Journal of Statistical Planning and Inference 241 (2026) 106326

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Privacy-preserving estimation for non-randomly distributed data

Xirui Liu [a], Ke Yang [b], Liwen Xu [c], Mixia Wu [b,*]

[a] School of Mathematical Sciences, Guizhou Normal University, Guiyang 550025, China
[b] School of Statistics and Data Science, Beijing University of Technology, Beijing 100124, China
[c] College of Sciences, North China University of Technology, Beijing 100144, China

ARTICLE INFO

Keywords:
Distributed estimation
Non-randomly distributed data
Privacy preserving
Bootstrap sample
Kullback–Leibler divergence

ABSTRACT

This paper investigates data distributed across various machines in a non-random manner. We introduce two innovative distributed estimators, tailored to accommodate varying levels of communication cost and data privacy protection. The proposed estimators adeptly addresses the challenges associated with the non-random distribution of data. Both methods are communication-efficient, necessitating only two rounds of communication between the Master and worker machines, and safeguard data privacy by solely sharing summary statistics. Under mild conditions, we establish the ℓ_2-error bound and the asymptotic distribution of the estimators. Theoretical analysis confirms that the proposed estimators are statistically efficient. Additionally, numerical simulations and two real-world applications demonstrate the good performance of the proposed methods.

1. Introduction

In various fields, individual entities such as local governments, hospitals, and research labs collect data independently. For instance, in financial studies, it is common for data to be stored across different bank branches. Some studies focus on integrating the raw data for analysis (Tang and Song, 2016; Huang et al., 2017). While these integrative methods can be effective, they are not always practical due to privacy concerns and the high communication costs. Breach of privacy resulting from data sharing has indeed become a growing concern in scientific studies, leading to the development of various privacy protection schemes (Cai et al., 2022; Chen et al., 2024). Aggregation technologies offer an alternative, enabling collaborative machine learning without requiring the sharing of raw data (Zhang et al., 2021).

Distributed learning approach, as a prominent framework for aggregation, seeks to train a global model by aggregating summary statistics from all clients. This approach addresses crucial topics such as M-estimation (Lin and Xi, 2011; Zhang et al., 2012; Liu and Ihler, 2014; Huang and Huo, 2019; Jordan et al., 2019), penalized regression (Chen and Xie, 2014; Lee et al., 2017; Battey et al., 2018), semi-parametric regression (Lv and Lian, 2022), quantile regression (Volgushev et al., 2019; Chen et al., 2019; Yang et al., 2023), principal component analysis (Garber et al., 2017; Fan et al., 2019).

The distributed algorithms discussed above assume that data across worker machines are randomly distributed. While this assumption simplifies theoretical analysis, it may prove overly restrictive in practice, as deviations from randomness can induce significant heterogeneity. To overcome the non-randomness across different datasets, several studies have been developed (Wang et al., 2021; Pan et al., 2022; Wang et al., 2022). These approaches result in statistically efficient estimators, regardless of whether the data are randomly distributed. However, these current methods require the transfer of a portion of the raw data, which risks privacy breaches.

* Corresponding author.
E-mail address: wumixia@bjut.edu.cn (M. Wu).

https://doi.org/10.1016/j.jspi.2025.106326
Received 8 October 2024; Received in revised form 16 June 2025; Accepted 8 July 2025
Available online 28 July 2025

In this paper, our objective is to develop distributed estimation methods specifically for non-randomly distributed data, with a focus on privacy protection. The proposed methods entails two rounds of communication between the Master and workers. In the first round, each worker computes the local maximum likelihood estimator (MLE) and sends the result to the Master. The Master then derives a KL estimate by minimizing the sum of KL divergences across all local models, using a pilot sample obtained through a parametric bootstrap procedure. In the second round, one-step update is applied to the KL estimate to achieve the optimal statistical efficiency. We employ two distinct update strategies for the KL estimate, to accommodate varying levels of communication cost and data privacy protection, resulting in two new distributed estimators: the Privacy-preserving Communication-Efficient Estimator (PCEE) and the Privacy-preserving Communication-Efficient Surrogate Estimator (PCESE). Both algorithms ensure data privacy by avoiding the transfer of raw data. Under mild conditions, we theoretically demonstrate that both of the proposed estimators achieve the same asymptotic efficiency as the global estimator. Their performance is studied through numerical simulations and a real data example.

The rest of this paper is organized as follows. Section 2 describes the problem setup and details of the proposed PCEE and PCESE methods. Theoretical properties of the PCEE and PCESE algorithms are shown in Section 3. In Section 4 and Section 5, we provide simulation results and real-world dataset examples to validate the finite sample performance of the proposed methods. Section 6 is the conclusion. The proof of the main results is given in Appendix.

2. The proposed estimation

2.1. Problem setup

Let $\{Z_i\}_{i=1}^N$ be a data set comprising N independent observations sampled from the distribution of $Z = (Y, X)$, where $Z_i = (Y_i, X_i)$. Here $Y_i \in \mathbb{R}^1$ is the response of interest and $X_i \in \mathbb{R}^p$ is the associated p-dimensional predictor. Conditional on X_i, assume that Y_i is randomly distributed with a probability density function $p\left(Y_i|X_i;\theta^*\right)$, where $\theta^* \in \Theta \subset \mathbb{R}^d$ denotes the true parameter vector. Let $l(Z;\theta) = -\log\, p(Y|X;\theta)$ be the loss function. Define the global loss function as

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} l(Z_i;\theta). \tag{1}$$

Then minimizing (1) results in an estimator of θ^*, i.e., $\hat{\theta} = \arg\min_{\theta\in\Theta} L(\theta)$, referred to as the global estimator. In distributed computing systems, the dataset $\{Z_i\}_{i=1}^N$ is typically partitioned across K worker nodes rather than being stored on a single machine. Specifically, let $\mathcal{S}_k$ be the index set of samples stored on the k-th worker, and let $n_k = |\mathcal{S}_k|$ denote the sample size of the k-th worker node such that $\cup_{k=1}^K \mathcal{S}_k = \{1,\dots,N\}$, $\sum_{k=1}^K n_k = N$, $\mathcal{S}_{k_1} \cap \mathcal{S}_{k_1} = \emptyset$ for any $k_1 \neq k_1$. The corresponding data subset stored on the k-th node can then be represented as $\{Z_i : i \in \mathcal{S}_k\} = \{Z_{k1},\dots,Z_{kn_k}\}$.

Clearly, for the case of the randomly distributed data, that is, $\mathcal{S}_1,\dots,\mathcal{S}_K$ are randomly partitioned from $\{1,\dots,N\}$ given the sample sizes $n_1,\dots,n_K$, and all local densities are the same as the global density function $p(Y|X;\theta)$. The global parameter vector θ^* can be estimated using well-established distributed estimation frameworks within the proposed approaches (see Zhang et al., 2012; Jordan et al., 2019; Huang and Huo, 2019).

However, real-world distributed systems, such as regional healthcare data partnerships or multi-center clinical trial networks, often exhibit non-random data distribution, meaning their characteristic alignment does not follow random partitioning. For the non-randomly distributed data, the sample index sets $\mathcal{S}_1,\dots,\mathcal{S}_K$ follow prescribed storage strategies. For example, the Census Income dataset analyzed in Section 5.2 is naturally divided by individuals' education levels. Since people with similar educational backgrounds tend to cluster geographically in both residential and occupational patterns, this partitioning introduces heterogeneity. Therefore, we consider a heterogeneous statistical environment where local densities may vary across workers.

Without loss of generality, we assume that the kth local density of Y given X and W_k is $p_k(Y|X,W_k;\beta_k)$, where W_k represents the worker-specific covariate vector for the k-th node, β_k is determined by a shared parameter vector θ and the k-th worker-specific parameter vector γ_k, denoted as $\beta_k = \beta_k(\theta,\gamma_k)$. Thus, the global density Y can be represented as

$$p(Y|X;\theta) = \sum_{k=1}^{K}\int p_k(Y|X,W_k;\beta_k) f_k(W_k|X)\,dW_k, \tag{2}$$

where $f_k(W_k|X)$ is the k-th local density of W_k given X. If W_k and X are independent, then $f_k(W_k|X) = f_k(W_k)$. This formulation generalizes several common assumptions found in the literature. For example, Zhu et al. (2021) assumed that all local densities share the same functional form and the covariate set as the global model, differing only in their parameters (i.e., $p_k(Y_|X,W_k;\beta_k) = p(Y|X;\beta_k)$ with $\beta_k = \theta+\gamma_k$). Gu and Chen (2023) consider the case that the local densities may differ in functional form and partial parameters, but share the same covariate set (i.e., $p_k(Y|X,W_k;\beta_k) = p_k(Y|X;\beta_k)$ with $\beta_k = (\theta,\gamma_k)$ to separate shared and worker-specific effects). In contrast, Duan et al. (2022) assumed that local densities may differ from the global model in both functional form and covariate sets, with parameter structure, where $p_k(Y|X,W_k;\beta_k) = p_k(Y|X,W_k;\beta_k)$ with $\beta_k = (\theta,\gamma_k)$.

Denote by $\{Y_{ki},X_{ki},W_{ki}\}_{i=1}^{n_k}$ the sample at the k-node. We will propose two distributed estimation methods for global parameter vector θ under non-randomly distributed data in the following subsections.

2.2. The PCEE algorithm

This approach involves two rounds of communication between the Master and worker machines.

In the first round, we compute a KL estimate of θ^* by minimizing the KL divergence between the global distribution $p(Y|X;\theta)$ and the estimated distribution $p_0(Y|X) = \sum_{k=1}^{K} \int p_k(Y|X, W_k; \hat{\beta}_k) f_k(W_k|X) dW_k$, that is

$$\begin{aligned} \theta^*_{KL} &= \arg\min_{\theta\in\Theta} \int p_0(Y|X) \log \frac{p_0(Y|X)}{p(Y|X;\theta)} dY \\ &= \arg\max_{\theta\in\Theta} \int p_0(Y|X)\ \log p(Y|X;\theta) dY, \end{aligned} \tag{3}$$

where $\hat{\beta}_k$ is the local MLE of β_k for the kth worker node, given by

$$\hat{\beta}_k = \arg\max_{\beta_k\in\mathbb{B}_k} \frac{1}{n_k} \sum_{i=1}^{n_k} \log p_k(Y_{ki}|X_{ki}, W_{ki}; \beta_k).$$

Since θ^*_{KL} is not computationally tractable in most cases, we adopt the parametric bootstrap procedure to estimate it. Firstly, a sample $\{\tilde{Y}_{ki}\}_{i=1}^{\tilde{n}_k}$ is generated from each local model $p_k(Y|\tilde{X}_{ki}, \tilde{W}_{ki}; \hat{\beta}_k)$ on the Master node, where $\tilde{n}_k = \frac{n_k}{N} \times n_0$, and n_0 is the pre-determined sample size of bootstrap pilot sample size, $\{(\tilde{X}_{ki}, \tilde{W}_{ki})\}_{i=1}^{\tilde{n}_k}$ are drawn from the density $f_k(X, W_k)$. Here $f_k(X, W_k)$ is the density of (X, W_k) on the kth worker node. Combining them together produces a pilot bootstrap sample on the Master, $\{\tilde{Z}_i\}_{i=1}^{n_0} = \bigcup\{\tilde{Z}_{ki}\}_{i=1}^{\tilde{n}_i}$, where $\tilde{Z}_{ki} = (\tilde{Y}_{ki}, \tilde{X}_{ki})$. Subsequently, we obtain an estimate of θ^*_{KL} by solving the optimization problem:

$$\hat{\theta}_{KL} = \arg\max_{\theta\in\Theta} \frac{1}{n_0} \sum_{k=1}^{K} \left(\sum_{i=1}^{\tilde{n}_k} \log p(\tilde{Y}_{ki}|\tilde{X}_{ki};\theta) \right) = \arg\min_{\theta\in\Theta} L_{\mathcal{M}}(\theta), \tag{4}$$

where $L_{\mathcal{M}}(\theta) = \frac{1}{n_0} \sum_{k=1}^{K} \sum_{i=1}^{\tilde{n}_k} l(\tilde{Z}_{ki};\theta)$.

