Supplementary Material for
"Consistent Fixed-Effects Selection in Ultra-high dimensional Linear Mixed Models with Error-Covariate Endogeneity"

Abhik Ghosh[1] and Magne Thoresen[2]

[1] *Indian Statistical Institute, India*

[2] *Department of Biostatsitics, University of Oslo, Norway*

1. Proof of Theorem 1

Define $P'_{n,\lambda}(0+) = \lim_{t\to 0+} P'_{n,\lambda}(t)$. Then, by an application of the Karush-Kuhn-Tucker (KKT) condition on the local minimizers $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\eta}}$, we get

$$\frac{\partial L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})}{\partial \beta_l} + v_l = 0, \qquad \text{for all } l \leq p,$$

where $v_i = P_{n,\lambda}(|\widehat{\beta}_l|)\text{sgn}(\widehat{\beta}_l)$ if $\widehat{\beta}_l \neq 0$, and $v_i \in [-P'_{n,\lambda}(0+), P'_{n,\lambda}(0+)]$ if $\widehat{\beta}_l = 0$. Therefore, by using the monotonicity and the limit of $P'_{n,\lambda}(t)$ from Condition (C3), we get

$$\left|\frac{\partial L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})}{\partial \beta_l}\right| \leq P'_{n,\lambda}(0+) = o(1). \tag{1.1}$$

Next, by the first order Taylor series expansion of $\frac{\partial L_n(\boldsymbol{\beta},\boldsymbol{\eta})}{\partial \beta_l}$ at $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$ around $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$, we get a $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})$ on the line segment joining $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$ and $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ such that

$$\frac{\partial L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \beta_l} = \sum_{j=1}^{p} \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \beta_j}(\widehat{\beta}_j - \beta_{0j}) + \sum_{k=1}^{m} \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \eta_k}(\widehat{\eta}_k - \eta_{0k}).$$

Therefore, in the event $\widehat{\boldsymbol{\beta}}_N = 0$ having probability tending to one [by Condition (C2)], we get

$$\begin{aligned}
&\left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \beta_l} \right| \\
&= \left| \sum_{j \in S} \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \beta_j}(\widehat{\beta}_j - \beta_{0j}) + \sum_{k=1}^{m} \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \eta_k}(\widehat{\eta}_k - \eta_{0k}) \right| \\
&\leq \max_{l,j \leq p} \left| \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \beta_j} \right| \left|\left| \widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S} \right|\right|_1 + \max_{l,k \leq p} \left| \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \eta_k} \right| \left|\left| \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right|\right|_1 \\
&\leq \max_{l,j \leq p} \left| \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \beta_j} \right| \sqrt{s} \left|\left| \widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S} \right|\right|_2 + \max_{l,k \leq p} \left| \frac{\partial^2 L_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\eta}})}{\partial \beta_l \eta_k} \right| \sqrt{m} \left|\left| \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right|\right|_2,
\end{aligned}$$

where the last step follows by Cauchy-Swartz inequality; here m is the dimension of $\boldsymbol{\eta}$. Now, by Conditions (C1) and (C2), we get

$$\left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \beta_l} \right| = o_P(1).$$

Then the theorem follows using (1.1) $\square$

2. Proof of Theorem 2

First let us note that, for the likelihood loss $L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = -l_n(\boldsymbol{\beta}, \boldsymbol{\eta})$, we have

$$\frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \beta_k} = -\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1}\boldsymbol{X}^{(k)},$$

for any $k \leq p$, where $\boldsymbol{X}^{(k)}$ denotes the k-th column of the matrix $\boldsymbol{X}$. Therefore, by an application of Strong law of Large Numbers, we have the following result in terms of the transformed regression model given in Equation (3.2) of the main paper:

$$\left|\frac{\partial L_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{\partial \beta_k}\right| \to E(\epsilon^* X_k^*), \quad \text{almost surely}, \quad \text{as } n \to \infty, \tag{2.2}$$

where ϵ^* and X_k^* represent the random variables corresponding to the transformed error $\boldsymbol{\epsilon}^* = \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1/2}\boldsymbol{\epsilon}$ and the k-th transformed covariate (column) in $\boldsymbol{X}^* = \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2)^{-1/2}\boldsymbol{X}$. Now, if X_k is endogenous, then clearly ϵ^* and X_k^* will be correlated and hence the limit in (2.2) will be non-zero. Then the proof follows directly from the results of Theorem 1. □

