Supplementary Materials for "Model Selection of Generalized Estimating Equation with Divergent Model Size"

Shicheng Wu[1], Xin Gao[1] and [2]Raymond J. Carroll

[1] *York University and* [2] *Texas A&M University*

Supplementary Material

This online supplementary file contains the proofs of Lemmas 1 - 8 in the main paper and some technical lemmas S2.1 - S2.8 and their proofs. In the following proofs, we assume $m_i = m$ for simplicity.

S1 Proofs of Lemmas 1 - 7

Proof of Lemma 1. This lemma is similar to Lemma A6 of Gao and Carroll (2017). Because $Q(\beta_s)$ satisfies the cumulant boundedness condition, its first and second moments are uniformly bounded. Given a model s, by Lemma A5 of Gao and Carroll (2017), $\Pr(\sum_{i=1}^{n}[Q_i(\beta_s) - \mathrm{E}\{Q_i(\beta_s)\}])/\mathrm{Var}\{Q_i(\beta_s)\} > (2np_n \log p_n)^{1/2}] = o(p_n^{-p_n})$. According to Bonferroni inequality,

$$\Pr(\max_{s\in S}\sum_{i=1}^{n}[Q_i(\beta_s)-\mathrm{E}\{Q_i(\beta_s)\}])>b_{var}(2np_n\log p_n)^{1/2})\leq o(p_n^{-p_n})2^{p_n}=0,$$

as there are 2^{p_n} models in the model space, and b_{var} is the upper bound for $\mathrm{Var}\{Q_i(\beta_s)\}$. Similar arguments apply to the result for each element of the score function, and its first and second derivatives.

□

Proof of Lemma 2. We know for true model $V_i(\beta_s^*)=V_i(\beta_F^*)$ for $s\in S_+$. Therefore we know that $||\widehat{V}_i^{-1}-V_i(\beta_s^*)^{-1}||_{\max}=||V_i(\widehat{\beta}_F)^{-1}-V_i(\beta_s^*)^{-1}||_{\max}=||V_i(\widehat{\beta}_F)^{-1}-V_i(\beta_F^*)^{-1}||_{\max}=O_p\{(p_n^3\log p_n/n)^{1/2}\}$, according to Lemma S2.5. In addition $||\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}||_{\max}\leq||\widehat{V}_i^{-1}-V_i(\beta_s^*)^{-1}||_{\max}+||V_i(\beta_s^*)^{-1}-V_i(\widehat{\beta}_s)^{-1}||_{\max}=O_p\{(p_n^3\log p_n/n)^{1/2}\}$. This implies:

$$\max|\lambda_{\max}\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\},\lambda_{\min}\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}|=O_p\{(p_n^3\log p_n/n)^{1/2}\},$$

$$\max|\lambda_{\max}\{\widehat{V}_i^{-1}-V_i(\beta_s^*)^{-1}\},\lambda_{\min}(\widehat{V}_i^{-1}-V_i(\beta_s^*)^{-1})|=O_p\{(p_n^3\log p_n/n)^{1/2}\}.$$

□

Proof of Lemma 3. From Taylor expansion around β_s^*, there exists $\widetilde{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that $(1/n)U(\widehat{\beta}_s)_{[r]}=0$. Therefore we know

$$\frac{1}{n}U(\beta_s^*)_{[r]}+\sum_k\frac{1}{n}U(\beta_s^*)^{(1)}_{[rk]}(\widehat{\beta}_s-\beta_s^*)_{[k]}+\sum_{k,l}\frac{1}{n}U(\widetilde{\beta}_s)^{(2)}_{[rkl]}(\widehat{\beta}_s-\beta_s^*)_{[k]}(\widehat{\beta}_s-\beta_s^*)_{[l]}=0.$$

According to Lemma 1 $\max_{s\in S}|(1/n)U(\beta_s^*)^{(1)}_{[rk]}+\Omega(\beta_s^*)_{[rk]}| = \max_{s\in S}|(1/n)U(\beta_s^*)^{(1)}_{[rk]}-\mathrm{E}[U(\beta_s^*)^{(1)}_{[rk]}]| = O_p\{(p_n \log p_n/n)^{1/2}\}$, then we have

$$
\begin{aligned}
&\sum_k \frac{1}{n}U(\beta_s^*)^{(1)}_{[rk]}(\widehat{\beta}_s-\beta_s^*)_{[k]} \\
&= \sum_k [-\Omega(\beta_s^*)_{[rk]} + \{\Omega(\beta_s^*)_{[rk]} + \frac{1}{n}U(\beta_s^*)^{(1)}_{[rk]}\}](\widehat{\beta}_s-\beta_s^*)_{[k]} \\
&= \sum_k [-\Omega(\beta_s^*)_{[rk]} + O_p\{(p_n \log p_n/n)^{1/2}\}](\widehat{\beta}_s-\beta_s^*)_{[k]}.
\end{aligned}
$$

Similarly from Lemma 1, $n^{-1}U(\widetilde{\beta}_s)^{(2)}_{[rkl]} = n^{-1}\mathrm{E}[U(\widetilde{\beta}_s)^{(2)}_{[rkl]}]\{1+o_p(1)\}$. From Assumption 4, $n^{-1}\mathrm{E}[U(\widetilde{\beta}_s)^{(2)}_{[rkl]}]$ is bounded. So $n^{-1}U(\widetilde{\beta}_s)^{(2)}_{[rkl]} = O_p(1)$. According to Theorem 1 $||\widehat{\beta}_s-\beta_s^*|| = O_p\{(p_n^2 \log p_n/n)^{1/2}\}$. Then

$$
\begin{aligned}
&|\sum_l n^{-1}U(\beta_s^*)^{(2)}_{[rkl]}(\widehat{\beta}_s-\beta_s^*)_{[l]}| \\
&\le \max_l |n^{-1}U(\beta_s^*)^{(2)}_{[rkl]}| \sum_l |(\widehat{\beta}_s-\beta_s^*)_{[l]}| \\
&= O_p(1) \times \sum_l |(\widehat{\beta}_s-\beta_s^*)_{[l]}| \\
&\le O_p(1) \times d_s^{1/2} \times ||\widehat{\beta}_s-\beta_s^*|| \\
&= O_p\{(p_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

Combining the second and the third order terms of Taylor expansion, we have

$$
\begin{aligned}
0 &= \frac{1}{n} U(\widehat{\beta}_s)_{[r]} \\
&= \frac{1}{n} U(\beta_s^*)_{[r]} - \sum_k [\Omega(\beta_s^*)_{[rk]} + O_p\{(p_n \log p_n / n)^{1/2}\}](\widehat{\beta}_s - \beta_s^*)_{[k]} \\
&+ \sum_k O_p\{(p_n^3 \log p_n / n)^{1/2}\}(\widehat{\beta}_s - \beta_s^*)_{[k]}.
\end{aligned}
$$

We can reformat it as

$$
\frac{1}{n} U(\beta_s^*) - \{\Omega(\beta_s^*) + Res_d\}(\widehat{\beta}_s - \beta_s^*) = 0,
$$

where Res_d is a matrix that all elements are at order of $O_p\{(p_n^3 \log p_n / n)^{1/2}\}$ uniformly. Let v_{min} be the corresponding eigenvector of smallest eigenvalue $\lambda_{\min}\{\Omega(\beta_s^*)\}$. According to matrix perturbation theory (Stewart, 1990), we have

$$
\begin{aligned}
&\lambda_{\min}\{\Omega(\beta_s^*) + Res_d\} \\
&= \lambda_{\min}\{\Omega(\beta_s^*)\} + v_{min}^T Res_d v_{min} + o(||Res_d||^2) \\
&\geq \lambda_{\min}\{\Omega(\beta_s^*)\} + d_s \times ||Res_d||_{\max} + o(1) \\
&= \lambda_{\min}\{\Omega(\beta_s^*)\} + O_p\{(p_n^5 \log p_n / n)^{1/2}\} + o(1)
\end{aligned}
$$

Since $\lambda_{\min}\{\Omega(\beta_s^*)\} > 0$ and $p_n^5 \log p_n / n \to 0$, we have $\lambda_{\min}\{\Omega(\beta_s^*) + Res_d\} > 0$ and therefore $\Omega(\beta_s^*) + Res_d$ is invertible. This entails

$$
\widehat{\beta}_s - \beta_s^* = \frac{1}{n}\{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*).
$$

□

Proof of Lemma 4. Considering a competing model s.

$$2Q(\beta_s^*) = \sum_{i=1}^{n}\{Y_i - \mu_i(\beta_s^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}$$

$$= \sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s) + \mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s) + \mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}$$

$$= 2Q(\widehat{\beta}_s) + \sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}$$

$$+ 2\sum_{i=1}^{n}\{\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\}.$$

Lemma S2.7 shows that the last term $\sum_{i=1}^{n}\{\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} = n||\beta_s^* - \widehat{\beta}_s||^2 o_p(1)$. We consider the second term. Applying Equation (S2.1) from Lemma S2.3 to the second term, we have

$$\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}$$

$$= \sum_{i=1}^{n}(\beta_s^* - \widehat{\beta}_s)^T\{D_i(\beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}^T \widehat{V}_i^{-1}\{D_i(\beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}(\beta_s^* - \widehat{\beta}_s)$$

$$= (\beta_s^* - \widehat{\beta}_s)^T\{\sum_{i=1}^{n} D_i(\beta_s^*)^T \widehat{V}_i^{-1} D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1} D_i(\beta_s^*)$$

$$+ \frac{1}{4}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}(\beta_s^* - \widehat{\beta}_s)$$

$$= n(\beta_s^* - \widehat{\beta}_s)^T\{\Omega(\beta_s^*) + \frac{1}{n}\sum_{i=1}^{n} D_i(\beta_s^*)^T\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1} D_i(\beta_s^*)$$

$$+ \frac{1}{4}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}(\beta_s^* - \widehat{\beta}_s)$$

$$= n(\beta_s^* - \widehat{\beta}_s)^T\{\Omega(\beta_s^*) + Res_3\}(\beta_s^* - \widehat{\beta}_s).$$

Let $Res_3 = Res_{31} + Res_{32} + Res_{33}$, with $Res_{31} = \sum_{i=1}^{n} D_i(\beta_s^*)^T\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}D_i(\beta_s^*)/n$, $Res_{32} = \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1}\, D_i(\beta_s^*)/n$, and $Res_{33} =$

$\sum_{i=1}^n D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T \widehat{V}_i^{-1}\ D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})/4n$. Let v be a d_s dimensional unit vector with $||v||^2 = 1$. Then

$$
\begin{aligned}
&|v^T Res_{31} v| = |v^T \frac{1}{n}\sum_{i=1}^n D_i(\beta_s^*)^T\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}D_i(\beta_s^*)v| \\
&\leq \max[|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}|] \max_{||v||^2=1}\{v^T\frac{1}{n}\sum_{i=1}^n D_i(\beta_s^*)^T D_i(\beta_s^*)v\} \\
&\leq O_p\{(p_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

From Lemma S2.4, we have $|v^T Res_{32} v| = |v^T(1/n)\sum_{i=1}^n D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T V_i^{-1} D_i(\beta_s^*)v| = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. For Res_{33},

$$
\begin{aligned}
&|v^T Res_{33} v| = |v^T\frac{1}{n}\sum_{i=1}^n D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T V_i^{-1} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T v| \\
&\leq \lambda_{\max_i}(V_i^{-1}) \max_{||v||^2=1}\{v^T\frac{1}{n}\sum_{i=1}^n D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})v\} \\
&= O_p(p_n^3 \log p_n/n).
\end{aligned}
$$

