Weighted Distributed Estimation under Heterogeneity

Jia Gu
Center for Statistical Science, Peking University, and
Song Xi Chen *
School of Mathematical Science and Guanghua School of Management
Peking University
September 15, 2022

Abstract

This paper considers distributed M-estimation under heterogeneous distributions among distributed data blocks. A weighted distributed estimator is proposed to improve the efficiency of the standard "Split-And-Conquer" (SaC) estimator for the common parameter shared by all the data blocks. The weighted distributed estimator is shown to be at least as efficient as the would-be full sample and the generalized method of moment estimators with the latter two estimators requiring full data access. A bias reduction is formulated to the WD estimator to accommodate much larger numbers of data blocks than the existing methods without sacrificing the estimation efficiency, and a similar debiased operation is made to the SaC estimator. The mean squared error (MSE) bounds and the asymptotic distributions of the WD and the two debiased estimators are derived, which shows advantageous performance of the debiased estimators when the number of data blocks is large.

*corresponding author
1 Introduction

Modern big data have brought new challenges to statistical inference. One such challenge is that despite the sheer volume of the data, a full communication among the data points may not be possible due to the expensive cost of data communication or the privacy concern. The distributed or the "split-and-conquer" (SaC) method has been proposed to divide the full data sample to smaller size data blocks as data communication is too costly to attain an estimation task in a timely fashion. The SaC estimator is also suited to the situations where the data are naturally divided to data blocks and data communication among the data blocks are prohibited due to privacy concern. The SaC estimation had been considered in [17] for the U-statistics, [30] for the M-estimation, [5] for the generalized linear models, [27] and [4] for the quantile regression, while [2] studied the high dimensional testing and estimation with sparse penalties. [3] studied the estimation efficiency and asymptotic distributions for the general asymptotic symmetric statistics [14] and found differences in the efficiency and the asymptotic distributions between the non-degenerate and degenerate cases.

Bootstrap resampling-based methods had been introduced to facilitate statistical inference. [13] proposed the bag-of-little bootstrap (BLB) method for the plug-in estimators by making up economically the full sample for the distributed inference. [23] suggested a sub-sampled double bootstrap method designed to improve the computational efficiency of the BLB. [3] proposed the distributed and the pseudo-distributed bootstrap methods with the former conducted the resampling within each data block while the latter directly resampled the distributed statistics.

Privacy has been a major concern in big data applications where people are naturally reluctant to share the raw data to form a pool of big data as practised in the traditional
full sample estimation. However, the data holders may like to contribute summary statistics without having to give away the full data information. Federated Learning or the distributed inference with a central host has been proposed to accommodate such reality [12, 16, 19, 28], where summary statistics of the data blocks or the gradients of the objective functions associated with the private data blocks are submitted to a central host for forming aggregated estimation or computation.

Homogeneous distribution among the data blocks are assumed in majority of the distributed inference studies with only a few exceptions [6, 32]. Federated Learning, on the other hand, was introduced to mitigate many challenges arising from classical distributed optimization. In particular, heterogeneous or Non-IID distributed data across different data blocks is one of the defining characteristics and challenges in the Federated Learning [12, 16]. Indeed, it is natural to expect the existence of heterogeneity, especially for data stored in different locations or generated by different stochastic mechanism, for instance mobile phones of different users. However, there has been little published works on the statistical properties of estimators considered in the Federated Learning.

This paper considers distributed estimation under heterogeneous distributions among the data blocks, which is closely related to the Federated Learning and especially the multi-task learning (MTL) [31]. We consider distributed M-estimation where there is a common parameter shared by the distributions of the data blocks and data-block specific heterogeneous parameters. Our treatment of the heterogeneity is made by explicit parameterization, which is different from the MTL where the heterogeneity is regularized by penalty terms. It is noted that [6] considered a heterogeneous setting, but under a fully parametric likelihood framework. Our study reveals that in the presence of the heterogeneity the full sample M-estimator of the common parameter obtained by requiring full data communication, can be less efficient than the SaC estimator. However, this phenomenon disappears if the objective function of the M-estimation satisfies a generalized second-order Bartlett’s identity, which are satisfied by the parametric and quasi likelihoods, and the least square estimation in the parametric regression.
We propose a weighted distributed (WD) estimator, which is asymptotically at least as efficient as the full sample and the SaC estimator when the number of data blocks \( K = o(N^{1/2}) \) where \( N \) is the full sample size. The mean-squared error (MSE) bound and the asymptotic distribution of the proposed WD estimator are derived, as well as the asymptotic equivalence between the WD and the generalized method of moment (GMM) estimator. We propose a debiased weighted distributed (dWD) estimator with a data splitting mechanism on each data block to remove the correlation between the empirical bias correction and the weights used to tackle the heterogeneity. The dWD is asymptotically as efficient as the WD estimator, but with a more relaxed constraint of \( K = o(N^{2/3}) \). The bias-correction is also applied to the SaC formulation leading to a more communication-efficient dSaC estimator, which is shown to be more accurate than the subsampled average mixture estimator (SAVGM) \([30]\) in the homogeneous case.

The paper is organized as follows. The estimation framework and necessary notations for the study are outlined in Section 2. The relative efficiency between the full sample and the SaC estimators under the heterogeneity is discussed in Section 3 to motivate the construction of the weighted distributed (WD) estimator. The WD estimator is introduced in Section 4 along with its efficiency, asymptotic distribution and MSE bound. Statistical properties of two debiased estimators dSaC and dWD are revealed in Section 5. Section 6 provides numerical verification to the theoretical results. Section 7 concludes with a discussion. Technical details are reported in the supplementary materials (SM).

2 Preliminaries

Suppose that there is a large data sample of size \( N \), which is divided into \( K \) data blocks of sizes \( \{n_k\}_{k=1}^K \) such that \( N = \sum_{k=1}^K n_k \) and let \( n = NK^{-1} \) be the average sample size of the data blocks. For the relative sample size among data blocks, we assume the following assumption.

**Assumption 1.** There exist constants \( 0 < c < 1 < C \) such that \( c \leq \frac{n_1}{n_{k_2}} \leq C \) for all pairs
of \((k_1, k_2)\), and if \(K\) is a fixed constant we further assume that \(\frac{n_k}{N} \to \gamma_k \in (0, 1)\) for a set of constants \(\{\gamma_k\}_{k=1}^{K}\).

The \(k\)-th data block consists of a sub-sample \(\{X_{k,i}\}_{i=1}^{n_k}\) which are independent and identically distributed (IID) random vectors from a probability space \((\Omega, \mathcal{F}, P)\) to \((\mathbb{R}^d, \mathcal{R}^d)\) with \(F_k\) as the distribution. The \(K\) distributions \(\{F_k\}\) share a common parameter \(\phi \in \mathbb{R}^{p_1}\), while each \(F_k\) has another parameter \(\lambda_k \in \mathbb{R}^{p_2}\) specific to \(F_k\) of the \(k\)-th data block. There are maybe other hidden parameters which define \(F_k\), which are however not directly involved in the semi-parametric M-estimation, and thus are not of interest in the study.

The parameters of interests in the \(k\)-th block are \(\theta_k = (\phi^T, \lambda_k^T)^T\), and the overall parameters of interests are \(\theta = (\phi^T, \lambda_1^T, \lambda_2^T, \ldots, \lambda_K^T)^T \in \mathbb{R}^{p_1 + Kp_2}\). Suppose there is a common objective function \(M(X; \phi, \lambda_k)\) that is convex with respect to the parameter \((\phi, \lambda_k)\) and facilitates the M-estimation of the parameters in each data block. In general, the criteria function can be made block specific, say \(M_k\) function. Indeed, the presence of the heterogeneous local parameters \(\{\lambda_k\}_{k=1}^{K}\) leads to different \(M_k(x, \phi) = M(x, \phi, \lambda_k)\) for the inference on \(\phi\), which connects to the multi-task learning (MTL).

In the \(k\)-th data block the true parameter \(\theta_k^* = (\phi^*_k, \lambda_k^*_k)^T\) is defined as the unique minimum of the expected objective function, namely

\[
\theta_k^* = (\phi^*_k, \lambda_k^*_k)^T = \arg\min_{\theta_k \in \Theta_k} \mathbb{E}_{F_k} M(X_{k,1}; \phi, \lambda_k). \tag{1}
\]

The true common parameter \(\phi^*\) appears in all \(\theta_k^*\), and the block-specific \(\{\lambda_k^*_k\}_{k=1}^{K}\) may differ from each other. The entire set of true parameters \(\theta^* = (\phi^*_1, \lambda_1^*, \ldots, \lambda_K^*)^T\), can be also identified as

\[
\theta^* = \arg\min_{\theta \in \Theta} \sum_{k=1}^{K} \gamma_k \mathbb{E}_{\theta_k^*} M(X_{k,1}; \phi, \lambda_k). \tag{2}
\]

If the data could be shared across the data blocks, we would attain the conventional
full sample M-estimator

\[ \hat{\theta}_{\text{full}} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{K} \sum_{i=1}^{n_k} M(X_{k,i}; \phi, \lambda_k), \]  

which serves as a benchmark for the distributed estimators. Let \( \psi_\phi(X_{k,i}; \phi, \lambda_k) = \frac{\partial M(X_{k,i}; \phi, \lambda_k)}{\partial \phi} \) and \( \psi_\lambda(X_{k,i}; \phi, \lambda_k) = \frac{\partial M(X_{k,i}; \phi, \lambda_k)}{\partial \lambda_k} \) be the score functions. The estimating equations for the full sample M-estimators are

\[
\begin{cases}
\sum_{k=1}^{K} \sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi, \lambda_k) = 0, \\
\sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi, \lambda_k) = 0 \quad k = 1, \ldots, K.
\end{cases}
\]  

The above full sample estimation is not attainable for the distributed situations due to privacy or the costs associated with the data communications. The distributed estimation first conducts local estimation on each data block, namely the local M-estimator

\[ \hat{\theta}_k = (\hat{\phi}_k, \hat{\lambda}_k) = \arg\min_{\theta_k \in \Theta_k} \sum_{i=1}^{n_k} M(X_{k,i}; \theta_k) \]

with the corresponding estimating equations

\[
\begin{cases}
\sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi_k, \lambda_k) = 0, \\
\sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi_k, \lambda_k) = 0.
\end{cases}
\]

Then, the "split-and-conquer" (SaC) estimator for the common parameter \( \phi \) is

\[ \hat{\phi}^{\text{SaC}} = \frac{1}{N} \sum_{k=1}^{K} n_k \hat{\phi}_k. \]

The heterogeneity among the distributions and the inference models among the data blocks bring new dimensions to the discussion of the relative efficiency and the estimation errors, which are the focus of this paper. We are to show that the conventionally weighted SaC estimator \([6]\) may not be the best formulation for the estimation of \( \phi \). Throughout
this paper, unless otherwise stated, \( \| \cdot \|_2 \) and \( \|\| \|_2 \) represent the \( L_2 \) norm of a vector and a matrix, respectively. Besides, we will use \( C \) and \( C_i \) to denote absolute positive constants independent of \((n_k, K, N)\).

An important question is the efficiency and the estimation errors of the SaC estimator \( \hat{\phi}^{SaC} \) relative to the full sample estimator \( \hat{\phi}_{full} \). For the homogeneous case, Chen and Peng (2021) [3] found that for the asymptotic symmetric statistics, the SaC estimator (6) attains the same efficiency of the full sample estimator in the non-degenerate case, but encounters an efficiency loss in the degenerate case due to a lack of communications among different data blocks. Zhang et al. (2013) [30] derived the mean square error (MSE) bound for the SaC estimator in the homogeneous case and showed that whenever \( K \leq \sqrt{N} \), the SaC estimator achieves the best possible rate of convergence when all \( N \) samples are accessible.

