Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space

By SHAN LUO[1] and ZEHUA CHEN[2]

[1]Department of Mathematics

Shanghai Jiao Tong University

[2]Department of Statistics and Applied Probability

National University of Singapore

Email: [1]sluo2012@gmail.com, [2]stachenz@nus.edu.sg.

Abstract

In this paper, we propose a method called sequential Lasso (SLasso) for feature selection in sparse high dimensional linear models. The SLasso selects features by sequentially solving partially penalized least squares problems where the features selected in earlier steps are not penalized. The SLasso uses extended BIC (EBIC) as the stopping rule. The procedure stops when EBIC reaches a minimum. The asymptotic properties of SLasso are considered when the dimension of the feature space is ultra-high and the number of relevant feature diverges. We show that, with probability converging to 1, the SLasso first selects all the relevant features before any irrelevant features can be selected, and that the EBIC decreases until it attains the minimum at the model consisting of exactly all the relevant features and then begins to increase. These results establish the selection consistency of SLasso. The SLasso estimators of the final model are ordinary least squares estimators. The selection consistency implies the oracle property of SLasso. The asymptotic distribution of the SLasso estimators with diverging number of relevant features is provided. The SLasso is compared with other methods by simulation studies, which demonstrates that SLasso is a desirable approach having an edge over the other methods. The SLasso together with the other methods are applied to a microarray data for mapping disease genes.

Key Words: extended BIC; feature selection; oracle property; selection consistency; sequential Lasso, sparse high dimensional linear models.

1 Introduction

Sparse high-dimensional regression (SHR) models arise in many important contemporary scientific fields. A SHR model is as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \ i = 1, \ldots, n, \tag{1.1}$$

where the number of features p is much larger than the sample size n, and only a relatively small number of the β_j's are non-zero. Feature selection is crucial in the analysis of SHR models. There are usually two goals of feature selection: (i) to build a model with desirable prediction properties and (ii) to identify the features with non-zero coefficients (for convenience, such features are referred to as relevant features in this article). These two goals are intertwined but not the same.

Regularized regression approaches to the analysis of SHR models have attracted considerable attention of the researchers. A regularized regression approach selects the features and estimates the coefficients simultaneously by minimizing a penalized sum of squares of the form:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.2}$$

where λ is a regulating parameter and p_λ is a penalty function such that the number of fitted non-zero coefficients can be regulated by λ; that is, only a certain number of β_j's are estimated non-zero when λ is set at a certain value. Various penalty functions have been proposed and studied, including Lasso [28]: $p_\lambda(|\beta_j|) = \lambda|\beta_j|$, SCAD [8], which smoothly clips a L_1 penalty (for small $|\beta_j|$) and a constant penalty (for large $|\beta_j|$), Adaptive Lasso [38]: $p_\lambda(|\beta_j|) = \lambda w_j|\beta_j|$ where w_j are given weights, and MCP [35], which smoothly approaches the L_1 penalty from a constant penalty (for large $|\beta_j|$'s) by an asymptote.

A so-called oracle property is of major concern for any feature selection method. The oracle property refers to two asymptotic natures: (a) selection consistency, i.e.,

the sparse relevant features can be exactly selected with probability converging to 1, and (b) the effects of relevant features can be consistently estimated the same as they would be were they obtained by knowing the relevant features in advance. For fixed p, it was showed in [17] [19] that Lasso is consistent in estimating the regression coefficients but, in general, it does not have the oracle property. A condition on the feature matrix was provided in [38] for Lasso to possess the oracle property. The condition was also discovered in [21] [37] and was dubbed as irrepresentability condition in [37]. When p is allowed to diverge to infinity faster than n (but not too fast), the selection consistency of Lasso under the irrepresentability condition was established in [21] [37]. To relax the irrepresentability condition, Adaptive Lasso was considered in [38] for fixed p using the ordinary least squares estimates as the weights in the penalty, and Adaptive Lasso was shown to have the oracle property. For diverging p, [5] showed that Adaptive Lasso with marginal least squares estimates as the weights has the oracle property if a partial orthogonality condition holds. The properties of SCAD were studied in [8] [9] [10] [33]. In these papers, the oracle property of SCAD was established for various models when p is fixed or diverging to infinity not too fast. The MCP penalty is similar to the SCAD penalty. The asymptotic properties of the MCP penalty were studied in [35]. To realize the oracle property of the various regularized regression methods in finite samples, a proper choice of the regulating parameter has to be made. A multi-fold cross validation (CV) is commonly used in these methods for the choice of the regulating parameter.

Sequential methods have also received attention in recent years for feature selection in SHR models. The traditional sequential procedures such as forward stepwise regression (FSR) were criticized for their greedy nature. However, it was discovered recently that the greedy nature is indeed a good one if the goal is to identify relevant features, see [29] [30], especially, in the presence of high spurious correlations due to

extremely high dimensionality of the feature space. In many practical problems, the identification of the relevant features is of primary interest. For example, in genetic quantitative trait loci (QTL) mapping and disease gene mapping, of interest are the markers which are either QTL or disease gene themselves or are in linkage disequilibrium with QTL or disease genes. In addition, sequential methods are computationally appealing. These revived interests in sequential approaches.

The properties of FSR for feature selection in SHR models were re-examined in [31]. It was shown that FSR has a so-called sure screening property when the procedure is carried out until a certain step before the number of steps reaches the sample size. The sure screening property means that the selected set contains the set of relevant features with probability converging to 1, see [11].

A different version of forward regression referred to as forward selection in [32] was re-considered and dubbed as orthogonal matching pursuit (OMP) in [34]. At each step of OMP, the response vector is projected onto the space spanned by the currently selected features, and the next feature is selected to maximize the correlation with the current residual. The procedure stops when the residual is reduced below a certain level specified by a stopping rule. The properties of OMP have been studied under quite strict conditions in, e.g., [2] [29] [30]. A thorough investigation of the properties of OMP is still lacking.

An adaptive forward-backward greedy algorithm (FoBa) was considered in [36]. The FoBa is a variant of OMP. At the forward step of FoBa, the same mechanism of OMP is used to select new features. A new feature is selected if the amount of decrease in the residual exceeds a specified threshold. Following each forward step, a backward step is carried out if the amount of increase in the residual caused by deleting one of the selected features is less than half of the amount of increase at the forward step. The procedure stops when the amount of decrease in the residual at the

forward step falls below the specified threshold. Some oracle inequalities have been derived under a so-called restricted isometry condition in [36], however, whether or not FoBa is selection consistent is left untouched.

A non-sequential but closely related procedure called compressive sampling matching pursuit (CoSaMP) was proposed in [22]. The procedure of CoSaMP assumes the knowledge of sparsity level k, i.e., the number of relevant features. For a given k, the CoSaMP starts with a sparse set consisting of the k features having highest correlations with the response vector, then updates the sparse set by iteration. At each iteration, the response vector is projected onto the space spanned by the sparse set, $2k$ additional features having highest correlations with the residual of the projection together with the sparse set are fitted to a least square regression model, the updated sparse set consists of the k features having the largest absolute fitted coefficients in the regression model. The number of iterations is either fixed or determined by a certain rule. Eventually, the true sparsity level is chosen by a certain method. Though the procedure of CoSaMP is appealing, its properties are not fully investigated.

A sequential procedure of a different nature called least angle regression (LAR) was proposed in [7]. The LAR continuously updates the estimate of the expected responses along a direction having equal angle with the features already selected and selects new features having the largest absolute correlation with the updated current residuals. The fitted regression coefficients at each step are shrunk. The LAR algorithm has been modified to compute the solution path of Lasso. There are also variants of LAR, e.g., the forward Lasso adaptive shrinkage (FLASH) considered in [24]. The fitted regression coefficients at each step of FLASH are not fully shrunk as in LAR or Lasso.

In this paper, we propose a sequential procedure called sequential Lasso (SLasso) with the emphasis on the goal of identifying relevant features. We give a conceptual

description of SLasso in the following. Its computation algorithm is given in §2. In summary, SLasso solves a sequence of partially penalized least squares problems. The features selected in an earlier step are not penalized in the subsequent steps. Let the vectors $\boldsymbol{y} = (y_1, \ldots, y_n)^\tau$, $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^\tau$, be standardized such that they have length $\sqrt{n}$ and are orthogonal to the vector with all elements 1. Thus in model (1.1), the intercept β_0 can be omitted. At the initial step, SLasso minimizes the following penalized sum of squares:

$$l_1 = \|\boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|,$$

where $\|\cdot\|$ is the L_2-norm, and λ_1 is the largest value of the penalty parameter such that at least one of the β_j's will be estimated non-zero. The features with non-zero estimated coefficients are selected and the set of their indices is denoted by s_{*1}. For $k \geq 1$, let s_{*k} be the index set of the features selected until step k. At step $k+1$, SLasso minimizes the following partially penalized sum of squares:

$$l_{k+1} = \|\boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j\|^2 + \lambda_{k+1} \sum_{j \notin s_{*k}} |\beta_j|,$$

where no penalty is imposed on the β_j's for $j \in s_{*k}$ and λ_{k+1} is the largest value of the penalty parameter such that at least one of the β_j's, $j \notin s_{*k}$, will be estimated non-zero. The selected set is then updated to s_{*k+1}. The EBIC proposed in [3] is used as the stopping rule. For each s_{*k}, the EBIC of the model with features in s_{*k} is computed. The procedure continues, if the EBIC keeps decreasing. If the EBIC attains a minimum at step k^*, the procedure stops and the set s_{*k^*} is taken as the final selected set.

