Supplementary Material for "Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models"

Shan Luo[1] AND Zehua Chen[2]

[1]Shanghai Jiao Tong University

[2]National University of Singapore

Email: [1]sluomath@sjtu.edu.cn, [2]stachenz@nus.edu.sg.

In this supplementary document, we provide the proofs for the theoretical results presented in Section 2.3 of the main article. First, the notations are reproduced below.

$\mathcal{Z}_0$ — the set of all relevant composite features;

$\mathcal{Z}_0^*$ — any subset of $\mathcal{Z}_0$;

s_0 — the set of all relevant simple features;

s_0^* — the set of relevant simple features contained in $\mathcal{Z}_0^*$;

$\Sigma_{\mathcal{Z}_{jk}}$ — the covariance matrix of $\mathcal{Z}_{jk} = (X_j, X_k, X_j X_k)$;

$\tilde{Y}(\mathcal{Z}_0^*) = Y - \alpha - \Sigma_{y,z_0^*}\Sigma_{z_0^*,z_0^*}^{-1}\mathcal{Z}_0^*$ — the residual of Y adjusting for the effects of $\mathcal{Z}_0^*$, that is, the difference between Y and its best linear predictor in terms of $\mathcal{Z}_0^*$.

$\Sigma_{y,z_{jk}}(\mathcal{Z}_0^*)$ — the covariance vector between $\tilde{Y}(\mathcal{Z}_0^*)$ and $\mathcal{Z}_{jk}$;

$R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk})$ — the multiple correlation coefficient between the $\tilde{Y}(z_0^*)$ and Z_{jk} given by $\Sigma_{yz_{jk}}(\mathcal{Z}_0^*)\Sigma_{\mathcal{Z}_{jk}}^{-1}\Sigma_{z_{jk}y}(\mathcal{Z}_0^*)$;

$r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) = (1/n)\tilde{\boldsymbol{y}}^\top(\mathcal{Z}_0^*)H(\mathcal{Z}_{jk})\tilde{\boldsymbol{y}}(\mathcal{Z}_0^*)$ — The sample version of $R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk})$.

For the sake of convenience, the theoretical results in Section 2.3 are restated before their proofs.

1 Proof of Lemma 1

Lemma 1. Assume the following conditions:

A1. $|s_0|^3 \ln p/n \to 0$.

A2. The eigenvalues of $\{\Sigma_{s,s} : |s| \leq 3|s_0|\}$ are bounded from below and above.

A3. Denote by Z_j's all the simple features. $\max_{j,l}\{\mathrm{E}\exp(t(Z_j - \mathrm{E}Z_j)(Z_l - \mathrm{E}Z_l), \mathrm{E}\exp(t\epsilon^2)\} \leq C$ for all $|t| \leq \eta$ for some constants η and C.

Suppose that $\mathcal{Z}_{j^*k^*}$ is the composite feature such that $R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{j^*k^*}) = \max_{(j,k)\in(\mathcal{Z}_0^*)^c} R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk})$. Then, as $n \to \infty$, uniformly for all $\mathcal{Z}_0^* \subset \mathcal{Z}_0$ with $|\mathcal{Z}_0^*| \leq 3|s_0|$, we have

$$P\left(r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{j^*k^*}) = \max_{(j,k)\in(Z_0^*)^c} r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk})\right) \to 1,$$

where $r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) = (1/n)\tilde{\boldsymbol{y}}^\top(Z_0^*)H(Z_{jk})\tilde{\boldsymbol{y}}(Z_0^*)$ with $\tilde{\boldsymbol{y}}(Z_0^*) = [I - H(Z_0^*)]\boldsymbol{y}$ is the sample version of $R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{j^*k^*})$.

