Supplementary material for "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

S1 Discussions about the technical assumptions in the main body of this article

Assumption 1 imposes some constraints for the regression coefficients. In addition, from this assumption, we could conclude that the parameter space $\Theta \subset \mathbb{R}^d$ for $\boldsymbol{\theta}$ is compact, which is frequently encountered in the literature on distributed data analysis, such as Zhang et al. (2013), Huang and Huo (2019) and Jordan et al. (2019). Assumption 2 states that the eigenvalues of $\mathbf{Q}_{(m)}$ is bounded below and above by some positive constants, which ensures that $\mathbf{Q}_{(m)}$ is a positive definite matrix. Assumption 3 imposes certain moment conditions, which align with the smoothness assumption proposed in Zhang et al. (2013) for the squared loss function. However, we require these conditions to hold only at the true value of regression coefficient vector. This requirement is milder than that of Zhang et al. (2013), which necessitates these conditions to be satisfied in a neighborhood around the true value of regression coefficient vector. In addition, Assumptions 2 and 3 together guarantee that the Hessian matrix $n^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}$ of the empirical loss function $l_{k,n}(\boldsymbol{\theta}_{(m)})$ is invertible and $\left\|n^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right\|$ is bounded with high probability. These results are important for deriving the upper bounds of the mean squared errors of the distributed estimators for $\boldsymbol{\theta}_{0(m)}$.

S2 Proofs of the theorems

Before proceeding to the proofs of the main theorems, we present two auxiliary lemmas that provide upper bounds for the sums of independent and identically distributed random vectors and matrices. These lemmas have appeared in the literature of Rosenblatt and Nadler (2016) and Huang and Huo (2019), where detailed proofs can be found.

Lemma 1. *Let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n \in \mathbb{R}^d$ be i.i.d. random vectors with $E\boldsymbol{\xi}_i = \mathbf{0}$. Suppose there exist some constant $a > 0$ and $l_0 \geq 2$ such that $E\left\|\boldsymbol{\xi}_i\right\|_2^{l_0} \leq a^{l_0}$. Let $\overline{\boldsymbol{\xi}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{\xi}_i$, then for $1 \leq l \leq l_0$, we have*

$$E\left\|\overline{\boldsymbol{\xi}}_n\right\|_2^l \leq \frac{c_{vct}(l,d)a^l}{n^{l/2}},$$

where $c_{vct}(l,d)$ is a constant depending solely on l and d.

Lemma 2. *Let $\mathbf{A}_1,\ldots,\mathbf{A}_n \in \mathbb{R}^{d\times d}$ be i.i.d. random matrices with $E\mathbf{A}_i = \mathbf{0}_{d\times d}$. Suppose there exist some constant $b>0$ and $l_0 \geq 2$ such that $E\|\mathbf{A}_i\|^{l_0} \leq b^{l_0}$. Let $\overline{\mathbf{A}}_n = n^{-1}\sum_{i=1}^n \mathbf{A}_i$, then for $1 \leq l \leq l_0$, we have*

$$E\left\|\overline{\mathbf{A}}_n\right\|^l \leq \frac{c_{mtr}(l,d)b^l}{n^{l/2}},$$

where $c_{mtr}(l,d)$ is a constant depending on l and d only.

Utilizing the aforementioned Lemmas 1 and 2, we are able to derive the key results that are crucial for our subsequent proofs. These results are encapsulated in the following lemma.

Lemma 3. *Suppose that Assumptions 2 and 3 are satisfied. Then, for $m=1,2,\ldots,M$, $k=1,2\ldots,K$, we have the following inequalities*

$$E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2^l \leq \frac{c_{vct}(l,d)c_3^l}{n^{l/2}}, \quad for \quad l=2,4,8, \tag{S1}$$

and

$$E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|^l \leq \frac{c_{mtr}(l,d)c_4^l}{n^{l/2}}, \quad for \quad l=2,4,8. \tag{S2}$$

Furthermore, the Hessian matrix corresponding to the mth candidate model and data on the kth machine satisfies

$$\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right\| = O_p(1) \tag{S3}$$

and

$$\left\| \left(\frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right)^{-1} \right\| = O_p(1). \tag{S4}$$

Moreover, if the number of machines satisfies $K = o(n^4)$, it follows that

$$\max_{1 \le k \le K} \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right\| = O_p(1) \tag{S5}$$

and

$$\max_{1 \le k \le K} \left\| \left(\frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right)^{-1} \right\| = O_p(1). \tag{S6}$$

Proof of Lemma 3: We begin by expressing $n^{-1} \mathbf{H}_k^{\mathrm{T}} \boldsymbol{\varepsilon}_k$ as

$$\frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_{k,i} \boldsymbol{h}_{k,i} - E\left(\varepsilon_{k,i} \boldsymbol{h}_{k,i} \right) \right].$$

From Lemma 1 and Assumption 3, we could obtain that

$$E \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \boldsymbol{\varepsilon}_k \right\|_2^l \le E \left\| \frac{1}{n} \mathbf{H}_k^{\mathrm{T}} \boldsymbol{\varepsilon}_k \right\|_2^l \le \frac{c_{vct}(l,d) c_3^l}{n^{l/2}},$$

for $l = 2, 4, 8$. Similarly, it is easy to see that

$$\frac{1}{n} \mathbf{H}_k^{\mathrm{T}} \mathbf{H}_k - \mathbf{Q} = \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{h}_{k,i} \boldsymbol{h}_{k,i}^{\mathrm{T}} - E\left(\boldsymbol{h}_{k,i} \boldsymbol{h}_{k,i}^{\mathrm{T}} \right) \right].$$

Then, we could get

$$E \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\|^l \le E \left\| \frac{1}{n} \mathbf{H}_k^{\mathrm{T}} \mathbf{H}_k - \mathbf{Q} \right\|^l \le \frac{c_{mtr}(l,d) c_4^l}{n^{l/2}},$$

for $l = 2, 4, 8$.

For any $t > 0$, according to Markov's inequality, we have that

$$\Pr\left\{ \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\| \ge t \right\} \le \frac{E \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\|^8}{t^8} \le \frac{c_{mtr}(8,d) c_4^8}{n^4 t^8}.$$

Let $t = \sqrt[8]{c_{mtr}(8,d)c_4^8/n^4\tau}$ with τ being a positive constant. Then, for any $0 < \tau < 1$, with probability at least $1-\tau$, it holds that

$$\left\| \frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\| \leq \sqrt[8]{\frac{c_{mtr}(8,d)c_4^8}{n^4\tau}}.$$

Moreover, as the sample size of data stored on each machine n tends to infinity, there exists an integer J_1 such that for all $n \geq J_1$,

$$\sqrt[8]{c_{mtr}(8,d)c_4^8/n^4\tau} < \underline{\lambda}_{(m)}.$$

Therefore, according to Assumption 2, it follows that

$$0 < -\sqrt[8]{\frac{c_{mtr}(8,d)c_4^8}{n^4\tau}} + \underline{\lambda}_{(m)} \preceq \frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} \preceq \sqrt[8]{\frac{c_{mtr}(8,d)c_4^8}{n^4\tau}} + \overline{\lambda}_{(m)}.$$

Consequently,

$$\left[\sqrt[8]{\frac{c_{mtr}(8,d)c_4^8}{n^4\tau}} + \overline{\lambda}_{(m)}\right]^{-1} \preceq \left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1} \preceq \left[-\sqrt[8]{\frac{c_{mtr}(8,d)c_4^8}{n^4\tau}} + \underline{\lambda}_{(m)}\right]^{-1}.$$

We then could conclude that

$$\left\| n^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} \right\| = O_p(1) \quad \text{and} \quad \left\| \left(n^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1} \right\| = O_p(1).$$

By the union bound, for any $\tau > 0$, with probability at least $1-\tau$,

$$\max_{1\leq k\leq K} \left\| \frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\| \leq \sqrt[8]{\frac{c_{mtr}(8,d)c_4^8 K}{n^4\tau}}.$$

Since $K = o(n^4)$, there exist an integer J_2 such that for all $n \geq J_2$,

$$\sqrt[8]{c_{mtr}(8,d)c_4^8 K/n^4\tau} < \underline{\lambda}_{(m)}.$$

Thus, we have that

$$0 < -\sqrt[8]{\frac{c_{mtr}(8,d)c_4^8 K}{n^4\tau}} + \underline{\lambda}_{(m)} \leq \min_{1\leq k\leq K} \lambda_{\min}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right) \leq \max_{1\leq k\leq K} \lambda_{\max}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right) \leq \sqrt[8]{\frac{c_{mtr}(8,d)c_4^8 K}{n^4\tau}} + \overline{\lambda}_{(m)}.$$

It immediately follows that

$$\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right\| = O_p(1) \quad \text{and} \quad \max_{1\leq k\leq K}\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\| = O_p(1).$$

This completes the proof of Lemma 3.

