Statistica Sinica Preprint No: SS-2025-0060	
Title	Distributed Focused Information Criterion and Focused Frequentist Model Averaging for Massive Data
Manuscript ID	SS-2025-0060
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0060
Complete List of Authors	Yifan Zhang, Xiaolin Chen and Yuzhan Xing
Corresponding Authors	Xiaolin Chen
E-mails	xlchen@amss.ac.cn
Notice: Accepted author version.	

Statistica Sinica

Distributed Focused Information Criterion and Focused Frequentist Model Averaging for Massive Data

Yifan Zhang, Xiaolin Chen and Yuzhan Xing

School of Statistics and Data Science, Qufu Normal University

Abstract: This article investigates the focused information criterion and focused frequentist model averaging estimators for linear regression model with massive distributed data under the local asymptotic framework. Three divide-and-conquer-type approaches—one one-shot and two iterative methods—are employed to estimate the regression coefficient vector of each candidate model. We establish corresponding estimators' limiting distributions and the upper bounds for their mean square errors. These distributed estimators are subsequently utilized to develop the distributed focused information criterion and distributed focused frequentist model averaging estimators for the focus parameter. We also rigorously derive the asymptotic distributions for the distributed estimators of the targeted parameter under each candidate model and the distributed focused frequentist model averaging estimators with both fixed and data-driven weights, along with the upper bounds of mean square errors of the model averaging estimators. Extensive simulation studies are conducted to validate the established methodologies and the corresponding theories, and a real-world dataset is analyzed to demonstrate their practical application.

Key words and phrases: Distributed focused information criterion; Distributed focused frequentist model averaging; Local asymptotic framework.

1. Introduction

In empirical research involving statistical modelling, practitioners are frequently confronted with the model uncertainty. The most popular schemes to tackle this issue are model selection and model averaging with different goals. The goal of model selection is to pick up the most plausible model among candidate models according to some criteria, such as AIC (Akaike, 1973), Mallows' C$_p$ (Mallows, 1973), BIC (Schwarz, 1978) and so on. Among various criteria, the focused information criterion (FIC) suggested by Claeskens and Hjort (2003) is a special one. Unlike most methods, which pursue a universally "best" model—an ideal often difficult to justify—the FIC introduces a flexible framework that enables the selection of different models depending on the specific parameter of interest.

Although model selection enjoys some sound theoretical properties and has been widely used in practice, it is still criticized for some glaring flaws, such as ignorance of uncertainty in post-selection inference, instability to small data perturbation and so on. Instead of determining a "best" model, model averaging approaches acknowledge the usefulness of all candidate models, and combine estimators or predictors from all competing models by the weighting averaging with the goal of reducing estimation variance while controlling for bias arising from omitted variables (Hansen, 2007). By averaging, the model averaging naturally incorporates the model uncertainty into the conclusions about the unknown parameters (Wan et al., 2010), and thus is a smoothed extension of model selection with the potential of substantially reducing risk. Over the past two decades, various frequentist model averaging strategies have been proposed, including criterion based model averaging (Buckland et al., 1997; Claeskens and Hjort, 2003), adaptive regression by mixing models (Yang, 2001), Mallows model averaging (Hansen, 2007; Wan et al., 2010), optimal mean squared error averaging (Liang et al., 2011), Jackknife model averaging (Hansen and Racine, 2012), model averaging by K-fold cross-validation (Zhang and Liu, 2023) and so on. Particularly, since the seminal work of Hansen (2007), the literature on optimal model averaging has experienced rapid growth, with a central focus on determining the optimal convex combination of candidate estimates, which is typically accomplished by minimizing some specific performance measures. This expanding body of research spans diverse areas including high-dimensional data analysis (Ando and Li, 2014; Zhang et al., 2020), post-averaging inference (Zhang and Liu, 2019; Yu et al., 2024), distributed data analysis (Fang et al., 2018; Zhang et al., 2023; Xia et al., 2025), and in-depth theoretical investigations of model averaging (Peng and Yang, 2022; Peng et al., 2025) and so on.

In contemporary times, massive datasets are ubiquitous and frequently stored distributively across many machines or locations. Because of the challenges of computation, storage, communication cost and privacy protection, it is necessary to develop distributed statistical methodologies for these massive distributed data. Thus, a burgeoning literature has emerged in recent years, covering a broad range of topics, including M-estimation (Zhang et al., 2013; Shamir et al., 2014; Jordan et al., 2019; Fan et al., 2023), survival analysis (Su et al., 2022; Bayle et al., 2025), empirical likelihood (Zhou et al., 2023; Wang et al., 2025), to name a few.

Most existing distributed statistical approaches can be broadly categorized into two classes: one-shot (OS) and iterative methods. The key idea of OS methods is to compute relevant statistics on each node machine in parallel and then transmit these statistics to the central machine, which assembles them into the final estimator (Zhang et al., 2013). Although the OS approaches offer the advantage of minimal communication cost, they may not achieve the best statistical efficiency (Shamir et al., 2014; Jordan et al., 2019). As an alternative, the iterative methods could improve the statistical efficiency of OS methods via alternating between local computations and global aggregations. For instance, Shamir et al. (2014) introduced the Distributed Approximate NEwton (DANE) algorithm, which is particularly well-suited for stochastic optimization and learning problems. Jordan et al. (2019) developed a Communication-efficient Surrogate Likelihood (CSL) method for estimation and inference in various statistical models, including regular parametric models, penalized regression, and Bayesian statistics. Additional iterative algorithms and a comprehensive review of these methods can be found in Gao et al. (2022).

Notwithstanding the substantial body of work on distributed statistical approaches, the theory and methods for model selection and averaging in massive distributed data scenarios remain underdeveloped. To the best of our knowledge, the relevant literature is limited to a few works, such as Fang et al. (2018), Zhang et al. (2023) and Xia et al. (2025). In Fang et al. (2018), the authors investigated distributed model averaging for linear regression with homoscedastic and heteroscedastic errors and generalized linear model, but did not provide corresponding theoretical properties. Zhang et al. (2023) explored the least-squares model averaging method for distributed data, and suggested two aggregation strategies with sound theoretical foundations. However, the approaches in Zhang et al. (2023)

have three notable limitations: (1) they require that all candidate models are nested; (2) they employ the OS estimation procedure, which sacrifices some statistical efficiency compared with the iterative methods; (3) their weights are selected via Mallows criterion, which inherently limits their applicability to heteroscedastic data. Xia et al. (2025) extended the work of Zhang et al. (2023) in two key aspects: by accommodating non-nested candidate models, and adopting an iterative estimation approach. Nevertheless, the work of Xia et al. (2025) still operates under the Mallows framework, which could not handle heteroscedastic data. Different from the existing distributed model averaging literature, we will investigate the model selection and model averaging for massive distributed data under the local asymptotic framework (Hjort and Claeskens, 2003) with the emphasis on the asymptotic distributions of distributed model averaging estimators. Compared with the existing approaches, our proposed methods enjoy several characteristics. Firstly, our distributed model selection and averaging approaches are tailored to some specific parameters of interest, while the existed ones are designed according to some global measures. Secondly, under rather mild assumptions, we derive the asymptotic distribution of our distributed model averaging estimators—the first such result in the literature to date. Thirdly, our methodology does not require the candidate models to be nested, and accommodates the heteroscedastic error scenario. Fourthly, the iterative distributed estimation procedures are employed to derive our distributed estimators, ensuring enhanced statistical efficiency.

The rest of this article is organized as follows. In the next section, details of our model framework, three divide-and-conquer-type estimation approaches, and large-sample properties of distributed estimators of regression coefficient vector of each candidate model are presented. Section 3 provides the distributed focused information criterion (DFIC) for model selection and distributed focused frequentist model averaging (DFFMA) inference under the local misspecification configuration. In Section 4, we offer some simulation studies to examine the performance of suggested DFIC and DFFMA procedures. Section 5 illustrates our methodologies by a real-world application. Some conclusion remarks are given in Section 6. The *Supplementary Material* includes discussions about our technical assumptions, proofs of the theorems, additional numerical simulation studies, and implement details and further analysis of the real-world example.