Proof of Lemma 1. It suffices to show that, uniformly for $(j,k), \mathcal{Z}_0^* \subset \mathcal{Z}_0$, we have

$$r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) = R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk})(1 + o_p(1)). \tag{1}$$

Denote

$$\boldsymbol{a} = \Sigma_{Z_{jk}y}(\mathcal{Z}_0^*), \quad \boldsymbol{b} = n^{-1}Z_{jk}^\top\tilde{\boldsymbol{y}}(\mathcal{Z}_0^*), \quad A = \Sigma_{\mathcal{Z}_{jk}}, \quad B = n^{-1}Z_{jk}^\top Z_{jk}.$$

Then, we have

$$\begin{aligned} r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) - R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk}) &= \boldsymbol{b}^\top B^{-1}\boldsymbol{b} - \boldsymbol{a}^\top A^{-1}\boldsymbol{a} \\ &= \boldsymbol{b}^\top(B^{-1} - A^{-1})\boldsymbol{b} + (\boldsymbol{b} - \boldsymbol{a})^\top A^{-1}\boldsymbol{b} + \boldsymbol{a}^\top A^{-1}(\boldsymbol{b} - \boldsymbol{a}). \end{aligned}$$

Therefore,

$$\begin{aligned}
&|r(\tilde{\boldsymbol{y}}(Z_0^*), Z_{jk}) - R(\tilde{Y}(\mathcal{Z}_0^*), \mathcal{Z}_{jk})| \\
\leq\ & |\boldsymbol{b}^\top(B^{-1} - A^{-1})\boldsymbol{b}| + |(\boldsymbol{b}-\boldsymbol{a})^\top A^{-1}\boldsymbol{b}| + |\boldsymbol{a}^\top A^{-1}(\boldsymbol{b}-\boldsymbol{a})| \\
\leq\ & \|B^{-1} - A^{-1}\|\|\boldsymbol{b}\|_2^2 + \|A^{-1}\|\|\boldsymbol{b}-\boldsymbol{a}\|_2(\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2) \\
=\ & \|B^{-1} - A^{-1}\|\|\boldsymbol{b}\|_2^2 + \lambda_{\min}^{-1}(A)\|\boldsymbol{b}-\boldsymbol{a}\|_2(\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2) \\
\leq\ & 2\|B^{-1} - A^{-1}\|(\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}-\boldsymbol{a}\|_2^2) + \lambda_{\min}^{-1}(A)\|\boldsymbol{b}-\boldsymbol{a}\|_2(2\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}-\boldsymbol{a}\|_2),
\end{aligned}$$

where, for a matrix C, $\|C\|$ denotes its matrix norm, that is $\|C\| = \lambda_{\max}(C)$ if C is symmetric, and $\|C\| = \sqrt{\lambda_{\max}(C^\top C)}$, otherwise, for a vector $\boldsymbol{c}$, $\|\boldsymbol{c}\|_2$ denotes its L_2-norm.

Under condition A2, $\|\boldsymbol{a}\|_2$ and $\lambda_{\min}^{-1}(A)$ are both bounded, it remains to show that

$$\|B^{-1} - A^{-1}\| = o_p(1), \quad \text{uniformly for all} \quad (j,k). \tag{2}$$

and that

$$\|\boldsymbol{b} - \boldsymbol{a}\|_2 = o_p(1), \quad \text{uniformly for all} \quad (j,k) \text{ and } \mathcal{Z}_0^* \text{ with } |\mathcal{Z}_0^*| \leq 3|s_0|. \tag{3}$$

Write $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, we have

$$\begin{aligned}
\|A^{-1} - B^{-1}\| &\leq \|A^{-1}\|\|B-A\|\|B^{-1}\| = \lambda_{\max}(A^{-1})\lambda_{\max}(B^{-1})\|B-A\| \\
&\leq 3\lambda_{\min}^{-1}(A)\lambda_{\min}^{-1}(B)\|B-A\|_\infty,
\end{aligned}$$

where $\|\cdot\|_\infty$ denotes the maximum of the absolute values of the entries of a matrix. Under condition A1 and A3, the entries of B converges uniformly to the corresponding entries of A, c.f., Lemma 3.1 of Luo and Chen (2014). Hence $\lambda_{\min}^{-1}(B)$ is bounded since $\lambda_{\min}^{-1}(A)$ is bounded, and $\|B-A\|_\infty \to 0$ uniformly for all (j,k). This establishes (2).

