

# Supplementary document to “Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space”

BY SHAN LUO<sup>1</sup> AND ZEHUA CHEN<sup>2</sup>

<sup>1</sup>Department of Mathematics

Shanghai Jiao Tong University

<sup>2</sup>Department of Statistics and Applied Probability

National University of Singapore

Email: <sup>1</sup>sluo2012@gmail.com, <sup>2</sup>stachenz@nus.edu.sg.

In this supplementary document, we provide (i) the proofs for Theorem 3.1, Lemma 3.1 and Theorem 3.3 in §3, (ii) the verification of the special cases in §3.4, and (iii) the details of the simulation settings in §4. To make the document self-contained to a certain extent, these results and their conditions are re-stated in this document.

## 1 Technical proofs

### 1.1 Theorem 3.1 and its proof

Assume the following conditions:

**A1**  $\max_{j \in s_0^c} |\gamma_n(j, s, \beta)| < q \max_{j \in s^-} |\gamma_n(j, s, \beta)|$ ,  $0 < q < 1$ .

**A2** (Partial positive cone condition). If  $s^- \neq \emptyset$ , let

$$\mathcal{A}_s = \{\tilde{j} : \tilde{j} \in s^-, |\gamma_n(\tilde{j}, s, \beta)| = \max_{j \in s^c} |\gamma_n(j, s, \beta)|\},$$

and  $\tilde{\mathbf{X}}(\mathcal{A}_s) = [I - \mathbf{H}(s)]\mathbf{X}(\mathcal{A}_s)$ . Then  $[\tilde{\mathbf{X}}^\tau(\mathcal{A}_s)\tilde{\mathbf{X}}(\mathcal{A}_s)]^{-1}\mathbf{1} > 0$ , where  $\mathbf{1}$  is the vector with all components 1.

**A3**  $\frac{\sqrt{n}}{\ln p_n} \lambda_{\min}[\frac{1}{n}\mathbf{X}^\tau(s_0)\mathbf{X}(s_0)] \min_{j \in s_0} |\beta_j| \rightarrow +\infty$ , as  $n \rightarrow \infty$ , where  $\lambda_{\min}$  denotes the smallest eigenvalue.

**Theorem 3.1** *Let  $s_{*1}, s_{*2}, \dots, s_{*k}, \dots$  be the sequence generated by the SLasso procedure. Suppose that assumptions A1-A3 hold. Let  $\ln p_n = O(n^\kappa)$ , where  $\kappa < 1/2$ . Then, there is a  $k^*$  such that*

$$Pr(s_{*k^*} = s_0) \rightarrow 1, \quad \text{as } n \rightarrow \infty,$$

where  $s_0$  is the exact index set of the relevant features.

*Proof.* By the KKT condition, at the  $(k+1)$ st step of the sequential Lasso, the solution  $\hat{\boldsymbol{\beta}}$  satisfies

$$2\tilde{\mathbf{X}}^\tau(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) = \lambda \partial \|\hat{\boldsymbol{\beta}}\|_1, \quad (1.1)$$

where  $\tilde{\mathbf{y}} = [\mathbf{I} - \mathbf{H}(s_{*k})]\mathbf{y}$ ,  $\tilde{\mathbf{X}} = [\mathbf{I} - \mathbf{H}(s_{*k})]\mathbf{X}(s_{*k}^c)$ , and  $\partial \|\hat{\boldsymbol{\beta}}\|_1$  is a sub gradient of  $\|\boldsymbol{\beta}\|_1$  at  $\hat{\boldsymbol{\beta}}$  whose components are 1, -1 or a number with absolute value less than or equal to 1 according as the components are positive, negative or zero.

For  $k = 0$ ,  $s_{*0}$  is taken as the empty set  $\phi$ . Obviously,  $s_{*0} \subset s_0$ . It suffices to show that, given  $s_{*k} \subset s_0$ ,  $P(s_{*k+1} \subset s_0) \rightarrow 1$ , uniformly for all  $k$  such that  $|s_{*k}| < |s_0|$ .

Let

$$\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) = \frac{1}{n} \mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \mathbf{y} = \gamma_n(j, s_{*k}, \boldsymbol{\beta}) + \frac{1}{n} \mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon}.$$

Define

$$\mathcal{A}_k = \{j : |\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})| = \max_{j \in s_{*k}^c} |\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})|\}.$$

We are going to show that, with probability converging to 1,  $\mathcal{A}_k \subset s_0$  and that  $\mathcal{A}_k$  is the set of non-zero elements of the solution to equation (1.1). We first show that

$\mathcal{A}_k \subset s_0$ , which is implied by  $|\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})| > \max_{l \in s_0^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\beta})|$  for  $j \in s_{*k}^-$  with probability converging to 1. The statement is established by showing

- (i)  $\frac{1}{n} \mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon} = O_p(n^{-1/2} \ln p)$  uniformly for all  $j \in s_{*k}^c$ .
- (ii) For  $j \in s_{*k}^-$ ,  $\max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| \geq C_n n^{-1/2} \ln p$  for  $C_n \rightarrow \infty$ .

Notice that  $\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon} \sim N(0, \sigma^2 \|\tilde{\mathbf{x}}_j\|_2^2)$  where  $\|\tilde{\mathbf{x}}_j\|_2^2 \leq \|\mathbf{x}_j\|_2^2 = n$ . Hence

$$\begin{aligned}
& P\left(\frac{1}{n} |\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon}| > \sigma n^{-1/2} \ln p\right) \\
&= P(|\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon}| > \sigma n^{1/2} \ln p) \\
&\leq P(|\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon}| > \sigma \|\tilde{\mathbf{x}}_j\|_2 \ln p) \\
&= P(|z| > \ln p) \leq \frac{2}{\ln p} \exp\left\{-\frac{(\ln p)^2}{2}\right\},
\end{aligned}$$

where  $z$  is a standard normal random variable. Thus, by Bonferroni inequality,

$$P\left(\max_{j \in s_{*k}^c} \frac{1}{n} |\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\epsilon}| > \sigma n^{-1/2} \ln p\right) \leq \frac{2}{\ln p} \exp\left\{-\frac{(\ln p)^2}{2} + \ln p\right\} \rightarrow 0. \quad (1.2)$$

Thus (i) is proved.