Unlike the method proposed by Wang et al. (2021), which requires transferring partial raw data as pilot samples, our method constructs the pilot sample from K bootstrap samples generated from the estimated densities $p_k(\cdot;\hat{\beta}_k)$ on the master node, thus effectively reducing privacy risks.

Algorithm 1 The PCEE algorithm

Round 1: Compute KL Estimate

1. **For** Worker $k = 1, \ldots, K$ **do**:
   Compute and broadcast $\hat{\beta}_k$; .
2. **For** Master **do**:
   (1) draw samples $\{\tilde{X}_{ki}, \tilde{W}_{ki}\}_{i=1}^{\tilde{n}_k}$ from the density $f_k(X, W_k)$ for $k = 1, \ldots, K$;
   (2) Generate the bootstrap samples $\{\tilde{Y}_{ki}\}_{i=1}^{\tilde{n}}$ from the estimated local model $p_k(Y|\tilde{X}_{ki}, \tilde{W}_{ki}; \hat{\beta}_k)$ for $k = 1, \ldots, K$;
   (3) Compute and broadcast the KL estimate:
   $\hat{\theta}_{KL} = \arg\max_{\theta\in\Theta} \frac{1}{n_0} \sum_{k=1}^{K} \sum_{i=1}^{\tilde{n}_k} \log p(\tilde{Y}_{ki}|\tilde{X}_{ki};\theta)$.

Round 2: Compute the PCEE Estimate

1. **For** Worker $k = 1, \ldots, K$ **do**:
   Compute and broadcast the gradient vector and Hessian matrix at the $\hat{\theta}_{KL}$:
   $\nabla L_k(\hat{\theta}_{KL}) = \frac{1}{n_k} \sum_{i\in S_k} \nabla l(Z_i;\hat{\theta}_{KL})$ and $\nabla^2 L_k(\hat{\theta}_{KL}) = \frac{1}{n_k} \sum_{i\in S_k} \nabla^2 l(Z_i;\hat{\theta}_{KL})$;
2. **For** Master **do**:
   (1) Sum the derivatives respectively:
   $\nabla L(\hat{\theta}_{KL}) = \frac{1}{K} \sum_k \nabla L_k(\hat{\theta}_{KL})$ and $\nabla^2 L(\hat{\theta}_{KL}) = \frac{1}{K} \sum_k \nabla^2 L_k(\hat{\theta}_{KL})$;
   (2) Compute the one-step Newton–Raphson update for $\hat{\theta}_{KL}$:
   $\hat{\theta}_{PCEE} = \hat{\theta}_{KL} - \left(\nabla^2 L(\hat{\theta}_{KL})\right)^{-1} \nabla L(\hat{\theta}_{KL})$.

In the second round, an additional update step is implemented. We take $\hat{\theta}_{KL}$ as the initial point and apply Newton–Raphson one-step update. Specifically, the Master broadcasts $\hat{\theta}_{KL}$ to the workers, then each worker compute the first-order derivatives $\nabla L_k(\hat{\theta}_{KL})$ and the second-order derivatives $\nabla^2 L_k(\hat{\theta}_{KL})$ and report these derivatives back to the Master, where

$$L_k(\hat{\theta}_{KL}) = \frac{1}{n_k} \sum_{i=1}^{n_k} l(Z_{ki};\hat{\theta}_{KL}).$$

Finally, the Master performs one-step update to refine $\hat{\theta}_{KL}$, resulting in a more accurate estimate:

$$\hat{\theta}_{PCEE} = \hat{\theta}_{KL} - \left(\frac{1}{K} \sum_{k=1}^{K} \nabla^2 L_k(\hat{\theta}_{KL}) \right)^{-1} \frac{1}{K} \sum_{k=1}^{K} \nabla L_k(\hat{\theta}_{KL}). \tag{5}$$

The detailed algorithm of the PCEE is described in Algorithm 1, where the density $f_k(X, W_k)$ on the is assumed to be known. If they are unknown, they can be replaced with the estimated densities $\hat{f}_k(X, W_k)$s on the kth node.

Note that computing $\hat{\theta}_{PCEE}$ requires each node worker to transfer a $d \times d$-dimensional Hessian matrix to the Master node. To further reduce both communication costs and second-order information exposure risks, we adopt a surrogate likelihood function approach following Jordan et al. (2019) to update $\hat{\theta}_{KL}$, which leads to our proposed PCESE algorithm in the following subsection.

2.3. The PCESE algorithm

The PCESE algorithm also involves two rounds of communication. The first round is the same with that of the PCEE algorithm, we obtain the initial estimate $\hat{\theta}_{KL}$ from (4). In the second round, we perform one-step update of $\hat{\theta}_{KL}$ using a surrogate likelihood function. By applying a Taylor expansion of $L(\theta)$ around the initial estimate $\hat{\theta}_{KL}$, we derive

$$L(\theta) = L(\hat{\theta}_{KL}) + \langle \nabla L(\hat{\theta}_{KL}), \theta - \hat{\theta}_{KL} \rangle + R_N(\theta).$$

where $R_N(\theta) = \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j L(\hat{\theta}_{KL})(\theta - \hat{\theta}_{KL})^{\otimes j}$ is the remainder term. However, transferring the higher-order derivatives across sites can be costly. Since the pilot bootstrap sample on the Master are independent and identically distributed, it holds that $\nabla^j L_{\mathcal{M}}(\hat{\theta}_{KL}) - \nabla^j L(\hat{\theta}_{KL}) = o_P(1)$ for any $j \in \{0, 1, \dots\}$. This motivates us to replace the global remainder term $R_N(\theta)$ with the pilot bootstrap sample version $R_{\mathcal{M}}(\theta)$, that is

$$R_{\mathcal{M}}(\theta) = L_{\mathcal{M}}(\theta) - L_{\mathcal{M}}(\hat{\theta}_{KL}) - \langle \nabla L_{\mathcal{M}}(\hat{\theta}_{KL}), \theta - \hat{\theta}_{KL} \rangle,$$

where $L_{\mathcal{M}}(\theta)$ is defined in (4). Consequently, the approximation version $\widetilde{L}(\theta)$ of $L(\theta)$ is given by

$$\widetilde{L}(\theta) = L(\hat{\theta}_{KL}) + \langle \nabla L(\hat{\theta}_{KL}), \theta - \hat{\theta}_{KL} \rangle + L_{\mathcal{M}}(\theta) - L_{\mathcal{M}}(\hat{\theta}_{KL}) - \langle \nabla L_{\mathcal{M}}, \theta - \hat{\theta}_{KL} \rangle.$$

Removing the additive constants term, we redefine $\widetilde{L}(\theta)$ as follows,

$$\widetilde{L}(\theta) = L_{\mathcal{M}}(\theta) - \langle \nabla L_{\mathcal{M}}(\hat{\theta}_{KL}) - \nabla L_N(\hat{\theta}_{KL}), \theta \rangle. \tag{6}$$

Thus, the PCESE estimator is obtained by minimizing $\widetilde{L}(\theta)$, that is,

$$\hat{\theta}_{PCESE} = \arg\min_{\theta \in \Theta} \widetilde{L}(\theta). \tag{7}$$

The minimization of $\widetilde{L}(\theta)$ is carried out on the Master machine, while the workers only need to compute and transmit the local MLE and gradients. Therefore, the PCESE is more communication-efficient than the PCEE.

The detailed algorithm of the PCESE is described in Algorithm 2.

Algorithm 2 The PCESE algorithm

Round 1: Obtain $\hat{\theta}_{KL}$, which is same as the Round 1 of the PCEE algorithm.

Round 2: Compute the PCESE Estimate

1. **For** Worker $k = 1, \dots, K$ **do**:

   Compute and broadcast the gradient vector $\nabla L_k(\hat{\theta}_{KL})$:

   $$\nabla L_k(\hat{\theta}_{KL}) = \sum_{i \in S_k} \nabla l(Z_i; \hat{\theta}_{KL}).$$

2. **For** Master **do**:

   (1) Compute the gradients $\nabla L_{\mathcal{M}}(\hat{\theta}_{KL})$ based on the pilot bootstrap sample and sum the gradients:

   $$\nabla L_{\mathcal{M}}(\hat{\theta}_{KL}) = \frac{1}{n_0} \sum_{k=1}^{K} \sum_{i=1}^{\tilde{n}_k} \nabla l(\tilde{Z}_{ki}; \hat{\theta}_{KL}),$$
   $$\nabla L(\hat{\theta}_{KL}) = \frac{1}{K} \sum_k \nabla L_k(\hat{\theta}_{KL});$$

   (2) Compute the surrogate loss function:

   $$\widetilde{L}(\theta) = L_{\mathcal{M}}(\theta) - \langle \nabla L_{\mathcal{M}}(\hat{\theta}_{KL}) - \nabla L_N(\hat{\theta}_{KL}), \theta \rangle;$$

   (3) Compute the surrogate estimate:

   $$\hat{\theta}_{PCESE} = \arg\min_{\theta \in \Theta} \widetilde{L}(\theta).$$

3. Theoretical results

In this section, we focus on the consistency and the asymptotic normality results of the proposed estimators, the proofs of theorems are given in Appendices. For convenience, we first pose some technical assumptions on the parameter space and the loss function.

Let $B_\rho = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 \le \rho\} \subset \Theta$ be a ball of radius ρ around the true parameter θ^*. Similarly, let $B_\eta^k = \{\beta_k : \|\beta_k - \beta_k^*\|_2 \le \eta\} \subset \mathbb{B}_k$ be a ball of radius η around the true parameter β_k^* for $k = 1, 2, \dots, K$. For a vector $v \in \mathbb{R}^d$, $\|v\|_2 = (\sum_{i=1}^d v_i^2)^{1/2}$. For a matrix $V \in \mathbb{R}^{d\times d}$, $\|V\|_2$ denote its maximum singular value, and $\|V\|_2 = \sup_{u\in\mathbb{R}^d:\|u\|\le 1} \|Vu\|_2$. The notations $\lambda_{min}(V)$ and $\lambda_{max}(V)$ denote the minimum and maximum eigenvalues of the matrix V, respectively.

Assumption 1. The global parameter space $\Theta \in \mathbb{R}^d$ and local parameter space $\Theta_k' \in \mathbb{R}^{d+q_k}$ are compact convex sets, and θ^* is an interior point of Θ, β_k^* is an interior point of Θ_k', $k = 1, \dots, K$.

Assumption 2. The population Hessian matrix $I(\theta) = \mathbb{E}(\nabla^2 l(Z;\theta))$ is nonsingular at θ^*: there exist two positive constants (λ_-, λ_+) such that $\lambda_- \le \lambda_{min}(I(\theta^*)) \le \lambda_{max}(I(\theta^*)) \le \lambda_+$.

