigma}$, as $N \to \infty$, where "$\to_p$" denotes convergence in probability.
(A3) The random error satisfies that $E[\dot{\phi}_h(\epsilon)|\boldsymbol{X}] = 0$.

Condition (A1) is elementary. Condition (A2) is a standard condition for proving estimation consistency and asymptotic normality. Condition (A3) ensures the consistency of the estimator, which is also used in Yao et al. (2012) and Zhao et al. (2014).

Theorem 1 *Suppose Conditions A1–A3 hold, we have*

(a) $\|\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}\| = O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}\|$;
(b) *if* $n/\sqrt{N} \to \infty$, *then* $\sqrt{N}(\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0) \to_d N(\mathbf{0}, \xi(h)\boldsymbol{\Sigma}^{-1})$, *where* $\xi(h) = \frac{E(\dot{\phi}_h^2(\epsilon))}{[E(\ddot{\phi}_h(\epsilon))]^2}$ *and* "$\to_d$" *denotes convergence in distribution.*

Theorem 1 shows that if we use the pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ as our initial estimator, the final estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{N}}$ can significantly match the accuracy of the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$, and it can achieve the optimal rate of convergence.

We use the Newton–Raphson algorithm on the master machine to solve (4). The iterative procedure is summarized in Algorithm 1.

Input: Bandwidth h, initial value $\boldsymbol{\beta}^{(0)}$ and the maximum number of iterations T;
for $t = 0, \cdots, T-1$ **do**
  Update $\boldsymbol{\beta}^{(t+1)}$ via

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\nabla^2 \widetilde{L}_N^h(\boldsymbol{\beta}^{(t)})\right]^{-1} \left[\nabla \widetilde{L}_N^h(\boldsymbol{\beta}^{(t)})\right];$$

end
Output: $\boldsymbol{\beta}^{(T)}$

Next, we discuss how to select the bandwidth h. Based on Theorem 1, the optimal bandwidth is $h_{opt} = \arg\min_h \xi(h)$. In practice, we replace $\xi(h)$ by its estimator

$$\hat{\xi}(h) = \left[\frac{1}{n}\sum_{i\in\mathcal{P}} \ddot{\phi}_h(\bar{\epsilon}_i)\right]^{-2} \left[\frac{1}{n}\sum\nolimits_{i\in\mathcal{P}}[\dot{\phi}_h(\bar{\epsilon}_{ki})]^2\right],$$

where $\bar{\epsilon}_i = Y_i - \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathcal{P}}$. Then, the optimal bandwidth could be found by

$$\hat{h}_{opt} = \arg\min_h \hat{\xi}(h). \tag{5}$$

This selection strategy is similar to those of Yao et al. (2012); Wang and Lin (2016). The detailed algorithm of the proposed distributed PMR estimator is given in Algorithm 2.

The first step

for $i = 1, \cdots, N$ **do**
  Generate $\delta_i \sim Bernoulli(1, n/N)$;
  if $\delta_i = 1$ **then**
    add $\boldsymbol{Z}_i$ to the pilot sample set $\mathcal{P}$;
  end
end
Based on $\mathcal{P}$, calculate the pilot estimator by (2).

The second step

Input: Initial value $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}_{\mathcal{P}}$ and the number of rounds of communication T;
for $t = 0, \cdots, T-1$ **do**
  Broadcast the current value $\boldsymbol{\beta}^{(t)}$ to worker machines;
  Calculate $\hat{\xi}(h)$ at grid points, update $h^{(t)}$ via (5) and broadcast it to worker machines;
  Calculate gradient vectors $\nabla Q_k^{h^{(t)}}(\boldsymbol{\beta}^{(t)})$, $k = 1, \cdots, K$ on worker machines;
  Transmit $\nabla Q_k^{h^{(t)}}(\boldsymbol{\beta}^{(t)})$s to the master machine and calculate the current value of $\nabla Q_{\mathcal{P}}^{h^{(t)}}(\boldsymbol{\beta}^{(t)}) - \nabla Q_N^{h^{(t)}}(\boldsymbol{\beta}^{(t)})$;
  Update the pilot sample surrogate objective function

$$\widetilde{L}_N^{(t)}(\boldsymbol{\beta}) = Q_{\mathcal{P}}^{h^{(t)}}(\boldsymbol{\beta}) - \langle \nabla Q_{\mathcal{P}}^{h^{(t)}}(\boldsymbol{\beta}^{(t)}) - \nabla Q_N^{h^{(t)}}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}\rangle;$$

  Update $\boldsymbol{\beta}^{(t+1)} = \arg\min_{\boldsymbol{\beta}\in R^p}\{\widetilde{L}_N^{(t)}(\boldsymbol{\beta})\}$ on the master machine by the Algorithm 1;
end
Output: $\boldsymbol{\beta}^{(T)}$.