Let  $\Delta(s_{*k}) = \boldsymbol{\mu}^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\mu}$  where  $\boldsymbol{\mu} = \mathbf{X} \boldsymbol{\beta}$ . We have the following inequalities

$$\Delta(s_{*k}) = \sum_{j \in s_{*k}^-} \beta_j \mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \boldsymbol{\mu} \leq n \|\boldsymbol{\beta}(s_{*k}^-)\|_1 \max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})|, \quad (1.3)$$

and

$$\begin{aligned}
\Delta(s_{*k}) &= \boldsymbol{\beta}^\tau(s_{*k}^-) \mathbf{X}^\tau(s_{*k}^-) [\mathbf{I} - \mathbf{H}(s_{*k})] \mathbf{X}(s_{*k}^-) \boldsymbol{\beta}(s_{*k}^-) \\
&\geq \lambda_{\min}(\mathbf{X}^\tau(s_{*k}^-) [\mathbf{I} - \mathbf{H}(s_{*k})] \mathbf{X}(s_{*k}^-)) \|\boldsymbol{\beta}(s_{*k}^-)\|_2^2 \\
&\geq \lambda_{\min}(\mathbf{X}^\tau(s_0) \mathbf{X}(s_0)) \|\boldsymbol{\beta}(s_{*k}^-)\|_2^2.
\end{aligned} \quad (1.4)$$

The second inequality above follows since  $s_{*k} \cup s_{*k}^- = s_0$  and  $(\mathbf{X}^\tau(s_{*k}^-) [\mathbf{I} - \mathbf{H}(s_{*k})] \mathbf{X}(s_{*k}^-))^{-1}$  is a sub-matrix of  $(\mathbf{X}^\tau(s_0) \mathbf{X}(s_0))^{-1}$  by the formula of the inverse of blocked matrices.

Combining (1.3) and (1.4) yields

$$\begin{aligned}
\max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| &\geq \lambda_{\min}\left(\frac{1}{n} \mathbf{X}^\tau(s_0) \mathbf{X}(s_0)\right) \frac{\|\boldsymbol{\beta}(s_{*k}^-)\|_2^2}{\|\boldsymbol{\beta}(s_{*k}^-)\|_1} \\
&\geq \lambda_{\min}\left(\frac{1}{n} \mathbf{X}^\tau(s_0) \mathbf{X}(s_0)\right) \min_{j \in s_0} |\beta_j| \\
&\equiv C_n n^{-1/2} \ln p, \quad \text{say,}
\end{aligned}$$

with  $C_n = \frac{n^{1/2}}{\ln p} \lambda_{\min}\left(\frac{1}{n} \mathbf{X}^\tau(s_0) \mathbf{X}(s_0)\right) \min_{j \in s_0} |\beta_j|$ . The second inequality above holds since  $|s_{*k}^-| \|\boldsymbol{\beta}(s_{*k}^-)\|_2^2 \geq \|\boldsymbol{\beta}(s_{*k}^-)\|_1^2 \geq |s_{*k}^-| \min_{j \in s_0} |\beta_{0j}| \|\boldsymbol{\beta}(s_{*k}^-)\|_1$ .  $C_n \rightarrow \infty$  by A3. Thus (ii) is proved.

By A1 and (ii),

$$\begin{aligned}
&\left| \max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| - \max_{j \in s_0^c} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| \right| \\
&> (1 - q) \max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| \geq (1 - q) C_n n^{-1/2} \ln p.
\end{aligned}$$

This fact and (i) then imply that  $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})$  must attain the maximum within  $s_{*k}^-$ . Therefore,  $\mathcal{A}_k \subset s_{*k}^- \subset s_0$ .

Without loss of generality, assume that  $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) > 0$  for all  $j \in \mathcal{A}_k$ . Consider  $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi})$  as a function of  $\boldsymbol{\xi}$ . Since the function is continuous, for each  $j \in \mathcal{A}_k$ , there exist a neighborhood  $\mathcal{N}_j = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\beta}\|_2 \leq \delta_j\}$  and a constant  $c_j > 0$  such that, for all  $\boldsymbol{\xi} \in \mathcal{N}_j$ ,  $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi}) - \max_{l \in \mathcal{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\xi})| > c_j$ . Here  $\mathcal{A}_k^c$  denotes the complement of  $\mathcal{A}_k$  in  $s_{*k}^c$  by an abuse of notation. Let  $\mathcal{N} = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\beta}\|_2 \leq \delta\}$  where  $\delta = \min \delta_j$ . Then for all  $\boldsymbol{\xi} \in \mathcal{N}$ ,  $\min_{j \in \mathcal{A}_k} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi}) - \max_{l \in \mathcal{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\xi})| > C$ , where  $C = \max c_j$ .

