Supplementary Materials of "Response Variable Selection in Multivariate Linear Regression"

Kshitij Khare and Zhihua Su

University of Florida

Supplementary material includes proofs of all propositions and theorems, implementation details, additional simulations and discussion of future research directions.

We start by introducing some required notation. Let $\mathbb{Y}_{\mathcal{D}} \in \mathbb{R}^{n \times r_{\mathcal{D}}}$, $\mathbb{Y}_{\mathcal{A}} \in \mathbb{R}^{n \times r_{\mathcal{A}}}$, $\mathbb{Y}_{\mathcal{S}} \in \mathbb{R}^{n \times r_{\mathcal{S}}}$ and $\mathbb{X} \in \mathbb{R}^{n \times p}$ denote the data matrices of $\mathbf{Y}_{\mathcal{D}}$, $\mathbf{Y}_{\mathcal{A}}$, $\mathbf{Y}_{\mathcal{S}}$ and $\mathbf{X}$. For example, the ith row of $\mathbb{Y}_{\mathcal{D}} \in \mathbb{R}^{n \times r_{\mathcal{D}}}$ contains the ith observation of $\mathbf{Y}_{\mathcal{D}}$. Let $\mathbb{E}_{\mathcal{D}} \in \mathbb{R}^{n \times r_{\mathcal{D}}}$, $\mathbb{E}_{\mathcal{A}} \in \mathbb{R}^{n \times r_{\mathcal{A}}}$ and $\mathbb{E}_{\mathcal{S}} \in \mathbb{R}^{n \times r_{\mathcal{S}}}$ denote the data matrices of $\boldsymbol{\varepsilon}_{\mathcal{D}}$, $\boldsymbol{\varepsilon}_{\mathcal{A}}$, $\boldsymbol{\varepsilon}_{\mathcal{S}}$, and $\mathbf{1}_n$ denotes an n-dimensional vector of 1's. Then $\mathbb{Y}_{\mathcal{D}} = \mathbf{1}_n^T \boldsymbol{\alpha}_{\mathcal{D}} + \mathbb{X}\boldsymbol{\beta}_{\mathcal{D}}^T + \mathbb{E}_{\mathcal{D}}$ and $\mathbb{Y}_{-\mathcal{D}} = \mathbf{1}_n^T \boldsymbol{\alpha}_{-\mathcal{D}} + \mathbb{E}_{-\mathcal{D}}$, where $\mathbb{Y}_{-\mathcal{D}} = \left[\mathbb{Y}_{\mathcal{A}}^T \, \mathbb{Y}_{\mathcal{S}}^T\right]^T$, $\mathbb{E}_{-\mathcal{D}} = [\mathbb{E}_{\mathcal{A}} \, \mathbb{E}_{\mathcal{S}}]$, and $\boldsymbol{\alpha}_{\mathcal{D}} \in \mathbb{R}^{r_{\mathcal{D}}}$ and $\boldsymbol{\alpha}_{-\mathcal{D}} \in \mathbb{R}^{r_{\mathcal{A}}}$ are parts of $\boldsymbol{\alpha}$ that correspond to the dynamic responses and non-dynamic responses respectively. Let $\mathbb{Y}_{\mathcal{D},c} \in \mathbb{R}^{n \times r_{\mathcal{D}}}$, $\mathbb{Y}_{\mathcal{A},c} \in \mathbb{R}^{n \times r_{\mathcal{A}}}$, $\mathbb{Y}_{\mathcal{S},c} \in \mathbb{R}^{n \times r_{\mathcal{S}}}$ and $\mathbb{X}_c \in \mathbb{R}^{n \times p}$ denote the centered data matrices of $\mathbf{Y}_{\mathcal{D}}$, $\mathbf{Y}_{\mathcal{A}}$, $\mathbf{Y}_{\mathcal{S}}$ and $\mathbf{X}$, i.e. $\mathbb{Y}_{\mathcal{D},c} = \mathbb{Y}_{\mathcal{D}} - \mathbf{1}_n \bar{\mathbf{Y}}_{\mathcal{D}}^T$, $\mathbb{Y}_{\mathcal{A},c} = \mathbb{Y}_{\mathcal{A}} - \mathbf{1}_n \bar{\mathbf{Y}}_{\mathcal{A}}^T$, $\mathbb{Y}_{\mathcal{S},c} = \mathbb{Y}_{\mathcal{S}} - \mathbf{1}_n \bar{\mathbf{Y}}_{\mathcal{S}}^T$ and $\mathbb{X}_c = \mathbb{X} - \mathbf{1}_n \bar{\mathbf{X}}^T$, where $\bar{\mathbf{Y}}_{\mathcal{D}}$, $\bar{\mathbf{Y}}_{\mathcal{A}}$, $\bar{\mathbf{Y}}_{\mathcal{S}}$ and $\bar{\mathbf{X}}$ denote the sample mean of $\mathbf{Y}_{\mathcal{D}}$,

$\mathbf{Y}_{\mathcal{A}}$, $\mathbf{Y}_{\mathcal{S}}$ and $\mathbf{X}$. Let $\mathbb{E}_{\mathcal{D},c}$ and $\mathbb{E}_{-\mathcal{D},c}$ denote the centered data matrix of $\mathbb{E}_{\mathcal{D}}$ and $\mathbb{E}_{-\mathcal{D}}$, then $\mathbb{Y}_{\mathcal{D},c} = \mathbb{X}_c\boldsymbol{\beta}_{\mathcal{D}}^T + \mathbb{E}_{\mathcal{D},c}$ and $\mathbb{Y}_{-\mathcal{D},c} = \mathbb{E}_{-\mathcal{D},c}$. Let $\mathbf{P}_{\mathbf{C}}$ denote the projection matrix into the column space $\mathbf{C}$ and $\mathbf{Q}_{\mathbf{C}} = \mathbf{I} - \mathbf{P}_{\mathbf{C}}$. Let $\|\cdot\|_2$, $\|\cdot\|_F$ and $\|\cdot\|_{\max}$ denote the spectral norm, Frobenius norm and entry-wise maximum norm of a matrix, respectively.

S1. Proof of Proposition 2

Proof. From Proposition 1, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ is given by

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_1), \quad \mathbf{V}_1 = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}).$$

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ is given by

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_2), \quad \mathbf{V}_2 = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},(\mathcal{A},\mathcal{S})}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})}^{-1}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S}),\mathcal{D}}).$$

We hope to prove that $\mathbf{V}_1 = \mathbf{V}_2$. Let $\mathbf{D} = \left(\mathbf{\Sigma}_{\mathcal{S}} - \mathbf{\Sigma}_{\mathcal{S},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}}\right)^{-1}$. Note that

$$
\begin{aligned}
& (\mathbf{\Sigma}_{\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{D}}) - (\mathbf{\Sigma}_{\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},(\mathcal{A},\mathcal{S})}\mathbf{\Sigma}_{(\mathcal{A},\mathcal{S})}^{-1}\mathbf{\Sigma}_{(\mathcal{A},\mathcal{S}),\mathcal{D}}) \\
= & \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}} & \mathbf{\Sigma}_{\mathcal{D},\mathcal{S}} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{A}} & \mathbf{\Sigma}_{\mathcal{A},\mathcal{S}} \\ \mathbf{\Sigma}_{\mathcal{S},\mathcal{A}} & \mathbf{\Sigma}_{\mathcal{S}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{A},\mathcal{D}} \\ \mathbf{\Sigma}_{\mathcal{S},\mathcal{D}} \end{pmatrix} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{D}} \\
= & \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}}^{-1} + \mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}}\mathbf{D}\mathbf{\Sigma}_{\mathcal{S},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1})\mathbf{\Sigma}_{\mathcal{A},\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}}\mathbf{D}\mathbf{\Sigma}_{\mathcal{S},\mathcal{D}} \\
& -\mathbf{\Sigma}_{\mathcal{D},\mathcal{S}}\mathbf{D}\mathbf{\Sigma}_{\mathcal{S},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{D}} + \mathbf{\Sigma}_{\mathcal{D},\mathcal{S}}\mathbf{D}\mathbf{\Sigma}_{\mathcal{S},\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{D}} \\
= & (\mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{S}})\mathbf{D}(\mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{S}})^T.
\end{aligned}
$$

Since $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid \mathbf{Y}_{\mathcal{A}}$, we have $\mathbf{\Sigma}_{\mathcal{D},\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A},\mathcal{S}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{S}} = 0$. Thus $\mathbf{V}_1 = \mathbf{V}_2$.

An alternative proof uses an equivalence stated in §3.1 in Dawid (1979), that $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid \mathbf{Y}_{\mathcal{A}}$ implies $\mathbf{Y}_{\mathcal{D}}$ given $\mathbf{Y}_{\mathcal{A}}$ and $\mathbf{Y}_{\mathcal{D}}$ given $(\mathbf{Y}_{\mathcal{A}}, \mathbf{Y}_{\mathcal{S}})$ have the same distribution. Note that $\mathbf{V}_1$ is the covariance matrix of $\mathbf{Y}_{\mathcal{D}}$ given $\mathbf{Y}_{\mathcal{A}}$, and $\mathbf{V}_2$ is the covariance matrix of $\mathbf{Y}_{\mathcal{D}}$ given $(\mathbf{Y}_{\mathcal{A}}, \mathbf{Y}_{\mathcal{S}})$. Thus $\mathbf{V}_1 = \mathbf{V}_2$.

S2. A generalization of Propositions 1 and 2

In this section we present a generalization of Propositions 1 and 2 for the setting when $r = r_n \to \infty$ as $n \to \infty$, but the number of dynamic variables $r_{\mathcal{D}}$ and the number of predictors p remains fixed. In this setting, the error matrix $\mathbb{E}_c = [\mathbb{E}_{\mathcal{D},c}\, \mathbb{E}_{\mathcal{A},c}]$ and the corresponding error covariance matrix $\mathbf{\Sigma}$

depend on n, but this dependence is suppressed for simplicity of notation.

Proposition S2. *Consider an asymptotic regime discussed above, where $r_{\mathcal{D}}$ and p remain fixed, but r is allowed to grow with n. Assume that the errors are normally distributed in models (2.2), (2.3), (2.4). Assume that $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{S}$ are given. Let $\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}}\boldsymbol{\Sigma}_{-\mathcal{D}}^{-1}\boldsymbol{\Sigma}_{-\mathcal{D},\mathcal{D}}$ and $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{A}} = \boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}$.*

(a) Suppose $r = o(n)$ and the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded (in n) away from zero and infinity. Then the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ (the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under (2.2)) is given by

$$\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}})\left(\mathbb{X}_c^T\mathbb{X}_c\right)^{1/2} \xrightarrow{d} MN_{r_{\mathcal{D}}\times p}\left(\mathbf{0}, \mathbf{I}_{r_{\mathcal{D}}}, \mathbf{I}_p\right),$$

and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ (the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under (2.4) with $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$) is given by

$$\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{A}}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},2} - \boldsymbol{\beta}_{\mathcal{D}})\left(\mathbb{X}_c^T\mathbb{X}_c\right)^{1/2} \xrightarrow{d} MN_{r_{\mathcal{D}}\times p}\left(\mathbf{0}, \mathbf{I}_{r_{\mathcal{D}}}, \mathbf{I}_p\right).$$

Here $MN_{r_{\mathcal{D}}\times p}$ denotes the matrix normal distribution on the space of $r_{\mathcal{D}} \times p$ matrices.

(b) Suppose $r_{\mathcal{A}} = o(n)$ and the eigenvalues of the principal (top) $r_{\mathcal{D}} + r_{\mathcal{A}}$ submatrix of $\boldsymbol{\Sigma}$ are uniformly bounded (in n) away from zero and infinity. Then the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ (the maximum likelihood

estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under (2.3)) is given by

$$\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{A}}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}) \left(\mathbb{X}_c^T \mathbb{X}_c\right)^{1/2} \xrightarrow{d} MN_{r_{\mathcal{D}} \times p}\left(\mathbf{0}, \mathbf{I}_{r_{\mathcal{D}}}, \mathbf{I}_p\right).$$

To see that Proposition 1 follows as a special case of this proposition when r is fixed, note that $\frac{1}{n}\mathbb{X}_c^T\mathbb{X}_c \xrightarrow{P} \boldsymbol{\Sigma}_{\mathbf{X}}$ and

- $\boldsymbol{\theta} \sim MN(\boldsymbol{\mu}, \mathbf{U}, \mathbf{V}) \Rightarrow \mathbf{D}\boldsymbol{\theta}\mathbf{C} \sim MN(\mathbf{D}\boldsymbol{\mu}\mathbf{C}, \mathbf{D}\mathbf{U}\mathbf{D}^T, \mathbf{C}^T\mathbf{V}\mathbf{C})$.
- $\boldsymbol{\theta} \sim MN(\boldsymbol{\mu}, \mathbf{U}, \mathbf{V}) \Rightarrow \text{vec}(\boldsymbol{\theta}) \sim N(\text{vec}(\boldsymbol{\mu}), \mathbf{V} \otimes \mathbf{U})$.

