TENSOR GENERALIZED ESTIMATING EQUATIONS FOR LONGITUDINAL IMAGING ANALYSIS

Xiang Zhang[1], Lexin Li[2], Hua Zhou[3], Yeqing Zhou[2], Dinggang Shen[4], and ADNI[5]

[1]*North Carolina State University,* [2]*University of California, Berkeley*
[3]*University of California, Los Angeles,* [4]*University of North Carolina, Chapel Hill and* [5]*the Alzheimer's Disease Neuroimaging Initiative*

Supplementary Materials

A Outline of the proofs

To facilitate the proof, we introduce the following notations. Denote $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_{\hat{\boldsymbol{B}}}$ the estimator from tensor GEE and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{\boldsymbol{B}_0}$ the true values. Recall that the CP decomposition ensures that $\boldsymbol{B}$ is uniquely determined by $\boldsymbol{\beta}_n \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$. Denote $\boldsymbol{J}(\boldsymbol{\beta}) = [\boldsymbol{J}_1,\ \boldsymbol{J}_2, \cdots, \boldsymbol{J}_D]$, and note that under tensor structure $\partial\theta_{ij}/\partial\boldsymbol{\beta} = \boldsymbol{J}(\boldsymbol{\beta})^{\intercal}\mathrm{vec}\boldsymbol{X}_{ij}$. Recall the generalized estimating equations can be written as

$$\boldsymbol{s}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n)).$$

The proof of Lemma 1 is similar to that of Theorem 1 by dropping the terms involving the working correlation matrix and thus is omitted here.

The main technique to prove Theorem 1 is the sufficient condition for existence and consistency of a root of equations proposed in Ortega and Rheinboldt (2000), which also has been used in Portnoy (1984) for M-estimator and in Wang (2011) for GEE estimator with vector covariates. To check this condition, Lemmas B.1–B.3 are proposed. Lemma B.1 provides a useful approximation to the generalized estimating equations $\boldsymbol{s}_n(\boldsymbol{\beta}_0)$ based on the condition (A4) of the working correlation matrix. This facilitates the later evaluations of the moments of the generalized estimating equations by treating the intra-subject correlation as known. Lemma B.2 further establishes the approximation of the negative gradients of the generalized estimating equations. Lemma B.3 refines this approximation of the negative gradients at one more step, providing the foundations for the Talyor expansion of generalized estimating equations at the true value.

Based on Theorem 1, the proof of Theorem 2 is obtained by evaluating the covariance matrix of the generalized estimating equations and applying the Lindeberg-Feller central limit theorem.

The proof of Theorem 3 follows two steps. We show that BIC neither overestimates nor underestimates the true rank. By combing these results, the rank selection consistency is established.

Theorem 4 is proved by construction. We show that the oracle estimator is an approximated solution to the SCAD regularized tensor GEE.

B Technical lemmas

Lemma B.1. *Under conditions (A1)-(A8),* $||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) - \boldsymbol{s}_n(\boldsymbol{\beta}_0)|| = O_p(1)$*, where* $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)$ *is* $\boldsymbol{s}_n(\boldsymbol{\beta}_0)$ *with* $\widehat{\boldsymbol{R}}$ *replaced by* $\tilde{\boldsymbol{R}}$*.*

Proof of Lemma B.1. Consider

$$\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n)).$$

Denote by $\{r_{i,j}\}_{1\le i,j\le m}$ the (i,j)-th element of $\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}$. By condition (A4), $r_{i,j} = O_p(n^{-1/2})$. By direct calculation,

$$\begin{aligned}
&\boldsymbol{s}_n(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} r_{j,m}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij} \\
=&\sum_{j=1}^{m}\sum_{k=1}^{m} r_{j,m}\Big[\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}\Big],
\end{aligned}$$

where $\epsilon_{ik}(\boldsymbol{\beta}_0) = \sigma_{ik}^{-1}(\boldsymbol{\beta}_0)(Y_{ik} - \mu_{ik}(\boldsymbol{\beta}_0))$. By condition (A5), (A6) and (A7),

$$\mathbb{E}\Big[||\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}||^2\Big] = O(n).$$

Therefore, $||\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}|| = O_p(\sqrt{n})$. Since $r_{i,j} = O_p(n^{-1/2})$, the proof is complete. □

Consider $\boldsymbol{D}_n(\boldsymbol{\beta}_n) = -\partial\boldsymbol{s}_n(\boldsymbol{\beta}_n)/\partial\boldsymbol{\beta}_n$, $\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) = -\partial\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n)/\partial\boldsymbol{\beta}_n$. Lemma B.2 establishes the approximation of the negative gradients of the estimating equations.

