Biometrika (2012), **99**, 1, *pp*. 1–21

Printed in Great Britain
Advance Access publication on 31 July 2012

Supplementary material for "Sparse Envelope Model: Efficient Estimation and Response Variable Selection in Multivariate Linear Regression"

BY Z. SU, G. ZHU
Department of Statistics, University of Florida, Gainesville, Florida, USA
zhihuasu@stat.ufl.edu gzhu22@ufl.edu

X. CHEN
Department of Statistics and Applied Probability, National University of Singapore, Singapore
stacx@nus.edu.sg

AND Y. YANG
Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada
yi.yang6@mcgill.ca

A. PROOFS

Proof of Proposition 1. We will prove this proposition by contradiction. Without loss of generality, let $\{Y_1, \dots, Y_{u-1}\}$ be the collection of all active response variables that are connected with a response that has non-zero regression coefficients, and let Y_u be a response which has regression coefficient zero and is not connected with any of the responses that have non-zero regression coefficients. We will show that Y_u is inactive.

Let $e_u \in \mathbb{R}^r$ be a vector of zeros but having 1 at its uth element and let $\Gamma^* = Q_{e_u}\Gamma$. Since Y_u has regression coefficients zero, $\beta = Q_{e_u}\beta$, giving $\mathcal{B} = Q_{e_u}\mathcal{B}$. Therefore

$$\mathcal{B} = Q_{e_u}\mathcal{B} \subseteq Q_{e_u}\mathrm{span}(\Gamma) = \mathrm{span}(\Gamma^*).$$

Because this Y_u is not connected with any of the responses that have non-zero regression coefficients, $\mathrm{cov}\{Y_u, (Y_1, \dots, Y_{u-1})^{\mathrm{T}} \mid X\} = 0$, so $\mathrm{cov}(Y_u, \Gamma^{*\mathrm{T}}Y \mid X) = 0$. Recall that $\mathrm{cov}(\Gamma_0^{\mathrm{T}}Y, \Gamma^{\mathrm{T}}Y \mid X) = 0$, so $\mathrm{cov}(\Gamma_0^{\mathrm{T}}Y, \Gamma^{*\mathrm{T}}Y \mid X) = 0$. Notice that

$$\mathrm{span}(\Gamma^*)^{\perp} = \mathrm{span}(\Gamma_0) + \mathrm{span}(P_{e_u}\Gamma) = \mathrm{span}(\Gamma_0) + \mathrm{span}(e_u),$$

where $\perp$ denotes orthogonal complement of a subspace. If $\tilde{\Gamma}$ is an orthogonal basis of $\mathrm{span}(\Gamma^*)^{\perp}$, then $\tilde{\Gamma} = P_{\Gamma_0}\tilde{\Gamma} + P_{e_u}\tilde{\Gamma}$. So

$$\mathrm{cov}(\tilde{\Gamma}^{\mathrm{T}}Y, \Gamma^{*\mathrm{T}}Y \mid X) = \mathrm{cov}(\tilde{\Gamma}^{\mathrm{T}}P_{\Gamma_0}Y + \tilde{\Gamma}^{\mathrm{T}}P_{e_u}Y, \Gamma^{*\mathrm{T}}Y \mid X) = 0.$$

Therefore $\mathrm{span}(\Gamma^*)$ is a reducing subspace of Σ that contains $\mathcal{B}$. As $\Gamma^* = Q_{e_u}\Gamma$, its dimension is smaller or equal to $\mathrm{span}(\Gamma)$. Since $\mathrm{span}(\Gamma)$ is the envelope subspace, $\mathrm{span}(\Gamma^*) = \mathrm{span}(\Gamma)$. This is because if not, $\mathrm{span}(\Gamma^*) \cap \mathrm{span}(\Gamma)$, which has a smaller dimension than $\mathrm{span}(\Gamma)$, is a reducing subspace of Σ that contains $\mathcal{B}$; and it contradicts the definition of the envelope subspace. Since $\mathrm{span}(\Gamma^*) = \mathrm{span}(\Gamma)$, the ith row of Γ must be zero, and Y_u is an inactive response. □

Now we discuss about the relationship between the two statements: (a) Y_i and Y_j are not connected and (b) Y_i and Y_j are independent given the rest of the responses and X. If we assume normality, (a) implies (b), but (b) does not imply (a). If normality is not assumed, they do not imply each other.

First we show that (b) does not imply (a). Statement (b) is based on the structure of Σ^{-1}: If Y_i and Y_j are independent given the rest of the responses and X, then the (i,j)th element in Σ^{-1} is zero. On the other hand, if Y_i and Y_j are connected is based on the structure of Σ. A sparse Σ^{-1} does not necessarily imply a sparse Σ. For example, suppose that Y_2 and Y_3 are independent given Y_1 and X, and

$$\Sigma^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 3 \end{pmatrix}.$$

Then

$$\Sigma = \begin{pmatrix} 6 & -3 & -2 \\ -3 & 2 & 1 \\ -2 & 1 & 1 \end{pmatrix},$$

and Y_2 and Y_3 are connected.

Now suppose that Y_i and Y_j are not connected. Without loss of generality, we assume that Y_1 and Y_r are not connected. For positive integers $k \geq 2$ and $l \geq 1$, let $Y_2, \ldots, Y_k$ be the responses that connect with Y_1, $Y_{k+1}, \ldots, Y_{k+l}$ be the responses that neither connect with Y_1 nor connect with Y_r, and $Y_{k+l+1}, \ldots, Y_{r-1}$ be the responses that connect with Y_r. Then Σ has a block diagonal structure as follows:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{k,1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{k,k} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{k+1,k+1} & \cdots & \sigma_{k+1,k+l} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \sigma_{k+l,k+1} & \cdots & \sigma_{k+l,k+l} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \sigma_{k+l+1,k+l+1} & \cdots & \sigma_{k+l+1,r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \sigma_{r,k+l+1} & \cdots & \sigma_{r,r} \end{pmatrix}.$$

The inverse matrix Σ^{-1} will preserve the block diagonal structure of Σ, so the $(1,r)$th element in Σ^{-1} is 0. Under the normality assumption, this implies Y_1 and Y_r are independent given the rest of the responses and X. If normality is not assumed, this does not imply the conditional independent of Y_1 and Y_r.

Proof of Theorem 2. To prove Theorem 2, denote the objective function in (6) by $f_{\mathrm{obj}}(A)$. It is sufficient to show that for any small $\epsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_n \mathrm{pr}\left\{ \inf_{\Delta \in \mathbb{R}^{(r-u)\times u}, \|\Delta\|_F = C} f_{\mathrm{obj}}(A + n^{-1/2}\Delta) > f_{\mathrm{obj}}(A) \right\} > 1 - \epsilon. \tag{A1}$$

If (A1) is established, then there exists a local minimizer $\widehat{A}$ of f_{obj} with arbitrarily large probability such that $\|\widehat{A} - A\|_F = O_p(n^{-1/2})$. Therefore $\widehat{A}$ is a $\sqrt{n}$-consistent estimator of A. As $P_\Gamma = G_A(I_u + A^{\mathrm{T}}A)^{-1}G_A^{\mathrm{T}}$ is a function of A only, $P_{\widehat{\Gamma}}$ is a $\sqrt{n}$-consistent estimator of P_Γ. As $\widehat{\beta} = P_{\widehat{\Gamma}}\widehat{\beta}_{\mathrm{ols}}$, and $\widehat{\beta}_{\mathrm{ols}}$ is a $\sqrt{n}$-consistent estimator of β, then $\widehat{\beta}$ is a $\sqrt{n}$-consistent estimator of β.

Now we only need to show (A1). We write

$$\begin{aligned} f_{\mathrm{obj}}(A) &= -2\log|G_A^{\mathrm{T}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A| + \log|G_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}G_A| + \sum_{i=1}^{r-u}\lambda_i\|a_i\|_2 \\ &\equiv f_1(A) + f_2(A) + f_3(A) + f_4(A), \end{aligned}$$

say, and we first focus on $f_1(A) = -2\log|G_A^{\mathrm{T}} G_A|$. Expand $f_1(A + n^{-1/2}\Delta)$, we have

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

where $\overset{\to\Delta}{df_1}(A)$ and $\overset{\to\Delta}{df_1^2}(A)$ are directional derivatives (Dattorro, 2005, p.706).

The first directional derivative is

$$\overset{\to\Delta}{df_1}(A) = \operatorname{tr}\left\{\frac{d}{dA} f_1(A)^{\mathrm{T}}\Delta\right\} = -4\operatorname{tr}\{(I_u + A^T A)^{-1} A^T \Delta\}.$$

The second directional derivative is

$$\begin{aligned}\overset{\to\Delta}{df_1^2}(A) &= -4\operatorname{tr}\left(\left[\frac{d}{dA}\operatorname{tr}\{(I_u + A^T A)^{-1}A^{\mathrm{T}}\Delta\}\right]^{\mathrm{T}}\Delta\right)\\ &= -4\operatorname{tr}\Big[\Big\{-A(I_u + A^T A)^{-1}(A^T\Delta + \Delta^T A)(I_u + A^T A)^{-1} + \Delta(I_u + A^T A)^{-1}\Big\}^T\Delta\Big]\\ &= 4\operatorname{tr}\Big\{(I_u + A^T A)^{-1}(A^T\Delta + \Delta^T A)(I_u + A^T A)^{-1}A^T\Delta - (I_u + A^T A)^{-1}\Delta^T\Delta\Big\}.\end{aligned}$$

Let

$$\Delta_* = \begin{pmatrix} 0 \\ \Delta \end{pmatrix};$$

then

$$\begin{aligned}&\overset{\to\Delta}{df_1^2}(A)\\ &= 4\operatorname{tr}\Big[(I_u + A^T A)^{-1}A^T\Delta(I_u + A^T A)^{-1}A^T\Delta\\ &\quad +(I_u + A^T A)^{-1}\Delta^{\mathrm{T}}\big\{A(I_u + A^T A)^{-1}A^T - I_{r-u}\big\}\Delta\Big]\\ &= 4\operatorname{tr}\Big[(I_u + A^T A)^{-1}A^T\Delta(I_u + A^T A)^{-1}A^T\Delta + (I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\big\{G_A(G_A^T G_A)^{-1}G_A^T - I_r\big\}\Delta_*\Big]\\ &= 4\operatorname{tr}\Big\{(I_u + A^T A)^{-1}A^T\Delta(I_u + A^T A)^{-1}A^T\Delta + (I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}(\Gamma\Gamma^{\mathrm{T}} - I_r)\Delta_*\Big\}\\ &= 4\operatorname{tr}\Big\{(I_u + A^T A)^{-1}A^T\Delta(I_u + A^T A)^{-1}A^T\Delta - (I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\}.\end{aligned}$$

We substitute $\overset{\to\Delta}{df_1}(A)$ and $\overset{\to\Delta}{df_1^2}(A)$ into the expansion for $f_1(A + n^{-1/2}\Delta)$ and get

$$\begin{aligned}f_1(A + n^{-1/2}\Delta) - f_1(A) = &-4n^{-1/2}\operatorname{tr}\{(I_u + A^T A)^{-1}A^T\Delta\}\\ &+2n^{-1}\operatorname{tr}\Big\{(I_u + A^T A)^{-1}A^T\Delta(I_u + A^T A)^{-1}A^T\Delta\\ &-(I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\}.\end{aligned}$$

With $f_2(A) = \log|G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A|$, the first directional derivative is

$$\overset{\to\Delta}{df_2}(A) = \operatorname{tr}\left\{\frac{d}{dA}f_2(A)^{\mathrm{T}}\Delta\right\} = 2\operatorname{tr}\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_*\}.$$

Let Σ_X, Σ_Y and Σ_{YX} be the variance matrix of X, the variance matrix of Y and the covariance matrix of Y and X in population, and let $\widehat{\Sigma}_X$, $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_{XY}$ be the corresponding sample versions. Then by Cook

