Variable Selection with the Strong Heredity Constraint and Its Oracle Property - Supplemental Materials

October 19, 2009

Part 1: Regularity Conditions

Regularity Conditions for Section 3.1

(C1) The observations $\{\boldsymbol{V}_i : i = 1, \ldots, n\}$ are independent and identically distributed with a probability density $f(\boldsymbol{V}, \boldsymbol{\theta})$, which has a common support. We assume the density f satisfies the following equations:

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_j}\right] = 0 \quad \text{for } j = 1, \ldots, \frac{p(p+1)}{2},$$

and

$$\begin{aligned} \boldsymbol{I}_{jk}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \theta_j} \log f(\boldsymbol{V}, \boldsymbol{\theta}) \frac{\partial}{\partial \theta_k} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right] \\ &= E_{\boldsymbol{\theta}}\left[- \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right]. \end{aligned}$$

(C2) The Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right)^{\mathsf{T}}\right]$$

is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

(C3) There exists an open set ω of Ω that contains the true parameter point $\boldsymbol{\theta}^*$ such that for almost all $\boldsymbol{V}$ the density $f(\boldsymbol{V}, \boldsymbol{\theta})$ admits all third derivatives $(\partial^3 f(\boldsymbol{V}, \boldsymbol{\theta}))/(\partial\theta_j \partial\theta_k \partial\theta_l)$ for all $\boldsymbol{\theta} \in \omega$ and any $j,\ k,\ l = 1, \ldots, p(p+1)/2$. Furthermore, there exist functions M_{jkl} such that

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(\boldsymbol{V}, \boldsymbol{\theta}) \right| \leq M_{jkl}(\boldsymbol{V}) \quad \text{for all } \boldsymbol{\theta} \in \omega,$$

where $m_{jkl} = E_{\boldsymbol{\theta}^*}[M_{jkl}(\boldsymbol{V})] < \infty$.

Regularity Conditions for Section 3.2

(C4) The observations $\{\boldsymbol{V}_{ni} : i = 1, \ldots, n\}$ are independent and identically distributed with a probability density $f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)$, which has a common support. We assume the density f_n satisfies the following equations:

$$E_{\boldsymbol{\theta}_n}\left[\frac{\partial \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)}{\partial\theta_{nj}} \right] = 0 \quad \text{for } j = 1, \ldots, q_n,$$

and

$$\begin{aligned} \boldsymbol{I}_{jk}(\boldsymbol{\theta}_n) &= E_{\boldsymbol{\theta}_n}\left[\frac{\partial}{\partial\theta_{nj}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n) \frac{\partial}{\partial\theta_{nk}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n) \right] \\ &= E_{\boldsymbol{\theta}_n}\left[- \frac{\partial^2}{\partial\theta_{nj}\partial\theta_{nk}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n) \right]. \end{aligned}$$

(C5) $I_n(\boldsymbol{\theta}_n) = E[(\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n})(\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n})^\top]$ satisfies $0 < C_1 < \lambda_{\min}\{I_n(\boldsymbol{\theta}_n)\} \leq \lambda_{\max}\{I_n(\boldsymbol{\theta}_n)\}$ $< C_2 < \infty$ for all n, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and the largest eigenvalues of a matrix respectively. Moreover, for any $j, k = 1, 2, \ldots, q_n$,

$$E_{\boldsymbol{\theta}_n}\left\{ \frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial\theta_{nj}} \frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial\theta_{nk}} \right\}^2 < C_3 < \infty,$$

and

$$E_{\boldsymbol{\theta}_n}\left\{ \frac{\partial^2 \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nk}} \right\}^2 < C_4 < \infty.$$

(C6) There exists a large open set $\omega_n \subset \Omega_n \in \mathbb{R}^{q_n}$ which contains the true parameter $\boldsymbol{\theta}_n^*$ such that for almost all $\boldsymbol{V}_{ni}$ the density admits all third derivatives $\partial^3 f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n)/\partial\theta_{nj}\partial\theta_{nk}\partial\theta_{nl}$

for all $\boldsymbol{\theta}_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that

$$\left| \frac{\partial^3 \log f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n)}{\partial \theta_{nj} \partial \theta_{nk} \partial \theta_{nl}} \right| \leq M_{njkl}(\boldsymbol{V}_{ni})$$

for all $\boldsymbol{\theta}_n \in \omega_n$ and

$$E_{\boldsymbol{\theta}_n} M_{njkl}^2(\boldsymbol{V}_{ni}) < C_5 < \infty$$

for all q_n, n, and j, k, l.