PROOF. Note that $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}} = \mathbb{Y}_{\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1} = \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1}$ and $\widetilde{\boldsymbol{\beta}}_{-\mathcal{D}} = \mathbb{Y}_{-\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1} = \mathbb{E}_{-\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1}$. The sample residual $\mathbf{R}_{\mathcal{D}}$ from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$ is $\mathbf{R}_{\mathcal{D}} = \mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{\mathcal{D},c}$ and the sample residual $\mathbf{R}_{-\mathcal{D}}$ from the regression of $\mathbf{Y}_{-\mathcal{D}}$ on $\mathbf{X}$ is $\mathbf{R}_{-\mathcal{D}} = \mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{-\mathcal{D},c}$. The matrix $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|-\mathcal{D}}$ contains the coefficients from the regression of $\mathbf{R}_{\mathcal{D}}$ on $\mathbf{R}_{-\mathcal{D}}$, i.e. $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|-\mathcal{D}} = \mathbb{E}_{\mathcal{D},c}^T\mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{-\mathcal{D},c}(\mathbb{E}_{-\mathcal{D},c}^T\mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{-\mathcal{D},c})^{-1}$. So $\widehat{\boldsymbol{\beta}}_{\mathcal{D}} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|-\mathcal{D}}\widetilde{\boldsymbol{\beta}}_{-\mathcal{D}} = \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c}^T\mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{-\mathcal{D},c}(\mathbb{E}_{-\mathcal{D},c}^T\mathbf{Q}_{\mathbb{X}_c}\mathbb{E}_{-\mathcal{D},c})^{-1}\mathbb{E}_{-\mathcal{D},c}^T\mathbb{X}_c(\mathbb{X}_c^T\mathbb{X}_c)^{-1}$. Note that

$$\mathbb{E} = [\mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{-\mathcal{D}}] \sim MN(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma})$$

and

$$\mathbb{E}_c \sim MN\left(\mathbf{0}, \mathbf{I}_n - 2\frac{\mathbf{J}_n}{n} + \frac{\mathbf{J}_n^2}{n^2}, \boldsymbol{\Sigma}\right),$$

where $\mathbf{J}_n$ is the $n \times n$ matrix with all entries 1. Using the properties of the matrix normal distribution listed above, it follows that

$$(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}) \left(\mathbb{X}_c^T \mathbb{X}_c\right)^{1/2} = \left(\tilde{\mathbb{E}}_{c,\mathcal{D}}^T - \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\right) + \left(\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \hat{\mathbf{C}}\right) \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T \tag{S2.1}$$

where

$$\tilde{\mathbb{E}}_c^T = \mathbb{E}_c^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1/2} \sim MN(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_p) \tag{S2.2}$$

(since $\mathbb{X}_c^T \mathbf{J}_n = \mathbf{0}$), and

$$\hat{\mathbf{C}} = \left(\frac{\mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{-\mathcal{D},c}}{n}\right) \left(\frac{\mathbb{E}_{-\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{-\mathcal{D},c}}{n}\right)^{-1}.$$

For arbitrary $\mathbf{a} \in \mathbb{R}^{r_{\mathcal{D}}}$ and $\mathbf{b} \in \mathbb{R}^p$ satisfying $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$, we have

$$\begin{aligned}
& \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} (\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}) \left(\mathbb{X}_c^T \mathbb{X}_c\right)^{1/2} \mathbf{b} \\
= \; & \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} \left(\tilde{\mathbb{E}}_{c,\mathcal{D}}^T - \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\right) \mathbf{b} + \\
& \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} \left(\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \hat{\mathbf{C}}\right) \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T \mathbf{b} \\
= \; & Z + W,
\end{aligned} \tag{S2.3}$$

where

$$\begin{aligned}
Z &= \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} \left(\tilde{\mathbb{E}}_{c,\mathcal{D}}^T - \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\right) \mathbf{b} \\
&= \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} \left[\mathbf{I}_{r_{\mathcal{D}}} - \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1}\right] \tilde{\mathbb{E}}_c^T \mathbf{b}
\end{aligned} \tag{S2.4}$$

and

$$W = \mathbf{a}^T \boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2} \left(\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \hat{\mathbf{C}}\right) \tilde{\mathbb{E}}_{c,-\mathcal{D}}^T \mathbf{b}. \tag{S2.5}$$

Using (S2.2), (S2.3), (S2.4) and properties of the matrix normal distribution listed above, it follows that

$$Z \sim N\left(0, (\mathbf{a}^T\mathbf{a})(\mathbf{b}^T\mathbf{b})\right) = N(0,1).$$

If we show that $W \xrightarrow{P} 0$, then it follows that

$$\mathbf{a}^T\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}})\left(\mathbb{X}_c^T\mathbb{X}_c\right)^{1/2}\mathbf{b} \xrightarrow{d} N(0,1)$$

for arbitrary $\mathbf{a} \in \mathbb{R}^{r_{\mathcal{D}}}$ and $\mathbf{b} \in \mathbb{R}^p$ with $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$. Using the Cramer-Wold device gives the required result for $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$.

We now complete this final step of the proof, i.e., show that $W \xrightarrow{P} 0$. Note by (S2.5) that

$$|W| \leq \|\mathbf{a}\|_2 \left\|\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2}\right\|_2 \left\|\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}}\boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \hat{\mathbf{C}}\right\|_2 \left\|\tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\mathbf{b}\right\|_2. \tag{S2.6}$$

Since $\|\mathbf{a}\|_2 = 1$, $\tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\mathbf{b} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{-\mathcal{D}}\right)$, and the eigenvalues of $\boldsymbol{\Sigma}$ (hence those of $\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}$ and $\boldsymbol{\Sigma}_{-\mathcal{D}}$) are uniformly bounded away from zero and infinity, it follows that

$$\|\mathbf{a}\|_2 \left\|\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}}^{-1/2}\right\|_2 \left\|\tilde{\mathbb{E}}_{c,-\mathcal{D}}^T\mathbf{b}\right\|_2 = O_P(1). \tag{S2.7}$$

For any $\mathbf{v} \in \mathbb{R}^r$ with $\|\mathbf{v}\|_2 \leq 1$, we have $\mathbb{E}\mathbf{v} \sim N\left(\mathbf{0}, \left(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}\right)\mathbf{I}_n\right)$. Since the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded above and below, it follows by the Hanson-Wright inequality (Rudelson and Vershynin, 2013, Theorem

1.1) that

$$P\left(\left|\mathbf{v}^T\left(\frac{\mathbb{E}^T\mathbb{E}}{n}-\mathbf{\Sigma}\right)\mathbf{v}\right|>\eta\right)\leq 2\exp\left[-C_0 n\min(\eta^2,\eta)\right] \tag{S2.8}$$

for an appropriate constant C_0 which does not depend on n. Using (Ghosh et al., 2018, Lemma B.2), it follows that

$$P\left(\left\|\frac{\mathbb{E}^T\mathbb{E}}{n}-\mathbf{\Sigma}\right\|_2>\eta\right)\leq 2\exp\left[-C_0 n\min(\eta^2,\eta)+2r\log 21\right]. \tag{S2.9}$$

Using $\eta=\sqrt{\frac{3r\log 21}{C_0 n}}$ and $r=o(n)$, we get

$$\left\|\frac{\mathbb{E}^T\mathbb{E}}{n}-\mathbf{\Sigma}\right\|_2=O_P\left(\sqrt{\frac{r}{n}}\right)=o_P(1). \tag{S2.10}$$

Hence

$$\begin{aligned}\left\|\frac{\mathbb{E}_c^T\mathbb{E}_c}{n}-\mathbf{\Sigma}\right\|_2 &= \left\|\frac{\mathbb{E}^T\mathbb{E}}{n}-\mathbf{\Sigma}-\frac{\mathbb{E}^T\mathbf{J}_n\mathbb{E}}{n^2}\right\|_2\\ &\leq \left\|\frac{\mathbb{E}^T\mathbb{E}}{n}-\mathbf{\Sigma}\right\|_2+\left\|\frac{\mathbb{E}^T\mathbf{J}_n\mathbb{E}}{n^2}\right\|_2\\ &\leq O_P\left(\sqrt{\frac{r}{n}}\right)+\frac{1}{n^2}\mathbf{1}_n^T\mathbb{E}\mathbb{E}^T\mathbf{1}_n\\ &= O_P\left(\sqrt{\frac{r}{n}}\right)+O_P\left(\frac{r}{n}\right).\end{aligned} \tag{S2.11}$$

The last equality follows from the fact that

$$\frac{1}{n}\mathbf{1}_n^T\mathbb{E}\mathbf{\Sigma}^{-1}\mathbb{E}^T\mathbf{1}_n\sim\chi_r^2,$$

the eigenvalues of $\mathbf{\Sigma}$ are uniformly bounded above and below, and $\chi_r^2=O_P(r)$.

Let $\tilde{\mathbb{X}}_c = \mathbb{X} - \mathbf{1}_n \boldsymbol{\mu}_{\mathbf{X}}^T$. By using the fact that p is fixed, $\mathbb{X}$ and $\mathbb{E}$ are independent (with matrix normal distributions specified above), the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded above and below, and the Hanson-Wright inequality (Rudelson and Vershynin, 2013, Theorem 1.1) (similar to (S2.8)), we get

$$P\left(\left|\mathbf{u}^T\left(\frac{\tilde{\mathbb{X}}_c^T\tilde{\mathbb{X}}_c}{n} - \boldsymbol{\Sigma}_{\mathbf{X}}\right)\mathbf{u}\right| > \eta\right) \leq 2\exp\left[-C_1 n \min(\eta^2, \eta)\right] \tag{S2.12}$$

for every $\mathbf{u} \in \mathbb{R}^p$ with $\|\mathbf{u}\|_2 \leq 1$, and an appropriate constant C_1 which does not depend on n, and

$$P\left(\left|\frac{(\tilde{\mathbb{X}}_c\mathbf{u} + \mathbb{E}\mathbf{v})^T(\tilde{\mathbb{X}}_c\mathbf{u} + \mathbb{E}\mathbf{v})}{n} - \mathbf{u}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u} - \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}\right| > \eta\right) \leq 2\exp\left[-C_2 n \min(\eta^2, \eta)\right] \tag{S2.13}$$

for every $\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^r$ with $\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1$, and an appropriate constant C_2 which does not depend on n. It follows from (S2.8), (S2.12) and (S2.13) that

$$P\left(\left|\mathbf{u}^T\left(\frac{\tilde{\mathbb{X}}_c^T\mathbb{E}}{n}\right)\mathbf{v}\right| > 3\eta\right) \leq 6\exp\left[-C_3 n \min(\eta^2, \eta)\right] \tag{S2.14}$$

where $C_3 = \min(C_0, C_1, C_2)$. Using (Ghosh et al., 2018, Lemma B.2), it follows that

$$P\left(\left\|\frac{\tilde{\mathbb{X}}_c^T\mathbb{E}}{n}\right\|_2 > 3\eta\right) \leq 6\exp\left[-C_3 n \min(\eta^2, \eta) + (r+p)\log 21\right]. \tag{S2.15}$$

Using $\eta = \sqrt{\frac{2r\log 21}{C_3 n}}$, $r = o(n)$ and the fact that p is fixed, we get

$$\left\| \frac{\tilde{\mathbb{X}}_c^T \mathbb{E}}{n} \right\|_2 = O_P\left(\sqrt{\frac{r}{n}}\right) = o_P(1). \tag{S2.16}$$

Note that $\mathbb{X}_c - \tilde{\mathbb{X}}_c = \mathbf{1}_n(\boldsymbol{\mu}_{\mathbf{X}} - \bar{\mathbf{X}})^T$, it follows that $\|\mathbb{X}_c - \tilde{\mathbb{X}}_c\|_2 = \sqrt{n}\|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\|_2 = O_P(1)$ (since p is fixed). Note from (S2.11) that $\left\| \frac{\mathbb{E}_c^T \mathbb{E}_c}{n} \right\|_2 = O_P(1)$. It follows from (S2.16) that

$$\begin{aligned}
\left\| \frac{\mathbb{X}_c^T \mathbb{E}_c}{n} \right\|_2 &= \left\| \frac{\mathbb{X}_c^T \mathbb{E}}{n} \right\|_2 \\
&\leq \left\| \frac{\tilde{\mathbb{X}}_c^T \mathbb{E}}{n} \right\|_2 + \left\| \frac{(\mathbb{X}_c - \tilde{\mathbb{X}}_c)^T \mathbb{E}_c}{n} \right\|_2 \\
&\leq \left\| \frac{\tilde{\mathbb{X}}_c^T \mathbb{E}}{n} \right\|_2 + \frac{1}{n} \left\| \mathbb{X}_c - \tilde{\mathbb{X}}_c \right\|_2 \|\mathbb{E}_c\|_2 \\
&= O_P\left(\sqrt{\frac{r}{n}}\right).
\end{aligned} \tag{S2.17}$$