Lemma B.2. *Under conditions (A1)-(A8), for some constant* $\triangle > 0$,

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \leq \triangle n^{-1/2}} |\lambda_{\max}[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \boldsymbol{D}_n(\boldsymbol{\beta}_n)]| = O_p(n^{1/2}),$$
$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \leq \triangle n^{-1/2}} |\lambda_{\min}[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \boldsymbol{D}_n(\boldsymbol{\beta}_n)]| = O_p(n^{1/2}).$$

Proof of Lemma B.2. Similar to Lemma C.1. of Wang (2011), it can be shown by direct calculation that

$$\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) = \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_n),$$

where

$$\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$
$$\tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n) = \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_i(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$
$$\tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_n) = -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\boldsymbol{K}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$
$$\tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_n) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{e}_j^{\intercal}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n))\mathbf{H}(\boldsymbol{\beta}_n),$$

with

$$\mathbf{C}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(Y_{i1} - \mu_{i1}(\boldsymbol{\beta}_n), \ldots, Y_{im} - \mu_{im}(\boldsymbol{\beta}_n)\Big),$$
$$\boldsymbol{F}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(\mu_{i1}^{(2)}(\boldsymbol{\beta}_n), \ldots, \mu_{im}^{(2)}(\boldsymbol{\beta}_n)\Big),$$
$$\boldsymbol{K}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n))\Big),$$

$\mathbf{e}_j^{\intercal}$ the length m vector with j-th element 1 and 0 everywhere else, and $\mathbf{H}(\boldsymbol{\beta}_n)$ is defined in condition (A8).

Let $\boldsymbol{D}_{ni}(\boldsymbol{\beta}_n)$ be defined the same as $\tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)$, but with $\tilde{\boldsymbol{R}}$ replaced by $\widehat{\boldsymbol{R}}$, for $i = 1, \ldots, 4$. It is sufficient to prove

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \leq \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^{\intercal}[\boldsymbol{D}_{ni}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2})$$

for any $\boldsymbol{u} \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$ such that $||\boldsymbol{u}|| = 1$, $i = 1, \ldots, 4$.

For $i = 1$, we have

$$
\begin{aligned}
&|\boldsymbol{u}^\intercal[\boldsymbol{D}_{n1}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| \\
\leq & n||\boldsymbol{u}||^2 \cdot ||\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}||_F \cdot \lambda_{\max}(\boldsymbol{A}_i(\boldsymbol{\beta}_n)) \cdot \lambda_{\max}\Big(n^{-1}\sum_{i=1}^{n} \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\Big).
\end{aligned}
$$

By condition (A3), (A4) and (A7), $|\boldsymbol{u}^\intercal[\boldsymbol{D}_{n1}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2})$ on the set $\boldsymbol{N}_n$.

For $i = 2$, we have

$$
\begin{aligned}
&|\boldsymbol{u}^\intercal[\boldsymbol{D}_{n2}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)]\boldsymbol{u}| \\
\leq & \frac{1}{2}|\boldsymbol{u}^\intercal \sum_{i=1}^{n} \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i1}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
& + \frac{1}{2}|\boldsymbol{u}^\intercal \sum_{i=1}^{n} \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i2}(\boldsymbol{\beta}_0)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
\triangleq & J_{n1} + J_{n2},
\end{aligned}
$$

where we decompose $\mathbf{C}_i(\boldsymbol{\beta}_n)$ as $\mathbf{C}_{i1}(\boldsymbol{\beta}_n) + \mathbf{C}_{i2}(\boldsymbol{\beta}_0)$,

$$
\begin{aligned}
\mathbf{C}_{i1}(\boldsymbol{\beta}_n) &= \mathrm{diag}\Big(\mu_{i1}(\boldsymbol{\beta}_0) - \mu_{i1}(\boldsymbol{\beta}_n), \ldots, \mu_{im}(\boldsymbol{\beta}_0) - \mu_{im}(\boldsymbol{\beta}_n)\Big), \\
\mathbf{C}_{i2}(\boldsymbol{\beta}_0) &= \mathrm{diag}\Big(Y_{i1} - \mu_{i1}(\boldsymbol{\beta}_0), \ldots, Y_{im} - \mu_{im}(\boldsymbol{\beta}_0)\Big).
\end{aligned}
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
2J_{n1} \leq & \sum_{i=1}^{n} ||\boldsymbol{u}^\intercal\boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i1}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)|| \\
& \times ||\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}|| \\
\leq & \sum_{i=1}^{n} ||\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}||^2 \times \lambda_{\max}\Big(\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\Big) \times \lambda_{\max}\Big(\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\Big) \times ||\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}||_F \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \\
\leq & \sum_{i=1}^{n} ||\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}||^2 \times ||\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}||_F \times \frac{\max_{i,j}\sigma_{ij}(\boldsymbol{\beta}_n)}{\min_{i,j}\sigma_{ij}^3(\boldsymbol{\beta}_n)} \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \\
\leq & ||\boldsymbol{u}||^2 \times \lambda_{\max}\Big(\sum_{i=1}^{n} \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\Big) \times ||\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}||_F \times \frac{\max_{i,j}\sigma_{ij}(\boldsymbol{\beta}_n)}{\min_{i,j}\sigma_{ij}^3(\boldsymbol{\beta}_n)} \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0||
\end{aligned}
$$

where $\widetilde{\boldsymbol{\beta}}_n$ is between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. Under conditions (A3), (A4) and (A7), it can be easily seen now $J_{n1} \leq CnO_p(n^{-1/2})O_p(n^{-1/2}) = O_p(1)$.