The minimization of l_{k+1} is equivalent to the minimization of

$$\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|^2 + \lambda_{k+1} \sum_{j \notin s_{*k}} |\beta_j|, \tag{1.3}$$

where $\tilde{\boldsymbol{y}}$ is the residual of $\boldsymbol{y}$ projected on the space spanned by the $\boldsymbol{x}_j$'s with $j \in s_{*k}$ and $\tilde{\boldsymbol{X}}$ is the residual matrix of the $\boldsymbol{x}_j$'s, $j \notin s_{*k}$, projected on the same space, see Proposition 2.2. The active features $\boldsymbol{x}_j$ in the minimization of (1.3) must attain $\max_{j' \notin s_{*k}} |\tilde{\boldsymbol{y}}^\tau \boldsymbol{x}_{j'}|$. Thus, the minimization of (1.3) further reduces to the minimization of

$$\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_{\text{TEMP}} \tilde{\boldsymbol{\beta}}_{\text{TEMP}}\|^2 + \lambda_{k+1} \sum_{j \in s_{\text{TEMP}}} |\beta_j|, \tag{1.4}$$

where $s_{\text{TEMP}} = \{j : |\tilde{\boldsymbol{y}}^\tau \boldsymbol{x}_j| = \max_{j' \notin s_{*k}} |\tilde{\boldsymbol{y}}^\tau \boldsymbol{x}_{j'}|\}$, $\tilde{\boldsymbol{X}}_{\text{TEMP}}$ and $\tilde{\boldsymbol{\beta}}_{\text{TEMP}}$ are, respectively, the corresponding projected residual matrix and the coefficient vector. If a partial positive cone condition (condition A2 in §3) is satisfied then s_{TEMP} is exactly the index set of the active $\boldsymbol{x}_j$'s. When s_{TEMP} is a singleton, the partial positive cone condition is automatically satisfied. For these results, see the proof of Theorem 3.1. The non-singleton case rarely occurs. Therefore, the minimization of (1.4) is rarely called. If the need for the minimization of (1.4) does arise, the active $\boldsymbol{x}_j$'s can be easily obtained by applying the R function `glmpath` [23] to $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}_{\text{TEMP}}$ and extracting the first feature (or features) with non-zero coefficient in the solution path. The results discussed above give rise to an efficient computation algorithm which is provided in §2.

We consider the properties of SLasso cum EBIC in the scenario that $p = \exp(cn^\kappa)$, $0 < \kappa < 1$, and the number of relevant features p_0 is also diverging to infinity at a proper rate. We establish the following properties. Let $s_{*1}, s_{*2}, \cdots, s_{*k}, \cdots$ be the sequence generated by SLasso. Under reasonable conditions, there is a $k = k^*$ such that $s_{*k^*} = s_0$ with probability converging to 1 as n goes to infinity, where s_0 is the exact index set of the relevant features (Theorem 3.1 and 3.2). Further, with probability converging to 1 uniformly for all $k < k^*$, $\text{EBIC}(s_{*k}) > \text{EBIC}(s_{*k+1})$ and $\text{EBIC}(s) > \text{EBIC}(s_0)$ for all s such that $p_0 < |s| \le k_0 p_0$ with any fixed $k_0 > 1$, $|s|$ denoting the number of features in s, (Theorem 3.3). These results imply the selection

consistency of the SLasso cum EBIC procedure. The asymptotic distribution of the SLasso estimators with diverging p_0 is given in Theorem 3.4, which justifies the second part of the oracle property.

The remainder of the article is arranged as follows. The basic properties of SLasso cum EBIC and its computation algorithm are given in §2. The theoretical properties of SLasso cum EBIC are studied in §3. Simulation studies comparing SLasso cum EBIC with various other methods are reported in §4. A real data analysis is provided in §5. The paper is concluded by a discussion of the similarities and differences between SLasso cum EBIC and other related methods in §6. Some technical details are provided in a supplementary document.

2 Basic properties and computation algorithm

We consider the scenario that both the total number of features and the number of relevant features diverge. We also allow the set of relevant features and their effects vary as n varies. For the sake of clarity, we do not index these quantities explicitly by n, but their dependence on n should be kept in mind. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be the design matrix. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\tau$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\tau$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\tau$. In matrix notation, model (1.1) is expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Let S denote the set of indices $\{1, 2, \cdots, p\}$. Let s be any subset of S. Denote by $\boldsymbol{X}(s)$ the matrix consisting of the columns of $\boldsymbol{X}$ with indices in s. Similarly, let $\boldsymbol{\beta}(s)$ denote the vector consisting of the corresponding components of $\boldsymbol{\beta}$. Let $\mathcal{R}(s)$ be the linear space spanned by the columns of $\boldsymbol{X}(s)$ and $\boldsymbol{H}(s)$ its corresponding projection matrix, i.e, $\boldsymbol{H}(s) = \boldsymbol{X}(s)[\boldsymbol{X}^\tau(s)\boldsymbol{X}(s)]^{-1}\boldsymbol{X}^\tau(s)$.

Proposition 2.1. *Let s_{*k} denote the index set of the features selected at the k-th step of SLasso. For $k \geq 1$ and any $\tilde{j} \in s_{*k}^c$, if $X(\{\tilde{j}\}) \in \mathcal{R}(s_{*k})$ then $\tilde{j} \notin s_{*k+1}$.*

Proof: If $\boldsymbol{X}(\{\tilde{j}\}) \in \mathcal{R}(s_{*k})$ then there exists an $\boldsymbol{a}_k$ such that $\boldsymbol{X}(\{\tilde{j}\}) = \boldsymbol{X}(s_{*k})\boldsymbol{a}_k$ and hence

$$\begin{aligned} l_{k+1} &= \|\boldsymbol{y} - \boldsymbol{X}(s_{*k})(\boldsymbol{\beta}(s_{*k}) + \beta_{\tilde{j}}\boldsymbol{a}_k) - \boldsymbol{X}(s_{*k}^c/\{\tilde{j}\})\boldsymbol{\beta}(s_{*k}^c/\{\tilde{j}\})\|_2^2 + \lambda(|\beta_{\tilde{j}}| + \sum_{j \in s_{*k}^c/\{\tilde{j}\}} |\beta|_j) \\ &= \|\boldsymbol{y} - \boldsymbol{X}(s_{*k})\tilde{\boldsymbol{\beta}}(s_{*k}) - \boldsymbol{X}(s_{*k}^c/\{\tilde{j}\})\boldsymbol{\beta}(s_{*k}^c/\{\tilde{j}\})\|_2^2 + \lambda(|\beta_{\tilde{j}}| + \sum_{j \in s_{*k}^c/\{\tilde{j}\}} |\beta|_j) \\ &\geq \|\boldsymbol{y} - \boldsymbol{X}(s_{*k})\tilde{\boldsymbol{\beta}}(s_{*k}) - \boldsymbol{X}(s_{*k}^c/\{\tilde{j}\})\boldsymbol{\beta}(s_{*k}^c/\{\tilde{j}\})\|_2^2 + \lambda \sum_{j \in s_{*k}^c/\{\tilde{j}\}} |\beta|_j. \end{aligned}$$

Thus when l_{k+1} is minimized $\beta_{\tilde{j}}$ must be 0, i.e., $\tilde{j} \notin s_{*k+1}$. □

Proposition 2.1 implies that, for any k, the matrix $X(s_{*k})$ is of full column rank. It also suggests that, in the SLasso procedure, any feature that is highly correlated with the features selected already will have little chance to be selected subsequently. This nature of SLasso is favorable when it is used for feature selection in ultra-high dimensional feature space where high spurious correlations present, see [11].

Proposition 2.2. *For $k \geq 1$, the minimization of l_{k+1} is equivalent to the minimization of*

$$\tilde{l}_{k+1} = \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|^2 + \lambda_{k+1} \sum_{j \in s_{*k}^c} |\beta_j|, \tag{2.1}$$

where $\tilde{\boldsymbol{y}} = [I - \boldsymbol{H}(s_{*k})]\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = [I - \boldsymbol{H}(s_{*k})]\boldsymbol{X}(s_{*k}^c)$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(s_{*k}^c)$.