□

Lemma 4. *Let* $\mathbf{A}$ *be a positive definite matrix with minimal eigenvalue being* λ*, and* $\{\mathbf{\Delta}_k\}_{k=1}^K$ *be a collection of matrices of the same size satisfying* $\max_{1\leq k\leq K}\|\mathbf{\Delta}_k\| \leq \psi$ *with* $\psi < \lambda$*. If* $\sum_{k=1}^K \mathbf{\Delta}_k = 0$*, it holds that*

$$\left\|\left[\frac{1}{K}\sum_{k=1}^K (\mathbf{A}+\mathbf{\Delta}_k)^{-1} - \mathbf{A}^{-1}\right]\mathbf{A}\right\| \leq \frac{\psi^2}{\lambda(\lambda-\psi)}.$$

Proof of Lemma 4: Firstly, we rewrite

$$\left\|\left[K^{-1}\sum_{k=1}^K (\mathbf{A}+\mathbf{\Delta}_k)^{-1} - \mathbf{A}^{-1}\right]\mathbf{A}\right\| = \left\|K^{-1}\sum_{k=1}^K (\mathbf{A}+\mathbf{\Delta}_k)^{-1}\mathbf{A} - \mathbf{I}\right\|.$$

For any $1 \leq k \leq K$, we have

$$(\mathbf{A}+\mathbf{\Delta}_k)^{-1}\mathbf{A} = \left[\mathbf{A}\left(\mathbf{I} + A^{-1}\mathbf{\Delta}_k\right)\right]^{-1}\mathbf{A} = \left(\mathbf{I} + \mathbf{A}^{-1}\mathbf{\Delta}_k\right)^{-1}.$$

Since

$$\left\|\mathbf{A}^{-1}\mathbf{\Delta}_k\right\| \leq \left\|\mathbf{A}^{-1}\right\|\left\|\mathbf{\Delta}_k\right\| \leq \frac{\psi}{\lambda} < 1,$$

the following Neumann series expansion holds

$$(\mathbf{I}+\mathbf{C})^{-1}=\sum_{r=0}^{\infty}(-1)^r\mathbf{C}^r,$$

for any $\|\mathbf{C}\|<1$. Setting $\mathbf{C}=\mathbf{A}^{-1}\boldsymbol{\Delta}_k$ in the expression above, we obtain that

$$(\mathbf{A}+\boldsymbol{\Delta}_k)^{-1}\mathbf{A}=\sum_{r=0}^{\infty}(-1)^r\left(\mathbf{A}^{-1}\boldsymbol{\Delta}_k\right)^r=\mathbf{I}-\mathbf{A}^{-1}\boldsymbol{\Delta}_k+\sum_{r=2}^{\infty}(-1)^r\left(\mathbf{A}^{-1}\boldsymbol{\Delta}_k\right)^r.$$

Under the condition $\sum_{k=1}^{K}\boldsymbol{\Delta}_k=0$, we achieve that

$$\frac{1}{K}\sum_{k=1}^{K}(\mathbf{A}+\boldsymbol{\Delta}_k)^{-1}\mathbf{A}-\mathbf{I}=\sum_{r=2}^{\infty}\frac{(-1)^r}{K}\sum_{k=1}^{K}\left(\mathbf{A}^{-1}\boldsymbol{\Delta}_k\right)^r.$$

According to the triangle inequality and the convexity of the operator norm, we could get that

$$\begin{aligned}&\left\|\sum_{r=2}^{\infty}\frac{(-1)^r}{K}\sum_{k=1}^{K}\left(\mathbf{A}^{-1}\boldsymbol{\Delta}_k\right)^r\right\|\\ \leq\ &\sum_{r=2}^{\infty}\frac{1}{K}\sum_{k=1}^{K}\left\|\left(\mathbf{A}^{-1}\boldsymbol{\Delta}_k\right)^r\right\|\\ \leq\ &\sum_{r=2}^{\infty}\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{A}^{-1}\right\|^r\left\|\boldsymbol{\Delta}_k\right\|^r\\ \leq\ &\sum_{r=2}^{\infty}\frac{\psi}{\lambda}\leq\frac{\psi^2}{\lambda(\lambda-\psi)}.\end{aligned}$$

Hence, the desired inequality follows.

□

Proof of Theorem 1:

Firstly, we establish the asymptotic distribution of the estimator $\hat{\boldsymbol{\theta}}_{1,(m)}$. According to the definitions of $\boldsymbol{\Pi}_{(m)}$, $\mathbf{S}_0$ and $\mathbf{S}_{(m)}$, we could obtain the following decomposition

$$\begin{aligned}
&\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{1(m)}-\boldsymbol{\theta}_{0(m)}\right)\\
=\;& \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left[\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{y}_k-\boldsymbol{\theta}_{0(m)}\right]\\
=\;& \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left[\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\mathbf{H}_{k(m)}^{\mathrm{T}}\left(\mathbf{H}_k\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\gamma}_0+\boldsymbol{\varepsilon}_k\right)\right]\\
=\;& \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right)\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\gamma}_0\right]\\
&+\frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right)\right]\\
\triangleq\;& T_1+T_2+T_3+T_4, \qquad\qquad (S7)
\end{aligned}$$

where

$$T_1=\frac{1}{K}\sum_{k=1}^{K}\left\{\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}-\mathbf{Q}_{(m)}^{-1}\right]\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right)\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta}\right\},$$

$$T_2=\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta},$$

$$T_3=\frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left\{\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}-\mathbf{Q}_{(m)}^{-1}\right]\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right)\right\},$$

and

$$T_4=\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{\sqrt{N}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}\right).$$

We proceed to analyze the asymptotic behavior of T_1 to T_4.

Initially, we show that $T_1 = o_p(1)$. By the c_r inequality, and the product inequality, we obtain

$$
\begin{aligned}
& \|T_1\|_2 \\
\leq\ & \frac{1}{K}\sum_{k=1}^{K}\left[\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right\|\|\boldsymbol{\delta}\|_2\right] \\
\leq\ & \max_{1\leq k\leq K}\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\| \\
& \times\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right\|\|\boldsymbol{\delta}\|_2\right].
\end{aligned}
$$

By the c_r inequality, Cauchy-Schwarz inequality and Inequality (S2), we could derive that

$$
\begin{aligned}
& E\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right\|\|\boldsymbol{\delta}\|_2\right] \\
=\ & E\left[\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1\right\|\|\boldsymbol{\delta}\|_2\right] \\
\leq\ & \underline{\lambda}_{(m)}^{-1}qc_2\left\{E\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)} - \mathbf{Q}_{(m)}\right\|^2\right\}^{1/2}\left\{E\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1\right\|^2\right\}^{1/2} \\
\leq\ & \underline{\lambda}_{(m)}^{-1}\left\{E\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)} - \mathbf{Q}_{(m)}\right\|^2\right\}^{1/2}\left\{2E\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1 - \mathbf{Q}\right\|^2 + 2\|\mathbf{Q}\|^2\right\}^{1/2} \\
=\ & O(n^{-1/2}).
\end{aligned}
$$

Hence,

$$
\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right\|\|\boldsymbol{\delta}\|_2 = o_p(1),
$$

which implies that

$$
T_1 = o_p(1), \tag{S8}
$$

according to (S6).

For T_2, by the weak law of large numbers, we obtain

$$\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H} \xrightarrow{p} \mathbf{Q}.$$

Hence,

$$T_2 \xrightarrow{p} \mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)}\right)\boldsymbol{\delta}. \tag{S9}$$

Next, we show that $T_3 = o_p(1)$. Similarly to the analysis of T_1, we have that

$$\begin{aligned}
&\left\|\frac{\sqrt{N}}{K}\sum_{k=1}^{K}\left\{\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1} - \mathbf{Q}_{(m)}^{-1}\right]\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right)\right\}\right\|_2 \\
\leq\ & \sqrt{\frac{n}{K}}\sum_{k=1}^{K}\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2 \\
\leq\ & \max_{1\leq k\leq K}\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\| \\
& \left[\sqrt{\frac{n}{K}}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2\right].
\end{aligned}$$

Therein,

$$\begin{aligned}
& E\left[\sqrt{\frac{n}{K}}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2\right] \\
=\ & \sqrt{N}E\left[\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2\right] \\
\leq\ & \underline{\lambda}_{(m)}^{-1}\sqrt{N}\left\{E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|^2\right\}^{1/2}\left\{E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2^2\right\}^{1/2} \\
=\ & O\left(\sqrt{\frac{K}{n}}\right).
\end{aligned}$$

Hence,

$$\sqrt{\frac{n}{K}}\sum_{k=1}^{K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}-\mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\|\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right\|_2=O\left(\sqrt{\frac{K}{n}}\right),$$

which implies that

$$T_3=o_p(1), \tag{S10}$$

according to (S6) and the condition $K=o(n)$,

For T_4, by the central limit theorem, we know that

$$\frac{1}{\sqrt{N}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}\xrightarrow{d}\boldsymbol{\zeta}.$$

Thus,

$$T_4\xrightarrow{d}\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\zeta}. \tag{S11}$$

From (S8)-(S11), applying Slutsky's theorem, we obtain

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

Thus, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{1(m)}$ is established.