3. Proof of Theorem 3

We will first show that our Assumptions (A), (I) and (M) together with (P) imply the following four results for the PFGMM loss function $L_n^P(\boldsymbol{\beta})$ given in Eq. (3.7) of the main paper.

(R1) $\left|\left|\nabla_S L_n^P(\boldsymbol{\beta}_{0S}, \mathbf{0})\right|\right| = O_P\left(\sqrt{\frac{s \log p}{n}}\right)$, where ∇_S denotes the gradient with respect to the (non-zero) elements of $\boldsymbol{\beta}$ in S. Note that $\sqrt{\frac{s \log p}{n}} = o(d_n)$ by our Assumptions.

(R2) For any $\epsilon > 0$, there exists a positive constant C_ϵ such that, for all sufficiently large n,

$$P\left(\lambda_{\min}\left[\nabla_S^2 L_n^P(\boldsymbol{\beta}_{0S}, \mathbf{0})\right] > C_\epsilon\right) > 1 - \epsilon.$$

(R3) For any $\epsilon > 0$, $\delta > 0$ and any non-negative sequence $\alpha_n = o(d_n)$, there exists a positive integer N such that, for all $n \geq N$,

$$P\left(\sup_{||\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}|| \leq \alpha_n} \left|\left|\nabla_S^2 L_n^P(\boldsymbol{\beta}_S, \mathbf{0}) - \nabla_S^2 L_n^P(\boldsymbol{\beta}_{0S}, \mathbf{0})\right|\right|_F \leq \delta\right) > 1 - \epsilon,$$

where $||\boldsymbol{A}||_F$ denotes the Frobenius norm of a matrix $\boldsymbol{A}$.

(R4) For any $\epsilon > 0$, there exists a positive constant C_ϵ such that, for all sufficiently large n,

$$P\left(\lambda_{\min}\left[\nabla_S^2 L_n^P(\boldsymbol{\beta}_{0S}, \mathbf{0})\right] > C_\epsilon\right) > 1 - \epsilon.$$

Then, Parts (a) and (b) of our Theorem 3 follow from Theorems B.1 and B.2 of Fan and Liao (2014). Note that the assumptions required on the penalty functions there are exactly the same as our Assumption (P); see Fan and Liao (2014) for details.

In the following, we will use the notations $\boldsymbol{\Pi}_S = \boldsymbol{\Pi}(\boldsymbol{\beta}_{0S})$ and

$$\widetilde{L_n^P}(\boldsymbol{\beta}_S) = \left[\frac{1}{n}\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y} - \boldsymbol{X}_S\boldsymbol{\beta}_S)\right]^T \boldsymbol{J}(\boldsymbol{\beta}_0)\left[\frac{1}{n}\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y} - \boldsymbol{X}_S\boldsymbol{\beta}_S)\right], \quad \boldsymbol{\beta}_S \in \mathbb{R}^s. \tag{3.3}$$

Note that $\widetilde{L_n^P}(\boldsymbol{\beta}_S) = L_n^P(\boldsymbol{\beta}_S, \mathbf{0})$. We will now prove results (R1)–(R4).

