Biometrika (2022), *pp.* 1–8
Printed in Great Britain

Supplementary Material for "Testing generalized linear models with high-dimensional nuisance parameter"

BY JINSONG CHEN

College of Applied Health Sciences, University of Illinois at Chicago, 1919 W Taylor St, Chicago, Illinois 60612, U.S.A.

jinsongc@uic.edu

QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

quefeng@email.unc.edu

HUA YUN CHEN

School of Public Health, University of Illinois at Chicago, 2121 W Taylor St, Chicago, Illinois 60612, U.S.A.

hychen@uic.edu

PROOF OF THEOREM 3

When the true model is under H_a, we disassemble $\hat{U}_n$ as

$$\begin{aligned}
\hat{U}_n &= \underbrace{\frac{1}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(y_j-\mu_j)w_i^T w_j\}}_{T1_{\hat{U}_n}} + \frac{1}{n}\sum_{i\neq j}^{n}\{(\mu_i-\hat{\mu}_{\o i})(\mu_j-\hat{\mu}_{\o j})w_i^T w_j\} \\
&+ 2\frac{1}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_j-\hat{\mu}_{\o j})w_i^T w_j\} \\
&= T1_{\hat{U}_n} + \underbrace{\frac{1}{n}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\o i})(\mu_j-\mu_{\o j})w_i^T w_j\}}_{T2_{\hat{U}_n}} + \underbrace{\frac{2}{n}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\o i})(\mu_{\o j}-\hat{\mu}_{\o j})w_i^T w_j\}}_{T3_{\hat{U}_n}} \\
&+ \underbrace{\frac{1}{n}\sum_{i\neq j}^{n}\{(\mu_{\o i}-\hat{\mu}_{\o i})(\mu_{\o j}-\hat{\mu}_{\o j})w_i^T w_j\}}_{T4_{\hat{U}_n}} + \underbrace{\frac{2}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_j-\mu_{\o j})w_i^T w_j\}}_{T5_{\hat{U}_n}} \\
&+ \underbrace{\frac{2}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_{\o j}-\hat{\mu}_{\o j})w_i^T w_j\}}_{T6_{\hat{U}_n}}.
\end{aligned}$$

The term $T2_{\hat{U}_n}$ is a U-statistic. Let $H_{T2_{\hat{U}_n}}(W_1, W_2) = (\mu_{\o 1}-\mu_1)(\mu_{\o 2}-\mu_2)w_1^T w_2$ and $EH_{T2_{\hat{U}_n}} = E\{H_{T2_{\hat{U}_n}}(W_1, W_2)\}$. By the independence of terms in $H_{T2_{\hat{U}_n}}(W_1, W_2)$, it can be seen that

$$E(T2_{\hat{U}_n}) = (n-1)E\{H_{T2_{\hat{U}_n}}(W_1, W_2)\} = (n-1)\{\tilde{\gamma}^T E(Z\omega_\theta^* W^T)E(W\omega_\theta^* Z^T)\tilde{\gamma} + 2\tilde{\gamma}^T E(Z\omega_\theta^* W^T)E(W\omega_\theta^* W^T)\beta + \beta^T E(W\omega_\theta^* W^T)E(W\omega_\theta^* W^T)\beta\} \asymp (n-1)||\beta||_2^2$$

by Assumption 8 and Theorem 2.

Let $H1_{T2_{\hat{U}_n}} = E\{H_{T2_{\hat{U}_n}}(W_1, W_2)|W_1\} = (\mu_{\o 1} - \mu_1)w_1^T E\{(\mu_{\o 2} - \mu_2)W\}$. Based on the sub-Gaussian property of X and Assumption 2, we have

$$P\left(\frac{|\mu_\o - \mu|}{\|\tilde{\theta}\|_2} > t\right) = P\left(\frac{|g^{-1}(X^T\theta_\o) - g^{-1}(X^T\theta)|}{\|\tilde{\theta}\|_2} > t\right) = P\left(\frac{|\tilde{\theta}^T X|}{\|\tilde{\theta}\|_2} > D_6^{-1}t\right) \le 2\exp\left(-\frac{t^2}{2D_6^2\sigma^2}\right),$$

and further

$$P\left(\frac{|\tilde{\theta}^T X|}{\|\beta\|_2} > D_6^{-1}D_{11}t\right) \le 2\exp\left(-\frac{t^2}{2D_6^2\sigma^2}\right)$$

because $\|\tilde{\theta}\|_2^2 = D_{10}\|\beta\|_2^2$ for a constant D_{10} by Theorem 2. It means for any $\varepsilon > 0$, there exists a $M > 0$ such that

$$P\left(\frac{|\mu_\o - \mu|}{\|\beta\|_2} > M\right) \le \varepsilon.$$

Thus, by definition, $|\mu_\o - \mu| = O_P(\|\beta\|_2)$. It implies

$$E\{(H1_{T2_{\hat{U}_n}})^2\} = \tilde{\theta}^T E(XW^T)E\{(\mu_\o - \mu)^2 WW^T\}E(WX^T)\tilde{\theta} = O(\|\beta\|_2^4).$$

