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

Jiaming Luan¹, Hongwei Wang¹, Kangning Wang¹* and Benle Zhang¹
¹Shandong 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

\[ \sqrt{N} \left( \hat{\beta}^{DR^2} - \beta_0 \right) = \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right)^{-1} \left( \sqrt{N} \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} (\hat{\beta}^R_k - \beta_0) \right). \]

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

\[ \hat{\beta}^R_k - \beta_0 = \Sigma_k^{-1} \int f^2(t) dt \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki} \zeta(\epsilon_{ki}) + O_p \left( \frac{1}{n_k} \right), \quad (1) \]

*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 $REE_j(\text{OS})$, $REE_j(\text{CSL})$, $REE_j(\text{GLAD})$ and $REE_j(\text{GLS})$, $j = 1, \cdots, 4$ 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 \( \epsilon_{ki} \). Note that
\[
\sqrt{N} \left( \sum_{k=1}^{K} w_k \left( \frac{X_{ki}^T X_k}{n_k} - \Sigma_k (\hat{\beta}_k^{R^2} - \beta_0) \right) \right) = O_p\left( \frac{K}{\sqrt{N}} \right),
\]
\( E(\zeta(\epsilon_{ki})) = 0 \) and
\[
\text{var}(\zeta(\epsilon_{ki})) = \frac{1}{n_k} \text{var}(2R(\epsilon_{ki}) - (n+1))
\]
\[
= \frac{1}{n_k^3} \sum_{i=1}^{n_k} (2i - (n+1))^2
\]
\[
= \frac{4(n_k + 1)^2}{n_k} \sum_{i=1}^{n_k} \left( \frac{i}{n_k + 1} - \frac{1}{2} \right)^2
\]
\[
\to 4 \int_0^1 (t - \frac{1}{2})^2 dt = \frac{1}{3},
\]
\[
\text{cov}(\zeta(\epsilon_{ki}), \zeta(\epsilon_{kj})) = \frac{1}{n_k} \text{cov}(2R(\epsilon_{ki}) - (n+1), 2R(\epsilon_{kj}) - (n+1))
\]
\[
= \frac{1}{n_k^2(n_k - 1)} \sum_{i=1}^{n_k} \sum_{j \neq i} (2i - (n+1))(2j - (n+1))
\]
\[
= \frac{4(n_k + 1)^2}{n_k^2(n_k - 1)} \int_0^1 (t - \frac{1}{2})^2 dt
\]
\[
\to 0, \text{ for } i \neq j.
\]

By the condition about \( K \) in Theorem 1 and (1), we can get that
\[
\sqrt{N} \left( \sum_{k=1}^{K} w_k \Sigma_k \left( \frac{1}{2} \int f^2(t) dt \right) \right) \to_d N\left( 0, \frac{1}{12(\int f^2(t) dt)^2} \left( \sum_{k=1}^{K} w_k \Sigma_k \right) \right).
\]
Further note that \( \sum_{k=1}^{K} w_k = 1 \), by condition (A1), we have \( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} = \sum_{k=1}^{K} w_k \Sigma_k = o_p(1) \). Then we can obtain that
\[
\sqrt{N} \left( \hat{\beta}_{DR}^2 - \beta_0 \right) = \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right)^{-1} \left( \sqrt{N} \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} (\hat{\beta}_{DR}^2 - \beta_0) \right) \\
\rightarrow_d N \left( 0, \frac{1}{12 \omega^2} \left( \sum_{k=1}^{K} w_k \Sigma_k \right)^{-1} \right).
\]

The proof is completed.

**Proof of Theorem 2.** Consider

\[
L_\lambda(\beta) = P_\lambda(\beta) - P_\lambda(\beta_0) \\
= (\beta - \hat{\beta}_{DR}^2)^T \left[ \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right] (\beta - \hat{\beta}_{DR}^2) - (\beta_0 - \hat{\beta}_{DR}^2)^T \left[ \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right] (\beta_0 - \hat{\beta}_{DR}^2) \\
+ \lambda \sum_{j=1}^{p} \lambda_j |\beta_j| - |\beta_{0,j}|.
\]

Denote \( u = (u_1, \ldots, u_p)^T = \sqrt{N}(\beta - \beta_0) \), we may write \( NL_\lambda(\beta) \) as
\[
NL_\lambda(\beta) = u^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) u + 2u^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \{ \sqrt{N}(\beta_0 - \hat{\beta}_{DR}^2) \} \right) \\
+ N \lambda \sum_{j=1}^{p} \lambda_j |\beta_j| - |\beta_{0,j}|,
\]
which is minimized by \( \hat{u}_\lambda = \sqrt{N}(\hat{\beta}_{DR}^3 - \beta_0) \). Let
\[
Z(u) = N \lambda \sum_{j=1}^{p} \lambda_j \left[ |\beta_{0,j} + u_j / \sqrt{N}| - |\beta_{0,j}| \right],
\]
and we write \( Z_j(u) = N \lambda \lambda_j \left[ |\beta_{0,j} + u_j / \sqrt{N}| - |\beta_{0,j}| \right] \), then
\[
Z_j(u) = \begin{cases} 
\sqrt{N} \lambda \lambda_j u_j \text{sign}(\beta_{0,j}), & \text{if } \beta_{0,j} \neq 0, \\
\sqrt{N} \lambda \lambda_j |u_j|, & \text{if } \beta_{0,j} = 0.
\end{cases}
\]
Now, the conditions in Theorem 2 assure the following

\[ Z_j(u) \rightarrow P(\beta_{0,j}, u_j) = \begin{cases} 0, & \text{if } \beta_{0,j} \neq 0, \\ 0, & \text{if } \beta_{0,j} = 0 \text{ and } u_j = 0 \\ \infty, & \text{if } \beta_{0,j} = 0 \text{ and } u_j \neq 0. \end{cases} \]