Next, we establish an upper bound for $E\left\|\hat{\boldsymbol{\theta}}_{1(m)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2$. Define event

$$\mathrm{E}_{(m)}\triangleq\left\{\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|\le\kappa\underline{\lambda}_{(m)}\right\}, \tag{S12}$$

with κ being some constant satisfying $0<\kappa<1$, and $\mathrm{E}_{(m)}^c$ denotes the complement of $\mathrm{E}_{(m)}$. Then, it is easy to see that

$$E\left\|\hat{\boldsymbol{\theta}}_{1(m)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2 \;=\; E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2\mathbf{I}_{\mathrm{E}_{(m)}}\right]+E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2\mathbf{I}_{\mathrm{E}_{(m)}^c}\right]$$

$$\triangleq \quad L_1 + L_2, \tag{S13}$$

where

$$L_1 = E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathbf{I}_{\mathrm{E}_{(m)}}\right],$$

and

$$L_2 = E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathbf{I}_{\mathrm{E}^c_{(m)}}\right].$$

Under the event $\mathrm{E}_{(m)}$, the eigenvalues of $n^{-1}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}$ satisfies

$$0 < (1-\kappa)\underline{\lambda}_{(m)} \preceq \frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)} \preceq \kappa\underline{\lambda}_{(m)} + \overline{\lambda}_{(m)},$$

which implies that

$$\left\|\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\right\| \leq [(1-\kappa)\underline{\lambda}_{(m)}]^{-1}. \tag{S14}$$

We first provide an upper bound for L_1. From (S7), we could derive that

$$E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right] \leq 4\left(L_{11} + L_{12} + L_{13} + L_{14}\right), \tag{S15}$$

where

$$L_{11} = E\left\|\frac{1}{K}\sum_{k=1}^{K}\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1} - \mathbf{Q}_{(m)}^{-1}\right]\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right)\mathbf{S}_0\left(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)}\right)\boldsymbol{\gamma}\right\|_2^2,$$

$$L_{12} = E\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\right)\mathbf{S}_0\left(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)}\right)\boldsymbol{\gamma}\right\|_2^2,$$

$$L_{13} = E\left\|\frac{1}{K}\sum_{k=1}^{K}\left\{\left[\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1} - \mathbf{Q}_{(m)}^{-1}\right]\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}_k\right)\right\}\right\|_2^2,$$

and

$$L_{14} = E\left\|\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)\right\|_2^2.$$

We handle L_{11} to L_{14} separately under $\mathrm{E}_{(m)}$. According to the c_r and Cauchy–Schwarz inequalities, (S2) and (S14), under Assumptions 1 and 2, we could get that

$$\begin{aligned}
& L_{11} \\
\leq\ & E\left\|\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\right\|^2\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|^2\left\|\mathbf{Q}_{(m)}^{-1}\right\|^2\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1\right\|^2\|\boldsymbol{\gamma}\|_2^2 \\
\leq\ & \frac{[(1-\kappa)\underline{\lambda}_{(m)}]^{-2}\underline{\lambda}_{(m)}^{-2}qc_2^2}{N}\left\{E\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|^4\right\}^{1/2}\left\{E\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1\right\|^4\right\}^{1/2} \\
=\ & O\left(\frac{1}{nN}\right),
\end{aligned}$$

$$\begin{aligned}
& L_{12} \\
\leq\ & E\left\|\mathbf{Q}_{(m)}^{-1}\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1\right)\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\gamma}\right\|_2^2 \\
\leq\ & \frac{2\underline{\lambda}_{(m)}^{-2}qc_2^2}{N}E\left[\left\|\frac{1}{n}\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1-\mathbf{Q}\right\|^2+\|\mathbf{Q}\|^2\right] \\
\leq\ & \frac{2\underline{\lambda}_{(m)}^{-2}\overline{\lambda}_{(m)}^{-2}qc_2^2}{N}+O\left(\frac{1}{nN}\right),
\end{aligned}$$

$$\begin{aligned}
& L_{13} \\
\leq\ & E\left\|\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\right\|^2\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|^2\left\|\mathbf{Q}_{(m)}^{-1}\right\|^2\left\|\frac{1}{n}\mathbf{H}_{m,1}^{\mathrm{T}}\boldsymbol{\varepsilon}_1\right\|_2^2 \\
\leq\ & [(1-\kappa)\underline{\lambda}_{(m)}]^{-2}\underline{\lambda}_{(m)}^{-2}\left\{E\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|^4\right\}^{1/2}\left\{E\left\|\frac{1}{n}\mathbf{H}_{m,1}^{\mathrm{T}}\boldsymbol{\varepsilon}_1\right\|_2^4\right\}^{1/2}
\end{aligned}$$

$$= O\left(\frac{1}{n^2}\right),$$

and

$$L_{14} \leq \left\|\mathbf{Q}_{(m)}^{-1}\right\|^2 E\left\|\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)\right\|_2^2 \leq \frac{\underline{\lambda}_{(m)}^{-2} c_{vct}(2,d) c_3^2}{N}.$$

Hence,

$$L_1 \leq \frac{2\underline{\lambda}_{(m)}^{-2}\overline{\lambda}_{(m)}^{-2} q c_2^2 + \underline{\lambda}_{(m)}^{-2} c_{vct}(2,d) c_3^2}{N} + O\left(\frac{1}{n^2}\right). \tag{S16}$$

Now, let's deal with L_2. From (S2), Assumptions 1 and 2 and Markov's inequality, we have

$$\begin{aligned}
& L_2 \\
= \;& E\left[\left\|\hat{\boldsymbol{\theta}}_{1(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathbf{I}_{\mathrm{E}_{(m)}^c}\right] \\
\leq \;& \left(pc_1^2 + \frac{q_{(m)} c_2^2}{N}\right) \Pr\left(\mathrm{E}_{(m)}^c\right) \\
\leq \;& \left(pc_1^2 + \frac{q_{(m)} c_2^2}{N}\right) \frac{E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|^8}{(\kappa\underline{\lambda}_{(m)})^8} \\
= \;& O\left(\frac{1}{n^4}\right).
\end{aligned} \tag{S17}$$

From (S13), (S16) and (S17), we could finally conclude that

$$E\left\|\hat{\boldsymbol{\theta}}_{1(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \leq \frac{4\underline{\lambda}_{(m)}^{-2}\left(2qc_2^2\overline{\lambda}_{(m)}^{-2} + c_{vct}(2,d)c_3^2\right)}{N} + O\left(\frac{1}{n^2}\right).$$

This completes the proof of Theorem 1.

□

Proof of Theorem 2:

We firstly establish the convergence rate of $\hat{\boldsymbol{\theta}}_{2(m)}^{(t+1)}$. According to the definitions of $\boldsymbol{\Pi}_{(m)}$, $\mathbf{S}_0$, and $\mathbf{S}_{(m)}$, we could obtain the following decomposition

$$
\begin{aligned}
&\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)} \\
= \; & \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} + \left(\mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)}\right)^{-1} \left(\frac{1}{K} \sum_{k=1}^{K} \mathbf{H}_{k(m)}^{\mathrm{T}} \boldsymbol{y}_k - \frac{1}{K} \sum_{k=1}^{K} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} \right) \\
= \; & \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} + \left(\mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)}\right)^{-1} \left[\frac{1}{K} \sum_{k=1}^{K} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \left(\boldsymbol{\theta}_{0(m)} - \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} \right) \right. \\
& \left. + \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{H}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{S}_0 \left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)} \right) \boldsymbol{\gamma} + \frac{1}{K} \sum_{k=1}^{K} \mathbf{H}_{k(m)}^{\mathrm{T}} \boldsymbol{\varepsilon}_k \right] \\
= \; & \left(\frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} \right)^{-1} \left\{ \left[\frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} - \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \mathbf{H}_{(m)} \right] \left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right) \right. \\
& \left. + \mathbf{S}_{(m)}^{\mathrm{T}} \left(\frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} \right) \mathbf{S}_0 \left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)} \right) \boldsymbol{\gamma} + \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\}. \qquad \text{(S18)}
\end{aligned}
$$

From the product and c_r inequalities, we achieve that

$$
\begin{aligned}
&\left\| \hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)} \right\|_2 \\
\leq \; & \left\| \left(\frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} \right)^{-1} \right\| \left\{ \left\| \frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} - \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \mathbf{H}_{(m)} \right\| \left\| \left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right) \right\|_2 \right. \\
& \left. + \left\| \frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} \right\| \|\boldsymbol{\gamma}\|_2 + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\|_2 \right\} \\
\leq \; & \left\| \left(\frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} \right)^{-1} \right\| \left\{ \left[\left\| \frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} - \mathbf{Q}_{(m)} \right\| + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \mathbf{H}_{(m)} - \mathbf{Q}_{(m)} \right\| \right] \right. \\
& \left. \times \left\| \left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right) \right\|_2 + \left[\left\| \frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} - \mathbf{Q} \right\| + \|\mathbf{Q}\| \right] \|\boldsymbol{\gamma}\|_2 + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\|_2 \right\}. \\
& \qquad \text{(S19)}
\end{aligned}
$$

Under Assumptions 1 and 2, and invoking (S1) to (S4), we could conclude that

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

which establishes the convergence rate of $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)}$.

Next, we establish the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)}$. Under Assumptions 1 and 2, and from (S1) to (S4) and (S18), we could deduce that

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)}\right) = O_p\left\{\sqrt{K}\left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)}\right)\right\} + T_1 + T_2, \tag{S20}$$

where

$$T_1 = \left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\left[\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta}\right],$$

and

$$T_2 = \left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\left(\frac{1}{\sqrt{N}}\mathbf{H}_{(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}\right).$$

By the law of large numbers, the central limit theorem, and Slutsky's theorem, it is straightforward to show that

$$T_1 \xrightarrow{p} \mathbf{A}_{(m)}\boldsymbol{\delta}, \quad T_2 \xrightarrow{d} \mathbf{B}_{(m)}\boldsymbol{\zeta}.$$

Furthermore, from (S20) and the Slutsky's theorem, under the condition that $\sqrt{K}\left(\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)}\right) = o_p(1)$, it follows that

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

This completes the derivation of the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)}$.

