Statistica Sinica **25** (2015), 975-992
doi:http://dx.doi.org/10.5705/ss.2013.001

REGULARIZING LASSO: A CONSISTENT VARIABLE SELECTION METHOD

Quefeng Li[1] and Jun Shao[1,2]

[1] *University of Wisconsin, Madison and* [2] *East China Normal University*

Abstract: LASSO for variable selection in linear regression has been studied by many authors. To achieve asymptotic selection consistency, it is well known that the LASSO method requires a strong irrepresentable condition. Even adding a thresholding step after LASSO is still too conservative, especially when the number of explanatory variables p is much larger than the number of observations n. Another well-known method, the sure independence screening (SIS), applies thresholding to an estimator of marginal covariate effect vector and, therefore, is not selection consistent unless the zero components of the marginal covariate effect vector are asymptotically the same as the zero components of the regression effect vector. Since the weakness of LASSO is caused by the fact that it utilizes the covariate sample covariance matrix that is not well behaved when p is larger than n, we propose a regularized LASSO (RLASSO) method for replacing the covariate sample covariance matrix in LASSO by a regularized estimator of covariate covariance matrix and adding a thresholding step. Using a regularized estimator of covariate covariance matrix, we can consistently estimate the regression effects and, hence, our method also extends and improves the SIS method that estimates marginal covariate effects. We establish selection consistency of RLASSO under conditions that the regression effect vector is sparse and the covariate covariance matrix or its inverse is sparse. Some simulation results for comparing variable selection performances of RLASSO and various other methods are presented. A data example is also provided.

Key words and phrases: High-dimensional data, LASSO, regularization, selection consistency, sparsity, thresholding.

1. Introduction

In many statistical applications, one investigates the effect of a vector $\boldsymbol{x}$ of p explanatory variables on a response variable y based on n independently observed data $\{y_i, \boldsymbol{x}_i, i = 1, \ldots, n\}$ following a linear model

$$y_i = \mu + \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma_i\varepsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Here y_i is the ith observed response, $\boldsymbol{x}_i$ is the p-dimensional observed explanatory variables associated with y_i, $\boldsymbol{x}_i$'s are independent and identically distributed (i.i.d.), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of unknown parameters called

regression effects, μ is an unknown parameter (intercept), σ_i's are positive unknown parameters, ε_i's are i.i.d. unobserved random errors with mean 0 and variance 1, $\boldsymbol{x}_i$'s and ε_i's are independent, and $\boldsymbol{A}'$ denotes the usual transpose of a vector or matrix $\boldsymbol{A}$. The theory of linear models is well established for traditional applications where the dimension p is fixed and the sample size $n > p$. With modern technologies, however, in many biological, medical, social, and economical studies, p is comparable with or much larger than n. The variable j, the jth component of $\boldsymbol{x}$, has no effect on the response if $\beta_j = 0$. When the number of variables p is large but many variables have no effect on the response, which is often true in applications. The identification of the zero components of $\boldsymbol{\beta}$ is usually made prior to statistical inference. Without loss of generality we assume that the $\boldsymbol{x}_i$'s have mean 0 and variance 1 and are standardized so that $\sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{0}$, and that the diagonal elements of $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'/n$ are equal to 1; this does not affect variable selection.

There is a rich literature on asymptotic theory for variable selection in the case where $n \to \infty$ and p is fixed, or $p \to \infty$ at a rate much slower than n. For variable selection when $p > n$ with $p = O(n^l)$ for some $l > 1$ or $O(e^{n^\nu})$ for some $\nu \in (0, 1)$ (ultra-high dimension), some excellent advances in asymptotic theory have been made. See, for example, the review paper Fan and Lv (2010).

