Supplemental Materials: Proof of Theorem 1

In this section, we provide the proof of Theorem 1. First recall the outcome Y is modeled as the linear model, $Y = X\Theta B + E$, where $E = [\epsilon_{ik}]$ is $n \times K$ matrix with $\epsilon_{ik}\, i.i.d. \sim N(0, \sigma^2)$. Let $\mathbf{Y} = \mathrm{vec}(Y')$, $\mathbf{\Theta} = \mathrm{vec}(\Theta')$ and $\epsilon = \mathrm{vec}(E')$. The model can be written as,

$$\mathbf{Y} = (X \otimes B')\mathbf{\Theta} + \epsilon. \tag{A.1}$$

It is convenient to centralize X. Let $X_c = [X_{.1} - \bar{X}_{.1}, ..., X_{.P} - \bar{X}_{.P}]$, where $\bar{X}_{.p}$ is the mean of the p-th column of X. Then (A.1) can be written as the centered form,

$$\mathbf{Y} = (X_c \otimes B')\mathbf{\Theta}_{[(J+1):(J(P+1))]} + \mu_{\mathbf{Y}} + \epsilon, \tag{A.2}$$

where $\mathbf{\Theta}_{[(J+1):(J(P+1))]}$ is the $\mathbf{\Theta}$ deleting the first J components, and $\mu_{\mathbf{Y}}$ is the mean vector which does not depend on X_c. Since it is enough to verify the consistency and sparsistency under the centered model (A.2), we slightly abuse the notation as follows. Denote $X_c \otimes B'$ as $\mathbb{X}$, and $\mathbf{\Theta}_{[(J+1):(J(P+1))]}$ as $\mathbf{\Theta}$. Let $\widehat{\mathbf{\Theta}}^A$, $\widehat{\mathbf{\Theta}}^I$ and $\widehat{\mathbf{\Theta}}^0$ be the vectorized estimators without the intercept. And let $\mathbf{\Theta}_p = \Theta'_{p.}$. $\widehat{\mathbf{\Theta}}^A_p$, $\widehat{\mathbf{\Theta}}^I_p$ and $\widehat{\mathbf{\Theta}}^0_p$ are defined similarly.

Proof. First consider the adaptive group lasso estimator $\widehat{\mathbf{\Theta}}^A$. Since $\lambda_n/n \to 0$, according to standard theory of M-estimator, $\widehat{\mathbf{\Theta}}^A \xrightarrow{p} \mathbf{\Theta}$. So we only need to prove the sparsistency.

Let $\mathcal{P}$ be the true nonzero set, i.e., $\mathcal{P} = \{p : \mathbf{\Theta}_p \neq 0\}$. Without loss of generality, assume the first a groups of $\mathbf{\Theta}$ is truly nonzero, that is, $\mathcal{P} = \{1, ..., a\}$. Let $\mathbb{X}_{\mathcal{P}}$ and $\mathbf{\Theta}_{\mathcal{P}}$ be the corresponding components of $\mathbb{X}$ and $\mathbf{\Theta}$ indexed by the nonzero groups, that is $\mathbb{X}_{\mathcal{P}} = X_{c[1:n,1:a]} \otimes B'$, $\mathbf{\Theta}_{\mathcal{P}} = \mathbf{\Theta}_{[1:(aJ)]}$. And define

$$\widetilde{\mathbf{\Theta}}^A_{\mathcal{P}} = \arg\min \|\mathbf{Y} - \mathbb{X}_{\mathcal{P}}\mathbf{\Theta}_{\mathcal{P}}\|^2 + \lambda_n \sum_{p \in \mathcal{P}} w(\widehat{\mathbf{\Theta}}^0_p)\|\mathbf{\Theta}_p\|. \tag{A.3}$$

Since $\lambda_n/n \to 0$, we also have $\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A} \xrightarrow{p} \mathbf{\Theta}_{\mathcal{P}}$. Let

$$\begin{aligned}
\widehat{\Sigma}_{\mathbb{X}\mathbf{Y}} &= \frac{1}{n}\mathbb{X}'\mathbf{Y}, \\
\widehat{\Sigma}_{\mathbb{X}\mathbb{X}} &= \frac{1}{n}\mathbb{X}'\mathbb{X}, \\
\widehat{\Sigma}_{\mathbb{X}\epsilon} &= \frac{1}{n}\mathbb{X}'\epsilon.
\end{aligned}$$