& Setodji (2003),

$$n^{1/2}(\widehat{\Sigma}_{YX} - \Sigma_{YX}) = n^{-1/2}(\mathbb{Y}_c^{\mathrm{T}}\mathbb{X} - n\Sigma_{YX}) + O_p(n^{-1/2}),$$
$$n^{1/2}(\widehat{\Sigma}_{X} - \Sigma_{X}) = n^{-1/2}(\mathbb{X}^{\mathrm{T}}\mathbb{X} - n\Sigma_{X}) + O_p(n^{-1/2}),$$
$$n^{1/2}(\widehat{\Sigma}_{Y} - \Sigma_{Y}) = n^{-1/2}(\mathbb{Y}_c^{\mathrm{T}}\mathbb{Y}_c - n\Sigma_{Y}) + O_p(n^{-1/2}),$$

where $\mathbb{Y}_c \in \mathbb{R}^{n\times r}$ is the centred data matrix of Y, whose ith row is $(Y_i - \bar{Y})^{\mathrm{T}}$. Since $\widehat{\Sigma}_{\mathrm{res}} = \widehat{\Sigma}_Y - \widehat{\Sigma}_{YX}\widehat{\Sigma}_X^{-1}\widehat{\Sigma}_{XY}$ and $\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1} = -\Sigma_X^{-1}(\widehat{\Sigma}_X - \Sigma_X)\Sigma_X^{-1} + O_p(n^{-1})$,

$$\begin{aligned}
\widehat{\Sigma}_{\mathrm{res}} &= (\widehat{\Sigma}_Y - \Sigma_Y + \Sigma_Y) - (\widehat{\Sigma}_{YX} - \Sigma_{YX} + \Sigma_{YX})(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1} + \Sigma_X^{-1})(\widehat{\Sigma}_{XY} - \Sigma_{XY} + \Sigma_{XY})\\
&= \Sigma + n^{-1/2}\Big\{ - n^{-1/2}(\mathbb{Y}_c^{\mathrm{T}}\mathbb{X} - n\Sigma_{YX})\Sigma_X^{-1}\Sigma_{XY} + n^{-1/2}\Sigma_{YX}\Sigma_X^{-1}(\mathbb{X}^{\mathrm{T}}\mathbb{X} - n\Sigma_X)\Sigma_X^{-1}\Sigma_{XY}\\
&\quad -n^{-1/2}\Sigma_{YX}\Sigma_X^{-1}(\mathbb{X}^{\mathrm{T}}\mathbb{Y}_c - n\Sigma_{XY}) + n^{-1/2}(\mathbb{Y}_c^{\mathrm{T}}\mathbb{Y}_c - n\Sigma_Y)\Big\} + O_p(n^{-1})\\
&\equiv \Sigma + n^{-1/2}(T_{1n} + T_{2n} + T_{3n} + T_{4n}) + O_p(n^{-1}),
\end{aligned}$$

where by the central limit theorem, each element in T_{1n}, T_{2n}, T_{3n} and T_{4n} converges in distribution to a normal random variable which has mean 0. As

$$\begin{aligned}
(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1} &= (G_A^{\mathrm{T}}\Sigma G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A - G_A^{\mathrm{T}}\Sigma G_A)(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + O_p(n^{-1})\\
&= (G_A^{\mathrm{T}}\Sigma G_A)^{-1} - n^{-1/2}(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(T_{1n} + T_{2n} + T_{3n} + T_{4n})G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\\
&\quad +O_p(n^{-1}),
\end{aligned}$$

$\overrightarrow{df_2}^{\mathbf{Z}^*}(\Gamma)$ can be expanded as

$$\begin{aligned}
&2\,\mathrm{tr}\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_*\}\\
&= 2\,\mathrm{tr}\{(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma\Delta_*\} + 2n^{-1/2}\,\mathrm{tr}\Big\{(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(T_{1n} + T_{2n} + T_{3n} + T_{4n})\Delta_*\\
&\quad -(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(T_{1n} + T_{2n} + T_{3n} + T_{4n})G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma\Delta_*\Big\} + O_p(n^{-1})\\
&= 2\,\mathrm{tr}\{(I_u + A^TA)^{-1}G_A^T\Delta_*\} + 2n^{-1/2}\,\mathrm{tr}\Big[(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(T_{1n} + T_{2n} + T_{3n} + T_{4n})\big\{I_r\\
&\quad -G_A(I_u + A^TA)^{-1}G_A^T\big\}\Delta_*\Big] + O_p(n^{-1})\\
&= 2\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} + 2n^{-1/2}\,\mathrm{tr}\big\{(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\big\} + O_p(n^{-1})\\
&= 2\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} + 2n^{-1/2}\,\mathrm{tr}\big\{\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\big\} + O_p(n^{-1}).
\end{aligned}$$

The second equality is because $\Gamma = G_A\Gamma_1$, so

$$\Gamma_1^{\mathrm{T}}G_A^{\mathrm{T}}G_A\Gamma_1 = I \Rightarrow \Gamma_1^{\mathrm{T}}(I_u + A^TA)\Gamma_1 = I \Rightarrow I_u + A^TA = (\Gamma_1^{\mathrm{T}})^{-1}(\Gamma_1)^{-1} \Rightarrow (I_u + A^TA)^{-1} = \Gamma_1\Gamma_1^{\mathrm{T}},$$

and

$$(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma = \big\{(\Gamma\Gamma_1^{-1})^{\mathrm{T}}\Sigma\Gamma\Gamma_1^{-1}\big\}^{-1}(\Gamma\Gamma_1^{-1})^{\mathrm{T}}\Sigma = \Gamma_1\Omega^{-1}\Omega\Gamma^{\mathrm{T}} = \Gamma_1\Gamma_1^{\mathrm{T}}G_A^{\mathrm{T}} = (I_u + A^TA)^{-1}G_A^T.$$

Using the Cauchy–Schwarz inequality for matrix trace (Magnus & Neudecker, 2007, p.227),

$$\begin{aligned}
\Big|\,\mathrm{tr}\big\{\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\big\}\Big| &\le \|\Delta_*\|_F\|\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\Gamma_0^{\mathrm{T}}\|_F\\
&= \|\Delta\|_F\|\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\|_F.
\end{aligned}$$

The second directional derivative of f_2 is

$$
\begin{aligned}
\overset{\rightarrow\Delta}{df_2^2}(A) &= 2\,\mathrm{tr}\left(\left[\frac{d}{dA}\,\mathrm{tr}\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_*\}\right]^{\mathrm{T}}\Delta\right) \\
&= 2\,\mathrm{tr}\big\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}\Delta_*^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_* \\
&\quad -(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_* + \Delta_*^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1}G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\Delta_*\big\} \\
&= 2\,\mathrm{tr}\big\{(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\Delta_*^{\mathrm{T}}\Sigma\Delta_* - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}(G_A^{\mathrm{T}}\Sigma\Delta_* + \Delta_*^{\mathrm{T}}\Sigma G_A)(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma\Delta_*\big\} \\
&\quad +O_p(n^{-1/2}) \\
&= 2\,\mathrm{tr}\Big[-(I_u + A^TA)^{-1}G_A^T\Delta_*(I_u + A^TA)^{-1}G_A^T\Delta_* \\
&\quad +(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\Delta_*^{\mathrm{T}}\Sigma\big\{I_r - G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma\big\}\Delta_*\Big] + O_p(n^{-1/2}) \\
&= 2\,\mathrm{tr}\Big[-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta \\
&\quad +\big\{(\Gamma_1^{-1})^{\mathrm{T}}\Omega\Gamma_1^{-1}\big\}^{-1}\Delta_*^{\mathrm{T}}\Sigma\big\{I_r - G_A(I_u + A^TA)^{-1}G_A^T\big\}\Delta_*\Big] + O_p(n^{-1/2}) \\
&= 2\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + \Gamma_1\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Sigma\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\big\} + O_p(n^{-1/2}) \\
&= 2\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + \Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\big\} + O_p(n^{-1/2}).
\end{aligned}
$$

Substitute $\overset{\rightarrow\Delta}{df_2}(A)$ and $\overset{\rightarrow\Delta}{df_2^2}(A)$ into the expansion for $f_2(A + n^{-1/2}\Delta)$, we get

$$
\begin{aligned}
&f_2(A + n^{-1/2}\Delta) - f_2(A) \\
&= 2n^{-1/2}\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} + 2n^{-1}\,\mathrm{tr}\big\{\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\big\} \\
&\quad +n^{-1}\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + \Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\big\} + o_p(n^{-1}) \\
&\geq 2n^{-1/2}\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} - 2n^{-1}\|\Delta\|_F\|\Gamma_1\Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\|_F \\
&\quad +n^{-1}\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + \Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\big\} + o_p(n^{-1}).
\end{aligned}
$$

Since that f_3 has similar structure as f_2, the derivation above can be applied parallel to f_2, just with $\widehat{\Sigma}_{\mathrm{res}}$ replaced $\widehat{\Sigma}_Y^{-1}$. Let $T_{5n} = -n^{-1/2}\Sigma_Y^{-1}(\mathbb{Y}_c^{\mathrm{T}}\mathbb{Y}_c - n\Sigma_Y)\Sigma_Y^{-1}$. By the central limit theorem, T_{5n} converges in distribution to a normal random variable with mean 0. After some straightforward algebra, we have

$$
\begin{aligned}
&f_3(A + n^{-1/2}\Delta) - f_3(A) \\
&= 2n^{-1/2}\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} + 2n^{-1}\,\mathrm{tr}\{\Gamma_1(\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma^{\mathrm{T}}T_{5n}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\} \\
&\quad +n^{-1}\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\big\} \\
&\quad +o_p(n^{-1}) \\
&\geq 2n^{-1/2}\,\mathrm{tr}\{(I_u + A^TA)^{-1}A^T\Delta\} - 2n^{-1}\|\Delta\|_F\|\Gamma_1(\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma^{\mathrm{T}}T_{5n}\Gamma_0\|_F \\
&\quad +n^{-1}\,\mathrm{tr}\big\{-(I_u + A^TA)^{-1}A^T\Delta(I_u + A^TA)^{-1}A^T\Delta + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\big\} \\
&\quad +o_p(n^{-1}).
\end{aligned}
$$

Now we expand $f_4(A) = \sum_{i=1}^{r-u}\lambda_i\|a_i\|_2$. Let δ_i^{T} be the ith row of Δ, then

$$
\begin{aligned}
f_4(A + n^{-1/2}\Delta) - f_4(A) &\geq \sum_{i=1}^{q-u}\Big(\lambda_i\|a_i + n^{-1/2}\delta_i\|_2 - \lambda_i\|a_i\|_2\Big) \\
&\geq -\frac{1}{2}(q-u)n^{-1/2}\lambda_{\max,n}\max_i\left(\|a_i\|_2^{-1}\|\delta_i\|_2\right)\{1 + o_p(1)\} \\
&= -\frac{1}{2}n^{-1}(q-u)n^{1/2}\lambda_{\max,n}\max_i\left(\|a_i\|_2^{-1}\|\delta_i\|_2\right)\{1 + o_p(1)\}.
\end{aligned}
$$

The second inequality is based on Taylor expansion at a_i. As $n^{1/2}\lambda_{\max,n} \to 0$ as $n \to \infty$, $n\{f_4(A + n^{-1/2}\Delta) - f_4(A)\} = o_p(1)$. Collecting all the results so far

$$
\begin{aligned}
& f_{\mathrm{obj}}(A + n^{-1/2}\Delta) - f_{\mathrm{obj}}(A) \\
\geq\ & -2n^{-1}\|\Delta\|_F \|\Gamma_1 \Omega^{-1}\Gamma^{\mathrm{T}}(T_{3n} + T_{4n})\Gamma_0\|_F - 2n^{-1}\|\Delta\|_F \|\Gamma_1(\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma^{\mathrm{T}} T_{5n}\Gamma_0\|_F \\
& + n^{-1}\operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 \\
& - 2(I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\} - \frac{1}{2}n^{-1}(q-u)n^{1/2}\lambda_{\max,n}\max_i\left(\|a_i\|_2^{-1}\|\delta_i\|_2\right) + o_p(n^{-1}).
\end{aligned}
$$