Consider the simultaneous estimating equations of the full sample M-estimation

\[
\Psi_N(X; \theta) = \begin{bmatrix}
\sum_{k=1}^{K} \sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi, \lambda_k) \\
\sum_{i=1}^{n_1} \psi_\lambda(X_{1,i}; \phi, \lambda_1) \\
\vdots \\
\sum_{i=1}^{n_K} \psi_\lambda(X_{K,i}; \phi, \lambda_K)
\end{bmatrix}.
\] (7)

Define

\[
\Psi_\theta(\theta_k) = (\Psi_\phi(\theta_k)^T, \Psi_\lambda(\theta_k)^T)^T = \mathbb{E}\nabla_{\theta_k} M(X_{k,1}; \theta_k),
\]

\[
\Psi_0(\theta_k) = \begin{pmatrix}
\Psi_0^\phi(\theta_k) \\
\Psi_0^\lambda(\theta_k)
\end{pmatrix} = \mathbb{E}\nabla^2_{\theta_k} M(X_{k,1}; \theta_k),
\]

\[
J_{\phi_0}(\theta_k) = \Psi_0^\phi(\theta_k) - \Psi_0^\lambda(\theta_k)^{-1}\Psi_0^\lambda(\theta_k) \quad \text{and}
\]

\[
S_0(X_{k,i}; \theta_k) = \psi_\phi(X_{k,i}; \theta_k) - \Psi_0^\phi(\theta_k)^{-1}\Psi_0^\lambda(\theta_k)^{-1}\psi_\lambda(X_{k,i}; \theta_k).
\]

Then we can apply Taylor’s expansion and obtain (see Section 1.1 in SM for details)

\[
\hat{\phi}_{full} - \phi^* = -\left\{ \sum_{k=1}^{K} \frac{n_k}{N} J_{\phi_0}(\theta_k^*) \right\}^{-1} \frac{1}{N} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} S_0(X_{k,i}; \theta_k^*) \right\} + o_p(N^{-1/2}),
\] (8)
For the local estimator \((\hat{\phi}_k, \hat{\lambda}_k)\) based on the \(k\)-th data block that solves (5), by replicating the same derivation leading to (8), we have

\[
\begin{align*}
\hat{\phi}_k - \phi^* &= -n_k^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1} \sum_{i=1}^{n_k} S_{\phi}(X_{k,i}; \theta_k^*) + o_p(n_k^{-1/2}), \\
\hat{\lambda}_k - \lambda_k^* &= -n_k^{-1}J_{\lambda|\phi}(\theta_k^*)^{-1} \sum_{i=1}^{n_k} S_{\lambda}(X_{k,i}; \theta_k^*) + o_p(n_k^{-1/2}),
\end{align*}
\]

where

\[
J_{\lambda|\phi}(\theta_k) = \Psi_{\lambda}(\theta_k) - \Psi_{\phi}(\theta_k)^{-1} \Psi_{\phi}(\theta_k) \quad \text{and} \quad S_{\lambda}(X_{k,i}; \theta_k) = \psi_{\lambda}(X_{k,i}; \theta_k) - \Psi_{\phi}(\theta_k)^{-1} \psi_{\phi}(X_{k,i}; \theta_k).
\]

The distributed inference setting is closely related to the Multi-Task Learning (MTL) which fits separate local parameters \(\phi_k \in \mathbb{R}^p\) to the data of different data blocks (tasks) through convex loss functions \(\{\ell_k\}\). In particular, the MTL is formulated as [24]:

\[
\min_{\Phi, \Omega} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ell_k(\phi_k^T X_{k,i}, Y_{k,i}) + \mathcal{R}(\Phi, \Omega) \right\},
\]

where \(\{(X_{k,i}, Y_{k,i}), i = 1, 2, \cdots , n_k\}\) are data in the \(k\)-th block, \(\Phi\) is the matrix with \(\{\phi_k\}_{k=1}^{K}\) as column vectors, \(\Omega \in \mathbb{R}^{K \times K}\) and \(\mathcal{R}(\cdot, \cdot)\) measures the extent of the heterogeneity among different data blocks. Choices of \(\mathcal{R}(\cdot, \cdot)\) include \(\mathcal{R}(\Phi, \Omega) = \delta_1 tr(\Phi \Omega \Phi^T) + \delta_2 \|\Phi\|_F^2\) for \(\delta_1, \delta_2 > 0\) and \(\Omega = I_{K \times K} - \frac{1}{K} 1_K 1_K^T\) such that \(tr(\Phi \Omega \Phi^T) = \sum_{k=1}^{K} \|\phi_k - \phi_K\|_2^2\) where \(\phi_K = \frac{1}{K} \sum_{k=1}^{K} \phi_k\), which leads to the mean-regularized MTL [7]. The second term of \(\mathcal{R}\) performs regularization on each local model, trying to control the magnitude of the estimates of \(\phi_k\).

The distributed framework is well connected to the MTL in two key aspects. One is that despite we use the same objective (loss) function \(M\) over the data blocks, the heterogeneity induced by local parameters \(\{\lambda_k\}_{k=1}^{K}\) and the distributions effectively define \(M_k(\phi, x) = M(x, \phi, \lambda_k)\), which is equivalent to the block specific loss functions \(\ell_k\) used in MTL. Another aspect is that although the MTL assumes different parameters \(\{\phi_k\}\) over the data blocks, it regularizes them toward a common one. In contrast, we assume there
is a common parameter $\phi$ shared by the heterogeneous distributions.

3 Full Sample versus SaC Estimation

It is naturally expected that the full sample estimator $\hat{\phi}_{full}$ should be at least as efficient as the distributed SaC estimator $\hat{\phi}^{SaC}$ since the former utilizes the full sample information including the communications among different data blocks. However, we are to show that this is not necessarily true in the presence of heterogeneity. To appreciate this point, we first list more regularity conditions needed in the analysis.

Assumption 2. (Identifiability) The parameters $\theta^*_k = (\phi^*, \lambda^*_k)$ is the unique minimizer of $M_k(\theta_k) = EM(\theta_k)$ for $\theta_k \in \Theta_k$.

Assumption 3. (Compactness) The parameter space $\Theta_k$ is a compact and convex set in $\mathbb{R}^p$ and the true parameter $\theta_k^*$ is an interior point of $\Theta_k$ and $\sup_{\theta_k \in \Theta_k} \|\theta_k - \theta_k^*\|_2 \leq r$ for all $k \geq 1$ and some $r > 0$. The true common parameter $\phi^*$ is an interior point of a compact and convex set $\Phi \subset \Theta_k$.

Assumption 4. (Local strong convexity) The population objective function on the $k$-th data block $M_k(\theta_k) = EM(X_{k,1}; \theta_k)$ is twice differentiable, and there exists a constant $\rho_- > 0$ such that $\nabla^2_{\theta_k} M_k(\theta_k) \succeq \rho_- I_{p \times p}$. Here $A \succeq B$ means $A - B$ is a positive semi-definite matrix.

These three assumptions are standard ones on the parameter space and population objective functions as those in Zhang et al. (2013) and Jordan et al. (2019) for the homogeneous case. In the heterogeneous setting, Duan et al. (2021) only requires the parameter space for the common parameter to be bounded, i.e. $\|\phi - \phi^*\| \leq r$ under a fully parametric setting, while in our assumption, we need the overall parameter space to be bounded. This stronger assumption is needed since we do not fully specify the distributions $\{F_k\}_{k=1}^K$ of the random variables and will be useful when we derive the MSE bound for the weighted distributed estimator which will be proposed in Section 4.
Assumption 5. (Smoothness) There are finite positive constants \( R, L, v \) and \( v_1 \) such that for all \( k \geq 1, \) \( \mathbb{E}\left\| \nabla_{\theta_k} M(X_{k,1}; \theta_k^*) \right\|_2^{2v_1} \leq R^{2v_1} \) and \( \mathbb{E}\left\| \nabla_{\theta_k}^2 M(X_{k,1}; \theta_k^*) - \nabla_{\theta_k}^2 M_k(\theta_k^*) \right\|_2^{2v} \leq L^2v. \) In addition, for any \( x \in \mathbb{R}^d, \) \( \nabla_{\theta_k}^2 M(x; \theta_k) \) and \( \nabla_{\theta_k} M(x; \theta_k)\nabla_{\theta_k} M(x; \theta_k)^T \) are \( G(x) - \) and \( B(x) - \) Lipschitz continuous, respectively, in the sense that

\[
\left\| \nabla_{\theta_k}^2 M(x; \theta_k) - \nabla_{\theta_k}^2 M(x; \theta_k') \right\|_2 \leq G(x)\|\theta_k - \theta_k'\|_2,
\]

\[
\left\| \nabla_{\theta_k} M(x; \theta_k)\nabla_{\theta_k} M(x; \theta_k)^T - \nabla_{\theta_k} M(x; \theta_k')\nabla_{\theta_k} M(x; \theta_k')^T \right\|_2 \leq B(x)\|\theta_k - \theta_k'\|_2,
\]

for all \( \theta_k, \theta_k' \in U_k := \{ \theta_k \| \theta_k - \theta_k' \|_2 \leq \rho \} \) for some \( \rho > 0, \) and \( \mathbb{E}G(X_{k,1})^{2v} \leq G^{2v}, \mathbb{E}B(X_{k,1})^{2v} \leq B^{2v} \) for some positive constants \( \mathbb{E} \) and \( B. \)

The Lipschitz continuity of the outer product of the first-order derivative is required to control the estimation error when we estimate the asymptotic covariance matrix of the local estimator \( \hat{\theta}_k, \) and it can be directly verified under the logistic regression case; see Section 1.2 in the SM for details.

Proposition 1. Under Assumptions [1] - [4] and Assumption [5] with \( v, v_1 \geq 1, \) and if \( K \) is fixed, then \( \hat{\theta}_k \xrightarrow{p} \theta_k^* \) and \( \hat{\theta}_{full} \xrightarrow{p} \theta^*; \) \( \hat{\phi}^{\text{SaC}} = \frac{1}{N} \sum_{k=1}^{K} n_k \hat{\phi}_k \) and \( \hat{\phi}_{full} \) are consistent to \( \phi^*. \)

Theorem 1. Under Assumptions [1] - [4] and Assumption [5] with \( v, v_1 \geq 2, \) if \( K \) is a fixed constant, the SaC estimator \( \hat{\phi}^{\text{SaC}} \) and the full sample estimator \( \hat{\phi}_{full} \) satisfy

\[
\sqrt{N}(\hat{\phi}^{\text{SaC}} - \phi^*) \xrightarrow{d} \mathcal{N}(0, \sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta_k^*)^{-1}\Sigma_k(\theta_k^*) J_{\phi|\lambda}(\theta_k^*)^{-1}), \tag{12a}
\]

\[
\sqrt{N}(\hat{\phi}_{full} - \phi^*) \xrightarrow{d} \mathcal{N}(0, \sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta_k^*)^{-1}(\sum_{k=1}^{K} \gamma_k \Sigma_k(\theta_k^*))(\sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta_k^*))^{-1}), \tag{12b}
\]

where \( J_{\phi|\lambda}(\theta_k^*) = \Psi_{\phi}(\theta_k^*) - \Psi_{\phi}(\theta_k^*) \Psi_{\phi}(\theta_k^*)^{-1} \Psi_{\phi}(\theta_k^*) \) and \( \Sigma_k = \text{Var}\{S_\phi(X_{k,1}; \theta_k^*)\}. \)