In the following, let's establish the upper bound for $E\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2$. Following the strategy analogous to the proof of Theorem 1, we decompose the target expectation as follows

$$\begin{aligned} & E\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \\ = \ & E\left[\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathrm{I}_{\mathrm{E}_{(\mathrm{m})}}\right] + E\left[\left\|\hat{\boldsymbol{\theta}}^{(t+1)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathrm{I}_{\mathrm{E}^{\mathrm{c}}_{(\mathrm{m})}}\right], \\ \triangleq \ & L_1 + L_2, \end{aligned} \tag{S21}$$

where the event $\mathrm{E}_{(m)}$ is defined in (S12).

We begin by deriving an upper bound for L_1. It is noted that under the event $\mathrm{E}_{(m)}$, inequality (S14) holds. from (S19), we derive

$$L_1 \leq 3\left[(1-\kappa)\underline{\lambda}_{(m)}\right]^{-2}(L_{11} + L_{12} + L_{13}),$$

where

$$\begin{aligned} & L_{11} \\ = \ & E\left[\left(\left\|\frac{1}{n}\mathbf{H}^{\mathrm{T}}_{1(m)}\mathbf{H}_{1(m)} - \mathbf{Q}_{(m)}\right\| + \left\|\frac{1}{N}\mathbf{H}^{\mathrm{T}}_{(m)}\mathbf{H}_{(m)} - \mathbf{Q}_{(m)}\right\|\right)\left\|\hat{\boldsymbol{\theta}}^{(t)}_{2,(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right]^2, \end{aligned}$$

$$L_{12} = E\left\|\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right\|^2 \|\boldsymbol{\gamma}\|_2^2,$$

and

$$L_{13} = E\left\|\frac{1}{N}\mathbf{H}^{\mathrm{T}}_{(m)}\boldsymbol{\varepsilon}\right\|_2^2.$$

We cope with each term separately. According to Assumptions 1 and 2, (S2), the Cauchy–Schwarz and the c_r inequalities, we could get that

$$\begin{aligned} L_{11} &\leq \left\{ E \left[\left\| \frac{1}{n} \mathbf{H}_{1(m)}^{\mathrm{T}} \mathbf{H}_{1(m)} - Q_m \right\| + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \mathbf{H}_{(m)} - \mathbf{Q}_{(m)} \right\| \right]^4 \right\}^{1/2} \\ &\quad \left\{ E \left\| \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right\|_2^4 \right\}^{1/2} \\ &= O\left(\frac{1}{n^2} E \left\| \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right\|_2^2 \right), \end{aligned}$$

$$\begin{aligned} L_{12} &\leq \frac{2qc_2^2}{N} \left(E \left\| \frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} - \mathbf{Q} \right\|^2 + \|\mathbf{Q}\|^2 \right) \\ &\leq \frac{2qc_2^2 \overline{\lambda}_{(m)}^2}{N} + O\left(\frac{1}{N^2} \right), \end{aligned}$$

and

$$L_{13} \leq \frac{c_{vct}(2,d) c_3^2}{N}.$$

Hence,

$$L_1 \leq \frac{2qc^2 \overline{\lambda}_{(m)}^2 + c_{vct}(2,d) c_3^2}{N} + O\left(\frac{1}{n^2} E \left\| \hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right\|_2^2 \right) + O\left(\frac{1}{N^2} \right). \tag{S22}$$

Now, let's turn to L_2. Based on (S2) and Markov's inequality, under Assumptions 1 and 2, we could arrive at

$$\begin{aligned} L_2 &\leq \left(pc_1^2 + \frac{q_{(m)} c_2^2}{N} \right) \Pr\left(\mathrm{E}_{(m)}^c \right) \\ &\leq \left(pc_1^2 + \frac{q_{(m)} c_2^2}{N} \right) \frac{E \left\| \frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} - \mathbf{Q}_{(m)} \right\|^8}{(\kappa \underline{\lambda}_{(m)})^8} \end{aligned}$$

$$= O\left(\frac{1}{n^4}\right). \tag{S23}$$

From (S21) to (S23), we could draw the conclusion that

$$\begin{aligned} & E\left\|\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \\ \leq & \frac{3(1-\kappa)^{-2}\underline{\lambda}_{(m)}^{-2}\left(2qc_2^2\overline{\lambda}_{(m)}^{-2} + c_{vct}(2,d)c_3^2\right)}{N} + O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}_{2,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right) + O\left(\frac{1}{n^4}\right). \end{aligned}$$

This completes the proof of Theorem 2.

□

Before the formal proof of Theorem 3, we firstly state a lemma that will be used in the subsequent proof.

Lemma 5. *Suppose that Assumptions 2 and 3 hold, and the number of machines satisfies $K = o(n^4)$, then, we have*

$$\left\|\mathbf{I}_{d_m} - \left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right\| = O_p\left(\frac{K^{1/4}}{n}\right), \tag{S24}$$

for $m = 1, 2, \ldots, M$.

Proof of Lemma 5: Under Assumption 3, it is easy to see that (S2) hold for $m = 1, 2, \ldots, M$. Therefore, for any $0 < \tau < 1$, with probability at least $1 - \tau$, we have

$$\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\| \leq \psi_n(d, \tau), \quad k = 1, 2, \ldots, K, \tag{S25}$$

where $\psi_n(d,\tau) = \sqrt[8]{c_{mtr}(8,d)/n^4\tau}$. Applying a union bound over $k = 1,2,\ldots,K$, we obtain

$$\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\| \leq K^{1/8}\psi_n(d,\tau). \tag{S26}$$

Consequently,

$$\begin{aligned}
&\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right\| \\
\leq\ &\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\| + \left\|\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right) - \mathbf{Q}_{(m)}\right\| \\
\leq\ &2\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\| \leq 2K^{1/8}\psi_n(m,\tau).
\end{aligned} \tag{S27}$$

Since $K = o(n^4)$, under Asssumption 2, there exist an integer J_3 such that for all $n \geq J_3$, it holds that

$$0 < \underline{\lambda}_{(m)} - \psi_n(m,\tau) \preceq \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right) \preceq \overline{\lambda}_{(m)} + \psi_n(m,\tau), \tag{S28}$$

and

$$2K^{1/8}\psi_n(m,\tau) < \underline{\lambda}_{(m)} - \psi_n(m,\tau). \tag{S29}$$

From (S26) to (S28) and Lemma 4, it follows that for all $n \geq J_3$, with probability at least $1-\tau$,

$$\begin{aligned}
&\left\|\mathbf{I}_{d_m} - \left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right\| \\
\leq\ &\frac{4K^{1/4}\psi_n(m,\tau)^2}{\left(\underline{\lambda}_{(m)} - \psi_n(m,\tau)\right)\left[\underline{\lambda}_{(m)} - (1+2K^{1/8})\psi_n(m,\tau)\right]}.
\end{aligned}$$

This implies

$$\left\|\mathbf{I}_{d_m}-\left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right\|=O_p\left(\frac{K^{1/4}}{n}\right),$$

which completes the proof.

□

Proof of Theorem 3: We first investigate the convergence rate of $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}$. According to the definitions of $\boldsymbol{\Pi}_{(m)}, \mathbf{S}_0$ and $\mathbf{S}_{(m)}$, we could obtain the following decomposition

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}-\boldsymbol{\theta}_{0(m)}\\
=\;&\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}+\left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\left(\boldsymbol{\theta}_{0(m)}-\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}\right)\right.\\
&\left.+\frac{1}{K}\sum_{k=1}^{K}\mathbf{S}_{(m)}^{\mathrm{T}}\mathbf{H}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\gamma}+\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{k(m)}\boldsymbol{\varepsilon}_k\right]\\
=\;&\left\{\mathbf{I}_{d_m}-\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right]\right\}\left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right)\\
&+\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_{(m)}^{\mathrm{T}}\boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\gamma}+\frac{1}{N}\mathbf{H}_{(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}\right].
\end{aligned} \tag{S30}$$

From (S24), we can deduce that

$$\begin{aligned}
&\left\|\left\{\mathbf{I}_{d_m}-\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right]\right\}\left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right)\right\|_2\\
=\;&O_p\left(\frac{K^{1/4}}{n}\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right\|_2\right).
\end{aligned} \tag{S31}$$

In addition, from (S6), we have

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right)^{-1} \right\| \leq \max_{1 \leq k \leq K} \left\| \left(\frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right)^{-1} \right\| = O_p(1). \tag{S32}$$

Under Assumptions 1 and 2, (S1) and (S2), it holds that

$$\begin{aligned} & \left\| \mathbf{S}_{(m)}^{\mathrm{T}} \left(\frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} \right) \mathbf{S}_0 \left(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}} \mathbf{\Pi}_{(m)} \right) \boldsymbol{\gamma} + \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\|_2 \\ \leq & \left\| \frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} \right\| \|\boldsymbol{\gamma}\|_2 + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\|_2 \\ \leq & \left(\left\| \frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} - \mathbf{Q} \right\| + \|\mathbf{Q}\| \right) \|\boldsymbol{\gamma}\|_2 + \left\| \frac{1}{N} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right\|_2 \\ = & O_p \left(\frac{1}{\sqrt{N}} \right). \end{aligned} \tag{S33}$$

Hence, from (S30) to (S33), we could conclude that

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

which establishes the desired convergence rate.