Proof of (R1):

By standard derivative calculations, we get $\nabla \widetilde{L_n^P}(\boldsymbol{\beta}_S) = 2\boldsymbol{A}_n(\boldsymbol{\beta}_S)\boldsymbol{J}(\boldsymbol{\beta}_0)\left[\frac{1}{n}\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y} - \boldsymbol{X}_S\boldsymbol{\beta}_S)\right]$, where $\boldsymbol{A}_n(\boldsymbol{\beta}_S) = -\frac{1}{n}\left(\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1}\boldsymbol{X}_S\right)$. Now, by Assumption (I4), we know that $||\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})|| = O_P(1)$. Also, by Assumption (I2), the elements in $\boldsymbol{J}(\boldsymbol{\beta}_0)$ are uniformly bounded in probability, and hence

$$\left|\left|\nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right|\right| \leq O_P(1)\left|\left|\frac{1}{n}\boldsymbol{\Pi}_S\widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S})\right|\right|. \tag{3.4}$$

Next, we study the difference of the random variables $\boldsymbol{Z}_1 = \left[\frac{1}{n}\boldsymbol{\Pi}_S\widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S})\right]$ and $\boldsymbol{Z}_2 = \left[\frac{1}{n}\boldsymbol{\Pi}_S\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S})\right]$. By Assumption (M1), we get

$$C_1\widetilde{\boldsymbol{V}}_z - \boldsymbol{V}(\boldsymbol{\theta},\sigma^2) = (C_1-1)\boldsymbol{I} + \boldsymbol{Z}^T(C_1\mathcal{M}-\sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}})\boldsymbol{Z} \geq 0.$$

That is $C_1\widetilde{\boldsymbol{V}}_z \geq \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)$. By the Woodbury formula, since $C_1\widetilde{\boldsymbol{V}}_z$ and $\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)$ are both positive definite, we get $\widetilde{\boldsymbol{V}}_z^{-1} \leq C_1\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}$. Therefore,

$$\begin{aligned}\boldsymbol{Z}_1 - \boldsymbol{Z}_2 &= \frac{1}{n}\boldsymbol{\Pi}_S\left[\widetilde{\boldsymbol{V}}_z^{-1} - \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}\right](\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S}) \\ &\leq \frac{(C_1-1)}{n}\boldsymbol{\Pi}_S\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S}).\end{aligned} \tag{3.5}$$

Further, by Assumption (M2), we have

$$C_1(\log n)\boldsymbol{V}(\boldsymbol{\theta},\sigma^2) - \widetilde{\boldsymbol{V}}_z = (C_1\log n - 1)\boldsymbol{I} + \boldsymbol{Z}^T(C_1\log n\sigma^{-2}\boldsymbol{\Psi}_{\boldsymbol{\theta}} - \mathcal{M})\boldsymbol{Z} \geq 0.$$

Then, $C_1(\log n)\boldsymbol{V}(\boldsymbol{\theta},\sigma^2) \geq \widetilde{\boldsymbol{V}}_z$, and as before we get $C_1(\log n)\widetilde{\boldsymbol{V}}_z^{-1} \geq \boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}$. Therefore,

$$\begin{aligned}\boldsymbol{Z}_2 - \boldsymbol{Z}_1 &= \frac{1}{n}\boldsymbol{\Pi}_S\left[\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1} - \widetilde{\boldsymbol{V}}_z^{-1}\right](\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S}) \\ &\leq \frac{C_1(C_1\log n-1)}{n}\boldsymbol{\Pi}_S\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S}).\end{aligned} \tag{3.6}$$

Combining (3.5) and (3.6), along with our basic IV assumption (Eq. (3.3) of the main paper), we have $|\boldsymbol{Z}_1 - \boldsymbol{Z}_2| = o_P(1)$. Therefore, from (3.4), we get

$$\left\|\nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right\| \leq O_P(1)\left\|\frac{1}{n}\boldsymbol{\Pi}_S\boldsymbol{V}(\boldsymbol{\theta},\sigma^2)^{-1}(\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{\beta}_{0S})\right\|$$
$$= O_P(1)\left\|\frac{1}{n}\sum_{i=1}^{n}(y_i^*-\boldsymbol{X}_{iS}^*\boldsymbol{\beta}_{0S})\Pi_{iS}^*\right\|. \tag{3.7}$$