Combining (S2.11), (S2.17) along with the fact that $\mathbb{X}^T\mathbb{X}/n \xrightarrow{P} \boldsymbol{\Sigma}_{\mathbf{X}}$, we get

$$\left\| \frac{1}{n}\mathbb{E}_c^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_c - \boldsymbol{\Sigma} \right\|_2 \leq \left\| \frac{1}{n}\mathbb{E}_c^T \mathbb{E}_c - \boldsymbol{\Sigma} \right\|_2 + \left\| \frac{\mathbb{E}_c^T \mathbb{X}_c}{n} \left(\frac{\mathbb{X}_c^T \mathbb{X}_c}{n} \right)^{-1} \frac{\mathbb{X}_c^T \mathbb{E}_c}{n} \right\|_2 = O_P\left(\sqrt{\frac{r}{n}}\right). \tag{S2.18}$$

Since the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded above and below, and $r = o(n)$, it follows that

$$\left\| \left(\frac{1}{n}\mathbb{E}_c^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_c \right)^{-1} - \boldsymbol{\Sigma}^{-1} \right\|_2 = O_P\left(\sqrt{\frac{r}{n}}\right).$$

Finally, using the form of the inverse of a partitioned matrix, we conclude that

$$\begin{aligned} & \left\| \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \hat{\mathbf{C}} \right\|_2 \\ = & \left\| \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \boldsymbol{\Sigma}_{-\mathcal{D}}^{-1} - \left(\frac{\mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{-\mathcal{D},c}}{n} \right) \left(\frac{\mathbb{E}_{-\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{-\mathcal{D},c}}{n} \right)^{-1} \right\|_2 \\ = & O_P\left(\sqrt{\frac{r}{n}} \right). \end{aligned}$$

The desired result for $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ now follows from (S2.6) and (S2.7). The result for $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ follows by noting that $\mathbf{Y}_{-\mathcal{D}} = (\mathbf{Y}_{\mathcal{A}}^T, \mathbf{Y}_{\mathcal{S}}^T)^T$. If $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$, then as shown in Section S1 above, $\boldsymbol{\Sigma}_{\mathcal{D}|-\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{A}}$. The result for $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ in part (b) follows by going through the arguments for the $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ proof above verbatim - but replacing $-\mathcal{D}$ by $\mathcal{A}$ in all relevant places.

S3. Proof of Proposition 3

PROOF. We first look into $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$. Note that $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}} = \mathbb{Y}_{\mathcal{D},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} = \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1}$ and $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} = \mathbb{Y}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} = \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1}$. The sample residual $\mathbf{R}_{\mathcal{D}}$ from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$ is $\mathbf{R}_{\mathcal{D}} = \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{D},c}$ and the sample residual $\mathbf{R}_{\mathcal{A}}$ from the regression of $\mathbf{Y}_{\mathcal{A}}$ on $\mathbf{X}$ is $\mathbf{R}_{\mathcal{A}} = \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c}$. The matrix $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}}$ contains the coefficients from the regression of $\mathbf{R}_{\mathcal{D}}$ on $\mathbf{R}_{\mathcal{A}}$, i.e. $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}} = \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1}$. So $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}} = \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1}$.

On the other hand, $\widehat{\boldsymbol{\beta}}_{\mathcal{D},3} = \mathbf{R}_{\mathcal{D}|\mathcal{A}}^T \mathbf{R}_{\mathbf{X}|\mathcal{A}} (\mathbf{R}_{\mathbf{X}|\mathcal{A}}^T \mathbf{R}_{\mathbf{X}|\mathcal{A}})^{-1}$, where $\mathbf{R}_{\mathbf{X}|\mathcal{A}} = \mathbf{Q}_{\mathbb{Y}_{\mathcal{A},c}} \mathbb{X}_c = \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c$, $\mathbf{R}_{\mathcal{D}|\mathcal{A}} = \mathbf{Q}_{\mathbb{Y}_{\mathcal{A},c}} \mathbb{Y}_{\mathcal{D},c} = \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{Y}_{\mathcal{D},c}$. Therefore $\widehat{\boldsymbol{\beta}}_{\mathcal{D},3} = \mathbb{Y}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c)^{-1} = \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c)^{-1}$.

We focus on $\widehat{\boldsymbol{\beta}}_{\mathcal{D},3}$ first. By Woodbury equality, we have

$$
\begin{aligned}
& (\mathbb{X}_c^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c)^{-1} \\
= \; & \left[\mathbb{X}_c^T \mathbb{X}_c - \mathbb{X}_c^T \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c \right]^{-1} \\
= \; & (\mathbb{X}_c^T \mathbb{X}_c)^{-1} + (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \mathbb{X}_c^T \mathbb{E}_{\mathcal{A},c} \left[\mathbb{E}_{\mathcal{A},c}^T \mathbb{E}_{\mathcal{A},c} - \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \mathbb{X}_c^T \mathbb{E}_{\mathcal{A},c} \right]^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
= \; & (\mathbb{X}_c^T \mathbb{X}_c)^{-1} + (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \mathbb{X}_c^T \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathcal{D},3} &= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbf{P}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} (\mathbf{I} - \mathbf{Q}_{\mathbb{X}_c}) \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c}^T (\mathbf{I} - \mathbf{P}_{\mathbb{E}_{\mathcal{A},c}}) \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&\quad + \mathbb{E}_{\mathcal{D},c}^T \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbf{Q}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c} \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&\quad + \mathbb{E}_{\mathcal{D},c}^T \mathbf{P}_{\mathbb{E}_{\mathcal{A},c}} \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \boldsymbol{\beta}_{\mathcal{D}} + \mathbb{E}_{\mathcal{D},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} - \mathbb{E}_{\mathcal{D},c} \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c} (\mathbb{E}_{\mathcal{A},c}^T \mathbf{Q}_{\mathbb{X}_c} \mathbb{E}_{\mathcal{A},c})^{-1} \mathbb{E}_{\mathcal{A},c}^T \mathbb{X}_c (\mathbb{X}_c^T \mathbb{X}_c)^{-1} \\
&= \widehat{\boldsymbol{\beta}}_{\mathcal{D},1}.
\end{aligned}
$$

S4. Proof of Proposition 4

PROOF. Following the definitions of $\mathbb{X}_c$, $\mathbb{Y}_{\mathcal{A},c}$, $\mathbb{Y}_{\mathcal{D},c}$, $\mathbf{R}_{\mathbf{X}|\mathcal{A}}$ and $\mathbf{R}_{\mathcal{D}|\mathcal{A}}$ in the proof of Proposition 3, the OLS estimator of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ is

$$
(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2) = \mathbb{Y}_{\mathcal{D},c}^T (\mathbb{X}_c, \mathbb{Y}_{\mathcal{A},c}) [(\mathbb{X}_c, \mathbb{Y}_{\mathcal{A},c})^T (\mathbb{X}_c, \mathbb{Y}_{\mathcal{A},c})]^{-1}.
$$

Using the structure on the matrix inverse, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= \mathbb{Y}_{\mathcal{D},c}^T\mathbb{X}_c[\mathbb{X}_c^T\mathbb{X}_c - \mathbb{X}_c^T\mathbb{Y}_{\mathcal{A},c}(\mathbb{Y}_{\mathcal{A},c}^T\mathbb{Y}_{\mathcal{A},c})^{-1}\mathbb{Y}_{\mathcal{A},c}^T\mathbb{X}_c]^{-1} \\
&\quad -\mathbb{Y}_{\mathcal{D},c}^T\mathbb{Y}_{\mathcal{A},c}(\mathbb{Y}_{\mathcal{A},c}^T\mathbb{Y}_{\mathcal{A},c})^{-1}\mathbb{Y}_{\mathcal{A},c}^T\mathbb{X}[\mathbb{X}_c^T\mathbb{X}_c - \mathbb{X}_c^T\mathbb{Y}_{\mathcal{A},c}(\mathbb{Y}_{\mathcal{A},c}^T\mathbb{Y}_{\mathcal{A},c})^{-1}\mathbb{Y}_{\mathcal{A},c}^T\mathbb{X}_c]^{-1} \\
&= \mathbb{Y}_{\mathcal{D},c}^T\mathbb{Y}_{\mathcal{A},c}(\mathbf{R}_{\mathbf{X}|\mathcal{A}}^T\mathbf{R}_{\mathbf{X}|\mathcal{A}})^{-1} - \mathbb{Y}_{\mathcal{D},c}^T\mathbb{Y}_{\mathcal{A},c}(\mathbb{Y}_{\mathcal{A},c}^T\mathbb{Y}_{\mathcal{A},c})^{-1}\mathbb{Y}_{\mathcal{A},c}^T\mathbb{X}(\mathbf{R}_{\mathbf{X}|\mathcal{A}}^T\mathbf{R}_{\mathbf{X}|\mathcal{A}})^{-1} \\
&= \mathbf{R}_{\mathcal{D}|\mathcal{A}}^T\mathbf{R}_{\mathbf{X}|\mathcal{A}}(\mathbf{R}_{\mathbf{X}|\mathcal{A}}^T\mathbf{R}_{\mathbf{X}|\mathcal{A}})^{-1} \\
&= \widehat{\boldsymbol{\beta}}_{\mathcal{D},3}.
\end{aligned}$$

S5. Proof of Proposition 5

Proof. Recall that $\boldsymbol{\Sigma}$ has the partition

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{S}} \\ \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{S}} \\ \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{S}} \end{pmatrix}.$$

We partition $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ into a 2×2 block matrix with respect to $\mathcal{D}$ and $(\mathcal{A}, \mathcal{S})$, then

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}^{-1} & -\boldsymbol{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}^{-1}\boldsymbol{\Sigma}_{\mathcal{D},(\mathcal{A},\mathcal{S})}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})}^{-1} \\ -\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})|\mathcal{D}}^{-1}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S}),\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1} & \boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})|\mathcal{D}}^{-1} \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}.$$

Notice that

$$\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} = \boldsymbol{\Sigma}_{\mathcal{D},(\mathcal{A},\mathcal{S})}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})}^{-1} = \boldsymbol{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}\boldsymbol{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}^{-1}\boldsymbol{\Sigma}_{\mathcal{D},(\mathcal{A},\mathcal{S})}\boldsymbol{\Sigma}_{(\mathcal{A},\mathcal{S})}^{-1} = -\boldsymbol{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}\boldsymbol{\Omega}_{12} = -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}.$$

Since $\mathbf{\Sigma}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} > 0$, the columns in $\mathbf{\Omega}_{12} = (\mathbf{\Omega}_{\mathcal{D},\mathcal{A}}, \mathbf{\Omega}_{\mathcal{D},\mathcal{S}})$ is zero if and only if the corresponding columns in $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} = (\mathbf{B}_{\mathcal{D}|\mathcal{A}}, \mathbf{B}_{\mathcal{D}|\mathcal{S}})$ is zero. Thus $\mathbf{B}_{\mathcal{D}|\mathcal{S}} = \mathbf{0}$ and each column in $\mathbf{B}_{\mathcal{D}|\mathcal{A}}$ is nonzero.

S6. Proof of Theorem 1

PROOF. Note that $f_1(\boldsymbol{\beta})$ is a strictly convex function for $n > p$, and hence has a unique global (and local) minimum $\widehat{\boldsymbol{\beta}}_{\text{step1}}$. Hence, to prove estimation consistency of $\widehat{\boldsymbol{\beta}}_{\text{step1}}$, it is sufficient to show that for any small $\epsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_{n\to\infty} \bar{P}\left(\inf_{\|\text{vec}(\mathbf{u})\|=C} f_1(\bar{\boldsymbol{\beta}} + n^{-1/2}\mathbf{u}) > f_1(\boldsymbol{\beta}) \right) > 1 - \epsilon. \tag{S6.19}$$

Note that

$$\begin{aligned}
& f_1(\bar{\boldsymbol{\beta}} + n^{-1/2}\mathbf{u}) - f_1(\bar{\boldsymbol{\beta}}) \\
= \;& \text{tr}\Big[\big\{ \mathbb{Y}_c - \mathbb{X}_c(\bar{\boldsymbol{\beta}} + n^{-1/2}\mathbf{u})^T \big\} \mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1} \big\{ \mathbb{Y}_c - \mathbb{X}_c(\bar{\boldsymbol{\beta}} + n^{-1/2}\mathbf{u})^T \big\}^T / n \Big] \\
& - \text{tr}\{(\mathbb{Y}_c - \mathbb{X}_c\bar{\boldsymbol{\beta}}^T)\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbb{Y}_c - \mathbb{X}_c\bar{\boldsymbol{\beta}}^T)^T/n\} + \lambda_1 \sum_{i=1}^{r} w_i \left\{ \|\bar{\boldsymbol{\beta}}_{i\cdot} + n^{-1/2}\mathbf{u}_{i\cdot}\| - \|\bar{\boldsymbol{\beta}}_{i\cdot}\| \right\} \\
\geq \;& -2n^{-3/2}\,\text{tr}\{\mathbb{X}_c^T(\mathbb{Y}_c - \mathbb{X}_c\bar{\boldsymbol{\beta}}^T)\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{u}\} + n^{-1}\,\text{tr}(\mathbf{u}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{u}\mathbf{S}_{\mathbf{X}}) + \\
& \lambda_1 \sum_{i\in\bar{D}} w_i \big(\|\bar{\boldsymbol{\beta}}_{i\cdot} + n^{-1/2}\mathbf{u}_{i\cdot}\| - \|\bar{\boldsymbol{\beta}}_{i\cdot}\| \big) \\
\equiv \;& (I) + (II) + (III).
\end{aligned} \tag{S6.20}$$

