J Syst Sci Complex (2020) 33: 510–526

# Sequential Feature Screening for Generalized Linear Models with Sparse Ultra-High Dimensional Data*

**ZHANG Junying · WANG Hang · ZHANG Riquan · ZHANG Jiajia**

DOI: 10.1007/s11424-020-8273-2
Received: 17 September 2018 / Revised: 25 January 2019


**Abstract** This paper considers the iterative sequential lasso (ISLasso) variable selection for generalized linear model with ultrahigh dimensional feature space. The ISLasso selects features by estimated parameter sequentially iteratively for the second order approximation of likelihood function where the features selected depend on regulatory parameters. The procedure stops when extended BIC (EBIC) reaches a minimum. Simulation study demonstrates that the new method is a desirable approach over other methods.

**Keywords** Extended BIC, generalized linear model, sequential lasso, sequential iteration, variable screening, variable selection.

## 1 Introduction

Data sets, in which the number of variables are comparable to or much larger than the sample size are frequently seen in many fields including genomics, health sciences, economics and machine learning. The data collected is usually of the type $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})_{i=1}^n$, where the $y_i'$s are independent observations of the response variable $Y$ given its covariates, or explanatory variables $(x_{i1}, x_{i2}, \cdots, x_{ip})$. Under sparsity conditions, feature selection becomes crucial in data analysis. Usually there are two goals of feature selection: (i) To build a model with desirable prediction properties and (ii) to identify the features with nonzero coefficients (for convenience,

ZHANG Junying · WANG Hang
*Department of Mathematics, Taiyuan University of Technology, Taiyuan* 030024*, China.*
Email: zhangjunying-zjy@163.com.
ZHANG Riquan
*Department of Statistics, East China Normal University, Shanghai* 200241*, China.*
Email: zhangriquan@163.com.
ZHANG Jiajia
*Department of Epidemiology and Biostatistics, University of South Carolina, Columbia* 20295*, USA.*
*The research was supported in part by the National Natural Science Foundation of China under Grant Nos. 11571112, 11501372, 11571148, 11471160, Doctoral Fund of Ministry of Education of China under Grant No. 20130076110004, Program of Shanghai Subject Chief Scientist under Grant No. 14XD1401600, and the 111 Project of China under Grant No. B14019.
◇ *This paper was recommended for publication by Editor DONG Yuexiao.*

such features are referred to as relevant features in this article). These two goals are intertwined but are not the same.

Generalized linear models (GLMs) provide a flexible parametric approach to estimating the covariate effects[1]:

$$f_n(Y;\boldsymbol{X},\beta)=\prod_{i=1}^{n}f_0(y_i;\theta_i)=\prod_{i=1}^{n}\exp[y_i\theta_i-b(\theta_i)], \tag{1}$$

where $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ is unknown $p$-dimensional vector of regression coefficients, $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$, $X_j=(X_{1j},X_{2j},\cdots,X_{nj})^{\mathrm{T}}$, $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$, $\{f_0(y_i;\theta_i):\theta_i\in R\}$ is a family of distributions in the regular exponential family with $\theta=(\theta_1,\theta_2,\cdots,\theta_n)^{\mathrm{T}}=\boldsymbol{X}\beta=(x_1^{\mathrm{T}}\beta,x_2^{\mathrm{T}}\beta,\cdots,x_n^{\mathrm{T}}\beta)$. $E(y_i|\boldsymbol{x}_i)=\mu_i=b'(\theta_i)$ and $b(\cdot)$ are some suitably chosen known functions. Thus, the log-likelihood function for $\beta$ is given by

$$l_n(\beta)=n^{-1}[Y^{\mathrm{T}}\boldsymbol{X}\beta-\mathbf{1}^{\mathrm{T}}b(\boldsymbol{X}\beta)], \tag{2}$$

where $b(\theta)=(b(\theta_1),b(\theta_2),\cdots,b(\theta_n))$ for $\theta=(\theta_1,\theta_2,\cdots,\theta_n)^{\mathrm{T}}$, and $\mathbf{1}$ is the $n$ dimensional column vector with all entries 1.

A regularized regression approach selects the features and estimates the coefficients simultaneously by maximizing a penalized likelihood:

$$Q(\beta)=l_n(\beta)-\sum_{j=1}^{p}p_\lambda(|\beta_j|), \tag{3}$$

where $\lambda$ is a regulating parameter and $p_\lambda(\cdot)$ is a penalty function such that the number of fitted nonzero coefficients can be regulated by $\lambda$. That is, only a certain number of $\beta_j$'s are estimated nonzero when $\lambda$ is set at a certain value. Various penalty functions have been proposed and studied, including Lasso[2], SCAD[3], adaptive Lasso[4] and MCP[5], and so on.

We consider $L_1$ penalty function in (3) with Lasso penalty function and sequentially select relevant variables. We call it as iterative sequential lasso (ISLasso). By selecting regular paremeter $\lambda$, we sequentially select important variables and only select one at each step. For the important variables selected at each step, we compute EBIC proposed in [6]. The procedure continues, if the EBIC keeps decreasing. When the EBIC attains a minimum at step $k$, the procedure stops and the set is taken as the final selected set. Its conceptual description and computation algorithm are given in the next section.

The proposed ISLasso extends the procedure of Slasso[7] to the generalize linear model. Slasso[7] uses Lasso penalty function sequentially to select relevant variables for the linear models, utilized a sequence of partially penalized least squares equation in variable selection while ISLasso addressed these problems a sequence of partially penalized maximizing likelihood functions. The main technical challenge is to deal with the likelihood function sequentially which has different numbers of parameters and different Taylor expansion. We propose iterative algorithm sequentially based Lasso penalty Taylor expansion of likelihood function, establish the convergence of the iteration algorithm and further show its screening effectiveness. The high efficiency of the proposed method is illustrated through extensive simulation studies.

The rest of the paper is organized as follows: In Section 2, we discuss general linear model sequential lasso feature selection procedure; We study basic properties and computation algorithm of ISLasso is given in Section 3. The theoretical properties of ISLasso are studied in Section 4. Simulation studies comparing ISLasso with other existing methods and a real data analysis are reported in Section 5. Proofs of theorems are presented in Section 6.

## 2 Feature Screening Procedure

We consider the sequential lasso penalized likelihood identifying relevant features,

$$Q_1(\beta) = l_n(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j|, \tag{4}$$

where $\lambda_1 \geq 0$ is a regularization parameter. We select $\lambda_1$ large enough such that at least one of the $\beta_j$'s will be estimated nonzero. The feature with nonzero estimated coefficients are selected and the set of their indices are denoted by $T_1$ at first step. $T_h$ is the indices set selected in the $h$th step. Let $s_h = |T_h|$ denote the sizes of elements in set $T_h$, $T_0$ denotes the true model and $s_0 = |T_0|$ is the true model size. We gain $T_{h+1}$ by solving

$$Q_{h+1}(\beta) = l_n(\beta) - \lambda_{h+1} \sum_{j \notin T_h} |\beta_j|, \tag{5}$$

where no penalty is imposed on the $\beta_j's$ for $j \in T_h$, and $\lambda_{h+1}$ is large enough such that at least one of the $\beta_j's$, $j \notin T_h$, will be estimated nonzero. The selected set is then updated to $T_{h+1}$.

Using the second order approximation of $l_n(\beta)$ at $\beta_0$, we can rewrite Equation (5) as follows:

$$Q_{h+1}(\beta) \simeq Q^*_{h+1}(\beta, \beta_0) = l^*_n(\beta, \beta_0) - \lambda_{h+1} \sum_{j \notin T_h} |\beta_j|, \tag{6}$$

where

$$l^*_n(\beta, \beta_0) = l_n(\beta_0) + (l'_n(\beta_0))^{\mathrm{T}}(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}} \Sigma(\beta_0) \boldsymbol{X})(\beta - \beta_0),$$

where $\Sigma(\beta_0) = \mathrm{diag}\{b''(\boldsymbol{x}_1^{\mathrm{T}}\beta_0), b''(\boldsymbol{x}_1^{\mathrm{T}}\beta_1), \cdots, b''(\boldsymbol{x}_n^{\mathrm{T}}\beta_0)\}$.