2. Model framework and distributed estimations for candidate models

2.1 The statistical model

Suppose that $\mathcal{D}^N = \{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^N$ is an independent and identically distributed sample with $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ being generated from the following linear regression model:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{l=1}^{q} \gamma_l z_{il} + \varepsilon_i, \tag{2.1}$$

where y_i is the response variable, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})^{\mathrm{T}}$ are vectors of explanatory variables (covariates) with dimensions p and q, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^{\mathrm{T}}$ are vectors of unknown regression coefficients corresponding to $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, ε_i is the random error satisfying $E(\varepsilon_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = 0$ and $E(\varepsilon_i^2|\boldsymbol{x}_i, \boldsymbol{z}_i) = \sigma^2(\boldsymbol{x}_i, \boldsymbol{z}_i)$. The error in (2.1) is allowed to be homogeneous or heteroskedastic. Here, the covariates consist of two groups: the core covariates, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$, which must be included in the model based on prior knowledge, and the auxiliary covariates, $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})^{\mathrm{T}}$, which may or may not be relevant and are subject to selection. The full model's covariate vector is denoted by $\boldsymbol{h}_i = \left(\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{z}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ with $d = p + q$ components, and the corresponding regression coefficient vector is $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$, $\mathbf{X} =$

$(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^{\mathrm{T}}$, $\mathbf{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N)^{\mathrm{T}}$, $\mathbf{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)^{\mathrm{T}} = (\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^{\mathrm{T}}$. Then, in matrix notations, the model (2.1) can be expressed as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$.

In this article, we develop the distributed model selection and averaging estimators under the local misspecification framework, also known as the local asymptotic or local-to-zero framework (Hjort and Claeskens, 2003; Liu, 2015). Under this framework, the true value of regression coefficient vector is $\boldsymbol{\theta}_0 = \left(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}}/\sqrt{N}\right)^{\mathrm{T}}$, with $\boldsymbol{\delta}$ being a q-dimentional constant vector. It is very convenient to study the asymptotic distribution of model averaging estimators within this framework (Hjort and Claeskens, 2003; Claeskens and Hjort, 2003; Zhang and Liang, 2011; Liu, 2015). We refer the reader to Hjort and Claeskens (2003) for a detailed discussion of this local-to-zero assumption.

In consideration of model uncertainty, suppose that there exists a set of M $(\leq 2^q)$ candidate models for inferential choice. Furthermore, assume that the mth candidate model includes all core covariates in $\boldsymbol{x}$ and $q_{(m)}$ $(\leq q)$ auxiliary covariates from $\boldsymbol{z}$. Thus, the mth candidate model includes $d_{(m)} = p + q_{(m)}$ covariates. As for those candidate models, we allow them to be either nested or non-nested. Let $\boldsymbol{\Pi}_{(m)}$ be the $q_{(m)} \times q$ selection matrix. For $m = 1, 2, \ldots, M$, the mth approximating model is $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{i(m)}^{\mathrm{T}}\boldsymbol{\gamma}_{(m)} +$

$\varepsilon_{i(m)} \triangleq \boldsymbol{h}_{i(m)}^{\mathrm{T}}\boldsymbol{\theta}_{(m)} + \varepsilon_{i(m)}$, where $\boldsymbol{z}_{i(m)} = \boldsymbol{\Pi}_{(m)}\boldsymbol{z}_i$ represents the auxiliary covariate vector for the mth candidate model, $\boldsymbol{\gamma}_{(m)} = \boldsymbol{\Pi}_{(m)}\boldsymbol{\gamma}$ denotes the corresponding regression coefficient vector, $\boldsymbol{h}_{i(m)} = (\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{z}_{i(m)}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_{(m)} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}_{(m)}^{\mathrm{T}})^{\mathrm{T}}$, and $\varepsilon_{i(m)}$ is the error term which contains both the approximation and random errors. Write $\mathbf{Z}_{(m)} = (\boldsymbol{z}_{1(m)}, \boldsymbol{z}_{2(m)}, \ldots, \boldsymbol{z}_{N(m)})^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}_{(m)} = (\varepsilon_{1(m)}, \varepsilon_{2(m)}, \ldots, \varepsilon_{N(m)})^{\mathrm{T}}$. Let $\mathbf{H}_{(m)} = \left(\boldsymbol{h}_{1(m)}, \boldsymbol{h}_{2(m)}, \ldots, \boldsymbol{h}_{N(m)}\right)^{\mathrm{T}} = (\mathbf{X}, \mathbf{Z}_{(m)})$ denote the design matrix of the mth candidate model. Without loss of generality, we assume that the Mth model is the full model, so that $\mathbf{H}_{(M)} = \mathbf{H}$. In matrix notations, the mth candidate model can then be expressed as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{(m)}\boldsymbol{\gamma}_{(m)} + \boldsymbol{\varepsilon}_{(m)} = \mathbf{H}_{(m)}\boldsymbol{\theta}_{(m)} + \boldsymbol{\varepsilon}_{(m)}$.

2.2 Distributed estimators of regression coefficient vector

Suppose that the entire dataset is divided evenly and uniformly at random among K local machines, so that each local machine stores a non-overlapping subset of the data with manageable size. Without loss of generality, we assume that each local machine contains n observations, such that $N = Kn$, where both K and n are positive integers. For $k = 1, 2, \ldots, K$, the dataset stored in the kth machine is denoted by $\mathcal{D}_k^n = \{(y_{k,i}, \boldsymbol{x}_{k,i}, \boldsymbol{z}_{k,i})\}_{i=1}^n$.

Denote $\boldsymbol{y}_k = (y_{k,1}, y_{k,2}, \ldots, y_{k,n})^{\mathrm{T}}$, $\mathbf{X}_k = (\boldsymbol{x}_{k,1}, \boldsymbol{x}_{k,2}, \ldots, \boldsymbol{x}_{k,n})^{\mathrm{T}}$, $\mathbf{Z}_k = (\boldsymbol{z}_{k,1}, \boldsymbol{z}_{k,2}, \ldots, \boldsymbol{z}_{k,n})^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}_k = (\varepsilon_{k,1}, \varepsilon_{k,2}, \ldots, \varepsilon_{k,n})^{\mathrm{T}}$. Let $\mathbf{Z}_{k(m)} = \mathbf{Z}_k\boldsymbol{\Pi}_{(m)}^{\mathrm{T}} =$

$(\boldsymbol{z}_{k,1(m)}, \boldsymbol{z}_{k,2(m)}, \dots, \boldsymbol{z}_{k,n(m)})^{\mathrm{T}}$ with $\boldsymbol{z}_{k,i(m)} = \boldsymbol{\Pi}_{(m)}\boldsymbol{z}_{k,i}$ and $\mathbf{H}_{k(m)} = \big(\mathbf{X}_k, \mathbf{Z}_{k(m)}\big) = (\boldsymbol{h}_{k,1(m)}, \boldsymbol{h}_{k,2(m)}, \dots, \boldsymbol{h}_{k,n(m)})^{\mathrm{T}}$ with $\boldsymbol{h}_{k,i(m)} = (\boldsymbol{x}_{k,i}^{\mathrm{T}}, \boldsymbol{z}_{k,i(m)}^{\mathrm{T}})^{\mathrm{T}}$, which represents the design matrix of the mth candidate model within the kth machine. Then, the matrix form of the mth candidate model in the kth machine is given by $\boldsymbol{y}_k = \mathbf{X}_k\boldsymbol{\beta} + \mathbf{Z}_{k(m)}\boldsymbol{\gamma}_{(m)} + \boldsymbol{\varepsilon}_{k(m)} = \mathbf{H}_{k(m)}\boldsymbol{\theta}_{(m)} + \boldsymbol{\varepsilon}_{k(m)}$.

In this subsection, we consider three popular distributed estimation approaches to estimate the regression coefficient vector of each candidate model: the OS estimation (Zhang et al., 2013), the CSL estimation (Jordan et al., 2019) and the DANE estimation (Shamir et al., 2014).

2.2.1 The OS estimation

The ordinary least squares estimator of the true regression coefficient vector of the mth candidate model in the kth machine is given by $\tilde{\boldsymbol{\theta}}_{k(m)} = (\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)})^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{y}_k$. The K estimators $\tilde{\boldsymbol{\theta}}_{k(m)}$s are then transmitted to the central machine and aggregated via the simple averaging to obtain the final OS estimator, denoted by

$$\hat{\boldsymbol{\theta}}_{1,(m)} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\boldsymbol{\theta}}_{k(m)}. \tag{2.2}$$

From the estimation process, it is evident that the OS approach requires only a single round of communication between each local machine and the central machine, and thus involves extremely low communication burden.

2.2.2 The CSL estimation

Although the OS approach enjoys the advantage of low communication cost, it is subject to several limitations in practical applications (Gao et al., 2022). For example, the OS approach requires the local sample size to be large enough on each local machine. To overcome the disadvantages and enhance the estimation precision of OS approach, we consider the CSL method (Jordan et al., 2019) here.