Since $(y_i, x_{i1}, ..., x_{ip})$ are *i.i.d.* and have finite fourth moment, we have $\frac{1}{n}X_c'X_c = \mathrm{Cov}(x)$, where $\mathrm{Cov}(x)$ is the covariance matrix of $(x_{i1}, ..., x_{ip})$. It can be seen that each element of $\widehat{\Sigma}_{\mathbb{X}\mathbb{X}}$ is a linear combination of elements in $\mathrm{Cov}(x)$. Therefore, $\mathbb{E}(\widehat{\Sigma}_{\mathbb{X}\mathbb{X}})$ exists and let $\Sigma_{\mathbb{X}\mathbb{X}} = \mathbb{E}(\widehat{\Sigma}_{\mathbb{X}\mathbb{X}})$. We have

$$\begin{aligned}
\widehat{\Sigma}_{\mathbb{X}\mathbb{X}} &= \Sigma_{\mathbb{X}\mathbb{X}} + O_p(n^{-1/2}), \\
\widehat{\Sigma}_{\mathbb{X}\epsilon} &= O_p(n^{-1/2}).
\end{aligned}$$

With some simple algebra, it is easy to show that $\mathbb{X}'\mu_{\mathbf{Y}} = 0$. Therefore,

$$\begin{aligned}
\widehat{\Sigma}_{\mathbb{X}\mathbf{Y}} &= \frac{1}{n}\mathbb{X}'(\mathbb{X}\mathbf{\Theta} + \mu_{\mathbf{Y}} + \epsilon) = \widehat{\Sigma}_{\mathbb{X}\mathbb{X}}\mathbf{\Theta} + \widehat{\Sigma}_{\mathbb{X}\epsilon} \\
&= (\Sigma_{\mathbb{X}\mathbb{X}} + O_p(n^{-1/2}))\mathbf{\Theta} + O_p(n^{-1/2}) = \Sigma_{\mathbb{X}\mathbb{X}_{\mathcal{P}}}\mathbf{\Theta}_{\mathcal{P}} + O_p(n^{-1/2}),
\end{aligned}$$

where $\Sigma_{\mathbb{X}\mathbb{X}_{\mathcal{P}}}$ are the sub-matrix of $\Sigma_{\mathbb{X}\mathbb{X}}$ with column index in nonzero groups, that is, $\Sigma_{\mathbb{X}\mathbb{X}_{\mathcal{P}}} = \mathbb{E}(\frac{1}{n}\mathbb{X}\mathbb{X}_{\mathcal{P}})$. Furthermore, define $\Sigma_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}} = \mathbb{E}(\frac{1}{n}\mathbb{X}_{\mathcal{P}}'\mathbb{X}_{\mathcal{P}})$, $\mathcal{P}^c = \{(a+1), ..., P\}$ and $\mathbb{X}_{\mathcal{P}^c} = X_{[,(a+1):P]} \otimes B'$. Then $\Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}$, $\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}$, $\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}$, $\widehat{\Sigma}_{\mathbb{X}\mathbb{X}_{\mathcal{P}}}$, $\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbf{Y}}$ and $\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbf{Y}}$ are similarly defined. Therefore,

$$\widehat{\Sigma}_{\mathbb{X}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A} = \Sigma_{\mathbb{X}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A}) + O_p(n^{-1/2}),$$

which implies

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A} = \Sigma_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A}) + O_p(n^{-1/2}), \tag{A.4}$$

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A} = \Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A}) + O_p(n^{-1/2}). \tag{A.5}$$

Define $\Sigma(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A}) = \mathrm{diag}\left(\frac{w(\widehat{\mathbf{\Theta}}_1^0)}{2\|\widetilde{\mathbf{\Theta}}_1^A\|}I_J, ..., \frac{w(\widehat{\mathbf{\Theta}}_d^0)}{2\|\widetilde{\mathbf{\Theta}}_d^A\|}I_J\right)$. Taking derivative of (A.3) with respect to $\mathbf{\Theta}_p$, we obtain the following condition,

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A} = n^{-1}\lambda_n\Sigma(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A})\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{A}. \tag{A.6}$$

Combining (A.4) and (A.6), we have

$$\boldsymbol{\Theta}_{\mathcal{P}} - \widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}} = n^{-1}\lambda_n \Sigma^{-1}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}} \Sigma(\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}})\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}} + O_p(n^{-1/2}).$$

Therefore, from (A.5),

$$\begin{aligned}\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}} &= \Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}(\boldsymbol{\Theta}_{\mathcal{P}} - \widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}}) + O_p(n^{-1/2}) \\ &= n^{-1}\lambda_n \Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}} \Sigma^{-1}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}} \Sigma(\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}})\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}} + O_p(n^{-1/2}). \qquad \text{(A.7)}\end{aligned}$$