Proof: Differentiating l_{k+1} with respect to $\boldsymbol{\beta}(s_{*k})$, we have

$$\frac{\partial l_{k+1}}{\partial \boldsymbol{\beta}(s_{*k})} = -2\boldsymbol{X}^\tau(s_{*k})\boldsymbol{y} + 2\boldsymbol{X}^\tau(s_{*k})\boldsymbol{X}(s_{*k})\boldsymbol{\beta}(s_{*k}) + 2\boldsymbol{X}^\tau(s_{*k})\boldsymbol{X}(s_{*k}^c)\boldsymbol{\beta}(s_{*k}^c).$$

Setting the above derivative to zero, we obtain

$$\hat{\boldsymbol{\beta}}(s_{*k}) = [\boldsymbol{X}^\tau(s_{*k})\boldsymbol{X}(s_{*k})]^{-1}\boldsymbol{X}^\tau(s_{*k})[\boldsymbol{y} - \boldsymbol{X}(s_{*k}^c)\boldsymbol{\beta}(s_{*k}^c)]. \tag{2.2}$$

Substituting (2.2) into $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ we have

$$\begin{aligned} l_{k+1} &= \|\boldsymbol{y} - \boldsymbol{X}(s_{*k})\boldsymbol{\beta}(s_{*k}) - \boldsymbol{X}_(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})\|^2 + \lambda_{k+1}\sum_{j\in s^c_{*k}}|\beta_j| \\ &= \|[\boldsymbol{y}-\boldsymbol{X}(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]-\boldsymbol{X}(s_{*k})[\boldsymbol{X}^\tau(s_{*k})\boldsymbol{X}(s_{*k})]^{-1}\boldsymbol{X}^\tau(s_{*k})[\boldsymbol{y}-\boldsymbol{X}(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]\|^2+\lambda_{k+1}\sum_{j\in s^c_{*k}}|\beta_j| \\ &= \|[\mathrm{I} - \boldsymbol{H}(s_{*k})][\boldsymbol{y} - \boldsymbol{X}(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]\|^2 + \lambda_{k+1}\sum_{j\in s^c_{*k}}|\beta_j| \\ &= \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|^2 + \lambda_{k+1}\sum_{j\in s^c_{*k}}|\beta_j|. \end{aligned}$$

□

As a by-product of the above proof, the components of $\hat{\boldsymbol{\beta}}(s_{*k})$ are almost surely nonzero since $\boldsymbol{y}$ is a vector of continuous random variables. This implies that $s_{*1} \subset s_{*2} \subset \cdots \subset s_{*k} \subset \cdots$; that is, the feature sets selected in the sequential steps are nested.

Proposition 2.3. *Let* $s_{TEMP} = \{j : j \in s^c_{*k}, |\tilde{\boldsymbol{y}}^\tau \boldsymbol{x}_j| = \max_{l\in s^c_{*k}}|\tilde{\boldsymbol{y}}^\tau \boldsymbol{x}_l|\}$. *If* s_{TEMP} *is a singleton, then the* $\boldsymbol{x}_j$ *with* $j \in s_{TEMP}$ *is the only feature with non-zero estimated coefficient in the minimization of (2.1); otherwise, the minimization of (2.1) is equivalent to the minimization of*

$$\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_{TEMP}\tilde{\boldsymbol{\beta}}_{TEMP}\|^2 + \lambda_{k+1}\sum_{j\in s_{TEMP}}|\beta_j|,$$

where $\tilde{\boldsymbol{X}}_{TEMP}$ *consists of the* $\tilde{\boldsymbol{x}}_j$ *with* $j \in s_{TEMP}$, $\tilde{\boldsymbol{\beta}}_{TEMP}$ *is the corresponding coefficient vector.*

This proposition follows from the proof of Theorem 3.1. Proposition 2.3 gives rise to the following computation algorithm.

SLasso cum EBIC algorithm:

- Initial Step: Standardize $\boldsymbol{y}$, $\boldsymbol{x}_j$, $j = 1, \ldots, p$, such that $\boldsymbol{y}^\tau \mathbf{1} = 0, \boldsymbol{x}_j^\tau \mathbf{1} = 0$ and $\boldsymbol{y}^\tau \boldsymbol{y} = n$, $\boldsymbol{x}_j^\tau \boldsymbol{x}_j = n$. Compute $\boldsymbol{x}_j^\tau \boldsymbol{y}$ for $j \in S$. Let

$$s_{\text{TEMP}} = \{j : |\boldsymbol{x}_j^\tau \boldsymbol{y}| = \max_{j' \in S} |\boldsymbol{x}_{j'}^\tau \boldsymbol{y}|\}.$$

  If s_{TEMP} is a singleton, let $s_{*1} = s_{\text{TEMP}}$, otherwise, apply `glmpath` to $\boldsymbol{y}$ and $\boldsymbol{X}(s_{\text{TEMP}})$ and extract the first feature with non-zero coefficient in the solution path, and let s_{*1} be the active set. Compute $\boldsymbol{I} - \boldsymbol{H}(s_{*1})$ and EBIC(s_{*1}).

- General Step: For $k \geq 1$, compute $\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{y}}$ for $j \in s_{*k}^c$, where $\tilde{\boldsymbol{y}} = [\boldsymbol{I} - \boldsymbol{H}(s_{*k})]\boldsymbol{y}$, $\tilde{\boldsymbol{x}}_j = [\boldsymbol{I} - \boldsymbol{H}(s_{*k})]\boldsymbol{x}_j$. Let

$$s_{\text{TEMP}} = \{j : |\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{y}}| = \max_{j' \in s_{*k}^c} |\tilde{\boldsymbol{x}}_{j'}^\tau \tilde{\boldsymbol{y}}|\}.$$

  If s_{TEMP} is a singleton, let $s_{*k+1} = s_{*k} \cup s_{\text{TEMP}}$, otherwise, apply `glmpath` to $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}(s_{\text{TEMP}})$ and extract the first feature with non-zero coefficient in the solution path, and let s_{*k+1} be s_{*k} union the active set. Compute EBIC(s_{*k+1}). If EBIC$(s_{*k+1}) >$ EBIC(s_{*k}), stop; otherwise, compute $\boldsymbol{I} - \boldsymbol{H}(s_{*k+1})$ and continue.

- When the process stops, the parameters in the selected model are estimated by their least squares estimates.

The EBIC for s_{*k}, $k = 1, 2, \ldots$, in the above algorithm is given by

$$\text{EBIC}(s_{*k}) = n \ln \left(\frac{\|[\boldsymbol{I} - \boldsymbol{H}(s_{*k})]\boldsymbol{y}\|_2^2}{n} \right) + |s_{*k}| \ln n + 2(1 - \frac{\ln n}{r \ln p}) \ln \binom{p}{|s_{*k}|},$$

where r is a positive number slightly bigger than 2, say $r = 2.1$. For more details on EBIC, see §3.3. The matrix $\boldsymbol{I} - \boldsymbol{H}(s_{*k+1})$ can be updated from $\boldsymbol{I} - \boldsymbol{H}(s_{*k})$ recursively. Suppose there are K active features with indices $\{j_m : m = 1, \ldots, K\}$ at step $k+1$. Denote by $J_m = \{j_1, \ldots, j_m\}$. Let $J_0 = \phi$. The recursive formula is given by

$$\boldsymbol{I} - \boldsymbol{H}(s_{*k} \cup J_m) = [\boldsymbol{I} - \boldsymbol{H}(s_{*k} \cup J_{m-1})] \left\{ \boldsymbol{I} - \frac{\boldsymbol{x}_{j_m} \boldsymbol{x}_{j_m}^\tau [\boldsymbol{I} - \boldsymbol{H}(s_{*k} \cup J_{m-1})]}{\boldsymbol{x}_{j_m}^\tau [\boldsymbol{I} - \boldsymbol{H}(s_{*k} \cup J_{m-1})] \boldsymbol{x}_{j_m}} \right\}, \quad (2.3)$$

The amount of computation in the above algorithm is minimal. The computation of the projection matrices does not involve any matrix inversion. The call for `glmpath` is in fact seldom invoked. As mentioned earlier that, at steps where the partial positive cone condition holds, the set s_{TEMP} is indeed the index set of the active features at those steps. Thus, the SLasso selects features at those steps by maximizing $|\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{y}}|$, which is the same as OMP.

3 The oracle property of SLasso cum EBIC

We assume in model (1.1) that the ϵ_i's are i.i.d. normal random variables with mean zero and variance σ^2. We consider the design matrix X either as a deterministic or a random matrix. Let $s_0 = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$. Assume $\ln p = O(n^\kappa)$ for some $\kappa > 0$ and $p_0 = |s_0| = O(n^c)$ for some $0 < c < 1$. We establish the oracle property of the SLasso cum EBIC procedure in this section. We first present the result that, with probability converging to 1, the SLasso selects all the relevant features before any irrelevant feature can be selected. Then we present the result that, with probability converging to 1, the SLasso 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 SLasso estimator is also provided. The proofs of these results are given in the supplementary document. Some special cases are discussed at the end of this section.

3.1 The case of deterministic design matrix

In the case of deterministic design matrix, suppose the columns of X are standardized. We now introduce some notations. For $s \subset S$, let $s^- = s^c \cap s_0$. If $s \subset s_0$ then s^- is the complement of s in s_0. For $s \subset s_0$, define

$$\gamma_n(j, s, \boldsymbol{\beta}) = \frac{1}{n} \boldsymbol{x}_j^\tau [\boldsymbol{I} - \boldsymbol{H}(s)] \boldsymbol{X} \boldsymbol{\beta}.$$

In fact, $\gamma_n(j, s, \boldsymbol{\beta})$ only depends on $\boldsymbol{\beta}(s^c)$. But for the ease of notation, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}(s^c)$ will be used interchangeably. Unless otherwise stated, $\boldsymbol{\beta}$ also denotes the unknown true value of the parameter vector. We make the following assumptions.