Note that $\tilde{Y}(\mathcal{Z}_0^*) = Y - \alpha - \Sigma_{y,z_0^*}\Sigma_{z_0^*,z_0^*}^{-1}\mathcal{Z}_0^*$ and

$$\begin{aligned}
\boldsymbol{a} &= \mathrm{Cov}(\mathcal{Z}_{jk}, \tilde{Y}(\mathcal{Z}_0^*)) = \Sigma_{z_{jk},y} - \Sigma_{z_{jk},z_0^*}\Sigma_{z_0^*,z_0^*}^{-1}\Sigma_{z_0^*,y}, \\
\boldsymbol{b} &= \frac{Z_{jk}^\top \boldsymbol{y}}{n} - \frac{Z_{jk}^\top Z_0^*}{n}\left(\frac{Z_0^{*\top} Z_0^*}{n}\right)^{-1}\frac{Z_0^{*\top}\boldsymbol{y}}{n}.
\end{aligned}$$

Again, under condition A1 and A3, $\boldsymbol{b}$ converges to $\boldsymbol{a}$ uniformly, which establishes (3). The lemma then follows.

$\square$

2 Proof of Theorem 1

The additional conditions for Theorem 1 are as follows.

B1. As $n \to +\infty$,

$$\sqrt{n} \min_{j \in s_0} |\xi_j| / \sqrt{|s_0| \ln p} \to +\infty,$$

where ξ_j's are the coefficients of the simple features in s_0.

B2. For all $\mathcal{Z}_0^* \subset \mathcal{Z}_0$ with $|\mathcal{Z}_0^*| \leq 3|s_0|$, denote $s_z^\star$ as the set of relevant simple features contained in $\mathcal{Z}_0^*$ and $s_z^{\star -} = s_0 \backslash s_z^\star$. Define $\tilde{\mathcal{Z}}_0^\star = \{Z_{jk} : Z_{jk} \notin \mathcal{Z}_0^*,\ Z_{jk} \cap s_z^{\star -} \neq \emptyset\}$. There exists a $0 < q_1 < 1$ such that

$$\max_{(j,k): \mathcal{Z}_{jk} \notin \tilde{\mathcal{Z}}_0^\star} R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk}) < q_1 \max_{(j,k): \mathcal{Z}_{jk} \in \tilde{\mathcal{Z}}_0^\star} R(\tilde{Y}(\mathcal{Z}_0^*), Z_{jk}).$$

B3. There exists a $0 < q_2 < 1$, such that for any $s \subset s_0$,

$$\max_{j \in s_0^c} |(\Sigma_{js_0} - \Sigma_{js} \Sigma_{ss}^{-1} \Sigma_{ss_0}) \boldsymbol{\xi}| < q_2 \max_{j \in s^-} |(\Sigma_{js_0} - \Sigma_{js} \Sigma_{ss}^{-1} \Sigma_{ss_0}) \boldsymbol{\xi}|.$$

Theorem 1. Assume conditions A1 – A3 and B1 – B3. Let s^* be the selected set of simple features by the procedures of SIGS. Then, we have, as $n \to \infty$, $P\left(s^* = s_0\right) \to 1$.

Proof of Theorem 1.

(I) We first show that all relevant simple features are retained in the set $\mathcal{Z}_0^*$ selected at the first stage of our SIGS procedure, that is, $P(s_0 \subset \mathcal{Z}_0^*) \to 1$.

Let $\mathcal{Z}_1^*, \cdots, \mathcal{Z}_l^*, \cdots$ be the sequence of the sets of composite featurers selected by the SIGS procedure. By the nature of the sequential procedure, $\mathcal{Z}_1^* \subset \cdots \subset \mathcal{Z}_l^* \subset \cdots$. For the ease of notation, let s_l^* denote $s_{z_l}^*$, the set of simple features contained in $\mathcal{Z}_l^*$. Recall that by definition $\tilde{\mathcal{Z}}_l^\star = \{Z_{jk} : Z_{jk} \notin \mathcal{Z}_l^*,\ Z_{jk} \cap s_l^{\star -} \neq \emptyset\}$.