For J_{n2}, recall that $\epsilon_{ij}(\boldsymbol{\beta}_0) = \sigma_{ij}^{-1}(\boldsymbol{\beta}_0)(Y_{ij} - \mu_{ij}(\boldsymbol{\beta}_0))$. By condition (A5),

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} \mathbb{E}[J_{n2}^2] = \sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} \mathrm{Tr}[\mathbb{E}(J_{n2}^{\mathsf{T}} J_{n2})] \\
&= \sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} \mathbb{E}[\epsilon_{ij}\epsilon_{ik}] \mathrm{Tr}\Big[\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{e}_j\mathbf{e}_j^{\mathsf{T}}\boldsymbol{F}_i(\boldsymbol{\beta}_n) \\
&\quad \cdot \mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \mathbf{e}_k\mathbf{e}_k^{\mathsf{T}}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n) \\
&\leq \sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} C\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} ||\mathbf{e}_j^{\mathsf{T}}\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)|| \cdot ||\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathbf{e}_k|| \\
&\quad \cdot ||\mathbf{e}_k^{\mathsf{T}}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)|| \\
&\quad \cdot ||\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{e}_j||.
\end{aligned}
$$

By conditions (A4), (A6) and (A7), $\sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} \mathbb{E}[J_{n2}^2] \leq Cn||\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1}||_F^2 = O(1)$. Using similar decompositions, we can verify the results for $\boldsymbol{D}_{n3}$ and $\boldsymbol{D}_{n4}$, which completes the proof. □

Based on Lemma B.2, we can further approximate $\tilde{\mathbf{D}}_n(\boldsymbol{\beta}_n)$ by $\tilde{\mathbf{D}}_{n1}(\boldsymbol{\beta}_n)$, which are easier to evaluate. Lemma B.3 provides this approximation.

Lemma B.3. *Under conditions (A1)-(A8), for some constant $\triangle > 0$ and $\boldsymbol{u} \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$ such that $||\boldsymbol{u}|| = 1$,*

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^{\mathsf{T}}[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2}), \tag{S1}$$

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^{\mathsf{T}}[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2}). \tag{S2}$$

Proof of Lemma B.3. To prove (S1), it is sufficient to show, for $i = 2, 3, 4$,

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^{\mathsf{T}}\tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)\boldsymbol{u}| = O_p(n^{1/2}).$$

For $\tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)$, if suffices to show

$$\sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} |\boldsymbol{u}^{\mathsf{T}} \sum_{i=1}^{n} \boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_i(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| = O_p(n^{1/2}).$$

By using the decomposition $\mathbf{C}_i(\boldsymbol{\beta}_n)$ as $\mathbf{C}_{i1}(\boldsymbol{\beta}_n)+\mathbf{C}_{i2}(\boldsymbol{\beta}_0)$, the proof is similar to the proof for $|\boldsymbol{u}^\intercal[\boldsymbol{D}_{n2}(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)]\boldsymbol{u}|$ in Lemma B.2. We can prove the results for $\tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_0)$ and $\tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_0)$ in the same way, which completes the proof of (S1).

To prove (S2), note that

$$\begin{aligned}
&|\boldsymbol{u}^\intercal[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)-\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}|\\
\leq&|\boldsymbol{u}^\intercal[\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)-\boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)]\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}|\\
&+|\boldsymbol{u}^\intercal\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i[\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)-\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)]\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}|\\
&+|\boldsymbol{u}^\intercal\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}[\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)-\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)]\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}|\\
&+|\boldsymbol{u}^\intercal\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\mathrm{vec}^\intercal\boldsymbol{X}_i[\boldsymbol{J}(\boldsymbol{\beta}_0)-\boldsymbol{J}(\boldsymbol{\beta}_n)]\boldsymbol{u}|.
\end{aligned}$$

The rest of the proof is similar to the proof of Lemma B.2 and thus is omitted here. □

C Proof of theorems

Proof of Theorem 1. Wang (2011) gave a sufficient condition for the existence and consistency of a sequence of roots $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{s}_n(\boldsymbol{\beta}_n)=0$, namely,

$$P\left(\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||=\triangle n^{-1/2}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal\boldsymbol{s}_n(\boldsymbol{\beta}_n)<0\right)\geq 1-\epsilon \tag{S3}$$

with $\forall\epsilon>0$ and a constant $\triangle>0$. To verify (S3), the main idea is to approximate $\boldsymbol{s}_n(\boldsymbol{\beta}_n)$ by $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n)$, whose moments are easier to evaluate.