Part 2: Proofs

Proof of Lemma 1

Let $\eta_n = n^{-1/2} + a_n$ and $\{\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$ be the ball around $\boldsymbol{\theta}^*$, where $\boldsymbol{\delta} = (u_1, \ldots, u_p, v_{12}, \ldots, v_{p-1,p})^\intercal = (\boldsymbol{u}^\intercal, \boldsymbol{v}^\intercal)^\intercal$. Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}^*).$$

Let $-L_n$ denote the first term of Q_n in (8). For $\boldsymbol{\delta}$ that satisfies $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) + n \sum_j \lambda_j^\beta \big(|\beta_j^* + \eta_n u_j| - |\beta_j^*|\big) \\
&\quad + n \sum_{k<k'} \lambda_{kk'}^\gamma \big(|\gamma_{kk'}^* + \eta_n v_{kk'}| - |\gamma_{kk'}^*|\big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) + n \sum_{j \in \mathcal{A}_1} \lambda_j^\beta \big(|\beta_j^* + \eta_n u_j| - |\beta_j^*|\big) \\
&\quad + n \sum_{(k,k') \in \mathcal{A}_2} \lambda_{kk'}^\gamma \big(|\gamma_{kk'}^* + \eta_n v_{kk'}| - |\gamma_{kk'}^*|\big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n \sum_{j \in \mathcal{A}_1} \lambda_j^\beta |u_j| - n\eta_n \sum_{(k,k') \in \mathcal{A}_2} \lambda_{kk'}^\gamma |v_{kk'}| \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n^2 \Big(\sum_{j \in \mathcal{A}_1} |u_j| + \sum_{(k,k') \in \mathcal{A}_2} |v_{kk'}| \Big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|) d \\
&= -[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal (\eta_n \boldsymbol{\delta}) - \frac{1}{2} (\eta_n \boldsymbol{\delta})^\intercal [\nabla^2 L_n(\boldsymbol{\theta}^*)] (\eta_n \boldsymbol{\delta}) (1 + o_p(1)) \\
&\quad - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|) d. \qquad (10)
\end{aligned}$$

We split (10) into three parts:

$$\begin{aligned}
A_1 &= -[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal(\eta_n\boldsymbol{\delta}) \\
A_2 &= -\frac{1}{2}(\eta_n\boldsymbol{\delta})^\intercal[\nabla^2 L_n(\boldsymbol{\theta}^*)](\eta_n\boldsymbol{\delta})(1+o_p(1)) \\
A_3 &= -n\eta_n^2(|\mathcal{A}_1|+|\mathcal{A}_2|)d
\end{aligned}$$

Then

$$\begin{aligned}
A_1 &= -\eta_n[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal\boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\Big(\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}^*)\Big)^\intercal\boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\Big(\sqrt{n}\ \frac{1}{n}\sum_{i=1}^{n}\nabla\log f(\boldsymbol{V}_i,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\Big)^\intercal\boldsymbol{\delta} \\
&= -O_p(\sqrt{n}\eta_n)\boldsymbol{\delta} \\
&= -O_p(n\eta_n^2)\boldsymbol{\delta},
\end{aligned}$$

$$\begin{aligned}
A_2 &= \frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[-\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1)) \\
&= \frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[I(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1)) \ \text{ by the weak law of large numbers.}
\end{aligned}$$

Thus,

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &\geq A_1+A_2+A_3 \\
&= -n\eta_n^2 O_p(1)\boldsymbol{\delta}+\frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[I(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1))-n\eta_n^2(|\mathcal{A}_1|+|\mathcal{A}_2|)d. \qquad (11)
\end{aligned}$$

Notice that A_2 dominates the rest terms A_1 and A_3 and is positive since $I(\boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ from (C2). Therefore, for any given $\epsilon>0$, there exists a large enough constant d such that

$$P\Big\{\inf_{||\delta||=d} Q_n(\boldsymbol{\theta}^*+\eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}^*)\Big\} \geq 1-\epsilon.$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer in the ball $\{\boldsymbol{\theta}^*+\eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$. Thus, there exists a local minimizer of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*||=O_p(\eta_n)$. $\square$