Thus, we estimate $\beta$ by the following iterative procedure

$$\widehat{\beta}^{(h+1)} = \arg\max_{\beta} Q^*_{h+1}(\beta, \widehat{\beta}^{(h)}), \tag{7}$$

where $\widehat{\beta}^{(h)} = \{\widehat{\beta}_h^{(1)}, \widehat{\beta}_h^{(2)}, \cdots, \widehat{\beta}_h^{(h)}\}$ and $\widehat{\beta}^{(l)}$ denotes the $l$-th iterative estimator. The detail explaination can be found in the next section.

We use the extended Bayes information criterion (EBIC)[6] as the stoping rule. The EBIC algorithm is given by

$$\mathrm{EBIC}(\widehat{\beta}(T_h)) = -n \ln \left( \frac{l(\widehat{\beta}(T_h))}{n} \right) + |T_h| \ln n + 2r \ln \binom{p}{|T_h|}, \quad r \geq 0.$$

Let $l_j'(\beta_0) = X_j^{\mathrm{T}}(Y - \mu(\beta_0))$ for $\mu(\beta_0) = \{\mu(\boldsymbol{x}_1\beta_0), \mu(\boldsymbol{x}_2\beta_0), \cdots, \mu(\boldsymbol{x}_n\beta_0)\}^{\mathrm{T}}$, $\widetilde{l}_j'(\beta_0) = \widetilde{X}_j^{\mathrm{T}}\Sigma^{-1}(\beta_0)(Y - \mu(\beta_0))$, $\widetilde{X}_j = \widetilde{H}(A)X_j$ for the set $A$ which are the subsets of $\{1, 2, \cdots, p\}$, and $j \notin A$, where

$$\widetilde{H}(A) = \Sigma(\beta_0) - \Sigma(\beta_0)\boldsymbol{X}(A)\big(\boldsymbol{X}(A^{\mathrm{T}})\Sigma(\beta_0)\boldsymbol{X}(A)\big)^{-1}\boldsymbol{X}^{\mathrm{T}}(A)\Sigma(\beta_0).$$

Without loss of generality, we assume that $T_h$ is obtained, $s_h = |T_h|$ and the initial value of $\beta_0$ can be estimated using $s_h$ dimensional vector in the $T_h$ model. Next $\beta_j$ ($j \notin T_h$) will be estimated with the initial value of 0.

**Proposition 2.1** *For $h \geq 1$ and the initial value $\beta_0$, the maximization of $Q_{h+1}^*(\beta, \beta_0)$ is equivalent to the maximization of*

$$\begin{aligned}
&\overline{Q}_{h+1}^*(\beta(T_h^c), \beta_0) \\
&= l_n(\beta_0) + (l_{T_h}')^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma(\beta_0)\boldsymbol{X}(T_h))^{-1}l_{T_h}' \\
&\quad + (\widetilde{l}'(T_h^c))^{\mathrm{T}}\beta(T_h^c) - \frac{1}{2}\beta(T_h^c)\widetilde{\boldsymbol{X}}^{\mathrm{T}}(T_h^c)\widetilde{H}(T_h)\widetilde{\boldsymbol{X}}(T_h^c)\beta(T_h^c) - \lambda_{h+1}\sum_{j\in T_h^c}|\beta_j|. \qquad (8)
\end{aligned}$$

*Proof* Differentiating $l_{h+1}$ with respect to $\beta(T_h)$, we have

$$\frac{\partial l_n^*(\beta, \beta_0)}{\partial \beta(T_h^c)} = l_{T_h}'(\beta_0) - \boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma(\beta_0)\boldsymbol{X}(\beta - \beta_0).$$

Setting the above derivative to 0, we obtain

$$\begin{aligned}
\widehat{\beta}(T_h) = &\big(\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma(\beta_0)\boldsymbol{X}(T_h)\big)^{-1}[l_{T_h}'(\beta_0) \\
&-\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma\boldsymbol{X}(T_h^c)(\beta(T_h^c) - \beta_0(T_h^c))]. \qquad (9)
\end{aligned}$$

Substituting (9) into (6) we have the type (8). ∎

Note the component of $\widehat{\beta}(T_h)$ are almost surely nonzero since $Y$ in $l_{T_h}'(\beta_0) = \boldsymbol{X}^{\mathrm{T}}(T_h)(Y - \mu(\beta_0))$ is a vector of continuous random variables for initial value $\beta_0$. Thus, the active variable set from the sequential step are nested: $T_1 \subset \cdots \subset T_h \subset \cdots$.

**Proposition 2.2** *Let $t* = \{\widehat{j} : \widehat{j} \in T_h^c, \widehat{j} = \arg\max_j |(\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j)^{-1}\widetilde{X}_j^{\mathrm{T}}\Sigma^{-1}(\beta_0)(Y - \mu(\beta_0) + \beta_{0j}|\}$, where $\beta_{0j}$ is the $j$-component of $\beta_j$. Then $X_{\widehat{j}}$ with $\widehat{j} \notin T_h$ is the only feature with nonzero estimated coefficient in the maximization of* (8)*, a.e., the set $t*$ only has one element $\widehat{j}$.*

The proof of Proposition 2.2 follows from the proof of Theorem 2.3. Next we consider the following ISLasso algorithm from Proposition 2.2.

**Step 0** The initial value $\widehat{\beta}^{(0)}$ sets $\mathbf{0}$.

**Step 1** Compute $\widehat{\beta}_{n_{(1)}}^{(1)} = \arg\max_j |(X_j^{\mathrm{T}}\Sigma(\mathbf{0})X_j)^{-1}(X_j^{\mathrm{T}}(Y - \mu(\mathbf{0})))|$ and $\widehat{\beta}_{(1)}^{(1)} = (0, \cdots, \widehat{\beta}_{n_{(1)}}^{(1)}, \cdots, 0)^{\mathrm{T}}$, then update $\widehat{\beta}^{(0)}$ by $\widehat{\beta}_{(1)}^{(1)}$.

**Iterative procedure** Let $\widehat{\beta}_{n_{(k_1)}}^{(1)} = \arg\max_j |(X_j^{\mathrm{T}}\Sigma(\widehat{\beta}_{(k_1-1)}^{(1)})X_j)^{-1}(X_j^{\mathrm{T}}(Y - \mu(\widehat{\beta}_{(k_1-1)}^{(1)})))|$.

**Convergence criteria** When $|\widehat{\beta}_{(k_1+1)}^{(1)} - \widehat{\beta}_{(k_1)}^{(1)}|$ falls below some tolerance level, the iteration stops and $T_1 = \{n_{(k_1)}\}$. Update $\widehat{\beta}^{(1)} = \widehat{\beta}_{(k_1+1)}^{(1)} = (0, \cdots, 0, \widehat{\beta}_{n_{(k_1)}}, 0 \cdots, 0$. Compute EBIC$(\widehat{\beta}^{(1)})$.

**Step 2** (**General Iteration Step**) For $h \geq 1$, $T_{h-1} = \{n_{(k_1)}, n_{(k_2)}, \cdots, n_{k_{(h-1)}}\}$, initial value $\widehat{\beta}^{(h-1)}$ have nonzero component index in $T_{h-1}$. Compute

$$\widehat{\beta}^{(h)}_{n_{(k_h)}} = \arg\max_j (\widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T_{h-1}) \widetilde{X}_j)^{-1} \widetilde{l}'_j(\widehat{\beta}^{(h)}_{(k_h-1)}).$$

By iteration step we obtain $T_h = T_{h-1} \bigcup \{n_h\}$ until $|\widehat{\beta}^{(h)}_{(n_h+1)} - \widehat{\beta}^{(h)}_{(n_k)}|$ falls below some tolerance level.

Next compute EBIC($\widehat{\beta}^{(h)}$), if EBIC($\widehat{\beta}^{(h)}$) >EBIC($\widehat{\beta}^{(h-1)}$), stop, the selected variables set is $T_h$; otherwie, continue.