We now derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}$. From (S24) and (S30), it follows that

$$\begin{aligned} & \sqrt{N} \left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)} \right) \\ = & \left[\frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{n} \mathbf{H}_{k(m)}^{\mathrm{T}} \mathbf{H}_{k(m)} \right)^{-1} \right] \left[\mathbf{S}_{(m)}^{\mathrm{T}} \left(\frac{1}{N} \mathbf{H}^{\mathrm{T}} \mathbf{H} \right) \mathbf{S}_0 \left(\mathbf{I}_q - \mathbf{\Pi}_{(m)}^{\mathrm{T}} \mathbf{\Pi}_{(m)} \right) \boldsymbol{\delta} + \frac{1}{\sqrt{N}} \mathbf{H}_{(m)}^{\mathrm{T}} \boldsymbol{\varepsilon} \right] \\ + & O_p \left\{ \frac{K^{3/4}}{\sqrt{n}} \left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)} \right) \right\}. \end{aligned} \tag{S34}$$

From (S6) and (S26), we could obtain that

$$\begin{aligned}
&\left\|\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}-\mathbf{Q}_{(m)}^{-1}\right\| \\
\leq\ & \max_{1\leq k\leq K}\left\|\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right\|\max_{1\leq k\leq K}\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}-\mathbf{Q}_{(m)}\right\|\left\|\mathbf{Q}_{(m)}^{-1}\right\| \\
=\ & O_p(K^{1/8}n^{-1/2}).
\end{aligned}$$

Thus, under $K=o(n^4)$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\xrightarrow{p}\mathbf{Q}_{(m)}^{-1}. \tag{S35}$$

Furthermore, by the law of large numbers and central limit theorem, we have

$$N^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}\xrightarrow{p}\mathbf{Q}, \tag{S36}$$

and

$$N^{-1/2}\mathbf{H}_{(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}\xrightarrow{d}\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\zeta}. \tag{S37}$$

According to the Slutsky's theorem, (S34) to (S37), under the condition $K^{3/4}n^{-1/2}\left(\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right)=o_p(1)$, we could arrive at

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

which completes the proof of the asymptotic distribution.

In the next step, we consider the mean squared error of the estimator $\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}$, the proof of which follows a structure analogous to those for $\hat{\boldsymbol{\theta}}_{2,(m)}^{(t+1)}$

and $\hat{\boldsymbol{\theta}}_{1,(m)}$. We begin by decomposing the mean squared error as

$$\begin{aligned} & E\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2 \\ = \;& E\left[\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathbf{I}_{\mathrm{E}_{(m)}}\right]+E\left[\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2 \mathbf{I}_{\mathrm{E}_{(m)}^c}\right] \\ \triangleq \;& L_1+L_2. \end{aligned} \tag{S38}$$

where $E_{(m)}$ is defined in (S13).

Let's cope with L_1 in (S38). It is easy to see that inequality (S14) holds under the event $E_{(m)}$. Then, according to (S2), we could derive that

$$\begin{aligned} & E\left\|\mathbf{I}_{d_m}-\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right]\right\|^4 \\ \leq \;& E\left\|\mathbf{I}_{d_{(m)}}-\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right]\right\|^4 \\ \leq \;& E\left\|\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\right\|^4\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\left[\frac{1}{N}\mathbf{H}_{(m)}^{\mathrm{T}}\mathbf{H}_{(m)}\right]\right\|^4 \\ \leq \;& 4[(1-\kappa)\underline{\lambda}_{(m)}]^{-4}\left\{E\left\|\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}-\mathbf{Q}_{(m)}\right\|^4+E\left\|\frac{1}{N}\mathbf{H}_{(m)}^{\mathrm{T}}\mathbf{H}_{(m)}-\mathbf{Q}_{(m)}\right\|^4\right\} \\ = \;& O\left(\frac{1}{n^2}\right). \end{aligned} \tag{S39}$$

From (S30) and (S39), it is easy to see that

$$\begin{aligned} & L_1 \\ \leq \;& 2E\left\{\left\|\mathbf{I}_{d_m}-\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)^{-1}\right]\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)}\right)\right]\right\|^2\right. \\ & \left.\times\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right\|_2^2\right\} \end{aligned}$$

$$
\begin{aligned}
&+2E\left\|\left(\frac{1}{n}\mathbf{H}_{1(m)}^{\mathrm{T}}\mathbf{H}_{1(m)}\right)^{-1}\right\|^{2}\left\|\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\mathbf{S}_{0}\left(\mathbf{I}_{q}-\mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)}\right)\boldsymbol{\gamma}+\frac{1}{N}\mathbf{H}_{(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_{2}^{2}\\
\leq\ & O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right\|_{2}^{2}\right)+4(1-\kappa)^{-2}\underline{\lambda}_{(m)}^{-2}(L_{11}+L_{12}),
\end{aligned}
$$

where

$$
L_{11}=E\left\|\mathbf{S}_{(m)}^{\mathrm{T}}\left(\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\mathbf{S}_{0}\left(\mathbf{I}_{q}-\mathbf{\Pi}_{(m)}^{\mathrm{T}}\mathbf{\Pi}_{(m)}\right)\boldsymbol{\gamma}\right\|_{2}^{2},
$$

and

$$
L_{12}=E\left\|\frac{1}{N}\mathbf{H}_{(m)}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_{2}^{2}.
$$

According to (S1) and (S2), we could find that

$$
\begin{aligned}
L_{11}\ &\leq\ \frac{2qc_2^2}{N}\left(E\left\|\frac{1}{N}\mathbf{H}^{\mathrm{T}}\mathbf{H}-\mathbf{Q}\right\|^{2}+\|\mathbf{Q}\|^{2}\right)\\
&\leq\ \frac{2\overline{\lambda}_{(m)}^{-2}qc_2^2}{N}+O\left(\frac{1}{N^2}\right),
\end{aligned}
$$

and

$$
L_{12}\leq\frac{c_{vct}(2,d)c_3^2}{N},
$$

Therefore,

$$
\begin{aligned}
&L_1\\
\leq\ & O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)}-\boldsymbol{\theta}_{0(m)}\right\|_{2}^{2}\right)+\frac{4(1-\kappa)^{-2}\underline{\lambda}_{(m)}^{-2}\left(2qc_2^2\overline{\lambda}_{(m)}^{-2}+c_{vct}(2,d)c_3^2\right)}{N}+O\left(\frac{1}{N^2}\right).
\end{aligned}
\tag{S40}
$$

We now proceed to L_2 in (S38). From Lemma 3, under Assumptions 1 and 2, it is easy to see that (S2) is valid. Consequently, according to

Markov's inequality, we could get that

$$\begin{aligned} & L_2 \\ \leq & \left(pc_1^2 + \frac{q_{(m)}c_2^2}{N}\right) \Pr\left(\mathrm{E}_{(m)}^c\right) \\ \leq & \left(pc_1^2 + \frac{q_{(m)}c_2^2}{N}\right) \frac{E\left\|\frac{1}{n}\mathbf{H}_{k(m)}^{\mathrm{T}}\mathbf{H}_{k(m)} - \mathbf{Q}_{(m)}\right\|^8}{(\kappa\underline{\lambda}_{(m)})^8} \\ = & O\left(\frac{1}{n^4}\right). \end{aligned} \tag{S41}$$

Having combined (S38), (S40) and (S41), it follows that

$$\begin{aligned} & E\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t+1)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 \\ \leq & \frac{4(1-\kappa)^{-2}\underline{\lambda}_{(m)}^{-2}\left(2qc_2^2\overline{\lambda}_{(m)}^{-2} + c_{vct}(2,d)c_3^2\right)}{N} + O\left(\frac{1}{n}E\left\|\hat{\boldsymbol{\theta}}_{3,(m)}^{(t)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2\right) + O\left(\frac{1}{n^4}\right). \end{aligned}$$

This completes the proof of Theorem 3.

□

Proof of Theorem 4: We begin by proving the first assertion of this theorem, the proof of which proceeds along lines similar to that in Liu (2015). Define the set

$$\boldsymbol{\gamma}_{(m)}^c = \left\{\boldsymbol{\gamma} : \gamma_j \notin \boldsymbol{\gamma}_{(m)}, j = 1, 2, \ldots, q\right\},$$

which consists of all parameters γ_j excluded from mth candidate model. Accordingly, we may write

$$\mu(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{0(m)}, \boldsymbol{\gamma}_{0(m)}^c), \quad \text{and} \quad \mu(\boldsymbol{\theta}_{0(m)}) = \mu(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{0(m)}, \mathbf{0}).$$

Given that $\|\boldsymbol{\gamma}_0\|_2 = O(N^{-1/2})$, the Taylor expansion of $\mu(\boldsymbol{\theta}_0)$ around $\boldsymbol{\gamma}_{0(m)^c}$ yields

$$\begin{aligned}
\mu(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{0(m)}, \boldsymbol{\gamma}_{0(m)^c}) &= \mu(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{0(m)}, \mathbf{0}) + \boldsymbol{\mu}_{\gamma^c_{0(m)}}^{\mathrm{T}} \gamma^c_{0(m)} + O(N^{-1}) \\
&= \mu(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{0(m)}, \mathbf{0}) + \boldsymbol{\mu}_a^{\mathrm{T}} \left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)}\right) \boldsymbol{\gamma}_0 + O(N^{-1}),
\end{aligned}$$

which implies

$$\mu(\boldsymbol{\theta}_0) - \mu(\mathbf{S}_{(m)}^{\mathrm{T}} \boldsymbol{\theta}_{0(m)}) = \boldsymbol{\mu}_a^{\mathrm{T}} \left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)}\right) \boldsymbol{\gamma}_0 + O(N^{-1}).$$

