Statistical Papers (2025) 66:1
https://doi.org/10.1007/s00362-024-01625-2

REGULAR ARTICLE

Composite quantile regression for a distributed system with non-randomly distributed data

Jun Jin[1] · **Chenyan Hao**[1] · **Yewen Chen**[2]

Received: 31 July 2024 / Revised: 13 September 2024 / Published online: 4 December 2024

Abstract

The composite quantile regression estimator is widely acknowledged for its robustness and efficiency, offering a compelling alternative to both ordinary least squares and quantile regression estimators in linear models. However, when data is not randomly distributed across different workers in distributed settings, existing methods for composite quantile regression become statistically inefficient. To address this limitation, we present a novel one-step upgraded pilot composite quantile regression method. Our proposed approach involves two essential steps. In the first step, we obtain a pilot estimator by leveraging a small random sample collected from different workers. Subsequently, in the second step, we perform one-step updating based on the pilot estimator, involving the summarization of sample moment quantities on each worker. The resulting estimator is theoretically proven to be as statistically efficient as the composite quantile regression estimator using the entire sample, without relying on restrictive assumptions about randomness. Furthermore, the resulting estimator inherits the robustness and efficiency advantages of the composite quantile regression estimator, while also being computationally efficient in terms of communication cost and storage usage. To validate the practical performance of our proposed method, we conduct numerical studies using simulated and real data, demonstrating its effectiveness in real-world scenarios.

Keywords Communication efficiency · Distributed system · Distributed data · Robust estimation · Statistical efficiency.

✉ Jun Jin
junjin@yzu.edu.cn

[1] College of Mathematical Sciences, Yangzhou University, Yangzhou, China

[2] College of Public Health, University of Georgia, Athens, GA, USA

1 Introduction

1.1 Background

The swift progress in scientific and technological fields has resulted in the ongoing production of big data from various domains, including social media, e-commerce, and health (Gopal and Yang 2013; Battey et al. 2018; Yu et al. 2024). This “data deluge" has resulted in data being spread across multiple servers. For instance, sensor data is collected by multiple sensor nodes situated at different locations, and the collected data is typically processed and stored using distributed systems (Li et al. 2018). Moreover, local processors operate at speeds that can be several thousand times higher than modern network data transmission rates, making traditional centralized systems inadequate to address the challenges posed by big data (Jaggi et al. 2014). Consequently, distributed systems have emerged as a powerful solution. A distributed system is composed of interconnected computers that function as a cohesive system through network communication. In this context, we focus on the widely used "master-and-worker" architecture of distributed systems (e.g., Hadoop, Spark, etc.). In this architecture, one computer functions as the master, with the others operating as workers, and no direct communication is required among the workers. The data is distributed across these workers, and most of the computations are performed on the workers, with the master primarily responsible for coordinating their activities (Jordan et al. 2018; Zhu et al. 2021; Pan et al. 2022; Wang et al. 2022, 2023).

1.2 Related works

For this “master-and-worker"-type distributed system, the primary communication costs arise between the master and workers. In recent years, researchers have proposed numerous distributed statistical optimization methods tailored specifically for this type of distributed system. These methods can be categorized into two main categories. The first category is known as the one-shot (OS) approach, which is alternatively referred to as divide-and-conquer in some literature. The OS approach involves two key steps. First, the workers independently compute local estimators in parallel using the available data on each worker; then, the master aggregates these local estimators to derive a global estimator through simply averaging. Numerous studies have explored the OS versions of classical statistical methods. For instance, researchers have applied this approach to linear regression (Zhang et al. 2013; Chang et al. 2017), generalized linear regression (Chen and Xie 2014; Battey et al. 2018), semiparametric regression (Zhao et al. 2016), quantile regression (Chen et al. 2019; Chen and Zhou 2020; Xu et al. 2020; Hou et al. 2023), support vector machines (Wang et al. 2019), principal component analysis (Fan et al. 2019), among others.

The OS estimator offers several advantageous characteristics. Firstly, it boasts straightforward implementation in practical scenarios. Secondly, it demonstrates highly efficient communication, necessitating just a single round of interaction between the master and workers. Lastly, given suitable regularity conditions, it can achieve notable statistical efficiency. However, the exceptional statistical efficiency of OS estimators relies on several critical conditions. To elaborate, there are three crucial factors: (1) It requires a constraint on the number of workers, denoted as K, such that $K = o(\sqrt{N})$, where N signifies the total sample size. (2) Simple averaging often proves inadequate for nonlinear estimators. (3) The data must be randomly distributed across different workers (Zhang et al. 2012; Rosenblatt and Nadler 2016; Wang et al. 2017; Jordan et al. 2018; Huang and Huo 2019). In a bid to relax the constraints mentioned earlier (particularly (1) and (2)), Shamir et al. (2014) and Wang et al. (2017) proposed the second category of approach, which is the so-called iterative approach. This method involves several exchanges between the master and workers, thus improving the efficiency of the resulting statistical estimator. Expanding upon this concept, Jordan et al. (2018) proposed a communication-efficient surrogate likelihood framework for addressing distributed statistical inference problems. Fan et al. (2021) conducted further investigations into two accurate statistical estimators by incorporating an additional regularization term into the approximate likelihood used in Jordan et al. (2018).

The ordinary least squares (OLS) estimator has proven to be a statistically efficient tool for linear models characterized by Gaussian errors or other light-tailed errors. However, it exhibits inconsistency when confronted with heavy-tailed errors and is particularly sensitive to the presence of outliers. As an alternative to OLS, quantile regression (QR) was introduced by Koenker (2005) and has gained recognition for its robustness against heavy-tailed distributions and outliers. Nonetheless, it is essential to note that the relative statistical efficiency of the QR estimator compared to the OLS estimator can be arbitrarily small. To address the limitations of QR, Zou and Yuan (2008) proposed composite quantile regression (CQR) as a solution for estimating regression coefficients in traditional linear regression models. They demonstrated that the CQR estimator is valid regardless of the error distribution and exhibits superior estimation efficiency when compared to the OLS estimator. This efficiency and robustness of CQR have captured the attention of researchers since its proposal. Zhao et al. (2017) explored CQR estimation for longitudinal data. Xu et al. (2017) delved into the CQR neural network model. More recently, Gu and Zou (2020) investigated CQR for ultrahigh-dimensional data, with Qu et al. (2022) extending this work to ultrahigh-dimensional heterogeneous data. Jin et al. (2023) introduced an optimal subsampling algorithm to address the estimation challenges of CQR in massive data. These strong theoretical properties of the CQR estimator have motivated its application to distributed big data scenarios. In recent years, researchers have explored CQR estimation for distributed big data using both the OS approach and the iterative approach. In the OS approach, Jiang et al. (2018) explored CQR for large-scale data in linear models. On the other hand, for the iterative approach, Wang et al. (2021) extended CQR to massive data using the communication-efficient surrogate likelihood method in Jordan et al. (2018).

1.3 Motivations and contributions

In the context of distributed systems, alongside communication costs, the arrangement of data across the system poses another significant concern. While both OS approaches and iterative approaches have demonstrated practical utility, they rely on a pivotal assumption that data is distributed randomly among distinct workers. Ideally, data would exhibit complete randomness in its allocation across workers (Jordan et al. 2018; Fan et al. 2021; Zhu et al. 2021; Pan et al. 2022). However, real-world implementations seldom adhere to such ideal conditions, as data allocation tends to prioritize convenience. For instance, practitioners often assign data to workers based on factors like location or time. As a result, when data deviates from a random distribution among workers, the estimators derived using the aforementioned methods might lose their consistency, leading to a significant compromise in statistical efficiency. Moreover, even though the OS approach is practically feasible for distributed CQR estimators, it could still suffer from statistical inefficiency (Jordan et al. 2018). Hence, delving into the CQR problem within the context of non-randomly distributed data in distributed systems represents a captivating avenue of research. To the best of our knowledge, no prior works have addressed this specific aspect, making it an unexplored area deserving of attention. Therefore, this paper introduces a novel approach named the "One-Step Upgraded Pilot CQR Method" to efficiently tackle the challenge of CQR estimation within a distributed system featuring non-randomly distributed large-scale data. Here, efficiency pertains to both computational and statistical aspects. In practice, data can be distributed across various workers in diverse convenient ways.

The proposed method comprises two fundamental steps. In the first step, a small pilot sample of size n is randomly gathered from different workers within the distributed system and stored on the master node. Notably, n is chosen to be substantially smaller than the total sample size N. Subsequently, a standard CQR estimator, known as the pilot estimator, is computed on the master. Since the pilot sample is obtained through entirely random selection, the pilot estimator demonstrates consistency. However, due to the significantly smaller size of n compared to N, the pilot estimator lacks statistical efficiency. To further enhance the statistical efficiency, the second step involves broadcasting the pilot estimator from the master to the workers. Leveraging the pilot estimator and local samples, each worker calculates local sample moment quantities. Then, the computed sample moment quantities in each worker is received by the master with little communication cost, since they are finite dimensional vectors or matrices. Finally, the master employs a one-step update process based on the pilot estimator and the computed sample moment quantities to derive the ultimate outcome—the one-step upgraded pilot CQR estimator. This final estimator balances the need for improved statistical efficiency with the practical constraints of communication costs within a distributed environment.

It's important to highlight that due to the non-smooth nature of the CQR objective function, standard Newton-Raphson-type algorithms cannot be directly applied for one-step updates. To address this challenge, we introduce a new one-step update approach that leverages the asymptotic expression of the CQR estimator. Notably, this approach doesn't rely on restrictive assumptions about randomness. We can demonstrate that the resulting one-step upgraded pilot CQR estimator is $\sqrt{N}$-consistent. Moreover, its asymptotic covariance matches that of the global estimator calculated using the entire dataset. This result demonstrates that the resulting estimator achieves asymptotically optimal statistical efficiency.

