Submitted to the Annals of Statistics

SUPPLEMENT TO "ENVELOPE-BASED SPARSE PARTIAL LEAST SQUARES"

BY GUANGYU ZHU AND ZHIHUA SU[*]

University of Rhode Island and University of Florida

APPENDIX A: ESTIMATION ALGORITHM

The optimization of (7) or (8) can be performed by a blockwise coordinate descend algorithm, which updates one row of $\mathbf{A}$ at a time and cycles through the rows of $\mathbf{A}$ until convergence. Suppose we want to update $\mathbf{a}_1^T$, the first row of $\mathbf{A}$. For the sake of simplifying the notations, we further assume that $\mathbf{a}_1^T$ is the first row in $\mathbf{G}_{\mathbf{A}}$. This does not lose any generality since if it were not the first row in $\mathbf{G}_{\mathbf{A}}$, we can permute the rows and columns of $\mathbf{G}_{\mathbf{A}}$, $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{X}}^{-1}$ to make it so without changing the value of the objective function. Let $\mathbf{A}_{-1} \in \mathbb{R}^{(p-d-1)\times d}$ denote the submatrix of $\mathbf{A}$ with the first row removed. We partition $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{X}}^{-1}$ as

$$\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = \begin{pmatrix} S_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}, \qquad \mathbf{S}_{\mathbf{X}}^{-1} = \begin{pmatrix} T_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{pmatrix},$$

where $S_{11} \in \mathbb{R}$ and $T_{11} \in \mathbb{R}$. Define $\mathbf{S}_{2|1} = \mathbf{S}_{22} - \mathbf{S}_{21}S_{11}^{-1}\mathbf{S}_{12}$, and $\mathbf{T}_{2|1} = \mathbf{T}_{22} - \mathbf{T}_{21}T_{11}^{-1}\mathbf{T}_{12}$. Let $\mathbf{a}_1^{(k)}$, $\mathbf{A}_{-1}^{(k)}$ and $\mathbf{G}_{\mathbf{A}^{(k)}}$ be the value of $\mathbf{a}_1$, $\mathbf{A}_{-1}$ and $\mathbf{G}_{\mathbf{A}}$ after k iterations, and let $\mathbf{G}$ denote the submatrix of $\mathbf{G}_{\mathbf{A}}$ with the first row removed. We can update $\mathbf{a}_1^T$ by solving the following optimization problem

$$\widehat{\mathbf{a}}_1 = \arg\min_{\mathbf{a}_1} L(\mathbf{a}_1) + \lambda w_1 \|\mathbf{a}_1\|_2, \tag{1}$$

where

$$\begin{aligned} L(\mathbf{a}_1) \quad = \quad & -2\log\{1 + \mathbf{a}_1^T(\mathbf{I}_d + \mathbf{A}_{-1}^T\mathbf{A}_{-1})^{-1}\mathbf{a}_1^T\} \\ & + \log\{1 + S_{11}(\mathbf{a}_1 + S_{11}^{-1}\mathbf{G}^T\mathbf{S}_{21})^T(\mathbf{G}^T\mathbf{S}_{2|1}\mathbf{G})^{-1}(\mathbf{a}_1 + S_{11}^{-1}\mathbf{G}^T\mathbf{S}_{21})\} \\ & + \log\{1 + T_{11}(\mathbf{a}_1 + T_{11}^{-1}\mathbf{G}^T\mathbf{T}_{21})^T(\mathbf{G}^T\mathbf{T}_{2|1}\mathbf{G})^{-1}(\mathbf{a}_1 + T_{11}^{-1}\mathbf{G}^T\mathbf{T}_{21})\}. \end{aligned}$$

To solve (1), we adopt the majorization-minimization (MM) principle (Hunter and Lange, 2004; Zou and Li, 2008) and construct a majorization function

[*] Supported by NSF Grant DMS-1407460.

for $L(\mathbf{a}_1)$. Let $L'(\mathbf{a}_1^{(k)})$ be the gradient of $L(\mathbf{a}_1)$ evaluated at $\mathbf{a}_1^{(k)}$, $\mathbf{G}^{(k)}$ be the value of $\mathbf{G}$ after k iterations and the constant δ be an upper bound of the eigenvalues of the Hessian matrix $L''(\mathbf{a}_1^{(k)})$. For example, we can take

$$\delta = (1+\epsilon)[4\lambda_{\max}\{(\mathbf{I}_d+{\mathbf{A}_{-1}^{(k)}}^T\mathbf{A}_{-1}^{(k)})^{-1}\}+2\lambda_{\max}\{S_{11}({\mathbf{G}^{(k)}}^T\mathbf{S}_{2|1}\mathbf{G}^{(k)})^{-1}\}+2\lambda_{\max}\{T_{11}({\mathbf{G}^{(k)}}^T\mathbf{T}_{2|1}\mathbf{G}^{(k)})^{-1}\}], \tag{2}$$

where $\epsilon > 0$ and $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix. Then the majorization function $Q(\mathbf{a}_1^{(k)})$ can be defined as

$$Q(\mathbf{a}_1) = L(\mathbf{a}_1^{(k)}) + (\mathbf{a}_1 - \mathbf{a}_1^{(k)})L'(\mathbf{a}_1^{(k)}) + \frac{1}{2}\delta(\mathbf{a}_1 - \mathbf{a}_1^{(k)})^T(\mathbf{a}_1 - \mathbf{a}_1^{(k)}).$$

The majorization function $Q(\mathbf{a}_1^{(k)})$ satisfies $Q(\mathbf{a}_1^{(k)}) \geq L(\mathbf{a}_1^{(k)})$, with the equality holds if and only if $\mathbf{a}_1 = \mathbf{a}_1^{(k)}$. Now instead of updating $\mathbf{a}_1$ from (1), we update $\mathbf{a}_1$ by

$$\mathbf{a}_1^{(k+1)} = \arg\min_{\mathbf{a}_1} Q(\mathbf{a}_1) + \lambda w_1\|\mathbf{a}_1\|_2. \tag{3}$$

The optimization problem (3) has an analytical solution

$$\mathbf{a}_1^{(k+1)} = \frac{1}{\delta}\left\{\delta\mathbf{a}_1^{(k)} - L'(\mathbf{a}_1^{(k)})\right\}\left\{1 - \frac{\lambda w_1}{\|\delta\mathbf{a}_1^{(k)} - L'(\mathbf{a}_1^{(k)})\|_2}\right\}_+, \tag{4}$$

where $(\cdot)_+$ denotes the positive part of a number. The update of other rows in $\mathbf{A}$ is similar. The details of the algorithm to solve (7) or (8) is summarized in Algorithm 1. The starting value of $\mathbf{A}$ can be obtained by fitting the predictor envelope model (3). It takes $O(pd+d^3)$ flops to update δ, and each update of $\mathbf{a}_i$ takes $O(d^2)$ flops. The details regarding the flop counts are given below.

Algorithm 1 The algorithm for solving the adaptive group lasso problem (7) or (8).

```
1. Get a starting value of $\mathbf{A}$
2. Repeat until convergence of $\mathbf{A}$
   For $i = 1$ to $i = p - d$
     (a) Compute $\delta$ using (2)
     (b) Keep updating $\mathbf{a}_i$ by (4) until the convergence of $\mathbf{a}_i$
```

- It costs $O(d^3 + pd)$ to update

$$\begin{aligned}\delta =&(1+\epsilon)[4\lambda_{\max}\{(\mathbf{I}_d + {\mathbf{A}_{-1}^{(k)}}^T\mathbf{A}_{-1}^{(k)})^{-1}\} + 2\lambda_{\max}\{S_{11}({\mathbf{G}^{(k)}}^T\mathbf{S}_{2|1}\mathbf{G}^{(k)})^{-1}\}\\ &+ 2\lambda_{\max}\{T_{11}({\mathbf{G}^{(k)}}^T\mathbf{T}_{2|1}\mathbf{G}^{(k)})^{-1}\}].\end{aligned}$$

The counts are as follows.

In order to update δ, we first need to update $\mathbf{B}_1 = (\mathbf{I}_d + \mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)})^{-1}$, $\mathbf{B}_2 = S_{11}(\mathbf{G}^{(k)^T}\mathbf{S}_{2|1}\mathbf{G}^{(k)})^{-1}$ and $\mathbf{B}_3 = T_{11}(\mathbf{G}^{(k)^T}\mathbf{T}_{2|1}\mathbf{G}^{(k)})^{-1}$.

(i) It takes $O(d^3)$ flops to update $\mathbf{B}_1 = (\mathbf{I}_d + \mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)})^{-1}$. Because $\mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)} = \mathbf{A}^{(k)^T}\mathbf{A}^{(k)} - \mathbf{a}_1^{(k)}\mathbf{a}_1^{(k)^T}$, once $\mathbf{A}^{(k)^T}\mathbf{A}^{(k)}$ is calculated in the last cycle, it takes $2d^2$ flops to get $\mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)}$ (d^2 for multiplication in $\mathbf{a}_1^{(k)}\mathbf{a}_1^{(k)^T}$ and d^2 in the subtraction $\mathbf{A}^{(k)^T}\mathbf{A}^{(k)} - \mathbf{a}_1^{(k)}\mathbf{a}_1^{(k)^T}$). We used a Cholesky decomposition for the matrix inversion, and it takes approximately d^3 flops to get $(\mathbf{I}_d + \mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)})^{-1}$ from $(\mathbf{I}_d + \mathbf{A}_{-1}^{(k)^T}\mathbf{A}_{-1}^{(k)})$.

(ii) It takes $O(d^3 + pd)$ to update $\mathbf{B}_2 = S_{11}(\mathbf{G}^{(k)^T}\mathbf{S}_{2|1}\mathbf{G}^{(k)})^{-1}$. Notice that

$$\begin{aligned}&\mathbf{G}^{(k)^T}\mathbf{S}_{2|1}\mathbf{G}^{(k)} = \mathbf{G}^{(k)^T}(\mathbf{S}_{22} - \mathbf{S}_{21}S_{11}^{-1}\mathbf{S}_{12})\mathbf{G}^{(k)}\\ =&\mathbf{G}_{\mathbf{A}^{(k)}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}^{(k)}} - (S_{11}^{-1/2}\mathbf{G}^{(k)^T}\mathbf{S}_{21} + S_{11}^{1/2}\mathbf{a}_1^{(k)})(S_{11}^{-1/2}\mathbf{G}^{(k)^T}\mathbf{S}_{21} + S_{11}^{1/2}\mathbf{a}_1^{(k)})^T.\end{aligned}$$

After $\mathbf{G}_{\mathbf{A}^{(k)}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}^{(k)}}$ is initially calculated (before the iterations), it takes $2dp + d$ flops to compute $S_{11}^{-1/2}\mathbf{G}^{(k)^T}\mathbf{S}_{21} + S_{11}^{1/2}\mathbf{a}_1^{(k)}$, d^2 flops to compute $(S_{11}^{-1/2}\mathbf{G}^{(k)^T}\mathbf{S}_{21} + S_{11}^{1/2}\mathbf{a}_1^{(k)})(S_{11}^{-1/2}\mathbf{G}^{(k)^T}\mathbf{S}_{21} + S_{11}^{1/2}\mathbf{a}_1^{(k)})^T$ and d^2 flops for the subtraction to get $\mathbf{G}^{(k)^T}\mathbf{S}_{2|1}\mathbf{G}^{(k)}$. The matrix inverse takes $O(d^3)$ flops. So the cost of updating $S_{11}(\mathbf{G}^{(k)^T}\mathbf{S}_{2|1}\mathbf{G}^{(k)})^{-1}$ is $O(d^3 + dp)$.

(iii) The count for $\mathbf{B}_3 = T_{11}(\mathbf{G}^{(k)^T}\mathbf{T}_{2|1}\mathbf{G}^{(k)})^{-1}$ is similar to that for $\mathbf{B}_2$.

After we have obtained $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$, it takes $O(d^3)$ flops to compute δ. This is because the cost of getting the maximum eigenvalue of $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$ is $O(d^3)$ flops each. Therefore the cost of getting $\delta = (1+\epsilon)[4\lambda_{\max}(\mathbf{B}_1 + 2\lambda_{\max}(\mathbf{B}_2) + 2\lambda_{\max}(\mathbf{B}_3)]$ is $O(d^3)$ flops.

So the total number of flops to update δ is in the order of $O(d^3 + pd)$.

- It takes $O(d^2)$ to update $\mathbf{a}_1^{(k+1)}$. The counts are as follows.

  To update $\mathbf{a}_1^{(k+1)}$, we first need to compute the value of $L'(\mathbf{a}_1^{(k)})$. The expression of $L'(\mathbf{a}_1^{(k)})$ is

$$\begin{aligned}L'(\mathbf{a}_1^{(k)}) \quad = \quad & \frac{-4\mathbf{B}_1\mathbf{a}_1^{(k)}}{1 + \mathbf{a}_1^{(k)^T}\mathbf{B}_1\mathbf{a}_1^{(k)}} + \frac{2\mathbf{B}_2(\mathbf{a}_1^{(k)} + \mathbf{v}_2)}{1 + (\mathbf{a}_1^{(k)} + \mathbf{v}_2)^T\mathbf{B}_2(\mathbf{a}_1^{(k)} + \mathbf{v}_2)}\\ & + \frac{2\mathbf{B}_3(\mathbf{a}_1^{(k)} + \mathbf{v}_3)}{1 + (\mathbf{a}_1^{(k)} + \mathbf{v}_3)^T\mathbf{B}_3(\mathbf{a}_1^{(k)} + \mathbf{v}_3)}\end{aligned}$$

where $\mathbf{v}_2 = S_{11}^{-1}{\mathbf{G}^{(k)}}^T\mathbf{S}_{21}$, $\mathbf{v}_3 = T_{11}^{-1}{\mathbf{G}^{(k)}}^T\mathbf{T}_{21}$. Notice that

$$\mathbf{v}_2 = S_{11}^{-1}\left({\mathbf{G}_{\mathbf{A}^{(k)}}}^T(S_{11},\mathbf{S}_{12})^T - \mathbf{a}_1^{(k)}S_{11}\right).$$

Once $\mathbf{G}_{\mathbf{A}^{(k)}}^T(S_{11},\mathbf{S}_{12})^T$ is initially calculated (before the iterations), we need only d flops to calculate $\mathbf{a}_1^{(k)}S_{11}$ and d flops to subtract it from $\mathbf{G}_{\mathbf{A}^{(k)}}^T(S_{11},\mathbf{S}_{12})^T$. We need another d flops to multiply S_{11}^{-1} to ${\mathbf{G}_{\mathbf{A}^{(k)}}}^T(S_{11},\mathbf{S}_{12})^T - \mathbf{a}_1^{(k)}S_{11}$. So it takes a total of $3d$ flops to update $\mathbf{v}_2$. The same count holds for updating $\mathbf{v}_3$.

Once we have $\mathbf{v}_2$ and $\mathbf{v}_3$, it takes $O(d^2)$ to calculate each of the three summands in $L'(\mathbf{a}_1^{(k)})$: $-4\mathbf{B}_1\mathbf{a}_1^{(k)}/[1+{\mathbf{a}_1^{(k)}}^T\mathbf{B}_1\mathbf{a}_1^{(k)}]$,
$2\mathbf{B}_2(\mathbf{a}_1^{(k)}+\mathbf{v}_2)/[1+(\mathbf{a}_1^{(k)}+\mathbf{v}_2)^T\mathbf{B}_2(\mathbf{a}_1^{(k)}+\mathbf{v}_2)]$
and $2\mathbf{B}_3(\mathbf{a}_1^{(k)}+\mathbf{v}_3)/[1+(\mathbf{a}_1^{(k)}+\mathbf{v}_3)^T\mathbf{B}_3(\mathbf{a}_1^{(k)}+\mathbf{v}_3)]$. So it takes $O(d^2)$ flops to obtain the value of $L'(\mathbf{a}_1^{(k)})$.

After we have the value of δ and $L'(\mathbf{a}_1^{(k)})$, it takes $O(d)$ flops to update $\mathbf{a}_1^{(k+1)}$. This can be observed from the formula

$$\mathbf{a}_1^{(k+1)} = \frac{1}{\delta}\left\{\delta\mathbf{a}_1^{(k)} - L'(\mathbf{a}_1^{(k)})\right\}\left\{1-\frac{\lambda w_1}{\|\delta\mathbf{a}_1^{(k)} - L'(\mathbf{a}_1^{(k)})\|_2}\right\}_+ .$$

So the total number of flops to update $\mathbf{a}_1^{(k+1)}$ has order $O(d^2)$.

The adaptive weights can be chosen as $w_i = \|\mathbf{a}_i^{(0)}\|_2^{-\gamma}$, where $\mathbf{a}_i^{(0)}$ is the initial value and also a $\sqrt{n}$-consistent estimator of $\mathbf{a}_i$. Parameter γ is a positive number and can be chosen by cross validation. As suggested by Chen and Huang (2012) and Zou (2006), it is sufficient to choose γ from a small candidate set such as $\{0.5, 1, 2, 4, 8\}$. For a fixed d, the tuning parameter λ can be chosen by cross validation or Bayesian information criterion (Zou and Chen, 2012, BIC). Let l_λ be the log likelihood and $p_{\mathcal{A},\lambda}$ be the number of selected active predictors. Then $\mathrm{BIC}(\lambda) = -2l_\lambda + (p_{\mathcal{A},\lambda} - d)d\log(n)$. And we select the λ that minimizes $\mathrm{BIC}(\lambda)$. Zou and Chen (2012) established the consistency of BIC. The selection of d can be performed using likelihood ratio testing, BIC, or cross validation.

The optimization of (8) is similar as (7), except we choose γ, d and λ by a three dimensional cross validation. This is because the likelihood based methods need moderate sample size to achieve good performance. The starting values of $\mathbf{a}$ can be calculated from a starting value of $\mathbf{\Gamma}$, which is discussed in Cook, Forzani and Su (2016).

APPENDIX B: ASYMPTOTIC VARIANCE UNDER NORMALITY

If we assume normality, we can get a closed form of the asymptotic variance for the oracle predictor envelope estimator, the E-SPLS estimator and the E-SGPLS estimator. The details are included as follows.

Let $\mathbf{A}_{\mathcal{A}}$ denote the first $p_{\mathcal{A}} - d$ rows of $\mathbf{A}$, $\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}$ be the sample covariance matrix of $\mathbf{X}_{\mathcal{A}}$, $\mathbf{S}_{\mathbf{X}_{\mathcal{A}}|\mathbf{Y}}$ be the sample covariance matrix of the residuals from the linear regression of $\mathbf{X}_{\mathcal{A}}$ on $\mathbf{Y}$, and $(\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}}$ be the upper left $p_{\mathcal{A}} \times p_{\mathcal{A}}$ block in $\mathbf{S}_{\mathbf{X}}^{-1}$. Define $\mathbf{G}_{\mathbf{A}_{\mathcal{A}}} = (\mathbf{I}_d, \mathbf{A}_{\mathcal{A}}^T)^T$. The symbol $\overset{d}{\to}$ denotes convergence in distribution.

Proposition 1 *Assume that the oracle predictor envelope model (9) holds, $\mathbf{X}$ has finite fourth moments and the errors are normally distributed. Then the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{A}}$ is $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O} = \mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}_{\mathcal{A},O}}(\mathbf{S}_{\mathbf{X}_{\mathcal{A}}})}\widehat{\boldsymbol{\beta}}_{\mathcal{A},\text{ols}}$, where*

$$\widehat{\mathbf{A}}_{\mathcal{A},O} = \arg \min_{\mathbf{A}_{\mathcal{A}} \in \mathbb{R}^{(p_{\mathcal{A}}-d)\times d}} -2\log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T \mathbf{G}_{\mathbf{A}_{\mathcal{A}}}| + \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T \mathbf{S}_{\mathbf{X}_{\mathcal{A}}|\mathbf{Y}} \mathbf{G}_{\mathbf{A}_{\mathcal{A}}}| + \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T (\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}} \mathbf{G}_{\mathbf{A}_{\mathcal{A}}}|.$$

Furthermore,

$$\sqrt{n}\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}) - \text{vec}(\boldsymbol{\beta}_{\mathcal{A}})\} \overset{d}{\to} N(0, \mathbf{V}_O),$$

where $\mathbf{V}_O = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}_{\mathcal{A}}} \otimes \boldsymbol{\Gamma}_{\mathcal{A}} \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}_{\mathcal{A}}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)$, and $\mathbf{T} = (\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}_{\mathcal{A}}}^{-1}\boldsymbol{\eta}^T + \boldsymbol{\Omega}^{-1}) \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}$.

The proof of this proposition is in Section C.3 of the Supplement.

Theorem 3 *Assume that the conditions in Theorem 2 hold. If we further assume normality in $\mathbf{X}$ and $\boldsymbol{\varepsilon}$, then the asymptotic variance of the E-SPLS estimator $\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})$ equals $\mathbf{V}_O$ in Proposition 1.*

Theorem 6 *(c) Assume that the same conditions in (b) hold, we have a closed form for the asymptotic variance of the E-SGPLS estimator $\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})$: $\mathbf{V} = \mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}} \mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}} \mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)$, where $\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}$ is the asymptotic variance of the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$ with $d = p_{\mathcal{A}}$, and $\mathbf{T} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}) + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \boldsymbol{\Omega}^{-1} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}$.*

The explicit form of $\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}$ is given in Lemma 3.

APPENDIX C: PROOFS

C.1. Proof of Theorem 1. We denote the objective function in (7) as $f_{\text{obj}}(\mathbf{A})$. To prove Theorem 1, we will show that for any small $\epsilon > 0$, there exists a sufficiently large constant C, such that

$$(5)\qquad \lim_{n\to\infty} P\left(\inf_{\mathbf{\Delta}\in\mathbb{R}^{(p-d)\times d},\|\mathbf{\Delta}\|_F=C} f_{\text{obj}}(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) > f_{\text{obj}}(\mathbf{A})\right) > 1-\epsilon.$$

If (5) holds, there exists a local minimizer $\widehat{\mathbf{A}}$ of f_{obj} such that $\|\widehat{\mathbf{A}}-\mathbf{A}\|_F = O_p(n^{-1/2})$. This establishes that $\widehat{\mathbf{A}}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{A}$. Because $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S_X})} = \widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S_X}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1}\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S_X}$, and $\mathbf{S_X}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{\Sigma_X}$, $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S_X})}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{P}_{\mathbf{\Gamma}(\mathbf{\Sigma_X})}$. Then the E-SPLS estimator $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S_X})}\widehat{\boldsymbol{\beta}}_{\text{ols}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$.