But, $E\left[(Y^*-\boldsymbol{X}_i^*\boldsymbol{\beta}_{0S})\Pi_i^*\right]=0$ by the choice of IV π_i^*. So, using the Bonferroni inequality and the exponential-tail Bernstein inequality along with Assumption (I1) and the normality of $(Y^*-\boldsymbol{X}_i^*\boldsymbol{\beta}_{0S})$, we get a positive constant C such that, for any $t>0$,

$$P\left(\max_{l\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}(y_i^*-\boldsymbol{X}_{iS}^*\boldsymbol{\beta}_{0S})F_{li}^*\right|>t\right) < p\max_{l\leq p}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(y_i^*-\boldsymbol{X}_{iS}^*\boldsymbol{\beta}_{0S})F_{li}^*\right|>t\right)$$
$$\leq \leq p\exp(-Ct^2/n).$$

Thus,
$$P\left(\max_{l\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}(y_i^*-\boldsymbol{X}_{iS}^*\boldsymbol{\beta}_{0S})F_{li}^*\right|>t\right)=O_P\left(\sqrt{\frac{\log p}{n}}\right).$$

Similarly, we can show
$$P\left(\max_{l\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}(y_i^*-\boldsymbol{X}_{iS}^*\boldsymbol{\beta}_{0S})H_{li}^*\right|>t\right)=O_P\left(\sqrt{\frac{\log p}{n}}\right).$$

Combining with (3.7) we get $\left\|\nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right\|=O_P\left(\sqrt{\frac{s\log p}{n}}\right)$, proving (R1). $\square$

Proof of (R2):

Note that, by standard derivative calculations, we have $\nabla^2\widetilde{L_n^P}(\boldsymbol{\beta}_{0S})=2\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})^T$. Fix any $\epsilon>0$. By Assumption (I2), there exists a constant $C>0$ such that $P(\lambda_{\min}[\boldsymbol{J}(\boldsymbol{\beta}_0)]>C)>1-\epsilon$ for all sufficiently large n. Also, by Assumption (I4), there exists a constant $C_2>0$ such that $\lambda_{\min}[\boldsymbol{A}\boldsymbol{A}^T]>C_2$, where $\boldsymbol{A}$ is as defined in Assumption (I4). Now, let us consider the events

$$G_1 = \{\lambda_{\min}[\boldsymbol{J}(\boldsymbol{\beta}_0)] > C\}, \quad G_2 = \left\{ \left|\left| \boldsymbol{A}_n(\boldsymbol{\beta}_{0S})\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})^T - \boldsymbol{A}\boldsymbol{A}^T \right|\right| < \frac{C_2}{2} \right\}.$$

On the event $G_1 \cap G_2$, we have

$$\begin{aligned} \lambda_{\min}\left[\nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right] &\geq 2\lambda_{\min}[\boldsymbol{J}(\boldsymbol{\beta}_0)]\lambda_{\min}\left[\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})^T\right] \\ &\geq \geq 2C\left\{\lambda_{\min}[\boldsymbol{A}\boldsymbol{A}^T] - \frac{C_2}{2}\right\} > CC_2. \end{aligned} \tag{3.8}$$

But, we already have $P(G_1) > 1-\epsilon$. And, by the definition of matrix $\boldsymbol{A}$, we have $P(G_2^c) < \epsilon$ for all sufficiently large n. Hence $P(G_1 \cap G_2) \geq 1 - P(G_1^c) - P(G_2^c) > 1 - 2\epsilon$, which completes the proof of (R2). $\square$

Proof of (R3):

