Statistica Sinica Preprint No: SS-2015-0413.R3	
Title	Regularization after retention in ultrahigh dimensional linear regression models
Manuscript ID	SS-2015-0413.R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0413
Complete List of Authors	Haolei Weng Yang Feng and Xingye Qiao
Corresponding Author	Yang Feng
E-mail	yang.feng@columbia.edu
Notice: Accepted version subject to English editing.	

REGULARIZATION AFTER RETENTION IN ULTRAHIGH DIMENSIONAL LINEAR REGRESSION MODELS

Haolei Weng, Yang Feng and Xingye Qiao

Columbia University, Columbia University and Binghamton University

Abstract: In ultrahigh dimensional setting, independence screening has been both theoretically and empirically proved a useful variable selection framework with low computation cost. In this work, we propose a two-step framework by using marginal information in a different perspective from independence screening. In particular, we retain significant variables rather than screening out irrelevant ones. The new method is shown to be model selection consistent in the ultrahigh dimensional linear regression model. To improve the finite sample performance, we then introduce a three-step version and characterize its asymptotic behavior. Simulations and real data analysis show advantages of our method over independence screening and its iterative variants in certain regimes.

Key words and phrases: Independence screening, lasso, penalized least square, retention, selection consistency, variable selection.

1. Introduction

High dimensional statistical learning has become increasingly important in many scientific areas. It mainly deals with statistical estimation and prediction in the setting where the dimensionality p is substantially larger than the sample size n.

An active philosophy of research imposes sparsity constraints on the model. Under this framework, variable selection plays a crucial role in three aspects: statistical accuracy, model interpretability and computational complexity.

Various penalized maximum likelihood methods have been proposed in recent years. Compared to traditional variable selection methods such as Akaike's information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz, 1978), these regularization techniques in general aim to improve stability and reduce computational cost. Examples include bridge regression (Frank and Friedman, 1993), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), MC+ (Zhang, 2010), among others. Theoretical results on parameter estimation (Knight and Fu, 2000), model selection (Zhao and Yu, 2006; Wainwright, 2009), prediction (Greenshtein and Ritov, 2004) and oracle properties (Fan and Li, 2001) have been developed under different model contexts. However, in the ultrahigh dimensional setting, where $\log p = O(n^{\xi})$ $(\xi > 0)$, the conditions for model selection/parameter estimation consistency associated with these techniques may easily fail due to high correlations between important and unimportant variables. Motivated by these concerns, Fan and Lv (2008) proposed a sure independence screening (SIS) method in the linear regression setting. The SIS method has been further extended to generalized linear models (Fan and Song, 2010), additive models (Fan et al., 2011), and model free scenarios (Zhu et al., 2011, Li et al., 2012, Li et al.,

2012). The main idea of independence screening methods is to utilize marginal information to screen out irrelevant variables. Fast computation and desirable statistical properties make them more attractive in large scale problems. After independence screening, other variable selection methods can be further applied to improve finite sample performances.

Besides using independence screening, there is a rich literature on multi-step variable selection methods. Examples include screen and clean (Wasserman and Roeder, 2009), LOL (Kerkyacharian et al., 2009), thresholded Lasso (Zhou, 2010), stepwise regression method using orthogonal greedy algorithm (Ing and Lai, 2011), sequential Lasso (Luo and Chen, 2011), UPS (Ji and Jin, 2012) and tilted correlation screening (Cho and Fryzlewicz, 2012).

In this paper, we consider variable selection consistency in the ultrahigh dimensional linear regression model and focus on the situations where there exist signals with weak marginal correlations. Under these scenarios, independence screening tends to either miss such signals or include many unimportant variables, which will undermine the variable selection performance. We propose a general two-step framework, in a different direction from independence screening, in terms of how the marginal information is used. The motivation of our method is that, instead of screening out noises, it may be relatively easy to identify a subset of signals. Therefore, we use marginal regression coefficient estimates to retain a set of important

predictors in the first step (called retention). In the second step (called regularization), we use penalized least square by imposing regularization only on the variables not retained in the retention step. In the theoretical development, we replace the assumption on the lower bound of marginal information for important variables (Fan and Lv, 2008) by an assumption on the upper bound of marginal information for irrelevant variables. From the practical point of view, a permutation-based method is introduced to choose the threshold in the retention step. To enhance the finite sample performance, we also introduce a three-step version to eliminate unimportant variables falsely selected during the retention step. We further derive its selection consistency result as a generalization from the two-step method. The main contribution of our paper is to provide an alternative way to conduct high dimensional variable selection, especially in the cases where independence screening tends to fail. More importantly, we characterize our method by asymptotic analysis. As a by-product, we also give theoretical comparison between Lasso and our method, to demonstrate its improvement over Lasso, under certain regularity conditions.

The rest of the paper is organized as follows. We introduce the model setup and review the techniques of Lasso and independence screening in Section 2. In Section 3, after introducing the two-step framework with its asymptotic properties delineated, we also propose a three-step version along with its associated theory. Simulation examples and real data analysis are presented in Section 4. We conclude the paper with

a short discussion in Section 5. All technical proofs and some additional simulation results are collected in the online supplementary materials.

2. Model Setup and Relevant Variable Selection Techniques

In this section, the model setup is introduced and two related model selection methods, Lasso and independence screening, are reviewed.

2.1. Model Setup and Notations

Let $V_1, \cdots, V_n$ be independently and identically distributed random vectors, where $V_i = (X_i^T, Y_i)^T$, following the linear regression model,