Following Hoeffding decomposition in the variance evaluation of U-statistic (Hoeffding, 1948), we have

$$\mathrm{Var}(T2_{\hat{U}_n}) = \frac{4(n-2)(n-1)}{n}\mathrm{Var}(H1_{T2_{\hat{U}_n}}) + \frac{2(n-1)}{n}\mathrm{Var}\{H_{T2_{\hat{U}_n}}(W_1, W_2)\}$$
$$= \frac{4(n-2)(n-1)}{n}\left[E\{(H1_{T2_{\hat{U}_n}})^2\} - EH^2_{T2_{\hat{U}_n}}\right] + \frac{2(n-1)}{n}\left[E[\{H_{T2_{\hat{U}_n}}(W_1, W_2)\}^2] - EH^2_{T2_{\hat{U}_n}}\right].$$

where $E[\{H_{T2_{\hat{U}_n}}(W_1, W_2)\}^2] = O[||\boldsymbol{\beta}||_2^4 \mathrm{tr}\{E(WW^T)^2\}]$. Then, we conclude that

$$E(T2_{\hat{U}_n}) \asymp (n-1)||\beta||_2^2 \ \text{ and } \ \mathrm{Var}(T2_{\hat{U}_n}) = O(\max[n||\beta||_2^4, ||\beta||_2^4 \mathrm{tr}\{E(WW^T)^2\}]). \quad \text{(S1)}$$

Let's examine the term $T3_{\hat{U}_n}$. Utilizing similar steps to the derivation in (A2), we have $\|\hat{\Sigma}^T_{z\omega w}\ddot{\gamma}\|_2^2 = O_P(\|\ddot{\gamma}\|_2^2)$.

It can be seen that $\|\tilde{\theta}^T\hat{\Sigma}_{x\omega w}\|_2^2 = O_P\{n^{-1}\max(n\|\beta\|_2^2, \|\beta\|_2^2\sqrt{2\Lambda_W^\varepsilon})\}$ because $\|\tilde{\theta}^T\hat{\Sigma}_{x\omega w}\|_2^2 = n^{-1}[T2_{\hat{U}_n} + 1/n\sum_{i=1}^n\{(\mu_i - \mu_{\o i})^2 w_i^T w_i\}]$ and $1/n\sum_{i=1}^n\{(\mu_i - \mu_{\o i})^2 w_i^T w_i\} = O_P(\|\beta\|_2^2\sqrt{2\Lambda_W^\varepsilon})$.

Applying Cauchy-Schwarz inequality, it can be seen that $|\tilde{\theta}^T\hat{\Sigma}_{x\omega w}\hat{\Sigma}^T_{z\omega w}\ddot{\gamma}| \le \|\tilde{\theta}^T\hat{\Sigma}_{x\omega w}\|_2\|\hat{\Sigma}^T_{z\omega w}\ddot{\gamma}\|_2$. Following the disassembly of the term $III_{\hat{U}_n}$ in the proof of Theorem 1, we can derive

$$T3_{\hat{U}_n} = o_P\{\max(n\|\beta\|_2, \sqrt{n}\|\beta\|_2\sqrt[4]{2\Lambda_W^\varepsilon}, \sqrt{2\Lambda_W^\varepsilon})\} \quad \text{(S2)}$$

Let b be a vector of size p_β not orthogonal to $\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}$, and satisfying $d_{11} \le \|b\|_1 \le D_{11}$ and $d_{11} \le \|b\|_\infty \le D_{11}$ for some positive constants d_{11} and D_{11}. Utilizing similar steps to the derivation in (A2), we have

$$\left|\frac{\ddot{\gamma}^T\Sigma_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| - \tau\frac{\|\ddot{\gamma}\|_1\|b\|_1}{\|\ddot{\gamma}\|_2\|b\|_2} \le \left|\frac{\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| \le \left|\frac{\ddot{\gamma}^T\Sigma_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| + \tau\frac{\|\ddot{\gamma}\|_1\|b\|_1}{\|\ddot{\gamma}\|_2\|b\|_2}.$$

It indicates that

$$d_{12} \le |\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}b|/(\|\ddot{\gamma}\|_2\|b\|_2) \le D_{12} \quad \text{(S3)}$$

for some positive constants d_{12} and D_{12}, and further implies that $d_{12} \le |\ddot{\gamma}^T \hat{\Sigma}_{z\omega w} \hat{\Sigma}_{z\omega w}^T \ddot{\gamma}| / \|\ddot{\gamma}\|_2^2 \le D_{12}$. Following the disassembly of the term $II_{\hat{U}_n}$ in the proof of Theorem 1, by Assumption 6, (S3) implies that

$$T4_{\hat{U}_n} = \max\{O_p(s_\gamma^{\o} \log p_\gamma), o_P(\sqrt{2\Lambda_W^\varepsilon})\}). \tag{S4}$$

We then examine the term $T5_{\hat{U}_n}$, which is a U-statistic. Let $H_{T5_{\hat{U}_n}}(W_1, W_2) = (\mu_{\o 1} - \mu_1)\varepsilon_2 w_1^T w_2$, $EH_{T5_{\hat{U}_n}} = E\{H_{T5_{\hat{U}_n}}(W_1, W_2)\}$, and $H1_{T5_{\hat{U}_n}} = E\{H_{T5_{\hat{U}_n}}(W_1, W_2)|W_1\} = (\mu_{\o 1} - \mu_1) w_1^T E(\varepsilon W)$. By the independence of terms in $H_{T5_{\hat{U}_n}}(W_1, W_2)$, it can be seen that $E(T5_{\hat{U}_n}) = 0$.