For the sequence $T_1, T_2, \cdots, T_h, \cdots$ and that, with probability converging to 1, there is an $h*$ such that $h* = T_0$. In Theorem 3.3 we provide the result that, with probability converging to 1, EBIC($T_{h*}$) increases when $h < h*$ and reaches its maximum at step $h*$, and that EBIC($T_h$) < EBIC($T_{h*}$) for any $h > h*$. The result is given in Theorem 3.3. This result implies that, with probability converging to 1, the procedure of ISLasso stops at step $h*$. The parameters in the selected model are estimated by maximum likelihood estimation. More over, we derive the asymptotic distribution of the ISLasso estimator of $\beta(T*)$ in Theorem 3.4.

**Theorem 2.3** *Assume conditions of Proposition* 2.2 *hold. Suppose that* $\ln p_n = O(n^\kappa), 0 < \kappa < 1 - 2\tau_1 - \alpha$. *Let* $T_1 \subset T_2 \subset \cdots \subset T_h$ *be the sets generated by the procedure of the ISLasso. Let* $h*$ *be given in Theorem* 2.3. *For* $\gamma > 1 - \frac{\ln n}{\ln p}$, *then:*

(i) *Uniformly, for* $h < h*$,

$$P\big(\mathrm{EBIC}_\gamma(\widehat{\beta}(T_{h+1}))) < \mathrm{EBIC}(\widehat{\beta}(T_h)))\big) \to 1.$$

(ii)

$$P\Big(\max_{|T_h|>p_0} \mathrm{EBIC}_\gamma(\widehat{\beta}(T_h)) > \mathrm{EBIC}_\gamma(\widehat{\beta}(T_0))\Big) \to 1.$$

*Proof* By Proposition 2.2, $T_h \subset T_0$ if $T_h \leq p_0$ with probability converging to 1. Let $D_h$ =EBIC$_\gamma(T_h)$−EBIC$_\gamma(T_{h+1})$. 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,

$$\begin{aligned} D_h &= n \ln \left( \frac{l_n(\widehat{\beta}_{h+1})}{l_n(\widehat{\beta_h})} \right) + (|T_h| - |T_{(h+1)}|)(\ln n + 2\gamma \ln p) \\ &= n \ln \left( 1 + \frac{l_n(\widehat{\beta_{h+1}}) - l_n(\widehat{\beta_h})}{l_n(\widehat{\beta_h}} ) \right) - (\ln n + 2\gamma \ln p) \\ &= G_h - (\ln n + 2\gamma \ln p), \end{aligned}$$

where $G_h = n \ln \big(1 + \frac{l_n(\widehat{\beta_{h+1}}) - l_n(\widehat{\beta_h})}{l_n(\widehat{\beta_h}}\big)$. Thus, it suffices to show that

$$P(G_h \leq \ln n + 2\gamma \ln p) \to 1, \quad \text{uniformly for } 1 \leq h \leq p_0,$$

which is implied by

$$\frac{\min_{1 \leq h \leq p_0} G_h}{\ln p} \to \infty, \quad \text{in probability.} \tag{10}$$

Now we bound the $l_n(\widehat{\beta_{h+1}}) - l_n(\widehat{\beta_h})$ and $l_n(\widehat{\beta_h})$ separately. By Conditions C3, C4, we have

$$\begin{aligned} l_n(\widehat{\beta}_{h+1}) - l_n(\widehat{\beta_h}) &= Y^{\mathrm{T}}\boldsymbol{X}(T_{(h+1)})\widehat{\beta}(T_{(h+1)}) - Y^{\mathrm{T}}\boldsymbol{X}(T_h)\widehat{\beta}(T_h) \\ &\quad -1^{\mathrm{T}}b(\boldsymbol{X}(T_{(h+1)})\widehat{\beta}(T_{(h+1)})) + 1^{\mathrm{T}}b(\boldsymbol{X}(T_h)\widehat{\beta}(T_h)) \\ &= Y^{\mathrm{T}}X_j\widehat{\beta}_j - \mu(\boldsymbol{X}(T_h)\widehat{\beta}(T_h))X_j\widehat{\beta}_j + (\widehat{\beta}_j)^2X_j^{\mathrm{T}}\Sigma X_j \\ &\geq C_3n^{-\tau/2}\cdot w_1n^{-\tau_1} + n\lambda_{\min}(\Sigma)\cdot w_1^2n^{-2\tau_1} \\ &\geq C_4n^{1-2\tau_1}, \end{aligned}$$

the first inequality holds by Conditions C3, C4, $C_4 = C_3n^{\tau_1-\tau/2-1} + \lambda_{\min}(\Sigma)\cdot w_1^2$ for $\tau_1 - \tau/2 - 1 < 0$, $1 - 2\tau_1 > 0$.

$$\begin{aligned} l_n(\widehat{\beta_h}) &= Y^{\mathrm{T}}\boldsymbol{X}((T_h)\widehat{\beta}(T_h)) - 1^{\mathrm{T}}b(\boldsymbol{X}(T_h)\widehat{\beta}(T_h)) \\ &= Y^{\mathrm{T}}\boldsymbol{X}((T_h)\widehat{\beta}(T_h)) - \mu(\boldsymbol{X}(T_h)\widehat{\beta}(T_h))X_j\widehat{\beta}_j \\ &\quad +\mu(\boldsymbol{X}(T_h)\widehat{\beta}(T_h))X_j\widehat{\beta}_j - 1^{\mathrm{T}}b(\boldsymbol{X}(T_h)\widehat{\beta}(T_h)) \\ &\leq C_4n^{-\tau/2} - M_2n + N_2\|X(T_h)\|_1\min_{j\in T_h}|\widehat{\beta}_j| \\ &\leq C_4n^{-\tau/2} - M_2n + N_2n\frac{\min_{j\in T_h}|\widehat{\beta}_j|}{\min_{1\leq i\leq n}|x_{il}|} \\ &\leq C_4n^{-\tau/2} - M_2n + N_2C_5n^{1+\alpha} = C_6n^{1+\alpha}, \end{aligned} \tag{11}$$

where $\|a\|_1$ denotes the maximum value of all column vector $L_1$ norm of the matrix $a$, $C_6 = N_2C_5 - M_2n^{-\alpha} + C_4n^{-\tau/2} - 1 - \alpha$. The last two inequalities in (11) holds since Conditions C4, C5, C6 and $\|X(T_h)\|_1 = |X_l|_1$, $l \in T_h$. Since $n|X_l|_2^2 \geq |X_l|_1^2 \geq n|X_l|_1\min_{1\leq i\leq n}|x_{il}|$.

Thus, we have

$$\begin{aligned} \frac{\min_{1\leq k\leq p_0}G_h}{\ln p} &\geq \frac{n}{\ln p}\ln\left(1 + \frac{C_4n^{1-2\tau_1}}{C_6n^{1+\alpha}}\right) \\ &\geq \frac{n}{\ln p}C_7n^{-2\tau_1-\alpha} = C_7\frac{n^{1-2\tau_1-\alpha}}{\ln p} \to \infty, \end{aligned}$$

where $C_7 = C_4/C_6$. The last inequality holds since $\ln(1+x) \geq \frac{x}{2}$ if $0 \leq x \leq 1$, and $\ln p = O(n^\kappa)$, $k < 1 - 2\tau_1 - \alpha$. The result of (i) is proved. The result of (ii) in Theorem 3.3 follows from the selection consistency of EBIC[6]. ∎

## 3 Theoretical Properties

We consider the design matrix $\boldsymbol{X}$ as a deterministic matrix. Let $T_0 = \{1 \leq j \leq p : \beta_j \neq 0\}$. Assume $\ln p = O(n^\kappa)$ for some $\kappa > 0$ and $s_0 = |T_0| = O(n^c)$ for some $0 < c < 1$, we first present the result that, with probability converging to 1, the ISLasso selects all the relevant features before any irrelevant feature can be selected. Then we present the result that, with probability converging to 1, the ISLasso procedure using the EBIC as the stopping rule stops exactly at the step when all the relevant features are selected. The asymptotic distribution of the ISLasso estimator is also provided. The proofs of these results are given in the supplementary document.