3 Simulation studies and application

3.1 Simulation studies

In this section, we conduct several simulation studies to investigate the finite sample performance of the proposed method. We generate the data by

$$Y_i = X_i^T \boldsymbol{\theta} + \epsilon_i,$$

where $X_i = (X_{i1}, X_{i2}, X_{i3})^T$, and each component $X_{ij}, (1 \leqslant i \leqslant N, j = 1, 2, 3)$ follows a uniform distribution $U(0, 1)$. We set $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}} = (1, 2, -1)^T$. Three different error distributions are considered to study the robustness of the estimator.

Example 1 $\varepsilon \sim N(0, 1)$.

Example 2 $\varepsilon \sim t(3)$, where $t(3)$ denotes the t-distribution with three degrees of freedom.

Example 3 The error follows the Cauchy distribution.

We set the total sample size $N = 50 \times 10^4$, and split the data into $K \in \{5, 20, 50, 100, 250, 500\}$ block. We consider three typical data storing strategies for each example. The first strategy stores data randomly so that the samples on different machines are i.i.d. In contrast, the other two strategies store data nonrandomly.

Strategy 1. (Randomly Distributed). We distribute all the samples $\{(Y_i, X_i), 1 \leqslant i \leqslant N\}$ in a completely random manner.
Strategy 2. (Completely Nonrandomly Distributed). Let $D_i = \sum_{j=1}^{3} X_{ij}$, and $D_{(i)}$ be the i-th order statistic, i.e, $D_{(1)} \leqslant \cdots \leqslant D_{(N)}$. Then, we store $(X_{(i)}, Y_{(i)})$ on the $([iK/N] + 1)$-th machine. Notably, all the samples are nonrandomly distributed in this strategy.
Strategy 3. (Partially Nonrandomly Distributed). Let $X_{(i)1}$ be the i-th order statistic of X_{i1} and store $(X_{(i)1}, X_{(i)2}, X_{(i)3}, Y_{(i)})$ on the $([iK/N] + 1)$-th machine. Thus, the observations of X_1 are stored nonrandomly and the observations of (X_2, X_3) are stored randomly.

We compare the proposed estimator (NEW) with the following: (a) the global modal regression estimator (GMR), (b) the global ordinary least squares estimator (GLO), (c) the pilot modal regression estimator (PMR), and (d) the average distributed modal regression estimator (ADMR). The experiment is repeated by 500 times. Let $\hat{\theta}_i^{(s)}$ be the estimator of θ_i in the s-th replication, and define the average estimation error (AEE) as $AEE(\hat{\theta}_i) = 500^{-1} \sum_{s=1}^{500} (\hat{\theta}_i^{(s)} - \theta_i)^2$.

Comparison of estimation efficiency and robustness. Tables 1, 2 and 3 present the relative AEEs of the estimators (GLO, NEW, PMR, ADMR) to the GMR for Examples 1–3, respectively. The pilot sample percentage $\rho = n/N$ is set as 5%. From these

