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, \ldots, 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 \(u\)th element and let \(\Gamma^* = Q_{e_u} \Gamma\). Since \(Y_u\) has regression coefficients zero, \(\beta = Q_{e_u}\beta\), giving \(B = Q_{e_u} B\). Therefore

\[\mathcal{B} = Q_{e_u} \mathcal{B} \subseteq Q_{e_u} \text{span}(\Gamma) = \text{span}(\Gamma^*).\]

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

\[\text{span}(\Gamma^*) = \text{span}(\Gamma_0) + \text{span}(P_{e_u} \Gamma) = \text{span}(\Gamma_0) + \text{span}(e_u),\]

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

\[\text{cov}(\tilde{\Gamma}^T Y, \Gamma^*^T Y | X) = \text{cov}(\tilde{\Gamma}^T Y, \Gamma^*^T Y | X) = 0.\]

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

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 \( \Sigma^{-1} \): If \( Y_i \) and \( Y_j \) are independent given the rest of the responses and \( X \), then the \((i, j)\)th element in \( \Sigma^{-1} \) is zero. On the other hand, if \( Y_i \) and \( Y_j \) are connected is based on the structure of \( \Sigma \). A sparse \( \Sigma^{-1} \) does not necessarily imply a sparse \( \Sigma \). 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 \( \Sigma \) has a block diagonal structure as follows:

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

The inverse matrix \( \Sigma^{-1} \) will preserve the block diagonal structure of \( \Sigma \), so the \((1, r)\)th element in \( \Sigma^{-1} \) 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 \( f_{\text{obj}}(A) \). It is sufficient to show that for any small \( \varepsilon > 0 \), there exists a sufficiently large constant \( C \), such that

\[
\lim_{n \to \infty} \inf_{\Delta \in \Re^{(r-n)\times n}, \|\Delta\|_F = C} \left\{ \frac{f_{\text{obj}}(A + n^{-1/2} \Delta)}{f_{\text{obj}}(A)} \right\} > 1 - \varepsilon. \tag{A1}
\]

If (A1) is established, then there exists a local minimizer \( \hat{A} \) of \( f_{\text{obj}} \) with arbitrarily large probability such that \( \|\hat{A} - A\|_F = O_p(n^{-1/2}) \). Therefore \( \hat{A} \) is a \( \sqrt{n} \)-consistent estimator of \( A \). As \( \hat{P}_T = G_A(I_n + A^T A)^{-1} G_A^T \) is a function of \( A \) only, \( \hat{P}_T \) is a \( \sqrt{n} \)-consistent estimator of \( P_T \). As \( \hat{\beta} = P_T \hat{\beta}_{\text{ols}} \), and \( \beta_{\text{ols}} \) is a \( \sqrt{n} \)-consistent estimator of \( \beta \), then \( \hat{\beta} \) is a \( \sqrt{n} \)-consistent estimator of \( \beta \).

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

\[
f_{\text{obj}}(A) = -2 \log |G_A^T G_A| + \log |G_A^T \hat{\Sigma}_{\text{res}} G_A| + \log |G_A^T \hat{\Sigma}_{Y}^{-1} G_A| + \sum_{i=1}^{r-n} \lambda_i \|a_i\|_2
\]

\[
\equiv f_1(A) + f_2(A) + f_3(A) + f_4(A),
\]
Sparse Envelope Model

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

\[
f_1(A + n^{-1/2} \Delta) = f_1(A) + n^{-1/2} \overrightarrow{\Delta} f_1(A) + \frac{1}{2} n^{-1} \overrightarrow{\Delta}^2 f_1(A) + o_p(n^{-1}),
\]

where \( \overrightarrow{\Delta} f_1(A) \) and \( \overrightarrow{\Delta}^2 f_1(A) \) are directional derivatives (Dattorro, 2005, p.706).

The first directional derivative is

\[
\overrightarrow{\Delta} f_1(A) = \text{tr} \left\{ \frac{d}{dA} f_1(A)^T \Delta \right\} = -4 \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta \right\}.
\]

The second directional derivative is

\[
\overrightarrow{\Delta}^2 f_1(A) = -4 \text{tr} \left[ \left( \frac{d}{dA} \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta \right\} \right)^T \Delta \right]
= -4 \text{tr} \left\{ I_u + A^T A \right\}^{-1} A^T \Delta \cdot \left\{ A(I_u + A^T A)^{-1} A^T - I_{r-u} \right\} \Delta
= 4 \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta - (I_u + A^T A)^{-1} \Delta^2 \right\}.
\]

Let

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

then

\[
\overrightarrow{\Delta} f_1(A) = 4 \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta \\
+ (I_u + A^T A)^{-1} \Delta^T (A(I_u + A^T A)^{-1} A^T - I_{r-u}) \Delta \right\}
= 4 \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta + (I_u + A^T A)^{-1} \Delta (I_u + A^T A)^{-1} \Gamma \Delta_0 \Gamma_0^T \Delta_* \right\}
= 4 \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta - (I_u + A^T A)^{-1} \Delta^2 \Gamma_0 \Gamma_0^T \Delta_* \right\}.
\]

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

\[
f_1(A + n^{-1/2} \Delta) - f_1(A) = -4n^{-1/2} \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta \right\}
+ 2n^{-1} \text{tr} \left\{ (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta \\
- (I_u + A^T A)^{-1} \Delta^2 \Gamma_0 \Gamma_0^T \Delta_* \right\}.
\]

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

\[
\overrightarrow{\Delta} f_2(A) = \text{tr} \left\{ \frac{d}{dA} f_2(A)^T \Delta \right\} = 2 \text{tr} \left\{ (G_A^T \tilde{\Sigma}_{\text{res}} G_A)^{-1} G_A^T \tilde{\Sigma}_{\text{res}} \Delta_* \right\}.
\]

Let \( \Sigma_X, \Sigma_Y \) and \( \Sigma_{XY} \) be the variance matrix of \( X \), the variance matrix of \( Y \) and the covariance matrix of \( Y \) and \( X \) in population, and let \( \hat{\Sigma}_X, \hat{\Sigma}_Y \) and \( \hat{\Sigma}_{XY} \) be the corresponding sample versions. Then by Cook
\[ n^{1/2}(\hat{\Sigma}_Y - \Sigma_Y) = n^{-1/2}(Y^T X - n\Sigma_Y) + O_p(n^{-1/2}), \]
\[ n^{1/2}(\hat{\Sigma}_X - \Sigma_X) = n^{-1/2}(X^T X - n\Sigma_X) + O_p(n^{-1/2}), \]
\[ n^{-1/2}(\hat{\Sigma}_Y - \Sigma_Y) = n^{-1/2}(Y^T Y - n\Sigma_Y) + O_p(n^{-1/2}), \]

where \( Y \in \mathbb{R}^{n \times r} \) is the centred data matrix of \( Y \), whose \( i \)th row is \((Y_i - \bar{Y})^T\). Since \( \hat{\Sigma}_{\text{res}} = \hat{\Sigma}_Y - \hat{\Sigma}_Y \hat{\Sigma}_X^{-1} \hat{\Sigma}_Y \) and \( \hat{\Sigma}_X^{-1} - \Sigma_X^{-1} = -\Sigma_X^{-1}(\hat{\Sigma}_X - \Sigma_X)\Sigma_X^{-1} + O_p(n^{-1}), \)
\[ \hat{\Sigma}_{\text{res}} = (\hat{\Sigma}_Y - \Sigma_Y + \Sigma_Y) - (\hat{\Sigma}_Y - \Sigma_Y + \Sigma_Y)(\hat{\Sigma}_X^{-1} - \Sigma_X^{-1} + \Sigma_X^{-1})(\hat{\Sigma}_Y - \Sigma_Y + \Sigma_Y) \]
\[ = \Sigma + n^{-1/2}\left\{ -n^{-1/2}(Y^T X - n\Sigma_Y)\Sigma_X^{-1}\Sigma_X^{-1}(X^T X - n\Sigma_X)\Sigma_X^{-1}\Sigma_X - n^{-1/2}(Y^T Y - n\Sigma_Y) \right\} + O_p(n^{-1}) \]
\[ = \Sigma + n^{-1/2}(T_1 + T_2 + T_3 + T_4) + O_p(n^{-1}), \]
where by the central limit theorem, each element in \( T_1, T_2, T_3 \) and \( T_4 \) converges in distribution to a normal random variable which has mean 0. As
\[ \left(G_A^T \hat{\Sigma}_{\text{res}} G_A \right)^{-1} = (G_A^T \Sigma G_A)^{-1} - (G_A^T \Sigma G_A)^{-1}(G_A^T \hat{\Sigma}_{\text{res}} G_A - G_A^T \Sigma G_A)(G_A^T \Sigma G_A)^{-1} + O_p(n^{-1}) \]
\[ = (G_A^T \Sigma G_A)^{-1} - n^{-1/2}(G_A^T \Sigma G_A)^{-1}G_A^T(T_1 + T_2 + T_3 + T_4)G_A(G_A^T \Sigma G_A)^{-1} \]
\[ + O_p(n^{-1}), \]
\[ \frac{dY}{df_2} (\Gamma) \] can be expanded as
\[ 2 \text{tr}\{ (G_A^T \hat{\Sigma}_{\text{res}} G_A)^{-1} G_A^T (\hat{\Sigma}_{\text{res}} \Delta) \} = 2 \text{tr}\{ (G_A^T \Sigma G_A)^{-1} G_A^T (T_1 + T_2 + T_3 + T_4) \Delta \} + 2n^{-1/2} \text{tr}\{ (G_A^T \Sigma G_A)^{-1} G_A^T (T_1 + T_2 + T_3 + T_4) \Delta \} \]
\[ = 2 \text{tr}\{ (I + A^T A)^{-1} G_A^T \Delta \} + 2n^{-1/2} \text{tr}\{ (G_A^T \Sigma G_A)^{-1} G_A^T (T_3 + T_4) \Gamma_0 \Gamma_0^T \Delta \} + O_p(n^{-1}) \]
\[ = 2 \text{tr}\{ (I + A^T A)^{-1} A^T A \} + 2n^{-1/2} \text{tr}\{ (G_A^T \Sigma G_A)^{-1} G_A^T (T_3 + T_4) \Gamma_0 \Gamma_0^T \Delta \} + O_p(n^{-1}) \]
\[ = 2 \text{tr}\{ (I + A^T A)^{-1} A^T A \} + 2n^{-1/2} \text{tr}\{ \Gamma_1 \Omega^{-1} A^T (T_3 + T_4) \Gamma_0 \Omega_0 \Delta \} + O_p(n^{-1}). \]