Under condition B2, Lemma 1 implies that, before all the relevant simple features are selected, at each step of the first stage of SIGS, at least one new relevant simple feature will be contained in the composite feature selected at that step. If the EBIC stopping rule is not activated, then there is an $l^* \leq |s_0|$ such that $s_0 \subset s_{l^*}^*$. In what follows, we show that the sequence will stop exactly at l^* by invoking the EBIC stopping rule with probability converging to 1. Let

$$\begin{aligned} D &= \mathrm{EBIC}(\mathcal{Z}_{l+1}^*) - \mathrm{EBIC}(\mathcal{Z}_l^*) \\ &= n \ln \frac{\|I - H(\mathcal{Z}_{l+1}^*)\boldsymbol{y}\|_2^2}{\|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2} + (|\mathcal{Z}_{l+1}^*| - |\mathcal{Z}_l^*|)\ln n + 2\gamma \left[\binom{N}{l+1} - \ln \binom{N}{l}\right] \\ &= n \ln \left(1 - \frac{\|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2 - \|I - H(\mathcal{Z}_{l+1}^*)\boldsymbol{y}\|_2^2}{\|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2}\right) \\ &\quad + (|\mathcal{Z}_{l+1}^*| - |\mathcal{Z}_l^*|)\ln n + 2\gamma \left[\binom{N}{l+1} - \ln \binom{N}{l}\right]. \end{aligned}$$

It suffices to show that if $\tilde{\mathcal{Z}}_l^\star \neq \emptyset$ then $D < 0$ and otherwise $D > 0$.

Case (i): $\tilde{\mathcal{Z}}_l^\star \neq \emptyset$

Note that under A2–A3, for any positive number ϵ,

$$\begin{aligned} P(\max_{|s| \leq 3|s_0|} \sup_{\|\boldsymbol{u}\|_2^2 = 1} |\boldsymbol{u}^\top(\hat{\Sigma}_{ss} - \Sigma_{ss})\boldsymbol{u}| > \epsilon) &\leq p^{6|s_0|} P(\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq \frac{\epsilon}{3|s_0|}) \\ &= C_0 \exp\{(4 + 6|s_0|\ln p) - C_1 n \frac{\epsilon^2}{9|s_0|^2}\}, \end{aligned} \tag{4}$$

where σ_{ij} is the covariance between two simple features and i, j range over all simple features. The right hand side of (4) converges to 0 under A1. Therefore, there exists positive constants c and d such that, with probability tending to 1,

$$c \leq \min_{(jk)} \lambda_{\min}(n^{-1} Z_{jk}^\top Z_{jk}) \leq \max_{(jk)} \lambda_{\max}(n^{-1} Z_{jk}^\top Z_{jk}) \leq d \tag{5}$$

$$c \leq \min_{\mathcal{Z}_0^* \subset \mathcal{Z}_0, |\mathcal{Z}_0^*| \leq 3|s_0|} \lambda_{\min}(n^{-1} Z_0^{*\top} Z_0^*) \leq \max_{\mathcal{Z}_0^* \subset \mathcal{Z}_0, |\mathcal{Z}_0^*| \leq 3|s_0|} \lambda_{\max}(n^{-1} Z_0^{*\top} Z_0^*) \leq d. \tag{6}$$

Here and after, all claims are made in the asymptotic sense that the probability of the claims converges to 1 as n goes to infinity. Now we focus on analysing the first term in D. Let Z^* be the composite feature selected at step $l+1$, that is, $Z_{l+1}^* = \begin{pmatrix} Z_l^* & Z^* \end{pmatrix}$. By using the formula for the inverse of partioned four-block matrix, we have the the following identity,

$$\begin{aligned} & \|[I - H(Z_l^*)]\boldsymbol{y}\|_2^2 - \|[I - H(Z_{l+1}^*)]\boldsymbol{y}\|_2^2 \\ = & \|[I - H(Z_l^*)]\boldsymbol{y}\|_2^2 - \|[I - H(\begin{pmatrix} Z_l^* & Z^* \end{pmatrix})]\boldsymbol{y}\|_2^2 \\ = & \boldsymbol{y}^\top [I - H(Z_l^*)] Z^* \left[Z^{*\top} [I - H(Z_l^*)] Z^* \right]^{-1} Z^{*\top} [I - H(\mathcal{Z}_l^*)]\boldsymbol{y}. \end{aligned}$$