Let $\mathcal{M}_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$ denote the index set for nonzero components of $\boldsymbol{\beta}$ and let $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$ denote the set of indices of nonzero components of $\boldsymbol{\beta}$ selected by a variable selection method using data. The selection method is selection-consistent if

$$P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\boldsymbol{\beta}}\right) \to 1, \tag{1.2}$$

where the limit is taken as $n \to \infty$ with $p = p_n$ that may also diverge to ∞ and the probability is with respect to the randomness of data $\{y_i, \boldsymbol{x}_i, i = 1, \ldots, n\}$. Selection-consistency is important, since it leads to oracle properties of estimation and inference procedures (see, e.g., Fan and Lv (2008)). Some results on selection-consistency have been established under conditions that do not generally hold. For example, the LASSO method (Tibshirani (1996)) requires a strong irrepresentable condition (see (4.1) in Section 4) for its selection-consistency. The sure independent screening (SIS) in Fan and Lv (2008) requires that

$$\min_{j \in \mathcal{M}_\beta} \left| \sum_{k \in \mathcal{M}_\beta} \beta_k \rho_{kj} \right| \geq c_0 n^{-\kappa} \tag{1.3}$$

for some constant $c_0 > 0$ and $0 \leq \kappa < 1/2$, where ρ_{kj} is the correlation coefficient between the jth and kth components of $\boldsymbol{x}$; under some regularity conditions, the SIS is screening consistent in the sense that $P\left(\mathcal{M}_{\boldsymbol{\beta}} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}}\right) \to 1$. However, (1.3)

can be questionable, see Model 4 in Section 5. The SIS is selection-consistent if, in addition to (1.3),

$$\max_{j \notin \mathcal{M}_\beta} \left| \sum_{k \in \mathcal{M}_\beta} \beta_k \rho_{kj} \right| = o(n^{-\kappa}) \tag{1.4}$$

holds. Condition (1.4) is rarely satisfied in practice, since it imposes a strong structure on the correlation coefficients ρ_{kj}. The SIS has a reputation of being screening consistent only, not selection consistent.

The purpose of this paper is to derive a variable selection method that is selection-consistent without requiring conditions (1.3) and (1.4), or the strong irrepresentable condition. These conditions are replaced by a sparsity condition on the covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{x}'\boldsymbol{x})$ (Bickel and Levina (2008) or Cai, Zhang, and Zhou (2010)), or a sparsity condition on the inverse of $\boldsymbol{\Sigma}$ (Cai, Liu, and Luo (2011)). The key idea is, since the LASSO utilizes a least squares minimization involving the covariate sample covariance matrix, which is not well behaved when p is larger than n, to replace the least squares component in the minimization of LASSO by a regularized least squares component using results in high-dimensional covariance matrix estimation (Bickel and Levina (2008), Cai, Zhang, and Zhou (2010), Cai, Liu, and Luo (2011)). A thresholding step is added to the resulting estimator to improve its variable selection performance.

The proposed procedure, the regularized LASSO (RLASSO), is introduced in Section 2. Section 3 contains results on the selection-consistency of the proposed method. A comparison of LASSO and RLASSO is given in Section 4. Section 5 provides some simulation results on the performance of the proposed method and several other variable selection methods, and a data example is given. The last section contains some discussions and recommendations. All proofs are given in a separate web appendix.

2. The Methodology

We first introduce a simple procedure that is selection-consistent. The idea is simple. If p is fixed, then we can select variables by thresholding the least squares estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{\mathrm{lse}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{S}^{-1}\boldsymbol{X}'\frac{\boldsymbol{y}}{n},$$

where $\boldsymbol{y}$ is the n-dimensional vector of y_i's, $\boldsymbol{X}$ is the $n \times p$ matrix whose ith row is $\boldsymbol{x}_i$, and $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}/n$. But when $p > n$, $\boldsymbol{S}$ is singular and even if we use a generalized inverse as its inverse, $\hat{\boldsymbol{\beta}}_{\mathrm{lse}}$ does not have a good behavior because $\boldsymbol{S}$ is not a good estimator of the covariate covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{S})$. If $\boldsymbol{\Sigma}$ is sparse in some sense, then we may estimate $\boldsymbol{\Sigma}$ by a regularized or sparse estimator $\hat{\boldsymbol{\Sigma}}$ that is L_2-consistent in the sense that $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 = o_p(1)$, where $\|\boldsymbol{A}\|_2$ is the L_2

norm of a matrix $\boldsymbol{A}$. Such an estimator can be obtained using results in high-dimensional covariance matrix estimation (Bickel and Levina (2008); Cai and Liu (2011)). Then we estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}_{\text{slse}} = \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}'\frac{\boldsymbol{y}}{n}, \tag{2.1}$$

and select those components of $\hat{\boldsymbol{\beta}}_{\text{slse}}$ whose absolute values are larger than a threshold.