Now construct  $\hat{\boldsymbol{\beta}}$  as follows. Let  $\hat{\boldsymbol{\beta}}(\mathcal{A}_k) = \omega [\tilde{\mathbf{X}}^\tau(\mathcal{A}_k) \tilde{\mathbf{X}}(\mathcal{A}_k)]^{-1} \mathbf{1}$  and  $\hat{\boldsymbol{\beta}}(\mathcal{A}_k^c) = 0$ , where  $\omega > 0$ . By A2,  $\hat{\boldsymbol{\beta}}(\mathcal{A}_k) > 0$ . Take  $\omega$  small enough such that  $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \in \mathcal{N}$ . Thus we have  $\min_{j \in \mathcal{A}_k} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) > \max_{l \in \mathcal{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|$ . On the other hand,

for any  $j \in \mathcal{A}_k$ ,

$$\begin{aligned}\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \omega \frac{1}{n} \tilde{\mathbf{X}}_j^\tau \tilde{\mathbf{X}}(\mathcal{A}_k) [\tilde{\mathbf{X}}^\tau(\mathcal{A}_k) \tilde{\mathbf{X}}(\mathcal{A}_k)]^{-1} \mathbf{1} \\ &= \max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \frac{\omega}{n}.\end{aligned}$$

Let  $\lambda = 2n[\max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \frac{\omega}{n}]$ . Then, we have

$$\begin{aligned}2\tilde{\mathbf{X}}_j^\tau(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) &= \lambda, \quad \text{for } j \in \mathcal{A}_k, \\ 2\tilde{\mathbf{X}}_j^\tau(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) &< \lambda, \quad \text{for } j \notin \mathcal{A}_k.\end{aligned}$$

Let  $\partial|\hat{\beta}_j| = 2\tilde{\mathbf{X}}_j^\tau(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})/\lambda$  for  $j \notin \mathcal{A}_k$ , and 1 for  $j \in \mathcal{A}_k$ . Then  $\partial\|\hat{\boldsymbol{\beta}}\|_1$  with these components is a sub gradient of  $\|\boldsymbol{\beta}\|_1$  at  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\beta}}$  solves equation (1.1). From the construction of  $\hat{\boldsymbol{\beta}}$ , all the features corresponding to the non-zero components of  $\hat{\boldsymbol{\beta}}$  belong to  $s_0$ . Hence  $s_{*k+1} \subset s_0$ . Thus we have shown that, given  $s_{*k} \subset s_0$ ,  $s_{*k+1} \subset s_0$  with probability converging to 1.

Since  $p_0 = |s_0|$  diverges as  $n \rightarrow \infty$ , we need to show that the above convergence is uniform for all  $k$  such that  $|s_{*k}| < p_0$ . Note that, under the assumptions,  $s_{*k+1} \subset s_0$  is equivalent to  $\min_{j \in \mathcal{A}_k} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) > \max_{l \in \mathcal{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\beta})|$  which is implied by  $P(\max_{j \in s_{*k}^c} \frac{1}{n} |\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \epsilon| > \sigma n^{-1/2} \ln p) \rightarrow 0$ . Therefore, when  $p_0$  is divergent, the uniform convergence is established if

$$P(\max_{0 \leq k < p_0} \max_{j \in s_{*k}^c} \frac{1}{n} |\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \epsilon| > \sigma n^{-1/2} \ln p) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

It follows from (1.2) and the Bonferroni inequality that

$$\begin{aligned}&P(\max_{0 \leq k < p_0} \max_{j \in s_{*k}^c} \frac{1}{n} |\mathbf{x}_j^\tau [\mathbf{I} - \mathbf{H}(s_{*k})] \epsilon| > \sigma n^{-1/2} \ln p) \\ &\leq \frac{2p_0}{\ln p} \exp\{-\frac{(\ln p)^2}{2} + \ln p\} \\ &\leq \frac{2}{\ln p} \exp\{-\frac{(\ln p)^2}{2} + 2 \ln p\} \rightarrow 0,\end{aligned}$$

since  $p_0 < p$ . The proof is completed.  $\square$

## 1.2 Lemma 3.1 and its proof

Assume that

**a1** The off-diagonal elements of  $\Sigma$  are bounded by a constant less than 1; that is, the correlation between any two features are bounded below from  $-1$  and above from  $1$ .

**a2**  $\sigma_{\max} \equiv \max_{1 \leq j, k \leq p} \sigma(z_j z_k) < \infty$  where  $\sigma(z_j z_k)$  denotes the standard deviation of  $z_j z_k$ .

**a3**  $\max_{1 \leq j, k \leq p} E \exp(t z_j z_k)$  and  $\max_{1 \leq j \leq p} E \exp(t z_j \epsilon)$  are finite for  $t$  in a neighborhood of zero.

**Lemma 3.1** Under assumptions *a1* - *a3*,

(i)  $P(\max_{1 \leq j, k \leq p} |\frac{1}{n} \sum_{i=1}^n x_{ij} x_{ik} - \Sigma_{jk}| > n^{-\frac{1}{3}} \sigma_{\max}) \rightarrow 0$ .

(ii)  $P(\max_{1 \leq j \leq p} |\frac{1}{n} \sum_{i=1}^n x_{ij} \epsilon_i| > n^{-\frac{1}{3}} \sigma) \rightarrow 0$ .

(iii) Let  $\Sigma_{jl|s} = \Sigma_{jl} - \Sigma_{js} \Sigma_{ss}^{-1} \Sigma_{sl}$  and  $\hat{\Sigma}_{jl|s} = \mathbf{x}_j^T [\mathbf{I} - \mathbf{H}(s)] \mathbf{x}_l / n$ . Then

$$\max_{1 \leq j, l \leq p_n} \max_{s: |s| \leq p_0} |\hat{\Sigma}_{jl|s} - \Sigma_{jl|s}| = o_p(1).$$

*Proof.* : For any  $j, k \in \{1, 2, \dots, p\}$  it follows from [1] that

$$P(|\sum_{i=1}^n x_{ij} x_{ik} - n \Sigma_{jk}| > \sqrt{n} \sigma(z_j z_k) \psi_n) \leq C[1 - \Phi(\psi_n)] \exp[\frac{\psi_n^3}{\sqrt{n}} \lambda(\frac{\psi_n}{\sqrt{n}})] \quad (1.5)$$

where  $C$  is a constant,  $\Phi(\cdot)$  is the cumulative distribution function of standard normal distribution,  $\lambda(\cdot)$  is the Cramer series for the distribution of  $z_j z_k$  which converges in a neighborhood of zero under assumption *a3*, and  $\psi_n$  is a sequence satisfying  $\psi_n = o(n^{1/2})$  and  $\psi_n \rightarrow \infty$ .