By direct calculation,

$$(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal\boldsymbol{s}_n(\boldsymbol{\beta}_n)=(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal\boldsymbol{s}_n(\boldsymbol{\beta}_0)-(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal\boldsymbol{D}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)\triangleq I_{n1}+I_{n2},$$

where $\boldsymbol{\beta}_n^*$ is between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. Further decompose I_{n1} into

$$I_{n1}=(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)+(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\intercal[\boldsymbol{s}_n(\boldsymbol{\beta}_0)-\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)]\triangleq I_{n11}+I_{n12}.$$

Note that $I_{n11}\leq\triangle n^{-1/2}\cdot||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)||$. By condition (A6),

$$\begin{aligned}
&\mathbb{E}[||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)||^2]\\
=&\mathbb{E}\Big\{\sum_{i=1}^n\boldsymbol{\epsilon}_i^\intercal\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{\epsilon}_i\Big\}\\
\leq&C\cdot\sum_{i=1}^n\mathrm{Tr}\Big(\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\Big)\\
=&C\sum_{i=1}^n\sum_{j=1}^m\cdot\mathrm{Tr}\Big(\mathrm{vec}^\intercal\boldsymbol{X}_{ij}\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_{ij}\Big)=O(n)
\end{aligned}$$

for some constant $C > 0$. This implies that $I_{n11} = \triangle n^{-1/2} O_p(n^{1/2}) = \triangle O_p(1)$. For I_{n12}, by Lemma B.1,

$$I_{n12} \leq ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \cdot ||\boldsymbol{s}_n(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)|| = o_p(1).$$

Therefore, I_{n1} is dominated in probability by I_{n11}.

For I_{n2} ,we decompose it into

$$\begin{aligned} I_{n2} =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ & -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal [\boldsymbol{D}_n(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ \triangleq & I_{n21} + I_{n22}. \end{aligned}$$

By Lemma B.2, it can be easily checked that $I_{n22} = o_p(1)$. Next, for I_{n21},

$$\begin{aligned} I_{n21} =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ & -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal [\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ & -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal [\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ \triangleq & I_{n21}^1 + I_{n21}^2 + I_{n21}^3. \end{aligned}$$

We next show that I_{n21} is dominated in probability by I_{n21}^1. Note that by conditions (A3), (A4) and (A7),

$$\begin{aligned} I_{n21}^1 =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal \Big[\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n) \mathrm{vec} \boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n) \tilde{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n) \mathrm{vec}^\intercal \boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n) \Big] (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\ \leq & -n^{-1} \triangle^2 \min_i \lambda_{\min}(\boldsymbol{A}_i(\boldsymbol{\beta}_n)) \lambda_{\min} \Big(\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_n) \mathrm{vec} \boldsymbol{X}_i \mathrm{vec}^\intercal \boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n) \Big) \lambda_{\min}(\tilde{\boldsymbol{R}}^{-1}) \\ \leq & -C \triangle^2, \end{aligned}$$

for some constant $C > 0$. By Lemma B.3, it can be checked directly that both I_{n21}^2 and I_{n21}^3 are $o_p(1)$.

Therefore, with high probability, the sign of $(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\intercal \boldsymbol{s}_n(\boldsymbol{\beta}_n)$ is determined by $I_{n11} + I_{n21}^1$ and is negative for sufficiently large $\triangle$, which completes the proof. □

Proof of Theorem 2. We first show that the normalized $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)$ has an asymptotic normal distribution. That is, for any $\boldsymbol{b} \in \mathbb{R}^{R \sum_{d=1}^D p_d}$ such that $||\boldsymbol{b}|| = 1$,

$$\boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0) \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) \to N(0,1), \tag{S4}$$

where $\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0) = \mathrm{Var}(\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0))$.