Table 1 Relative AEEs for normal distribution

K		Strategy 1			Strategy 2			Strategy 3		
	GLO	PMR	ADMR	NEW	PMR	ADMR	NEW	PMR	ADMR	NEW
θ_1										
5	1.02	9.48	1.08	1.06	11.35	1.19	1.18	9.42	3.03	1.06
20	0.96	10.45	1.04	1	11.16	1.26	1.02	11.69	6.11	1.07
50	1.01	9.79	1.06	1.06	8.71	1.35	1.01	9.97	7.18	0.99
100	0.96	8.16	1.01	0.99	9.09	1.22	0.98	8.71	12.93	1.09
250	0.94	10.04	1.04	1.01	9.45	1.29	1.06	10.87	16.38	1.06
500	1.02	9.95	1.06	1.07	10.36	1.36	1.04	11.21	14.66	1
θ_2										
5	0.98	9.99	1.07	1.04	9.45	1.12	1.11	10.82	1.28	1.13
20	1.02	11.7	1.08	1.07	12.43	1.39	1	11.09	1.19	1.14
50	0.99	8.86	1.03	1.02	8.66	1.23	1.05	10.13	1.21	1.09
100	0.92	9.98	1.01	1.03	9.57	1.33	1.02	10.35	1.17	1.08
250	0.97	12.38	1.05	1.02	11.18	1.27	1.03	11.59	1.3	1.09
500	1.02	8.95	1.04	0.94	9.03	1.38	1.03	9.88	1.17	1.1
θ_3										
5	0.88	11.04	0.96	0.96	12.16	1.05	1.03	11.22	1.13	1.04
20	0.93	10	1.03	1.05	11.27	1.21	1	10.55	1.23	1.06
50	0.98	9.44	1.05	1.03	10.84	1.24	1.04	10.84	1.18	1.08
100	1.05	10.69	1.11	1.11	10.11	1.43	1.16	11.07	1.37	1.08
250	0.98	11.81	1.07	1.08	9.31	1.38	1.03	11.73	1.32	0.99
500	0.99	11.23	1.15	1.06	10.92	1.35	1.07	10.13	1.27	1.08

tables, we draw the following conclusions. First, our new estimator is as accurate as the global estimator in all examples and storage strategies, because the corresponding relative AEEs are always close to 1. Second, because the pilot modal regression estimator only uses the pilot sample, it always performs much worse than our new method, and the corresponding relative AEEs are always much higher than 1; in fact, the AEEs of the PMR are about 10 times those of the NEW. Third, the DC-based method also performs worse than our method in all the settings, especially when the data are distributed nonrandomly, i.e., Strategies 2 and 3. Furthermore, regarding the robustness, when the error distribution is standard normal, our new method performs comparably to the GLO method. When the error distribution is $t(3)$, the GLO performs worse than our method. GLO does not work at all when the error distribution is Cauchy, while our method works well in such situations.

Effect of pilot sample size. To illustrate the influence of pilot sample size n, different values of the pilot percentage are considered. We take $\rho = 0.5\%, 1\%, 2\%, 5\%$, 10%, and 20% for illustration. Figures 1, 2 and 3 present the AEEs of different estimators for the three examples. We can see our estimator always performs as well as the global estimator even with only 1% of the total sample being the pilot sample. The pilot estimator, however, is very sensitive to the pilot percentage. When

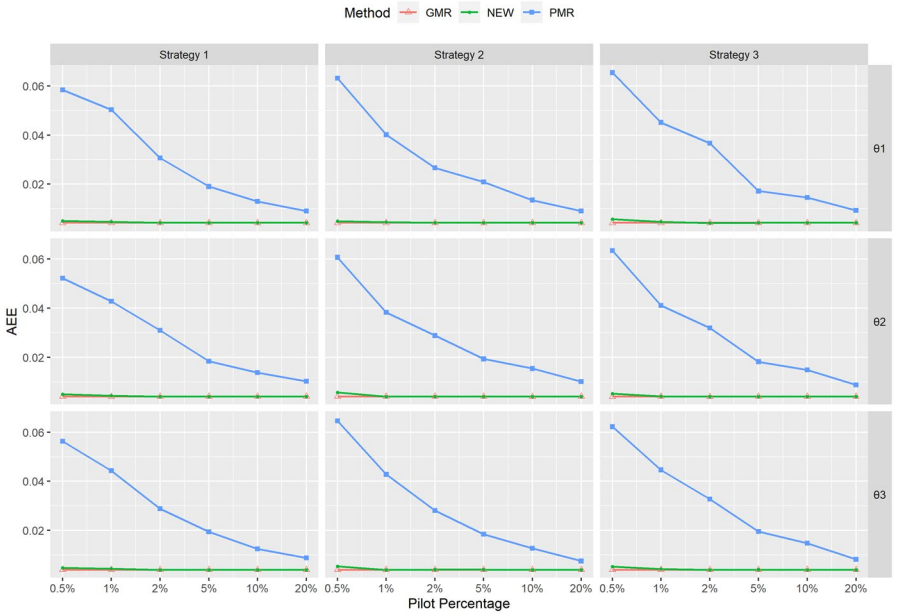

Fig. 1 AEEs for normal distribution

the pilot percentage is small, the AEEs of pilot estimator are much lager than ours. Even when 20% of the total sample are used as the pilot sample, the pilot estimator's AEEs are still bigger than ours.

