Supplementary to "Robust Distributed Estimation and Variable Selection for Massive Datasets via Rank Regression"

Jiaming Luan[1], Hongwei Wang[1], Kangning Wang[1]*and Benle Zhang[1]
[1]Shandong Technology and Business University,
No. 191, Binhai Middle Road, Laishan District, Yantai 264005, China

1 Figures of simulation results

Figures 1-6 are about here.

2 Technical proofs

Proof of Theorem 1**.** By direct calculation, we can obtain that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}^{DR^2}-\boldsymbol{\beta}_0\right)=\left(\sum_{k=1}^{K}w_k\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\right)^{-1}\left(\sqrt{N}\sum_{k=1}^{K}w_k\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}(\hat{\boldsymbol{\beta}}_k^{R^2}-\boldsymbol{\beta}_0)\right).$$

For the robust local R^2 estimators $\hat{\boldsymbol{\beta}}_k^{R^2}, k=1,\cdots,K$, by the Theorem 1 in Leng (2010), we know that they admit the following asymptotic rule

$$\hat{\boldsymbol{\beta}}_k^{R^2}-\boldsymbol{\beta}_0=\boldsymbol{\Sigma}_k^{-1}\frac{1}{2\int f^2(t)dt}\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_{ki}\zeta(\epsilon_{ki})+O_p\left(\frac{1}{n_k}\right), \tag{1}$$

*The corresponding author. Email: wkn1986@126.com. The authors are listed in the alphabetical order. The research was supported by NNSF project of China (11901356, 11901149), wealth management project (2019ZBKY047) of Shandong Technology and Business University.