Write

$$[I - H(Z_l^*)] Z^* \boldsymbol{u} = Z^* \boldsymbol{u} + Z_l^* \boldsymbol{v} = \begin{pmatrix} Z_l^* & Z^* \end{pmatrix} \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{u} \end{pmatrix} = Z_{l+1}^* \boldsymbol{\eta},$$

where $\boldsymbol{v} = -(Z_l^{*\top} Z_l^*)^{-1} Z_l^{*\top} Z^* \boldsymbol{u}$ and $\|\boldsymbol{\eta}\|_2 \geq \|\boldsymbol{u}\|_2$. For any $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_2 = 1$, we have

$$\begin{aligned} \|[I - H(\mathcal{Z}_l^*)] Z^* \boldsymbol{u}\|_2^2 & \geq \inf_{\boldsymbol{\eta}} \frac{\|Z_{l+1}^* \boldsymbol{\eta}\|_2^2}{\|\boldsymbol{\eta}\|_2^2} \geq \lambda_{\min}(Z_{l+1}^{*\top} Z_{l+1}^*) \geq nc, \\ \|[I - H(Z_l^*)] Z^* \boldsymbol{u}\|_2^2 & \leq \|Z^* \boldsymbol{u}\|_2^2 \leq \lambda_{\max}(Z^{*\top} Z^*) \leq nd. \end{aligned}$$

Therefore, the eigenvalues of $\left[Z^{*\top} [I - H(Z_l^*)] Z^* \right]^{-1}$ is within $[1/(nd), 1/(nc)]$. Consequently,

$$\|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2 - \|[I - H(\mathcal{Z}_{l+1}^*)]\boldsymbol{y}\|_2^2 \geq n^{-1} d^{-1} \|\boldsymbol{y}^\top [I - H(\mathcal{Z}_l^*)] Z^*\|_2^2.$$

Let $Z^{**} \in \mathcal{Z}_l^{*c}$ be the composite feature which contains the simple feature $\boldsymbol{z} \in s_0 \cap \mathcal{Z}_l^{*c}$ such that $[\boldsymbol{y}^\top (I - H(\mathcal{Z}_l^*))\boldsymbol{z}]^2 = \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} [\boldsymbol{y}^\top (I - H(\mathcal{Z}_l^*))\boldsymbol{z}_j]^2$. Then, we have

$$\begin{aligned} n^{-2} c^{-1} \|\boldsymbol{y}^\top [I - H(\mathcal{Z}_l^*)] Z^*\|_2^2] & \geq r(\tilde{\boldsymbol{y}}(\mathcal{Z}_l^*), Z^*) \geq r(\tilde{\boldsymbol{y}}(\mathcal{Z}_l^*), Z^{**}) \geq n^{-2} d^{-1} \|\boldsymbol{y}^\top [I - H(\mathcal{Z}_l^*)] Z^{**}\|_2^2 \\ & \geq n^{-2} d^{-1} \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} [\boldsymbol{y}^\top (I - H(\mathcal{Z}_l^*))\boldsymbol{z}_j]^2. \end{aligned}$$