3.2 Real data analysis

In this section, we apply the proposed method to analyze the greenhouse gas (GHG) data. This data set is from the UCI machine learning repository, and consists of 954, 840 observations. The response variable is the GHG concentration of synthetic observations, and the predictors are GHG concentrations of tracers emitted from a region outside of California and 14 distinct regions in California. Our goal is to predict the GHG concentrations of synthetic observations.

We compare our estimator with the global modal regression estimator (GMR) and the pilot modal regression estimator (PMR) by the prediction accuracy. We randomly split the data set into two part: a training set and a testing set. The training set consists with 500, 00 observations and is evenly split into $K = 100$ subsets to mimic a distributed system. The coefficients are estimated using the training data set D_{train} and the average prediction error (APE) , the average of $\{(\hat{Y}_i - Y_i)^2, i \in D_{\text{test}}\}$, is calculated based on the test data set D_{test}.

Table 4 summarizes the APEs of the three estimators with six pilot percentages from 0.5 to 20%. We can see that the prediction errors of our estimator are similar to

Table 2 Relative AEEs for t(3) distribution

K		Strategy 1			Strategy 2			Strategy 3		
	GLO	PMR	ADMR	NEW	PMR	ADMR	NEW	PMR	ADMR	NEW
θ_1										
5	1.42	9.58	1.05	1.03	9.5	1.07	1.05	10.03	2.49	1.03
20	1.68	10.63	1.02	1.03	10.34	1.2	1.08	11.69	6.83	1.13
50	1.56	10.22	1.16	1.12	11.09	1.33	1.07	10.39	9.65	1.03
100	1.57	11.08	1.11	1.09	11.8	1.32	1.03	10.91	14.59	1.13
250	1.26	9.31	1.12	1.08	10.49	1.18	1.08	8.74	15.72	1.01
500	1.34	10.59	1.06	1.05	10.04	1.31	1.04	10.22	13.3	1.07
θ_2										
5	1.37	9.06	1.05	1.04	9.11	1.1	1.08	9.71	1.12	1.04
20	1.36	10.39	1.04	1.06	9.96	1.17	1.14	10.83	1.23	1.07
50	1.36	10.15	1.06	1.05	10.23	1.24	1.03	11.46	1.22	1.1
100	1.42	10.52	1.15	1.10	10.51	1.28	1.04	10.69	1.19	1.07
250	1.46	9.91	1.08	1.08	9.94	1.31	1.1	9.84	1.22	1.07
500	1.48	10.25	1.09	1.04	10.6	1.43	1.04	10.15	1.34	1.14
θ_3										
5	1.47	11.09	1.03	1.02	10.99	1.1	1.07	11.43	1.17	1.08
20	1.48	9.78	1.02	1.04	11.31	1.16	1.07	10.71	1.01	1.1
50	1.44	11.7	1.07	1.05	10.53	1.29	1.05	10.4	1.29	1.1
100	1.68	11.86	1.11	1.02	12.66	1.5	1.19	11.65	1.36	1.11
250	1.34	9.82	1.16	1.12	10.34	1.39	1.11	9.85	1.2	1.04
500	1.27	9.33	1.09	1.09	10.17	1.27	1.07	10.45	1.28	1.03

the global estimator's. When we only take 0.5% of the total sample as the pilot sample, our method still works well, while the PMR performs much worse.

4 Summary and discussions

In this article, we proposed a distributed pilot modal regression estimator for nonrandomly distributed data. This estimator has three advantages: (1) it can achieve both efficiency and robustness by introducing a tuning parameter that is automatically selected by a data-driven approach; (2) it is communication-efficient and is statistically as efficient as the global estimator; (3) by using the pilot sample, it can adapt even though the data are stored nonrandomly. However, we did not consider the high-dimensional data, which are very common in the era of massive data. This will comprise our future research work.