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

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

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

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

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

Assumption A1 is implied by the following condition

$$\|\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{X}}(s^-)[\tilde{\boldsymbol{X}}^\tau(s^-)\tilde{\boldsymbol{X}}(s^-)]^{-1}\|_1 < 1 - \eta, \forall j \in s_0^c, \tag{3.1}$$

where $\tilde{\boldsymbol{x}}_j = [\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{x}_j$ and $0 < \eta < 1$. The claim above follows because

$$\begin{aligned}
|\gamma_n(j, s, \boldsymbol{\beta})| &= \frac{1}{n}|\boldsymbol{x}_j^\tau[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{\mu}| \\
&= |\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{X}}(s^-)[\tilde{\boldsymbol{X}}^\tau(s^-)\tilde{\boldsymbol{X}}(s^-)]^{-1}\frac{1}{n}\tilde{\boldsymbol{X}}^\tau(s^-)[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{\mu}| \\
&\leq \|\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{X}}(s^-)[\tilde{\boldsymbol{X}}^\tau(s^-)\tilde{\boldsymbol{X}}(s^-)]^{-1}\|_1 \frac{1}{n}\|\tilde{\boldsymbol{X}}^\tau(s^-)[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{\mu}\|_\infty \\
&< (1 - \eta)\frac{1}{n}\|\tilde{\boldsymbol{X}}^\tau(s^-)[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{\mu}\|_\infty = (1 - \eta)\frac{1}{n}\max_{j\in s^-} |\boldsymbol{x}_j^\tau[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{\mu}| \\
&= (1 - \eta)\max_{j\in s^-} |\gamma_n(j, s, \boldsymbol{\beta})|,
\end{aligned}$$

where the strict inequality holds by (3.1).

Under assumption A1, the $\mathcal{A}_s$ in A2 is a subset of s_0. Assumption A2 holds if and only if

$$\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{X}}(\mathcal{A}_s\backslash\{j\})[\tilde{\boldsymbol{X}}^\tau(\mathcal{A}_s\backslash\{j\})\tilde{\boldsymbol{X}}(\mathcal{A}_s\backslash\{j\})]^{-1}\mathbf{1} < 1, \forall j \in \mathcal{A}_s. \tag{3.2}$$

We establish the equivalence of A2 and (3.2) below. Let $\boldsymbol{A} = \tilde{\boldsymbol{X}}(\mathcal{A}_s \backslash \{j\})$ and $\boldsymbol{b} = \tilde{\boldsymbol{x}}_j$. Since a permutation of the rows and columns does not change the sum of the rows, it suffices to verify that the sum of the last row of $\begin{pmatrix} \boldsymbol{A}^\tau \boldsymbol{A} & \boldsymbol{A}^\tau \boldsymbol{b} \\ \boldsymbol{b}^\tau \boldsymbol{A} & \boldsymbol{b}^\tau \boldsymbol{b} \end{pmatrix}^{-1}$ is positive if and only if $\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\mathbf{1} < 1$. Let $\boldsymbol{E} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\boldsymbol{A}^\tau$ and $\boldsymbol{F} = \boldsymbol{I} - \boldsymbol{b}(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau$. By the formula for the inverse of blocked matrices, we have

$$\begin{pmatrix} \boldsymbol{A}^\tau \boldsymbol{A} & \boldsymbol{A}^\tau \boldsymbol{b} \\ \boldsymbol{b}^\tau \boldsymbol{A} & \boldsymbol{b}^\tau \boldsymbol{b} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{A}^\tau \boldsymbol{F} \boldsymbol{A})^{-1} & -(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\boldsymbol{A}^\tau \boldsymbol{b}(\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1} \\ -(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{F} \boldsymbol{A})^{-1} & (\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1} \end{pmatrix}.$$

and

$$\begin{aligned} (\boldsymbol{A}^\tau \boldsymbol{F} \boldsymbol{A})^{-1} &= [\boldsymbol{A}^\tau \boldsymbol{A} - \boldsymbol{A}^\tau \boldsymbol{b}(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}]^{-1} \\ &= (\boldsymbol{A}^\tau \boldsymbol{A})^{-1} + (\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\boldsymbol{A}^\tau (\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}. \end{aligned}$$

Substituting the expression of $(\boldsymbol{A}^\tau \boldsymbol{F} \boldsymbol{A})^{-1}$ into the first block of the last row of the above matrix, we obtain

$$-(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{F} \boldsymbol{A})^{-1} = -(\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}.$$

Thus the sum of the last row becomes

$$(\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1} - (\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1}\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\mathbf{1} = (\boldsymbol{b}^\tau \boldsymbol{E} \boldsymbol{b})^{-1}[1 - \boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\mathbf{1}]$$

which is greater than 0 if and only if $\boldsymbol{b}^\tau \boldsymbol{A}(\boldsymbol{A}^\tau \boldsymbol{A})^{-1}\mathbf{1} < 1$.

Condition (3.1) is a conditional version of the exact recovery condition (ERC) assumed in [29] while conditioning on the subset s of the relevant features. Condition (3.2) is similar to but much weaker than the *irrepresentability condition*. The above arguments suggest that Conditions A1 and A2 might be weaker than the ERC and the *irrepresentability condition*. This is indeed the case. We will demonstrate this by special cases where the conditions for the selection consistency of the SLasso hold but the ERC and the *irrepresentability condition* are not satisfied. If $\lambda_{\min}(\frac{1}{n}\boldsymbol{X}^\tau(s_0)\boldsymbol{X}^\tau(s_0))$ is bounded away from zero, which is a common assumption in the case of ultra-high

dimensional feature space, then Condition A3 is equivalent to $\frac{\sqrt{n}}{\ln p_n}\min_{j\in s_0}|\beta_j| \to \infty$. If $\ln p = O(n^\kappa)$ with $\kappa < 1/2$ and $\min_{j\in s_0}|\beta_j| \geq Cn^{-\delta}$ for some constant C and $\delta < 1/2-\kappa$, A3 is then satisfied.

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

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

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

3.2 The case of random design matrix

Assume $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\tau$, $i = 1, \ldots, n$, are i.i.d. copies of a random vector $\boldsymbol{z} = (z_1, \ldots, z_p)^\tau$. Without loss of generality, assume that $E\boldsymbol{z} = 0$ and $\mathrm{Var}(\boldsymbol{z}) = \Sigma$ with diagonal elements 1 and off-diagonal elements independent of n. Assume that

a1 The off-diagonal elements of Σ 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 $\max_{1\leq j,k\leq p} E\exp(tz_j z_k)$ and $\max_{1\leq j\leq p} E\exp(tz_j\epsilon)$ are finite for t in a neighborhood of zero.

For any $s, \tilde{s} \subset S$, denote by $\Sigma_{s\tilde{s}}$ the sub matrix of Σ with row indices in s and column indices in $\tilde{s}$. Define

$$\Gamma(j, s, \boldsymbol{\beta}) = (\Sigma_{jS} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sS})\boldsymbol{\beta}.$$

The following assumptions are imposed:

A1$'$ For any $s \subset s_0$, $s \neq s_0$, $\max_{j \in s_0^c} |\Gamma(j, s, \boldsymbol{\beta})| < \max_{j \in s^-} |\Gamma(j, s, \boldsymbol{\beta})|$.

A2$'$ Let $\mathcal{A}_s = \{j : j \in s^c, |\Gamma(j, s, \boldsymbol{\beta})| = \max_{l \in s^c} |\Gamma(l, s, \boldsymbol{\beta})|\}$. Then

$$(\Sigma_{\mathcal{A}_s \mathcal{A}_s} - \Sigma_{\mathcal{A}_s s} \Sigma_{ss}^{-1} \Sigma_{s \mathcal{A}_s})^{-1} \mathbf{1} > 0.$$

A3$'$ $\frac{n^{1/2}}{\ln p} \lambda_{\min}(\Sigma_{s_0 s_0})(\min_{j \in s_0} |\beta_j|) \to +\infty$ as $n \to +\infty$.

The assumptions A1$'$ - A3$'$ are in fact the assumptions A1 - A3 with the empirical variances and covariances of the features replaced by their theoretical counterparts. In order to establish the selection consistency of SLasso in the case of random feature matrix, we need to pass from assumptions A1$'$ - A3$'$ to assumptions A1 - A3. The following lemma ensures that if A1$'$ - A3$'$ hold then A1 - A3 hold with probability converging to 1 as n goes to infinity.

Lemma 3.1. *Under assumptions a*1 *- a*3*,*

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

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

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

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

The proof of the lemma is given in the supplementary document.

Theorem 3.2. *Let* $\ln p = O(n^\kappa)$*,* $\kappa < 1/3$*, and* $p_0 = O(n^c)$*,* $\kappa/2 < c < 1/6$*. Assume that conditions a*1 *- a*3 *and A*1$'$*-A*3$'$ *are satisfied. Then, there is a* k^* *such that*

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

The theorem is in fact a corollary of Lemma 3.1. It follows from the lemma immediately that if $a1$ - $a3$ and A1$'$-A3$'$ are satisfied then A1-A3 hold with probability converging to 1. Thus the selection consistency of SLasso with random feature matrix is established.

3.3 Property of the stopping rule and the asymptotic distribution of the SLasso estimators

For a linear model with features in s, the EBIC proposed in [3] is defined as

$$\mathrm{EBIC}_\gamma(s) = n \ln\left(\frac{\|[\boldsymbol{I} - \boldsymbol{H}(s)]\boldsymbol{y}\|_2^2}{n}\right) + |s| \ln n + 2\gamma \ln \binom{p}{|s|}, \ \gamma \geq 0.$$

The properties of EBIC for sparse high-dimensional linear models are investigated in [3] and [20]. It is shown that, if $\gamma > 1 - \ln n/(2 \ln p)$, EBIC is selection consistent in the sense that

$$P(\min_{|s| \leq k_0 p_0} \mathrm{EBIC}_\gamma(s) > \mathrm{EBIC}_\gamma(s_0)) \to 1, \text{ as } n \to \infty,$$

where $k_0 > 1$ is any fixed number.