Assumption 3. There exist constants G, H and a function $M(z)$ such that

$$\mathbb{E}\left[\|\nabla l(Z;\theta)\|_2^8\right] \le G^8, \quad \mathbb{E}\left[\|\nabla^2 l(Z;\theta) - I(\theta)\|_2^8\right] \le H^8, \text{ for all } \theta \in B_\rho,$$
$$\|\nabla^{20} l(Z;\theta) - \nabla^2 l(Z;\theta')\|_2 \le M(Z)\|\theta - \theta'\|_2, \text{ for all } \theta, \theta' \in B_\rho,$$

where $M(Z)$ satisfies $\mathbb{E}\left[M(Z)^8\right] \le M^8$ for constant $M > 0$.

Assumption 4. For any positive constants δ and ρ, there exist $\varepsilon > 0$ and $\xi > 0$, such that

$$\liminf_{N\to\infty} \mathbb{P}\{\inf_{\|\theta-\theta^*\|_2\ge\delta} \left(L(\theta) - L(\theta^*) \right) \ge \varepsilon\} = 1;$$
$$\liminf_{n_k\to\infty} \mathbb{P}\{\inf_{\|\beta_k-\beta_k^*\|_2\ge\rho} \left(L_{0k}(\beta_k) - L_{0k}(\beta_k^*) \right) \ge \xi\} = 1,$$

where $L_{0k}(\beta_k) = \frac{1}{n_k}\sum_{i=1}^{n_k}\{-\log p_k(Y_{ki}|X_{ki}, W_{ki};\beta_k)\}$.

Assumption 5. As $N \to \infty$, it has $n_k \to \infty$ for all k.

Assumption 1 specifies the relationship between the parameter space and the true parameter. Assumption 2 imposes a local identifiability condition, guaranteeing that θ^* is a local minimum. Assumption 3 puts a constraint on smoothness for $l(Z;\theta)$. Assumption 4 is a identifiability condition, necessary for establishing the consistency of the estimator (Van der Vaart, 2000). Assumption 5 is a necessary condition that the ML estimator $\hat\beta_k$ for each local model is consistent.

Theorem 1. *Under Assumptions 1–5, one has*

$$\hat\theta_{PCEE} - \theta^* = -I(\theta^*)^{-1}\frac{1}{N}\sum_{k=1}^K\sum_{i\in S_k}\nabla l(Z_{ki};\theta^*) + O_P(n_0^{-1}).$$

If $n_0/\sqrt{N} \to \infty$ and $N \to \infty$, then the asymptotic normality of the PCEE estimator holds,

$$\sqrt{N}(\hat\theta_{PCEE} - \theta^*) \to \mathcal{N}(0, \Sigma)$$

where $\Sigma = I(\theta^*)^{-1}\mathbb{E}[\nabla l(Z;\theta^*)\nabla l(Z;\theta^*)^\top]I(\theta^*)^{-1}$.

It can be concluded from Theorem 1 that the error $\hat\theta_{PCEE} - \theta^*$ comprises two components: the variance term and the bias term. Remarkably, the variance order matches that of the global estimator $\widehat\theta$, which is $O_P(N^{-1/2})$. The bias term converges to zero at a rate of $1/n_0$. Therefore Theorem 1 indicates that the convergence rate of the ℓ_2 estimation error for PCEE estimator is $\|\hat\theta_{PCEE} - \theta^*\|_2 = O_P(\sqrt{1/N} + 1/n_0)$. If the pilot bootstrap sample size is sufficiently large, the estimation efficiency will match that of the global estimator.

Theorem 2. *Suppose that Assumptions 1–5 holds, it has*

$$\|\hat\theta_{PCESE} - \widehat\theta\|_2 \le C_2\left(\|\hat\theta_{KL} - \widehat\theta\|_2 + \|\widehat\theta - \theta^*\|_2 + \|\nabla^2 L_{\mathcal{M}}(\theta^*) - \nabla^2 L(\theta^*)\|_2\right)\|\hat\theta_{KL} - \widehat\theta\|_2,$$

with probability at least $1 - C_1 K n_0^{-4}$, where $\widehat\theta$ is the global estimator and constants C_1, C_2 are independent of (K, n_0, N).

Under Assumptions 1–5, it can be inferred that $\|\widehat\theta - \theta^*\|_2 = O_P(N^{-1/2})$ and $\|\nabla^2 L_{\mathcal{M}}(\theta^*) - \nabla^2 L(\theta^*)\|_2 = O_P(n_0^{-1/2})$, where the second equality follows from Lemma B.2 in Appendix B. Thus, one has $\|\hat\theta_{PCESE} - \widehat\theta\|_2 = O_P(n_0^{-1})$.

Theorem 3. *Under Assumptions 1–5, the PCESE estimator defined in (7) satisfies*

$$\hat\theta_{PCESE} - \theta^* = -I(\theta^*)^{-1}\frac{1}{N}\sum_{k=1}^K\sum_{i\in S_k}\nabla l(Z_{ki};\theta^*) + O_P(n_0^{-1}).$$

If $n_0/\sqrt{N} \to \infty$ *and* $N \to \infty$*, then the asymptotic normality of the PCESE estimator holds,*

$$\sqrt{N}(\hat{\theta}_{PCESE} - \theta^*) \to \mathcal{N}(0, \Sigma),$$

where $\Sigma = I(\theta^*)^{-1}\mathbb{E}[\nabla l(Z;\theta^*)\nabla l(Z;\theta^*)^\top]I(\theta^*)^{-1}$.

Theorems 1 and 3 show that $\hat{\theta}_{PCESE}$ and $\hat{\theta}_{PCEE}$ share the same asymptotic distribution. In other words, for sufficiently large values of N and n_0, $\hat{\theta}_{PCESE}$ can perform as effectively as $\hat{\theta}_{PCEE}$, while maintaining lower communication costs. Moreover, both the $\hat{\theta}_{PCESE}$ and $\hat{\theta}_{PCEE}$ can be as efficient as the global estimator in the sense of that they enjoy the same asymptotic distribution.

4. Numerical simulation

In this section, we conduct simulations to demonstrate the performance of the two proposed estimators. We consider two regimes: (1) the number of Workers, K, is fixed at $K = 10$, and the whole sample sizes, N, ranges from 1×10^4 to 7×10^4; (2) the whole sample sizes, N, is fixed at $N = 2 \times 10^4$, while the number of Workers, K, varies from 2 to 200. The sample size of each worker is the same as $n = N/K$. And Unbalanced case is given in Section 5.2. Across all data generative models, we fix the dimension of covariate $d = 5$. We set the pilot bootstrap sample size $n_0 = N \times 5\%$. We consider three data-generating mechanisms:

(i) **Logistic Regression Model**. For each sample $i \in \mathcal{S}_k$, the covariate X_i is generated from a multivariate normal distribution with mean zero and covariance matrix $\alpha_k \Sigma_0$, where $\Sigma_0 = (\sigma_{j_1 j_2}) \in \mathbb{R}^{5\times5}$ with $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ and α_k is a constant. Given X_i, the response Y_i is generated from a logistic regression model. That is, $Y_i \in \{0, 1\}$ is a binary response variable with

$$P\left(Y_i = 1 \mid X_i\right) = \exp(X_i^\top \theta) / \left\{1 + \exp(X_i^\top \theta)\right\}. \tag{8}$$

The true parameter $\theta = (0, 0, -0.1, 0.1, 0)$.

(ii) **Poisson Regression Model**. Same as the mechanism (i), the covariate X_i is generated from a multivariate normal distribution with mean zero and covariance matrix $\alpha_k \Sigma_0$, where $\Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{5\times5}$ with $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ and α_k is a constant. Given X_i, the response Y_i is generated from a Poisson distribution as $P\left(Y_i = m \mid X_i, \theta\right) = \lambda_i^m \exp\left(-\lambda_i\right)/m!$, where $\lambda_i = \exp\left(X_i^\top \theta\right)$ and $\theta = (0, 0, -0.2, 0.5, 0)$.

(iii) **Mis-specified Model**. Suppose an experimenter postulates the logistic regression model for the sample of observations $Z_1, \ldots, Z_n$. However, the true underlying model is the probit regression model. The true parameter $\theta = (0, 0, -0.1, 0.1, 0)$ is the same as the mechanism (i).

To study the performance of the proposed estimators in a non-randomly distributed mechanism, we consider four data-allocating cases. In the first case, data on different workers are randomly distributed. This is the ideal Case which is assumed by most distributed algorithms. The Case II allows the data stored on different workers to be heterogeneous, while the regression relationship remains the same. The Case III allows that the data allocation mechanism is related to the X. For Case IV, the data allocation mechanism is related to both the predictors Y and the covariates X, resulting in a non-random distribution of the entire dataset across K workers. The four data allocation mechanisms are as follows:

Case I. Set $\alpha_k = 1$, the data on different workers are randomly distributed.

Case II. Set $\alpha_k = 1/k$ for $k = 1, \ldots, K$. Thus, the randomness of the distributed data is violated, while the conditional regression relationship $Y_i \mid X_i$ remains the same for various workers.

Case III. The storage location of each observation depends on its first covariate. Specifically, let $X_{(i)1}$ be the ith order index of X_{i1}, $i = 1, \ldots, N$, and the (i)th observation $(X_{(i)}, Y_{(i)})$ is allocated on the kth local machine if it satisfies $(i) \in [(k-1)n + 1, kn]$.

Case IV. Let $U_i = Y_i + X_i^T \gamma$ with $\gamma = (1, \ldots, 1)^\top \in \mathbb{R}^d$. sorting the whole sample according to U_i. Let $U_{(1)} \le \cdots \le U_{(N)}$ be the order index statistic of the U_is. The (i)th observation $(X_{(i)}, Y_{(i)})$ is assigned on the kth local machine if it satisfies $(i) \in [(k-1)n + 1, kn]$.

To make the evaluation reliable, we compare the PCEE and the PCESE estimators with:

- GLO: the global estimator which minimizes the global loss function (1).
- AVG: the average estimator proposed by Zhang et al. (2012).
- CSL: the Communication-efficient Surrogate Likelihood estimator proposed by Jordan et al. (2019).
- DOS: Distributed One-Step estimator proposed by Huang and Huo (2019).
- DLSA: the Distributed Least-Square Approximation estimator proposed by Zhu et al. (2021).
- OSUP: One-Step Upgraded Pilot estimator proposed by Wang et al. (2021).

Among these distributed estimators, only the OSUP estimator is designed for non-randomly distributed data, which requires transferring a proportion of raw data as a pilot sample. The simulation is repeated $B = 500$ times. We examine the MSE of each estimator, namely, $\mathrm{MSE}(\hat{\theta}) = B^{-1} \sum_{b=1}^{B} \|\hat{\theta}_{(b)} - \theta\|_2^2$.

4.1. Effects of the total sample size

Figs. 1–2 depict the MSE for the 8 estimators versus the varying total sample size $N = 10^4\delta$ ($\delta = 1, \ldots, 7$) in logistic regression and Poisson regression models, respectively. Based on our analysis, we could draw the following conclusions.