A proof of the method's selection-consistency, under some conditions, is a special case of the general results we establish in Theorems 1 and 2. The method is computationally simple, if $\hat{\boldsymbol{\Sigma}}$ is obtained by thresholding elements of $\boldsymbol{S}$. In what follows we derive a general method for variable selection, and show that it is selection-consistent and has better finite sample properties than the simple method of thresholding $\hat{\boldsymbol{\beta}}_{\text{slse}}$ in (2.1).

Since $\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\beta}_M$, the true $\boldsymbol{\beta}$ is a solution to

$$\min_{\boldsymbol{\beta}}\left(\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2} - \boldsymbol{\beta}_M'\boldsymbol{\beta}\right)$$

for any fixed $\boldsymbol{\beta}_M$ and $\boldsymbol{\Sigma}$, and the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{\text{lse}}$ is a solution to

$$\min_{\boldsymbol{\beta}}\left(\frac{\boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}}{n}\right). \tag{2.2}$$

When $p > n$, the estimator $\hat{\boldsymbol{\beta}}_{\text{slse}}$ in (2.1) improves on the least squares estimator by replacing $\boldsymbol{S}$ in (2.2) with an L_2-consistent sparse estimator $\hat{\boldsymbol{\Sigma}}$.

The LASSO is a solution to

$$\min_{\boldsymbol{\beta}}\left(\frac{\boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}}{n} + \lambda_n\|\boldsymbol{\beta}\|_1\right), \tag{2.3}$$

where $\|\boldsymbol{\beta}\|_1$ is the L_1-norm of the vector $\boldsymbol{\beta}$ and $\lambda_n \geq 0$ is a tuning parameter. Using the same penalty idea, we consider regularizing the LASSO by replacing $\boldsymbol{S}$ in (2.3) by an L_2-consistent sparse estimator $\hat{\boldsymbol{\Sigma}}$. This leads to the minimization problem

$$\min_{\boldsymbol{\beta}}\left(\frac{\boldsymbol{\beta}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}}{n} + \lambda_n\|\boldsymbol{\beta}\|_1\right). \tag{2.4}$$

In addition to the regularization step in the estimation of $\boldsymbol{\Sigma}$ and the L_1 penalization, our proposed method thresholds the solution to (2.4). If $\tilde{\boldsymbol{\beta}}$ is a solution to (2.4), estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\beta}_1 I(|\tilde{\beta}_1| > t_n), \ldots, \tilde{\beta}_p I(|\tilde{\beta}_p| > t_n)\right)', \tag{2.5}$$

where $\tilde{\beta}_j$ is the jth component of $\tilde{\boldsymbol{\beta}}$, $I(A)$ is the indicator function of the event A, and t_n is an appropriate threshold. Our method can be viewed as adding two regularization steps to the LASSO, and it will be referred to as the regularized LASSO (RLASSO).

If we choose $\lambda_n = 0$ in (2.4), then RLASSO reduces to thresholding $\hat{\boldsymbol{\beta}}_{\text{slse}}$. If we ignore the regularization step in the estimation of $\boldsymbol{\Sigma}$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, then RLASSO is thresholding the LASSO estimator as discussed in Meinshausen and Yu (2009); and if the last step of thresholding is also ignored, then RLASSO is LASSO. If we choose $\hat{\boldsymbol{\Sigma}}$ to be the $p \times p$ identity matrix, which can be viewed as a particular type of regularization ignoring all correlations among components of $\boldsymbol{x}_i$, and if we also choose $\lambda_n = 0$, then RLASSO is SIS. If we choose $\hat{\boldsymbol{\Sigma}}$ to be the inverse of the graphical LASSO estimator of $\boldsymbol{\Sigma}^{-1}$, and if we ignore the last step of thresholding, then RLASSO is the "Scout" method proposed by Witten and Tibshirani (2009).