For the sequence $s_{*1}, s_{*2}, \cdots, s_{*k}, \cdots$ selected by the procedure of SLasso, we have shown that $s_{*1} \subset s_{*2} \subset \cdots \subset s_{*k} \subset \cdots$ and that, with probability converging to 1, there is a k^* such that $s_{*k^*} = s_0$. In this sub section, we provide the result that, with probability converging to 1, $\mathrm{EBIC}(s_{*k})$ decreases when $k < k^*$ and reaches its minimum at step k^*, and that $\mathrm{EBIC}(s_{*k}) > \mathrm{EBIC}(s_{*k^*})$ for any $k > k^*$. The result is given in Theorem 3.3. This result implies that, with probability converging to 1, the procedure of SLasso cum EBIC stops at step k^*. The parameters in the selected model are estimated by their least squares estimates. Theorems 3.1 - 3.3 imply that the estimators are obtained as if s_0 were known in advance. This implies that the SLasso cum EBIC procedure possesses the oracle property. Since p_0 diverges, the asymptotic theory on ordinary least squares estimators with fixed p_0 does not apply. We derive the asymptotic distribution of the SLasso estimator of $\boldsymbol{\beta}(s_0)$ in Theorem 3.4.

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* $\lambda_{\min}(\frac{1}{n}\boldsymbol{X}(s_0)^\tau \boldsymbol{X}(s_0)) \min_{j \in s_0} |\beta_j|$

$\geq Cn^{-1/6+\delta}$, *where* δ *is an arbitrarily small positive number. Let* $s_{*1} \subset s_{*2} \subset \cdots \subset s_{*k} \subset \cdots$ *be the sets generated by the procedure of SLasso. Let* k^* *be as given in Theorem 3.1 and 3.2. Then*

(i) *Uniformly, for* $k < k^*$,

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

(ii) $P(\min_{p_0<|s_{*k}|\leq k_0 p_0} EBIC_\gamma(s_{*k}) > EBIC_\gamma(s_0)) \to 1$, *when* $\gamma > 1 - \frac{\ln n}{2\ln p}$, *where* $k_0 > 1$ *is an arbitrarily fixed constant.*

The proof of the theorem is given in the supplementary document.

In the stopping rule, γ is taken as $1 - \frac{\ln n}{r \ln p}$ where r is slightly bigger than 2. This choice of γ is to keep the EBIC selection consistent at one hand and to achieve the largest power for the identification of relevant feature at another hand. A brief justification is given as follows. For a sample of size n, define the positive discovery rate (PDR_n) and the false discovery rate (FDR_n) as follows.

$$\text{PDR}_n = \frac{|s_{*k^*} \cap s_0|}{|s_0|}, \quad \text{FDR}_n = \frac{|s_{*k^*} \cap s_0^c|}{|s_{*k^*}|}. \tag{3.3}$$

The asymptotic property $P(s_{*k^*} = s_0) \to 1$ is equivalent to that $\text{FDR}_n \to 0$ and $\text{PDR}_n \to 1$ simultaneously. For any $\gamma > 1 - \frac{\ln n}{2\ln p_n}$, the above convergences are guaranteed, but the convergence rates are different for different γ values. For a bigger γ, both FDR_n and PDR_n are smaller. By choosing the γ as small as possible in its consistent range, the PDR_n is maximized while the FDR_n still converges to zero.

Let s^* be the set selected by SLasso cum EBIC and $\hat{\boldsymbol{\beta}}(s^*)$ the SLasso estimator of $\boldsymbol{\beta}(s^*)$ (which is indeed the least squares estimator). Let $\boldsymbol{a} = (a_1, a_2, \ldots,)$ be an infinite sequence of constants. For any index set s, let $\boldsymbol{a}(s)$ denote the vector with components $a_j, j \in s$. We have the following theorem.

Theorem 3.4. *Let $\boldsymbol{z}_i^\tau$ be the ith row vector of $\boldsymbol{X}(s_0)$, $i = 1, \ldots, n$. Assume that*

$$\lim_{n\to\infty} \max_{1\le i\le n} \boldsymbol{z}_i^\tau [\boldsymbol{X}(s_0)^\tau \boldsymbol{X}(s_0)]^{-1} \boldsymbol{z}_i \to 0. \tag{3.4}$$

Then, for any fixed sequence $\boldsymbol{a}$,

$$\frac{\boldsymbol{a}(s^*)^\tau [\hat{\boldsymbol{\beta}}(s^*) - \boldsymbol{\beta}(s^*)]}{\sqrt{\boldsymbol{a}(s^*)^\tau [\boldsymbol{X}(s^*)^\tau \boldsymbol{X}(s^*)]^{-1} \boldsymbol{a}(s^*)}} \to_d N(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}(Y_i)$.

Let $\hat{\boldsymbol{\beta}}(s_0) = [\boldsymbol{X}(s_0)^\tau \boldsymbol{X}(s_0)]^{-1} \boldsymbol{X}(s_0)^\tau \boldsymbol{y}$. Then under (3.4),

$$V_n = \frac{\boldsymbol{a}(s_0)^\tau [\hat{\boldsymbol{\beta}}(s_0) - \boldsymbol{\beta}(s_0)]}{\sqrt{\boldsymbol{a}(s_0)^\tau [\boldsymbol{X}(s_0)^\tau \boldsymbol{X}(s_0)]^{-1} \boldsymbol{a}(s_0)}} \to_d N(0, \sigma^2),$$

which follows from the Linderberg's central limit theorem, see Corollary 1.3 in [26]. Its proof is by checking the validity of the conditions for this corollary, which is straightforward and is omitted here. Let $U_n = \frac{\boldsymbol{a}(s^*)^\tau [\hat{\boldsymbol{\beta}}(s^*) - \boldsymbol{\beta}(s^*)]}{\sqrt{\boldsymbol{a}(s^*)^\tau [\boldsymbol{X}(s^*)^\tau \boldsymbol{X}(s^*)]^{-1} \boldsymbol{a}(s^*)}}$. Since $P(U_n \neq V_n) = P(s^* \neq s_0) \to 0$, we have that $U_n - V_n \to 0$ in probability. Thus, by Slutsky's theorem, $U_n = V_n + (U_n - V_n) \to_d N(0, \sigma^2)$. Theorem 3.4 implies that any fixed dimensional sub vector of $\hat{\boldsymbol{\beta}}(s^*)$ has an asymptotic multivariate normal distribution.

3.4 Special cases

We give two special cases which demonstrate that the conditions required for the oracle property of SLasso is weaker than the well-known *irrepresentability condition*.

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

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

where I is the identity matrix of dimension p, $\mathbf{1}$ is a p-vector of all elements 1, and $0 < \rho \le \rho_0 < 1$. In this case, for the *irrepresentability condition* to be satisfied, some restriction must be imposed. But such restriction is not needed for sequential Lasso.

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

(i) $|\beta_1| > |\beta_2| > \cdots > |\beta_{p_0}| = Cn^{-1/2+\delta}$ for some constant C and an arbitrarily small positive δ;

(ii) The correlation matrix Σ has the following structure:

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

In this case, the *irrepresentability condition* is violated but the conditions for the oracle property of SLasso are satisfied.

The verification for the claims on the two special cases are provided in the supplementary document.

4 Simulation Study

In our simulation study, we compare SLasso with adaptive Lasso (ALasso) [5], SCAD [16],[33], SIS+SCAD [11], forward stepwise regression (FSR) [31] and a forward-backward greedy algorithm (FoBa) [36]. The first three competing methods have also been shown to have the oracle property (or selection consistency) under certain conditions. The FSR and FoBa are included in the comparison because of their close relationship with SLasso. We do not include CoSaMP in the simulation study since no concrete approach for the selection of models was given in [22]. For ALasso, SCAD, SIS+SCAD and FoBa, cross-validation is used to determine the final selected model as in the cited references. For SLasso and FSR, EBIC is used as the stopping rule. The R packages parcor ([18]), ncvreg ([1]) , SIS ([12]) and foba are used for the computation of ALasso, SCAD , SIS+SCAD and FoBa respectively.

We take two groups of settings: group A and group B. In group A, we consider the diverging pattern $(n, p, p_0) = (n, [4n^{0.16}], [5e^{n^{0.3}}])$ for $n = 100, 200$, which is in

consistence with the assumption in the theory of SLasso. The coefficients are generated as independent random variables distributed as $(-1)^u(4n^{-0.15} + |z|)$, where $u \sim Bernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. The absolute values of the coefficients are roughly of order $O(n^{-0.15})$. The variance of the error term in the linear model is determined by the following equation:

$$h = \frac{\boldsymbol{\beta}^\tau \Sigma \boldsymbol{\beta}}{\boldsymbol{\beta}^\tau \Sigma \boldsymbol{\beta} + \sigma^2} = 0.8,$$

where Σ is the variance-covariance matrix of relevant features. Five settings of the covariance structure for the design matrix X are considered. They are named GA1, GA2, $\cdots$, GA5. In group B, three settings named GB1, GB2 and GB3, which are adapted from the cited references, are considered. In these settings, the triplet (n, p, p_0) does not follow the diverging pattern. The details of the above simulation settings are provided in the supplementary document.