Following Hoeffding decomposition in the variance evaluation of U-statistic (Hoeffding, 1948), we have

$$\mathrm{Var}(T5_{\hat{U}_n}) = \frac{4(n-2)(n-1)}{n}\left[E\{(H1_{T5_{\hat{U}_n}})^2\} - EH_{T5_{\hat{U}_n}}^2\right] + \frac{2(n-1)}{n}\left[E[\{H_{T5_{\hat{U}_n}}(W_1, W_2)\}^2] - EH_{T5_{\hat{U}_n}}^2\right].$$

The dominating term in $\mathrm{Var}(T5_{\hat{U}_n})$ is $\frac{2(n-1)}{n} E[\{H_{T5_{\hat{U}_n}}(W_1, W_2)\}^2]$ and $E[\{H_{T5_{\hat{U}_n}}(W_1, W_2)\}^2] = O(||\beta||_2^2 \Lambda_W^\varepsilon)$, we have

$$E(T5_{\hat{U}_n}) = 0 \text{ and } \mathrm{Var}(T5_{\hat{U}_n}) = O(||\beta||_2^2 \Lambda_W^\varepsilon). \tag{S5}$$

For the term $T6_{\hat{U}_n}$, We have:

$$\begin{aligned} \frac{T6_{\hat{U}_n}}{n} &= \left[\frac{1}{n}\sum_{i=1}^n \{(\mu_{\o i} - \hat{\mu}_{\o i}) w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^n \{w_i(y_i - \mu_i)\}\right] - \frac{1}{n^2}\sum_{i=1}^n \{(y_i - \mu_i)(\mu_{\o i} - \hat{\mu}_{\o i}) w_i^T w_i\} \\ &= \underbrace{(\hat{\gamma}_{\o} - \gamma_{\o})^T \left[\frac{1}{n}\sum_{i=1}^n \{z_i \omega_{\gamma^* i} w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^n \{w_i(y_i - \mu_i)\}\right]}_{T6_1} - \underbrace{\frac{1}{n^2}\sum_{i=1}^n \{(y_i - \mu_i)(\mu_{\o i} - \hat{\mu}_{\o i}) w_i^T w_i\}}_{T6_2}. \end{aligned}$$

Denote $\varsigma = \frac{1}{n}\sum_{i=1}^n \{w_i(y_i - \mu_i)\}$, it can be seen $\|\varsigma\|_2^2 = O_P(n^{-1}\sqrt{2\Lambda_W^\varepsilon})$ because $\|\varsigma\|_2^2 = n^{-1}[I_{\hat{U}_n} + 1/n\sum_{i=1}^n \{(y_i - \mu_i)^2 w_i^T w_i\}]$, $I_{\hat{U}_n} = O_P(\sqrt{2\Lambda_W^\varepsilon})$, and $1/n\sum_{i=1}^n \{(y_i - \mu_i)^2 w_i^T w_i\} = O_P(\sqrt{2\Lambda_W^\varepsilon})$. Following the same argument as (A2) and applying Cauchy-Schwarz inequality, we have

$$|T6_1| = |\ddot{\gamma}^T \hat{\Sigma}_{z\omega w} \varsigma| \le \|\ddot{\gamma}^T \hat{\Sigma}_{z\omega w}\|_2 \|\varsigma\|_2 \le D_{13} \|\ddot{\gamma}\|_2 \|\varsigma\|_2$$

for a positive constant D_{13}, which implies $T6_1 = O_P(n^{-1}\sqrt{s_\gamma^{\o} \log p_\gamma}\sqrt[4]{2\Lambda_W^\varepsilon})$ by Assumption 6. For $T6_2$, we have

$$nT6_2 \le \frac{1}{n}\sum_{i=1}^n \{|(y_i - \mu_i)(\mu_{\o i} - \hat{\mu}_{\o i})| w_i^T w_i\} \le \|\underline{y} - \underline{\mu}\|_\infty \|\underline{\mu}_{\o} - \underline{\hat{\mu}_{\o}}\|_\infty \frac{1}{n}\sum_{i=1}^n (w_i^T w_i) = o_P(\sqrt{2\Lambda_W^\varepsilon}),$$

since $\|\underline{y} - \underline{\mu}\|_\infty = O_P(1)$ and $\|\underline{\mu}_{\o} - \underline{\hat{\mu}_{\o}}\|_\infty = o_P(1)$. In summary, combining the bounds of $T6_1$ and $T6_2$, we have

$$T6_{\hat{U}_n} = \max\{O_P(\sqrt{s_\gamma^{\o} \log p_\gamma}\sqrt[4]{2\Lambda_W^\varepsilon}), o_P(\sqrt{2\Lambda_W^\varepsilon})\} \tag{S6}$$

Based on (S1), (S2), (S4), (S5), (S6), we can see that the dominating terms in $\hat{U}_n$ is $T2_{\hat{U}_n}$ or $T4_{\hat{U}_n}$ depending on the magnitude of $n\|\beta\|_2^2$ and $s_\gamma^{\o} \log p_\gamma$.