In summary, this paper contributes in the following ways:

1. Novel One-Step Upgraded Pilot CQR Method: We introduce a new approach for one-step upgraded pilot CQR estimation. The uniqueness lies in the one-step update method, which ingeniously utilizes the asymptotic expression of the CQR estimator. This method inherits the merits of CQR while effectively addressing computational challenges associated with large-scale data.
2. Desirable Characteristics of the Method: The newly introduced method possesses several favorable qualities. Firstly, in terms of communication cost and storage usage, the method is computationally efficient. It requires only three rounds of communication between the master and the workers, with small amounts of data being transmitted. Secondly, the resulting estimator demonstrates statistical efficiency, with its asymptotic mean squared error being equivalent to that of the global estimator. Third, the resulting estimator is robust against data distribution across workers, as its consistency is guaranteed even in scenarios where data are non-randomly distributed among workers.Thirdly, the new method imposes no limitation on the number of workers K.
3. Applicability Beyond Distributed Systems: Although tailored for distributed systems, it's noteworthy that the proposed method can also find application in single computers constrained by memory limitations.

The remainder of this article is organized as follows. In Sect. 2, we present the problem statement and introduce the related methods. Section 3 provides a detailed explanation of the proposed method, along with the related algorithm and theoretical properties. In Sects. 4, 5, we conduct numerical experiments to validate the theoretical properties. Conclusions and discussions are presented in Sect. 6. All proofs are included in the Appendix.

2 Problem statement

2.1 Standard composite quantile regression

Let $\{\boldsymbol{X}_i, Y_i\}_{i=1}^N$ denote an independent and identically distributed random sample consisting of N observations from a linear regression model:

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{1}$$

Here Y_i is the univariate response, $\boldsymbol{X}_i \in \mathbb{R}^p$ is a covariate vector, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter vector, and ε_i is the random error. The cumulative distribution function of ε_i is denoted as $F(\cdot)$, and its corresponding probability density function is represented as $f(\cdot)$. For the remainder of this article, we assume that ε_i is independent of covariates $\boldsymbol{X}_i$ with $E(\varepsilon_i) = 0$. Additionally, $\boldsymbol{X}_i$ is assumed to have finite moments, whereas Y_i does not have this restriction.

Given a single quantile level $\tau \in (0, 1)$, an estimator of the $100\tau\%$ quantile of Y_i in model (1) is defined as

$$\left(\hat{b}_\tau, \hat{\boldsymbol{\beta}}_\tau^{\mathrm{T}}\right) = \arg\min_{b_\tau, \boldsymbol{\beta}} \sum_{i=1}^{N} \rho_\tau \left(Y_i - b_\tau - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right),$$

where $\hat{b}_\tau$ is the estimator of the b_τ, $\hat{\boldsymbol{\beta}}_\tau$ is the estimator of the $\boldsymbol{\beta}$, and $\rho_\tau(t) = \tau t - t\mathrm{I}(t < 0)$ is the check loss function, with $\mathrm{I}(\cdot)$ being the indicator function. The true $100\tau\%$ quantile of Y_i given $\boldsymbol{X}_i$ is $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + b_\tau$, where $b_\tau = F^{-1}(\tau)$ is the $100\tau\%$ quantile of the distribution of the error term ε_i, as discussed by Koenker (2005). In this context, it is important to note that the only regression parameter that depends on the quantile level is the b_τ, while the true $\boldsymbol{\beta}$ are the same for all quantiles. Motivated by the desire to integrate information from various quantile levels, Zou and Yuan (2008) proposed CQR method to estimate $\boldsymbol{\beta}$ as

$$\left(\hat{\boldsymbol{b}}_{\mathrm{CQR}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{T}}\right) = \arg \min_{(b_1,\ldots,b_M,\boldsymbol{\beta})} \sum_{m=1}^{M}\sum_{i=1}^{N} \rho_{\tau_m}\left\{Y_i - b_m - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right\}, \tag{2}$$

where $\hat{\boldsymbol{b}}_{\mathrm{CQR}} = (\hat{b}_1, \ldots, \hat{b}_M)^{\mathrm{T}}$, $\hat{b}_m$ represents the estimator for b_m, b_m is the τ_m quantile of ε_i, and $0 < \tau_1 < \tau_2 < \cdots < \tau_M < 1$ denote M quantile levels. Typically, for a given positive integer M, using equally spaced quantiles at $\tau_m = m/(M+1)$ for $m = 1, 2, \ldots, M$ is recommended in Zou and Yuan (2008). By integrating information from multiple quantile levels, the CQR method achieves higher estimation efficiency compared to single-level quantile regression, while being more robust compared to least squares procedure.

It is worth noting that if a single computer has sufficient memory to load all the data, Eq. (2) can be solved using a linear programming algorithm; see Zou and Yuan (2008). However, for big data, solving the Eq. (2) is infeasible due to the limited operational memory of one single computer. Furthermore, linear programming methods are not well-suited for big data in a distributed system. This is primarily because existing linear programming algorithms involve extensive inter-computer iterations, resulting in substantial communication overhead in distributed systems. Consequently, creating a computationally efficient approach for solving the problem (2) has become a key focus.

2.2 Distributed composite quantile regression

Assume that N is sufficiently large that storing the data on a single computer becomes impractical. Thus, the data must be divided among K local workers. Define $\mathcal{S} = \{1, \cdots, N\}$ and let $\mathcal{S}_k$ denote the set of indices for the samples held by the kth worker, with $k = 1, \ldots, K$. Define $n_k = |\mathcal{S}_k|$ as the number of samples assigned to the kth worker. It follows that $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$, with $\sum_{k=1}^{K} |\mathcal{S}_k| = N$ and $\mathcal{S}_{k_1} \cap \mathcal{S}_{k_2} = \emptyset$ for any $k_1 \neq k_2$. The OS approach and the iterative approach are two common strategies to deal with distributed big data.

The basic idea of the OS approach is to calculate the local CQR estimator, denoted as $\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$, on each worker by solving the optimization problem:

$$\left(\hat{\boldsymbol{b}}_{\mathrm{CQR},k}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}^{\mathrm{T}}\right) = \arg \min_{b_1,\ldots,b_M,\ \boldsymbol{\beta}} \sum_{m=1}^{M} \sum_{i \in \mathcal{S}_k} \rho_{\tau_m} \left\{ Y_i - b_m - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right\}. \tag{3}$$

Subsequently, the kth worker transmits the corresponding $\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$ to the master. The master then aggregates the received local estimators and outputs the resulting OS estimator, denoted as $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}$. This is achieved by performing a simple averaging operation: $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$. The OS approach exhibits minimal communication cost, as it requires only one round of communication between the master and the workers. Furthermore, when the data are randomly distributed among the workers, the local CQR estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$ should be $\sqrt{n_k}$-consistent. This property ensures that the OS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}$ is also consistent (Zhang et al. 2012; Fan et al. 2019).

However, the OS approach has certain limitations. Firstly, the consistency of the OS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{OS}$ depends on the consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$, which in turn depends on how the dataset is randomly distributed among the K workers. Secondly, when the data are distributed non-randomly, the OS estimator will loss estimation efficiency. This intuitive argument finds theoretical confirmation in Proposition 1.

Proposition 1 *For each sample* $i \in \mathcal{S}_k$, $\boldsymbol{X}_i$ *is generated independently with mean zero and covariance matrix* $E\left(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \mid i \in \mathcal{S}_k\right) = \boldsymbol{\Sigma}_k$. *Assume all the sample sizes of different workers are equal and* $K = o(\sqrt{N})$, *we have*

(P1.1) *the covariance matrix of* $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ *is*

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}\right) = \frac{1}{N} R \left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_k\right)^{-1} \left\{1 + o_p(1)\right\},$$

where $R = \dfrac{\sum_{m,m'=1}^{M} \min(\tau_m, \tau_{m'})(1 - \max(\tau_m, \tau_{m'}))}{\left(\sum_{m}^{M} f(b_m)\right)^2}$*; see (Zou and Yuan 2008).*

(P1.2) *the covariance matrix of* $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}$ *is*

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}\right) = \frac{1}{N} R \left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_k^{-1}\right) \left\{1 + o_p(1)\right\}.$$

Thus, we can obtain that $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}\right) - \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}\right)$ *is a semi-positive definite matrix,* $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}\right)\left\{1 + o_p(1)\right\}$ *if and only if all the* $\boldsymbol{\Sigma}_k$*s are equal.*

Furthermore, in the iterative approach, to avoid the transmission of the Hessian matrix between the master and workers, the local Hessian matrix is usually utilized as a substitute for the global Hessian matrix. Consequently, it is necessary to impose the following homogeneity assumption:

$$\left\| \frac{1}{n_k} \sum_{i \in \boldsymbol{S}_k} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \right\| \leqslant \delta \to 0$$

where δ is a small homogeneity parameter. This assumption implies that the data sampled from each worker should be randomly drawn from the full dataset. Failing to satisfy this assumption, the local Hessian matrix may not adequately substitute the global Hessian matrix, potentially resulting in statistically inefficient estimators. (Wang et al. 2017; Jordan et al. 2018; Fan et al. 2019; Wang et al. 2022, 2023).

Both the OS method and the iterative method rely on one crucial assumption that data is randomly distributed among different workers. However, this assumption may be impractical in real-world scenarios, as the distribution scheme is often tailored to various practical considerations such as time, region, and other factors. To address this limitation, this paper aims to tackle the problem (2) in a non-randomly distributed manner while ensuring both statistical efficiency and computational efficiency.