Now we prove (5). We calculate $f_{\text{obj}}(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) - f_{\text{obj}}(\mathbf{A})$ using Taylor expansion. Since the form of f_{obj} is a little complicated, we write it into four parts

$$\begin{aligned} f_{\text{obj}}(\mathbf{A}) &= -2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}| + \sum_{i=1}^{p-d}\lambda w_i\|\mathbf{a}_i\|_2 \\ &\equiv f_1(\mathbf{A}) + f_2(\mathbf{A}) + f_3(\mathbf{A}) + f_4(\mathbf{A}). \end{aligned}$$

We first expand $f_1(\mathbf{A}+n^{-1/2}\mathbf{\Delta})$,

$$f_1(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) = f_1(\mathbf{A}) + n^{-1/2}\overset{\rightarrow\mathbf{\Delta}}{df_1}(\mathbf{A}) + \frac{1}{2}n^{-1}\overset{\rightarrow\mathbf{\Delta}}{df_1^2}(\mathbf{A}) + o_p(n^{-1}),$$

where $\overset{\rightarrow\mathbf{\Delta}}{df_1}(\mathbf{A})$ and $\overset{\rightarrow\mathbf{\Delta}}{df_1^2}(\mathbf{A})$ are the first and second directional derivatives (Dattorro, 2016). The first directional derivative is

$$\overset{\rightarrow\mathbf{\Delta}}{df_1}(\mathbf{A}) = \operatorname{tr}\left\{\left[\frac{df_1(\mathbf{A})}{d\mathbf{A}}\right]^T\mathbf{\Delta}\right\} = -4\operatorname{tr}[(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}].$$

The second directional derivative is

$$\begin{aligned} \overset{\rightarrow\mathbf{\Delta}}{df_1^2}(\mathbf{A}) =& \operatorname{tr}\left(\left[\frac{\overset{\rightarrow\mathbf{\Delta}}{df_1}(\mathbf{A})}{d\mathbf{A}}\right]^T\mathbf{\Delta}\right) \\ =& -4\operatorname{tr}\left\{[-\mathbf{A}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{\Delta}+\mathbf{\Delta}^T\mathbf{A})(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1} + \mathbf{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}]^T\mathbf{\Delta}\right\} \\ =& 4\operatorname{tr}\left\{(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{\Delta}+\mathbf{\Delta}^T\mathbf{A})(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta} - (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{\Delta}^T\mathbf{\Delta}\right\} \\ =& 4\operatorname{tr}\left\{(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}\right. \\ & \left. -(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{\Delta}_*^T(\mathbf{I}_p - \mathbf{G}_{\mathbf{A}}(\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}})\mathbf{\Delta}_*\right\} \\ =& 4\operatorname{tr}\left\{(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta} - (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\right\} \end{aligned}$$

where

$$\boldsymbol{\Delta}_* = \begin{pmatrix} 0_{d\times d} \\ \boldsymbol{\Delta} \end{pmatrix} \in \mathbb{R}^{p\times d}.$$

Substitute $\overset{\rightarrow\boldsymbol{\Delta}}{df_1}(\mathbf{A})$ and $\overset{\rightarrow\boldsymbol{\Delta}}{df_1^2}(\mathbf{A})$ into the expansion for f_1. We obtain

$$\begin{aligned} & f_1(\mathbf{A}+n^{-1/2}\boldsymbol{\Delta}) - f_1(\mathbf{A}) \\ = & -4n^{-1/2}\operatorname{tr}[(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}] + 2n^{-1}\operatorname{tr}\Big\{(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta} \\ & - (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\Big\} + o_p(n^{-1}). \end{aligned}$$

Now we expand $f_2(\mathbf{A}) = \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}|$. The first directional derivative is

$$\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A}) = \operatorname{tr}\Big\{\Big[\frac{df_2(\mathbf{A})}{d\mathbf{A}}\Big]^T\boldsymbol{\Delta}\Big\} = 2\operatorname{tr}[(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*].$$

Since $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}$, $(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ is a $\sqrt{n}$-consistent estimator of $(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}$. Then we have

$$(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} + n^{-1/2}\mathbf{T}_n + O_p(n^{-1}),$$

where $\operatorname{vec}(\mathbf{T}_n)$ converges in distribution to a normal random vector with mean 0. Substitute the expression to $\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A})$, we have

$$\begin{aligned} \overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A}) =& 2\operatorname{tr}[(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*] \\ =& 2\operatorname{tr}[(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*] + 2n^{-1/2}\operatorname{tr}(\mathbf{T}_n\boldsymbol{\Delta}_*) + O_p(n^{-1}). \end{aligned}$$

Because that $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} = \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\eta}$, we have

$$\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} = \boldsymbol{\Gamma}(\boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\boldsymbol{\eta}^T\boldsymbol{\Omega})\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.$$

Then by Woodbury matrix identity, we have

$$(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Gamma})^{-1} = (\boldsymbol{\Omega}-\boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\boldsymbol{\eta}^T\boldsymbol{\Omega})^{-1} = \boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}(\boldsymbol{\Sigma}_{\mathbf{Y}}-\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T = \boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T.$$

Since $\boldsymbol{\Gamma} = \mathbf{G}_{\mathbf{A}}\boldsymbol{\Gamma}_1$, we have $(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1} = \boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}_1^T$, and

$$\begin{aligned} (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} &= \boldsymbol{\Gamma}_1(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}_1^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}_1^{-1})^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \\ &= \boldsymbol{\Gamma}_1(\boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)^{-1}(\boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^T\mathbf{G}_{\mathbf{A}}^T = (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{G}_{\mathbf{A}}^T. \end{aligned}$$

The last equality is because $\mathbf{I}_d = \boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1^T\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}\boldsymbol{\Gamma}_1$. This gives $\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}} = (\boldsymbol{\Gamma}_1^T)^{-1}(\boldsymbol{\Gamma}_1)^{-1}$. Then

$$\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^T = (\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}})^{-1} = (\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}. \tag{6}$$

So the first directional derivative is

$$\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A}) = 2\operatorname{tr}[(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}] + 2n^{-1/2}\operatorname{tr}(\mathbf{T}_n\boldsymbol{\Delta}_*) + O_p(n^{-1}).$$

By Cauchy-Schwartz inequality (Harville, 1998),

$$\Big|\operatorname{tr}(\mathbf{T}_n\boldsymbol{\Delta}_*)\Big| \le \|\boldsymbol{\Delta}\|_F\|\mathbf{T}_n\|_F.$$

The second directional derivative of f_2 is

$$\begin{aligned}\overset{\rightarrow\boldsymbol{\Delta}}{df_2^2}(\mathbf{A}) =& \operatorname{tr}\Big(\Big[\frac{\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A})}{d\mathbf{A}}\Big]^T\boldsymbol{\Delta}\Big)\\ =&2\operatorname{tr}\Big\{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\boldsymbol{\Delta}_*^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*\\ &- (\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_* + \boldsymbol{\Delta}_*^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*\Big\}.\end{aligned}$$

Since $\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} + o_p(1)$, after some straightforward algebra, we have

$$\begin{aligned}\overset{\rightarrow\boldsymbol{\Delta}}{df_2^2}(\mathbf{A}) =&2\operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\\ &- (\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\Big\} + o_p(1).\end{aligned}$$

Substitute $\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\mathbf{A})$ and $\overset{\rightarrow\boldsymbol{\Delta}}{df_2^2}(\mathbf{A})$ into the expansion for f_2, we get

$$\begin{aligned}&f_2(\mathbf{A} + n^{-1/2}\boldsymbol{\Delta}) - f_2(\mathbf{A})\\ =&2n^{-1/2}\operatorname{tr}[(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}] + 2n^{-1}\operatorname{tr}(\mathbf{T}_n\boldsymbol{\Delta}_*)\\ &+ n^{-1}\operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\\ &- (\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\Big\} + o_p(n^{-1})\\ \ge&2n^{-1/2}\operatorname{tr}[(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}] - 2n^{-1}\|\boldsymbol{\Delta}\|_F\|\mathbf{T}_n\|_F\\ &+ n^{-1}\operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\\ &- (\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\Big\} + o_p(n^{-1}).\end{aligned}$$

Notice that f_3 and f_2 have the same structure, except that $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ is replaced by $\mathbf{S}_{\mathbf{X}}^{-1}$. Because $\mathbf{S}_{\mathbf{X}}^{-1}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$, $(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}$ is a $\sqrt{n}$-consistent estimator of $(\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}$. Then we have

$$(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1} = (\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1} + n^{-1/2}\mathbf{W}_n + O_p(n^{-1}),$$

where $\mathrm{vec}(\mathbf{W}_n)$ converges in distribution to a normal random vector with mean 0. We then do the same expansion to f_3 and get

$$\begin{aligned}
&f_3(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) - f_3(\mathbf{A})\\
=&2n^{-1/2}\,\mathrm{tr}[(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}] + 2n^{-1}\,\mathrm{tr}(\mathbf{W}_n\mathbf{\Delta}_*)\\
&+n^{-1}\,\mathrm{tr}\Big\{\mathbf{\Omega}\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1 - (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}\Big\} + o_p(n^{-1})\\
\geq&2n^{-1/2}\,\mathrm{tr}[(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}] - 2n^{-1}\|\mathbf{\Delta}\|_F\|\mathbf{W}_n\|_F\\
&+n^{-1}\,\mathrm{tr}\Big\{\mathbf{\Omega}\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1 - (\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Delta}\Big\} + o_p(n^{-1}).
\end{aligned}$$

Now we expand $f_4(\mathbf{A})$. Let $\boldsymbol{\delta}_i^T$ be the ith row of $\mathbf{\Delta}$, then

$$\begin{aligned}
f_4(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) - f_4(\mathbf{A}) &= \sum_{i=1}^{p-d}\Big\{\lambda w_i\|\mathbf{a}_i+n^{-1/2}\boldsymbol{\delta}_i\|_2 - \lambda w_i\|\mathbf{a}_i\|_2\Big\}\\
&\geq \sum_{i=1}^{p_{\mathcal{A}}-d}\Big\{\lambda w_i\|\mathbf{a}_i+n^{-1/2}\boldsymbol{\delta}_i\|_2 - \lambda w_i\|\mathbf{a}_i\|_2\Big\}\\
&\geq -(p_{\mathcal{A}}-d)n^{-1/2}\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\|\boldsymbol{\delta}_i\|_2.
\end{aligned}$$

The third inequality is based on the triangular inequality that $-\|-n^{-1/2}\boldsymbol{\delta}_i\|_2 \leq \|\mathbf{a}_i+n^{-1/2}\boldsymbol{\delta}_i\|_2 - \|\mathbf{a}_i\|_2$. As $\sqrt{n}\lambda_{\mathcal{A}} \to 0$, $f_4(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) - f_4(\mathbf{A}) = o_p(n^{-1})$.

Combine the results for f_1, f_2, f_3 and f_4, we have

$$\begin{aligned}
&f_{\mathrm{obj}}(\mathbf{A}+n^{-1/2}\mathbf{\Delta}) - f_{\mathrm{obj}}(\mathbf{A})\\
\geq\;&-2n^{-1}\|\mathbf{\Delta}\|_F\|\mathbf{T}_n\|_F - 2n^{-1}\|\mathbf{\Delta}\|_F\|\mathbf{W}_n\|_F\\
&+n^{-1}\,\mathrm{tr}\Big\{(\mathbf{\Omega}^{-1}+\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1 + \mathbf{\Omega}\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1\\
&-2(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\mathbf{\Delta}_*^T\Big\} - n^{-1}(p_{\mathcal{A}}-d)\sqrt{n}\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\|\boldsymbol{\delta}_i\|_2 + o_p(n^{-1}).
\end{aligned}$$

Let $\mathbf{M} = (\mathbf{\Omega}^{-1}+\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\otimes\mathbf{\Omega}_0 + \mathbf{\Omega}\otimes\mathbf{\Omega}_0^{-1} - 2\mathbf{I}_d\otimes\mathbf{I}_{p-d}$, and let m_1 be the smallest eigenvalue of $\mathbf{M}$. The matrix $\mathbf{M}$ appears in proposition 4.4 in

Cook, Helland and Su (2013). By Shapiro (1986), $\mathbf{M}$ is a positive definite matrix, and $m_1 > 0$. Then we have

$$\begin{aligned}
& \mathrm{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 + \boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*^T\Big\} \\
= \;& \mathrm{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 + \boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 - 2\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\Big\} \\
= \;& \mathrm{vec}(\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_1)^T\mathbf{M}\mathrm{vec}(\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1) \\
\geq \;& m_1\|\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\|_F^2 \\
= \;& m_1\,\mathrm{tr}(\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T) \\
= \;& m_1\,\mathrm{tr}\Big\{\big[\mathbf{I}_p - \mathbf{G}_{\mathbf{A}}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{G}_{\mathbf{A}}^T\big]\boldsymbol{\Delta}_*(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\Big\} \\
= \;& m_1\,\mathrm{tr}\Big\{\boldsymbol{\Delta}^T\big[\mathbf{I}_{p-d} - \mathbf{A}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\big]\boldsymbol{\Delta}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\Big\} \\
= \;& m_1\,\mathrm{tr}\Big\{\boldsymbol{\Delta}^T(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\Big\} \\
= \;& m_1\mathrm{vec}(\boldsymbol{\Delta})^T[(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1} \otimes (\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}]\mathrm{vec}(\boldsymbol{\Delta}) \\
\geq \;& m_1 m_2^2\|\boldsymbol{\Delta}\|_F^2,
\end{aligned}$$

where m_2 is the smallest eigenvalue of $(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}$. When $\|\boldsymbol{\Delta}\|_F > C$ for sufficiently large C, the terms with order $\|\boldsymbol{\Delta}\|_F^2$ dominate the terms with order $\|\boldsymbol{\Delta}\|_F$. Then $f_{\mathrm{obj}}(\mathbf{A} + n^{-1/2}\boldsymbol{\Delta}) - f_{\mathrm{obj}}(\mathbf{A}) > 0$ with probability tending to 1, and conclusion (5) follows.

C.2. Proof of Theorem 2. We prove Theorem 2 by contradiction. Suppose that $\widehat{\mathbf{a}}_i \neq 0$ for $i = p_{\mathcal{A}} - d + 1, \ldots, p - d$. Let $\mathbf{e}_i \in \mathbb{R}^d$ denote a vector of 0 with a 1 in the ith place. Then

$$\frac{df_{\mathrm{obj}}(\mathbf{A})}{d\mathbf{a}_i^T}\Big|_{\mathbf{a}_i = \widehat{\mathbf{a}}_i} = 0, \tag{7}$$

where

$$\begin{aligned}
\frac{df_{\mathrm{obj}}(\mathbf{A})}{d\mathbf{a}_i^T}\Big|_{\mathbf{a}_i = \widehat{\mathbf{a}}_i} =& - 4\mathbf{e}_i^T\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{I}_d + \widehat{\mathbf{A}}^T\widehat{\mathbf{A}})^{-1} + 2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1} \\
&+ 2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1} + \lambda w_i\frac{\widehat{\mathbf{a}}_i^T}{\|\widehat{\mathbf{a}}_i\|_2}.
\end{aligned}$$

Since $\widehat{\mathbf{A}}$, $\mathbf{S}_{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ are $\sqrt{n}$-consistent estimators of $\mathbf{A}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}$, then

$$\begin{aligned}
& -4\mathbf{e}_i^T\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{I}_d+\widehat{\mathbf{A}}^T\widehat{\mathbf{A}})^{-1}+2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1}+2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1}\\
= & -4\mathbf{e}_i^T\mathbf{G}_{\mathbf{A}}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}+2\mathbf{e}_i^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}+2\mathbf{e}_i^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\\
& +O_p(n^{-1/2})\\
= & -4\mathbf{e}_i^T\mathbf{G}_{\mathbf{A}}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}+2\mathbf{e}_i^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Gamma}\boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1^{-T}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Gamma}\boldsymbol{\Gamma}_1^{-1})^{-1}\\
& +2\mathbf{e}_i^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1^{-T}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}_1)^{-1}+O_p(n^{-1/2})\\
= & -4\mathbf{e}_i^T\mathbf{G}_{\mathbf{A}}(\mathbf{I}_d+\mathbf{A}^T\mathbf{A})^{-1}+2\mathbf{e}_i^T\mathbf{G}_{\mathbf{A}}\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^T+2\mathbf{e}_i^T\mathbf{G}_{\mathbf{A}}\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^T+O_p(n^{-1/2})\\
= & \; O_p(n^{-1/2}).
\end{aligned}$$

The last equality is because of (6). Then

$$\sqrt{n}\left\|-4\mathbf{e}_i^T\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{I}_d+\widehat{\mathbf{A}}^T\widehat{\mathbf{A}})^{-1}+2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1}+2\mathbf{e}_i^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}}(\widehat{\mathbf{G}}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{G}}_{\mathbf{A}})^{-1}\right\|_2=O_p(1). \tag{8}$$

On the other hand, for $i=p_{\mathcal{A}}-d+1,\ldots,p-d$,

$$\sqrt{n}\lambda w_i\left\|\frac{\widehat{\mathbf{a}}_i^T}{\|\widehat{\mathbf{a}}_i\|_2}\right\|_2=\sqrt{n}\lambda w_i\geq\sqrt{n}\lambda_{\mathcal{I}}\to\infty. \tag{9}$$

With (7), (8) and (9) are contradictory to each other. So we have $P(\widehat{\mathbf{a}}_i=0)\to 1$ for $i=p_{\mathcal{A}}-d+1,\ldots,p-d$.

C.3. Proof of Proposition 1. Without the oracle information, the objective function for $\mathbf{A}$ is (5). By considering the oracle information and substituting the sparse structure of $\mathbf{A}$ to (5), we obtain the objective function of $\mathbf{A}_{\mathcal{A}}$ as displayed in Proposition 1.

Since the oracle predictor envelope model (9) is over-parameterized, we use Theorem 4.1 in Shapiro (1986) to derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$. We use $\boldsymbol{\xi}$ to denote the parameters in (9), and

$$\boldsymbol{\xi}=(\boldsymbol{\mu}_{\mathbf{X}}^T,\boldsymbol{\mu}_{\mathbf{Y}}^T,\,\mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})^T,\mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})^T,\,\mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})^T)^T,$$

where vech is the vector-half operator that stacks the lower triangle of a symmetric matrix to a vector. We use $\boldsymbol{\phi}$ to denote the parameters under the oracle predictor envelope model (9), and

$$\boldsymbol{\phi}=\{\boldsymbol{\mu}_{\mathbf{X}}^T,\boldsymbol{\mu}_{\mathbf{Y}}^T,\,\mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}),\mathrm{vec}^T(\boldsymbol{\eta}),\mathrm{vec}^T(\boldsymbol{\Gamma}_{\mathcal{A}}),\,\mathrm{vech}^T(\boldsymbol{\Omega}),\,\mathrm{vech}^T(\boldsymbol{\Omega}_0)\}^T.$$

Let $\mathbf{C}_m\in\mathbb{R}^{m(m+1)/2\times m^2}$ and $\mathbf{E}_m\in\mathbb{R}^{m(m+1)/2\times m^2}$ be contraction and expansion matrices that connect vec and vech: for a symmetric matrix $\mathbf{M}\in\mathbb{R}^{m\times m}$, $\mathrm{vec}(\mathbf{M})=\mathbf{E}_m\,\mathrm{vech}(\mathbf{M})$ and $\mathrm{vech}(\mathbf{M})=\mathbf{C}_m\mathrm{vec}(\mathbf{M})$.

The gradient matrix $\mathbf{H} = \partial\boldsymbol{\xi}/\partial\boldsymbol{\phi}^T$ is

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{r+p+r(r+1)/2} & 0 \\ 0 & \mathbf{H}_{22} \end{pmatrix},$$

where $\mathbf{H}_{22}$ is a $\{p_{\mathcal{A}}r + p(p+1)/2\} \times \{d(r - p_{\mathcal{I}}) + p(p+1)/2\}$ matrix

$$\mathbf{H}_{22} = \begin{pmatrix} \mathbf{I}_r \otimes \boldsymbol{\Gamma}_{\mathcal{A}} & \boldsymbol{\eta}^T \otimes \mathbf{I}_{p_{\mathcal{A}}} & 0 & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\mathbf{L}_1 & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_d & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-d} \end{pmatrix},$$

and $\mathbf{L}_1 = \mathbf{I}_d \otimes (\mathbf{I}_{p_{\mathcal{A}}}, 0)^T \in \mathbb{R}^{dp \times dp_{\mathcal{A}}}$. The Fisher information of $\boldsymbol{\xi}$ is given by

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1} & \mathbf{A}_{12} & 0 & 0 & 0 \\ \mathbf{A}_{21} & \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1})\mathbf{E}_r & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p \end{pmatrix},$$

where $\mathbf{A}_{12} = \mathbf{A}_{21}^T = -\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}$. Denote the estimator of $\boldsymbol{\xi}$ under the envelope parameterization by $\widehat{\boldsymbol{\xi}}$, and denote the estimator of $\boldsymbol{\phi}$ under the standard parameterization by $\widehat{\boldsymbol{\phi}}$. Then

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\phi}}) - \mathrm{vec}(\boldsymbol{\phi})\} \xrightarrow{d} N(0, \mathbf{J}^{-1}).$$

Let $l_{(\mathbf{X},\mathbf{Y})}$ denote the joint likelihood function of $\mathbf{X}$ and $\mathbf{Y}$,

$$\begin{aligned} l_{(\mathbf{X},\mathbf{Y})} = & -\frac{n(r+p)}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| \\ & - \frac{1}{2}\mathrm{tr}\big[(\mathbb{X} - 1_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbb{X} - 1_n\boldsymbol{\mu}_{\mathbf{X}}^T)^T\big] - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\ & - \frac{1}{2}\mathrm{tr}\Big\{\big[(\mathbb{Y} - 1_n\boldsymbol{\mu}_{\mathbf{Y}}^T) - (\mathbb{X} - 1_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\beta}\big]\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \\ & \big[(\mathbb{Y} - 1_n\boldsymbol{\mu}_{\mathbf{Y}}^T) - (\mathbb{X} - 1_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\beta}\big]^T\Big\}, \end{aligned}$$

where $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n \times p}$, $\mathbb{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^T \in \mathbb{R}^{n \times r}$ are data matrices, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. samples of $\mathbf{X}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d. samples of $\mathbf{Y}$. Then the oracle predictor envelope estimator is obtained by minimizing the objective function $l_{\max} - l_{(\mathbf{X},\mathbf{Y})}$, where $l_{\max}$ is the maximized log-likelihood function under the standard model. The objective function $l_{\max} - l_{(\mathbf{X},\mathbf{Y})}$ satisfies Conditions 1–4 in Section 3 of Shapiro (1986). Furthermore, since $\mathbf{J}$ is full rank, rank$(\mathbf{H}^T\mathbf{J}\mathbf{H})$ =rank$(\mathbf{H})$. Then all the conditions in Proposition 4.1 of Shapiro (1986) are satisfied. Using this Proposition, we have

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\xi}}) - \mathrm{vec}(\boldsymbol{\xi})\} \xrightarrow{d} N(0, \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T),$$

where † denotes the Moore-Penrose generalized inverse. After some straightforward calculations similar to those in Theorem 5.1 of Cook, Li and Chiaromonte (2010), we have

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})\} \xrightarrow{d} N(0, \mathbf{V}_O),$$

where $\mathbf{V}_O = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}_{\mathcal{A}}} \otimes \boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}_{\mathcal{A}}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)$, and $\mathbf{T} = (\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}_{\mathcal{A}}}^{-1}\boldsymbol{\eta}^T + \boldsymbol{\Omega}^{-1}) \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}$.

C.4. Proof of Theorem 3. Let $\widehat{\mathbf{A}}_{\mathcal{A}}$ denote the E-SPLS estimator of $\mathbf{A}_{\mathcal{A}}$. Suppose we can prove $\widehat{\mathbf{A}}_{\mathcal{A}} = \widehat{\mathbf{A}}_{\mathcal{A},O} + o_p(a_n)$, for a sequence a_n that $a_n = o(n^{-1/2})$, then $\widehat{\mathbf{G}}_{\mathbf{A}_{\mathcal{A}}} = \widehat{\mathbf{G}}_{\mathbf{A}_{\mathcal{A},O}} + o_p(a_n)$, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} = \widehat{\boldsymbol{\beta}}_{\mathcal{A},O} + o_p(a_n)$, and $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})\}$ and $\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})\}$ converge to the same asymptotic distribution. Regarding the choice of a_n, we can take $a_n = (n^{-1/2}\lambda_{\mathcal{A}})^{1/2}$.

Since the E-SPLS estimator enjoys the selection consistency, the objective function to estimate $\mathbf{A}_{\mathcal{A}}$ is

$$f_{\mathrm{obj},\mathcal{A}}(\mathbf{A}_{\mathcal{A}}) = -2\log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}| + \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T\mathbf{S}_{\mathbf{X}_{\mathcal{A}}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}| + \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T(\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}}\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}| + \sum_{i=1}^{p_{\mathcal{A}}-d}\lambda w_i\|\mathbf{a}_i\|_2.$$

In order to prove $\widehat{\mathbf{A}}_{\mathcal{A}} = \widehat{\mathbf{A}}_{\mathcal{A},O} + o_p(a_n)$, it is sufficient to show that for any small $\epsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_n P\left(\inf_{\boldsymbol{\Delta}\in\mathbb{R}^{(p_{\mathcal{A}}-d)\times d}, \|\boldsymbol{\Delta}\|_F = C} f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O} + a_n\boldsymbol{\Delta}) > f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O})\right) > 1 - \epsilon. \tag{10}$$

We name the four summands in $f_{\mathrm{obj},\mathcal{A}}(\mathbf{A}_{\mathcal{A}})$ as $f_1(\mathbf{A}_{\mathcal{A}})$, $f_2(\mathbf{A}_{\mathcal{A}})$, $f_3(\mathbf{A}_{\mathcal{A}})$ and $f_4(\mathbf{A}_{\mathcal{A}})$, i.e.

$$f_1(\mathbf{A}_{\mathcal{A}}) = -2\log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}|, \qquad f_2(\mathbf{A}_{\mathcal{A}}) = \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T\mathbf{S}_{\mathbf{X}_{\mathcal{A}}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}|,$$

$$f_3(\mathbf{A}_{\mathcal{A}}) = \log|\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}^T(\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}}\mathbf{G}_{\mathbf{A}_{\mathcal{A}}}|, \qquad f_4(\mathbf{A}_{\mathcal{A}}) = \sum_{i=1}^{p_{\mathcal{A}}-d}\lambda w_i\|\mathbf{a}_i\|_2.$$

Notice that the objective function of $\mathbf{A}_{\mathcal{A}}$ under the oracle predictor envelope model (9) is $f_1(\mathbf{A}_{\mathcal{A}}) + f_2(\mathbf{A}_{\mathcal{A}}) + f_3(\mathbf{A}_{\mathcal{A}})$. So

$$\frac{df_1(\mathbf{A}_{\mathcal{A}})}{d\mathbf{A}_{\mathcal{A}}}\Big|_{\mathbf{A}_{\mathcal{A}}=\widehat{\mathbf{A}}_{\mathcal{A},O}} + \frac{df_2(\mathbf{A}_{\mathcal{A}})}{d\mathbf{A}_{\mathcal{A}}}\Big|_{\mathbf{A}_{\mathcal{A}}=\widehat{\mathbf{A}}_{\mathcal{A},O}} + \frac{df_3(\mathbf{A}_{\mathcal{A}})}{d\mathbf{A}_{\mathcal{A}}}\Big|_{\mathbf{A}_{\mathcal{A}}=\widehat{\mathbf{A}}_{\mathcal{A},O}} = 0. \tag{11}$$

Now we compute $f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}+a_n\boldsymbol{\Delta})-f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O})$ using Taylor expansion.

$$\begin{aligned}
& f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}+a_n\boldsymbol{\Delta})-f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}) \\
= \;& a_n\left\{\overset{\rightarrow\boldsymbol{\Delta}}{df_1}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_3}(\widehat{\mathbf{A}}_{\mathcal{A},O})\right\} \\
& +\frac{1}{2}a_n^2\left\{\overset{\rightarrow\boldsymbol{\Delta}}{df_1^2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_2^2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_3^2}(\widehat{\mathbf{A}}_{\mathcal{A},O})\right\} \\
& +f_4(\widehat{\mathbf{A}}_{\mathcal{A},O}+a_n\boldsymbol{\Delta})-f_4(\widehat{\mathbf{A}}_{\mathcal{A},O})+o_p(a_n^2).
\end{aligned}$$

Because of (11), we have

$$\overset{\rightarrow\boldsymbol{\Delta}}{df_1}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_3}(\widehat{\mathbf{A}}_{\mathcal{A},O})=0. \tag{12}$$

Let $\boldsymbol{\Delta}_{*\mathcal{A}}=(0,\boldsymbol{\Delta}^T)^T\in\mathbb{R}^{p_{\mathcal{A}}\times d}$. Following similar calculations as those in proving (5), we have

$$\begin{aligned}
&\overset{\rightarrow\boldsymbol{\Delta}}{df_1^2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_2^2}(\widehat{\mathbf{A}}_{\mathcal{A},O})+\overset{\rightarrow\boldsymbol{\Delta}}{df_3^2}(\widehat{\mathbf{A}}_{\mathcal{A},O}) \\
&=2\,\mathrm{tr}\Big\{(\boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1+\boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1 \\
&\quad-2(\mathbf{I}_d+\mathbf{A}_{\mathcal{A}}^T\mathbf{A}_{\mathcal{A}})^{-1}\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\Big\}+o_p(1),
\end{aligned} \tag{13}$$

and

$$f_4(\widehat{\mathbf{A}}_{\mathcal{A},O}+a_n\boldsymbol{\Delta})-f_4(\widehat{\mathbf{A}}_{\mathcal{A},O})\geq\sum_{i=1}^{p_{\mathcal{A}}-d}-(p_{\mathcal{A}}-d)a_n\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\|\boldsymbol{\delta}_i\|_2. \tag{14}$$

Since $a_n=(n^{-1/2}\lambda_{\mathcal{A}})^{1/2}$ and $\lambda_{\mathcal{A}}=o(n^{-1/2})$, then $a_n\lambda_{\mathcal{A}}=o(a_n^2)$. The right-hand side of (14) is dominated by a_n^2.