Note that $\mathbb{X}_c^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c) = \mathbb{X}^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c)$, and the $(k,l)^{th}$ entry of $\mathbb{X}^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c)$ is given by $\sum_{j=1}^n X_{jk}(\varepsilon_{jl} - \sum_{j=1}^n \varepsilon_{jl}/n)$. Recall that $\boldsymbol{\varepsilon}_j =$

$\mathbf{Y}_j - \bar{\boldsymbol{\beta}}\mathbf{X}_j$, and $\{\boldsymbol{\varepsilon}_j\}_{j=1}^n$ are IID with mean $\mathbf{0}$ and covariance matrix $\bar{\boldsymbol{\Sigma}}$ under $\bar{P}$. Let $(\boldsymbol{\Sigma}_{\mathbf{X}})_{kk}$ denote the (k,k)th element of $\boldsymbol{\Sigma}_{\mathbf{X}}$, and $(\boldsymbol{\mu}_{\mathbf{X}})_k$ denote the kth element of $\boldsymbol{\mu}_{\mathbf{X}}$. Since $\frac{1}{n}\sum_{j=1}^n X_{jk}^2 \to (\boldsymbol{\Sigma}_{\mathbf{X}})_{kk} + (\boldsymbol{\mu}_{\mathbf{X}})_k^2 > 0$, it follows by Lindeberg's central limit theorem that $\sum_{j=1}^n X_{jk}(\varepsilon_{il} - \sum_{j=1}^n \varepsilon_{jl}/n) = \sum_{j=1}^n X_{jk}\varepsilon_{il} - \sum_{j=1}^n \varepsilon_{jl}/n \sum_{j=1}^n X_{jk} = O_{\bar{P}}(\sqrt{n})$. Since $\mathbb{X}^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c)$ is a $p \times r$ matrix, and p, r are fixed, it follows that $\|\mathrm{vec}\{\mathbb{X}^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c)\}\| = O_{\bar{P}}(\sqrt{n})$. For any matrix $\mathbf{A}$, let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the square-root of the maximum and minimum eigenvalue of $\mathbf{A}^T\mathbf{A}$ respectively. Since $\mathbf{S}_{\mathbf{Y}|\mathbf{X}} \to \bar{\boldsymbol{\Sigma}}$ as $n \to \infty$, it follows that $\lambda_{\max}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}) = O_{\bar{P}}(1)$, and $\lambda_{\min}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}) = O_{\bar{P}}(1)$. From these facts, we get

$$(I) \geq -\frac{1}{2n^{3/2}}\|\mathrm{vec}(\mathbf{u})\|\|\mathrm{vec}\{\mathbb{X}^T(\mathbb{Y}_c - \bar{\boldsymbol{\beta}}\mathbb{X}_c)\}\|\lambda_{\max}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}) = -\frac{\|\mathrm{vec}(\mathbf{u})\|}{n}O_{\bar{P}}(1), \tag{S6.21}$$

and

$$(II) \geq \frac{\lambda_{\min}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1})\lambda_{\min}(\mathbf{S}_{\mathbf{X}})\|\mathrm{vec}(\mathbf{u})\|^2}{n} = \frac{\|\mathrm{vec}(\mathbf{u})\|^2}{n}O_{\bar{P}}(1). \tag{S6.22}$$

Recall that $w_i = \|\widehat{\boldsymbol{\beta}}_{i\cdot}\|^{-\gamma_1}$. Since $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ is a $\sqrt{n}$-consistent estimator for $\bar{\boldsymbol{\beta}}$ and $\|\bar{\boldsymbol{\beta}}_{i\cdot}\| > 0$ if $i \in \bar{\mathcal{D}}$, it follows that $w_i = O_{\bar{P}}(1)$ for $i \in \bar{\mathcal{D}}$. Hence, by the triangle inequality and the fact that $\lambda_1 = o\left(\frac{1}{\sqrt{n}}\right)$, we get that

$$-(III) \leq \frac{\lambda_1}{\sqrt{n}}\sum_{i\in\bar{\mathcal{D}}} w_i\|\mathbf{u}_{i\cdot}\| \leq \frac{\lambda_1 r_{\bar{\mathcal{D}}}\|\mathrm{vec}(\mathbf{u})\|}{\sqrt{n}}\max_{1\leq i\leq r_{\bar{\mathcal{D}}}} w_{i\cdot} = \frac{\|\mathrm{vec}(\mathbf{u})\|}{n}o_{\bar{P}}(1). \tag{S6.23}$$

Note that (S6.23) is dominated by (S6.21) and (S6.22), and by choosing $\|\text{vec}(\mathbf{u})\| = C$ sufficiently large, the positive term (S6.22) dominates (S6.21). Thus the statement in (S6.19) holds.

We now use the method of contradiction to prove dynamic response selection consistency. Suppose that $\|(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot}\| > 0$ for some $i \notin \bar{\mathcal{D}}$. Since $\widehat{\boldsymbol{\beta}}_{\text{step1}}$ is the global minimizer of $f_1(\boldsymbol{\beta})$, it follows by the first derivative condition that

$$\left(2\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\widehat{\boldsymbol{\beta}}_{\text{step1}}\mathbf{S}_{\mathbf{X}} - 2\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{YX}}\right)_{i\cdot} + \lambda_1 w_i \frac{(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot}}{\|(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot}\|} = 0.$$

Since $\widehat{\boldsymbol{\beta}}_{\text{step1}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ are $\sqrt{n}$-consistent estimators of $\bar{\boldsymbol{\beta}}$, and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, $\mathbf{S}_{\mathbf{X}}$ are consistent estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, it follows that

$$2\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\widehat{\boldsymbol{\beta}}_{\text{step1}}\mathbf{S}_{\mathbf{X}} - 2\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{YX}} = 2\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}(\widehat{\boldsymbol{\beta}}_{\text{step1}} - \widehat{\boldsymbol{\beta}}_{\text{ols}})\mathbf{S}_{\mathbf{X}} = O_p(n^{-1/2}),$$

Since $(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot} \neq \mathbf{0}$, there exists a k such that $|(\widehat{\beta}_{\text{step1}})_{ik}|/\|(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot}\| \geq 1/\sqrt{r}$. Recall that $w_i = 1/\|\widehat{\boldsymbol{\beta}}_{i\cdot,\text{ols}}\|^{\gamma_1}$. As $\widehat{\boldsymbol{\beta}}_{i\cdot,\text{ols}}$ is $\sqrt{n}$-consistent, then $\|\widehat{\boldsymbol{\beta}}_{i\cdot,\text{ols}}\|^{-\gamma_1} = \Omega_{\bar{P}}(n^{\gamma_1/2})$. Since $n^{(1+\gamma_1)/2}\lambda_1 \to \infty$, $\sqrt{n}\lambda_1 w_i|\widehat{\beta}_{ik}|/\|\widehat{\boldsymbol{\beta}}_{i\cdot}\|$ tends to infinity as $n \to \infty$. This is a contradiction, so $(\widehat{\boldsymbol{\beta}}_{\text{step1}})_{i\cdot} = \mathbf{0}$ with probability tending to 1 for any $i \notin \bar{\mathcal{D}}$. Combining this with the estimation consistency result for $\widehat{\boldsymbol{\beta}}_{\text{step1}}$, we get $\bar{P}(\widehat{\mathcal{D}} = \bar{\mathcal{D}}) \to 1$ as $n \to \infty$.

Selection consistency of ancillary response can be established following exactly the same procedure as the dynamic response.

Now we prove the estimation consistency of $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$. Let T^c denote the complement of set T. Notice that

$$\begin{aligned}
& \bar{P}(\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\| \geq \varepsilon) \\
= \; & \bar{P}\left(\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\| \geq \varepsilon \mid \widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\right) \bar{P}\left(\widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\right) \\
& + \bar{P}\left(\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\| \geq \varepsilon \mid \{\widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\}^c\right) \bar{P}\left(\{\widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\}^c\right) \\
\geq \; & \bar{P}\left(\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\| \geq \varepsilon \mid \widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\right) \bar{P}\left(\widehat{\mathcal{D}} = \mathcal{D}, \widehat{\mathcal{A}} = \mathcal{A}\right)
\end{aligned}$$

By dynamic response selection consistency and ancillary response selection consistency, $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ is a $\sqrt{n}$-consistent estimator of $\bar{\boldsymbol{\beta}}_{\mathcal{D}}$.

S7. Proof of Theorem 2

Proof. Note that $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} = \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} - \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}|\widehat{\mathcal{A}}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\mathrm{oracle}} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$, where $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$, $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}|\widehat{\mathcal{A}}}$, $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}$, $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}}$ and $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$ are OLS estimators. By the selection consistency of the dynamic and ancillary responses in Theorem 1, it follows that

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},\mathrm{oracle}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{\mathrm{oracle}}), \quad \mathbf{V}_{\mathrm{oracle}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}).$$

and

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}), \quad \mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}).$$

Since $\mathbf{V} = \mathbf{V}_{\mathrm{oracle}}$ and $\bar{P}(\widehat{\mathcal{D}} = \bar{\mathcal{D}}) \to 1$, we have $\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},\mathrm{oracle}})\| = o_{\bar{P}}(n^{-1/2})$.

S8. Assumptions for high-dimensional consistency

In this section, we provide and discuss the regularity assumptions that are needed for establishing high-dimensional selection consistency of the proposed procedure in Theorem 3. Let $T_n = \{(i,j) : \bar{\Omega}_{ij} \neq 0\}$ be the set of indices of nonzero elements in $\bar{\mathbf{\Omega}}$, $t_n = |T_n|_c$ the cardinality of T_n, and s_n the minimum absolute value of nonzero entries in $\bar{\mathbf{\Omega}}$.

The first three assumptions, Assumptions 1- 3, are required for the consistency of the CONCORD estimator $\widehat{\mathbf{\Omega}}$.

Assumption 1. *The errors $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n$ are independent and identically sampled from a sub-Gaussian distribution with mean zero and covariance matrix $\bar{\mathbf{\Sigma}}$. The eigenvalues of $\bar{\mathbf{\Sigma}}$ are uniformly bounded above by $\bar{k}$ and uniformly bounded below by $\underline{k} = 1/\bar{k}$. The predictors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ are IID from a sub-Gaussian distribution.*

Assumption 2 (Incoherence condition)**.** *Let $\mathbf{\Gamma}$ denote a $\binom{r_n}{2}$-dimensional square matrix. Then the number of rows or columns of $\mathbf{\Gamma}$ is the same as the number of the edges connecting vertices $\{1, \ldots, r_n\}$. The element $\Gamma_{(i,j),(t,s)}$ denotes the element whose row corresponds to the edge connecting vertices i and j and column corresponds to the edge connecting vertices s and t. Let*

$1_{\{\cdot\}}$ *be the indicator function. For* $1 \leq i < j \leq r_n$ *and* $1 \leq t < s \leq r_n$,

$$\Gamma_{(i,j),(t,s)} = \bar{\Sigma}_{js} 1_{\{i=t\}} + \bar{\Sigma}_{it} 1_{\{j=s\}} + \bar{\Sigma}_{is} 1_{\{j=t\}} + \bar{\Sigma}_{jt} 1_{\{i=s\}},$$

and $\boldsymbol{\gamma}$ *denote a* $\binom{r_n}{2}$*-dimensional vector such that for* $1 \leq i < j \leq r_n$, $\gamma_{(i,j)} = \bar{\Omega}_{ij}$. *Then*

$$\max_{(i,j) \in T_n^c} \left| \boldsymbol{\Gamma}_{(i,j),T_n} \boldsymbol{\Gamma}_{T_n,T_n}^{-1} sign(\boldsymbol{\gamma}_{T_n}) \right| < 1,$$

where $sign(\mathbf{x}) = (sign(x_i))_{i=1}^k$ *is the sign function for any* k*-dimensional vector* $\mathbf{x}$.

Assumption 3. $r_n = O(n^\kappa)$ *for some* $\kappa > 0$, $t_n \sqrt{\log r_n / n} = o\left(\min(1, s_n^2)\right)$, *and for some constant* $c > 0$, $\lambda = c\left(\sqrt[4]{\log r_n / n}\right)$, *where* λ *is the penalty in* (3.13).

Similar assumptions are made in Khare et al. (2015) for consistency of the CONCORD estimator in the IID setting. Note that the setting here is slightly different from Khare et al. (2015), as $\boldsymbol{\Omega}$ is the precision matrix of the residuals, which are not IID.

The next assumption controls the rate at which the true number of dynamic and ancillary variables can grow as n increases.

Assumption 4. $r_{\bar{\mathcal{D}}}^2 r_{\bar{\mathcal{A}}} + r_{\bar{\mathcal{D}}} r_{\bar{\mathcal{A}}}^2 = o\left(\frac{n}{\log r_n}\right)$.