3 One-step upgraded pilot composite quantile regression

3.1 One-step upgraded pilot estimator

As previously mentioned, the objective of this subsection is to propose a novel approach for standard CQR in distributed systems with non-randomly distributed data. The proposed procedure involves two steps. In the first step, a random sample of size n is drawn from the distributed system, called the pilot sample, where n is much smaller than N (i.e., $n \ll N$). Specifically, we begin by determining the pilot sample sizes for each worker (i.e., $\tilde{n}_1, \cdots, \tilde{n}_K$), drawing from a multinomial distribution with probabilities $\boldsymbol{p} = (n_1/N, \cdots, n_K/N)^{\mathrm{T}}$. Subsequently, we perform simple random sampling without replacement within each worker k, selecting indexes $\mathcal{P}_k$ from $\mathcal{S}_k$, such that the size of each $\mathcal{P}_k$ is $\tilde{n}_k$. The union of all $\mathcal{P}_k$ is denoted as $\mathcal{P}$ and represents the index set of the pilot sample. It is important to note that $\mathcal{P}_k \subset \mathcal{S}_k$ and $\mathcal{P} \subset \mathcal{S}$. The samples associated with $\mathcal{P}_k$ are then transmitted from the kth worker to the master. Given the small size of n, the communication cost associated with transferring the pilot sample is practically acceptable. Then, the pilot sample $\{(\boldsymbol{X}_i, Y_i), i \in \mathcal{P}\}$ can be stored on the master and a pilot CQR estimator can be calculated by

$$\left(\hat{\boldsymbol{b}}_{\mathcal{P}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{\mathcal{P}}^{\mathrm{T}}\right) = \arg \min_{b_1, \ldots, b_M, \, \boldsymbol{\beta}} \sum_{m=1}^{M} \sum_{i \in \mathcal{P}} \rho_{\tau_m} \left\{ Y_i - b_m - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right\}, \tag{4}$$

where $\hat{\boldsymbol{b}}_{\mathcal{P}} = (\hat{b}_{\mathcal{P},1}, \ldots, \hat{b}_{\mathcal{P},M})^{\mathrm{T}}$. It is noteworthy that the $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is $\sqrt{n}$-consistent irrespective of how data are distributed across workers. This is because the pilot sample is obtained through a fully random process. Although $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is $\sqrt{n}$-consistent, it is not statistically efficient, as an estimator that is truly statistically efficient should achieve $\sqrt{N}$-consistent given the full dataset.

To this end, in the second step, we propose a novel one-step method to further improve the statistical efficiency. It is worth recalling that $\hat{\boldsymbol{\beta}}_{\text{CQR}}$, the standard CQR estimator based on the entire sample, will be referred to as the global estimator. Under suitable conditions, $\hat{\boldsymbol{\beta}}_{\text{CQR}}$ has been shown to be $\sqrt{N}$-consistent and asymptotically normal (Koenker 2005; Zou and Yuan 2008). Specifically, denote by $\boldsymbol{X}$ the design matrix and assume that $\lim_{N\to\infty} \frac{1}{N}\boldsymbol{X}^{\text{T}}\boldsymbol{X} = \mathbf{C}$, where $\mathbf{C}$ is a $p \times p$ positive definite matrix, we have that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\text{CQR}} - \boldsymbol{\beta}\right) \xrightarrow{d} \left[\left(\sum_{m=1}^{M} f\left(b_m\right)\right)\mathbf{C}\right]^{-1} \mathbf{z} \sim N\left(0, \left(\sum_{m=1}^{M} f\left(b_m\right)\right)^{-2} \mathbf{C}^{-1}\Sigma_{\mathbf{z}}\mathbf{C}^{-1}\right)$$

where $\xrightarrow{d}$ represents convergence in distribution, $\mathbf{z} = \sqrt{N}E(\boldsymbol{X}^{\text{T}})\left[\sum_{m=1}^{M}\left(I\left(\varepsilon_i < b_m\right)\right.\right.$ $\left.\left.-\tau_m\right)\right]$, $\boldsymbol{\Sigma}_{\mathbf{z}} = \mathbf{C}\left[\sum_{m,m'=1}^{M} \min\left(\tau_m, \tau_{m'}\right)\left(1 - \max\left(\tau_m, \tau_{m'}\right)\right)\right]$ and $f(\cdot)$ is the probability density function of ε_i. Therefore,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\text{CQR}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \mathbf{C}^{-1}\frac{\sum_{m,m'=1}^{M} \min\left(\tau_k, \tau_{k'}\right)\left(1 - \max\left(\tau_m, \tau_{m'}\right)\right)}{\left(\sum_{m=1}^{M} f\left(b_m\right)\right)^2}\right). \tag{5}$$

A more comprehensive and detailed proof can be found in (Zou and Yuan 2008).

Thus, the asymptotic expression for $\hat{\boldsymbol{\beta}}_{\text{CQR}}$ is as follows:

$$\hat{\boldsymbol{\beta}}_{\text{CQR}} = \boldsymbol{\beta} + \left(\sum_{m=1}^{M} f(b_m)\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\text{T}}\right)^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i^{\text{T}}\left[\sum_{m=1}^{M}\left(I\left(\varepsilon_i < b_m\right) - \tau_m\right)\right] + o_p\left(\frac{1}{\sqrt{N}}\right). \tag{6}$$

We then replace $\boldsymbol{\beta}$ in Eq. (6) with the pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ and ε_i with $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\mathcal{P}}$. Moreover, it should be noted that the error density $f(\cdot)$ and b_m are also generally unknown. However, we can replace $f(\cdot)$ with $\hat{f}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(\hat{\varepsilon}_i - \cdot\right)$. Here $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is a selected bandwidth. Similarly, we can replace b_m with $\hat{b}_{\mathcal{P},m}$. Consequently, we can obtain the estimation of $f(b_m)$ by $\hat{f}(\hat{b}_{\mathcal{P},m})$ for $m = 1, \ldots, M$. This results in the formulation of our one-step updated pilot estimator $\hat{\boldsymbol{\beta}}_{\text{CQR}}^{(1)}$ presented as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)} = \hat{\boldsymbol{\beta}}_{\mathcal{P}} + \left(\sum_{m=1}^{M} \hat{f}(\hat{b}_{\mathcal{P},m}) \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^{\mathrm{T}} \left[\sum_{m=1}^{M} \left(I\left(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m} \right) - \tau_m \right) \right]. \tag{7}$$

The one-step updated pilot estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ shares a similar spirit with classical one-step estimators (Fan and Chen 1999). Classical one-step estimators rely on the assumption of sufficient smoothness of the loss function and employ the Hessian matrix for the one-step updated. However, the loss function for CQR is not smooth, rendering the Hessian matrix cannot be naturally defined and computed. Furthermore, the classical one-step method is $\sqrt{N}$-consistent, while our pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is $\sqrt{n}$-consistent. To tackle these problems, we draw inspiration from the asymptotic expressions of CQR estimator and propose Eqs. (6, 7) as a solution. Our approach does not depend on the use of the Hessian matrix. We can computationally confirm that $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ exhibits both $\sqrt{N}$-consistent and asymptotic normality. We have also found that its asymptotic covariance aligns with that of the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ when the conditions are met. This important finding suggests that $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ possesses statistical efficiency. Notably, this nice property primarily depends on the consistency of $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ and is independent of the data distribution across different workers.

Remark 1 (R1.1) This study employs the Gaussian kernel $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$. The bandwidth is set to $h = 0.9AN^{-1/5}$ as selected by Jiang et al. (2018), where $A = \min\{\hat{\sigma}, \mathrm{IQR}/1.34\}$, with $\hat{\sigma}$ representing the sample sd of $\left\{\hat{\varepsilon}_i : i \in \mathcal{P}\right\}$, and IQR referring to the interquartile range of $\left\{\hat{\varepsilon}_i : i \in \mathcal{P}\right\}$. (R1.2) Zou and Yuan (2008) recommended $M = 19$ for linear model. Hence, we adopt the $M = 19$ in our study. (R1.3) The remarkable aspect of $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ is that it exhibits desirable properties while incurring a communication cost that is practically acceptable. Specifically, the estimation process necessitates only three rounds of master-and-worker communications.

The detailed algorithm of the proposed estimator is described in Algorithm 1.

Algorithm 1 One-step updated pilot composite quantile regression algorithm

- **Step 1: Pilot Sampling and Estimation**
  1. Set the pilot sample size n.
  2. Generate the pilot sample sizes assigned to workers, i.e., $\tilde{n}_1, \cdots, \tilde{n}_K$, from multinomial distribution $M(n, n_1/N, \cdots, n_K/N)$.
  3. **for** $k = 1, 2, \cdots, K$ **do**:
     1) The master broadcasts $\tilde{n}_k$ to the kth worker.
     2) The kth worker selects n_k samples without replacement form S_k, and denote it by $\mathcal{P}_k$.
     3) The kth worker transfers $\{(\boldsymbol{X}_i, Y_i), i \in \mathcal{P}_k\}$ to the master, initiating the first round of communication between workers and the master.
  4. Compute the pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ and $\hat{\boldsymbol{b}}_{\mathcal{P}}$ using the pilot samples $\mathcal{P} = \cup_k \mathcal{P}_k$ on the master according to Equation 4. Additionally, the bandwidth h can be calculated on the master.
- **Step 2: One-Step Updating**
  1. The master broadcasts $\hat{\boldsymbol{\beta}}_{\mathcal{P}}, \hat{\boldsymbol{b}}_{\mathcal{P}}, M = 19, \tau_m$ and h to each worker. This marks the second phase of communication between workers and the master.
  2. **for** $k = 1, 2, \cdots, K$ **do**:
     1) Based on the local samples $\mathcal{S}_k$ on the k worker, calculate the sample moment quantities, $\sum_{i \in \mathcal{S}_k} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}$, $\sum_{i \in \mathcal{S}_k} \boldsymbol{X}_i^{\mathrm{T}} [\sum_{m=1}^{M} (I(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}) - \tau_m)]$, $\sum_{i \in \mathcal{S}_k} K_h(\hat{\varepsilon}_i - \hat{b}_{\mathcal{P},m})$ for $m = 1, 2, \cdots, M$.
     2) Report all sample moment quantities to the master, marking the third phase of communication between workers and the master.
  3. The master calculates the one-step updated pilot estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ using Equation 7.

3.2 Theoretical results

In this subsection, we explore the asymptotic properties of our proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$. To establish these properties, we require the fulfillment of the following standard technical conditions.

(A1) There is a $p \times p$ positive definite matrix $\mathbf{C}$ such that $\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} = \mathbf{C}$ and $\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} = \mathbf{C}$.

(A2) The random error ε has cumulative distribution function $F(\cdot)$ and density function $f(\cdot)$ with uniformly bounded first derivative, and for each p-vector $\boldsymbol{u}$,

$$\lim_{n\to\infty} \frac{1}{N} \sum_{i=1}^{N} \int_0^{v+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{u}} \sqrt{n} \left[F\left(a + \frac{t}{\sqrt{n}}\right) - F(a) \right] dt = \frac{1}{2} f(a) \left(v, \boldsymbol{u}^{\mathrm{T}}\right) \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \Sigma \end{bmatrix} \left(v, \boldsymbol{u}^{\mathrm{T}}\right)^{\mathrm{T}}$$

(A3) The density function $f(\cdot)$ is positive and continuous at the τ_m-th quantile b_m.