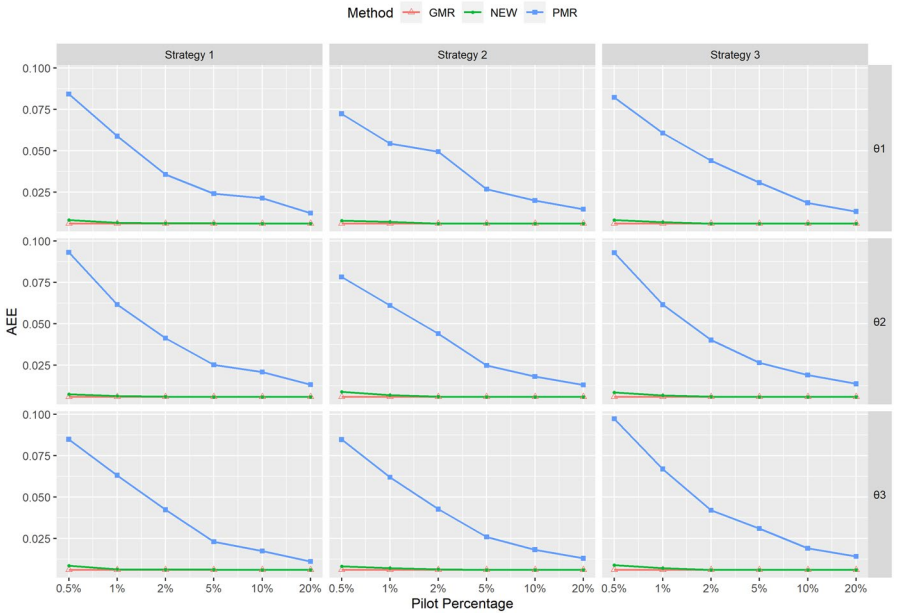

Fig. 2 AEEs for t(3) distribution

In addition to the modal regression-based estimator, there are many other robust methods, such as the Huber regression. We did not use it in this study for two reasons. First, the Huber regression would lose some efficiency when there are no outliers or the error distribution is normal, while the modal regression-based estimator can achieve both robustness and efficiency by introducing a tuning parameter. Second, the Huber loss function is not twice differentiable and thus not able to be applied in our algorithm. Thus, we studied the modal regression-based robust estimation for distributed data. Nonetheless, the algorithm for Huber regression in the nonrandomly distributed framework is an exciting study and is worth further exploration.

Acknowledgements The research was supported by NNSF project of China (12101056 and 11901356).

Appendix

Proof of Theorem 1 First, we compute the order of $\|\nabla^2 Q_{\mathrm{N}}^h(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathcal{P}}^h(\boldsymbol{\beta}_0)\|_\infty$. Let $\boldsymbol{\Sigma} = E(\boldsymbol{X}\boldsymbol{X}^T)$, we have that

$$\begin{aligned}&\|\nabla^2 Q_{\mathrm{N}}^h(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathcal{P}}^h(\boldsymbol{\beta}_0)\|_\infty \\ &= O_p\left(\left\|\boldsymbol{\Sigma} - \frac{1}{N}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right\|_\infty\right) + O_p\left(\left\|\frac{1}{n}\widetilde{\boldsymbol{X}}_{\mathcal{P}}\widetilde{\boldsymbol{X}}_{\mathcal{P}}^T - \boldsymbol{\Sigma}\right\|_\infty\right) + O_p(N^{-1/2}),\end{aligned}$$

where $\widetilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{X}}_{\mathcal{P}} = (\boldsymbol{X}_i, i \in \mathcal{P})^{\mathrm{T}}$. It is easy to obtain