The second equality is because \( \Gamma = G_A \Gamma_1 \), so
\[ \Gamma_1 G_A G_A \Gamma_1 = I \Rightarrow \Gamma_1 (I + A^T A) \Gamma_1 = I \Rightarrow I + A^T A = (\Gamma_1)^{-1} \Gamma_1^{-1} \Rightarrow (I + A^T A)^{-1} = \Gamma_1 \Gamma_1^{-1}, \]
and
\[ (G_A^T \Sigma G_A)^{-1} G_A^T \Sigma = \{(\Gamma_1^{-1})^T \Sigma \Gamma_1^{-1}\}^{-1} (\Gamma_1^{-1})^T \Sigma = \Gamma_1 \Omega^{-1} \Omega^T \Gamma_1 = \Gamma_1 \Gamma_1^{-1} G_A^T = (I + A^T A)^{-1} G_A^T. \]

Using the Cauchy–Schwarz inequality for matrix trace (Magnus & Neudecker, 2007, p.227),
\[ \left| \text{tr}\{ \Gamma_1 \Omega^{-1} A^T (T_3 + T_4) \Gamma_0 \Omega_0 \Delta \} \right| \leq \| \Delta \|_F \| \Gamma_1 \Omega^{-1} A^T (T_3 + T_4) \Gamma_0 \Omega_0 \|_F \]
\[ = \| \Delta \|_F \| \Gamma_1 \Omega^{-1} \Omega^T (T_3 + T_4) \Gamma_0 \|_F. \]
The second directional derivative of $f_2$ is

$$\nabla^2 f_2(A) = 2 \text{tr} \left( \frac{d}{dA} \text{tr} \left( (G_A \hat{\Sigma}_{\text{res}} G_A)^{-1} G_A \hat{\Sigma}_{\text{res}} \Delta \right) \right)$$

$$= 2 \text{tr} \left( (G_A \hat{\Sigma}_{\text{res}} G_A)^{-1} \Delta^T G_A \hat{\Sigma}_{\text{res}} \Delta \right)$$

$$- (G_A \hat{\Sigma}_{\text{res}} G_A)^{-1} \left( (G_A \hat{\Sigma}_{\text{res}} \Delta + \Delta^T G_A \hat{\Sigma}_{\text{res}} G_A) (G_A \hat{\Sigma}_{\text{res}} G_A)^{-1} G_A \hat{\Sigma}_{\text{res}} \Delta \right)$$

$$= 2 \text{tr} \left( (G_A \hat{\Sigma}_{G \Sigma G} A)^{-1} \Delta^T \Sigma \Delta - (G_A \hat{\Sigma}_{G \Sigma G} A)^{-1} (G_A \hat{\Sigma}_{G \Sigma G} A + \Delta^T \Sigma G_A) (G_A \hat{\Sigma}_{G \Sigma G} A)^{-1} G_A \hat{\Sigma}_{G \Sigma G} A \right)$$

$$+ O_p(n^{-1/2})$$

$$= 2 \text{tr} \left( - (I_u + A^T A)^{-1} G_A^T \Delta (I_u + A^T A)^{-1} G_A \Delta \right)$$

$$+ (G_A \hat{\Sigma}_{G \Sigma G} A)^{-1} \Delta^T \Sigma \Delta + O_p(n^{-1/2})$$

$$= 2 \text{tr} \left( - (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta \right)$$

$$+ (G_A \hat{\Sigma}_{G \Sigma G} A)^{-1} \Delta^T \Sigma \Delta + O_p(n^{-1/2})$$

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

$$f_2(A + n^{-1/2} \Delta) - f_2(A)$$

$$= 2n^{-1/2} \text{tr} \left( (I_u + A^T A)^{-1} A^T \Delta \right) + 2n^{-1} \text{tr} \left( \Gamma_1 \Omega^{-1} \Gamma_1^T (T_{3n} + T_{4n}) \Gamma_0 \Gamma_0^T \Delta \right)$$

$$+ n^{-1} \text{tr} \left( - (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta + \Omega^{-1} \Gamma_1^T \Gamma_1^T \Delta \right)$$

$$+ O_p(n^{-1/2})$$

$$\geq 2n^{-1/2} \text{tr} \left( (I_u + A^T A)^{-1} A^T \Delta \right) - 2n^{-1} \text{tr} \left( \| \Gamma_1 \Omega^{-1} \Gamma_1^T (T_{3n} + T_{4n}) \Gamma_0 \Gamma_0^T \Delta \right)$$

$$+ n^{-1} \text{tr} \left( - (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta + \Omega^{-1} \Gamma_1^T \Gamma_1^T \Delta \right)$$

$$+ O_p(n^{-1/2})$$

Since that $f_3$ has similar structure as $f_2$, the derivation above can be applied parallel to $f_2$, just with $\hat{\Sigma}_{\text{res}}$ replaced by $\Sigma_Y^{-1}$. Let $T_{3n} = - n^{-1/2} \Sigma_Y^{-1} (\Sigma_Y^{1/2} \Sigma_Y^{1/2} - n \Sigma_Y) \Sigma_Y^{-1}$. By the central limit theorem, $T_{3n}$ converges in distribution to a normal random variable with mean 0. After some straightforward algebra, we have

$$f_3(A + n^{-1/2} \Delta) - f_3(A)$$

$$= 2n^{-1/2} \text{tr} \left( (I_u + A^T A)^{-1} A^T \Delta \right) + 2n^{-1} \text{tr} \left( \Gamma_1 \Omega^{-1} \Gamma_1^T (T_{3n} + T_{4n}) \Gamma_0 \Gamma_0^T \Delta \right)$$

$$+ n^{-1} \text{tr} \left( - (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta + \Omega^{-1} \Gamma_1^T \Gamma_1^T \Delta \right)$$

$$+ O_p(n^{-1/2})$$

$$\geq 2n^{-1/2} \text{tr} \left( (I_u + A^T A)^{-1} A^T \Delta \right) - 2n^{-1} \text{tr} \left( \| \Gamma_1 \Omega^{-1} \Gamma_1^T (T_{3n} + T_{4n}) \Gamma_0 \Gamma_0^T \Delta \right)$$

$$+ n^{-1} \text{tr} \left( - (I_u + A^T A)^{-1} A^T \Delta (I_u + A^T A)^{-1} A^T \Delta + \Omega^{-1} \Gamma_1^T \Gamma_1^T \Delta \right)$$

$$+ O_p(n^{-1/2})$$

Now we expand $f_4(A) = \sum_{i=1}^{q-u} \lambda_i ||a_i||_2^2$. Let $\delta_i^T$ be the ith row of $\Delta$, then

$$f_4(A + n^{-1/2} \Delta) - f_4(A) \geq \sum_{i=1}^{q-u} \left( \lambda_i ||a_i + n^{-1/2} \delta_i||_2^2 - \lambda_i ||a_i||_2^2 \right)$$

$$\geq - \frac{1}{2} (q-u)n^{-1/2} \lambda_{\text{max},n} \max_i \left( ||a_i||_2^{-1} ||\delta_i||_2 \right) \left( 1 + o_p(1) \right)$$

$$= - \frac{1}{2} (q-u)n^{-1/2} \lambda_{\text{max},n} \max_i \left( ||a_i||_2^{-1} ||\delta_i||_2 \right) \left( 1 + o_p(1) \right).$$
The second inequality is based on Taylor expansion at $a_i$. As $n^{1/2} \lambda_{\text{max}, n} \to 0$ as $n \to \infty$, $n\{f_d(A + n^{-1/2} \Delta) - f_d(A)\} = o_p(1)$. Collecting all the results so far

\[
\begin{align*}
    f_{\text{obj}}(A + n^{-1/2} \Delta) - f_{\text{obj}}(A) \\
    \geq -2n^{-1}||\Delta||_F ||\Gamma_i^{1}\Omega^{1}\Gamma^{1}(T_{3 n} + T_{4 n})\Gamma_0||_F - 2n^{-1}||\Delta||_F ||\Gamma_1^{1}(\Omega + \eta \Sigma X \eta^T)\Gamma_r T_{5 n} \Gamma_0||_F \\
    + n^{-1} \text{tr} \left\{ \Omega^{1}\Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 - 2(I_u + A^T A)^{-1} \Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 \right\} \\
    - 2(I_u + A^T A)^{-1} \Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 - \frac{1}{2} n^{-1} (q - u) n^{1/2} \lambda_{\text{max, n}} \max_i \left( ||a_i||^{-1}_2 ||\delta_i||_2 \right) + o_p(n^{-1}).
\end{align*}
\]