Combining (12), (13) and (14),

$$\begin{aligned}
& f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}+a_n\boldsymbol{\Delta})-f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}) \\
= \;& a_n^2\,\mathrm{tr}\Big\{(\boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1+\boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1 \\
& -2(\mathbf{I}_d+\mathbf{A}_{\mathcal{A}}^T\mathbf{A}_{\mathcal{A}})^{-1}\boldsymbol{\Delta}_{*\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}^T\Big\}+o_p(a_n^2) \\
= \;& a_n^2\mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1)^T\mathbf{T}\mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1)+o_p(a_n^2),
\end{aligned}$$

where $\mathbf{T} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T) \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}$ is defined in Proposition 1, and it is invertible. Because

$$\begin{aligned}\mathbf{T} &= \boldsymbol{\Omega}^{-1} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} \\ &\geq \boldsymbol{\Omega}^{-1} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} \\ &= \left(\boldsymbol{\Omega}^{-1/2} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}^{1/2} - \boldsymbol{\Omega}^{1/2} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}^{-1/2}\right)^2 + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}},\end{aligned}$$

$\mathbf{T}$ is semi-positive definite. Since $\mathbf{T}$ is invertible, $\mathbf{T}$ is then positive definite. Let m_1 denote its smallest eigenvalue, then

$$\mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1)^T\mathbf{T}\mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1) \geq m_1\|\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1\|_F^2.$$

Following the discussion at the end of the proof of Theorem 1, we also have

$$\|\boldsymbol{\Gamma}_{\mathcal{A},0}^T\boldsymbol{\Delta}_{*\mathcal{A}}\boldsymbol{\Gamma}_1\|_F \geq m_2^2\|\boldsymbol{\Delta}\|_F^2,$$

where m_2 denote the smallest eigenvalue of $(\mathbf{I}_d + \mathbf{A}_{\mathcal{A}}^T\mathbf{A}_{\mathcal{A}})^{-1}$. Substituting these results to $f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O} + a_n\boldsymbol{\Delta}) - f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O})$, we have

$$f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O} + a_n\boldsymbol{\Delta}) - f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}) \geq a_n^2 m_1 m_2^2\|\boldsymbol{\Delta}\|_F^2 + o_p(a_n^2).$$

Since $f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O} + a_n\boldsymbol{\Delta}) - f_{\mathrm{obj},\mathcal{A}}(\widehat{\mathbf{A}}_{\mathcal{A},O}) > 0$ with probability tending to 1, we have established (10).

C.5. Proof of Theorem 4. We first show that

$$\|\mathbf{S}_{\mathbf{X},\mathrm{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\|_F = O_p[\{(p_n + s_1)\log p_n/n\}^{1/2}], \tag{15}$$

$$\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\mathrm{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}\|_F = O_p[\{(p_n + s_2)\log p_n/n\}^{1/2}]. \tag{16}$$

Let $\|\cdot\|_{\max}$ denotes the max norm of a matrix, which is the maximum of the absolute values of all elements in the matrix. To establish (15) and (16), it is sufficient to show that there exist positive constants $C_{\mathbf{X}}$ and $C_{\mathbf{X}|\mathbf{Y}}$ such that

$$\|\mathbf{S}_{\mathbf{X}}-\boldsymbol{\Sigma}_{\mathbf{X}}\|_{\max} \leq C_{\mathbf{X}}\{\log(p_n)/n\}^{1/2}, \qquad \|\mathbf{S}_{\mathbf{X}|\mathbf{Y}}-\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|_{\max} \leq C_{\mathbf{X}|\mathbf{Y}}\{\log(p_n)/n\}^{1/2}.$$

Let $\mathbf{V} = (\mathbf{X}^T - \boldsymbol{\mu}_{\mathbf{X}}^T, \boldsymbol{\varepsilon}^T)^T \in \mathbb{R}^{r+p_n}$, $\boldsymbol{\mu}_{\mathbf{V}}$ denote the mean of $\mathbf{V}$, $\boldsymbol{\Sigma}_{\mathbf{V}}$ denote the covariance matrix of $\mathbf{V}$ and $\sigma^2 = \max(\sigma_1^2, \sigma_2^2)$. Then each $\mathbf{V}_i/\sqrt{\boldsymbol{\Sigma}_{\mathbf{V},ii}}$, $i = 1, \ldots, r + p_n$, follows a sub-Gaussian distribution with parameter σ^2, where $\mathbf{V}_i$ is the ith element in $\mathbf{V}$. Let $\bar{\mathbf{V}}$ denote the sample mean of $\mathbf{V}$ and $\mathbf{S}_{\mathbf{V}}$ denote the sample covariance matrix of $\mathbf{V}$.

For a matrix $\mathbf{A}$, let $\mathbf{A}_{ij}$ denote the (i,j)th element in $\mathbf{A}$. Then for $\delta > 0$,

$$\begin{aligned} P(|\mathbf{S}_{\mathbf{V},ij} - \mathbf{\Sigma}_{\mathbf{V},ij}| > \delta) \quad &\leq \quad P\left(\left|\left\{\frac{1}{n}\sum_{k=1}^{n}(\mathbf{V}_k - \boldsymbol{\mu}_{\mathbf{V}})(\mathbf{V}_k - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ij} - \mathbf{\Sigma}_{\mathbf{V},ij}\right| > \frac{\delta}{2}\right) \\ &+ \quad P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ij}\right| > \frac{\delta}{2}\right). \end{aligned}$$

From Lemma 1 in Ravikumar et al. (2011), there exist positive constants C_1 and C_2 such that

(17)
$$P\left(\left|\left\{\frac{1}{n}\sum_{k=1}^{n}(\mathbf{V}_k - \boldsymbol{\mu}_{\mathbf{V}})(\mathbf{V}_k - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ij} - \mathbf{\Sigma}_{\mathbf{V},ij}\right| > \frac{\delta}{2}\right) \leq C_1 \exp(-C_2 n\delta^2)$$

for all $\delta \in \big(0, 8\bar{k}(1+4\sigma^2)\big)$. Let $(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})_i$ denote the ith element in the vector $\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}}$. Then we have

$$\mathbf{E}\big\{e^{t(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})_i}\big\} \leq \prod_{k=1}^{n}\exp\left(\frac{t^2\sigma^2}{2n^2}\right) = \exp\left(\frac{t^2\sigma^2}{2n}\right).$$

By the Chernoff's bound, we have

$$P\left(\left|(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})_i\right| > \delta\right) \leq 2\exp\left(-\frac{n\delta^2}{2\sigma^2}\right).$$

For any $\delta \in (0, 1/2)$, we have $\delta^2 < \delta/2$ and

(18)
$$\begin{aligned} &P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ij}\right| > \frac{\delta}{2}\right) \\ \leq &P\left(\left|\sqrt{\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ii}}\sqrt{\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{jj}}\right| > \frac{\delta}{2}\right) \\ \leq &P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ii}\right| > \frac{\delta}{2}\right) + P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{jj}\right| > \frac{\delta}{2}\right) \\ \leq &P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{ii}\right| > \delta^2\right) + P\left(\left|\left\{(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})^T\right\}_{jj}\right| > \delta^2\right) \\ \leq &P\left(\left|(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})_i\right| > \delta\right) + P\left(\left|(\bar{\mathbf{V}} - \boldsymbol{\mu}_{\mathbf{V}})_j\right| > \delta\right) \\ \leq &4\exp\left(-\frac{n\delta^2}{2\sigma^2}\right). \end{aligned}$$

By (17) and (18), there exists positive constants C_3 and C_4 such that

$$P(|\mathbf{S}_{\mathbf{V},ij} - \mathbf{\Sigma}_{\mathbf{V},ij}| > \delta) \leq C_3 \exp(-C_4 n\delta^2).$$

Let $\delta = C\sqrt{\log(p_n)/n}$ for some $C > 0$. By the union sum inequality, there exists a positive constant $C_{\mathbf{V}}$ such that

$$\|\mathbf{S}_{\mathbf{V}} - \mathbf{\Sigma}_{\mathbf{V}}\|_{\max} \leq C_{\mathbf{V}}\{\log(p_n)/n\}^{1/2}. \tag{19}$$

with probability tending to 1 as n tends to infinity. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be n samples of $\mathbf{X}$, $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be n samples of $\mathbf{Y}$ and $\mathbf{e}_1, \dots, \mathbf{e}_n$ be n samples of $\boldsymbol{\varepsilon}$. Then $\mathbf{Y}_i = \boldsymbol{\mu} + \boldsymbol{\beta}^T(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}) + \mathbf{e}_i$. Note that we only observe $\mathbf{X}_i$ and $\mathbf{Y}_i$, but we cannot observe $\mathbf{e}_i$. Define $\bar{\mathbf{e}} = \sum_{i=1}^n \mathbf{e}_i/n$, $\mathbf{S}_{\boldsymbol{\varepsilon}} = \sum_{i=1}^n (\mathbf{e}_i - \bar{\mathbf{e}})^T(\mathbf{e}_i - \bar{\mathbf{e}})/n$ and $\mathbf{S}_{\mathbf{X}\boldsymbol{\varepsilon}} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^T(\mathbf{e}_i - \bar{\mathbf{e}})/n$. Then from (19), we have

$$\begin{aligned} \|\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}\|_{\max} &\leq C_{\mathbf{V}}\{\log(p_n)/n\}^{1/2} \quad \|\mathbf{S}_{\mathbf{X}\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\mathbf{X}\boldsymbol{\varepsilon}}\|_{\max} \leq C_{\mathbf{V}}\{\log(p_n)/n\}^{1/2} \\ \|\mathbf{S}_{\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\|_{\max} &\leq C_{\mathbf{V}}\{\log(p_n)/n\}^{1/2}. \end{aligned}$$

By Proposition 1 in Li, Chun and Zhao (2012), if $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{N} \in \mathbb{R}^{d_2 \times d_3}$, then

$$\|\mathbf{M}\mathbf{N}\|_{\max} \leq d_2\|\mathbf{M}\|_{\max}\|\mathbf{N}\|_{\max}. \tag{20}$$

Since $\mathbf{S}_{\mathbf{XY}} = \mathbf{S}_{\mathbf{X}}\boldsymbol{\beta}^T + \mathbf{S}_{\mathbf{X}\boldsymbol{\varepsilon}}$ and $\boldsymbol{\beta}$ has $p_{\mathcal{A}}r$ nonzero elements, then

$$\begin{aligned} \|\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\max} &\leq \|\mathbf{S}_{\mathbf{X}}\boldsymbol{\beta}^T - \mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\|_{\max} + \|\mathbf{S}_{\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\|_{\max} \\ &\leq p_{\mathcal{A}}\|\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}\|_{\max}\|\boldsymbol{\beta}^T\|_{\max} + \|\mathbf{S}_{\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\|_{\max}. \end{aligned}$$

So there exists a positive constant C_5 such that

$$\|\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\max} \leq C_5\{\log(p_n)/n\}^{1/2}.$$

Since $\boldsymbol{\varepsilon}$ has finite fourth moment and $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, $\mathbf{Y}$ has finite fourth moment. Then $\mathbf{S}_{\mathbf{Y}}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{\Sigma}_{\mathbf{Y}}$, i.e. $\mathbf{S}_{\mathbf{Y}} = \mathbf{\Sigma}_{\mathbf{Y}} + O_p(n^{-1/2})$, and $\mathbf{S}_{\mathbf{Y}}^{-1} = \mathbf{\Sigma}_{\mathbf{Y}}^{-1} + O_p(n^{-1/2})$. Then for any positive constant C_6

$$\|\mathbf{S}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Y}}\|_{\max} \leq C_6[\log(p_n)/n]^{1/2}$$

with probability tending to 1 as n tends to infinity. Because $\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = \mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{XY}}^T$, we have

$$\begin{aligned} \mathbf{S}_{\mathbf{X}|\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}} =& (\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}) - (\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}^T - \mathbf{\Sigma}_{\mathbf{XY}}(\mathbf{S}_{\mathbf{Y}}^{-1} - \mathbf{\Sigma}_{\mathbf{Y}}^{-1})\mathbf{\Sigma}_{\mathbf{XY}}^T \\ & - \mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}(\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})^T - (\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})(\mathbf{S}_{\mathbf{Y}}^{-1} - \mathbf{\Sigma}_{\mathbf{Y}}^{-1})\mathbf{\Sigma}_{\mathbf{XY}}^T \\ & - \mathbf{\Sigma}_{\mathbf{XY}}(\mathbf{S}_{\mathbf{Y}}^{-1} - \mathbf{\Sigma}_{\mathbf{Y}}^{-1})(\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})^T - (\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})\mathbf{\Sigma}_{\mathbf{Y}}^{-1}(\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})^T \\ & - (\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})(\mathbf{S}_{\mathbf{Y}}^{-1} - \mathbf{\Sigma}_{\mathbf{Y}}^{-1})(\mathbf{S}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})^T. \end{aligned}$$

Note that $\|\mathbf{\Sigma}_{\mathbf{X}}\|_{\max} \leq \bar{k}$ and $\|\mathbf{\Sigma}_{\mathbf{XY}}\|_{\max} = \|\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}\|_{\max} \leq p_{\mathcal{A}}\bar{k}\|\boldsymbol{\beta}\|_{\max}$. By repeatedly using (20) in the expansion of $\mathbf{S}_{\mathbf{X}|\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}$, there exists a positive integer $C_{\mathbf{X}|\mathbf{Y}}$ such that $\|\mathbf{S}_{\mathbf{X}|\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|_{\max} \leq C_{\mathbf{X}|\mathbf{Y}}\{\log(p_n)/n\}^{1/2}$. Take $C_{\mathbf{X}} = C_{\mathbf{V}}$, then we have $\|\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}\|_{\max} \leq C_{\mathbf{X}}\{\log(p_n)/n\}^{1/2}$. Then we have established (15) and (16).

Let $a_n = \sqrt{(p_n + s)\log(p_n)/n}$. Now we show $\|\widehat{\mathbf{A}} - \mathbf{A}\|_F = O_p\left(\sqrt{(p_n + s)\log(p_n)/n}\right)$. We denote the objective function in (8) as $f_{\text{obj},2}$. It is sufficient to show that for any small $\epsilon > 0$, there exists a sufficiently large constant C such that

$$\lim_{n\to\infty} P\left(\inf_{\mathbf{\Delta}\in\mathbb{R}^{(p_n-d)\times d}, \|\mathbf{\Delta}\|_F = C} f_{\text{obj},2}(\mathbf{A} + a_n\mathbf{\Delta}) > f_{\text{obj},2}(\mathbf{A})\right) > 1 - \epsilon. \tag{21}$$

Let $\mathbf{\Delta}_* = (0_{d\times d}, \mathbf{\Delta}^T)^T \in \mathbb{R}^{p_n\times d}$. Following the calculations as in Theorem 1, we compute the Taylor expansion of $f_{\text{obj},2}(\mathbf{A} + a_n\mathbf{\Delta})$ at $\mathbf{A}$, and get

$$\begin{aligned}
& f_{\text{obj},2}(\mathbf{A} + a_n\mathbf{\Delta}) - f_{\text{obj},2}(\mathbf{A}) \\
\geq\ & 2a_n \operatorname{tr}\Big[(\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}})\mathbf{\Delta}_* \\
& + \{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{\Delta}_* \\
& + \{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}})\mathbf{\Delta}_*\Big] \\
& + 2a_n \operatorname{tr}\Big[(\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1} - \mathbf{\Sigma}_{\mathbf{X}}^{-1})\mathbf{\Delta}_* \\
& + \{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{\Delta}_* \\
& + \{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1} - \mathbf{\Sigma}_{\mathbf{X}}^{-1})\mathbf{\Delta}_*\Big] \\
& + a_n^2 \operatorname{tr}\Big\{(\mathbf{\Omega}^{-1} + \boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1 + \mathbf{\Omega}\mathbf{\Gamma}_1^T\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{\Delta}_*\mathbf{\Gamma}_1 \\
& - 2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{\Delta}_*^T\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\mathbf{\Delta}_*^T\Big\} - a_n^2(p_{\mathcal{A}} - d)a_n^{-1}\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\|\boldsymbol{\delta}_i\|_2 + o_p(a_n^2).
\end{aligned}$$

Let $\|\cdot\|$ denote the spectral norm of a matrix. For any two matrices $\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$ and $\mathbf{N} \in \mathbb{R}^{d_2\times d_3}$,

$$\|\mathbf{MN}\|_F \leq \|\mathbf{M}\|\|\mathbf{N}\|_F. \tag{22}$$

Then by (22) and Cauchy-Schwartz inequality, the first term in the preceding display can be lower bounded as follows: $\operatorname{tr}\{(\mathbf{G}_{\mathbf{A}}^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} -$

$\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}})\boldsymbol{\Delta}_*\} \geq -\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|_F\|\boldsymbol{\Delta}\|_F\|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\|\|\mathbf{G}_{\mathbf{A}}\|$. Since

$$\begin{aligned}
& \{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_* \\
= \; & -(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}} - \mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_* \\
& +o_p(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}} - \mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}) \\
= \; & (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}(\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1})\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_* \\
& +o_p(a_n),
\end{aligned}$$

the second term can be lower bounded as follows

$$\begin{aligned}
& \operatorname{tr}[\{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_*] \\
& \geq -\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}\|_F\|\boldsymbol{\Delta}\|_F\|\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}})}\|\|\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|\|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|.]
\end{aligned}$$

Following the same derivations and also using the fact that for any matrix $\mathbf{A}$, $\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$, the third term can be lower bounded by $\{(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1} - (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\}\mathbf{G}_{\mathbf{A}}^T(\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}})\boldsymbol{\Delta}_* \geq -\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}\|_F\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|_F\|\boldsymbol{\Delta}\|_F\|\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}})}\|\|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|$. Based on the convergence rate of $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1}$, this lower bound is in the order of a_n^2. The bounds for the fourth till the sixth terms can be developed similarly. Let $m_1 = \|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\|\|\mathbf{G}_{\mathbf{A}}\|$, $m_2 = \|\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}})}\|\|\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|\|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|$, $m_3 = \|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\|\|\mathbf{G}_{\mathbf{A}}\|$ and $m_4 = \|\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})}\|\|\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\|\|(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\|$. We also notice that since $\lambda_{\mathcal{A}} = o(a_n)$, then $a_n^{-1}\lambda_{\mathcal{A}} = o(1)$, and $a_n^2(p_{\mathcal{A}} - d)a_n^{-1}\lambda_{\mathcal{A}} \max_{1\leq i\leq p_{\mathcal{A}}-d}\|\boldsymbol{\delta}_i\|_2 = o_p(a_n^2)$. Collecting all these bounds and results together, we have

$$\begin{aligned}
& f_{\text{obj},2}(\mathbf{A} + a_n\boldsymbol{\Delta}) - f_{\text{obj},2}(\mathbf{A}) \\
\geq \; & -2a_n\big(m_1\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\|_F\|\boldsymbol{\Delta}\|_F + m_2\|\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}\|_F\|\boldsymbol{\Delta}\|_F \\
& +m_3\|\mathbf{S}_{\mathbf{X},\text{spice}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\|_F\|\boldsymbol{\Delta}\|_F + m_4\|\mathbf{S}_{\mathbf{X},\text{spice}} - \boldsymbol{\Sigma}_{\mathbf{X}}\|_F\|\boldsymbol{\Delta}\|_F\big) \\
& +a_n^2\operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 + \boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*^T\Big\} + o_p(a_n^2).
\end{aligned}$$

According to the proof at the end of Theorem 1, there exist a positive constant k such that

$$\begin{aligned}
& \operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 + \boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*^T\Big\} \\
\geq \; & k\|\boldsymbol{\Delta}\|_F^2.
\end{aligned}$$

Then with sufficiently large C, the second order term of $\|\boldsymbol{\Delta}\|_F$ dominates the first order term, and $f_{\text{obj},2}(\mathbf{A}+a_n\boldsymbol{\Delta})-f_{\text{obj},2}(\mathbf{A})>0$ with probability tending to 1. Then we have established (21). So $\|\widehat{\mathbf{A}}-\mathbf{A}\|_F=O_p\left(a_n\right)$, where $a_n=\sqrt{(p_n+s)\log(p_n)/n}$.

Since $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X},\text{spice}}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{XY}}=\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{S}_{\mathbf{XY}}$,

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F=\|\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{S}_{\mathbf{XY}}-\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})\|_F+\|(\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}-\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}})})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\|_F.$$

Since $\|\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}\|=1$ and $\mathbf{S}_{\mathbf{XY}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Sigma}_{\mathbf{XY}}$, then

$$\begin{aligned}
&\|\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{S}_{\mathbf{XY}}-\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})\|_F\le\|\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{S}_{\mathbf{XY}}-\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\|_F\\
\le\ &\|(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}-\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\boldsymbol{\Sigma}_{\mathbf{XY}}\|_F+\|\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{S}_{\mathbf{XY}}-\boldsymbol{\Sigma}_{\mathbf{XY}})\|_F+\|(\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}-\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})(\mathbf{S}_{\mathbf{XY}}-\boldsymbol{\Sigma}_{\mathbf{XY}})\|_F\\
=\ &O_p(a_n),
\end{aligned}$$

and

$$\|(\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}-\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}})})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\|_F\le\|\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}-\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}})}\|\|\boldsymbol{\beta}\|_F.$$

Because $\|\mathbf{S}_{\mathbf{X},\text{spice}}-\boldsymbol{\Sigma}_{\mathbf{X}}\|_F=O_p\left(a_n\right)$, $\|\widehat{\mathbf{A}}-\mathbf{A}\|_F=O_p\left(a_n\right)$, then $\|\mathbf{P}_{\widehat{\mathbf{G}}_{\mathbf{A}}(\mathbf{S}_{\mathbf{X},\text{spice}})}-\mathbf{P}_{\mathbf{G}_{\mathbf{A}}(\boldsymbol{\Sigma}_{\mathbf{X}})}\|=O_p\left(a_n\right)$. So we have $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F=O_p(a_n)$.

C.6. Proof of Theorem 5. Because of the consistency results in Theorem 4 that $\|\widehat{\mathbf{A}}-\mathbf{A}\|_F=O_p(\sqrt{(p_n+s)\log(p_n)/n})$, then $\widehat{\mathbf{a}}_i$, $i=1,\dots,p_{\mathcal{A}}-d$, converges to $\mathbf{a}_i$ with the rate $\sqrt{n/[(p_n+s)\log(p_n)]}$. Therefore $P(\widehat{\mathbf{a}}_i\neq 0)\to 1$ for $i=1,\dots,p_{\mathcal{A}}-d$.

Now we prove the selection consistency $P(\widehat{\mathbf{a}}_i=0, i=p_{\mathcal{A}}-d+1,\dots,p_n-d)\to 1$. Let $f_{\text{obj},2}$ be the objective function for $\mathbf{A}$ defined in (8). Let

$$\mathbf{A}_{\mathcal{A}}=\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_{p_{\mathcal{A}}-d}\end{pmatrix}\in\mathbb{R}^{(p_{\mathcal{A}}-d)\times d}\quad\text{and}\quad\mathbf{A}_{\mathcal{I}}=\begin{pmatrix}\mathbf{a}_{p_{\mathcal{A}}-d+1}\\ \vdots\\ \mathbf{a}_{p_n-d}\end{pmatrix}\in\mathbb{R}^{p_{\mathcal{I}}\times d}.$$

Then $\mathbf{A}=(\mathbf{A}_{\mathcal{A}}^T,\mathbf{A}_{\mathcal{I}}^T)^T$. Let $a_n=\sqrt{(p_n+s)\log(p_n)/n}$,

$$\widehat{\mathbf{A}}_0=\begin{pmatrix}\widehat{\mathbf{A}}_{\mathcal{A}}\\ 0\end{pmatrix}\in\mathbb{R}^{(p_n-d)\times d}\quad\text{and}\quad\boldsymbol{\Delta}_0=\begin{pmatrix}0\\ \widehat{\mathbf{A}}_{\mathcal{I}}\end{pmatrix}\in\mathbb{R}^{(p_n-d)\times d}.$$

We need to show that for any constant $C>0$, if $\|\widehat{\mathbf{A}}_{\mathcal{A}}-\mathbf{A}_{\mathcal{A}}\|_F=O_p(a_n)$, then

$$\lim_{n\to\infty}P\left(\inf_{\|\widehat{\mathbf{A}}_{\mathcal{I}}\|_F\le a_nC,\widehat{\mathbf{A}}_{\mathcal{I}}\neq 0}f_{\text{obj},2}(\widehat{\mathbf{A}}_0+\boldsymbol{\Delta}_0)>f_{\text{obj},2}(\widehat{\mathbf{A}}_0)\right)>1-\epsilon,\tag{23}$$

for any small $\epsilon > 0$.