(A4) The kernel function $K(\cdot)$ is symmetric about 0, and it also satisfies $\int K^2(u)du < \infty$, $\int |u|^3 K(u)du < \infty$. Its first derivative, $\dot{K}(\cdot)$, is also bounded. Additionally, let the bandwidth h satisfy $h = O\left(N^{-1/5}\right)$.

(A5) The pilot sample size n meets the conditions $n/N^{4/5} \to 0$ and $n/\sqrt{N} \to \infty$.

Remark 2 Assumptions (A1) and (A2) are essential conditions for the validity of Eqs. (5, 6). Prior studies by Zou and Yuan (2008) and Kai et al. (2010) have established the significance of these assumptions. Assumption (A3) describes a typical requirement for the density function in (Pan et al. 2022). Similarly, Assumption (A4) is a customary requirement pertaining to the kernel function. Lastly, Assumption (A5) imposes a restriction on the n. As per Assumption (A5), we understand that n needs to be much smaller in comparison to N, specifically $n/N^{4/5} \to 0$. Additionally, Assumption (A5) indicates that the n should not be too small, ensuring that $n/\sqrt{N} \to \infty$. If n is defined as $[\sqrt{N} \log N]$, with $[t]$ denoting the integer part of a real number t, Assumption (A5) is consequently satisfied. We then have the following theorem.

The following theorem describes the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$.

Theorem 1 *Suppose that consitions* (A1)–(A5) *are satisfied. Then,*

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \mathbf{C}^{-1}\frac{\sum_{m,m'=1}^{M} \min\left(\tau_k, \tau_{k'}\right)\left(1 - \max\left(\tau_m, \tau_{m'}\right)\right)}{\left(\sum_{m=1}^{M} f\left(b_m\right)\right)^2}\right). \tag{8}$$

Refer to Appendix for the detailed proof. Based on Eqs. (5, 8), it is evident that both the proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ and the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ have the same asymptotic distribution. This indicates that $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ possesses the same asymptotic efficiency as $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$.

4 Simulation studies

In this section, we rigorously evaluate the performance of the proposed method through extensive simulations. All numerical experiments were carried out using the `R` software on a desktop computer running Windows 10, equipped with an Intel I9 processor and 32 GB RAM. To ensure statistical robustness, all the simulation results are based on 500 independent replications. In accordance with the guidance provided by Zou and Yuan (2008), we set the number of quantiles M to be 19.

4.1 Data generation and storage strategies

Full sample $\{\boldsymbol{X}_i, Y_i\}_{i=1}^{N}$ of size $N = 10^5$ are generated from a linear regression model

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

with the true parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^{\mathrm{T}} = (-2, -1, 0, 1, 2)^{\mathrm{T}}$. Let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3},$
$X_{i4}, X_{i5})^{\mathrm{T}}$ is generated from a multivariate normal distribution with mean 0 and covariance $\boldsymbol{\Sigma} = \left(\sigma_{j_1 j_2}\right) \in \mathbb{R}^{5\times 5}$, where $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ and $j_1, j_2 = 1, \ldots, 5$. The K is set to 10, 20, 50, 100, 250, respectively. Furthermore, we consider the following four different error distributions for ε_i to illustrate the robustness.

Example 1: the standard normal distribution, $N(0, 1)$.
Example 2: the Student t distribution with three degrees of freedom, $t(3)$.
Example 3: the mixture normal distribution, $0.9N(0, 1) + 0.1N(0, 10)$.
Example 4: the standard Cauchy distribution, Cauchy(0,1).

For each example, we examine three distinct data storage strategies. The first strategy stores the entire sample in a completely random manner. Conversely, the other second strategies are to distribute data in a non-random manner.

Strategy 1 (Randomly distributed sample): All samples $\{\boldsymbol{X}_i, Y_i\}_{i=1}^{N}$ are allocated to K workers randomly, with each worker receiving an identical sample size, specifically $n_k = [N/K]$. The pilot sample is created by drawing $[n/K]$ samples at random from each worker, where $[x]$ denotes the integer part of x.
Strategy 2 (Non-randomly distributed sample): The storage position of each observation is determined by its covariates, resulting in non-random storage of all samples. To be specific, let us define $Z_i = \sum_{j=1}^{5} X_{ij}$ and $Z_{(i)}$ as the order statistics of Z_i, where $Z_{(1)} < Z_{(2)} < \cdots < Z_{(N)}$. Subsequently, we assign the (i)th observation $(\boldsymbol{X}_{(i)}^{\mathrm{T}}, Y_{(i)})$ to the $\kappa(i)$ th worker, where $\kappa(i) = [iK/N] + 1$. The pilot sample is created by drawing $[n/K]$ samples at random from each worker.
Strategy 3 (Non-randomly distributed sample): The distribution of samples across K workers is uneven. Specifically, at each time, K random numbers are drawn from a normal distribution $N(N/K, 9N/(64K))$. These numbers are then scaled and rounded to integers, ensuring their summation equals N, representing the sample size n_k on K workers. Subsequently, $[n \times n_k/N]$ samples are randomly drawn from each worker, and these samples are concatenated to form the pilot sample.

4.2 Competing estimators and performance metrics

For a reliable evaluation, we compare our proposed One-Step Updated Pilot Estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)}$ (referred to as OSUPE) with the following competing estimators.

(i) The Global Estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ (referred to as GE) is calculated using Eq. (2).

(ii) The One-Shot Estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}}$ (referred to as OSE) is proposed in Jiang et al. (2018). Here $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{\mathrm{OS}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$, and $\hat{\boldsymbol{\beta}}_{\mathrm{CQR},k}$ is calculated using Eq. (3).

(iii) The Communication-efficient Surrogate Likelihood Estimator (referred to as CSLE) is proposed in Jordan et al. (2018).
(iv) The Pilot Estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ (referred to as PE) is calculated using Eq. (4).

To evaluate the accuracy of various estimators, we define the following performance metrics. The Mean Squared Errors (denoted by MSE) are defined as

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{500}\sum_{b=1}^{500}\|\hat{\boldsymbol{\beta}}^{(b)} - \boldsymbol{\beta}\|^2,$$

where the $\|\cdot\|$ represents the Euclidean norm. Let $\hat{\boldsymbol{\beta}}^{(b)}$ represent a specific estimator derived from the bth random replication. For instance, this could refer to the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$. Furthermore, to assess the statistical efficiency of various estimators $\hat{\boldsymbol{\beta}}$, including OSUPE, OSE, CSLE, and PE, in comparison to the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$, the Ratio of Mean Squared Errors (denoted by RMSE) are defined as

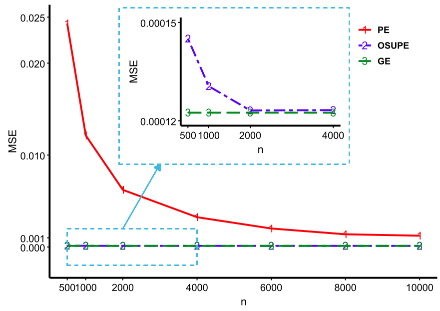
$$\mathrm{RMSE}(\hat{\boldsymbol{\beta}}) = \frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{CQR}})}.$$

Considering that the global estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}}$ is anticipated to have the highest estimation accuracy, it is reasonable to anticipate that the RMSE will be ≥ 1 for all other estimators. A RMSE value close to 1 indicates that the estimator in question exhibits statistical efficiency similar to the global estimator. These metrics provide quantitative measures to assess the accuracy of different estimators. By evaluating the MSE and RMSE, we gain insights into the performance of each estimator.

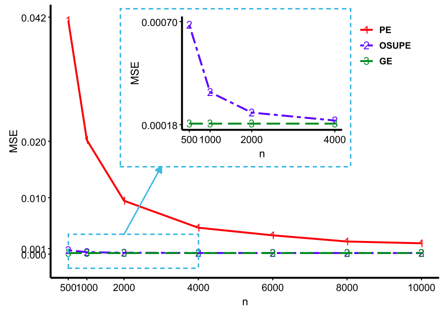
4.3 Effect of the pilot sample size and comparison with GE

As the pilot sample size n may influence the performance of the resulting estimator, we prioritize investigating its influence. Accordingly, we systematically tested multiple values of n in ascending order. Specifically, simulations were conducted with n set to 500, 1000, 2000, 4000, 6000, 8000 and 10000, considering four types of error distributions and three data storage strategies. The MSE values for our new OSUPE, GE, and PE estimators were presented in Figs. 1, 2, 3. Notably, the results for different values of K exhibited similarity (The expression for OSUPE in (7) shows the robustness to varying numbers of workers K; this conclusion is further supported by simulations in Sect. 4.4). For conciseness and efficient use of space, we have fixed $K = 50$ for presenting the results.

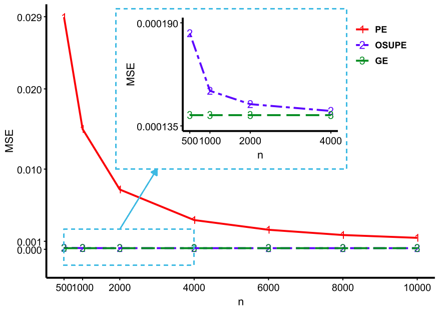
Based on the analysis of Figs. 1, 2, 3, it becomes evident that the MSE values of our method remain nearly constant when $n \geqslant 2000$. Consequently, it has been established that fixing the pilot sample size at $n = [\sqrt{N}\log N] = 3640$ is the appropriate course of action. This finding provides valuable guidance for selecting n in practical applications, as discussed in Remark 2.