Define \( V(\Sigma, A) = (A^T)^{-1}\Sigma A^{-1} \) as a mapping from \( \mathbb{S}^{p_1 \times p_1}_{++} \times \mathbb{G}L(\mathbb{R}^{p_1}) \) to \( \mathbb{S}^{p_1 \times p_1}_{++}, \) where \( \mathbb{S}^{p_1 \times p_1}_{++} \) and \( \mathbb{G}L(\mathbb{R}^{p_1}) \) denote the symmetric positive definite matrices and invertible real matrices of order \( p_1, \) respectively. Since \( \sum_{k=1}^{K} \gamma_k = 1 \) and \( \gamma_k > 0, \) the asymptotic variance of \( \hat{\phi}^{\text{SaC}} \) can be interpreted as a convex combination of function values \( \{V(\Sigma_k(\theta_k^*), J_{\phi|\lambda}(\theta_k^*))\}_{k=1}^{K}. \)
and that of \( \hat{\phi}_{\text{full}} \) can be seen as \( V(\sum_{k=1}^{K} \gamma_k \Sigma_k(\theta^*_k), \sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta^*_k)) \). However, \( V(\cdot, \cdot) \) is not convex with respect to its arguments \((\Sigma, A)\), which means that the inequality

\[
\sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta^*_k)^{-1} \sum_{k=1}^{K} \gamma_k \Sigma_k(\theta^*_k) - \sum_{k=1}^{K} \gamma_k J_{\phi|\lambda}(\theta^*_k)^{-1} \sum_{k=1}^{K} \gamma_k \Sigma_k(\theta^*_k) J_{\phi|\lambda}(\theta^*_k)^{-1}
\]

does not always hold. In other words, \( \hat{\phi}_{\text{full}} \) is not necessarily more efficient than \( \hat{\phi}_{\text{SaC}} \).

To gain understanding of Theorem 1 and to motivate the weighted distributed estimator, we consider the errors-in-variables model. Suppose that one observes \( K \) blocks of independent data samples \( \{(X_{k,i}, Y_{k,i})\}_{i=1}^{n} \) for \( k = 1, 2, ..., K \) and \( N = nK \), where \( (X_{k,i}, Y_{k,i}) \) are IID and generated from the following model:

\[
\begin{align*}
X_k &= Z_k + e_k, \\
Y_k &= \phi^* + \lambda^*_k Z_k + f_k,
\end{align*}
\]

(13)

where \( \{Z_k\}_{k=1}^{K} \) are random variables whose measurements \( \{(X_k, Y_k)\}_{k=1}^{K} \) are subject to errors \( \{(e_k, f_k)\}_{k=1}^{K} \), and \((e, f)\) is bivariate normally distributed with zero mean and covariance matrix \( \sigma^2 I_2 \) and is independent of \( Z_k \). Here, \( \phi^* \) is the common parameter across all data blocks while \( \lambda^*_k (\lambda^*_k > 0) \) represents the block specific parameter. We assume that \( Var(e) = Var(f) \) to avoid any identification issue arisen when \( Z \) is also normally distributed [20]. There is a considerable literature on the regression problem with measurement errors, as summarised in [8, 22].

We consider the approach displayed in Example 5.26 of [26] which constructs a kind of marginal likelihood followed by centering to make a bona fide score equation, as detailed in Section 1.3 of the SM. The M-function is

\[
M(X_k, \theta_k) = \frac{1}{2\sigma^2(1 + \lambda^2_k)} (\lambda_k X_k - (Y_k - \phi))^2,
\]

(14)

with the score equation satisfying \( \mathbb{E} \nabla M(X_{k,1}, Y_{k,1}|Z_{k,1}, \theta^*_k) = 0_{2 \times 1} \).
For simplicity we assume $K = 2$, then from Theorem 1 we have

\[
\begin{align*}
\text{Var}(\hat{\phi}_{\text{full}}) & \approx \left\{ \frac{\sigma^2 EZ^2}{\text{var}(Z)} \frac{2}{1+\lambda_1^*} + \frac{\sigma^4 (EZ)^2}{\text{var}^2(Z)} \frac{2}{(1+\lambda_1^*)^2 + (1+\lambda_2^*)^2} \right\} \frac{1}{N}, \\
\text{Var}(\hat{\phi}_{\text{SaC}}) & \approx \left\{ \frac{\sigma^2 EZ^2}{\text{var}(Z)} \frac{1+\lambda_1^*}{2} + \frac{\sigma^4 (EZ)^2}{\text{var}^2(Z)} \frac{1}{2} \right\} \frac{1}{N}. 
\end{align*}
\]  

(15)

Note that the coefficients to $\frac{\sigma^2 EZ^2}{\text{var}(Z)}$ in the first terms of the variances are harmonic and arithmetic means of $\{1 + \lambda_1^*, 1 + \lambda_2^*\}$, respectively. By the mean inequality the coefficient in the first term of $\text{Var}(\hat{\phi}_{\text{SaC}})$ is larger than that in $\text{Var}(\hat{\phi}_{\text{full}})$. The second term of the variances involves $(EZ)^2$ as a multiplicative factor. Thus, if the unobserved $Z$ has zero mean, the full-sample estimator would be at least as good as the SaC estimator in terms of variation when the sample size goes to infinity. However, the story may change when $EZ \neq 0$, because the second term of $\text{Var}(\hat{\phi}_{\text{full}})$ has a factor which is the square of a ratio between the quadratic mean and the arithmetic mean of $(\frac{1}{1+\lambda_1^*}, \frac{1}{1+\lambda_2^*})$. The factor is larger than or equal to 1 if and only if $\lambda_1^* = \lambda_2^*$ namely the homogeneous case. In the heterogeneous case, by adjusting $\frac{\sigma^4 (EZ)^2}{\text{var}^2(Z)} / \frac{\sigma^4 EZ^2}{\text{var}(Z)}$, we can find cases such that $\lambda_1^* \neq \lambda_2^*$ such that the full sample estimator has a larger variance than the SaC estimator. Simulation experiments presented in Section 6 display such cases.

4 Weighted Distributed Estimator

The previous section shows that the full sample estimator $\hat{\phi}_{\text{full}}$ under heterogeneity may be less efficient than the simple averaged $\hat{\phi}_{\text{SaC}}$. This phenomenon suggests that the conventional wisdom in the homogeneous context case may not be applicable to the heterogeneous case. One may also wonder if the simple SaC estimator can be improved under the heterogeneity. Specifically, how to better aggregate the local estimator $\hat{\phi}_k$ for more efficiency estimation to the common parameter $\phi$ is the focus of this section.
4.1 Formulation and Results

Consider a class of estimators formed by linear combinations of the local estimators \( \hat{\phi}_k \):

\[
\{ \hat{\phi}^{SaC}_w \mid \hat{\phi}^{SaC}_w = \sum_{k=1}^{K} W_k \hat{\phi}_k, W_k \in \mathbb{R}^{p_1 \times p_1}, \sum_{k=1}^{K} W_k = I_{p_1} \}. 
\]

We want to minimize the asymptotic variance of \( \hat{\phi}^{SaC}_w \) with respect to \( \{ W_k \}_{k=1}^{K} \). According to a generalization of Theorem 1

\[
\text{AsyVar}(\hat{\phi}^{SaC}_w) = \sum_{k=1}^{K} n_k^{-1} W_k A_k^{-1} \Sigma_k (A_k^T)^{-1} W_k^T, 
\]

where \( A_k = J_{\phi|\lambda}(\theta^*_k) \) and \( \Sigma_k = \text{Var}\{S_\phi(X_{k,i}; \theta^*_k)\} \). It is noted that the asymptotic variance is defined via the asymptotic normality of the M-estimation. For the time being, \( A_k \) and \( \Sigma_k \) are assumed known and denote \( H_k = A_k^{-1} \Sigma_k (A_k^T)^{-1} \). We choose the trace operator as a measure on the size of the asymptotic covariance matrix and this leads to the minimization problem

\[
\text{Minimize} \quad tr \left( \sum_{k=1}^{K} n_k^{-1} W_k H_k W_k^T \right) \quad \text{s.t.} \quad \sum_{k=1}^{K} W_k = I_{p_1}, 
\]

which is a convex optimization problem. It can be solved via the Lagrangian multiplier method which gives \( W_k^* = (\sum_{s=1}^{K} n_s H_s^{-1})^{-1} n_h H_k^{-1} \). If we replace the trace with the Frobenius norm in the objective function (17), the same solution is attained as shown in Section 1.4 of the SM. The SaC estimator under the optimal weights \( W_k^* \) is called the weighted distributed (WD) estimator and denoted as \( \hat{\phi}^{WD} \). By construction, the WD estimator is at least as efficient as the SaC estimator \( \hat{\phi}^{WD} \). To compare the relative efficiency between \( \hat{\phi}_{full} \) and \( \hat{\phi}^{WD} \), we note that

\[
\text{AsyVar}(\hat{\phi}_{full}) = \left\{ (\sum_{k=1}^{K} n_k A_k)^T (\sum_{k=1}^{K} n_k \Sigma_k)^{-1} (\sum_{k=1}^{K} n_k A_k) \right\}^{-1} \quad \text{and} \\
\text{AsyVar}(\hat{\phi}^{WD}) = \left( \sum_{k=1}^{K} n_k A_k^T \Sigma_k^{-1} A_k \right)^{-1}. 
\]

(18)
Define \( F(\Sigma, A) = A^T\Sigma^{-1}A \). If we can show the convexity of \( F \), an application of Jensen’s inequality will establish the relative efficiency of the two estimators. In fact, we have the following lemma.

**Lemma 1.** Suppose \( H \) and \( K \) are positive definite matrices of order \( p \), and \( X \) and \( Y \) are arbitrary \( p \times m \) matrices. Then,

\[
Q = X^T H^{-1}X + Y^T K^{-1}Y - (X + Y)^T (H + K)^{-1} (X + Y) \succeq 0.
\]

The lemma implies that

\[
\text{\begin{align*}
&\sum_{k=1}^{K} n_k A_k^T \sum_{k=1}^{K} n_k \Sigma_k \sum_{k=1}^{K} n_k A_k \\
\leq &\sum_{k=1}^{K} n_k A_k^T \Sigma_k^{-1} A_k,
\end{align*}}
\]

which means that the WD estimator is at least as efficient as the full sample estimator, and can be more efficient than \( \hat{\phi}_{full} \). That is to say, the simultaneous estimating equations (7), which are obtained from the first-order derivative of the the simple summation of local objectives \( \sum_{i=1}^{n_k} M(X_{k,i}; \theta_k) \), are not the best formulation of the M-estimation problem, since the formulation itself does not utilize the heterogeneity existed in the data blocks. In contrast, the WD estimator exploits the potential efficiency gain from the heterogeneity by re-weighting of the local estimators, and this is why the full sample estimator may not be as efficient as the WD estimator.