In our setup, for the mth candidate model, let $l_{k,n}(\boldsymbol{\theta}_{(m)}) = n^{-1}\sum_{i=1}^{n}(y_{k,i} - \boldsymbol{h}_{k,i(m)}^{\mathrm{T}}\boldsymbol{\theta}_{(m)})^2$ and $l_N(\boldsymbol{\theta}_{(m)}) = N^{-1}\sum_{i=1}^{N}(y_i - \boldsymbol{h}_{i(m)}^{\mathrm{T}}\boldsymbol{\theta}_{(m)})^2 = K^{-1}\sum_{k=1}^{K} l_{k,n}(\boldsymbol{\theta}_{(m)})$ denote squared losses based on the data $\mathcal{D}_k^n$ and $\mathcal{D}^N$, respectively. Following Jordan et al. (2019), given the current estimator $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)}$, the surrogate loss function is $l_{1,n}(\boldsymbol{\theta}_{(m)}) - \left\langle \nabla l_{1,n}\left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)}\right) - \nabla l_N\left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)}\right), \boldsymbol{\theta}_{(m)}\right\rangle$, where $\nabla l_{1,n}(\boldsymbol{\theta}_{(m)})$ and $\nabla l_N(\boldsymbol{\theta}_{(m)})$ are the gradient vectors of $l_{1,n}(\boldsymbol{\theta}_{(m)})$ and $l_N(\boldsymbol{\theta}_{(m)})$, respectively. Through minimizing this surrogate loss function, the updating formula of the estimation process is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} &= \left[\mathbf{I}_{d_{(m)}} - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}_{1,i(m)}\boldsymbol{h}_{1,i(m)}^{\mathrm{T}}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_{i(m)}\boldsymbol{h}_{i(m)}^{\mathrm{T}}\right)\right]\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}_{1,i(m)}\boldsymbol{h}_{1,i(m)}^{\mathrm{T}}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_{i(m)}y_i\right),\end{aligned}$$

where $\mathbf{I}_{d_{(m)}}$ is the $d_{(m)} \times d_{(m)}$ identity matrix. After some algebraic manipulations, $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)}$ can be written as

$$\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} = \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} + \left(\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)}\right), \tag{2.3}$$

where $\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)} = \boldsymbol{y}_k - \mathbf{H}_{k(m)}\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)}$. From this iterative update formula, it is easy to see that the CSL estimation only requires the transmission of certain vectors and proceeds without the need of matrix communications.

2.2.3 The DANE estimation

In the CSL estimation procedure, only the Hessian matrix from the first local machine is utilized, while those from other local machines are discarded. Thus, the CSL procedure obviously wastes a lot of computing power and loses much information. It is noted that the DANE approach (Shamir et al., 2014) could address this problem, and possesses promising performance for quadratic loss function. Thus, we adapt it to our context.

The core idea of DANE is to minimize the surrogate loss on each local machine and then aggregate the results via simple averaging in each iteration. In our context, for the mth candidate model, given the current global estimator $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}$, the regression coefficient vector of the mth candidate model is locally estimated by

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k,3,(m)}^{(t+1)} &= \underset{\boldsymbol{\theta}_{(m)}}{\arg\min}\left\{l_{k,n}(\boldsymbol{\theta}_{(m)}) - \boldsymbol{\theta}_{(m)}^{\mathrm{T}}\left[\nabla l_{k,n}\left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}\right) - \nabla l_N\left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}\right)\right]\right\} \\ &= \hat{\boldsymbol{\theta}}_{3,(m)}^{(t)} + \left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)}\right),\end{aligned}$$

where $\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)} = \boldsymbol{y}_k - \mathbf{H}_{k(m)}\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}$. The central machine then collects those local estimators and aggregates them to get the updated global estimator $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)} = K^{-1}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_{k,3,(m)}^{(t+1)}$. After some algebraic manipulations, $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}$ could be expressed as

$$\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)} = \hat{\boldsymbol{\theta}}_{3,(m)}^{(t)} + \left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)}\right) \quad (2.4)$$

or

$$\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)} = \hat{\boldsymbol{\theta}}_{3,(m)}^{(t)} + \frac{1}{K}\sum_{k=1}^{K}\left\{\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_{k(m)}^{(t)}\right]\right\}. \quad (2.5)$$

Equation (2.4) indicates that the DANE estimation leverages more Hessian matrix information compared with the CSL method, while Equation (2.5) shows that there is no need to transmit matrices during the DANE's iteration.

2.3 Theoretical properties

Throughout this paper, we denote the Euclidean norm of a p-dimentional vector $\boldsymbol{v}$ by $\|\boldsymbol{v}\|_2 = (\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v})^{1/2}$, the infinity norm by $\|\boldsymbol{v}\|_\infty = \max_{1\le j\le p}\{|v|_j\}$, and the spectral norm of matrix $\mathbf{A}$ by $\|\mathbf{A}\| = \{\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\}^{1/2}$, where $\lambda_{\max}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)$ denotes the largest eigenvalue of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. We also write $\lambda_1 \preceq \mathbf{A} \preceq \lambda_2$ to indicate that all eigenvalues of $\mathbf{A}$ are bounded below by λ_1 and above by λ_2.

Denote by

$$\mathbf{S}_0 = \begin{pmatrix} \mathbf{0}_{p\times q} \\ \mathbf{I}_q \end{pmatrix} \quad \text{and} \quad \mathbf{S}_{(m)} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q_{(m)}} \\ \mathbf{0}_{q\times p} & \mathbf{\Pi}_{(m)}^{\mathrm{T}} \end{pmatrix},$$

the selection matrices of dimensions $d \times q$ and $d \times d_{(m)}$, respectively, where $\mathbf{0}_{\cdot\times\cdot}$ represents a zero matrix, and $\mathbf{I}_{\cdot}$ denotes the identity matrix. Then, the true value of regression coefficient vector of the mth candidate model is given by $\boldsymbol{\theta}_{0(m)} = \mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_0$. Define $\mathbf{Q} = E(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathrm{T}})$ and $\mathbf{\Omega} = E\left(\varepsilon_i^2\boldsymbol{h}_i\boldsymbol{h}_i^{\mathrm{T}}\right)$. Under homoscedasticity, $\mathbf{\Omega}$ reduces to $\mathbf{\Omega} = \sigma^2\mathbf{Q}$ with $\sigma^2 = E(\varepsilon_i^2|\boldsymbol{h}_i)$. For $m = 1, 2, \ldots, M$, denote $\mathbf{Q}_{(m)} = \mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{Q}\mathbf{S}_{(m)}$, $\mathbf{\Omega}_{(m)} = \mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{\Omega}\mathbf{S}_{(m)}$, $\mathbf{V}_{(m)} = \mathbf{Q}_{(m)}^{-1}\mathbf{\Omega}_{(m)}\mathbf{Q}_{(m)}^{-1}$, $\mathbf{A}_m = \mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{Q}\mathbf{S}_0(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)})$ and $\mathbf{B}_{(m)} = \mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}$. Let $\boldsymbol{\zeta}$ denote a d-dimentional random variable such that $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$.

Assumption 1. *There exist positive constants c_1 and c_2 such that $\|\boldsymbol{\beta}_0\|_\infty \leq c_1$ and $\|\boldsymbol{\delta}\|_\infty \leq c_2$.*

Assumption 2. *For $m = 1, 2, \ldots, M$, there exist positive constants $\overline{\lambda}_{(m)}$ and $\underline{\lambda}_{(m)}$ such that $\underline{\lambda}_{(m)} \leq \overline{\lambda}_{(m)}$ and $\underline{\lambda}_{(m)} \preceq \mathbf{Q}_{(m)} \preceq \overline{\lambda}_{(m)}$.*

Assumption 3. *There exist positive constants c_3 and c_4 such that $E\|\varepsilon_i \boldsymbol{h}_i\|_2^8 \leq c_3^8$ and $E\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathrm{T}} - \mathbf{Q}\|^8 \leq c_4^8$.*

To save space, the discussions about these assumptions are deferred to the *Supplementary Material*. The following results state the properties of the OS estimator for the coefficient vector of the mth candidate model.

Theorem 1. *Suppose that Assumptions 1 to 3 hold, and $K = o(n)$. Then, as $N \to \infty$, we have*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{1,(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta} \sim \mathcal{N}\left(\mathbf{A}_{(m)}\boldsymbol{\delta}, \mathbf{V}_{(m)}\right).$$

Besides, the mean squared error of the OS estimator $\hat{\boldsymbol{\theta}}_{1,(m)}$ satisfies that

$$E\left\|\hat{\boldsymbol{\theta}}_{1,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \leq \frac{u_{1,(m)}}{N} + O\left(\frac{1}{n^2}\right), \tag{2.6}$$

where $u_{1,(m)} > 0$ is a constant that depends on the mth candidate model but not on N.