From Assumption 2, the eigenvalues of $\Omega(\beta_s^*)$ are bounded from zero and infinity. We have

$$
\sup_{||v||=1}|v^T\{\Omega(\beta_s^*) + Res_3\}v| = \sup_{||v||=1}|v^T\Omega(\beta_s^*)v|(1 + O_p\{(p_n^3 \log p_n/n)^{1/2}\}).
$$

Combining the above equation and Lemma S2.7, we have $2[Q(\widehat{\beta}_s) - Q(\beta_s^*)] = -n(\beta_s^* - \widehat{\beta}_s)^T\{\Omega(\beta_s^*) + Res_3\}(\beta_s^* - \widehat{\beta}_s) + n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\} = -n(\beta_s^* - \widehat{\beta}_s)^T\Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\}$. According to Lemma 3, $\widehat{\beta}_s - \beta_s^* =$

$\{\Omega(\beta_s^*) + Res_d\}^{-1}U(\beta_s^*)$. We rewrite the equation as

$$2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} = -n(\beta_s^* - \widehat{\beta}_s)^T\Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\}$$

$$= -\frac{1}{n}U(\beta_s^*)^T\{\Omega(\beta_s^*) + Res_d^T\}^{-1}\Omega(\beta_s^*)\{\Omega(\beta_s^*) + Res_d\}^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

$$= -\frac{1}{n}U(\beta_s^*)^T[\{\Omega(\beta_s^*) + Res_d^T\}\Omega(\beta_s^*)^{-1}\{\Omega(\beta_s^*) + Res_d\}]^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

$$= -\frac{1}{n}U(\beta_s^*)^T\{\Omega(\beta_s^*) + Res_d + Res_d^T + Res_d^T\Omega(\beta_s^*)^{-1}Res_d\}^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

Let $Res_s = Res_d + Res_d^T + Res_d^T\Omega(\beta_s^*)^{-1}Res_d$ and we have

$$2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} = -1/nU(\beta_s^*)^T\{\Omega(\beta_s^*) + Res_s\}^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

We estimate the order of the matrix Res_s as follows:

$$\sup_{||v||=1} v^T(Res_d + Res_d^T)v \leq 2\sup_{||v||=1} v^T Res_d v$$

$$\leq \max_{k,r}|[Res_d]_{[kr]}| \times d_s \times ||v||^2$$

$$= O_p\{(p_n^5\log p_n)^{1/2}\};$$

$$\inf_{||v||=1} v^T(Res_d + Res_d^T)v \geq 2\inf_{||v||=1} v^T Res_d v$$

$$\geq -\max_{k,r}|[Res_d]_{[kr]}| \times d_s \times ||v||^2$$

$$= -O_p\{(p_n^5\log p_n)^{1/2}\};$$

$$\sup_{||v||=1} v^T(Res_d^T\Omega(\beta_s^*)^{-1}Res_d)v$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}d_s \times ||Res_d||_{\max}^2 \times d_s \times ||v||^2$$

$$= O_p\{p_n^5\log p_n/n\};$$

$$\inf_{||v||=1} v^T(Res_d^T\Omega(\beta_s^*)^{-1}Res_d)v$$

$$\geq -\lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}d_s \times ||Res_d||_{\max}^2 \times d_s \times ||v||^2$$

$$= -O_p\{p_n^5 \log p_n/n\}.$$

Thus we have $\sup_{||v||=1} |v^T Res_s v| = O_p\{(p_n^5 \log p_n/n)^{1/2}\} = o_p(1)$. This implies that the eigenvalues of Res_s are of the order of $o_p(1)$. It follows that

$$\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_s\}^{-1}]v$$

$$= \sup_{||v||^2=1} v^T\Omega(\beta_s^*)^{-1/2}[I - \{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}]\Omega(\beta_s^*)^{-1/2}v$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}\lambda_{\max}(I - [\{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}])||v||^2$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}(1 - \lambda_{\min}[\{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}])$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}[\frac{\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}}{1 + \lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}}].$$

Furthermore,

$$\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}$$

$$= \sup_{||v||=1} v^T\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}v$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}\lambda_{\max}\{Res_s\}||v||^2 = o_p(1)$$

Thus $\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_s\}^{-1}]v = o_p(1)$. Therefore,

$$2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} = -1/nU(\beta_s^*)^T\{\Omega(\beta_s^*) + Res_s\}^{-1}U(\beta_s^*)\{1 + o_p(1)\}$$

$$= -\frac{1}{n}U(\beta_s^*)^T\Omega(\beta_s^*)^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

□

Proof of Lemma 5. We first consider the true and overfitting situation. By Lemma 4 shows that $|Q(\widehat{\beta}_s) - Q(\beta_s^*)| = (n/2)(\beta_s^* - \widehat{\beta}_s)^T\Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\}$. Theorem 1 shows that $||\beta_s^* - \widehat{\beta}_s|| = O_p\{(p_n^2 \log p_n/n)^{1/2}\}$. And Assumption 2 indicates that all eigenvalue of $\Omega(\beta_s^*)$ is bounded.

$$
\begin{aligned}
&|Q(\widehat{\beta}_s) - Q(\beta_s^*)| = \frac{n}{2}(\beta_s^* - \widehat{\beta}_s)^T\Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\} \\
&\leq \frac{n}{2}\lambda_{\max}\{\Omega(\beta_s^*)\}||\beta_s^* - \widehat{\beta}_s||^2\{1 + o_p(1)\} \\
&= O_p(p_n^2 \log p_n).
\end{aligned}
$$

Then we consider the underfitting situation.

$$
\begin{aligned}
&|2Q(\widehat{\beta}_s) - 2Q(\beta_s^*)| = |\sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} - 2Q(\beta_s^*)| \\
&= |\sum_{i=1}^{n}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} - 2Q(\beta_s^*)| \\
&\leq |\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| + 2|\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}|.
\end{aligned}
$$

We consider the first term from the above formula. According to Taylor expansion, there exists a $\check{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that $\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\check{\beta})(\beta_s^* - \widehat{\beta}_s)$. Then we have

$$
\begin{aligned}
&|\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| \\
&= \sum_{i=1}^{n}(\beta_s^* - \widehat{\beta}_s)^T D_i(\check{\beta})^T\widehat{V}_i^{-1}D_i(\check{\beta})(\beta_s^* - \widehat{\beta}_s) \\
&\leq n\lambda_{\max}\{\widehat{V}_i^{-1}\} \times ||\beta_s^* - \widehat{\beta}_s||^2 \times \max_{||v||^2=1}\{v^T\frac{1}{n}\sum_{i=1}^{n}D_i(\check{\beta})^T D_i(\check{\beta})v\} \\
&= O_p(p_n^2 \log p_n).
\end{aligned}
$$

Next we consider the second term. Lemma S2.6 implies that $\max_j\{(1/n)\sum_{i=1}^n |Y_{ij} - \mu_{ij}(\beta_s^*)|\} = O_p(1)$. Assumption 3 implies that $\max_i\{||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max}\} = O_p(1)$. Combining these results, we have

$$
\begin{aligned}
&|\sum_{i=1}^n \{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}| \\
&= |\sum_{i=1}^n (\beta_s^* - \widehat{\beta}_s)^T D_i(\check{\beta})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}| \\
&\le ||\beta_s^* - \widehat{\beta}_s|| \times ||\sum_{i=1}^n D_i(\check{\beta})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}|| \\
&\le ||\beta_s^* - \widehat{\beta}_s|| \times p_n^{1/2} \max_k |\sum_{i=1}^n [D_i(\check{\beta})^T]_{[k,]} \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}| \\
&\le np_n^{1/2}||\beta_s^* - \widehat{\beta}_s|| \times \frac{1}{n}\sum_{i=1}^n m^2||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max} \times \max_j |Y_{ij} - \mu_{ij}(\beta_s^*)| \\
&\le np_n^{1/2}m^2||\beta_s^* - \widehat{\beta}_s|| \times \max_i\{||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max}\} \times \max_j\{\frac{1}{n}\sum_{i=1}^n |Y_{ij} - \mu_{ij}(\beta_s^*)|\} \\
&= O_p\{(p_n^3 \log p_n/n)^{1/2}\},
\end{aligned}
$$

□

Proof of Lemma 6. From Assumption 2, both $\Omega(\beta_T^*)^{-1}$ and $\Omega(\beta_s^*)^{-1}$ are positive definite. The $\Omega(\beta_T^*)$ is a sub-block of $\Omega(\beta_s^*)$. We define $\Omega = \Omega(\beta_T^*)$ and define block matrix $\Omega(\beta_s^*) = \begin{bmatrix} \Omega & \check{\Omega} \\ \check{\Omega}^T & \widetilde{\Omega} \end{bmatrix}$, where $\widetilde{\Omega}$ is a positive definite $d_s \times d_s$ matrix and $\check{\Omega}$ a $d_s \times (d_s - d_T)$ matrix. For any $d_T \times 1$ vector ι_1 and $(d_s - d_T) \times 1$ vector ι_2, we can show that the matrix $M_{s/T}$ is non-negative definite by the formula below:

$$
\begin{bmatrix} \iota_1 \\ \\ \iota_2 \end{bmatrix}^T
(
\begin{bmatrix} \Omega & \check{\Omega} \\ \\ \check{\Omega}^T & \widetilde{\Omega} \end{bmatrix}^{-1}
-
\begin{bmatrix} \Omega^{-1} & 0 \\ \\ 0 & 0 \end{bmatrix}
)
\begin{bmatrix} \iota_1 \\ \\ \iota_2 \end{bmatrix}
$$

$$
= \begin{bmatrix} \iota_1 \\ \\ \iota_2 \end{bmatrix}^T
\begin{bmatrix} \Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & -\Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \\ \\ -(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & (\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \end{bmatrix}
\begin{bmatrix} \iota_1 \\ \\ \iota_2 \end{bmatrix}
$$

$$
= (\check{\Omega}^T\Omega^{-1}\iota_1 - \iota_2)^T(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}(\check{\Omega}^T\Omega^{-1}\iota_1 - \iota_2)
$$

$$
\geq 0.
$$

□

Proof of Lemma 7. Let $\eta_s = n^{-1/2}W^{-1/2}(\beta_s^*)U(\beta_s^*)$. According to Lemma S2.8, it satisfies the exponential moment condition,

$$
\log[\mathrm{E}\{\exp(t^T\eta_s)\}] \leq a^2||t||^2/2,
$$

with $t \in R^{d_s}$, $||t||^2 \leq p_n^2 \log p_n$ and some constant $a^2 > 1$. We scale the vector η as $\eta^* = \eta/a$, so that $\log[\mathrm{E}\{\exp(t^T\eta^*)\}] \leq ||t||^2/2$ with $||t|| \leq \{a^2p_n^2(\log p_n)\}^{1/2} = a\times\rho$. Given matrix $B_{s/T} = W^{1/2}(\beta_s^*)M_{s/T}(\beta_s^*)W^{1/2}(\beta_s^*)$ and $\mathrm{Tr}(B_{s/T}) = d_s^* - d_T^*$, we define $B_{s/T}^* = B_{s/T}/\tau$ where $\tau = \lambda_{max}(B_{s/T})$. Therefore $\lambda_{max}(B_{s/T}^*) = 1$. We scale the quadratic form $\Delta_s^* = \Delta_s/a^2\tau = (\eta^*)^TB_s^*\eta^*$. Let $\Delta_{s/T} = \eta^TB\eta$ and $\Delta_{s/T}^* = \Delta/a^2\tau = (\eta^*)B^*\eta^*$. Define $P_G = \mathrm{Tr}[B^*] = (d_s^* - d_T^*)/\tau$ and $V_G^2 = \mathrm{Tr}[(B^*)^2]$. Using the inequality for the trace of matrix product, we obtain $V_G \leq (2P_G)^{1/2} = O(p_n)$. We apply the large deviation result from Corollary 4.2 of Spokoiny and Zhilova (2013). For $3/2\rho^2 > K > V_G/3$,