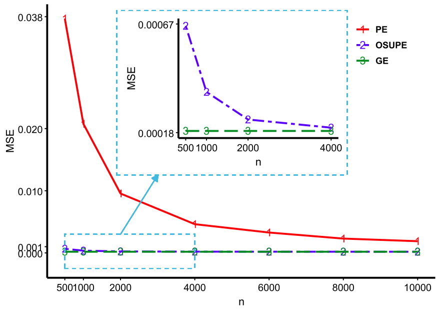
(a) Example 1 ($N(0, 1)$)



(b) Example 2 ($t(3)$)



(c) Example 3 ($0.9N(0, 1) + 0.1N(0, 10)$)



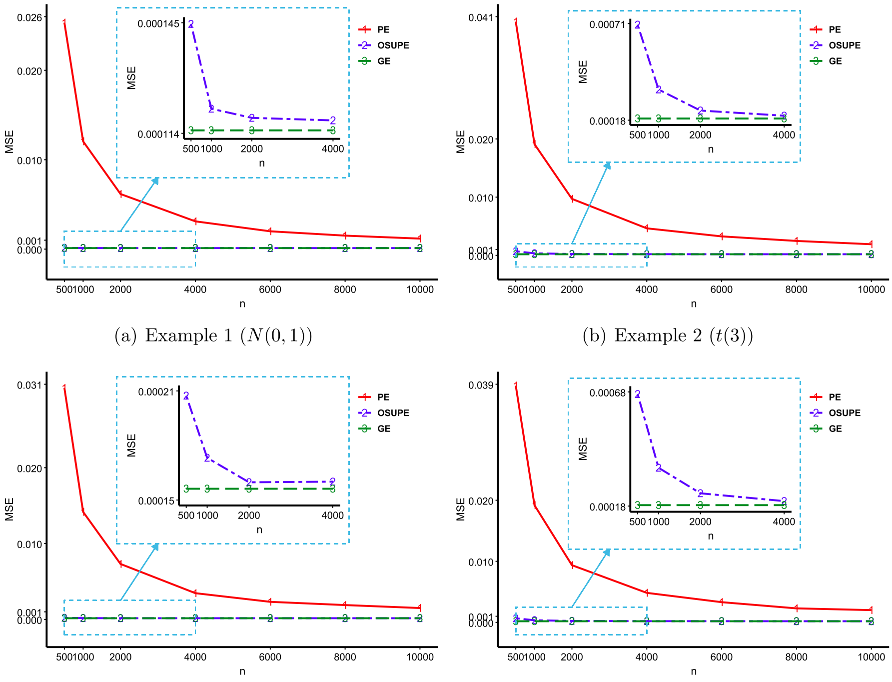
(d) Example 4 (Cauchy(0,1))

Fig. 1 In Strategy 1, the MSE values of PE, OSUPE and GE for different pilot sample size n with the workers size being fixed at $K = 50$

Furthermore, by carefully analyzing Figs. 1, 2, 3, along with their respective zoomed-in sections, we make a significant observation: for all experiments where $n \geq [\sqrt{N} \log N]$, the MSE values of our OSUPE remains within a minute difference of 0.0001 compared to that of GE. This finding strongly suggests that, for $n \geq [\sqrt{N} \log N]$, there is little to no difference between the performance of the new estimator and the global estimator. In stark contrast, the PE exhibits notably inferior performance compared to our OSUPE. This outcome aligns with our expectations since the pilot estimator is known to be $\sqrt{n}$-consistent, while n is considerably smaller than the total sample size N. These observations serve to validate the theoretical results presented in Theorem 1.

4.4 Comparison with competing estimators

Based on the preceding discussion, we have established that when the pilot sample size satisfies $n \geq [\sqrt{N} \log N]$, the MSE values of our new estimator remain almost constant. Consequently, we have chosen $n = [\sqrt{N} \log N] = 3640$ to conduct a comprehensive performance comparison between our new estimator and three other competing estimators (PE, OSE, CSLE). Furthermore, to further explore the specific impact of different worker numbers K on each distributed estimator, we have set K to



(a) Example 1 ($N(0,1)$)

(b) Example 2 ($t(3)$)

(c) Example 3 ($0.9N(0,1) + 0.1N(0,10)$)

(d) Example 4 (Cauchy(0,1))

Fig. 2 In Strategy 2, the MSE values of PE, OSUPE and GE for different pilot sample size n with the workers size being fixed at $K = 50$

the following values: 10, 20, 50, 100, 250, and 500. The RMSE (ratio-MSE) values of four competing estimates (PE, OSE, CSLE, OSUPE) with respect to the global estimator are recorded in Tables 1, 2, 3.

Based on the simulation results presented in Tables 1, 2, 3, several significant conclusions can be drawn. Firstly, when comparing our OSUPE with OSE, we observe that, in the case of randomly stored data (i.e., Strategy 1), both estimators perform similarly, as evidenced by their RMSE values being exactly or approximately equal to 1. However, when the data are not randomly distributed (Strategies 2 and 3), our new estimator significantly outperforms OSE. We find that the RMSE values for the OSE exceed 1 significantly, suggesting that the OSE demonstrates considerably lower statistical efficiency compared to the global estimator in this scenario. In contrast, our proposed estimator consistently maintains RMSE values very close to 1, suggesting that its statistical efficiency remains comparable to that of the global estimator. Moreover, an expected finding is that the RMSE values of the OSE increase with an increase in the number of workers K. Conversely, our method exhibits robustness in response to varying numbers of workers K.

Secondly, when comparing our OSUPE with CSLE, we observe notable differences based on different error distributions and data strategies. For Strategy 1, our OSUPE outperforms CSLE when the errors follow the $t(3)$ and $0.9N(0, 1) + 0.1N(0, 10)$

Table 1 In Strategy 1, the RMSEs of the four alternative estimators relative to the global estimator. If the figures are smaller than 10, they are rounded to four decimal places; if they fall between 10 and 100, they are rounded to three decimal places. Otherwise, they are represented approximately as $aEb = a \times 10^b$, where a is rounded to two decimal places

Error type	Method	K					
		10	20	50	100	250	500
$N(0, 1)$	PE	24.665	25.060	25.008	25.438	24.884	27.352
	OSE	1.0094	1.0101	1.0118	1.0126	1.0261	1.0650
	CSLE	1.0058	1.0093	1.0101	1.0105	1.0110	1.0124
	OSUPE	1.0106	1.0105	1.0107	1.0094	1.0168	1.0114
$t(3)$	PE	26.094	26.994	25.978	27.323	27.042	23.742
	OSE	1.0002	1.0061	1.0290	1.0412	1.1057	1.2078
	CSLE	2.2358	2.2793	2.3101	2.3605	2.3910	2.4124
	OSUPE	1.1175	1.0809	1.1000	1.0959	1.0984	1.0781
$(0.9N(0, 1) + 0.1N(0, 10))$	PE	25.601	25.205	25.512	25.482	24.486	24.872
	OSE	1.0032	1.0051	1.0107	1.0119	1.0127	1.0200
	CSLE	2.5562	2.5819	2.6048	2.6133	2.6221	2.6230
	OSUPE	1.0060	1.0087	1.0150	1.0114	1.0036	1.0017
Cauchy(0,1)	PE	24.524	24.625	25.264	26.233	25.222	22.898
	OSE	1.0143	1.0136	1.0115	1.0292	1.0863	1.1076
	CSLE	9.19E5	9.25E5	9.54E5	9.55E5	9.61E5	9.63E5
	OSUPE	1.1147	1.0975	1.0889	1.0847	1.0537	1.0781

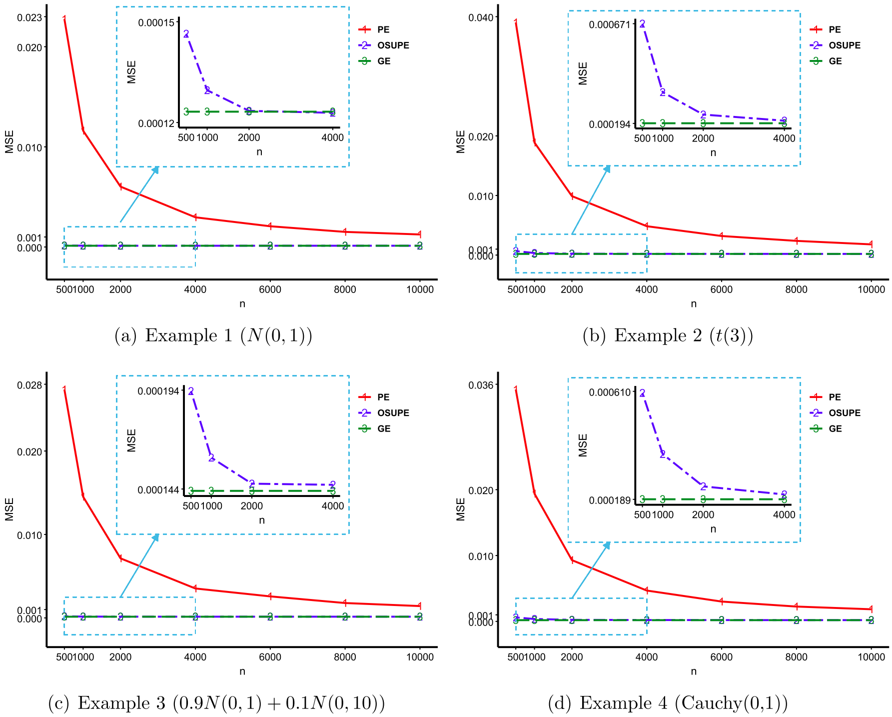


Fig. 3 In Strategy 3, the MSE values of PE, OSUPE and GE for different pilot sample size n with the workers size being fixed at $K = 50$

distributions. Additionally, CSLE experiences breakdown when confronted with the Cauchy error distribution. This vulnerability is attributed to CSLE being built upon mean regression, making it sensitive to heavy-tailed distributions and outliers. However, CSLE slightly outperforms OSUPE when the error follows a standard normal distribution. This is because composite quantile regression may sacrifice estimation efficiency for normal random errors. Furthermore, when the data are non-randomly distributed (Strategies 2 and 3), our new OSUPE consistently exhibits superior performance to CSLE for each considered error distribution. In fact, there are instances where CSLE fails to perform, as evidenced by excessively large RMSE values. This limitation of CSLE can be attributed to its exclusive reliance on the local Hessian matrix, which significantly differs from the global one and can lead to inadequate performance in non-randomly distributed data scenarios.