Denote $\boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) = \sum_{i=1}^n Z_{ni}$, where

$$Z_{ni} = \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0),$$

and $\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0) = \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_0)(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0))$. Note that $\mathbb{E}(Z_{ni}) = 0$, $\mathrm{Var}(\sum_{i=1}^n Z_{ni}) = 1$. To prove (S4), it suffices to check the Lyapunov condition. That is, for some $\delta > 0$,

$$\sum_{i=1}^n \mathbb{E}\Big(|Z_{ni}|^{2+\delta}\Big) \to 0,$$

as $n \to \infty$. By Cauchy-Schwarz inequality,

$$Z_{ni}^2 \le \lambda_{\max}(\tilde{\boldsymbol{R}}^{-2})\lambda_{\max}(\boldsymbol{A}_i(\boldsymbol{\beta}_0))||\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0)||^2\gamma_{ni},$$

where $\gamma_{ni} \triangleq \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{b}$. To evaluate $\max_{1\le i\le n}\gamma_{ni}$, we need to evaluate $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0))$. Note that

$$\begin{aligned}\boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\boldsymbol{b} \ge& C\boldsymbol{b}^\intercal\Big(\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big)\boldsymbol{b}\\ \ge& C\lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big),\end{aligned}$$

which implies $\lambda_{\min}\Big(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\Big) \ge \lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big)$. By condition (A3), $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)) = O(n^{-1})$ and hence $\max_{1\le i\le n}\gamma_{ni} = o(1)$.

It follows that, for any $\delta > 0$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\Big(|Z_{ni}|^{2+\delta}\Big) &\le \sum_{i=1}^n \mathbb{E}\Big(C^{1+\delta/2}\gamma_{ni}^{1+\delta/2}||\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0)||^{2+\delta}\Big)\\ &\le C(\max_{1\le i\le n}\gamma_{ni})^{\delta/2}\sum_{i=1}^n \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{b}\\ &\le C(\max_{1\le i\le n}\gamma_{ni})^{\delta/2}\lambda_{\max}\Big(\sum_{i=1}^n \boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\intercal\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big)\lambda_{\min}^{-1}\big(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\big)\\ &= o(1)O(n)O(n^{-1}) = o(1),\end{aligned}$$

which completes the proof of (S4).

To prove Theorem 2, note that because $\boldsymbol{s}_n(\widehat{\boldsymbol{\beta}}_n) = 0$, we have $\boldsymbol{s}_n(\boldsymbol{\beta}_0) = \boldsymbol{D}_n(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$, for some $\boldsymbol{\beta}_n^*$ between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Hence,

$$\begin{aligned}
&\boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) \\
=&\boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&+ \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\boldsymbol{D}_n(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&+ \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) - \boldsymbol{s}_n(\boldsymbol{\beta}_0)] \\
=&J_{n1} + J_{n2}(\boldsymbol{\beta}_n^*) + J_{n3}(\boldsymbol{\beta}_0).
\end{aligned}$$

By (S4), it is sufficient to prove that both $\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle n^{-1/2}} |J_{n2}(\boldsymbol{\beta}_n)|$ and $|J_{n3}(\boldsymbol{\beta}_0)|$ are $o_p(1)$.

For J_{n3}, recall that $||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) - \boldsymbol{s}_n(\boldsymbol{\beta}_0)|| = O_p(1)$ from Lemma B.1. Using the previous result that $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)) = O(n^{-1})$, it can be easily checked that $J_{n3}^2 = o_p(1)$ and hence $|J_{n3}| = o_p(1)$.

For J_{n2}, we have

$$\begin{aligned}
&\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle n^{-1/2}} |J_{n2}(\boldsymbol{\beta}_n)| \\
\le &\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle n^{-1/2}} \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\boldsymbol{D}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&+ \sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle n^{-1/2}} \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&+ \sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle n^{-1/2}} \boldsymbol{b}^\intercal \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
\triangleq &I_{n1} + I_{n2} + I_{n3}.
\end{aligned}$$

Notice that

$$I_{n1} \le C \times |\lambda_{\max}(\boldsymbol{D}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n))| \times \lambda_{\min}^{-1/2}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)) \times ||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0||.$$

By Lemma 3, we have $\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \le \triangle\sqrt{p/n}} |\lambda_{\max}(\boldsymbol{D}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n))| = O_p(\sqrt{np_n})$. Therefore, $I_{n1} = O_p(\sqrt{np_n})O(n^{-1/2})O_p(\sqrt{p/n}) = O_p(pn^{-1/2}) = o_p(1)$. Similarly, by Lemma B.3, we have $I_{n2} = o_p(1)$ and $I_{n3} = o_p(1)$. Therefore J_{n1} has the same asymptotic distribution as in $(S4)$, which completes the proof. □

Proof of Theorem 3. Denote the rank-R tensor GEE estimator by $\widehat{\boldsymbol{B}}_{(R)}$. For Gaussian response, the BIC can be written as

$$BIC(R) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \langle \boldsymbol{X}_{ij}, \widehat{\boldsymbol{B}}_{(R)}\rangle)^2 + \lambda_n R$$

where $\lambda_n = O(\log n)$.

The proof follows two steps: we need show that BIC neither overestimate nor underestimate the rank. By combining these two results, the consistency of BIC is established.