Let $\boldsymbol{\Delta} = \boldsymbol{\Delta}_0/a_n$, then $\|\boldsymbol{\Delta}\|_F \leq C$. Let $\boldsymbol{\Delta}_* = (0_{d\times d}, \boldsymbol{\Delta}^T)^T \in \mathbb{R}^{p_n\times d}$. Using the Taylor expansion of $f_{\text{obj},2}(\widehat{\mathbf{A}}_0 + a_n\boldsymbol{\Delta})$ at $\widehat{\mathbf{A}}_0$, we have

$$\begin{aligned}
& f_{\text{obj},2}(\widehat{\mathbf{A}}_0 + a_n\boldsymbol{\Delta}) - f_{\text{obj},2}(\widehat{\mathbf{A}}_0) \\
= \;& 2a_n \operatorname{tr}\big[(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\boldsymbol{\Delta}_* \\
& +(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\boldsymbol{\Delta}_* - 2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\big] \\
& +a_n^2 \operatorname{tr}\big[(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 \\
& +\boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 - 2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\big] \\
& +\lambda \sum_{i=p_{\mathcal{A}-d+1}}^{p_n-d} \mathbf{w}_i\|\widehat{\mathbf{a}}_i\|_2 + o_p(a_n^2).
\end{aligned}$$

First

$$\lambda \sum_{i=p_{\mathcal{A}-d+1}}^{p_n-d} w_i\|\widehat{\mathbf{a}}_i\|_2 \geq \lambda_{\mathcal{I}} \sum_{i=p_{\mathcal{A}-d+1}}^{p_n-d} \|\widehat{\mathbf{a}}_i\|_2.$$

Using equations (15) and (16), $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} = \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} + O_p(a_n)$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}} = \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} + O_p(a_n)$.

$$\begin{aligned}
& \operatorname{tr}\big[(\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\boldsymbol{\Delta}_* + (\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\boldsymbol{\Delta}_* \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\big] \\
= \;& \operatorname{tr}\big[(\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\boldsymbol{\Delta}_* + (\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}})^{-1}\mathbf{G}_{\mathbf{A}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Delta}_* \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Delta}\big] + O_p(a_n) \\
= \;& O_p(a_n).
\end{aligned}$$

Based on the calculation in Theorem 4, we also have

$$\begin{aligned}
& \operatorname{tr}\Big\{(\boldsymbol{\Omega}^{-1} + \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 + \boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*\boldsymbol{\Gamma}_1 \\
& -2(\mathbf{I}_d + \mathbf{A}^T\mathbf{A})^{-1}\boldsymbol{\Delta}_*^T\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_*^T\Big\} \\
\geq \;& k\|\boldsymbol{\Delta}\|_F^2,
\end{aligned}$$

where k is a positive constant.

Combining all these results,

$$f_{\text{obj},2}(\widehat{\mathbf{A}}_0 + a_n\boldsymbol{\Delta}) - f_{\text{obj},2}(\widehat{\mathbf{A}}_0) \geq O_p(a_n^2) + a_n^2 k\|\boldsymbol{\Delta}\|_F^2 + a_n^2\frac{\lambda_{\mathcal{I}}}{a_n}\sum_{i=p_{\mathcal{A}-d+1}}^{p_n-d} \|\widehat{\mathbf{a}}_i/a_n\|_2.$$

Since $\widehat{\mathbf{A}}_{\mathcal{I}} \neq 0$ and $\|\boldsymbol{\Delta}\|_F \leq C$, we have $0 < \sum_{i=p_{\mathcal{A}}-d+1}^{p_n-d} \|\widehat{\mathbf{a}}_i/a_n\|_2 \leq C$. Because $a_n = o(\lambda_{\mathcal{I}})$, then $\lambda_{\mathcal{I}}/a_n \to \infty$. Therefore $f_{\text{obj},2}(\widehat{\mathbf{A}}_0 + a_n\boldsymbol{\Delta}) > f_{\text{obj},2}(\widehat{\mathbf{A}}_0)$ with probability tending to 1. In other words, we have established (23).

C.7. Proof of Theorem 6. Before we prove Theorem 6, we first introduce some results on manifold theory, as well as the properties of the oracle estimator.

Outline of manifold theory. We first introduce a few concepts and results on Stiefel manifold and Grassmann manifold, which will play an important role in this proof. A Stiefel manifold, denoted by $St(p, d)$, is the set of all $p \times d$ semi-orthogonal matrices. In other words, $St(p, d) = \{\mathbf{G} \in \mathbb{R}^{p\times d} : \mathbf{G}^T\mathbf{G} = \mathbf{I}_d\}$. A Grassmann manifold is $\mathcal{G}(p, d) = \{\text{span}(\mathbf{G}) : \mathbf{G} \in St(p, d)\}$.

Next we define neighborhood in manifolds, which is a key concept used in our proofs. In preparation, we first introduce projection operator and tangent space in manifolds. The projection operator onto a Stiefel manifold $R : \mathbb{R}^{p\times d} \to St(p, d)$ is defined as

$$R(\mathbf{M}) = \underset{\mathbf{G}\in St(p,d)}{\arg\min} \|\mathbf{M} - \mathbf{G}\|_F^2,$$

where $\mathbf{M} \in \mathbb{R}^{p\times d}$ is an arbitrary matrix. By Proposition 7 in Manton (2002), if $\mathbf{M} = \mathbf{LDR}^T$ is a singular value decomposition of $\mathbf{M}$, then $R(\mathbf{M}) = \mathbf{L}\mathbf{I}_{p,d}\mathbf{R}^T \in St(p, d)$, where $\mathbf{I}_{p,d} \in \mathbb{R}^{p\times d}$ is obtained by replacing all the nonzero elements in $\mathbf{D}$ by 1. The tangent space $T_{\mathbf{\Gamma}}(p, d)$ of $\mathbf{G} \in St(p, d)$ is defined as

$$T_{\mathbf{G}}(p, d) = \{\mathbf{Z} \in \mathbb{R}^{p\times d} : \mathbf{Z} = \mathbf{GA}+\mathbf{G}_0\mathbf{B}, \text{ where } \mathbf{A} \in \mathbb{R}^{d\times d}, \mathbf{A}+\mathbf{A}^T = 0, \mathbf{B} \in \mathbb{R}^{(p-d)\times d}\},$$

$\mathbf{G}_0$ is the completion of $\mathbf{G}$ such that $(\mathbf{G}, \mathbf{G}_0)$ is an orthogonal matrix (Manton, 2002). Then for a point $\mathbf{G}$ on a Stiefel manifold, the neighborhood around $\mathbf{G}$ is $R(\mathbf{G} + \delta\mathbf{M})$, where $\mathbf{M} \in \mathbb{R}^{p\times d}$ is an arbitrary matrix, δ is a scalar, and R is the projection operator. Then for a point $\text{span}(\mathbf{G})$ on a Grassmann manifold, the neighborhood around $\text{span}(\mathbf{G})$ is $\text{span}\{R(\mathbf{G} + \delta\mathbf{M})\}$. By Lemma 8 in (Manton, 2002), $\mathbf{M}$ can be decomposed as $\mathbf{M} = \mathbf{GA}+\mathbf{G}_0\mathbf{B}+\mathbf{GC}$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is skew symmetric, $\mathbf{B} \in \mathbb{R}^{(p-d)\times d}$, and $\mathbf{C} \in \mathbb{R}^{d\times d}$ is symmetric. Notice that $\mathbf{GA} + \mathbf{G}_0\mathbf{B} \in T_{\mathbf{G}}(p, d)$. Chen, Zou and Cook (2010) showed that $\text{span}\{R(\mathbf{G}+\delta\mathbf{M})\}$ depends on $\mathbf{M}$ only through $\mathbf{G}_0\mathbf{B}$, so the neighborhood around $\text{span}(\mathbf{G})$ on a Grassmann manifold only depends on the tangent space.

Now we develop Taylor expansion of a differentiable function on a manifold, which is essential to our techniques for handling a general objective function. In preparation, we first explain directional derivative. Let

$f(\mathbf{G}) : \mathbb{R}^{p\times d} \to \mathbb{R}$ be a differentiable function, and let $\mathbf{M} \in \mathbb{R}^{p\times d}$ be an arbitrary matrix. The directional derivative of $f(\mathbf{G})$ in the direction $\mathbf{M}$ evaluated at $\mathbf{G}$ is

$$\overset{\rightarrow\mathbf{M}}{df}(\mathbf{G}) = \frac{d}{d\delta}\bigg|_{\delta=0} f(\mathbf{G}+\delta\mathbf{M}).$$

In other words, the directional derivative $\overset{\rightarrow\mathbf{M}}{df}(\mathbf{G})$ can be thought of the slope of $f(\mathbf{G})$ along the "line" $\{f(\mathbf{G}+\delta\mathbf{M}) : \delta \in \mathbb{R}\}$ at $\delta = 0$ (Dattorro, 2016). Similarly, the second directional derivative of $f(\mathbf{G})$ in the direction $\mathbf{M}$ evaluated at $\mathbf{G}$ is

$$\overset{\rightarrow\mathbf{M}}{df^2}(\mathbf{G}) = \frac{d^2}{d\delta^2}\bigg|_{\delta=0} f(\mathbf{G}+\delta\mathbf{M}).$$

By Dattorro (2016), if f is second-order differentiable, then for some open interval of δ, f has the following second-order Taylor series expansion

$$f(\mathbf{G}+\delta\mathbf{M}) = f(\mathbf{G}) + \delta\,\overset{\rightarrow\mathbf{M}}{df}(\mathbf{G}) + \frac{1}{2!}\delta^2\,\overset{\rightarrow\mathbf{M}}{df^2}(\mathbf{G}) + o(\delta^2).$$

Since a neighborhood on a Stiefel manifold can be written as

$$R(\mathbf{G}+\delta\mathbf{M}) = \mathbf{G}+\delta\mathbf{M} - \frac{1}{2}\delta^2\mathbf{G}\mathbf{M}^T\mathbf{M} + o(\delta^2)$$

(Manton, 2002, Proposition 12). Let $\mathbf{M}^* = \mathbf{M} - \frac{1}{2}\delta\mathbf{G}\mathbf{M}^T\mathbf{M} + o(\delta)$, then for a second-order differentiable function f defined on a Stiefel manifold, we have the following Taylor expansion

$$f\{R(\mathbf{G}+\delta\mathbf{M})\} = f(\mathbf{G}) + \delta\,\overset{\rightarrow\mathbf{M}^*}{df}(\mathbf{G}) + \frac{1}{2}\delta^2\,\overset{\rightarrow\mathbf{M}^*}{df^2}(\mathbf{G}) + o(\delta^2).$$

Properties of oracle estimators. We first define the oracle model for a standard generalized linear model and present the asymptotic variance of its estimators.

In the generalized linear model (12), suppose we know in advance about which predictors are active and which are inactive, the oracle model is defined as

$$\log(f(Y|\theta)) = y\theta - b(\theta) + c(y) \quad \theta(\alpha,\boldsymbol{\beta}_{\mathcal{A}}) = (b')^{-1}\{g^{-1}(\zeta)\}, \quad \zeta = \alpha + \boldsymbol{\beta}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}. \tag{24}$$

Then α and $\boldsymbol{\beta}_{\mathcal{A}}$ can be estimated by fitting a generalized linear model of Y on $\mathbf{X}_{\mathcal{A}}$, and the Fisher information matrix for the parameters $(\alpha, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$ is

$$\mathrm{E}_{\mathbf{X},Y}\left(-\mathcal{D}''(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T \end{pmatrix}\right),$$

where the expectation above is taken with respect to the joint distribution $(\mathbf{X}, Y)$. Define $\mu = b'(\theta)$. Then $\mu = g^{-1}(\zeta)$. The first derivative of $\mathcal{D}(\zeta)$ is

$$\begin{aligned}\mathcal{D}'(\zeta) &= \frac{d\mathcal{D}(\zeta)}{d\zeta} = \frac{d\mathcal{D}(\zeta)}{d\theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\zeta} \\ &= \mathcal{C}'(\theta)\frac{1}{b''(\theta)}\frac{1}{g'(\mu)} \\ &= \{Y - b'(\theta)\}\frac{1}{b''(\theta)}\frac{1}{g'(\mu)} \\ &= (Y-\mu)\frac{1}{b''(\theta)}\frac{1}{g'(\mu)}.\end{aligned}$$

The second derivative of $\mathcal{D}(\zeta)$ is

$$\begin{aligned}\mathcal{D}''(\zeta) &= \frac{d^2\mathcal{D}(\zeta)}{d\zeta^2} = \frac{d\mathcal{D}'(\zeta)}{d\mu}\frac{d\mu}{d\zeta} \\ &= \left[-\frac{1}{b''(\theta)g'(\mu)} + (Y-\mu)\frac{d}{d\mu}\left(\frac{1}{b''(\theta)g'(\mu)}\right)\right]\frac{1}{g'(\mu)}.\end{aligned}$$

Then the Fisher information matrix for the parameters $(\alpha, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$ is

$$\begin{aligned} & \mathrm{E}_{\mathbf{X},Y}\left(-\mathcal{D}''(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\end{pmatrix}\right) = \mathrm{E}_{\mathbf{X}}\left[\mathrm{E}_{Y|\mathbf{X}}\left(-\mathcal{D}''(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\end{pmatrix}\right)\right] \\ = {} & \mathrm{E}_{\mathbf{X}}\left[\frac{1}{b''(\theta)g'(\mu)}\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\end{pmatrix}\right] \\ = {} & \mathrm{E}_{\mathbf{X}}\left[\frac{1}{b''[(b')^{-1}\{g^{-1}(\zeta)\}][g'\{g^{-1}(\zeta)\}]^2}\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\end{pmatrix}\right].\end{aligned}$$

The second equality is because $\mathrm{E}_{Y|\mathbf{X}}(Y-\mu) = 0$. Let

$$I(\zeta) = \frac{1}{b''(\theta)g'(\mu)} = \frac{1}{b''[(b')^{-1}\{g^{-1}(\zeta)\}][g'\{g^{-1}(\zeta)\}]^2}.$$

Then

$$\mathcal{D}''(\zeta) = -I(\zeta) + \frac{Y-\mu}{g'(\mu)}\frac{d}{d\mu}\left(\frac{1}{b''(\theta)g'(\mu)}\right).$$

From now on, when it does not raise confusion, we omit the subscript $\mathbf{X}$ in the expectation to make the notation simpler. So the Fisher information matrix for the parameters $(\alpha, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$ is denoted as

$$\mathrm{E}\left[I(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\end{pmatrix}\right].$$

Let $W(\zeta) = I(\zeta)/\mathrm{E}\{I(\zeta)\}$. Then

$$\begin{aligned}
&\mathrm{E}\left[I(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T \end{pmatrix}\right] = \mathrm{E}\{I(\zeta)\}E\left[W(\zeta)\begin{pmatrix} 1 & \mathbf{X}_{\mathcal{A}}^T \\ \mathbf{X}_{\mathcal{A}} & \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T \end{pmatrix}\right] \\
= \; &\mathrm{E}\{I(\zeta)\}\begin{bmatrix} \mathrm{E}\{W(\zeta)\} & \mathrm{E}\{W(\zeta)\mathbf{X}_{\mathcal{A}}^T\} \\ \mathrm{E}\{W(\zeta)\mathbf{X}_{\mathcal{A}}\} & \mathrm{E}\{W(\zeta)\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\} \end{bmatrix} \\
= \; &\mathrm{E}\{I(\zeta)\}\begin{pmatrix} 1 & \mathrm{E}(W\mathbf{X}_{\mathcal{A}}^T) \\ \mathrm{E}(W\mathbf{X}_{\mathcal{A}}) & \mathrm{E}(W\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T) \end{pmatrix}.
\end{aligned}$$

To facilitate the calculations in subsequent derivations, we transform the original parameters to orthogonal parameters (Huzurbazar, 1956; Cox and Reid, 1987). Let $a = \alpha + \boldsymbol{\beta}_{\mathcal{A}}^T\mathrm{E}(W\mathbf{X}_{\mathcal{A}})$, then $\zeta = \alpha + \boldsymbol{\beta}_{\mathcal{A}}^T\mathbf{X} = a + \boldsymbol{\beta}_{\mathcal{A}}^T(\mathbf{X}_{\mathcal{A}} - \mathrm{E}(W\mathbf{X}_{\mathcal{A}}))$. After transforming the original parameter $(\alpha, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$ to $(a, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$, the Fisher information matrix for the new parameterization $(a, \boldsymbol{\beta}_{\mathcal{A}}^T)$ is

$$\begin{aligned}
&\mathrm{E}\{I(\zeta)\}\begin{pmatrix} 1 & 0 \\ -\mathrm{E}(W\mathbf{X}_{\mathcal{A}}) & \mathbf{I}_{p_{\mathcal{A}}} \end{pmatrix}\begin{pmatrix} 1 & \mathrm{E}(W\mathbf{X}_{\mathcal{A}}^T) \\ \mathrm{E}(W\mathbf{X}_{\mathcal{A}}) & \mathrm{E}(W\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T) \end{pmatrix}\begin{pmatrix} 1 & -\mathrm{E}(W\mathbf{X}_{\mathcal{A}}^T) \\ 0 & \mathbf{I}_{p_{\mathcal{A}}} \end{pmatrix} \\
= \; &\mathrm{E}\{I(\zeta)\}\begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}(W)}} \end{pmatrix},
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}(W)}} = \mathrm{E}\{W[\mathbf{X}_{\mathcal{A}} - \mathrm{E}(W\mathbf{X}_{\mathcal{A}})][\mathbf{X}_{\mathcal{A}} - \mathrm{E}(W\mathbf{X}_{\mathcal{A}})]^T\}$. Therefore the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{A}}$ has asymptotic distribution

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})\} \xrightarrow{d} N(0, \mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}),$$

where $\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}} = [\mathrm{E}\{I(\zeta)\}\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}(W)}}]^{-1}$.

Based on the discussion, we have the following Lemma.

Lemma 2 *Assume the oracle model* (24) *holds, and* $\mathbf{X}$ *follows a normal distribution with mean* $\boldsymbol{\mu}_{\mathbf{X}}$ *and covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{X}}$*. Then the Fisher information matrix for* $(a, \boldsymbol{\beta}_{\mathcal{A}}^T, \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})^T)^T$ *is*

$$\mathbf{J} = \begin{pmatrix} \mathrm{E}\{I(\zeta)\} & 0 & 0 \\ 0 & \mathrm{E}\{I(\zeta)\}\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}(W)}} & 0 \\ 0 & 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p \end{pmatrix}.$$

The parameter $\boldsymbol{\Sigma}_{\mathbf{X}}$ is asymptotically independent of the other parameters because the distribution of $\mathbf{X}$ is ancillary in the estimation and its parameters are estimated independently.

Now we define the oracle envelope model under the context of generalized linear regression and discuss the properties of its estimators. Under the envelope model (13), suppose we have the information of which predictors are active and which predictors are inactive, we define the oracle envelope model as

(25)
$$\begin{aligned}\log(f(Y \mid \theta)) =& Y\theta - b(\theta) + c(Y), \quad \theta(\zeta) = (b')^{-1}\{g^{-1}(\zeta)\},\\ \zeta(\alpha, \boldsymbol{\Gamma}, \boldsymbol{\eta}) =& \alpha + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\mathbf{X}, \quad \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \quad \boldsymbol{\Gamma} = \begin{pmatrix}\boldsymbol{\Gamma}_{\mathcal{A}}\\ 0\end{pmatrix}.\end{aligned}$$

The oracle envelope model seems to have the same form as the sparse envelope model (15). The difference is that the oracle envelope model (25) has the additional information of which predictors are active and which predictors are inactive, while under the sparse envelope model (15), this information is unknown and needs to be estimated.

Denote the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{A}}$ under the oracle envelope model as $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$. The next lemma gives the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$ under normality. In preparation, under the oracle envelope model, $\boldsymbol{\Gamma}_0$ can have the block diagonal structure

$$\boldsymbol{\Gamma}_0 = \begin{pmatrix}\boldsymbol{\Gamma}_{\mathcal{A},0} & 0\\ 0 & \mathbf{I}_{p-p_{\mathcal{A}}}\end{pmatrix}$$

where $\boldsymbol{\Gamma}_{\mathcal{A},0} \in \mathbb{R}^{p_{\mathcal{A}}\times(p_{\mathcal{A}}-d)}$ is the completion of $\boldsymbol{\Gamma}_{\mathcal{A}}$. As $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0$, when $\boldsymbol{\Gamma}_0$ has this block diagonal structure, we denote the corresponding $\boldsymbol{\Omega}_0$ as $\widetilde{\boldsymbol{\Omega}}_0$, and $\widetilde{\boldsymbol{\Omega}}_0$ can be partitioned accordingly into

$$\widetilde{\boldsymbol{\Omega}}_0 = \begin{pmatrix}\widetilde{\boldsymbol{\Omega}}_{\mathcal{A},0} & \widetilde{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{I},0}\\ \widetilde{\boldsymbol{\Omega}}_{\mathcal{I}\mathcal{A},0} & \widetilde{\boldsymbol{\Omega}}_{\mathcal{I},0}\end{pmatrix}.$$

Lemma 3 *Assume the oracle envelope model (25) holds and the distribution of $\mathbf{X}$ has finite fourth moments. Then $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O} - \boldsymbol{\beta}_{\mathcal{A}})$ converges to a normal distribution with mean 0. If we further assume that $\mathbf{X}$ follows a normal distribution with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$. Then the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$ has a closed form, i.e.*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \mathbf{V}),$$

where $\mathbf{V} = \mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}}\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}} + (\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_{\mathcal{A},0}^T)$, $\mathbf{T} = (\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_{\mathcal{A},0}^T)\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}(\boldsymbol{\eta}^T\otimes \boldsymbol{\Gamma}_{\mathcal{A},0}) + \boldsymbol{\Omega}\otimes\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \boldsymbol{\Omega}^{-1}\otimes\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} - 2\mathbf{I}_d\otimes\mathbf{I}_{p_{\mathcal{A}}-d}$, and $\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}} = [\mathrm{E}\{I(\zeta)\}\cdot \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{A}}(W)}]^{-1}$.

Proof: We use Shapiro (1986) to derive the asymptotic variance of the oracle envelope model estimator of $(a, \boldsymbol{\beta}_{\mathcal{A}}^T, \text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})^T)^T$. The parameters of the oracle envelope model are

$$\begin{aligned}\boldsymbol{\phi} &= \{a, \boldsymbol{\eta}^T, \text{vec}^T(\boldsymbol{\Gamma}_{\mathcal{A}}), \text{vech}^T(\boldsymbol{\Omega}), \text{vech}^T(\boldsymbol{\Omega}_0)\}^T \\ &\equiv \{\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T, \boldsymbol{\phi}_3^T, \boldsymbol{\phi}_4^T, \boldsymbol{\phi}_5^T\}^T,\end{aligned}$$

and the parameters under model (24) are

$$\begin{aligned}\mathbf{h}(\boldsymbol{\phi}) &= \{a, \boldsymbol{\beta}_{\mathcal{A}}^T, \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T \\ &\equiv (h_1^T(\boldsymbol{\phi}), \mathbf{h}_2^T(\boldsymbol{\phi}), \mathbf{h}_3^T(\boldsymbol{\phi}))^T.\end{aligned}$$

According to Proposition 4.1 in Shapiro (1986), the asymptotic covariance of $\mathbf{h}(\widehat{\boldsymbol{\phi}})$ is $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T$, where $\mathbf{J}$ is the Fisher information matrix of $\mathbf{h}(\boldsymbol{\phi})$ under the standard model (24) and is given in Lemma 2, and $\mathbf{H} = \partial\mathbf{h}/\partial\boldsymbol{\phi}^T$ is the gradient matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Gamma}_{\mathcal{A}} & \boldsymbol{\eta}^T \otimes \mathbf{I}_q & 0 & 0 \\ 0 & 0 & \mathbf{H}_{32} & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_d & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-d} \end{pmatrix}.$$

The expression of $\mathbf{H}_{32}$ is $\mathbf{H}_{32} = 2\mathbf{C}_p[(\boldsymbol{\Gamma}\boldsymbol{\Omega}) \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes (\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)]\mathbf{L}$, where $\mathbf{L} = \mathbf{I}_d \otimes \mathbf{I}_{p,p_{\mathcal{A}}}$ and $\mathbf{I}_{p,p_{\mathcal{A}}} \in \mathbb{R}^{p \times p_{\mathcal{A}}}$ contains the first $p_{\mathcal{A}}$ columns of $\mathbf{I}_p$. After some straightforward algebra, we find that the asymptotic variance of $\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O})$ is $\mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}}\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}\mathbf{P}_{\boldsymbol{\Gamma}_{\mathcal{A}}} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)$, where $\mathbf{T} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}^T)\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_{\mathcal{A},0}) + \boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \boldsymbol{\Omega}^{-1} \otimes \widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}$.

Now we return to the proof of Theorem 6.