The next two assumptions, are again standard for sparsity selection consistency in both the frequentist and Bayesian paradigms. They essentially provide lower bounds for the "minimum signal strength" in the context of the dynamic and ancillary response selection steps. As a specific example, if $r_{\bar{\mathcal{D}}}$ and $r_{\bar{\mathcal{A}}}$ are uniformly bounded in n, then the two assumptions are satisfied, for example, if $\sqrt{\log(r_n)/n} = o(1)$, the entries of $\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}$ are uniformly bounded below (very mild requirement given $r_{\bar{\mathcal{D}}} = O(1)$), and the eigenvalues of $\bar{\boldsymbol{\Sigma}}$ are uniformly bounded (a standard assumption in high-dimensional asymptotics).

Assumption 5. $\sqrt{r_{\bar{\mathcal{D}}} \log(r_n)/n} + \sqrt{t_n r_{\bar{\mathcal{A}}}} (\log(r_n)/n)^{3/4} = o\left(\min_{1 \le i \le r_{\bar{\mathcal{D}}}} \|\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}},i\cdot}\|_{\max}\right)$.

Assumption 6. *Recall that* $\bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}},\bar{\mathcal{S}}} = -\bar{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1} \bar{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},(\bar{\mathcal{A}},\bar{\mathcal{S}})}$. *Let* $\bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}},\bar{\mathcal{S}}} = \left(\bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}}, \bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{S}}}\right)$. *Then*

$$\sqrt{\frac{\left(r_{\bar{\mathcal{D}}}^2 r_{\bar{\mathcal{A}}} + r_{\bar{\mathcal{D}}} r_{\bar{\mathcal{A}}}^2\right) \log(r_n)}{n}} = o\left(\min_{1 \le i \le r_{\bar{\mathcal{A}}}} \|\bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}},\cdot i}\|_{\max}\right).$$

The last two assumptions control the behavior of various group-specific penalty parameters.

Assumption 7. $\lambda_1 \max_{1 \le i \le r_{\bar{\mathcal{D}}}} w_i = O\left(\sqrt{\log(r_n)/n}\right)$ *and* $(t_n + \sqrt{r_{\bar{\mathcal{D}}}} + r_{\bar{\mathcal{A}}}) \sqrt{\log(r_n)/n} = o\left(\lambda_1 \min_{r_{\bar{\mathcal{D}}}+1 \le i \le r_n} w_i\right)$.

Assumption 8. $\lambda_2 \max_{1 \le i \le r_{\bar{\mathcal{A}}}} \tilde{w}_i = O\left(\sqrt{(r_{\bar{\mathcal{D}}} + r_{\bar{\mathcal{A}}}) \log(r_n)/n}\right)$ *and* $\sqrt{\left(r_{\bar{\mathcal{D}}}^2 r_{\bar{\mathcal{A}}} + r_{\bar{\mathcal{D}}} r_{\bar{\mathcal{A}}}^2\right) \log(r_n)/n} = o\left(\lambda_2 \min_{r_{\bar{\mathcal{A}}}+1 \le i \le r_n} \tilde{w}_i\right)$.

S9. Proof of Theorem 3

PROOF. We first establish the consistency of the CONCORD estimator $\widehat{\boldsymbol{\Omega}}$. Let $\mathbf{E}$ be the $r_n \times n$ matrix with columns $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n$. Then

$$\mathbf{S}_{\mathbf{Y}|\mathbf{X}} = \frac{1}{n}\mathbf{E}(\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbf{E}^T. \tag{S9.24}$$

Hence $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ is not the covariance matrix of the errors. To derive the consistency of the CONCORD estimator of $\boldsymbol{\Omega}^{-1}$, we need to derive a concentration inequality for $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$.

Let $\mathbf{W}$ be a d-dimensional random vector with mean $\boldsymbol{\mu}_{\mathbf{W}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}$. Assume that $\mathbf{W} - \boldsymbol{\mu}_{\mathbf{W}}$ follows a sub-Gaussian distribution. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are IID samples of $\mathbf{W}$. Let $\bar{\mathbf{W}} = \sum_{i=1}^n \mathbf{W}_i/n$. Then $\bar{\mathbf{W}} - \boldsymbol{\mu}_{\mathbf{W}}$ also follows a sub-Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}/n$. Note that the sample covariance matrix is

$$\begin{aligned}\mathbf{S}_{\mathbf{W}} =& \frac{1}{n}\sum_{k=1}^{n}(\mathbf{W}_k - \bar{\mathbf{W}})(\mathbf{W}_k - \bar{\mathbf{W}})^T \\ =& \frac{1}{n}\sum_{k=1}^{n}(\mathbf{W}_k - \boldsymbol{\mu}_{\mathbf{W}})(\mathbf{W}_k - \boldsymbol{\mu}_{\mathbf{W}})^T - (\bar{\mathbf{W}} - \boldsymbol{\mu}_{\mathbf{W}})(\bar{\mathbf{W}} - \boldsymbol{\mu}_{\mathbf{W}})^T.\end{aligned}$$

Let $(\mathbf{W}_k - \boldsymbol{\mu}_{\mathbf{W}})_i$ denote the ith element of the random vector $\mathbf{W}_k - \boldsymbol{\mu}_{\mathbf{W}}$. From Lemma 1 in Ravikumar et al. (2011), there exist positive constants

b_1, C_1, C_2, C_3 and C_4, such that for $\delta \in (0, b_1)$,

$$\begin{aligned}
P(|\mathbf{S}_{\mathbf{W},ij} - \mathbf{\Sigma}_{\mathbf{W},ij}| > \delta) \leq & P\left[\left|\left\{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_k - \bar{\mathbf{W}})(\mathbf{W}_k - \bar{\mathbf{W}})^T\right\}_{ij} - \mathbf{\Sigma}_{\mathbf{W},ij}\right| > \frac{\delta}{2}\right] \\
& + P\left[\left|\left\{(\bar{\mathbf{W}} - \boldsymbol{\mu}_{\mathbf{W}})(\bar{\mathbf{W}} - \boldsymbol{\mu}_{\mathbf{W}})^T\right\}_{ij} - \frac{1}{n}\mathbf{\Sigma}_{\mathbf{W},ij}\right| > \frac{\delta}{2}\right] + 1_{\{\frac{1}{n}\mathbf{\Sigma}_{\mathbf{W},ij} > \frac{\delta}{2}\}} \\
& \leq C_1 \exp(-C_2 n\delta^2) + C_3 \exp(-C_4 n^3 \delta^2) \\
& \leq C_5 \exp(-C_6 n\delta^2),
\end{aligned}$$

where $C_5 > C1$, and $C_6 = C_2$. In the second inequality, when n is sufficiently large, $1_{\{\frac{1}{n}\mathbf{\Sigma}_{\mathbf{W},ij} > \frac{\delta}{2}\}} = 0$. This is because $\mathbf{\Sigma}$ is upper bounded by $\bar{k}$ and the dimension of $\mathbf{\Sigma}_{\mathbf{X}}$ is fixed. Let $\delta = \sqrt{C_7}\{\log(d)/n\}^{1/2}$ for some $d > 0$. Then

$$P(|\mathbf{S}_{\mathbf{W},ij} - \mathbf{\Sigma}_{\mathbf{W},ij}| > \sqrt{C_7}\{\log(d)/n\}^{1/2}) \leq d^{-C_6 C_7}$$

for large enough d (or large enough n, if d grows with n). Since C_7 can be any positive constant, we take C_7 such that $C_6 C_7 > 2$.

Using the union sum inequality, we have

$$\begin{aligned}
P(\|\mathbf{S}_{\mathbf{W}} - \mathbf{\Sigma}_{\mathbf{W}}\|_{\max} > \delta) = & P(\cup_{i,j=1}^{d} |\mathbf{S}_{\mathbf{W},ij} - \mathbf{\Sigma}_{\mathbf{W},ij}| > \delta) \\
& \leq \sum_{i=1}^{d}\sum_{j=1}^{d} P(|\mathbf{S}_{\mathbf{W},ij} - \mathbf{\Sigma}_{\mathbf{W},ij}| > \delta) \\
& \leq d^2 d^{-C_6 C_7} = d^{-(C_6 C_7 - 2)}.
\end{aligned}$$

Thus with probability at least $1 - d^{-\eta}$, where $\eta = C_6 C_7 - 2 > 0$, we have

$$\|\mathbf{S}_{\mathbf{W}} - \mathbf{\Sigma}_{\mathbf{W}}\|_{\max} \leq C_7 \{\log(d)/n\}^{1/2}$$

for large enough d. Henceforth, it will be understood that statements regarding relevant high probability events hold for large enough n (depending on η). Now we take $\mathbf{W} = (\mathbf{X}^T, \boldsymbol{\varepsilon}^T)^T$, then $\mathbf{W}$ is a $(p+r_n)$-dimensional random vector with mean $\boldsymbol{\mu}_{\mathbf{W}} = (\boldsymbol{\mu}_{\mathbf{X}}^T, \mathbf{0}^T)^T$, where $\mathbf{0}$ is an r_n dimensional zero vector. Since $\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\varepsilon}$ are both sub-Gaussian random vector, then $\mathbf{W} - \boldsymbol{\mu}_{\mathbf{W}}$ is a sub-Gaussian random vector with mean $\mathbf{0}$ and block diagonal covariance matrix with diagonal blocks $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}$. Thus there exists a constant $C_0 > 0$ such that $\|\mathbf{S}_{\mathbf{W}} - \mathbf{\Sigma}_{\mathbf{W}}\|_{\max} \leq C_0 \{\log(r_n + p)/n\}^{1/2}$ with probability at least $1 - (p + r_n)^{-\eta}$. Since p is fixed, we can find C_0^* such that

$$\|\mathbf{S}_{\mathbf{W}} - \mathbf{\Sigma}_{\mathbf{W}}\|_{\max} \leq C_0^* \{\log(r_n)/n\}^{1/2},$$

with probability at least $1 - r_n^{-\eta}$. Note that $\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}$, $\mathbf{S}_{\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\mathbf{S}_{\boldsymbol{\varepsilon},\mathbf{X}}$ are sub-matrices of $\mathbf{S}_{\mathbf{W}} - \mathbf{\Sigma}_{\mathbf{W}}$, where $\mathbf{S}_{\mathbf{X}} = \mathbb{X}_c^T \mathbb{X}_c^T / n$, $\mathbf{S}_{\boldsymbol{\varepsilon}} = \mathbf{E}^T \mathbf{E}/n$ and $\mathbf{S}_{\boldsymbol{\varepsilon},\mathbf{X}} = \mathbf{E}^T \mathbb{X}/n$. Hence with probability at least $1 - r_n^{-\eta}$

$$\begin{aligned}
\|\mathbf{S}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{X}}\|_{\max} &\leq C_0^* \{\log(r_n)/n\}^{1/2}, \\
\|\mathbf{S}_{\boldsymbol{\varepsilon}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\|_{\max} &\leq C_0^* \{\log(r_n)/n\}^{1/2}, \\
\|\mathbf{S}_{\boldsymbol{\varepsilon},\mathbf{X}}\|_{\max} &\leq C_0^* \{\log(r_n)/n\}^{1/2}.
\end{aligned} \tag{S9.25}$$

Because that $\mathbf{S}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\boldsymbol{\mathcal{E}}} - \mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}}$, where $\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}} = \mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}}^T$, we have

$$\begin{aligned}\mathbf{S}_{\mathbf{Y}|\mathbf{X}} - \boldsymbol{\Sigma} =& \mathbf{S}_{\boldsymbol{\mathcal{E}}} - \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}}} + (\mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}} - \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}} + \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}}(\mathbf{S}_{\mathbf{X}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}} \\ &+ \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}} - \boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}}) + (\mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}} - \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}})(\mathbf{S}_{\mathbf{X}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}} \\ &+ \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}}(\mathbf{S}_{\mathbf{X}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})(\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}} - \boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}}) + (\mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}} - \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}} - \boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}}) \\ &+ (\mathbf{S}_{\boldsymbol{\mathcal{E}},\mathbf{X}} - \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}},\mathbf{X}})(\mathbf{S}_{\mathbf{X}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})(\mathbf{S}_{\mathbf{X},\boldsymbol{\mathcal{E}}} - \boldsymbol{\Sigma}_{\mathbf{X},\boldsymbol{\mathcal{E}}}).\end{aligned}$$

Using the fact that for $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{B} \in \mathbb{R}^{d_2 \times d_3}$, $\|\mathbf{AB}\|_{\max} \leq d_2 \|\mathbf{A}\|_{\max}\|\mathbf{B}\|_{\max}$, we have

$$\|\mathbf{S}_{\mathbf{Y}|\mathbf{X}} - \boldsymbol{\Sigma}\|_{\max} \leq C_{\text{res}}\{\log(r_n)/n\}^{1/2}, \tag{S9.26}$$

with probability at least $1 - r_n^{-\eta}$ for some $C_{\text{res}} > 0$.