The following Theorem 2 presents properties of the CSL estimator for $\boldsymbol{\theta}_{0(m)}$.

Theorem 2. *Given the current estimator $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)}$, under Assumptions 1 to 3 and $K = o(n^4)$, we have*

$$\|\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)}\|_2 = O_p(n^{-1/2})\|\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)}\|_2 + O_p(N^{-1/2}).$$

Additionally, when the current estimator $\hat{\boldsymbol{\theta}}^{(t)}_{2,(m)}$ *satisfies* $\sqrt{K}(\hat{\boldsymbol{\theta}}^{(t)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}) = o_p(1)$*, we have*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta} \sim \mathcal{N}\left(\mathbf{A}_{(m)}\boldsymbol{\delta}, \mathbf{V}_{(m)}\right),$$

as $N \to \infty$*. Furthermore, for the mean squared error of* $\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)}$*, it holds that*

$$E\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \;\leq\; \frac{u_{2,(m)}}{N} + O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}^{(t)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right) + O\left(\frac{1}{n^4}\right),$$

where $u_{2,(m)} > 0$ *is a constant that depends on the mth candidate model but not on* N*.*

Theorem 2 provides the convergence properties of the CSL estimator under each candidate model. Given that $\|\hat{\boldsymbol{\theta}}^{(0)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\|_2 = O_p(n^{-1/2})$, a condition that is easily satisfied, it follows that $\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\|_2 = O_p(n^{-(t+1)/2}) + O_p(N^{-1/2})$. Thus, for sufficiently large t, $\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\|_2 = O_p(N^{-1/2})$. This shows that as t goes to ∞, the estimation error of the CSL estimator will reach the order of $O_p(N^{-1/2})$. What's more, the CSL estimator will have the same asymptotic distribution as that of the oracle OLS estimator with the entire dataset. Suppose that $E\left\|\hat{\boldsymbol{\theta}}^{(0)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 = O(n^{-1})$, which holds, for example, for the least squares estimator. Then, the upper bound for the mean squared error of the CSL estimator will attain the order of $O(N^{-1} + n^{-4})$ after a sufficient large number of iterations.

The properties of the DANE estimator are given in the next theorem.

Theorem 3. *Suppose that Assumptions 1 to 3 hold, given the current estimator* $\hat{\boldsymbol{\theta}}^{(t)}_{3,(m)}$*, if* $K = o(n^4)$*, it follows that*

$$\|\hat{\boldsymbol{\theta}}^{(t+1)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}\|_2 = O_p(K^{1/4}n^{-1})\|\hat{\boldsymbol{\theta}}^{(t)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}\|_2 + O_p(N^{-1/2}).$$

Moreover, if $K^{3/4}n^{-1/2}(\hat{\boldsymbol{\theta}}^{(t)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}) = o_p(1)$*, we have*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}^{(t+1)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta} \sim \mathcal{N}\left(\mathbf{A}_{(m)}\boldsymbol{\delta}, \mathbf{V}_{(m)}\right),$$

as $N \to \infty$*. In addition, the mean squared error of* $\hat{\boldsymbol{\theta}}^{(t+1)}_{3,(m)}$ *satisfies that*

$$E\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \leq \frac{u_{3,(m)}}{N} + O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}^{(t)}_{3,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right) + O\left(\frac{1}{n^4}\right),$$

where $u_{3,(m)} > 0$ *is a constant that depends on the mth candidate model but not on* N*.*

Theorem 3 offers the convergence properties of the DANE estimator. These theoretical properties are similar to those of the CSL estimator, except the convergence rate. Thus, the implications of Theorem 3 are similar to those of Theorem 2 except a faster convergence rate in the number of iterations t.

3. The DFIC and DFFMA inference

In this section, we develop the DFIC and DFFMA inference for the focus parameter, based on the distributed estimations investigated in Section 2.

3.1 The DFIC inference

Denote by $\mu_0 = \mu(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\beta}_0, \boldsymbol{\delta}/\sqrt{N})$ the focus or targeted parameter with $\mu(\cdot)$ being a smooth d-dimensional real-valued function. In practice, the focus parameter may be a single coefficient of interest, the ratio of two different coefficients and so on. Assume that the partial derivatives of $\mu(\boldsymbol{\theta})$ are continuous in a neighborhood of $(\boldsymbol{\beta}_0^{\mathrm{T}}, \mathbf{0}_{q\times 1}^{\mathrm{T}})^{\mathrm{T}}$. Denote $\boldsymbol{\mu}_c = \frac{\partial\mu(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\gamma}=\mathbf{0}}$, $\boldsymbol{\mu}_a = \frac{\partial\mu(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\gamma}=\mathbf{0}}$, and $\boldsymbol{\mu}_{\boldsymbol{\theta}} = (\boldsymbol{\mu}_c^{\mathrm{T}}, \boldsymbol{\mu}_a^{\mathrm{T}})^{\mathrm{T}}$.

For the mth candidate model, we use the symbol $\hat{\boldsymbol{\theta}}_{(m)}$ to denote one of the three distributed estimators presented in Section 2 to simplify the subsequent presentation. In this symbol, we also discard the superscript which indicates the number of iterations as for the two iterative distributed estimators. Then, our focus parameter μ_0 can be estimated by $\hat{\mu}_{(m)} = \mu(\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)}) = \mu([\mathbf{I}_p, \mathbf{0}_{p\times q}]\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)}, [\mathbf{0}_{q\times p}, \mathbf{I}_q]\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)})$.

The following theorem states the theoretical properties of $\hat{\mu}_{(m)}$, including the asymptotic normality and upper bounds of the mean squared errors.

Theorem 4. *Suppose that* $\sqrt{N}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta}$, *as* $N \to \infty$. *Then, as* $N \to \infty$, *we have*

$$\sqrt{N}\left(\hat{\mu}_{(m)} - \mu_0\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{C}_{(m)}\boldsymbol{\delta}, \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{P}_{(m)}\boldsymbol{\Omega}\mathbf{P}_{(m)}\boldsymbol{\mu}_{\boldsymbol{\theta}}\right), \tag{3.7}$$

where $\mathbf{P}_{(m)} = \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{Q}\mathbf{S}_{(m)}\right)^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}$ *and* $\mathbf{C}_{(m)} = \left(\mathbf{P}_{(m)}\mathbf{Q} - \mathbf{I}_d\right)\mathbf{S}_0$. *Be-*

sides, if the Assumption 1 hold, there exists a constant $c_5 > 0$ *such that*

$$E\left\|\hat{\mu}_{(m)} - \mu_0\right\|_2^2 \leq c_5 E\left\|\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 + \frac{c_5(q - q_{(m)})c_2^2}{N}. \tag{3.8}$$

In this theorem, it is assumed that $\sqrt{N}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta}$. This is a high-level assumption, which holds for all the three distributed estimators in Section 2 under Assumptions 1 to 3 and some other mild assumptions. Theorem 4 shows that if the distributed estimator $\hat{\boldsymbol{\theta}}_{(m)}$ is asymptotic normality, then the distributed estimator of $\hat{\mu}_{(m)}$ is also asymptotic normality. This result aligns with that in Theorem 1 of Liu (2015), which achieves similar conclusions for data with small or moderate size without divide-and-conquer techniques. Additionally, the upper bounds of the mean squared errors in this theorem, together with upper bounds for $\hat{\boldsymbol{\theta}}_{(m)}$ in Theorems 1 to 3, yields the upper bounds of the mean squared error of $\hat{\mu}_{(m)}$.