Proof of Theorem 6, part (a). As $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$, to prove the $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\beta}}$, we only need to prove the $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\eta}}$. We use an iterative algorithm to estimate $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\Gamma}}$, and the starting value is $\sqrt{n}$-consistent. So we will show that at each iteration, the estimators are $\sqrt{n}$-consistent. Then the estimators we obtain at convergence are $\sqrt{n}$-consistent. Suppose we have the $\sqrt{n}$-consistent estimator $\widetilde{\boldsymbol{\Gamma}}_{(k)}$ at the kth step $(k = 1, 2 \ldots)$, we can get $\widetilde{\alpha}_{(k)}$ and $\widetilde{\boldsymbol{\eta}}_{(k)}$ by using scoring method McCullagh and Nelder (1989) to fit the generalized linear model of Y on $\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{X}$. Then $\widetilde{\alpha}_{(k)}$ and $\widetilde{\boldsymbol{\eta}}_{(k)}$ are $\sqrt{n}$-consistent. The rest part of the proof shows that given the $\sqrt{n}$-consistent estimators $\widetilde{\alpha}_{(k)}$ and $\widetilde{\boldsymbol{\beta}}_{(k)} = \widetilde{\boldsymbol{\Gamma}}_{(k)}\widetilde{\boldsymbol{\eta}}_{(k)}$, the estimator $\widetilde{\boldsymbol{\Gamma}}_{(k+1)}$ obtained by minimizing the objective function (16) is also $\sqrt{n}$-consistent.

Denote the objective function in (16) by $f_{\text{obj}}(\boldsymbol{\Gamma})$. Let $\widetilde{\boldsymbol{\Gamma}}_{(k)0}$ be a completion of $\widetilde{\boldsymbol{\Gamma}}_{(k)}$. Since f_{obj} depends on $\boldsymbol{\Gamma}$ only through its span, it is sufficient to show that for any small $\epsilon > 0$, there exists a sufficiently large constant C, such that

(26)

$$\lim_{n\to\infty} P\left(\inf_{\substack{\mathbf{Z}\in T_{\widetilde{\mathbf{\Gamma}}_{(k)}}(p,d), \\ \mathbf{Z}=\widetilde{\mathbf{\Gamma}}_{(k)}\mathbf{A}+\widetilde{\mathbf{\Gamma}}_{(k)0}\mathbf{B},\, \|\mathbf{B}\|_F=C}} f_{\text{obj}}(R(\widetilde{\mathbf{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})) > f_{\text{obj}}(\widetilde{\mathbf{\Gamma}}_{(k)})\right) > 1-\epsilon.$$

We write f_{obj} into four parts

$$\begin{aligned} f_{\text{obj}}(\mathbf{\Gamma}) &= -\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}(\mu_{V_{(k)}(W_{(k)})} + \widetilde{\boldsymbol{\eta}}_{(k)}^T\mathbf{\Gamma}^T(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})) \\ &\quad + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| + \sum_{i=1}^{p}\lambda_i\|\boldsymbol{\gamma}_i\|_2 \\ &\equiv f_1(\mathbf{\Gamma}) + f_2(\mathbf{\Gamma}) + f_3(\mathbf{\Gamma}) + f_4(\mathbf{\Gamma}), \end{aligned}$$

where $W_{(k)}$ and $V_{(k)}$ denote the weight and pseudo-response at kth iteration, and $\widetilde{\boldsymbol{\eta}}_{(k)}$ is the estimator of $\boldsymbol{\eta}$ at kth iteration. Let $W_{(k),i}$ and $V_{(k),i}$ be the weight and pseudo-response for the ith observation at kth iteration, $\mu_{\mathbf{X}(W_{(k)})} = \sum_{i=1}^n W_{(k),i}\mathbf{X}_i/n$, $\mathbf{S}_{\mathbf{X}(W_{(k)})} = \sum_{i=1}^n W_{(k),i}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T/n$, $\mathbf{S}_{\mathbf{X}V_{(k)}(W_{(k)})} = \sum_{i=1}^n W_{(k),i}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})(V_{(k),i} - \mu_{V_{(k)}(W_{(k)})})/n$. With fixed $\widetilde{\mathbf{\Gamma}}_{(k)}$, $\widetilde{\boldsymbol{\eta}}_{(k)}$ is the estimated coefficients from the generalized linear model of Y on $\widetilde{\mathbf{\Gamma}}_{(k)}^T\mathbf{X}$. Then $\widetilde{\boldsymbol{\eta}}_{(k)} = (\widetilde{\mathbf{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\mathbf{\Gamma}}_{(k)})^{-1}\widetilde{\mathbf{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}V_{(k)}(W_{(k)})}$, and it is treated as a function of $\widetilde{\mathbf{\Gamma}}_{(k)}$ in the objective function. Let $\mu_{V_{(k)}(W_{(k)})} = \sum_{i=1}^n W_{(k),i}V_{(k),i}/n$ and $\widetilde{\zeta}_{(k),i}(\mathbf{\Gamma}) = \mu_{V_{(k)}(W_{(k)})} + \widetilde{\boldsymbol{\eta}}_{(k)}^T\mathbf{\Gamma}^T(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})$. Then we update the weights and pseudo-responses for the next iteration $W_{(k+1),i} = nI(\widetilde{\zeta}_{(k),i})/\sum_{i=1}^n I(\widetilde{\zeta}_{(k),i})$ and $V_{(k+1),i} = \widetilde{\zeta}_{(k),i} + \{Y_i - g^{-1}(\widetilde{\zeta}_{(k),i})\}/W_{(k),i}$.

Expand $f_1\{R(\widetilde{\mathbf{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\}$, we have

$$\begin{aligned} f_1\{R(\widetilde{\mathbf{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} =& f_1(\widetilde{\mathbf{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z}^*) \\ =& f_1(\widetilde{\mathbf{\Gamma}}_{(k)}) + n^{-1/2}\overset{\to\mathbf{Z}^*}{df_1}(\widetilde{\mathbf{\Gamma}}_{(k)}) + \frac{1}{2}n^{-1}\overset{\to\mathbf{Z}^*}{df_1^2}(\widetilde{\mathbf{\Gamma}}_{(k)}) + o_p(n^{-1}), \end{aligned}$$

where $\mathbf{Z}^* = \mathbf{Z} - (1/2)n^{-1/2}\widetilde{\mathbf{\Gamma}}_{(k)}\mathbf{Z}^T\mathbf{Z} + o_p(n^{-1/2})$. The first directional derivative is

$$
\begin{aligned}
\overset{\rightarrow \mathbf{Z}^*}{df_1}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= \operatorname{tr}\left\{\frac{df_1(\boldsymbol{\Gamma})}{d\boldsymbol{\Gamma}}^T \mathbf{Z}^*\right\}\Big|_{\boldsymbol{\Gamma}=\widetilde{\boldsymbol{\Gamma}}_{(k)}} \\
&= -\frac{2}{n}\operatorname{tr}\left\{\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)\Big[\widetilde{\boldsymbol{\eta}}_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\right. \\
&\quad +(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}(W^{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\mathbf{S}^T_{\mathbf{X}\widetilde{V}_{(k)}(W^{(k)})} \\
&\quad -\widetilde{\boldsymbol{\eta}}_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W^{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\boldsymbol{\Gamma}^T_{(k)}\mathbf{S}_{\mathbf{X}(W^{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}(W^{(k)})} \\
&\quad \left. -(\boldsymbol{\Gamma}^T_{(k)}\mathbf{S}_{\mathbf{X}(W^{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\widetilde{\boldsymbol{\eta}}^T_{(k)}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}(W^{(k)})}\Big]\mathbf{Z}^*\right\} \\
&= -\frac{2}{n}\operatorname{tr}\left\{\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)\widetilde{\boldsymbol{\eta}}_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*\right\} \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*\widetilde{\boldsymbol{\eta}}_{(k)} \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T[(\widetilde{\boldsymbol{\Gamma}}_{(k)}\mathbf{A} + \widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}) \\
&\quad -\frac{1}{2}n^{-1/2}\widetilde{\boldsymbol{\Gamma}}_{(k)}\mathbf{Z}^T\mathbf{Z}]\widetilde{\boldsymbol{\eta}}_{(k)} + O_p(n^{-1}) \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\widetilde{\boldsymbol{\eta}}_{(k)} + O_p(n^{-1}).
\end{aligned}
$$

The third and last equality are because that $\widetilde{\boldsymbol{\eta}}_{(k)}$ is the maximum likelihood estimator of the coefficients from the generalized linear model of Y on $\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{X}$, with $\widetilde{\boldsymbol{\Gamma}}^T_{(k)}$ being fixed. Then $\sum_{i=1}^n \mathcal{D}'(\widetilde{\zeta}_{(k),i})\widetilde{\boldsymbol{\Gamma}}^T_{(k)}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) = 0$. Taking transpose on both sides, we have

$$
\sum_{i=1}^n \mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k)} = 0.
$$

Let $\widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}$ and $\widehat{\alpha}_{\mathrm{GLM}}$ denote the maximum likelihood estimator of $\boldsymbol{\beta}$ and α under the standard generalized linear model (12). Then $\widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}$ and $\widehat{\alpha}_{\mathrm{GLM}}$ are $\sqrt{n}$-consistent estimator of α and $\boldsymbol{\beta}$. Let $\widetilde{\alpha}_{(k)} = \mu_{V_{(k)}(W_{(k)})} - \widetilde{\boldsymbol{\eta}}^T_{(k)}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}$ and $\widetilde{\boldsymbol{\beta}}_{(k)} = \widetilde{\boldsymbol{\Gamma}}_{(k)}\widetilde{\boldsymbol{\eta}}_{(k)}$ be the estimator of α and $\boldsymbol{\beta}$ at kth iteration. Then $\widetilde{\alpha}_{(k)}$ and $\widetilde{\boldsymbol{\beta}}_{(k)}$ are also $\sqrt{n}$-consistent estimator of α and $\boldsymbol{\beta}$. Then there exists U_{1n}

and $\mathbf{U}_{2n}$ such that

$$\widetilde{\alpha}_{(k)} - \widehat{\alpha}_{\mathrm{GLM}} = n^{-1/2}U_{1n} + O_p(n^{-1}),$$

$$\widetilde{\boldsymbol{\beta}}_{(k)} - \widehat{\boldsymbol{\beta}}_{\mathrm{GLM}} = n^{-1/2}\mathbf{U}_{2n} + O_p(n^{-1}).$$

Notice that $\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ can be written as $\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) = \widetilde{\alpha}_{(k)} + \widetilde{\boldsymbol{\beta}}_{(k)}^T\mathbf{X}_i$. Let $\widehat{\zeta}_{\mathrm{GLM},i} = \widehat{\alpha}_{\mathrm{GLM}} + \widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}^T\mathbf{X}_i$. Then we have

$$\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) - \widehat{\zeta}_{\mathrm{GLM},i} = n^{-1/2}(U_{1n} + \mathbf{U}_{2n}^T\mathbf{X}_i) + O_p(n^{-1}).$$

Expand $\mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)$, we get

$$\mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right) = \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i}) + n^{-1/2}\mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})(U_{1n} + \mathbf{U}_{2n}^T\mathbf{X}_i) + O_p(n^{-1}). \tag{27}$$

We also have $\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i}) = 0$ and $\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})\mathbf{X}_i = 0$ because the derivative of $\mathcal{D}(\zeta(\alpha,\boldsymbol{\beta}))$ at $\widehat{\alpha}_{\mathrm{GLM}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}$ is 0. So

$$\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) = 0. \tag{28}$$

Let $\widehat{W}$ be the weight at convergence under the standard model, then $\boldsymbol{\mu}_{\mathbf{X}(\widehat{W})}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\mu}_{\mathbf{X}(W)}$. Let $\mu_i = g^{-1}(\widehat{\zeta}_{\mathrm{GLM},i})$. Then

$$\begin{aligned}
(29)\quad & \frac{1}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) \\
= \quad & -\frac{1}{n}\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) \\
& + \frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu_i}{g'(\mu_i)}\frac{d}{d\mu_i}\Big(\frac{1}{b''((b')^{-1}(\mu_i))g'(\mu_i)}\Big)(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) \\
= \quad & -\frac{1}{n}\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) + O_p(n^{-1/2}) \\
= \quad & -\frac{1}{n}\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}(\widehat{W})}) - \frac{1}{n}\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})(\boldsymbol{\mu}_{\mathbf{X}(\widehat{W})} - \boldsymbol{\mu}_{\mathbf{X}(W_{(k)})}) \\
= \quad & O_p(n^{-1/2}).
\end{aligned}$$

In the second equality, we use central limited theorem and the expectation of the second summand is zero. More specifically, the second summand converges in distribution with rate of $\sqrt{n}$ to

$$\mathrm{E}_{\mathbf{X},Y}\left[\frac{Y-\mu}{g'(\mu)}\frac{d}{d\mu}\left(\frac{1}{b''(\theta)g'(\mu)}\right)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}(W)})\right]=0.$$

In the last equality, $\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(\widehat{W})})=0$ is because of $I(\widehat{\zeta}_{\mathrm{GLM},i})=c\widehat{W}_i$, where $\widehat{W}_i$'s are the sample weights, and $c=\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})/n$. And the second term has order $O_p(n^{-1/2})$ since $\sum_{i=1}^n I(\widehat{\zeta}_{\mathrm{GLM},i})/n$ converges to $\mathrm{E}\{I(\zeta)\}$ in distribution.

Let $\mathbf{V}_{\boldsymbol{\beta}}$ be the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}$. Using (29), we have

$$\begin{aligned}(30)\qquad & \frac{1}{n}\sum_{i=1}^n \mathcal{D}''(\hat{\zeta}_{\mathrm{GLM},i})\mathbf{X}_i^T(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\\ = \quad & \frac{1}{n}\sum_{i=1}^n \mathcal{D}''(\hat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})+O_p(n^{-1/2})\\ = \quad & -\mathbf{V}_{\boldsymbol{\beta}}^{-1}+O_p(n^{-1/2}).\end{aligned}$$

The last equality is because $\sum_{i=1}^n(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T/n$ converges in distribution to $\mathrm{E}_{\mathbf{X},Y}[\mathcal{D}''(\zeta)\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}^T]=-\mathrm{E}_{\mathbf{X}}[I(\zeta)\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}^T]=-\mathrm{E}\{I(\zeta)\}\mathrm{E}[W\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}\{\mathbf{X}-\mathrm{E}(W\mathbf{X})\}^T]=-\mathrm{E}\{I(\zeta)\}\boldsymbol{\Sigma}_{\mathbf{X}(W)}=-\mathbf{V}_{\boldsymbol{\beta}}^{-1}$ with rate of $\sqrt{n}$.

Using (27), (28), (29) and (30), the first directional derivative of f_1 is

$$\begin{aligned}\overset{\rightarrow\mathbf{Z}^*}{df_1}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= -\frac{2}{n}\operatorname{tr}\Big\{\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\widetilde{\boldsymbol{\eta}}_{(k)}\Big\}\\ &\quad -\frac{2}{n}\operatorname{tr}\Big\{n^{-1/2}\sum_{i=1}^n\big[\mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})U_{1n}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\\ &\quad +\mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})\mathbf{U}_{2n}^T\mathbf{X}_i(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\big]\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\widetilde{\boldsymbol{\eta}}_{(k)}\Big\}+O_p(n^{-1})\\ &= -\frac{2}{n}n^{-1/2}\sum_{i=1}^n\mathbf{U}_{2n}^T(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\mathcal{D}''(\widehat{\zeta}_{\mathrm{GLM},i})(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\widetilde{\boldsymbol{\eta}}_{(k)}\\ &\quad +O_p(n^{-1})\\ &= 2n^{-1/2}\mathbf{U}_{2n}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta}+O_p(n^{-1})\\ &= 2n^{-1/2}\operatorname{tr}(\boldsymbol{\eta}\mathbf{U}_{2n}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\mathbf{B})+O_p(n^{-1})\\ &\geq -2n^{-1/2}\|\mathbf{B}\|_F\|\boldsymbol{\eta}\mathbf{U}_{2n}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\|_F+O_p(n^{-1}).\end{aligned}$$

The second directional derivative of f_1 is

$$\begin{aligned}
&\overset{\rightarrow\mathbf{Z}^*}{df_1^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) = \operatorname{tr}\left\{\left[\frac{d\,\overset{\rightarrow\mathbf{Z}^*}{df_1}(\boldsymbol{\Gamma})}{d\boldsymbol{\Gamma}}\right]^T\mathbf{Z}^*\right\}\bigg|_{\boldsymbol{\Gamma}=\widetilde{\boldsymbol{\Gamma}}_{(k)}}\\
= \quad &\frac{2}{n}\sum_{i=1}^n \operatorname{tr}\Bigg\{-\mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)\Big[\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\\
&-\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\\
&-\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\mathbf{Z}^{*T}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\Big]^T\mathbf{Z}^*\\
&+\widetilde{\boldsymbol{\eta}}_{(k)}^T\mathbf{Z}^{*T}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\left(-\mathcal{D}''(\widetilde{\zeta}_{(k),i})\right)\Big[(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*\widetilde{\boldsymbol{\eta}}_{(k)}\\
&+(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\mathbf{Z}^{*T}\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}\\
&-(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\mathbf{Z}^*\widetilde{\boldsymbol{\eta}}_{(k)}\\
&-(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\mathbf{Z}^{*T}\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)}\widetilde{\boldsymbol{\eta}}_{(k)}\Big]\Bigg\}.
\end{aligned}$$

We focus on the first three terms, which all contains $-\mathcal{D}'\left(\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})\right)$. Since $\widetilde{\zeta}_{(k),i}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ is $\sqrt{n}$-consistent, using (28), the first three terms can be written as

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^n \operatorname{tr}\Bigg\{-\mathcal{D}'(\hat{\zeta}_{\mathrm{GLM},i})\Big[\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\\
&-\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})^T\mathbf{Z}^*(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\\
&-\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\mathbf{Z}^{*T}(\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}(W_{(k)})})\mathbf{S}_{\mathbf{X}\widetilde{V}_{(k)}(W_{(k)})}^T\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\Big]^T\mathbf{Z}^*\Bigg\}\\
&+o_p(1)\\
&=o_p(1).
\end{aligned}$$

Then

$$\begin{aligned}
\overset{\rightarrow\mathbf{Z}^*}{df_1^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= 2\big\{\boldsymbol{\eta}^T\mathbf{Z}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{Z}\boldsymbol{\eta} + \mathrm{E}\{-\mathcal{D}''(\zeta)\}\boldsymbol{\eta}^T\mathbf{Z}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}_{\mathbf{X}V(W)}\\
&\quad -\boldsymbol{\eta}^T\mathbf{Z}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{Z}\boldsymbol{\eta}\\
&\quad -\mathrm{E}\{-\mathcal{D}''(\zeta)\}\boldsymbol{\eta}^T\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\mathbf{Z}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}V(W)}\big\}\\
&\quad +o_p(1)\\
&= 2\big\{\boldsymbol{\eta}^T\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\eta} + \boldsymbol{\eta}^T\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta}\\
&\quad +\mathrm{E}\{-\mathcal{D}''(\zeta)\}\boldsymbol{\eta}^T\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\mathbf{A}^T\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}V(W)}\\
&\quad -\boldsymbol{\eta}^T\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\eta}\\
&\quad -\mathrm{E}\{-\mathcal{D}''(\zeta)\}\boldsymbol{\eta}^T\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\mathbf{A}^T\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}V(W)}\big\}\\
&\quad +o_p(1)\\
&= 2\boldsymbol{\eta}^T\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta} + o_p(1)\\
&= 2\mathrm{vec}(\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta})^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathrm{vec}(\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta}) + o_p(1)\\
&= 2\mathrm{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)]\mathrm{vec}(\mathbf{B}) + o_p(1).
\end{aligned}$$

The first equality uses the facts $\widetilde{\boldsymbol{\eta}}_{(k)} = (\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}(W_{(k)})}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}V_{(k)}(W_{(k)})}$ and $\mathbf{S}_{\mathbf{X}(W_{(k)})} = \boldsymbol{\Sigma}_{\mathbf{X}(W_{(k)})}+o_p(1) = \mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathrm{E}\{I(\zeta)\}^{-1}+o_p(1) = \mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathrm{E}\{-\mathcal{D}''(\zeta)\}^{-1}+o_p(1)$. The second equality uses the fact that $\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta}}\boldsymbol{\Gamma}_0 = 0$. This is because that when $\mathbf{X}$ follows an elliptically contoured distribution, $\mathrm{span}(\boldsymbol{\Gamma})$ also reduces $\mathbf{V}_{\boldsymbol{\beta}}$ (Cook and Zhang, 2015).

We substitute $\overset{\rightarrow\mathbf{Z}^*}{df_1}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ and $\overset{\rightarrow\mathbf{Z}^*}{df_1^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ into the expansion for f_1 and get

$$\begin{aligned}
& f_1\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_1(\widetilde{\boldsymbol{\Gamma}}_{(k)})\\
\geq\ & -2n^{-1}\|\mathbf{B}\|_F\|\boldsymbol{\eta}\mathbf{U}_{2n}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\|_F\\
& +n^{-1}\mathrm{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)]\mathrm{vec}(\mathbf{B}) + o_p(n^{-1}).
\end{aligned}$$

Now we expand $f_2\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\}$,

$$f_2\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)}+n^{-1/2}\mathbf{Z})\} = f_2(\widetilde{\boldsymbol{\Gamma}}_{(k)})+n^{-1/2}\overset{\rightarrow\mathbf{Z}^*}{df_2}(\widetilde{\boldsymbol{\Gamma}}_{(k)})+\frac{1}{2}n^{-1}\overset{\rightarrow\mathbf{Z}^*}{df_2^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)})+o_p(n^{-1}).$$

Let $\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ be a completion of $\widetilde{\boldsymbol{\Gamma}}_{(k)}$. Since $\widetilde{\boldsymbol{\Gamma}}_{(k)}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Gamma}$, and $\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ is a continuous function of $\widetilde{\boldsymbol{\Gamma}}_{(k)}$, then $\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Gamma}_0$. Because that $\mathbf{S}_{\mathbf{X}}$, $\widetilde{\boldsymbol{\Gamma}}_{(k)}$ and $\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ are $\sqrt{n}$-consistent estimators of $\boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$. Then $(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ is a $\sqrt{n}$-consistent estimator of $(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0$, and we have

$$(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k),0} = (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0 + n^{-1/2}\mathbf{T}_n + O_p(n^{-1}),$$

where $\mathbf{T}_n$ converges in distribution to a normal random matrix with mean 0. Then the first directional derivative of f_2 is

$$\begin{aligned}
\overset{\rightarrow\mathbf{Z}^*}{df_2}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= \operatorname{tr}\left\{\frac{df_2(\boldsymbol{\Gamma})}{d\boldsymbol{\Gamma}}^T\mathbf{Z}^*\right\}\Big|_{\boldsymbol{\Gamma}=\widetilde{\boldsymbol{\Gamma}}_{(k)}} = \operatorname{tr}\left\{2(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\mathbf{Z}^*\right\}\\
&= 2\operatorname{tr}\left\{(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\mathbf{Z}\right\} - n^{-1/2}\operatorname{tr}\left\{(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)}\mathbf{Z}^T\mathbf{Z}\right\}\\
&\quad + O_p(n^{-1})\\
&= 2\operatorname{tr}\left\{\mathbf{A} + (\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\right\} - n^{-1/2}\operatorname{tr}(\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}) + O_p(n^{-1})\\
&= 2\operatorname{tr}\left\{(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k),0}\mathbf{B}\right\} - n^{-1/2}\operatorname{tr}(\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}) + O_p(n^{-1})\\
&= 2\operatorname{tr}\left\{(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0\mathbf{B} + n^{-1/2}\mathbf{T}_n\mathbf{B}\right\} - n^{-1/2}\operatorname{tr}(\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}) + O_p(n^{-1})\\
&= 2n^{-1/2}\operatorname{tr}(\mathbf{T}_n\mathbf{B}) - n^{-1/2}\operatorname{tr}(\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}) + O_p(n^{-1}).
\end{aligned}$$

The third equality in $\overset{\rightarrow\mathbf{Z}^*}{df_2}(\boldsymbol{\Gamma})$ is because that $\mathbf{A} + \mathbf{A}^T = 0$ implies $\operatorname{tr}(\mathbf{A}) = 0$. The second directional derivative is

$$\begin{aligned}
\overset{\rightarrow\mathbf{Z}^*}{df_2^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= \operatorname{tr}\left(\left[\frac{d\operatorname{tr}\left\{2(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{Z}^*\right\}}{d\boldsymbol{\Gamma}}\right]^T\mathbf{Z}^*\right)\Bigg|_{\boldsymbol{\Gamma}=\widetilde{\boldsymbol{\Gamma}}_{(k)}}\\
&= 2\operatorname{tr}\Big[\big\{-\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)}(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\mathbf{Z}^* + \mathbf{Z}^{*T}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\\
&\quad + \mathbf{S}_{\mathbf{X}}\mathbf{Z}^*(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\big\}^T\mathbf{Z}^*\Big]\\
&= 2\operatorname{tr}\Big\{(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\mathbf{Z}^T\mathbf{S}_{\mathbf{X}}\mathbf{Z}\\
&\quad - (\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\mathbf{Z} + \mathbf{Z}^T\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})(\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}^T_{(k)}\mathbf{S}_{\mathbf{X}}\mathbf{Z}\Big\} + o_p(1)\\
&= 2\operatorname{tr}\Big[(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}(\boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Gamma}_0\mathbf{B})^T\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Gamma}_0\mathbf{B}) - (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\big\{\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\Gamma}\mathbf{A}\\
&\quad + \boldsymbol{\Gamma}_0\mathbf{B}) + (\boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Gamma}_0\mathbf{B})^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}\big\}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Gamma}_0\mathbf{B})\Big] + o_p(1)\\
&= 2\operatorname{tr}\left\{(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\mathbf{B}^T\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0\mathbf{B} - \mathbf{A}\mathbf{A}\right\} + o_p(1)\\
&= 2\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{B}^T\boldsymbol{\Omega}_0\mathbf{B} - \mathbf{A}\mathbf{A}) + o_p(1).
\end{aligned}$$