$$\Pr(\Delta^*_{s/T} > P_G + K) \leq 10.4\exp(-K/6).$$

Choosing $K = \{(d^*_s - d^*_T)/\tau\}\{\gamma_n/a^2 - 1\}$ leads to $K = O(p_n \log p_n)$. Given $\rho^2 = O(p_n^2 \log p_n)$ and $V_G = O_p(p_n)$. We have $3/2\rho^2 > K > V_G/3$. Let $\check{\tau} = (d^*_s - d^*_T)/(d_s - d_T)$. We have

$$\begin{aligned}
&\Pr\{\max_{s\in S_+, s\neq T} \Delta_{s/T} > (d^*_s - d^*_T)\gamma_n\} \\
&\leq \sum_{s\in S_+, s\neq T} \Pr\{\Delta^*_{s/T} > [(d^*_s - d^*_T)\gamma_n/(a^2\tau)]\} \\
&\leq \sum_{s\in S_+, s\neq T} \Pr\{\Delta^*_{s/T} > P_G + P_G(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \sum_{s\in S_+, s\neq T} \Pr\{\Delta^*_{s/T} > P_G + K\} \\
&\leq \sum_{s\in S_+, s\neq T} 10.4\exp\{-\frac{(d_s - d_T)\check{\tau}}{6\tau}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \sum_{d_s = d_T + 1}^{p_n} C^{d_s - d_T}_{p_n - d_T} 10.4\exp\{-\frac{(d_s - d_T)\check{\tau}}{6\tau}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \sum_{m'=1}^{p_n - d_T} C^{m'}_{p_n - d_T} 10.4\exp\{-\frac{m'}{6\omega}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \{1 + 10.4\exp(-\frac{\gamma_n/a^2 - 1}{6\omega})\}^{p_n - d_T} - 1.
\end{aligned}$$

As a^2 can be chosen as close to 1 as possible with increasing sample size n, the choices of $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$ or $\gamma_n = 6\omega(1 + \log\log p_n)\log p_n$ lead to $\lim_{n\to\infty}(1 + 10.4\exp\{-(\gamma_n/a^2 - 1)/(6\omega)\})^{p_n - d_T} = 1$.

This entails $\Pr\{\max_{s\in S_+, s\neq T} \Delta_{s/T} > (d_s^* - d_T^*)\gamma_n\} \to 0$. $\square$

Proof of Lemma 8. First we decompose the difference between d_s^* and $\widehat{d}_s$ as follows:

$$
\begin{aligned}
d_s^* - \widehat{d}_s &= \mathrm{Tr}\{W_s(\beta_s^*)\Omega_s^{-1}(\beta_s^*) - W_s(\widehat{\beta}_s)\Omega_s^{-1}(\widehat{\beta}_s)\} \\
&= \mathrm{Tr}[\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\{\Omega_s^{-1}(\beta_s^*) - \Omega_s^{-1}(\widehat{\beta}_s)\}] \\
&+ \mathrm{Tr}[W_s(\widehat{\beta}_s)\{\Omega_s^{-1}(\beta_s^*) - \Omega_s^{-1}(\widehat{\beta}_s)\}] \\
&+ \mathrm{Tr}[\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\Omega_s^{-1}(\widehat{\beta}_s)].
\end{aligned}
$$

Next we will prove that $|\mathrm{Tr}[\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\Omega_s^{-1}(\beta_s^*)]| = O_p\{(p_n^5 \log p_n/n)^{1/2}\}$. Let the subscript $[j,k]$ denote the (j,k)th element of a matrix. The covariance matrix of the score vector can be expressed as

$$
\begin{aligned}
W_s(\beta)_{[jk]} =& n^{-1}\mathrm{Cov}\{U(\beta)_{[j]}, U(\beta)_{[k]}\} \\
=& n^{-1}\sum_{i=1}^{n}\mathrm{Cov}\{U_i(\beta)_{[j]}, U_i(\beta)_{[k]}\} \\
=& n^{-1}\sum_{i=1}^{n}[\mathrm{E}\{U_i(\beta)_{[j]}U_i(\beta)_{[k]}\} - \mathrm{E}\{U_i(\beta)_{[j]}\}\mathrm{E}\{U_i(\beta)_{[k]}\}].
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}_{[jk]} =& n^{-1}\sum_{i=1}^{n}[\mathrm{E}\{U_i(\beta_s^*)_{[j]}U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[j]}U_i(\widehat{\beta}_s)_{[k]}\}] \\
& - n^{-1}\sum_{i=1}^{n}[\mathrm{E}\{U_i(\beta_s^*)_{[j]}\}\mathrm{E}\{U_i(\beta_s^*)_{[k]}\} - \mathrm{E}\{U_i(\widehat{\beta}_s)_{[j]}\}\mathrm{E}\{U_i(\widehat{\beta}_s)_{[k]}\}].
\end{aligned}
$$

For the first component, we have

$$
\begin{aligned}
& U_i(\beta_s^*)_{[j]} U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[j]} U_i(\widehat{\beta}_s)_{[k]} \\
& = U_i(\widehat{\beta}_s)_{[j]} \{U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[k]}\} + \{U_i(\beta_s^*)_{[j]} - U_i(\widehat{\beta}_s)_{[j]}\} U_i(\widehat{\beta}_s)_{[k]} \\
& + \{U_i(\beta_s^*)_{[j]} - U_i(\widehat{\beta}_s)_{[j]}\}\{U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[k]}\}.
\end{aligned}
$$

From Taylor expansion, there exist a $\tilde{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that

$$
\begin{aligned}
|U_i(\beta_s^*)_{[j]} - U_i(\widehat{\beta}_s)_{[j]}| \times |U_i(\beta_s^*)_{[k]}| &= |(\beta_s^* - \widehat{\beta}_s)^T U_i(\tilde{\beta}_s)_{[j]}^{(1)}| \times |U_i(\beta_s^*)_{[k]}| \\
&\leq ||\beta_s^* - \widehat{\beta}_s|| \times ||U_i(\tilde{\beta}_s)_{[j]}^{(1)}|| \times |U_i(\beta_s^*)_{[k]}| \\
&= O_p\{(p_n^3 \log p_n / n)^{1/2}\},
\end{aligned}
$$

where $||\beta_s^* - \widehat{\beta}_s|| = O_p\{(p_n^2 \log p_n/n)^{1/2}\}$ according to Theorem 1, and $|U_i(\beta_s^*)_{[k]}|$ is bounded and each element of p_n vector $U_i(\tilde{\beta}_s)_{[j]}^{(1)}$ is bounded according to Lemma S2.2. Similarly we get $\{U_i(\beta_s^*)_{[j]} - U_i(\widehat{\beta}_s)_{[j]}\} U_i(\widehat{\beta}_s)_{[k]} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$ and $\{U_i(\beta_s^*)_{[j]} - U_i(\widehat{\beta}_s)_{[j]}\}\{U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[k]}\} = O_p \{p_n^3 \log p_n/n\}$. Thus $|U_i(\beta_s^*)_{[j]} U_i(\beta_s^*)_{[k]} - U_i(\widehat{\beta}_s)_{[j]} U_i(\widehat{\beta}_s)_{[k]}| = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$, and $1/n \sum_{i=1}^n [\mathrm{E}\{U_i(\beta_s^*)_{[j]}\}\mathrm{E}\{U_i(\beta_s^*)_{[k]}\} - \mathrm{E}\{U_i(\widehat{\beta}_s)_{[j]}\}\mathrm{E}\{U_i(\widehat{\beta}_s)_{[k]}\}] = O_p \{(p_n^3 \log p_n/n)^{1/2}\}$. Combining the terms above, we have $\max_{j,k}\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}_{[jk]} = O_p \{(p_n^3 \log p_n/n)^{1/2}\}$. Based on Assumption 2, all eigenvalues of matrices $W_s(\beta_s^*)$, $W_s(\widehat{\beta}_s)$ and $\Omega_s^{-1}(\widehat{\beta}_s)$ are bounded. According to Von

Neumann's Trace Inequality, we have

$$
\begin{aligned}
&|\operatorname{Tr}[\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\Omega_s^{-1}(\widehat{\beta}_s)]| \\
&\leq |\sum_{r=1} \lambda_r\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\lambda_r\{\Omega_s^{-1}(\widehat{\beta}_s)\}| \\
&= O_p\{(p_n^5 \log p_n/n)^{1/2}\},
\end{aligned}
$$

where λ_r denotes the rth eigenvalues ordered from the least to the greatest. Next we have

$$
\begin{aligned}
\operatorname{Tr}[W_s(\widehat{\beta}_s)\{\Omega_s^{-1}(\beta_s^*) - \Omega_s^{-1}(\widehat{\beta}_s)\}] &= \operatorname{Tr}[W_s(\widehat{\beta}_s)\Omega_s^{-1}(\widehat{\beta}_s)\{\Omega_s(\widehat{\beta}_s) - \Omega_s(\beta_s^*)\}\Omega_s^{-1}(\beta_s^*)] \\
&= \operatorname{Tr}[\{\Omega_s(\widehat{\beta}_s) - \Omega_s(\beta_s^*)\}\Omega_s^{-1}(\beta_s^*)W_s(\widehat{\beta}_s)\Omega_s^{-1}(\widehat{\beta}_s)].
\end{aligned}
$$

Similarly we obtain $\max_{j,k}|[\Omega_s(\beta_s^*) - \Omega_s(\widehat{\beta}_s)]_{[jk]}| = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. According to Assumption 2, all eigenvalues of $\Omega_s^{-1}(\beta_s^*)W_s(\widehat{\beta}_s)\Omega_s^{-1}(\widehat{\beta}_s)$ are positive and bounded. This entails $|\operatorname{Tr}[W_s(\widehat{\beta}_s)\{\Omega_s^{-1}(\beta_s^*) - \Omega_s^{-1}(\widehat{\beta}_s)\}]| = O_p\{(p_n^5 \log p_n/n)^{1/2}\}$. Following similar argument, we have $|\operatorname{Tr}[\{W_s(\beta_s^*) - W_s(\widehat{\beta}_s)\}\{\Omega_s^{-1}(\beta_s^*) - \Omega_s^{-1}(\widehat{\beta}_s)\}]| = O_p\{(p_n^5 \log p_n/n)^{1/2}\}$. Combining all the results above, we have $|d_s^* - \widehat{d}_s| = O_p\{(p_n^5 \log p_n/n)^{1/2}\}$. As all the asymptotic orders are established uniformly for all model s in the model space, the consistency result of $\widehat{d}_s$ is uniform over the space of all models. $\square$

S2 Some Technical Lemmas

Lemma S2.1. *Let the joint distribution of the observations from the ith cluster follow a canonical multivariate exponential family with the likelihood* $L_i(Y_i, \theta_i) = \exp\{\sum_{j=1}^{m_i} Y_{ij}\theta_{ij} - b(\theta_i) + c(Y_i)\}$, *where* $\theta_i = (\theta_{i1}, \theta_{i2} \dots \theta_{im_i})^T$. *Assume the parameter space for* θ_i *is a compact subspace of* $\mathcal{R}^{m_i}$ *with the true value* θ_i^* *being an interior point of the parameter space, the function* $b(\theta_i)$ *is three times differentiable and bounded on the parameter space and* $Y_{ij}s$ *are sub-Gaussian random variables. Under Assumption 3,* $Q_i(\beta)$, $U_i(\beta)_{[k]}$, $U_i(\beta)^{(1)}_{[kl]}$, *and* $U_i(\beta)^{(2)}_{[klr]}$ *satisfy the cumulant boundedness condition uniformly for all model* $s \in S$ *and all* β *in the neighborhood* $||\beta - \beta_s^*|| \leq (p_n^2 \log p_n/n)^{1/2}$.