Notice that

$$
\begin{aligned}
& \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 - 2(I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\} \\
=\ & \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 - 2\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\Big\} \\
=\ & \operatorname{vec}(\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1)^{\mathrm{T}}(\Omega \otimes \Omega_0^{-1} + \Omega^{-1}\otimes\Omega_0 - 2I_u \otimes I_{r-u} + \eta\Sigma_X\eta^{\mathrm{T}} \otimes \Omega_0^{-1})\operatorname{vec}(\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1) \\
\equiv\ & \operatorname{vec}(\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1)^{\mathrm{T}} K \operatorname{vec}(\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1) \\
\geq\ & m\|\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\|_F^2,
\end{aligned}
$$

where m is the smallest eigenvalue of K. The matrix K appears in (5.7) in Cook et al. (2010), by Shapiro (1986), K is a positive definite matrix and $m > 0$. Since

$$
\begin{aligned}
\|\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\|_F^2 =\ & \operatorname{tr}(\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0) \\
=\ & \operatorname{tr}\{\Gamma_0^{\mathrm{T}}\Delta_*(I_u + A^{\mathrm{T}}A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\} \\
=\ & \operatorname{tr}\{\Delta_*(I_u + A^{\mathrm{T}}A)^{-1}\Delta_*^{\mathrm{T}}(I_r - \Gamma\Gamma^{\mathrm{T}})\} \\
=\ & \operatorname{tr}\big[(I_u + A^{\mathrm{T}}A)^{-1}\Delta_*^{\mathrm{T}}\{I_r - G_A(I_u + A^{\mathrm{T}}A)^{-1}G_A^{\mathrm{T}}\}\Delta_*\big] \\
=\ & \operatorname{tr}\big[(I_u + A^{\mathrm{T}}A)^{-1}\Delta^{\mathrm{T}}\{I_{r-u} - A(I_u + A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\}\Delta\big] \\
=\ & \operatorname{tr}\{(I_u + A^{\mathrm{T}}A)^{-1}\Delta^{\mathrm{T}}(I_u + A^{\mathrm{T}}A)^{-1}\Delta\} \\
=\ & \operatorname{vec}(\Delta)^{\mathrm{T}}\{(I_u + A^{\mathrm{T}}A)^{-1}\otimes(I_u + A^{\mathrm{T}}A)^{-1}\}\operatorname{vec}(\Delta) \\
\geq\ & m_0^2\|\Delta\|_F^2,
\end{aligned}
$$

where m_0 is the smallest eigenvalue of $(I_u + A^{\mathrm{T}}A)^{-1}$, we have

$$
\begin{aligned}
& \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 - 2(I_u + A^T A)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\} \\
\geq\ & m m_0^2\|\Delta\|_F^2.
\end{aligned}
$$

Then the terms with order $\|\Delta\|_F^2$ dominate the terms with order $\|\Delta\|_F$. When $\|\Delta\|_F = C$ for sufficiently large C, the conclusion (A1) follows. □

Proof of Theorem 3. We will prove this theorem by contradiction. Suppose that $\|\widehat{a}_i\|_2 > 0$ for $i = q + 1 - u, \ldots, r - u$. The first derivative of f_{obj} with respect to a_i should be 0 evaluated at the local minimum $\widehat{a}_i$. The derivative of f_{obj} with respect to a_i^{T} $(i = q + 1 - u, \ldots, r - u)$ is

$$
\frac{\partial f_{\mathrm{obj}}}{\partial a_i^{\mathrm{T}}} = -4e_i^{\mathrm{T}}G_A(I_u + A^{\mathrm{T}}A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}G_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}G_A(G_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}G_A)^{-1} + \frac{\lambda_i a_i^{\mathrm{T}}}{\|a_i\|_2},
$$

where e_i be the ith column of I_r. Then

$$
-4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A)^{-1} + \lambda_i\frac{\widehat{a}_i^{\mathrm{T}}}{\|\widehat{a}_i\|_2} = 0. \tag{A2}
$$

Because $\widehat{\Sigma}_{\mathrm{res}}$, $\widehat{\Sigma}_Y$ and $\widehat{A}$ are $\sqrt{n}$-consistent estimators of Σ, Σ_Y and A, $\Sigma = \Gamma\Omega\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$ and $\Sigma_Y = \Gamma(\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$,

$$
\begin{aligned}
&-4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A)^{-1}\\
&= -4e_i^{\mathrm{T}}G_A(I_u + A^{\mathrm{T}}A)^{-1} + 2e_i^{\mathrm{T}}\Sigma G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + 2e_i^{\mathrm{T}}\Sigma_Y^{-1}G_A(G_A^{\mathrm{T}}\Sigma_Y^{-1}G_A)^{-1} + O_p(n^{-1/2})\\
&= -4a_i^{\mathrm{T}}(I_u + A^{\mathrm{T}}A)^{-1} + 2e_i^{\mathrm{T}}G_A(I_u + A^{\mathrm{T}}A)^{-1} + 2e_i^{\mathrm{T}}G_A(I_u + A^{\mathrm{T}}A)^{-1} + O_p(n^{-1/2})\\
&= -4a_i^{\mathrm{T}}(I_u + A^{\mathrm{T}}A)^{-1} + 2a_i^{\mathrm{T}}(I_u + A^{\mathrm{T}}A)^{-1} + 2a_i^{\mathrm{T}}(I_u + A^{\mathrm{T}}A)^{-1} + O_p(n^{-1/2})\\
&= O_p(n^{-1/2}).
\end{aligned}
$$

Then $n^{1/2}\left\{-4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A)^{-1}\right\} = O_p(1)$.

On the other hand, let m be the element in a_i that has the largest absolute value, then $|m|/\|a_i\|_2 > \sqrt{u}$, where $|\cdot|$ denotes absolute value. Because we have $n^{1/2}\lambda_{\min,n} \to \infty$, there is at least one element in $n^{1/2}\lambda_i a_i^{\mathrm{T}}/\|a_i\|_2$ that tends to infinity. With (A2), this is a contradiction of

$$
n^{1/2}\left\{-4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1}\widehat{G}_A)^{-1}\right\} = O_p(1).
$$

Therefore for $i = q+1-u, \ldots, r-u$, $a_i = 0$ with probability tending to 1. □

Proof of Proposition 2 *and Proposition* 3. For the proof of Proposition 3, the derivation of the maximum likelihood estimator of β_D and its asymptotic variance under model (13) follows from standard theory on regression. Now we start to proof Proposition 2. We need to justify the results for model (12). First we derive the maximum likelihood estimator of β_D. As $Y = (Y_D^{\mathrm{T}}, Y_S^{\mathrm{T}})^{\mathrm{T}}$, we can partition the centred matrix $\mathbb{Y}_c$ accordingly into $\mathbb{Y}_c = (\mathbb{Y}_{c,D}, \mathbb{Y}_{c,S})$. We also partition the matrix Σ^{-1} into

$$
\Sigma^{-1} = \begin{pmatrix} M_1 & M_2 \\ M_2^{\mathrm{T}} & M_3 \end{pmatrix}.
$$

The log likelihood function under model (12) is

$$
\begin{aligned}
l = &-\frac{n(r+p)}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma_X| - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\{(\mathbb{X} - 1_n\mu_X)\Sigma_X^{-1}(\mathbb{X} - 1_n\mu_X)^{\mathrm{T}}\}\\
&-\frac{1}{2}\mathrm{tr}\{(\mathbb{Y}_{c,D} - 1_n\alpha^{\mathrm{T}} - \mathbb{X}\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})\Sigma^{-1}(\mathbb{Y}_{c,D} - 1_n\alpha - \mathbb{X}\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})^{\mathrm{T}}\}.
\end{aligned}
$$

It is easy to show that $\mu_X = \bar{X}$, $\widehat{\Sigma}_X = (\mathbb{X} - 1_n\mu_X)^{\mathrm{T}}(\mathbb{X} - 1_n\mu_X)/n$, and $\widehat{\alpha} = \bar{Y}$. Substituting these estimates in, the partially maximized log likelihood is

$$
\begin{aligned}
l = &-\frac{n(r+p)}{2}\log(2\pi) - \frac{n}{2}\log|\widehat{\Sigma}_X| - \frac{np}{2} - \frac{n}{2}\log|\Sigma|\\
&-\frac{1}{2}\mathrm{tr}\{(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})\Sigma^{-1}(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})^{\mathrm{T}}\}\\
= &-\frac{n(r+p)}{2}\log(2\pi) - \frac{n}{2}\log|\widehat{\Sigma}_X| - \frac{np}{2} - \frac{n}{2}\log|\Sigma|\\
&-\frac{1}{2}\mathrm{tr}\{(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}})M_1(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}})^{\mathrm{T}} + 2(\mathbb{Y}_{c,D} - \mathbb{X}\beta_D^{\mathrm{T}})M_2\mathbb{Y}_{c,S}^{\mathrm{T}} + \mathbb{Y}_{c,S}M_3\mathbb{Y}_{c,S}^{\mathrm{T}}\}.
\end{aligned}
$$

Take the derivative of l with respect to β_D and Σ, we get

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_D} &= -M_1(\beta_D\mathbb{X}_c^{\mathrm{T}} - \mathbb{Y}_{c,D}^{\mathrm{T}})\mathbb{X}_c - M_2\mathbb{Y}_{c,S}^{\mathrm{T}}\mathbb{X}_c,\\
\frac{\partial l}{\partial \Sigma} &= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})^{\mathrm{T}}(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}}, \mathbb{Y}_{c,S})\Sigma^{-1}.
\end{aligned}
$$

Set the derivatives to 0 and we get $\widehat{\beta}_D = \mathbb{Y}_{c,D}^{\mathrm{T}}\mathbb{X}_c(\mathbb{X}_c^{\mathrm{T}}\mathbb{X}_c)^{-1} - M_1^{-1}M_2\mathbb{Y}_{c,S}^{\mathrm{T}}\mathbb{X}_c(\mathbb{X}_c^{\mathrm{T}}\mathbb{X}_c)^{-1} = \widehat{\beta}_{D,\mathrm{ols}} - \widehat{\Sigma}_{DS}\widehat{\Sigma}_S^{-1}\widehat{\beta}_{S,\mathrm{ols}}$ and $\widehat{\Sigma} = \frac{1}{n}(\mathbb{Y}_{c,D} - \mathbb{X}_c\widehat{\beta}_D^{\mathrm{T}}, \mathbb{Y}_{c,S})^{\mathrm{T}}(\mathbb{Y}_{c,D} - \mathbb{X}_c\widehat{\beta}_D^{\mathrm{T}}, \mathbb{Y}_{c,S})$. Since $\beta_S = 0$, $\widehat{\Sigma}_S = S_S^{-1}$, where S_S^{-1} is the sample covariance matrix of Y_S. We can build an equation with $\widehat{\Sigma}_{DS}$. Notice that

$$\begin{aligned}\widehat{\Sigma}_{DS} &= \frac{1}{n}(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_D^{\mathrm{T}})^{\mathrm{T}}\mathbb{Y}_{c,S} \\ &= \frac{1}{n}(\mathbb{Y}_{c,D} - \mathbb{X}_c\beta_{D,\mathrm{ols}}^{\mathrm{T}} + \mathbb{X}_c\widehat{\Sigma}_{DS}\widehat{\Sigma}_S^{-1}\widehat{\beta}_{S,\mathrm{ols}})^{\mathrm{T}}\mathbb{Y}_{c,S} \\ &= \frac{1}{n}(Q_{\mathbb{X}_c}\mathbb{Y}_{c,D} + P_X\mathbb{Y}_{c,S}S_c^{-1}\widehat{\Sigma}_{DS}^{\mathrm{T}})^{\mathrm{T}}\mathbb{Y}_{c,S}.\end{aligned}$$

Solve for $\widehat{\Sigma}_{DS}$, we get

$$\widehat{\Sigma}_{DS} = \mathbb{Y}_{c,D}^{\mathrm{T}}Q_{\mathbb{X}_c}\mathbb{Y}_{c,S}(\mathbb{Y}_{c,S}^{\mathrm{T}}Q_{\mathbb{X}_c}\mathbb{Y}_{S,c})^{-1}S_S^{-1}.$$

Substitute it into $\widehat{\beta}_D$, we get $\widehat{\beta}_D = \widehat{\beta}_{D,\mathrm{ols}} - \widehat{\beta}_{D|S}\widehat{\beta}_{S,\mathrm{ols}}$, where $\widehat{\beta}_{D|S} = \mathbb{Y}_{c,D}^{\mathrm{T}}Q_{\mathbb{X}_c}\mathbb{Y}_{c,S}(\mathbb{Y}_{c,S}^{\mathrm{T}}Q_{\mathbb{X}_c}\mathbb{Y}_{c,S})^{-1}$ contains the coefficients from the regression of R_D on R_S.

To compute the asymptotic variance of the maximum likelihood estimators, we compute the Fisher information matrix for $\{\mathrm{vec}(\beta_D)^{\mathrm{T}}, \mathrm{vech}(\Sigma)^{\mathrm{T}}\}$, where vech is the operator that stacks the lower triangle of a symmetric matrix into a vector column-wise. For an $a \times a$ symmetric matrix M, let C_a and E_a be the contraction matrix and expansion matrix that connect the vec operator and vech operator: $\mathrm{vech}(M) = C_a\,\mathrm{vec}(M)$ and $\mathrm{vec}(M) = E_a\,\mathrm{vech}(M)$. After some straightforward algebra, the Fisher information matrix J is

$$\begin{pmatrix} \Sigma_X \otimes (\Sigma_D - \Sigma_{DS}\Sigma_S^{-1}\Sigma_{DS}^{\mathrm{T}})^{-1} & 0 \\ 0 & \frac{1}{2}E_r^{\mathrm{T}}(\Sigma^{-1}\otimes\Sigma^{-1})E_r^{\mathrm{T}} \end{pmatrix}.$$

The inverse of the upper left block of J relates to the asymptotic variance of $\mathrm{vec}(\widehat{\beta}_D)$. Therefore

$$n^{1/2}\{\mathrm{vec}(\widehat{\beta}_{D,1}) - \mathrm{vec}(\beta_D)\} \to N(0, \Sigma_X^{-1}\otimes\Sigma_{D|S})$$

in distribution as $n \to \infty$. □

Proof of Proposition 4 *and Proposition* 5. The proof of Proposition 5 follows from the standard theory of the envelope model in Cook et al. (2010).