The methods are compared in terms of PDR, FDR, model size (Msize) and prediction mean square error (PMSE). The definition of PDR and FDR are given in (3.3). The MSize is the number of features selected. The PMSE is the average squared differences between the observations in a sample and their predicted values obtained using the model built from another independent sample. Thus, for the computation of PMSE, we generate two independent samples with the same sample size n in each replicate of the settings. One sample is used for the selection of features and the estimation of the coefficients, and the other one is used to compute the PMSE. For the settings in group A, these quantities are averaged over 200 replicates, for those in group B, they are averaged over 500 replicates. The results for group A are reported in Tables 6.1 and 6.2. The results for group B are reported in Table 6.3.

The findings of the simulation study are summarized as follows. First, consider the performance in prediction. For settings of Group A, ALasso, SCAD, FSR and SLasso

have comparable PMSE, however, FSR and SLasso have smaller (in GA1 and GA2, much smaller) MSize. The other two, SIS+SCAD and FoBa, have much larger PMSE. SIS+SCAD has smaller MSize than all the others and FoBa has much larger MSize than all the others. FSR and SLasso always have about the same MSize, but SLasso has smaller PMSE than FSR except in setting GA2. For settings of Group B, in GB1 and GB2, the three methods, SCAD, FSR and SLasso, have comparable PMSE which are much smaller then the other three methods. SCAD has larger MSize than FSR and SLasso whose MSize are about the same. In GB3, ALasso, FoBa and FSR have much smaller PMSE but also have much larger MSize than SCAD, SIS+SCAD and SLasso.

Now, consider the performance in the identification of relevant features. First, let our attention be drawn to the performance of FoBa. In all the settings, FoBa has extremely high FDR which are much higher than all the other methods. In settings of Group A with sample size 100 and 200, its minimum FDR is respectively 0.909 and 0.949. In the settings of Group B, its minimum FDR is 0.852. Whatever high PDR it might achieve cannot be justified with such high FDRs. Let alone the fact that its PDR is even lower than at least one of the other methods in all the settings except in GA2 with sample size 200 and GB3. FoBa fails the goal of identifying relevant features. The poor performance of FoBa is not surprising. Though it has essentially the same mechanism for the selection of features as SLasso, it has an improper stopping rule. FoBa stops at a forward step when the decrease in the residual, $\frac{1}{n}\|\tilde{\boldsymbol{y}}\|_2^2$, falls below a threshold of order $O(\sigma^2 \ln p/n)$ without penalizing the increase in the number of features. The backward step is hardly activated since it is in force only when the increase in the residual by deleting one of the selected features is less than half of the amount of decrease at the forward step. See §1 and §6. In a SHR model, $\frac{1}{n}\|\tilde{\boldsymbol{y}}\|_2^2$ can be reduced to zero, it is easy for a feature, relevant or not,

to reduce $\frac{1}{n}\|\tilde{\boldsymbol{y}}\|_2^2$ by an amount larger than the threshold. Thus, features with high spurious correlation with the response can be easily selected. Note that the order of threshold, $O(\sigma^2 \ln p/n)$, decreases as n increases, which suggests that, the larger the sample size, the more features can be selected and hence the higher the FDR. This is in fact demonstrated in the simulation study, when the sample size goes from 100 to 200 in the settings of Group A, the minimum FDR of FoBa goes from 0.909 to 0.949. It is interestingly contrasted with SLasso cum EBIC whose FDR decreases as n increases, which is also demonstrated in the simulation study with the settings of Group A.

In what follows, we compare the performance of other methods. Under settings GA1 and GA2, which are common settings in many simulation studies, SLasso and FSR are better than the other methods. They have high PDR and very low FDR. SLasso and FSR are comparable while FSR is slightly better. ALasso and SCAD have higher PDR than SLasso and FSR, but their FDR are too much higher. Averaged over the 4 simulations, FSR, SLasso, ALasso and SCAD have averaged PDR 0.847, 0.823, 0.950 and 0.913 respectively, and averaged FDR 0.042, 0.067, 0.741 and 0.459 respectively. The difference in PDR between FSR, SLasso and ALasso, SCAD, Foba is not too much, but the difference in FDR is strikingly large. FSR and SLasso are absolutely much better than SIS+SCAD in terms of both PDR and FDR.

Under settings GA3 - GA5, SCAD is absolutely better than all the other methods. The performance of SLasso is close to SCAD. SLasso is absolutely better than SIS+SCAD and FSR. Though ALasso has a slightly higher PDR in a few cases, its FDR is too high to be acceptable in terms of the identification of relevant features. It is not surprising that SLasso is absolutely better than FSR. In settings GA3 - GA5, all the irrelevant features are equally and highly correlated with the relevant features. In these situations, FSR is more prone to error compared with SLasso, since FSR

tends to select features which are highly correlated with the features already selected though they might have less correlation with the current residuals, see the discussion in §6.

Under settings GB1 and GB2, the pattern is similar to that under settings GA1 and GA2. Under setting GB3, though the condition (which is sufficient but not necessarily necessary) for the selection consistency of SLasso is not satisfied, SLasso performs better than all the other methods: it has comparable or higher PDR than other methods and has the lowest (much lower than the others) FDR.

The simulation study demonstrates that (i) the SLasso cum EBIC method is one of the best feature selection methods for the purpose of prediction, and (ii) in terms of the identification of relevant features the performance of SLasso is satisfactory and robust: it always has a very low FDR and it is always close to the best, though it is not the best over all the simulation settings. On the contrast, the performance of SCAD and FSR are erratic over the settings. They are the best in certain settings but perform much worse in other settings.

5 Real Data Analysis

The data, which was reported in [25], consists of the expression levels of over 31,042 different probes from 120 F_2 male rats generated from an intercross experiment. A cross of SR/JrHsd male rats and SHRSP female rats was performed to generate F_1 and the F_1 rats were intercrossed to generate the F_2 rats. The probes that were not expressed in the eye or that lacked sufficient variation were excluded. A probe was considered expressed if its maximum expression value observed among the 120 F2 rats was greater than the 25th percentile of the entire set of RMA (robust multi-chip averaging) expression values. A probe was considered "sufficiently variable" if it exhibited at least 2-fold variation in expression level among the 120 F2 rats. A total

of 18,976 probes that met these criteria were retained. Among the 18,976 probes, there is one, 1389163_at, from gene TRIM32. This gene was found to cause Bardet-Biedl syndrom [4]. Of interest is to find the probes among the remaining 18, 975 probes that are most related to TRIM32. This has been studied by using different methods in the literature, see [5], [16], [6], [13] and [27]. In this section, we apply the five methods considered in our simulation study, i.e., ALasso, SCAD, SIS+SCAD, FSR and SLasso, to the above problem. The response variable is the expression level of probe 1389163_at. The features are the expression levels of the remaining 18, 975 probes. The expression levels are standardized to have mean 0 and standard deviation 1 in the analysis.

Following the same strategy of [5], the probes are first screened according to their variances and the top 3,000 probes with the largest variances are retained for further selection. But, unlike in [5] where these 3,000 probes were further reduced to 200 probes that are marginally most correlated with TRIM32, the concerned five methods are directly applied to the 3,000 probes. The numbers of probes selected from these 3,000 probes by ALasso, SCAD, SIS+SCAD, FSR and SLasso are 21, 28, 5, 3 and 2 respectively. The ID of the selected probes are reported in Table 6.4. The two probes selected by SLasso, i.e., 1383110_at and 1392692_at, are also selected by FSR and ALasso. But they are not selected by SCAD and SIS+SCAD. The additional probe selected by FSR, 1389584_at, is also selected by SCAD and ALasso. There is an intersection of 7 probes selected by ALasso and SCAD. There is no intersection of the probes selected by SIS+SCAD with any other methods.

It is interesting to note that one of the probes selected by SLasso, i.e., 1383110_at, is also detected by other methods (Lasso, Scaled Lasso, Scaled MC) and the other one, 1392692_at, is also detected by Lasso, as reported in [27]. Combining all these findings together and taking into account the low FDR of SLasso evidenced in the

simulation studies, we have a strong belief that the two probes selected by SLasso are associated with TRIM32.

6 Discussion

The properties of SLasso shows that SLasso and the orthogonal matching pursuit (OMP) differ only at steps where the partial positive cone condition is violated. When the partial positive cone condition is satisfied at each step, SLasso is equivalent to OMP. Since the set s_{TEMP} at the steps of SLasso is rarely non-singleton, the procedure of SLasso is essentially the same as OMP, and hence, as a by-product of the paper, we reveal new properties of OMP other than those discovered in [2] [29] [30] under much weaker conditions.