Since $\sigma_n \to 0$, $\sigma_n\sqrt{n} \to \infty$, $\widehat{\boldsymbol{\Theta}}^0 = \boldsymbol{\Theta} + O_p(n^{-1/2})$, we have

$$\frac{\|\widehat{\boldsymbol{\Theta}}^0_p\|_\infty}{\sigma_n} \xrightarrow{p} \infty, w(\widehat{\boldsymbol{\Theta}}^0_p) \xrightarrow{p} 0, \text{ for } p \in \mathcal{P}, \qquad \text{(A.8)}$$

$$\frac{\|\widehat{\boldsymbol{\Theta}}^0_p\|_\infty}{\sigma_n} \xrightarrow{p} 0, w(\widehat{\boldsymbol{\Theta}}^0_p) \xrightarrow{p} 1, \text{ for } p \in \mathcal{P}^c. \qquad \text{(A.9)}$$

Since $\lambda_n/\sqrt{n} \to \infty$, for each $p \in \mathcal{P}^c$, combining (A.7), (A.8) and (A.9),

$$\frac{n}{w(\widehat{\boldsymbol{\Theta}}^0_p)\lambda_n}\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}_{\mathcal{P}}}\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}}\| \xrightarrow{p} 0.$$

Therefore,

$$\mathbb{P}\{\forall p \in \mathcal{P}^c, \frac{2n}{w(\widehat{\boldsymbol{\Theta}}^0_p)\lambda_n}\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}_{\mathcal{P}}}\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}}\| \leq 1\} \to 1. \qquad \text{(A.10)}$$

We define a PJ-length vector $\widetilde{\boldsymbol{\Theta}}^A$ as the combination of $\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}}$ and a $(P-a)J$-length vector of zero. Since (A.6) and (A.10), with probability approaching to 1, $\widetilde{\boldsymbol{\Theta}}^A$ satisfies

$$\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^A = n^{-1}\lambda_n\frac{w(\widehat{\boldsymbol{\Theta}}^0_p)}{2\|\widetilde{\boldsymbol{\Theta}}^A_p\|}\widetilde{\boldsymbol{\Theta}}^A_p, \text{ for } p \in \mathcal{P},$$

$$\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^A\| \leq (2n)^{-1}\lambda_n w(\widehat{\boldsymbol{\Theta}}^0_p), \text{ for } p \in \mathcal{P}^c,$$

where $\widetilde{\boldsymbol{\Theta}}^A_p$ is $(p-1)J+1$ to pJ elements of $\widetilde{\boldsymbol{\Theta}}^A_{\mathcal{P}}$. The above condition is exactly the optimality condition for the adaptive group lasso (Yuan & Lin, 2006), which justifies the sparsistency.

Now we consider integrative group lasso estimator $\widehat{\boldsymbol{\Theta}}^I$. Similarly, according to standard theory of M-estimator, $\widehat{\boldsymbol{\Theta}}^I \xrightarrow{p} \boldsymbol{\Theta}$. We only prove the sparsistency below.

Define

$$\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} = \arg\min \|\mathbf{Y} - \mathbb{X}_{\mathcal{P}}\mathbf{\Theta}_{\mathcal{P}}\|^2 + \lambda_n \sum_{p \in P} w(\mathbf{\Theta}_p)\|\mathbf{\Theta}_p\|. \tag{A.11}$$

Similarly to the previous proof, we have

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} = \Sigma_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I}) + O_p(n^{-1/2}), \tag{A.12}$$

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} = \Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I}) + O_p(n^{-1/2}). \tag{A.13}$$

Define $M = [m_{pj}]$ with

$$m_{pj} = \begin{cases} \frac{1}{2}w(\widetilde{\mathbf{\Theta}}_p^I)(||\widetilde{\mathbf{\Theta}}_p^I||)^{-1}\tilde{\theta}_{p[j]}^I & \text{if } |\tilde{\theta}_{p[j]}^I| \neq \|\widetilde{\mathbf{\Theta}}_p^I\|_\infty, \\ \frac{1}{2}w(\widetilde{\mathbf{\Theta}}_p^I)(||\widetilde{\mathbf{\Theta}}_p^I||)^{-1}\tilde{\theta}_{p[j]}^I - \frac{1}{2\sigma_n}w(\widetilde{\mathbf{\Theta}}_p^I)||\widetilde{\mathbf{\Theta}}_p^I||\text{sgn}(\tilde{\theta}_{p[j]}^I) & \text{otherwise,} \end{cases}$$

where $\tilde{\theta}_{p[j]}^I$ is the j-th component of $\widetilde{\mathbf{\Theta}}_p^I$ and $\widetilde{\mathbf{\Theta}}_p^I$ is the $(p-1)J+1$ to pJ elements of $\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I$. Define $\mathbf{M}(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I) = \text{vec}(M')$. Taking derivative of (A.11) with respect to $\mathbf{\Theta}_{\mathcal{P}}$, we obtain

$$\widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} = n^{-1}\lambda_n\mathbf{M}(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I). \tag{A.14}$$