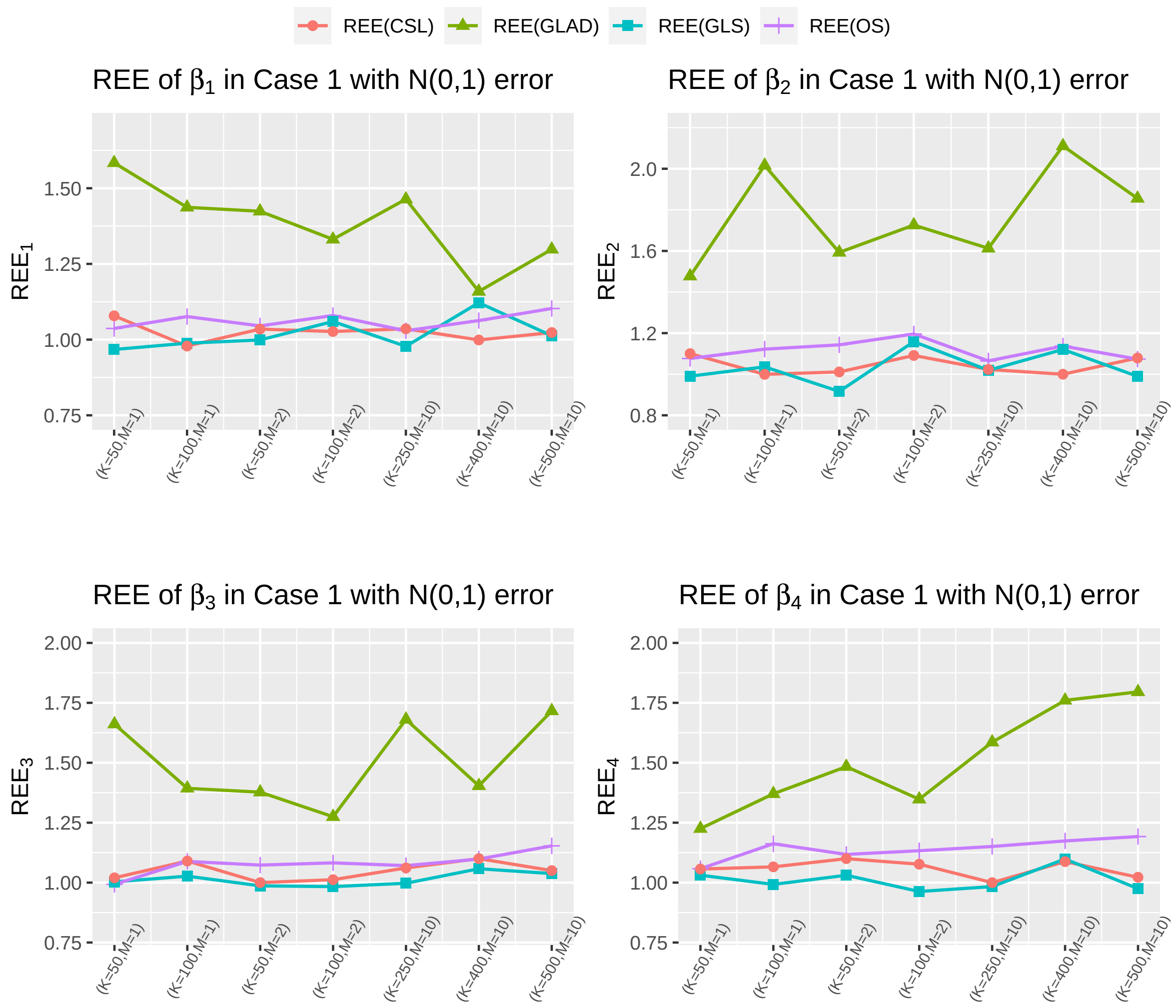

Figure 1: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 1 with $N(0, 1)$ random error.