Now take  $\psi_n = n^{\frac{1}{6}-\delta}$  for  $0 < \delta < \frac{1}{6} - \frac{\kappa}{2}$ . Then  $\lambda(\frac{\psi_n}{\sqrt{n}})$  is bounded and  $\frac{\psi_n^3}{\sqrt{n}}$  goes to 0 as  $n$  converges to  $\infty$ . Thus (1.7) leads to

$$\begin{aligned} & P(|\sum_{i=1}^n x_{ij}x_{ik} - n\Sigma_{jk}| > n^{\frac{2}{3}-\delta}\sigma_{\max}) \\ & \leq P(|\sum_{i=1}^n x_{ij}x_{ik} - n\Sigma_{jk}| > n^{\frac{2}{3}-\delta}\sigma(z_j z_k)) \\ & \leq C_1[1 - \Phi(n^{\frac{1}{6}-\delta})] \\ & \leq \frac{C_1}{n^{\frac{1}{6}-\delta}} \exp(-\frac{1}{2}n^{\frac{1}{3}-2\delta}), \end{aligned}$$

where  $C_1$  is a generic constant. Let  $p = \exp(an^\kappa)$  where  $a > 0$  and  $\kappa < \frac{1}{3}$ . By Bonferroni inequality,

$$P(\max_{1 \leq j, k \leq p} \left| \sum_{i=1}^n x_{ij}x_{ik} - n\Sigma_{jk} \right| > n^{\frac{2}{3}-\delta}\sigma_{\max}) = o(n^{-\frac{1}{6}+\delta}) \rightarrow 0.$$

Hence (i) is proved. The proof of (ii) is similar and is omitted.

Note that, for  $\mathbf{x}_j$ ,  $\mathbf{x}_l$  and  $\mathbf{X}(s)$ ,  $\frac{1}{n}\mathbf{x}_j^\tau(I - \mathbf{X}(s)[X^\tau(s)\mathbf{X}(s)]^{-1}X^\tau(s))\mathbf{x}_l$  is a continuous function of the means  $\frac{1}{n}\sum_{i=1}^n x_{ij}x_{il}$ ,  $\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}$ ,  $\frac{1}{n}\sum_{i=1}^n x_{il}x_{ik}$  and  $\frac{1}{n}\sum_{i=1}^n x_{ik}x_{im}$ ,  $k, m \in s$ . Let  $\bar{\mathbf{X}}_{jls}$  denote the vector consisting of these means and  $\boldsymbol{\mu}_{jls}$  its expectation. The function depends on  $|s|$  but not on  $n$ . Let  $g_{|s|}(\bar{\mathbf{X}}_{jls})$  denote this function. We then have  $g_{|s|}(\boldsymbol{\mu}_{jls}) = \Sigma_{jl|s}$ .

By assumption a1, the range of  $\boldsymbol{\mu}_{jls}$  for all  $j, l, s$  with fixed  $|s|$  is compact. Hence  $g_{|s|}$  is also uniformly continuous for all  $(j, l, s)$  with fixed  $|s|$ . Thus for any  $\eta > 0$  there is a  $\zeta > 0$  such that if  $\|\bar{\mathbf{X}}_{jls} - \boldsymbol{\mu}_{jls}\|_\infty \leq \zeta$  then  $|g_{|s|}(\bar{\mathbf{X}}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| \leq \eta$ , where  $\zeta$  does not depend on  $(j, l, s)$ . From the proof of (i), we can choose a  $n_0$  such that when  $n > n_0$ ,

$$P(\max_{1 \leq j, k \leq p} \left| \frac{1}{n} \sum_{i=1}^n x_{ij}x_{ik} - \Sigma_{jk} \right| > \zeta) = o(n^{-\frac{1}{6}+\delta}).$$

Thus we have

$$P(\max_{j,l} |g_{|s|}(\bar{\mathbf{X}}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| > \eta) = o(n^{-\frac{1}{6}+\delta}).$$

By Bonferroni inequality,

$$P(\max_{j,l} \max_{s:|s|\leq p_0} |g_{|s|}(\bar{\mathbf{X}}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| > \eta) \leq o(n^{-\frac{1}{6}+\delta})p_0 \rightarrow 0,$$

for  $p_0 = O(n^{\frac{1}{6}-\delta})$ . (iii) is proved. □

### 1.3 Theorem 3.3 and its proof

**Theorem 3.3** Assume conditions A1 and A2. Suppose that  $\ln p_n = O(n^\kappa)$ ,  $\kappa < 1/3$ ,  $p_0 = O(n^c)$ ,  $c < 1/6$ , and there is a constant  $C$  such that  $\lambda_{\min}(\frac{1}{n}\mathbf{X}(s_0)^\tau \mathbf{X}(s_0)) \min_{j \in s_0} |\beta_j| \geq Cn^{-1/6+\delta}$ , where  $\delta$  is an arbitrarily small positive number. Let  $s_{*1} \subset s_{*2} \subset \dots \subset s_{*k} \subset \dots$  be the sets generated by the procedure of SLasso. Then

(i) Uniformly, for  $k$  such that  $|s_{*k}| < p_0$ ,

$$P(\text{EBIC}_\gamma(s_{*k+1}) < \text{EBIC}_\gamma(s_{*k})) \rightarrow 1, \text{ when } \gamma > 0.$$

(ii)  $P(\min_{p_0 < |s| \leq k_0} \text{EBIC}_\gamma(s) > \text{EBIC}_\gamma(s_0)) \rightarrow 1$ , when  $\gamma > 1 - \frac{\ln n}{2 \ln p}$ , where  $k_0 > p_0$  is an arbitrarily fixed integer.

Note that the additional conditions above imply condition A3. Result (ii) follows from the selection consistency of EBIC, see [2]. We only need to prove (i).