Proof of Lemma S2.1. The cumulant generating function of Y_i can be formulated as

$$
\begin{aligned}
C_{Y_i}(t_i) &= \log \mathrm{E}\{\exp(t_i^T \cdot Y_i)\} \\
&= \log \int \exp\{\sum_{j=1}^{m} t_{ij}Y_{ij}\} \exp\{\sum_{j=1}^{m} \theta_{ij}Y_{ij} - b(\theta_i) + c(Y_i)\} dY_i \\
&= \log \int \exp\{(\sum_{j=1}^{m} t_{ij} + \theta_{ij})Y_{ij} - b(\theta_i + t_i) + c(Y_i)\} \exp\{b(\theta_i + t_i) - b(\theta_i)\} dY_i \\
&= b(\theta_i + t_i) - b(\theta_i),
\end{aligned}
$$

with $b(\theta_i + t_i) = b(\theta_{i1} + t_{i1}, \theta_{i2} + t_{i2} \dots \theta_{im} + t_{im})$. Next we consider the cumulant boundedness condition of $U_i(\beta_s^*)_{[k]}$. Let $[R^{-1}]_{[jh]}$ be the (j, h) element of matrix R^{-1}. We define $f^{ik}_{jh}(\beta) = \{\partial\mu_{ij}(\beta)^T/\partial\beta_{[k]}\}A_{ij}(\beta)^{-1/2}[R^{-1}]_{[jh]}A_{ih}(\beta)^{-1/2}$ and $\tilde{f}^{ik}_{jh}(\beta) = \{\partial\mu_{ij}(\beta)^T/\partial\beta_{[k]}\}\, A_{ij}(\beta)^{-1/2}[R^{-1}]_{[jh]}A_{ih}(\beta)^{-1/2}\mu_{ih}(\beta)$. We rewrite $U_i(\beta)_{[k]}$ as follows:

$$
\begin{aligned}
U_i(\beta)_{[k]} &= \frac{\partial\mu_i(\beta)^T}{\partial\beta_{[k]}}V_i(\beta)^{-1}\{Y_i - \mu_i(\beta)\} \\
&= \sum_{j=1}^{m}\sum_{h=1}^{m}\frac{\partial\mu_{ij}(\beta)^T}{\partial\beta_{[k]}}A_{ij}(\beta)^{-1/2}[R^{-1}]_{[jh]}A_{ih}(\beta)^{-1/2}\{Y_{ih} - \mu_{ih}(\beta)\} \\
&= \sum_{j=1}^{m}\sum_{h=1}^{m} f^{ik}_{jh}(\beta)Y_{ih} - \tilde{f}^{ik}_{jh}(\beta).
\end{aligned}
$$

This entails the following expressions:

$$
U_i(\beta)^{(1)}_{[kl]} = \sum_{j=1}^{m}\sum_{h=1}^{m}\frac{\partial f^{ik}_{jh}(\beta)}{\partial\beta_{[l]}}Y_{ih} - \frac{\partial\tilde{f}^{ik}_{jh}(\beta)}{\partial\beta_{[l]}},
$$

$$
U_i(\beta)^{(2)}_{[klr]} = \sum_{j=1}^{m}\sum_{h=1}^{m}\frac{\partial^2 f^{ik}_{jh}(\beta)}{\partial\beta_{[l]}\partial\beta_{[r]}}Y_{ih} - \frac{\partial^2\tilde{f}^{ik}_{jh}(\beta)}{\partial\beta_{[l]}\partial\beta_{[r]}},
$$

We first consider the cumulant generating function of $U_i(\beta)_{[k]}$

$$
\begin{aligned}
&C_{U_i(\beta)_{[k]}}(t) = \log \mathrm{E}[\exp\{tU_i(\beta)_{[k]}\}] \\
&= \log\int \exp\{t\sum_{j=1}^{m}\sum_{h=1}^{m} f^{ik}_{jh}(\beta)Y_{ih} - \tilde{f}^{ik}_{jh}(\beta)\}\exp\{\sum_{j=1}^{m}\theta_{ij}Y_{ij} - b(\theta_i) + c(Y_i)\}dY_i \\
&= \log\int \exp[\sum_{h=1}^{m}\{\theta_{ih} + t\sum_{j=1}^{m} f^{ik}_{jh}(\beta)\}Y_{ih} - b(\theta_i + t^{(0)}) + c(Y_i)] \\
&\qquad \exp\{t\sum_{j=1}^{m}\sum_{h=1}^{m} -\tilde{f}^{ik}_{jh}(\beta) + b(\theta_i + t^{(0)}) - b(\theta_i)\}dY_i \\
&\qquad = b(\theta_i + t^{(0)}) - b(\theta_i) - t\sum_{j=1}^{m}\sum_{h=1}^{m}\tilde{f}^{ik}_{jh}(\beta),
\end{aligned}
$$

with $t^{(0)} = (t\sum_{j=1}^m f_{j1}^{ik}(\beta), t\sum_{j=1}^m f_{j2}^{ik}(\beta) \dots t\sum_{j=1}^m f_{jm}^{ik}(\beta))^T$. Similarly we can calculate the cumulant generating function of $U_i(\beta)_{[kl]}^{(1)}$ and $U_i(\beta)_{[klr]}^{(2)}$ below:

$$C_{U_i(\beta_s^*)_{[kl]}^{(1)}}(t) = b(\theta_i + t^{(1)}) - b(\theta_i) - t\sum_{j=1}^m \sum_{h=1}^m \frac{\partial \tilde{f}_{jh}^{ik}(\beta)}{\partial \beta_{[l]}},$$

$$C_{U_i(\beta_s^*)_{[klr]}^{(2)}}(t) = b(\theta_i + t^{(2)}) - b(\theta_i) - t\sum_{j=1}^m \sum_{h=1}^m \frac{\partial^2 \tilde{f}_{jh}^{ik}(\beta)}{\partial \beta_{[l]}\partial \beta_{[r]}},$$

with

$$t^{(1)} = (t\sum_{j=1}^m \partial f_{j1}^{ik}(\beta)/\partial\beta_{[l]}, t\sum_{j=1}^m \partial f_{j2}^{ik}(\beta)/\partial\beta_{[l]}, \dots, t\sum_{j=1}^m \partial f_{jm}^{ik}(\beta)/\partial\beta_{[l]})^T,$$

$$\begin{aligned} t^{(2)} =&(t\sum_{j=1}^m \partial^2 f_{j1}^{ik}(\beta)/\partial\beta_{[l]}\partial\beta_{[r]}, t\sum_{j=1}^m \partial^2 f_{j2}^{ik}(\beta)/\partial\beta_{[l]}\partial\beta_{[r]}, \dots, \\ & t\sum_{j=1}^m \partial^2 f_{jm}^{ik}(\beta)/\partial\beta_{[l]}\partial\beta_{[r]})^T. \end{aligned}$$

From Assumption 3 and Lemma S2.2, it is known that $f_{jh}^{ik}(\beta)$, $\tilde{f}_{jh}^{ik}(\beta)$, and their derivatives are uniformly bounded when $||\beta - \beta_s^*|| \leq (p_n^2 \log p_n/n)^{1/2}$. Thus there exists a positive number b_t such that $\max\{||t^{(0)}||_{\max}, ||t^{(1)}||_{\max}, ||t^{(2)}||_{\max}\} < b_t|t|$. Then we have $\partial b(\theta_i + t^{(i)})/\partial t \leq m b_t ||\nabla b(\theta_i + t^{(i)})||_{max}$. As the function $b(.)$ is three times differentiable and bounded on the parameter space, we have $\partial b(\theta_i + t^{(i)})/\partial t$, $\partial^2 b(\theta_i + t^{(i)})/(\partial t)^2$ and $\partial^3 b(\theta_i + t^{(i)})/(\partial t)^3$ all uniformly bounded. Similarly, we can verify that the $U_i(\beta_s)_{[k]}$, $U_i(\beta_s)_{[kl]}^{(1)}$ and $U_i(\beta_s)_{[klr]}^{(2)}$ satisfy cumulant boundedness condition in Definition 1.

Let V_{ijh} be the (j,h) element of matrix $\widehat{V}_i^{-1}$. We can rewrite the $Q_i(\beta)$ as follows:

$$
\begin{aligned}
2Q_i(\beta) &= \{Y_i - \mu_i(\beta)\}^T \widehat{V}_i^{-1} \{Y_i - \mu_i(\beta)\} \\
&= \sum_{j=1}^{m}\sum_{h=1}^{m} \{Y_{ij} - \mu_{ij}(\beta)\} V_{ijh} \{Y_{ih} - \mu_{ih}(\beta)\} \\
&= \sum_{j=1}^{m}\sum_{h=1}^{m} V_{ijh} Y_{ij} Y_{ih} + V_{ijh}\mu_{ij}(\beta)\mu_{ih}(\beta) - (V_{ijh} + V_{ihj}) Y_{ij}\mu_{ih}(\beta).
\end{aligned}
$$

Next we consider the cumulant generating function of $Q_i(\beta)$:

$$
\begin{aligned}
&C_{Q_i(\beta)}(t) = \log \mathrm{E}[\exp\{Q_i(\beta)t\}] \\
&= \log \int \exp\{t \sum_{j=1}^{m}\sum_{h=1}^{m} V_{ijh} Y_{ij} Y_{ih} + V_{ijh}\mu_{ij}(\beta)\mu_{ih}(\beta) - (V_{ijh} + V_{ihj}) Y_{ij}\mu_{ih}(\beta)\} \\
&\exp\{\sum_{j=1}^{m} \theta_{ij} Y_{ij} - b(\theta_i) + c(Y_i)\} dY_i \\
&= t \sum_{j=1}^{m}\sum_{h=1}^{m} V_{ijh}\mu_{ij}(\beta)\mu_{ih}(\beta) + b(\theta_i + \bar{t}) - b(\theta_i) + \log E\{\exp(t \sum_{j=1}^{m}\sum_{h=1}^{m} V_{ijh} Y_{ij} Y_{ih})\},
\end{aligned}
$$

with $\bar{t} = (-t\sum_{h=1}^{m}(V_{ih1}+V_{i1h})\mu_{i1}(\beta), -t\sum_{h=1}^{m}(V_{ih2}+V_{i2h})\mu_{i2}(\beta), \ldots, -t\sum_{h=1}^{m} (V_{ihm} + V_{imh})\mu_{im}(\beta))^T$. For the last term, we have $\log E\{\exp(t\sum_{j=1}^{m}\sum_{h=1}^{m} V_{ijh} Y_{ij} Y_{ih})\} \leq \max_{i,j,h} \log E[\exp\{tm^2(\max_{i,j,h}|V_{ijh})|Y_{ij}^2\}]$. If Y_{ij}s are sub-Gaussian random variables with uniformly bounded mean and variances, there exist constants b_s and b_g such that for $\forall i\,, j$,

$$
\max_{i,j}[E\{\exp(b_s Y_{ij}^2)\}] < b_g
$$

(Rudelson and Vershynin, 2013). Let $b_6 = b_s/(m^2 \max_{ijh}|V_{ijh}|)$, then if

$|t| < b_6$, $\partial C_{Q_i(\beta)}(t)/\partial t$ is bounded. By similar argument, we can prove that $\partial^2 C_{Q_i(\beta)}(t)/(\partial t)^2$ and $\partial^3 C_{Q_i(\beta)}(t)/(\partial t)^3$ are bounded.