From (A.12) and (A.14) we have

$$\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} = n^{-1}\lambda_n\Sigma_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}^{-1}\mathbf{M}(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I) + O_p(n^{-1/2}).$$

Following (A.13),

$$\begin{aligned} \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I} &= \Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}(\mathbf{\Theta}_{\mathcal{P}} - \widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I}) + O_p(n^{-1/2}) \\ &= n^{-1}\lambda_n\Sigma_{\mathbb{X}_{\mathcal{P}^c}\mathbb{X}_{\mathcal{P}}}\Sigma_{\mathbb{X}_{\mathcal{P}}\mathbb{X}_{\mathcal{P}}}^{-1}\mathbf{M}(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I) + O_p(n^{-1/2}). \end{aligned}$$

Note that when $\sigma_n \to 0$, $\mathbf{M}(\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I) \xrightarrow{p} \mathbf{0}$, where $\mathbf{0}$ is a zero vector of length aJ, since $\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^I \xrightarrow{p} \mathbf{\Theta}_{\mathcal{P}}$.

Therefore, since $\lambda_n/\sqrt{n} \to \infty$, for $p \in \mathcal{P}^c$,

$$\frac{n}{\lambda_n}\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I}\| \xrightarrow{p} 0.$$

We obtain,

$$\mathbb{P}\{\forall p \in \mathcal{P}^c, \frac{2n}{\lambda_n}\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}_{\mathcal{P}}}\widetilde{\mathbf{\Theta}}_{\mathcal{P}}^{I}\| \leq 1\} \to 1.$$

Define a PJ-length vector $\widetilde{\boldsymbol{\Theta}}^I$ as the combination of $\widetilde{\Theta}^I_{\mathcal{P}}$ and a $(P-a)J$ length vector of zeros. With probability approaching one, $\widetilde{\boldsymbol{\Theta}}^I$ satisfies the following conditions,

$$\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I = n^{-1}\lambda_n M'_{p.}, \text{ for } p \in \mathcal{P}, \tag{A.15}$$

$$\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I\| \leq (2n)^{-1}\lambda_n, \text{ for } p \in \mathcal{P}^c. \tag{A.16}$$

To finish the proof, we will show in the following that (A.15) and (A.16) are equivalent to the optimality condition of the integrative group lasso. The proof of (A.15) is trivial and omitted. So we only prove the second one in details. The centered integrative group lasso criterion is

$$\arg\min \|\mathbf{Y} - \mathbb{X}\boldsymbol{\Theta}\|^2 + \lambda_n \sum_p w(\boldsymbol{\Theta}_p)\|\boldsymbol{\Theta}_p\|. \tag{A.17}$$

Taking derivative of (A.17) with respect to $\boldsymbol{\Theta}_p$ at 0 in u direction,

$$\begin{aligned}
\left.\frac{d}{dt}\right|_{t=0+} (\boldsymbol{\Theta}'\widehat{\Sigma}_{\mathbb{X}\mathbb{X}}\boldsymbol{\Theta} - 2\widehat{\Sigma}'_{\mathbb{X}\mathbf{Y}}\boldsymbol{\Theta}) &= 2\widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\boldsymbol{\Theta}u' - 2\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}}u', \\
\left.\frac{d}{dt}\right|_{t=0+} (w(\boldsymbol{\Theta}_p)\|\boldsymbol{\Theta}_p\|) &= \lim_{t\to 0}\frac{w(ut)\|ut\|}{t} = \lim_{t\to 0} w(ut) = 1.
\end{aligned}$$

$\widetilde{\boldsymbol{\Theta}}^I$ is the minimizer only if for any direction u, the directional derivative is greater and equal to 0, which is equivalent to

$$\min_u(2\widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I u' - 2\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}}u' + n^{-1}\lambda_n) \geq 0. \tag{A.18}$$

Since the left hand side of (A.18) is minimized as $-2\|\widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}} - \widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I\| + n^{-1}\lambda_n$ when $u = -(\widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I - \widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}})/\|\widehat{\Sigma}_{\mathbb{X}_p\mathbb{X}}\widetilde{\boldsymbol{\Theta}}^I - \widehat{\Sigma}_{\mathbb{X}_p\mathbf{Y}}\|$, (A.18) is equivalent to (A.16).

□