Proof of Theorem 1

We first consider $P(\hat{\beta}_{\mathcal{A}_1^c} = 0) \to 1$. It is sufficient to show for any $j \in \mathcal{A}_1^c$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} \quad < \quad 0 \quad \text{for } -\epsilon_n < \hat{\beta}_j < 0 \tag{12}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} \quad > \quad 0 \quad \text{for } \; 0 < \hat{\beta}_j < \epsilon_n \tag{13}$$

with probability tending to 1 where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (13), notice

$$\begin{aligned} \frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} &= -\frac{L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} + n\lambda_j^{\beta}\text{sgn}(\hat{\beta}_j) \\ &= -\frac{L_n(\boldsymbol{\theta}^*)}{\partial \beta_j} - \sum_{k=1}^{\frac{p(p+1)}{2}} \frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial \beta_j \partial \theta_k}(\hat{\theta}_k - \theta_k^*) \\ &\quad - \sum_{k=1}^{\frac{p(p+1)}{2}} \sum_{l=1}^{\frac{p(p+1)}{2}} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}})}{\partial \beta_j \partial \theta_k \partial \theta_l}(\hat{\theta}_k - \theta_k^*)(\hat{\theta}_l - \theta_l^*) + n\lambda_j^{\beta}\text{sgn}(\hat{\beta}_j) \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$. By (C1)–(C3) and the condition $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = O_p(n^{-1/2})$,

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} \quad = \quad \sqrt{n}\Big\{O_p(1) + \sqrt{n}\lambda_j^{\beta}\text{sgn}(\hat{\beta}_j)\Big\}.$$

As $\sqrt{n}\lambda_j^{\beta} \to \infty$ for $j \in \mathcal{A}_1^c$ from the assumption, the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j}$ is dominated by $\text{sgn}(\hat{\beta}_j)$. Therefore,

$$P\Big[\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} > 0 \quad \text{for } \; 0 < \hat{\beta}_j < \epsilon_n\Big] \to 1 \quad \text{as } n \to \infty.$$

(12) can be shown in the same way.

Next, we prove $P(\hat{\boldsymbol{\gamma}}_{\mathcal{A}_2^c} = 0) \to 1$.

- For (k, k') where $(k, k') \in \mathcal{A}_2^c$ and $k, k' \in \mathcal{A}_1$: we can prove $P(\hat{\boldsymbol{\gamma}}_{kk'} = 0) \to 1$ by a similar reasoning.

- For (k, k') where $(k, k') \in \mathcal{A}_2^c$ and either k or k' is in $\mathcal{A}_1^c$: without loss of generality, assume that $\beta_k^* = 0$. Notice that $\hat{\beta}_k = 0$ implies $\hat{\gamma}_{kk'} = 0$, because if $\hat{\gamma}_{kk'} \neq 0$, then the value of the loss function does not change but the value of the penalty function will increase. Since we already have $P(\hat{\beta}_k = 0) \to 1$, we can conclude $P(\hat{\gamma}_{kk'} = 0) \to 1$ as well.

$\square$

Proof of Theorem 2

Let $Q_n(\boldsymbol{\theta}_{\mathcal{A}})$ denote the objective function Q_n only on the $\mathcal{A}$-component of $\boldsymbol{\theta}$, that is, $Q_n(\boldsymbol{\theta})$ with $\boldsymbol{\theta}_{\mathcal{A}^c}$. Based on Lemma 1 and Theorem 1, we have $P\big(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0\big) \to 1$. Thus,

$$P\Big[\arg\min_{\boldsymbol{\theta}_{\mathcal{A}}} Q_n(\boldsymbol{\theta}_{\mathcal{A}}) = \big(\mathcal{A}\text{-component of } \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta})\big)\Big] \to 1.$$

It means that $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ should satisfy

$$\left.\frac{\partial Q_n(\boldsymbol{\theta}_{\mathcal{A}})}{\partial \theta_j}\right|_{\boldsymbol{\theta}_{\mathcal{A}}=\hat{\boldsymbol{\theta}}_{\mathcal{A}}} = 0, \quad \forall j \in \mathcal{A} \tag{14}$$

with probability tending to 1.