Step 1: To show BIC does not overestimate the rank, it suffices to show that for any $R > R_0$,

$$\begin{aligned}&\Pr\big(BIC(R) - BIC(R_0) > 0\big)\\ =&\Pr\big(\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) + (R-R_0)\lambda_n > 0\big) \to 1\end{aligned}$$

as $n \to \infty$, where

$$\ell(\widehat{\boldsymbol{B}}_{(R)}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \langle \boldsymbol{X}_{ij}, \widehat{\boldsymbol{B}}_{(R)}\rangle)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \theta_{ij}(\widehat{\boldsymbol{B}}_{(R)}))^2$$

by the identity link function.

Denote $\widehat{\boldsymbol{\beta}}_{(R)} = \mathrm{vec}(\widehat{\boldsymbol{B}}_{(R)1}, \ldots, \widehat{\boldsymbol{B}}_{(R)D})$, where $[\![\widehat{\boldsymbol{B}}_{(R)1}, \ldots, \widehat{\boldsymbol{B}}_{(R)D}]\!]$ is the CP-decomposition of $\widehat{\boldsymbol{B}}_{(R)}$. That is, $\widehat{\boldsymbol{\beta}}_{(R)}$ is the vector of free parameters in $\widehat{\boldsymbol{B}}_{(R)}$. By previous theorems, there exists one tensor GEE estimator $\widehat{\boldsymbol{\beta}}_{(R)}$ that is a root-n consistent estimator for $\boldsymbol{\beta}_{0(R)}$ for $R \geq R_0$, where $\boldsymbol{\beta}_{0(R)}$ is simply $\boldsymbol{\beta}_0$ with additional 0's in ranks $R_0+1, \ldots, R$ and $\boldsymbol{\beta}_{0(R_0)} = \boldsymbol{\beta}_0$. If we can show that $\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) = O_p(1)$, the proof is completed by the fact that $R - R_0 > 0$ and λ_n is a diverging sequence. Notice that by subtracting the same term,

$$\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) = \big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big).$$

Denote $L(\boldsymbol{\beta}) = \mathbb{E}[\ell(\boldsymbol{\beta})]$, where the expectation is taken w.r.t. Y_{ij}. We have

$$\begin{aligned}&\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\\ =&\big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) + \big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big).\end{aligned}$$

Therefore, it suffices to show that

$$\big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) = O_p(1), \tag{S5}$$

$$\big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) = O_p(1). \tag{S6}$$

To show (S5), by the definition of $\boldsymbol{\beta}_0$, we have $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0$. By Taylor expansion at $\boldsymbol{\beta}_0$ and Proposition 2.3 in Zhou et al. (2013),

$$L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0) = Cn||\widehat{\boldsymbol{\beta}}_{(R_0)} - \boldsymbol{\beta}_0||^{\intercal} I(\tilde{\boldsymbol{B}}_0)||\widehat{\boldsymbol{\beta}}_{(R_0)} - \boldsymbol{\beta}_0||$$

where $I(\tilde{\boldsymbol{B}}_0)$ is determined by some $\tilde{\boldsymbol{\beta}}_0 \in \{\boldsymbol{\beta} : ||\boldsymbol{\beta}-\boldsymbol{\beta}_0|| \leq \Delta n^{-1/2}\}$ via CP-decomposition. Under the condition (A3*), this term is $O_p(1)$.

Next we bound the term in (S6). By direct algebra, it can be shown that

$$\begin{aligned}
&\big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}[Y_{ij}]\big(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m} \big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) \\
\leq&\sum_{i=1}^{n}\sum_{j=1}^{m} \big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)Cn^{-1/2}
\end{aligned}$$

by the condition that $\partial\theta_{ij}/\partial\boldsymbol{\beta}$ are uniformly bounded, $|\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)| \leq Cn^{-1/2}$ for some constant C. Denote $g_i(\boldsymbol{u}) = \sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)Cn^{-1/2}$. Notice that $\{g_i(\boldsymbol{u})\}_{i=1}^{n}$ are independent mean zero random variables. Under the condition that $\mathrm{Var}(\boldsymbol{Y}_i)$ has bounded eigenvalues, it can be easily verified that $\mathrm{Var}(g_i(\boldsymbol{u})) = O(n^{-1})$. Therefore, $\sum_{i=1}^{n} g_i(\boldsymbol{u}) = O_p(1)$.

Using similar techniques, it can be shown that $\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_{0(R)}) = O_p(1)$ for $R > R_0$ as well. Therefore, for $R > R_0$, the term $BIC(R) - BIC(R_0)$ is asymptotically dominated by $(R - R_0)\log(n)$, which is always positive.