Hence,

$$\begin{aligned} & \|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2 - \|[I - H(\mathcal{Z}_{l+1}^*)]\boldsymbol{y}\|_2^2 \\ \geq & n^{-1} c d^{-2} \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} [\boldsymbol{y}^\top (I - H(\mathcal{Z}_l^*))\boldsymbol{z}_j]^2 \\ = & n^{-1} c d^{-2} \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} [\boldsymbol{\mu}^\top (I - H(\mathcal{Z}_l^*))\boldsymbol{z}_j]^2 [1 + o_p(1)], \end{aligned} \tag{7}$$

where $\boldsymbol{\mu} = E\boldsymbol{y} = Z\boldsymbol{\xi}$. We now derive a lower bound for the rightmost side of (7). We can write

$$\begin{aligned}
\boldsymbol{\mu}^\top[I - H(\mathcal{Z}_l^*)]\boldsymbol{\mu} &= \sum_{j \in s_0 \cap (\mathcal{Z}_l^*)^c} \xi_j \boldsymbol{z}_j^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{\mu} \\
&\leq \sum_{j \in s_0 \cap \mathcal{Z}_l^{*c}} |\xi_j| |\boldsymbol{z}_j^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{\mu}| \\
&\leq \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} |\boldsymbol{\mu}^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{z}_j| \sum_{j \in s_0 \cap (\mathcal{Z}_l^*)^c} |\xi_j| \\
&= \max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} |\boldsymbol{\mu}^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{z}_j| \ \|\boldsymbol{\xi}_{s_0 \cap \mathcal{Z}_l^{*c}}\|_1.
\end{aligned}$$

On the other hand, denoting $\tilde{Z} = Z_{s_0 \cap \mathcal{Z}_l^{*c}}$ and $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}_{s_0 \cap \mathcal{Z}_l^{*c}}$, we have

$$\begin{aligned}
\boldsymbol{\mu}^\top[I - H(\mathcal{Z}_l^*)]\boldsymbol{\mu} &= \|[I - H(\mathcal{Z}_l^*)]\boldsymbol{\mu}\|_2^2 = \|[I - H(\mathcal{Z}_l^*)]\tilde{Z}\tilde{\boldsymbol{\xi}}\|_2^2 \\
&= \left\| (\tilde{Z} \quad Z_l^*) \begin{pmatrix} \tilde{\boldsymbol{\xi}} \\ -(Z_l^{*\top} Z_l^*)^{-1} Z_l^{*\top} \tilde{Z}\tilde{\boldsymbol{\xi}} \end{pmatrix} \right\|_2^2 \\
&= \begin{pmatrix} \tilde{\boldsymbol{\xi}} \\ -(Z_l^{*\top} Z_l^*)^{-1} Z_l^{*\top} \tilde{Z}\tilde{\boldsymbol{\xi}} \end{pmatrix}^\top \widetilde{Z}_l^\top \widetilde{Z}_l \begin{pmatrix} \tilde{\boldsymbol{\xi}} \\ -(Z_l^{*\top} Z_l^*)^{-1} Z_l^{*\top} \tilde{Z}\tilde{\boldsymbol{\xi}} \end{pmatrix} \\
&\geq \lambda_{\min}(\widetilde{Z}_l^\top \widetilde{Z}_l)\|\tilde{\boldsymbol{\xi}}\|_2^2,
\end{aligned}$$

where $\widetilde{Z}_l = (\tilde{Z} \quad Z_l^*)$. Thus, we have

$$\max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} |\boldsymbol{\mu}^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{z}_j| \ \|\tilde{\boldsymbol{\xi}}\|_1 \geq \lambda_{\min}(\widetilde{Z}_l^\top \widetilde{Z}_l)\|\tilde{\boldsymbol{\xi}}\|_2^2 \tag{8}$$

Since $\|\tilde{\boldsymbol{\xi}}\|_2^2 = \sum_j |\xi_j|^2 \geq \sum_j |\xi_j| \min_j |\xi_j| = \|\tilde{\boldsymbol{\xi}}\|_1 \min_j |\xi_j|$, we have

$$\max_{j \in s_0 \cap \mathcal{Z}_l^{*c}} |\boldsymbol{\mu}^\top [I - H(\mathcal{Z}_l^*)]\boldsymbol{z}_j| \geq \lambda_{\min}(\widetilde{\mathcal{Z}}_l^\top \widetilde{\mathcal{Z}}_l) \min_{j \in s_0} |\xi_j|. \tag{9}$$