First of all, when the data are randomly distributed across local sites in Case I, all the estimators exhibit similar performance with the GLO estimator. In terms of communication complexity, the DOS, OSUP, DLSA and PCEE methods require at least $O(p^2)$ bits

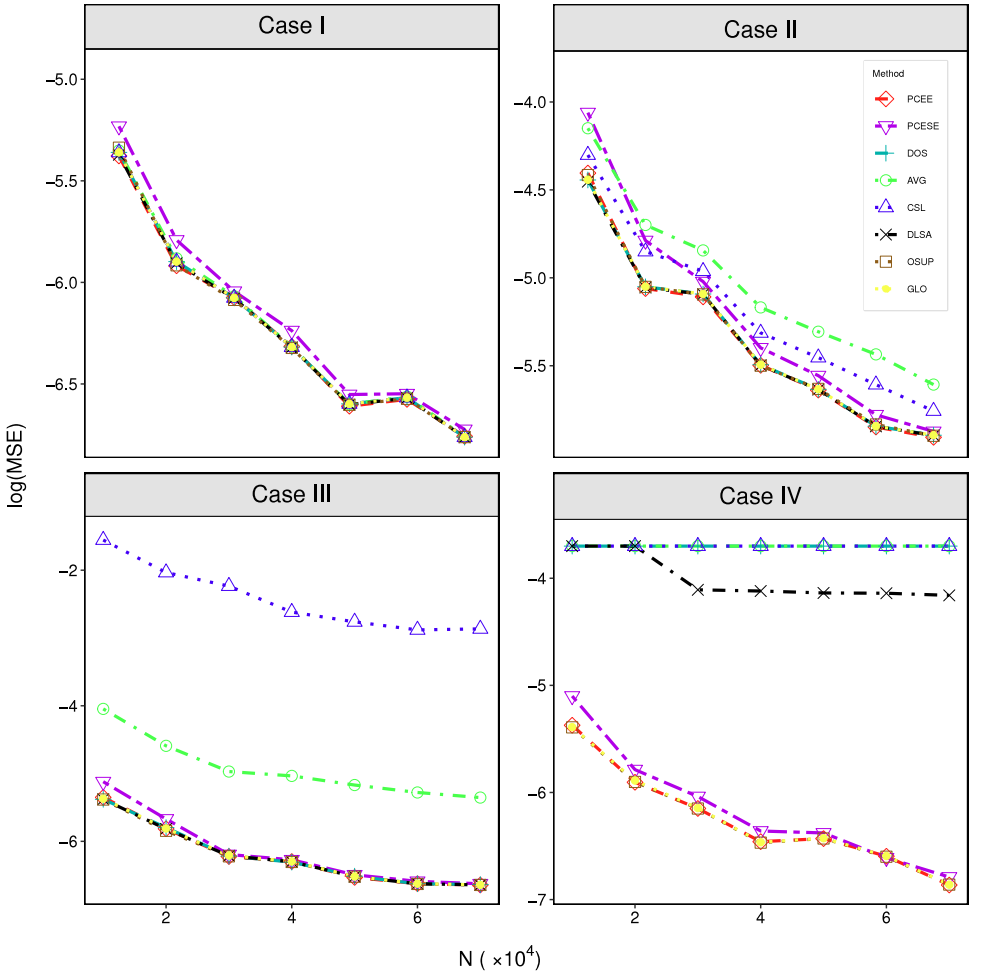

Fig. 1. The logarithm of the MSE for all estimators in logistic regression varies with the full sample size N, with the number of worker K is fixed at $K = 10$. In all cases, each point represents the average of 500 replications. In Case I, the whole data is split randomly. The Case II, the data is heterogeneous. The data allocation mechanism in Case III depends solely on predictors X, while in Case IV, it depends on both predictors X and the response Y.

of communication, whereas the AVG, CSL, and our proposed PCESE methods only require $O(p)$ bits. However, for heterogeneous data in Case II, the discrepancies among various estimators become more pronounced. Among them, the PCEE, PCESE, DOS and OSUP estimators perform similarly, while the CSL and AVG estimators is inferior to the others because their MSEs do not approach that of the GLO estimator as the N increases again. And the DLSA estimator is sensitive to the regression model, it performs as well as the GLO estimator under the logistic regression model, but worst under the Poisson regression model.

In Cases III and IV, where the randomness condition is strongly violated, it is more clear that the proposed estimators, the PCEE and the PCESE, are superior to other distributed estimators. In Case III, where the data allocation mechanism depends on X_i, the PCEE and PCESE uniformly outperform both the CSL and the AVG estimators. In Case IV, where data allocation depends on both X_i and Y_i, only our methods and the OSUP estimator achieve performance comparable to the GLO estimator, significantly outperforming the DOS, CSL, AVG and DLSA estimators. It is worthy noting that the OSUP estimator requires transferring a proportion of raw data from work machines to the Master machine, while our methods require only the transmission of statistics. The privacy preservation and communication efficiency highlight the advantages over the OSUP. When the size of the pilot bootstrap sample, n_0, is small, the approximation error is relatively significant. However, this issue can be mitigated by generating a larger pilot bootstrap sample on the Master. Increasing n_0 will not elevate the communication cost of the proposed methods, but it will increase the communication cost of the OSUP. This highlights a significant advantage of our methods over the OSUP. Figs. 1–2 also show that the PCESE slightly underperforms the PCEE and OSUP, with the difference in MSEs between the two estimators approaching zero as N increases, except in the scenario involving Poisson regression variables in Case IV. However, the PCESE outperforms the PCEE and OSUP in communication costs by eliminating the need for workers to transmit their Hessian matrices to the Master.

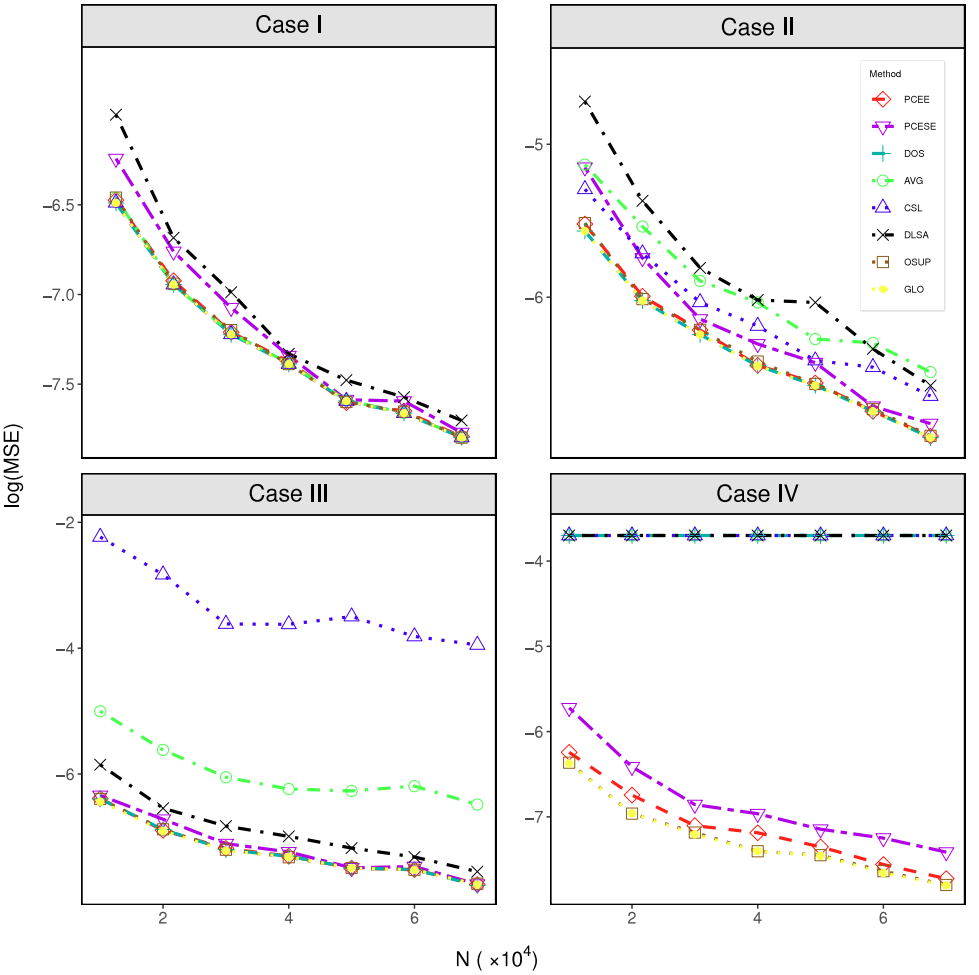

Fig. 2. The logarithm of the MSE for all estimators in Poisson regression varies with the full sample size N, with the number of worker K is fixed at $K = 10$. In all cases, each point represents the average of 500 replications. In Case I, the whole data is split randomly. The Case II, the data is heterogeneous. The data allocation mechanism in Case III depends solely on predictors X, while in Case IV, it depends on both predictors X and the response Y.

4.2. Effects of number of local machines

In this subsection, we examine the performance of the proposed and other competing estimators versus the different number of local machines. Tables 1–2 report the logarithm of MSEs of the different estimators versus the number of machines, K, ranging from 2 to 200 for Logistic regression and Poisson regression models, respectively. The total sample size, N, is fixed at $N = 2 \times 10^4$. In all cases, each cell in the table corresponds to the average of 500 replications. From Tables 1–2, it can be observed that our methods perform almost as well as the GLO estimator in all scenarios. In cases II-IV, the AVG, CSL, DOS, and DLSA estimators fail in succession.

4.3. Effects of mis-specified model

To further evaluate the robustness of our methods, we conducted a model misspecification analysis. Specifically, given X_i, the binary response Y was generated from a probit model with

$$P\left(Y_i = 1 \mid X_i\right) = \Phi(X_i^{\top}\theta),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. However, the parameter estimation was performed using a logistic regression model (8), introducing model misspecification.

Fig. 3 presents the log-MSEs of the all estimators. Notably, all estimators show higher MSE values compared to the correctly specified case in Fig. 1, reflecting the inherent bias caused by model misspecification. Nevertheless, the proposed PCEE and PCESE

Table 1
The logarithm of MSE of the different estimators in Poisson regression model versus the number of machines K ranging for 2 to 200.

Case	Method	K				
		2	10	50	100	200
I	PCEE	−6.185	−6.179	−6.162	−6.104	−6.177
	PCESE	−6.065	−6.086	−6.053	−5.979	−5.845
	AVG	−3.620	−4.440	−4.314	−4.339	−4.332
	CSL	−6.141	−6.152	−5.953	−5.582	−5.379
	DOS	−6.170	−6.176	−6.168	−6.100	−6.174
	DLSA	−6.116	−6.112	−6.148	−6.156	−6.217
	OSUP	−6.173	−6.162	−6.172	−6.098	−6.177
	GLO	−6.168	−6.169	−6.158	−6.086	−6.163
II	PCEE	−5.629	−4.913	−4.219	−3.542	−3.067
	PCESE	−5.534	−4.680	−3.866	−3.123	−2.494
	AVG	−3.582	−4.230	−3.807	−3.533	−3.025
	CSL	−5.249	−4.680	−3.935	−3.617	−3.069
	DOS	−5.620	−4.925	−4.255	−3.635	−3.210
	DLSA	−5.545	−4.880	−4.281	−3.686	−3.326
	OSUP	−5.611	−4.959	−4.166	−3.661	−3.058
	GLO	−5.620	−4.920	−4.253	−3.633	−3.213
III	PCEE	−6.219	−6.133	−6.178	−6.155	−6.147
	PCESE	−6.133	−5.967	−6.076	−6.010	−6.029
	AVG	−3.606	−3.557	−1.774	−0.547	0.399
	CSL	−4.790	−1.994	2.128	8.005	20.436
	DOS	−6.230	−6.105	−2.839	0.803	3.619
	DLSA	−6.142	−6.009	−6.170	−6.136	−6.156
	OSUP	−6.219	−6.144	−6.183	−6.153	−6.161
	GLO	−6.222	−6.132	−6.176	−6.164	−6.145
IV	PCEE	−6.171	−6.187	−6.083	−6.194	−6.065
	PCESE	−6.110	−6.065	−5.995	−6.061	−5.962
	AVG	−1.439	9.899	5.355	5.288	5.264
	CSL	−0.149	26.927	58.653	46.244	42.569
	DOS	−2.617	21.516	17.993	17.643	17.531
	DLSA	−5.800	12.159	−2.712	−2.574	−2.422
	OSUP	−6.171	−6.187	−6.088	−6.204	−6.062
	GLO	−6.164	−6.176	−6.080	−6.192	−6.061

Note: The total sample size N is fixed at $N = 2 \times 10^4$. In all cases, each cell in the table corresponds to the average of 500 replications.

estimators maintain performance comparable to the global estimator, particularly for large sample size N. Furthermore, Additionally, iterative updating further improves the accuracy of the proposed estimators.