Suppose that the columns of $\boldsymbol{X}$ are standardized. For some subset $T$ of the set $\{1, 2, \cdots, p\}$, let $T_- = T^c \bigcap T_0$, where $T^c$ denotes the complement of the set $T$. If $T \subset T_0$, $T_-$ is the complement of $T$ in $T_0$. For $T \subset T_0$, define

$$\eta_n(T, j) = (\widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T) \widetilde{X}_j)^{-1} \widetilde{l}_j'(\beta(T)). \tag{12}$$

The scalar $\eta_n(T, j)$ depends on $\beta(T)$ and the column vector $\widetilde{X}_j$.

To gain theoretical insights about the ISLasso algorithm, the following standard technical conditions are needed.

**C1** $a \max_{j \in T^-} |\eta_n(T, j)| \geq \max_{j \in T_0^c} |\eta_n(T, j)|$, $0 < a < 1$.

**C2** There exists $w_1 > 0$ and some non-negative constant $\tau_1$ such that

$$\min_{j \in T_0} |\beta_j| \geq w_1 n^{-\tau_1}.$$

**C3** $C_1 t_0 n^{-1} \leq \lambda_{\min}(\boldsymbol{X}^{\mathrm{T}}(T_h) \Sigma(\beta(T_k)) \boldsymbol{X}(T_h))^{-1} \leq \lambda_{\max}(\boldsymbol{X}^{\mathrm{T}}(T_h) \Sigma(\beta(T_h)) \boldsymbol{X}(T_h))^{-1} \leq C_2 T_0 n^{-1}$, for some positive constants $C_1, C_2$ and positive integer $h$.

**C4** There exist positive constants $C_3, C_4$ such that $C_4 n^{-\tau/2} \geq \min_{j \in t_-} |X_j^{\mathrm{T}}(Y - \mu(\beta(t)))| \geq C_3 n^{-\tau/2}$.

**C5** There exist positive constants $M_1, M_2$ and $N_1, N_2$, such that

$$M_1 \leq \min\{b(\theta_1), b(\theta_2), \cdots, b(\theta_n)\} \leq \max\{b(\theta_1), b(\theta_2), \cdots, b(\theta_n)\} \leq M_2,$$
$$N_1 \leq \min\{\mu(\theta_1), \mu(\theta_2), \cdots, \mu(\theta_n)\} \leq \max\{\mu(\theta_1), \mu(\theta_2), \cdots, \mu(\theta_n)\} \leq N_2.$$

**C6** $\frac{\min_{j \in T_k} |\widehat{\beta}_j|}{\min_{1 \leq i \leq n} |X_{il}|} \leq C_5 n^{\alpha}$, for $0 < \alpha < 1$, $C_5 > 0$.

**C7** Assume $\max_{i=1}^n E|Y_i - b'(\theta_{0i})|^3 = O(1)$ and $\sum_{i=1}^n (x_i^{\mathrm{T}} B_n^{-1} x_i)^{3/2} \to 0$ as $n \to \infty$ where $(\theta_{01}, \theta_{02}, \cdots, \theta_{0n}) = \theta_0$, $B_n = X^{\mathrm{T}}(T_0) \Sigma(\theta_0) X(T_0)$ and $X(T_0)$ is the matrix consisting of the columns of $X$ with indices in $T_0$.

Condition C2 assumes that the feature in active set has enough strong signals. Condition C3 is similar to the assumption of a positive definite matrix of $\sum(\cdot)$. Condition C7 is similar as Condition 6 in [8] which is related to the Lyapunov condition.

Let $\mathcal{E}(t)$ be the the linear space spanned by the columns of $X(t)$.

**Proposition 3.1** *Let $T_h$ denote the index set of the features selected at the $h$th step of ISLasso. For $h \geq 1$ and any $j \in T_h^c$, if $X_j \in \mathcal{E}(T_h)$, $j \notin T_{h+1}$.*

*Proof* If $X_j \in \mathcal{E}(T_h)$ then there exists $\boldsymbol{b}_h$ such that $X_j = \boldsymbol{X}(T_h)\boldsymbol{b}_h$ and hence

$$\begin{aligned}
Q_{h+1}(\beta) &= Y^{\mathrm{T}} \boldsymbol{X} \beta - \mathbf{1}^{\mathrm{T}} b(\boldsymbol{X}\beta) - \lambda_{k+1} \sum_{j \notin T_k} |\beta_j| \\
&= Y^{\mathrm{T}}(\boldsymbol{X}(T_h)(\beta(T_h) + \beta_j \boldsymbol{b}_h) + \boldsymbol{X}(T_h^c/j)\beta(T_h/j)) \\
&\quad -\mathbf{1}^{\mathrm{T}} b(\boldsymbol{X}(T_h)(\beta(T_h) + \beta_j \boldsymbol{b}_h)) + \boldsymbol{X}(T_h^c/j)\beta(T_h/j) - \lambda_{k+1}\left(|\beta_j| + \sum_{j \in (T_h^c/j)} |\beta_j|\right) \\
&\leq Y^{\mathrm{T}}(\boldsymbol{X}(T_h)\widetilde{\beta}(T_h) + \boldsymbol{X}(T_h^c/j)\beta(T_h^c/j)) \\
&\quad -\mathbf{1}^{\mathrm{T}} b(\boldsymbol{X}(T_h)\widetilde{\beta}(T_h) + \boldsymbol{X}(T_h^c/j)\beta^c(T_h/j)) - \lambda_{k+1} \sum_{j \in (T_h^c/j)} |\beta_j|.
\end{aligned}$$

Thus, when $Q_{k+1}(\beta)$ maximization, $\beta_j$ must be 0. That is, $j \notin T_{h+1}$. ∎

Proposition 3.1 implies that, any feature highly correlated with the features selected will have little chance to be selected subsequently. This nature of ISLasso is similar to that of SLasso[7] which is favorable when it is used for feature selection in ultra-high dimensional feature space where high spurious correlations present, see Fan and Lü[9].

**Theorem 3.2** (Select consistency) *Let $T_1, T_2, \cdots, T_h, \cdots$ be the sequence generated by the ISLasso procedure. The conditions* C1–C3 *hold. Let $\ln p = o(n^k)$, where $k < \frac{1}{2}$. Then, there is an $h*$ such that*

$$Pr(T_{h*} = T_0) \to 1, \quad n \to \infty,$$

*where $T_0$ is the exact index set of the relevant features.*

*Proof* By the KKT condition, at the $(h+1)$th step of the sequential Lasso, the solution $\widehat{\beta}$ satisfies

$$\widetilde{l}_j(\beta(T_h) - \widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T_h) \widetilde{X}_j \widehat{\beta}_j = \lambda \partial \|\beta\|_1, \tag{13}$$

where $\partial\|\widehat{\beta}\|_1$ is a sub gradient of $\|\beta\|_1$ at $\widehat{\beta}$ 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.

Firstly, we show $T_h \subset T_0$, $P(T*_{h+1} \subset T_0) \to 1$, uniformly for all $h$ such that $|s_h| < |s_0|$. For $j \in T_h^c$, let

$$\begin{aligned}\widehat{\eta}_n(T_h, \beta_j) &= l_n(\beta(T_h^c \bigcup j)) - l_n(\beta(T_h)) \\ &= (\widetilde{l}_j(\beta(T_h))^{\mathrm{T}} \beta_j - \frac{1}{2} \beta_j^{\mathrm{T}} \widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T_h) \widetilde{X}_j \beta_j \\ &\quad -(l'_{T_h}(T_h))^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}}(T_h) \Sigma(T_h) \boldsymbol{X}(T_h))^{-1} l'_{T_h}(\beta(T_h)).\end{aligned}$$

Substitute $\xi_j = (\widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T_h) \widetilde{X}_j)^{-1} \widetilde{l}_j(\beta(T_h))$ for $\beta_j$, that

$$\begin{aligned}\widehat{\eta}_n(T_h, \xi_j) &= \frac{1}{2} (\widetilde{l}_j(\beta(T_h))^{\mathrm{T}} (\widetilde{X}_j^{\mathrm{T}} \widetilde{H}(T_h) \widetilde{X}_j)^{-1} \widetilde{l}_j(\beta(T_h)) \\ &\quad -(l'_{T_h}(\beta(T_h))^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}}(T_h) \Sigma(T_h) \boldsymbol{X}(T_h))^{-1} l'_{T_h}(\beta(T_h)) \\ &= \frac{1}{2} (l'_{T_h}(\beta(T_h)) \eta_n(T_h, j) - S1,\end{aligned}$$

where $\eta_n(s*_k, \xi_j)$ is as in the type of (12),

$$S1 = (l'_{T_h}(\beta(T_h))(\boldsymbol{X}^{\mathrm{T}}(T_h) \Sigma(T_k) \boldsymbol{X}(T_h))^{-1} l'_{T_h}(\beta(T_h)).$$