The concentration inequality in (S9.26), along with Assumptions 1-3 ensure that the proof of (Khare et al., 2015, Theorem 2) goes through for the current setting. It follows that for any $\eta > 0$, there exists a constant C_η such that for large enough n,

$$\left\|\widehat{\boldsymbol{\Omega}} - \bar{\boldsymbol{\Omega}}\right\|_2 < C_\eta \sqrt{t_n} \sqrt[4]{\frac{\log r_n}{n}}, \tag{S9.27}$$

and $\widehat{\boldsymbol{\Omega}}$ recovers the zeros and non-zeros in $\boldsymbol{\Omega}$ with probability at least $1 - r_n^{-\eta}$. This establishes the high-dimensional accuracy of $\widehat{\boldsymbol{\Omega}}$ as an estimator of $\boldsymbol{\Omega}$.

Before proving Theorem 3, we first establish two inequalities (S9.28) and (S9.31) which will be useful in subsequent analysis. It follows by the

triangle inequality that for every $\eta > 0$,

$$\begin{aligned}
&\left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1}\right\|_2 \\
\leq\ &\left\|\bar{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right\|_2 + \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} - \bar{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right\|_2 \\
\leq\ &\bar{k}^2 + \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\left((\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} - \bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right)\right\|_2 + \left\|\left(\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}} - \bar{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\right)\bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right\|_2 \\
\leq\ &\bar{k}^2 + \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\right\|_2 \left\|(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} - \bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right\|_2 + \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}} - \bar{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\right\|_2 \left\|\bar{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\right\|_2 \\
\overset{(a)}{\leq}\ &\bar{k}^2 + C_{1,\eta}\sqrt{t_n}\sqrt[4]{\frac{\log r_n}{n}} \qquad \text{(S9.28)}
\end{aligned}$$

for some constant $C_{1,\eta}$, with probability at least $1 - r_n^{-\eta}$ for large enough n. Here (a) follows from Assumption 1, (S9.27), the fact that $\widehat{\mathbf{\Omega}}$ recovers the zeros of $\mathbf{\Omega}$ with high probability, and $t_n\sqrt{\log r_n/n} \to 0$ as $n \to \infty$.

Note that each row of $\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}$ has p entries, and $p = O(1)$. Let $\|\cdot\|_\infty$ denote the row-sum norm of a matrix. Using the third relation in (S9.25), for every $\eta > 0$, there exists a constant $C_{2,\eta}$ such that

$$\|\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_\infty = \|\mathbf{S}_{\boldsymbol{\varepsilon}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_\infty \leq C_{2,\eta}\sqrt{\frac{\log r_n}{n}} \qquad \text{(S9.29)}$$

with probability at least $1 - r_n^{-\eta}$ for large enough n. Since $\bar{\mathbf{\Omega}}_{\bar{\mathcal{S}},\bar{\mathcal{D}}} = \mathbf{0}$, we get

$$\begin{aligned}
\left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\right\|_\infty &\leq \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\right\|_\infty + \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\right\|_\infty \\
&\leq \left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}}\right\|_\infty \left\|\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\right\|_\infty + \left\|\widehat{\mathbf{\Omega}}_{\bar{\mathcal{A}},\bar{\mathcal{D}}}(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\right\|_\infty \left\|\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{A}}}\mathbf{X}}\right\|_\infty .
\end{aligned} \qquad \text{(S9.30)}$$

From (S9.27) and Assumption 1, $\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}}\|_\infty$ is bounded by $t_n/\underline{k}$ and $\left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}(\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}\right\|_\infty$ is bounded by $\bar{k}^3 r_{\bar{\mathcal{A}}}$ with probability tending to 1. Then it follows by (S9.27), (S9.29), (S9.30), the fact that $\widehat{\mathbf{\Omega}}$ recovers the zeros and nonzeros in $\bar{\mathbf{\Omega}}$ with high probability that for every $\eta > 0$, there exists a constant $C_{3,\eta} > 0$ such that

$$\left\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\right\|_\infty \leq C_{3,\eta}(t_n + r_{\bar{\mathcal{A}}})\sqrt{\frac{\log r_n}{n}} \tag{S9.31}$$

with probability at least $1 - 2r_n^{-\eta}$ for large enough n.

Now let $\widetilde{\boldsymbol{\beta}}$ be the solution of the restricted problem

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\widetilde{\boldsymbol{\beta}}_{-\bar{\mathcal{D}}\cdot}=\mathbf{0}} \tilde{f}_1(\boldsymbol{\beta}).$$

We will show that $\widetilde{\boldsymbol{\beta}}$ is the minimizer for (3.14) with high probability. As $\partial f_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_{i\cdot} = -2\mathbf{e}_i^T\widehat{\mathbf{\Omega}}(\mathbf{S}_{\mathbf{YX}} - \boldsymbol{\beta}\mathbf{S}_{\mathbf{X}}) + \lambda_1 w_i \partial\|\boldsymbol{\beta}_{i\cdot}\|/\partial\boldsymbol{\beta}_{i\cdot}$, where $\mathbf{e}_i$ is an r_n-dimensional vector of 0 except for a 1 in the ith element, this is equivalent to show that

$$-2\mathbf{e}_i^T\widehat{\mathbf{\Omega}}(\mathbf{S}_{\mathbf{YX}} - \widetilde{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}}) + \lambda_1 w_i \frac{\widetilde{\boldsymbol{\beta}}_{i\cdot}}{\|\widetilde{\boldsymbol{\beta}}_{i\cdot}\|} = \mathbf{0}, \qquad \text{for} \quad i \in \bar{\mathcal{D}}, \tag{S9.32}$$

$$\|2\mathbf{e}_i^T\widehat{\mathbf{\Omega}}(\mathbf{S}_{\mathbf{YX}} - \widetilde{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}})\| \leq \lambda_1 w_i, \qquad \text{for} \quad i \notin \bar{\mathcal{D}}. \tag{S9.33}$$

Condition (S9.32) holds because of the definition of $\widetilde{\boldsymbol{\beta}}$. Let

$$\mathbf{t}_{\bar{\mathcal{D}}} = \left(\begin{array}{ccc} w_1 \frac{\widetilde{\boldsymbol{\beta}}_{1\cdot}^T}{\|\widetilde{\boldsymbol{\beta}}_{1\cdot}^T\|} & \cdots & w_{r_{\bar{\mathcal{D}}}} \frac{\widetilde{\boldsymbol{\beta}}_{r_{\bar{\mathcal{D}}}\cdot}^T}{\|\widetilde{\boldsymbol{\beta}}_{r_{\bar{\mathcal{D}}}\cdot}^T\|} \end{array} \right)^T \in \mathbb{R}^{r_{\bar{\mathcal{D}}} \times p},$$

and partition $\widehat{\mathbf{\Omega}}$ as

$$\widehat{\mathbf{\Omega}} = \begin{pmatrix} \widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}} & \widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}} \\ \widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}} & \widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}} \end{pmatrix},$$

then condition (S9.32) can be written as

$$-2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}(\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{D}}}\mathbf{X}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}) - 2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}} + \lambda_1 \mathbf{t}_{\bar{\mathcal{D}}} = \mathbf{0}.$$

Notice that

$$\mathbf{S}_{\mathbf{YX}} - \widetilde{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}} = \begin{pmatrix} \mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{D}}}\mathbf{X}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}} \\ \mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}} \end{pmatrix},$$

and denote $\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}} = \widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}} - \widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}$. Then for every $i \notin \bar{\mathcal{D}}$,

$$\begin{aligned} & \|2\mathbf{e}_i^T\widehat{\mathbf{\Omega}}(\mathbf{S}_{\mathbf{YX}} - \widetilde{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}})\|_2 \\ = \; & \|2\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}(\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{D}}}\mathbf{X}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}) + 2\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_2 \\ = \; & \|2\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}} + \lambda_1\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\|_2 \\ \leq \; & 2\|\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_2 + \lambda_1\|\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\|_2, \end{aligned} \tag{S9.34}$$

where $\tilde{\mathbf{e}}_i \in \mathbb{R}^{r_n - r_{\bar{\mathcal{D}}}}$ and $\mathbf{e}_i^T = (\mathbf{0}^T,\ \tilde{\mathbf{e}}_i^T)$. Here $\mathbf{0}$ is a $r_{\bar{\mathcal{D}}}$-dimensional zero vector. Note that $\|\mathbf{t}_{\bar{\mathcal{D}}}\|_2 \leq \sqrt{pr_{\bar{\mathcal{D}}}}\max(w_1, \cdots, w_{r_{\bar{\mathcal{D}}}})$ and $\|\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\|_2 \leq \|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\|_2 \leq \|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}},\bar{\mathcal{D}}}\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}^{-1}\|_2\|\mathbf{t}_{\bar{\mathcal{D}}}\|_2$. Also,

$$2\|\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_2 \leq 2\|\widehat{\mathbf{\Omega}}_{-\bar{\mathcal{D}}|\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\|_\infty.$$

By (S9.28), (S9.31), (S9.34) and Assumption 7, it follows that $\|2\tilde{\mathbf{e}}_i^T\widehat{\mathbf{\Omega}}(\mathbf{S}_{\mathbf{YX}} -$

$\widetilde{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}})\|_2 \leq \lambda_1 w_i$ for every $i \notin \bar{\mathcal{D}}$, which implies that $\widehat{\boldsymbol{\beta}}_{\text{step1}} = \widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{step1,i\cdot} = \mathbf{0}$ for every $i \notin \bar{\mathcal{D}}$ with probability at least $1 - 3r_n^{-\eta}$ for large enough n.

From (S9.32), we have

$$-2\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}(\boldsymbol{\beta}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}} + \mathbf{S}_{\boldsymbol{\mathcal{E}}_{\bar{\mathcal{D}}}\mathbf{X}}) - 2\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}} + \lambda_1 \mathbf{t}_{\bar{\mathcal{D}}} = \mathbf{0}.$$

Thus

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{step1,\bar{\mathcal{D}}} - \bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}} &= \widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},-\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{Y}_{-\bar{\mathcal{D}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} + \widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{S}_{\boldsymbol{\mathcal{E}}_{\bar{\mathcal{D}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} - \frac{\lambda_1}{2}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}^{-1} \\
&= \widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{A}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} + \widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{S}_{\boldsymbol{\mathcal{E}}_{\bar{\mathcal{D}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} - \frac{\lambda_1}{2}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}^{-1}
\end{aligned}$$

with probability at least $1 - r_n^{-\eta}$ for large enough n. The last equality above follows from the selection consistency of the CONCORD estimator $\widehat{\boldsymbol{\Omega}}$. By the sub-Gaussianity of the errors and the predictors, (S9.27), Assumption 3, and Assumption 4, it follows that for every $\eta > 0$, there exists a constant $C_{4,\eta}$ such that

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\text{step1},\bar{\mathcal{D}}} - \bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}\|_{\max} &= \|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{A}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} + \widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{S}_{\boldsymbol{\mathcal{E}}_{\bar{\mathcal{D}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} - \frac{\lambda_1}{2}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}^{-1}\|_{\max} \\
&\leq \|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{S}_{\mathbf{Y}_{\bar{\mathcal{A}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\|_{\max} + \|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{S}_{\boldsymbol{\mathcal{E}}_{\bar{\mathcal{D}}}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\|_{\max} + \|\frac{\lambda_1}{2}\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}^{-1}\mathbf{t}_{\bar{\mathcal{D}}}\mathbf{S}_{\mathbf{X}}^{-1}\|_{\max} \\
&\leq C_{4,\eta}\left(\sqrt{t_n}\sqrt[4]{\frac{\log r_n}{n}}\sqrt{r_{\mathcal{A}}}\sqrt{\frac{\log r_n}{n}} + \sqrt{\frac{r_{\bar{\mathcal{D}}}\log r_n}{n}}\right) \qquad \text{(S9.35)}
\end{aligned}$$

with probability at least $1 - 3r_n^{-\eta}$ for large enough n. Using Assumption 5, we get $\widehat{\boldsymbol{\beta}}_{\text{step1},j\cdot} \neq \mathbf{0}$ for every $j \in \bar{\mathcal{D}}$ with probability at least $1 - 3r_n^{-\eta}$ for large enough n. Since we already established that $\widehat{\boldsymbol{\beta}}_{\text{step1},j\cdot} = \mathbf{0}$ for every

$j \notin \bar{\mathcal{D}}$ with probability at least $1 - 3r_n^{-\eta}$ for large enough n, part (a) of the result follows.