We now prove Proposition 4. The derivation of the maximum likelihood estimator of $\beta_{\mathcal{A}}$ is similar to the derivation of the maximum likelihood estimator of β under the envelope model in Cook et al. (2010).

To derive the asymptotic variance, we apply Proposition 4·1 in Shapiro (1986), as there is overparameterization in the oracle envelope model. First we check the assumptions in Proposition 4·1. We will match our notations with Shapiro's. Shapiro's x is our $\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A},1})^{\mathrm{T}}, \mathrm{vech}(\widehat{\Sigma}_1)^{\mathrm{T}}\}^{\mathrm{T}}$, where $\widehat{\Sigma}_1$ is the estimator under the oracle model (12). Using techniques similar to those in the proof of Theorem 2 in Su & Cook (2012), we can verify that when the errors have finite fourth moments, x is asymptotically normally distributed. Shapiro's ξ is our $\{\mathrm{vec}(\beta_{\mathcal{A}})^{\mathrm{T}}, \mathrm{vech}(\Sigma)^{\mathrm{T}}\}^{\mathrm{T}}$. Let l be the log-likelihood function in (A3) and let $l_{\max}$ be its maximum value. We define the minimum discrepancy function as $f_{\mathrm{MDF}} = l_{\max} - l$. Since f_{MDF} is derived from the normal likelihood function, it satisfies the four conditions in Section 3 of Shapiro (1986). Our $\{\mathrm{vec}(\eta)^{\mathrm{T}}, \mathrm{vec}(\Gamma_{\mathcal{A}})^{\mathrm{T}}, \mathrm{vech}(\Omega)^{\mathrm{T}}, \mathrm{vech}(\Omega_0)^{\mathrm{T}}\}^{\mathrm{T}}$ is Shapiro's θ. Therefore the function g that connects ξ and θ: $\xi = g(\theta)$ is twice differentiable. All the assumptions in Proposition 4·1 are satisfied. Let $\widehat{\Sigma}_O$ be the estimator of Σ under the oracle envelope model (14), then $n^{1/2}[\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A},O})^{\mathrm{T}}, \mathrm{vech}(\widehat{\Sigma}_O)^{\mathrm{T}}\}^{\mathrm{T}} - \{\mathrm{vec}(\beta_{\mathcal{A}})^{\mathrm{T}}, \mathrm{vech}(\Sigma)^{\mathrm{T}}\}^{\mathrm{T}}]$ is asymptotically normally distributed with zero mean and some covariance matrix. So far in this proof, we did not use the normality of the errors, but just require that the errors have finite fourth moments.

Using the normality of the errors gives us closed-form expressions for the asymptotic variance of $\mathrm{vec}(\widehat{\beta}_{\mathcal{A},O})$. Proposition 4·1 indicates that the asymptotic variance has the form $H(H^{\mathrm{T}}JH)^{\dagger}H^{\mathrm{T}}$, where $\dagger$ denotes Moore–Penrose inverse, J is the Fisher information displayed at the end of the proof for Proposition 2, and H is the Jacobian matrix $\partial\xi/\partial^{\mathrm{T}}\theta$

$$H = \begin{pmatrix} I_p \otimes \Gamma_{\mathcal{A}} & \eta^{\mathrm{T}} \otimes I_q & 0 & 0 \\ 0 & 2C_r(I_r \otimes \Gamma\Omega - \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}} \otimes \Gamma)L & C_r(\Gamma \otimes \Gamma)E_u & C_r(\Gamma_0 \otimes \Gamma_0)E_{r-u} \end{pmatrix},$$

where $L = (K_{qu}^{\mathrm{T}}, 0)^{\mathrm{T}} \in \mathbb{R}^{ru \times qu}$, and $K_{qu} \in \mathbb{R}^{qu \times qu}$ is a commutation matrix (Magnus & Neudecker, 1979). After some algebra similar to that in S4 in the supplementary materials of Cook et al. (2010), we can get the closed-form for the asymptotic variance of $\mathrm{vec}(\widehat{\beta}_{\mathcal{A},O})$:

$$n^{1/2}\{\mathrm{vec}(\widehat{\beta}_{\mathcal{A},O}) - \mathrm{vec}(\beta_{\mathcal{A}})\} \to N(0, V_O)$$

in distribution, where $V_O = \Sigma_X^{-1} \otimes \Gamma_{\mathcal{A}}\Omega\Gamma_{\mathcal{A}}^{\mathrm{T}} + (\eta^{\mathrm{T}} \otimes \Gamma_{\mathcal{A},0})T(\eta \otimes \Gamma_{\mathcal{A},0}^{\mathrm{T}})$, and $T = \eta\Sigma_X\eta^{\mathrm{T}} \otimes \widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega \otimes \widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} + \Omega^{-1} \otimes \widetilde{\Omega}_{0,\mathcal{A}} - 2I_u \otimes I_{q-u}$.

Note: We ignored μ_X, α and Σ_X in J and H matrices. This does not affect the results because they are not involved in the parameterization of β and Σ, and their maximum likelihood estimates are asymptotically independent of the estimates of β and Σ. □

Proof of Theorem 4. Let $\widehat{A}_{\mathcal{A}}$ denote the nonzero rows in the sparse envelope estimator $\widehat{A}$, and $\widehat{A}_O$ denote the nonzero rows in the oracle envelope estimator. As $P_{\Gamma} = G_A(G_A^{\mathrm{T}}G_A)^{-1}G_A^{\mathrm{T}}$, for a sequence $a_n = o(n^{-1/2})$, if $\widehat{A}_{\mathcal{A}} = \widehat{A}_O + O_p(a_n)$, then $P_{\widehat{\Gamma}} = P_{\widehat{\Gamma}_O} + O_p(a_n)$. Therefore $\widehat{\beta} - \widehat{\beta}_O = (P_{\widehat{\Gamma}} - P_{\widehat{\Gamma}_O})\widehat{\beta}_{\mathrm{ols}} = (P_{\widehat{\Gamma}} - P_{\widehat{\Gamma}_O})(\widehat{\beta}_{\mathrm{ols}} - \beta) + (P_{\widehat{\Gamma}} - P_{\widehat{\Gamma}_O})\beta = O_p(a_n)o_p(1) + O_p(a_n) = O_p(a_n)$. So $n^{1/2}(\widehat{\beta} - \widehat{\beta}_O) \to 0$ in probability. By Slutsky's theorem $n^{1/2}(\widehat{\beta} - \beta)$ has the same asymptotic distribution as $n^{1/2}(\widehat{\beta}_O - \beta)$. From the proof of Proposition 4, we know that $n^{1/2}(\widehat{\beta}_O - \beta)$ is asymptotically normally distributed with zero mean if the errors have finite fourth moment, and we can obtain the closed-form of the asymptotic variance if normality is assumed. Therefore the conclusion of Theorem 4 follows if we can prove $\widehat{A}_{\mathcal{A}} = \widehat{A}_O + O_p(a_n)$ for $a_n = o(n^{-1/2})$. Since $n^{1/2}\lambda_{\max,n} \to 0$, $\lambda_{\max,n} = o(n^{-1/2})$. For simplicity, we just take $a_n = (n^{-1/2}\lambda_{\max,n})^{1/2}$.

Let B be a $(q-u) \times u$ matrix, and

$$G_B = \begin{pmatrix} I_u \\ B \end{pmatrix} \in \mathbb{R}^{q \times u}.$$

Define

$$f_{\mathrm{obj},\mathcal{A}}(B) = -2\log|G_B^{\mathrm{T}}G_B| + \log|G_B^{\mathrm{T}}\widehat{\Sigma}_{Y_{\mathcal{A}}|X}G_B| + \log|G_B^{\mathrm{T}}(\widehat{\Sigma}_Y^{-1})_{\mathcal{A}}G_B| + \sum_{i=1}^{q-u}\lambda_i\|b_i\|_2,$$

where b_i is the ith row of B. Because of the selection consistency of the sparse envelope model, $\widehat{A}_{\mathcal{A}} = \arg\min_{B \in \mathbb{R}^{(q-u)\times u}} f_{\mathrm{obj},\mathcal{A}}(B)$. Then it is enough to show that for arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_n \mathrm{pr}\left\{\inf_{\Delta \in \mathbb{R}^{(q-u)\times u}, \|\Delta\|_F = C} f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O + a_n\Delta) > f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O)\right\} > 1 - \epsilon. \tag{A3}$$

If (A3) holds, $\widehat{A}_{\mathcal{A}} = \widehat{A}_O + O_p(a_n)$ for $a_n = o(n^{-1/2})$. Now we show (A3). Similar to the proof of Theorem 2, we expand $f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O + a_n\Delta)$ and compute $f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O + a_n\Delta) - f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O)$. We divide $f_{\mathrm{obj},\mathcal{A}}(B)$ into four parts according to the three additions: $f_{\mathrm{obj},\mathcal{A}}(B) \equiv f_{1,\mathcal{A}}(B) + f_{2,\mathcal{A}}(B) + f_{3,\mathcal{A}}(B) + f_{4,\mathcal{A}}(B)$. The first directional derivatives of $f_{1,\mathcal{A}}(B)$, $f_{2,\mathcal{A}}(B)$ and $f_{3,\mathcal{A}}(B)$ at $\widehat{A}_O$ are

$$\overrightarrow{d}^{\Delta} f_{1,\mathcal{A}}(\widehat{A}_O) = \mathrm{tr}\left\{\frac{d}{dB}f_{1,\mathcal{A}}(B)^{\mathrm{T}}\Big|_{B=\widehat{A}_O}\Delta\right\}, \qquad \overrightarrow{d}^{\Delta} f_{2,\mathcal{A}}(\widehat{A}_O) = \mathrm{tr}\left\{\frac{d}{dB}f_{2,\mathcal{A}}(B)^{\mathrm{T}}\Big|_{B=\widehat{A}_O}\Delta\right\},$$

$$\overrightarrow{d}^{\Delta} f_{3,\mathcal{A}}(\widehat{A}_O) = \mathrm{tr}\left\{\frac{d}{dB}f_{3,\mathcal{A}}(B)^{\mathrm{T}}\Big|_{B=\widehat{A}_O}\Delta\right\}.$$

Since $\widehat{A}_O$ is a minimizer of $f_{1,\mathcal{A}}(B) + f_{2,\mathcal{A}}(B) + f_{3,\mathcal{A}}(B)$,

$$\frac{d}{dB} f_{1,\mathcal{A}}(B)\Big|_{B=\widehat{A}_O} + \frac{d}{dB} f_{3,\mathcal{A}}(B)\Big|_{B=\widehat{A}_O} + \frac{d}{dB} f_{3,\mathcal{A}}(B)\Big|_{B=\widehat{A}_O} = 0.$$

Then $\overset{\to\Delta}{df}_{1,\mathcal{A}}(\widehat{A}_O) + \overset{\to\Delta}{df}_{2,\mathcal{A}}(\widehat{A}_O) + \overset{\to\Delta}{df}_{3,\mathcal{A}}(\widehat{A}_O) = 0$.