### 4.2 Likelihood and Quasi-likelihood

The above results lead us to wonder whether we can attain more efficient distributed estimators than the full sample estimator under the heterogeneity if we restrict to a fully parametric setting. When the distribution of \( X_{k,i} \) is fully parametric with density function...
\( f(\cdot; \phi, \lambda_k) \), the Fisher information matrix in the \( k \)-th data block is
\[
I(\theta_k) = I(\phi, \lambda_k) = \begin{pmatrix}
I_{\phi \phi} & I_{\phi \lambda_k} \\
I_{\lambda_k \phi} & I_{\lambda_k \lambda_k}
\end{pmatrix} = -\mathbb{E} \begin{pmatrix}
\frac{\partial^2}{\partial \phi^2} \log f(X_k; \theta_k) & \frac{\partial^2}{\partial \phi \lambda_k} \log f(X_k; \theta_k) \\
\frac{\partial^2}{\partial \lambda_k \phi} \log f(X_k; \theta_k) & \frac{\partial^2}{\partial \lambda_k^2} \log f(X_k; \theta_k)
\end{pmatrix},
\]
and the partial information matrix as
\[
I_{\phi \mid \lambda_k} = I_{\phi \phi} - I_{\phi \lambda_k}^{-1} I_{\lambda_k \phi}.
\]
Now, the objective function for the M-estimation (also the maximum likelihood estimation (MLE)) is
\[
M(X_k, i; \phi, \lambda_k) = -\log f(X_k, i; \theta_k^\ast).
\]
Routine derivations show that
\[
\Sigma_k = \text{Var}\{S(\phi_k, \lambda_k, \theta_k^\ast)\} = I_{\phi \mid \lambda_k}
\]
and
\[
A_k = J_{\phi \mid \lambda_k}(\theta_k^\ast) = I_{\phi \mid \lambda_k}.
\]
Thus,
\[
\text{AsyVar}(\hat{\phi}_{full}) = \text{AsyVar}(\hat{\phi}^{WD}) = \left( \sum_{k=1}^{K} n_k I_{\phi \mid \lambda_k} \right)^{-1}
\]
and
\[
\text{AsyVar}(\hat{\phi}^{SaC}) = \frac{1}{N^2} \sum_{k=1}^{K} n_k I_{\phi \mid \lambda_k}^{-1}.
\]
A direct application of Lemma 1 shows that
\[
\text{AsyVar}(\hat{\phi}_{full}) = \text{AsyVar}(\hat{\phi}^{WD}) \preceq \text{AsyVar}(\hat{\phi}^{SaC}).
\]
Thus, the full sample MLE can automatically adjust for the heterogeneity and is at least as efficient as SaC estimator \( \hat{\phi}^{SaC} \). Besides, the weighted distributed estimators \( \hat{\phi}^{WD} \) can fully recover the efficiency gap of the SaC estimator.

The same relationship among \( \hat{\phi}_{full}, \hat{\phi}^{SaC} \) and \( \hat{\phi}^{WD} \) also holds for the maximum quasi-likelihood estimator (MQLE) with independent observations (see Section 1.5 in the SM for details). If one looks into the asymptotic variances of the MLE and MQLE, it can be found that the underlying reason for (19) is that the two special M-estimation functions satisfy the second order Bartlett’s identity \[1, 18\]:
\[
\mathbb{E} \nabla M(X_k, \theta_k^\ast) \nabla M(X_k, \theta_k^\ast)^T = \mathbb{E} \nabla^2 M(X_k, \theta_k^\ast).
\]
By the variance formula of the asymptotic distribution of the M-estimator and Lemma 1
we readily have that the Bartlett’s identity can be relaxed by inserting a factor \( \gamma \neq 0 \) such that

\[
\mathbb{E} \nabla M(X_k, \theta_k^*) \nabla^T M(X_k, \theta_k^*) = \gamma \mathbb{E} \nabla^2 M(X_k, \theta_k^*). 
\]

(20)

An important example for such a case is the least square estimation for the parametric regression with homoscedastic and non-autocorrelated disturbances (see Section 1.6 in the SM for details). Otherwise the full sample least square estimator may not be efficient and there is an opportunity for the weighted distributed least square estimation. In summary, as long as the objective function \( M(x_k, \theta_k) \) satisfies [20], then \( \hat{\phi}_{full} \) attains the same asymptotic efficiency as \( \hat{\phi}_{WD} \), and \( \hat{\phi}_{SaC} \) is at most as efficient as the former two estimators.

### 4.3 Relative to Generalized Method of Moment Estimation

To provide a benchmark on the efficiency of the WD estimation, we consider the generalized method of moment (GMM) estimator [9]. The GMM estimator possess certain optimal property for semiparametric inference that the weighted distributed estimation can compare with, despite the GMM requires more data sharing than the distributed inference would require.

The score functions of the M-estimation on each data block can be aggregated and combined to form the moment equations

\[
\begin{aligned}
\sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi, \lambda_k) &= 0, \\
\sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi, \lambda_k) &= 0, \quad k = 1, \ldots, K.
\end{aligned}
\]

(21)

There are \( pK \) estimating equations, where the dimension of \( \theta^* \) is \( pK - (K - 1)p_1 \). Thus, the parameter is over-identified which offers potential in efficiency gain for the GMM [9]. The GMM estimation based on the moment restrictions (21) is asymptotically equivalent to solving the following problem:

\[
\hat{\theta}_{GMM} = \arg\min_{\theta_k=(\phi, \lambda_k) \in \Theta_k, 1 \leq k \leq K} \tilde{\psi}_N^T(\theta)W_0\tilde{\psi}_N(\theta),
\]

(22)
where \( W_0 = \text{Var}(\tilde{\psi}_N(\theta^*))^{-1} \) is the optimal weighting matrix \([9, 29]\) and

\[
\tilde{\psi}_N(\theta) = \left( \sum_{i=1}^{n_1} \psi_\phi(X_{1,i}; \theta_1)^T, \sum_{i=1}^{n_1} \psi_\lambda(X_{1,i}; \theta_1)^T, \ldots, \sum_{i=1}^{n_K} \psi_\phi(X_{K,i}; \theta_K)^T, \sum_{i=1}^{n_K} \psi_\lambda(X_{K,i}; \theta_K)^T \right)^T
\]

The asymptotic variance of the GMM estimator \([9]\) is \( \text{AsyVar}(\hat{\theta}_{GMM}) = (G_0^T W_0 G_0)^{-1} \), where

\[
G_0^T = \mathbb{E}\{\frac{\partial \tilde{\psi}_N(\theta^*)}{\partial \theta}\}. \quad A \text{ derivation given in Section 1.7 of the SM shows that}
\]

\[
\text{AsyVar}(\hat{\phi}_{GMM}) = \left\{ \sum_{k=1}^{K} n_k J_{\phi|\lambda} \Sigma_k^{-1} J_{\phi|\lambda} \right\}^{-1}. \quad (23)
\]

Thus, the weighted distributed estimator’s efficiency is the same as that of the GMM estimator. This is very encouraging to the proposed WD estimator as it attains the same efficiency as the GMM without requiring much data sharing among the blocks, which avoids the expenses of the data transmission and preserves the privacy of the data.

### 4.4 Estimation of Weights in one round communication

To formulate the WD estimator, the optimal weights \( W_k^* = (\sum_{s=1}^{K} n_s H_s^{-1})^{-1} n_k H_k^{-1} \) have to be estimated. By the structure of \( W_k^* \), we only need to separately estimate \( H_k \), the leading principal submatrix of order \( p_1 \) of the asymptotic covariance matrix \( \tilde{H}_k \) of \( \hat{\theta}_k \). It is noted that

\[
\tilde{H}_k = (\nabla \Psi_\theta(\theta_k^*))^{-1} \mathbb{E}\{\psi_{\theta_k}(X_{k,1}; \theta_k^*) \psi_{\theta_k}(X_{k,1}; \theta_k^*)^T\} (\nabla \Psi_\theta(\theta_k^*))^{-1} = \begin{pmatrix} H_k & * \\ * & * \end{pmatrix},
\]

where \( \Psi_\theta(\theta_k) = \mathbb{E}\psi_{\theta_k}(X_{k,1}; \theta_k) \). We can construct the sandwich estimator \([25]\) to estimate \( \tilde{H}_k \) and then \( H_k \). The distributive procedure to attain the WD estimator is summarized in the Algorithm 1.

The Step 4 in the algorithm is necessary since there is no guarantee that after weighting the estimator \( \tilde{\phi}^{WD} \) still belongs to the set \( \Phi \) as required in Assumption \([3]\). However the event \( \{\tilde{\phi}^{WD} \in \Phi\} \) should happen with probability approaching one. Hence, the \( \phi^{SAC} I(\tilde{\phi}^{WD} \not\in \Phi) \)
Algorithm 1: Weighted Distributed M-estimator

| Input: \( \{X_{k,i}, k = 1, \ldots, K; i = 1, \ldots, n_k\} \) |
| Output: \( \hat{\phi}^{WD}, \hat{\lambda}_k \) |

1. Obtain the initial estimates \( \hat{\theta}_k = (\hat{\phi}_k, \hat{\lambda}_k) \) based on data block \( k \);
2. Calculate \( \hat{H}_k(\hat{\theta}_k) \) in each block, which is the leading principal sub-matrix of order \( p_1 \) of \((\nabla_{\theta_k} \hat{\Psi}_{\theta_k})^{-1}(n_k^{-1} \sum_{i=1}^{n_k} \hat{\psi}_{\theta_k}(X_{k,i}; \hat{\theta}_k)\hat{\psi}_{\theta_k}(X_{k,i}; \hat{\theta}_k)^T)(\nabla_{\theta_k} \hat{\Psi}_{\theta_k})^{-T} \) where \( \hat{\Psi}_{\theta_k} = n_k^{-1} \sum_{i=1}^{n_k} \hat{\psi}_{\theta_k}(X_{k,i}; \hat{\theta}_k) \);
3. Send \((\hat{\phi}_k, \hat{H}_k(\hat{\theta}_k))^{-1}\) to a central server and construct \( \tilde{\phi}^{WD} := \{\sum_{k=1}^{K} n_k \hat{H}_k(\hat{\theta}_k)^{-1}\}^{-1} \sum_{k=1}^{K} n_k (\hat{H}_k(\hat{\theta}_k))^{-1} \hat{\phi}_k \);
4. \( \hat{\phi}^{WD} := \tilde{\phi}^{WD} I(\tilde{\phi}^{WD} \in \Phi) + \hat{\phi}^{SA_C} I(\tilde{\phi}^{WD} \notin \Phi) \), where \( \hat{\phi}^{SA_C} := N^{-1} \sum_{k=1}^{K} n_k \hat{\phi}_k \).

Under Assumptions 1 - 4, and Assumption 5 with Theorem 2.

By the definition of \( H_k(\theta_k) \), we have that

\[
\|H_k\|_2 \leq \|\Psi_\theta(\theta_k^*)^{-1}\Sigma_{S,k}(\theta_k^* )\Psi_\theta(\theta_k^*)^{-1}\|_2 \leq \|\Psi_\theta(\theta_k^*)^{-1}\|_2^2 \|\Sigma_{S,k}(\theta_k^*)\|_2 \leq \frac{\rho_\sigma}{\rho^2},
\]

which implies \( H_k(\theta_k^*)^{-1} \geq \frac{\rho^2}{\rho_\sigma} I_{p_1 \times p_1} \). On the other hand, the above inequality leads to \( \|\Psi_\theta(\theta_k^*)^{-1}\|_2 \geq \sqrt{\frac{c}{\rho_\sigma}} \), and this indicate a finite upper bound for the norm of the Hessian matrix, just as that assumed in Jordan et al. (2019) and Duan et al. (2021).

Theorem 2. Under Assumptions 1 - 4, and Assumption 5 with \( v, v_1 \geq 2 \), the mean-squared error of the WD estimator \( \hat{\phi}^{WD} \) satisfies

\[
\mathbb{E}\|\hat{\phi}^{WD} - \phi^*\|_2^2 \leq \frac{C_1}{nK} + \frac{C_2}{n^2} + \frac{C_3}{n^3} + \frac{C_4}{n^4} + \frac{C_5 K}{n^v},
\]

where \( v, v_1 \geq 2 \), and Assumption 5 with \( v, v_1 \geq 2 \), the mean-squared error of the WD estimator \( \hat{\phi}^{WD} \) satisfies

We need the following assumption in order to establish the MSE bound and asymptotic properties of the proposed WD estimator.