In the light of Theorem 4, we could derive the DFIC for picking out the "best" model for the target parameter μ_0 via minimizing the asymptotic mean square error (AMSE) of $\hat{\mu}_{(m)}$. Through some algebraic manipulations, we could obtain the AMSE as

$$\mathrm{AMSE}\left(\hat{\mu}_{(m)}\right) = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\left(\mathbf{C}_{(m)}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}\mathbf{C}_{(m)}^{\mathrm{T}} + \mathbf{P}_{(m)}\boldsymbol{\Omega}\mathbf{P}_{(m)}\right)\boldsymbol{\mu}_{\boldsymbol{\theta}}.$$

It is noted that $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\mathbf{C}_{(m)}$, $\mathbf{P}_{(m)}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$ are all unknown. Thus, previous to ultimately present the DFIC, we need to estimate these unknown parameters. For $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, we define $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \partial\mu(\hat{\boldsymbol{\theta}}_{(M)})/\partial\boldsymbol{\theta}$, where $\hat{\boldsymbol{\theta}}_{(M)}$ is the distributed estimator of $\boldsymbol{\theta}_0$ based on the full model. Recall that $\hat{\boldsymbol{\theta}}_{(M)}$ is a consistent estimator under mild conditions, it follows that $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\mu}_0$. Recall that $\mathbf{P}_{(m)} = \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{Q}\mathbf{S}_{(m)}\right)^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}$ and $\mathbf{C}_{(m)} = \left(\mathbf{P}_{(m)}\mathbf{Q} - \mathbf{I}_d\right)\mathbf{S}_0$. To estimate $\mathbf{P}_{(m)}$ and $\mathbf{C}_{(m)}$, we only need to estimate $\mathbf{Q}$. A natural estimator of $\mathbf{Q}$ is $\hat{\mathbf{Q}} = K^{-1}\sum_{k=1}^{K}\hat{\mathbf{Q}}_k$, where $\hat{\mathbf{Q}}_k = n^{-1}\sum_{i=1}^{n}\boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathrm{T}}$. In practice, transmitting all $\hat{\mathbf{Q}}_k$s from local machines to the central machine is time-consuming. To reduce communication cost, we use only the estimators from a single local machine, e.g., $\hat{\mathbf{Q}}_1$ from the first one. Then, $\mathbf{P}_{(m)}$ and $\mathbf{C}_{(m)}$ could be estimated by $\hat{\mathbf{P}}_{(m)} = \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\mathbf{Q}}_1\mathbf{S}_{(m)}\right)^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}$ and $\hat{\mathbf{C}}_{(m)} = \left(\hat{\mathbf{P}}_{(m)}\hat{\mathbf{Q}}_1 - \mathbf{I}_d\right)\mathbf{S}_0$. Similarly, $\boldsymbol{\Omega}$ is also estimated from data on the first machine, denoted by $\hat{\boldsymbol{\Omega}}_1$. As explained in Hjort and Claeskens (2003), a consistent estimator of $\boldsymbol{\delta}$ does not exist under the local asymptotic framework. Following the approach of Liu (2015), we instead construct an asymptotically unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}^T$. Let $\hat{\boldsymbol{\delta}} = \mathbf{S}_0^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(M)}$ and $\boldsymbol{\zeta}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0^{\mathrm{T}}\mathbf{Q}^{-1}\boldsymbol{\zeta}$. From the results in Theorems 1 to 3, it is easy to see that, under mild assumptions, $\hat{\boldsymbol{\delta}} \xrightarrow{d} \boldsymbol{\zeta}_{\boldsymbol{\delta}} \sim \mathcal{N}\left(\boldsymbol{\delta}, \mathbf{S}_0^{\mathrm{T}}\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0\right)$. Thus, we could conclude that $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^T - \mathbf{S}_0^T\hat{\mathbf{Q}}_1^{-1}\hat{\boldsymbol{\Omega}}_1\hat{\mathbf{Q}}_1^{-1}\mathbf{S}_0$ is an asymptotically unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}^T$. Finally, substituting all estimators into

AMSE $\left(\hat{\mu}_{(m)}\right)$, we obtain the DFIC for the mth candidate model as

$$\mathrm{DFIC}_{(m)} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{\mathrm{T}} \left[\hat{\mathbf{C}}_{(m)} \left(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}} - \mathbf{S}_0^{\mathrm{T}}\hat{\mathbf{Q}}_1^{-1}\hat{\boldsymbol{\Omega}}_1^{-1}\hat{\mathbf{Q}}_1^{-1}\mathbf{S}_0\right) \hat{\mathbf{C}}_{(m)}^{\mathrm{T}} + \hat{\mathbf{P}}_{(m)}\hat{\boldsymbol{\Omega}}_1\hat{\mathbf{P}}_{(m)}\right] \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}.$$

From its construction, it is easy to see that $\mathrm{DFIC}_{(m)}$ is an asymptotically unbiased estimator of AMSE $\left(\hat{\mu}_{(m)}\right)$. Then, the model with the smallest $\mathrm{DFIC}_{(m)}$ value is selected as the "best" model for the parameter of interest.

3.2 The DFFMA inference

In this subsection, we extend the idea of DFIC to the averaging estimator of the focus parameter. Let $\boldsymbol{w} = (w_1, w_2, \ldots, w_M)^{\mathrm{T}}$ be a weight vector lying in the unit simplex $\mathcal{W}$ of $\mathbb{R}^M$, i.e., $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. The distributed averaging estimator of μ with fixed weight vector $\boldsymbol{w}$ is then given by $\hat{\mu}(\boldsymbol{w}) = \sum_{m=1}^M w_m \hat{\mu}_{(m)}$, where $\hat{\mu}_{(m)}$s are defined in Section 3.1. The following theorem offers the asymptotic normality of $\hat{\mu}(\boldsymbol{w})$.

Theorem 5. *For $m = 1, 2, \ldots, M$, suppose that $\sqrt{N}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta}$ as $N \to \infty$, then it holds that*

$$\sqrt{N}\left(\hat{\mu}(\boldsymbol{w}) - \mu_0\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{C}_{\boldsymbol{w}}\boldsymbol{\delta}, \mathbf{V}_{\boldsymbol{w}}\right),$$

where $\mathbf{C}_{\boldsymbol{w}} = \sum_{m=1}^M w_m \mathbf{C}_{(m)}$, $\mathbf{V}_{\boldsymbol{w}} = \sum_{m=1}^M w_m^2 \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{P}_{(m)}\boldsymbol{\Omega}\mathbf{P}_{(m)}\boldsymbol{\mu}_{\boldsymbol{\theta}} + 2\sum\sum_{m\neq l} w_m w_l \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{P}_{(m)}\boldsymbol{\Omega}\mathbf{P}_{(l)}\boldsymbol{\mu}_{\boldsymbol{\theta}}$.

The proof of Theorem 5 is similar to that in Liu (2015), so we omit it here. With the assistance of Theorem 5, we could acquire the unavailable optimal weight vector which minimizes AMSE$\,(\hat{\mu}(\boldsymbol{w}))$ over $\boldsymbol{w} \in \mathcal{W}$. Specifically, it is easy to obtain that AMSE$(\hat{\mu}(\boldsymbol{w})) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{w}$, where $\boldsymbol{\Phi}$ is an $M \times M$ matrix with the (m, l)th element being

$$\boldsymbol{\Phi}_{m,l} = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \left(\mathbf{C}_{(m)}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}\mathbf{C}_{(l)}^{\mathrm{T}} + \mathbf{P}_{(m)}\boldsymbol{\Omega}\mathbf{P}_{(l)}\right) \boldsymbol{\mu}_{\boldsymbol{\theta}}. \tag{3.9}$$

Thus, the unavailable optimal weight vector is $\boldsymbol{w}^{(o)} = \arg\min_{w\in\mathcal{W}} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{w}$.

It is noted that $\boldsymbol{w}^{(o)}$ has no closed-form expression. Thus, the data-dependent weight vector could only be achieved by minimizing the sample version of $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{w}$. By plugging $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$, $\hat{\mathbf{C}}_{(m)}$s, $\hat{\boldsymbol{\delta}}$, $\hat{\mathbf{P}}_{(m)}$s and $\hat{\boldsymbol{\Omega}}_1$ in (3.9), we could get $\boldsymbol{\Phi}$'s estimator, denoted by $\hat{\boldsymbol{\Phi}}$, with the (m, l)th element being

$$\hat{\boldsymbol{\Phi}}_{m,l} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{\mathrm{T}} \left(\hat{\mathbf{C}}_{(m)}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\mathrm{T}}\hat{\mathbf{C}}_{(l)}^{\mathrm{T}} + \hat{\mathbf{P}}_{(m)}\hat{\boldsymbol{\Omega}}_1\hat{\mathbf{P}}_{(l)}\right) \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}.$$

Then, the data-dependent weight vector is $\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Phi}}\boldsymbol{w}$. Eventually, we could offer our DFFMA estimator with data-dependent weight vector as $\hat{\mu}(\hat{\boldsymbol{w}}) = \sum_{m=1}^{M} \hat{w}_m\hat{\mu}_{(m)}$.

The following theorem affords the large-sample properties of $\hat{\mu}(\hat{\boldsymbol{w}})$, including the asymptotic normality and the upper bound of the mean square error.