Thus, we have that

\[ NL_\lambda(\beta) \rightarrow_d u^T \left( \sum_{k=1}^{K} w_k \Sigma_k \right) u + 2u^T \left( \sum_{k=1}^{K} w_k \Sigma_k \right) \{ \sqrt{N}(\beta_0 - \hat{\beta}^{DR_2}) \} + \sum_{j=1}^{p} P(\beta_{0,j}, u_j). \]

Applying the arguments in Knight (1998), we have

\[ \hat{u}_{\lambda, A} = \sqrt{N}(\hat{\beta}^{DR_3} - \beta_{01}) \rightarrow_d \left( \sum_{k=1}^{K} w_k \Sigma_k \right)^{-1} \left\{ \left( \sum_{k=1}^{K} w_k \Sigma_k \right) \sqrt{N}(\beta_0 - \hat{\beta}^{DR_2}) \right\} \sim N \left( 0, \frac{1}{12\omega^2} \left( \sum_{k=1}^{K} w_k \Sigma_k \right)^{-1} \right). \]

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 \( P_\lambda(\beta) \) can be calculated as

\[ \sqrt{N} \frac{\partial P_\lambda(\beta)}{\partial \beta_j} |_{\beta = \hat{\beta}^{DR_3}} = 2 \left[ \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right] \left( \hat{\beta}^{DR_3} - \hat{\beta}^{DR_2} \right) + \sqrt{N} \lambda \lambda_j \text{sign}(\beta_j), \]

where \( \left[ \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right] \) is the \( j \)th row of the matrix \( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \). By Theorem 2.1 and the \( \sqrt{N} \) consistency of \( \hat{\beta}^{DR_3}_\lambda \), we can get that \( \sqrt{N}(\hat{\beta}^{DR_3}_\lambda - \hat{\beta}^{DR_2}) = \sqrt{N}(\hat{\beta}^{DR_3}_\lambda - \beta_0) - \sqrt{N}(\hat{\beta}^{DR_2}_\lambda - \beta_0) = O_p(1) \), consequently, \( 2 \left[ \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right] \left( \hat{\beta}^{DR_3}_\lambda - \hat{\beta}^{DR_2}_\lambda \right) = O_p(1) \). 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 \( \lambda \), we know that \( |\sqrt{N}\lambda \lambda_j \text{sign}(\beta_j)| \geq |\sqrt{N}\lambda b_N| \rightarrow \infty \) for \( j > d \). Thus equation \( \sqrt{N} \frac{\partial P_\lambda(\beta)}{\partial \beta_j} |_{\beta = 0} \) can not hold, which implies that \( P \left( \hat{\beta}^{DR_3}_{\lambda,j} = 0 \right) \rightarrow 1 \) 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 \( j^* \in A \) and \( \hat{\beta}^{DR^3}_{\lambda,j^*} = 0 \), we have

\[
RSS(\lambda) = \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)
\]

\[
\geq \hat{\lambda}_{\min} \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)^T \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)
\]

\[
\geq \hat{\lambda}_{\min} (\hat{\beta}^{DR^2}_{j^*})^2,
\]

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

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

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

\[
RSS(\lambda_N) = \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda} \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda} \right)
\]

\[
\leq \hat{\lambda}_{\max} \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda} \right)^T \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda} \right)
\]

\[
= \hat{\lambda}_{\max} \left[ \left( \hat{\beta}^{DR^3}_{\lambda_N} - \beta_0 \right)^2 + \left( \hat{\beta}^{DR^2}_{\lambda} - \beta_0 \right)^2 \right] + o_p(1),
\]

where \( \hat{\lambda}_{\max} \) is the largest eigenvalue of \( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \). 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

$$N \left[ RSS(\lambda) - RSS(\lambda_N) \right]$$

$$= N \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)$$

$$- N \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda_N} \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{DR^3}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda_N} \right)$$

$$\geq \inf_{\lambda \in \mathbb{R}_+^+} N \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{DR^3}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)$$

$$- N \left( \hat{\beta}^{R}_\lambda - \hat{\beta}^{DR^2}_\lambda \right)^T \left( \sum_{k=1}^{K} w_k \frac{X_k^T X_k}{n_k} \right) \left( \hat{\beta}^{R}_{\lambda_N} - \hat{\beta}^{DR^2}_{\lambda_N} \right)$$

$$= O_p(1).$$

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**