The calculations on the second directional derivatives of $f_{1,\mathcal{A}}(B)$, $f_{2,\mathcal{A}}(B)$ and $f_{3,\mathcal{A}}(B)$ at $\widehat{A}_O$ and the expansion of $f_{4,\mathcal{A}}(B)$ are parallel to those in Theorem 2. Assembling all those terms together, we have

$$\begin{aligned}
& f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O + a_n\Delta) - f_{\mathrm{obj},\mathcal{A}}(\widehat{A}_O) \\
\geq\ & a_n^2 \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\widetilde{\Omega}_{0,\mathcal{A}}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1 \\
& -2(I_u + A_{\mathcal{A}}^T A_{\mathcal{A}})^{-1}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Big\} - \frac{1}{2}a_n(q-u)\lambda_{\max,n}\max_i\left(\|a_i\|_2^{-1}\|\delta_i\|_2\right) + o_p(a_n^2),
\end{aligned}$$

where $A_{\mathcal{A}} \in \mathbb{R}^{(q-u)\times u}$ contains the nonzero rows in A and $\Delta_{*\mathcal{A}} = (0_{u\times u}, \Delta^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q\times u}$. Based on the definition of a_n, we have $\lambda_{\max,n} = o_p(a_n)$. So the second term is dominated by the first term. Then (A3) is established if we can show that the trace in the first term is positive. We have

$$\begin{aligned}
& \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\widetilde{\Omega}_{0,\mathcal{A}}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1 \\
& -2(I_u + A_{\mathcal{A}}^T A_{\mathcal{A}})^{-1}\Delta_{*\mathcal{A}}^{\mathrm{T}}\Gamma_{\mathcal{A},0}\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Big\} \\
=\ & \operatorname{vec}(\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1)^{\mathrm{T}}\big\{\Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\otimes\widetilde{\Omega}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2I_u\otimes I_{q-u}\big\}\operatorname{vec}(\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1) \\
\geq\ & \operatorname{vec}(\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1)^{\mathrm{T}}\big\{\Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\otimes\widetilde{\Omega}_{0,\mathcal{A}}^{-1} - 2I_u\otimes I_{q-u}\big\}\operatorname{vec}(\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1) \\
\geq\ & m\|\Gamma_{\mathcal{A},0}^{\mathrm{T}}\Delta_{*\mathcal{A}}\Gamma_1^{\mathrm{T}}\|_F \\
\geq\ & m m_0^2\|\Delta\|_F^2,
\end{aligned}$$

where m_0 is the smallest eigenvalue of $(I_u + A_{\mathcal{A}}^{\mathrm{T}}A_{\mathcal{A}})^{-1}$, and m is the smallest eigenvalue of $\Omega^{-1}\otimes\widetilde{\Omega}_{0,\mathcal{A}} + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\otimes\widetilde{\Omega}_{0,\mathcal{A}}^{-1} - 2I_u\otimes I_{q-u}$, which is a positive definite matrix by Shapiro (1986). The derivation of the last inequality is the same as the derivation of a similar inequality at the end of the proof of Theorem 2. □

Proof of Theorem 5. First, we show that

$$\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F = O_p[\{(r_n + s_1)\log r_n/n\}^{1/2}], \tag{A4}$$

$$\|\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1} - \Sigma_Y^{-1}\|_F = O_p[\{(r_n + s_2)\log r_n/n\}^{1/2}]. \tag{A5}$$

Because that Y is sub-gaussian plus a constant and the residuals are not independent, Y and the residuals do not satisfy the assumptions required for establishing the consistency of the sparse permutation invariant covariance estimator. However the sparse permutation invariant covariance estimator depends on the data only through a bound of the sample covariance matrix. Therefore as long as we can show that

$$\max_{i,j}|\widehat{\Sigma}_{Y,ij} - \Sigma_{Y,ij}| \leq C_Y\{\log(r_n)/n\}^{1/2}, \quad \max_{i,j}|\widehat{\Sigma}_{\mathrm{res},ij} - \Sigma_{ij}| \leq C_{\mathrm{res}}\{\log(r_n)/n\}^{1/2} \tag{A6}$$

for some $C_Y > 0$, $C_{\mathrm{res}} > 0$, (A4) and (A5) hold.

We begin by showing (A6). Let W be an m-dimensional random vector with mean μ_W and covariance matrix Σ_W, and $W - \mu_W$ follow a sub-gaussian distribution. Suppose $W_1, \ldots, W_n$ are n independent and identically distributed samples of W, then $\bar{W} = \sum_{i=1}^n W_i$ and

$$\widehat{\Sigma}_W = \frac{1}{n}\sum_{k=1}^n (W_k - \bar{W})(W_k - \bar{W})^{\mathrm{T}} = \frac{1}{n}\sum_{k=1}^n (W_k - \mu_W)(W_k - \mu_W)^{\mathrm{T}} - (\bar{W} - \mu_W)(\bar{W} - \mu_W)^{\mathrm{T}}.$$

From Ravikumar et al. (2011), there exists positive constants C_i, such that for $\delta \in (0, b_1)$,

$$\begin{aligned}
\mathrm{pr}(|\widehat{\Sigma}_{W,ij} - \Sigma_{W,ij}| > \delta) \leq\ & \mathrm{pr}\left[\left|\left\{\frac{1}{n}\sum_{k=1}^{n}(W_k - \mu_W)(W_k - \mu_W)^{\mathrm{T}}\right\}_{ij} - \Sigma_{W,ij}\right| > \frac{\delta}{2}\right] \\
& + \mathrm{pr}\left[\left|\left\{(\bar{W} - \mu_W)(\bar{W} - \mu_W)^{\mathrm{T}}\right\}_{ij}\right| > \frac{\delta}{2}\right] \\
\leq\ & C_1 \exp(-C_2 n\delta^2) + C_3 \exp(-C_4 n \delta^2)
\end{aligned}$$

where $|\cdot|$ denotes absolute value. Let $\delta = C_5\{\log(m)/n\}^{1/2}$ for some $C_5 > 0$. Using the union sum inequality, as $n \to \infty$, we have with probability tending to 1,

$$\max_{i,j} |\widehat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_6\{\log(m)/n\}^{1/2},$$

where C_6 is a positive number.

Now we take $W = (X^{\mathrm{T}}, \boldsymbol{\varepsilon}^{\mathrm{T}})^{\mathrm{T}}$, then W is a $p + r_n$ dimensional random vector with mean $(\mu_X^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$, where the 0 is an r_n dimensional vector. It has a block diagonal covariance matrix with diagonal blocks being Σ_X and Σ. Then by the preceding conclusion, we can find constant C_0 such that $\max_{i,j} |\widehat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_0\{\log(r_n + p)/n\}^{1/2}$. Since p is fixed, we can find C_0^* such that $\max_{i,j} |\widehat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_0^*\{\log(r_n)/n\}^{1/2}$. Then

$$\begin{aligned}
\max_{i,j} |\widehat{\Sigma}_{X,ij} - \Sigma_{X,ij}| &\leq C_0^*\{\log(r_n)/n\}^{1/2}, \\
\max_{i,j} |\widehat{\Sigma}_{\varepsilon,ij} - \Sigma_{ij}| &\leq C_0^*\{\log(r_n)/n\}^{1/2}, \\
\max_{i,j} |\widehat{\Sigma}_{\varepsilon X,ij}| &\leq C_0^*\{\log(r_n)/n\}^{1/2}.
\end{aligned}$$

Since $\widehat{\Sigma}_Y = \beta\widehat{\Sigma}_X\beta^{\mathrm{T}} + \widehat{\Sigma}_\varepsilon + \beta\widehat{\Sigma}_{X\varepsilon} + \widehat{\Sigma}_{\varepsilon X}\beta^{\mathrm{T}}$, we have

$$\max_{i,j} |\widehat{\Sigma}_{Y,ij} - \Sigma_{Y,ij}| \leq C_Y\{\log(r_n)/n\}^{1/2},$$

for some $C_Y > 0$.

As $\widehat{\Sigma}_{\mathrm{res}} = \widehat{\Sigma}_\varepsilon - \widehat{\Sigma}_{\varepsilon X}\widehat{\Sigma}_X^{-1}\widehat{\Sigma}_{X\varepsilon}$,

$$\begin{aligned}
\widehat{\Sigma}_{\mathrm{res}} - \Sigma = & (\widehat{\Sigma}_{\varepsilon X} - \Sigma_{\varepsilon X})\Sigma_X^{-1}\Sigma_{X\varepsilon} + \Sigma_{\varepsilon X}(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1})\Sigma_{X\varepsilon} + \Sigma_{\varepsilon X}\Sigma_X^{-1}(\widehat{\Sigma}_{X\varepsilon} - \Sigma_{X\varepsilon}) \\
& + (\widehat{\Sigma}_{\varepsilon X} - \Sigma_{\varepsilon X})(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1})\Sigma_{X\varepsilon} + \Sigma_{\varepsilon X}(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1})(\widehat{\Sigma}_{X\varepsilon} - \Sigma_{X\varepsilon}) \\
& + (\widehat{\Sigma}_{\varepsilon X} - \Sigma_{\varepsilon X})\Sigma_X^{-1}(\widehat{\Sigma}_{X\varepsilon} - \Sigma_{X\varepsilon}) + (\widehat{\Sigma}_{\varepsilon X} - \Sigma_{\varepsilon X})(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1})(\widehat{\Sigma}_{X\varepsilon} - \Sigma_{X\varepsilon}) \\
& + \widehat{\Sigma}_\varepsilon - \Sigma.
\end{aligned}$$

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

$$\max_{i,j} |\widehat{\Sigma}_{\mathrm{res},ij} - \Sigma_{ij}| \leq C_{\mathrm{res}}\{\log(r_n)/n\}^{1/2},$$

for some $C_{\mathrm{res}} > 0$. Therefore (A4) and (A5) hold.

We denote the objective function in (7) as $f_{\mathrm{obj},2}$. Let $a_n = \{(r_n + s)\log r_n/n\}^{1/2}$. Theorem 5 holds if for arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_n \mathrm{pr}\left\{\inf_{\Delta \in \mathbb{R}^{(q-u)\times u}, \|\Delta\|_F = C} f_{\mathrm{obj},2}(A + a_n\Delta) > f_{\mathrm{obj},2}(A)\right\} > 1 - \epsilon. \tag{A7}$$

Following the techniques and notations in the proof of Theorem 2, we expand $f_{\mathrm{obj},2}(A + a_n\Delta) - f_{\mathrm{obj},2}(A)$ and get

$$\begin{aligned}
&f_{\mathrm{obj},2}(A + a_n\Delta) - f_{\mathrm{obj},2}(A)\\
\geq\ & 2a_n \operatorname{tr}\big[(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma)\Delta_* + \{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}\}G_A^{\mathrm{T}}\Sigma\Delta_*\\
&+\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}\}G_A^{\mathrm{T}}(\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma)\Delta_* + (G_A^{\mathrm{T}}\Sigma_{Y,\mathrm{sp}}^{-1}G_A)^{-1}G_A^{\mathrm{T}}(\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1} - \Sigma_Y^{-1})\Delta_*\\
&+\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma_Y^{-1}G_A)^{-1}\}G_A^{\mathrm{T}}\Sigma_Y^{-1}\Delta_*\\
&+\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma_Y^{-1}G_A)^{-1}\}G_A^{\mathrm{T}}(\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1} - \Sigma_Y^{-1})\Delta_*\big]\\
&+a_n^2 \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\\
&-2(I_u + A^TA)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\} - \frac{1}{2}a_n(q-u)\lambda_{\max,n}\max_i\big(\|a_i\|_2^{-1}\|\delta_i\|_2\big) + o_p(a_n^2).
\end{aligned}$$

Notice that

$$\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma = -\Sigma(\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1})\Sigma + o_p(\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}).$$

Let $\|\cdot\|$ be the spectral norm of a matrix. For two matrices $A \in \mathbb{R}^{d_1\times d_2}$ and $B \in \mathbb{R}^{d_2\times d_3}$, $\|AB\|_F \leq \|A\|\|B\|_F$. So

$$\|\Sigma(\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1})\Sigma\|_F \leq \|\Sigma\|^2\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F \leq \bar{k}^2\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F,$$

and $\|\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma\|_F = O_p[\{(r_n + s)\log r_n/n\}^{1/2}]$. Then

$$\operatorname{tr}\big[(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma)\Delta_*\big] \geq -\bar{k}^2\|\Delta\|_F\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F\|(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\|\|G_A\|_F.$$

Now

$$\begin{aligned}
&(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}\\
=\ & -(G_A^{\mathrm{T}}\Sigma G_A)^{-1}(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A - G_A^{\mathrm{T}}\Sigma G_A)(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + o_p(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A - G_A^{\mathrm{T}}\Sigma G_A)\\
=\ & -(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}(\widehat{\Sigma}_{\mathrm{res,sp}} - \Sigma)G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + o_p[\{(r_n + s_1)\log r_n/n\}^{1/2}]\\
=\ & -(G_A^{\mathrm{T}}\Sigma G_A)^{-1}G_A^{\mathrm{T}}\Sigma(\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1})\Sigma G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + o_p[\{(r_n + s_1)\log r_n/n\}^{1/2}]
\end{aligned}$$

so

$$\begin{aligned}
&\operatorname{tr}\big[\{(G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}G_A)^{-1} - (G_A^{\mathrm{T}}\Sigma G_A)^{-1}\}G_A^{\mathrm{T}}\Sigma\Delta_*\big]\\
\geq\ & -u^{1/2}\bar{k}^2\|\Delta\|_F\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F\|(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\|\|G_A\|_F.
\end{aligned}$$