We can also disassemble $\hat{R}_n$ as

$$
\begin{aligned}
\hat{R}_n = {} & \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(y_j-\mu_j)^2(w_i^Tw_j)^2\}}_{T1_{\hat{R}_n}} + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(y_j-\mu_j)(\mu_j-\mu_{\varnothing j})(w_i^Tw_j)^2\}}_{T2_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(y_j-\mu_j)(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T3_{\hat{R}_n}} + \underbrace{\frac{2}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(\mu_j-\mu_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T4_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(\mu_j-\mu_{\varnothing j})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T5_{\hat{R}_n}} + \underbrace{\frac{2}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T6_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_i-\mu_{\varnothing i})(\mu_j-\mu_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T7_{\hat{R}_n}} + \underbrace{\frac{8}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_i-\mu_{\varnothing i})(\mu_j-\mu_{\varnothing j})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T8_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_i-\mu_{\varnothing i})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T9_{\hat{R}_n}} + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})(\mu_j-\mu_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T10_{\hat{R}_n}} \\
& + \underbrace{\frac{8}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})(\mu_j-\mu_{\varnothing j})(\mu_{\varnothing j}^0-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T11_{\hat{R}_n}} + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T12_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(y_j-\mu_j)(\mu_i-\mu_{\varnothing i})(\mu_j-\mu_{\varnothing j})(w_i^Tw_j)^2\}}_{T13_{\hat{R}_n}} + \underbrace{\frac{8}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(y_j-\mu_j)(\mu_i-\mu_{\varnothing i})(\mu_j-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T14_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(y_j-\mu_j)(\mu_i-\hat{\mu}_{\varnothing i})(\mu_j-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T15_{\hat{R}_n}} + \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\varnothing i})^2(\mu_j-\mu_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T16_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\varnothing i})^2(\mu_j-\mu_{\varnothing j})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T17_{\hat{R}_n}} + \underbrace{\frac{2}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\varnothing i})^2(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T18_{\hat{R}_n}} \\
& + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_i-\mu_{\varnothing i})(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})(\mu_j-\mu_{\varnothing j})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{T19_{\hat{R}_n}} + \underbrace{\frac{4}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_{0i}-\mu_{\varnothing i})(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T20_{\hat{R}_n}} \\
& + \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_{\varnothing i}-\hat{\mu}_{\varnothing i})^2(\mu_{\varnothing j}-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{T21_{\hat{R}_n}},
\end{aligned}
$$

which means the terms $T1_{\hat{R}_n}$, $T2_{\hat{R}_n}$, $T4_{\hat{R}_n}$, $T7_{\hat{R}_n}$, $T13_{\hat{R}_n}$, and $T16_{\hat{R}_n}$ determines the magnitude of $\hat{R}_n$.

Furthermore, utilizing similar derivations used for Theorem 1, it can be found that the dominating terms in $\hat{R}_n$ are $T1_{\hat{R}_n}$ and $T16_{\hat{R}_n}$ depending on the magnitude of $||\beta||_2$. That is

$$
\begin{cases}
\hat{R}_n = T1_{\hat{R}_n}\{1+o(1)\} & \text{if} \quad ||\beta||_2 = o(1) \\
\hat{R}_n = (T1_{\hat{R}_n} + T16_{\hat{R}_n})\{D_{14}+o(1)\} & \text{if} \quad ||\beta||_2 \asymp 1
\end{cases},
$$

for a positive constant D_{14}. It is straight to see that

$$
E(T1_{\hat{R}_n}) = \Lambda_W^{\varepsilon} \text{ and } \mathrm{Var}(T1_{\hat{R}_n}) = O(n^{-1}\mathrm{tr}[E\{\mathrm{var}(\varepsilon)WW^T\}^4]),
$$

and

$$E(T16_{\hat{R}_n}) = \mathrm{tr}[E\{(\mu-\mu_\o)^2WW^T\}^2] \text{ and } \mathrm{Var}(T16_{\hat{R}_n}) = O(n^{-1}\mathrm{tr}[E\{(\mu-\mu_\o)^2WW^T\}^4]),$$

which implies

$$\begin{cases} \frac{\hat{R}_n}{\Lambda_W^\varepsilon} \xrightarrow{p} 1 & \text{if} \quad ||\beta||_2 = o(1) \\ \frac{\hat{R}_n}{D_{14}[\Lambda_W^\varepsilon + E\{(\mu-\mu_\o)^2WW^T\}^2]} \xrightarrow{p} 1 & \text{if} \quad ||\beta||_2 \asymp 1 \end{cases}. \tag{S7}$$

It can be seen that $E\{(\mu-\mu_\o)^2WW^T\}^2 = O(\Lambda_W^\varepsilon)$ when $||\beta||_2 \asymp 1$.