$$Y_i = X_i^T \beta + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $X_i = (X_i^1, \cdots, X_i^p)^T$ is a p-dimensional vector distributed as $N(0, \Sigma)$, $\beta = (\beta_1, \cdots, \beta_p)^T$ is the true coefficient vector, $\varepsilon_1, \cdots, \varepsilon_n$ are independently and identically distributed as $N(0, \sigma^2)$, and $\{X_i\}_{i=1}^n$ are independent of $\{\varepsilon_i\}_{i=1}^n$. Denote the support index set of β by $S = \{j : \beta_j \neq 0\}$ and the cardinality of S by s. For any set A, let A^c be its complement set. For any k dimensional vector w and any subset $K \subseteq \{1, \cdots, k\}$, w_K denotes the subvector of w indexed by K, and let $\|w\|_1 = \sum_{i=1}^k |w_i|$, $\|w\|_2 = (\sum_{i=1}^k w_i^2)^{1/2}$, $\|w\|_\infty = \max_{i=1,\cdots,k} |w_i|$. For any $k_1 \times k_2$ matrix M, any subsets $K_1 \subseteq \{1, \cdots, k_1\}$ and $K_2 \subseteq \{1, \cdots, k_2\}$, $M_{K_1 K_2}$ represents the submatrix of M consisting of entries indexed by the Cartesian product $K_1 \times K_2$. Let M_{K_2} be the columns of M indexed by K_2 and M^j be the jth column of M. Denote $\|M\|_2 = \{\Lambda_{\max}(M^T M)\}^{1/2}$ and $\|M\|_\infty = \max_{i=1,\cdots,k_1} \sum_{j=1}^{k_2} |M_{ij}|$. When $k_1 = k_2 = k$,

let $\rho(M) = \max_{i=1,\cdots,k} M_{ii}$, $\Lambda_{\min}(M)$ and $\Lambda_{\max}(M)$ be the minimum and maximum eigenvalues of M respectively, and $\Sigma_{S^c|S} = \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}$.

In the ultrahigh dimensional scenario, assuming β is sparse, we are interested in recovering the sparsity pattern S of β. For technical convenience, we consider a stronger result called sign consistency (Zhao and Yu, 2006), namely $\text{pr}(\text{sign}(\hat{\beta}) = \text{sign}(\beta)) \to 1$, as $n \to \infty$, where $\text{sign}(\cdot)$ maps positive numbers to 1, negative numbers to -1 and zero to zero. In asymptotic analysis, we denote the sparsity level by s_n and dimension by p_n to allow them to grow with the number of observations. For conciseness, we sometimes use signals and noises to represent relevant predictors S and irrelevant predictors S^c or their corresponding coefficients, respectively.

2.2. Lasso in Random Design

The least absolute shrinkage and selection operator (aka Lasso) (Tibshirani, 1996) solves

$$\hat{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\}.$$

For fixed design, model selection consistency has been well studied in Zhao and Yu (2006) and Wainwright (2009). They characterized the dependency between relevant and irrelevant predictors by an irrepresentable condition, which proved to be both sufficient and (almost) necessary for sign consistency. For random design, Wainwright (2009) established precise sufficient and necessary conditions on (n, p_n, s_n) for sparse recovery. We state a corollary from his general results with a particular scaling of

the triplet for further use in the sequel. Here are some key conditions.

Condition 0. $\log p_n = O(n^{a_1}), s_n = O(n^{a_2}), a_1 > 0, a_2 > 0, a_1 + 2a_2 < 1.$

Condition 1. $\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0.$

Condition 2. $\|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty \leq 1 - \gamma, \quad \gamma \in (0, 1].$

Condition 3. $\rho(\Sigma_{S^c|S}) = o(n^\delta), \ 0 < \delta < 1 - a_1 - 2a_2.$

Condition 4. $\min_{j \in S} |\beta_j| \geq Cn^{(\delta + a_1 + 2a_2 - 1)/2}$ for a sufficient large C, where δ is the same as in Condition 3.

Condition 2 is the population analog of the irrepresentable condition in Zhao and Yu (2006), in which $(\Sigma_{SS})^{-1}\Sigma_{SS^c}$ is the regression coefficient matrix by regressing noises on signals. Hence, $\|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty$ can be viewed as a reasonable measurement of the dependency between signals and noises. In the ultrahigh dimensional scenario, noises are likely to be highly correlated with signals, which could make this condition fail. To relax this condition, the corresponding matrix in the regularization step for our method (to be formally defined in Section 3) will be a submatrix of $\Sigma_{S^cS}(\Sigma_{SS})^{-1}$ with fewer number of columns. As a result, the corresponding quantity in Condition 2 is reduced. In Condition 3, $\Sigma_{S^c|S}$ is the conditional covariance matrix of X_{S^c} given X_S. This condition imposes another kind of eigenvalue-type dependency constraint. In addition to the dependency conditions between signals and noises, the signals should be linearly independent and the minimum signal can not decay too fast as shown by Conditions 1 and 4, respectively.

Proposition 1. *Under the scaling specified in Condition 0, if the covariance matrix* Σ *and the true parameter* β *satisfy Conditions 1-4, and* $s_n \to \infty$, $p_n - s_n \to \infty$, $\lambda_n \asymp n^{(\delta+a_1-1)/2}$, *we have sign consistency*

$$\mathrm{pr}(\hat{\beta} \text{ is unique, and } \mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta)) \to 1 \quad \text{as } n \to \infty.$$

2.3. Independence Screening

Sure independence screening was proposed by Fan and Lv (2008) in the linear regression model framework. It conducts variable selection according to magnitude of marginal correlations. Specifically, assume that the columns in the design matrix $X = (X^1, \cdots, X^{p_n})$ have been standardized with mean zero and variance one. Denote the response vector $Y = (Y_1, \cdots, Y_n)^T$ and the rescaled sample correlation between each predictor X^j and Y by $\hat{\beta}_j^M = Y^T X^j$ $(1 \le j \le p_n)$. Then the selected submodel by sure independence screening is

$$\widehat{\mathcal{M}}_{d_n} = \{1 \le j \le p_n : |\hat{\beta}_j^M| \text{ belongs to the } d_n \text{ largest values}\},$$

where d_n is a positive integer smaller than n. This simple procedure turns out to enjoy the sure screening property as reviewed below. Consider $\log p_n = O(n^a), a \in (0, 1-2\kappa)$, where $0 < \kappa < 1/2$. Under the conditions

$$\mathrm{var}(Y_i) = O(1),\ \Lambda_{\max}(\Sigma) = O(n^\tau),\ \min_{j\in S}|\beta_j| \ge cn^{-\kappa}, \tau \ge 0,$$

$$\min_{j\in S}|\mathrm{cov}(\beta_j^{-1}Y_1, X_1^j)| \ge c > 0, \tag{21}$$

Fan and Lv (2008) showed that if $2\kappa + \tau < 1$, then there exists some $\theta \in (2\kappa + \tau, 1)$ such that for $d_n \asymp n^\theta$, we have for some $C > 0$,

$$\mathrm{pr}(S \subseteq \widehat{\mathcal{M}}_{d_n}) = 1 - O(p_n \exp(-Cn^{1-2\kappa}/\log n)).$$

The condition in (21) imposes a lower bound for magnitudes of the marginal correlations between response and signals. However, in some cases, signals are marginally uncorrelated with the response, then this condition is not satisfied. Although Fan and Lv (2008) introduced an iterative version to overcome this issue, the associated theoretical property is still unknown. We will drop this assumption and focus on instead the situation where the marginal correlations between noises and the response are not large.