Hence,

$$\begin{aligned}
\frac{\|[I - H(Z_l^*)]\boldsymbol{y}\|_2^2 - \|[I - H(Z_{l+1}^*)]\boldsymbol{y}\|_2^2}{\|[I - H(\mathcal{Z}_l^*)]\boldsymbol{y}\|_2^2} &\geq n^{-1} c d^{-2} \frac{\left[\lambda_{\min}(\widetilde{\mathcal{Z}}_l^\top \widetilde{\mathcal{Z}}_l) \min_{j \in s_0} |\xi_j|\right]^2}{\lambda_{\max}(Z_{s_0}^\top Z_{s_0})\|\boldsymbol{\xi}\|_2^2}[1 + o_p(1)] \\
&\geq c^3 d^{-3} (\max_{j \in s_0} |\xi_j|)^{-2} |s_0|^{-1} \min_{j \in s_0} |\xi_j|^2 [1 + o_p(1)].
\end{aligned}$$

Denote the rightmost side of the above inequality by R_l. Therefore, the first term in D is smaller than $-nR_l$. The remaing term in D is less than $3\ln n + 2\gamma \ln N$ where $N = p(p-1)/2$. Hence, under condition B1, $D < 0$ as $n \to \infty$.

Case (ii): $\tilde{\mathcal{Z}}_l^\star = \emptyset$.

The proof for the selection consistency of EBIC in *Case II:* $s_{0n} \subset s$ of Luo and Chen (2013) can be followed verbatim to show that $D > 0$ if $\widehat{\mathcal{Z}_l^\star} = \emptyset$. Hence, the details are omitted.

(II) We now show that $P(s^* = s_0) \to 1$. We have shown in **(I)** that all relevant simple features are in the set of simple features contained in the composite features selected in the first stage with probability tending to 1, that is $P(s_0 \subset \mathcal{Z}_0^*) \to 1$. The second stage is in fact a SLasso procedure applied to the feature set $\mathcal{Z}_0^*$. Therefore, the selection consistency of the second stage can be proved in the same way as that of the SLasso procedure.

Alternatively, we can proceed as follows. Let $s_1^* \subset \cdots \subset s_l^* \subset \cdots$ be the sequence of the feature sets selected in the second stage of the SIGS procedure without activating the EBIC stopping rule. Under condition B3, the true set of relevant simple features s_0 is one of those sets. That is, the procedure selects all relevant simple features before any irrelevant simple features could be selected. Then following the proof in **(I)**, we can establish that, for any $s_l^\star \subsetneq s_0$,

$$\begin{aligned} \max_{j \in s_l^{\star -}} |\boldsymbol{\mu}^\top [I - H(s_l^\star)] \boldsymbol{z}_j| &\geq \lambda_{\min}(Z_{s_0}^\top Z_{s_0}) \min_{j \in s_0} |\xi_j| \\ \frac{\|[I - H(s_{l+1}^\star)]\boldsymbol{y}\|_2^2 - \|[I - H(s_l^\star)]\boldsymbol{y}\|_2^2}{\|[I - H(s_l^\star)]\boldsymbol{y}\|_2^2} &\geq c^3 d^{-3} (\max_{j \in s_0} |\xi_j|)^{-2} |s_0|^{-1} \min_{j \in s_0} |\xi_j|^2. \end{aligned}$$

Hence, $\text{EBIC}(s_{l+1}^*) - \text{EBIC}(s_l^*) < 0$ when $s_l^\star \subsetneq s_0$. Therefore, the sequence $\text{EBIC}(s_l^*)$ is decreasing until it reaches s_0. That if $s_l^\star = s_0$ then the above difference is greater than zero can be shown by following the proof in Luo and Chen (2013) as in *Case (ii)* of **(I)**.

□

References

Luo, S. and Z. Chen (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference 143*(3), 494–504.

Luo, S. and Z. Chen (2014). Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association 109* (507), 1229–1240.