Step 2: To show BIC does not underestimate the rank, it suffices to show that for any $R < R_0$,

$$\begin{aligned}
&\Pr\big(BIC(R) - BIC(R_0) > 0\big) \\
=&\Pr\big(\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) + (R - R_0)\lambda_n > 0\big) \to 1
\end{aligned}$$

as $n \to \infty$. Notice that $n^{-1}(R - R_0)\lambda_n \to 0$ as $n \to \infty$. Therefore, if we can show that $n^{-1}\{\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)})\} \geq c$ for some constant $c > 0$, the proof is completed. Intuitively, we need to show that for any underestimated estimator, the increase of the population loss function to the one with correct rank is bounded away from zero.

Notice that

$$\begin{aligned}&\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\\ =&\big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big) + \big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big).\end{aligned}$$

Denote $\widehat{\boldsymbol{\beta}}_{(R),R_0}$ the augmented vector of $\widehat{\boldsymbol{\beta}}_{(R)}$ with 0's at the those rank $R+1,\ldots,R_0$ so that it has the same length as $\boldsymbol{\beta}_0$. By similar arguments in Step 1, for $R < R_0$

$$L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0) = Cn||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0||^{\intercal} I(\tilde{\boldsymbol{B}}_0)||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0||.$$

Notice that there exists some positive constant c_1 such that $||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0|| \geq c_1$. This is true because the elements of $\boldsymbol{\beta}_0$ at those locations for rank $R+1,\ldots,R_0$ cannot be all zeros. By the condition (A3*) that the smallest eigenvalue of $I(\boldsymbol{B})$ is bounded away from 0, it can be seen that $n^{-1}\{L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\} \geq c_2$ for some constant $c_2 > 0$ that does not depend on R.

Similar as in Step 1,

$$\begin{aligned}&\big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big)\\ =&\sum_{i=1}^{n}\sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big).\end{aligned}$$

By the condition that the first derivative of $\theta_{ij}(\boldsymbol{\beta})$ is bounded away from infinity, $\boldsymbol{X}_{ij}$ are uniformly bounded and p is fixed, $\mathrm{Var}[\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big)] = O(1)$. Therefore $\sum_{i=1}^{n}\sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) = O_p(\sqrt{n})$ and $n^{-1}\{\big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big)\} = o_p(1)$. Combined with previous result, $n^{-1}\{\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)})\}$ dominates the term in $n^{-1}\{\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\}$for sufficiently large n and is bounded away from 0, which completes the proof. □

Proof of Theorem 4. Write the SCAD regularized tensor GEE as

$$n^{-1}\boldsymbol{s}_n(\boldsymbol{\beta}_n) - \boldsymbol{q}_{\rho_n}(|\boldsymbol{\beta}_n|) \times \mathrm{sign}(\boldsymbol{\beta}_n),$$

where $\boldsymbol{q}_{\rho_n}(|\boldsymbol{\beta}_n|) = (q_{\rho_n}(|\beta_{n1}|),\ldots,q_{\rho_n}(|\beta_{nR\sum_{d=1}^{D}p_d}|))^{\intercal}$ is a $R\sum_{d=1}^{D}p_d$-dimensional vector of the subgradients of SCAD penalty, $q_{\rho_n}(\beta) = \rho_n\big\{1_{\{|\beta|\leq\rho_n\}} + (\lambda\rho_n - |\beta|)_+/(\lambda-1)1_{\{|\beta|>\rho_n\}}\big\}$, $\mathrm{sign}(\boldsymbol{\beta}_n) = (\mathrm{sign}(\beta_{n1}),\ldots,\mathrm{sign}(\beta_{nR\sum_{d=1}^{D}p_d}))^{\intercal}$, the symbol "×" denotes component-wise product, β_{nj} is the jth element of $\boldsymbol{\beta}_n$, $j = 1,\ldots,R\sum_{d=1}^{D}p_d$. Write the support of $\boldsymbol{\beta}_0$ as $\mathcal{J} = \{j : \beta_{0j} > 0\}$.

We prove the theorem by showing the the oracle estimator, $\widehat{\boldsymbol{\beta}}_n^O$, is an approximated solution to the regularized tensor GEE. Denote the jth element of $\widehat{\boldsymbol{\beta}}_n^O$ as $\widehat{\beta}_{nj}^O$. By the definition of the oracle estimator, $\widehat{\beta}_{nj}^O = 0$ for $j \notin \mathcal{J}$. Similar as the definition in Wang et al. (2012), an approximated solution to the regularized tensor GEE, $\widehat{\boldsymbol{\beta}}_n$, is defined to satisfy

$$\Pr\left(n^{-1}s_{nj}(\boldsymbol{\beta}_n) - q_{\rho_n}(|\beta_{nj}|)\text{sign}(\beta_{nj}) = 0,\ j \in \mathcal{J}\right) \to 1, \tag{S7}$$

$$\Pr\left(|n^{-1}s_{nj}(\boldsymbol{\beta}_n) - q_{\rho_n}(|\beta_{nj}|)\text{sign}(\beta_{nj})| \le \rho_n/\log n,\ j \notin \mathcal{J}\right) \to 1. \tag{S8}$$

The reason for this definition of the approximated solution is that the regularized tensor GEE involves non-smooth points, so the exact solution may not exist. It suffices to show that $\widehat{\boldsymbol{\beta}}_n^O$ satisfies both (S7) and (S8).