Assumption 6. (Boundedness) There exists constants \( \rho_\sigma, c > 0 \) such that for \( k \geq 1 \),

\[
\|\Sigma_{S,k}(\theta_k^*)\|_2 \leq \rho_\sigma, H_k \geq c I_{p_1 \times p_1},
\]

where \( \Sigma_{S,k}(\theta_k) = \mathbb{E}\psi_{\theta_k}(X_{k,1}; \theta_k)\psi_{\theta_k}(X_{k,1}; \theta_k)^T \).

Algorithm 1: Weighted Distributed M-estimator
for \( n = NK^{-1} \) and \( \bar{v} = \text{min}\{v, \frac{v_1}{2}\} \).

The \( v \) and \( v_1 \) appeared in Assumption 5 quantify the moments of the first two orders of the derivatives of the \( M \) function and their corresponding Lipschitz functions. When the number of data blocks \( K = O(n^{\text{min}(1, \frac{v-1}{2})}) \), the convergence rate of MSE of \( \hat{\phi}^{WD} \) is \( \mathcal{O}(nK^{-1}) \), which is the same as the standard full sample estimator. However, when there are too many data blocks such that \( K \gg n \), the convergence rate is reduced to \( \mathcal{O}(n^{-2}) \). Furthermore, if the derivatives of the \( M \) function and their corresponding Lipschitz functions are heavy-tailed, say \( \bar{v} < 3 \), the convergence rate is further reduced to \( \mathcal{O}(Kn^{-\bar{v}}) \).

**Theorem 3.** Under Assumptions 1 - 4 and 6, and Assumption 5 with \( v, v_1 \geq 2 \), if \( K = o(n) \),

\[
(\hat{\phi}^{WD} - \phi^*)^T \left\{ \sum_{k=1}^{K} n_k H_k(\theta_k^*)^{-1} \right\} (\hat{\phi}^{WD} - \phi^*) \overset{d}{\to} \chi^2_{p_1}.
\]

Although \( \{H_k(\theta_k^*)\}_{k=1}^{K} \) have bounded spectral norms, \( \sum_{k=1}^{K} \frac{n_k}{N} H_k(\theta_k^*)^{-1} \) may not converge to a fixed matrix in presence of heterogeneity. Thus, we can only obtain the asymptotic normality of the standardized \( \sqrt{N} \left\{ \sum_{k=1}^{K} \frac{n_k}{N} H_k(\theta_k^*)^{-1} \right\}^{1/2} (\hat{\phi}^{WD} - \phi^*) \). This is why Theorem 3 is formulated in a limiting chi-squared distribution form.

The asymptotic normality implies that we can construct confidence regions for \( \phi \) with confidence level 1 - \( \alpha \) as

\[
\{ \phi \mid (\hat{\phi}^{WD} - \phi)^T \left\{ \sum_{k=1}^{K} n_k \hat{H}_k(\hat{\theta}_k)^{-1} \right\} (\hat{\phi}^{WD} - \phi) \leq \chi^2_{p_1,\alpha} \} \tag{24}
\]

after replacing \( \sum_{k=1}^{K} n_k H_k(\theta_k^*)^{-1} \) with its sample counterpart \( \sum_{k=1}^{K} n_k \hat{H}_k(\hat{\theta}_k)^{-1} \), where \( \chi^2_{p_1,\alpha} \) is the upper \( \alpha \) quantile of the \( \chi^2_{p_1} \) distribution. The block-specific parameter \( \lambda_k \) can also be of interest. Then given the WD estimator of the common parameter \( \phi^* \), a question is that whether a more efficient estimator of \( \lambda_k^* \) can be obtained. Specifically, we plug in the WD estimator to each data block and re-estimate \( \lambda_k \). The corresponding updated estimator is denoted as \( \hat{\lambda}_k^{(2)} \). Actually, the answer is that \( \hat{\lambda}_k^{(2)} \) is not necessarily more efficient than \( \hat{\lambda}_k \).

Due to space limit, more discussions on this aspect are available in Section 1.8 in SM.
5 Debiased Estimator for diverging K

It is noted that $K = o(\sqrt{N})$ is required in both Theorems 2 and 3 to validate the $O(N^{-1})$ leading order MSE and limiting chi-squared distribution of the WD estimator. The reason is that the bias of the local estimator $\hat{\theta}_k$ is at order $O_p(n_k^{-1})$, which can accumulate across the data blocks by the weighted averaging. This leads to the bias of $\sqrt{N}(\hat{\phi}^{WD} - \phi^*)$ being at order $O_p(KN^{-1/2})$, which is not necessarily diminishing to zero unless $K = o(\sqrt{N})$. It is worth mentioning that Duan et al. (2021) [6] needed the same $K = o(\sqrt{N})$ order in their MLE framework to obtain the $\sqrt{N}$-convergence since Li et al. (2003) [15] showed that the MLE is asymptotically biased when $K/n \to C \in (0, +\infty)$. This calls for a debias step for the local estimators before aggregation to allow for larger $K$, which is needed especially in the Federated Learning scenario where the number of users (data blocks) can be much larger than the size of local data.

To facilitate the bias correction operation, we have to simplify the notations. Suppose $F(\theta)$ is a $p \times 1$ vector function, $\nabla F(\theta)$ is the usual Jacobian whose $l$-th row contains the partial derivatives of the $l$-th element of $F(\theta)$. Then the matrices of higher derivatives are defined recursively so that the $j$-th element of the $l$-th row of $\nabla^v L(\theta)$ (a $p \times p^v$ matrix) is the $1 \times p$ vector $f^v_{ij}(\theta) = \partial f^v_{ij-1}(\theta)/\partial \theta^T$, where $f^v_{ij}$ is the $l-$th row and $j$-th element of $\nabla^v F(\theta)$. We use $\otimes$ to denote a usual Kronecker product. Using Kronecker product we can express $\nabla^v F(\theta) = \frac{\partial^v F(\theta)}{\partial \theta^1 \otimes \partial \theta^2 \otimes \cdots \otimes \partial \theta^p}$. Besides, let $M_{n,k}(\theta_k) = n_k^{-1} \sum_{i=1}^{n_k} M(X_{k,i}; \theta_k)$,

$$H_{3,k}(\theta_k) = \mathbb{E} \nabla^2 \theta_k \psi_{\theta_k}(X_{k,1}; \theta_k), \quad Q_k(\theta_k) = \{ -\mathbb{E} \nabla \psi_{\theta_k}(X_{k,1}; \theta_k) \}^{-1},$$

$$d_{i,k}(\theta_k) = Q_k(\theta_k) \psi_{\theta_k}(X_{k,i}; \theta_k) \quad \text{and} \quad v_{i,k}(\theta_k) = \nabla \psi_{\theta_k}(X_{k,i}; \theta_k) - \nabla \psi_{\theta_k}(\theta_k).$$

According to [21], the leading order bias of $\hat{\theta}_k$ is

$$\text{Bias}(\hat{\theta}_k) = n_k^{-1} Q_k(\theta_k) \left( \mathbb{E} v_{i,k}(\theta_k) d_{i,k}(\theta_k) + \frac{1}{2} H_{3,k}(\theta_k) \mathbb{E} \{ d_{i,k}(\theta_k) \otimes d_{i,k}(\theta_k) \} \right). \quad (25)$$

Let $B_k(\theta_k) = Q_k(\theta_k) \left( \mathbb{E} v_{i,k}(\theta_k) d_{i,k}(\theta_k) + \frac{1}{2} H_{3,k}(\theta_k) \mathbb{E} \{ d_{i,k}(\theta_k) \otimes d_{i,k}(\theta_k) \} \right)$, whose the first
$p_1$ dimension associated with $\phi$ are denoted as $B^1_k(\theta_k)$. The empirical estimator of $B_k(\theta_k)$ is

$$
\hat{B}_k(\theta_k) = \hat{Q}_k(\theta_k)(n_k^{-1}\sum_{i=1}^{n_k} \hat{v}_{i,k}(\theta_k)\hat{d}_{i,k}(\theta_k) + \frac{1}{2}\hat{H}_{3,k}(\theta_k)n_k^{-1}\sum_{i=1}^{n_k}(\hat{d}_{i,k}(\theta_k) \otimes \hat{d}_{i,k}(\theta_k)))
$$

(26)

where $\hat{H}_{3,k}(\theta_k) = n_k^{-1}\sum_{i=1}^{n_k} \nabla_{\theta_k}^2\psi_{\theta_k}(X_{k,i};\theta_k)$, $\hat{Q}_k(\theta_k) = \{-n_k^{-1}\sum_{i=1}^{n_k} \nabla_{\theta_k}\psi_{\theta_k}(X_{k,i};\theta_k)\}^{-1}$, $\hat{d}_{i,k}(\theta_k) = \hat{Q}_k(\theta_k)\psi_{\theta_k}(X_{k,i};\theta_k)$ and $\hat{v}_{i,k}(\theta_k) = \nabla_{\theta_k}\psi_{\theta_k}(X_{k,i};\theta_k)$. Applying bias correction to each data block, we have the bias-corrected local estimator

$$
\hat{\theta}_{k,bc} := \hat{\theta}_k - n_k^{-1}\hat{B}_k(\hat{\theta}_k)1\xi_{k,bc},
$$

(27)

where $\xi_{k,bc} = \{\hat{\theta}_k - n_k^{-1}\hat{B}_k(\hat{\theta}_k) \in \Theta_k\}$. The indicator function here is to ensure that $\hat{\theta}_{k,bc}$ is within the parameter space.

After the local debiased estimators are obtained, we need to aggregate them with estimated weights. However, a direct aggregation will invalidate the bias correction procedure due to the correlation between the estimated weights and the local debiased estimator as they are constructed with the same dataset. The accumulation of the dependence over a large number of data blocks can make the bias correction fail.

To remove such correlation between the local estimators and the corresponding estimated local weights $\hat{W}_k = \{\sum_{s=1}^{K}\hat{H}_s(\hat{\theta}_s)^{-1}\}^{-1}\hat{H}_k(\hat{\theta}_k)^{-1}$, we first divide each local dataset $\{X_{k,i}\}_{i=1}^{n_k}$ into two equal-sized subsets $D^s_k = \{X_{k,i}^{(s)}\}_{i=1}^{n_k/2}, s = 1, 2$. Then, for $s = 1, 2$ we calculate the local M-estimators $\hat{\theta}_{k,s}$ and obtain $\hat{H}_{k,s}(\hat{\theta}_{k,s})$, which is the leading principal sub-matrix of order $p_1$ of

$$(\nabla_{\theta_k}\hat{\Psi}_{\theta_k})^{-1}\left(\frac{1}{n_k/2}\sum_{i=1}^{n_k/2}\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})^T(\nabla_{\theta_k}\hat{\Psi}_{\theta_k})^{-T}\right),$$

where $\hat{\Psi}_{\theta_k} = \frac{1}{n_k/2}\sum_{i=1}^{n_k/2}\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})$. We then perform the local bias correction to $\{\hat{\theta}_{k,s}\}$ based on data in subset $D^s_k$ to attain the debiased estimators $\{\hat{\theta}_{k,s,bc}\}$. At last, two debiased
weighted distributed estimators of the form

$$\tilde{\phi}_s^{dWD} := \left\{ \sum_{k=1}^{K} n_k \hat{H}_{k,s}(\hat{\theta}_{k,s})^{-1} \right\}^{-1} \sum_{k=1}^{K} n_k (\hat{H}_{k,s}(\hat{\theta}_{k,s}))^{-1} \tilde{\phi}_{k,2-s}^{bc}$$

for \( s = 1, 2 \) are averaged to obtain the final debiased WD (dWD) estimator, whose procedure is summarized in Algorithm 2. That the weight estimation and the debiasing are conducted on different data splits remove the correlation, and realize the gain of bias-correction procedure.