$\square$

Lemma S2.2. *Under Assumptions 1 - 4, for all* β *in the neighborhood of* $||\beta - \beta_s^*|| \leq (p_n^2 \log p_n/n)^{1/2}$, $A_{ij}(\beta)$ *are uniformly bounded away from zero and infinity as* $n \to \infty$. *Furthermore,* $|\partial A_{ij}(\beta)/\partial \beta_{[k]}|$, $|\partial^2 A_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}|$, $|A_{ij}^{-1/2}(\beta)|$, $|\partial A_{ij}^{-1/2}(\beta)/ \partial \beta_{[k]}|$, $|\partial^2 A_{ij}^{-1/2}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}|$, $|\Lambda_{ij}(\beta)|$, $|\partial \Lambda_{ij}(\beta)/\partial \beta_{[k]}|$, $|\mu_{ij}(\beta)|$, $|\partial \mu_{ij}(\beta)/\partial \beta_{[k]}|$, $|\partial^2 \mu_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}|$, $|\partial^3 \mu_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}\partial \beta_{[r]}|$ *are all uniformly bounded for all* $i \in \{1, 2 \ldots n\}$, $j \in \{1, 2 \ldots m\}$, $k, l, r \in \{1, 2 \ldots p_n\}$.

Proof of Lemma S2.2. We first consider the boundedness of the third derivative of $\mu_{ij}(\beta)$. As $|\partial^3 \mu_{ij}(\beta) \ /\partial \beta_{[k]}\partial \beta_{[l]}\partial \beta_{[r]}| = X_{ijk}X_{ijl}X_{ijr}\partial^3 g^{-1}\{\zeta_{ij}(\beta)\} /\{\partial \zeta_{ij}(\beta)\}^3$ and $\partial^3 g^{-1}\{\zeta_{ij}(\beta_s^*)\} \ /\{\partial \zeta_{ij}(\beta_s^*)\}^3$ is bounded and continuous, therefore $|\partial^3 \mu_{ij}(\beta) \ /\partial \beta_{[k]}\partial \beta_{[l]}\partial \beta_{[r]}|$ is bounded for $||\beta - \beta_s^*|| \leq (p_n^2 \log p_n/n)^{1/2}$. Similarly we can also prove the boundedness of $|\mu_{ij}(\beta)|$, $|\partial \mu_{ij}(\beta)/\partial \beta_{[k]}|$, $|\partial^2 \mu_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}|$. Furthermore, $\partial^2 A_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]} = [\partial \nu\{\mu_{ij}(\beta)\}/\partial \mu_{ij}(\beta)] [\partial^2 \mu_{ij}(\beta)/\partial \beta_{[k]}\partial \beta_{[l]}] + [\partial^2 \nu\{\mu_{ij}(\beta)\}/\{\partial \mu_{ij}(\beta)\}^2] [\partial \mu_{ij}(\beta)/\partial \beta_{[l]}] [\partial \mu_{ij}(\beta)/\partial \beta_{[k]}]$

and

$$
\begin{aligned}
&\partial^2 A_{ij}^{-1/2}(\beta)/\partial\beta_{[k]}\partial\beta_{[l]}\\
&= (3/4)A_{ij}^{-5/2}(\beta)[\partial A_{ij}(\beta)/\partial\beta_{[l]}][\partial A_{ij}(\beta)/\partial\beta_{[k]}]\\
&- (1/2)A_{ij}^{-3/2}(\beta)[\partial^2 A_{ij}(\beta)/\partial\beta_{[k]}\partial\beta_{[l]}]
\end{aligned}
$$

are also bounded. Similarly we can prove the boundedness of the other terms. □

Let B and $\widetilde{B}$ denote $d_s \times m$ matrices. Let $D_i^{(1)}(\beta, \check{\beta}, B)$ be an $m \times d_s$ matrix and its jth row and kth column entry is $D_i^{(1)}(\beta, \check{\beta}, B)_{[jk]} = (\beta - \check{\beta})^T\{\partial^2\mu_{ij}(B_{[,j]})/\ \partial\beta_{[k]}\partial\beta\}$. Let $D_i^{(2)}(\beta, \widetilde{\beta}, B, \widetilde{B})$ be an $m \times d_s$ matrix with the jth row and kth column entry as $D_i^{(2)}(\beta, \widetilde{\beta}, B, \widetilde{B})_{[jk]} = (\beta - \widetilde{\beta})^T\{\partial^3\mu_{ij}(\widetilde{B}_{[,j]})/\partial\beta_{[k]}\partial\beta\partial\beta^T\}(B_{[,j]} - \beta)$.

Lemma S2.3. *Let $B_s^* = (\beta_s^*, \beta_s^* \dots \beta_s^*)$. Under Assumptions 1 - 4, for all model $s \in S$, $i \in \{1, 2 \dots n\}$, there exist $d_s \times m$ matrices $B_s^{\mu_i}$, $B_s^{\widetilde{\mu}_i}$, $B_s^{\check{\mu}_i}$, and $B_s^{D_i}$ such that each column of these four matrices is between $\widehat{\beta}_s$ and β_s^* and they satisfy:*

$$\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*) = D_i(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})(\widehat{\beta}_s - \beta_s^*); \quad \text{(S2.1)}$$

$$\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) + \frac{1}{2}D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s); \quad \text{(S2.2)}$$

$$D_i(\widehat{\beta}_s) = D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i}); \tag{S2.3}$$

$$D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i}) = D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}) - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*). \tag{S2.4}$$

The max norms of the matrices have the following uniform bounds for all model $s \in S$, $i \in \{1, 2 \ldots n\}$:

$$||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$|D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$||D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})||_{\max} = O_p(p_n^3 \log p_n/n).$$

Proof of Lemma S2.3. From Taylor expansion, there exists a $\beta_s^{\mu_{ij}}$ between β_s^* and $\widehat{\beta}_s$ such that

$$\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*) = \frac{\partial \mu_{ij}(\beta_s^*)}{\partial \beta^T}(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2}(\widehat{\beta}_s - \beta_s^*)^T \frac{\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})}{\partial \beta \partial \beta^T}(\widehat{\beta}_s - \beta_s^*). \tag{S2.5}$$

Let $B_s^{\mu_i} = (\beta_s^{\mu_{i1}}, \beta_s^{\mu_{i2}} \ldots, \beta_s^{\mu_{im}})$ and each column of $B_s^{\mu_i}$ is between β_s^* and

$\widehat{\beta}_s$. Define

$$D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{\mu_i}) = \begin{bmatrix} (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2\mu_{i1}(\beta_s^{\mu_{i1}})/\partial\beta\partial\beta^T\} \\ (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2\mu_{i2}(\beta_s^{\mu_{i2}})/\partial\beta\partial\beta^T\} \\ \cdots \\ (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2\mu_{im}(\beta_s^{\mu_{im}})/\partial\beta\partial\beta^T\} \end{bmatrix}.$$

Then Equation (S2.5) can be reformulated as

$$\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*) = D_i(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})(\widehat{\beta}_s - \beta_s^*).$$

Similarly if we perform Taylor Expansion at $\mu_i(\widehat{\beta}_s)$, we obtain

$$\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) + \frac{1}{2}D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s).$$

By similar argument, there exists a $\beta_s^{D_{ij}}$ between β_s^* and $\widehat{\beta}_s$ such that $\partial\mu_{ij}(\widehat{\beta}_s)/\partial\beta_{[k]} = \partial\mu_{ij}(\beta_s^*)/\partial\beta_{[k]} + (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2\mu_{ij}(\beta_s^{D_{ij}})/\partial\beta_{[k]}\partial\beta\}$. Define $B_s^{D_i} = (\beta_s^{D_{i1}}, \beta_s^{D_{i2}} \ldots, \beta_s^{D_{im}})$ and we can reformulate the equation above as

$$D_i(\widehat{\beta}_s) = D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{D_i}).$$

According to Taylor Expansion, there exists a $\beta_s^{\check{\mu}_{ij}}$ between β_s^* and $\beta_s^{\mu_{ij}}$ such that

$$
\begin{aligned}
&D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})_{[jk]} - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)_{[jk]} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^2 \mu_{ij}(\beta_s^{\widetilde{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta} - (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^2 \mu_{ij}(\beta_s^*)}{\partial \beta_{[k]} \partial \beta} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^3 \mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta \partial \beta^T} (\beta_s^{\widetilde{\mu}_{ij}} - \beta_s^*).
\end{aligned}
$$

Define $B_s^{\check{\mu}_i} = (\beta_s^{\check{\mu}_{i1}}, \beta_s^{\check{\mu}_{i2}} \ldots, \beta_s^{\check{\mu}_{im}})$. Then the equation above can be simplified as

$$
D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i}) = D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}) - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*).
$$

Next we estimate the orders of $D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})$ and $D_i^{(2)}(\beta_s^*, \widehat{\beta}_s B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})$. According to Cauchy-Schwarz inequality, we have

$$
|D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})_{[jk]}| = |(\widehat{\beta}_s - \beta_s^*)^T \frac{\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})}{\partial \beta_{[k]} \partial \beta}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times || \frac{\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})}{\partial \beta_{[k]} \partial \beta} ||.
$$

Here $||\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})/\partial \beta_{[k]} \partial \beta|| = [\sum_{l=1}^{p_n} \{\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})/\partial \beta_{[k]} \partial {\beta_{[l]}}^2\}]^{1/2} = O_p(p_n^{1/2})$. Thus $D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{\mu_i})_{[jk]} = O_p(p_n^{1/2} ||\beta_s^* - \widehat{\beta}_s||)$, for all i, j, and s. Similarly we have

$$
\begin{aligned}
&|D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})_{[jk]}| = |(\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^3 \mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta \partial \beta^T} (B_s^{\widetilde{\mu}_{ij}} - \beta_s^*)| \\
&\leq p_n ||\beta_s^* - \widehat{\beta}_s|| \times ||B_s^{\widetilde{\mu}_{ij}} - \beta_s^*|| \times || \frac{\partial^3 \mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta \partial \beta^T} ||_{\max} \\
&= O_p(p_n^3 \log p_n / n).
\end{aligned}
$$

$\square$

Lemma S2.4. *Let β_s, $\dot{\beta}_s$, $\widetilde{\beta}_s$, $\check{\beta}_s$ and every column of B_i be a $d_s \times 1$ vector that falls within a $(p_n^2 \log p_n/n)^{1/2}$ neighborhood of β_T^*, $i = 1, 2 \ldots n$. Under Assumptions 1 - 4, for any unit vector $||v||^2 = 1$ and any model $s \in S_+$ we have the following bounds:*

$$
\begin{aligned}
&\max_{||v||^2=1} v^T\{\frac{1}{n}\sum_{i=1}^n D_i(\beta_s)^T D_i(\beta_s)\}v = O(1),\\
&\max_{||v||^2=1} v^T\{\frac{1}{n}\sum_{i=1}^n D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)^T D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)\}v = O_p(p_n^3 \log p_n/n),\\
&\max_{||v||^2=1} v^T\{\frac{1}{n}\sum_{i=1}^n D_i(\dot{\beta}_s)^T V_i(\widetilde{\beta}_s)^{-1} D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)\}v = O_p\{(p_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

Proof of Lemma S2.4. First we have the following bound:

$$
\begin{aligned}
&\max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n D_i(\beta_s)^T D_i(\beta_s) v| = \max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n X_i^T \Lambda_i(\beta_s)^2 X_i v|\\
&\leq \max_i \lambda_{\max}\{\Lambda_i(\beta_s)^2\}\lambda_{\max}\{\frac{1}{n}\sum_{i=1}^n X_i^T X_i\}||v||^2\\
&\leq \max_{i,j}\{\Lambda_{ij}(\beta_s)^2\}\lambda_{\max}\{\frac{1}{n}\sum_{i=1}^n X_i^T X_i\}||v||^2\\
&= O(1).
\end{aligned}
$$