Then, by the delta method, we could obtain that

$$\begin{aligned}
& \sqrt{N}\left(\hat{\mu}_{(m)} - \mu_0\right) \\
= \;& \sqrt{N}\left(\hat{\mu}(\hat{\boldsymbol{\theta}}_{(m)}) - \mu(\mathbf{S}_{(m)}^{\mathrm{T}} \boldsymbol{\theta}_{0(m)})\right) - \sqrt{N}\left(\mu(\boldsymbol{\theta}_0) - \mu(\mathbf{S}_{(m)}^{\mathrm{T}} \boldsymbol{\theta}_{0(m)})\right) \\
\xrightarrow{d} \;& \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)}(\mathbf{A}_{(m)}\boldsymbol{\delta} + \mathbf{B}_{(m)}\boldsymbol{\zeta}) - \boldsymbol{\mu}_a^{\mathrm{T}}\left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta} \\
= \;& \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)} \mathbf{A}_{(m)}\boldsymbol{\delta} - \boldsymbol{\mu}_a^{\mathrm{T}}\left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta} + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)} \mathbf{B}_{(m)}\boldsymbol{\zeta} \\
= \;& \left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{Q} \mathbf{S}_{(m)}\right)^{-1} \mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{Q} \mathbf{S}_0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_0\right)\left(\mathbf{I}_q - \boldsymbol{\Pi}_{(m)}^{\mathrm{T}} \boldsymbol{\Pi}_{(m)}\right)\boldsymbol{\delta} \\
& + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)}\left(\mathbf{Q}_{(m)}\right)^{-1} \mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\zeta} \\
= \;& \left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{Q} \mathbf{S}_{(m)}\right)^{-1} \mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{Q} \mathbf{S}_0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_0\right)\boldsymbol{\delta} \\
& + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{S}_{(m)}\left(\mathbf{S}_{(m)}^{\mathrm{T}} \mathbf{P}_{(m)} \mathbf{S}_{(m)}\right)^{-1} \mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\zeta} \\
= \;& \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}\left(\mathbf{P}_{(m)}\mathbf{Q} - \mathbf{I}_d\right)\mathbf{S}_0\boldsymbol{\delta} + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{P}_{(m)}\boldsymbol{\zeta} \sim \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).
\end{aligned}$$

Furthermore, let's consider the upper bound of $E\left\|\hat{\mu}_{(m)} - \mu_0\right\|_2^2$. Since the partial derivatives of $\mu(\boldsymbol{\theta}_0)$ are continuous in a neighborhood of $\left(\boldsymbol{\beta}_0^{\mathrm{T}}, \mathbf{0}_{q\times 1}^{\mathrm{T}}\right)^{\mathrm{T}}$, there exists a positive constant c_5 such that

$$\left\| \frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{\theta} = \boldsymbol{\theta}^* \right\|_2^2 \leq \frac{c_5}{2},$$

for all $\boldsymbol{\theta}^*$ in a neighborhood of $\left(\boldsymbol{\beta}_0^{\mathrm{T}}, \mathbf{0}_{q\times 1}^{\mathrm{T}}\right)^{\mathrm{T}}$. Therefore,

$$\begin{aligned} & \left\|\mu(\boldsymbol{\theta}_0) - \mu(\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)})\right\|_2^2 \\ = & \left\|\boldsymbol{\mu}_{\boldsymbol{\theta}_{(m)}^*}^{\mathrm{T}} \left(\boldsymbol{\theta}_0 - \mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)}\right)\right\|_2^2 \\ \leq & \; C \left\|\boldsymbol{\gamma}_{(m)^c}\right\|_2^2 \\ \leq & \; \frac{(q-q_m)c_2^2 c_5}{2N}, \end{aligned} \tag{S42}$$

where $\boldsymbol{\theta}_{(m)}^*$ lies between $\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)}$ and $\boldsymbol{\theta}_0$. Similarly, we could also get

$$\begin{aligned} & \left\|\hat{\mu}(\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)}) - \mu\left(\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)}\right)\right\|_2^2 \\ = & \left\|\boldsymbol{\mu}_{\boldsymbol{\theta}_{(m)}^{**}}^{\mathrm{T}} \mathbf{S}_{(m)}^{\mathrm{T}} \left(\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}\right)\right\|_2^2 \\ \leq & \; \frac{c_5}{2}\left\|\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2, \end{aligned} \tag{S43}$$

with $\boldsymbol{\theta}_{(m)}^{**}$ between $\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)}$ and $\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)}$. From (S42) and (S43), it follows that

$$\begin{aligned} & E\left\|\hat{\mu}_{(m)} - \mu_0\right\|_2^2 \\ \leq & \; 2E\left\|\hat{\mu}(\mathbf{S}_{(m)}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{(m)}) - \mu(\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)})\right\|_2^2 + 2\left\|\mu(\mathbf{S}_{(m)}^{\mathrm{T}}\boldsymbol{\theta}_{0(m)}) - \mu(\boldsymbol{\theta}_0)\right\|_2^2 \\ \leq & \; c_5 E\left\|\hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)}\right\|_2^2 + \frac{c_5(q-q_m)c_2^2}{N}. \end{aligned}$$

This completes the proof of the second part of this theorem.

□

Proof of Theorem 6:

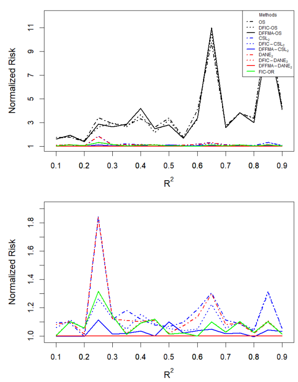
As stated in the main body of our article, the proof of the first two conclusions of this theorem is similar to that in Liu (2015). Therefore, we omit it for simplicity and only need to coupe with the last assertion of Theorem 6, i.e., the upper bound for the mean squared error of the DFFMA estimator with data-driven weights. In fact, according to the c_r inequality and constraints $0 \leq \hat{w}_{(m)} \leq 1$, we could easily obtain the following result

$$
\begin{aligned}
& E \left\| \hat{\mu}\left(\hat{\boldsymbol{w}}\right) - \mu_0 \right\|_2^2 \\
= \; & E \left\| \sum_{m=1}^{M} \hat{w}_{(m)} \hat{\mu}_{(m)} - \mu_0 \right\|_2^2 \\
\leq \; & M \sum_{m=1}^{M} \hat{w}_{(m)}^2 E \left\| \hat{\mu}_{(m)} - \mu_0 \right\|_2^2 \\
\leq \; & c_5 M \max_{1 \leq m \leq M} \left\{ E \left\| \hat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{0(m)} \right\|_2^2 + \frac{(q - q_{(m)}) c_2^2}{N} \right\}.
\end{aligned}
$$

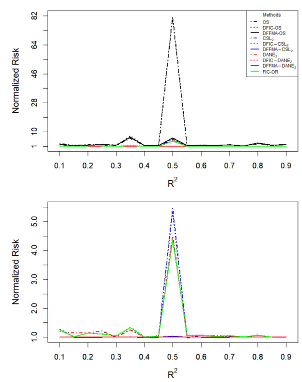
□

S3 Additional simulation studies

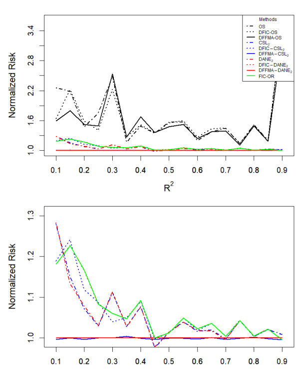
S3.1 Simulation results of the heteroscedastic scenarios of various settings in the main body of this article



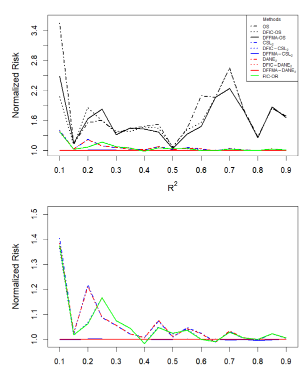
(a) $N = 1 \times 2^{12}$.



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

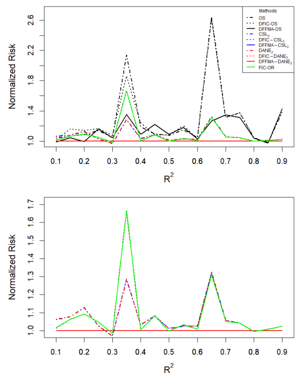


(c) $N = 3 \times 2^{12}$.

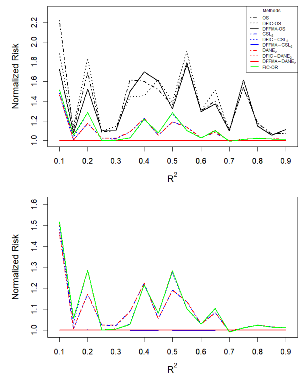


(d) $N = 4 \times 2^{12}$.