Table 3 Relative AEEs for Cauchy distribution

K		Strategy 1			Strategy 2			Strategy 3		
	GLO	PMR	ADMR	NEW	PMR	ADMR	NEW	PMR	ADMR	NEW
θ_1										
5	5621	11.39	1.03	1.05	11.87	1.04	1.1	11.35	3.16	1.18
20	10936	10.56	1.06	1.06	10.84	1.11	1.14	10.02	6.44	1.09
50	11105	8.53	1.04	1.02	8.73	1.15	1.15	9.99	9.23	0.99
100	3248	10.21	1.06	1.07	9.5	1.31	1.11	9.26	11.23	1.1
250	4016	12.55	1.14	1.11	11.04	1.23	1.21	10.52	15.77	1.11
500	29892	12.3	1.06	1.05	10.73	1.41	1.13	9.21	11.7	1.04
θ_2										
5	41754	10.82	1.09	1.06	10.83	1.3	1.15	10.74	1.19	1.18
20	8268	8.95	1.09	1.07	9.67	1.24	1.09	10.31	1.26	1.07
50	6844	11.49	1.10	1.10	11.71	1.42	1.18	11.92	1.33	1.08
100	4130	8.94	1.18	1.14	11.41	1.33	1.19	9.9	1.23	1.07
250	2589	11.57	1.16	1.10	12.1	1.46	1.26	12.55	1.28	1.07
500	25315	12.82	1.09	1.11	11.07	1.33	1.16	10.96	1.36	1.22
θ_3										
5	16563	12.06	1.07	1.11	11.91	1.21	1.2	11.86	1.22	1.16
20	11624	9.1	1.05	0.99	11.02	1.22	1.05	10.2	1.24	1.12
50	10049	8.15	1.13	1.03	9.54	1.24	1.06	9.14	1.25	1.07
100	6169	11.51	1.17	1.11	9.65	1.17	1.08	10.19	1.28	1.18
250	1944	12.99	1.16	1.12	11.56	1.22	1.18	12.15	1.36	1.13
500	23525	9.99	1.07	1.01	9.86	1.27	1.08	9.76	1.36	1.15

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_{ij}X_{ik} - \mathbf{\Sigma}_{jk}\right| > t\right) \leqslant \exp(-c_1 \min(t^2, t)N),$$

where c_1 is a constant that depends on $\mathbf{\Sigma}$. By a union bound over all (j, k) pairs,

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T - \mathbf{\Sigma}\right| > t\right) \leqslant \exp(2\log p - c_1 \min(t^2, t)N).$$

Thus, letting $t = C\sqrt{\frac{\log p}{N}}$, we have $O_p(\|\mathbf{\Sigma} - \frac{1}{N}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\|_\infty) = O_p(N^{-1/2})$. By a similar argument, $O_p(\|\frac{1}{n}\widetilde{\boldsymbol{X}}_{\mathcal{P}}\widetilde{\boldsymbol{X}}_{\mathcal{P}}^T - \mathbf{\Sigma}\|_\infty) = O_p(n^{-1/2})$. Then, we can get that

$$\left\|\nabla^2 Q_{\mathrm{N}}^h(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathcal{P}}^h(\boldsymbol{\beta}_0)\right\|_\infty = O_p\left(n^{-1/2}\right).$$

By applying Lemma 6 in Zhang et al. (2013) with $F_1 = \widetilde{L}_{\mathrm{N}}^h(\boldsymbol{\beta})$ in the notation therein, we can also obtain

$$\left\|\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}\right\| = O_p\left(\left\|\nabla\widetilde{L}_{\mathrm{N}}^h(\hat{\boldsymbol{\beta}}_{\mathrm{N}})\right\|\right).$$

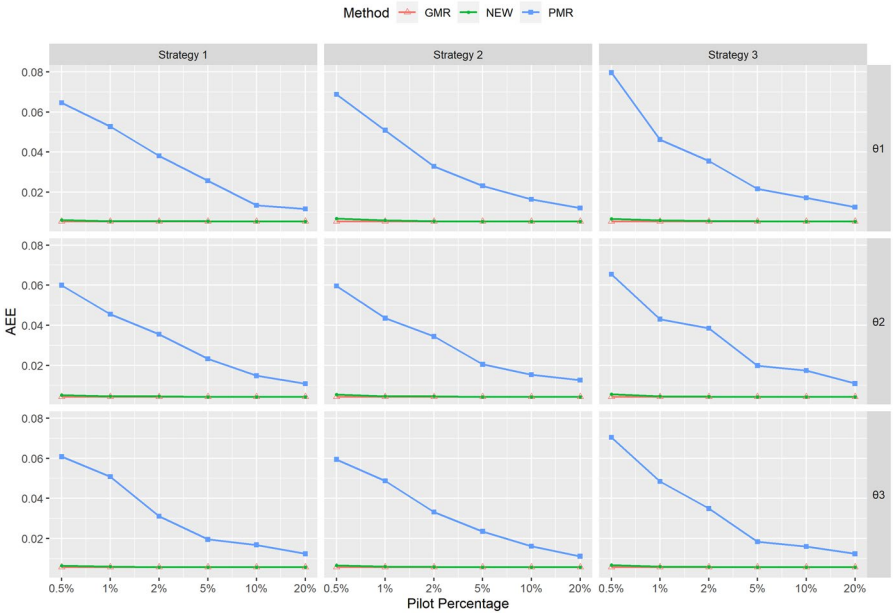

Fig. 3 AEEs of three estimators for Cauchy distribution

Table 4 APE of the global estimator (GMR), pilot modal regression estimator (PMR) and the proposed estimator (NEW) for analysis of the greenhouse gas data set