Since the mechanism for the selection of features in FoBa and CoSaMP are essentially the same as OMP, our results also reveal that the backward steps in FoBa are indeed not needed and that the iterative procedure of CoSaMP to get the best sparse set at a given sparsity level is not really necessary. The crucial issue in all these procedures is actually the stopping rule. Some ad hoc stopping rules are adopted for OMP in [2]. These rules compare the norm (L_2 or L_∞) of the residual $\tilde{\boldsymbol{y}}$ with an upper bound of the norm of error $\boldsymbol{\epsilon}$. The procedure continues until the norm of the residual falls below the upper bound. For example, when $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, $\|\tilde{\boldsymbol{y}}\|_2$ is compared with $\sigma\sqrt{n + 2\sqrt{n \ln n}}$. With this rule, the OMP can correctly select s_0 with probability $1 - 1/n$, if $\min_{j \in s_0} |\beta_j| \geq \frac{2\sigma\sqrt{n+2\sqrt{n \ln n}}}{1-(2p_0-1)\text{MI}}$ and $\text{MI} < \frac{1}{2p_0-1}$, where MI is the mutual incoherence defined by $\text{MI} = \max_{i \neq j} |\text{CORR}(\boldsymbol{x}_i, \boldsymbol{x}_j)|$. The limitations of this rule are obvious. First, its effectiveness relies on the strict condition $\text{MI} < \frac{1}{2p_0-1}$ which essentially imposes mutual independence among the features when $p_0 \to \infty$, as assumed in our setting. Second, the lower bound required for $\min_{j \in s_0} |\beta_j|$ is too large compared with the requirement in SLasso that $\min_{j \in s_0} |\beta_j| > Cn^{-\delta}$ where $\delta < 1/2 - \kappa$

for some $\kappa < 1/2$, see condition A3 and the remark that follows in §3.1. Third, it requires an accurate estimate of σ which cannot be easily obtained without knowing the true model. The other ad hoc stopping rules considered in [2] have the same limitations. A similar stopping rule is considered in [36] for FoBa. The rule compares the decrease in $\frac{1}{n}\|\tilde{\boldsymbol{y}}\|_2^2$ with a threshold of order $O(\sigma^2 \ln p/n)$ at the forward steps. Similar comments can be made on this stopping rule. A common nature of the above stopping rules is that only the contribution of the features to the decrease of the residual is taken into account, the contribution is not penalized by the increase in the number of features. This common nature is a crucial drawback. It has the potential to select more irrelevant features, which is demonstrated for FoBa in the simulation studies. As a contrast, by using EBIC as the stopping rule, a feature can be selected only when its contribution to the decrease of the residual is large enough to compensate the increase in the number of features. This is perhaps the main reason why SLasso cum EBIC is selection consistent under much weaker conditions.

Let $g_1(j) = |\boldsymbol{x}_j^\tau[\boldsymbol{I} - \boldsymbol{H}(s_{*k})]\boldsymbol{y}|$, where $\boldsymbol{H}(s_{*k})$ is the projection matrix of the space spanned by the features in s_{*k}. SLasso selects the next features among the features that maximize $g_1(j)$ after the sub model s_{*k} is selected. This is to be compared with FSR that selects the next feature by minimizing $\mathrm{RSS}(j) = \boldsymbol{y}^\tau[\boldsymbol{I} - \boldsymbol{H}(s_{*k} \cup \{j\})]\boldsymbol{y}$ which is equivalent to maximizing $g_2(j) = \frac{|\boldsymbol{x}_j^\tau[\boldsymbol{I}-\boldsymbol{H}(s_{*k})]\boldsymbol{y}|}{\sqrt{\boldsymbol{x}_j^\tau[\boldsymbol{I}-\boldsymbol{H}(s_{*k})]\boldsymbol{x}_j}}$. The equivalence follows from (2.3). SLasso selects the next feature that has the highest correlation with the current residual but the FSR selects the next feature that has the highest inflated correlation with the current residual by an inflating factor $[\boldsymbol{x}_j^\tau[\boldsymbol{I} - \boldsymbol{H}(s_{*k})]\boldsymbol{x}_j]^{-1/2}$. The more correlated the $\boldsymbol{x}_j$ is with the features in s_{*k}, the larger the inflating factor. If two features have the same absolute correlation with the current residual, the FSR will select the one that is more correlated with the features in s_{*k}. If one feature has a lower correlation with the current residual but is more correlated with the features

in s_{*k} than another feature, it might turn out that this feature has a higher inflated correlation and is selected by FSR. Obviously, this is a disadvantage of FSR in terms of the identification of relevant features, especially when high spurious correlations present.

Like SLasso, solution path of Lasso, LAR and variants of LAR also select the next feature that has the highest correlation with the current residual. But, in these methods, the current residual is obtained from a shrunk estimate of $E\boldsymbol{y}$, i.e., they select $\boldsymbol{x}_j$ that maximizes $g_3(j) = |\boldsymbol{x}_j^\tau[\boldsymbol{y} - \boldsymbol{X}(s_{*k})\tilde{\boldsymbol{\beta}}(s_{*k})]|$ where $\tilde{\boldsymbol{\beta}}(s_{*k})$ is a shrunk estimate. In the shrunk estimate, the effects on $\boldsymbol{y}$ of the features in s_{*k} are not fully counted. This leaves more chance for those features that have high spurious correlations with the features in s_{*k} to be selected in subsequent steps than in the case of SLasso. This is a potential disadvantage for the identification of relevant features.

Acknowledgment

References

[1] Breheny, P. (2011). ncvreg: Regularization paths for SCAD- and MCP-penalized regression models. R package version 2.3-2. http://cran.r-project.org/web/packages/ncvreg/index.html.

[2] Cai, T.T and Wang, L. (2011). Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise. *IEEE Transactions on Information Theory*, **57**, 4680-4688.

[3] Chen, J. H. , Chen, Z. H. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika.* **95**, 759-771.

[4] Chiang, A. P., Beck, J.S., Yen, H.-J. , Tayeh, M.K., Scheetz, T.E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T.L., Stone, E. M., and Sheffield, V. C. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a BardetCBiedl syndrome gene (BBS11). *Proc. Nat.Acad.Sci*, **103**, 6287-6292.

[5] Huang, J., Ma, S. and Zhang, C-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.

[6] Hwang, W.Y. , Zhang, H.H. , and Ghosal, S. (2009). FIRST: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Statistics and its interface*, **2**, 341-348.

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-499.

[8] Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348-1360.

[9] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

[10] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

[11] Fan, J., Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Stat. Soc. B.* **70**, 849-911.

[12] Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). SIS: Sure Independence Screening. R package version 0.6. http://cran.r-project.org/web/packages/SIS/index.html.

[13] Fan, J.Q., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. American Statistical Association*, **116**, 544-557.

[14] Fill, J.A. (1983). Convergence rates related to the strong law of large numbers. *Ann. Prob.* **11**, 123-142.

[15] Friedman, J., Hastie, T. and Tibshirani, R. (2010). glmnet: Lasso and elastic-net regularized generalized linear models R package version 1.7.1. http://cran.r-project.org/web/packages/glmnet/index.html.

[16] Kim,Y.D., Choi, H., and Oh, H-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. American. Statistical. Association*, **103**, 1665-1673.

[17] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann.Statist.* **28**, 1356-1378.

[18] Kraemer, N. and Schaefer, J. (2010). parcor: Regularized estimation of partial correlation matrices. R package version 0.2-2. http://cran.r-project.org/web/packages/parcor/index.html.

[19] Leng, C., Lin, Y. and Wahba, G. (2004). A note on the Lasso and related procedures in model selection. *Statistica Sinica* **16**, 1273-1284.

[20] 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.

[21] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

[22] Needell, D. Tropp, J. A. (2009). CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis* **26**, 301-321.

[23] Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc. B* **69**, 659-677.

[24] Radchenko, P. and James, G. (2011). Improved variable selection with forward-LASSO adaptive shrinkage. *The Annals of Applied Statistics* **5**, 427-448.

[25] Scheetz, T.E., Kim, K.-Y.A., Swiderski, R.E., Philip1, A.R., Braun, T.A., Knudtson, K.L., Dorrance, A.M., et al. (2006). Regularization of gene expression in the mammalian eye and its relevance to eye disease. *Proc.Nat.Acad.Sci*, **103**, 14429-14434.

[26] Shao, J. (2003). *Mathematical Statistics*. 2nd edition. Springer.

[27] Sun, T.N., and Zhang, C.H. (2011). Scaled sparse linear regression. arXiv:1104.4595v1.

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. B (Methodological)* **58**, 267-288.

[29] Tropp, J.A. (2004). Greed is good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory.* **50**,1-21.

[30] Tropp, J.A., Gilbert, A.C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory.* **53**, 4655-4666.

[31] Wang, H. (2009). Forward Regression for Ultra-High Dimensional Variable Screening. *J. Amer. Statist. Assoc.* **104**, 1512-1524.

[32] Weisberg, S. (1980). *Applied Linear Regression*. Wiley, New York.

[33] Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

[34] Pati, Y. C., Rezaiifar, R. and Krishnaprasad, P.S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, 40-44. IEEE.

[35] Zhang, C-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**, 894-942.

[36] Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *Information Theory, IEEE Transactions*, **57(7)**, 4689-4708.

[37] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research* **7**, 2541-2567.