Finally, it is evident that our OSUPE consistently outperforms the PE by a significant margin. This outcome was anticipated, given that the PE is $\sqrt{n}$-consistent, whereas the OSUPE is $\sqrt{N}$-consistent.

5 Real-world data application

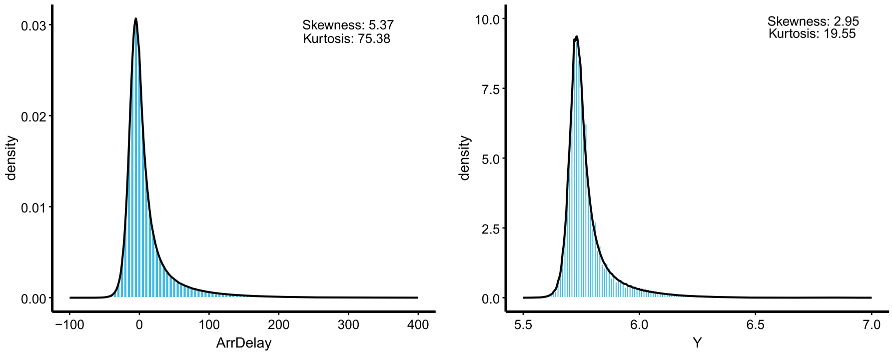
The real dataset used in this study is extracted from the airline on-time performance data of the 2009 ASA Data Expo. This dataset is publicly available at the fol-

Table 2 In the Strategy 2, the RMSEs of the four alternative estimators relative to the global estimator. If the figures are smaller than 10, they are rounded to four decimal places; if they fall between 10 and 100, they are rounded to three decimal places. Otherwise, they are expressed approximately in the form of $aEb = a \times 10^b$, where a is rounded to two decimal places

Error type	Method	K					
		10	20	50	100	250	500
$N(0, 1)$	PE	23.428	23.892	24.221	23.712	25.740	24.217
	OSE	1.5709	2.2994	4.1742	9.0983	26.032	52.824
	CSLE	1.4218	1.6328	2.0132	3.4697	8.0288	12.636
	OSUPE	1.0149	1.0170	1.0222	1.0074	1.0083	1.0211
$t(3)$	PE	25.835	24.890	24.606	24.652	25.340	24.370
	OSE	1.5809	2.2813	4.6602	9.5336	24.548	61.242
	CSLE	3.5682	3.6678	4.0142	6.1378	8.7626	13.050
	OSUPE	1.0943	1.0789	1.0851	1.0655	1.0633	1.0851
$(0.9N(0, 1) + 0.1N(0, 10))$	PE	27.782	26.417	25.711	25.254	25.985	26.296
	OSE	1.6048	2.4667	4.9086	9.6560	24.824	59.062
	CSLE	11.388	12.131	12.988	14.843	24.158	35.855
	OSUPE	1.0019	1.0278	1.0048	1.0161	1.0025	1.0053
Cauchy(0,1)	PE	25.800	25.409	25.074	24.139	23.852	25.099
	OSE	1.5689	2.2468	4.9692	8.0791	26.192	49.077
	CSLE	1.14E7	2.11E6	2.77E6	3.21E6	3.63E7	5.01E7
	OSUPE	1.0879	1.0698	1.0793	1.0860	1.0735	1.0571

Table 3 In the Strategy 3, the RMSEs of the four alternative estimators relative to the global estimator. If the figures are smaller than 10, they are rounded to four decimal places; if they fall between 10 and 100, they are rounded to three decimal places. Otherwise, they are expressed approximately in the form of $aEb = a \times 10^b$, where a is rounded to two decimal places

Error type	Method	K					
		10	20	50	100	250	500
$N(0, 1)$	PE	25.209	24.402	23.681	24.682	22.816	24.572
	OSE	1.6542	2.3760	4.4373	5.2070	66.840	1.24E4
	CSLE	1.4543	1.7476	2.5998	2.6322	2.7843	15.376
	OSUPE	1.0222	1.0215	1.0064	1.0071	1.0046	1.0112
$t(3)$	PE	24.120	24.135	23.037	22.236	23.474	22.396
	OSE	1.4601	13.011	4.20E2	3.03E3	1.64E3	8.43E3
	CSLE	2.1574	6.6270	1.22E2	1.29E2	3.31E2	1.12E3
	OSUPE	1.1098	1.0588	1.0852	1.0642	1.0772	1.0671
$(0.9N(0, 1) + 0.1N(0, 10))$	PE	24.197	22.815	21.854	23.145	20.972	21.577
	OSE	1.3293	4.5428	12.853	32.480	3.14E3	9.56E4
	CSLE	8.8021	8.8491	9.1203	15.224	1.01E2	2.65E4
	OSUPE	1.0018	1.0017	1.0001	1.0095	1.0160	1.0105
Cauchy(0,1)	PE	28.049	29.204	28.028	27.603	26.301	25.987
	OSE	1.3725	21.029	54.635	8.37E2	3.21E3	4.47E7
	CSLE	8.64E6	9.98E6	1.03E7	1.04E7	1.08E7	1.11E7
	OSUPE	1.0089	1.0708	1.1029	1.0909	1.0692	1.0877



(a) ArrDelay (b) log(ArrDelay-min(ArrDelay)+1)

Fig. 4 Distribution of actual delays and log-transformed actual delays based on the full data

lowing link: https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009. The full dataset includes flight arrival and departure details for all commercial flights within the USA, spanning from October 1987 to April 2008. With a substantial total of 123,534,969 observations and 29 variables, the dataset is quite large, amounting to approximately 11 GB in size. The dataset has been used by Schifano et al. (2016) to demonstrate the performance of their methods for analyzing big data. One of the primary objectives of analyzing this dataset is to develop a model for predicting airline delays. Due to the excessively long computing time, we only considered the data from the year 2007, which consists of $7,455,458$ observations with complete data. Following data cleaning, over 7 million observations are preserved, resulting in $N = 7,252,496$. In Fig. 4a, we present the kernel density curve of actual arrival delays, revealing a significant departure from the normal distribution. The distribution of actual delays exhibits pronounced heavy-tailedness and skewness. Even after applying a log transformation to the actual arrival delays (as shown in Fig. 4b), the heavy-tailed phenomenon persists. This observation implies that mean regression may not be suitable for this problem. As previously mentioned, in scenarios with heavily heavy-tailed and skewed error distributions, composite quantile regression is a robust and efficient alternative to ordinary least squares regression and quantile regression. This makes it a highly effective and efficient choice for data analysis.

We consider arrival delay (ArrDelay) as a continuous variable by modeling log(ArrDelay-min(ArrDelay)+1), denoted as Y, as a linear function of other covariate variables: X_1, departure hour (rang 0 to 24); X_2, the distance between airports (in thousands of miles); X_3, day/night status (1 if departure between 8 p.m. and 5 a.m., 0 otherwise); and X_4, weekend/weekday status (1 if departure occurred during the weekend, 0 otherwise). Similar models were also established by Schifano et al. (2016) and Jiang et al. (2018).

Assess the performance of our one-step updated pilot composite quantile regression estimator (OSUPE) and conduct a comparative analysis with the OS-based simple averaging composite quantile regression estimators (OSE), the communication-efficient surrogate likelihood estimator (CSLE) and the global composite quantile regression (GE). For the purpose of comparison, we evaluate the performance of these estimators based on their out-of-sample prediction. The data are randomly split into a training set (D_{train}) of size $70\% \times N = 5,076,747$ and a testing set (D_{test}) of size $30\% \times N = 2,175,749$. The D_{train} is further divided into K subsets ($K = 50, 100, 500, 1000$), hence the sample size of $K-1$ subsets is $[5,076,747/K]$ and the remaining one is $5,076,747 - (K-1) \times [5,076,747/K]$. We then estimate the model coefficients using OSUPE, CSLE, GE, and OSE on the D_{train}, respectively. Based on the discussion in Sect. 4, here we set the number of quantiles $M = 19$, $n = [\sqrt{70\% \times N} \log(70\% \times N)]$. To compare the prediction accuracy of different estimators, we compute both the mean squared prediction error (MSE) and the mean absolute deviation (MAD) of the predictions based on D_{test} as

$$\text{MSE} = \frac{1}{N_{\text{test}}} \sum_{i \in D_{\text{test}}} \left(Y_i - \hat{Y}_i\right)^2, \text{MAD} = \frac{1}{N_{\text{test}}} \sum_{i \in D_{\text{test}}} \left|Y_i - \hat{Y}_i\right|,$$

where $\hat{Y}_i$ is the predicted value of Y_i, and N_{test}=2,175,749. Furthermore, to evaluate the computational efficiency of the proposed estimator, we record the computing time for different estimators. For fair comparison, we counted the average MSE (AMSE), the average MAD (AMAD), and the average CPU time used by 100 repetitions of each estimator. All methods were programmed using the R software and executed on a computer with the same configuration as described in Sect. 4. All results are reported in Table 4. After a thorough analysis of Table 4, the following conclusions can be drawn. In terms of estimation accuracy, the AMSE and AMAD values of OSE, CSLE, OSUOE, and the global estimator (GE) are very close, with CSLE performing slightly less favorably. This finding is consistent with the conclusions from Sect. 4, which suggests that these methods are statistically effective when data is randomly distributed across different workers. Concerning computational efficiency, although OSUOE requires three communication rounds, its time cost is comparable to that of OSE. In summary, all of these findings validate the excellent performance of our proposed new method and confirm the theoretical findings.

6 Conclusion

In this paper, we aim to design an effective algorithm for addressing large-scale CQR issues. We demonstrate that the statistical effectiveness of both the one-shot method and the iterative method relies on a crucial assumption: that data is assigned to different workers in a random manner. To overcome this limitation, we propose a new one-step upgraded pilot CQR method that not only inherits the strengths of composite quantile regression itself but also effectively addresses the computational challenges posed by large-scale data. Moreover, our theoretical results demonstrate its asymptotic equivalence to the oracle global estimator, without any restrictive assumptions about randomness. Both evaluations on simulation data and real-world data consistently confirm the good performance of our proposed approach.