Notice that

\[
\begin{align*}
    \text{tr} \left\{ \Omega^{1}\Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 - 2(I_u + A^T A)^{-1} \Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 \right\} \geq m\|\Gamma_0^{T}\Delta_{x} \Gamma_1\|_F^2,
\end{align*}
\]

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

\[
\begin{align*}
    \|\Gamma_0^{T}\Delta_{x} \Gamma_1\|_F^2 &= \text{tr}(\Gamma_0^{T}\Delta_{x} \Gamma_1\Gamma_1^{T}\Gamma_0) \\
    &= \text{tr}(\Gamma_0^{T}\Delta_{x}(I_u + A^T A)^{-1}\Delta_{x}^{T}\Gamma_0) \\
    &= \text{tr}(\Delta_{x}(I_u + A^T A)^{-1}\Delta_{x}^{T}(I_r - \Gamma_r^{T})) \\
    &= \text{tr}([I_u + A^T A]^{-1}\Delta_{x}^{T}(I_r - G_A(I_u + A^T A)^{-1}G_A^{T}) \Delta_{x}) \\
    &= \text{tr}([I_u + A^T A]^{-1}\Delta_{x}^{T}(I_r - A(I_u + A^T A)^{-1}A^{T}) \Delta_{x}) \\
    &= \text{tr}([I_u + A^T A]^{-1}\Delta_{x}^{T}(I_u + A^T A)^{-1}\Delta_{x}) \\
    &= \text{vec}(\Delta)^{T}\{[I_u + A^T A]^{-1} \otimes (I_u + A^T A)^{-1}\} \text{vec}(\Delta) \geq m_0^2\|\Delta\|_F^2,
\end{align*}
\]

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

\[
\begin{align*}
    \text{tr} \left\{ \Omega^{1}\Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_i^{1}\Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 - 2(I_u + A^T A)^{-1} \Delta^{2}_{x} \Gamma_0 \Omega_0^{1}\Delta_{x} \Gamma_1 \right\} \geq m m_0^2\|\Delta\|_F^2.
\end{align*}
\]

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 $f_{\text{obj}}$ with respect to $a_i$ should be $0$ evaluated at the local minimum $\widehat{a}_i$. The derivative of $f_{\text{obj}}$ with respect to $a_i$ ($i = q + 1 - u, \ldots, r - u$) is

\[
\frac{\partial f_{\text{obj}}}{\partial a_i} = -4e_i^{T}G_A(I_u + A^T A)^{-1} + 2e_i^{T}\tilde{\Sigma}_{\text{res}}G_A(G_A^{T}\tilde{\Sigma}_{\text{res}}G_A)^{-1} + 2e_i^{T}\tilde{\Sigma}_{Y}^{-1}G_A(G_A^{T}\tilde{\Sigma}_{Y}^{-1}G_A)^{-1} + \frac{\lambda_i a_i^{T}}{||a_i||_2},
\]

where $e_i$ be the $i$th column of $I_r$. Then

\[
-4e_i^{T}G_A(I_u + A^T A)^{-1} + 2e_i^{T}\tilde{\Sigma}_{\text{res}}G_A(G_A^{T}\tilde{\Sigma}_{\text{res}}G_A)^{-1} + 2e_i^{T}\tilde{\Sigma}_{Y}^{-1}G_A(G_A^{T}\tilde{\Sigma}_{Y}^{-1}G_A)^{-1} + \lambda_i \frac{\widehat{a}_i^{T}}{||\widehat{a}_i||_2} = 0.
\]

(A2)
Because $\hat{\Sigma}_{\text{res}}, \hat{\Sigma}_Y$ and $\hat{A}$ are $\sqrt{n}$-consistent estimators of $\Sigma, \Sigma_Y$ and $A$, $\Sigma = \Gamma \Omega \Gamma^\top + \Gamma_0 \Omega_0 \Gamma_0^\top$ and $\Sigma_Y = \Gamma (\Omega + \eta \Sigma_X \eta^\top) \Gamma^\top + \Gamma_0 \Omega_0 \Gamma_0^\top$.

$$-4e_i^\top \hat{G}_A (I_n + A^\top \hat{A})^{-1} + 2e_i^\top \hat{\Sigma}_{\text{res}} \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_{\text{res}} \hat{G}_A) - 1 + 2e_i^\top \hat{\Sigma}_Y \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_Y \hat{G}_A)^{-1} - 2e_i^\top \hat{\Sigma}_Y \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_Y \hat{G}_A)^{-1} = O_p(n^{-1/2})$$

Then $n^{1/2} \left\{ -4e_i^\top \hat{G}_A (I_n + A^\top \hat{A})^{-1} + 2e_i^\top \hat{\Sigma}_{\text{res}} \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_{\text{res}} \hat{G}_A) - 1 + 2e_i^\top \hat{\Sigma}_Y \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_Y \hat{G}_A)^{-1} \right\} = O_p(1)$. 

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 $n^{1/2} \lambda_{\text{min},n} \rightarrow \infty$, there is at least one element in $n^{1/2} \lambda_i a_i^\top / \|a_i\|_2$ that tends to infinity. With (A2), this is a contradiction of $n^{1/2} \{ -4e_i^\top \hat{G}_A (I_n + A^\top \hat{A})^{-1} + 2e_i^\top \hat{\Sigma}_{\text{res}} \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_{\text{res}} \hat{G}_A) - 1 + 2e_i^\top \hat{\Sigma}_Y \hat{G}_A (\hat{G}_A^\top \hat{\Sigma}_Y \hat{G}_A)^{-1} \} = O_p(1)$.

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

\[ \square \]

**Proof of Proposition 2 and Proposition 3.** For the proof of Proposition 3, the derivation of the maximum likelihood estimator of $\beta_D$ 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 $\beta_D$. As $Y = (Y_D^T, Y_S^T)^T$, we can partition the centred matrix $\Sigma$ accordingly into $\Sigma = (\Sigma_{c,D}, \Sigma_{c,S})$. We also partition the matrix $\Sigma^{-1}$ into

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

The log likelihood function under model (12) is

$$l = -\frac{n(r + p)}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma_X| - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \{ (X - n\mu_X) \Sigma^{-1} (X - n\mu_X)^\top \}$$

$$- \frac{1}{2} \text{tr} \{ (\Sigma_{c,D} - n\beta_D)^T, \Sigma_{c,S})^{-1} (\Sigma_{c,D} - n\beta_D, \Sigma_{c,S})^T \}.$$  

It is easy to show that $\mu_X = \bar{X}, \Sigma_X = (X - n\mu_X)^T (X - n\mu_X) / n$, and $\hat{\alpha} = \hat{Y}$. Substituting these estimates in, the partially maximized log likelihood is

$$l = -\frac{n(r + p)}{2} \log(2\pi) - \frac{n}{2} \log |\hat{\Sigma}_X| - \frac{np}{2} - \frac{n}{2} \log |\Sigma|$$

$$- \frac{1}{2} \text{tr} \{ (\Sigma_{c,D} - X \beta_D^T, \Sigma_{c,S})^{-1} (\Sigma_{c,D} - X \beta_D^T, \Sigma_{c,S})^T \}$$

$$= -\frac{n(r + p)}{2} \log(2\pi) - \frac{np}{2} \log |\hat{\Sigma}_X| - \frac{np}{2} - \frac{n}{2} \log |\Sigma|$$

$$- \frac{1}{2} \text{tr} \{ (\Sigma_{c,D} - X \beta_D^T) M_1 (\Sigma_{c,D} - X \beta_D^T)^T + 2(\Sigma_{c,D} - X \beta_D^T) M_2 \Sigma_{c,S} + \Sigma_{c,S} M_3 \Sigma_{c,S} \}.$$  