3. Method and Theory

3.1. The New Two-Step Estimator

In this section, we propose a two-step method named regularization after retention (RAR). In the first step, we use marginal information to retain important signals, and in the second step, we conduct a penalized least square with penalty imposed only on the variables not retained in the first step.

Step 1. (Retention) Calculate the marginal regression coefficient estimate for each predictor,

$$\hat{\beta}_j^M = \frac{\sum_{i=1}^n (X_i^j - \bar{X}^j) Y_i}{\sum_{i=1}^n (X_i^j - \bar{X}^j)^2} \quad (1 \le j \le p),$$

where $\bar{X}^j = n^{-1}\sum_{i=1}^n X_i^j$. Then define a retention set by $\hat{R} = \{1 \le j \le p : |\hat{\beta}_j^M| \ge \gamma_n\}$, for a positive constant γ_n.

Step 2. (Regularization) The final estimator is

$$\check{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda_n \sum_{j\in\hat{R}^c} |\beta_j| \right\}.$$

Note that the difference between the retention step and independence screening is that independence screening aims at screening out as many noises as possible, while the retention step tries to detect and retain as many signals as possible. The threshold γ_n needs to be chosen carefully so that no noise is retained. In the desired situation when $\hat{R} \subseteq S$, meaning all the variables in $\hat{R}$ are signals, one only needs to impose sparsity on $\hat{R}^c$ to recover the entire sparsity pattern. The advantage is that the estimation accuracy of $\beta_{\hat{R}}$ is not compromised due to regularization.

Moreover, it turns out that this well-learned information can relax the consistency conditions of Lasso. On the other hand, we need extra regularity conditions to guarantee $\hat{R} \subseteq S$ with high probability. We will show that under the scaling $\log p_n = O(n^\xi)$ $(\xi > 0)$, our estimator $\check{\beta}$ achieves sign consistency. The two steps will be studied separately in Section 3.2 and Section 3.3.

3.2. Asymptotics in the Retention Step

Let the marginal regression coefficients $\beta_j^M = \mathrm{cov}(X_1^j, Y_1)$. For simplicity, we assume the covariance matrix Σ for X_1 has unit diagonal elements and the variance of random

error is $\sigma^2 = 1$. We first present several conditions.

Condition 5. $\|\Sigma\beta\|_\infty = O(n^{(1-2\kappa)/8})$, where $0 < \kappa < \frac{1}{2}$ is a constant.

Condition 6. $\beta_S^T \Sigma_{SS} \beta_S = O(1)$.

Proposition 2. *Under Conditions 5 and 6, we have for any $c_* > 0$, there exists $c_2 > 0$,*

$$\mathrm{pr}(\max_{1\le j\le p_n} |\hat{\beta}_j^M - \beta_j^M| > c_* n^{-\kappa}) = O(p_n \exp(-c_2 n^{(1-2\kappa)/4})). \tag{11}$$

The essential part of the proof for Proposition 2 follows an exponential inequality for the quasi-maximum likelihood estimator in Fan and Song (2010). Condition 5 puts an upper bound on the maximum marginal correlation between covariates and the response, and is a technical condition required to achieve the convergence rate in (11). Condition 6 bounds $\mathrm{var}(Y_1)$ as in Fan and Lv (2008) and Fan and Song (2010). The rationale of this condition is as follows. Imagine we would like to study the relationship between blood pressure $(Y_1, Y_2, \cdots, Y_n)$ of n patients using gene expression data $X_{n\times p}$. As n increases, we are measuring more gene expression predictors (p increases) and the number of important predictors s_n also increases. However, the distribution of blood pressure remains unchanged, which actually puts an implicit restriction on the overall contribution of the s_n important predictors $(\beta_S^T \Sigma_{SS} \beta_S)$ asymptotically. Proposition 2 provides a uniform concentration result for the marginal coefficient estimates and it leads to the following desirable property of $\hat{R}$ when the retention threshold is chosen properly.

Corollary 1. *Let* $\zeta_n = \|\Sigma_{S^cS}\beta_S\|_\infty$ *and* c_1 *be a positive constant. Under Conditions 5-6, and when the threshold* $\gamma_n = \zeta_n + c_1 n^{-\kappa}$*, there exists a constant* $c_3 > 0$ *so that we have the following sure retention property,*

$$\mathrm{pr}(\hat{R} \subseteq S) = 1 - O(p_n \exp(-c_3 n^{(1-2\kappa)/4})). \tag{12}$$

Here, ζ_n is the maximum magnitude of the covariances between noises and the response, which may change as s_n increases. The choice of the threshold γ_n is essential for sure retention.

Equation (12) may not be informative if the threshold γ_n is set too high so that $\hat{R}$ is an empty set. Before quantifying how large $\hat{R}$ is, define the marginal strong signal set $R = \{j \in S : |\beta_j^M| > \zeta_n + 2c_1 n^{-\kappa}\}$. On the set $\{\max_{1\le j\le p_n} |\hat{\beta}_j^M - \beta_j^M| \le c_1 n^{-\kappa}\}$, we have $\{|\beta_j^M| > \zeta_n + 2c_1 n^{-\kappa}\} \subseteq \{|\hat{\beta}_j^M| > \zeta_n + c_1 n^{-\kappa}\}$ holds for any j. Thus,

$$\mathrm{pr}(R \subseteq \hat{R}) \ge 1 - O(p_n \exp(-c_3 n^{(1-2\kappa)/4})). \tag{13}$$

Equation (13) indicates that our retention set $\hat{R}$ contains the marginal strong signal set R with high probability when the dimensionality p_n satisfying $\log p_n = o(n^{(1-2\kappa)/4})$. It will be clear from the conditions in the next subsection that the size of R plays an important role in achieving sign consistency for $\check{\beta}$.