Define

$$\mathcal{A}_h = \Big\{ j : |\widehat{\eta}_n(T_h, \xi_j)| = \max_{l \in T_h^c} |\widehat{\eta}_n(T_h, \xi_l)| \Big\}.$$

By Conditions C3 and C4, we have

$$|S1| = O(t_0 n^{1-\tau}). \tag{14}$$

Let $\Phi_t = (\widetilde{l}_{T_-}(\beta(T_h))(\widetilde{\boldsymbol{X}}(T_-)^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{\boldsymbol{X}}(T_-))^{-1}\widetilde{l}_{T_-}(\beta(T_h))$, and

$$\Phi_t \le n\|\widetilde{l}_{T_-}(\beta(T_h)\|_1 \max_{j\in T_h^c}|\eta_n(S_h,j)|. \tag{15}$$

On the other hand,

$$\begin{aligned}\Phi_t &= (Y-\mu(\beta(T_h)))^{\mathrm{T}}\widetilde{H}(T_h)\Sigma^{1/2}(\beta(T_h))\boldsymbol{X}^{\mathrm{T}}(T_-)\\ &\quad\times[\boldsymbol{X}_{T_-}^{\mathrm{T}}\Sigma^{1/2}(\beta(T_h))\widetilde{H}(T_h)\Sigma^{1/2}\boldsymbol{X}_{T_-}]^{-1}\boldsymbol{X}(T_-)\Sigma^{1/2}(\beta(S_h))\widetilde{H}(T_k)(Y-\mu(\beta(S_h)))\\ &\ge \lambda_{\min}[\boldsymbol{X}^{\mathrm{T}}(t_0)\Sigma(\beta(T_h))\boldsymbol{X}(T_0)]^{-1}\|\boldsymbol{X}(T_-)\Sigma^{1/2}(\beta(T_h))\widetilde{H}(T_h)(Y-\mu(\beta(T_h)))\|_2^2\\ &= \lambda_{\min}[\boldsymbol{X}^{\mathrm{T}}(T_0)\Sigma(\beta(T_h))\boldsymbol{X}(T_0)]^{-1}\|\widetilde{l}_{T_-}(\beta(T_h)\|_2^2.\end{aligned} \tag{16}$$

By the types (15) and (16), we have

$$\begin{aligned}\max_{j\in T_h^c}|\eta_n(T_h,j)| &\ge \lambda_{\min}[\boldsymbol{X}_{T_0}^{\mathrm{T}}\Sigma(\beta(T_h))\boldsymbol{X}_{T_0}]^{-1}\frac{\|\widetilde{l}_{T_-}(\beta(T_h)\|_2^2}{\|\widetilde{l}_{T_-}(\beta(T_h)\|_1}\\ &\ge C_2C_3T_0n^{-1-\tau}.\end{aligned} \tag{17}$$

Above the type (17) is attained by Conditions C3 and C4. By the types (14) and (17), we know that $\mathcal{A}_h \subset T^c *_h \subset t_0$.

Now construct $\widehat{\beta}_j$ as follows. Let $\widehat{\beta}_j = [\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j]^{-1}(\widetilde{l}_j(\beta(T_h)+\frac{w}{n})$ for $j\in\mathcal{A}_h$, where $w$ satisfies that type (13), a.e., $|\widetilde{l}_j(\beta(T_h)-\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j\widehat{\beta}_j| = \lambda$ for $j\in\mathcal{A}_h$, $|\widetilde{l}_j(\beta(T_h)-\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j\widehat{\beta}_j| < \lambda$ for $j\notin\mathcal{A}_h$. The proof is completed. ∎

**Theorem 3.3** *Assume Conditions* C3, C4 *hold. For* $0<|T_h|\le p_0$, *if* $\frac{1}{2}\lambda_{\min}(\widetilde{H}(T_k)) \ge \lambda_{\max}\Sigma(\widetilde{\theta})$. *then*

$$l_n(\widehat{\beta}(T_h)) \le l_n(\widehat{\beta}(T_{h+1})).$$

*Proof* We recall that

$$\begin{aligned}\overline{Q}_{h+1}^*(\beta_j,\beta(t)) &= l_n(\beta(t)) + (l_t'(\beta(t))^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}(t)\Sigma(\beta(t)\boldsymbol{X}(t)^{-1}l_t'(\beta(t))\\ &\quad+(\widetilde{l}_j'(\beta(t))^{\mathrm{T}}\beta_j - \frac{1}{2}\beta_j^{\mathrm{T}}\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(t)\widetilde{X}_j\beta_j - \lambda_{k+1}|\beta_j|,\end{aligned}$$

where $j\notin t$, $t$ is the subset of $\{1,2,\cdots,p\}$.

It is seen that

$$\begin{aligned}l(\widehat{\beta}(T_h)) &\le l(\widehat{\beta}(T_{(h+1)})) - Y^{\mathrm{T}}X_j\widehat{\beta}_j - 1^{\mathrm{T}}b(\boldsymbol{X}(T_h)\widehat{\beta}(T_h)) + 1^{\mathrm{T}}b(\boldsymbol{X}(T_{(h+1)})\widehat{\beta}(T_{(h+1)}))\\ &\quad+(l'(\widehat{\beta}(T_h)))^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}(\widehat{\beta}(T_h))\Sigma(\widehat{\beta}(T_h))\boldsymbol{X}(\widehat{\beta}(T_h))^{-1}l'(\widehat{\beta}(T_h))\\ &\quad+(\widetilde{l}_j'(\widehat{\beta}(T_h))^{\mathrm{T}}\widehat{\beta}_j - \frac{1}{2}\widehat{\beta}_j^{\mathrm{T}}\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j\widehat{\beta}_j - \lambda_{k+1}|\widehat{\beta}_j|\\ &= l(\widehat{\beta}(T_{(h+1)})) + S_1 + S_2 - S_3 - \lambda_{h+1}|\widehat{\beta}_j|,\end{aligned}$$

where

$$\begin{aligned}S_1 &= (\widetilde{l}_j'(\widehat{\beta}(T_h))^{\mathrm{T}}\widehat{\beta}_j - Y^{\mathrm{T}}X_j\widehat{\beta}_j - 1^{\mathrm{T}}b(\boldsymbol{X}(T_h)\widehat{\beta}(T_h)) + 1^{\mathrm{T}}b(\boldsymbol{X}(T_{(h+1)})\widehat{\beta}(T_{(h+1)})),\\ S_2 &= (l'(\widehat{\beta}(T_h)))^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}(\widehat{\beta}(T_h))\Sigma(\widehat{\beta}(T_h))\boldsymbol{X}(\widehat{\beta}(T_h))^{-1}l'(\widehat{\beta}(T_h)),\\ S_3 &= \frac{1}{2}\widehat{\beta}_j^{\mathrm{T}}\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j\widehat{\beta}_j.\end{aligned}$$

Since

$$(\widetilde{l}_j'(\widehat{\beta}(T_h))^{\mathrm{T}}\widehat{\beta}_j = Y^{\mathrm{T}}X_j\widehat{\beta}_j - \mu(\boldsymbol{X}(T_h)\widehat{\beta}(T_h))X_j\widehat{\beta}_j \\ -(Y-\mu(\beta_0))^{\mathrm{T}}\boldsymbol{X}(T_h)\big(\boldsymbol{X}^{\mathbf{T}}(T_h)\Sigma\boldsymbol{X}(T_h)\big)^{-1}\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma X_j\widehat{\beta}_j.$$

We have

$$S_1 = X_j^{\mathrm{T}}\Sigma(\widetilde{\theta})X_j\widehat{\beta}_j^2 - (Y-\mu(\beta_0))^{\mathrm{T}}\boldsymbol{X}(T_h)\big(\mathbf{X}^{\mathrm{T}}(T_h)\Sigma\boldsymbol{X}(T_h)\big)^{-1}\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma X_j\widehat{\beta}_j,$$

where $\widetilde{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$, $\boldsymbol{x}_i^{\mathrm{T}}(T_h)\widehat{\beta}(T_h) \le \theta_i \le \boldsymbol{x}_i^{\mathrm{T}}(T_{(h+1)})\widehat{\beta}(T_{(h+1)})$, and $\Sigma(\widetilde{\theta}) = \mathrm{diag}\{b''(\theta_1), b''(\theta_2), \cdots, b''(\theta_n)\}$.