To validate the accuracy improvement from iterative updating, we conducted an experiment with a fixed total sample size ($N = 10\,000$) distributed across $K = 10$ local machines. The pilot sample size was determined by $n_0 = N \times \pi$, where $\pi = 0.01$ represents the bootstrap pilot sampling proportion, yielding $n_0 = 100$. With a covariate dimension of $p = 5$, we compared the MSEs of the GLO estimator against multi-step PCEE and PCESE estimators. Table 3 presents the MSEs of the GLO and multi-step PCEE, PCESE estimators. Table 3 shows that iterative updating significantly enhances accuracy. Notably, while all multi-step estimators outperform their initial versions, the 2-step PCEE and 4-step PCESE estimators achieve MSEs comparable to the GLO estimator in this experimental setup.

4.4. Effects of the pilot sample size

We conduct a simulation study to assess the impact of pilot sample size ($n_0 = N \times \pi$, where $\pi = 1\%, 5\%, 10\%, 20\%$) on the MSEs of our proposed distributed estimators. The experiment maintains fixed sample size $N = 10\,000$ distributed across $K = 10$ local machines, while varying the bootstrap sampling proportion π. The empirical results in Table 4 demonstrate that the performance gap between estimators decreases monotonically with increasing pilot sampling proportion π. Notably, at $\pi = 5\%$, MSE values of both proposed estimators approach that does the GLO. This convergence suggests that the proposed methods attain asymptotic equivalence to the global estimator with moderate pilot sampling.

4.5. Compared with federated learning algorithm

Federated Learning (FL) approaches also address the challenge of data non-randomness, particularly under constraints of privacy and limited communication. Among them, the FedAvg algorithm (McMahan et al., 2017) addresses non-IID (non-independent and identically distributed) data by employing an iterative model averaging strategy. In this section, we compare the performance of

Table 2
The logarithm of MSE of the different estimators in Poisson regression model versus the number of machines K ranging for 2 to 200.

Case	Method	K				
		2	10	50	100	200
I	PCEE	−6.185	−6.179	−6.162	−6.104	−6.177
	PCESE	−6.065	−6.086	−6.053	−5.979	−5.845
	AVG	−3.620	−4.440	−4.314	−4.339	−4.332
	CSL	−6.141	−6.152	−5.953	−5.582	−5.379
	DOS	−6.170	−6.176	−6.168	−6.100	−6.174
	DLSA	−6.116	−6.112	−6.148	−6.156	−6.217
	OSUP	−6.173	−6.162	−6.172	−6.098	−6.177
	GLO	−6.168	−6.169	−6.158	−6.086	−6.163
II	PCEE	−5.629	−4.913	−4.219	−3.542	−3.067
	PCESE	−5.534	−4.680	−3.866	−3.123	−2.494
	AVG	−3.582	−4.230	−3.807	−3.533	−3.025
	CSL	−5.249	−4.680	−3.935	−3.617	−3.069
	DOS	−5.620	−4.925	−4.255	−3.635	−3.210
	DLSA	−5.545	−4.880	−4.281	−3.686	−3.326
	OSUP	−5.611	−4.959	−4.166	−3.661	−3.058
	GLO	−5.620	−4.920	−4.253	−3.633	−3.213
III	PCEE	−6.219	−6.133	−6.178	−6.155	−6.147
	PCESE	−6.133	−5.967	−6.076	−6.010	−6.029
	AVG	−3.606	−3.557	−1.774	−0.547	0.399
	CSL	−4.790	−1.994	2.128	8.005	20.436
	DOS	−6.230	−6.105	−2.839	0.803	3.619
	DLSA	−6.142	−6.009	−6.170	−6.136	−6.156
	OSUP	−6.219	−6.144	−6.183	−6.153	−6.161
	GLO	−6.222	−6.132	−6.176	−6.164	−6.145
IV	PCEE	−6.171	−6.187	−6.083	−6.194	−6.065
	PCESE	−6.110	−6.065	−5.995	−6.061	−5.962
	AVG	−1.439	9.899	5.355	5.288	5.264
	CSL	−0.149	26.927	58.653	46.244	42.569
	DOS	−2.617	21.516	17.993	17.643	17.531
	DLSA	−5.800	12.159	−2.712	−2.574	−2.422
	OSUP	−6.171	−6.187	−6.088	−6.204	−6.062
	GLO	−6.164	−6.176	−6.080	−6.192	−6.061

Note: The total sample size N is fixed at $N = 2 \times 10^4$. In all cases, each cell in the table corresponds to the average of 500 replications.

Table 3
MSEs of multiple iterative estimators.

Estimator		Multi-step			
		1-step	2-step	3-step	4-step
GLO	0.0240				
PCEE		0.0436	0.0240	0.0240	0.0240
PCESE		0.1002	0.0439	0.0347	0.0276

Table 4
MSEs of different estimators with varying pilot sample sizes.

Estimator		Pilot percentage			
		1%	5%	10%	20%
GLO	0.0031				
PCEE		0.0181	0.0032	0.0031	0.0031
PCESE		0.0731	0.0039	0.0034	0.0032

the proposed methods with FedAvg using two metrics: (a) MSE, (b) the number of communication rounds. The total sample size is fixed at $N = 5 \times 10^4$, while the number of local machines K varies from 5 to 50. Table 5 presents the results for all methods. The proposed methods achieve the best estimation accuracy with two communication rounds. In contrast, FedAvg employs local stochastic gradient descent (SGD) with centralized model averaging. As a first-order method, it converges more slowly than our second-order approaches, requiring more communication rounds.

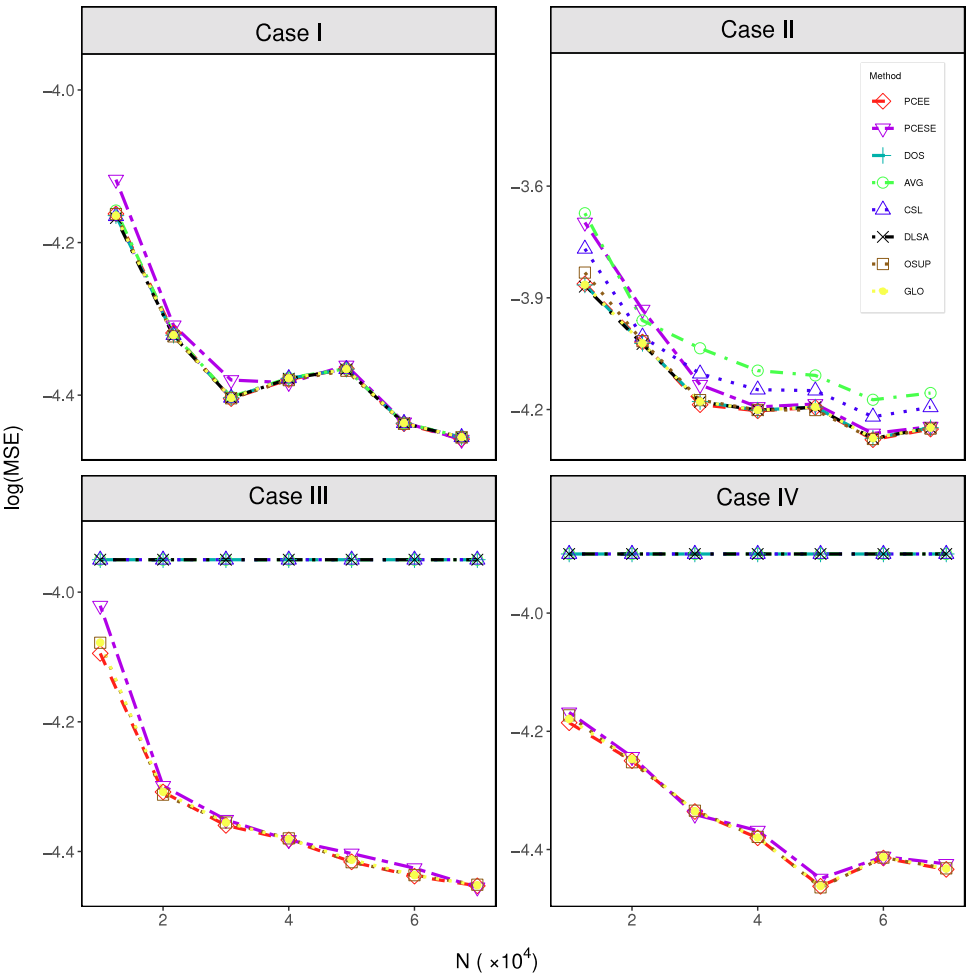

Fig. 3. The logarithm of the MSE for all estimators in mis-specified models varies with the full sample size N, with the number of worker K is fixed at $K = 10$. In all cases, each point represents the average of 500 replications. n all cases, each point corresponds to the average of 500 replications. In Case I, the whole data is split randomly. The Case II, the data is heterogeneous. The data allocation mechanism in Case III depends solely on predictors X, while in Case IV, it depends on both predictors X and the response Y.

Table 5
The results of comparison between different methods.

Estimator		K				Communication
		5	10	20	50	round
GLO	0.0006					0
FedAvg		0.0012	0.0019	0.0023	0.0026	100
PCEE		0.0006	0.0006	0.0006	0.0006	2
PCESE		0.0006	0.0006	0.0006	0.0006	2

5. Real data examples

5.1. U.S. Airline dataset

We apply the PCEE and PCESE methods to the U.S. Airline Dataset (http://stat-computing.org/dataexpo/2009) to check their practical performances. Detailed flight information about American Airlines in 2008 is used. It is a large-scale dataset with 238,9217 observations. The response variable "Delayed" is a binary indicator, which is defined as a flight being fifteen minutes or more later than its scheduled departure time. Following Wang et al. (2021), we take four covariates as the regressors, including the distance of the flight, departure time, departure day of the week, and departure month. The complete variable details are described in Table 6.

Table 6
Description of the response variable and covariates utilized in the U.S. Airlines dataset.

	Variable	Description
Response	Delayed	Dummy variable used to indicate whether the flight is delayed or not.
Predictors	Distance	Distance between airport and destination airport
	Departure time	Categorical variables with 4 levels
	Departure day of week	Categorical variables with 7 levels
	Departure month	Categorical variables with 4 levels

Table 7
Classification error rates of the different estimators.

K	2	50	100
PCEE	0.252	0.257	0.258
PCESE	0.252	0.295	0.327
AVG	0.252	0.519	0.523
CSL	0.252	–	–
DOS	0.280	–	–
DLSA	0.303	0.322	0.526
OSUP	0.252	0.252	0.252
GLO	0.252	0.252	0.252

Note: The notation "–" indicates the method fails to compute the outcome.