*Proof.* By Theorem 3.1,  $s_{*k} \subset s_0$  if  $|s_{*k}| \leq p_0$  with probability converging to 1. Let  $D_k = \text{EBIC}_\gamma(s_{*k}) - \text{EBIC}_\gamma(s_{*k+1})$ . Note that, under the assumption on  $p$  and  $p_0$ ,  $\ln \binom{p}{j} = j \ln p(1 + o(1))$ , uniformly for all  $j \leq p_0$ . Thus, we can replace  $\ln \binom{p}{j}$  by  $j \ln p$  in the definition of EBIC. Hence,

$$\begin{aligned} D_k &= n \ln \left( \frac{\|(\mathbf{I} - \mathbf{H}(s_{*k}))\mathbf{y}\|_2^2}{\|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2} \right) + (|s_{*k}| - |s_{*k+1}|)(\ln n + 2\gamma \ln p) \\ &= n \ln \left( 1 + \frac{\|(\mathbf{I} - \mathbf{H}(s_{*k}))\mathbf{y}\|_2^2 - \|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2}{\|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2} \right) - |\mathcal{A}_k|(\ln n + 2\gamma \ln p) \\ &= T_k - |\mathcal{A}_k|(\ln n + 2\gamma \ln p), \text{ say,} \end{aligned}$$



where  $\mathcal{A}_k = s_{*k+1}/s_{*k} = \{l : |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\beta})| = \max_{j \in s_{*k}^c} |\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})|\}$ . Without loss of generality, assume  $|\mathcal{A}_k| = 1$ . It suffices to show that

$$P(T_k \leq \ln n + 2\gamma \ln p) \rightarrow 0, \text{ uniformly for } 1 \leq k < p_0,$$

which is implied by

$$\frac{\min_{1 \leq k \leq p_0} T_k}{\ln p} \rightarrow \infty, \text{ in probability.} \quad (1.6)$$

It remains to prove (1.6). For any given set  $s$ , vectors  $\mathbf{u}$  and  $\mathbf{v}$ , let  $\Delta(s, \mathbf{u}) = \mathbf{u}^\tau [\mathbf{I} - \mathbf{H}(s)] \mathbf{u}$  and  $\Delta(s, \mathbf{u}, \mathbf{v}) = \mathbf{u}^\tau [\mathbf{I} - \mathbf{H}(s)] \mathbf{v}$ . Thus

$$\begin{aligned} & \|(\mathbf{I} - \mathbf{H}(s_{*k}))\mathbf{y}\|_2^2 - \|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2 \\ &= [\Delta(s_{*k}, \mu) - \Delta(s_{*k+1}, \mu)] + 2[\Delta(s_{*k}, \mu, \boldsymbol{\epsilon}) - \Delta(s_{*k+1}, \mu, \boldsymbol{\epsilon})] + [\Delta(s_{*k}, \boldsymbol{\epsilon}) - \Delta(s_{*k+1}, \boldsymbol{\epsilon})]. \\ & \|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2 \\ &= \Delta(s_{*k+1}, \mu) + 2\Delta(s_{*k+1}, \mu, \boldsymbol{\epsilon}) + \Delta(s_{*k+1}, \boldsymbol{\epsilon}). \end{aligned}$$

First we show that

$$\begin{aligned} & \|(\mathbf{I} - \mathbf{H}(s_{*k}))\mathbf{y}\|_2^2 - \|(\mathbf{I} - \mathbf{H}(s_{*k+1}))\mathbf{y}\|_2^2 \\ & \geq n(\lambda_{\min}[\frac{1}{n}\mathbf{X}(s_0)^\tau \mathbf{X}(s_0)] \min_{j \in s_0} |\beta_j|)^2 (1 + o_p(1)), \end{aligned} \quad (1.7)$$

uniformly for  $k$ . This inequality follows from

- (a)  $\max_{1 \leq k < p_0} [\Delta(s_{*k}, \boldsymbol{\epsilon}) - \Delta(s_{*k+1}, \boldsymbol{\epsilon})] = O_p(\ln p_0)$ ;
- (b)  $\min[\Delta(s_{*k}, \mu) - \Delta(s_{*k+1}, \mu)] \geq n(\lambda_{\min}[\frac{1}{n}\mathbf{X}(s_0)^\tau \mathbf{X}(s_0)] \min_{j \in s_0} |\beta_j|)^2$ ;
- (c)  $\Delta(s_{*k}, \mu, \boldsymbol{\epsilon}) - \Delta(s_{*k+1}, \mu, \boldsymbol{\epsilon}) = o_p(\Delta(s_{*k}, \mu) - \Delta(s_{*k+1}, \mu))$ , uniformly for  $k$ .

The claims (a), (b) and (c) are proved in the following.

Note that  $\Delta(s_{*k}, \epsilon) - \Delta(s_{*k+1}, \epsilon)$  follows a  $\chi^2$  distribution with degrees of freedom  $|\mathcal{A}_k|$ . Thus, by Bonferroni inequality, we have, for any  $a > 0$ ,

$$\begin{aligned} & P\left(\max_{1 \leq k < p_0} [\Delta(s_{*k}, \epsilon) - \Delta(s_{*k+1}, \epsilon)] \geq a\right) \\ & \leq p_0 P(\chi_1^2 \geq a) = 2p_0[1 - \Phi(\sqrt{a})] \leq \frac{C}{\sqrt{a}} \exp\left(-\frac{a}{2} + \ln p_0\right). \end{aligned}$$

Let  $a = 4 \ln p_0$ , then the above probability converges to zero, and thus (a) is proved.