**Algorithm 2: debiased Weighted Distributed (dWD) Estimator**

**Input:** \( \{X_{k,i}, k = 1, \ldots, K; i = 1, \ldots, n_k\} \)

**Output:** \( \hat{\phi}_s^{dWD} \)

1. For each data block, split the data set into two non-overlapping equal-sized subsets and denote those subsets as \( D_k^s = \{X_{k,i}^{(s)}\}_{i=1}^{n_k/2}, s = 1, 2 \);
2. Obtain the initial estimates \( \hat{\theta}_{k,s} = (\hat{\phi}_{k,s}, \hat{\lambda}_{k,s}) \) based on data from \( D_k^s, s = 1, 2 \);
3. Calculate \( \hat{H}_{k,s}(\hat{\theta}_{k,s}) \) in each block \( (s = 1, 2) \), which is the leading principal sub-matrix of order \( p_1 \) of \( (\nabla_{\theta_k} \hat{\Psi}_k^{(s)})^{-1} (2n_k^{-1} \sum_{i=1}^{n_k/2} \psi_{\theta_k}(X_{k,i}^{(s)}; \hat{\theta}_{k,s}) \psi_{\theta_k}(X_{k,i}^{(s)}; \hat{\theta}_{k,s})^T) (\nabla_{\theta_k} \hat{\Psi}_k^{(s)})^{-T} \) where \( \hat{\Psi}_k = 2n_k^{-1} \sum_{i=1}^{n_k/2} \psi_{\theta_k}(X_{k,i}^{(s)}; \hat{\theta}_{k,s}) \);
4. Calculate the bias corrected estimators in each block \( (k = 1, 2, \ldots, K; s = 1, 2) \):
   \[
   \hat{\phi}_{k,s}^{bc} := \hat{\theta}_{k,s} - 2n_k^{-1} \hat{B}_{k,s}(\hat{\theta}_{k,s}) \hat{\varepsilon}_{k,bc,s} \text{ where } \hat{\varepsilon}_{k,bc,s} := \{\hat{\theta}_{k,s} - 2n_k^{-1} \hat{B}_{k,s}(\hat{\theta}_{k,s}) \in \Theta_k\}.
   \]
   Denote the first \( p_1 \) dimensions of \( \hat{\phi}_{k,s}^{bc} \) as \( \tilde{\phi}_{k,s}^{bc} \);
5. Send \( \{\tilde{\phi}_{k,s}^{bc}, \hat{H}_{k,1}(\hat{\theta}_{k,s})^{-1}, s = 1, 2\} \) to a central server and construct
   \[
   \tilde{\phi}_s^{dWD} := \left\{ \sum_{k=1}^{K} n_k \hat{H}_{k,s}(\hat{\theta}_{k,s})^{-1} \right\}^{-1} \sum_{k=1}^{K} n_k (\hat{H}_{k,s}(\hat{\theta}_{k,s}))^{-1} \tilde{\phi}_{k,2-s}^{bc};
   \]
6. \( \hat{\phi}_s^{dWD} := \tilde{\phi}_s^{dWD} I(\tilde{\phi}_s^{dWD} \in \Phi) + K^{-1} \sum_{k=1}^{K} n_k \hat{\phi}_{k,2-s}^{bc} I(\tilde{\phi}_s^{dWD} \notin \Phi) \) for \( s = 1, 2 \);
7. \( \hat{\phi}_s^{dWD} = \frac{1}{2} \sum_{s=1}^{2} \hat{\phi}_s^{dWD} \).

To provide theoretical guarantee on the bias correction, we need an assumption on the third derivative of the M-function \( M \) (see [30]), which strengthens part of Assumption 5.

**Assumption 7. (Strong smoothness)** For each \( x \in \mathbb{R}^p \), the third order derivatives of \( M(x; \theta_k) \) with respect to \( \theta_k \) exist and are \( A(x) - \) Lipschitz continuous, i.e.

\[
\| (\nabla^2_{\theta_k} \psi_{\theta_k}(x; \theta_k) - \nabla^2_{\theta_k} \psi_{\theta_k}(x; \theta'_k))(u \otimes u) \|_2 \leq A(x) \| \theta_k - \theta'_k \|_2 \| u \|_2,
\]
for all $\theta_k, \theta_k' \in U_k$ defined in Assumption 5 and $u \in \mathbb{R}^p$, where $EA(X_{k,i})^{2v} \leq A^v$ for some $v > 0$ and $A < \infty$.

**Theorem 4.** Under Assumptions 1 - 4 and 6 - 7, and Assumption 5 with $v, v_1 \geq 4$, 

$$E\|\hat{\phi}_d^{WD} - \phi^*\|^2_2 \leq \frac{C_1}{nK} + \frac{C_2}{n^2K} + \frac{C_3}{n^3} + \frac{C_4K}{n^p},$$

where $\bar{v} = \min\{v, \frac{v_1}{2}\}$.

The main difference between the upper bounds in Theorem 4 and that of Theorem 2 for the WD estimator is the disappearance of the $O(n^{-2})$ term for the WD estimator, which has been dissolved and absorbed into the $O((n^2K)^{-1})$ and $O(n^{-3})$ terms for the dWD estimator. As shown next, this translates to more relaxed $K = o(n^2)$ as compared with $K = o(n)$ for the WD estimator in Theorem 3.

**Theorem 5.** Under the conditions required by Theorem 4, if $K = o(n^2)$,

$$(\hat{\phi}_d^{WD} - \phi^*)^T \{\sum_{k=1}^K n_k H_k(\hat{\theta}^*_k)^{-1}\} (\hat{\phi}_d^{WD} - \phi^*) \xrightarrow{d} \chi^2_{p_1}.$$ 

Note that the reason why Theorem 5 is formulated in the chi-squared distribution form is the same as that when we formulate Theorem 3 and similar confidence region with confidence level $1 - \alpha$ can be constructed as

$$\{\phi | (\hat{\phi}_d^{WD} - \phi)^T \{\sum_{k=1}^K n_k H_k(\hat{\theta}^*_k)^{-1}\} (\hat{\phi}_d^{WD} - \phi) \leq \chi^2_{p_1, \alpha}\}. \quad (28)$$

The fact that the confidence regions of dWD and WD estimators use the same standardizing matrix $\sum_{k=1}^K n_k H_k(\hat{\theta}^*_k)^{-1}$ reflects that the dWD and WD estimators have the same estimation efficiency. However, the debiased version has more relaxed constraint on $K = O(n^2)$ (which is equivalent to $K = o(N^{2/3})$) than that of the WD estimator at $K = o(n) (K = o(\sqrt{N}))$.

A more communication-efficient estimator of the common parameter can be defined as
the following debiased SaC (dSaC) estimator:

\[
\hat{\phi}^{dSaC} = N^{-1} \sum_{k=1}^{K} n_k (\hat{\phi}_k - n_k^{-1} \hat{D}_k(\hat{\theta}_k) 1_{\epsilon_{k,bc}}),
\]

(29)

which only performs bias correction and may be preferable when the heterogeneity is not large. The asymptotic property of the dSaC estimator is summarized in the following proposition.

**Theorem 6.** Under the conditions required by Theorem 4, if \( K = o(n^2) \),

\[
E\|\hat{\phi}^{dSaC} - \phi^*\|^2 \leq \frac{C_1}{nK} + \frac{C_2}{n^2K} + \frac{C_3}{n^3} \quad \text{and}
\]

\[
N^2(\hat{\phi}^{dSaC} - \phi^*)^T \left\{ \sum_{k=1}^{K} n_k H_k(\hat{\theta}_k^*) \right\}^{-1} (\hat{\phi}^{dSaC} - \phi^*) \overset{d}{\rightarrow} \chi^2_{p_1}.
\]

The corresponding confidence region with confidence level \( 1 - \alpha \) can be constructed as

\[
\{ \phi | N^2(\hat{\phi}^{dSaC} - \phi)^T \left\{ \sum_{k=1}^{K} n_k H_k(\hat{\theta}_k) \right\}^{-1} (\hat{\phi}^{dSaC} - \phi) \leq \chi^2_{p_1,\alpha} \}.
\]

(30)

It is noted that the dSaC and SaC estimators have the same asymptotic distribution. Hence, the confidence regions based on the SaC estimator can be constructed as (30) with \( \hat{\phi}^{dSaC} \) replaced by \( \hat{\phi}^{SaC} \).

To compare with the subsampled average mixture method (SAVGM) estimator proposed in [30] which also performs local bias correction but under the homogeneous setting, we have the following corollary to Theorem 6.

**Corollary 1.** Under the homogeneous case such that \( \{X_{k,i}, k = 1,\ldots,K, i = 1,\ldots,n_i\} \) are IID distributed, and the assumptions required by Theorem 4

\[
E\|\hat{\phi}^{dSaC} - \theta_1^*\|^2 \leq \frac{2E\|\nabla \theta_1 \Psi_1(\theta_1)^{-1} \psi_{\theta_1}(X_{1,1};\theta_1)\|^2}{nK} + \frac{C_1}{n^2K} + \frac{C_2}{n^3},
\]

(31)

where \( \theta_1^* \) is the true parameter for all the \( K \) data blocks.

The SAVGM estimator resamples \( \lfloor rn_k \rfloor \) data points from each data block \( k \) for a \( r \in \)
(0, 1) to obtain a local estimator $\hat{\theta}_{k,r}^{SaC}$ based on the sub-samples. Then, the SAVGM estimator is

$$
\hat{\theta}_{SAVGM} = \frac{\hat{\theta}_k^{SaC} - r\hat{\theta}_{k,r}^{SaC}}{1 - r},
$$

whose MSE bound as given in Theorem 4 of [30] is

$$
E\|\hat{\theta}_{SAVGM} - \theta^*_1\|^2 \leq \frac{2 + 3r}{(1 - r)^2} \frac{E\|\nabla_{\theta_1} \phi_1(\theta^*_1)^{-1}\psi_1(X_{1,1}; \theta^*_1)\|^2}{nK} + \frac{C_1}{n^2K} + \frac{C_2}{n^3}.
$$

Thus, the MSE bound (33) of the SAVGM estimator has an inflated factor $\frac{2 + 3r}{2(1 - r)^2} > 1$ for $r \in (0, 1)$, when compared with that of the dSaC estimator, although it is computationally more efficient than the dSaC and dWD estimators as it only draws one subsample in its resampling. For more comparisons between the dSaC estimator and one-step estimators proposed by Huang and Huo (2019) [10], see Section 1.10 in SM.

### 6 Simulation Results

We report results from simulation experiments designed to verify two sets theoretical findings made in the previous sections. One was to confirm the finding in Section 2 that the full sample estimator $\hat{\phi}_{full}$ is not necessarily more efficient than the SaC estimator $\hat{\phi}^{SaC}$. The other was to evaluate the numerical performance of the newly proposed weighted distributive (WD), debaised SaC (dSaC) and debiased WD (dWD) estimators of the common parameter and compare them with the existing SaC and subsampled average mixture method (SAVGM) (with subsampling rate $r = 0.05$) estimators. Although the SAVGM estimator [30] was proposed under the homogeneous setting, but since its main bias correction is performed locally on each data block $k$ as shown in (32), similar theoretical bounds as formula (33) can be derived without much modifications on the original proof. Throughout the simulation experiments, the results of each simulation setting were based on $B = 500$ number of replications and were conducted in R paralleled with a single 10-core Intel(R) Core(TM) i9-10900K @3.7 GHz processor.