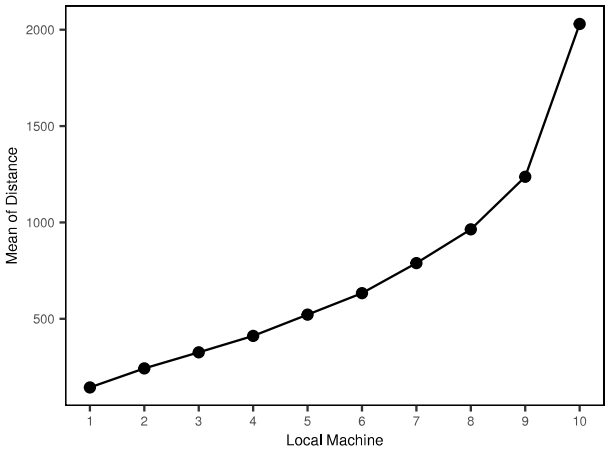

Fig. 4. The mean of distance across local data.

The goal is to predict whether a flight is delayed or not using a logistic regression model. To accomplish this, we start by removing observations associated with canceled flights, reducing the dataset's sample size to 2,319,121. The total sample size of the training set and test set are respectively 200,0000 and 319,121. We train the model on the training set and evaluate the classification error on the test set. We split the training data set into K subsets according to the same way as in the Case IV in Section 4, where the K ranges from 2 to 100. To investigate the heterogeneity across different local data files, we use the covariate "Distance" as an example and illustrate the trends of sample means in each data file in Fig. 4 It is evident that the mean values of the covariate "Distance" vary significantly across data files, indicating the non-randomness of data distribution among the files.

Next, we apply our methods for logistic regression model to the non-randomly distributed datasets. The classification error of the proposed methods and the other competing 6 estimators is compared. Table 7 exhibits the classification error of different methods. The pilot bootstrap sample proportion is set as 0.1%. When K is extremely small ($K = 2$), all methods behave similarly. As K increases, the CSL and DOS methods developed for randomly distributed data fail, while the classification error rates of the our estimators are lower than those of the AVG and DLSA estimators and close to those of the OSUP and GLO estimators. However, our methods adopt the bootstrap pilot sample, avoiding the transfer of raw data in the OSUP estimator.

5.2. Census Income Dataset

In this section, we demonstrate the application of our proposed methods using the Census Income Dataset (Kohavi et al., 1996), a widely recognized benchmark dataset extracted from the 1994 U.S. Census database. The dataset contains 48,842 observations with a binary response variable indicating whether an individual's annual income exceeds $50,000. We define individuals exceeding this

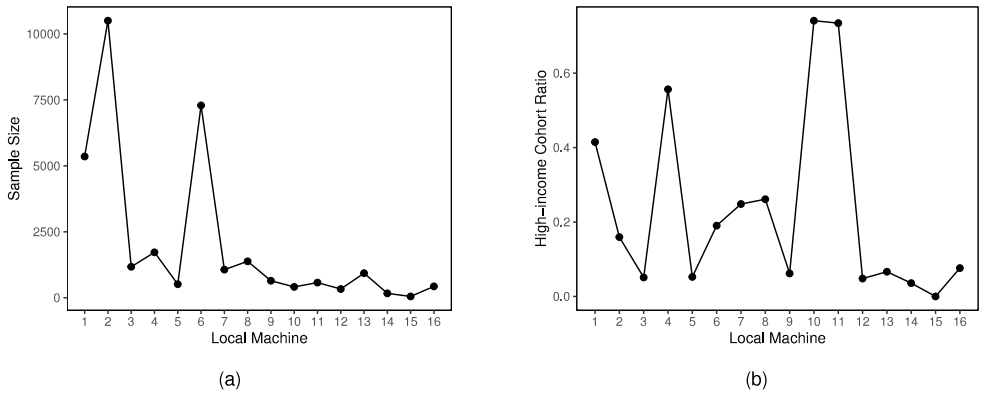

Fig. 5. The sample sizes (a) and high-income cohort proportions (b) across data partitions. Significant disparities in both sample sizes and high-income cohort proportions emphasize the dataset's heterogeneous nature.

Table 8
Estimation results for the logistic regression under the GLO, PCEE, and PCESE methods. In each method, we report the estimated coefficients (Estimate), standard error (SE), and p-values for all the variables.

	GLO			PCEE			PCESE		
	Estimate	SE	p-Value	Estimate	SE	p-Value	Estimate	SE	p-Value
Intercept	−1.3271	0.0150	0.0000	−1.3100	0.0148	0.0000	−1.3105	0.0148	0.0000
θ_1	0.5851	0.0145	0.0000	0.5723	0.0144	0.0000	0.5690	0.0144	0.0000
θ_2	0.0414	0.0140	0.0030	0.0396	0.0139	0.0044	0.0355	0.0139	0.0107
θ_3	0.2612	0.0122	0.0000	0.2569	0.0121	0.0000	0.2640	0.0121	0.0000
θ_4	0.5837	0.0151	0.0000	0.5642	0.0149	0.0000	0.5554	0.0149	0.0000

threshold as the high-income cohort. We select for variables as the covarites: Age (x_1), Fnlwgt (x_2), LosCap (x_3) and hours worked per week (x_4). The variable Fnlwgt is the number of people the observation represents and LosCap is capital loss quantifies financial losses from investment activities.

The dataset, obtained from the UCI Machine Learning Repository (Lichman, 2013), has been pre-partitioned into a training set (32,561 observations) and a validation set (16,281 observations). We segmented the entire training set into $K = 16$ distinct subsets according to educational attainment levels. Fig. 5 illustrates the characteristics of the dataset, revealing substantial heterogeneity in both sample sizes and proportions of high-income cohorts between data partitions.

Next, we construct a logistic regression model to identify the variables that help predict individuals' income status. Table 8 gives the estimation results of the GLO, PCEE, and PCESE estimators, including the estimated coefficients, standard errors (SE), and the p-values. All methods exhibit similar regression result. Notably, the PCESE estimator exhibits reduced statistical power for detecting the Final Weight effect. This is because larger samples typically improve detection power for non-zero effects.

Table 9 presents a comprehensive comparison of classification errors and computation times across methods. The empirical findings show that our proposed estimators, the PCEE and PCESE, attain classification accuracy levels that are comparable to those achieved by the GLO and the OSUP estimators. In terms of computational efficiency, the PCEE and PCESE exhibit a substantial reduction in computation time compared to the GLO estimator. This efficiency improvement is primarily attributed to the distributed inference framework employed by our proposed methods, which allows for parallel processing and reduces the computational burden on any single node. However, when compared to the OSUP estimator, the PCEE and PCESE exhibit a moderate increase in computation time. This increase is largely due to the additional computational overhead associated with generating bootstrap pilot samples on the central server, a necessary step in our methodologies to ensure robustness and accuracy. It is important to note, however, that the OSUP approach necessitates the transmission of the Hessian matrix and partial real data from each local node to the master node. This transmission not only introduces additional communication overhead but may also raise privacy-preserving concerns, as sensitive data or model parameters could potentially be exposed during the transfer. In summary, while the PCEE and PCESE estimators offer a balance between classification accuracy and computational efficiency, their performance relative to OSUP highlights the trade-offs involved in choosing between different methodologies, particularly with respect to privacy and computational overhead.

Furthermore, while the DLSA estimator performs adequately in the specific context of this study, it, along with the CSL and the DOS estimators, exhibits inconsistent performance across a wide range of non-random data distribution scenarios, as detailed in Section 5.1. This inconsistency underscores the importance of carefully considering the underlying data distribution characteristics when selecting an appropriate estimator for a given application.

Table 9
Classification error and computation time (in seconds) of various estimators.

Estimator	GLO	OSUP	CSL	DOS	DLSA	PCEE	PCESE
Error	0.2381	0.2381	0.7634	0.7638	0.2343	0.2384	0.2379
Time	0.4420	0.0178	0.0179	0.0200	0.0110	0.0613	0.0609

6. Conclusion

In this study, we address the statistical estimation problem for non-randomly distributed data. To tackle the challenges posed by non-randomness, we propose two distributed estimators: the PCEE and the PCESE. These methods are designed to accommodate the nature of non-randomly distributed data, requiring only two rounds of communication between the workers and the Master. In the first round, each worker computes its local MLE and broadcasts it to the Master. The Master then generates a pilot bootstrap sample using a parametric bootstrap procedure and computes the KL estimator. In the second round, we perform one-step update on the KL estimator using the derivative information collected from all workers. We adopt two update strategies-Newton–Raphson one-step update and a surrogate likelihood function- and obtain two distinct estimators, PCEE and PCESE, respectively. Theoretical and simulation results confirm that our proposed estimators share the same asymptotic distribution as the global estimator, underscoring their validity and effectiveness.

Our investigation has also revealed several promising research directions. First, we assume that the local density functions $p_k\left(Y|X, W_k; \beta_k\right)$ is known with the parameter vector β_k being unknown. If the density is unkown, it can be estimated using non-parametric methods such as kernel density estimation to mitigate model mis-specification risks. Second, while KL divergence was employed in this work, alternative distance metrics such as the Wasserstein distance may offer advantages in more complex models, such as linear mixed-effect models (Srivastava and Xu, 2021). Thirdly, our methods do not fully address network quality issues, such as delays, packet loss, and bandwidth limitations, which are critical in real-world deployments with potentially unstable distributed networks. Lastly, despite reducing the risk of individual information leakage, our methods remain vulnerable to reconstruction attacks, highlighting the need for integrating rigorous privacy-preserving techniques, such as differential privacy, into distributed statistical inference frameworks. We plan to explore these avenues in our future research endeavors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research is supported by National Natural Science Foundation of China (No:12271034) and the Open Fund Project of Key Laboratory of Market Regulation, China (No:2023SYSKF02003). The authors are grateful to the associate editor and two referees for their highly valuable suggestions which greatly improved the quality of this article.

Appendix A. Proof of main results

Let $\delta_\rho = \min\{\rho, \frac{\rho\lambda_-}{4M}\}$. We first define some "good events":

$$\mathcal{E}_0 := \left\{ \|\widehat{\theta} - \theta^*\|_2 \le \min\left\{ \frac{\rho\lambda_-}{8M}, \frac{(1-\rho)\lambda_- \delta_\rho}{8\lambda_+} \right\} \right\}, \quad \mathcal{E}_1 = \left\{ \frac{1}{n} \sum_{i \in S_k} M(Z_i) \le 2M \right\},$$

$$\mathcal{E}_2 = \left\{ \|\nabla^2 L_{\mathcal{M}}(\theta^*) - I(\theta^*)\|_2 \le \frac{\rho\lambda_-}{2} \right\}, \quad \mathcal{E}_3 = \left\{ \|\nabla^2 L(\theta^*) - I(\theta^*)\|_2 \le \frac{\rho\lambda_-}{2} \right\},$$

$$\mathcal{E}_4 = \left\{ \|\nabla L_{\mathcal{M}}(\theta^*)\|_2 \le \frac{(1-\rho)\lambda_- \delta_\rho}{2} \right\}.$$

Lemma 1. *Let $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$. Under the Assumptions 1–5, it has*

$$\mathbb{P}(\mathcal{E}^c) \le \left(c_1 + c_2 (\log 2d)^8 H^8 + c_3 G^8\right) \frac{K}{n_0^4},$$

where c_1, c_2, c_3 are constants independent of (N, K, n_0).

Then, we establish the upper bound of the error $\hat{\theta}_{PCESE} - \widehat{\theta}$.

Lemma 2. *Under event $\mathcal{E}$ and Assumptions 1–5, it has*

$$\|\hat{\theta}_{PCESE} - \widehat{\theta}\|_2 \le \frac{2\|\nabla \widetilde{L}(\widehat{\theta})\|_2}{(1-\rho)\lambda_-}. \tag{9}$$

Lemma 3. *Under Assumptions 1–5, it has $\hat{\theta}_{KL} - \theta^* = O_P(n_0^{-1/2})$.*

A.1. Proof of Theorem 1

Proof.