Take the derivative of $l$ with respect to $\beta_D$ and $\Sigma$, we get

$$\frac{\partial l}{\partial \beta_D} = -M_1 (X \beta_D^T - Y_{c,D}^T) X_c - M_2 X_{c,S} X_c,$$

$$\frac{\partial l}{\partial \Sigma} = -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} (\Sigma_{c,D} - X \beta_D^T, \Sigma_{c,S})^T (\Sigma_{c,D} - X \beta_D^T, \Sigma_{c,S}) \Sigma^{-1}.$$
Set the derivatives to 0 and we get
\[ \hat{\beta}_D = Y_{c,D}^T \Sigma_c (X_c^T X_c)^{-1} - M_1^{-1} M_2 \Sigma_{c,S} X_c (X_c^T X_c)^{-1} = \hat{\beta}_{D,obs} - \tilde{\Sigma}_{DS} \Sigma_{S,obs} \] and
\[ \hat{\Sigma} = \frac{1}{n} (Y_{c,D} - X_c \hat{\beta}_{D,obs}) (Y_{c,D} - X_c \hat{\beta}_{D,obs})^T. \] Since \( \beta_S = 0 \), \( \hat{\Sigma}_S = S_S^{-1} \),
where \( S_S^{-1} \) is the sample covariance matrix of \( Y_S \). We can build an equation with \( \hat{\Sigma}_{DS} \). Notice that
\[ \hat{\Sigma}_{DS} = \frac{1}{n} (Y_{c,D} - X_c \hat{\beta}_{D,obs})^T \Sigma_{c,S} \]
\[ = \frac{1}{n} (Y_{c,D} - X_c \hat{\beta}_{D,obs} + X_c \Sigma_{DS} \hat{\Sigma}_S^{-1} \hat{\beta}_{S,obs})^T \Sigma_{c,S} \]
\[ = \frac{1}{n} (Q X_{c,D} + P X_{c,S} S_c^{-1} \hat{\Sigma}_{DS})^T \Sigma_{c,S}. \]

Solve for \( \hat{\Sigma}_{DS} \), we get
\[ \hat{\Sigma}_{DS} = Y_{c,D}^T Q X_{c,S} \Sigma_{c,S}^{-1} S_{c,S}^{-1} Y_{c,S}. \]
Substitute it into \( \hat{\beta}_D \), we get \( \hat{\beta}_D = \hat{\beta}_{D,obs} - \hat{\beta}_{D,S} \hat{\beta}_{S,obs} \), where \( \hat{\beta}_{D,S} = Y_{c,D}^T Q X_{c,S} \Sigma_{c,S}^{-1} S_{c,S}^{-1} Y_{c,S} \)
contains the coefficients from the regression of \( \Sigma \).

To compute the asymptotic variance of the maximum likelihood estimators, we compute the Fisher information matrix for \( \{ \text{vec}(\beta_D)^T, \text{vech}(\Sigma)^T \} \), where \( \text{vec} \) 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 \( \text{vec} \) operator and \( \text{vech} \) operator: \( \text{vech}(M) = C_a \text{vec}(M) \) and \( \text{vec}(M) = E_a \text{vech}(M) \). After some straightforward algebra, the Fisher information matrix \( J \) is
\[ J = \left( \begin{array}{cc} \Sigma_X \otimes (\Sigma_D - \Sigma_{DS} \Sigma_S^{-1} \hat{\Sigma}_{DS})^{-1} & 0 \\ 0 & \frac{1}{n} E_a^T (\Sigma^{-1} \otimes \Sigma^{-1}) E_a \end{array} \right). \]

The inverse of the upper left block of \( J \) relates to the asymptotic variance of \( \text{vec}(\hat{\beta}_D) \). Therefore
\[ n^{1/2} \{ \text{vec}(\hat{\beta}_{D,1}) - \text{vec}(\beta_D) \} \rightarrow N(0, \Sigma_X^{-1} \otimes \Sigma_D[S]). \]
in distribution as \( n \rightarrow \infty \).

**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 \( \beta_A \) is similar to the derivation of the maximum likelihood estimator of \( \beta \) 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 \( \{ \text{vec}(\hat{\beta}_{A,1})^T, \text{vech}(\hat{\Sigma}_1)^T \} \), while \( \hat{\Sigma}_1 \) 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 \( \xi \) is our \( \{ \text{vec}(\hat{\beta}_A)^T, \text{vech}(\Sigma)^T \} \). Let \( l \) be the log-likelihood function in (A3) and let \( l_{max} \) be its maximum value. We define the minimum discrepancy function as \( f_{\text{MDF}} = l_{max} - l \). Since \( f_{\text{MDF}} \) is derived from the normal likelihood function, it satisfies the four conditions in Section 3 of Shapiro (1986). Our \( \{ \text{vec}(n)^T, \text{vech}(\Gamma_A)^T, \text{vech}(\Omega)^T \} \) is Shapiro’s \( \theta \). Therefore the function \( g \) that connects \( \xi \) and \( \theta: \xi = g(\theta) \) is twice differentiable. All the assumptions in Proposition 4-1 are satisfied. Let \( \hat{\Sigma}_O \) be the estimator of \( \Sigma \) under the oracle envelope model (14), then
\[ n^{1/2} \{ \text{vec}(\hat{\beta}_{A,O})^T, \text{vech}(\hat{\Sigma}_O)^T \} - \{ \text{vec}(\hat{\beta}_A)^T, \text{vech}(\Sigma)^T \} \] 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 \( \text{vec}(\hat{\beta}_{A,O}) \). Proposition 4-1 indicates that the asymptotic variance has the form \( H(H^T J H)^T H^T \), where \( H \) denotes Moore–Penrose inverse, \( J \) is the Fisher information displayed at the end of the proof for Proposition 4.
Sparse Envelope Model

9

tion 2, and $H$ is the Jacobian matrix $\partial \xi / \partial \theta$

$$H = \begin{bmatrix} I_p \otimes \Gamma_A & \eta^T \otimes I_q \\ 0 & 2C_r(I_r \otimes \Gamma - \Gamma_0 \Gamma_0^T \otimes \Gamma) L C_r(\Gamma \otimes \Gamma_0) E_u \end{bmatrix},$$

where $L = (K_{qu}^T, 0)^T \in \mathbb{R}^{n_0 \times qu}$, 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 $\text{vec}(\hat{\beta}_{A,O})$:

$$n^{1/2} \{ \text{vec}(\hat{\beta}_{A,O}) - \text{vec}(\beta_A) \} \to N(0, V_0)$$
in distribution, where $V_0 = \Sigma_X^{-1} \otimes \Gamma_A \Gamma_A^T + (\eta^T \otimes \Gamma_A) T(\eta \otimes \Gamma_A), T = \eta \Sigma_X \eta^T \otimes \tilde{\Omega}_{0,4}^{-1} \tilde{\Omega}_{0,4}^{-1} \tilde{\Omega}_{0,4}^{-1} + \Omega \otimes \tilde{\Omega}_{0,4}^{-1} \tilde{\Omega}_{0,4}^{-1} - 2I_u \otimes I_q - u$.

Note: We ignored $\mu_X$, $\alpha$ and $\Sigma_X$ in $J$ and $H$ matrices. This does not affect the results because they are not involved in the parameterization of $\beta$ and $\Sigma$, and their maximum likelihood estimates are asymptotically independent of the estimates of $\beta$ and $\Sigma$. \hfill \Box

Proof of Theorem 4. Let $\hat{A}_A$ denote the nonzero rows in the sparse envelope estimator $\hat{A}_A$, and $\hat{A}_O$ denote the nonzero rows in the oracle envelope estimator. As $P_1 = G_A(G_0^T G_A)^{-1} G_0^T$, for a sequence $a_n = o(n^{-1/2})$, if $\hat{A}_A = \hat{A}_O + O_p(a_n)$, then $P_1 = P_{\hat{A}_O} + O_p(a_n)$. Therefore $\beta - \hat{\beta}_O = (P_1 - P_{\hat{A}_O}) \beta_{obs} = (P_1 - P_{\hat{A}_O})(\beta_{obs} - \beta) + (P_1 - P_{\hat{A}_O}) \beta = O_p(a_n) \eta(1) + O_p(a_n) = O_p(a_n)$. So $n^{1/2}(\hat{\beta}_O - \beta) \to 0$ in probability. By Slutsky’s theorem $n^{1/2}(\hat{\beta}_O - \beta)$ has the same asymptotic distribution as $n^{1/2}(\tilde{\Omega}_O - \beta)$. From the proof of Proposition 4, we know that $n^{1/2}(\tilde{\Omega}_O - \beta)$ 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 $\hat{A}_A = A_O + O_p(a_n)$ 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 $G_B = \begin{bmatrix} I_u \\ B \end{bmatrix} \in \mathbb{R}^{q \times u}$.

Define

$$f_{\text{obj}, A}(B) = -2 \log |G_B^T G_B| + \log |G_B^T \hat{\Sigma}_Y| X G_B| + \log |G_B^T (\hat{\Sigma}_Y^{-1})_A G_B| + \sum_{i=1}^{q-u} \lambda_i ||b_i||_2,$$

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

$$\lim \text{pr} \left\{ \inf_{\Delta \in \mathbb{R}^{(q-u) \times u}, \|\Delta\|_F = C} f_{\text{obj}, A}(\hat{A}_O + a_n \Delta) > f_{\text{obj}, A}(\hat{A}_O) \right\} > 1 - \varepsilon. \quad (A3)$$

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

$$\begin{align*}
\frac{d f_{1, A}}{d B}(A_O) &= \text{tr}\left\{ \frac{d}{d B} f_{1, A}(B)^T \right\}_{|B = A_O} \Delta, \\
\frac{d f_{2, A}}{d B}(A_O) &= \text{tr}\left\{ \frac{d}{d B} f_{2, A}(B)^T \right\}_{|B = A_O} \Delta, \\
\frac{d f_{3, A}}{d B}(A_O) &= \text{tr}\left\{ \frac{d}{d B} f_{3, A}(B)^T \right\}_{|B = A_O} \Delta.
\end{align*}$$
Since \( \hat{A}_O \) is a minimizer of \( f_1(A) + f_2(A) + f_3(A) \),
\[
\frac{d}{dB} f_1(A) \bigg|_{B = \hat{A}_O} + \frac{d}{dB} f_3(A) \bigg|_{B = \hat{A}_O} + \frac{d}{dB} f_3(A) \bigg|_{B = \hat{A}_O} = 0.
\]
Then \( df_1(A) + df_2(A) + df_3(A) = 0 \).