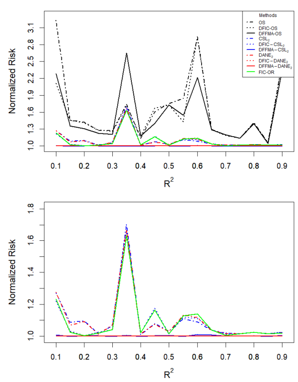
Figure S1: Normalized risks for settings of varying N with heteroscedastic errors.



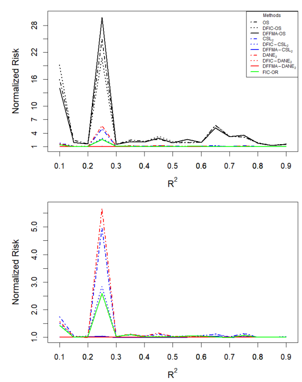
(a) $K = 4$.



(b) $K = 8$.

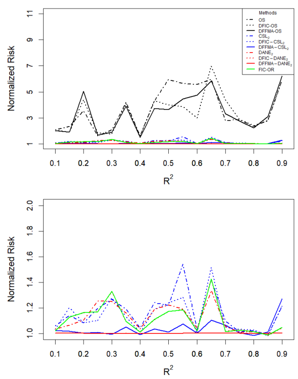


(c) $K = 16$.

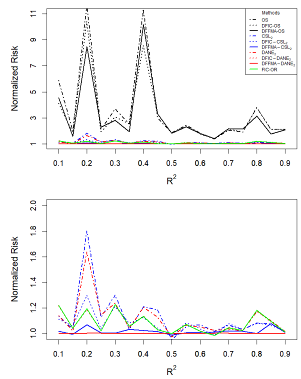


(d) $K = 32$.

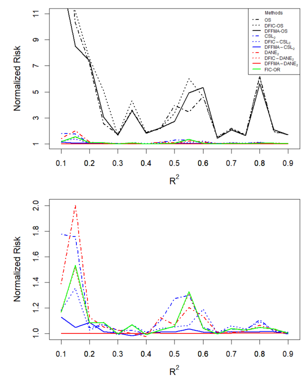
Figure S2: Normalized risks for settings of varying K with heteroscedastic errors.



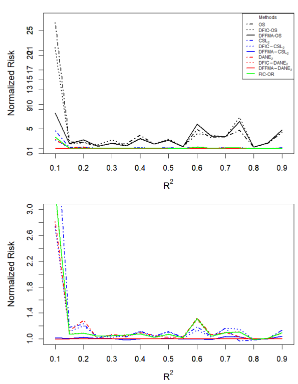
(a) $a = 0.8$.



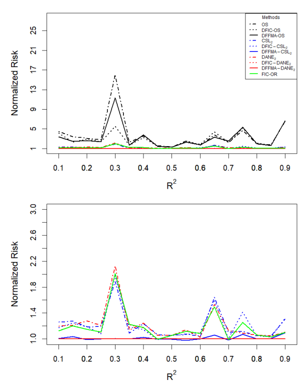
(b) $a = 1$.



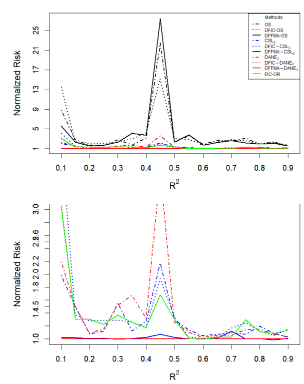
(c) $a = 1.2$.



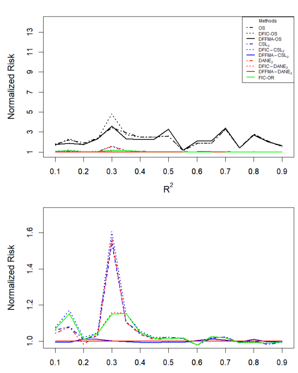
(d) $\rho = 0.2$.



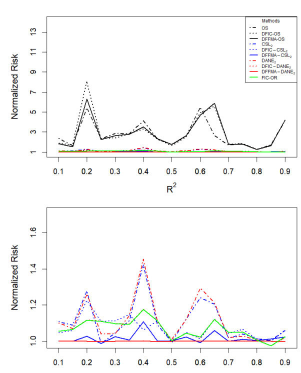
(e) $\rho = 0.4$.



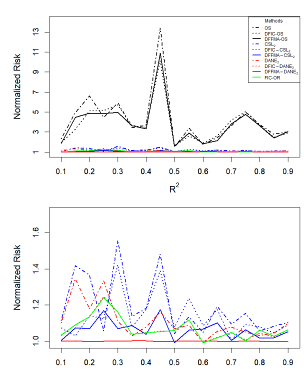
(f) $\rho = 0.6$.



(g) $q = 5$.



(h) $q = 10$.



(i) $q = 15$.

Figure S3: Normalized risks for different settings of a, ρ, and q with heteroscedastic errors.

In Figures S1, S2 and S3, we offer numerical results about the evaluation of performances of various approaches by varying values of N, K, q, a and ρ under the heteroscedastic error case. In those figures, similar observations are observed as those in the homoscedastic error case.

Table S1: Empirical CPs for different methods with heteroscedastic errors.

$N(\times 2^{12})$	K	R^2	Methods		
			DFFMA-OS	DFFMA-CSL$_2$	DFFMA-DANE$_2$
1	4	0.3	0.908	0.904	0.904
		0.6	0.906	0.906	0.906
		0.9	0.882	0.888	0.888
	8	0.3	0.882	0.892	0.892
		0.6	0.896	0.896	0.896
		0.9	0.896	0.906	0.906
	16	0.3	0.904	0.914	0.910
		0.6	0.914	0.922	0.918
		0.9	0.896	0.904	0.902
2	4	0.3	0.904	0.904	0.904
		0.6	0.898	0.892	0.892
		0.9	0.900	0.894	0.894
	8	0.3	0.920	0.918	0.918
		0.6	0.898	0.898	0.898
		0.9	0.900	0.894	0.894
	16	0.3	0.876	0.888	0.886
		0.6	0.884	0.892	0.892
		0.9	0.878	0.882	0.882

Table S1 reports the results of empirical coverage probabilities (CPs) of 90% confidence interval over 500 data repetitions under the heteroscedastic error scenario. It is easy to see that the empirical CPs of the three distributed model averaging estimators are all close to the nominal level 90%.

S3.2 Influence of number of iterations on the performance of iterative distributed methods

In this and next sections, simulated data were generated from the same regression model as that in the main body of this article with different parameters.

Here, we examine how the performance of our iterative approaches-DFIC-CSL$_t$, DFIC-DANE$_t$, DFFMA-CSL$_t$, and DFFMA-DANE$_t$-evolves as the number of iterations increases. Specifically, the parameters (N, K, R^2, a, ρ, q) were set to be $(2^{12}, 16, 0.8, 1, 0.2, 10)$, while the number of iterations t varied from 1 to 8. The results of normalized risks for the homoscedastic and heteroscedastic scenarios are presented in Figures S4 and S5, respectively.

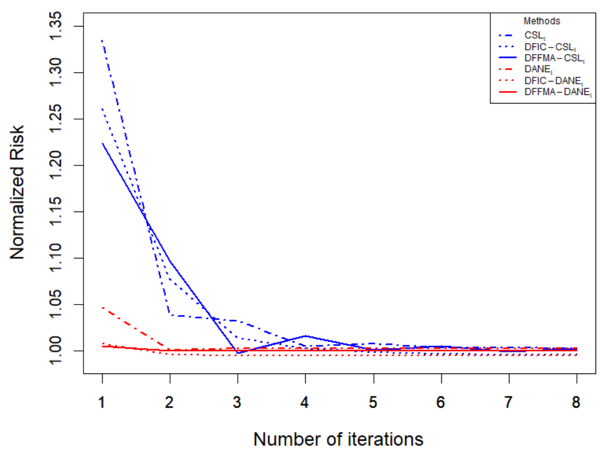


Figure S4: Normalized risks for the homoskedastic error scenario.

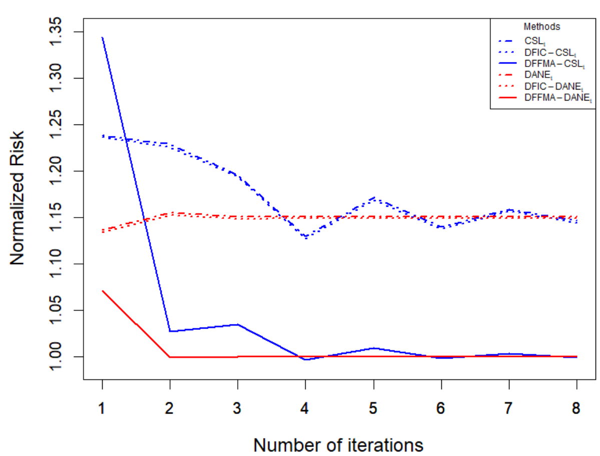


Figure S5: Normalized risks for the heteroscedastic error scenario.

The numerical results of Figures S4 and S5 indicate that, in general, the normalized risks of all methods decrease as the number of iterations increases. All approaches converged rapidly, typically requiring only two iterations in this setting. Therefore, in Sections 4 and 5 of the main body of this article and subsequent analysis in this *Supplementary Material*, the number of iterations for the iterative distributed methods was set to be two. Compared to the CSL-based methods, the DANE-based counterparts exhibit faster convergence rates. Furthermore, DFFMA methods not only converge more rapidly than both DFIC model selection estimators and full-model-based distributed estimators, but also achieve a lower final normalized risk within a few iterations.