In general, there are two ways to obtain regularized estimator of $\boldsymbol{\Sigma}$, depending on whether $\boldsymbol{\Sigma}$ is sparse or its inverse $\boldsymbol{\Sigma}^{-1}$ is sparse. If $\boldsymbol{\Sigma}$ is sparse, we can apply thresholding $\boldsymbol{S}$ as proposed by Bickel and Levina (2008), or the adaptive thresholding method in Cai and Liu (2011). Both methods provide L_2-consistent estimators of $\boldsymbol{\Sigma}$. Take $\hat{\sigma}_{ij}$ to be the (i, j)th element of $\boldsymbol{S}$. The adaptive thresholding method estimates $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}^*_{ij})_{p\times p}$, where $\hat{\sigma}^*_{ij}$ is $\hat{\sigma}_{ij}$ being thresholded at $\delta\left\{(\log p/n^2)\sum_{k=1}^n [x_{ki}x_{kj} - \hat{\sigma}_{ij}]^2\right\}^{1/2}$. We choose the tuning parameter (δ, λ_n, t_n) of this procedure by minimizing the Bayesian Information Criterion (BIC):

$$(\hat{\delta}, \hat{\lambda}_n, \hat{t}_n) = \underset{\delta\geq 2, \lambda_n, t_n}{\operatorname{argmin}} -2\ell(\hat{\boldsymbol{\beta}}_{\delta,\lambda_n,t_n}) + s\log(n), \tag{2.6}$$

where $\ell(\hat{\boldsymbol{\beta}}_{\delta,\lambda_n,t_n})$ is the log-likelihood based on $\hat{\boldsymbol{\beta}}_{\delta,\lambda_n,t_n}$ under a particular choice of (δ, λ_n, t_n), and s is the number of non-zero elements in $\hat{\boldsymbol{\beta}}_{\delta,\lambda_n,t_n}$.

If $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse, we can obtain a regularized estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ and estimate $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Omega}}^{-1}$. For example, Friedman, Hastie, and Tibshirani (2008) proposed the graphical LASSO estimator of $\boldsymbol{\Omega}$; Cai, Liu, and Luo (2011) proposed the CLIME estimator of $\boldsymbol{\Omega}$ and proved its consistency under the L_1-norm and the L_2-norm.

In applications, one has to make a judgment on which of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ is sparse in order to apply the proposed RLASSO. We discuss this issue in Section **??** after some theoretical and empirical results are presented.

3. Asymptotic Results

In this section we establish the selection-consistency of the RLASSO. For asymptotic results when $p \to \infty$ at a rate faster than n, intuitively we need tail conditions on ε_i and x_j, the jth component of $\boldsymbol{x}$, a sparsity condition on

the vector $\boldsymbol{\beta}$, and a sparsity condition on the covariance matrix $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

First, consider that $\boldsymbol{\Sigma}$ is sparse. To measure the sparsity of $\boldsymbol{\Sigma}$, we use

$$r_q = \max_{1\le i\le p} \sum_{j=1}^{p} |\rho_{ij}|^q, \tag{3.1}$$

where ρ_{ij} is the (i,j)th element of $\boldsymbol{\Sigma}$ and $0 \le q < 1$ is a constant not depending on p or n. This measure was considered by Bickel and Levina (2008). When $q = 0$, r_0 is simply the maximum of the numbers of nonzero components of rows of $\boldsymbol{\Sigma}$. If $r_q \to \infty$ at a rate much slower than p (e.g. (C4) in Theorem 1), then $\boldsymbol{\Sigma}$ is considered sparse. In this subsection, we estimate $\boldsymbol{\Sigma}$ by the adaptive thresholding estimator in Cai and Liu (2011).