Let $L_n(\boldsymbol{\theta}_{\mathcal{A}})$ and $P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}})$ denote the log-likelihood function of $\boldsymbol{\theta}_{\mathcal{A}}$ and the penalty function of $\boldsymbol{\theta}_{\mathcal{A}}$ respectively so that we have

$$Q_n(\boldsymbol{\theta}_{\mathcal{A}}) = -L_n(\boldsymbol{\theta}_{\mathcal{A}}) + nP_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}}).$$

From (14), now we have

$$\nabla_{\mathcal{A}} Q_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) = -\nabla_{\mathcal{A}} L_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) + n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) = \mathbf{0}, \tag{15}$$

with probability tending to 1.

- Consider the first term in (15). By the Taylor expansion of $-\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}_{\mathcal{A}})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$,

$$\begin{aligned}
-\nabla_{\mathcal{A}} L_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) &= -\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) - \big[\nabla^2_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\big]\big(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}\big) \\
&= \sqrt{n}\Big[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \big(-\frac{1}{n}\nabla^2_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) - o_p(1)\big)\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}})\Big] \\
&= \sqrt{n}\Big[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\Big].
\end{aligned}$$

- Consider the second term in (15). By the Taylor expansion of $n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$,

$$\begin{aligned}
n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) &= n\left\{\begin{bmatrix} \lambda^{\beta}_j \mathrm{sgn}(\beta_j) \\ \lambda^{\gamma}_{kk'} \mathrm{sgn}(\gamma_{kk'}) \end{bmatrix}_{j\in\mathcal{A}_1,(k,k')\in\mathcal{A}_2} + o_p(1)(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}})\right\} \\
&= \sqrt{n}o_p(1)
\end{aligned}$$

because $\sqrt{n}a_n = o(1)$ and $\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}\| = O_p(n^{-1/2})$.

Thus,

$$0 = \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\right].$$

It follows

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) = \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})^{-1}\ \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathcal{A}} \log f(\boldsymbol{V}_i, \boldsymbol{\theta}_{\mathcal{A}}) + o_p(1).$$

Therefore, by central limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \to_d N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}^*_{\mathcal{A}})).$$

□

Proof of Lemma 2

Let $\eta_n = \sqrt{q_n}(n^{-1/2} + a_n)$ and $\{\boldsymbol{\theta}^*_n + \eta_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\| \le d\}$ be the ball around $\boldsymbol{\theta}^*_n$, where $\boldsymbol{\delta} = (u_1, \ldots, u_{p_n}, v_{12}, \ldots, v_{p_n-1,p_n})^\intercal = (\boldsymbol{u}^\intercal, \boldsymbol{v}^\intercal)^\intercal$. It is sufficient to show that for any $\epsilon > 0$, there is a large constant d such that

$$P\left\{\inf_{\|\delta\|=d} Q_n(\boldsymbol{\theta}^*_n + \eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}^*_n)\right\} \ge 1 - \epsilon,$$

because it implies that with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\theta}^*_n + \eta_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \le d\}$. Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}^*_n + \eta_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}^*_n).$$

Let $-L_n$ and nP_n denote the first and the second terms of Q_n in (9). For any $\boldsymbol{\delta}$ satisfying $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) + nP_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) - nP_n(\boldsymbol{\theta}_n^*) \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&\quad + n\Big\{\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}(|\beta_j + \eta_n u_j| - |\beta_j|) + \sum_{(k,k')\in\mathcal{A}_{n2}} \lambda_{n,kk'}^{\gamma}(|\gamma_{kk'} + \eta_n v_{kk'}| - |\gamma_{kk'}|)\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n\Big\{\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}|u_j| + \sum_{(k,k')\in\mathcal{A}_{n2}} \lambda_{n,kk'}^{\gamma}|v_{kk'}|\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n\Big\{\sum_{j\in\mathcal{A}_{n1}} a_n|u_j| + \sum_{(k,k')\in\mathcal{A}_{n2}} a_n|v_{kk'}|\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n(\sqrt{s_n}a_n)d \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n^2 d.
\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &\geq -\nabla^{\intercal}L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta}) - \frac{1}{2}(\eta_n\boldsymbol{\delta})^{\intercal}\nabla^2 L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta}) - \frac{1}{6}\nabla^{\intercal}\big\{\boldsymbol{\delta}^{\intercal}\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{\delta}\big\}\boldsymbol{\delta}\eta_n^3 - n\eta_n^2 d \\
&\equiv A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}$ and $\boldsymbol{\theta}_n^*$. We first consider A_1.