[38] Zou, H. (2006). The adaptive lasso and its oracle properties *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Table 6.1: Comparison of SLasso with ALasso, SCAD, SIS+SCAD and FSR in terms of PDR, FDR, PMSE and model size (MSize) averaged over 200 simulation replicates in Group A settings (sample size $n = 100$, numbers in parentheses are standard deviations)

Setting	Methods	MSize	PDR	FDR	PMSE
GA1	ALasso	37.8(12.5)	.999(.009)	.767(.072)	13.714(2.837)
	SCAD	14.9(3.8)	.998(.022)	.430(.142)	10.523(2.943)
	SIS+SCAD	5.0(0.1)	.488(.098)	.219(.157)	25.135(5.091)
	FoBa	87.3(1.3)	.998(.020)	.909(.002)	32.583(6.759)
	FSR	8.3(1.4)	.963(.149)	.063(.098)	11.250(5.782)
	SLasso	8.4(1.3)	.968(.134)	.071(.098)	11.129(5.233)
GA2	ALasso	26.8(15.8)	.861(.112)	.674(.142)	17.07(3.334)
	SCAD	13.7(6.1)	.724(.189)	.513(.179)	18.178(4.162)
	SIS+SCAD	4.9(0.3)	.484(.066)	.210(.086)	16.68(3.142)
	FoBa	87.2(1.6)	.802(.169)	.926(.015)	48.407(13.017)
	FSR	5.0(1.7)	.579(.201)	.057(.104)	17.554(3.995)
	SLasso	4.6(1.6)	.511(.176)	.091(.133)	19.236 (4.323)
GA3	ALasso	15.1(8.1)	.999(.009)	.356(.242)	4.676(0.82)
	SCAD	8.1(0.7)	.979(.099)	.029(.099)	4.45(0.933)
	SIS+SCAD	5.0(0.3)	.543(.068)	.126(.100)	17.033(3.178)
	FoBa	87.2(1.4)	.862(.042)	.921(.004)	15.121(3.794)
	FSR	8.2(0.8)	.745(.190)	.270(.169)	5.658(1.290)
	SLasso	8.2(1.0)	.932(.153)	.092(.143)	4.892(1.266)
GA4	ALasso	11.2(5.2)	1.00(.000)	.188(.225)	2.312(0.419)
	SCAD	8.1(0.1)	.999(.009)	.000(.000)	2.582(1.049)
	SIS+SCAD	4.9(0.4)	.527(.076)	.132(.115)	10.02(2.652)
	FoBa	86.9(1.4)	.873(.032)	.920(.003)	7.679(2.419)
	FSR	6.7(3.4)	.779(.398)	.049(.086)	5.766(6.555)
	SLasso	6.5(3.5)	.784(.409)	.028(.059)	5.745(6.525)
GA5	ALasso	11.2(5.2)	1.00(.000)	.190(.226)	5.196(0.842)
	SCAD	8.1(0.1)	.999(.012)	.000(.000)	5.147(1.031)
	SIS+SCAD	5.1(0.1)	.504(.034)	.190(.046)	10.324(1.601)
	FoBa	87.1(1.5)	.871(.021)	.920(.002)	17.665(4.927)
	FSR	7.3(2.0)	.782(.167)	.124(.104)	7.500(2.734)
	SLasso	7.5(1.7)	.911(.188)	.027(.057)	6.440(2.768)

Table 6.2: Comparison of SLasso with ALasso, SCAD, SIS+SCAD and FSR in terms of PDR, FDR, PMSE and model size (MSize) averaged over 200 simulation replicates in Group A settings (sample size $n = 200$, numbers in parentheses are standard deviations)

Setting	Methods	MSize	PDR	FDR	PMSE
GA1	ALasso	49.0(18.0)	1.00(.000)	.791(.077)	10.937(1.463)
	SCAD	13.6(4.8)	1.00(.000)	.283(.183)	8.638(0.897)
	SIS+SCAD	8.7(0.5)	.793(.077)	.181(.086)	12.355(3.309)
	FoBa	176.2(1.7)	1.00(.000)	.949(.000)	28.439(4.261)
	FSR	9.4(0.7)	1.00(.000)	.035(.060)	8.688(1.024)
	SLasso	9.4(0.7)	1.00(.000)	.035(.061)	8.683(1.025)
GA2	ALasso	40.6(19.8)	.941(.072)	.735(.140)	15.297(2.03)
	SCAD	23.7(7.3)	.931(.110)	.612(.127)	14.159(2.928)
	SIS+SCAD	8.1(0.8)	.661(.028)	.255(.076)	14.715(1.715)
	FoBa	176.2(1.6)	.943(.102)	.952(.005)	41.228(8.005)
	FSR	7.9(1.7)	.846(.179)	.035(.070)	13.541(2.915)
	SLasso	7.8(2.1)	.796(.190)	.073(.100)	14.462(3.207)
GA3	ALasso	25.5(15.9)	.956(.071)	.507(.283)	4.205(0.539)
	SCAD	9.1(1.1)	.972(.121)	.031(.124)	3.963(0.62)
	SIS+SCAD	8.9(0.4)	.864(.064)	.128(.046)	4.498(1.987)
	FoBa	176.3(1.5)	.882(.034)	.955(.002)	12.544(2.252)
	FSR	9.2(0.9)	.708(.206)	.311(.183)	4.688(0.672)
	SLasso	9.2(1.0)	.873(.209)	.148(.190)	4.272(0.714)
GA4	ALasso	13.3(6.4)	1.00(.000)	.215(.242)	2.186(0.267)
	SCAD	9.0(.0)	1.00 (.000)	.000(.000)	2.320(0.753)
	SIS+SCAD	8.7(0.7)	.449(.064)	.535(.061)	3.327(1.679)
	FoBa	174.6(1.5)	.888(.011)	.954(.001)	6.875(1.229)
	FSR	9.3(0.6)	.993(.043)	.037(.074)	2.207(0.297)
	SLasso	9.2(0.6)	1.00(.000)	.023(.052)	2.183(0.268)
GA5	ALasso	15.7(9.5)	.986(.044)	.276(.284)	5.303(0.622)
	SCAD	9.0 (0.1)	.999(.011)	.000(.000)	5.199(0.71)
	SIS+SCAD	7.8(0.8)	.681(.066)	.206(.070)	7.975(1.258)
	FoBa	175.1(1.6)	.886(.017)	.954(.001)	16.644(2.932)
	FSR	9.4(0.6)	.943(.086)	.091(.100)	5.545(0.799)
	SLasso	9.3(0.6)	1.00(.000)	.024(.054)	5.241(0.608)

Table 6.3: Comparison of SLasso with ALasso, SCAD, SIS+SCAD and FSR in terms of PDR, FDR, PMSE and model size (MSize) averaged over 500 simulation replicates in Group B settings (numbers in parentheses are standard deviations)

Setting	Methods	MSize	PDR	FDR	PMSE
GB1	ALasso	45.3(13.4)	.858(.072)	.695(.079)	5.514(0.853)
	SCAD	13.0(2.8)	.731(.078)	.135(.115)	4.237(0.893)
	SIS+SCAD	3.0(0.7)	.198(.042)	.016(.062)	21.581(1.881)
	FoBa	87.0(1.6)	.858(.076)	.852(.013)	9.750(2.282)
	FSR	11.2(1.3)	.705(.066)	.050(.067)	4.009(0.973)
	SLasso	11.2(2.1)	.677(.109)	.083(.082)	4.503(2.654)
GB2	ALasso	56.5(14.7)	.881(.077)	.752(.059)	6.802(1.371)
	SCAD	17.2(4.0)	.792(.065)	.277(.136)	3.858(0.997)
	SIS+SCAD	4.7(0.6)	.259(.041)	.175(.127)	15.327(2.527)
	FoBa	86.8(1.6)	.855(.056)	.852(.010)	12.469(3.468)
	FSR	11.0(2.2)	.696(.128)	.040(.062)	4.349(1.625)
	SLasso	10.5(2.6)	.660(.155)	.053(.073)	4.966(2.630)
GB3	ALasso	69.6(5.9)	.852(.051)	.877(.012)	7.893(4.293)
	SCAD	8.8(2.6)	.583(.104)	.308(.125)	28.737(10.868)
	SIS+SCAD	4.3(0.7)	.000(.000)	1.00(.000)	58.334(10.717)
	FoBa	82.5(1.7)	.997(.018)	.879(.004)	7.701(3.389)
	FSR	18.2(3.0)	.785(.122)	.561(.075)	8.638(7.522)
	SLasso	9.8 (3.5)	.754(.31)	.262(.146)	19.470(20.808)

Table 6.4: The ID of the probes selected by ALasso, SCAD, SIS+SCAD, FSR and SLasso in the analysis of the rat data

Method	Probes ID				
ALasso	1387060_at	1388538_at	1380070_at	1370052_at	1382452_at
	1379079_at	1397489_at	1374131_at	1383110_at	1389584_at
	1392692_at	1379971_at	1385687_at	1369353_at	1374106_at
	1383673_at	1379495_at	1383749_at	1382835_at	1395415_at
	1383996_at				
SCAD	1394689_at	1370434_a_at	1375724_at	1378765_at	1375139_at
	1388538_at	1370052_at	1382452_at	1377781_at	1383841_at
	1380311_at	1379460_at	1385921_at	1384886_at	1384136_at
	1387111_at	1390789_at	1376693_at	1389584_at	1389231_at
	1390788_a_at	1367741_at	1374106_at	1387455_a_at	1383749_at
	1379803_at	1383996_at	1382633_at		
SIS+SCAD	1377546_at	1396809_at	1381430_at	1393543_at	1372481_at
FSR	1383110_at	1392692_at	1389584_at		
SLasso	1383110_at	1392692_at			