Table 4 The estimates and CPU time (seconds) for the airline on-time data

K	Methods	Departure hour	Distance	Day	Weekend	AMSE	AMAD	Time
1	GE	0.3243	1.3134	−1.9925	0.6011	0.0278	0.1347	3597.42
50	OSE	0.3243	1.3130	−1.9922	0.6052	0.0279	0.1349	76.2918
	CSLE	0.3389	1.1574	−2.3987	0.5980	0.0314	0.1425	191.873
	OSUPE	0.3243	1.3135	−1.9923	0.6008	0.0279	0.1348	84.7251
100	OSE	0.3244	1.3106	−1.9934	0.6036	0.0281	0.1351	41.7334
	CSLE	0.3389	1.1575	−2.3987	0.5980	0.0314	0.1425	152.414
	OSUPE	0.3243	1.3136	−1.9928	0.6005	0.0279	0.1348	45.6278
500	OSE	0.3244	1.3126	−1.9940	0.6033	0.0281	0.1352	13.6535
	CSLE	0.3389	1.1578	−2.3988	0.5978	0.0314	0.1425	64.5359
	OSUPE	0.3244	1.3135	−1.9931	0.6010	0.0278	0.1348	15.3762
1000	OSE	0.3243	1.3141	−1.9937	0.6032	0.0281	0.1352	10.3683
	CSLE	0.3389	1.1583	−2.3990	0.5978	0.0314	0.1425	20.6514
	OSUPE	0.3243	1.3135	−1.9927	0.6008	0.0278	0.1348	11.2531

Furthermore, there are several potential avenues for future research that warrant exploration. Firstly, the transmission of pilot samples represents a trade-off that the one-step upgraded pilot CQR estimator must navigate to overcome challenges arising from non-randomness. As a result, the proposed estimator demonstrates suitability for problems characterized by small-to-moderate dimensions. However, its application to ultra-high-dimensional problems may prove impractical due to considerable communication costs. Therefore, investigating alternative strategies for feature selection or screening, leveraging the proposed estimator, would be a pertinent focus of inquiry.

Secondly, while the proposed distributed CQR procedure adeptly addresses independent big data, real-world applications often involve dependent data, such as time series, longitudinal, or network data. Hence, adapting the method to effectively accommodate dependent data poses a significant and pertinent research direction.

Thirdly, the current methodology centers on the linear model, presenting an essential yet constrained context. Expanding the method's applicability to nonlinear models or semiparametric models, such as single index models, would present an intriguing and valuable extension that warrants future exploration.

Appendix

A.1 Two Lemmas

Lemma 1 *Assume that conditions* (A1)–(A5) *hold, it follows that* $\hat{f}(\hat{b}_{\mathcal{P},m}) - f(b_m) = O_p\left(n^{-1/2}\right)$, *where* $m = 1, 2, \cdots, M$.

Proof. Define $K_h(x) = h^{-1}K(x/h)$, and let $\dot{K}_h(x) = h^{-2}\dot{K}(x/h)$ represent the first derivative of $K_h(x)$. Since $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathcal{P}} = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i - \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ and $\hat{f}(\hat{b}_{\mathcal{P},m}) = \frac{1}{N}\sum_{i=1}^{N} K_h(\hat{\varepsilon}_i - \hat{b}_{\mathcal{P},m})$, using the conventional Taylor expansion, it follows that

$$
\begin{aligned}
\hat{f}(\hat{b}_{\mathcal{P},m}) - f(b_m) =& N^{-1}\sum_{i=1}^{N} K_h\left\{\varepsilon_i - \boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \boldsymbol{\beta}) - \hat{b}_{\mathcal{P},m}\right\} - f(b_m)\\
=& N^{-1}\sum_{i=1}^{N}\left\{K_h(\varepsilon_i - \hat{b}_{\mathcal{P},m}) - f(b_m)\right\}\\
&+ N^{-1}\sum_{i=1}^{N}\dot{K}_h(\varepsilon_i - \hat{b}_{\mathcal{P},m} - \delta_i)\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \boldsymbol{\beta})\\
=& A_1 + A_2,
\end{aligned}
$$

where δ_i is between 0 and $\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \boldsymbol{\beta})$. We will now examine the order of A_1 and A_2 individually.

Firstly, following Theorem 11.6 of Jiang (2010), we obtain the result $A_1 = O_p(h^2) + O_p(nh)^{-1/2}$. Given that h is optimally chosen as $O\left(N^{-1/5}\right)$, it follows that the A_1 is $O_p(N^{-2/5})$. Secondly, observe that $N^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i$ is of order $O_p(1)$ and $\dot{K}(\cdot)$ is bounded as per Assumption (A3). The pilot estimator $\hat{\boldsymbol{\beta}}_{\mathcal{P}}$ is derived from a pilot sample of size n. Hence, under the fulfillment of Assumptions (A1) and (A2), it can be deduced from Zou and Yuan (2008) that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})=O_p(1)$ and $\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta}$ is $O_p\left(n^{-1/2}\right)$. Thus, it can be confirmed that $A_2=O_p\left(n^{-1/2}\right)$. From Assumption (A5), we know that $\hat{f}(\hat{b}_{\mathcal{P},m})-f(b_m)=O_p\left(n^{-1/2}\right)$ and the proof of Lemma 1 is completed.

Lemma 2 *Assume that (1) $n^2/N\to\infty$ and $n/N\to 0$, as $N\to\infty$; (2) $E\,\|\boldsymbol{X}_i\|^3<\infty$. Assuming Assumption (A5) is also satisfied, then*

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I(\hat{\varepsilon}_i<\hat{b}_{\mathcal{P},m})-I\left(\varepsilon_i<b_m\right)\right]\right\}\\
&=\sum_{m=1}^{M}f(b_m)\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})+o_p(N^{-1/2}).
\end{aligned}
$$

Proof. Let $\dot{f}(\cdot)$, $f(\cdot)$ and $F(\cdot)$ represent the the derivative function, the density function of ε_i, and the distribution function of $f(\cdot)$, respectively. According to Assumption (A3), there is a constant $\delta_0>0$ such that $\sup_{|x|<\delta_0}f(x)\le C_f$ and $\sup_{|x|<\delta_0}|\dot{f}(x)|\le C_f$. To establish the lemma, let us define

$$
\begin{aligned}
V_i=&\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I\left(\hat{\varepsilon}_i<\hat{b}_{\mathcal{P},m}\right)-I\left(\varepsilon_i<b_m\right)\right]\right\}\\
=&\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I\left(\varepsilon_i<\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})+\hat{b}_{\mathcal{P},m}\right)-I\left(\varepsilon_i<b_m\right)\right]\right\}.
\end{aligned}
$$

Write $\bar{V}=\bar{V}_a+\bar{V}_b$, where $\bar{V}=N^{-1}\sum_{i=1}^{N}V_i$, $\bar{V}_a=N^{-1}\sum_{i\in\mathcal{P}}V_i$ and $\bar{V}_b=N^{-1}\sum_{i\notin\mathcal{P}}V_i$. Using a technique analogous to that in the proof of (A.8) from Li et al. (2015), we obtain $n^{-1/2}\sum_{i\in\mathcal{P}}V_i=O_p(1)$, and hence,

$$
\bar{V}_a=o_p(N^{-1/2}). \tag{A.1-1}
$$

Next, let us examine $\bar{V}_b$. Define $\mathcal{F}$ as the σ-field created by the $\{(\boldsymbol{X}_i,Y_i):i\in\mathcal{P}\}\bigcup\{\boldsymbol{X}_i:1\le i\le N\}$. Then,

$$
E\left(V_i\mid\mathcal{F}\right)=\boldsymbol{X}_i\left\{\sum_{m=1}^{M}\left[F(\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})+\hat{b}_{\mathcal{P},m})-F(b_m)\right]\right\}.
$$

Utilizing Taylor expansion, and for any $D>0$, it follows that $E|\sum_{m=1}^{M}[F(\|\boldsymbol{X}_i\|\cdot\|\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta}\|+\hat{b}_{\mathcal{P},m})-F(b_m)]|\le P(\sqrt{n}\|\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta}\|\ge D)+MC_fE\|\boldsymbol{X}_i\|\cdot n^{-1/2}D$, indicating that

$$E\left|\sum_{m=1}^{M}\left[F(\|\boldsymbol{X}_i\|\cdot\|\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta}\|+\hat{b}_{\mathcal{P},m})-F(b_m)\right]\right|\to 0 \tag{A.1-2}$$

as $n\to\infty$, since $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})=O_p(1)$. Thus, using the Taylor expansion in conjunction with Eq. A.1-2,

$$\begin{aligned}
&E\left|N^{-1/2}\sum_{i\notin\mathcal{P}}V_i-E\left(V_i\mid\mathcal{F}\right)\right|^2\\
&=\frac{N-n}{N}E\left[V_i-E\left(V_i\mid\mathcal{F}\right)\right]^2\\
&\le\frac{N-n}{N}E\left\|\boldsymbol{X}_i\right\|^2\left\{\sum_{m=1}^{M}\left[I\left(\varepsilon_i<\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})+\hat{b}_{\mathcal{P},m}\right)-I\left(\varepsilon_i<b_m\right)\right]\right\}\\
&\le\frac{N-n}{N}E\left\|\boldsymbol{X}_i\right\|^3\cdot\sum_{m=1}^{M}\left[F(\|\boldsymbol{X}_i\|\cdot\|\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta}\|+\hat{b}_{\mathcal{P},m})-F(b_m)\right]\\
&\to 0,
\end{aligned}$$

as $N\to\infty$, it follows that

$$\bar{V}_b=N^{-1}\sum_{i\notin\mathcal{P}}E\left(V_i\mid\mathcal{F}\right)+o_p\left(N^{-1/2}\right). \tag{A.1-3}$$

Using Taylor expansion again, we derive

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i\notin\mathcal{P}}E\left(V_i\mid\mathcal{F}\right)-\sum_{m=1}^{M}f\left(b_m\right)\left(\frac{1}{N}\sum_{i\notin\mathcal{P}}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})\right|\\
=&O_p\left(n^{-1}\right)=o_p\left(N^{-1/2}\right),
\end{aligned}$$

here δ lies within the interval of 0 and $\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})$, which, together with Eqs. A.1-1, A.1-3, implies that

$$\bar{V}=\sum_{m=1}^{M}f\left(b_m\right)\left(\frac{1}{N}\sum_{i\notin\mathcal{P}}X_iX_i^{\mathrm{T}}\right)(\hat{\boldsymbol{\beta}}_{\mathcal{P}}-\boldsymbol{\beta})+o_p\left(N^{-1/2}\right).$$

Since $N^{-1}\sum_{i\in\mathcal{P}}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}=O_p(n/N)$, the proof of this lemma is completed.