When $n\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon} = o(1)$, $\|\beta\|_2 = o(1)$, and $s_\gamma^\o \log p_\gamma/\sqrt{2\Lambda_W^\varepsilon} = o(1)$, the terms $T2_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$, $T3_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$, $T4_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$, $T5_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$, and $T6_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$ vanish. When $n\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon} \to \infty$ and $\sqrt{n}\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon} = O(1)$, the term $T2_{\hat{U}_n}/\sqrt{2\Lambda_W^\varepsilon}$ has a diverging mean and bounded variance.

In summary, combining results (S1), (S2), (S4), (S5), (S6), and (S7), we have

$$\lim_{n\to\infty} \inf_{||\theta||_2 = O(1)} P\left(\frac{|\hat{U}_n|}{\sqrt{2\hat{R}_n}} > z_{1-\alpha/2}\right) = \begin{cases} \alpha \ \text{ if } \ \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = o(1), \|\beta\|_2 = o(1), \text{ and } \frac{s_\gamma^\o \log p_\gamma}{\sqrt{2\Lambda_W^\varepsilon}} = o(1); \\ 1 \ \text{ if } \ \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} \to \infty \text{ and } \frac{\sqrt{n}\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = O(1). \end{cases}$$

LEMMA S.1

LEMMA S1. *Under Assumptions 1, 2, and 3,*

$$\frac{\frac{1}{n}\sum_{i\neq j}^n\{(y_i-\mu_i)(y_j-\mu_j)w_i^Tw_j\}}{\sqrt{2\Lambda_W^\varepsilon}} \to N(0,1) \text{ as } n \to \infty.$$

Proof: Denote $\sqrt{\Lambda_W^\varepsilon} = \sigma$, $S_{nkl} = \frac{1}{n\sigma}\{\sum_{i=2}^{k-1}\sum_{j=1}^{i-1}(\varepsilon_i\varepsilon_j w_i^T w_j) + \sum_{j=1,l<k}^{l}(\varepsilon_k\varepsilon_j w_k^T w_j)\}$. It can be seen that $\{S_{nkl}, \mathscr{F}_{nkl}, (1 \le k \le n, 1 \le l < k)\}$ is a martingale. The martingale difference is: $R_{nkl} = S_{nkl} - S_{n,k,l-1} = (\varepsilon_k\varepsilon_l w_k^T w_l)/(n\sigma)$.

We first build

$$\text{for all } \varsigma > 0, \ \sum_k\sum_{l<k} E\{R_{nkl}^2 I(|R_{nkl}| > \varsigma)\} \to 0, \tag{S8}$$

because

$$\sum_k\sum_{l<k} E\{R_{nkl}^2 I(|R_{nkl}| > \varsigma)\} \le \frac{1}{\varsigma^2}\sum_k\sum_{l<k} E(R_{nkl}^4) \ \text{ and}$$
$$\sum_k\sum_{l<k} E(R_{nkl}^4) = \frac{1}{n^4\sigma^4}\sum_k\sum_{l<k} E\{(\varepsilon_k\varepsilon_l w_k^T w_l)^4\} = O(n^{-2}).$$

It is also straight to see that

$$\sum_k\sum_{l<k} R_{nkl}^2 \to 1 \tag{S9}$$

Combining (S8) and (S9), following martingale central limit theorem (Hall & Heyde, 1980), we concludes the asymptotic normality.

LEMMA S.2

Considering an L_q-ball of radius R_q, given by

$$\mathbb{B}_q(R_q) = \left\{\gamma_\o \in \mathbb{R}^{p_\gamma} \,\middle|\, \sum_{i=1}^{p_\gamma} |\gamma_{\o i}|^q \le R_q\right\},$$

where $q \in [0, 1]$ controls the relative "sparsifiability" of $\gamma_{\emptyset}$, with larger values corresponding to lesser sparsity. In a special case that $q = 0$, this set corresponds to an exact sparsity constraint. For a threshold $\eta > 0$, define the thresholded subset $S_\eta = \{j||\gamma_{\emptyset j}| > \eta\}$, and denote s_η the number of elements in S_η.

*Assumption S*1. Under H_a, the estimate $\hat{\gamma}_{\emptyset}$ satisfies $||\hat{\gamma}_{\emptyset} - \gamma_{\emptyset}||_2 = O_P\{R_q^{1/2}(\log p_\gamma/n)^{1/2-q/4}\}$, and $R_q^{1/2}(\log p_\gamma/n)^{1/2-q/4} = o(1)$.

*Assumption S*2. Under H_a, with probability close to 1, $\|\ddot{\gamma}^{S_\eta^c}\|_1 \le 3\|\ddot{\gamma}^{S_\eta}\|_1 + 4\|\gamma_{\emptyset}^{S_\eta^c}\|_1$, where $\ddot{\gamma} = \hat{\gamma}_{\emptyset} - \gamma_{\emptyset}$.