Apply these inequalities to the terms in the first four lines in $f_{\mathrm{obj},2}(A + a_n\Delta) - f_{\mathrm{obj},2}(A)$, then

$$\begin{aligned}
&f_{\mathrm{obj},2}(A + a_n\Delta) - f_{\mathrm{obj},2}(A)\\
\geq\ & 2M_1a_n\|\Delta\|_F\|\widehat{\Sigma}_{\mathrm{res,sp}}^{-1} - \Sigma^{-1}\|_F + 2M_2a_n\|\Delta\|_F\|\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1} - \Sigma_Y^{-1}\|_F\\
&+a_n^2 \operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1\\
&-2(I_u + A^TA)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\} - \frac{1}{2}a_n(q-u)\lambda_{\max,n}\max_{i=1,\ldots,q-u}\big(\|a_i\|_2^{-1}\|\delta_i\|_2\big) + o_p(a_n^2),
\end{aligned}$$

where $M_1 = -2u^{1/2}\bar{k}^2\|(G_A^{\mathrm{T}}\Sigma G_A)^{-1}\|\|G_A\|_F$ and $M_2 = -2\|(G_A^{\mathrm{T}}\Sigma_Y^{-1}G_A)^{-1}\|\|G_A\|_F$. Because $\lambda_{\max,n} = o[\{(r_n + s)\log r_n/n\}^{1/2}] = o_p(a_n)$ and

$$\begin{aligned}
&\operatorname{tr}\Big\{\Omega^{-1}\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 + (\Omega + \eta\Sigma_X\eta^{\mathrm{T}})\Gamma_1^{\mathrm{T}}\Delta_*^{\mathrm{T}}\Gamma_0\Omega_0^{-1}\Gamma_0^{\mathrm{T}}\Delta_*\Gamma_1 - 2(I_u + A^TA)^{-1}\Delta_*^{\mathrm{T}}\Gamma_0\Gamma_0^{\mathrm{T}}\Delta_*\Big\}\\
\geq\ & m\|\Delta\|_F^2,
\end{aligned}$$

for some $m > 0$ by Theorem 2, the second order term of $\|\Delta\|_F$ dominates the first order term of $\|\Delta\|_F$ in $f_{\mathrm{obj},2}(A + a_n\Delta) - f_{\mathrm{obj},2}(A)$. Therefore (A7) holds, and $\|\widehat{A} - A\|_F = O_p[\{(r_n + s)\log r_n/n\}^{1/2}]$. As $P_\Gamma = G_A(I_u + A^{\mathrm{T}}A)^{-1}G_A^{\mathrm{T}}$ is a simple and continuous function of A, then $\|P_{\widehat{\Gamma}} - P_\Gamma\|_F = O_p[\{(r_n + s)\log r_n/n\}^{1/2}]$.

Since

$$\widehat{\beta}_{\mathrm{ols}} - \beta = \widehat{\Sigma}_{\varepsilon X}\widehat{\Sigma}_X^{-1} - \Sigma_{\varepsilon X}\Sigma_X^{-1} = (\widehat{\Sigma}_{\varepsilon X} - \Sigma_{\varepsilon X})\Sigma_X^{-1} + \Sigma_{\varepsilon X}(\widehat{\Sigma}_X^{-1} - \Sigma_X^{-1}) + o_p[\{(r_n + s)\log r_n/n\}^{1/2}],$$

there exists a constant C_{ols} such that

$$\max_{i,j} |\widehat{\beta}_{\mathrm{ols},ij} - \beta_{ij}| \le C_{\mathrm{ols}}\{\log(r_n)/n\}^{1/2}.$$

Because $\|\widehat{\beta}_{\mathrm{ols}} - \beta\|_F \le (pr_n)^{1/2}\|\widehat{\beta}_{\mathrm{ols}} - \beta\|_{\max}$, then

$$\|\widehat{\beta} - \beta\|_F \le \|(P_{\widehat{\Gamma}} - P_\Gamma)\widehat{\beta}_{\mathrm{ols}}\|_F + \|P_\Gamma(\widehat{\beta}_{\mathrm{ols}} - \beta)\|_F \le \|(P_{\widehat{\Gamma}} - P_\Gamma)\widehat{\beta}_{\mathrm{ols}}\|_F + \|\widehat{\beta}_{\mathrm{ols}} - \beta\|_F.$$

Therefore the sparse envelope estimator $\widehat{\beta}$ converges to β with rate $\{(r_n + s)\log r_n/n\}^{1/2}$. □

Proof of Theorem 6. Let

$$\delta = \min_{i=1,\ldots,q-u} \|a_i\|_2 > 0,$$

then δ is the smallest norm of the non-sparse rows in A. Since $\|\widehat{\beta} - \beta\|_F = O_p[\{(r_n + s)\log r_n/n\}^{1/2}]$, and $\{(r_n + s)\log r_n/n\}^{1/2} \to 0$, then $\|\widehat{\beta} - \beta\|_F < \delta/2$ with probability tending to 1. This implies $\|\hat{a}_i - a_i\|_2 < \delta/2$ for $i = 1, \ldots, r_n$. For $i = 1, \ldots, q$, $\|\hat{a}_i\|_2 > \|a_i\|_2 - \delta/2 > 0$. Therefore the sparse envelope estimator identifies the nonzero rows with probability tending to 1.

For a_i, $i = q - u + 1, \ldots, r_n - u$, suppose $\hat{a}_i \neq 0$, taking the derivative of $f_{\mathrm{obj},2}$ with respect to a_i and evaluate at $\hat{a}_i$, we have

$$-4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}\widehat{G}_A)^{-1}$$
$$+\lambda_i \frac{\widehat{a}_i^{\mathrm{T}}}{\|\widehat{a}_i\|_2} = 0.$$

Because $-4e_i^{\mathrm{T}}G_A(I_u + A^{\mathrm{T}}A)^{-1} + 2e_i^{\mathrm{T}}\Sigma G_A(G_A^{\mathrm{T}}\Sigma G_A)^{-1} + 2e_i^{\mathrm{T}}\Sigma_Y^{-1}G_A(G_A^{\mathrm{T}}\Sigma_Y^{-1}G_A)^{-1} = 0$, we have

$$\| - 4e_i^{\mathrm{T}}\widehat{G}_A(I_u + \widehat{A}^{\mathrm{T}}\widehat{A})^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res,sp}}\widehat{G}_A)^{-1} + 2e_i^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}\widehat{G}_A(\widehat{G}_A^{\mathrm{T}}\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}\widehat{G}_A)^{-1}\|_F$$
$$= O_p[\{(r_n + s)\log r_n/n\}^{1/2}].$$

But

$$\left\|\lambda_i \frac{\widehat{a}_i^{\mathrm{T}}}{\|\widehat{a}_i\|_2}\right\|_F = \lambda_i \ge \lambda_{\min,n}.$$

Since $\{(r_n + s)\log r_n/n\}^{1/2} = o(\lambda_{\min,n})$, this is a contradiction. Therefore we have $\mathrm{pr}(\hat{a}_i = 0) \to 1$ for $i = q - u + 1, \ldots, r_n - u$. □

B. Convergence Analysis of Algorithm 1

In this section, we prove the strict descent property of our blockwise coordinate descent algorithm. The proof relies on the following two lemmas.

LEMMA B1. *The loss function $L(a_i \mid \widetilde{A}_{-i})$ as defined in* (9) *has a bounded second derivative*

$$\frac{d^2}{da_i^2}L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i} \preceq \{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\}\, I,$$

where $I \in \mathbb{R}^{u\times u}$ and $M_1 \preceq M_2$ means that $M_2 - M_1$ is a semi-positive definite matrix.

LEMMA B2. *One can find a quadratic majorization function Q for the loss function $L(a_i \mid \widetilde{A}_{-i})$* in (9), i.e.,

$$Q(a_i) = L(\widetilde{a}_i \mid \widetilde{A}_{-i}) + (a_i - \widetilde{a}_i)^{\mathrm{T}} \frac{d}{da_i} L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i} + 1/2\delta_i (a_i - \widetilde{a}_i)^{\mathrm{T}}(a_i - \widetilde{a}_i), \tag{A1}$$

such that $Q(a_i) = L(\widetilde{a}_i \mid \widetilde{A}_{-i})$ when $a_i = \widetilde{a}_i$ and $Q(a_i) > L(\widetilde{a}_i \mid \widetilde{A}_{-i})$ when $a_i \neq \widetilde{a}_i$.

Proof of Lemma B1. The second derivative of $L(a_i \mid \widetilde{A}_{-i})$ is

$$\frac{d^2}{da_i^2} L(a_i|\widetilde{A}_{-i}) = -4T_1 + 2T_2 + 2T_3,$$

where

$$T_1 = \frac{(1 + a_i^{\mathrm{T}} B_1 a_i)\, B_1 - 2B_1 a_i a_i^{\mathrm{T}} B_1}{(1 + a_i^{\mathrm{T}} B_1 a_i)^2}, \tag{A2}$$

$$T_2 = \frac{\{1 + (a_i + v_2)^{\mathrm{T}} B_2 (a_i + v_2)\} B_2 - 2B_2 (a_i + v_2)(a_i + v_2)^{\mathrm{T}} B_2}{\{1 + (a_i + v_2)^{\mathrm{T}} B_2 (a_i + v_2)\}^2},$$

$$T_3 = \frac{\{1 + (a_i + v_3)^{\mathrm{T}} B_3 (a_i + v_3)\} B_3 - 2B_3 (a_i + v_3)(a_i + v_3)^{\mathrm{T}} B_3}{\{1 + (a_i + v_3)^{\mathrm{T}} B_3 (a_i + v_3)\}^2}.$$

We only prove that T_1 defined in (A2) can be bounded as $-\gamma_{\max}(B_1) I \preceq T_1 \preceq \gamma_{\max}(B_1) I$, since the proofs for bounding T_2 and T_3 are very similar. We write T_1 as

$$T_1 = \frac{(1 + a_i^{\mathrm{T}} B_1 a_i)\, B_1 - 2B_1 a_i a_i^{\mathrm{T}} B_1}{(1 + a_i^{\mathrm{T}} B_1 a_i)^2} = \frac{B_1^{1/2} \left\{ \left(1 + a_i^{\mathrm{T}} B_1^{1/2} B_1^{1/2} a_i\right) I - 2B_1^{1/2} a_i a_i^{\mathrm{T}} B_1^{1/2} \right\} B_1^{1/2}}{(1 + a_i^{\mathrm{T}} B_1 a_i)^2}. \tag{A3}$$

Replace $x = B_1^{1/2} a_i$ in (A3), we get

$$T_1 = \frac{B_1^{1/2} \left\{ (1 + x^{\mathrm{T}} x) I - 2xx^{\mathrm{T}} \right\} B_1^{1/2}}{(1 + x^{\mathrm{T}} x)^2}.$$

We now prove that

$$-I \preceq \frac{(1 + x^{\mathrm{T}} x) I - 2xx^{\mathrm{T}}}{(1 + x^{\mathrm{T}} x)^2} \preceq I.$$

Denote $z = x/\|x\|$ and denote $M = (z^{\mathrm{T}} z) I - zz^{\mathrm{T}}$. It is easy to see that $0 \preceq M \preceq I$. As

$$(x^{\mathrm{T}} x) I - xx^{\mathrm{T}} = \|x\|^2 \left(\frac{x^{\mathrm{T}}}{\|x\|} \frac{x}{\|x\|} I - \frac{x}{\|x\|} \frac{x^{\mathrm{T}}}{\|x\|} \right) = \|x\|^2 M,$$

we have $(x^{\mathrm{T}} x) I - xx^{\mathrm{T}} \succeq 0$. Then

$$(1 + x^{\mathrm{T}} x) I - 2xx^{\mathrm{T}} \succeq (1 + x^{\mathrm{T}} x) I - 2(x^{\mathrm{T}} x) I = (1 - x^{\mathrm{T}} x) I \succeq -(1 + x^{\mathrm{T}} x) I. \tag{A4}$$

We also have

$$(1 + x^{\mathrm{T}} x) I - 2xx^{\mathrm{T}} \preceq (1 + x^{\mathrm{T}} x) I. \tag{A5}$$

Therefore combining (A4), (A5) and $1 + x^{\mathrm{T}} x \geq 1$, we have

$$-I \preceq \frac{(1 + x^{\mathrm{T}} x) I - 2xx^{\mathrm{T}}}{(1 + x^{\mathrm{T}} x)^2} \preceq I.$$

Therefore

$$-\gamma_{\max}(B_1) I \preceq -B_1 \preceq T_1 \preceq B_1 \preceq \gamma_{\max}(B_1) I.$$

Similarly we can prove that $-\gamma_{\max}(B_2)I \preceq T_2 \preceq \gamma_{\max}(B_2)I$, and $-\gamma_{\max}(B_3)I \preceq T_3 \preceq \gamma_{\max}(B_3)I$. Hence the lemma is proved. □