Fix any $\epsilon > 0$, $\delta > 0$ and any non-negative sequence $\alpha_n = o(d_n)$. For all $\boldsymbol{\beta}_S$ satisfying $||\boldsymbol{\beta_S} - \boldsymbol{\beta}_{0S}|| < d_n/2$, we have $\beta_{S,k} \neq 0$ for all $k \leq s$. Thus, $\boldsymbol{J}(\boldsymbol{\beta}_S) = \boldsymbol{J}(\boldsymbol{\beta}_{0S})$. Also

$$P\left(\sup_{||\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}|| \leq \alpha_n} ||\boldsymbol{A}_n(\boldsymbol{\beta}_S) - \boldsymbol{A}_n(\boldsymbol{\beta}_{0S})||_F \leq \delta\right) > 1-\epsilon.$$

Combining we get

$$P\left(\sup_{||\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}|| \leq \alpha_n} \left|\left|\nabla_S^2 \widetilde{L_n^P}(\boldsymbol{\beta}_S) - \nabla_S^2 \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right|\right|_F \leq \delta\right) > 1-\epsilon,$$

which completes the proof of (R3). $\square$

Proof of (R4):

The proof follows in the same line of argument as in Appendix C.1.2 of Fan and Liao (2014) and hence left out for brevity. $\square$

Proof of Parts (a)-(b) of Theorem 3:

Under the results (R1)–(R4) along with Assumption (P), we can apply Theorem B.2 of Fan and Liao (2014) for our PFGMM loss to conclude Part (a) of Theorem 3, and we also get that $P(\widehat{S} \subset S) \to 1$. Further, from Theorem B.1 of Fan and Liao (2014), we have $\left|\left|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\right|\right| = o_P(d_n)$. Then,

$$
\begin{aligned}
P(S \not\subseteq \widehat{S}) &= P(\text{There exists a } j \in S \text{ such that } \widehat{\beta}_j = 0) \\
&\leq P(\text{There exists a } j \in S \text{ such that } |\widehat{\beta}_j - \beta_{0j}| \geq |\beta_{0j}|) \\
&\leq P(\max_{j \in S} |\widehat{\beta}_j - \beta_{0j}| \geq d_n) \\
&\leq P(||\widehat{\beta}_j - \beta_{0j}|| \geq d_n) = o(1). \qquad (3.9)
\end{aligned}
$$

Therefore, $P(S \subset \widehat{S}) \to 1$, and hence $P(\widehat{S} = S) \to 1$. □

Proof of Part (c) of Theorem 3:

We start with the KKT condition for $\widehat{\boldsymbol{\beta}}_S$ which gives

$$-P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S) = \nabla \widetilde{L_n^P}(\widehat{\boldsymbol{\beta}}_S),$$

where sgn denote the sign function, $\circ$ denotes the element-wise product and

$$P_n'(|\widehat{\boldsymbol{\beta}}_S|) = (P_{n,\lambda}(|\widehat{\beta}_{S,1}|), \ldots, P_{n,\lambda}(|\widehat{\beta}_{S,s}|))^T.$$

By the Mean-Value Theorem, we can get $\boldsymbol{\beta}^*$ lying on the segment joining $\boldsymbol{\beta}_{0S}$ and $\widehat{\boldsymbol{\beta}}_S$ such that

$$\nabla \widetilde{L_n^P}(\widehat{\boldsymbol{\beta}}_S) = \nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}) + \nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}).$$

Therefore, denoting $\boldsymbol{D} = \left[\nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}^*) - \nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right] (\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S})$, we get

$$\nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}_{0S})(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) + \boldsymbol{D} = -P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S) - \nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}).$$

Now, take any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$. Then, since $\nabla^2 \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}) = \boldsymbol{\Sigma} + o_P(1)$ by definition, using the consistency of $\widehat{\boldsymbol{\beta}}_S$ we have from the above equation that

$$\sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) = -\sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{\Gamma}^{-1/2} \nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}) - \sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{\Gamma}^{-1/2} \left[P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S) + \boldsymbol{D} \right]. \quad (3.10)$$