As $\mu_{ij}(\beta_s)$ is differentiable to the third order, we rewrite $\partial^2\mu_{ij}(\beta_s)/\partial\beta_s\partial\beta_s^T = \{\partial\Lambda_{ij}(\beta_s)/\partial\beta_s\}X_{ij}^T$. Here $\{\partial\Lambda_{ij}(\beta_s)/\partial\beta_s\}$ is a $d_s \times 1$ column vector and X_{ij}^T is a $1 \times d_s$ row vector. We have $D_i^{(1)}(\beta_s, \check{\beta}_s, B)_{[j,]} = (\beta_s - \check{\beta}_s)^T\{\partial^2\mu_{ij}(B_{[,j]})/\partial\beta_s\partial\beta_s^T\} =$

$(\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{ij}(B_{[,j]})/\partial\beta_s\}X_{ij}^T$. Therefore we have

$$D_i^{(1)}(\beta_s, \check{\beta}_s, B) = \begin{bmatrix} (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{i1}(B_{[,1]})/\partial\beta_s\}X_{i1}^T \\ (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{i2}(B_{[,2]})/\partial\beta_s\}X_{i2}^T \\ \cdots \\ (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{im}(B_{[,m]})/\partial\beta_s\}X_{im}^T \end{bmatrix}.$$

Let $diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]\}$ represent a diagonal matrix with the jth diagonal entry equal to $(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]$. Then we can reformat $D_i^{(1)}(\beta_s, \check{\beta}_s, B) = diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]\}X_i$. From Assumption 2, we have the boundedness of $\lambda_{\max}\{n^{-1}\sum_{i=1}^n X_i^T X_i\}$. This entails

$$\begin{aligned}
&\max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)^T D_i^{(1)}(\beta_s, \check{\beta}_s, B_i) v| \\
&= \max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n X_i^T diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T \frac{\partial\Lambda_{ij}(B_{i[,j]})}{\partial\beta_s}\}^2 X_i v| \\
&\leq \max_i \lambda_{\max}\{diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T \frac{\partial\Lambda_{ij}(B_{i[,j]})}{\partial\beta_s}\}^2\}\lambda_{\max}(\frac{1}{n}\sum_{i=1}^n X_i^T X_i)||v||^2 \\
&\leq \max_{i,j}\{(\beta_s - \check{\beta}_s)^T \frac{\partial\Lambda_{ij}(B_{i[,j]})}{\partial\beta_s}\}^2 \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n X_i^T X_i) \\
&\leq ||\beta_s - \check{\beta}_s||^2 \times \max_{i,j}\{||\frac{\partial\Lambda_{ij}(B_{i[,j]})}{\partial\beta_s}||^2\} \times \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n X_i^T X_i) \\
&\leq ||\beta_s - \check{\beta}_s||^2 \times p_n \times \max_{i,j,k}[\{\frac{\partial\Lambda_{ij}(B_{i[,j]})}{\partial\beta_{s_{[k]}}}\}^2] \times \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n X_i^T X_i) \\
&= O_p(p_n^3 \log p_n / n);
\end{aligned}$$

$$
\begin{aligned}
&\max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n D_i(\dot{\beta}_s)^T V_i(\widetilde{\beta}_s)^{-1} D_i^{(1)}(\beta_s, \check{\beta}_s, B_i) v| \\
&= \max_{||v||^2=1} |v^T \frac{1}{n}\sum_{i=1}^n X_i^T \Lambda_i(\dot{\beta}_s) V_i(\widetilde{\beta}_s)^{-1} diag_{j=1}^m \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\} X_i v| \\
&\leq \max_{i,j}\{\Lambda_{ij}(\dot{\beta}_s)\} \times \max_i \lambda_{\max}\{V_i(\widetilde{\beta}_s)^{-1}\} \times \max_{i,j}\{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\} \\
&\times \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n X_i^T X_i) \times ||v||^2 \\
&\leq ||\beta_s - \check{\beta}_s|| \times \max_{i,j} ||\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}|| \times O(1) \\
&\leq ||\beta_s - \check{\beta}_s|| \times (p_n^{1/2}) \times \max_{i,j,k} ||\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_{s_{[k]}}}|| \times O(1) \\
&= O_p\{(p_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

$\square$

Lemma S2.5. *Under Assumption 1 - 4, the estimated inverse working covariance matrices can be decomposed into the sum of several matrices of the same dimension* $V_i(\widehat{\beta}_s)^{-1} = V_i^{-1}(\beta_s^*) + V_i^{(1)}(\widehat{\beta}_s, \beta_s^*) + V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})$, *where* $B_s^{A_i} = (\beta_s^{A_{i1}}, \ldots, \beta_s^{A_{im}})$, *and each* $\beta_s^{A_{ij}}$, $j = 1, \ldots, m$, *is a vector between* $\widehat{\beta}_s$ *and* β_s^*. *Let* $\check{V}_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i}) = V_i^{(1)}(\widehat{\beta}_s, \beta_s^*) + V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})$. *The bounds*

$$
||V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},
$$

$$
||\check{V}_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},
$$

$$
||V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})||_{\max} = O_p\{p_n^3 \log p_n/n\}
$$

are uniformly held for all model $s \in S$, and $i = 1, 2 \ldots n$.

Proof of Lemma S2.5. According to Taylor expansion, there exists a $\beta_s^{A_{ij}}$ between $\widehat{\beta}_s$ and β_s^* such that

$$A_{ij}^{-1/2}(\widehat{\beta}_s) = A_{ij}^{-1/2}(\beta_s^*)+(\widehat{\beta}_s-\beta_s^*)^T\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta}+\frac{1}{2}(\widehat{\beta}_s-\beta_s^*)^T\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial\beta\partial\beta^T}(\widehat{\beta}_s-\beta_s^*).$$

For the jth row and hth column of matrix $V_i^{-1}(\widehat{\beta}_s) - V_i^{-1}(\beta_s^*)$, we apply the formula above and obtain

$$\begin{aligned}
&[V_i^{-1}(\widehat{\beta}_s) - V_i^{-1}(\beta_s^*)]_{[jh]} = A_{ij}^{-1/2}(\widehat{\beta}_s)[R^{-1}]_{[jh]}A_{ih}^{-1/2}(\widehat{\beta}_s) - A_{ij}^{-1/2}(\beta_s^*)[R^{-1}]_{[jh]}A_{ih}^{-1/2}(\beta_s^*)\\
&= (\widehat{\beta}_s - \beta_s^*)^T[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta}\}\\
&+ (\widehat{\beta}_s - \beta_s^*)^T[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial\beta\partial\beta^T} + \frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial\beta\partial\beta^T}\\
&+ \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta^T}\}(\widehat{\beta}_s - \beta_s^*) + O_p(||\widehat{\beta}_s - \beta_s^*||^3).
\end{aligned}$$

Denote the first term in the expansion as $V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)_{[jh]}$ and the remaining three terms as $V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})_{[jh]}$. Based on the Cauchy-Schwarz inequality, the bounds determined in Lemma S2.2 and Assumption 2, we have

$$\begin{aligned}
&|V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)_{[jh]}|\\
&\le ||\widehat{\beta}_s - \beta_s^*|| \times ||[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta}\}||\\
&\le p_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times \max_l |[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_{[l]}} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta_{[l]}}\}|\\
&= O_p(p_n^{1/2}||\widehat{\beta}_s - \beta_s^*||)\\
&= O_p\{(p_n^3 \log p_n/n)^{1/2}\};
\end{aligned}$$

and

$$|V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})_{[jh]}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \max_l ||[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial\beta_{[l]}\partial\beta^T}$$

$$+ \frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial\beta_{[l]}\partial\beta^T} + \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_{[l]}}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta^T}\}(\widehat{\beta}_s - \beta_s^*)|| + O_p(||\widehat{\beta}_s - \beta_s^*||^3)$$

$$\leq ||\widehat{\beta}_s - \beta_s^*||^2 \times p_n \max_l \max_r |[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial\beta_{[l]}\partial\beta_{[r]}}$$

$$+ \frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial\beta_{[l]}\partial\beta_{[r]}} + \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_{[l]}}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta_{[r]}}\}| + O_p(||\widehat{\beta}_s - \beta_s^*||^3)$$

$$= O_p(||\widehat{\beta}_s - \beta_s^*||^2 \times p_n) = O_p(p_n^3 \log p_n / n).$$

□

Lemma S2.6. *Under Assumptions 1 - 4, for all model $s \in S$,*

$$\max_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| = O_p(1).$$

Proof of Lemma S2.6. First we consider the true and overfitting models $s \in S_+$. As Y_{ij}'s are independent and their variances are uniformly bounded, we have

$$var\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\} = \mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|^2\} - [\mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\}]^2$$

$$\leq \mathrm{E}[\{Y_{ij} - \mu_{ij}(\beta_s^*)\}^2] = var(Y_{ij})$$

$$\leq b_2.$$

By the Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\}.$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{|Y_{ij}-\mu_{ij}(\beta_s^*)|\} \le \frac{1}{2n}\sum_{i=1}^{n}\mathrm{E}[\{Y_{ij}-\mu_{ij}(\beta_s^*)\}^2+1]$$
$$\le \frac{1}{2n}\{\sum_{i=1}^{n} var(Y_{ij})+1\}$$
$$\le (b_2+1)/2.$$

For all $j=1,2,\dots,m$, we have $n^{-1}\sum_{i=1}^{n}|Y_{ij}-\mu_{ij}(\beta_s^*)|=O_p(1)$.

For underfitting models $s\in S_-$, Lemma S2.2 implies that both $\mu_{ij}(\beta_T^*)$ and $\mu_{ij}(\beta_s^*)$ are bounded. Thus $n^{-1}\sum_{i=1}^{n}|\mu_{ij}(\beta_T^*)-\mu_{ij}(\beta_s^*)|=O(1)$. For $j=1,2,\dots,m$, we have $n^{-1}\sum_{i=1}^{n}|Y_{ij}-\mu_{ij}(\beta_s^*)|\le n^{-1}\sum_{i=1}^{n}|Y_{ij}-\mu_{ij}(\beta_T^*)|+ n^{-1}\sum_{i=1}^{n}|\mu_{ij}(\beta_T^*)-\mu_{ij}(\beta_s^*)|=O_p(1)$. □

Lemma S2.7. *Under Assumptions 1 - 4, for the true and overfitting models,*

$$\sum_{i=1}^{n}\{\mu_i(\beta_s^*)-\mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}=n||\beta_s^*-\widehat{\beta}_s||^2o_p(1),$$

where the $o_p(1)$ term holds for all models $s\in S_+$.