Theorem 6. *For $m = 1, 2, \ldots, M$, suppose that $\sqrt{N}(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}) \xrightarrow{d} \mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta}$, as $N \to \infty$, then we have $\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Phi}}\boldsymbol{w} \xrightarrow{d} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}^*\boldsymbol{w}$, where $\boldsymbol{\Phi}^*$ is* *an $M \times M$ matrix with the (m,l)th element being*

$$\mathbf{\Phi}^*_{m,l} = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \left(\mathbf{C}_{(m)} \boldsymbol{\zeta}_{\boldsymbol{\delta}} \boldsymbol{\zeta}_{\boldsymbol{\delta}}^{\mathrm{T}} \mathbf{C}_{(l)}^{\mathrm{T}} + \mathbf{P}_{(m)} \mathbf{\Omega} \mathbf{P}_{(l)} \right) \boldsymbol{\mu}_{\boldsymbol{\theta}}.$$

And, it also holds that $\hat{\boldsymbol{w}} \xrightarrow{d} \boldsymbol{w}^*$, where $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{w}^{\mathrm{T}} \mathbf{\Phi}^* \boldsymbol{w}$. Furthermore,

$$\sqrt{N} \left(\hat{\mu}(\hat{\boldsymbol{w}}) - \mu_0 \right) \xrightarrow{d} \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{Q}^{-1} \boldsymbol{\zeta} + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \left(\sum_{m=1}^{M} w_m^* \mathbf{C}_{(m)} \right) \boldsymbol{\zeta}_{\boldsymbol{\delta}}.$$

Besides these results, suppose that Assumption 1 is valid, it holds that

$$E \left\| \hat{\mu}(\hat{\boldsymbol{w}}) - \mu_0 \right\|_2^2 \leq c_5 M \max_{1 \leq m \leq M} \left\{ E \left\| \hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)} \right\|_2^2 + \frac{(q - q_{(m)}) c_2^2}{N} \right\},$$

where c_5 is defined in Theorem 4.

These asymptotic properties in Theorem 6 coincide with those of frequentist model averaging estimator for data with small or moderate size without divide-and-conquer techniques in Liu (2015). The proof of these results is similar to that in Liu (2015), so we omit it in this article for simplicity. In addition, similar to the discussion in Liu (2015), it is not appropriate to directly construct confidence interval for μ_0 based on Theorem 6. Here, we follow the operations of Hjort and Claeskens (2003) and Liu (2015) to get the valid confidence interval. Specially, denote by $\hat{\varphi}^2 = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{\mathrm{T}} \hat{\mathbf{Q}}_1^{-1} \hat{\mathbf{\Omega}}_1 \hat{\mathbf{Q}}_1^{-1} \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and $b_{\boldsymbol{\delta}}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{\mathrm{T}} (\sum_{m=1}^{M} \hat{w}_m \hat{\mathbf{C}}_{(m)}) \hat{\boldsymbol{\delta}}$. A $100(1-\alpha)\%$

level confidence interval for μ_0 could be constructed as follows

$$\mathrm{CI}_N = \left[\hat{\mu}(\hat{\boldsymbol{w}}) - \frac{b_{\boldsymbol{\delta}}(\hat{\boldsymbol{w}}) + z_{1-\alpha/2}\hat{\varphi}}{\sqrt{N}}, \hat{\mu}(\hat{\boldsymbol{w}}) - \frac{b_{\boldsymbol{\delta}}(\hat{\boldsymbol{w}}) - z_{1-\alpha/2}\hat{\varphi}}{\sqrt{N}}\right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

4. Simulation studies

In this section, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed DFIC model selection and DFFMA procedures, comparing with some other competing approaches. Specifically, eleven procedures are examined, including approaches of FIC model selection and focus frequentist model averaging with whole data, denoted by FIC-OR and FFMA-OR; three distributed estimations under the full model, denoted by OS, CSL$_t$ and DANE$_t$; three DFIC model selection and three DFFMA methods, denoted by DFIC-OS, DFIC-CSL$_t$, DFIC-DANE$_t$, DFFMA-OS, DFFMA-CSL$_t$ and DFFMA-DANE$_t$. For all iterative distributed approaches, the subscript "t" indicates the number of iterations. As suggested by one reviewer, we also examined two distributed model averaging methods proposed in Xia et al. (2025); However, the performance of these methods are substantially inferior to ours under our context, so the specific numerical results are omitted here. Additionally, it is noted that the methods of OS, DFIC-OS and DFFMA-OS behave significantly inferior to the other eight methods. To facilitate graphical presentation, two plots with different axis scales are provided for each scenario, one of which excludes the methods of OS, DFIC-OS and DFFMA-OS for clearer visualization of the main trends of the other methods. What's more, to conserve space, we have included the numerical results for heteroscedastic scenarios and further simulation studies in the *Supplementary Material*.

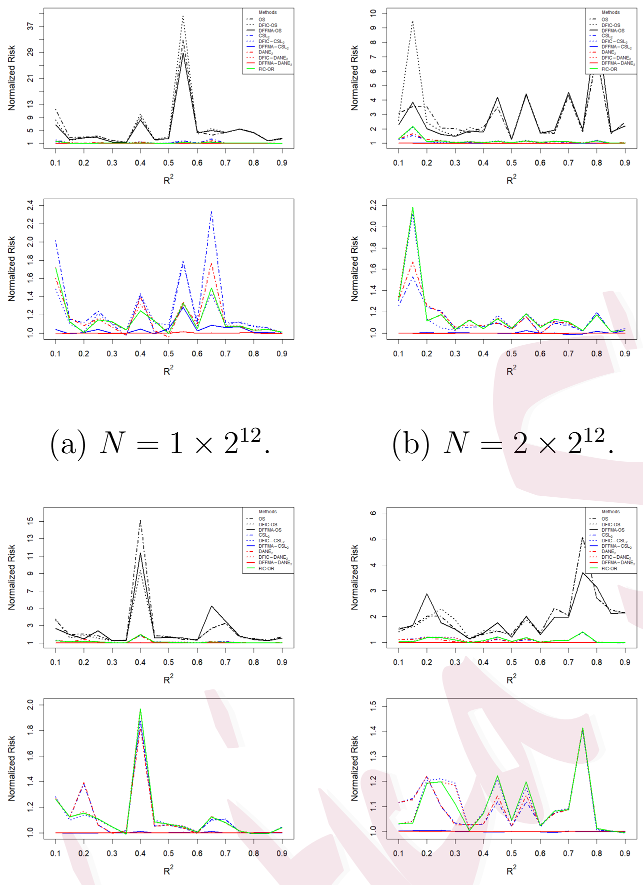
The independent and identically distributed observations were generated from the linear regression model

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{l=1}^{q} \gamma_l z_{il} + \varepsilon_i, \quad i = 1, 2, \ldots, N,$$

where $x_{i1} = 1$ and $(x_{i2}, \ldots, x_{ip}, z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ follows the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix of unit variances and constant correlation ρ. The error term ε_i follows the normal distribution $\mathcal{N}(0, \sigma_i^2)$ with $\sigma_i = 1$ for the homoscedastic scenario and $\sigma_i = x_{i2}$ for the heteroscedastic scenario. The regression coefficient vector was set to be

$$\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)^{\mathrm{T}} = u \left(\frac{a}{64} (1, 1, \ldots, 1)_{1 \times p}, \frac{1}{\sqrt{N}} \left(1, \frac{q-1}{q}, \ldots, \frac{1}{q} \right) \right)^{\mathrm{T}},$$

where u controls the population $R^2 = (var(y_i) - var(\varepsilon_i))/var(y_i)$. The dimension p of core regressors is fixed to be 3 in all of the subsequent simulation studies. For candidate models, only nested models were considered, ranging from the narrow model (with core covariates only) to the full model

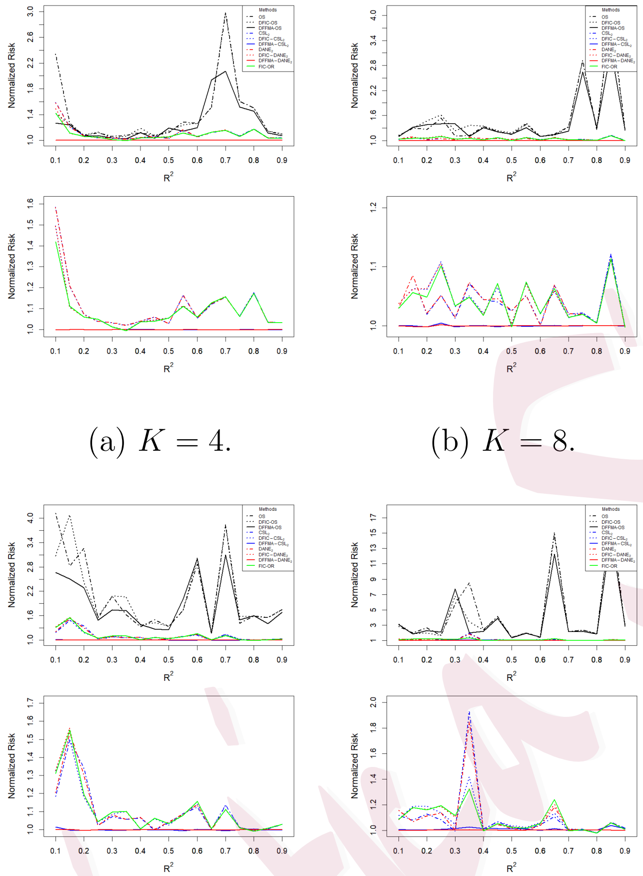


(a) $N = 1 \times 2^{12}$. (b) $N = 2 \times 2^{12}$.