25
In the first simulation experiment, we simulated the errors-in-variables Model (13) with the objective function (14) to compare the performance of the full sample, the SaC and the WD estimators: \( \hat{\phi}_{\text{full}}, \hat{\phi}_{\text{SaC}} \) and \( \hat{\phi}^{WD} \). The simulation was carried out by first generating IID \( \{Z_{i,k}\} \) from \( \mathcal{N}(\mu_Z, \sigma_Z^2) \), and then upon given a \( Z_{i,k} \), \( (X_{k,i}, Y_{i,k})^T \) were independently drawn from \( \mathcal{N}((Z_{i,k}, \phi^* + \lambda_k^*Z_{i,k})^T, \sigma^2I_{2\times 2}) \). We chose \( \phi^* = 1 \), \( K = 2 \), \( \sigma^2 = 1 \) and \( n_1 = n_2 = 5 \times 10^4 = N/2 \), and \( \lambda_1^*, \lambda_2^*, \mu_Z \) and \( \sigma_Z^2 \) were those reported in Table 1 under four scenarios.

As discussed in Section 2, the relative efficiency of \( \hat{\phi}_{\text{full}} \) to \( \hat{\phi}_{\text{SaC}} \) depends on the ratio \( \sigma^2(\mathbb{E}Z)^2/(\text{var}(Z)\mathbb{E}Z^2) \) as shown in (15). We designed four scenarios according to the above ratio under \( \lambda_1^* \neq \lambda_2^* \) and \( \mathbb{E}Z \neq 0 \), respectively, which represented the settings where the full sample estimator \( \hat{\phi}_{\text{full}} \) would be less (Scenario 1) or more (Scenario 2) efficient than the SaC estimator as predicted by the ratio, but not as efficient as the weighted distributed estimator \( \hat{\phi}^{WD} \). Scenario 3 (\( \lambda_1^* = \lambda_2^*, \mathbb{E}Z = 0 \)) was the case when \( \hat{\phi}_{\text{full}} \) and \( \hat{\phi}^{WD} \) would be asymptotically equivalent, and both estimators would be more efficient than \( \hat{\phi}_{\text{SaC}} \). Scenario 4 was the homogeneous case with \( \lambda_1^* = \lambda_2^* \) in which all the three estimators would have the same asymptotic efficiency. For all the four scenarios, the ARE column of the Table 1 confirmed the relative efficiency as predicted by the asymptotic variances in (15), and was well reflected in the comparison of the RMSEs, as the bias is of smaller order as compared with that of the SD and thus negligible.

In the second simulation experiment, we evaluated the numerical performance of the five estimators for the common parameter \( \phi \) under a logistic regression model. For each of \( K \) data block with \( K \in \{10, 50, 100, 250, 500, 1000, 2000\} \), \( \{(X_{k,i}; Y_{k,i})\}_{i=1}^n \subset \mathbb{R}^p \times \{0, 1\} \) were independently sampled from the following model:

\[
X_{k,i} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}_{p \times 1}, 0.75^2I_{p \times p}) \quad \text{and} \quad P(Y_{k,i} = 1|X_{k,i}) = \frac{\exp(X_{k,i}^T\theta_k^*)}{1 + \exp(X_{k,i}^T\theta_k^*)},
\]

where \( \theta_k^* = (\phi^*, \lambda_k^T)^T \), \( \phi^* = 1 \), \( \lambda_k^* = (\lambda_{k,1}, \lambda_{k,2}, \ldots, \lambda_{k,p_2})^T \) and \( \lambda_{k,j} = (-1)^j10(1 - \frac{2(k-j)}{K-1}) \). The sample sizes of the data blocks were equal at \( n = NK^{-1} \) with \( N = 2 \times 10^6 \). Two levels of the dimension \( p_2 = 4 \) and 10 of the nuisance parameter \( \lambda_k \) were considered. A derivation
Table 1: Average root mean squared error (RMSE) and the standard deviation (SD), multiplied by $10^2$, of the full sample estimator $\hat{\phi}_{\text{full}}$, the SaC estimator $\hat{\phi}_{\text{SaC}}$, and the WD estimator $\hat{\phi}_{\text{WD}}$ under four scenarios for the errors-in-variables model (12) for $N = 10^5$, $K = 2$ and $n_1 = n_2$. AREs (asymptotic relative efficiency) of $\hat{\phi}_{\text{full}}$ to $\hat{\phi}_{\text{SaC}}$ are calculated from (15).

| Scenario | $(\lambda_1^*, \lambda_2^*)$ | ARE | RMSE | SD | RMSE | SD | RMSE | SD |
|----------|-------------------------------|-----|------|----|------|----|------|----|
| Scenario 1 ($\mu_Z = 1, \sigma_Z^2 = 0.1$) | (0.25,3.25) | 0.89 | 4.55 | 4.51 | 4.12 | 4.09 | 3.91 | 3.89 |
| | (0.5,3.5) | 0.93 | 4.65 | 4.65 | 4.35 | 4.35 | 4.08 | 4.08 |
| | (0.75,3.75) | 0.97 | 4.52 | 4.52 | 4.40 | 4.38 | 4.13 | 4.13 |
| Scenario 2 ($\mu_Z = 3, \sigma_Z^2 = 0.5$) | (0.25,2.25) | 1.18 | 2.95 | 2.95 | 3.24 | 3.24 | 2.89 | 2.89 |
| | (0.75,2.75) | 1.28 | 3.28 | 3.26 | 3.65 | 3.64 | 3.17 | 3.16 |
| | (1.25,3.25) | 1.31 | 3.71 | 3.71 | 4.16 | 4.07 | 3.64 | 3.61 |
| Scenario 3 ($\mu_Z = 0, \sigma_Z^2 = 0.5$) | (0.25,2.25) | 1.97 | 0.41 | 0.41 | 0.61 | 0.61 | 0.41 | 0.41 |
| | (0.75,2.75) | 1.92 | 0.51 | 0.51 | 0.70 | 0.70 | 0.51 | 0.51 |
| | (1.25,3.25) | 1.68 | 0.64 | 0.64 | 0.82 | 0.82 | 0.64 | 0.64 |
| Scenario 4 ($\mu_Z = 4, \sigma_Z^2 = 0.5$) | (0.5,0.5) | 1 | 3.25 | 3.24 | 3.31 | 3.28 | 3.30 | 3.26 |
| | (1.0,1.0) | 1 | 3.53 | 3.53 | 3.59 | 3.59 | 3.59 | 3.59 |
| | (1.5,1.5) | 1 | 4.06 | 4.03 | 4.08 | 4.07 | 4.06 | 4.06 |

of the bias correction formula for the logistic model is given in Section 1.9 of the SM.

Figure [2] reports the root mean square errors (RMSEs) and absolute bias of the estimators. It is observed that the weighted distributed estimator WD, and the two debiased estimators dSaC and dWD had smaller RMSE than those of the SaC and SAVGM for almost all the simulation settings. Between the SaC and SAVGM, the SAVGM fared better in the lower dimensional case of $p_2 = 4$, but was another way around for $p_2 = 10$. It was evidence that the WD estimator had much smaller RMSEs than the SaC and SAVGM estimators for all the block number $K$, realizing its theoretical promises. In most cases the WD estimator had smaller bias than the SaC estimator although it was not debiased. It also had smaller RMSEs than the debiased SaC estimator dSaC for almost all cases of the block numbers for $p_2 = 4$, while in the higher dimensional $p_2 = 10$ the WD estimator was advantageous for $K \leq 250$. The latter indicated the need for conducting the bias correction to the WD estimator. Both bias corrected dWD and dSaC were very effective in reducing the bias of the WD and SaC estimators, respectively, especially for larger $K$ when
the bias was more severe. The debiased WD attained the smallest RMSEs and the bias in all settings, suggesting the need for conducting both weighting and the bias correction in the distributed inference especially for large $K$. These empirical results were consistent with Theorems 2 and 4, namely the leading RMSE term of the WD estimator changes from $O((Kn)^{-1})$ to $O(n^{-2})$ when $K$ surpasses the local sample size $n$, while the leading RMSEs of the dWD is still $O((nK)^{-1})$ until $K >> n^2$.

We also evaluated the coverage probabilities and widths of the $1 - \alpha$ ($\alpha = 0.01, 0.05, 0.1$) confidence intervals (CIs) of the common parameter based on the asymptotic normality as given after Theorems 3 and 5. The SAVGM estimator was not included as its asymptotic distribution was not made available in [30]. Table 2 reports the empirical coverage and the average width of the CIs. It is observed that for the lower dimensional nuisance parameter case of $p_2 = 4$ the four types of the CIs all had quite adequate coverage levels when $K \leq 100$. However, for $K \geq 250$, the SaC CIs first started to lose coverage, followed by those of the WD, while the CIs of the debiased SaC (dSaC) and debiased weighted distributed (dWD) estimators can hold up to the promised coverage for all cases of $K$. The outstanding performance of the dSaC and dWD CIs was largely replicated for the higher dimensional nuisance parameter case of $p_2 = 10$, while the other two non-debiased estimator based CIs had their coverage quickly slipped below the nominal coverage levels. Although the dSaC CIs had comparable coverages with the dWD CIs, their widths were much wider than those of the dWD. This was largely due to the fact that the weighted averaging conducted in the weighted distributed estimation reduced the variation and hence the width of the CIs. The widths of the WD CIs were largely the same with those of the dWD, and yet the coverage levels of the dWD CIs were much more accurate indicating the importance of the bias correction as it shifted the CIs without inflating the width.

In addition to the simulation experiments on the statistical properties of the estimators, the computation efficiency of the estimators was also evaluated. Table 3 reports the average CPU time per simulation run based on 500 replications of the five estimators for a range of $K$ and dimension $p_2$ of the nuisance parameter for the logistic regression model with the
total sample size \( N = 2 \times 10^6 \).

The computation speed of the dSaC and dWD estimators were relatively slower than those of the SaC, WD and SAVGM estimators. The WD estimator was quite fast, which means that the re-weighting used less computing time than the bias-reduction. In comparison, the dWD estimator was the slowest as a cost for attaining the best RMSE among the five estimators in all settings. It is observed in Table 3 that the overall computation time for each estimator first decreased and then increased as \( K \) became larger. The decrease in time was because the benefit of the distributed computation, while the increase was due to the increase in the number of optimization associated with the M-estimation performed as \( K \) got larger. However, it is worth mentioning that these results did not account for the potential time expenditure in data communication among different data blocks.

7 Discussion

This paper investigates several distributed M-estimators in the presence of heterogeneous distributions among the data blocks. The weighted distributed (WD) estimator is able to improve the estimation efficiency of the "Split-And-Conquer" (SaC) estimator for the common parameter. Two debiased estimators (dWD and dSaC) are proposed to allow for larger numbers of data blocks \( K \). The statistical properties of these three estimators are shown to be advantageous over the SaC and SAVGM estimators. In particular, the WD estimator has good performance for smaller \( K \) relative to \( n \), and the debiased WD estimator that conducted both bias correction and weighting offers good estimation accuracy for large \( K \).

An important issue for the distributed estimation is the size of \( K \) relative to the local average sample size \( n \). This is especially true in Federated Learning setting where the number of clients (data blocks) are usually very large. Both SaC and WD estimators require \( K = o(\sqrt{N}) \) to preserve the \( O(N^{-1}) \) convergence rate for its MSE and the \( \sqrt{N} \) rate for the asymptotic variance. The debiased dWD and dSaC relax the restriction to
$K = o(N^{2/3})$ without compromising the convergence rate. The dSaC may be used as a computationally cheaper version of the dWD at the cost of larger variations and wider confidence regions when compared with dWD.