3.3. Sign Consistency in the Regularization Step

In the retention step, we can detect part of signals with high probability, including the marginal strong signal set R. Incorporating this information into the regular-

ization step, namely not penalizing the retained signals, we can show that the sign consistency of ℓ_1 regularized least square holds in weaker conditions.

Condition 7. $\log p_n = O(n^{a_1})$, $s_n = O(n^{a_2})$, where $0 < a_1 < (1-2\kappa)/4$ with κ the same as in Condition 5, $a_2 > 0$, and $\max(a_1, a_2) + a_2 < 1$.

Condition 8. $\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0$.

Condition 9. $\|\{\Sigma_{S^cS}(\Sigma_{SS})^{-1}\}_{S\cap R^c}\|_\infty \leq 1-\gamma, \quad \gamma \in (0,1]$.

Condition 10. $\min_{j\in S}|\beta_j| \geq Cn^{-\delta+a_2/2}$ for a sufficient large C, where $0 < \delta < \{1-\max(a_1,a_2)\}/2$.

Theorem 1. *Under Conditions 5-10, if $s_n \to \infty$ and $\lambda_n \asymp n^{-\delta}$, our two-step estimator $\check{\beta}$ achieves sign consistency,*

$$\mathrm{pr}(\check{\beta} \text{ is unique and}, \mathrm{sign}(\check{\beta}) = \mathrm{sign}(\beta)) \to 1, \text{ as } n \to \infty.$$

As can be seen in the supplement, our proof follows the essential techniques of the proof in Wainwright (2009). That is why Conditions 8-10 share similarity with Conditions 1-4. The key difference is to prove that with high probability, the estimator in the second step recovers the signs when S_1 is not penalized, uniformly for all sets S_1 satisfying $R \subseteq S_1 \subseteq S$. Since the retention set $\hat{R}$ in the first step satisfies $R \subseteq \hat{R} \subseteq S$ with high probability from Corollary 1 and (13), the final two-step estimator achieves sign consistency.

Condition 9 is a weaker version of Condition 2. Each row of $\Sigma_{S^cS}\Sigma_{SS}^{-1}$ can be regarded as the regression coefficients (population version) by regressing the corre-

sponding noise on signals. Thus, Condition 2 requires that for each noise, the sum of the absolute values of its regression coefficients is less than $1-\gamma$. In contrast, the corresponding sum in Condition 9 excludes coefficients corresponding to the retained signals. As a result, we allow larger regression coefficients for the retained signals. Note that regression coefficients measure the dependency between response and regressors. In this sense, our method allows stronger dependency between noises and the retained signals. How much we gain by conducting the first step largely depends on the size of the strong signal set R. The larger R is, the greater improvement our method can make over Lasso.

3.4. The Redemption in the Third Step

The success of RAR highly depends on the quality of retained variables in the retention step. If the retention set contains noise variables, those variables would remain in the final model selected by RAR since they are not penalized in the second step. This could happen if the threshold for retention is chosen too small. To improve the robustness towards the choice of the threshold and the finite sample performance of our procedure, we propose to add one extra step, the redemption step, aiming to remove these falsely retained variables. In addition, we study the theoretical property of the three-step procedure.

Denote by Q the additional signals detected in the regularization step, that is $Q = \{j \in \hat{R}^c : \check{\beta}_j \neq 0\}$.

Step 3. (Redemption) Calculate the following penalized least square problem

$$\tilde{\beta} = \underset{\beta_{(\hat{R}\cup Q)^c}=0}{\arg\min} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j\in\hat{R}} X_{ij}\beta_j - \sum_{k\in Q} X_{ik}\beta_k)^2 + \lambda_n^* \sum_{j\in\hat{R}} |\beta_j| \right\},$$

where λ_n^* is the penalty parameter, which is in general different from λ_n in the second step.

The idea is to regularize only the coefficients in the retained set $\hat{R}$ while keeping the signals identified in Q. Note that variables not selected in the regularization step are no longer considered (that is, $\tilde{\beta}_{(\hat{R}\cup Q)^c} = 0$). Therefore, the redemption step has a much lower effective parameter dimension than the regularization step, and has little extra computational cost. The three-step estimator $\tilde{\beta}$ is called regularization after retention plus (RAR+).

Under certain regularity conditions, the three-step estimator $\tilde{\beta}$ achieves sign consistency. To this end, we define a strong noise set $Z = \{j \in S^c : |\beta_j^M| \geq \gamma_n - c_1 n^{-\kappa}\}$ with its cardinality z_n. Recall the strong signal set $R = \{j \in S : |\beta_j^M| \geq \gamma_n + c_1 n^{-\kappa}\}$. The new regularity conditions are as follows.

Condition 11. $\Lambda_{\min}(\Sigma_{S\cup Z, S\cup Z}) \geq C_{\min} > 0$.

Condition 12. $\max_{S\subset Q\subset S\cup Z} \|\{\Sigma_{Q^cQ}(\Sigma_{QQ})^{-1}\}_{S\cap R^c}\|_\infty \leq 1-\gamma$, where $\gamma > 0$.

Condition 13. $\|\Sigma_{ZS}\Sigma_{SS}^{-1}\|_\infty \leq 1-\alpha$, where $\alpha > 0$.

Theorem 2. *Under Conditions 5-7 and 10-13, if* $z_n/s_n \to 0, s_n \to \infty$ *and* $\lambda_n \asymp$

$n^{-\delta}, \lambda_n^* \asymp n^{-\delta}$, *our three-step estimator* $\tilde{\beta}$ *achieves sign consistency,*

$$\text{pr}(\tilde{\beta} \text{ is unique and } \text{sign}(\tilde{\beta}) = \text{sign}(\beta)) \to 1, \text{ as } n \to \infty.$$

Compared with Theorem 1, the strong noise set Z appears in Conditions 11-13. Theorem 2 is a generalization of Theorem 1 in the sense that if Z is empty, Theorem 2 reduces to Theorem 1. It provides a justification for RAR+ under a flexible choice of the threshold for retention; different choices of the threshold could lead to different Z's. RAR+ is able to tolerate false retention at a level quantified by Conditions 11-13, which essentially require the possible noises selected in the retention step cannot be highly correlated with the signals. We will demonstrate the improvement of RAR+ over RAR regarding the robustness towards the choice of the threshold using simulation studies in Section 4.1.