To measure the sparsity of $\boldsymbol{\beta}$, we consider

$$s_h = \sum_{j=1}^{p} |\beta_j|^h \tag{3.2}$$

for some $h \in [0, 1)$. In the special case that $h = 0$ in (3.2), s_0 is the number of non-zero components of $\boldsymbol{\beta}$.

Let $\hat{\boldsymbol{\beta}}_M = \boldsymbol{X}'\boldsymbol{y}/n$. In the following, a quantity is said to be a constant if it does not depend on n or p, but may depend on some unknown population parameters. For two sequences a_n and b_n, $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$.

Lemma 1. *Assume there exist positive constants m and M such that*
(C1) $\max_{1\le j\le p} \mathrm{E}[\exp(tx_j^2)] \le M$ *and* $\mathrm{E}[\exp(t\varepsilon_i^2)] \le M$ *for all* $|t| \le m$;
(C2) $\max_{1\le i\le n} \sigma_i \le M < \infty$ *and* $\|\boldsymbol{\beta}\|_\infty \le M < \infty$.
Then there exist positive constants C_1, C_2, and C_3 such that, for all $0 < t \le C_3 s_h$,

$$P\Big(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t\Big) \le 2p^2 \exp\Big(-\frac{C_1 n t^2}{s_h^2}\Big) + 4p\exp(-C_2 n t^2).$$

Lemma 2. *Assume* (C1), (C2) *and*
(C3) $\log p \asymp n^\tau$ *and* $\min_{j,k} \mathrm{Var}(x_j x_k) \ge m$, *where* $0 < \tau < 1/3$ *and* $m > 0$ *are constants.*
For any λ_n in (2.4) *such that $\lambda_n v_p \to 0$, where $v_p = \|\boldsymbol{\Sigma}^{-1}\|_1$, there exist positive constants C_4, C_5, and C_6 such that the solution to* (2.4) *satisfies*

$$P\Big(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\Big) \le 2p^2 \exp\Big(-\frac{C_4 n t^2}{(s_h v_p)^2}\Big) + 4p\exp\Big(-\frac{C_5 n t^2}{v_p^2}\Big) + C_6 n^{-1/2} p^{-(\delta-2)} \Big(\frac{r_q v_p}{t}\Big)^{1/(1-q)}.$$

The following result establishes the selection-consistency of the RLASSO. In the rest of this section, $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$ denotes the index set of nonzero components of $\hat{\boldsymbol{\beta}}$ defined in (2.5).

Theorem 1. *Assume* (C1)−(C3) *and*

(C4) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, $v_p \asymp n^{\alpha_3}$, *where* α_1, α_2 *and* α_3 *are positive constants satisfying* $\alpha_1 + \alpha_3 < (1-\tau)/2$ *and* $\alpha_2 + \alpha_3 < (1-q)/2$. *If, in* (2.4)−(2.5), $\lambda_n = M_1(n^{-1}\log p)$ *for some constant* $M_1 > 0$ *and* $t_n = M_2 n^{-\eta}$ *for some constant* $M_2 > 0$ *with* $0 < \eta < \min\{(1-\tau)/2 - \alpha_1 - \alpha_3, (1-q)/2 - \alpha_2 - \alpha_3\}$, *then there exists a positive constant* C_7 *such that*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta},a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}\right) \\ &\quad = O\Big[\exp\left(-C_7(\log n)^{-2} n^{1-2\alpha_1-2\alpha_3-2\eta}\right) \\ &\qquad + (\frac{1}{p})^{\delta-2}\Big((\log n)(\frac{1}{n})^{(1-q)/2-\alpha_2-\alpha_3-\eta}\Big)^{1/(1-q)}\Big], \end{aligned}$$

where $\mathcal{M}_{\boldsymbol{\beta},d_n}$ *denotes the index set of components of* $\boldsymbol{\beta}$ *whose absolute values are larger than* d_n *and* $a_n - 1 \asymp (\log n)^{-1}$. *If* $h = 0$ *in* (3.2) *and we additionally assume*