Proof of Lemma S2.7. From Equation (S2.2) of Lemma S2.3, we have

$$\sum_{i=1}^{n}\{\mu_i(\beta_s^*)-\mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$
$$=\sum_{i=1}^{n}\{D_i(\widehat{\beta}_s)(\beta_s^*-\widehat{\beta}_s)+\frac{1}{2}D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})(\beta_s^*-\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$
$$=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}+V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$+(\beta_s^*-\widehat{\beta}_s)^T U(\widehat{\beta}_s)$$

$$+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$=Res_1+Res_2.$$

We expand the residual terms as follows.

$$Res_1=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}$$

$$=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\beta_s^*)\}$$

$$+(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i^{-1}(\widehat{\beta}_s)\}\{\mu_i(\beta_s^*)-\mu_i(\widehat{\beta}_s)\}$$

$$=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-(V_i^*)^{-1}+(V_i^*)^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\beta_s^*)\}+Res_{11}$$

$$=Res_{11}+(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-(V_i^*)^{-1}\}\{Y_i-\mu_i(\beta_s^*)\}$$

$$-(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^n D_i(\widehat{\beta}_s)^T\{V_i(\widehat{\beta}_s)^{-1}-(V_i^*)^{-1}\}\{Y_i-\mu_i(\beta_s^*)\}$$

$$=Res_{11}+Res_{12}-Res_{13}.$$

The first term can be further decomposed:

$$
\begin{aligned}
Res_{11} &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{D_i(\widehat{\beta}_s) + \frac{1}{2} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})\}(\beta_s^* - \widehat{\beta}_s) \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) \\
&+ \frac{1}{2}(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \widehat{V}_i^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s) \\
&- \frac{1}{2}(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T V_i(\widehat{\beta}_s)^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s) \\
&= Res_{111} + Res_{112} + Res_{113}.
\end{aligned}
$$

We obtain the bounds for each of the residual terms:

$$
\begin{aligned}
|Res_{111}| &= |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&\leq n \max\{|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|\} \\
&\times (\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^n D_i(\widehat{\beta}_s)^T D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) \\
&\leq n||\beta_s^* - \widehat{\beta}_s||^2 \times \max\{|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|\} \\
&\times \lambda_{\max}\{\frac{1}{n} \sum_{i=1}^n D_i(\widehat{\beta}_s)^T D_i(\widehat{\beta}_s)\} \\
&= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n / n)^{1/2}\};
\end{aligned}
$$

$$
\begin{aligned}
|2Res_{112}| &= |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T \widehat{V}_i^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s)| \\
&= |n(\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&\leq n||\beta_s^* - \widehat{\beta}_s||^2 \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\widehat{\beta}_s) v\} \\
&= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n / n)^{1/2}\}.
\end{aligned}
$$

Following the same technique on Res_{112}, we obtain $|Res_{113}| = n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. Applying Lemma S2.3 and S2.5 to Res_{12}, we have

$$
\begin{aligned}
Res_{12} &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i(\widehat{\beta}_s)^T (\widehat{V}_i^{-1} - (V_i^*)^{-1})\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n \{D_i(\beta_s^*)^T + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})\}\{V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) \\
&+ V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n \{D_i(\beta_s^*)^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) + D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i}) \\
&+ D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= Res_{121} + Res_{122} + Res_{123} + Res_{124}.
\end{aligned}
$$

For Res_{121}, define the $d_s \times 1$ vector $\Gamma = \sum_{i=1}^n D_i(\beta_s^*)^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}$. Res_{121} can be reformulated as $(\widehat{\beta}_s - \beta_s^*)^T \Gamma$. The kth element of Γ is denoted as Γ_k. The kth row of $D_i(\beta_s^*)^T$ is denoted as $[D_i(\beta_s^*)^T]_{[k,]}$.

$$
\begin{aligned}
\Gamma_k &= \sum_{i=1}^{n} [D_i(\beta_s^*)^T]_{[k,]} V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} [D_i(\beta_s^*)^T]_{[kj]} V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)_{[j\bar{j}]}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} D_i(\beta_s^*)_{[jk]}(\widehat{\beta}_F - \beta_F^*)^T [R^{-1}]_{[j\bar{j}]}\{A_{i\bar{j}}^{-1/2}(\beta_F^*)\frac{\partial A_{ij}^{-1/2}(\beta_F^*)}{\partial \beta} \\
&+ A_{ij}^{-1/2}(\beta_F^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_F^*)}{\partial \beta}\}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\} \\
&= (\widehat{\beta}_F - \beta_F^*)^T \Pi_k(\beta_s^*).
\end{aligned}
$$

Note that for true and overfitting model, $\mu_i(\beta_F^*) = \mu_i(\beta_s^*)$. Here $\Pi_k(\beta_s^*) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}[A_{i\bar{j}}^{-1/2}(\beta_s^*)\{\partial A_{ij}^{-1/2}(\beta_s^*)/\partial\beta\} + A_{ij}^{-1/2}(\beta_s^*)$ $\{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)\ /\partial\beta\}]\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\}$ represents a $d_s \times 1$ vector. The rth element of $\Pi_k(\beta_s^*)$ is denoted as $\Pi_{kr}(\beta_s^*)$. Then we have

$$
\begin{aligned}
\Pi_{kr}(\beta_s^*) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}\{A_{i\bar{j}}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta_{[r]}} \\
&+ A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)}{\partial \beta_{[r]}}\}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\}.
\end{aligned}
$$

Given that $\mathrm{E}[Y_{i\bar{j}}] = \mu_{i\bar{j}}(\beta_s^*)$, then $\mathrm{E}[n^{-1}\Pi_{kr}(\beta_s^*)] = 0$. According to Lemma S2.2, $A_{ih}^{-1/2}(\beta_s^*)$ and its first derivative are uniformly bounded for all i, h and all model s. Therefore there exists a b_Π for all model s such that:

$$Var\{\frac{1}{n}\Pi_{kr}(\beta_s^*)\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m}\sum_{h=1}^{m}\sum_{\bar{h}=1}^{m} D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}[D_i(\beta_s^*)]_{hk}[R^{-1}]_{h\bar{h}}$$

$$\{A_{i\bar{j}}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_r} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)}{\partial\beta_r}\}$$

$$\{A_{i\bar{h}}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta_r} + A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{h}}^{-1/2}(\beta_s^*)}{\partial\beta_r}\}\mathrm{Cov}(Y_{i\bar{j}}, Y_{i\bar{h}}) \le \frac{b_\Pi}{n}.$$

According to Chebyshev's inequality,

$$\Pr\{|\frac{1}{n}\Pi_{kr}(\beta_s^*)| \ge (\frac{b_\Pi}{n}p_n^2\log p_n)^{1/2}\} \le \frac{1}{p_n^2\log p_n}.$$

When $p_n \to \infty$, according to Bonferroni inequality,

$$\Pr\{\max_{k,r}|\frac{1}{n}\Pi_{kr}(\beta_s^*)| \ge (\frac{b_\Pi}{n}p_n^2\log p_n)^{1/2}\} \le \frac{p_n^2}{p_n^2\log p_n} = (\log p_n)^{-1} \to 0,$$

Or equivalently we have

$$\max_{kr}|\Pi_{kr}(\beta_s^*)| = O_p\{(p_n^2\log p_n)^{1/2}\}.$$

According to Cauchy-Schwarz inequality

$$|Res_{121}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\Gamma|| \leq p_n^{1/2} \times ||\widehat{\beta}_s - \beta_s^*|| \times \max_k |\Gamma_k|,$$

$$|\Gamma_k| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\Pi_k|| \leq p_n^{1/2} \times ||\widehat{\beta}_s - \beta_s^*|| \times \max_{kr} |\Pi_r(\beta_s^*)| = ||\widehat{\beta}_s - \beta_s^*|| O_p\{(p_n^3 \log p_n)^{1/2}\}.$$

Therefore we have

$$|Res_{121}| = n||\widehat{\beta}_s - \beta_s^*||^2 \times O_p\{(p_n^4 \log p_n/n)^{1/2}\}.$$

For the term Res_{122}, Lemma S2.5 implies that the largest elements of matrix $||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}$ is $O_p(p_n^3 \log p_n/n)$. Lemma S2.2 implies that all elements from $D_i(\beta_s^*)$ are bounded. Lemma S2.6 demonstrates that $\sum_{i=1}^n |Y_{ij} - \mu_{ij}(\beta_s^*)|/n$ are bounded for all $j \in \{1, 2 \ldots m\}$. Therefore we have

$$\begin{aligned}
|Res_{122}| &= |(\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^n D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}| \\
&\leq n||\widehat{\beta}_s - \beta_s^*|| \times ||\frac{1}{n}\sum_{i=1}^n D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}|| \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times \max_k |\frac{1}{n}\sum_{i=1}^n [D_i(\beta_s^*)^T]_{[k,]} V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}| \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times m||D_i(\beta_s^*)||_{\max}||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m |Y_{ij} - \mu_{ij}(\beta_s^*)|\} \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times m^2||D_i(\beta_s^*)||_{\max}||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}\max_j\{\frac{1}{n}\sum_{i=1}^n |Y_{ij} - \mu_{ij}(\beta_s^*)|\}
\end{aligned}$$

$$= O(np_n^{1/2})||\widehat{\beta}_s - \beta_s^*|| \times ||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}$$

$$= O(np_n^{1/2} p_n^3 \log p_n/n)||\widehat{\beta}_s - \beta_s^*||$$

$$= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{(p_n^5 \log p_n/n)^{1/2}\}.$$

Similarly we can estimate the orders of Res_{123} and Res_{124}.

$$Res_{123} = (\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}$$

$$\leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}||$$

$$\leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \times \max_k |\sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})_{[k,]}^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}|$$

$$\leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \times ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times ||V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)||_{\max} \times m^2 \max_j \sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)|$$

$$= O(np_n^{1/2}) \times ||\widehat{\beta}_s - \beta_s^*|| \times ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times O_p\{(p_n^3 \log p_n/n)^{1/2}\}$$

$$= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{(p_n^5 \log p_n/n)^{1/2}\};$$

$$Res_{124} = (\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}$$

$$\leq ||\widehat{\beta}_s - \beta_s^*|| p_n^{1/2} ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times ||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max} m^2 \max_j \sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)|$$

$$= O_p\{np_n^{3/2}\}||\widehat{\beta}_s - \beta_s^*||^3 \times ||\widehat{\beta}_F - \beta_F^*||$$

$$= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{p_n^{3.5} \log p_n/n\}.$$

According to Lemma S2.5, both $||V_i^* - \widehat{V}_i||_{\max}$ and $||V_i^* - V_i(\widehat{\beta}_s)||_{\max}$ have the same order. Similar to $|Res_{12}|$, we have $|Res_{13}| = n||\beta_s^* - \widehat{\beta}_s||^2 o_p(1)$.

Next we analyze the other residual terms.

$$
\begin{aligned}
2Res_2 &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\} \\
&+ (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= Res_{21} + Res_{22}.
\end{aligned}
$$

$$
\begin{aligned}
Res_{21} &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n \{D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})\}^T \\
&\{V_i^{-1}(\beta_F^*) + \check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n \{D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)V_i^{-1}(\beta_F^*) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})V_i^{-1}(\beta_F^*) \\
&+ D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)\check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i}) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})\check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= Res_{211} + Res_{212} + Res_{213} + Res_{214}.
\end{aligned}
$$

By similar arguments as above, we are able to show Res_{211}, Res_{212}, Res_{213}, and Res_{214} are all of the order $n||\widehat{\beta}_s - \beta_s^*||^2 o_p(1)$. For Res_{22}, there exists a $\check{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that $\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)$.

This entails

$$
\begin{aligned}
|Res_{22}| &= |(\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| \\
&= |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&= |n(\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&\le n||\beta_s^* - \widehat{\beta}_s||^2 \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^n D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s) v\} \\
&= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n / n)^{1/2}\}.
\end{aligned}
$$

Combining all the orders for each of the terms, results in the lemma follows.