References

[1] Bartlett, M. (1953). Approximate confidence intervals. *Biometrika*, 40:12–19.

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

[3] Chen, S. X. and Peng, L. (2021). Distributed statistical inference for massive data. *The Annals of Statistics*, 49:2851–2869.

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

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

[6] Duan, R., Ning, Y., and Chen, Y. (2021). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, to appear.

[7] Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117.

[8] Fuller, W. (1987). *Measurement Error Models*. Wiley.

[9] Hansen, L. (1982). Large sample properties generalized method of moments estimators. *Econometrica*, 50:1029–1054.
[10] Huang, C. and Huo, X. (2019). A distributed one-step estimator. *Mathematical Programming*, 174:41–76.

[11] Jordan, M., Lee, J., and Yang, Y. (2019). Communication-efficient distributed statistical learning. *Journal of the American Statistical Association*, 114:668–681.

[12] Kairouz, P., McMahan, H., Avent, B., Bellet, A., Bennis, M., Bhagoji, A., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D’Oliveira, R., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P., and Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14:1–210.

[13] Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. (2011). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 76:795–816.

[14] Lai, T. and Wang, J. (1993). Edgeworth expansions for symmetric statistics with applications to bootstrap methods. *Statistica Sinica*, 3:517–542.

[15] Li, H., Lindsay, B., and Waterman, R. (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society Series B*, 65:191–208.

[16] Li, T., Sahu, A., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60.

[17] Lin, N. and Xi, R. (2010). Fast surrogates of U-statistics. *Computational Statistics & Data Analysis*, 54:16–24.

[18] McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, 11:59–67.

[19] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of Machine Learning Research*, 54:1273–1282.
[20] Reiersol, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18:375–389.

[21] Rilstone, P., Srivastava, V., and Ullah, A. (1996). The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics*, 124:369–395.

[22] Schafer, D. and Purdy, K. (1996). Likelihood analysis for error-in-variables regression with replicate measurements. *Biometrika*, 83:813–824.

[23] Sengupta, S., Volgushev, S., and Shao, X. (2015). A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111:1222–1232.

[24] Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. (2017). Federated Multi-Task Learning. *Advances in Neural Information Processing Systems*.

[25] Stefanski, L. and Boos, D. (2002). The Calculus of M-Estimation. *The American Statistician*, 56:29–38.

[26] van der Vaart, A. (1999). *Asymptotic Statistics*. Cambridge University Press.

[27] Volgushev, S., Chao, S.-K., and Cheng, G. (2017). Distributed inference for quantile regression processes. *Annals of Statistics*, 47.

[28] Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10:1–19.

[29] Yaron, A., Hansen, L., and Heaton, J. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics*, 14:262–80.

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

[31] Zhang, Y. and Yang, Q. (2021). A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*. 
[32] Zhao, T., Cheng, G., and Liu, H. (2014). A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44:1400–1437.

Figure 1: Average simulated bias (a, c) and the root mean square errors (RMSE) (b,d) of the weighted distributed (WD) (red circle), the SaC (blue triangle), the debiased SaC (dSaC) (green square), the debiased WD (dWD) (purple cross), the subsampled average mixture SAVGM (pink square cross) estimators, with respect to the number of data block $K$ for the logistic regression model with the dimension $p_2$ of the nuisance parameter $\lambda_k$ being 4 and 10, respectively with the full sample size $N = 2 \times 10^6$. 
Table 2: Coverage probabilities and widths (in parentheses, multiplied by 100) of the $1 - \alpha$ confidence intervals for the common parameter $\phi$ in the logistic regression model based on the asymptotic normality of the SaC, the WD, the debiased SaC and the debiased WD estimators with respect to the number of data blocks $K$ for two dimensions of the heterogeneous parameter $p_2$ with the full sample size $N = 2 \times 10^6$.

(a) $p_2 = 4$

| K  | 1 - $\alpha$ | SaC      |  | WD      |  | dSaC    |  | dWD     |  |
|----|--------------|----------|---|----------|---|---------|---|---------|---|
|    |              | 0.99     | 0.95| 0.90     |   | 0.99    | 0.95| 0.90    |   |
| 10 |              | 0.99     | 0.96| 0.92     |   | 0.99    | 0.97| 0.91    |   |
|    |              |         |    | (2.45)   |   | (2.03)  |   | (2.45)  |   |
|    |              |         |    | (1.87)   |   | (1.55)  |   | (1.87)  |   |
|    |              |         |    | (1.57)   |   | (1.30)  |   | (1.57)  |   |
| 50 |              | 0.99     | 0.95| 0.91     |   | 0.98    | 0.93| 0.89    |   |
|    |              |         |    | (2.36)   |   | (1.97)  |   | (2.36)  |   |
|    |              |         |    | (1.80)   |   | (1.50)  |   | (1.80)  |   |
|    |              |         |    | (1.51)   |   | (1.26)  |   | (1.51)  |   |
| 100|              | 0.98     | 0.94| 0.91     |   | 0.99    | 0.95| 0.91    |   |
|    |              |         |    | (2.36)   |   | (1.96)  |   | (2.36)  |   |
|    |              |         |    | (1.79)   |   | (1.51)  |   | (1.79)  |   |
|    |              |         |    | (1.50)   |   | (1.25)  |   | (1.50)  |   |
| 250|              | 0.99     | 0.93| 0.85     |   | 0.99    | 0.95| 0.90    |   |
|    |              |         |    | (2.36)   |   | (1.96)  |   | (2.36)  |   |
|    |              |         |    | (1.79)   |   | (1.50)  |   | (1.79)  |   |
|    |              |         |    | (1.50)   |   | (1.25)  |   | (1.50)  |   |
| 500|              | 0.91     | 0.77| 0.66     |   | 0.99    | 0.95| 0.88    |   |
|    |              |         |    | (2.36)   |   | (1.96)  |   | (2.36)  |   |
|    |              |         |    | (1.80)   |   | (1.51)  |   | (1.80)  |   |
|    |              |         |    | (1.51)   |   | (1.25)  |   | (1.51)  |   |
| 1000|             | 0.65     | 0.41| 0.28     |   | 0.99    | 0.94| 0.88    |   |
|     |              |         |    | (2.38)   |   | (1.96)  |   | (2.38)  |   |
|     |              |         |    | (1.81)   |   | (1.52)  |   | (1.81)  |   |
|     |              |         |    | (1.52)   |   | (1.25)  |   | (1.52)  |   |
| 2000|             | 0.01     | 0.01| 0.00     |   | 0.99    | 0.91| 0.81    |   |
|     |              |         |    | (2.42)   |   | (1.96)  |   | (2.42)  |   |
|     |              |         |    | (1.84)   |   | (1.55)  |   | (1.84)  |   |
|     |              |         |    | (1.55)   |   | (1.25)  |   | (1.55)  |   |

(b) $p_2 = 10$

| K  | 1 - $\alpha$ | SaC      |  | WD      |  | dSaC    |  | dWD     |  |
|----|--------------|----------|---|----------|---|---------|---|---------|---|
|    |              | 0.99     | 0.95| 0.90     |   | 0.99    | 0.95| 0.90    |   |
| 10 |              | 0.99     | 0.94| 0.88     |   | 1.00    | 0.96| 0.92    |   |
|    |              |         |    | (3.05)   |   | (2.41)  |   | (3.05)  |   |
|    |              |         |    | (2.32)   |   | (1.84)  |   | (2.32)  |   |
|    |              |         |    | (1.95)   |   | (1.54)  |   | (1.95)  |   |
| 50 |              | 0.99     | 0.93| 0.87     |   | 0.99    | 0.95| 0.88    |   |
|    |              |         |    | (2.94)   |   | (2.29)  |   | (2.94)  |   |
|    |              |         |    | (2.24)   |   | (1.88)  |   | (2.24)  |   |
|    |              |         |    | (1.88)   |   | (1.46)  |   | (1.88)  |   |
| 100|              | 0.97     | 0.89| 0.84     |   | 0.97    | 0.93| 0.87    |   |
|    |              |         |    | (2.93)   |   | (2.28)  |   | (2.93)  |   |
|    |              |         |    | (2.23)   |   | (1.87)  |   | (2.23)  |   |
|    |              |         |    | (1.87)   |   | (1.46)  |   | (1.87)  |   |
| 250|              | 0.89     | 0.72| 0.63     |   | 0.98    | 0.92| 0.87    |   |
|    |              |         |    | (2.94)   |   | (2.28)  |   | (2.94)  |   |
|    |              |         |    | (2.24)   |   | (1.88)  |   | (2.24)  |   |
|    |              |         |    | (1.88)   |   | (1.46)  |   | (1.88)  |   |
| 500|              | 0.51     | 0.28| 0.18     |   | 0.93    | 0.81| 0.70    |   |
|    |              |         |    | (2.97)   |   | (2.29)  |   | (2.97)  |   |
|    |              |         |    | (2.26)   |   | (1.90)  |   | (2.26)  |   |
|    |              |         |    | (1.90)   |   | (1.46)  |   | (1.90)  |   |
| 1000|             | 0.00     | 0.00| 0.00     |   | 0.66    | 0.37| 0.28    |   |
|     |              |         |    | (3.04)   |   | (2.30)  |   | (3.04)  |   |
|     |              |         |    | (2.31)   |   | (1.94)  |   | (2.31)  |   |
|     |              |         |    | (1.94)   |   | (1.47)  |   | (1.94)  |   |
| 2000|             | 0.00     | 0.00| 0.00     |   | 0.02    | 0.00| 0.00    |   |
|     |              |         |    | (3.22)   |   | (2.34)  |   | (3.22)  |   |
|     |              |         |    | (2.45)   |   | (2.06)  |   | (2.45)  |   |
|     |              |         |    | (2.06)   |   | (1.49)  |   | (2.06)  |   |
Table 3: Average CPU time for each replication based on \( B = 500 \) replications for the SaC, the SAVGM, the WD, the debiased SaC and the debiased WD estimators for the logistic regression model with respect to \( K \) and the dimension \( p_2 \) of the nuisance parameter. Total sample size \( N = 2 \times 10^6 \).

| K   | \( SaC \) | \( SAVGM \) | \( WD \) | \( dSaC \) | \( dWD \) |
|-----|-----------|-------------|--------|---------|---------|
|     | \( p_2 = 4 \)     |             |        |         |         |
| 10  | 15.65     | 15.97       | 18.50  | 20.00   | 21.95   |
| 50  | 9.63      | 9.95        | 10.66  | 12.37   | 14.59   |
| 100 | 8.09      | 8.63        | 8.76   | 10.50   | 12.05   |
| 250 | 8.49      | 9.69        | 9.07   | 10.84   | 12.82   |
| 500 | 9.68      | 11.58       | 10.25  | 11.97   | 14.84   |
| 1000| 11.67     | 13.81       | 12.32  | 13.93   | 19.08   |
| 2000| 15.78     | 19.68       | 16.57  | 18.11   | 28.55   |
|     | \( p_2 = 10 \)     |             |        |         |         |
| 10  | 34.60     | 35.19       | 43.84  | 50.47   | 55.35   |
| 50  | 20.13     | 20.18       | 24.16  | 29.99   | 33.69   |
| 100 | 15.60     | 16.20       | 17.74  | 23.63   | 24.47   |
| 250 | 10.77     | 12.61       | 11.88  | 18.22   | 20.39   |
| 500 | 11.55     | 14.50       | 12.56  | 18.80   | 23.73   |
| 1000| 15.23     | 18.27       | 16.28  | 22.38   | 32.24   |
| 2000| 23.42     | 27.99       | 24.62  | 30.43   | 48.05   |