The calculations on the second directional derivatives of \( f_1(A) + f_2(A) \) at \( \hat{A}_O \) and the expansion of \( f_4(A) \) are parallel to those in Theorem 2. Assembling all those terms together, we have
\[
f_{obj}(\hat{A}_O + \alpha \Delta) - f_{obj}(\hat{A}_O) \geq \alpha^2 \sum_{i,j} \left\{ \Omega^{-1} \Gamma_i \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} + (\Gamma + \eta \Sigma X \eta^T) \Gamma_i \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} \right. \\
- 2(I_u + A_i^T A_i) \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} \} \geq \frac{1}{2} m \eta \Omega_{\Delta,A} F^2,
\]
where \( A_i \in \mathbb{R}^{(q-u) \times u} \) contains the nonzero rows in \( A \) and \( \Delta_{A} = (0_{u \times u}, \Delta^T) \in \mathbb{R}^{q \times u} \). Based on the definition of \( \eta \), we have \( \lambda_{\max,n} = \rho(p(a_n)) \). 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
\[
\text{tr} \left\{ \Omega^{-1} \Gamma_i \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} + (\Gamma + \eta \Sigma X \eta^T) \Gamma_i \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} \right. \\
- 2(I_u + A_i^T A_i) \tilde{\Delta}_{i,j}^T \Gamma_i \hat{A}_{i,j} \} \geq m \eta \Omega_{\Delta,A} F^2,
\]
where \( m_0 \) is the smallest eigenvalue of \( (I_u + A_i^T A_i)^{-1} \), and \( m \) is the smallest eigenvalue of \( \Omega^{-1} \odot \tilde{\Omega}_{0,A} + (\Omega + \eta \Sigma \hat{X} \eta^T) \odot \tilde{\Omega}_{0,A}^2 - 2I_u \odot \tilde{\Omega}_{u,-u} \). 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
\[
\| \hat{\Sigma}_{res,sp}^{-1} - \Sigma^{-1} \|_F = O_p(\{(r_n + s_1) \log r_n/n\}^{1/2}), \tag{A4}
\]
\[
\| \hat{\Sigma}_{Y,sp}^{-1} - \Sigma_Y^{-1} \|_F = O_p(\{(r_n + s_2) \log r_n/n\}^{1/2}). \tag{A5}
\]
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
\[
\max_{i,j} |\hat{\Sigma}_{Y,ij} - \Sigma_{Y,ij}| \leq C_Y \{(r_n + s_1) \log r_n/n\}^{1/2}, \quad \max_{i,j} |\hat{\Sigma}_{res,ij} - \Sigma_{ij}| \leq C_{res} \{(r_n + s_2) \log r_n/n\}^{1/2} \tag{A6}
\]
for some \( C_Y > 0, C_{res} > 0 \), (A4) and (A5) hold.

We begin by showing (A6). Let \( W \) be an \( m \)-dimensional random vector with mean \( \mu_W \) and covariance matrix \( \Sigma_W \), 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 \( \bar{W} = \frac{1}{n} \sum_{i=1}^n W_i \) and
\[
\hat{\Sigma}_W = \frac{1}{n} \sum_{k=1}^n (W_k - \bar{W})(W_k - \bar{W})^T = \frac{1}{n} \sum_{k=1}^n (W_k - \mu_W)(W_k - \mu_W)^T - (\bar{W} - \mu_W)(\bar{W} - \mu_W)^T.
\]
Sparse Envelope Model

From Ravikumar et al. (2011), there exists positive constants $C_\epsilon$, such that for $\delta \in (0, b_1)$,
\[
\Pr(|\hat{\Sigma}_{W,ij} - \Sigma_{W,ij}| > \delta) \leq \Pr \left[ \left\{ \frac{1}{n} \sum_{k=1}^{n} (W_k - \mu_W)(W_k - \mu_W)^T \right\}_{ij} - \Sigma_{W,ij} > \frac{\delta}{2} \right] \\
+ \Pr \left[ \left\{ (\hat{W} - \mu_W)(\hat{W} - \mu_W)^T \right\}_{ij} > \frac{\delta}{2} \right] \\
\leq C_1 \exp(-C_2n\delta^2) + C_3 \exp(-C_4n\delta^2)
\]
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,
\[
\max_{i,j} |\hat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_6 \log(m/n)^{1/2},
\]
where $C_6$ is a positive number.

Now we take $W = (X^\top, \epsilon^\top)^\top$, then $W$ is a $p + r_n$ dimensional random vector with mean $(\mu_X^\top, 0^\top)^\top$, where the 0 is an $r_n$ dimensional vector. It has a block diagonal covariance matrix with diagonal blocks being $\Sigma_X$ and $\Sigma$. Then by the preceding conclusion, we can find constant $C_0$ such that $\max_{i,j} |\hat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_0 \log((r_n + p)/n)^{1/2}$. Since $p$ is fixed, we can find $C_0'$ such that $\max_{i,j} |\hat{\Sigma}_{W,ij} - \Sigma_{W,ij}| \leq C_0' \log((r_n)/n)^{1/2}$. Then
\[
\max_{i,j} |\hat{\Sigma}_{X,ij} - \Sigma_{X,ij}| \leq C_0' \log((r_n)/n)^{1/2},
\]
\[
\max_{i,j} |\hat{\Sigma}_{e,ij} - \Sigma_{e,ij}| \leq C_0' \log((r_n)/n)^{1/2},
\]
\[
\max_{i,j} |\hat{\Sigma}_{eX,ij} - \Sigma_{eX,ij}| \leq C_0' \log((r_n)/n)^{1/2}.
\]
Since $\hat{\Sigma}_Y = \beta\hat{\Sigma}_X\beta^\top + \hat{\Sigma}_e + \beta\hat{\Sigma}_{eX} + \hat{\Sigma}_{eX}\beta^\top$, we have
\[
\max_{i,j} |\hat{\Sigma}_{Y,ij} - \Sigma_{Y,ij}| \leq C_Y \log((r_n)/n)^{1/2},
\]
for some $C_Y > 0$.

As $\hat{\Sigma}_{res} = \hat{\Sigma}_e - \hat{\Sigma}_{eX}\Sigma_X^{-1}\hat{\Sigma}_{eX}$,
\[
\hat{\Sigma}_{res} - \Sigma = (\hat{\Sigma}_{eX} - \Sigma_{eX})\Sigma_X^{-1}\Sigma_X + \Sigma_{eX}(\hat{\Sigma}_X^{-1} - \Sigma_X^{-1})\Sigma_X + \Sigma_{eX}\Sigma_X^{-1}(\hat{\Sigma}_X - \Sigma_X) + (\hat{\Sigma}_e - \Sigma_e)(\hat{\Sigma}_X^{-1} - \Sigma_X^{-1})(\hat{\Sigma}_X - \Sigma_X) + \hat{\Sigma}_e - \Sigma.
\]
Using the fact that for $A \in \mathbb{R}^{d_1 \times d_2}$, $B \in \mathbb{R}^{d_2 \times d_3}$, $\|AB\|_{\text{max}} \leq d_2 \|A\|_{\text{max}} \|B\|_{\text{max}}$, where $\|\cdot\|_{\text{max}}$ is the matrix max norm, we have
\[
\max_{i,j} |\hat{\Sigma}_{res,ij} - \Sigma_{ij}| \leq C_{res} \log((r_n)/n)^{1/2},
\]
for some $C_{res} > 0$. Therefore (A4) and (A5) hold.