	0.5%	1%	2%	5%	10%	20%
NEW	16.173	16.168	16.142	16.132	16.131	16.131
PMR	16.844	16.737	16.620	16.617	16.421	16.309
GMR	16.130					

A simple calculation yields

$$\nabla\widetilde{L}^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) = \nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) - \nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) + \nabla Q^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}),$$

and note that $\nabla Q^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) = \mathbf{0}$, we obtain

$$\nabla\widetilde{L}^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) = \big(\nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) - \nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}})\big) - \big(\nabla Q^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) - \nabla Q^h_{\mathrm{N}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}})\big).$$

By the integral form of Taylor's expansion, we have

$$\nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) - \nabla Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) = \boldsymbol{H}_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}) \text{ and } \nabla Q^h_{N}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) - \nabla Q^h_{N}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}) = \boldsymbol{H}_{N}(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}),$$

where $\boldsymbol{H}_{\mathcal{P}} = \int_0^1 \nabla^2 Q^h_{\mathcal{P}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}} + t(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}))dt$ and $\boldsymbol{H}_{N} = \int_0^1 \nabla^2 Q^h_{N}(\hat{\boldsymbol{\beta}}_{\mathcal{P}} + t(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}))dt$ satisfy $\|\boldsymbol{H}_{\mathcal{P}} - \nabla^2 Q^h_{\mathcal{P}}(\boldsymbol{\beta}_0)\| = O_p(\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| + \|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0\|)$ and $\|\boldsymbol{H}_{N} - \nabla^2 Q^h_{N}(\boldsymbol{\beta}_0)\| = O_p(\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| + \|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0\|)$, respectively. Thus, we have

$$
\begin{aligned}
&\left\|\nabla\widetilde{L}_{\mathrm{N}}^{h}(\hat{\boldsymbol{\beta}}_{\mathrm{N}})\right\| \\
&= \left\|(\boldsymbol{H}_{\mathcal{P}} - \nabla^2 Q_{\mathcal{P}}^{h}(\boldsymbol{\beta}_0))(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}) - (\boldsymbol{H}_{N} - \nabla^2 Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0))(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}})\right. \\
&\quad \left.+(\nabla^2 Q_{\mathcal{P}}^{h}(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0))(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}})\right\| \\
&\leqslant \|\boldsymbol{H}_{\mathcal{P}} - \nabla^2 Q_{\mathcal{P}}^{h}(\boldsymbol{\beta}_0)\|\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| + \|\boldsymbol{H}_{N} - \nabla^2 Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0)\|\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| \\
&\quad + \|\nabla^2 Q_{\mathcal{P}}^{h}(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0)\|\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| \\
&= (O_p(\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\|) + O_p(\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\|) + \|\nabla^2 Q_{\mathcal{P}}^{h}(\boldsymbol{\beta}_0) - \nabla^2 Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0)\|)\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| \\
&= O_p\left(\frac{1}{\sqrt{n}}\right)\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\|.
\end{aligned}
$$

Now, we complete the proof of (a).

Together with

$$
\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0 = -\frac{1}{E(\ddot{\phi}_h(\epsilon))}\boldsymbol{\Sigma}^{-1}\nabla Q_{\mathrm{N}}^{h}(\boldsymbol{\beta}_0) + O_p\Big(\frac{1}{N}\Big),
$$

we can get that

$$
\begin{aligned}
&\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0 \\
&= (\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}) + (\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0) \\
&= -\frac{1}{E(\ddot{\phi}_h(\epsilon))}\boldsymbol{\Sigma}^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\dot{\phi}_h(\epsilon_i)\Big) + O_p\Big(\frac{1}{N} + n^{-1/2}\|\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}\|\Big).
\end{aligned}
$$