We substitute $\overset{\to\mathbf{Z}^*}{df_2}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ and $\overset{\to\mathbf{Z}^*}{df_2^2}(\widetilde{\boldsymbol{\Gamma}}_{(k)})$ into the expansion for $f_2\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\}$ and get

$$\begin{aligned} f_2\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_2(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= 2n^{-1}\operatorname{tr}(\mathbf{T}_n\mathbf{B}) - n^{-1}\operatorname{tr}(\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}) \\ &\quad + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{B}^T\boldsymbol{\Omega}_0\mathbf{B} - \mathbf{A}\mathbf{A}) + o_p(n^{-1}) \\ &\geq -2n^{-1}\|\mathbf{B}\|_F\|\mathbf{T}_n\|_F - n^{-1}\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \\ &\quad + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{B}^T\boldsymbol{\Omega}_0\mathbf{B}) + o_p(n^{-1}). \end{aligned}$$

Since that f_3 has similar structure as f_2, the derivation above can be applied to f_3, with $\mathbf{S}_{\mathbf{X}}$ replaced by $\mathbf{S}_{\mathbf{X}}^{-1}$. Since $(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k),0}$ is a $\sqrt{n}$-consistent estimator of $(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}_0$, then we have

$$(\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)})^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k)}^T\mathbf{S}_{\mathbf{X}}^{-1}\widetilde{\boldsymbol{\Gamma}}_{(k),0} = (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}_0 + n^{-1/2}\mathbf{T}_{2n} + O_p(n^{-1}),$$

where $\mathbf{T}_{2n}$ converges in distribution to a normal random matrix with mean 0. After some straightforward algebra, we have

$$\begin{aligned} f_3\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_3(\widetilde{\boldsymbol{\Gamma}}_{(k)}) &= 2n^{-1}\operatorname{tr}(\mathbf{T}_{2n}\mathbf{B}) - n^{-1}\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \\ &\quad + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}\mathbf{B}^T\boldsymbol{\Omega}_0^{-1}\mathbf{B}) + o_p(n^{-1}) \\ &\geq -2n^{-1}\|\mathbf{B}\|_F\|\mathbf{T}_{2n}\|_F - n^{-1}\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \\ &\quad + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}\mathbf{B}^T\boldsymbol{\Omega}_0^{-1}\mathbf{B}) + o_p(n^{-1}). \end{aligned}$$

With $f_4(\boldsymbol{\Gamma}) = \sum_{i=1}^p \lambda_i\|\boldsymbol{\gamma}_i\|_2$, we have

$$\begin{aligned} & f_4\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_4(\widetilde{\boldsymbol{\Gamma}}_{(k)}) \\ = {} & \sum_{i=1}^{p}\Big\{\lambda_i\|\mathbf{e}_i^T(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z}^*)\|_2 - \lambda_i\|\mathbf{e}_i^T\widetilde{\boldsymbol{\Gamma}}_{(k)}\|_2\Big\} \\ \geq {} & \sum_{i=1}^{p_{\mathcal{A}}}\Big\{\lambda_i\|\mathbf{e}_i^T(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z}^*)\|_2 - \lambda_i\|\mathbf{e}_i^T\widetilde{\boldsymbol{\Gamma}}_{(k)}\|_2\Big\} \\ \geq {} & -\frac{1}{2}p_{\mathcal{A}}n^{-1/2}\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\Big(\|\mathbf{e}_i^T\widetilde{\boldsymbol{\Gamma}}_{(k)}\|_2^{-1}\|\mathbf{e}_i^T\mathbf{Z}^*\|_2 + O_p(n^{-1/2})\Big) \\ = {} & -\frac{1}{2}p_{\mathcal{A}}n^{-1/2}\lambda_{\mathcal{A}}\max_{1\leq i\leq p_{\mathcal{A}}-d}\{\|\mathbf{e}_i^T\boldsymbol{\Gamma}\|_2^{-1}\|\mathbf{e}_i^T(\mathbf{Z} - \frac{1}{2}n^{-1/2}\boldsymbol{\Gamma}\mathbf{Z}^T\mathbf{Z})\|_2\} + o_p(n^{-1}). \end{aligned}$$

The third inequality is based on Taylor expansion at $\mathbf{e}_i^T\widetilde{\boldsymbol{\Gamma}}_{(k)}$. Because $\sqrt{n}\lambda_{\mathcal{A}} \to 0$, $f_4\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_4(\widetilde{\boldsymbol{\Gamma}}_{(k)}) = o_p(n^{-1})$.

Collecting all the results so far

$$\begin{aligned} & f_{\text{obj}}\{R(\widetilde{\boldsymbol{\Gamma}}_{(k)} + n^{-1/2}\mathbf{Z})\} - f_{\text{obj}}(\widetilde{\boldsymbol{\Gamma}}_{(k)}) \\ \geq\ & -2n^{-1}\|\mathbf{B}\|_F\|\boldsymbol{\eta}\mathbf{U}_{2n}^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0\|_F - 2n^{-1}\|\mathbf{B}\|_F\|\mathbf{T}_n\|_F - 2n^{-1}\|\mathbf{B}\|_F\|\mathbf{T}_{2n}\|_F \\ & + n^{-1}\text{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)]\text{vec}(\mathbf{B}) \\ & + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}\mathbf{B}^T\boldsymbol{\Omega}_0^{-1}\mathbf{B}) + n^{-1}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{B}^T\boldsymbol{\Omega}_0\mathbf{B}) - 2n^{-1}\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \\ & -\frac{n^{-1}}{2}p_{\mathcal{A}}\sqrt{n}\lambda_{\mathcal{A}}\max_{1\le i\le p_{\mathcal{A}}-d}\{\|\mathbf{e}_i^T\boldsymbol{\Gamma}\|_2^{-1}\|\mathbf{e}_i^T(\mathbf{Z}-\frac{1}{2}n^{-1/2}\boldsymbol{\Gamma}\mathbf{Z}^T\mathbf{Z})\|_2\} + o_p(n^{-1}). \end{aligned}$$

Notice that

$$\begin{aligned} & \text{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)]\text{vec}(\mathbf{B}) + \operatorname{tr}(\boldsymbol{\Omega}\mathbf{B}^T\boldsymbol{\Omega}_0^{-1}\mathbf{B}) + \operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{B}^T\boldsymbol{\Omega}_0\mathbf{B}) - 2\operatorname{tr}(\mathbf{B}^T\mathbf{B}) \\ =\ & \text{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0) + \boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0 - 2\mathbf{I}_d\otimes\mathbf{I}_{p-d}]\text{vec}(\mathbf{B}) \\ =\ & \text{vec}(\mathbf{B})^T\mathbf{M}\text{vec}(\mathbf{B}) \\ \geq\ & m\|\mathbf{B}\|_F^2, \end{aligned}$$

where m is the smallest eigenvalue of $\mathbf{M}$. Notice that $\mathbf{M}$ appears in Proposition 4 in Cook and Zhang (2015), by Shapiro (1986), $\mathbf{M}$ is a positive definite matrix and $m > 0$. When $\|\mathbf{B}\|_F > C$ for sufficiently large C, the terms with order $\|\mathbf{B}\|_F^2$ dominate the terms with order $\|\mathbf{B}\|_F$ and the conclusion follows.

Proof of Theorem 6, part (b). From part (a) of Theorem 6 we know that the E-SGPLS estimators $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\alpha}$, $\widehat{\boldsymbol{\eta}}$ are $\sqrt{n}$-consistent estimators of $\boldsymbol{\Gamma}$, α, $\boldsymbol{\eta}$. Let

$$\hat{\zeta}_i = \widehat{\alpha} + \widehat{\boldsymbol{\eta}}^T\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}_i.$$

We will prove the selection consistency by contradiction. Suppose that no row in $\widehat{\boldsymbol{\Gamma}}$ is zero, i.e., $\|\widehat{\boldsymbol{\gamma}}_i\|_2 > 0$ for $p_{\mathcal{A}} + 1 \le i \le p$. The derivative of f_{obj} with respect to $\boldsymbol{\gamma}_i$ should be 0 evaluated at $\widehat{\boldsymbol{\gamma}}_i$. By Adragni et al. (2012), the derivative of f_{obj} on the Grassmann manifold is

$$\begin{aligned} \left(\frac{\partial f_{\text{obj}}}{\partial \boldsymbol{\Gamma}}\right)^T \boldsymbol{\Gamma}_0 = & \Big\{ -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\zeta_i)\mathbf{X}_i\boldsymbol{\eta}^T + 2\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1} \\ & + 2\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma})^{-1} + \mathbf{v}(\boldsymbol{\Gamma})\Big\}^T \boldsymbol{\Gamma}_0, \end{aligned}$$

where

$$\mathbf{v}(\boldsymbol{\Gamma}) = \begin{pmatrix} \lambda_1 \frac{\boldsymbol{\gamma}_1^T}{\|\boldsymbol{\gamma}_1\|_2} \\ \vdots \\ \lambda_p \frac{\boldsymbol{\gamma}_p^T}{\|\boldsymbol{\gamma}_p\|_2} \end{pmatrix}.$$

Let $\widehat{W}$ be the weights when $\mathbf{\Gamma} = \widehat{\mathbf{\Gamma}}$. Then

$$(\mathbf{I}_p - \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^T)\Big\{ -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i\widehat{\boldsymbol{\eta}}^T + 2\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}})^{-1} + 2\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}})^{-1} + \mathbf{v}(\widehat{\mathbf{\Gamma}})\Big\} = 0. \tag{31}$$

Notice that $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{\Gamma}}\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{GLM}}$ are both $\sqrt{n}$-consistent estimators of $\boldsymbol{\beta}$, and $\widehat{\alpha}$ and $\widehat{\alpha}_{\mathrm{GLM}}$ are both $\sqrt{n}$-consistent estimators of α. Since both $\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i/n$ and $\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})\mathbf{X}_i/n$ are $\sqrt{n}$-consistent estimator of $\mathrm{E}(\mathcal{D}'(\zeta)\mathbf{X})$, we have

$$\frac{1}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})\mathbf{X}_i + O_p(n^{-1/2}) = O_p(n^{-1/2}).$$

The last equality is because that $\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{\mathrm{GLM},i})\mathbf{X}_i = 0$, as we discussed before (28).

Then

$$\begin{aligned} & (\mathbf{I}_p - \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^T)\Big\{-\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i\widehat{\boldsymbol{\eta}}^T + 2\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}})^{-1} + 2\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}})^{-1}\Big\} \\ = & (\mathbf{I}_p - \mathbf{\Gamma}\mathbf{\Gamma}^T)\left\{2\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Gamma})^{-1} + 2\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{\Gamma})^{-1}\right\} + O_p(n^{-1/2}) \\ = & (\mathbf{I}_p - \mathbf{\Gamma}\mathbf{\Gamma}^T)4\mathbf{\Gamma} + O_p(n^{-1/2}) \\ = & O_p(n^{-1/2}). \end{aligned}$$

Therefore

$$\sqrt{n}(\mathbf{I}_p - \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^T)\Big\{ -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i\widehat{\boldsymbol{\eta}}^T + 2\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}})^{-1} + 2\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{\Gamma}})^{-1}\Big\} = O_p(1). \tag{32}$$

On the other hand,

$$
\begin{aligned}
(\mathbf{I}_p - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T)\mathbf{v}(\widehat{\boldsymbol{\Gamma}}) &= \left(\mathbf{I}_p - \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1^T\widehat{\boldsymbol{\gamma}}_1 & \cdots & \widehat{\boldsymbol{\gamma}}_1^T\widehat{\boldsymbol{\gamma}}_p \\ \vdots & \vdots & \vdots \\ \widehat{\boldsymbol{\gamma}}_p^T\widehat{\boldsymbol{\gamma}}_1 & \cdots & \widehat{\boldsymbol{\gamma}}_p^T\widehat{\boldsymbol{\gamma}}_p \end{pmatrix}\right)\begin{pmatrix} \lambda_1 \dfrac{\widehat{\boldsymbol{\gamma}}_1^T}{\|\widehat{\boldsymbol{\gamma}}_1\|_2} \\ \vdots \\ \lambda_p \dfrac{\widehat{\boldsymbol{\gamma}}_p^T}{\|\widehat{\boldsymbol{\gamma}}_p\|_2} \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 \dfrac{\widehat{\boldsymbol{\gamma}}_1^T}{\|\widehat{\boldsymbol{\gamma}}_1\|_2} - \sum_{i=1}^p \lambda_i \dfrac{\widehat{\boldsymbol{\gamma}}_1^T\widehat{\boldsymbol{\gamma}}_i\widehat{\boldsymbol{\gamma}}_i^T}{\|\widehat{\boldsymbol{\gamma}}_i\|_2} \\ \vdots \\ \lambda_p \dfrac{\widehat{\boldsymbol{\gamma}}_p^T}{\|\widehat{\boldsymbol{\gamma}}_p\|_2} - \sum_{i=1}^p \lambda_i \dfrac{\widehat{\boldsymbol{\gamma}}_p^T\widehat{\boldsymbol{\gamma}}_i\widehat{\boldsymbol{\gamma}}_i^T}{\|\widehat{\boldsymbol{\gamma}}_i\|_2} \end{pmatrix}.
\end{aligned}
$$

Let $\lambda_{j^*} = \max\{\lambda_i, i > p_{\mathcal{A}}\}$. The j^*th row of $(\mathbf{I}_p - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T)\mathbf{v}(\widehat{\boldsymbol{\Gamma}})$ is

$$
\begin{aligned}
&\sqrt{n}\left(\lambda_{j^*}\frac{\widehat{\boldsymbol{\gamma}}_{j^*}^T}{\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2} - \sum_{i=1}^p \lambda_i \frac{\widehat{\boldsymbol{\gamma}}_{j^*}^T\widehat{\boldsymbol{\gamma}}_i\widehat{\boldsymbol{\gamma}}_i^T}{\|\widehat{\boldsymbol{\gamma}}_i\|_2}\right) \\
= \;& \sqrt{n}\lambda_{j^*}\frac{\widehat{\boldsymbol{\gamma}}_{j^*}^T}{\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2}\left(1 - \sum_{i=1}^{p_{\mathcal{A}}} \frac{\lambda_i}{\lambda_{j^*}}\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2\|\widehat{\boldsymbol{\gamma}}_i\|_2 - \sum_{i=p_{\mathcal{A}}+1}^{p} \frac{\lambda_i}{\lambda_{j^*}}\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2\|\widehat{\boldsymbol{\gamma}}_i\|_2\right)
\end{aligned}
$$

Since $\sqrt{n}\lambda_{\mathcal{A}} \to 0$ and $\sqrt{n}\lambda_{\mathcal{I}} \to \infty$, then $\lambda_i/\lambda_{j^*} \to 0$ for $1 \le j \le p_{\mathcal{A}}$, and $0 < \lambda_i/\lambda_{j^*} \le 1$ for $1 + p_{\mathcal{A}} \le j \le p$. From part (a), $\widehat{\boldsymbol{\Gamma}}$ is a consistent estimator of $\boldsymbol{\Gamma}$, so $\widehat{\boldsymbol{\gamma}}_i = \boldsymbol{\gamma}_i + O_p(n^{-1/2})$ for $1 \le j \le p_{\mathcal{A}}$ and $\widehat{\boldsymbol{\gamma}}_i = O_p(n^{-1/2})$ for $1 + p_{\mathcal{A}} \le j \le r$. Then as $n \to \infty$,

$$
\sum_{i=1}^{p_{\mathcal{A}}} \frac{\lambda_i}{\lambda_{j^*}}\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2\|\widehat{\boldsymbol{\gamma}}_i\|_2 \to 0, \quad \sum_{i=p_{\mathcal{A}}+1}^{p} \frac{\lambda_i}{\lambda_{j^*}}\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2\|\widehat{\boldsymbol{\gamma}}_i\|_2 \to 0.
$$

Let m be the element in $\widehat{\boldsymbol{\gamma}}_{j^*}$ that has the largest absolute value, then $m/\|\widehat{\boldsymbol{\gamma}}_{j^*}\|_2 > \sqrt{d}$. Because we have $\sqrt{n}\lambda_{j^*} \to \infty$, there is at least one element in the j^*th row of $(\mathbf{I}_p - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T)\mathbf{v}(\widehat{\boldsymbol{\Gamma}})$ tends to infinity or negative infinity. Together with (32), this is a contradiction to (31). Hence, there exists at least one i, $p_{\mathcal{A}} + 1 \le i \le p$, such that $\|\widehat{\boldsymbol{\gamma}}_i\|_2 = 0$ with probability tending to 1. Without loss of generality, we assume that $\|\widehat{\boldsymbol{\gamma}}_p\|_2 = 0$. We will prove that $\widehat{\boldsymbol{\gamma}}_i = 0$, for $p_{\mathcal{A}} + 1 \le i \le p - 1$ by induction.

Suppose that for $p_{\mathcal{A}} + 1 \le i \le p - 1$, $\|\widehat{\boldsymbol{\gamma}}_i\|_2 > 0$. Let

$$
\boldsymbol{\Gamma}^{(1)} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_{p-1} \end{pmatrix} \in \mathbb{R}^{(p-1)\times d}, \quad \widehat{\boldsymbol{\Gamma}}^{(1)} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \vdots \\ \widehat{\boldsymbol{\gamma}}_{p-1} \end{pmatrix} \in \mathbb{R}^{(p-1)\times d},
$$

then

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}^{(1)} \\ 0 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} \widehat{\boldsymbol{\Gamma}}^{(1)} \\ 0 \end{pmatrix}.$$

Let $\boldsymbol{\Gamma}_0^{(1)} \in \mathbb{R}^{(p-1)\times(p-d-1)}$ be the completion of $\boldsymbol{\Gamma}^{(1)}$. Given $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$ is determined up to an orthogonal transformation. For simplicity, we take the $\boldsymbol{\Gamma}_0$ that has the following form

$$\boldsymbol{\Gamma}_0 = \begin{pmatrix} \boldsymbol{\Gamma}_0^{(1)} & 0 \\ 0 & 1 \end{pmatrix}. \tag{33}$$

As $\boldsymbol{\Omega}_0$ carries the coordinates of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Gamma}_0$, based on the structure (33), we can partition $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$ accordingly into

$$\boldsymbol{\Omega}_0 = \begin{pmatrix} \boldsymbol{\Omega}_0^{(1)} & \boldsymbol{\Omega}_0^{(12)} \\ \boldsymbol{\Omega}_0^{(21)} & \Omega_0^{(2)} \end{pmatrix},$$

where $\boldsymbol{\Omega}_0^{(1)} \in \mathbb{R}^{(p-d-1)\times(p-d-1)}$, $\boldsymbol{\Omega}_0^{(12)} \in \mathbb{R}^{(p-d-1)}$, $\boldsymbol{\Omega}_0^{(21)} = (\boldsymbol{\Omega}_0^{(12)})^T$ and $\Omega_0^{(2)} \in \mathbb{R}$. Then

$$\boldsymbol{\Sigma}_{\mathbf{X}} \quad = \quad \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T = \begin{pmatrix} \boldsymbol{\Gamma}^{(1)}\boldsymbol{\Omega}(\boldsymbol{\Gamma}^{(1)})^T + \boldsymbol{\Gamma}_0^{(1)}\boldsymbol{\Omega}_0^{(1)}(\boldsymbol{\Gamma}_0^{(1)})^T & \boldsymbol{\Gamma}_0^{(1)}\boldsymbol{\Omega}_0^{(12)} \\ \boldsymbol{\Omega}_0^{(21)}(\boldsymbol{\Gamma}_0^{(1)})^T & \Omega_0^{(2)} \end{pmatrix}.$$

Let $\mathbf{S}_{\mathbf{X}}^{(1)} \in \mathbb{R}^{(p-1)\times(p-1)}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}^{(1)} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the first $p-1$ rows and $p-1$ columns of $\mathbf{S}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, then $\boldsymbol{\Sigma}_{\mathbf{X}}^{(1)} = \boldsymbol{\Gamma}^{(1)}\boldsymbol{\Omega}(\boldsymbol{\Gamma}^{(1)})^T + \boldsymbol{\Gamma}_0^{(1)}\boldsymbol{\Omega}_0^{(1)}(\boldsymbol{\Gamma}_0^{(1)})^T$. Since $\mathbf{S}_{\mathbf{X}}^{(1)}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Sigma}_{\mathbf{X}}^{(1)}$, $\mathbf{S}_{\mathbf{X}}^{(1)} = \boldsymbol{\Sigma}_{\mathbf{X}}^{(1)} + O_p(n^{-1/2})$. Let $(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)} \in \mathbb{R}^{(p-1)\times(p-1)}$ and $(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})^{(1)} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the first $p-1$ rows and first $p-1$ columns of $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$. Then

$$(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})^{(1)} = \boldsymbol{\Gamma}^{(1)}\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^{(1)})^T + \boldsymbol{\Gamma}_0^{(1)}[\boldsymbol{\Omega}_0^{(1)} - \boldsymbol{\Omega}_0^{(12)}(\Omega_0^{(2)})^{-1}\boldsymbol{\Omega}_0^{(21)}]^{-1}(\boldsymbol{\Gamma}_0^{(1)})^T$$

and $(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)} = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})^{(1)} + O_p(n^{-1/2})$. We define the objective function $f_{\text{obj}}^{(1)}$ on the Grassmann manifold $\mathcal{G}(p-1, d)$ as

$$f_{\text{obj}}^{(1)}(\mathbf{G}) = -\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}(\zeta_i) + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{X}}^{(1)}\mathbf{G}| + \log|\mathbf{G}^T(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\mathbf{G}| + \sum_{i=1}^{p-1}\lambda_i\|\mathbf{g}_i\|_2,$$

where $\mathbf{G} \in \mathbb{R}^{(p-1)\times d}$, $\mathbf{G}^T\mathbf{G} = \mathbf{I}_d$, $\zeta_i = \alpha + \boldsymbol{\eta}^T\mathbf{G}^T\mathbf{X}^{(1)}$, $\mathbf{X}^{(1)} \in \mathbb{R}^{p-1}$ denotes the first $p-1$ elements in $\mathbf{X}$, and $\mathbf{g}_i$ is the ith row of $\mathbf{G}$, for $i = 1, \cdots, p-1$.

Since $\widehat{\boldsymbol{\Gamma}}$ is a local minimum of f_{obj}, $\widehat{\boldsymbol{\Gamma}}^{(1)}$ is a local minimum of $f^{(1)}_{\text{obj}}$. Taking the derivative of $f^{(1)}_{\text{obj}}$ with respect to $\mathbf{G}$ on the Grassmann manifold and evaluate it on $\widehat{\boldsymbol{\Gamma}}^{(1)}$, we have

$$\Big(-\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i^{(1)}\widehat{\boldsymbol{\eta}}^T + 2\mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1}$$
$$+ 2(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1} + \mathbf{v}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})\Big)^T\widehat{\boldsymbol{\Gamma}}_0^{(1)} = 0,$$

where

$$\mathbf{v}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)}) = \begin{pmatrix} \lambda_1\frac{\widehat{\boldsymbol{\gamma}}_1^T}{\|\widehat{\boldsymbol{\gamma}}_1\|_2} \\ \vdots \\ \lambda_{p-1}\frac{\widehat{\boldsymbol{\gamma}}_{p-1}^T}{\|\widehat{\boldsymbol{\gamma}}_{p-1}\|_2} \end{pmatrix}.$$

This is equivalent to

$$\left\{\mathbf{I}_{p-1} - \widehat{\boldsymbol{\Gamma}}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\right\}\Big[-\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i^{(1)}\widehat{\boldsymbol{\eta}}^T + 2\mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1}$$
$$+2(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1} + \mathbf{v}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})\Big] = 0.$$

Following the previous discussion of obtaining (32), we have

$$\begin{aligned} & 2\Big\{\mathbf{I}_{p-1} - \widehat{\boldsymbol{\Gamma}}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\Big\}\Big[-\frac{1}{n}\sum_{i=1}^{n}\mathcal{D}'(\widehat{\zeta}_i)\mathbf{X}_i^{(1)}\widehat{\boldsymbol{\eta}}^T + \mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\mathbf{S}_{\mathbf{X}}^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1} \\ & +(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\left\{(\widehat{\boldsymbol{\Gamma}}^{(1)})^T(\mathbf{S}_{\mathbf{X}}^{-1})^{(1)}\widehat{\boldsymbol{\Gamma}}^{(1)}\right\}^{-1}\Big] \\ = \; & 2\Big\{\mathbf{I}_{p-1} - \boldsymbol{\Gamma}^{(1)}(\boldsymbol{\Gamma}^{(1)})^T\Big\}\Big[\boldsymbol{\Sigma}_{\mathbf{X}}^{(1)}\boldsymbol{\Gamma}^{(1)}\left\{(\boldsymbol{\Gamma}^{(1)})^T\boldsymbol{\Sigma}_{\mathbf{X}}^{(1)}\boldsymbol{\Gamma}^{(1)}\right\}^{-1} \\ & +(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})^{(1)}\boldsymbol{\Gamma}^{(1)}\left\{(\boldsymbol{\Gamma}^{(1)})^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})^{(1)}\boldsymbol{\Gamma}^{(1)}\right\}^{-1}\Big] + O_p(n^{-1/2}) \\ = \; & O_p(n^{-1/2}). \end{aligned}$$

Parallel to the derivation with $\sqrt{n}(\mathbf{I}_p - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T)\mathbf{v}(\widehat{\boldsymbol{\Gamma}})$, we can show that at least one element in the $p-1$ vector $\sqrt{n}\{\mathbf{I} - \widehat{\boldsymbol{\Gamma}}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})^T\}\mathbf{v}^{(1)}(\widehat{\boldsymbol{\Gamma}}^{(1)})$ goes to ∞ or $-\infty$ as n increases. This is a contradiction. So there is at least one i, $p_{\mathcal{A}}+1 \le i \le p-1$, such that $\|\widehat{\boldsymbol{\gamma}}_i\|_2 = 0$ with probability tending to 1.