3.5. Connections to SIS-Lasso and Adaptive Lasso

In this section, we highlight the connections of RAR with sure independence screening followed by Lasso (SIS-lasso) and the adaptive Lasso method (Ada-lasso). In the first step, both RAR and SIS-lasso calculate and rank the marginal regression coefficient estimates. In the second step, the estimator for RAR can be written as

$$\check{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + 0 \sum_{j \in \hat{R}} |\beta_j| + \lambda_n \sum_{j \in \hat{R}^c} |\beta_j| \right\}, \tag{14}$$

while the estimator for SIS-lasso is

$$\arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda_n \sum_{j \in \hat{S}} |\beta_j| + \infty \sum_{j \in \hat{S}^c} |\beta_j| \right\}, \tag{15}$$

where $\hat{S}^c$ is the set of the screened-out variables in Step 1 of SIS-lasso.

Both methods relax the consistency condition of Lasso $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty \leq 1-\gamma$. SIS-lasso reduces $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty$ by removing rows of $\Sigma_{S^cS}\Sigma_{SS}^{-1}$ corresponding to the screened-out noises. RAR reduces $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty$ by removing columns of $\Sigma_{S^cS}\Sigma_{SS}^{-1}$ corresponding to the retained signals. Although the number of removed rows by SIS is typically larger than that of removed columns by RAR, it does not necessarily mean that the amount of reduction by SIS will be greater than that by RAR. For example, if there exist signals highly correlated to noises (i.e., scenario 1(A) in Section 4.1), retaining signals with the largest marginal correlations will substantially decrease $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty$, while removing noises with small marginal correlations does not change $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty$ at all.

(14) and (15) lead to a natural comparison with the adaptive Lasso (Zou, 2006) estimator:

$$\arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j| \right\}, \tag{16}$$

where the weight w_j is usually chosen as $1/|\beta_{j,init}|^\gamma$ for some $\gamma > 0$ using an initial estimator $\beta_{j,init}$. For fixed design, Zou (2006) proved that the adaptive Lasso estimator achieves variable selection consistency under very mild conditions when p is fixed. In the high dimensional regime, Huang et al. (2008) showed variable selection consistency with $w_j = 1/|\hat{\beta}_j^M|$ under the partial orthogonality condition (i.e., signals are weakly correlated to noises). A more general theoretical treatment is given in

Zhou et al. (2009) under the restricted eigenvalue conditions (Bickel et al., 2009) for both fixed and random designs.

All of (14), (15) and (16) aim at improving Lasso by adaptively adjusting the penalty level for each predictor. The major difference between (14)-(15) and (16) is that (16) uses "soft" weights while both (14) and (15) use "thresholded" weights. For (16), it is possible that there exists $\beta_{j,init} \approx 0$ for some signal j with small marginal correlation, leading to a very large weight for that variable, which makes the consistent selection difficult. Due to the specific thresholding choices, a similar observation can be found for (15). In contrast, (14) can still succeed in sparse recovery for such a difficult case. Extensive simulation studies for comparing RAR, SIS-lasso and adaptive Lasso are conducted in Section 4.1.

3.6. A Permutation Method for Choosing the Retention Threshold

Theorems 1 and 2 provide a theoretical guideline for choosing the retention threshold γ_n, which depends on some unknown parameters. In practice, we propose to select γ_n by a permutation-based method. Denote m randomly permuted response vectors by $Y^{(1)}, \cdots, Y^{(m)}$. Let the marginal regression coefficients from the permuted data be

$$D_k^j = \frac{\sum_{i=1}^n (X_i^j - \bar{X}^j) Y_i^{(k)}}{\sum_{i=1}^n (X_i^j - \bar{X}^j)^2}, 1 \le j \le p, 1 \le k \le m.$$

Then we set the tentative threshold $\gamma_n = \max_{k,j} |D_k^j|$. Intuitively, if noises are not strongly correlated with response, the maximum absolute value of marginal regression

coefficients from permutation should be a reasonable threshold. If this tentative threshold leads to a retention set with size larger than $\lceil n^{1/2} \rceil$, we then retain only the top $\lceil n^{1/2} \rceil$ variables with the largest magnitudes of the marginal coefficients $|\hat{\beta}_j^M|$. This ensures that there are at most $\lceil n^{1/2} \rceil$ variables not penalized in the second step. Note that it is necessary to impose an upper bound on the retention set size since the predictors in the retention set are not regularized during the second step. We will show in the next section that the permutation method with the size upper bound $\lceil n^{1/2} \rceil$ works well in a range of simulation settings.

4. Numerical Studies

In this section, we compare the performance of RAR and RAR+ with some popular variable selection methods on an array of simulated examples and a real data set. To demonstrate the flexibility of our proposed framework, we also investigate modified versions of RAR and RAR+, denoted by RAR(MC+) and RAR+(MC+), in the way of replacing the ℓ_1 penalty by the nonconvex penalty MC+.

4.1. Simulations

We compare the variable selection performances of Lasso, SCAD, MC+, Ada-lasso, SIS-lasso, SIS-MC+, iterative sure independence screening (ISIS-lasso, ISIS-MC+), screen and clean (SC-lasso, SC-forward, SC-marginal), RAR, RAR+, RAR(MC+) and RAR+(MC+) in the ultrahigh dimensional linear regression setting. We set $n = 100, 200, 300, 400, 500$ and $p_n = \lfloor 100 \exp(n^{0.2}) \rfloor$, where $\lfloor k \rfloor$ is the largest integer not

exceeding k. The number of repetitions is 200 for each triplet (n, s_n, p_n). We calculate the proportion of exact sign recovery. All the Lasso procedures are implemented using the R package glmnet (Friedman et al., 2010). SCAD and MC+ are implemented using the R package ncvreg (Breheny and Huang, 2011).