For (S7), note that by consistency in Theorem 1, $||\widehat{\boldsymbol{\beta}}_{n\mathcal{J}}^O - \boldsymbol{\beta}_{0\mathcal{J}}|| = O_p(n^{-1/2})$, where $\widehat{\boldsymbol{\beta}}_{n\mathcal{J}}^O = \{\widehat{\beta}_{nj}^O : j \in \mathcal{J}\}$ and similarly for $\boldsymbol{\beta}_{0\mathcal{J}}$. For fixed p, there exists some constant $C > 0$ that $\min_j \beta_{0j} > C$. Therefore, $\Pr(\min_{j\in\mathcal{J}} \widehat{\beta}_{nj}^O > C) \to 1$ as $n \to \infty$. By the fact $\rho_n = o(1)$, $\Pr(\min_{j\in\mathcal{J}} \widehat{\beta}_{nj}^O > \lambda\rho_n) \to 1$. By the definition of the oracle estimator, $s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) = 0$. Therefore (S7) holds for the oracle estimator.

For (S8), by the definition of the oracle estimator, $q_{\rho_n}(|\widehat{\beta}_{nj}^O|)\text{sign}(\widehat{\beta}_{nj}^O) = 0$ for $j \notin \mathcal{J}$. Therefore, it suffices to show $\Pr\left(|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| \le n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 1$. Note that $|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| \le |s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) - \tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| + |\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)|$. By Lemma B.1 and the consistency of the oracle estimator established in Theorem 1,

$$\max_{j\notin\mathcal{J}} \Pr(|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) - \tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| > n\rho_n/\log n) \to 0.$$

Therefore, we only need to verify $\Pr\left(|\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$.

Consider the Taylor expansion

$$\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O) = \tilde{s}_{nj}(\boldsymbol{\beta}_0) + \nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0),$$

where $\nabla_j(\boldsymbol{\beta}) = \partial\tilde{s}_{nj}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $\boldsymbol{\psi}_j(\boldsymbol{\beta}) = \partial^2\tilde{s}_{nj}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal$, $\boldsymbol{\beta}_n^*$ is between $\widehat{\boldsymbol{\beta}}_n^O$ and $\boldsymbol{\beta}_0$.

We first show $\Pr\left(|\tilde{s}_{nj}(\boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$. Note that

$$n^{-1}\tilde{s}_{nj}(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \boldsymbol{e}_j^\intercal \boldsymbol{J}^\intercal(\boldsymbol{\beta}_0)\text{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0) \triangleq n^{-1}\sum_{i=1}^n Z_i.$$

Note that Z_i are independent random variables with mean zero. By condition (A4)-(A7), it can be directly verified that $\mathbb{E}(|Z_i|^l) \le l!C_1^{l-2}C_2$ for some constants $C_1 > 0$ and

$C_2 > 0$. Therefore, $\Pr\left(|n^{-1}\tilde{s}_{nj}(\boldsymbol{\beta}_0)| > \rho_n/\log n\right) \leq \exp[-Cn\rho_n^2/(\log n)^2] \to 0$ is implied by the Bernstein's inequality for any $j \notin \mathcal{J}$. By the condition that $n\rho_n^2/(\log n)^2 \to \infty$, the proof of this step is completed.

We next show $\Pr\left(|\nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$. Similar to the decomposition used in the proof of Lemma B.2, we write $\nabla_j(\boldsymbol{\beta}_0) = \sum_{m=1}^4 \tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0)$, where $\tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0) = \boldsymbol{e}_j^\intercal \tilde{\boldsymbol{D}}_{nm}(\boldsymbol{\beta}_0)$ for $m = 1, \ldots, 4$. By condition (A4)-(A8), the elements of $n^{-1}\tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0)$ are uniformly bounded by a positive constant for $j \notin \mathcal{J}$ and $m = 1, \ldots, 4$. Therefore, $|\nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| = O_p(n^{1/2}) = o_p(n\rho_n/\log n)$, which compiles the proof.

Finally, we show $\Pr\left(|(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$. It can be directly verified that the elements of $n^{-1}\boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)$ are uniformly bounded by a positive constant. Therefore, $|(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| = O_p(1)$, which completes the proof. □

References

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*, volume 30. SIAM.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *The Annals of Statistics*, pages 1298–1309.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39(1):389–417.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.