By the relationship between  $\mathbf{I} - \mathbf{H}(s_{*k})$  and  $\mathbf{I} - \mathbf{H}(s_{*k+1})$ , we have

$$\begin{aligned} & \Delta(s_{*k}, \mu) - \Delta(s_{*k+1}, \mu) \\ & = \boldsymbol{\mu}^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \mathbf{X}(\mathcal{A}_k) [\mathbf{X}(\mathcal{A}_k)^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \mathbf{X}(\mathcal{A}_k)]^{-1} \mathbf{X}(\mathcal{A}_k)^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \boldsymbol{\mu} \\ & \geq \|\mathbf{X}(\mathcal{A}_k)^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \boldsymbol{\mu}\|_2^2 \lambda_{\max}^{-1}(\mathbf{X}(\mathcal{A}_k)^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \mathbf{X}(\mathcal{A}_k)) \\ & \geq |\mathcal{A}_k| n^2 \gamma_n^2(k^*, s_{*k}, \boldsymbol{\beta}) (n |\mathcal{A}_k|)^{-1} = n \gamma_n^2(k^*, s_{*k}, \boldsymbol{\beta}), \end{aligned}$$

where the last inequality holds since, for  $j \in \mathcal{A}_k$ ,  $\mathbf{x}_j^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \boldsymbol{\mu} = n \gamma_n(k^*, s_{*k}, \boldsymbol{\beta}) = \max_{j \in s_{*k}^c} |\mathbf{x}_j^\tau (\mathbf{I} - \mathbf{H}(s_{*k})) \boldsymbol{\mu}|$ . From (ii) in the proof of Theorem 3.1, for all  $k$ ,

$$\gamma_n(k^*, s_{*k}, \boldsymbol{\beta}) \geq \lambda_{\min}\left[\frac{1}{n} \mathbf{X}(s_0)^\tau \mathbf{X}(s_0)\right] \min_{j \in s_0} |\beta_j|.$$

Thus (b) is proved.

To show (c), note that  $\frac{\Delta(s_{*k}, \mu, \epsilon) - \Delta(s_{*k+1}, \mu, \epsilon)}{\sqrt{\Delta(s_{*k}, \mu) - \Delta(s_{*k+1}, \mu)}} = Z_k$  follows a standard normal distribution. By the same argument as in the proof of (a), we have  $\max_k |Z_k| = O_p(\ln p_0)^2$ . Thus (c) follows from (b).

Next we show that, for all  $k$ ,

$$\|(\mathbf{I} - \mathbf{H}(s_{*k+1})) \mathbf{y}\|_2^2 \leq C n p_0^2, \quad (1.8)$$

for some constant  $C$ . We have

$$\begin{aligned} & \|(\mathbf{I} - \mathbf{H}(s_{*k+1})) \mathbf{y}\|_2^2 \\ & = \Delta(s_{*k+1}, \mu) + \Delta(s_{*k+1}, \epsilon) + 2\Delta(s_{*k+1}, \mu, \epsilon) \\ & \leq \Delta(s_{*k+1}, \mu) + \Delta(s_{*k+1}, \epsilon) + 2\sqrt{\Delta(s_{*k+1}, \mu)\Delta(s_{*k+1}, \epsilon)}. \end{aligned} \quad (1.9)$$

Note that

$$\Delta(s_{*k+1}, \epsilon) \sim \chi_{n-k-1}^2 = (n-k-1)[1 + o_p(1)] \leq n[1 + o_p(1)]. \quad (1.10)$$

If  $k = p_0 - 1$ ,  $\Delta(s_{*k+1}, \mu) = 0$ , otherwise,

$$\begin{aligned} \Delta(s_{*k+1}, \mu) &= \boldsymbol{\beta}(s_{*k+1}^-)^\tau \mathbf{X}(s_{*k+1}^-)^\tau [\mathbf{I} - \mathbf{H}(s_{*k+1})] \mathbf{X}(s_{*k+1}^-) \boldsymbol{\beta}(s_{*k+1}^-) \\ &\leq \boldsymbol{\beta}(s_{*k+1}^-)^\tau \mathbf{X}(s_{*k+1}^-)^\tau \mathbf{X}(s_{*k+1}^-) \boldsymbol{\beta}(s_{*k+1}^-) \\ &\leq C \sum_{j,l \in s_{*k+1}^-} \mathbf{x}_j^\tau \mathbf{x}_l \leq (p_0 - k - 1)^2 C n \leq p_0^2 C n, \end{aligned} \quad (1.11)$$

where  $C$  is the upper bound of  $|\beta_j|^2$ . Thus (1.8) follows from (1.9), (1.10) and (1.11).

Combining (1.7) and (1.8), we have

$$\begin{aligned} \frac{\min_{1 \leq k < p_0} T_k}{\ln p} &\geq \frac{n}{\ln p} \ln \left( 1 + \frac{n[\lambda_{\min}(\frac{1}{n} \mathbf{X}(s_0)^\tau \mathbf{X}(s_0)) \min_{j \in s_0} |\beta_j|]^2 (1 + o_p(1))}{C n p_0^2} \right) \\ &\geq \frac{n}{2C p_0^2 \ln p} [\lambda_{\min}(\frac{1}{n} \mathbf{X}(s_0)^\tau \mathbf{X}(s_0)) \min_{j \in s_0} |\beta_j|]^2 (1 + o_p(1)) \\ &\rightarrow \infty, \end{aligned}$$

by the assumptions additional to A1 and A2. The last inequality holds since  $\ln(1+x) \geq \frac{x}{2}$  if  $0 \leq x \leq 1$ . In fact,

$$\frac{[\lambda_{\min}(\frac{1}{n} \mathbf{X}(s_0)^\tau \mathbf{X}(s_0)) \min_{j \in s_0} |\beta_j|]^2}{C p_0^2} \leq 1.$$

□

## 2 Special cases

**Special case I:** Let the correlation matrix of  $\mathbf{z}$  be given by

$$\Sigma = (1 - \rho)I + \rho \mathbf{1}\mathbf{1}^\tau,$$

where  $I$  is the identity matrix of dimension  $p$ ,  $\mathbf{1}$  is a  $p$ -vector of all elements 1, and  $0 < \rho \leq \rho_0 < 1$ . Note that  $\rho$  is allowed to depend on  $n$ . But for the ease of notation

we don't make this dependence explicit. In this case, the assumptions A1'-A3' are satisfied with  $\min_{j \in s_0} |\beta_j| = Cn^{-1/2+\delta}$  for some constant C and an arbitrarily small positive  $\delta$ . The claim is verified in the following.