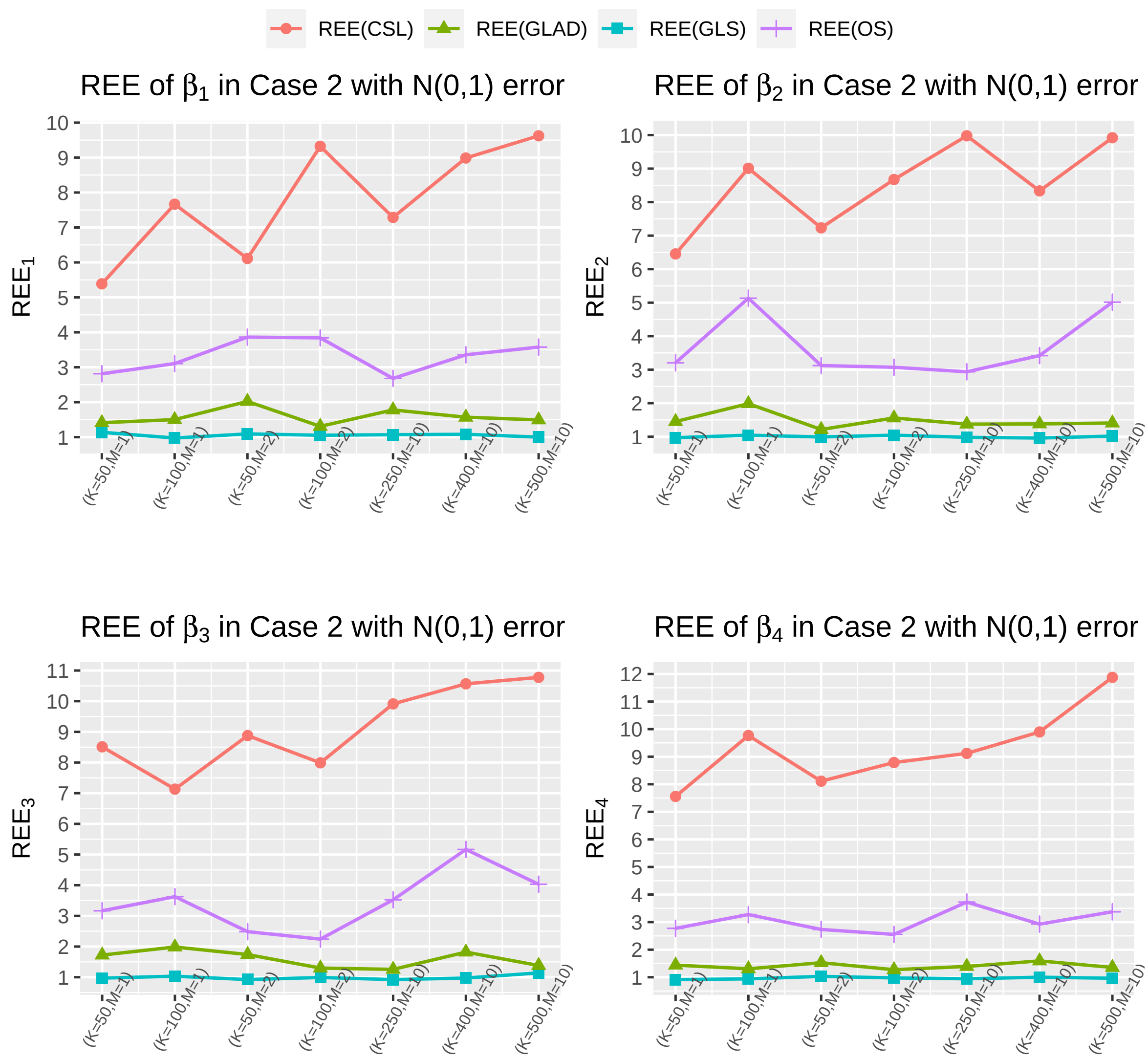

Figure 2: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 2 with $N(0, 1)$ random error.

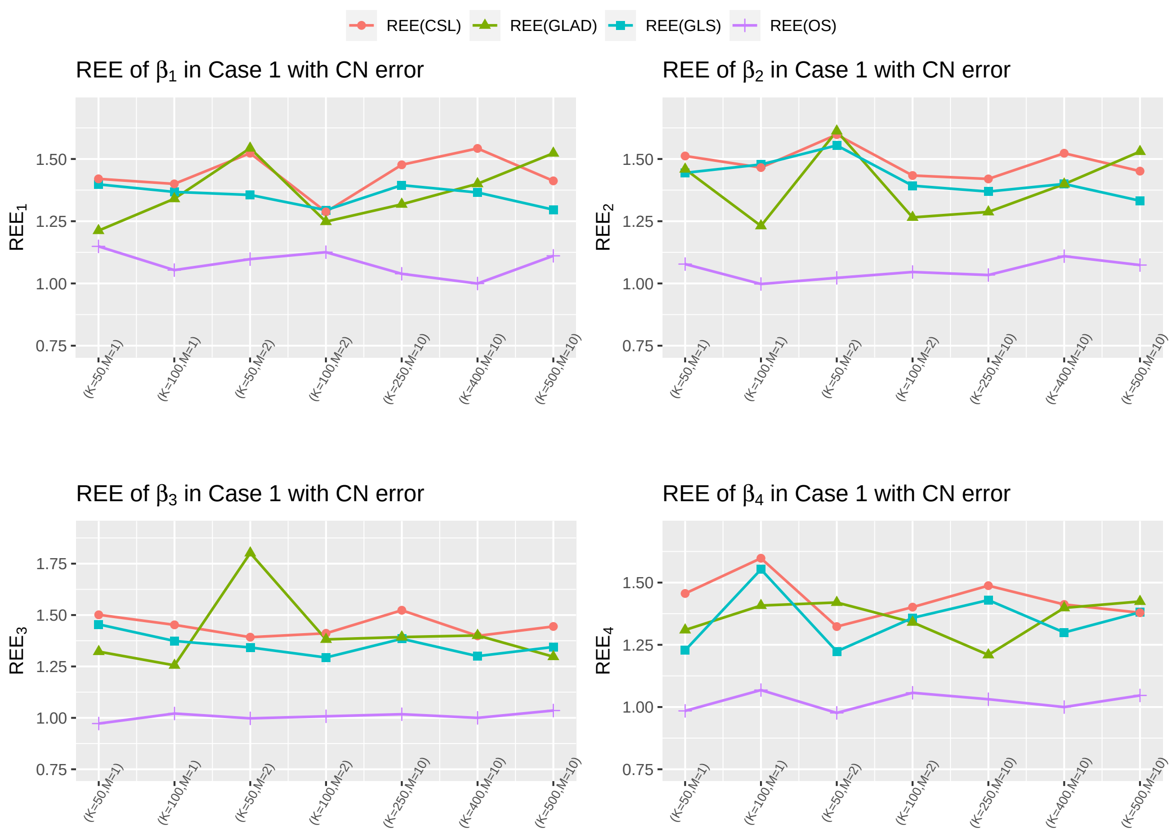

Figure 3: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 1 with contaminated normal random error.