Since data driven methods for tuning parameter selection introduce extra randomness into the entire variable selection process, we report the oracle performance of each method for fair comparison. Specifically, for Lasso, SCAD, MC+, Ada-lasso, the regularization steps of SIS-lasso, SIS-MC+, RAR, RAR+, RAR(MC+) and RAR+(MC+), the cleaning stage of SC-lasso, SC-forward and SC-marginal (with significance level as a tuning parameter), we check if there exists at least one estimator with exact sign recovery on the solution path. For SIS-lasso, SIS-MC+, ISIS-lasso and ISIS-MC+, we select the top $\lfloor n/\log n \rfloor$ variables with the largest absolute marginal correlation in the first step. For Ada-lasso, following Huang et al. (2008), we choose the weights $w_j = 1/|\hat{\beta}_j^M|$. For RAR(MC+) and RAR+(MC+), we fix the concavity parameter $\gamma = 1.5$ and compute the solution path by only varying penalty parameter λ. We consider different simulation settings in the following.

Scenario 1. The covariance matrix Σ is

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & I \end{pmatrix}, \text{ where } \Sigma_{11} = (1-r)I + rJ \in \mathbb{R}^{2s_n \times 2s_n},$$

in which I is the identity matrix and J is the matrix of all 1s.

Table 1: Sign recovery proportion over 200 simulation rounds.

Scenario 1 (A)	(100, 1232)	(200, 1791)	(300, 2285)	(400, 2750)	(500, 3199)
Lasso	0.000	0.000	0.050	0.205	0.545
SCAD	0.000	0.010	0.120	0.495	0.815
MC+	0.000	0.235	0.640	0.895	0.990
SIS-lasso	0.000	0.000	0.000	0.030	0.010
ISIS-lasso	0.000	0.000	0.040	0.185	0.500
Ada-lasso	0.000	0.000	0.000	0.025	0.030
SIS-MC+	0.000	0.000	0.000	0.045	0.015
ISIS-MC+	0.000	0.040	0.305	0.610	0.875
SC-lasso	0.000	0.000	0.005	0.040	0.150
SC-forward	0.000	0.000	0.010	0.120	0.390
SC-marginal	0.000	0.000	0.000	0.000	0.000
RAR_1	0.010	0.170	0.395	0.395	0.295
RAR_5	0.000	0.315	0.630	0.700	0.600
RAR_{30}	0.005	0.255	0.750	0.875	0.835
RAR(MC+)_{30}	0.000	0.280	0.750	0.880	0.840
RAR+_1	**0.020**	**0.460**	**0.925**	**0.990**	**1.000**
RAR+_5	0.000	0.415	0.880	0.975	0.995
RAR+_{30}	0.005	0.280	0.780	0.965	0.990
RAR+(MC+)_{30}	0.000	0.290	0.770	0.965	0.995

Scenario 1 (B)	(100, 1232)	(200, 1791)	(300, 2285)	(400, 2750)	(500, 3199)
Lasso	0.000	0.000	0.135	0.580	0.855
SCAD	0.000	0.140	0.815	0.990	0.995
MC+	0.055	0.805	**1.000**	**1.000**	**1.000**
SIS-lasso	0.000	0.000	0.010	0.140	0.270
ISIS-lasso	0.000	0.000	0.110	0.575	0.850
Ada-lasso	0.000	0.015	0.190	0.370	0.455
SIS-MC+	0.000	0.050	0.165	0.235	0.350
ISIS-MC+	0.025	0.585	0.960	**1.000**	**1.000**
SC-lasso	0.000	0.000	0.020	0.275	0.680
SC-forward	0.000	0.005	0.125	0.650	0.910
SC-marginal	0.000	0.010	0.010	0.020	0.020
RAR_1	0.130	0.025	0.010	0.000	0.000
RAR_5	0.190	0.105	0.030	0.000	0.000
RAR_{30}	0.160	0.250	0.055	0.005	0.000
RAR(MC+)_{30}	0.195	0.250	0.055	0.005	0.000
RAR+_1	0.195	0.855	0.980	0.980	0.985
RAR+_5	0.255	0.885	0.995	**1.000**	0.995
RAR+_{30}	0.205	0.925	0.995	**1.000**	**1.000**
RAR+(MC+)_{30}	**0.290**	**0.965**	**1.000**	**1.000**	**1.000**

Table 2: Sign recovery proportion over 200 simulation rounds.

Scenario 2 (C)	(100, 1232)	(200, 1791)	(300, 2285)	(400, 2750)	(500, 3199)
Lasso	0.000	0.000	0.000	0.000	0.025
SCAD	0.000	0.020	0.110	0.355	0.615
MC+	**0.025**	0.325	0.775	0.955	**1.000**
SIS-lasso	0.000	0.000	0.000	0.000	0.005
ISIS-lasso	0.000	0.000	0.000	0.005	0.020
Ada-lasso	0.000	0.000	0.000	0.000	0.005
SIS-MC+	0.000	0.000	0.005	0.000	0.035
ISIS-MC+	0.000	0.020	0.145	0.410	0.740
SC-lasso	0.000	0.000	0.000	0.000	0.015
SC-forward	0.005	0.025	0.175	0.495	0.730
SC-marginal	0.005	0.000	0.005	0.000	0.000
RAR_{1}	0.000	0.120	0.335	0.340	0.245
RAR_{5}	0.000	0.195	0.550	0.585	0.490
RAR_{30}	0.000	0.175	0.635	0.720	0.775
RAR(MC+)_{30}	**0.025**	0.340	0.745	0.760	0.785
RAR+_{1}	0.000	0.275	0.770	0.905	0.960
RAR+_{5}	0.000	0.245	0.785	0.920	0.980
RAR+_{30}	0.000	0.180	0.675	0.905	0.975
RAR+(MC+)_{30}	**0.025**	**0.355**	**0.805**	**0.965**	**1.000**

Scenario 2 (D)	(100, 1232)	(200, 1791)	(300, 2285)	(400, 2750)	(500, 3199)
Lasso	0.000	0.000	0.000	0.000	0.000
SCAD	0.000	0.040	0.205	0.470	0.680
MC+	**0.055**	0.470	0.725	0.885	0.975
SIS-lasso	0.000	0.000	0.000	0.000	0.000
ISIS-lasso	0.000	0.000	0.005	0.035	0.055
Ada-lasso	0.000	0.000	0.000	0.000	0.000
SIS-MC+	0.000	0.000	0.025	0.020	0.085
ISIS-MC+	0.010	0.115	0.415	0.655	0.830
SC-lasso	0.000	0.000	0.000	0.015	0.045
SC-forward	0.000	0.090	0.370	0.570	0.690
SC-marginal	0.000	0.000	0.005	0.010	0.000
RAR_{1}	0.025	0.110	0.050	0.005	0.000
RAR_{5}	0.015	0.195	0.095	0.020	0.010
RAR_{30}	0.000	0.230	0.190	0.060	0.010
RAR(MC+)_{30}	0.045	0.345	0.205	0.060	0.010
RAR+_{1}	0.035	0.270	0.620	0.830	0.930
RAR+_{5}	0.015	0.290	0.625	0.830	0.935
RAR+_{30}	0.000	0.265	0.595	0.820	0.935
RAR+(MC+)_{30}	0.050	**0.505**	**0.880**	**0.960**	**1.000**