We now prove part (b). Note from part (a) that $\widehat{\mathcal{D}} = \bar{\mathcal{D}}$ on an event with probability at least $1-6r_n^{-\eta}$ for large enough n. We will restrict to this high probability event throughout the subsequent argument. Recall that $\mathbf{R} = \mathbb{Y}_c - \mathbb{X}_c\widehat{\boldsymbol{\beta}}_{\text{step1}}$ is the matrix of residuals obtained by using $\widehat{\boldsymbol{\beta}}_{\text{step1}}$. Let $\mathbf{C} = \mathbf{R}\mathbf{R}^T/n$. We proceed to establish a useful concentration bound for $\mathbf{C}$ around the error covariance matrix $\boldsymbol{\Sigma}$. Note that $\mathbf{R} = \mathbb{Y}_c - \mathbb{X}_c\bar{\boldsymbol{\beta}}^T + \mathbb{X}_c\bar{\boldsymbol{\beta}}^T - \mathbb{X}_c\widehat{\boldsymbol{\beta}}_{\text{step1}}^T$. It follows that

$$
\begin{aligned}
& \|\mathbf{C} - \boldsymbol{\Sigma}\|_{\max} \\
\leq\ & \left\| \frac{1}{n}\sum_{\ell=1}^{n}(\boldsymbol{\varepsilon}_\ell - \bar{\boldsymbol{\varepsilon}})(\boldsymbol{\varepsilon}_\ell - \bar{\boldsymbol{\varepsilon}})^T - \boldsymbol{\Sigma} \right\|_{\max} + \frac{1}{n}\left\| (\widehat{\boldsymbol{\beta}}_{\text{step1}} - \bar{\boldsymbol{\beta}})\mathbb{X}_c^T\mathbb{X}_c(\widehat{\boldsymbol{\beta}}_{\text{step1}} - \bar{\boldsymbol{\beta}})^T \right\|_{\max} + \\
& \frac{2}{n}\left\| (\widehat{\boldsymbol{\beta}}_{\text{step1}} - \bar{\boldsymbol{\beta}})\mathbb{X}_c^T(\mathbb{Y}_c - \mathbb{X}_c\bar{\boldsymbol{\beta}}^T) \right\|_{\max} \\
\leq\ & \left\| \frac{1}{n}\sum_{\ell=1}^{n}(\boldsymbol{\varepsilon}_\ell - \bar{\boldsymbol{\varepsilon}})(\boldsymbol{\varepsilon}_\ell - \bar{\boldsymbol{\varepsilon}})^T - \boldsymbol{\Sigma} \right\|_{\max} + \left\| (\widehat{\boldsymbol{\beta}}_{\text{step1}} - \bar{\boldsymbol{\beta}})\mathbf{S}_{\mathbf{X}}(\widehat{\boldsymbol{\beta}}_{\text{step1}} - \bar{\boldsymbol{\beta}})^T \right\|_{\max} + \\
& 2\left\| (\widehat{\boldsymbol{\beta}}_{\text{step1},\bar{\mathcal{D}}} - \bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\mathbf{S}_{\mathbf{X}}\boldsymbol{\varepsilon}_{\bar{\mathcal{D}}} \right\|_{\max}.
\end{aligned}
$$

Since the errors $\{\boldsymbol{\varepsilon}_i\}_{i=1}^n$ and the predictors $\{\mathbf{X}_i\}_{i=1}^n$ are both IID sub-Gaussian with respective covariance matrices having uniformly bounded eigenvalues (by Assumption 1), and independent of each other, it follows by a straightforward application of Theorem 1.1 in Rudelson and Vershynin (2013), part (a), the uniform boundedness of p and (S9.35) that for every

$\eta > 0$, there exists a constant $C_{5,\eta}$ such that

$$\|\mathbf{C} - \mathbf{\Sigma}\|_{\max} \leq C_{5,\eta}\sqrt{\frac{r_{\bar{\mathcal{D}}} \log r_n}{n}} \tag{S9.36}$$

with probability at least $1 - 7r_n^{-\eta}$ for large enough n.

Now let $\widetilde{\mathbf{B}}$ be the solution of the following restricted problem

$$\widetilde{\mathbf{B}} = \arg\min_{\mathbf{B}_{\cdot\bar{\mathcal{S}}}=\mathbf{0}} \tilde{f}_2(\mathbf{B}).$$

We will show that $\widetilde{\mathbf{B}}$ is the minimizer for (3.15) with high probability. As $\partial \tilde{f}_2(\mathbf{B})/\partial \mathbf{B}_{\cdot i} = 2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}\left(\mathbf{B}\mathbf{C}_{-\bar{\mathcal{D}},i} - \mathbf{C}_{\bar{\mathcal{D}},i}\right) + \lambda_2 \tilde{w}_i \partial \|\mathbf{B}_{\cdot i}\|_2 / \partial \mathbf{B}_{\cdot i}$, this is equivalent to show that

$$2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}\mathbf{C}_{-\bar{\mathcal{D}},i} - \mathbf{C}_{\bar{\mathcal{D}},i}\right) + \lambda_2 \tilde{w}_i \frac{\widetilde{\mathbf{B}}_{\cdot i}}{\|\widetilde{\mathbf{B}}_{\cdot i}\|} = \mathbf{0}, \qquad \text{for} \quad i \in \bar{\mathcal{A}}, \tag{S9.37}$$

$$\|2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}\mathbf{C}_{-\bar{\mathcal{D}},i} - \mathbf{C}_{\bar{\mathcal{D}},i}\right)\| \leq \lambda_2 \tilde{w}_i, \qquad \text{for} \quad i \in \bar{\mathcal{S}}. \tag{S9.38}$$

Condition (S9.37) holds because of the definition of $\widetilde{\mathbf{B}}$. Let

$$\widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} = \left(\begin{array}{ccc} \tilde{w}_1 \frac{\widetilde{\mathbf{B}}_{\cdot 1}}{\left\|\widetilde{\mathbf{B}}_{\cdot 1}\right\|} & \cdots & \tilde{w}_{r_{\bar{\mathcal{A}}}} \frac{\widetilde{\mathbf{B}}_{\cdot r_{\bar{\mathcal{A}}}}}{\left\|\widetilde{\mathbf{B}}_{\cdot r_{\bar{\mathcal{A}}}}\right\|} \end{array} \right) \in \mathbb{R}^{r_{\bar{\mathcal{D}}} \times r_{\bar{\mathcal{A}}}},$$

then condition (S9.37) can be written as

$$2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}\mathbf{C}_{-\bar{\mathcal{D}},\bar{\mathcal{A}}} - \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\right) + \lambda_2 \widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} = \mathbf{0}.$$

Since $\widetilde{\mathbf{B}}_{\cdot\bar{\mathcal{S}}} = \mathbf{0}$, it follows that

$$2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}_{\cdot\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}} - \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\right) + \lambda_2 \widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} = \mathbf{0}. \tag{S9.39}$$

Now for every $i \in \bar{\mathcal{S}}$, it follows from (S9.39) that

$$\begin{aligned} & 2\left\|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}\mathbf{C}_{-\bar{\mathcal{D}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}\right)\right\|_2 \\ = \; & 2\left\|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}\left(\widetilde{\mathbf{B}}_{\cdot\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}\right)\right\|_2 \\ \leq \; & 2\left\|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}\left(\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\mathbf{C}_{\bar{\mathcal{A}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}\right)\right\|_2+\lambda_2\left\|\widetilde{\mathbf{t}}_{\bar{\mathcal{A}}}\right\|_2\left\|\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\mathbf{C}_{\bar{\mathcal{A}},i}\right\|_2. \end{aligned} \tag{S9.40}$$

Note that $\left\|\widetilde{\mathbf{t}}_{\bar{\mathcal{A}}}\right\|_2 \leq \sqrt{r_{\bar{\mathcal{D}}}r_{\bar{\mathcal{A}}}}\max_{1\leq i\leq r_{\bar{\mathcal{A}}}}\tilde{w}_i$. Also, $\bar{\boldsymbol{\Omega}}_{\bar{\mathcal{D}},\bar{\mathcal{S}}}=\mathbf{0}$, implies that $\boldsymbol{\varepsilon}_{\bar{\mathcal{D}}}$ and $\boldsymbol{\varepsilon}_{\bar{\mathcal{S}}}$ are conditionally independent given $\boldsymbol{\varepsilon}_{\bar{\mathcal{A}}}$, which further implies that

$$\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{S}}}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}}}^{-1}\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}},\bar{\mathcal{S}}}=\mathbf{0}.$$

Hence, it follows by the triangle inequality and $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2$ that

$$\begin{aligned} & \left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\mathbf{C}_{\bar{\mathcal{A}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}\right\|_2 \\ = \; & \left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\mathbf{C}_{\bar{\mathcal{A}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}+\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},i}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}}}^{-1}\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}},i}\right\|_2 \\ \leq \; & \left\|\mathbf{C}_{\bar{\mathcal{D}},i}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},i}\right\|_2+\left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\right\|_2\left\|\mathbf{C}_{\bar{\mathcal{A}},i}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}},i}\right\|_2+ \\ & \left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\right\|_2\left\|\mathbf{C}_{\bar{\mathcal{A}}}^{-1}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}}}^{-1}\right\|_2\left\|\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}},i}\right\|_2+\left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}-\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\right\|_2\left\|\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}}}^{-1}\bar{\boldsymbol{\Sigma}}_{\bar{\mathcal{A}},i}\right\|_2. \end{aligned} \tag{S9.41}$$

Note that $\|\mathbf{U}\|_2 \leq \sqrt{ab}\|\mathbf{U}\|_{\max}$ for any $a \times b$ matrix $\mathbf{U}$. It follows by Assumption 1, Assumption 4, (S9.27), (S9.36) and (S9.41) that for every $\eta > 0$, there exists a constant $C_{6,\eta}$ (not depending on i or n) such that

$$2\left\|\widehat{\boldsymbol{\Omega}}_{\bar{\mathcal{D}}}\right\|_2\left\|\mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}}\mathbf{C}_{\bar{\mathcal{A}}}^{-1}\mathbf{C}_{\bar{\mathcal{A}},i}-\mathbf{C}_{\bar{\mathcal{D}},i}\right\|_2 < C_{6,\eta}\sqrt{\frac{\left(r_{\bar{\mathcal{D}}}^2 r_{\bar{\mathcal{A}}}+r_{\bar{\mathcal{D}}}r_{\bar{\mathcal{A}}}^2\right)\log r_n}{n}} \tag{S9.42}$$

with probability at least $1 - 8r_n^{-\eta}$ for large enough n. Hence, by (S9.40), (S9.42) and Assumption 8, we get

$$\left\| 2\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}} \left(\widetilde{\mathbf{B}} \mathbf{C}_{-\bar{\mathcal{D}},i} - \mathbf{C}_{\bar{\mathcal{D}},i} \right) \right\|_2 \leq \lambda_2 \tilde{w}_i \tag{S9.43}$$

for every $i \in \bar{\mathcal{S}}$, which implies $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}}$ and $\widehat{\mathbf{B}}_{\cdot\bar{\mathcal{S}}} = \mathbf{0}$ with probability at least $1 - 8r_n^{-\eta}$ for large enough n, where $\widehat{\mathbf{B}}$ is the minimizer of $\tilde{f}_2(\mathbf{B})$.

Now, by (S9.39) it follows that

$$\widehat{\mathbf{B}}_{\cdot\bar{\mathcal{A}}} = \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \mathbf{C}_{\bar{\mathcal{A}}}^{-1} - \frac{\lambda_2}{2} (\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} \widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} \mathbf{C}_{\bar{\mathcal{A}}}^{-1}.$$

Since

$$\bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}} = \bar{\mathbf{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \bar{\mathbf{\Sigma}}_{\bar{\mathcal{A}}}^{-1},$$

we have

$$\begin{aligned}
& \left\| \widehat{\mathbf{B}}_{\cdot\bar{\mathcal{A}}} - \bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}} \right\|_{\max} \\
\leq \; & \left\| \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \mathbf{C}_{\bar{\mathcal{A}}}^{-1} - \bar{\mathbf{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \bar{\mathbf{\Sigma}}_{\bar{\mathcal{A}}}^{-1} \right\|_2 + \frac{\lambda_2}{2} \left\| (\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} \right\|_2 \left\| \widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} \right\|_2 \left\| \mathbf{C}_{\bar{\mathcal{A}}}^{-1} \right\|_2 \\
\leq \; & \left\| \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} - \bar{\mathbf{\Sigma}}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \right\|_2 \left\| \bar{\mathbf{\Sigma}}_{\bar{\mathcal{A}}}^{-1} \right\|_2 + \left\| \mathbf{C}_{\bar{\mathcal{D}},\bar{\mathcal{A}}} \right\|_2 \left\| \mathbf{C}_{\bar{\mathcal{A}}}^{-1} - \bar{\mathbf{\Sigma}}_{\bar{\mathcal{A}}}^{-1} \right\|_2 + \frac{\lambda_2}{2} \left\| (\widehat{\mathbf{\Omega}}_{\bar{\mathcal{D}}})^{-1} \right\|_2 \left\| \widetilde{\mathbf{t}}_{\bar{\mathcal{A}}} \right\|_2 \left\| \mathbf{C}_{\bar{\mathcal{A}}}^{-1} \right\|_2 .
\end{aligned}$$

It follows by (S9.27), (S9.36), and Assumptions 4 and 8 that for every $\eta > 0$, there exists a constant $C_{7,\eta}$ such that

$$\left\| \widehat{\mathbf{B}}_{\cdot\bar{\mathcal{A}}} - \bar{\mathbf{B}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}} \right\|_{\max} \leq C_{7,\eta} \sqrt{\frac{\left(r_{\bar{\mathcal{D}}}^2 r_{\bar{\mathcal{A}}} + r_{\bar{\mathcal{D}}} r_{\bar{\mathcal{A}}}^2 \right) \log r_n}{n}}$$

with probability at least $1 - 8r_n^{-\eta}$ for large enough n. We conclude from Assumption 6 that $\widehat{\mathbf{B}}_{\cdot i} \neq \mathbf{0}$ for every $1 \leq i \leq r_{\bar{\mathcal{A}}}$ with probability at least

$1 - 8r_n^{-\eta}$ for large enough n. Since we already proved that $\widehat{\mathbf{B}}_{.\bar{\mathcal{S}}} = \mathbf{0}$ and with probability at least $1 - 8r_n^{-\eta}$ for large enough n, the result in part (b) follows by recalling that we have restricted to the event $\widehat{\mathcal{D}} = \bar{\mathcal{D}}$ which holds with probability at least $1 - 6r_n^{-\eta}$ for large enough n.