To tackle the first term in (3.10), we recall that $\nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}) = 2\boldsymbol{A}_n(\boldsymbol{\beta}_{0S})\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{B}_n$, where the random component $\boldsymbol{B}_n = \left[\frac{1}{n}\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1}(\boldsymbol{y} - \boldsymbol{X}_S\boldsymbol{\beta}_S)\right]$ is normally distributed with

$$\mathrm{Var}(\sqrt{n}\boldsymbol{B}) = \frac{\sigma^2}{n}\boldsymbol{\Pi}_S \widetilde{\boldsymbol{V}}_z^{-1} \boldsymbol{V}(\boldsymbol{\theta}, \sigma^2) \widetilde{\boldsymbol{V}}_z^{-1} \boldsymbol{\Pi}_S \to \boldsymbol{\Upsilon}, \quad \text{as } \ n \to \infty.$$

So, by the central limit theorem, for any unit vector $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^{2s}$,

$$\sqrt{n}\widetilde{\boldsymbol{\alpha}}^t \boldsymbol{\Upsilon}^{-1/2} \boldsymbol{B}_n \xrightarrow{\mathcal{D}} N(0,1).$$

Further, by definition $||\boldsymbol{A}_n(\boldsymbol{\beta}_0) - \boldsymbol{A}|| = o_P(1)$. Hence, by Slutsky's theorem, we have

$$\sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{\Gamma}^{-1/2} \nabla \widetilde{L_n^P}(\boldsymbol{\beta}_{0S}) \xrightarrow{\mathcal{D}} N(0,1). \quad (3.11)$$

Next, for the second term in (3.10), we apply Lemma C.2 of Fan and Liao (2014) to get, under Assumption (P),

$$\left|\left| P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S) \right|\right| = O_P\left(\max_{||\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}|| \le d_n/4} \zeta(\boldsymbol{\beta}) \sqrt{\frac{s \log p}{n}} + \sqrt{s} P_{n,\lambda}'(d_n) \right).$$

Also, by Assumptions (I4)–(I5), we have $\lambda_{\min}(\boldsymbol{\Gamma}^{-1/2}) = O_P(1)$. Hence, applying Assumptions (A1)–(A2), we get

$$\lambda_{\min}(\sqrt{n}\boldsymbol{\Gamma}^{-1/2})\left|\left|P_n'(|\widehat{\boldsymbol{\beta}}_S|)\circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S)\right|\right|$$
$$\leq O_P(\sqrt{n})O_P\left(\max_{||\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}||\leq\frac{d_n}{4}}\zeta(\boldsymbol{\beta})\sqrt{\frac{s\log p}{n}}+\sqrt{s}P_{n,\lambda}'(d_n)\right)=o_P(1).$$

Further, by continuity of $\nabla^2\widetilde{L_n^P}(\boldsymbol{\beta}_S)$, one can easily show that

$$\left|\left|\nabla^2\widetilde{L_n^P}(\boldsymbol{\beta}^*)-\nabla^2\widetilde{L_n^P}(\boldsymbol{\beta}_{0S})\right|\right|=o_P\left(\frac{1}{\sqrt{s\log p}}\right).$$

Also, we have $||\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}||=O_P\left(\sqrt{\frac{s\log p}{n}}+\sqrt{s}P_{n,\lambda}'(d_n)\right)$. Then, combining the above equations with Assumption (A1), we have $||\boldsymbol{D}||=o_P(n^{-1/2})$. Hence, we get

$$\sqrt{n}\boldsymbol{\alpha}^t\boldsymbol{\Gamma}^{-1/2}\left[P_n'(|\widehat{\boldsymbol{\beta}}_S|)\circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S)+\boldsymbol{D}\right]=o_P(1). \tag{3.12}$$

Therefore, using (3.11) and (3.12) in (3.10) with the help of Slutsky's theorem, we get the desired asymptotic normality result completing the proof of the theorem. □

References

[1] Fan J. and Liao Y. (2014). Endogeneity in high dimensions. *Annals of Statistics*, 42(3), 872–917.