LEMMA S2. *Under H_a, if Assumptions 1, 2, 7, 8, S1, and S2 hold, and $R_q(\log p_\gamma)^{-q/2}n^{q/2-1}\log p_\beta = o(1)$, then*

$$\lim_{n\to\infty} \inf_{||\theta||_2=O(1)} P\left(\frac{|\hat{U}_n|}{\sqrt{2\hat{R}_n}} > z_{1-\alpha/2}\right) = \begin{cases} \alpha \text{ if } \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = o(1), \|\beta\|_2 = o(1) \text{ and } \frac{R_q(\log p_\gamma)^{1-q/2}n^{q/2}}{\sqrt{2\Lambda_W^\varepsilon}} = o(1); \\ 1 \text{ if } \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} \to \infty \text{ and } \frac{\sqrt{n}\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = O(1). \end{cases}$$

Proof: Negahban & Ravikumar (2010) and Negahban et al. (2012) provided a unified framework to establish consistency and convergence rates for regularized high-dimensional M-estimators, which paved the way to build the convergence rate for the generalized linear model under the weak sparsity. Based on their works, we will show that the estimate $\hat{\gamma}_{\emptyset}$ under H_a satisfies the rate in Assumption S1: $||\hat{\gamma}_{\emptyset} - \gamma_{\emptyset}||_2 = O_P\{R_q^{1/2}(\log p_\gamma/n)^{1/2-q/4}\}$.

We have $|S_\eta| \le R_q\eta^{-q}$ for $\eta > 0$ because $R_q \ge \sum_{j=1}^{p_\gamma}|\gamma_{\emptyset j}|^q \ge \sum_{j\in S_\eta}|\gamma_{\emptyset j}|^q \ge \eta^q|S_\eta|$, and $\|\gamma_{\emptyset}^{S_\eta^c}\|_1 = \sum_{j\in S_\eta^c}|\gamma_{\emptyset j}|^q|\gamma_{\emptyset j}|^{1-q} \le R_q\eta^{1-q}$. Combining these results with Assumption S2, we have

$$\|\ddot{\gamma}\|_1 \le 4\|\ddot{\gamma}^{S_\eta}\|_1 + 4\|\gamma_{\emptyset}^{S_\eta^c}\|_1 \le \sqrt{|S_\eta|}\|\ddot{\gamma}\|_2 + 4R_q\eta^{1-q} \le 4\sqrt{R_q}\eta^{-q/2}\|\ddot{\gamma}\|_2 + 4R_q\eta^{1-q}. \tag{S10}$$

The expression (3) can be represented as:

$$\hat{\gamma}_{\emptyset} = \underset{\gamma}{\operatorname{argmin}}\left[\mathcal{L}\{\gamma; (\mathsf{Y}, \mathsf{Z})\} + \zeta||\gamma||_1\right],$$

where Y is the vector of response and Z is $n \times p_\gamma$ design matrix. By the first-order Taylor series expansion at $\gamma_{\emptyset}$ and in the direction Δ, we define

$$\delta\mathcal{L}(\Delta, \gamma_{\emptyset}) = \mathcal{L}\{\gamma_{\emptyset} + \Delta; (\mathsf{Y}, \mathsf{Z})\} - \mathcal{L}\{\gamma_{\emptyset}; (\mathsf{Y}, \mathsf{Z})\} - \langle\bigtriangledown\mathcal{L}\{\gamma_{\emptyset}; (\mathsf{Y}, \mathsf{Z})\}, \Delta\rangle,$$

where $\langle\cdot, \cdot\rangle$ is the Euclidean inner product and $\bigtriangledown\mathcal{L}$ is the gradient.

By Proposition 2 in Negahban & Ravikumar (2010) and Assumption 3, it can be seen that there exists positive constants D_{15} and D_{16} such that:

$$P\left\{\delta\mathcal{L}(\Delta, \gamma_{\emptyset}) \ge D_{15}\|\Delta\|_2\left(\|\Delta\|_2 - D_{16}\sqrt{\frac{\log p_\gamma}{n}}\|\Delta\|_1\right)\right\} \to 1 \text{ for } \|\Delta\|_2 \le 1.$$

Let $\eta = \zeta/D_{15}$ and $\zeta = D_{17}\sqrt{\log p_\gamma/n}$ for a positive constant D_{17}. If Δ satisfies Assumption S2, we can derive from (S10) that

$$\begin{aligned} \delta\mathcal{L}(\Delta, \gamma_{\emptyset}) &\ge D_{15}\|\Delta\|_2^2 - D_{15}D_{16}\sqrt{\frac{\log p_\gamma}{n}}\|\Delta\|_2(4\sqrt{R_q}\eta^{-q/2}\|\Delta\|_2 + 4R_q\eta^{1-q}) \\ &= \left(D_{15} - 4D_{15}D_{16}\sqrt{\frac{\log p_\gamma}{n}}\sqrt{R_q}\eta^{-q/2}\right)\|\Delta\|_2^2 - 4D_{15}D_{16}\sqrt{\frac{\log p_\gamma}{n}}R_q\eta^{1-q}\|\Delta\|_2 \\ &= \left\{D_{15} - 4D_{15}^{1+q/2}D_{16}D_{17}\sqrt{R_q}\left(\sqrt{\frac{\log p_\gamma}{n}}\right)^{1-q/2}\right\}\|\Delta\|_2^2 - 4D_{15}^qD_{16}D_{17}R_q\left(\frac{\log p_\gamma}{n}\right)^{1-q/2}\|\Delta\|_2. \end{aligned}$$