(C5) $\liminf_{n\to\infty} n^{\kappa} \min_{j\in\mathcal{M}_{\boldsymbol{\beta}}} |\beta_j| > 0$ *with* $\kappa < \min\{(1-\tau)/2 - \alpha_1 - \alpha_3, (1-q)/2 - \alpha_2 - \alpha_3\}$, *and if* $t_n = M_2 n^{-\eta}$ *with* $\kappa < \eta < \min\{(1-\tau)/2 - \alpha_1 - \alpha_3, (1-q)/2 - \alpha_2 - \alpha_3\}$, *then*

$$\begin{aligned} &P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} \neq \mathcal{M}_{\boldsymbol{\beta}}\right) \\ &\quad = O\left[\exp\left(-C_8 n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (\frac{1}{p})^{\delta-2}(\frac{1}{n})^{((1-q)/2-\alpha_2-\alpha_3-\eta)/(1-q)}\right]. \end{aligned}$$

Theorem 1 is more general than the selection-consistency defined by (1.2). With probability tending to 1, the RLASSO eliminates all components of $\boldsymbol{\beta}$ whose absolute values are no larger than t_n/a_n, and retains all components of $\boldsymbol{\beta}$ whose absolute values are no smaller than $t_n a_n$. Since $a_n - 1 \asymp (\log n)^{-1}$, the RLASSO asymptotically eliminates or retains variables according to whether the components of $\boldsymbol{\beta}$ is smaller or larger than the threshold t_n. Thus, if $\boldsymbol{\beta}$ is sparse in the sense that s_0 (s_h in (3.2) with $h = 0$) diverges much slower than p, then the RLASSO is selection-consistent in the sense of (1.2), provided that the minimum of nonzero components of $\boldsymbol{\beta}$ does not decay too fast.

Condition (C2) holds in most applications. The bounded $\|\boldsymbol{\beta}\|_\infty$ condition can be relaxed by $\|\boldsymbol{\beta}\|_\infty^{1-h} s_h \sqrt{n^{-1}\log p} \to 0$ at some rate. It is needed because we estimate $\boldsymbol{\beta}$ through the estimation of $\boldsymbol{\beta}_M$ and $\boldsymbol{\Sigma}$.

Condition (C1) requires that the distributions of x_j's and ε_i have exponential tails. Asymptotic results can also be established when the distributions of x_j's and ε_i have polynomial tails.

Lemma 3. *Assume that there exist constants $M > 0$ and $l > 1$ such that*
(C1′) $\max_{1\le j\le p} \mathrm{E}x_j^{4l} \le M$ *and* $\mathrm{E}\varepsilon_i^{4l} \le M$.
If (C2) *also holds, then there exist some positive constants C_9 and C_{10} such that*

$$P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t\right) \le C_9 p^2 s_h^{2l} t^{-2l} n^{-l} + C_{10} p t^{-2l} n^{-l}.$$

Lemma 4. *Assume* (C1′), (C2), *and*

(C3′) $p \asymp n^\tau$, *where* $\tau < \min\{l/2, l-1\}$.

For any λ_n in (2.4) *such that $\lambda_n v_p \to 0$, there exist positive constants C_{11}, C_{12}, and C_{13} such that*

$$\begin{aligned} P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \le{}& C_{11} p^2 s_h^{2l} v_p^{2l} t^{-2l} n^{-l} + C_{12} p v_p^{2l} t^{-2l} n^{-l} \\ &+ C_{13}\left(n^{-1/2} p^{-(\delta-2)}\left(\frac{r_q v_p}{t}\right)^{1/(1-q)} + n^{-(l-1-\tau)/2}\right). \end{aligned}$$