We denote the objective function in (7) as $f_{obj,2}$. Let $a_n = \{(r_n + s)\log r_n/n\}^{1/2}$. Theorem 5 holds if for arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant $C$, such that
\[
\lim_{n \to \infty} \Pr \left\{ \inf_{\Delta \in \mathcal{R}^{(r_n+s)\times n}, \|\Delta\|_p = C} f_{obj,2}(A + a_n\Delta) > f_{obj,2}(A) \right\} > 1 - \epsilon.
\]
Following the techniques and notations in the proof of Theorem 2, we expand $f_{\text{obj,2}}(A + a_n \Delta) - f_{\text{obj,2}}(A)$ and get

$$f_{\text{obj,2}}(A + a_n \Delta) - f_{\text{obj,2}}(A) \geq 2a_n \text{tr} \left[ (G_A^T \Sigma G_A)^{-1} G_A^T (\widehat{\Sigma}_{\text{res,sp}} - \Sigma) \Delta_s + \{(G_A^T \widehat{\Sigma}_{\text{res,sp}} G_A)^{-1} - (G_A^T \Sigma G_A)^{-1}\} G_A^T \Sigma \Delta_s \right]$$

$$+ \{(G_A^T \widehat{\Sigma}_{\text{res,sp}} G_A)^{-1} - (G_A^T \Sigma G_A)^{-1}\} G_A^T \Sigma \Delta_s + (G_A^T \Sigma \Sigma_{\text{sp}} G_A)^{-1} G_A^T (\widehat{\Sigma}_{\text{Y,sp}} - \Sigma_{\text{Y}})^{-1} \Delta_s$$

$$+ \{(G_A^T \widehat{\Sigma}_{\text{Y,sp}} G_A)^{-1} - (G_A^T \Sigma_{\text{Y}} G_A)^{-1}\} G_A^T \Sigma_{\text{Y}} \Delta_s$$

$$+ a_n^2 \left( \Omega^{-1} \Gamma_1^T \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_1^T \Delta_1 \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1, \Gamma_1 \right)$$

$$- 2(I_u + A^T A)^{-1} \Delta_1 \Gamma_0 \Gamma_0^{-1} \Delta_1, - \frac{1}{2} a_n (q - u) \lambda_{\text{max},n} \max_i \left( \|a_i\|^{-1}_2 \|\delta_i\|_2 \right) + o_p(a_n^2).$$

Notice that

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

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

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

and $\|\widehat{\Sigma}_{\text{res,sp}} - \Sigma\|_F = O_p((r_n + s_1) \log r_n / n)^{1/2}$. Then

$$\text{tr} \left[ (G_A^T \Sigma G_A)^{-1} G_A^T (\widehat{\Sigma}_{\text{res,sp}} - \Sigma) \Delta_s \right] \geq -\hat{k}^2 \|\Delta_s\|_F \|\widehat{\Sigma}_{\text{res,sp}} - \Sigma^{-1}\|_F \| (G_A^T \Sigma G_A)^{-1} \| G_A\|_F.$$

Now

$$\left((G_A^T \widehat{\Sigma}_{\text{res,sp}} G_A)^{-1} - (G_A^T \Sigma G_A)^{-1}\right)$$

$$= - (G_A^T \Sigma G_A)^{-1} (G_A^T \widehat{\Sigma}_{\text{res,sp}} G_A - G_A^T \Sigma G_A) (G_A^T \Sigma G_A)^{-1} + o_p((r_n + s_1) \log r_n / n)^{1/2}$$

$$= - (G_A^T \Sigma G_A)^{-1} G_A (G_A^T \Sigma G_A)^{-1} + o_p((r_n + s_1) \log r_n / n)^{1/2}$$

$$= - (G_A^T \Sigma G_A)^{-1} G_A (\widehat{\Sigma}_{\text{res,sp}} - \Sigma G_A (G_A^T \Sigma G_A)^{-1} + o_p((r_n + s_1) \log r_n / n)^{1/2}$$

$$\text{tr} \left[ (G_A^T \widehat{\Sigma}_{\text{res,sp}} G_A)^{-1} - (G_A^T \Sigma G_A)^{-1}\right] G_A^T \Sigma \Delta_s$$

$$\geq -u^{1/2} \hat{k}^2 \|\Delta_s\|_F \|\widehat{\Sigma}_{\text{res,sp}} - \Sigma^{-1}\|_F \| (G_A^T \Sigma G_A)^{-1} \| G_A\|_F.$$

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

$$f_{\text{obj,2}}(A + a_n \Delta) - f_{\text{obj,2}}(A) \geq 2M_1 a_n \|\Delta\|_F \|\widehat{\Sigma}_{\text{res,sp}} - \Sigma^{-1}\|_F + 2M_2 a_n \|\Delta\|_F \|\widehat{\Sigma}_{\text{Y,sp}} - \Sigma_{\text{Y}}^{-1}\|_F$$

$$+ a_n^2 \left( \Omega^{-1} \Gamma_1^T \Gamma_0 \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_1^T \Delta_1 \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1, \Gamma_1 \right)$$

$$- 2(I_u + A^T A)^{-1} \Delta_1 \Gamma_0 \Gamma_0^{-1} \Delta_1, - \frac{1}{2} a_n (q - u) \lambda_{\text{max},n} \max_i \left( \|a_i\|^{-1}_2 \|\delta_i\|_2 \right) + o_p(a_n^2),$$

where $M_1 = -2u^{1/2} \hat{k}^2 \|G_A^T \Sigma G_A\|^{-1} \|G_A\|_F$ and $M_2 = -2\|G_A^T \Sigma_{\text{Y}}^{-1} G_A\|^{-1} \|G_A\|_F$. Because

$$\lambda_{\text{max},n} = o((r_n + s) \log r_n / n)^{1/2} = o_p(a_n)$$

and

$$\text{tr} \left[ \Omega^{-1} \Gamma_1^T \Gamma_0 \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1 + (\Omega + \eta \Sigma X \eta^T) \Gamma_1^T \Delta_1 \Gamma_0 \Gamma_0 \Gamma_0^{-1} \Delta_1, \Gamma_1 \right] - 2(I_u + A^T A)^{-1} \Delta_1 \Gamma_0 \Gamma_0^{-1} \Delta_1$$

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

Since
\[
\hat{\beta}_{\text{ols}} - \beta = \Sigma_{eX} \Sigma_{\bar{X}}^{-1} - \Sigma_{eX} \Sigma_{X}^{-1} = (\Sigma_{eX} - \Sigma_{eX}) \Sigma_{\bar{X}}^{-1} + \Sigma_{eX} (\Sigma_{X}^{-1} - \Sigma_{\bar{X}}^{-1}),
\]
there exists a constant $C_{\text{ols}}$ such that
\[
\max_{i,j} |\hat{\beta}_{\text{ols},ij} - \beta_{ij}| \leq C_{\text{ols}} \{\log(r_n)/n\}^{1/2}.
\]
Because $\|\hat{\beta}_{\text{ols}} - \beta\|_F \leq (pr_n)^{1/2} \|\hat{\beta}_{\text{ols}} - \beta\|_{\text{max}}$, then
\[
\|\hat{\beta} - \beta\|_F \leq \|(P_T - P_T)\hat{\beta}_{\text{ols}}\|_F + \|P_T(\hat{\beta}_{\text{ols}} - \beta)\|_F \leq \|(P_T - P_T)\hat{\beta}_{\text{ols}}\|_F + \|\hat{\beta}_{\text{ols}} - \beta\|_F.
\]
Therefore the sparse envelope estimator $\hat{\beta}$ converges to $\beta$ with rate $\{(r_n + s) \log r_n/n\}^{1/2}$. \square

Proof of Theorem 6. Let
\[
\delta = \min_{i=1,\ldots,q-u} \|a_i\|_2 > 0,
\]
then $\delta$ is the smallest norm of the non-sparse rows in $A$. Since $\|\hat{\beta} - \beta\|_F = O_p[(r_n + s) \log r_n/n]^{1/2}$, and $\{(r_n + s) \log r_n/n\}^{1/2} \to 0$, then $\|\hat{\beta} - \beta\|_F < \delta/2$ 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 $f_{\text{obj,2}}$ with respect to $a_i$ and evaluate at $\hat{a}_i$, we have
\[
-4e_i^T \hat{G}_A(I_u + \hat{A}^T \hat{A})^{-1} + 2e_i^T \Sigma_{\text{res,sp}} \hat{G}_A(\hat{G}_A^T \Sigma_{\text{res,sp}} \hat{G}_A)^{-1} + 2e_i^T \Sigma_{Y,sp} \hat{G}_A(\hat{G}_A^T \Sigma_{Y,sp} \hat{G}_A)^{-1}
+ \lambda_i \frac{\hat{a}_i^T \hat{a}_i}{\|\hat{a}_i\|_2} = 0.
\]
Because $-4e_i^T \hat{G}_A(I_u + A^T A)^{-1} + 2e_i^T \Sigma_A G_A(G_A^T \Sigma_A G_A)^{-1} + 2e_i^T \Sigma_{Y,sp} \Sigma_A G_A(G_A^T \Sigma_{Y,sp} \Sigma_A G_A)^{-1} = 0$, we have
\[
\| -4e_i^T \hat{G}_A(I_u + A^T A)^{-1} + 2e_i^T \Sigma_{\text{res,sp}} \hat{G}_A(\hat{G}_A^T \Sigma_{\text{res,sp}} \hat{G}_A)^{-1} + 2e_i^T \Sigma_{Y,sp} \hat{G}_A(\hat{G}_A^T \Sigma_{Y,sp} \hat{G}_A)^{-1} \|_F
= O_p[(r_n + s) \log r_n/n]^{1/2}.
\]
But
\[
\|\lambda_i \frac{\hat{a}_i^T \hat{a}_i}{\|\hat{a}_i\|_2}\|_F = \lambda_i \geq \lambda_{\min,n}.
\]
Since $\{(r_n + s) \log r_n/n\}^{1/2} = o(\lambda_{\min,n})$, this is a contradiction. Therefore we have $\text{pr}(\hat{a}_i = 0) \to 1$ for $i = q - u + 1, \ldots, r_n - u$. \square