Under the assumptions $n/\sqrt{N} \to \infty$ and $\|\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \hat{\boldsymbol{\beta}}_{\mathcal{P}}\| = O_p(n^{-1/2})$, we can obtain that $\sqrt{N}(\frac{1}{N} + n^{-1/2}\|\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \hat{\boldsymbol{\beta}}_{\mathrm{N}}\|) = o_p(1)$. Thus, we have that

$$
\begin{aligned}
&\sqrt{N}(\widetilde{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}_0) \\
&= -\frac{1}{E(\ddot{\phi}_h(\epsilon))}\boldsymbol{\Sigma}^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{X}_i\dot{\phi}_h(\epsilon_i)\right) + o_p(1) \\
&\to_d N(\mathbf{0}, \xi(h)\boldsymbol{\Sigma}^{-1}),
\end{aligned}
$$

where $\xi(h) = \frac{E(\dot{\phi}_h^2(\epsilon))}{[E(\ddot{\phi}_h(\epsilon))]^2}$. The proof of (b) is completed. □

Declarations

Conflict of interest The authors declare that there are no conflict of interests.

References

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

Chen, X., Liu, W., Zhang, Y. (2019). Quantile regression under memory constraint. *Annals of Statistics, 47,* 3244–3273.

Chen, Y., Genovese, C., Tibshirani, R., Wasserman, L. (2016). Nonparametric modal regression. *Annals of Statistics, 44,* 489–514.

Duchi, J., Jordan, M., Wainwright, M., Zhang, Y. (2014). Optimality guarantees for distributed statistical estimation. arXiv preprint arXiv:1405.0782.

Fan, J., Wang, D., Wang, K., Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *Annals of statistics, 47,* 3009.

Fan, J., Guo, Y., Wang, K. (2021). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.1969238.

Feng, Y., Fan, J., Suykens, J. (2020). A statistical learning approach to modal regression. *Journal of Machine Learning Research, 21*(2), 1–35.

Gopal, S., Yang, Y. (2013). Distributed training of large-scale logistic models. In: *International Conference on Machine Learning*, pp. 289–297.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Jordan, M. I., Lee, J. D., Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association, 14,* 668–681.

Koenker, R., Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*: *Journal of the Econometric Society, 46,* 33–50.

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

Pan, R., Ren, T., Guo, B., Li, F., Li, G., Wang, H. (2021). A note on distributed quantile regression by pilot sampling and one-step updating. *Journal of Business and Economic Statistics*. https://doi.org/10.1080/07350015.2021.1961789.

Shamir, O., Srebro, N., Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. *International Conference on Machine Learning, 32,* 1000–1008.

Tu, J., Liu, W., Mao, X., Chen, X. (2021). Variance reduced median-of-means estimator for byzantine-robust distributed inference. *Journal of Machine Learning Research, 22*(84), 1–67.

Wang, F., Huang, D., Zhu, Y., Wang, H. (2020). Efficient estimation for generalized linear models on a distributed system with nonrandomly distributed data. arXiv preprint arXiv:2004.02414.

Wang, J., Kolar, M., Srebro, N., Zhang, T. (2017). Efficient distributed learning with sparsity. *International Conference on Machine Learning, 70,* 3636–3645.

Wang, K., Li, S. (2021). Robust distributed modal regression for massive data. *Computational Statistics and Data Analysis, 160,* 107225.

Wang, K., Lin, L. (2016). Robust structure identification and variable selection in partial linear varying coefficient models. *Journal of Statistical Planning and Inference, 174,* 153–168.

Wang, K., Li, S., Sun, X., Lin, L. (2019). Modal regression statistical inference for longitudinal data semivarying coefficient models: Generalized estimating equations, empirical likelihood and variable selection. *Computational Statistics and Data Analysis, 133,* 257–276.

Wang, K., Wang, H., Li, S. (2022). Renewable Quantile Regression for Streaming Datasets. *Knowledge-based Systems, 235,* 107675.

Wang, K., Zhang, B., Sun, Xiao, Li, S. (2022). Efficient statistical estimation for a non-randomly distributed system with application to large-scale data neural network. *Expert Systems With Applications, 197,* 116698.

Yao, W., Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics, 41,* 656–671.

Yao, W., Lindsay, B., Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics, 24,* 647–663.

Zhang, Y., Duchi, J. C., Wainwright, M. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research, 14,* 3321–3363.

Zhao, W., Zhang, R., Liu, J., Lv, Y. (2014). Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Annals of the Institute of Statistical Mathematics, 66,* 165–191.

Zhu, X., Li, F., Wang, H. (2021). Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics, 30*(4), 1004–1018.