Proof of Lemma B2. For any a_i and a_i^*, let $d_i = a_i - a_i^*$ and define $g(t) = L(a_i^* + td_i \mid \widetilde{A}_{-i})$ such that

$$g(0) = L(a_i^* \mid \widetilde{A}_{-i}), \quad g(1) = L(a_i \mid \widetilde{A}_{-i}).$$

By Taylor expansion, there exists a $b \in (0,1)$ such that

$$g(1) = g(0) + g'(0) + 1/2g''(b). \tag{A6}$$

By Lemma B1,

$$\begin{aligned} g''(b) &= d_i^{\mathrm{T}} \frac{d^2}{da_i^2} L(a_i \mid \widetilde{A}_{-i})\Big|_{a_i = a_i^* + bd_i} d_i \\ &\le \{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\}\, d_i^{\mathrm{T}} d_i \\ &\le \delta_i d_i^{\mathrm{T}} d_i, \end{aligned} \tag{A7}$$

where $\delta_i = (1+\varepsilon^*)\{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\}$ and $\varepsilon^* > 0$. When $d_i \neq 0$ the inequality in (A7) strictly holds. Plugging (A7) into (A6) gives (A1). □

Proof of Theorem 1. By Lemma B2, after updating $\widetilde{a}_i$ using

$$\widetilde{a}_{i,\text{new}} = \frac{1}{\delta_i}\left\{\delta_i \widetilde{a}_i - \frac{d}{da_i} L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i}\right\}\left\{1 - \frac{\lambda\omega_i}{\left\|\delta_i\widetilde{a}_i - \frac{d}{da_i}L(a_i|\widetilde{A}_{-i})\big|_{a_i=\widetilde{a}_i}\right\|_2}\right\}_+, \tag{A8}$$

we have

$$\begin{aligned} L(\widetilde{a}_{i,\text{new}} \mid \widetilde{A}_{-i}) + \lambda\omega_i\|\widetilde{a}_{i,\text{new}}\|_2 &\le Q(\widetilde{a}_{i,\text{new}}) + \lambda\omega_i\|\widetilde{a}_{i,\text{new}}\|_2 \\ &\le Q(\widetilde{a}_i) + \lambda\omega_i\|\widetilde{a}_i\|_2 \\ &= L(\widetilde{a}_i) + \lambda\omega_i\|\widetilde{a}_i\|_2. \end{aligned}$$

Moreover, if $\widetilde{a}_i(\text{new}) \neq \widetilde{a}_i$, then the first inequality becomes

$$L(\widetilde{a}_{i,\text{new}} \mid \widetilde{A}_{-i}) + \lambda\omega_i\|\widetilde{a}_{i,\text{new}}\|_2 < Q(\widetilde{a}_{i,\text{new}}) + \lambda\omega_i\|\widetilde{a}_{i,\text{new}}\|_2.$$

Therefore, the objective function strictly decreases after updating all blocks in a cycle, unless the solution stays unchanged after each blockwise coordinate update. If this is the case, we can show that the solution must satisfy the Karush–Kuhn–Tucker conditions, which indicates that the algorithm has converged to the stationary point. To see this, if $\widetilde{a}_{i,\text{new}} = \widetilde{a}_i$ for all i, then by (A8) we have

$$\widetilde{a}_i = \frac{1}{\delta_i}\left\{\delta_i \widetilde{a}_i - \frac{d}{da_i} L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i}\right\}\left\{1 - \frac{\lambda\omega_i}{\left\|\delta_i\widetilde{a}_i - \frac{d}{da_i}L(a_i|\widetilde{A}_{-i})\big|_{a_i=\widetilde{a}_i}\right\|_2}\right\}$$

if

$$\left\|\delta_i\widetilde{a}_i - \frac{d}{da_i}L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i}\right\|_2 > \lambda\omega_i,$$

and $\widetilde{a}_i = 0$ otherwise. By straightforward algebra we obtain the Karush–Kuhn–Tucker conditions:

$$\begin{aligned} \frac{d}{da_i}L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i} + \lambda\omega_i \cdot \frac{\widetilde{a}_i}{\|\widetilde{a}_i\|_2} = 0, &\qquad \widetilde{a}_i \neq 0, \\ \left\|\frac{d}{da_i}L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i}\right\|_2 \le \lambda\omega_i, &\qquad \widetilde{a}_i = 0, \end{aligned}$$

where $i = 1, \dots, r-u$. Therefore, if the objective function stays unchanged after a cycle, the solution satisfies the Karush–Kuhn–Tucker conditions and necessarily converges to the stationary point of the problem. □

Now we show a figure that empirically confirms the convergence of Algorithm 1. We used the following settings to generate the figure. We set $p = 5$, $u = 2$, $n = 50$ and $r = 200$. The first $q/2$ rows in $\Gamma_{\mathcal{A}}$ were $\{(2/q)^{1/2}, 0\}^{\mathrm{T}}$ and the remaining $q/2$ rows were $\{0, (2/q)^{1/2}\}^{\mathrm{T}}$. Then we used the structure in (5) to construct Γ and Γ_0. The errors were generated from the multivariate normal distribution with mean 0 and covariance matrix $\Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$, where $\Omega = I_u$ and Ω_0 was a block diagonal matrix with the upper left block being $25I_{q-u}$ and lower right block being $4I_{r-q}$. The elements in η were independent $N(0, 4^2)$ variates. The predictors X were normally distributed with mean 0 and covariance matrix $\Sigma_X = 4I_p$. Figure 1 plotted the log of the objective value in (7) minus the optimal point versus the number of iterations. We added 10^{-3} to avoid taking logarithm of zero at the optimal point. For comparison, we used a subgradient method rather than the majorization-minimization method to get the solution of (9). We included a line for the subgradient method in the figure. The same convergence criterion and starting value were used for Algorithm 1 and the subgradient method. Figure 1 shows that Algorithm 1 takes less iterations to converge. The subgradient method is not a descent method, as the objective value is not monotonically decreasing. On the other hand, the objective value strictly decreases with Algorithm 1, which confirms Theorem 1.

Fig. 1. Comparison of convergence for Algorithm 1 (solid) and the subgradient method (dashed).

C. SIMULATIONS

In this section, we investigate the performance of the sparse envelope estimator under three cases: the first has $u < r < p < n$, the second varies the signal level σ_X, and the third has different values of u, i.e., the dimension of the envelope subspace.

In the first case with $u < r < p < n$, we set $n = 250$, $r = 100$, $u = 2$ and $q = 5$. The matrix $(\Gamma_{\mathcal{A}}, \Gamma_{\mathcal{A},0})$ was obtained by orthogonalizing a q by q matrix of independent uniform $(0, 1)$ variates. Then we used the structure in (5) to construct Γ and Γ_0. The elements in η were taken to be independent normal variates with mean 0 and variance 0·16. The error covariance matrix Σ followed the structure $\Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$, where $\Omega = I_u$ and Ω_0 was a block diagonal matrix with the upper left block being $9I_{q-u}$ and lower right block being $4I_{r-q}$. The predictors X were normally distributed with mean 0 and covariance matrix $\Sigma_X = \sigma_X^2 I_p$, where σ_X^2 =0·4. We varied p from 100 to 180. For each value of

p, 200 replications were generated. The selection performance is summarized in Table 1. The standard deviation of a randomly chosen element in β is displayed in Fig. 2. When $r < p < n$, the sparse envelope model still gives substantial efficient gains compared to the standard model.

Table 1. *Average true positive rate (%), true negative rate (%) and accuracy (%) of sparse envelope estimator, hard thresholding estimator and F test*

	sparse envelope			hard thresholding			F test		
p	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.
100	98·8	99·7	85·0	90·4	99·8	66·0	59·8	99·9	0·0
120	98·6	99·6	81·0	90·4	99·7	64·0	60·6	99·9	1·0
140	98·0	99·3	78·0	86·4	99·0	36·0	58·6	100·0	0·0
160	92·8	98·6	67·0	79·2	98·5	23·0	55·2	100·0	1·0
180	82·2	98·1	49·0	40·4	99·3	2·0	49·2	100·0	0·0

Fig. 2. Comparison of the standard deviations for sparse envelope estimator (solid) and standard estimator (dashed).

In the second simulation, we varied the signal level σ_X and investigated the selection performance and efficiency gains of the sparse envelope estimator. In the simulation that generated Table 1, we fixed $p = 160$ and varied σ_X from 0·05 to 0·6. The selection performance is summarized in Table 2, and the standard deviation of a randomly chosen element in β is displayed in Fig. 3. We notice that the sparse envelope model is more advantageous when the signal is weak. When the signal is stronger, both the sparse envelope estimator and the standard estimator improve. But for all signal levels, the sparse envelope estimator is more efficient than the standard estimator.

In the third case, we set $r = 100$, $q = 24$, $p = 50$, $n = 200$ and varied u from 2 to 20. The matrix $(\Gamma_{\mathcal{A}}, \Gamma_{\mathcal{A},0})$ was obtained by orthogonalizing a $q \times q$ matrix of independent standard normal variates. Then we used the structure in (5) to construct Γ and Γ_0. The elements in η were independent normal variates with mean 0 and variance 0·25, and the error covariance matrix had the structure $\Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$ with $\Omega = I_u$ and $\Omega_0 = 25I_{r-u}$. The predictors X were generated from a multivariate normal distribution with mean 0 and covariance matrix I_p. The selection performance under different u is summarized in Table 3, and the standard deviation of a randomly chosen element in β is displayed in Fig. 4. We notice that when u is small, there is a bigger immaterial part and therefore we expect a more substantial efficiency gain by using the sparse envelope estimator.

Table 2. *Average true positive rate (%), true negative rate (%) and accuracy (%) of sparse envelope estimator, hard thresholding estimator and F test*

	sparse envelope			hard thresholding			F test		
σ_X^2	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.
0·05	54·6	96·1	0·0	25·8	98·7	0·0	2·2	100·0	0·0
0·1	70·6	96·2	10·0	46·4	98·1	5·0	17·6	100·0	0·0
0·2	85·2	97·7	39·0	65·6	98·1	14·0	30·2	100·0	0·0
0·3	87·8	97·9	48·0	72·6	98·1	20·0	44·8	100·0	0·0
0·4	92·8	98·6	67·0	79·4	98·5	23·0	55·2	100·0	1·0
0·5	98·2	99·8	93·0	89·8	99·5	54·0	61·8	100·0	1·0
0·6	100·0	100·0	100·0	98·0	99·9	88·0	65·0	100·0	3·0

Fig. 3. Comparison of the standard deviations for sparse envelope estimator (solid) and standard estimator (dashed).

Table 3. *Average true positive rate (%), true negative rate (%) and accuracy (%) of sparse envelope estimator, hard thresholding estimator and F test*

	sparse envelope			hard thresholding			F test		
u	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.	T.P.R.	T.N.R.	Accu.
2	35·8	99·9	0·0	20·8	100·0	0·0	4·1	100·0	0·0
5	72·6	99·9	0·0	54·7	100·0	0·0	20·5	99·9	0·0
10	95·9	100·0	27·5	88·2	100·0	0·0	65·1	99·7	0·0
15	99·9	100·0	98·8	98·3	100·0	58·8	94·8	99·7	21·2
20	100·0	100·0	100·0	100·0	100·0	100·0	99·9	99·7	77·5

D. THE SMALLEST LAMBDA THAT YIELDS THE NULL MODEL

We define λ^* as the smallest λ value such that all the elements in A are zero. By the Karush–Kuhn–Tucker conditions of the optimization problem (8),

$$\frac{d}{da_i} L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i} + \lambda w_i \cdot \frac{\widetilde{a}_i}{\|\widetilde{a}_i\|_2} = 0, \qquad \widetilde{a}_i \neq 0,$$
$$\left\| \frac{d}{da_i} L(a_i|\widetilde{A}_{-i})\Big|_{a_i=\widetilde{a}_i} \right\|_2 \leq \lambda w_i, \qquad \widetilde{a}_i = 0,$$

Fig. 4. Comparison of the standard deviations for sparse envelope estimator (solid) and standard estimator (dashed).

for $i = 1, \ldots, r - u$. Then we can find that

$$\lambda^* = \max_{i=1,\ldots,r-u} \left\| \frac{d}{da_i} L(a_i|A_{-i} = 0)\Big|_{a_i=0} \right\|_2 /w_i, \quad w_i \neq 0.$$