Figure 4: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 2 with contaminated normal random error.

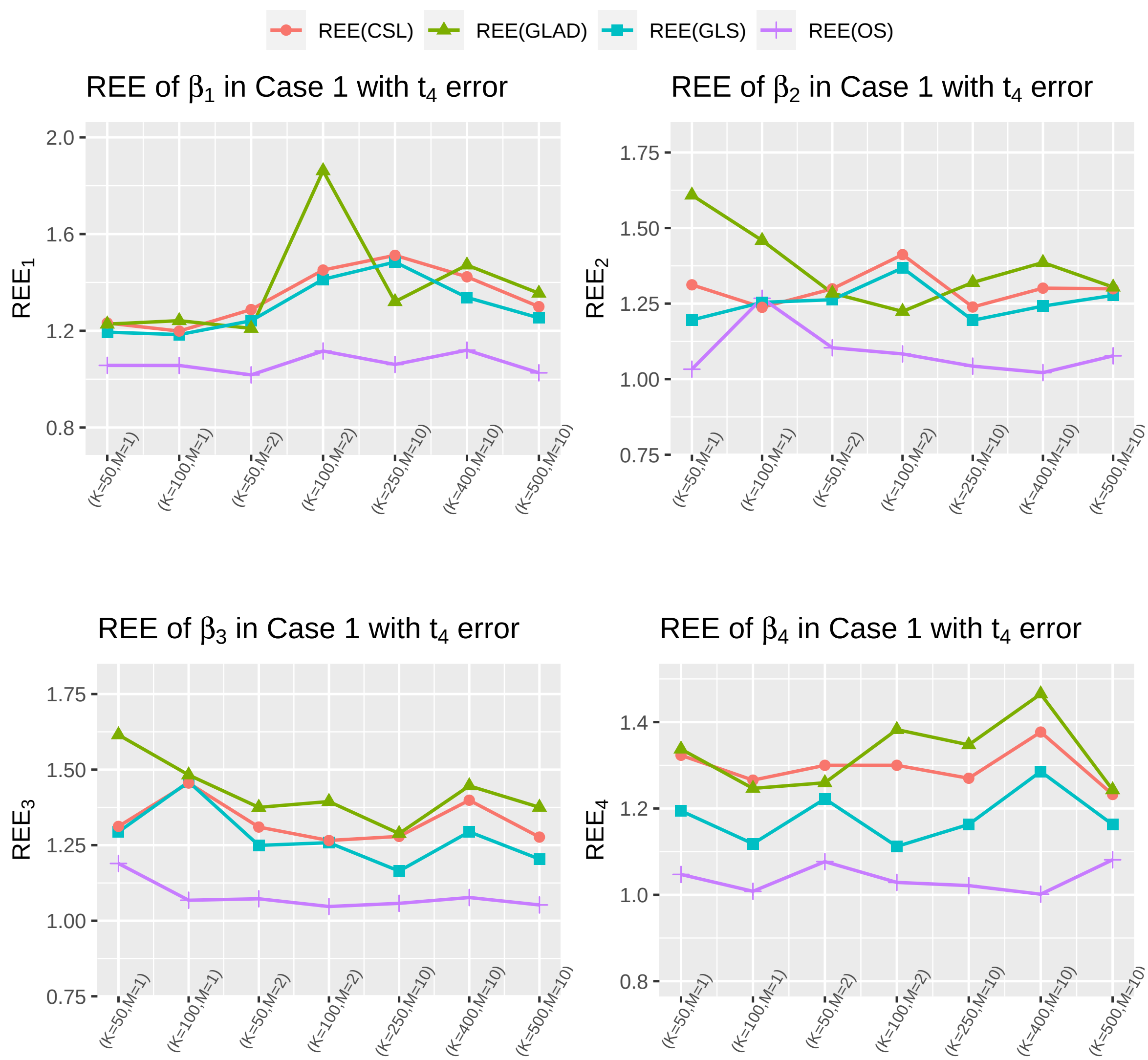

Figure 5: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 1 with t_4 random error.