$$\begin{aligned}
|A_1| &= |-\nabla^{\intercal}L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta})| \\
&\leq \eta_n\|\nabla^{\intercal}L_n(\boldsymbol{\theta}_n^*)\|\|\boldsymbol{\delta}\| \\
&= O_p(\eta_n\sqrt{nq_n})d = O_p(n\eta_n^2)d.
\end{aligned}$$

Next, since we have

$$\Big\|\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}_n^*) + \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)\Big\| = o_p\Big(\frac{1}{q_n}\Big) \tag{16}$$

by Chebyshev's inequality and (C5), we can show that

$$\begin{aligned}
A_2 &= -\frac{1}{2}\eta_n^2\Big[\boldsymbol{\delta}^{\intercal}\nabla^2 L_n(\boldsymbol{\theta}_n^*)\boldsymbol{\delta}\Big] \\
&= -\frac{1}{2}\boldsymbol{\delta}^{\intercal}\left[\frac{1}{n}\Big\{\nabla^2 L_n(\boldsymbol{\theta}_n^*) - E(\nabla^2 L_n(\boldsymbol{\theta}_n^*))\Big\}\right]\boldsymbol{\delta}\cdot n\eta_n^2 - \frac{1}{2}\boldsymbol{\delta}^{\intercal}\frac{1}{n}E(\nabla^2 L_n(\boldsymbol{\theta}_n^*))\boldsymbol{\delta}\cdot n\eta_n^2 \\
&= \frac{1}{2}n\eta_n^2\boldsymbol{\delta}^{\intercal}\boldsymbol{I}_n(\boldsymbol{\theta}_n^*)\boldsymbol{\delta} - \frac{1}{2}n\eta_n^2 d^2 o_p(1).
\end{aligned}$$

Moreover, by Cauchy-Schwarz inequality, (C6), and the conditions $\sqrt{n}a_n \to 0$ and $q_n^5/n \to 0$,

$$
\begin{aligned}
|A_3| &= \Big| -\frac{1}{6}\nabla^{\intercal}\big\{\boldsymbol{\delta}^{\intercal}\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{\delta}\big\}\boldsymbol{\delta}\eta_n^3 \Big| \\
&= \frac{1}{6}\eta_n^3 \Big| \sum_{i=1}^{n}\sum_{j,k,l=1}^{q_n} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\theta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\delta_j\delta_k\delta_l \Big| \\
&\le \eta_n^3 \sum_{i=1}^{n}\Big(\sum_{j,k,l=1}^{q_n} M_{njkl}^2(\boldsymbol{V}_{ni})\Big)^{1/2}\|\boldsymbol{\delta}\|^3 \\
&= n\eta_n^3\ O_p\big(q_n^{3/2}\big)\ \big(q_n O(1)\big)^{1/2}\|\boldsymbol{\delta}\|^2 \\
&= n\eta_n^2\ O_p\big(\eta_n q_n^2\big)d^2 \\
&= n\eta_n^2 o_p(1)d^2.
\end{aligned}
$$

A_2 dominates the rest terms A_1, A_3 and A_4 for a sufficiently large $\boldsymbol{\delta}$, and is positive because $\boldsymbol{I}_n(\boldsymbol{\theta}_n^*)$ is positive definite by (C5). $\square$

Proof of Theorem 3

Proof of (a)

We first prove $P(\hat{\beta}_{nj} = 0) \to 1$ for $j \in \mathcal{A}_{n1}^c$ as $n \to \infty$. It is enough to show that with probability tending to 1, for any $j \in \mathcal{A}_{n1}^c$,

$$
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} < 0 \quad \text{for } -\epsilon_n < \hat{\beta}_{nj} < 0 \tag{17}
$$

$$
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} > 0 \quad \text{for } 0 < \hat{\beta}_{nj} < \epsilon_n \tag{18}
$$

where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (18), we consider a Taylor expansion of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_n^*$.