For any  $s \subset S$ , the sub correlation matrix  $\Sigma_{ss}$  has eigenvalues  $1 - \rho$  and  $1 + (|s| - 1)\rho$  with multiplicities  $|s| - 1$  and 1 respectively. The eigenvector corresponding to  $1 + (|s| - 1)\rho$  is  $\mathbf{1}$  with dimension  $|s|$ . The smallest eigenvalue is  $1 - \rho$ . Thus A3' follows immediately.

Now suppose  $s \subset s_0$ . For any  $j, k \in s^c$ , we have

$$\begin{aligned} \Sigma_{jk} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sk} &= \Sigma_{jk} - \rho^2\mathbf{1}^\tau\Sigma_{ss}^{-1}\mathbf{1} = \Sigma_{jk} - \frac{\rho^2|s|}{1 + (|s| - 1)\rho} \\ &= \begin{cases} \frac{(1 - \rho)(\rho|s| + 1)}{1 + (|s| - 1)\rho} \equiv a, & \text{if } j = k \\ \frac{\rho(1 - \rho)}{1 + (|s| - 1)\rho} \equiv b, & \text{if } j \neq k. \end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} \gamma_n(j, s, \boldsymbol{\beta}) &= \sum_{k \in s^-} \beta_k (\Sigma_{jk} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sk}) \\ &= \begin{cases} (a - b)\beta_j + b \sum_{k \in s^-} \beta_k = b \sum_{k \in s^-} \beta_k + (1 - \rho)\beta_j, & \text{for } j \in s^-, \\ b \sum_{k \in s^-} \beta_k, & \text{for } j \in s_0^c. \end{cases} \end{aligned}$$

Thus

$$\max_{j \in s^-} |\gamma_n(j, s, \boldsymbol{\beta})| = \begin{cases} |b \sum_{k \in s^-} \beta_k| + (1 - \rho) \max_{j \in s^-} \beta_j & \text{if } \sum_{k \in s^-} \beta_k > 0, \\ |b \sum_{k \in s^-} \beta_k| + (1 - \rho) |\min_{j \in s^-} \beta_j| & \text{if } \sum_{k \in s^-} \beta_k < 0. \end{cases}$$

Obviously,  $\max_{j \in s^-} |\gamma_n(j, s, \boldsymbol{\beta})| > \max_{j \in s_0^c} |\gamma_n(j, s, \boldsymbol{\beta})|$  and hence A1' is satisfied. Finally, we have

$$\begin{aligned} &\Sigma_{\mathcal{A}_s \mathcal{A}_s} - \Sigma_{\mathcal{A}_s s} \Sigma_{ss}^{-1} \Sigma_{s \mathcal{A}_s} \\ &= (1 - \rho)I + \rho \mathbf{1} \mathbf{1}^\tau - \rho^2 \mathbf{1} \mathbf{1}^\tau \Sigma_{ss}^{-1} \mathbf{1} \mathbf{1}^\tau \\ &= (1 - \rho)I + \rho \mathbf{1} \mathbf{1}^\tau - \frac{\rho^2 |s|}{1 + (|s| - 1)\rho} \mathbf{1} \mathbf{1}^\tau \\ &= (1 - \rho)I + \frac{\rho(1 - \rho)}{1 + (|s| - 1)\rho} \mathbf{1} \mathbf{1}^\tau. \end{aligned}$$

Let  $\nu$  be the number of elements in  $\mathcal{A}_s$ . The eigenvalue of the above matrix corresponding to the eigenvector  $\mathbf{1}$  is

$$1 - \rho + \frac{\nu\rho(1 - \rho)}{1 + (|s| - 1)\rho} = a + (\nu - 1)b.$$

Hence

$$(\Sigma_{\mathcal{A}_s\mathcal{A}_s} - \Sigma_{\mathcal{A}_s s}\Sigma_{ss}^{-1}\Sigma_{s\mathcal{A}_s})^{-1}\mathbf{1} = \frac{1}{a + (\nu - 1)b}\mathbf{1} > 0,$$

i.e., A2' holds.

Note that, in the above argument, we only need  $\rho = \rho_n \leq \rho_0 < 1$ . But, for the *irrepresentability condition* to hold, the following restriction must be in place:

$$\rho_n < \frac{1}{1 + c|s_0|}$$

for some constant  $c$ , see [5]. If  $|s_0| \rightarrow \infty$ ,  $\rho_n$  must go to zero, i.e., eventually, all the features must be statistically uncorrelated.

**Special case II.** Without loss of generality, let  $s_0 = \{1, \dots, p_0\}$ . Assume that

- (i)  $|\beta_1| > |\beta_2| > \dots > |\beta_{p_0}| = Cn^{-1/2+\delta}$  for some constant  $C$  and an arbitrarily small positive  $\delta$ ;
- (ii) The correlation matrix  $\Sigma$  has the following structure:

$$\Sigma_{s_0 s_0} = I, \quad \Sigma_{j s_0} = \frac{1}{p_0} \text{sign} \beta(s_0)^\tau, \text{ for } j \in s_0^c.$$