If M is an $r \times r$ symmetric matrix and U is a set such that $U = \{1, \ldots, u\}$, let $M_{U,U}$ denote the upper left block of M that has dimension $u \times u$, $M_{U,u+i}$ denote the $u \times 1$ vector that includes the first u elements of the $(u + i)$th column, and $M_{u+i|U} = M_{u+i,u+i} - M^{\mathrm{T}}_{U,u+i} M^{-1}_{U,U} M_{U,u+i}$. Then after some straightforward calculations,

$$\frac{d}{da_i} L(a_i|A_{-i} = 0)\Big|_{a_i=0} = 2(\widehat{\Sigma}_{\mathrm{res}})_{u+i,u+i}/(\widehat{\Sigma}_{\mathrm{res}})_{u+i|U} + 2(\widehat{\Sigma}_Y^{-1})_{u+i,u+i}/(\widehat{\Sigma}_Y^{-1})_{u+i|U} - 4.$$

Therefore we have

$$\lambda^* = \max_{i=1,\ldots,r-u} \left\| 2(\widehat{\Sigma}_{\mathrm{res}})_{u+i,u+i}/(\widehat{\Sigma}_{\mathrm{res}})_{u+i|U} + 2(\widehat{\Sigma}_Y^{-1})_{u+i,u+i}/(\widehat{\Sigma}_Y^{-1})_{u+i|U} - 4 \right\|_2 /w_i, \quad w_i \neq 0.$$

E. Comparison of Akaike information criterion, Bayesian information criterion and likelihood ratio testing on selection of u

The simulation settings are the same as those used in Fig. 1. We used the Akaike information criterion, Bayesian information criterion and likelihood ratio testing with significance level $\alpha = 0.01$ to select u. For each sample size, 500 replications were generated. Results are summarised in Fig. 5. The selection performances for all three criteria are quite close, with Bayesian information criterion slightly better for larger sample sizes. This is because as n tends to infinity, Bayesian information criterion selects the true dimension with probability approaching 1 while likelihood ratio testing selects the true dimension at the nominal level $1 - \alpha$. Akaike information criterion tends to select a larger dimension, because asymptotically Akaike information criterion has positive probability in selecting a model that contains the true model. A similar pattern is also observed in Su & Cook (2013) when comparing these three criteria. Since Bayesian information criterion is quite stable with all sample sizes, we use it to select u for the data analysis in Section 3·2.

Fig. 5. Comparison of Akaike information criterion (dashed), Bayesian information criterion (solid) and likelihood ratio testing (dotted) on selection of u. The horizontal axis displays the sample size, and the vertical displays the fraction of the times that the estimated u is equal to 2.

F. CONVERGENCE OF THE SPARSE ENVELOPE ESTIMATOR $\widehat{\beta}$ IN HIGH DIMENSIONAL SCENARIO

The simulation settings used in Figure 6 are the same as those used in Table 2 of the paper. Because Theorem 5 indicates $\|\widehat{\beta}-\beta\|_F = O_p[\{(r_n+s)\log r_n/n\}^{1/2}]$, we plotted the average of $[n/\{(r_n+s)\log r_n\}]^{1/2}\|\widehat{\beta}-\beta\|_F$ over 200 replications versus n. The bootstrap estimator of $\|\widehat{\beta}-\beta\|_F$ is computed based on the average of 200 bootstrap samples. With each bootstrap sample, we obtained the sparse envelope estimator $\widehat{\beta}_{\mathrm{boot}}$ and computed $\|\widehat{\beta}_{\mathrm{boot}}-\widehat{\beta}\|_F$. Figure 6 indicates that $\|\widehat{\beta}_{\mathrm{boot}}-\widehat{\beta}\|_F$ is a good approximation to $\|\widehat{\beta}-\beta\|_F$. Figure 6 also shows that $\|\widehat{\beta}-\beta\|_F$ is much smaller than $\|\widehat{\beta}_{\mathrm{ols}}-\beta\|_F$. This is a result of the efficiency gains from the envelope construction.

G. NOTATION TABLE

The notations in this table includes all the notation in the main text as well as those in the Supplementary material.

Fig. 6. Comparison of sparse envelope estimator (solid), bootstrap estimator (solid with asterisks) and standard estimator (dashed).

REFERENCES

COOK, R. D., LI, B. & CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statist. Sinica* **20**, 927–1010.

COOK, R. D. & SETODJI, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Am. Statist. Assoc.* **98**, 340–351.

DATTORRO, J. (2005). *Convex Optimization & Euclidean Distance Geometry*. Palo Alto: Meboo Publishing USA.

MAGNUS, J. R. & NEUDECKER, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7**, 381–394.

MAGNUS, J. R. & NEUDECKER, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–980.

SHAPIRO, A. (1986). Asymptotic theory of overparameterized structural models. *J. Am. Statist. Assoc.* **81**, 142–149.

SU, Z. & COOK, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika* **99**, 687–702.

SU, Z. & COOK, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statist. Sinica* **23**, 213–230.

[*Received April* 2012. *Revised September* 2012]

A	$A = \Gamma_2\Gamma_1^{-1}$
A_{-i}	the submatrix of A with row a_i^{T} removed
G	the sub matrix of G_A with row a_i^{T} removed
G_A	$G_A = (I_u, A^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{r\times u}$
$L(A)$	loss function in the optimization of A
	$L(A) = -2\log\lvert G_A^{\mathrm{T}} G_A\rvert + \log\lvert G_A^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{res}} G_A\rvert + \log\lvert G_A^{\mathrm{T}}\widehat{\Sigma}_Y^{-1} G_A\rvert$
$Q(a_i)$	majorization function in the optimization of a_i
X	predictors
Y	responses
$Y_{\mathcal{A}}$	active response, i.e., the responses having non-zero rows in Γ
$Y_{\mathcal{I}}$	inactive response, i.e., the responses having zero rows in Γ
Y_D	dynamic response, i.e., the responses having non-zero rows in β
Y_S	static response, i.e., the responses having zero rows in β
a_i	the transpose of the ith row in A
d	number of dynamic responses
n	sample size
p	number of predictors
q	number of active responses
r, r_n	number of responses, when r increases with n, r is written as r_n
$r_{\mathcal{A}}$	number of active responses
$r_{\mathcal{I}}$	number of inactive responses
r_D	number of dynamic responses
r_S	number of static responses
s_1	nonzero elements in the lower triangle (not including the diagonal elements) of $\Sigma_{\mathrm{res}}^{-1}$
s_2	nonzero elements in the lower triangle (not including the diagonal elements) of Σ_Y^{-1}
s	$\max\{s_1, s_2\}$
u	dimension of the envelope $\mathcal{E}_\Sigma(\mathcal{B})$
α	intercept
β	regression coefficients
$\beta_{\mathcal{A}}$	$\beta_{\mathcal{A}} = \Gamma_{\mathcal{A}}\eta$
β_D	the nonzero coefficients in β
$\widehat{\beta}$	sparse envelope estimator of β
$\widehat{\beta}_{\mathcal{A}}$	sparse envelope estimator of $\beta_{\mathcal{A}}$
$\widehat{\beta}_{\mathcal{A},2}$	active envelope estimator of $\beta_{\mathcal{A}}$
$\widehat{\beta}_{\mathcal{A},O}$	oracle envelope estimator of $\beta_{\mathcal{A}}$
$\widehat{\beta}_{\mathrm{ols}}$	ordinary least squares estimator of β
$\mathcal{B}$	span of β
$\mathcal{E}_\Sigma(\mathcal{B})$	the envelope subspace
η	coordinates of β with respect to Γ
$\mathcal{G}(r,u)$	$r\times u$ Grassmann manifold, i.e., the set of all u-dimensional subspaces in an r-dimensional space
$\gamma_{\max}(\cdot)$	largest eigenvalue of a matrix
$\gamma_{\min}(\cdot)$	smallest eigenvalue of a matrix
Γ	orthogonal basis of the envelope $\mathcal{E}_\Sigma(\mathcal{B})$
$\Gamma_{\mathcal{A}}$	the non-zero rows in Γ
$\Gamma_{\mathcal{A},0}$	completion of $\Gamma_{\mathcal{A}}$
Γ_0	orthogonal basis of the orthogonal complement of $\mathcal{E}_\Sigma(\mathcal{B})$
$\tilde{\Gamma}_0$	orthogonal basis of the orthogonal complement of $\mathcal{E}_\Sigma(\mathcal{B})$ with a block diagonal structure
Γ_1	the first u rows in Γ
Γ_2	the last $r-u$ rows in Γ

$\widehat{\Gamma}$	sparse envelope estimator of Γ
$\widehat{\Gamma}_{\mathcal{A}}$	sparse envelope estimator of $\Gamma_{\mathcal{A}}$
$\widehat{\Gamma}_{\mathcal{A},2}$	active envelope estimator of $\Gamma_{\mathcal{A}}$
$\widehat{\Gamma}_{\mathcal{A},O}$	oracle envelope estimator of $\Gamma_{\mathcal{A}}$
λ_i	tuning parameter in the optimization of a_i, λ_i can be written as $\lambda_i = \lambda\omega_i$, where λ is the common tuning parameter and ω_i's are the weights. In this paper, $\omega_i = 1/\|a_i\|_2^{\nu}$.
$\lambda_{\max,n}$	$\max(\lambda_1, \ldots, \lambda_{q-u})$ at sample size n
$\lambda_{\min,n}$	$\min(\lambda_{q-u+1}, \ldots, \lambda_{r-u})$ at sample size n
Ω	coordinates of Σ with respect to Γ
Ω_0	coordinates of Σ with respect to Γ_0
$\widetilde{\Omega}_0$	coordinates of Σ with respect to $\widetilde{\Gamma}_0$
$\widetilde{\Omega}_{0,\mathcal{A}}$	upper left block of $\widetilde{\Omega}_0$, which has dimension $(q-u)\times(q-u)$
$\widetilde{\Omega}_{0,\mathcal{A}\mathcal{I}}$	upper right block of $\widetilde{\Omega}_0$, which has dimension $(q-u)\times(r-q)$
$\widetilde{\Omega}_{0,\mathcal{I}}$	lower right block of $\widetilde{\Omega}_0$, which has dimension $(r-q)\times(r-q)$
$\widetilde{\Omega}_{0,\mathcal{I}\mathcal{A}}$	lower left block of $\widetilde{\Omega}_0$, which has dimension $(r-q)\times(q-u)$
$\widetilde{\Omega}_{0,\mathcal{A}\mid\mathcal{I}}$	$\widetilde{\Omega}_{0,\mathcal{A}} - \widetilde{\Omega}_{0,\mathcal{A}\mathcal{I}}\widetilde{\Omega}_{0,\mathcal{I}}^{-1}\widetilde{\Omega}_{0,\mathcal{I}\mathcal{A}}$
μ_X	mean of the predictors X
Σ	variance of the error vector ε
Σ_X	variance of the predictors X
Σ_Y	variance of the responses Y
$\widehat{\Sigma}_X$	sample covariance matrix of X
$\widehat{\Sigma}_Y$	sample covariance matrix of Y
$\widehat{\Sigma}_{Y,\mathrm{sp}}^{-1}$	sparse permutation invariant covariance estimator of Σ_Y^{-1}
$\widehat{\Sigma}_{Y_{\mathcal{A}}\mid X}$	sample covariance matrix of the residuals from the regression of $Y_{\mathcal{A}}$ on X
$(\widehat{\Sigma}_Y^{-1})_{\mathcal{A}}$	the rows and columns in $\widehat{\Sigma}_Y^{-1}$ that have the same indices as $Y_{\mathcal{A}}$ in Y
$\widehat{\Sigma}_{\mathrm{res}}$	sample covariance matrix of the residuals from the regression of Y on X
$\widehat{\Sigma}_{\mathrm{res,sp}}^{-1}$	sparse permutation invariant covariance estimator of $\Sigma_{\mathrm{res}}^{-1}$
ε	error vector
$\otimes$	Kronecker product
$\perp\!\!\!\perp$	$V_1 \perp\!\!\!\perp V_2$ means V_1 and V_2 are independent
$\sim$	equality in distribution
$\perp$	orthogonal complement
$\dagger$	Moore–Penrose generalized inverse
$\|\cdot\|$	spectral norm of a matrix
$\|\cdot\|_2$	L_2 norm of a vector
$\|\cdot\|_F$	Frobenius norm of a matrix
P	projection matrix
Q	$I - P$
$\mathrm{vec}(\cdot)$	stack a matrix into a vector columnwise
$\mathrm{vech}(\cdot)$	stack the lower left triangle of a symmetric matrix into a vector
C_a, E_a	contraction matrix and expansion matrix: if M is an $a \times a$ symmetric matrix, $\mathrm{vech}(M) = C_a\,\mathrm{vec}(M),\ \mathrm{vec}(M) = E_a\,\mathrm{vech}(M)$