Since $\lambda_{\min}(\boldsymbol{X}(T_h)\big(\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma\boldsymbol{X}(T_h)\big)^{-1}\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma) \le \lambda_{\min}(\Sigma) = \rho_1$, and from condition C4 we get $(Y-\mu(\beta_0))^{\mathrm{T}}\boldsymbol{X}(T_h)\big(\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma\boldsymbol{X}(T_h)\big)^{-1}\boldsymbol{X}^{\mathrm{T}}(T_h)\Sigma X_j\widehat{\beta}_j = o(n^{-\tau/2})$.

Now we bound the $S_2$. By Conditions C3, C4, $S_2 = O(n^{-\tau/2-1})$.

Thus, $S_1+S_2-S_3 = X_j^{\mathrm{T}}\Sigma(\widetilde{\theta})X_j\widehat{\beta}_j^2 - \frac{1}{2}\widehat{\beta}_j^{\mathrm{T}}\widetilde{X}_j^{\mathrm{T}}\widetilde{H}(T_h)\widetilde{X}_j\widehat{\beta}_j + O(n^{-\tau/2}) \le 0$, when $\frac{1}{2}\lambda_{\min}(\widetilde{H}(T_h)) \ge \lambda_{\max}\Sigma(\widetilde{\theta})$. The result of Proposition 2.1 is obtained. ∎

By Theorem 3.3, we show that the sequence $\{\widehat{\beta}(T_h)\}$ based on ISLasso increases the value of $l_n(\cdot)$ and necessarily converges to a local maximum of $l_n(\cdot)$.

Let $T*$ be the set selected by ISLasso procedure, $\widehat{\beta}(T*)$ be ISLasso estimator of $\beta(T*)$ by maximization likelihood $l_n(\beta(T*))$. We have the following theorem.

**Theorem 3.4** *Under the conditions of Theorem* 3.3 *and if Condition* C7 *holds, then*

$$A_n[X^{\mathrm{T}}(T*)\Sigma(\beta(T*))X(T*)]^{-1/2}(\widehat{\beta}(T*) - \beta(T*)) \to_d N(\mathbf{0}, G),$$

*where $A_n$ is a $q \times |T*|$ matrix such that $A_nA_n^{\mathrm{T}} \to G$, $G$ is a $q \times q$ symmetric positive definite matrix.*

As the proof is similar as those of the proof of Theorem 4 in [8], we omit the detail here.

## 4 Simulation Study

We conduct a comprehensive simulation to evaluate the performance of the proposed methods and compare them with the existing methods. To be specific, we consider three classes of models: Linear regression, logistic regression and Poisson regression and we compare three existing screening methods including ISLasso, SIS[3] and SIS-MLR[10]. For SIS and SIS-MLR we add the iterative SIS (ISIS), because the SIS, ISIS and SIS-MLR only select the relevant variable crudely, we then continue select variables by the Lasso penalty (SIS-Lasso, ISIS-Lasso, SIS-MLR-Lasso).

For each class of models, we consider three correlation structures between candidate features. The first one is when the features are ideally independent with each other, under which the task of feature selection is the most straightforward. The second is when the features have an autocorrelation structure. The neighboring features are correlated but distant ones are virtually uncorrelated. This type of correlation is commonly used in modeling the data with a natural

order. The last one set every feature either relevant or correlated with some relevant features. This structure makes it challenging to distinguish the relevant features from irrelevant features.

We assess the performances of screening methods based on $M = 500$ simulation replications. Let $\widehat{\mathcal{M}}_k$ denote the model selected in the $k$th replication by any specific method. We measure its retaining capacity (RC) of relevant features by

$$\mathrm{RC} = M^{-1} \sum_{i=1}^{M} I(\widehat{\mathcal{M}}_i \subset T_0).$$

Averaged model size (AMS) by

$$\mathrm{AMS} = M^{-1} \sum_{i=1}^{M} |\widehat{\mathcal{M}}_i|.$$

We characterize the model selectivity in terms of positive selection rate (PSR) and false discovery rate (FDR):

$$\mathrm{PSR} = \frac{\sum_{i=1}^{M} |T_0 \cap \widehat{\mathcal{M}}_i|}{M|T_0|}, \quad \mathrm{FDR} = \frac{\sum_{i=1}^{M} |\widehat{\mathcal{M}}_i - T_0|}{M|\widehat{\mathcal{M}}_i|}.$$

The PSR and FDR depict two different aspects of a selection result: A high PSR means most relevant features are identified, while a low FDR indicates few irrelevant features are missselected. As further references, we also report averaged computational time for conducting each study (time, in seconds).

### 4.1 Linear Regression

In this setting, data$(Y, \boldsymbol{x})$ are generated from $Y = \boldsymbol{x}^{\mathrm{T}}\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. The model parameters and covariates used in each of the three correlation setups are as follows:

**(S1)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are independently generated as a random sample from $N(0, 1)$. $|T_0| = 8$, $T_0$ is a simple random sample from $\{1, 2, \cdots, p\}$. $\beta(T_0)$ are independent samples from $4\sqrt{n}(\log n + |Z|)U$, $\beta(t_0^c) = 0$, where $P(U = 1) = 0.6$, $P(U = -1) = 0.4$, $Z \sim N(0, 1)$. We set $(n, p, \sigma) = (100, 200, 3)$.

**(S2)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are joint normal, marginally $N(0, 1)$ with $\mathrm{cov}(X_j, X_{j-1}) = 2/3$, $\mathrm{cov}(X_j, X_{j-2}) = 1/3$ for $j \geq 3$; $\mathrm{cov}(X_j, X_h) = 0$ for $|j - h| \geq 3$. $T_0 = \{1, 3, 5, 7, 9\}$, $\beta(T_0) = (5, 3.5, 2.8, 2.5, 2.2)^{\mathrm{T}}$. We set $(n, p, \sigma) = (100, 200, 5)$.

**(S3)** Covariate $\{X_1, X_2, \cdots, X_p\}$ are joint normal, marginally $N(0, 1)$ with $\mathrm{cov}(X_j, X_{j-1}) = 2/3$, $\mathrm{cov}(X_j, X_{j-2}) = 1/3$ for $j \geq 3$; $\mathrm{cov}(X_j, X_h) = 0$ for $|j - h| \geq 3$. $T_0 = \{1, 2, 3, 4\}$, $\beta(T_0) = (2.5, 2.5, 2.5, 2.5)^{\mathrm{T}}$. We set $(n, p, \sigma) = (100, 1000, 1)$.

These structures are also chosen because they have been discussed in many papers (see, e.g., [11, 12]). Thus, it prevents potential bias in favor of the new method.