(c) $N = 3 \times 2^{12}$. (d) $N = 4 \times 2^{12}$.

Figure 1: Normalized risks for settings of varying N with homoscedastic errors.

(with all covariates). To estimate the regression coefficient vector of each candidate model, the initial values of the CSL and DANE methods were set to be the OS estimators. Here, we consider the focus parameter $\mu = \beta_1 + \beta_2 + \cdots + \beta_p$, i.e., sum of the regression coefficients of the core covariates.



(a) $K = 4$.

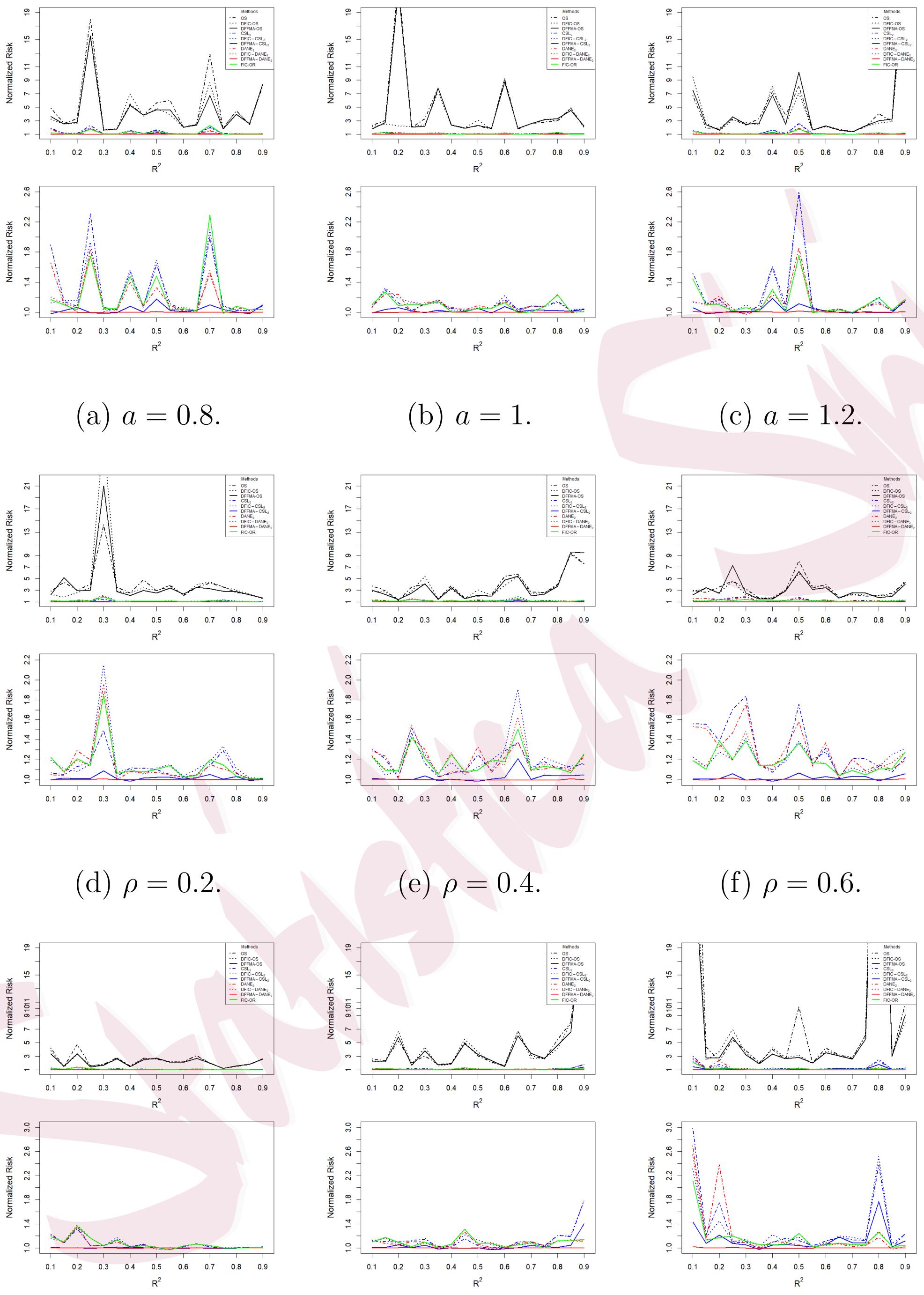
(b) $K = 8$.

(c) $K = 16$.

(d) $K = 32$.

Figure 2: Normalized risks for settings of varying K with homoscedastic errors.

To appraise the aforementioned approaches, we compute the average risks (expected squared errors) over 200 data repetitions and report the normalized risks, which are obtained through dividing by the risk of FFMA-OR. A smaller value of the normalized risk is indicative of the superior statistical



(a) $a = 0.8$. (b) $a = 1$. (c) $a = 1.2$.

(d) $\rho = 0.2$. (e) $\rho = 0.4$. (f) $\rho = 0.6$.

(g) $q = 5$. (h) $q = 10$. (i) $q = 15$.

Figure 3: Normalized risks for different settings of a, ρ, and q with homoscedastic errors.

performances. In our *Supplementary Material*, we carefully evaluate the influence of the number of iterations on the performance of iterative distributed methods. It is observed that all suggested iterative distributed methods converge rapidly, typically requiring only two iterations in that setting. Therefore, in the subsequent analysis, the number of iterations for these methods was set to be two.

We firstly evaluate the performance of various approaches by varying values of N, K, q, a and ρ. Different values of u were chosen to generate R^2 ranging from 0.1 to 0.9 in increments of 0.05 under various scenarios. Specifically, we consider the following experimental scenarios: (i) varying $N \in \{1 \times 2^{12}, 2 \times 2^{12}, 3 \times 2^{12}, 4 \times 2^{12}\}$ with $(a, \rho, q, K) = (1, 0.2, 10, 16)$; (ii) varying $K \in \{4, 8, 16, 32\}$ with $(N, a, \rho, q) = (3 \times 2^{12}, 1, 0.2, 10)$; (iii) varying $q \in \{5, 10, 15\}$ with $(N, K, a, \rho) = (2^{12}, 16, 1, 0.2)$; (iv) varying $a \in \{0.8, 1, 1.2\}$ with $(N, K, q, \rho) = (2^{12}, 16, 10, 0.2)$; (v) varying $\rho \in \{0.2, 0.4, 0.6\}$ with $(N, K, q, a) = (2^{12}, 16, 10, 1)$. The normalized risks for different methods under varying signal levels are shown in Figures 1 to 3 for all considered scenarios.

The numerical results of these figures reveal several key insights. To begin with, the normalized risks generally increase with the number of machines K but decrease with the sample size N. The DFFMA estimators

Table 1: Empirical CPs for different methods with homoscedastic errors.

$N(\times 2^{12})$	K	R^2	Methods		
			DFFMA-OS	DFFMA-CSL$_2$	DFFMA-DANE$_2$
1	4	0.3	0.894	0.902	0.902
		0.6	0.924	0.926	0.926
		0.9	0.904	0.902	0.902
	8	0.3	0.882	0.884	0.886
		0.6	0.890	0.890	0.890
		0.9	0.882	0.886	0.886
	16	0.3	0.898	0.912	0.912
		0.6	0.898	0.904	0.904
		0.9	0.894	0.908	0.912
2	4	0.3	0.922	0.918	0.918
		0.6	0.898	0.898	0.898
		0.9	0.872	0.876	0.876
	8	0.3	0.894	0.904	0.904
		0.6	0.888	0.894	0.896
		0.9	0.900	0.900	0.900
	16	0.3	0.874	0.876	0.876
		0.6	0.928	0.920	0.924
		0.9	0.922	0.914	0.912

typically exhibit lower risks compared to their counterparts of DFIC model selection estimators and distributed estimators based on the full model. Next, for small R^2s (indicating a high level of noise in the model), DFFMA estimators generally outperform their counterparts of DFIC model selection estimators and distributed estimators based on the full model, showing that DFFMA methods tend to be more robust and provide a form of insurance against selecting a poorly performing candidate model. Finally, as for the distributed estimation means, OS estimator, along with the DFIC-OS and

DFFMA-OS estimators, are unstable and exhibit low statistical efficiency, while iterative approaches perform much better.