It implies the restricted strong convexity condition (Definition 2 in Negahban et al. (2012)) is satisfied when $(\log p_\gamma/n)^{1-q/2}R_q \leq 1$.

Based on Lemma 6 in Negahban et al. (2012), Assumptions 2, 3, and 7 implies

$$P\left[\| \bigtriangledown \mathcal{L}\{\gamma_\varnothing; (\mathsf{Y},\mathsf{Z})\}\|_\infty \geq D_{18}\sqrt{\log p_\gamma/n}\right] \to 0, \tag{S11}$$

for a positive constant D_{18}.

In summary, by the restricted strong convexity condition and the bound in (S11), and with the choice of $\zeta = D_{17}\sqrt{\log p_\gamma/n}$, the estimate $\hat{\gamma}_\varnothing$ has the convergence rate of $||\hat{\gamma}_\varnothing - \gamma_\varnothing||_2 = O_P\{R_q^{1/2}(\log p_\gamma/n)^{1/2-q/4}\}$ based on Theorem 1 in Negahban et al. (2012).

Let b be a vector of size p_β not orthogonal to $\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}$, not orthogonal to $\tilde{\theta}^T\Sigma_{x\omega w}$, and satisfying $d_{19} \leq \|b\|_1 \leq D_{19}$ and $d_{19} \leq \|b\|_\infty \leq D_{19}$ for constants d_{19} and D_{19}. Assuming $[s_\eta\{\log\max(p_\gamma, p_\beta)\}]/n = o(1)$, we have

$$\tau\frac{\|\ddot{\gamma}\|_1\|b\|_1}{\|\ddot{\gamma}\|_2\|b\|_2} \leq \tau\frac{4\|\ddot{\gamma}^{S_\eta}\|_1 + 4\|\gamma_\varnothing^{S_\eta^c}\|_1}{\|\ddot{\gamma}^{S_\eta}\|_2} \leq o(1) + \tau\frac{4\|\gamma_\varnothing^{S_\eta^c}\|_1}{\|\ddot{\gamma}^{S_\eta}\|_2} \leq o(1) + D_{20}\tau\sqrt{R_q}\eta^{-q/2} = o(1),$$

for a constant D_{20}, because $\|\gamma_\varnothing^{S_\eta^c}\|_1 \leq R_q\eta^{1-q}$ and the rate $R_q(\log p_\gamma/n)^{1-q/2}$ is minimax-optimal over L_q-balls (Raskutti et al., 2011; Negahban et al., 2012). We further derive

$$\left|\frac{\ddot{\gamma}^T\Sigma_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| - \tau\frac{\|\ddot{\gamma}\|_1\|b\|_1}{\|\ddot{\gamma}\|_2\|b\|_2} \leq \left|\frac{\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| \leq \left|\frac{\ddot{\gamma}^T\Sigma_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| + \tau\frac{\|\ddot{\gamma}\|_1\|b\|_1}{\|\ddot{\gamma}\|_2\|b\|_2},$$

which indicates

$$d_{21} \leq \left|\frac{\ddot{\gamma}^T\hat{\Sigma}_{z\omega w}b}{\|\ddot{\gamma}\|_2\|b\|_2}\right| \leq D_{21}, \tag{S12}$$

for constants d_{21} and D_{21}. Equation (S12) here agrees with equation (S3) used for proving Theorem 3. Under the weakly sparsity assumption on $\gamma_\varnothing$, (S4) can be represented as

$$T4_{\hat{U}_n} = \max\{O_p(R_q(\log p_\gamma)^{1-q/2}n^{q/2}), o_P(\sqrt{2\Lambda_W^\varepsilon})\}. \tag{S13}$$

The rest proof follows the same argument as in the proof of Theorem 3. We omit the details here.

Comparison with the method in the literature

Wu et al. (2020) proposed the adaptive interaction sum of powered score test for the model setting as ours. To maintain high statistical power across a wide range of alternatives, their approach largely depends on a good choice of the power index. However, in practice, the optimal choice is unknown and often relies on some ad-hoc method to find it. To compare this method with ours, we apply it to Scenario 1 of our simulation studies. We implement their method by using their code shared at https://github.com/ChongWu-Biostat/aispu. We followed their recommendation of the choice of power index: "*In summary, we recommend use* $\Gamma = \{1, 2, \ldots, 6, \infty\}$ *as our default setting*" stated on page 10 of their paper. The results are shown in Figure S1 and S2 in below. In the linear regression setting, our method produced higher power, while the method of Wu et al. (2020) provided higher power in the logistic regression setting. However, the method of Wu et al. paid with price of severely inflating type I error to achieve a good performance in power. In both linear and logistic regressions, for choosing $\alpha = 0.05$, the empirical type I error rates of the method of Wu et al. (2020) are larger than 0.25. In contrast, the empirical type I error rates of our method are close to the theoretical value of 0.05.