(A). $r = 0.6, \sigma = 3.5, s_n = 4, \beta_S = (3, -2, 2, -2)^T, \beta = (\beta_S^T, 0_{p-4}^T)^T$. The absolute correlations between response and predictors are (0.390, 0.043, 0.304, 0.043, 0.130, 0.130, 0.130, 0.130, 0, 0, $\cdots$)^T.

(B). $r = 0.6, \sigma = 1.2, s_n = 5, \beta_S = (1, 1, -1, 1, -1)^T, \beta = (\beta_S^T, 0_{p-5}^T)^T$. The absolute correlations between response and predictors are (0.498, 0.498, 0.100, 0.498, 0.100, 0.299, 0.299, 0.299, 0.299, 0.299, 0, 0, $\cdots$)^T.

Scenario 2. The covariance matrix Σ is

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & I \end{pmatrix}, \text{ where } \Sigma_{11} = \begin{pmatrix} 1 & r_0 & r_1 & r_3 \\ r_0 & 1 & r_2 & r_4 \\ r_1 & r_2 & 1 & 0 \\ r_3 & r_4 & 0 & 1 \end{pmatrix}$$

(C). $r_0 = 0.8, r_1 = -r_2 = r_3 = -r_4 = -0.1, \sigma = 2.5, s_n = 2, \beta_S = (2.5, -2)^T, \beta = (\beta_S^T, 0_{p-2}^T)^T$. The absolute correlations between response and predictors are (0.309, 0.000, 0.154, 0.154, 0, 0, $\cdots$)^T.

(D). $r_0 = 0.75, r_1 = r_2 = r_3 = -r_4 = 0.2, \sigma = 2.5, s_n = 2, \beta_S = (2.5, -2)^T, \beta = (\beta_S^T, 0_{p-2}^T)^T$. The absolute correlations between response and predictors are (0.333, 0.0417, 0.033, 0.300, 0, 0, $\cdots$)^T.

The simulation results are shown in Tables 1 and 2 in which the (n, p_n) pair sequence is listed on the top row of each scenario in the Tables. The subscript for

RAR, RAR+, RAR(MC+) and RAR+(MC+) in the tables denotes the number of permutations in the retention step. For RAR(MC+) and RAR+(MC+), we only show the results with 30 permutations by noting that their improvement over RAR and RAR+ respectively is insensitive to the permutation numbers.

For Scenario 1, SIS-lasso fails to recover the sparsity pattern in both 1(A) and 1(B), due to that some signals and noises have correlations in similar magnitude with the response. ISIS-lasso substantially improves the performance of SIS-lasso and has similar performance as Lasso. The possible reason why it does not show clear advantage over Lasso is that the discrete stochastic process of the iterative algorithm may induce too much randomness. Ada-lasso is outperformed by Lasso for both 1(A) and 1(B), with the possible reason being that the weights are close to infinity for signals with small marginal correlation. Both RAR and RAR+ work very well in 1(A). For 1(B), RAR fails due to that there are noises with very large marginal correlation while RAR+ still has competitive performance. It is clear from Table 1 that the performance of RAR+ is more stable than that of RAR when the number of permutations changes, which verifies the theoretical results in Theorem 2. In addition, note that RAR+ with any number of permutations provides better performance than any non-RAR methods in both 1(A) and 1(B) across almost all (n, p_n) pairs. We also observe that the non-convex methods (SCAD and MC+) outperform Lasso. Accordingly, RAR(MC+) and RAR+(MC+) typically have better performance than

RAR and RAR+, respectively. Similar phenomenon can be observed regarding the comparison of SIS-lasso v.s. SIS-MC+ and ISIS-lasso v.s. ISIS-MC+. Moreover, the performance of screen and clean is inferior to that of RAR+, for all three versions considered. This is possibly because the approach splits the data into three parts and uses different parts for screen and clean, hence the sample size in each step is significantly reduced, leading to the loss of power for detecting signals (see Wasserman and Roeder (2009) for the detailed implementation).

We design Scenario 2, which is more challenging for Lasso to have sign consistency. For both 2(C) and 2(D), Lasso, SIS-lasso, ISIS-lasso and Ada-lasso all perform poorly. In contrast, RAR and RAR+ have similar performances as Scenario 1. An interesting observation is that MC+ outperforms the RAR+, probably due to that the ℓ_1 penalized step embedded in the procedure of RAR+ could be harmed by the high correlation among covariates. As expected, by using the MC+ penalty in the regularization step, RAR+(MC+) further improves both RAR+ and MC+. Similar as in Scenario 1, RAR+ outperforms the screen and clean approach.

To provide a more comprehensive comparison between different methods, we also calculate the oracle relative estimation error (the smallest relative estimation error of all estimators on the solution path) $\|\hat{\beta} - \beta\|^2/\|\beta\|^2$ for estimator $\hat{\beta}$ and its corresponding model size $\|\hat{\beta}\|_0$ for all scenarios. We observe that RAR+(MC+) has the smallest oracle estimation error with model size closest to the true model

Table 3: Average prediction mean square error and the average model size over 200 repetitions. The standard deviations of the error or model size are enclosed in parentheses. "Usize" denotes the size of the union of the selected variables across the 200 repetitions.