**Table 1** Simulation results under linear regression

| Setup | 1st Step | 2nd Step | RC | PSR | FDR | AMS | TIME |
|---|---|---|---|---|---|---|---|
| S1 | SIS | - | 0.19 | 0.83 | 0.78 | 34.0 | 10.54 |
| | SIS | LASSO | 0.18 | 0.83 | 0.12 | 8.2 | 11.05 |
| | ISIS | - | 0.95 | 0.99 | 0.74 | 31.0 | 58.02 |
| | ISIS | LASSO | 0.87 | 0.98 | 0.11 | 8.6 | 60.28 |
| | SIS-MLR | - | 0.99 | 1.0 | 0.74 | 31.0 | 11.37 |
| | SIS-MLR | LASSO | 0.93 | 0.99 | 0.08 | 8.8 | 12.87 |
| | ISLasso | - | 0.99 | 1.00 | 0.07 | 8.2 | 6.25 |
| S2 | SIS | - | 0.58 | 0.91 | 0.81 | 24.0 | 4.85 |
| | SIS | LASSO | 0.35 | 0.83 | 0.02 | 4.2 | 5.20 |
| | ISIS | - | 0.63 | 0.92 | 0.81 | 24.0 | 25.13 |
| | ISIS | LASSO | 0.36 | 0.85 | 0.11 | 4.6 | 25.43 |
| | SIS-MLR | - | 0.77 | 0.95 | 0.80 | 24.0 | 9.85 |
| | SIS-MLR | LASSO | 0.47 | 0.87 | 0.09 | 4.9 | 10.27 |
| | ISLasso | - | 0.87 | 0.96 | 0.18 | 6.2 | 4.33 |
| S3 | SIS | - | 0.01 | 0.32 | 0.95 | 21.0 | 0.8 |
| | SIS | LASSO | 0.01 | 0.23 | 0.81 | 21.3 | 0.96 |
| | ISIS | - | 0.70 | 0.83 | 0.84 | 21.0 | 5.34 |
| | ISIS | LASSO | 0.51 | 0.68 | 0.63 | 8.6 | 5.83 |
| | SIS-MLR | - | 0.99 | 1.00 | 0.81 | 21.0 | 3.27 |
| | SIS-MLR | LASSO | 0.67 | 0.89 | 0.55 | 8.5 | 3.93 |
| | ISLasso | - | 0.77 | 0.92 | 0.78 | 4.4 | 0.95 |

Under Setup S1, features are independent. All screening methods except for SIS have a high retaining capacity. The poor performance of SIS is likely attributable to its use of feature correlation in isolation. Note that the high FDR for screening methods are not of concern, as they simply implies that more elaborated second stage analysis is further needed. Our proposed ISLasso remained satisfactorily but does not outperform in any obvious manner. Under Setup S3, the most challenging one, the strong collinearity among features badly deteriorates the performances of SIS and MLR (screening method). The ISIS recoveries from the failure of SIS to a large degree. It is particularly noticeable that after ISLasso, the resulting selection outcomes outperform other combinations by a big margin. In summary, under linear model, the

proposed method ISLasso is among the best performers under Setup S1 and S2. It outperforms other methods under Setup S3.

### 4.2 Logistic Regression

We now consider the situation when the response variable $Y$ is binary and $\pi = P(Y = 1|x)$ follows $\text{logit}(\pi) = x^{\mathrm{T}}\beta$. That is, $(Y, x)$ satisfy a logistic regression model. The covariates and parameters are specified as follows:

**(S1)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are independently generated as a random sample from $N(0, 1)$. $|T_0| = 8$, $T_0$ is a simple random sample from $\{1, 2, \cdots, p\}$. $\beta_{T_0}$ are independent samples from $4\sqrt{n}\log n + |Z|/4)U$. $\beta(T_0^c) = 0$, where $P(U = 1) = 0.5$, $P(U = -1) = 0.5$, $Z \sim N(0, 1)$.

**(S2)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are independently generated as a random sample from $N(0, 1)$. $\text{cov}(X_j, X_{j-1}) = 2/3$ for $j \geq 3$. $\text{cov}(X_j, X_{j-2}) = 1/3$; $\text{cov}(X_j, X_h) = 0$ for $|j - h| \geq 3$. $T_0 = \{1, 3, 5, 7, 9\}$, and $\beta(T_0) = (2, 1.8, 1.6, 1.4, 1.2)^{\mathrm{T}}$.

**(S3)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are independently generated as a random sample from $N(0, 1)$. $\text{cov}(X_j, X_h) = 0.15$, for $j \in T_0$; $\text{cov}(X_j, X_h) = 0.3$, $j \in T_0^c$. $T_0 = \{1, 2, 3, 4\}$ and $\beta(T_0) = ((1.5, 1.5, 1.5, 1.5)^{\mathrm{T}}$.

We used $n = 400$ and $p = 1000$ for all three setups to allow appropriate parameter estimation. Logistic model is often used to predict the outcome of a future observation with given $x$. When $P(Y = 1|x) \geq 0.5$ according to the fitted model, we predict the future outcome $Y$ as 1 or 0 otherwise. The prediction error (P.err) of each selected model is evaluated by the proportion of incorrect predictions based on an independent testing data. We compute other performance measures in the same way as those for linear models. The results are given in Table 2. Similar to the linear model situation, under Setup S1, we again notice that all screening methods have good performances in terms of retaining relevant features. Under Setup S2, the SIS does not perform satisfactorily. Other two methods all work well for the feature screening, among which ISLasso performs the best. We note that both ISIS-Lasso combination have satisfactory performances. Under Setup S3, the correlation structure is least favorable to feature selection. The comparison between ISIS-Lasso, SIS-MLR, SIS-MLR-Lasso are not apparent in terms of prediction error. Yet we notice significant improvement of ISLasso over ISIS-Lasso for a higher PSR and a lower FDR.

### 4.3 Poisson Regression

Under Possion regression, we set $n = 200$, $p = 1000$. The model parameters and covariates used in each of the three correlation setups are as follows:

**(S1)** Covariates $\{X_1, X_2, \cdots, X_p\}$ are independently generated as a random sample from $N(0, 1)$. $|s_0| = 8$, $T_0$ is a simple random sample from $\{1, 2, \cdots, p\}$. $\beta(T_0)$ are independent samples from $4\sqrt{n}\log n + |Z|/8)U$, $\beta(T_0^c) = 0$, where $P(U = 1) = 0.8$, $P(U = -1) = 0.2$, $Z \sim N(0, 1)$.

**(S2)** Covariates $X_j$ $(j = 1, 2, \cdots, p)$ are independently generated as a random sample from $N(0, 1)$. $\text{cov}(X_j, X_{j-1}) = 2/3$, $\text{cov}(X_j, X_{j-2}) = 1/3$ for $j \geq 3$; $\text{cov}(X_j, X_h) = 0$ for $|j - h| \geq 3$. $T_0 = \{1, 3, 5, 7, 9\}$ and $\beta(T_0) = (2, 1.8, 1.6, 1.4, 1.2)^{\mathrm{T}}$.

**(S3)** Covariates $X_j$ $(j = 1, 2, \cdots, p)$ are independently generated as a random sample from $N(0, 1)$. $\text{cov}(X_j, X_h) = 0.15$ for $j \in s_0$; $\text{cov}(X_j, X_h) = 0.3$ for $j \in s_0^c$. $T_0 = \{1, 2, 3, 4\}$ and $\beta(T_0) = (0.7, 0.7, 0.7, 0.7))^{\mathrm{T}}$.

**Table 2** Simulation results under logistic regression

| Setup | 1st Step | 2nd Step | RC | PSR | FDR | AMS | TIMES |
|---|---|---|---|---|---|---|---|
| S1 | SIS | - | 0.92 | 0.99 | 0.45 | 14.9 | 3.45 |
| | SIS | LASSO | 0.92 | 0.99 | 0.01 | 8.0 | 3.76 |
| | ISIS | - | 1.0 | 1.0 | 0.47 | 5.0 | 16.16 |
| | ISIS | LASSO | 0.99 | 1.00 | 0.03 | 8.6 | 16.35 |
| | SIS-MLR | - | 0.99 | 0.99 | 0.02 | 15.0 | 3.63 |
| | SIS-MLR | LASSO | 0.99 | 1.00 | 0.04 | 8.2 | 3.88 |
| | ISLasso | - | 0.99 | 1.00 | 0.03 | 8.4 | 4.07 |
| S2 | SIS | - | 0.09 | 0.73 | 0.76 | 15.0 | 3.16 |
| | SIS | LASSO | 0.08 | 0.65 | 0.28 | 4.2 | 3.53 |
| | ISIS | - | 0.86 | 0.96 | 0.68 | 15.0 | 24.37 |
| | ISIS | LASSO | 0.75 | 0.93 | 0.22 | 6.2 | 24.82 |
| | SIS-MLR | - | 0.97 | 0.98 | 0.67 | 15.0 | 4.76 |
| | SIS-MLR | LASSO | 0.87 | 0.96 | 0.18 | 6.2 | 5.08 |
| | ISLasso | - | 0.99 | 1.00 | 0.28 | 4.9 | 3.98 |
| S3 | SIS | - | 0.01 | 0.43 | 0.89 | 15.0 | 3.17 |
| | SIS | LASSO | 0.01 | 0.32 | 0.78 | 6.7 | 3.51 |
| | ISIS | - | 0.61 | 0.82 | 0.78 | 15.0 | 21.50 |
| | ISIS | LASSO | 0.37 | 0.65 | 0.65 | 7.4 | 21.96 |
| | SIS-MLR | - | 0.77 | 0.92 | 0.78 | 15.0 | 5.06 |
| | SIS-MLR | LASSO | 0.56 | 0.85 | 0.50 | 7.0 | 5.15 |
| | ISLasso | - | 0.99 | 0.82 | 0.50 | 7.1 | 6.45 |

The simulation results are given in Table 3. In this modeling context, we observe that the SIS even fails under Setup S1, while the performance of ISIS is also unsatisfactory. In comparison, the ISLasso achieves high retaining capability of relevant features.