In addition, the proposed method has computational advantage over the method by Wu et al. (2020). For example, for 10 random data sets generated from linear regression (n=200 and p=2000) in simulation Scenario 1, the average running time was 5.69 seconds for the proposed method, and increased to 3.75

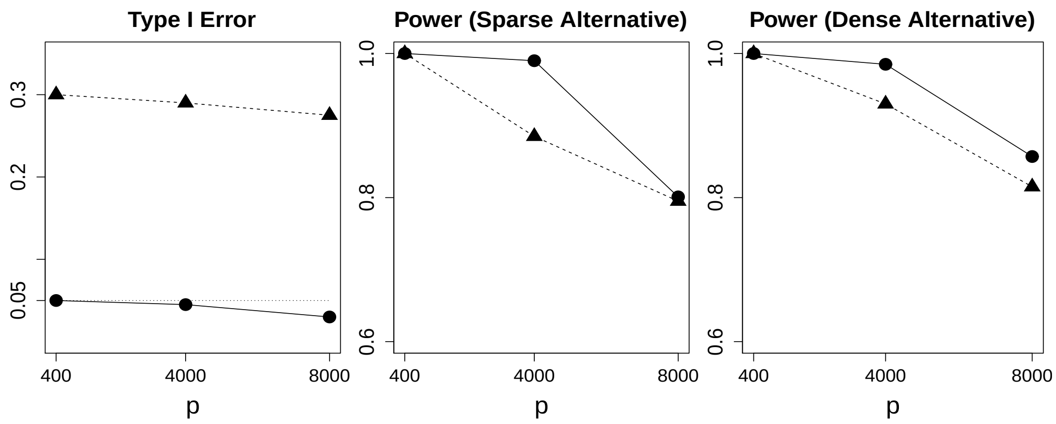

Fig. S1: Testing Results for Data Simulated from Linear Regression in Scenario 1. Solid Circle: results for the proposed method; Solid Triangle: results for the method by Wu et al. (2020).

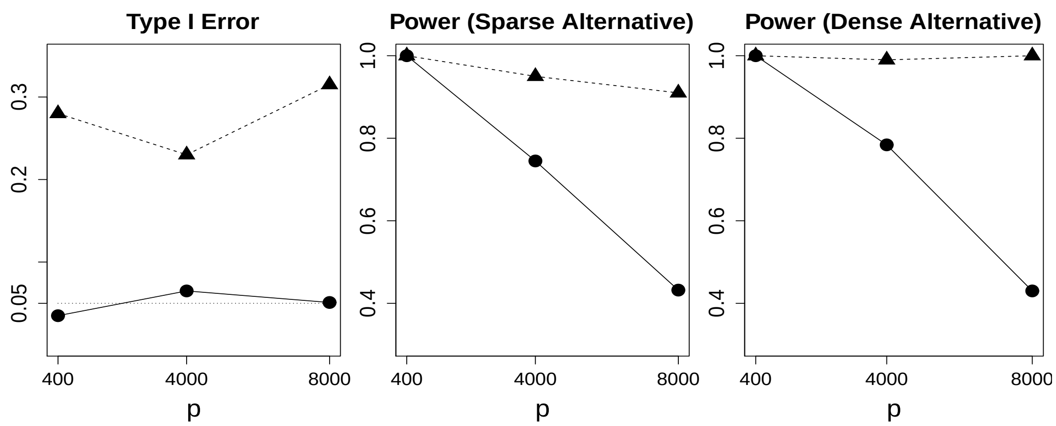

Fig. S2: Testing Results for Data Simulated from Logistic Regression in Scenario 1. Solid Circle: results for the proposed method; Solid Triangle: results for the method by Wu et al. (2020).

minutes for the method of Wu et al. (Intel(R) Xeon(R) CPU E5-1650 v3 @ 3.50GHz 3.50 GHz and 32.0 GB RAM).

REFERENCES

HALL, P. & HEYDE, C. C. (1980). *Martingale Limit Theory and its Application.* London: Academic Press.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

NEGAHBAN, S. & RAVIKUMAR, P. (2010). A unified framework for high-dimentional analysis of M-estimators with decomposable regularizers. *arXiv*: 1010.27318v1.

NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M., YU, B. (2012). A unified framework for high-dimentional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

RASKUTTI, G., WAINWRIGHT, M., YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Trans. Inform. Theory* **57**, 6976–6994.

WU, C., XU, G., SHEN, X., CHENG, G. (2020). A regularization-based adaptive test for high-dimensional generalized linear models. *Journal of Machine Learning Research* **21**, 1–67.