Obviously,

$$\Sigma_{j s_0} \Sigma_{s_0 s_0}^{-1} \text{sign} \beta(s_0) = 1,$$

i.e., the *irrepresentability condition* does not hold. In the following, we show that conditions A1'-A3' hold. Let  $s_{*0} = \phi$ . Suppose  $s_{*k} = \{1, \dots, k\}$  for  $k < p_0$ . For any

$$j \in s_0^c,$$

$$\begin{aligned} \Gamma(j, s_{*k}, \boldsymbol{\beta}) &= [(\Sigma_{js_{*k}}, \Sigma_{js_{*k}^-}, \Sigma_{js_0^c}) - \Sigma_{js_{*k}} \Sigma_{s_{*k}s_{*k}}^{-1} (\Sigma_{s_{*k}s_{*k}}, \Sigma_{s_{*k}s_{*k}^-}, \Sigma_{s_{*k}s_0^c})] \begin{pmatrix} \boldsymbol{\beta}(s_{*k}) \\ \boldsymbol{\beta}(s_{*k}^-) \\ \boldsymbol{\beta}(s_0^c) \end{pmatrix} \\ &= \Sigma_{js_{*k}^-} \boldsymbol{\beta}(s_{*k}^-) = \sum_{j \in s_{*k}^-} |\beta_j|/p_0 < |\beta_{k+1}| = \Gamma(k+1, s_{*k}, \boldsymbol{\beta}) \\ &= \max_{j \in s_{*k}^-} |\Gamma(j, s_{*k}, \boldsymbol{\beta})|. \end{aligned}$$

Thus A1' is satisfied. The validity of A2' is obvious since  $\mathcal{A}_{s_{*k}}$  contains only one element for each  $k < p_0$ . A3' reduces to  $\frac{\sqrt{n}}{\ln p} \min_{j \in s_0} |\beta_j| \rightarrow \infty$  which holds obviously.

### 3 Covariance structure of the design matrix in the simulation studies

The covariance structures of the design matrix  $X$  for the settings in group A (GA) and group B (GB) are given as follows:

**GA1.** All the  $p$  features are generated as i.i.d. standard normal random variables.

**GA2.** The features have a power decay correlation structure, i.e.,  $\rho_{ij} = 0.5^{|i-j|}$ , for  $i, j = 1, \dots, p$ .  $s_0 = \{1, \dots, p_0\}$ .

**GA3.** The features  $X_1, \dots, X_p$  are determined as follows. Let  $Z_1, \dots, Z_p$  and  $W_1, \dots, W_{p_0}$  be i.i.d. standard normal random variables. Then

$$\mathbf{x}_j = \frac{Z_j + W_j}{\sqrt{2}}, \text{ for } j \in s_0; \quad \mathbf{x}_j = \frac{Z_j + \sum_{k \in s_0} Z_k}{\sqrt{1 + p_0}} \text{ for } j \notin s_0.$$

**GA4.** The relevant features have a constant pairwise correlation, i.e.,  $\rho_{ij} = 0.5$ , for  $i, j \in s_0$ . For  $j \notin s_0$ ,  $\mathbf{x}_j$  is generated as:

$$\mathbf{x}_j = \epsilon_j + \frac{\sum_{k \in s_0} X_k}{p_0},$$

where  $\epsilon_j$ 's are i.i.d. with distribution  $N(0, 0.08)$ . The variance 0.08 is chosen such that the second term, which is correlated with relevant features, dominates the variance of  $\mathbf{x}_j$ .

**GA5.** The set  $s_0$  is taken as  $\{1, 2, \dots, p_0\}$ . The features in  $s_0$  has the power decay correlation  $\rho_{ij} = 0.5^{|i-j|}$ . The irrelevant features are generated in the same way as in GA4.

**GB1.** The setting is taken from [3]. All the features have constant pair-wise correlation  $\rho_{ij} = 0.5$ .  $(n, p, p_0) = (100, 200, 15)$ .  $\sigma = 1.5$ . The coefficients of the relevant features are specified as  $|\beta_j| = 2.5$  for  $1 \leq j \leq 5$ , 1.5 for  $6 \leq j \leq 10$ ; 0.5 for  $11 \leq j \leq 15$ . The signs of the coefficients are determined as  $(-1)^{u_i}$  where the  $u_i$ 's are i.i.d. Bernoulli random variables with probability of success  $p = 0.5$ .

**GB2.** The setting is also taken from [3]. It is the same as in GB1 that  $(n, p, p_0) = (100, 200, 15)$  and  $\sigma = 1.5$ . But the covariance structure of the features is specified such that the partially orthogonality condition in [3] is satisfied. Specifically, while  $s_0$  is taken as  $\{1, \dots, 5, 11, \dots, 15, 21, \dots, 25\}$ , the correlations are specified as  $\rho_{ij} = 0.5^{|i-j|}$  if either both  $i$  and  $j$  are less than or equal to 25 or both  $i$  and  $j$  are bigger than or equal to 25,  $\rho_{ij} = 0$  otherwise. The coefficients are specified as  $|\beta_j| = 2.5$  for  $1 \leq j \leq 5$ , 1.5 for  $10 \leq j \leq 15$ ; 0.5 for  $21 \leq j \leq 25$ . The signs of the coefficients are determined in the same way as in GB1.

**GB3.** The setting is taken from [4].  $(n, p, p_0) = (100, 1000, 10)$  and  $\sigma = 1$ . The relevant features are generated as i.i.d. standard normal variables with coefficients  $(3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$ . The irrelevant features are generated

as

$$\mathbf{x}_j = 0.25Z_j + \sqrt{0.75} \sum_{k \in s_0} X_k, j \notin s_0,$$

where  $Z_j$ 's are i.i.d. standard normal and independent from the relevant features. In this setting, the condition for the selection consistency of SLasso is not satisfied.

## References

- [1] Fill, J.A.(1983). Convergence rates related to the strong law of large numbers. *Ann. Prob.* **11**, 123-142.
- [2] Luo, S. and Chen, Z. (2011). 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** 494-504.
- [3] Huang, J., Ma, S. and Zhang, C-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.
- [4] Ing, C-K. , and Lai, T.L (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, **21**, 1473-1513.
- [5] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research* **7**, 2541-2567.