Secondly, we consider some numerical studies on confidence interval construction. In this investigation, N and K vary across $\{1,2\} \times 2^{12}$ and $\{4,8,16\}$ respectively, with (a,ρ,q) being fixed at $(1,0.2,10)$. Different us were chosen to produce the desired $R^2 = 0.3$, 0.6 and 0.9. We examined the methods of DFFMA-OS, DFFMA-CSL$_2$, and DFFMA-DANE$_2$ and construted the 90% confidence intervals for μ_0. The results of empirical coverage probabilities (CPs) over 500 data repetitions are presented in Table 1. From these results, we observe that the empirical CPs of the three distributed model averaging estimators are all close to the nominal level, demonstrating the validity of our methodology and theoretical analysis.

In this section, we treated the FFMA-OR as the gold standard, and as one reviewer noted, its performance is highly informative. After examining the average risks of FFMA-OR estimators, we found that they are at most on the order of 10^{-3} in our designed scenarios. This observation, together with the risk ratios of our methods to FFMA-OR being close to 1, provides compelling evidences for the efficacy of our suggested methods.

5. A real-world example

In this section, we illustrate our suggested DFIC model selection and DFFMA procedures by reanalyzing a large, publicly available dataset: the U.S. Airline Dataset, which has been analyzed in Zhang et al. (2023) under the fixed parameter configuration. This dataset comprises detailed records of all commercial flights within the United States, from October 1987 to April 2008. The total sizes of this data are approximately 1.7GB and 12GB for compressed and uncompressed versions, respectively.

In our illustrative analysis, we mainly consider the influence of some regressors on arrival delay (ArrDelay, in hours). Specifically, these explanatory variables include three continuous covariates-departure delay (DepDelay, in hours), scheduled elapsed time (CRSElapsedTime, in hours), and distance between airports (Distance, in units of 1,000 miles)-and three categorical covariates: Month (coded as 1 for winter months—December, January, February—and 0 otherwise), DayofWeek (1 for weekends, 0 for weekdays), and DepTime (1 for departures between 8 p.m. and 5 a.m., 0 otherwise). Therefore, our full model is specified as

$$\begin{aligned}\text{ArrDelay} &= \beta_1 + \beta_2\text{DepDelay} + \gamma_1\text{Month} + \gamma_2\text{DepTime} + \gamma_3\text{Distance} \\ &\quad + \gamma_4\text{DayofWeek} + \gamma_5\text{CRSElapsedTime} + \varepsilon. \qquad (5.10)\end{aligned}$$

To conserve space, we present the detailed modeling process for these variables within the *Supplementary Material*. To analyze this dataset based on model (5.10), we utilized 116,212,331 observations from this Airline Dataset after removing records with missing data. In order to align with the setup in the simulation, the focus parameter is set to be $\mu = \beta_1 + \beta_2$, representing the sum of the coefficients of intercept and the core covariate DepDelay. As for the candidate models, we consider a suite of nested candidate models, leading to a total of six distinct models.

To truly implement various methods mentioned in a distributed manner, as suggested by the associate editor, we conducted our analysis on the Spark system. To save space, implementation details are deferred to the *Supplementary Material*. Since the true value of the focus parameter μ is unknown, we treat the FFMA-OR estimator computed using the full dataset as a benchmark. Table 2 reports the squared errors of each estimator relative to this reference, where DMAP-SA and DMAP-SL$_t$ are the methods suggested in Xia et al. (2025). As shown in this table, our proposed methods-DFIC-CSL$_t$, DFIC-DANE$_t$, DFFMA-CSL$_t$ and DFFMA-DANE$_t$-significantly outperform other strategies; notably, DFFMA-DANE$_t$ exhibits even better performance than the impractical FIC-OR. Besides, some inferences for the focus parameter $\mu = \beta_1 + \beta_2$ are provided in the *Supplementary*

Table 2: Relative squared errors of various methods ($\times 10^7$).

Method	**OS**	**CSL$_2$**	**DANE$_2$**
Relative Squared Error	33.274	1.6323	0.15727
Method	**DFIC-OS**	**DFIC-CSL$_2$**	**DFIC-DANE$_2$**
Relative Squared Error	33.268	1.6319	0.15716
Method	**DFFMA-OS**	**DFFMA-CSL$_2$**	**DFFMA-DANE$_2$**
Relative Squared Error	33.256	0.99620	0.01064
Method	**DMAP-SA**	**DMAP-SL$_2$**	**FIC-OR**
Relative Squared Error	10.689	7.6028	0.11092

Material.

Besides $\mu = \beta_1 + \beta_2$, we also conducted inference for the focus parameter $\mu = \beta_2$. The corresponding results are outlined in the *Supplementary Material*.

6. Conclusion remarks

In this article, we studied the DFIC and DFFMA inference for the targeted parameter in linear regression model with massive distributed data. This study lays the groundwork for several potential research directions. For example, during the implement of our distributed estimations, the estimators from all candidate models need to be transmitted in each interaction between the central and node machines. The more candidate models we combine, the greater the statistical efficiency of our aggregation estimators; however, this comes at the cost of increased communication expenses. Therefore, if we could select a subset of candidate models in advance and then implement our DFIC and DFFMA inferences based on the models in this subset, the transmission costs would be reduced. Additionally, this article is only concerned with the linear regression model and regular complete data. It would be greatly desirable to extend the suggested DFIC and DFFMA procedures to more general models and complex data. These ideas warrant further investigations along our directions.

Supplementary material

The *Supplementary Material* includes discussions about our technical assumptions, proofs of the theorems, additional numerical simulation studies, and implementation details and further analysis of the real-world example.

Acknowledgements

Chen's research is supported by the National Social Science Fund of China (25BTJ038) .

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó Location Budapest, Hungary.

Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265.

Bayle, P., Fan, J., and Lou, Z. (2025). Communication-efficient distributed estimation and inference for Cox's model. *Journal of the American Statistical Association*, (just-accepted):1–20.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916.

Fan, J., Guo, Y., and Wang, K. (2023). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 118(542):1000–1010.

Fang, F., Yin, X., and Zhang, Q. (2018). Divide and conquer algorithms for model averaging with massive data. *Journal of Systems Science and Mathematical Sciences*, 38(7):764.

Gao, Y., Liu, W., Wang, H., Wang, X., Yan, Y., and Zhang, R. (2022). A review of distributed statistical inference. *Statistical Theory and Related Fields*, 6(2):89–99.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681.

Liang, H., Zou, G., Wan, A. T., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106(495):1053–1066.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186(1):142–159.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Peng, J., Li, Y., and Yang, Y. (2025). On optimality of Mallows model averaging. *Journal of the American Statistical Association*, 120(550):1152–1163.

Peng, J. and Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, 229(2):246–262.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32, pages 1000–1008.

Su, W., Yin, G., Zhang, J., and Zhao, X. (2022). Divide and conquer for accelerated failure time model with massive time-to-event data. *Canadian Journal of Statistics*, 51(2):400–419.

Wan, A. T., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2):277–283.

Wang, Q., Du, J., and Sheng, Y. (2025). Distributed empirical likelihood inference with or without byzantine failures. *Statistics and Computing*, 35(5):1–20.

Xia, X., He, S., and Pang, N. (2025). Communication-efficient model averaging prediction for massive data with asymptotic optimality. *Statistical Papers*, 66(2):1–45.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yu, D., Lian, H., Sun, Y., Zhang, X., and Hong, Y. (2024). Post-averaging inference for optimal model averaging estimator in generalized linear models. *Econometric Reviews*, 43(2-4):98–122.

Zhang, H., Liu, Z., and Zou, G. (2023). Least squares model averaging for distributed data. *Journal of Machine Learning Research*, 24(215):1–59.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1):174–200.

Zhang, X. and Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory*, 35(4):816–841.

Zhang, X. and Liu, C.-A. (2023). Model averaging prediction by K-fold cross-validation. *Journal of Econometrics*, 235(1):280–301.

Zhang, X., Zou, G., Liang, H., and Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115(530):972–984.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(68):3321–3363.

Zhou, L., She, X., and Song, P. X.-K. (2023). Distributed empirical likelihood approach to integrating unbalanced datasets. *Statistica Sinica*, 33(3):2209–2231.

Yifan Zhang

School of Statistics and Data Science, Qufu Normal University E-mail: yfzhang@188.com

Xiaolin Chen

School of Statistics and Data Science, Qufu Normal University E-mail: xlchen@amss.ac.cn

Yuzhan Xing

School of Statistics and Data Science, Qufu Normal University E-mail: yzXing1221@163.com