By the definition of the PCEE estimator, one has

$$\hat{\theta}_{PCEE} - \theta^* = \hat{\theta}_{KL} - \theta^* - \left(\nabla^2 L(\hat{\theta}_{KL})\right)^{-1} \nabla L(\hat{\theta}_{KL}),$$

the standardized estimator $\nabla^2 L(\hat{\theta}_{KL})(\hat{\theta}_{PCEE} - \theta^*)$ is equal to

$$\nabla^2 L(\hat{\theta}_{KL})(\hat{\theta}_{KL} - \theta^*) - \left(\nabla L(\hat{\theta}_{KL}) - \nabla L(\theta^*)\right) - \nabla L(\theta^*).$$

By Taylor expansion, the second term can be replaced by

$$\int_0^1 \nabla^2 L((1-\kappa_1)\theta^* + \kappa_1 \hat{\theta}_{KL}) d\kappa_1 \, (\hat{\theta}_{KL} - \theta^*).$$

Thus, it holds that

$$\begin{aligned}
&\nabla^2 L(\hat{\theta}_{KL})(\hat{\theta}_{KL} - \theta^*) - \left(\nabla L(\hat{\theta}_{KL}) - \nabla L(\theta^*)\right) - \nabla L(\theta^*) \\
&= \left(\nabla^2 L(\hat{\theta}_{KL}) - \int_0^1 \nabla^2 L((1-\kappa_1)\theta^* + \kappa_1 \hat{\theta}_{KL}) d\kappa_1 \right) \left(\hat{\theta}_{KL} - \theta^*\right) - \nabla L(\theta^*).
\end{aligned}$$

By Assumption 3, one has

$$\left\| \nabla^2 L(\hat{\theta}_{KL}) - \int_0^1 \nabla^2 L((1-\kappa_1)\theta^* + \kappa_1 \hat{\theta}_{KL}) d\kappa_1 \right\|_2 \leq M \|\hat{\theta}_{KL} - \theta^*\|_2.$$

Let $u_n = \left(\nabla^2 L(\hat{\theta}_{KL})\right)^{-1} \left(\nabla^2 L(\hat{\theta}_{KL}) - \int_0^1 \nabla^2 L((1-\kappa_1)\theta^* + \kappa_1 \hat{\theta}_{KL}) d\kappa_1 \right) (\hat{\theta}_{KL} - \theta^*)$. By Lemma B.3 and the Continuous mapping Theorem, we have

$$\|u_n\|_2 = O_P(\|\hat{\theta}_{KL} - \theta^*\|_2^2) = O_P(n_0^{-1}).$$

Hence, the first part of Theorem 1 is proven. Then, it holds that

$$\begin{aligned}
\sqrt{N}(\hat{\theta}_{PCEE} - \theta^*) &= -\nabla^2 L(\hat{\theta}_{KL})^{-1} \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{i \in S_k} \nabla l(Z_{ki}, \theta^*) + O_P(\sqrt{N} * u_n) \\
&= -\nabla^2 L(\hat{\theta}_{KL})^{-1} \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{i \in S_k} \nabla l(Z_{ki}, \theta^*) + O_P\left(\frac{\sqrt{N}}{n_0}\right).
\end{aligned}$$

Applying the Law of Large Numbers and the Continuous mapping Theorem, one has

$$\nabla^2 L(\hat{\theta}_{KL}) \xrightarrow{P} I(\theta^*).$$

By the Central Limit Theorem, we obtain

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{i \in S_k} \nabla l(Z_i, \theta^*) \xrightarrow{P} \mathcal{N}(0, \ \mathbb{E}[\nabla l(Z; \theta^*) \nabla l(Z; \theta^*)^T]).$$

By the condition that $\frac{\sqrt{n_0}}{N} \longrightarrow \infty$, combining the above results yields the claimed asymptotic distribution of $\sqrt{N}(\hat{\theta}_{PCEE} - \theta^*)$.

A.2. Proof of Theorem 2

To prove Theorem 2, it suffices to bound the $\|\nabla \widetilde{L}(\widehat{\theta})\|_2$. A simple algebraic operation yields that

$$\nabla \widetilde{L}(\widehat{\theta}) = \nabla L_{\mathcal{M}}(\widehat{\theta}) - \nabla L_{\mathcal{M}}(\hat{\theta}_{KL}) - (\nabla L(\widehat{\theta}) - \nabla L(\hat{\theta}_{KL})), \tag{10}$$

where we use the fact that $\nabla L(\widehat{\theta}) = 0$. Denote $H_{\mathcal{M}} = \int_0^1 \nabla^2 L_{\mathcal{M}}(\hat{\theta}_{KL} + \kappa_2(\widehat{\theta} - \hat{\theta}_{KL}))\, d\kappa_2$, $H_N = \int_0^1 \nabla^2 L(\hat{\theta}_{KL} + \kappa_3(\widehat{\theta} - \hat{\theta}_{KL}))\, d\kappa_3$. By Taylor expansion, one has

$$\begin{aligned}
\nabla \widetilde{L}(\widehat{\theta}) &= H_{\mathcal{M}}(\widehat{\theta} - \hat{\theta}_{KL}) - H_N(\widehat{\theta} - \hat{\theta}_{KL}) \\
&= \left(H_{\mathcal{M}} - \nabla^2 L_{\mathcal{M}}(\theta^*) - (H_N - \nabla^2 L(\theta^*)) + \nabla^2 L_{\mathcal{M}}(\theta^*) - \nabla^2 L(\theta^*)\right)(\widehat{\theta} - \hat{\theta}_{KL}).
\end{aligned}$$

By the Cauchy–Schwarz inequality, it holds that

$$\begin{aligned}
\|\nabla \widetilde{L}(\widehat{\theta})\|_2 &\leq \|H_1 - \nabla^2 L_{\mathcal{M}}(\theta^*)\|_2 \|\widehat{\theta} - \hat{\theta}_{KL}\|_2 + \|H_N - \nabla^2 L(\theta^*)\|_2 \|\widehat{\theta} - \hat{\theta}_{KL}\|_2 \\
&\quad + \|\nabla^2 L_{\mathcal{M}}(\theta^*) - \nabla^2 L(\theta^*)\|_2 \|\widehat{\theta} - \hat{\theta}_{KL}\|_2 \\
&\leq \left(2M\|\widehat{\theta} - \hat{\theta}_{KL}\|_2 + 2M\|\widehat{\theta} - \theta^*\|_2 + \|\nabla^2 L_{\mathcal{M}}(\theta^*) - \nabla^2 L(\theta^*)\|_2\right) \\
&\quad \cdot \|\widehat{\theta} - \hat{\theta}_{KL}\|_2.
\end{aligned}$$

The Theorem 2 follows by combining the preceding results.

A.3. Proof of Theorem 3

Proof. Recall that $\widehat{\theta}$ is the minimizer of $L(\theta)$, one has

$$0 = \nabla L(\widehat{\theta}) = \nabla L(\theta^*) + H'_N(\widehat{\theta} - \theta^*),$$

where $H'_N = \int_0^1 \nabla^2 L(\widehat{\theta} + \kappa_4(\widehat{\theta} - \theta^*))\, d\kappa_4$. By linear algebra, it holds that

$$\widehat{\theta} - \theta^* = -I(\theta^*)^{-1}\nabla L(\theta^*) - \left(H'_N - \nabla L(\theta^*)\right)(\widehat{\theta} - \theta^*) - \left(\nabla^2 L(\theta^*) - \nabla L(\theta^*)\right)(\widehat{\theta} - \theta^*).$$

Thus, by Assumption 4 and Lemma B.1, the last two terms are proved to be $O_P(1/N)$.

Furthermore, we decompose the error $\hat{\theta}_{PCESE} - \theta^*$ into two parts:

$$\hat{\theta}_{PCESE} - \theta^* = \hat{\theta}_{PCESE} - \widehat{\theta} + \widehat{\theta} - \theta^*.$$

Combining Theorem 2 and Slutsky's theorem, Theorem 3 can be proved.

Appendix B. Proof of the auxiliary lemmas

We first introduce two important lemmas in Zhang et al. (2012) and a theorem in Han and Liu (2016).

Lemma B.1 (*Zhang et al., 2012*). *Under the events* $\cap_{i=1}^4 \mathcal{E}_i$, *we have*

$$\|\hat{\theta}_1 - \theta^*\|_2 \le \frac{2\|\nabla F_1(\theta^*)\|_2}{(1-\rho)\lambda_-}. \tag{11}$$

Lemma B.2 (*Zhang et al., 2012*). *Suppose Assumptions 1–5 holds, there exist constants* c_1, c_2 *and* c_3 *such that for constant* q,

$$\mathbb{E}\left[\left\|\nabla L_{\mathcal{M}}(\theta^*)\right\|_2^q\right] \le c_1 \frac{G^q}{n_0^{q/2}},$$

$$\mathbb{E}\left[\left\|\nabla^2 L_{\mathcal{M}}(\theta^*) - I(\theta^*)\right\|_2^q\right] \le c_2 \frac{\log^{q/2}(2d) H^q}{n_0^{q/2}},$$

$$\mathbb{E}\left[\left\|\nabla^2 L(\theta^*) - I(\theta^*)\right\|_2^q\right] \le c_3 \frac{\log^{q/2}(2d) H^q}{N^{q/2}}.$$

B.1. Proof of Lemma 1

Proof. Recall that $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$.

Let $C' = \min\left\{\frac{\rho\lambda_-}{8M}, \frac{(1-\rho)\lambda_- \delta_\rho}{8\lambda_+}\right\}$. For $\mathcal{E}_0$, by Lemmas B.1, B.2, and Markov's inequality, we obtain that

$$\mathbb{P}(\|\widehat{\theta} - \theta^*\|_2^8 > C') \le \frac{\mathbb{E}\|\widehat{\theta} - \theta^*\|_2^8}{C'} \le \frac{2\mathbb{E}\|\nabla L(\theta^*)\|_2^8}{C'(1-\rho)\lambda_-} \le \frac{2C_0 G^8}{C'(1-\rho)\lambda_- N^4}.$$

Likewise, we have

$$\mathbb{P}(\mathcal{E}_1^c) \le \frac{\mathbb{E}\left[\left|\frac{1}{n_0}\sum_{i=1}^{n_0} M(z_i) - \mathbb{E}[M(z)]\right|^8\right]}{M^8} \le C_1 \frac{1}{n_0^4},$$

$$\mathbb{P}(\mathcal{E}_2^c) \le \frac{2^8\mathbb{E}\left[\left\|\nabla^2 L_{\mathcal{M}}(\theta^*) - I(\theta^*)\right\|_2^8\right]}{\rho^8\lambda_-^8} \le C_2 \frac{\log^4(2d) H^8}{n_0^4},$$

$$\mathbb{P}(\mathcal{E}_3^c) \le \frac{2^8\mathbb{E}\left[\left\|\nabla^2 L(\theta^*) - I(\theta^*)\right\|_2^8\right]}{\rho^8\lambda_-^8} \le C_3 \frac{\log^4(2d) H^8}{N^4},$$

$$\mathbb{P}(\mathcal{E}_4^c) \le \frac{2^8\mathbb{E}\left[\left\|\nabla L_{\mathcal{M}}(\theta^*)\right\|_2^8\right]}{(1-\rho)^8\lambda_-^8\delta_\rho^8} \le C_4 \frac{G^8}{n_0^4}.$$

Applying the union bound yields that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &= \mathbb{P}(\mathcal{E}_0^c \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c \cup \mathcal{E}_4^c) \le \mathbb{P}(\mathcal{E}_0^c) + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c) + \mathbb{P}(\mathcal{E}_4^c) \\
&\le C_0 \frac{G^8}{N^4} + C_1 \frac{1}{n_0^4} + C_2 \frac{\log^4(2d) H^8}{n_0^4} + C_3 \frac{\log^4(2d) H^8}{N^4} + C_4 \frac{G^8}{n_0^4} \\
&\le \left(c_1 + c_2 (\log 2d)^8 H^8 + c_3 G^8\right) \frac{K}{n_0^4}
\end{aligned}$$

where constants $C_i, i = 1, \ldots, 4$, and $c_j, j = 1, 2, 3$, are independent of (K, n_0, N).