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 B.1.** The loss function $L(a_i \mid \hat{A}_{-i})$ as defined in (9) has a bounded second derivative
\[
\frac{d^2}{da_i^2} L(a_i \mid \hat{A}_{-i}) \bigg|_{a_i = \hat{a}_i} \leq \{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\} I,
\]
where $I \in \mathbb{R}^{n \times n}$ 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 \( L(a_i \mid \tilde{A}_{-i}) \) in (9), i.e.,

\[
Q(a_i) = L(\tilde{a}_i \mid \tilde{A}_{-i}) + (a_i - \tilde{a}_i)^T \frac{d}{da_i} L(a_i \mid \tilde{A}_{-i}) \bigg | _{a_i = \tilde{a}_i} + 1/2 \delta_i (a_i - \tilde{a}_i)^T (a_i - \tilde{a}_i),
\]

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

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

\[
\frac{d^2}{da_i^2} L(a_i \mid \tilde{A}_{-i}) = -4T_1 + 2T_2 + 2T_3,
\]

where

\[
T_1 = \frac{(1 + a_i^T B_1 a_i) B_1 - 2B_1 a_i a_i^T B_1}{(1 + a_i^T B_1 a_i)^2},
\]

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

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

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

\[
T_1 = \frac{(1 + a_i^T B_1 a_i) B_1 - 2B_1 a_i a_i^T B_1}{(1 + a_i^T B_1 a_i)^2} = \frac{B_1^{1/2} \left\{ \left(1 + a_i^T B_1^{1/2} a_i^{1/2} \right) I - 2B_1^{1/2} a_i a_i^T B_1^{1/2} \right\} B_1^{1/2}}{(1 + a_i^T B_1 a_i)^2}.
\]

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

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

We now prove that

\[
-I \leq \frac{(1 + x^T x) I - 2xx^T}{(1 + x^T x)^2} \leq I.
\]

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

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

we have \((x^T x) I - xx^T \geq 0\). Then

\[
(1 + x^T x) I - 2xx^T \geq (1 + x^T x) I - 2(x^T x) I = (1 - x^T x) I \geq -(1 + x^T x) I.
\]

We also have

\[
(1 + x^T x) I - 2xx^T \leq (1 + x^T x) I.
\]

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

\[
-I \leq \frac{(1 + x^T x) I - 2xx^T}{(1 + x^T x)^2} \leq I.
\]

Therefore

\[
-\gamma_{\text{max}}(B_1) I \leq -B_1 \leq T_1 \leq B_1 \leq \gamma_{\text{max}}(B_1) I.
\]
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 $g(t) = L(a_i^* + td_i \mid \tilde{A}_i)$ such that

$$g(0) = L(a_i^* \mid \tilde{A}_i), \quad g(1) = L(a_i \mid \tilde{A}_i).$$

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

$$g(1) = g(0) + g'(0) + 1/2 g''(b).$$

By Lemma B1,

$$g''(b) = d_i^2 \frac{d^2}{da_i^2} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = a_i^* + bd_i} d_i \leq \{4\gamma_{\max}(B_1) + 2\gamma_{\max}(B_2) + 2\gamma_{\max}(B_3)\} d_i^2 d_i$$

$$\leq \delta_i d_i^2 d_i,$$

(7) 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 (7) strictly holds. Plugging (7) into (6) gives (1). □

Proof of Theorem 1. By Lemma C2, after updating $\tilde{a}_i$ using

$$\tilde{a}_{i,\text{new}} = \frac{1}{\delta_i} \left\{ \delta_i \tilde{a}_i - \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \right\} \left\{ 1 - \frac{\lambda \omega_i}{\| \delta_i \tilde{a}_i - \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \|_2} \right\},$$

(8) we have

$$L(\tilde{a}_{i,\text{new}} \mid \tilde{A}_i) + \lambda \omega_i \| \tilde{a}_{i,\text{new}} \|_2 \leq Q(\tilde{a}_{i,\text{new}}) + \lambda \omega_i \| \tilde{a}_{i,\text{new}} \|_2$$

$$\leq Q(\tilde{a}_i) + \lambda \omega_i \| \tilde{a}_i \|_2$$

$$= L(\tilde{a}_i) + \lambda \omega_i \| \tilde{a}_i \|_2.$$  

Moreover, if $\tilde{a}_{i,\text{new}} \neq \tilde{a}_i$, then the first inequality becomes

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

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 $\tilde{a}_{i,\text{new}} = \tilde{a}_i$ for all $i$, then by (8) we have

$$\tilde{a}_i = \frac{1}{\delta_i} \left\{ \delta_i \tilde{a}_i - \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \right\} \left\{ 1 - \frac{\lambda \omega_i}{\| \delta_i \tilde{a}_i - \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \|_2} \right\}$$

if

$$\| \delta_i \tilde{a}_i - \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \|_2 > \lambda \omega_i,$$

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

$$\frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} + \lambda \omega_i \| \tilde{a}_i \|_2 = 0, \quad \tilde{a}_i \neq 0,$$

$$\| \frac{d}{da_i} L(a_i \mid \tilde{A}_i) \bigg|_{a_i = \tilde{a}_i} \|_2 \leq \lambda \omega_i, \quad \tilde{a}_i = 0.$$
where $i = 1, \ldots, r - u$. 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 $\Gamma_A$ were \((2/q)^{1/2}, 0\)^T and the remaining $q/2$ rows were \(0, (2/q)^{1/2}\)^T. Then we used the structure in (5) to construct $\Gamma$ and $\Gamma_0$. 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 $\Omega_0$ was a block diagonal matrix with the upper left block being $25 \cdot I_{q-u}$ and lower right block being $4I_{r-q}$. The elements in $\eta$ were independent $N(0, 4^2)$ variates. The predictors $X$ were normally distributed with mean 0 and covariance matrix $\Sigma_X = \sigma_X^2 I_p$, where $\sigma_X^2 = 0.4$. We varied $p$ from 100 to 180. For each value of $p$, 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).](image)

### 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 $\sigma_X$, 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 $(\Gamma_A, \Gamma_A, 0)$ was obtained by orthogonalizing a $q$ by $q$ matrix of independent uniform $(0, 1)$ variates. Then we used the structure in (5) to construct $\Gamma$ and $\Gamma_0$. The elements in $\eta$ were taken to be independent normal variates with mean 0 and variance $0.16$. The error covariance matrix $\Sigma$ followed the structure $\Sigma = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$, where $\Omega = I_u$ and $\Omega_0$ 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 $\sigma_X^2 = 0.4$. We varied $p$ from 100 to 180. For each value of

...
Sparse Envelope Model

$p$, 200 replications were generated. The selection performance is summarized in Table 1. The standard deviation of a randomly chosen element in $\beta$ 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>98.8</td>
<td>99.7</td>
<td>85.0</td>
<td>90.4</td>
<td>99.8</td>
<td>66.0</td>
<td>59.8</td>
<td>99.9</td>
<td>0.0</td>
</tr>
<tr>
<td>120</td>
<td>98.6</td>
<td>99.6</td>
<td>81.0</td>
<td>90.4</td>
<td>99.7</td>
<td>64.0</td>
<td>60.6</td>
<td>99.9</td>
<td>1.0</td>
</tr>
<tr>
<td>140</td>
<td>98.0</td>
<td>99.3</td>
<td>78.0</td>
<td>86.4</td>
<td>99.0</td>
<td>36.0</td>
<td>58.6</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>160</td>
<td>92.8</td>
<td>98.6</td>
<td>67.0</td>
<td>79.2</td>
<td>98.5</td>
<td>23.0</td>
<td>55.2</td>
<td>100.0</td>
<td>1.0</td>
</tr>
<tr>
<td>180</td>
<td>82.2</td>
<td>98.1</td>
<td>49.0</td>
<td>40.4</td>
<td>99.3</td>
<td>2.0</td>
<td>49.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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 $\sigma_X$ 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 $\sigma_X$ from 0.05 to 0.6. The selection performance is summarized in Table 2, and the standard deviation of a randomly chosen element in $\beta$ 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 $(\Gamma_A, \Gamma_A, 0)$ was obtained by orthogonalizing a $q \times q$ matrix of independent standard normal variates. Then we used the structure in (5) to construct $\Gamma$ and $\Gamma_0$. The elements in $\eta$ 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 $\Omega_0 = 25I_r - I_u$. 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 $\beta$ 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>54.6</td>
<td>96.1</td>
<td>0.0</td>
<td>25.8</td>
<td>98.7</td>
<td>0.0</td>
<td>2.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.1</td>
<td>70.6</td>
<td>96.2</td>
<td>10.0</td>
<td>46.4</td>
<td>98.1</td>
<td>5.0</td>
<td>17.6</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.2</td>
<td>85.2</td>
<td>97.7</td>
<td>39.0</td>
<td>65.6</td>
<td>98.1</td>
<td>14.0</td>
<td>30.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.3</td>
<td>87.8</td>
<td>97.9</td>
<td>48.0</td>
<td>72.6</td>
<td>98.1</td>
<td>20.0</td>
<td>44.8</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.4</td>
<td>92.8</td>
<td>98.6</td>
<td>67.0</td>
<td>79.4</td>
<td>98.5</td>
<td>23.0</td>
<td>55.2</td>
<td>100.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.5</td>
<td>98.2</td>
<td>99.8</td>
<td>93.0</td>
<td>89.8</td>
<td>99.5</td>
<td>54.0</td>
<td>61.8</td>
<td>100.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.6</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>98.0</td>
<td>99.9</td>
<td>88.0</td>
<td>65.0</td>
<td>100.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>35.8</td>
<td>99.9</td>
<td>0.0</td>
<td>20.8</td>
<td>100.0</td>
<td>0.0</td>
<td>4.1</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>72.6</td>
<td>99.9</td>
<td>0.0</td>
<td>54.7</td>
<td>100.0</td>
<td>0.0</td>
<td>20.5</td>
<td>99.9</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>95.9</td>
<td>100.0</td>
<td>27.5</td>
<td>88.2</td>
<td>100.0</td>
<td>0.0</td>
<td>65.1</td>
<td>99.7</td>
<td>0.0</td>
</tr>
<tr>
<td>15</td>
<td>99.9</td>
<td>100.0</td>
<td>98.8</td>
<td>98.3</td>
<td>100.0</td>
<td>58.8</td>
<td>94.8</td>
<td>99.7</td>
<td>21.2</td>
</tr>
<tr>
<td>20</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>58.8</td>
<td>100.0</td>
<td>100.0</td>
<td>77.5</td>
</tr>
</tbody>
</table>