A.2 Proof of Proposition 1

The results presented in Proposition 1 can be obtained through straightforward calculations. Hence, we omit the detailed derivation.

A.3 Proof of Theorem 1

To establish the conclusion of the theorem, it is enough to demonstrate that $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}} - \hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)} = o_p\left(N^{-1/2}\right)$. Based on Eqs. (6), (7), we can express $\hat{\boldsymbol{\beta}}_{\mathrm{CQR}} - \hat{\boldsymbol{\beta}}_{\mathrm{CQR}}^{(1)} = E_1 + E_2 + E_3$. Here,

$$
\begin{aligned}
E_1 &= Q\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}) - \tau_m\right]\right\},\\
E_2 &= -Q\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}) - I(\varepsilon_i < b_m)\right]\right\},\\
E_3 &= \frac{1}{\sum_{m=1}^{M} f(b_m)}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}) - I(\varepsilon_i < b_m)\right]\right\},
\end{aligned}
$$

where $Q = (\sum_{m=1}^{M} f(b_m))^{-1} - (\sum_{m=1}^{M} f(\hat{b}_{\mathcal{P},m}))^{-1}$.

Thus, to confirm the theorems conclusion, it is sufficient to show $E_1 = o_p(N^{-1/2})$, $E_2 = o_p(N^{-1/2})$, and $E_3 = o_p(N^{-1/2})$. Let us start with E_1. Utilizing Lemma 1 and Assumption (A1), it is evident that

$$
\sum_{m=1}^{M} f(\hat{b}_{\mathcal{P},m}) - \sum_{m=1}^{M} f(b_m) = O_p\left(n^{-1/2}\right) = o_p(1).
$$

Furthermore, Assumption (A3) implies that $\sum_{m=1}^{M} f(b_m) > 0$, which in turn leads to

$$
(\sum_{m=1}^{M} f(b_m))^{-1} - \sum_{m=1}^{M} f(\hat{b}_{\mathcal{P},m}))^{-1} = o_p(1).
$$

According to the Central Limit Theorem, we derive

$$
\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i^{\mathrm{T}}\left\{\sum_{m=1}^{M}\left[I(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}) - \tau_m\right]\right\} = O_p\left(N^{-1/2}\right).
$$

Consequently, we have $E_1 = o_p\left(N^{-1/2}\right)$.

We then analyze E_2. Based on the theorem assumptions and Lemma 2, it is clear that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i^{\mathrm{T}} \left\{ \sum_{m=1}^{M} \left[I\left(\hat{\varepsilon}_i < \hat{b}_{\mathcal{P},m}\right) - I\left(\varepsilon_i < b_m\right) \right] \right\} \\
=&\sum_{m=1}^{M} f(b_m) \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \right) (\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \boldsymbol{\beta}) + o_p(N^{-1/2}) \\
=&O_p\left(n^{-1/2}\right),
\end{aligned}$$

where the last equality is because $\hat{\boldsymbol{\beta}}_{\mathcal{P}} - \boldsymbol{\beta} = O_p\left(n^{-1/2}\right)$. According to Lemma 1 and Assumption (A3), we can infer that $\sum_{m=1}^{M} f(b_m))^{-1} - \sum_{m=1}^{M} f(\hat{b}_{\mathcal{P},m}))^{-1} = O_p\left(n^{-1/2}\right)$. This, together with Assumption (A1), implies that $E_2 = O_p\left(n^{-1}\right)$. Moreover, according to Assumption (A5), it follows that $E_2 = o_p\left(N^{-1/2}\right)$.

Finally, we investigate E_3. According to the theorem's assumptions and Lemma 2, it is apparent that $E_3 = o_p\left(N^{-1/2}\right)$. This completes the proof of Proposition 1.

Acknowledgements We would like to thank the Reviewers for their constructive comments which led to an improved presentation of the manuscript. This research was supported by the National Natural Science Foundation of China (#11471264), Postdoctoral Research Foundation of China (#2024M752706), Natural Science Research of Jiangsu Higher Education Institutions of China (#24KJB110030) and General Project of Philosophy and Social Science Research in Colleges and Universities of Jiangsu Province (#2024SJYB1531).

References

Battey H, Fan J, Liu H, Lu J, Zhu Z (2018) Distributed testing and estimation under sparse high dimensional models. Ann Stat 46:1352–1382

Chang KW, Lin CJ, Zhou M (2017) Distributed learning with limited communication. J Mach Learn Res 18:1–38

Chen Y, Xie M (2014) Distributed variable selection for high-dimensional generalized linear models with application to pancancer genomic data. Bioinformatics 30:1620–1626

Chen L, Zhou Y (2020) Quantile regression in big data: a divide and conquer based strategy. Comput Stat Data Anal 144:106892

Chen X, Liu W, Zhang Y (2019) Quantile regression under memory constraint. Ann Stat 47:3244–3273

Fan J, Chen J (1999) One-step local quasi likelihood estimation. J R Stat Soc: Ser B (Stat Methodol) 61:927–943

Fan J, Wang D, Wang K, Zhu Z (2019) Distributed estimation of principal eigenspaces. Ann Stat 47:3009–3031

Fan J, Guo Y, Wang K (2021) Communication-efficient accurate statistical estimation. J Am Stat Assoc 2021:1–11

Gopal S, Yang Y (2013) Distributed training of large-scale logistic models. Int Conf Mach Learn 2013:289–290

Gu Y, Zou H (2020) Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration. IEEE Trans Inf Theory 66:7132–7154

Hou Z, Ma W, Wang L (2023) Sparse and debiased lasso estimation and inference for high-dimensional composite quantile regression with distributed data. TEST 2023:1–21

Huang C, Huo X (2019) A distributed one-step estimator. Math Program 174:41–76

Jaggi M, Smith V, Takc M, Terhorst J, Krishnan S, Hofmann T, Jordan MI (2014) Communication-efficient distributed dual coordinate ascent. Adv Neural Info Process Syst 2014:3068–3076

Jiang J (2010) Large sample techniques for statistics. Springer, Cham

Jiang R, Hu X, Yu K, Qian W (2018) Composite quantile regression for massive datasets. Statistics 52:980–1004

Jin J, Liu S, Ma T (2023) Optimal subsampling algorithms for composite quantile regression in massive data. Statistics 57:811–843

Jordan MI, Lee JD, Yang Y (2018) Communication-efficient distributed statistical inference. J Am Stat Assoc 15:1–14

Kai B, Li R, Zou H (2010) Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. J R Stat Soc: Ser B (Stat Methodol) 72:49–69

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Li G, Li Y, Tsai CL (2015) Quantile correlations and quantile autoregressive modeling. J Am Stat Assoc 110:246–261

Li X, Liu J, Li XY (2018) A survey on data management in Internet of Things. IEEE Internet Things J 5:3758–3771

Pan R, Ren T, Guo B, Li F, Li G, Wang H (2022) A note on distributed quantile regression by pilot sampling and one-step updating. J Bus Econ Stat 40:1691–1700

Qu L, Hao M, Sun L (2022) Sparse composite quantile regression with ultrahigh-dimensional heterogeneous data. Stat Sin 32:459-C475

Rosenblatt JD, Nadler B (2016) On the optimality of averaging in distributed statistical learning. Info Inference J IMA 5:379–404

Schifano ED, Wu J, Wang C, Yan J, Chen MH (2016) Online updating of statistical inference in the big data setting. Technometrics 58:393–403

Shamir O, Srebro N, Zhang T (2014) Communication-efficient distributed optimization using an approximate Newton-type method. Int Conf Mach Learn 2014:1000–1008

Wang J, Kolar M, Srebro N, Zhang T (2017) Efficient distributed learning with sparsity. Int Conf Mach Learn 2017:3636–3645

Wang X, Yang Z, Chen X, Liu W (2019) Distributed inference for linear support vector machine. J Mach Learn Res 20:1–41

Wang K, Li S, Zhang B (2021) Robust communication-efficient distributed composite quantile regression and variable selection for massive data. Comput Stat Data Anal 161:107262

Wang K, Zhang B, Alenezi F, Li S (2022) Communication-efficient surrogate quantile regression for non-randomly distributed system. Inf Sci 588:425–441

Wang K, Jia J, Polat K, Sun X, Alhudhaif A, Alenezi F (2023) Smooth quantile regression and distributed inference for non-randomly stored big data. Expert Syst Appl 215:119418

Xu Q, Deng K, Jiang C, Sun F, Huang X (2017) Composite quantile regression neural network with applications. Expert Syst Appl 76:129–139

Xu Q, Cai C, Jiang C, Sun F, Huang X (2020) Block average quantile regression for massive dataset. Stat Pap 61:141–165

Yu J, Ai M, Ye Z (2024) A review on design inspired subsampling for big data. Stat Pap 65:467–510

Zhang Y, Wainwright MJ, Duchi JC (2012) Communication-efficient algorithms for statistical optimization. Adv Neural Info Process Syst 2012:1502–1510

Zhang T, Duchi J, Wainwright M (2013) Divide-and-conquer linear programming: Scaling up discrete choice models. Adv Neural Info Process Syst 2013:2366–2374

Zhao T, Cheng G, Liu H (2016) A partially linear framework for massive heterogeneous data. Ann Stat 44:1400–1437

Zhao W, Lian H, Song X (2017) Composite quantile regression for correlated data. Comput Stat Data Anal 109:15–33

Zhu X, Li F, Wang H (2021) Least-square approximation for a distributed system. J Comput Graph Stat 30:1004–1018

Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann Stat 36:1108–1126