	Lasso	Ada-lasso	SCAD	MC+
Error	0.72 (0.34)	0.70 (0.29)	0.73 (0.35)	0.75 (0.37)
Size	63.0 (19.18)	72.5 (37.58)	53.6 (11.61)	48.8 (12.57)
Usize	1406	2065	1357	1357
	SIS-lasso	ISIS-lasso	SIS-MC+	ISIS-MC+
Error	0.83 (0.41)	0.85 (0.44)	0.95 (0.51)	0.91(0.45)
Size	24.0 (6.98)	20.4 (5.95)	7.5 (4.27)	7.8 (4.55)
Usize	343	273	205	173
	RAR	RAR+	RAR+(MC+)	SC-forward
Error	**0.69** (0.24)	0.71 (0.27)	0.72 (0.29)	0.81 (3.51)
Size	47.36 (27.43)	6.5 (1.17)	4.6 (2.32)	**1.29** (0.74)
Usize	1496	**56**	75	212

size in most cases. It is interesting to note that for some settings when the sample size is small, the one-step methods including SCAD and MC+ have smaller oracle estimation error than RAR+(MC+), however, they usually have a much larger model size than the truth. The detailed results for all four scenarios can be found in Tables 3-10 of the supplementary material.

4.2. Real Data Application

We compare the performances of Lasso, SCAD, MC+, SIS-lasso, ISIS-lasso, SIS-MC+, ISIS-MC+, Ada-lasso, SC-lasso, SC-forward, SC-marginal, RAR, RAR+, and

RAR+(MC+) on the data set reported by Scheetz et al. (2006). For this data set, 120 twelve-week old male rats were selected for tissue harvesting from the eyes. The microarrays used to analyze the RNA from the eyes of these rats contain over $31,042$ different probes (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multi-chip averaging method (Irizarry et al., 2003) to obtain summary expression values for each probe. Gene expression levels were analyzed on a logarithmic scale. Following Fan et al. (2001), we only focus on the $18,975$ probes that are expressed in the eye tissue. We are interested in finding the genes that are related to the gene TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang et al., 2006), and is a genetically heterogeneous disease of multiple organ systems including the retina.

The dataset includes $n = 120$ samples and $p = 18,975$ variables. We randomly partition the data into a training set of 96 observations and a test set of 24 observations. We use 5-fold cross validation for tuning parameter selection on the training set for the last regularization step of each method (cleaning step in the screen and clean method) and calculate the prediction mean square error on the test set. For the second step of RAR+ and RAR+(MC+), generalized information criterion is employed (Fan and Tang, 2013). The whole procedure is repeated 200 times. To evaluate the stability of different methods, we also calculate the size of the union of the selected variables across the 200 repetitions for each method. A summary of

prediction error (Error), selected model size (Size) and the size of the union of the selected variables (Usize) over 200 repetitions are reported in Table 3 to evaluate the performance of different methods.

As shown in Table 3, RAR performs the best in terms of prediction error. It selects fewer variables than Lasso does, but has larger variation in terms of model selection (based on standard deviation of model size and Usize). On average, RAR+ selects a more parsimonious model with slightly larger prediction error than RAR and Ada-lasso. It is worth noting that RAR+ has the smallest Usize among all considered methods. Independence screening-based methods lead to sparser models than RAR, but they have larger prediction errors on average. For SIS-lasso, ISIS-lasso, SIS-MC+ and ISIS-MC+, we select the top 60 variables in the screening step. We also try other thresholds and they lead to similar results. The reason may be that there exist signals that are weakly correlated with the response so that even ISIS-lasso misses them. In addition, we observe that RAR+(MC+) selects fewer variables with a slightly larger prediction error, when compared with RAR+, respectively. Similar observations can be made when comparing the nonconvex penalties SCAD and MC+ with Lasso, SIS-MC+ with SIS-lasso and ISIS-MC+ with ISIS-lasso. Note that SC-forward has the smallest model size, however, it has a much larger prediction error than our methods. Moreover, the standard deviation of its prediction error is very large, which might be due to large variation of selected variable set across 200 repetitions (implied by the

large Usize). We omit the results for SC-Lasso and SC-marginal as they have larger prediction errors than SC-forward.

5. Discussion

The proposed regularization after retention method is a general framework for model selection and estimation. In the retention step, there exist alternatives to obtain the retention set beyond those using marginal information. For example, we can use forward regression with early stopping. In the regularization step, it would be interesting to study the corresponding theoretical results when penalty functions other than the ℓ_1 norm (e.g., SCAD (Fan and Li, 2001), MC+ (Zhang, 2010)) are used.

The theoretical justification of the permutation approach for choosing the threshold γ_n in the retention step is an open problem. Parameter estimation consistency and persistency of the new framework could be an interesting future work. Theoretical extension for sub-Gaussian distributions of X and ε is possible. It might be also worth considering extensions to other models including generalized linear models, additive models, and semi-parametric models.

Supplementary Materials

The online supplementary material contains the proof of all theoretical results and additional simulation results.

Acknowledgment

We thank the Editor, the AE and two anonymous referees for many constructive comments that have greatly improved the scope of the paper. Feng is partially supported by NSF grant DMS-1554804. Qiao is partially supported by a grant from the Simons Foundation.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**, 716-723.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705-1732.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 232-253.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., 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 475 ubiquitin ligase, as a bardetcbiedl syndrome gene (bbs11). *PNAS*, **103**, 6287-6292.

Cho, H. and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**,

593-622.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106**, 544-557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849-911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* **38**, 3567-3604.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 531-552.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1-22.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971-988.

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

Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21**, 1473-1513.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Ji, Pengsheng, and Jin, Jiashun (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics* **40**, 73-103.

Kerkyacharian, G., Mougeot, M., Picard, D. and Tribouley, K. (2009). Learning out of leaders. *Multiscale, Nonlinear and Adaptive Approximation, Springer, Berlin.* 295-324.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**, 1356-1378.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 18461877.

Li, R., Zhong, W., and Zhu, L. (2012), Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129-1139.

Luo, S. and Chen, Z. (2011). Sequential lasso for feature selection with ultra-high dimensional feature space. *arXiv preprint arXiv:1107.2734.*

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., Dibona, G. F., Huang, J., Casavant, T. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *PNAS* **103**, 14429-14434.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery. Information Theory, *IEEE Transactions on* **55**, 2183-2202.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics* **37**, 2178.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894-942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7**, 2541-2563.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-Free Feature Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association* **106**, 1464-1475.

Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. Manuscript.

Zhou, S., Van de geer, S. and Buhlmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515.*

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320.

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

E-mail: hw2375@columbia.edu, yang.feng@columbia.edu

Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY 13902, U.S.A.

E-mail: qiao@math.binghamton.edu