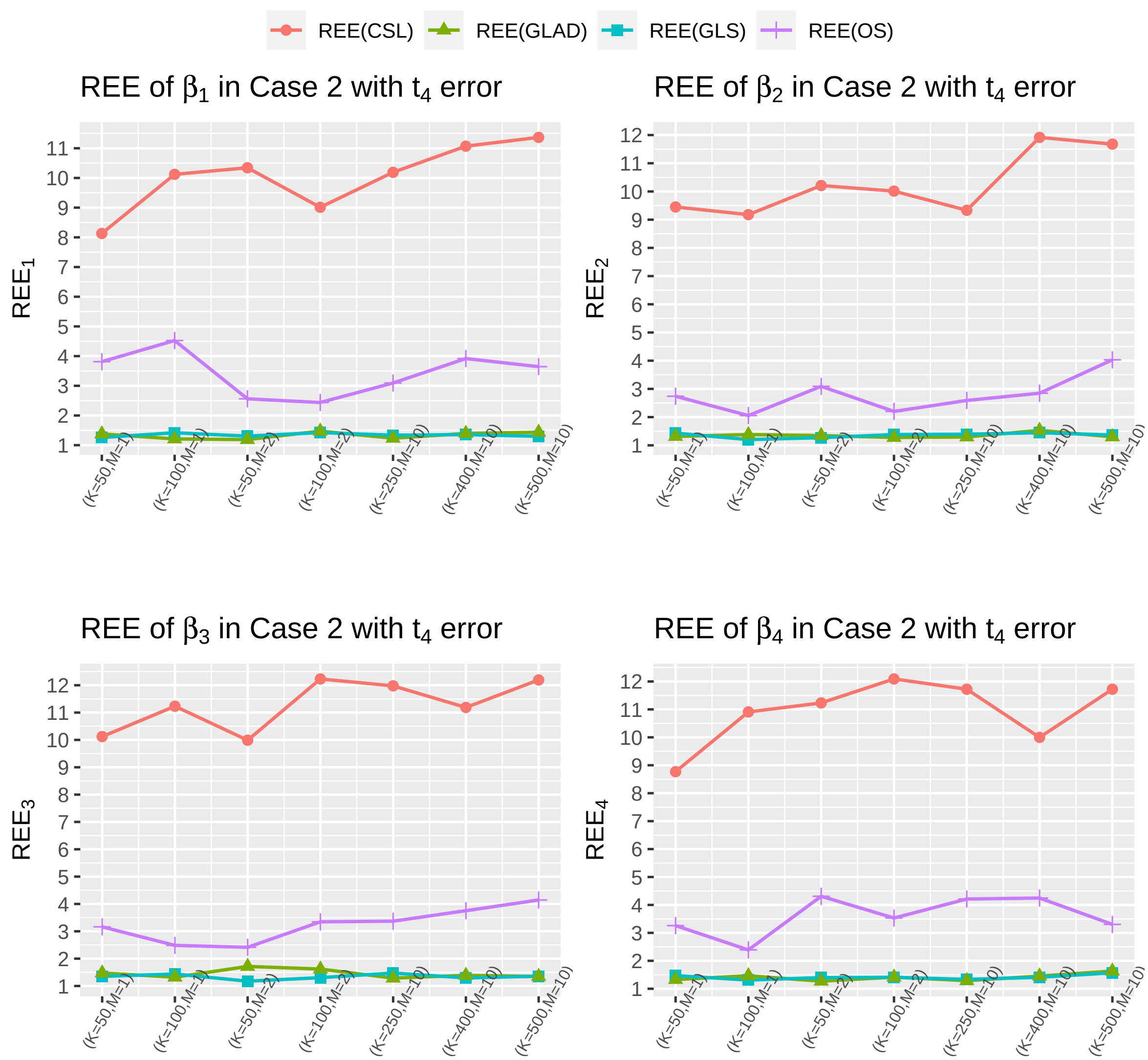

Figure 6: Relative estimation efficiency $REE_j(OS)$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$, $j = 1, \cdots, 4$ versus number of machines K and sample size M under Case 2 with t_4 random error.