$$
\begin{aligned}
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} &= -\frac{\partial L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} + n\lambda_{nj}^{\beta}\mathrm{sgn}(\hat{\beta}_{nj}) \\
&= -\frac{\partial L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}} - \sum_{k=1}^{q_n} \frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}(\hat{\theta}_{nk} - \theta_{nk}^*) \\
&\quad - \sum_{k=1}^{q_n}\sum_{l=1}^{q_n} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}\partial \theta_{nk}\partial \theta_{nl}}(\hat{\theta}_{nk} - \theta_{nk}^*)(\hat{\theta}_{nl} - \theta_{nl}^*) \\
&\quad + n\lambda_{nj}^{\beta}\mathrm{sgn}(\hat{\beta}_{nj}) \\
&\equiv I_1 + I_2 + I_3 + I_4
\end{aligned} \tag{19}
$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $\boldsymbol{\theta}_n^*$ and $\hat{\boldsymbol{\theta}}_n$. By Chebyshev's inequality,

$$
I_1 = -\sum_{i=1}^{n} \frac{\partial \log f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n^*)}{\partial \beta_{nj}} = O_p(\sqrt{n}) = O_p(\sqrt{nq_n}).
$$

Next,

$$
\begin{aligned}
I_2 &= -\sum_{k=1}^{q_n} \frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}(\hat{\theta}_{nk} - \theta_{nk}^*) \\
&= -\sum_{k=1}^{q_n}\left[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}} - E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}\Big]\right](\hat{\theta}_{nk} - \theta_{nk}^*) - \sum_{k=1}^{q_n} E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}\Big](\hat{\theta}_{nk} - \theta_{nk}^*) \\
&\equiv K_1 + K_2.
\end{aligned}
$$

By Cauchy-Schwarz inequality and (C5),

$$
\begin{aligned}
|K_1| &\le \left[\sum_{k=1}^{q_n}\Big\{\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}} - E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}\Big]\Big\}^2\right]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\| \\
&= O_p(\sqrt{nq_n})O_p(\sqrt{q_n/n}) \\
&= O_p(\sqrt{nq_n})o_p(1) = o_p(\sqrt{nq_n}).
\end{aligned}
$$

Again, by Cauchy-Schwarz inequality and (C5),

$$
\begin{aligned}
|K_2| &= n\Big|\sum_{k=1}^{q_n} \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)_{(j,k)}(\hat{\theta}_{nk} - \theta_{nk}^*)\Big| \\
&\le n\Big[\sum_{k=1}^{q_n} \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)_{(j,k)}^2\Big]^{1/2}\Big[\sum_{k=1}^{q_n}(\hat{\theta}_{nk} - \theta_{nk}^*)^2\Big]^{1/2} \\
&= n\; O(1)O_p(\sqrt{q_n/n}) = O_p(\sqrt{nq_n}).
\end{aligned}
$$

Therefore, $I_2 = O_p(\sqrt{nq_n})$.

$$
\begin{aligned}
I_3 &= -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}(\hat{\theta}_{nk}-\theta^*_{nk})(\hat{\theta}_{nl}-\theta^*_{nl}) \\
&= -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}} - E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big]\Big](\hat{\theta}_{nk}-\theta^*_{nk})(\hat{\theta}_{nl}-\theta^*_{nl}) \\
&\quad -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n} E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big](\hat{\theta}_{nk}-\theta^*_{nk})(\hat{\theta}_{nl}-\theta^*_{nl}) \\
&\equiv K_3 + K_4.
\end{aligned}
$$

By Cauchy-Schwarz inequality and (C6),

$$
\begin{aligned}
|K_4| &\le \Big[\sum_{k=1}^{q_n}\sum_{l=1}^{q_n} n^2\Big\{E\big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\big]\Big\}^2\Big]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_n\|^2 \\
&\le \Big[q_n^2 n^2 C_5\Big]^{1/2} O_p(q_n/n) \\
&= O_p(q_n^2) = O_p(\sqrt{nq_n})O_p(\sqrt{q_n^3/n}) = O_p(\sqrt{nq_n})o_p(1) \\
&= o_p(\sqrt{nq_n}).
\end{aligned}
$$

By Cauchy-Schwarz inequality and (C6),

$$
\begin{aligned}
|K_3| &\le \Big[\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\Big\{\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}} - E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big]\Big\}^2\Big]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_n\|^2 \\
&= \Big[nq_n^2 O_p(1)\Big]^{1/2} O_p(q_n/n) \\
&= o_p(\sqrt{nq_n}).
\end{aligned}
$$

Thus, $I_1 + I_2 + I_3 = O_p(\sqrt{nq_n})$. Therefore, returning to (19),

$$
\begin{aligned}
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} &= O_p(\sqrt{nq_n}) + n\lambda^{\beta}_{nj}\mathrm{sgn}(\hat{\beta}_{nj}) \\
&= \sqrt{nq_n}\Big\{O_p(1) + \sqrt{\frac{n}{q_n}}\lambda^{\beta}_{nj}\mathrm{sgn}(\hat{\beta}_{nj})\Big\}.
\end{aligned}
$$

Since $\sqrt{n/q_n}b_n \to \infty$, $\mathrm{sgn}(\hat{\beta}_{nj})$ dominates the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}}$ when n is large. Therefore, for $0 < \hat{\beta}_{nj} < \epsilon_n$, $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} > 0$ with probability tending to 1 as $n \to \infty$. (17) can be shown in the same way.