**Table 3** Simulation results under Poisson regression

| Setup | 1st Step | 2nd Step | RC | PSR | FDR | AMS | TIME |
|---|---|---|---|---|---|---|---|
| S1 | SIS | - | 0.08 | 0.76 | 0.70 | 21.0 | 3.80 |
| | SIS | LASSO | 0.06 | 0.74 | 0.31 | 8.9 | 4.57 |
| | ISIS | - | 0.95 | 0.99 | 0.63 | 21.0 | 19.82 |
| | ISIS | LASSO | 0.95 | 0.96 | 0.21 | 10.6 | 20.41 |
| | SIS-MLR | - | 0.94 | 0.99 | 0.62 | 21.0 | 4.05 |
| | SIS-MLR | LASSO | 0.92 | 0.99 | 0.11 | 9.0 | 4.46 |
| | ISLasso | - | 0.92 | 0.99 | 0.11 | 9.0 | 2.37 |
| S2 | SIS | - | 0..01 | 0.53 | 0.86 | 21.4 | 3.48 |
| | SIS | LASSO | 0.01 | 0.47 | 0.64 | 7.2 | 3.84 |
| | ISIS | - | 0.88 | 0.96 | 0.75 | 21.0 | 24.42 |
| | ISIS | LASSO | 0.64 | 0.89 | 0.41 | 8.2 | 26.50 |
| | SIS-MLR | - | 0.93 | 0.98 | 0.76 | 21.0 | 4.50 |
| | SIS-MLR | LASSO | 0.85 | 0.96 | 0.30 | 7.3 | 4.78 |
| | ISLasso | - | 0.87 | 0.95 | 0.22 | 6.5 | 1.27 |
| S3 | SIS | - | 0.00 | 0.21 | 0.96 | 21.0 | 3.54 |
| | SIS | LASSO | 0.00 | 0.14 | 0.93 | 9.0 | 4.15 |
| | ISIS | - | 0.59 | 0.74 | 0.86 | 21.0 | 24.04 |
| | ISIS | LASSO | 0.41 | 0.61 | 0.69 | 8.8 | 24.72 |
| | SIS-MLR | - | 0.93 | 0.98 | 0.81 | 21.0 | 4.69 |
| | SIS-MLR | LASSO | 0.59 | 0.85 | 0.58 | 8.6 | 5.02 |
| | ISLasso | - | 0.93 | 0.98 | 0.58 | 8.6 | 3.29 |

### 4.4 Real Data Example

We apply the ISLasso screening method in a genetic example. In [13], expression levels of 12600 genes were measured from prostate specimens of 52 prostate cancer patients and 50 healthy controls. One objective in this study is to build a gene expression-based classification rule to predict the identity of unknown prostate samples. Such a classification tool is helpful in early detection of prostate cancer, and it can provides a better opportunity for curative surgery. Identifying influential genes to the disease outcome also provides deeper understanding on prostate tumours from a genetic aspect. By performing a permutation-based correlation test, Singh, et al.[13] suggested 456 potential genes that are likely differently expressed between

tumour and normal samples. Based on the information of 6033 genes from the complete dataset, Efron[14] further suggested 377 genes for prediction though an empirical Bayes approach, while Chen and Chen[15] spotted 3 genes by the EBIC-based LASSO. In this example, we reanalyze the dataset by building a logistic regression

$$\text{logit}\big\{P(Y=1|X)\big\} = X\beta,$$

where $Y$ is the binary status of the prostate cancer (with $Y=1$ for a tumor sample, $Y=0$ for a normal sample) and $x$ is the 12600 gene expression levels. Accordingly, we predict $Y=1$ when $P(Y=1|x)$ is estimated over 0.75 and predict $Y=0$, otherwise. The ISLasso is used due to its decent performance in our simulation studies. Specifically, we randomly select a set of 10 subjects from each of the tumour and normal sample groups as the testing set, and treat the rest as the training set.

**Table 4** Results for analyzing the prostate data

| screeing | Ams | Sensitivity | Specificity | P.err |
|---|---|---|---|---|
| SIS | 2.2 | 0.71 | 0.96 | 0.17 |
| ISIS | 2.3 | 0.68 | 0.96 | 0.18 |
| SIS-MLR | 2.7 | 0.77 | 0.94 | 0.14 |
| ISLasso | 0.92 | 0.99 | 0.11 | 0.20 |

## 5 Summary and Conclusions

In this paper, we developed a new feature screening approach, ISLasso, for the ultra-high dimensional regression analysis. The proposed method has a good potential to improve the existing marginal-information-based methods by taking the joint effects between features into consideration. We established the screening consistency of the new method and further developed an iterative algorithm. The comprehensive simulation studies demonstrated that ISLasso has an excellent capability of retaining relevant features under simulation setting. These characteristics all merit ISLasso a promising approach in the analysis of ultra-high dimensional data.

## References

[1] McCullagh P and Nelder J, *Generalized Linear Models*, Chapman and Hall, London, 1996.

[2] Tibshirani R, Regression shrinkage and selection via LASSO, *Journal of the Royal Statistical Society, Series B*, 1996, **58**: 267–288.

[3] Fan J and Li R, Variable selection via nonconcave penalized likelihood and it oracle properties, *Journal of the American Statistical Association*, 2001, **96**: 1348–1360.

[4] Zou H, The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association*, 2006, **101**: 1418–1429.

[5] Zhang C H, Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, 2010, **38**: 894–942.

[6] Chen J H and Chen Z H, Extended Bayesian information criteria for model selection with large model spaces, *Biometrika*, 2008, **95**: 759–771.

[7] Luo S and Chen Z, Sequential Lasso cum ebic for feature selection with ultra-high dimensional feature space, *Joural of the American Statistical Association*, 2014, **109**: 1229–12400.

[8] Fan J and Lü J, Nonconcave penalized likelihood with NP-dimensionality, *IEEE Transactions on Information Theory*, 2011, **57**: 5467–5484.

[9] Fan J and Lü J, Sure independence screening for ultrahigh dimentional feature space (with discussion), *Journal of the Royal Statistical Society Series B*, 2008, **70**: 849–911.

[10] Fan J and Song R, Sure independence screening in generalized linear models with NP-dimensionality, *The Annals of Statistics*, 2010, **38**: 3567–3604.

[11] Fan J, Samworth R, and Wu Y, Ultrahigh dimentional variable selection: Beyond the linear model, *Journal of Machine Learning Research*, 2009, **10**: 1829–1853.

[12] Wang H and Xia Y, Shrinkage estimation of the varying coefficient model, *Journal of the American Statistical Association*, 2009, **104**: 747–757.

[13] Singh D, Febbo P G, Ross K, et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 2012, **1**(2): 1203–1209.

[14] Efron B, Empirical bayes estimates for large-scale predition problem, *Journal of the American Statistical Association*, 2009, **104**: 1015–1028.

[15] Chen J and Chen Z, Extended bic for small-$n$-large-$p$ sparse glm, *Statistica Sinica*, 2012, **22**: 555–574.

[16] Efron B, Hastie T, Johnstone I, et al., Least angle regression (with discussion), *The Annals of Statistics*, 2004, **32**: 407–499.