□

Lemma S2.8. *For $s \in S_+$, let $\eta = n^{-1/2} W(\beta_s^*)^{-1/2} \sum_{i=1}^n U_i(\beta_s^*)$. The random vectors $U_1(\beta_s^*), U_2(\beta_s^*), \ldots, U_n(\beta_s^*)$ are independently distributed random vectors of dimension d_s with zero mean and satisfy the cumulant boundedness condition. Under Assumptions 1 - 4, $\log E[e^{t^T \eta}] \le a^2 t^T t/2$ for $||t||^2 \le p_n^2 \log p_n$ and some constant $a^2 > 1$.*

Proof of Lemma S2.8. For $s \in S_+$, $\mathrm{E}\{U_i(\beta_s^*)\} = \mathrm{E}[D_i(\beta_s^*)^T V_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}] = 0$. Let $W_i = \mathrm{Cov}\{U_i(\beta_s^*)\}$ be the covariance matrix of $U_i(\beta_s^*)$ and $W = \sum_{i=1}^n W_i/n$. The cumulant generating function of U is

$$
\begin{aligned}
C_{U_i(\beta_s^*)}(t) &= \log \mathrm{E}\{e^{t^T U_i(\beta_s^*)}\} \\
&= C(0) + t^T C^{(1)}(0) + \frac{1}{2} t^T \mathrm{Cov}\{U_i(\beta_s^*)\} t + \frac{1}{6} \sum_{lrk} t_l t_r t_k C_{lrk}^{(3)}(t^*),
\end{aligned}
$$

with a t^* such that $||t^*|| \leq ||t||$. Let $\eta_1 = n^{-1/2} \sum_{i=1}^n U_i(\beta_s^*)$, then the cumulant generating function of η_1 is

$$
\begin{aligned}
C_{\eta_1}(t) &= \sum_{i=1}^{n} C_{U_i(\beta_s^*)}(\frac{t}{n^{1/2}}) \\
&= \sum_{i=1}^{n} \{\frac{1}{2n} t^T \mathrm{Cov}\{U_i(\beta_s^*)\} t + \sum_{lrk} \frac{1}{6n^{3/2}} t_l t_r t_k C_{lrk}^{(3)}(\frac{t^*}{n^{1/2}})\} \\
&= C_1 + C_2.
\end{aligned}
$$

First, C_1 can be simplified as $C_1 = \frac{1}{2} t^T \{\frac{1}{n} \sum_{i=1}^n \mathrm{Cov}(U_i)\} t = \frac{1}{2} t^T (\frac{1}{n} \sum_{i=1}^n W_i) t = \frac{1}{2} t^T W t$. Next, C_2 has the bound as follows:

$$
C_2 \leq \frac{b_c}{6n^{1/2}} p_n^{3/2} ||t||^3 = \frac{b_c}{6} (\frac{p_n^3 ||t||^2}{n})^{1/2} ||t||^2 = O_p[\{p_n^5 \log p_n / n\}^{1/2}] ||t||^2 = o_p(1) ||t||^2.
$$

This entails

$$
C_{\eta_1}(t) \leq \frac{1}{2} a^2 t^T W t,
$$

for some a with $a^2 > 1$ and $||t|| \leq \{(p_n^2 \log p_n)^{1/2}\}$. Let $\eta = W^{-1/2} \eta_1$, then the cumulant generating function of η is $\log \mathrm{E}[e^{t^T \eta}] \leq a^2 t^T t / 2$. $\square$

SIMULATION RESULTS

We conduct additional simulations on clustered binary responses and clustered Gaussian responses. Table 1 and 2 illustrate the simulation results of QIC and GIC for binary and Gaussian responses under 100 simulated datasets. We let the multiplicative factor c to vary from 1 to 4 and examine the PSR and FDR under different values of c. It is observed that when $c = 1$, or 2, the proposed GIC achieves high PSR and low FDR. In comparison, when c increases to 3 or 4, the GIC has much lower PSR. Table 3 provides additional simulation results for binary and Gaussian responses with the cluster size $m = 20$. The performance of the proposed GIC improves with higher PSR and lower FDR with larger cluster size.

References

Gao, X. and R. J. Carroll (2017). Data integration with high dimensionality. *Biometrika 104*(2), 251–272.

Rudelson, M. and R. Vershynin (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability 18*, no. 82, 9.

Spokoiny, V. and M. Zhilova (2013). Sharp deviation bounds for quadratic forms. *Math. Methods Statist. 22*(2), 100–113.

Stewart, G. W. (1990). *Matrix Perturbation Theory.* New York: Citeseer.

Table 1: The PSR and FDR of GIC and QIC for Binary Response under different valus of the multiplicative factor c

		n 1000	p 1000			n 1000	p 500			n 500	p 500	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
	PSR	PSR	FDR	FDR	PSR	PSR	FDR	FDR	PSR	PSR	FDR	FDR
QIC(I)	1.0000	0.0000	0.7093	0.0241	1.0000	0.0000	0.4984	0.0585	0.9974	0.0084	0.5677	0.0735
QIC(E)	1.0000	0.0000	0.7099	0.0241	1.0000	0.0000	0.4988	0.0582	0.9974	0.0084	0.5677	0.0735
QIC(A)	1.0000	0.0000	0.7093	0.0241	1.0000	0.0000	0.4984	0.0585	0.9974	0.0084	0.5677	0.0735
QIC(U)	1.0000	0.0000	0.7234	0.0179	1.0000	0.0000	0.5151	0.0707	0.9976	0.0082	0.6163	0.0677
GIC(I,c=1)	0.9982	0.0081	0.0596	0.0476	0.9988	0.0069	0.1194	0.0932	0.9182	0.0710	0.0250	0.0548
GIC(I,c=2)	0.9554	0.0552	0.0012	0.0065	0.9822	0.0339	0.0147	0.0233	0.7378	0.0881	0.0000	0.0000
GIC(I,c=3)	0.8398	0.0854	0.0000	0.0000	0.9317	0.0898	0.0018	0.0068	0.6288	0.0794	0.0000	0.0000
GIC(I,c=4)	0.7568	0.0802	0.0000	0.0000	0.8566	0.1222	0.0008	0.0047	0.5524	0.0906	0.0000	0.0000
GIC(E,c=1)	0.9982	0.0081	0.0574	0.0454	0.9986	0.0071	0.1217	0.0939	0.9194	0.0715	0.0265	0.0561
GIC(E,c=2)	0.9554	0.0552	0.0012	0.0065	0.9822	0.0339	0.0147	0.0233	0.7410	0.0880	0.0000	0.0000
GIC(E,c=3)	0.8414	0.0858	0.0000	0.0000	0.9323	0.0900	0.0018	0.0068	0.6316	0.0833	0.0000	0.0000
GIC(E,c=4)	0.7560	0.0810	0.0000	0.0000	0.8592	0.1211	0.0006	0.0043	0.5546	0.0906	0.0000	0.0000
GIC(A,c=1)	0.9980	0.0083	0.0584	0.0471	0.9988	0.0069	0.1168	0.0915	0.9180	0.0708	0.0246	0.0550
GIC(A,c=2)	0.9554	0.0552	0.0014	0.0068	0.9822	0.0339	0.0146	0.0230	0.7378	0.0881	0.0000	0.0000
GIC(A,c=3)	0.8414	0.0837	0.0000	0.0000	0.9317	0.0898	0.0023	0.0088	0.6278	0.0784	0.0000	0.0000
GIC(A,c=4)	0.7568	0.0802	0.0000	0.0000	0.8558	0.1215	0.0008	0.0047	0.5536	0.0910	0.0000	0.0000
GIC(U,c=1)	0.9990	0.0066	0.0445	0.0399	0.9998	0.0020	0.0990	0.0885	0.9498	0.0538	0.0242	0.0381
GIC(U,c=2)	0.9820	0.0296	0.0017	0.0085	0.9911	0.0200	0.0098	0.0193	0.7978	0.0848	0.0000	0.0000
GIC(U,c=3)	0.9060	0.0773	0.0000	0.0000	0.9588	0.0651	0.0025	0.0085	0.6782	0.0909	0.0000	0.0000
GIC(U,c=4)	0.8146	0.0820	0.0000	0.0000	0.9143	0.0969	0.0006	0.0034	0.6094	0.0841	0.0000	0.0000

The true parameters size d_T is 50 and the cluster size m is 10. For working correlations, "I" represents independent, "E" represents exchangeable, "A" represents AR1, and "U" represents unstructured.

Table 2: The PSR and FDR of GIC and QIC for Gaussian Response under different valus of the multiplicative factor c

		n 1000	p 1000			n 1000	p 500			n 500	p 500	
	mean PSR	std PSR	mean FDR	std FDR	mean PSR	std PSR	mean FDR	std FDR	mean PSR	std PSR	mean FDR	std FDR
QIC(I)	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5308	0.0652	1.0000	0.0000	0.5401	0.0607
QIC(E)	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5381	0.0648	1.0000	0.0000	0.5395	0.0599
QIC(A)	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5308	0.0652	1.0000	0.0000	0.5401	0.0607
QIC(U)	1.0000	0.0000	0.7281	0.0136	1.0000	0.0000	0.7077	0.0354	1.0000	0.0000	0.7109	0.0324
GIC(I,c=1)	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0784	0.0415	1.0000	0.0000	0.0871	0.0449
GIC(I,c=2)	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0008	0.0038
GIC(I,c=3)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC(I,c=4)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(E,c=1)	1.0000	0.0000	0.0961	0.0511	1.0000	0.0000	0.0709	0.0404	1.0000	0.0000	0.0705	0.0450
GIC(E,c=2)	1.0000	0.0000	0.0028	0.0095	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0015	0.0065
GIC(E,c=3)	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC(E,c=4)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(A,c=1)	1.0000	0.0000	0.1073	0.0471	1.0000	0.0000	0.0784	0.0415	1.0000	0.0000	0.0860	0.0461
GIC(A,c=2)	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0008	0.0038
GIC(A,c=3)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC(A,c=4)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(U,c=1)	1.0000	0.0000	0.0226	0.0295	1.0000	0.0000	0.0116	0.0220	1.0000	0.0000	0.0272	0.0355
GIC(U,c=2)	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0004	0.0027
GIC(U,c=3)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC(U,c=4)	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000

The true parameters size d_T is 50 and the cluster size m is 10. For working correlations, "I" represents independent, "E" represents exchangeable, "A" represents AR1, and "U" represents unstructured.

Table 3: The PSR and FDR of GIC and QIC for Binary and Gaussian Responses with the cluster size $m = 20$

	Binary				Gaussian			
		n 500	p 500			n 500	p 500	
	mean	std	mean	std	mean	std	mean	std
	PSR	PSR	FDR	FDR	PSR	PSR	FDR	FDR
QIC(I)	1.0000	0.0000	0.5482	0.0629	1.0000	0.0000	0.5262	0.0628
QIC(E)	1.0000	0.0000	0.5490	0.0634	1.0000	0.0000	0.5262	0.0628
QIC(A)	1.0000	0.0000	0.5490	0.0634	1.0000	0.0000	0.5262	0.0628
QIC(U)	1.0000	0.0000	0.5978	0.0553	1.0000	0.0000	0.7294	0.0207
GIC(I,c=1)	0.9986	0.0051	0.0452	0.0375	1.0000	0.0000	0.0779	0.0406
GIC(I,c=2)	0.9708	0.0380	0.0019	0.0070	1.0000	0.0000	0.0036	0.0086
GIC(I,c=3)	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(I,c=4)	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(E,c=1)	0.9986	0.0051	0.0456	0.0375	1.0000	0.0000	0.0775	0.0409
GIC(E,c=2)	0.9710	0.0379	0.0019	0.0070	1.0000	0.0000	0.0036	0.0086
GIC(E,c=3)	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(E,c=4)	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(A,c=1)	0.9986	0.0051	0.0448	0.0375	1.0000	0.0000	0.0789	0.0411
GIC(A,c=2)	0.9706	0.0379	0.0019	0.0070	1.0000	0.0000	0.0032	0.0083
GIC(A,c=3)	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(A,c=4)	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC(U,c=1)	0.9992	0.0039	0.0346	0.0370	1.0000	0.0000	0.0147	0.0292
GIC(U,c=2)	0.9916	0.0187	0.0020	0.0065	1.0000	0.0000	0.0016	0.0067
GIC(U,c=3)	0.9562	0.0525	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC(U,c=4)	0.9012	0.0733	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000

The true parameters size d_T is 50 and the cluster size m is 20. For working correlations, "I" represents independent, "E" represents exchangeable, "A" represents AR1, and "U" represents unstructured.