D. THE SMALLEST LAMBDA THAT YIELDS THE NULL MODEL

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

$$\frac{d}{da_i} L(\alpha_i | \tilde{A} - \tilde{a}) \bigg|_{a_i = \tilde{a}_i, \| \tilde{a}_i \|_2 = 0} = 0, \quad \tilde{a}_i \neq 0,$$

$$\left\| \frac{d}{da_i} L(\alpha_i | \tilde{A} - \tilde{a}) \bigg|_{a_i = \tilde{a}_i} \right\|_2 \leq \lambda \omega_i, \quad \tilde{a}_i = 0,$$
Sparse Envelope Model

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

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

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 $M_{u+i|U} = M_{u+i,u+i} - M_{U,u+i} M_{U,u+i}^{-1} M_{U,u+i}$. Then after some straightforward calculations,

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

Therefore we have

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

**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 \( \hat{\beta} \) 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 \( \| \hat{\beta} - \beta \|_F = O_P((r_n + s) \log r_n/n)^{1/2} \), we plotted the average of \( [n/(r_n + s) \log r_n]^1/2 \| \hat{\beta} - \beta \|_F \) over 200 replications versus \( n \). The bootstrap estimator of \( \| \hat{\beta} - \beta \|_F \) is computed based on the average of 200 bootstrap samples. With each bootstrap sample, we obtained the sparse envelope estimator \( \hat{\beta}_{\text{boot}} \) and computed \( \| \hat{\beta}_{\text{boot}} - \beta \|_F \). Figure 6 indicates that \( \| \hat{\beta}_{\text{boot}} - \beta \|_F \) is a good approximation to \( \| \hat{\beta} - \beta \|_F \). Figure 6 also shows that \( \| \hat{\beta} - \beta \|_F \) is much smaller than \( \| \hat{\beta}_{\text{ols}} - \beta \|_F \). 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


[Received April 2012. Revised September 2012]
$A = \Gamma_2 \Gamma_1^{-1}$

$A_{-i}$ the submatrix of $A$ with row $a_i^T$ removed

$G$ the sub matrix of $G_A$ with row $a_i^T$ removed

$G_A$ $G_A = (I_u, A^T)^T \in \mathbb{R}^{r \times u}$

$L(A)$ loss function in the optimization of $A$

$L(A) = -2 \log |G_A^T G_A| + \log |G_A^T \hat{\Sigma}_{res} G_A| + \log |G_A^T \hat{\Sigma}_{-1} G_A|$

$Q(a_i)$ majorization function in the optimization of $a_i$

$X$ predictors

$Y$ responses

$Y_A$ active response, i.e., the responses having non-zero rows in $\Gamma$

$Y_I$ inactive response, i.e., the responses having zero rows in $\Gamma$

$Y_D$ dynamic response, i.e., the responses having non-zero rows in $\beta$

$Y_S$ static response, i.e., the responses having zero rows in $\beta$

$a_i$ the transpose of the $i$th 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 $r_n$

$r_A$ number of active responses

$r_I$ 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 $\Sigma^{-1}_{res}$

$s_2$ nonzero elements in the lower triangle (not including the diagonal elements) of $\Sigma^{-1}_{Y}$

$s = \max\{s_1, s_2\}$

$u$ dimension of the envelope $E_{\Sigma}(B)$

$\alpha$ intercept

$\beta$ regression coefficients

$\beta_A$ $\beta_A = \Gamma_A \eta$

$\beta_D$ the nonzero coefficients in $\beta$

$\hat{\beta}$ sparse envelope estimator of $\beta$

$\hat{\beta}_A$ sparse envelope estimator of $\beta_A$

$\hat{\beta}_{A,2}$ active envelope estimator of $\beta_A$

$\hat{\beta}_{A,O}$ oracle envelope estimator of $\beta_A$

$\hat{\beta}_{ols}$ ordinary least squares estimator of $\beta$

$B$ span of $\beta$

$E_{\Sigma}(B)$ the envelope subspace

$\eta$ coordinates of $\beta$ with respect to $\Gamma$

$G(r, u)$ $r \times u$ 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

$\Gamma$ orthogonal basis of the envelope $E_{\Sigma}(B)$

$\Gamma_A$ the non-zero rows in $\Gamma$

$\Gamma_{A,0}$ completion of $\Gamma_A$

$\Gamma_0$ orthogonal basis of the orthogonal complement of $E_{\Sigma}(B)$

$\tilde{\Gamma}_0$ orthogonal basis of the orthogonal complement of $E_{\Sigma}(B)$ with a block diagonal structure

$\Gamma_1$ the first $u$ rows in $\Gamma$

$\Gamma_2$ the last $r - u$ rows in $\Gamma$
Sparse Envelope Model

\( \hat{\Gamma} \) sparse envelope estimator of \( \Gamma \)
\( \hat{\Gamma}_A \) sparse envelope estimator of \( \Gamma_A \)
\( \hat{\Gamma}_{A,2} \) active envelope estimator of \( \Gamma_A \)
\( \hat{\Gamma}_{A,O} \) oracle envelope estimator of \( \Gamma_A \)
\( \lambda_i \) tuning parameter in the optimization of \( a_i \), \( \lambda_i \) can be written as \( \lambda_i = \lambda \omega_i \), where \( \lambda \) is the common tuning parameter and \( \omega_i \)'s are the weights. In this paper, \( \omega_i = 1/\|a_i\|_2^p \).
\( \lambda_{\max,n} \) \( \max(\lambda_1, \ldots, \lambda_{q-u}) \) at sample size \( n \)
\( \lambda_{\min,n} \) \( \min(\lambda_{q-u+1}, \ldots, \lambda_{r-u}) \) at sample size \( n \)
\( \Omega \) coordinates of \( \Sigma \) with respect to \( \Gamma \)
\( \Omega_0 \) coordinates of \( \Sigma \) with respect to \( \Gamma_0 \)
\( \tilde{\Omega}_{0,A} \) upper left block of \( \tilde{\Omega}_0 \), which has dimension \((q-u) \times (q-u)\)
\( \tilde{\Omega}_{0,A}^T \) upper right block of \( \tilde{\Omega}_0 \), which has dimension \((q-u) \times (r-q)\)
\( \tilde{\Omega}_{0,I} \) lower right block of \( \tilde{\Omega}_0 \), which has dimension \((r-q) \times (r-q)\)
\( \tilde{\Omega}_{0,A,I} \) lower left block of \( \tilde{\Omega}_0 \), which has dimension \((r-q) \times (q-u)\)
\( \tilde{\Sigma}_{X}^{-1} \) sparse permutation invariant covariance estimator of \( \Sigma_X^{-1} \)
\( \hat{\Sigma}_{Y|X} \) sample covariance matrix of the residuals from the regression of \( Y_A \) on \( X \)
\( (\hat{\Sigma}_{Y|X})_A \) the rows and columns in \( \hat{\Sigma}_{Y|X}^{-1} \) that have the same indices as \( Y_A \) in \( Y \)
\( \hat{\Sigma}_{\text{res}} \) sample covariance matrix of the residuals from the regression of \( Y_A \) on \( X \)
\( \hat{\Sigma}_{\text{res},sp}^{-1} \) sparse permutation invariant covariance estimator of \( \Sigma_{\text{res}}^{-1} \)
\( \varepsilon \) error vector
\( \otimes \) Kronecker product
\( \mathbb{I} \) \( V_1 \mathbb{I} V_2 \) 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
\( Q \) \( I - P \)
\( \text{vec} \) stack a matrix into a vector columnwise
\( \text{vech} \) 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,
\( \text{vech}(M) = C_a \text{vec}(M), \text{vec}(M) = E_a \text{vech}(M) \)