Next, we prove $P(\hat{\boldsymbol{\gamma}}_{n\mathcal{A}^c_{n2}} = 0) \to 1$.

- For (k,k') where $(k,k') \in \mathcal{A}_{n2}^c$ and $k,k' \in \mathcal{A}_{n1}$: we can prove $P(\hat{\boldsymbol{\gamma}}_{n,kk'}=0) \to 1$ by a similar reasoning.

- For (k,k') where $(k,k') \in \mathcal{A}_{n2}^c$ and either k or k' is in $\mathcal{A}_{n1}^c$: without loss of generality, assume that $\beta_{nk}^*=0$. Notice that $\hat{\beta}_{nk}=0$ implies $\hat{\gamma}_{n,kk'}=0$, because if $\hat{\gamma}_{n,kk'} \neq 0$, then the value of the loss function does not change but the value of the penalty function will increase. Since we already have $P(\hat{\beta}_{nk}=0) \to 1$, we can conclude $P(\hat{\gamma}_{n,kk'}=0) \to 1$ as well.

Proof of (b)

We want to show that with probability tending to 1,

$$\begin{aligned}
\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n}) &= \sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&= \sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\{\frac{1}{n}\nabla L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})+ \\
&\quad o_p(n^{-1/2})\Big\} \qquad (20) \\
&= \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\sum_{i=1}^n\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big] \\
&\quad +o_p(\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\mathbf{1}_{(s_n\times 1)}) \\
&= \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\sum_{i=1}^n\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big]+o_p(1) \\
&\equiv \sum_{i=1}^n\boldsymbol{Y}_{ni}+o_p(1) \\
&\to_d N(\mathbf{0},\boldsymbol{G}), \qquad (21)
\end{aligned}$$

where $\boldsymbol{Y}_{ni}=\frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big]$. We will show (20) and (21) in (I) and (II) respectively.

(I) We want to show $\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n})=\frac{1}{n}\nabla L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})+o_p(\frac{1}{\sqrt{n}})$. We know that with probability tending to 1,

$$\mathbf{0}=\nabla_{\mathcal{A}_n}Q_n(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n})=-\nabla_{\mathcal{A}_n}L_n(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n})+n\nabla_{\mathcal{A}_n}P_{\lambda_n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}).$$

By Taylor expansion at $\boldsymbol{\theta} = \boldsymbol{\theta}^*_{n\mathcal{A}_n}$

$$\begin{aligned}
\mathbf{0} &= -\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) - \Big[\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\frac{1}{2}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\mathsf{T}}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) + n\nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) &= -\frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&= \frac{1}{n}\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\frac{1}{2}\frac{1}{n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\mathsf{T}}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}).
\end{aligned}$$

Therefore, it is sufficient to show that

$$\begin{aligned}
&-\frac{1}{2}\frac{1}{n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\mathsf{T}}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) - \nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&+ \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
\equiv\ & B_1 + B_2 + B_3 \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

First, by Cauchy-Schwarz inequality and (C6),

$$\begin{aligned}
\|B_1\|^2 &\leq \frac{1}{n^2}\|\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\|^2\|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^4 \\
&\leq \frac{1}{n^2}\sum_{j,k,l\in\mathcal{A}_n}\Big\{\sum_{i=1}^n M_{njkl}(\boldsymbol{V}_{ni})\Big\}^2\|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^4 \\
&= \frac{1}{n^2}\sum_{j,k,l\in\mathcal{A}_n} n^2 O_p(1) O_p(\frac{q_n^2}{n}) \\
&= O_p(q_n^5/n^2) \\
&= o_p(1/n).
\end{aligned}$$