S10. Proof of Theorem 4

Note that $\text{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \text{vec}(\bar{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) = \text{vec}(\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},1}) - \text{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})$ when $\widehat{\mathcal{D}} = \bar{\mathcal{D}}, \widehat{\mathcal{A}} = \bar{\mathcal{A}}$. Note that Assumptions 1-8 along with $r_{\bar{\mathcal{D}}} = O(1)$ ensure that

- $\bar{P}(\widehat{\mathcal{D}} = \bar{\mathcal{D}}, \widehat{\mathcal{A}} = \bar{\mathcal{A}}) \to 1$ as $n \to \infty$.
- $r_{\bar{\mathcal{A}}} = o(n)$.
- The eigenvalues of $\boldsymbol{\Sigma}$ (and all its principal submatrices) are uniformly bounded (in n) away from zero and infinity.

Since the error distribution is assumed to be normal, the required result follows immediately from the above facts and part (b) of Proposition S2.

S11. Tuning

Based on previous studies in Chen and Huang (2012) and Zou (2006), it is sufficient to select γ_1 and γ_2 from a small set like $\{0.5, 1, 2\}$. Our experience found that larger γ's usually leads to smaller models. To be more

conservative, all our numerical experiments are conducted at $\gamma_1 = 0.5$ and $\gamma_2 = 0.5$. The tuning parameters λ_1 and λ_2 can be selected using likelihood based methods like AIC, BIC or nonparameteric methods such as cross-validation. We used cross validation in both steps to select λ_1 and λ_2. The warm-start trick (Friedman et al., 2010) is implemented to increase computation efficiency.

In high-dimensional setting, when applying the CONCORD estimator, we standardize each columns of $\mathbf{R}$ first, and select the tuning parameter λ by 5-fold cross validation. For a sequence $\lambda_1 < \cdots < \lambda_k$, we select the tuning parameter that yields an estimator of $\boldsymbol{\Omega}$ that minimizes

$$-\sum_{i=1}^{r_n} \omega_{ii} + \operatorname{tr}(\boldsymbol{\Omega}^2 \mathbf{S}_{\mathbf{Y}|\mathbf{X}}).$$

Then resulting estimator is then rescaled back by the standard deviations of the original variables in $\mathbf{R}$ to obtain $\widehat{\boldsymbol{\Omega}}_{\text{con}}$.

S12. Additional Simulation

The first simulation investigates the performance of the response variable selection algorithm in large sample setting. We fixed $r = 10$, $r_{\mathcal{D}} = 6$, $r_{\mathcal{A}} = 2$, $p = 8$, and generated the data from model (2.6). Elements in $\boldsymbol{\beta}_{\mathcal{D}}$ were independent $N(0, 0.5^2)$ variates, the intercept was $\boldsymbol{\alpha} = \mathbf{0}$, and elements in $\mathbf{X}$ were independent $N(0, 0.5^2)$ variate. We varied the strength

of the association between $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ by generating the covariance matrix $\mathbf{\Sigma}$ such that $\rho^2_{\max}$, the squared largest canonical correlation between $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$, is about 0.9, 0.5 and 0.06 for each sample size 100, 200, 300, 500, 800 and 1200. Then 200 datasets were generated for each scenario. Given a dataset, we selected the dynamic, ancillary and static responses using the algorithm in Section 3.2. To evaluate the selection performance, we computed true positive rates $\text{TPR}_{\mathcal{D}}$, $\text{TPR}_{\mathcal{A}}$ and $\text{TPR}_{\mathcal{S}}$ for all three categories of the responses: $\text{TPR}_{\mathcal{D}} = |\bar{\mathcal{D}} \cap \widehat{\mathcal{D}}|_c / |\bar{\mathcal{D}}|_c$, $\text{TPR}_{\mathcal{A}} = |\bar{\mathcal{A}} \cap \widehat{\mathcal{A}}|_c / |\bar{\mathcal{A}}|_c$ and $\text{TPR}_{\mathcal{S}} = |\bar{\mathcal{S}} \cap \widehat{\mathcal{S}}|_c / |\bar{\mathcal{S}}|_c$, where for a set S, $|S|_c$ denotes its cardinality. We took the average of true positive rates over 200 replications. The results are in Table 1. The results confirm the selection consistency stated in

Table 1: Summary of selection performance as well as efficiency comparison

n	$\text{TPR}_{\mathcal{D}}$	$\text{TPR}_{\mathcal{A}}$	$\text{TPR}_{\mathcal{S}}$	R_{median}	$\text{TPR}_{\mathcal{D}}$	$\text{TPR}_{\mathcal{A}}$	$\text{TPR}_{\mathcal{S}}$	R_{median}	$\text{TPR}_{\mathcal{D}}$	$\text{TPR}_{\mathcal{A}}$	$\text{TPR}_{\mathcal{S}}$	R_{median}
	$\rho^2_{\max} = 0.9$				$\rho^2_{\max} = 0.5$				$\rho^2_{\max} = 0.06$			
50	0.993	0.915	0.655	2.868	0.989	0.950	0.635	1.292	1.000	0.743	0.868	0.990
100	1.000	1.000	0.875	3.719	1.000	0.995	0.778	1.245	0.998	0.525	0.870	0.965
200	1.000	1.000	0.895	4.809	1.000	1.000	0.833	1.427	0.975	0.450	0.620	0.960
500	1.000	1.000	0.948	5.303	1.000	1.000	0.848	1.371	1.000	0.965	0.823	1.003
1200	1.000	1.000	1.000	6.936	1.000	1.000	0.975	1.354	1.000	1.000	1.000	1.012

Theorem 1: When n is large, the dynamic, ancillary and static responses are correctly selected with probability tending to 1. Among the three rates, $\text{TPR}_{\mathcal{D}}$ is the largest and close to 1 even when $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ are weekly correlated $\rho^2_{\max} = 0.06$. Since the dynamic responses contain the most important information, this is a desirable property. The algorithm is also

very effective in the selection of the ancillary response, which is the second most important category. When $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ are weekly correlated, it becomes more difficult to select the ancillary responses. But the selection performance improves when n increases. We measured the efficiency gain of a randomly picked element, say β_{ij}, by the efficiency ratio R_{ij} defined as

$$R_{ij} = \frac{\text{var}(\tilde{\beta}_{ij})}{\text{var}(\hat{\beta}_{ij})}, \tag{S12.44}$$

where $\text{var}(\tilde{\beta}_{ij})$ and $\text{var}(\hat{\beta}_{ij})$ are the variances of the OLS estimator $\tilde{\beta}_{ij}$ and our estimator $\hat{\beta}_{ij}$ calculated based on 200 replications. Then R_{median} is the median of all the R_{ij} for the nonzero elements in $\boldsymbol{\beta}$. We notice that when $\rho^2_{\max} = 0.9$, we achieve substantial efficiency gains. When $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ are moderated correlated that $\rho^2_{\max} = 0.5$, the efficiency gain is reduced. However, when $R_{\text{median}} = 1.292$, by performing the response variable selection procedure, we can reduce the sample size by 23% while achieving the same efficiency as the using all the response variables, while is considered a worthwhile gain in many applications. When $\rho_{\max} = 0.06$, $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ are weekly correlated. The estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ has about the same efficiency as $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, which is computed using all the response variables.

For each dataset, we also computed the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\text{oracle}}$, and compared it with $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$. The standard deviations of each element in

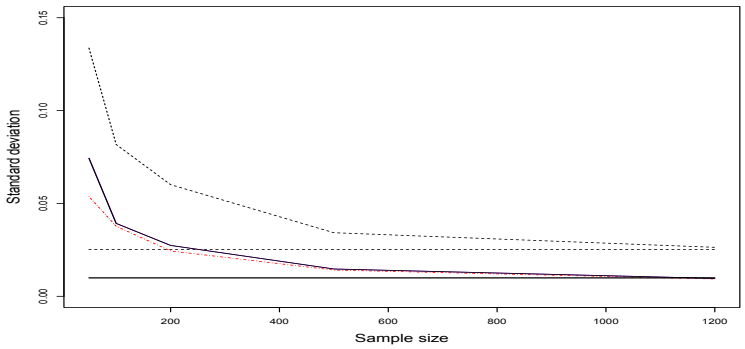

Figure 1: Comparison of $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\text{oracle}}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$. Solid lines mark $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$, the dash-dotted line marks the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\text{oracle}}$, and dashed lines mark $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$. The horizontal lines mark the asymptotic standard deviations of the corresponding estimators. Note that the asymptotic standard deviation of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\text{oracle}}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ is the same by Theorem 2.

$\boldsymbol{\beta}_{\mathcal{D}}$ for the three estimators were calculated. The results for a randomly selected element are summarized in Figure 1. Note that the OLS estimator $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ only uses the dynamic responses $\mathbf{Y}_{\mathcal{D}}$ for estimation. When the sample size is 50, the ratio of the standard deviations of $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ versus $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ is 1.80, which means that by including the ancillary responses $\mathbf{Y}_{\mathcal{A}}$ we reduces the sample size by about 70% compared to the regular OLS estimator using all response variables. We also notice that $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\text{oracle}}$ have similar standard deviations, especially when the sample size is large, which confirms the optimal estimation rate stated in Theorem 2.

S13. Generation of $\boldsymbol{\Sigma}$ for Simulations in Section 4.1

To generate the $\boldsymbol{\Sigma}$ for the simulations in Section 4.1, we started with a matrix $\mathbf{A}$ with diagonal elements 1 and off-diagonal elements 0.9. Then the elements in $\mathbf{A}^{-1}$ that correspond to $\boldsymbol{\Omega}_{\mathcal{D},\mathcal{S}}$ were set to zero to obtain matrix $\mathbf{B}$. Note that $\mathbf{B}$ has the same sparsity structure as $\boldsymbol{\Omega}$ in (2.6), but $\mathbf{B}$ may not be positive definite. To achieve positive definiteness, we added a positive definite matrix to $\mathbf{B}$. More specifically, we took $\mathbf{C} = \mathbf{B} + 0.1\mathbf{M}\mathbf{M}^T$, where each element in $\mathbf{M}$ that corresponds to $\boldsymbol{\Omega}_{-\mathcal{D},-\mathcal{D}}$ was an independent uniform $(0, 1)$ variate and other elements were zero. The matrix $\mathbf{C}$ preserves the sparsity structure of $\mathbf{B}$ and is positive definite due to the properties of

eigenvalues of $\mathbf{B}$. Then we took $\mathbf{\Sigma}_0 = \mathbf{C}^{-1}$. We checked the eigenvalue of $\mathbf{\Sigma}_0$, and found that while the largest eigenvalue is upper bounded by 1, the smallest eigenvalue goes to 0 as $r \to \infty$. We also observed that the second smallest eigenvalue is always greater than 0.01 as $r \to \infty$. To ensure that the eigenvalues of $\mathbf{\Sigma}$ are uniformly bounded as required in Assumption 1, we made a slight modification to $\mathbf{\Sigma}_0$. We performed a spectral decomposition of the matrix $\mathbf{\Sigma}_0$. Let $(\lambda_i, \mathbf{v}_i)$ denote the ith eigenvalue-eigenvector pair (with eigenvalues organized in descending order). Then we take $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + (0.01 - \lambda_r)\mathbf{v}_r\mathbf{v}_r^T$, and the eigenvalues of $\mathbf{\Sigma}$ are bounded below by 0.01 and bounded above by 1.

S14. Further work

The concept of response variable selection can be extended to other contexts where multivariate responses are involved, like generalized linear regression, reduced rank regression (Izenman, 1975), partial least squares (Wold, 1966) and envelope models (Cook et al., 2010). It may also be extended to linear regression models with a matrix-variate or tensor variate response (Li et al., 2011), which can have applications in neuroimaging data.

References

Chen, L. and J. Z. Huang (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc. 107*(500), 1533–1545.

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

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(1), 1–15.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Ghosh, S., K. Khare, and G. Michailidis (2018). High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association 114*(526), 735–748.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis 5*(2), 248–264.

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(4), 803–825.

Li, Y., H. Zhu, D. Shen, W. Lin, J. H. Gilmore, and J. G. Ibrahim (2011). Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 559–578.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Rudelson, M. and R. Vershynin (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab 18*(82), 1–9.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. Krishnaiah (Ed.), *In Multivariate Analysis*, Volume 59, pp. 391–420. Academic Press, NY.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.