By induction, we have $\|\widehat{\gamma}_i\|_2 = 0$ with probability tending to 1 for $p_{\mathcal{A}}+1 \leq i \leq p$.

Proof of Theorem 6, part (c). Now we derive the asymptotic distribution of the E-SGPLS estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$, and prove that it has the same asymptotic variance as the oracle envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$. If for some $a_n = o(n^{-1/2})$, $\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}} = \widehat{\boldsymbol{\Gamma}}_O + O_p(a_n)$ and $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}}_O + O_p(a_n)$, since $\boldsymbol{\beta}_{\mathcal{A}} = \boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\eta}$, then $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} = \widehat{\boldsymbol{\beta}}_{\mathcal{A},O}+O_p(a_n)$. By Slutsky's theorem, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}})$ has the same asymptotic distribution as $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O} - \boldsymbol{\beta}_{\mathcal{A}})$. Therefore the conclusion in part (c) follows if we can prove $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_O} + O_p(a_n)$. We take $a_n = (n^{-1/2}\lambda_{\mathcal{A}})^{1/2}$. Because $\sqrt{n}\lambda_{\mathcal{A}} \to 0$, we have $a_n = o(n^{-1/2})$. Let

$$f_{\mathrm{obj},\mathcal{A}}(\alpha, \boldsymbol{\eta}, \mathbf{G}) = -\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}(\alpha + \boldsymbol{\eta}^T\mathbf{G}^T\mathbf{X}_{i,\mathcal{A}}) + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}\mathbf{G}| + \log|\mathbf{G}^T(\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}}\mathbf{G}| + \sum_{i=1}^{p_{\mathcal{A}}}\lambda_i\|\mathbf{g}_i\|_2,$$

where $\mathbf{G} \in \mathbb{R}^{p_{\mathcal{A}}\times d}$, and $\mathbf{g}_i$ is the ith row of $\mathbf{G}$. The function $f_{\mathrm{obj},\mathcal{A}}(\alpha, \boldsymbol{\eta}, \mathbf{G})$ is the objective function of all the parameters, i.e., $\alpha, \boldsymbol{\eta}$ and $\boldsymbol{\Gamma}_{\mathcal{A}}$. When maximizing the joint likelihood function of Y and $\mathbf{X}$, the estimators of the parameters $\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ can be written as closed-form functions of $\alpha, \boldsymbol{\eta}$ and $\boldsymbol{\Gamma}_{\mathcal{A}}$: $\widehat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{X}}$, $\widehat{\boldsymbol{\Omega}} = \boldsymbol{\Gamma}_{\mathcal{A}}^T\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}\boldsymbol{\Gamma}_{\mathcal{A}}$ and $\widehat{\boldsymbol{\Omega}}_0 = \boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}_0$. After substituting $\widehat{\boldsymbol{\mu}}_{\mathbf{X}}$, $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Omega}}_0$ in the likelihood, we obtain the objective function $f_{\mathrm{obj},\mathcal{A}}(\alpha, \boldsymbol{\eta}, \mathbf{G})$. Because of the selection consistency of the E-SGPLS estimator, $(\widehat{\alpha}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A}}) = \arg\min f_{\mathrm{obj},\mathcal{A}}(\alpha, \boldsymbol{\eta}, \mathbf{G})$. Let $\widehat{\alpha}_O$, $\widehat{\boldsymbol{\eta}}_O$ and $\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}$ be the oracle envelope estimator of α, $\boldsymbol{\eta}$ and $\boldsymbol{\Gamma}_{\mathcal{A}}$. Then $(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})$ is a local minimum of the objective function

$$\begin{aligned} J_O(\alpha, \boldsymbol{\eta}, \mathbf{G}) = & -\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}(\alpha + \boldsymbol{\eta}^T\mathbf{G}^T\mathbf{X}_{i,\mathcal{A}}) + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}\mathbf{G}| + \log|\mathbf{G}^T(\mathbf{S}_{\mathbf{X}}^{-1})_{\mathcal{A}}\mathbf{G}| \\ \equiv & f_{O1}(\alpha, \boldsymbol{\eta}, \mathbf{G}) + f_{O2}(\mathbf{G}) + f_{O3}(\mathbf{G}). \end{aligned} \tag{34}$$

Let $\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}0,O} \in \mathbb{R}^{p_{\mathcal{A}}\times(p_{\mathcal{A}}-d)}$ be a completion of $\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}$. To establish Theorem 6 part (c), it is enough to show that for arbitrarily small $\epsilon > 0$, there exist sufficiently large constants C_0, C_1 and C_2, such that

$$\lim_{n\to\infty} P\left[\inf_{\substack{|\delta_0|=C_0,\boldsymbol{\Delta}_1\in\mathbb{R}^{d\times 1},\|\boldsymbol{\Delta}_1\|_F=C_1\\ \mathbf{Z}\in T_{\boldsymbol{\Gamma}(p_{\mathcal{A}},d)},\mathbf{Z}=\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\mathbf{A}+\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}0,O}\mathbf{B},\|\mathbf{B}\|_F=C_2}} f_{\mathrm{obj},\mathcal{A}}\{\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\boldsymbol{\Delta}_1, R(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\} > f_{\mathrm{obj},\mathcal{A}}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})\right] > 1-\epsilon. \tag{35}$$

Because $(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})$ is a local minimum of objective function (34), the derivative of J_O should be 0 evaluated at the $(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})$, i.e.

$$\begin{aligned} &\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i}) = 0, \quad \sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}^T\mathbf{X}_{i,\mathcal{A}} = 0,\\ &\left(\frac{\partial f_{O1}}{\partial \mathbf{G}} + \frac{\partial f_{O2}}{\partial \mathbf{G}} + \frac{\partial f_{O3}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\Big|_{\mathbf{G}=\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}} = 0, \end{aligned} \tag{36}$$

where $\hat{\zeta}_{O,i} = \widehat{\alpha}_O + \widehat{\boldsymbol{\eta}}_O^T\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}^T\mathbf{X}_{i,\mathcal{A}}$.

We write $f_{\mathrm{obj},\mathcal{A}}$ as $f_{\mathrm{obj},\mathcal{A}}(\alpha, \boldsymbol{\eta}, \mathbf{G}) = f_{\mathcal{A}1}(\alpha, \boldsymbol{\eta}, \mathbf{G}) + f_{\mathcal{A}2}(\mathbf{G}) + f_{\mathcal{A}3}(\mathbf{G}) + f_{\mathcal{A}4}(\mathbf{G})$, where $f_{\mathcal{A}1}$, $f_{\mathcal{A}2}$, $f_{\mathcal{A}3}$ and $f_{\mathcal{A}4}$ corresponds to the four summands in $f_{\mathrm{obj},\mathcal{A}}$. Similar to the proof of Theorem 6 part (a), we expand $f_{\mathrm{obj},\mathcal{A}}\big(\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\boldsymbol{\Delta}_1, R(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\big)$ and compute

$$f_{\mathrm{obj},\mathcal{A}}\big(\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\boldsymbol{\Delta}_1, R(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\big) - f_{\mathrm{obj},\mathcal{A}}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}).$$

After some calculations that are parallel to those in Theorem 6 part (a), we have

$$\begin{aligned} \overset{\to\delta_0}{d}f_{\mathcal{A}1}(\widehat{\alpha}_O) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\delta_0 = 0,\\ \overset{\to\boldsymbol{\Delta}_1}{d}f_{\mathcal{A}1}(\widehat{\boldsymbol{\eta}}_O) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\mathbf{X}_{i,\mathcal{A}}^T\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\boldsymbol{\Delta}_1 = 0, \end{aligned}$$

$$\begin{aligned}
\overset{\rightarrow\mathbf{Z}^*}{df}_{\mathcal{A}1}(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) &= \operatorname{tr}\Big\{\Big(\frac{\partial f_{\mathcal{A}1}}{\partial \mathbf{G}}\Big)^T\mathbf{Z}^*\Big|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}}\Big\} = -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\operatorname{tr}(\widehat{\boldsymbol{\eta}}_O\mathbf{X}_{i,\mathcal{A}}^T\mathbf{Z}^*) \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\mathbf{X}_{i,\mathcal{A}}^T\mathbf{Z}^*\widehat{\boldsymbol{\eta}}_O \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}'(\widehat{\zeta}_{O,i})\mathbf{X}_{i,\mathcal{A}}^T\widehat{\mathbf{\Gamma}}_{\mathcal{A}0,O}\mathbf{B}\widehat{\boldsymbol{\eta}}_O + O_p(a_n) \\
&= \operatorname{tr}\Big\{\Big[\Big(\frac{\partial f_{O1}}{\partial \mathbf{G}}\Big)^T\mathbf{G}_0\Big]\Big|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}}\mathbf{B}\Big\} + O_p(a_n),
\end{aligned}$$

where $\mathbf{Z}^* = \mathbf{Z} - (1/2)a_n\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}\mathbf{Z}^T\mathbf{Z} + o_p(a_n)$, $\mathbf{Z} = \widehat{\mathbf{\Gamma}}_{\mathcal{A},O}\mathbf{A} + \widehat{\mathbf{\Gamma}}_{\mathcal{A}0,O}\mathbf{B}$, $\widehat{\mathbf{\Gamma}}_{\mathcal{A}0,O} \in \mathbb{R}^{p_{\mathcal{A}}\times(p_{\mathcal{A}}-d)}$ is a completion of $\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}$ and $\mathbf{X}_{i,\mathcal{A}}$ denotes the ith observation of $\mathbf{X}_{\mathcal{A}}$. The penultimate equation is based on the second equation in (36).

The second order directional derivatives of $f_{\mathcal{A}1}$ are

$$\begin{aligned}
\overset{\rightarrow\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) &= \operatorname{tr}\Big\{\Big[\frac{\partial \overset{\rightarrow\mathbf{Z}^*}{df}_{\mathcal{A}1}(\mathbf{G})}{\partial \mathbf{G}}\Big|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}}\Big]^T\mathbf{Z}^*\Big\} \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})(\mathbf{X}_{i,\mathcal{A}}^T\mathbf{Z}^*\widehat{\boldsymbol{\eta}}_O)^2 \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)[\mathbf{X}_{i,\mathcal{A}}^T(\mathbf{\Gamma}_{\mathcal{A}}\mathbf{A} + \mathbf{\Gamma}_{\mathcal{A},0}\mathbf{B})\boldsymbol{\eta}]^2 + o_p(1), \\
\overset{\rightarrow\delta_0}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})\delta_0^2 = -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)\delta_0^2 + o_p(1), \\
\overset{\rightarrow\mathbf{\Delta}_1}{df^2_{\mathcal{A}1}}(\widehat{\boldsymbol{\eta}}_O) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})(\mathbf{X}_{i,\mathcal{A}}^T\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}\mathbf{\Delta}_1)^2 \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)(\mathbf{X}_{i,\mathcal{A}}^T\mathbf{\Gamma}_{\mathcal{A}}\mathbf{\Delta}_1)^2 + o_p(1), \\
\overset{\rightarrow\delta_0,\rightarrow\mathbf{\Delta}_1}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})\delta_0(\mathbf{X}_{i,\mathcal{A}}^T\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}\mathbf{\Delta}_1) \\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)\delta_0(\mathbf{X}_{i,\mathcal{A}}^T\mathbf{\Gamma}_{\mathcal{A}}\mathbf{\Delta}_1) + o_p(1),
\end{aligned}$$

$$\begin{aligned}
\overset{\to\delta_0,\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O,\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})\delta_0(\mathbf{X}_{i,\mathcal{A}}^T\mathbf{Z}^*\widehat{\boldsymbol{\eta}}_O)\\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})\delta_0\Big[\mathbf{X}_{i,\mathcal{A}}^T(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\mathbf{A}+\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}0,O}\mathbf{B})\widehat{\boldsymbol{\eta}}_O\Big]+o_p(1)\\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)\delta_0\Big[\mathbf{X}_{i,\mathcal{A}}^T(\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{A}+\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B})\boldsymbol{\eta}\Big]+o_p(1),\\
\overset{\to\boldsymbol{\Delta}_1,\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\boldsymbol{\eta}}_O,\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}) &= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})(\mathbf{X}_{i,\mathcal{A}}^T\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\boldsymbol{\Delta}_1)(\mathbf{X}_{i,\mathcal{A}}^T\mathbf{Z}^*\widehat{\boldsymbol{\eta}}_O)\\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\widehat{\zeta}_{O,i})(\mathbf{X}_{i,\mathcal{A}}^T\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\boldsymbol{\Delta}_1)\Big[\mathbf{X}_{i,\mathcal{A}}^T(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}\mathbf{A}+\widehat{\boldsymbol{\Gamma}}_{\mathcal{A}0,O}\mathbf{B})\widehat{\boldsymbol{\eta}}_O\Big]+o_p(1)\\
&= -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)(\mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\Delta}_1)\Big[\mathbf{X}_{i,\mathcal{A}}^T(\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{A}+\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B})\boldsymbol{\eta}\Big]+o_p(1).
\end{aligned}$$

Then the second directional derivative of $f_{\mathcal{A}1}$ with respect to all parameters is

$$\begin{aligned}
&\overset{\to\delta_0}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O)+\overset{\to\boldsymbol{\Delta}_1}{df^2_{\mathcal{A}1}}(\widehat{\boldsymbol{\eta}}_O)+\overset{\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})+2\overset{\to\delta_0,\to\boldsymbol{\Delta}_1}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O,\widehat{\boldsymbol{\eta}}_O)+2\overset{\to\delta_0,\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\alpha}_O,\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})\\
&+2\overset{\to\boldsymbol{\Delta}_1,\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\widehat{\boldsymbol{\eta}}_O,\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O})\\
=\;& -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)\big[\delta_0+\mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\Delta}_1+\mathbf{X}_{i,\mathcal{A}}^T(\boldsymbol{\Gamma}_{\mathcal{A}}\mathbf{A}+\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B})\boldsymbol{\eta}\big]^2+o_p(1)\\
=\;& -\frac{2}{n}\sum_{i=1}^n \mathcal{D}''(\zeta_i)\big[\delta_0+\mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1+\mathbf{A}\boldsymbol{\eta})+\mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}\big]^2+o_p(1).
\end{aligned}$$

We substitute $\overset{\to\mathbf{Z}^*}{df}_{\mathcal{A}1}(\boldsymbol{\Gamma})$ and $\overset{\to\mathbf{Z}^*}{df^2_{\mathcal{A}1}}(\boldsymbol{\Gamma})$ into the expansion for $f_{\mathcal{A}1}$ and get

$$\begin{aligned}
& f_{\mathcal{A}1}(\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\boldsymbol{\Delta}_1, R(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})) - f_{\mathcal{A}1}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}) \\
= \quad & a_n \operatorname{tr}\left\{ \left[\left(\frac{\partial f_{O1}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\right]\Bigg|_{\mathbf{G}=\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}} \mathbf{B}\right\} - a_n^2 \frac{1}{n}\sum_{i=1}^n \Big\{\mathcal{D}''(\zeta_i)[\delta_0 + \mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + \mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}]^2\Big\} + o_p(a_n^2) \\
= \quad & a_n \operatorname{tr}\left\{ \left[\left(\frac{\partial f_{O1}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\right]\Bigg|_{\mathbf{G}=\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}} \mathbf{B}\right\} \\
& + a_n^2 \mathrm{E}\{I(\zeta)\}[\delta_0 + \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}}^T \boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}}^T \boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}]^2 \\
& + a_n^2 (\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta})^T \boldsymbol{\Gamma}_{\mathcal{A}}^T \mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1} \boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + a_n^2 \operatorname{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_{\mathcal{A},0}^T)\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_{\mathcal{A},0})]\operatorname{vec}(\mathbf{B}) + o_p(a_n^2).
\end{aligned}$$

The second equality is because

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n -\mathcal{D}''(\zeta_i)\big[\delta_0 + \mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + \mathbf{X}_{i,\mathcal{A}}^T\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}\big]^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^n -\mathcal{D}''(\zeta_i)\big[\delta_0^* + (\mathbf{X}_{i,\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}})^T\boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + (\mathbf{X}_{i,\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}})^T\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}\big]^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^n \left\{ I(\zeta_i) - \frac{y_i - \mu_i}{g'(\mu_i)}\frac{d}{d\mu_i}\left(\frac{1}{b''((b')^{-1}(\mu_i))g'(\mu_i)}\right)\right\} \\
& \big[\delta_0^* + (\mathbf{X}_{i,\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}})^T\boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + (\mathbf{X}_{i,\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}})^T\boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}\big]^2 \\
= \quad & \mathrm{E}\{I(\zeta)\}\delta_0^{*2} + (\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta})^T \boldsymbol{\Gamma}_{\mathcal{A}}^T \mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1} \boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + \operatorname{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_{\mathcal{A},0}^T)\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_{\mathcal{A},0})]\operatorname{vec}(\mathbf{B}) + o_p(1).
\end{aligned}$$

where $\delta_0^* = \delta_0 + \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}}^T \boldsymbol{\Gamma}_{\mathcal{A}}(\boldsymbol{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}}^T \boldsymbol{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}$. The last equality uses similar reasoning as in (29) and (30), as well as the facts $\sum_{i=1}^n I(\zeta_i)(\mathbf{X}_{i,\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}(W)}})^T = 0$ and $\boldsymbol{\Gamma}_{\mathcal{A}}^T\mathbf{V}_{O,\boldsymbol{\beta}_{\mathcal{A}}}^{-1}\boldsymbol{\Gamma}_{\mathcal{A},0} = 0$.

Similar to the proof of Theorem 6 part (a), we have

$$\begin{aligned}
f_{\mathcal{A}2}\{R(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\} - f_{\mathcal{A}2}(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) &= 2a_n \operatorname{tr}\left\{(\widehat{\mathbf{\Gamma}}^T_{\mathcal{A},O}\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}\widehat{\mathbf{\Gamma}}_{\mathcal{A},O})^{-1}\widehat{\mathbf{\Gamma}}^T_{\mathcal{A},O}\mathbf{S}_{\mathbf{X}_{\mathcal{A}}}\widehat{\mathbf{\Gamma}}_{\mathcal{A}0,O}\mathbf{B}\right\} \\
&\quad -a_n^2\|\mathbf{B}\|_F^2 + a_n^2 \operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{B}^T\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}}\mathbf{B}) + o_p(a_n^2), \\
&= a_n \operatorname{tr}\left\{\left[\left(\frac{\partial f_{O2}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\right]\Bigg|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}} \mathbf{B}\right\} \\
&\quad -a_n^2\|\mathbf{B}\|_F^2 + a_n^2 \operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{B}^T\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}}\mathbf{B}) + o_p(a_n^2),
\end{aligned}$$

$$\begin{aligned}
f_{\mathcal{A}3}\{R(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\} - f_{\mathcal{A}3}(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) &= 2a_n \operatorname{tr}\left\{\left[\widehat{\mathbf{\Gamma}}^T_{\mathcal{A},O}(\mathbf{S}^{-1}_{\mathbf{X}})_{\mathcal{A}}\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}\right]^{-1}\widehat{\mathbf{\Gamma}}^T_{\mathcal{A},O}(\mathbf{S}^{-1}_{\mathbf{X}})_{\mathcal{A}}\widehat{\mathbf{\Gamma}}_{\mathcal{A}0,O}\mathbf{B}\right\} \\
&\quad -a_n^2\|\mathbf{B}\|_F^2 + a_n^2 \operatorname{tr}\left(\mathbf{\Omega}\mathbf{B}\widetilde{\mathbf{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}}\mathbf{B}\right) + o_p(a_n^2) \\
&= a_n \operatorname{tr}\left\{\left[\left(\frac{\partial f_{O3}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\right]\Bigg|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}} \mathbf{B}\right\} \\
&\quad -a_n^2\|\mathbf{B}\|_F^2 + a_n^2 \operatorname{tr}\left(\mathbf{\Omega}\mathbf{B}\widetilde{\mathbf{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}}\mathbf{B}\right) + o_p(a_n^2),
\end{aligned}$$

$$\begin{aligned}
& f_{\mathcal{A}4}\{R(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})\} - f_{\mathcal{A}4}(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) \\
\geq\ & -\frac{1}{2}a_n p_{\mathcal{A}}\lambda_{\mathcal{A}} \max\{\|\mathbf{e}_i^T\mathbf{\Gamma}_{\mathcal{A}}\|_2^{-1}\|\mathbf{e}_i^T(\mathbf{Z} - 0.5a_n\mathbf{\Gamma}_{\mathcal{A}}\mathbf{Z}^T\mathbf{Z})\|_2\}(1 + o_p(1)).
\end{aligned}$$

Collecting all the results so far

$$\begin{aligned}
& f_{\text{obj},\mathcal{A}}(\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\mathbf{\Delta}_1, R(\widehat{\mathbf{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})) - f_{\text{obj},\mathcal{A}}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\mathbf{\Gamma}}_{\mathcal{A},O}) \\
\geq\ & a_n \operatorname{tr}\left\{\left[\left(\frac{\partial f_{O1}}{\partial \mathbf{G}} + \frac{\partial f_{O2}}{\partial \mathbf{G}} + \frac{\partial f_{O3}}{\partial \mathbf{G}}\right)^T \mathbf{G}_0\right]\Bigg|_{\mathbf{G}=\widehat{\mathbf{\Gamma}}_{\mathcal{A},O}} \mathbf{B}\right\} \\
& + a_n^2 \mathrm{E}\{I(\zeta)\}[\delta_0 + \boldsymbol{\mu}^T_{\mathbf{X}_{\mathcal{A}(W)}}\mathbf{\Gamma}_{\mathcal{A}}(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + \boldsymbol{\mu}^T_{\mathbf{X}_{\mathcal{A}(W)}}\mathbf{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}]^2 \\
& + a_n^2(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta})^T\mathbf{\Gamma}_{\mathcal{A}}^T\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}\mathbf{\Gamma}_{\mathcal{A}}(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + a_n^2 \operatorname{vec}(\mathbf{B})^T[(\boldsymbol{\eta} \otimes \mathbf{\Gamma}^T_{\mathcal{A},0})\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_{\mathcal{A},0})]\operatorname{vec}(\mathbf{B}) \\
& -2a_n^2\|\mathbf{B}\|_F^2 + a_n^2 \operatorname{tr}(\mathbf{\Omega}^{-1}\mathbf{B}^T\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}}\mathbf{B}) + a_n^2 \operatorname{tr}(\mathbf{\Omega}\mathbf{B}\widetilde{\mathbf{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}}\mathbf{B}) \\
& -\frac{1}{2}a_n p_{\mathcal{A}}\lambda_{\mathcal{A}} \max\{\|\mathbf{e}_i^T\mathbf{\Gamma}_{\mathcal{A}}\|_2^{-1}\|\mathbf{e}_i^T\mathbf{Z} - 0.5a_n\mathbf{\Gamma}_{\mathcal{A}}\mathbf{Z}^T\mathbf{Z}\|_2\}(1 + o_p(1)) \\
=\ & a_n^2 \mathrm{E}\{I(\zeta)\}[\delta_0 + \boldsymbol{\mu}^T_{\mathbf{X}_{\mathcal{A}(W)}}\mathbf{\Gamma}_{\mathcal{A}}(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) + \boldsymbol{\mu}^T_{\mathbf{X}_{\mathcal{A}(W)}}\mathbf{\Gamma}_{\mathcal{A},0}\mathbf{B}\boldsymbol{\eta}]^2 \\
& + a_n^2(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta})^T\mathbf{\Gamma}_{\mathcal{A}}^T\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}\mathbf{\Gamma}_{\mathcal{A}}(\mathbf{\Delta}_1 + \mathbf{A}\boldsymbol{\eta}) \\
& + a_n^2 \operatorname{vec}(\mathbf{B})^T\big\{(\boldsymbol{\eta} \otimes \mathbf{\Gamma}^T_{\mathcal{A},0})\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_{\mathcal{A},0}) + \mathbf{\Omega} \otimes \widetilde{\mathbf{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}} \\
& + \mathbf{\Omega}^{-1} \otimes \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_{\mathcal{A}}-d}\big\}\operatorname{vec}(\mathbf{B}) + o_p(a_n^2).
\end{aligned}$$

Let $\mathbf{T}_{\mathcal{A}} = (\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}^T_{\mathcal{A},0})\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_{\mathcal{A},0})+\boldsymbol{\Omega}\otimes\widetilde{\boldsymbol{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}}+\boldsymbol{\Omega}^{-1}\otimes\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}-2\mathbf{I}_d\otimes\mathbf{I}_{p_{\mathcal{A}}-d}$. According to Lemma 3, $\mathbf{T}_{\mathcal{A}}$ is positive definite. Therefore

$$\begin{aligned}
& \mathrm{vec}(\mathbf{B})^T[(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}^T_{\mathcal{A},0})\mathbf{V}^{-1}_{O,\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_{\mathcal{A},0})+\boldsymbol{\Omega}\otimes\widetilde{\boldsymbol{\Omega}}^{-1}_{0,\mathcal{A}|\mathcal{I}} \\
& +\boldsymbol{\Omega}^{-1}\otimes\widetilde{\boldsymbol{\Omega}}_{0,\mathcal{A}}-2\mathbf{I}_d\otimes\mathbf{I}_{p_{\mathcal{A}}-d}]\mathrm{vec}(\mathbf{B}) \\
= \; & \mathrm{vec}(\mathbf{B})^T\mathbf{T}_{\mathcal{A}}\mathrm{vec}(\mathbf{B}) \\
\geq \; & m\|\mathbf{B}\|_F^2,
\end{aligned}$$

where m is the smallest eigenvalue of $\mathbf{T}_{\mathcal{A}}$ and $m > 0$ due to positive definiteness. Since $\mathrm{E}\{I(\zeta)\} \geq 0$, we have

$$f_{\mathrm{obj},\mathcal{A}}(\widehat{\alpha}_O + a_n\delta_0, \widehat{\boldsymbol{\eta}}_O + a_n\boldsymbol{\Delta}_1, R(\widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O} + a_n\mathbf{Z})) - f_{\mathrm{obj},\mathcal{A}}(\widehat{\alpha}_O, \widehat{\boldsymbol{\eta}}_O, \widehat{\boldsymbol{\Gamma}}_{\mathcal{A},O}) > 0$$

with probability tending to 1, which establishes (35).