where $\zeta(\epsilon_{ki}) = \frac{1}{n_k}\{2R(\epsilon_{ki}) - (n+1)\}$, $R(\epsilon_{ki})$ is the rank statistic of ϵ_{ki}. Note that $\sqrt{N}\left(\sum_{k=1}^{K} w_k(\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k} - \boldsymbol{\Sigma}_k)(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}_0)\right) = O_p(\frac{K}{\sqrt{N}})$, $E(\zeta(\epsilon_{ki})) = 0$ and

$$\begin{aligned}
\mathrm{var}(\zeta(\epsilon_{ki})) &= \frac{1}{n_k}\mathrm{var}(2R(\epsilon_{ki}) - (n+1)) \\
&= \frac{1}{n_k^3}\sum_{i=1}^{n_k}(2i - (n+1))^2 \\
&= \frac{4(n_k+1)^2}{n_k^3}\sum_{i=1}^{n_k}(\frac{i}{n+1} - \frac{1}{2})^2 \\
&\to 4\int_0^1 (t - \frac{1}{2})^2 dt = \frac{1}{3},
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(\zeta(\epsilon_{ki}), \zeta(\epsilon_{kj})) &= \frac{1}{n_k}\mathrm{cov}(2R(\epsilon_{ki}) - (n+1), 2R(\epsilon_{kj}) - (n+1)) \\
&= \frac{1}{n_k^3(n_k-1)}\sum_{i=1}^{n_k}\sum_{j\neq i}(2i - (n+1))(2j - (n+1)) \\
&= \frac{4(n_k+1)^2}{n_k^2(n_k-1)}\int_0^1 (t - \frac{1}{2})^2 dt \\
&\to 0, \text{ for } i \neq j.
\end{aligned}$$

By the condition about K in Theorem 1 and (1), we can get that

$$\begin{aligned}
&\sqrt{N}\left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}_0)\right) \\
&= \sqrt{N}\left(\sum_{k=1}^{K} w_k(\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k} - \boldsymbol{\Sigma}_k)(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}_0)\right) + \sqrt{N}\left(\sum_{k=1}^{K} w_k\boldsymbol{\Sigma}_k(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}_0)\right) \\
&= \sqrt{N}\left(\sum_{k=1}^{K} w_k\boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_k^{-1}\frac{1}{2\int f^2(t)dt}\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_{ki}\zeta(\epsilon_{ki}) + O_p\left(\frac{1}{n_k}\right)\right)\right) + O_p(\frac{K}{\sqrt{N}}) \\
&= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\frac{1}{2\int f^2(t)dt}\sum_{i=1}^{n_k}\boldsymbol{X}_{ki}\zeta(\epsilon_{ki}) + O_p(\frac{K}{\sqrt{N}}) \\
&\to_d N\left(\boldsymbol{0}, \frac{1}{12(\int f^2(t)dt)^2}\left(\sum_{k=1}^{K} w_k\boldsymbol{\Sigma}_k\right)\right).
\end{aligned}$$

Further note that $\sum_{k=1}^{K} w_k = 1$, by condition (A1), we have $\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k} - \sum_{k=1}^{K} w_k \boldsymbol{\Sigma}_k = o_p(1)$. Then we can obtain that

$$
\begin{aligned}
\sqrt{N}\left(\hat{\boldsymbol{\beta}}^{DR^2} - \boldsymbol{\beta}_0\right) &= \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right)^{-1} \left(\sqrt{N}\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}_0)\right) \\
&\to_d N\left(\mathbf{0}, \frac{1}{12\omega^2}\left(\sum_{k=1}^{K} w_k \boldsymbol{\Sigma}_k\right)^{-1}\right).
\end{aligned}
$$

The proof is completed.

Proof of Theorem 2**.** Consider

$$
\begin{aligned}
L_\lambda(\boldsymbol{\beta}) &= P_\lambda(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}_0) \\
&= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{DR^2})^T \left[\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{DR^2}) - (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{DR^2})^T \left[\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right](\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{DR^2}) \\
&\quad + \lambda \sum_{j=1}^{p} \lambda_j \left[|\beta_j| - |\beta_{0,j}|\right].
\end{aligned}
$$