B.2. Proof of Lemma 2

Proof. Inequality (9) of Lemma 2 follows directly from the application of Lemma B.1. By substituting $\hat{\theta}_1$, θ^* and F_1 with $\hat{\theta}_{PCESE}$, $\widehat{\theta}$, and $\widetilde{L}$, respectively, one has

$$\|\hat{\theta}_{PCESE} - \widehat{\theta}\|_2 \leq \frac{2\|\nabla \widetilde{L}(\widehat{\theta})\|_2}{(1-\rho)\lambda_-}.$$

Hence, in order to apply their result, it suffices to verify the conditions on the first and the second-order derivatives. Specifically, under the event $\mathcal{E}$, it holds that

$$\|\nabla^2 \widetilde{L}(\widehat{\theta}) - I(\theta^*)\|_2 \leq \frac{\rho\lambda_-}{2}, \quad \text{and} \quad \|\nabla \widetilde{L}(\widehat{\theta})\| \leq \frac{(1-\rho)\lambda_-\delta_\rho}{2}.$$

Notice that the Hessian matrix of $\nabla^2 \widetilde{L}(\theta) = \nabla^2 L_{\mathcal{M}}(\theta)$. Under the events $\mathcal{E}_1$ and $\mathcal{E}_2$, one has

$$\|\nabla^2 L_{\mathcal{M}}(\widehat{\theta}) - I(\theta^*)\|_2 \leq 2M \ \|\widehat{\theta} - \theta^*\|_2 + \|\nabla^2 L_{\mathcal{M}}(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho\mu_-}{4} + \frac{\rho\mu_-}{4} = \frac{\rho\mu_-}{2}.$$

To bound the gradient term, we apply the Taylor expansion to obtain that

$$\nabla L_{\mathcal{M}}(\widehat{\theta}) - \nabla L_{\mathcal{M}}(\theta^*) = H'_{\mathcal{M}}(\widehat{\theta} - \theta^*),$$

where $H'_{\mathcal{M}} = \int_0^1 \nabla^2 L_{\mathcal{M}}(\widehat{\theta} + \kappa_5(\theta^* - \widehat{\theta}))d\kappa_5$. Under the event $\mathcal{E}$, applying the Cauchy–Schwartz inequality and the triangle inequality yields that

$$\begin{aligned}
\|\nabla L_{\mathcal{M}}(\widehat{\theta})\|_2 &\leq \|\nabla L_{\mathcal{M}}(\theta^*)\|_2 + \|H'_{\mathcal{M}} - I(\theta^*)\|_2 \|\widehat{\theta} - \theta^*\|_2 + \|I(\theta^*)\|_2 \|\widehat{\theta} - \theta^*\|_2 \\
&\leq \frac{(1-\rho)\lambda_-\delta_\rho}{4} + 2M\|\widehat{\theta} - \theta^*\|_2^2 + \lambda_+\|\widehat{\theta} - \theta^*\|_2 \\
&\leq \frac{(1-\rho)\lambda_-\delta_\rho}{2}.
\end{aligned}$$

B.3. Proof of Lemma 3

Proof. By the definition of θ^* and θ^*_{KL}, it can be shown that

$$\begin{aligned}
\theta^* &= \arg\min_{\theta\in\Theta} \mathbb{E}[\ell(Z;\theta)], \\
\theta^*_{KL} &= \arg\min_{\theta\in\Theta} \mathbb{E}[\ell(\tilde{Z};\theta)].
\end{aligned}$$

Next, it can be proved that $\Delta_1 = \mathbb{E}[\ell(Z;\theta)] - \mathbb{E}[\ell(\tilde{Z};\theta)] = o_P(1)$.

$$\begin{aligned}
\Delta_1 &= \mathbb{E}[\ell(Z;\theta)] - \mathbb{E}[\ell(\tilde{Z};\theta)] \\
&= \int p(Y|X;\theta)\log p(Y|X;\theta)dY - \int p_0(Y|X)\log p(Y|X;\theta)dY, \\
&= \int \big(p(Y|X;\theta) - p_0(Y|X)\big)\log p(Y|X;\theta)dY,
\end{aligned}$$

It suffices to bound the term $p(Y|X;\theta) - p_0(Y|X)$.

$$\begin{aligned}
\Delta_2 &= p(Y|X;\theta) - p_0(Y|X) \\
&= \sum_{k=1}^{K} \int \big(p_k(Y|X,W_k;\beta_k) - p_k(Y|X,W_k;\hat{\beta}_k)\big) f_k(W_k|X)dW_k \\
&\leq \sum_{k=1}^{K} c_k\|\hat{\beta}_k - \beta_k\|_2.
\end{aligned}$$

The last inequality holds because, under Assumptions 1–5, we have that $|p_k(Y|X,W_k;\beta_k) - p_k(Y|X,W_k;\hat{\beta}_k)| \leq c_k\|\hat{\beta}_k - \beta_k\|_2$ for some constants c_k. By Theorem 5.7 in Van der Vaart (2000), we obtain that $\Delta_2 = O_P(n_k^{-1/2}) = o_P(1)$. Given that K is a fixed integer, we know that $\Delta_1 = o_P(1)$. Furthermore, the Law of Large Numbers guarantees that $\frac{1}{n_0}\sum_{i\in S_{\mathcal{M}}} l(\tilde{Z}_i;\theta) - \mathbb{E}[\ell(\tilde{Z};\theta)] = o_P(1)$. Synthesizing these results yields that

$$L_{\mathcal{M}}(\theta) - \mathbb{E}[\ell(Z;\theta)] = o_P(1).$$

Reapplying Theorem 5.7 in Van der Vaart (2000) ultimately establishes that $\hat{\theta}_{KL} - \theta^* = O_P(n_0^{-1/2})$.

References

Battey, H., Fan, J.Q., Liu, H., Lu, J.W., Zhu, Z.W., 2018. Distributed testing and estimation under sparse high dimensional models. Ann. Stat. 46 (3), 1352–1382.

Cai, T.X., Liu, M.L., Xia, Y., 2022. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. J. Amer. Statist. Assoc. 117 (540), 2105–2119.

Chen, X., Liu, W.D., Zhang, Y.C., 2019. Quantile regression under memory constraint. Ann. Stat. 47 (6), 3244–3273.

Chen, X.Y., Xie, M.G., 2014. A split-and-conquer approach for analysis of extraordinarily large data. Statist. Sinica 1655–1684.

Chen, Y., Zhang, Q., Ma, S., Fang, K., 2024. Heterogeneity-aware clustered distributed learning for multi-source data analysis. J. Mach. Learn. Res. 25 (211), 1–60.

Duan, R., Ning, Y., Chen, Y., 2022. Heterogeneity-aware and communication-efficient distributed statistical inference. Biom. 109 (1), 67–83.

Fan, J., Wang, D., Wang, K., Zhu, Z., 2019. Distributed estimation of principal eigenspaces. Ann. Stat. 47 (6), 3009–3031.

Garber, D., Shamir, O., Srebro, N., 2017. Communication-efficient algorithms for distributed stochastic principal component analysis. In: International Conference on Machine Learning. pp. 1203–1212.

Gu, J., Chen, S.X., 2023. Distributed statistical inference under heterogeneity. J. Mach. Learn. Res. 24 (387), 1–57.

Han, J., Liu, Q., 2016. Bootstrap model aggregation for distributed statistical learning. Adv. Neural Inf. Process. Syst. 29, 1795—-1803.

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

Huang, Y., Zhang, Q., Zhang, S., Huang, J., Ma, S., 2017. Promoting similarity of sparsity structures in integrative analysis with penalization. J. Amer. Statist. Assoc. 112 (517), 342–350.

Jordan, M.I., Lee, J.D., Yang, Y., 2019. Communication-efficient distributed statistical inference. J. Amer. Statist. Assoc. 114 (526), 668–681.

Kohavi, R., et al., 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Kdd. Vol. 96, pp. 202–207.

Lee, J.D., Liu, Q., Sun, Y.K., Taylor, J.E., 2017. Communication-efficient sparse regression. J. Mach. Learn. Res. 18 (1), 115–144.

Lichman, M., 2013. Uci Machine Learning Repository. University of california, school of information and computer science, Irvine, Ca, URL: http://archive.ics.uci.edu/ml.

Lin, N., Xi, R., 2011. Aggregated estimating equation estimation. Stat. Interface 4 (1), 73–83.

Liu, Q., Ihler, A.T., 2014. Distributed estimation, information loss and exponential families. Adv. Neural Inf. Process. Syst. 27, 1098—-1106.

Lv, S.G., Lian, H., 2022. Debiased distributed learning for sparse partial linear models in high dimensions. J. Mach. Learn. Res. 23 (1), 54–85.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR, pp. 1273–1282.

Pan, R., Ren, T., Guo, B.S., Li, F., Li, G.D., Wang, H.S., 2022. A note on distributed quantile regression by pilot sampling and one-step updating. J. Bus. Econom. Statist. 40 (4), 1691–1700.

Srivastava, S., Xu, Y., 2021. Distributed bayesian inference in linear mixed-effects models. J. Comput. Graph. Statist. 30 (3), 594–611.

Tang, L., Song, P.X., 2016. Fused lasso approach in regression coefficients clustering–learning parameter heterogeneity in data integration. J. Mach. Learn. Res. 17 (113), 1–23.

Van der Vaart, A.W., 2000. Asymptotic Statistics. Cambridge University Press, Cambridge.

Volgushev, S., Chao, S.K., Cheng, G., 2019. Distributed inference for quantile regression processes. Ann. Stat. 47 (3), 1634–1662.

Wang, K.N., Zhang, B.L., Sun, X.F., Li, S.M., 2022. Efficient statistical estimation for a non-randomly distributed system with application to large-scale data neural network. Expert Syst. Appl. 197, 116698.

Wang, F.F., Zhu, Y.Q., Huang, D.Y., Qi, H.B., Wang, H.S., 2021. Distributed one-step upgraded estimation for non-uniformly and non-randomly distributed data. Comput. Statist. Data Anal. 162, 107265.

Yang, Y.H., Wang, L., Liu, J.M., Li, R., Lian, H., 2023. Communication-efficient estimation of quantile matrix regression for massive datasets. Comput. Statist. Data Anal. 187, 107812.

Zhang, Y.C., Wainwright, M.J., Duchi, J.C., 2012. Communication-efficient algorithms for statistical optimization. Adv. Neural Inf. Process. Syst. 25, 1502–1510.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y., 2021. A survey on federated learning. Knowl.-Based Syst. 216, 106775.

Zhu, X.N., Li, F., Wang, H.S., 2021. Least-square approximation for a distributed system. J. Comput. Graph. Statist. 30 (4), 1004–1018.