APPENDIX D: ADDITIONAL SIMULATIONS

D.1. The effect of inactive predictors on efficiency. The following simulation was set up to provide numerical support for the Remark on page 12. We fixed $p = 10$, $r = 3$, $p_{\mathcal{A}} = 4$, $d = 2$ and set the first four predictors to be active predictors. The matrix $\boldsymbol{\Gamma}_{\mathcal{A}}$ was obtained by orthogonalizing a $p_{\mathcal{A}}$ by d matrix of independent standard normal random variates. The elements in $\boldsymbol{\eta}$ were independent $N(0,4)$ variates. The predictors were generated from multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} = 25\mathbf{I}_d$ and

$$\boldsymbol{\Omega}_0 = \begin{pmatrix} \mathbf{I}_{p_{\mathcal{A}}-d} & 2.7\mathbf{O} \\ 2.7\mathbf{O}^T & 9\mathbf{I}_{p_{\mathcal{I}}} \end{pmatrix}.$$

The matrix $\mathbf{O}$ was obtained by orthogonalizing a $p_{\mathcal{A}} - d$ by $p_{\mathcal{I}}$ matrix of independent standard normal random variates. We computed the eigenvalues of $\boldsymbol{\Omega}_0$, and they ranged from 0.17 to 9.83. The canonical correlation between the two sources of the immaterial part $\mathbf{Q}_{\boldsymbol{\Gamma}_{\mathcal{A}}}\mathbf{X}_{\mathcal{A}}$ and $\mathbf{X}_{\mathcal{I}}$ was 0.9. The intercept was a zero vector. And the errors followed a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{M}^T\mathbf{M}$, where the elements in $\mathbf{M}$ were independent uniform $(0,1)$ variates. We used normal errors just for easy comparison, since we can obtain the theoretical value of the asymptotic standard deviation for the standard estimator and the envelope estimator under normality. We generated 200 datasets for each sample size of 20, 40, 60, 80, 100, 150 and 200. For each dataset, we fit the standard model with only the active predictor (SMA), the standard model with all the predictors (SM), the predictor envelope model only using the active predictor (EMA), the oracle predictor envelope model (EMO) which uses all the predictors and the E-SPLS model. Then we computed the estimation standard deviation for every elements in $\boldsymbol{\beta}$ for all the estimators. The results for a randomly selected element in $\boldsymbol{\beta}$ are summarized in Figure 1. The other elements follow the same pattern.

From the results, we can see that for the standard model, SMA is more efficient than SM. This means that if we include the inactive predictors in the standard model, we will lose efficiency. The asymptotic standard deviation is 0.073 for the SM estimator and 0.032 for the SMA estimator. This justifies the standard practice in predictor selection: if a predictor is found to be inactive, we eliminate it from the subsequent analysis. However, under the envelope model, the situation is different. We notice that the E-SPLS estimator and the EMO estimator (both include the inactive predictors) are more efficient than the EMA estimator, which excludes the inactive predictors. The asymptotic standard deviation is 0.014 for the EMA estimator and 0.007 for both the EMO estimator and E-SPLS estimator. This is due to the structure of the covariance matrix of $\boldsymbol{\Sigma}_{\mathbf{X}}$ and its connection with $\boldsymbol{\beta}$. In the envelope model, since $\boldsymbol{\beta}$ is related only to the material part of $\mathbf{X}$, identifying the immaterial part of $\mathbf{X}$ is important in subsequent analysis. The inactive predictors make the identification easier through its canonical correlation with the active predictors.

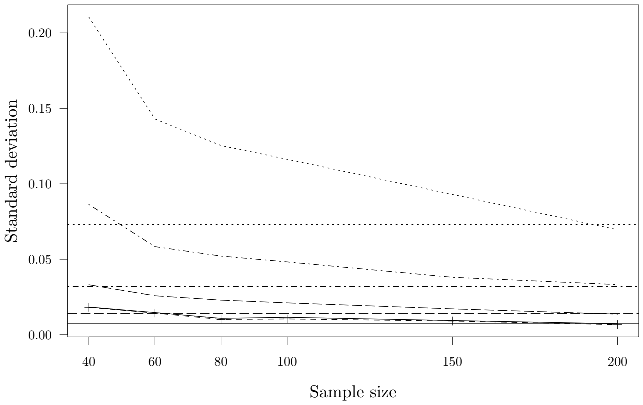

Fig 1. *Comparison of estimation standard deviations. Line $\cdots$ marks the SM estimator. Line $-\cdot-$ marks the SMA estimator. Line $--$ marks the EMA estimator. Line — marks the E-SPLS estimator. Line $--$ with $+$ marks the EMO estimator. The horizontal lines mark the asymptotic standard deviation of the corresponding estimator. (The asymptotic standard deviations of the EMO estimator and the E-SPLS estimator are the same.)*

D.2. E-SPLS estimator under model violation. We investigate the performance of the E-SPLS estimator under model violation. We set $p = 10$, $r = 2$ and $n = 20$. We also let $p_{\mathcal{A}} = p$ so that there is no sparsity in $\boldsymbol{\beta}$. The elements in $\boldsymbol{\beta}$ were independent standard normal variates. The predictor vector $\mathbf{X}$ was normally distributed with mean 0 and covariance matrix $\mathbf{M}_1\mathbf{M}_1^T$, where elements in $\mathbf{M}_1 \in \mathbb{R}^{p\times p}$ were independent uniform $(0,1)$ variates. The errors were sampled from normal distribution with mean 0 and covariance matrix $\mathbf{M}_2\mathbf{M}_2^T$, where elements in $\mathbf{M}_2 \in \mathbb{R}^{r\times r}$ were independent uniform $(0,3)$ variates. The intercept was 0. Note that neither the envelope assumption (3) nor the sparsity assumption (6) holds in this setting. We generated 200 datasets from this setting and computed the OLS, SIMPLS, SPLS and E-SPLS estimators from each dataset. For each model the average of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ was calculated for d varying from 1 to p. The criterion $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ measures the sum of the MSE for all elements in $\boldsymbol{\beta}$. The results are summarized in Figure 2. The E-SPLS estimator has the smallest $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ at all dimensions. The E-SPLS estimator attains its smallest $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ 4.19 at $d = 5$. It achieves 60.43% reduction compared to the OLS estimator ($\|\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}\|_F = 10.59$), 11.21% reduction compared to the SPLS estimator ($\|\widehat{\boldsymbol{\beta}}_{\text{spls}} - \boldsymbol{\beta}\|_F = 4.72$) and 7.04% reduction compared to the SIMPLS estimator($\|\widehat{\boldsymbol{\beta}}_{\text{pls}} - \boldsymbol{\beta}\|_F = 4.51$). Although the E-SPLS estimator has bias in this case, its variance is much smaller than that of the OLS estimator, and overall its MSE is much smaller than that of the OLS estimator.

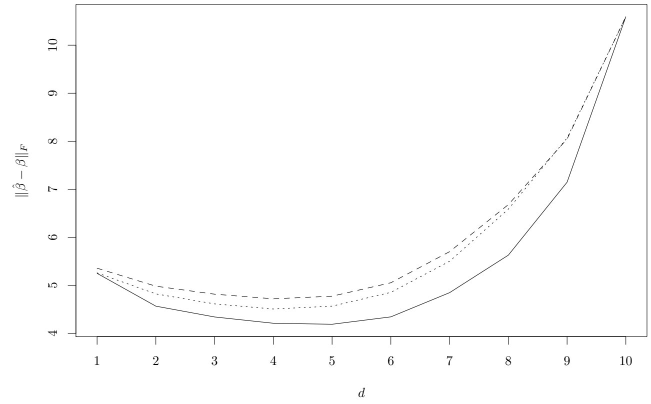

FIG 2. *Line — marks the E-SPLS estimator. Line – – marks the SPLS estimator. Line ··· marks the SIMPLS estimator.*

We also compared the true positive rate (TPR) between SPLS and E-

SPLS. Since we do not have sparsity in the setting, TNR is not relevant. The results are included in Table 1. From this table, we notice that the E-SPLS estimator identifies more active variables than the SPLS estimator for all dimensions.

TABLE 1
TPR for the SPLS estimator and E-SPLS estimator

d	1	2	3	4	5	6	7	8	9	10
SPLS	0.440	0.707	0.740	0.787	0.826	0.862	0.905	0.937	0.977	1.000
E-SPLS	0.962	0.919	0.924	0.952	0.960	0.976	0.973	0.984	0.997	1.000

D.3. E-SPLS estimator with varying r. The scenario of $r \to \infty$ does not change the E-SPLS model, but will change its objective function if $n < r$. We first assume that $n > p$. The original objective function is

$$\begin{aligned}\widehat{\mathbf{A}} = \arg \min_{\mathbf{A} \in \mathbb{R}^{(p-d)\times d}} & -2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}| \\ & + \log|\mathbf{G}_{\mathbf{A}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}| + \lambda\sum_{i=1}^{p-d} w_i\|\mathbf{a}_i\|_2,\end{aligned} \tag{37}$$

where $\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = \mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{XY}}^T$, $\mathbf{S}_{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{Y}}$ are sample covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{S}_{\mathbf{XY}}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. When $r > n$, $\mathbf{S}_{\mathbf{Y}}$ is not invertible. We fitted a multivariate response lasso regression of $\mathbf{X}$ on $\mathbf{Y}$, and computed the residual covariance matrix $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}^*$. Then we replace $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ by $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}^*$ in (37). Therefore, when $n < r$ the objective function becomes

$$\begin{aligned}\widehat{\mathbf{A}} = \arg \min_{\mathbf{A} \in \mathbb{R}^{(p-d)\times d}} & -2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}^*\mathbf{G}_{\mathbf{A}}| \\ & + \log|\mathbf{G}_{\mathbf{A}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}| + \lambda\sum_{i=1}^{p-d} w_i\|\mathbf{a}_i\|_2.\end{aligned} \tag{38}$$

We performed the following simulation to investigate the performance of the E-SPLS estimator with varying r. We set $n = 40$, $p = 10$, $d = 2$, $p_{\mathcal{A}} = 5$ and varied r from 1, 5, 10, 20, 40, 60, 80 and 100. The matrix $\mathbf{\Gamma}_{\mathcal{A}}$ was obtained by orthogonalizing a $p_{\mathcal{A}}$ by d matrix of independent standard normal random variates. The elements in $\boldsymbol{\eta}$ were independent standard normal variates with mean 0 and variance 4. The predictor vector $\mathbf{X}$ followed a multivariate normal distribution with $\boldsymbol{\mu}_{\mathbf{X}} = 0$ and covariance matrix having the structure $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, where $\mathbf{\Omega} = 4\mathbf{I}_d$ and $\mathbf{\Omega}_0$ was a block diagonal matrix with the upper left block being $\mathbf{I}_{p_{\mathcal{A}}-d}$ and lower right block being

$25\mathbf{I}_{p_{\mathcal{I}}}$. The errors were normally distributed with mean 0 and covariance matrix $\mathbf{M}^T\mathbf{M}$, where the elements in $\mathbf{M}$ were independent uniform $(0,2)$ variates. The elements in $\boldsymbol{\mu}$ were independent uniform $(0,1)$ variates. For each value of r, we generated 200 replications and computed $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ for the E-SPLS estimator, the SPLS estimator and the SIMPLS estimator. The average of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ estimates the sum of the mean squared errors from all the elements in $\boldsymbol{\beta}$. The results are summarized in Figure 3.

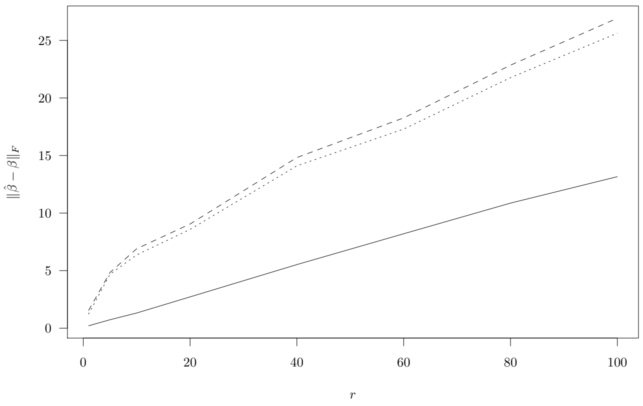

Fig 3. *Comparison of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$. Line — marks the E-SPLS estimator, line – – marks SPLS and line $\cdots$ marks the SIMPLS estimator.*

We notice that the SPLS estimator and the SIMPLS estimator are very close, while the E-SPLS estimator has the smallest $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$. As r increases, we notice that the difference between E-SPLS and SPLS becomes larger. This is because the coefficient matrix $\boldsymbol{\beta}$ has dimension $p \times r$. When r increases, the dimension of $\boldsymbol{\beta}$ becomes larger which amplifies $\|\boldsymbol{\beta}\|_F$. We adjusted for the increase of $\|\boldsymbol{\beta}\|_F$ and plotted $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$ in Figure 4. From Figure 4, we notice that E-SPLS estimator still has the smallest mean squared error adjusted by $\|\boldsymbol{\beta}\|_F$ for all value of r. For example, when $r = 100$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$ is 0.943 for the SIMPLS estimator, 0.991 for the SPLS estimator and 0.485 for the E-SPLS estimator. We also investigated selection performance. Table 2 shows that the E-SPLS estimator has a better selection performance than the SPLS estimator for all values of r in this setting. Also, as r increases, the selection performance of E-SPLS becomes better. This might be because there are more responses that carry the information on which predictors are active and which predictors are inactive. To be more specific, $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$, where $\mathbf{b}_j$ denotes the jth col-

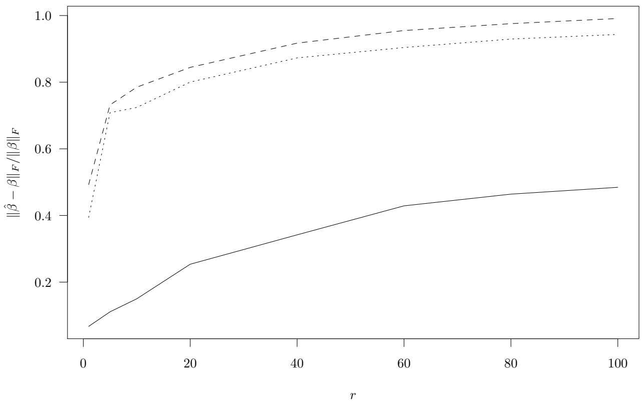

FIG 4. *Comparison of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$. Line — marks the E-SPLS estimator, line – – marks SPLS and line $\cdots$ marks the SIMPLS estimator.*

umn of $\boldsymbol{\beta}$. If the ith predictor is inactive, then the ith element in each $\mathbf{b}_j$ is zero. When we have more columns in $\boldsymbol{\beta}$, it is easier to identify the inactive predictor since the ith element in each $\mathbf{b}_j$, $j = 1, \ldots, r$, is zero.

TABLE 2
Comparison of selection performances of the E-SPLS estimator and the SPLS estimator.

	E-SPLS			SPLS		
r	TPR	TNR	Accuracy	TPR	TNR	Accuracy
1	96.80	47.10	30.50	31.30	74.20	0.00
5	100.00	98.00	92.00	49.50	63.30	0.00
10	100.00	98.90	95.00	57.20	50.90	0.00
20	99.90	100.00	99.50	51.00	44.30	0.00
40	99.50	99.80	96.50	47.50	37.70	0.00
60	99.60	99.90	97.50	34.20	45.20	0.00
80	99.90	99.60	98.00	39.20	37.30	0.00
100	100.00	99.60	98.50	32.00	40.50	0.00

Now we consider the case $n < p$. The matrices $\mathbf{S}^*_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{X}}$ both become singular. The inverse of $\mathbf{S}_{\mathbf{X}}$ is needed in the objective function (38) and the inverse of $\mathbf{S}^*_{\mathbf{X}|\mathbf{Y}}$ is needed in the estimation algorithm. We can use a covariance estimator such as the SPICE estimator (Rothman et al. 2008) to get the inverse of these matrices. We denote the SPICE estimators of the inverses as $\mathbf{S}^{-1}_{\mathbf{X},\text{spice}}$ and ${\mathbf{S}^*_{\mathbf{X}|\mathbf{Y},\text{spice}}}^{-1}$. And $\mathbf{S}^*_{\mathbf{X}|\mathbf{Y},\text{spice}}$ is the inverse of ${\mathbf{S}^*_{\mathbf{X}|\mathbf{Y},\text{spice}}}^{-1}$.

We repeated the simulation that generated Figures 3, 4 and Table 2, but changed p to 60 and the error covariance matrix as $\mathbf{M}^T\mathbf{M}$, where the elements in $\mathbf{M}$ were independent uniform $(0,1)$ variates. We plotted $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F$ and $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$ in Figures 5 and 6. The results are similar to the case of $n > p$: The estimated mean squared errors of the SPLS estimator and the SIMPLS estimator are similar, but the E-SPLS has a much smaller estimated mean squared error. For example, when $r = 100$, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$ is 0.989 for the SIMPLS estimator, 1.001 for the SPLS estimator and 0.297 for the E-SPLS estimator. The selection performance is summarized in Table 3. For all values of r, the E-SPLS estimator has a better selection performance than the SPLS estimator.

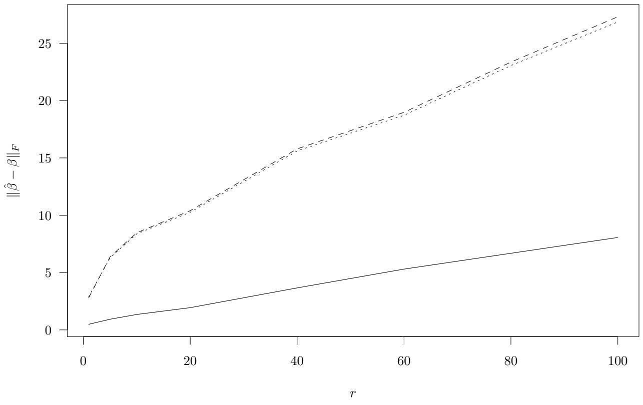

FIG 5. *Comparison of $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F$. Line — marks the E-SPLS estimator, line – – marks the SPLS estimator and line $\cdots$ marks the SIMPLS estimator.*

TABLE 3
Comparison of selection performances of the E-SPLS estimator and the SPLS estimator.

	E-SPLS			SPLS		
r	TPR	TNR	Accuracy	TPR	TNR	Accuracy
1	95.70	87.95	0.50	34.30	60.82	0.00
5	100.00	99.80	89.50	63.70	42.99	0.00
10	100.00	99.84	91.50	65.20	42.07	0.00
20	100.00	99.56	77.50	60.80	44.48	0.00
40	99.90	98.15	33.50	54.50	49.76	0.00
60	99.80	96.76	16.50	47.90	49.04	0.00
80	99.80	97.29	18.00	46.80	50.44	0.00
100	99.70	97.78	31.50	44.20	48.97	0.00

From the numerical results discussed above, we believe that the E-SPLS

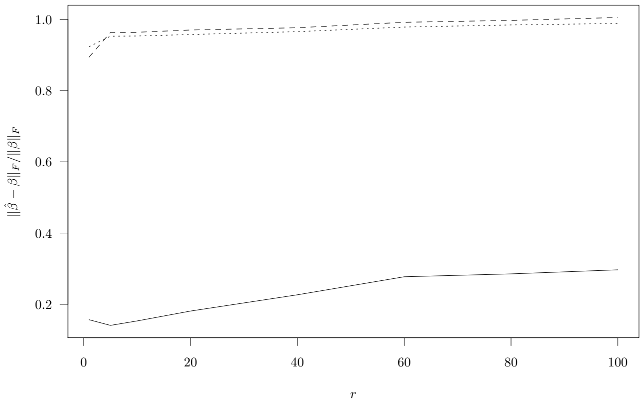

FIG 6. *Comparison of* $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F/\|\boldsymbol{\beta}\|_F$. *Line — marks the E-SPLS estimator, line – – marks the SPLS estimator and line $\cdots$ marks the SIMPLS estimator.*

estimator can maintain its advantage in estimation and variable selection with varying r. The results also indicate that the E-SPLS estimator may enjoy selection consistency when r tends to infinity with n under some mild or moderate conditions.

D.4. E-SGPLS with non-canonical link. We conducted a simulation to investigate the performance of variable selection and efficiency gains of the E-SGPLS estimator under the general link function. We repeated the simulation that produced Figure 7 and Table 5, but changed the link function to probit link, i.e. $P(Y=1 \mid \mathbf{X}) = \Phi(\alpha+\boldsymbol{\beta}^T\mathbf{X})$. We fit the data with both the standard probit model, envelope-based sparse probit model, i.e. model (13) with $g(\cdot)=\Phi^{-1}(\cdot)$, as well as the oracle envelope model, which is model (13) with $g(\cdot)=\Phi^{-1}(\cdot)$ and the extra information on which predictors are active. Under this setting, the estimator of the envelope-based sparse probit model is the E-SGPLS estimator. The standard deviation of the estimator of $\boldsymbol{\beta}$ was calculated based on 200 replications. The standard deviations of a randomly chosen element in $\boldsymbol{\beta}$ are displayed in Figure 7. The results for other elements in $\boldsymbol{\beta}$ follow the same pattern. From the plot, we noticed that the envelope-based sparse probit model achieves substantial efficiency gain compared with the standard probit model. For example, at sample size 1000, the standard deviation of the standard probit model estimator is 3.90 times that of the envelope-based sparse probit model. At sample size 400, the difference between the envelope-based sparse probit model and the oracle envelope model becomes indistinguishable. The variable selection performance of the envelope-based sparse probit model is summarized in Table 4. We noticed that the envelope-based sparse probit model is selection consistent and enjoys the oracle property as indicated in Theorem 6.

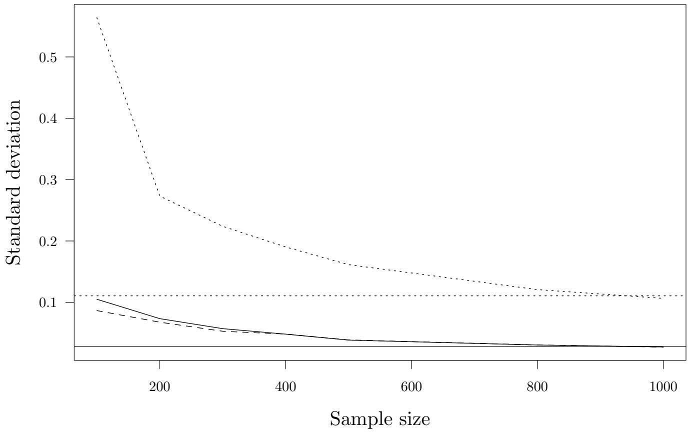

Fig 7. *Comparison of standard deviations of three estimators: Line — marks the envelope-based sparse probit model, line – – marks the oracle envelope model and line ··· marks standard probit model estimator.*

Table 4

TPR, TNR and accuracy of the envelope-based sparse probit model.

n	TPR	TNR	Accuracy
100	94.70	99.10	76.00
200	98.80	99.80	95.50
300	99.40	99.50	98.00
400	100.00	100.00	100.00
500	100.00	100.00	100.00
1000	100.00	100.00	100.00

REFERENCES

Adragni, K. P., Cook, R. D., Wu, S. et al. (2012). GrassmannOptim: An R Package for Grassmann Manifold Optimization. *Journal of Statistical Software* **50** 1–18.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association* **107** 1533–1545.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38** 3696–3723.

Cook, R. D., Forzani, L. and Su, Z. (2016). A note on fast envelope estimation. *J. Multivar. Anal.* **150** 42–54.

Cook, R. D., Helland, I. and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 851–877.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* **20** 927–1010.

Cook, R. D. and Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association* **110** 599–611.

Cox, D. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **49** 1–39.

Dattorro, J. (2016). *Convex Optimization & Euclidean Distance Geometry*. Palo Alto: Meboo Publishing.

Harville, D. A. (1998). Matrix algebra from a statistician's perspective. *Technometrics* **40** 164–164.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58** 30–37.

Huzurbazar, V. (1956). Sufficient statistics and orthogonal parameters. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **17** 217–220.

Li, B., Chun, H. and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association* **107** 152–167.

Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing* **50** 635–650.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Boca Raton: CRC press.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81** 142–149.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical Association* **101** 1418–1429.

Zou, C. and Chen, X. (2012). On the consistency of coordinate-independent sparse estimation with BIC. *Journal of Multivariate Analysis* **112** 248–255.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36** 1509–1533.

Department of Computer Science and Statistics
Tyler 252
University of Rhode Island
Kingston, RI, 02881 E-mail: guangyuzhu@uri.edu

Department of Statistics
102 Griffin-Floyd Hall
University of Florida
Gainesville, FL 32611E-mail: zhihuasu@ufl.edu