Denote $\boldsymbol{u} = (u_1, \cdots, u_p)^T = \sqrt{N}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, we may write $NL_\lambda(\boldsymbol{\beta})$ as

$$
\begin{aligned}
NL_\lambda(\boldsymbol{\beta}) &= \boldsymbol{u}^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right)\boldsymbol{u} + 2\boldsymbol{u}^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\{\sqrt{N}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{DR^2})\}\right) \\
&\quad + N\lambda \sum_{j=1}^{p} \lambda_j \left[|\beta_j| - |\beta_{0,j}|\right],
\end{aligned}
$$

which is minimized by $\hat{\boldsymbol{u}}_\lambda = \sqrt{N}(\hat{\boldsymbol{\beta}}_\lambda^{DR^3} - \boldsymbol{\beta}_0)$. Let

$$
Z(\boldsymbol{u}) = N\lambda \sum_{j=1}^{p} \lambda_j \left[|\beta_{0,j} + u_j/\sqrt{N}| - |\beta_{0,j}|\right],
$$

and we write $Z_j(\boldsymbol{u}) = N\lambda\lambda_j \left[|\beta_{0,j} + u_j/\sqrt{N}| - |\beta_{0,j}|\right]$, then

$$
Z_j(\boldsymbol{u}) = \begin{cases} \sqrt{N}\lambda\lambda_j u_j \mathrm{sign}(\beta_{0,j}), & \text{if } \beta_{0,j} \neq 0, \\ \sqrt{N}\lambda\lambda_j |u_j|, & \text{if } \beta_{0,j} = 0. \end{cases}
$$

Now, the conditions in Theorem 2 assure the following

$$Z_j(\boldsymbol{u}) \to P(\beta_{0,j}, u_j) = \begin{cases} 0, & \text{if } \beta_{0,j} \neq 0, \\ 0, & \text{if } \beta_{0,j} = 0 \text{ and } u_j = 0 \\ \infty, & \text{if } \beta_{0,j} = 0 \text{ and } u_j \neq 0, \end{cases}$$

Thus, we have that

$$NL_\lambda(\boldsymbol{\beta}) \to_d \boldsymbol{u}^T \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k \right) \boldsymbol{u} + 2\boldsymbol{u}^T \left(\left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k \right) \{ \sqrt{N}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{DR^2}) \} \right) + \sum_{j=1}^p P(\beta_{0,j}, u_j).$$

Applying the arguments in Knight (1998), we have

$$\begin{aligned} \hat{\boldsymbol{u}}_{\lambda,\mathcal{A}} = \sqrt{N}(\hat{\boldsymbol{\beta}}^{DR^3}_{\lambda,\mathcal{A}} - \boldsymbol{\beta}_{01}) &\to_d \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k \right)^{-1}_{\mathcal{A}\mathcal{A}} \left\{ \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k \right) \sqrt{N}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{DR^2}) \right\}_{\mathcal{A}} \\ &\sim N\left(\boldsymbol{0}, \frac{1}{12\omega^2} \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k \right)^{-1}_{\mathcal{A}\mathcal{A}} \right). \end{aligned}$$

The asymptotic normality is established. What is more, if $\hat{\beta}^{DR^3}_{\lambda,j} \neq 0$ for some $j > d$, the partial derivative of $P_\lambda(\boldsymbol{\beta})$ can be calculated as

$$\sqrt{N}\frac{\partial P_\lambda(\boldsymbol{\beta})}{\partial \beta_j}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{DR^3}_\lambda} = 2\left[\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k} \right]^T_j (\hat{\boldsymbol{\beta}}^{DR^3}_\lambda - \hat{\boldsymbol{\beta}}^{DR^2}) + \sqrt{N}\lambda\lambda_j \text{sign}(\beta_j),$$