Second, because $a_n = o(1/\sqrt{nq_n})$ from the condition of the theorem,

$$\begin{aligned}
\|B_2\|^2 &= \Big\|(\lambda^\beta_{n1}\mathrm{sgn}(\beta^*_{n1}), \ldots, \lambda^\gamma_{n,(p_n-1,p_n)}\mathrm{sgn}(\gamma^*_{n,(p_n-1,p_n)}))^\intercal\Big\|^2 \\
&\leq s_n\Big[\max\{\lambda^\beta_{nj}, \lambda^\gamma_{n,kk'} : j \in \mathcal{A}_{n1}, (k,k') \in \mathcal{A}_{n2}\}\Big]^2 \\
&= s_n a_n^2 = s_n o(1/nq_n) \\
&= o(1/n).
\end{aligned}$$

Third, based on (16), it can be shown that

$$\begin{aligned}
\|B_3\|^2 &\leq \|\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^2\|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^2 \\
&= o_p(1/q_n^2)O_p(q_n/n) = o_p(1/nq_n) \\
&= o_p(1/n).
\end{aligned}$$

Therefore,

$$B_1 + B_2 + B_3 = o_p(n^{-1/2}).$$

(II) Now we show $\sum_{i=1}^n \boldsymbol{Y}_{ni} + o_p(1) \to_d N(\mathbf{0}, \boldsymbol{G})$ where

$$\boldsymbol{Y}_{ni} = \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big[\nabla_{\mathcal{A}_n} L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big].$$

It is enough to show that $\boldsymbol{Y}_{ni}$, $i = 1, \ldots, n$ satisfies the conditions for Lindeberg-Feller central limit theorem (van der Vaart, 1998). For any given $\epsilon > 0$, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\sum_{i=1}^n E\Big[\|\boldsymbol{Y}_{ni}\|^2 I\{\|\boldsymbol{Y}_{ni}\| > \epsilon\}\Big] &= nE\Big[\|\boldsymbol{Y}_{n1}\|^2 I\{\|\boldsymbol{Y}_{n1}\| > \epsilon\}\Big] \\
&\leq n\Big[E\|\boldsymbol{Y}_{n1}\|^4\Big]^{1/2}\Big[E(1\{\|\boldsymbol{Y}_{n1}\| > \epsilon\})\Big]^{1/2} \\
&= n\ B_4^{1/2}\ B_5^{1/2}.
\end{aligned}$$

$$\begin{aligned}
B_4 &= \frac{1}{n^2}E\|\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\nabla_{\mathcal{A}_n} L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^4 \\
&\leq \frac{1}{n^2}\|\boldsymbol{A}_n^\intercal\boldsymbol{A}_n\|^2\|\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^2 E[\nabla^\intercal_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})]^2 \\
&= \frac{1}{n^2}\lambda^2_{\max}(\boldsymbol{A}_n^\intercal\boldsymbol{A}_n)\lambda^2_{\max}(\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}))O(s_n^2) \\
&= O(q_n^2/n^2).
\end{aligned}$$

By Markov inequality,

$$\begin{aligned} B_5 &= P(\|\boldsymbol{Y}_{n1}\| > \epsilon) \\ &\leq \frac{E\|\boldsymbol{Y}_{n1}\|^2}{\epsilon^2} \\ &= O(q_n/n). \end{aligned}$$

Therefore,

$$\sum_{i=1}^{n} E\Big[\|\boldsymbol{Y}_{ni}\|^2 1\{\|\boldsymbol{Y}_{ni}\| > \epsilon\}\Big] = nO(q_n/n)O(\sqrt{q_n/n}) = o(1).$$

Moreover,

$$\begin{aligned} \sum_{i=1}^{n} \text{Cov}(\boldsymbol{Y}_{ni}) &= n\text{Cov}(\boldsymbol{Y}_{n1}) \\ &= \boldsymbol{A}_n \boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) E[\nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \nabla^{\intercal}_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})] \boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \boldsymbol{A}_n^{\intercal} \\ &= \boldsymbol{A}_n \boldsymbol{A}_n^{\intercal} \to G. \end{aligned}$$

Since $\boldsymbol{Y}_{ni}$, $i = 1, \ldots, n$ satisfies the conditions for Lindeberg-Feller central limit theorem, we conclude $\sum_{i=1}^{n} \boldsymbol{Y}_{ni} + o_p(1) \to_d N(\boldsymbol{0}, \boldsymbol{G})$. $\square$