S3.3 Numerical studies on comparison of DFFMA and distributed estimators using the full model with identical communication costs

In the distributed model averaging, the cost of data transmission is typically heavy. The communication burden stems intuitively from the need for each site to simultaneously calculate and transmit some relevant estimators from all candidate models. Furthermore, if iterative estimation is employed, additional transmission cost arises for every candidate model, such as transmitting the gradient vector. As suggested by one reviewer, it is meaningful to compare the performance of the DFFMA-CSL$_t$ and DFFMA-DANE$_t$ estimators against CSL_t and DANE_t estimators based on the full model in a fairer and more insightful manner by restricting the number of transmission bits. Thus, in this section, we conducted some numerical investigations.

The simulated data were generated from the linear regression model used in Section 4 of main body of this article, in which the parameters (N, K, a, ρ) were set to be $(2^{12}, 16, 1, 0.2)$. Different values of u were chosen to generate $R^2 = 0.2$ and 0.8. In addition, we set the dimension q of auxiliary covariates to be 5, which resulted in six nested candidate models. Consequently, in each iteration, DFFMA-CSL_t and DFFMA-DANE_t incur approximately six times the communication cost between the central and local machines compared to CSL_t and DANE_t, respectively. Thus, we expect that CSL_{6t} and DANE_{6t} will incur the communication costs roughly equivalent to those of DFFMA-CSL_t and DFFMA-DANE_t, respectively. We vary t across $\{1, 2, 3\}$, and report the normalized risks for those four methods in Tables S2 and S3. The results indicate that, under the same communication cost, our proposed DFFMA methods consistently achieve superior statistical performance compared to their counterparts based on the full model, especially when R^2 is small and the errors are heteroscedastic. The underlying reason is that the proposed DFFMA methods converge more rapidly to a lower risk value as the communication cost (i.e., the number of iterations) increases. This pattern aligns with the observation from numerical results about the influence of number of iterations on iterative approaches.

Table S2: Normalized risks with homoscedastic errors.

Method	R^2	
	0.2	0.8
CSL_6	1.236061	1.088820
DFFMA-CSL_1	1.076776	1.196446
CSL_{12}	1.240099	1.088947
DFFMA-CSL_2	0.995879	1.020814
CSL_{18}	1.240129	1.088961
DFFMA-CSL_3	1.000228	1.012942
DANE_6	1.240129	1.088962
DFFMA-DANE_1	1.002965	0.993413
DANE_{12}	1.240129	1.088962
DFFMA-DANE_2	0.999170	1.003599
DANE_{18}	1.240129	1.088962
DFFMA-DANE_3	1.000039	1.002899

Table S3: Normalized risks with heteroscedastic errors.

Method	R^2	
	0.2	0.8
CSL_6	1.206742	1.029512
DFFMA-CSL_1	1.086634	1.191796
CSL_{12}	1.207133	1.028894
DFFMA-CSL_2	1.009011	1.026441
CSL_{18}	1.207158	1.028862
DFFMA-CSL_3	1.004148	1.003453
$DANE_6$	1.207161	1.028857
DFFMA-$DANE_1$	0.995178	1.009238
$DANE_{12}$	1.207161	1.028857
DFFMA-$DANE_2$	1.002539	1.000047
$DANE_{18}$	1.207161	1.028857
DFFMA-$DANE_3$	1.002337	1.000001

S4 Additional details on the analysis of the U.S. Airline dataset and further results

S4.1 Formulation of our modeling details

In this subsection, we formulate our full regression model and identify which variables are core and which are auxiliary. We accomplish this job by evaluating the marginal correlation between each regressor and the response. More specifically, we randomly drew a random subsample with 1,000,000 observations from the entire data and computed marginal Pearson correlation coefficient between each covariate and the response. The values of these correlation coefficients are presented in Table S4. It is easy to see that the covariate DepDelay exhibits a markedly stronger absolute correlation with ArrDelay than any other covariate. We therefore treated it as the core variable in our regression model, with the others accordingly assigned as auxiliary ones. Finally, our full model is specified as

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

Table S4: Marginal correlation coefficient between each covariate variable and the response.

Covariates	DepDelay	DepTime	DayofWeek	Month	CRSElapsedTime	Distance
Marginal Correlation Coefficient	0.857	0.145	-0.043	0.037	-0.012	0.002

S4.2 Implementation details of various distributed approaches on the Spark system

To implement the distributed procedures, we deployed a Spark-on-YARN cluster on the Alibaba Cloud E-MapReduce platform, which is an industry-standard architecture for distributed computation `https://www.alibabacloud.com/products/emapreduce`. The cluster comprises one master node and two worker nodes, each equipped with 8 virtual CPUs, 32 GB of RAM, and an 80 GB SSD local drive. The dataset was stored in the Hadoop Distributed File System (HDFS) and randomly partitioned into 1,139 subsets, each containing approximately 100,000 observations. Each executor was allocated 14 GB of memory and 4 CPU cores, with an additional 2 GB overhead memory. The algorithm was developed using the Spark R API (SparkR) and executed on a Spark system (Version 3.5.3). We executed the distributed estimation strategies outlined in Section 4, with the exception of the FIC-OR and FFMA-OR methods, which were instead run on a single machine with 6 virtual cores and 16 GB of RAM. The number of iterations for the distributed iterative methods was set to $t = 2$.

S4.3 Additional results for the focus parameter $\mu = \beta_1 + \beta_2$

Table S5: Estimate and confidence interval for $\mu = \beta_1 + \beta_2$.

Method	EST	LB	UB
OS	1.00273	1.00256	1.00291
CSL_2	1.00049	1.00032	1.00066
$DANE_2$	1.00103	1.00086	1.00121
DFIC-OS	1.00272	1.00254	1.00290
DFIC-CSL_2	1.00050	1.00033	1.00067
DFIC-$DANE_2$	1.00078	1.00062	1.00094
DFFMA-OS	1.00270	1.00254	1.00286
DFFMA-CSL_2	1.00122	1.00106	1.00138
DFFMA-$DANE_2$	1.00087	1.00071	1.00103
DMAP-SA	1.00193	/	/
DMAP-SL_2	1.00003	/	/
FIC-OR	1.00101	1.00084	1.00118
FFMA-OR	1.00090	1.00075	1.00106

As a supplement to the analysis in Section 5 of the main body of our article, we report the estimators and 90% confidence intervals for the focus parameter $\mu = \beta_1 + \beta_2$ from various methods. The results are listed in Table S5, where EST, LB, and UB denote the point estimate, lower bound, and upper bound of the 90% confidence interval, respectively. It should be noted that, as no distributional results are available for DMAP-SA and DMAP-SL_2, confidence intervals cannot be provided and are consequently absent. From Table S5, we could see that all methods produce the similar results. And, the DFFMA-$DANE_2$ yields the estimate closest to the gold standard FFMA-OR.

S4.4 Inferences for the focus parameter $\mu = \beta_2$

Table S6: Relative squared errors of various methods ($\times 10^7$).

Method	**OS**	**CSL_2**	**$DANE_2$**
Relative Squared Error	33.124	0.45780	0.38125
Method	**DFIC-OS**	**DFIC-CSL_2**	**DFIC-$DANE_2$**
Relative Squared Error	31.454	0.37675	0.37443
Method	**DFFMA-OS**	**DFFMA-CSL_2**	**DFFMA-$DANE_2$**
Relative Squared Error	29.582	0.04804	0.00403
Method	**DMAP-SA**	**DMAP-SL_2**	**FIC-OR**
Relative Squared Error	27.892	0.08091	0.10001

Table S7: Estimate and confidence interval for $\mu = \beta_2$.

Method	EST	LB	UB
OS	0.94709	0.94701	0.94718
CSL_2	0.94543	0.94535	0.94552
$DANE_2$	0.94548	0.94540	0.94556
DFIC-OS	0.94704	0.94696	0.94713
DFIC-CSL_2	0.94555	0.94546	0.94563
DFIC-$DANE_2$	0.94546	0.94538	0.94554
DFFMA-OS	0.94699	0.94692	0.94706
DFFMA-CSL_2	0.94534	0.94527	0.94541
DFFMA-$DANE_2$	0.94529	0.94522	0.94536
DMAP-SA	0.94694	/	/
DMAP-SL_2	0.94518	/	/
FIC-OR	0.94517	0.94509	0.94525
FMA-OR	0.94527	0.94520	0.94534

In this subsection, we consider the inferences for the focus parameter $\mu = \beta_2$, representing the coefficient of the core predictor DepDelay, which allows us to investigate how the departure delay influences the arrival delay time. 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 S6 reports the relative squared error of each estimator relative to this reference. Similar to the results for the focus parameter $\mu = \beta_1 + \beta_2$, our suggested DFIC-CSL$_t$ and DFFMA-CSL$_t$ demonstrate superior performances over all other strategies, including the impractical FIC-OR.

In Table S7, we report the estimators and 90% confidence intervals for the focus parameter $\mu = \beta_2$ from various methods. From these results, we could draw the similar conclusions as those for the inference of $\mu = \beta_1 + \beta_2$. In addition, these results indicate that for every one-hour increase in DepDelay, the arrival delay time increases by approximately 0.9453 hours.

References

Huang, C. and Huo, X. (2019). A distributed one-step estimator. *Mathematical Programming*, 174:41–76.

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.

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

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404.

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.