where $\left[\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right]_j$ is the jth row of the matrix $\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}$. By Theorem 2.1 and the $\sqrt{N}$ consistency of $\hat{\boldsymbol{\beta}}^{DR^3}_\lambda$, we can get that $\sqrt{N}(\hat{\boldsymbol{\beta}}^{DR^3}_\lambda - \hat{\boldsymbol{\beta}}^{DR^2}) = \sqrt{N}(\hat{\boldsymbol{\beta}}^{DR^3}_\lambda - \boldsymbol{\beta}_0) - \sqrt{N}(\hat{\boldsymbol{\beta}}^{DR^2} - \boldsymbol{\beta}_0) = O_p(1)$, consequently, $2\left[\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right]^T_j (\hat{\boldsymbol{\beta}}^{DR^3}_\lambda - \hat{\boldsymbol{\beta}}^{DR^2}) = O_p(1)$. If $\hat{\beta}^{DR^3}_{\lambda,j} \neq 0$, $\text{sign}(\beta_j) = -1$ or 1, then by the convergence rate about the tuning parameter λ, we know that $|\sqrt{N}\lambda\lambda_j \text{sign}(\beta_j)| \geqslant |\sqrt{N}\lambda b_N| \to \infty$ for $j > d$. Thus equation $\sqrt{N}\frac{\partial P_\lambda(\boldsymbol{\beta})}{\partial \beta_j}|_{\boldsymbol{\beta}} = 0$ can not hold, which implies that $P\left(\hat{\beta}^{DR^3}_{\lambda,j} = 0\right) \to 1$ for any $j \in \{d+1, \cdots, p\}$. Therefore, combining with the asymptotic normality in (b), (a) can be proved. The proof is completed.

Proof of Theorem 3. Firstly, for $\lambda \in \mathbb{R}_-^+$, we suppose $j^* \in \mathcal{A}$ and $\hat{\beta}_{\lambda,j^*}^{DR^3} = 0$, we have

$$
\begin{aligned}
RSS(\lambda) &= \left(\hat{\boldsymbol{\beta}}_\lambda^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_\lambda^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
&\geq \hat{\lambda}_{\min} \left(\hat{\boldsymbol{\beta}}_\lambda^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\hat{\boldsymbol{\beta}}_\lambda^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
&\geq \hat{\lambda}_{\min} (\hat{\beta}_{j^*}^{DR^2})^2,
\end{aligned}
$$

where $\hat{\lambda}_{\min}$ is the smallest eigenvalue of $\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}$. So combining (A1) and Theorem 1 together yields

$$
\hat{\lambda}_{\min} (\hat{\beta}_{j^*}^{DR^2})^2 \xrightarrow{p} \lambda_{\min}^0 \beta_{0,j^*}^2 > 0.
$$

Furthermore, for $\lambda_N = \log(N)/N$, we have

$$
\begin{aligned}
RSS(\lambda_N) &= \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
&\leq \hat{\lambda}_{\max} \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
&= \hat{\lambda}_{\max} \left[\left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \boldsymbol{\beta}_0\right)^2 + \left(\hat{\boldsymbol{\beta}}^{DR^2} - \boldsymbol{\beta}_0\right)^2\right] + o_p(1),
\end{aligned}
$$

where $\hat{\lambda}_{\max}$ is the largest eigenvalue of $\sum_{k=1}^K w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}$. So by (A1), Theorems 1 and 2, we have

$$
RSS(\lambda_N) = o_p(1),
$$

and furthermore, $df_\lambda \frac{\log(N)}{N} = o(1)$, for arbitrary $\lambda \in \mathbb{R}^+$. This implies

$$
P\left(\inf_{\lambda \in \mathbb{R}_-^+} DBIC(\lambda) > DBIC(\lambda_N)\right) \to 1.
$$

For $\lambda \in \mathbb{R}_+^+$, firstly we have

$$
P(df_\lambda - df_{\lambda_N} \geq 1) \to 1.
$$

What is more, one can verify

$$
\begin{aligned}
& N\left[RSS(\lambda) - RSS(\lambda_N)\right] \\
= & N\left(\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
& \quad - N\left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
\geq & \inf_{\lambda \in \mathbb{R}_+^+} N\left(\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
& \quad - N\left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{R} - \hat{\boldsymbol{\beta}}^{DR^2}\right)^T \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right) \left(\hat{\boldsymbol{\beta}}_{\lambda_N}^{R} - \hat{\boldsymbol{\beta}}^{DR^2}\right) \\
= & O_p(1).
\end{aligned}
$$

This implies that $P\{N[DBIC(\lambda) - DBIC(\lambda_N)] \to +\infty\} \to 1$. So

$$
P\left(\inf_{\lambda \in \mathbb{R}_+^+} DBIC(\lambda) > DBIC(\lambda_N)\right) \to 1.
$$

The proof is completed.

References

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. Statistica Sinica, 20, 167-181.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. The Annals of Statistics, 26, 755-770.