Theorem 2. *Assume* (C1′), (C2), (C3′), *and* (C4′) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, $v_p \asymp n^{\alpha_3}$ *where α_1, α_2 and α_3 are positive constants satisfying $\alpha_1 + \alpha_3 < 1/2 - \tau/l$ and $\alpha_2 + \alpha_3 < (1/2 + [\delta-2]\tau)(1-q)$. If $\lambda_n = M_3 n^{-1/2}$ for some $M_3 > 0$, $t_n = M_4 n^{-\eta}$ for some $M_4 > 0$, and $0 < \eta < \min\{1/2 - \tau/l - \alpha_1 - \alpha_3, (1/2 + [\delta-2]\tau)(1-q) - \alpha_2 - \alpha_3\}$, and $a_n - 1 \asymp (\log n)^{-1}$, then*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta}, a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta}, t_n/a_n}\right) \\ &\quad = O\Big\{\left(n^{-2l(1/2-\tau/l-\alpha_1-\alpha_3-\eta)} + n^{-\{(1/2+[\delta-2]\tau)(1-q)-\alpha_2-\alpha_3-\eta\}/(1-q)}\right)(\log n)^{2l} \\ &\qquad + n^{-(l-1-\tau)/2}\Big\}. \end{aligned}$$

If $h = 0$ in (3.2), (C5) *holds for some constant $\kappa < \min\{1/2 - \tau/l - \alpha_1 - \alpha_3, (1/2 + [\delta-2]\tau)(1-q) - \alpha_2 - \alpha_3\}$, and if $t_n = M_4 n^{-\eta}$ with $M_4 > 0$ and $\kappa < \eta < \min\{1/2 - \tau/l - \alpha_1 - \alpha_3, (1/2 + [\delta-2]\tau)(1-q) - \alpha_2 - \alpha_3\}$, then*

$$\begin{aligned} P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} \ne \mathcal{M}_{\boldsymbol{\beta}}\right) = O\Big(&n^{-2l(1/2-\tau/l-\alpha_1-\alpha_3-\eta)} + n^{-\{(1/2+[\delta-2]\tau)(1-q)-\alpha_2-\alpha_3-\eta\}/(1-q)} \\ &+ n^{-(l-1-\tau)/2}\Big). \end{aligned}$$

Consider now that $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse. In general, there is no relationship between the sparsity of $\boldsymbol{\Sigma}$ and the sparsity of its inverse $\boldsymbol{\Omega}$. For a sparsity measure of $\boldsymbol{\Omega}$, we still use the notation r_q in (3.1), but with ρ_{ij}'s replaced by the (i,j)th element of $\boldsymbol{\Omega}$. All asymptotic results here are based on this sparsity measurement of $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}$ is estimated by the CLIME in Cai, Liu, and Luo (2011).

Theorem 3. *Assume* (C1), (C2), (C3″) $\|\mathbf{\Omega}\|_1 \le M$ *and* $\log p \asymp n^{\tau}$, *where* $0 < \tau < 1/4$, *and (C4″)* $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, *where* $\alpha_1 < (1-\tau)/2$ *and* $\alpha_2 < (1-q)/2 - \alpha_1$ *are positive constants. If* $\lambda_n = M_5(n^{-1}\log p)$ *for some constant* $M_5 > 0$, $t_n = M_6 n^{-\eta}$ *with a constant* $M_6 > 0$ *and* $0 < \eta < \min\{(1-\tau)/2-\alpha_1, (1-q)/2-\alpha_1-\alpha_2\}$, *and* $a_n - 1 \asymp (\log n)^{-1}$, *then there exist constants* C_{17} *and* C_{18} *such that*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta},a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}\right) \\ &\quad = O\Big[\exp\left(-C_{17}\left[(\log n)^{-1} n^{(1-q)/2-\alpha_1-\alpha_2-\eta}\right]^{2/(1-q)}\right) \\ &\qquad + \exp\left(-C_{18}(\log n)^{-2} n^{2(1/2-\alpha_1-\eta)}\right)\Big]. \end{aligned}$$

If (3.2) *holds for* $h = 0$ *and* (C5) *holds for some constant* $\kappa < \min\{(1-\tau)/2 - \alpha_1, (1-q)/2-\alpha_1-\alpha_2\}$, *and if* $t_n = M_6 n^{-\eta}$ *with* $M_6 > 0$ *and* $\kappa < \eta < \min\{(1-\tau)/2-\alpha_1, (1-q)/2-\alpha_1-\alpha_2\}$, *then*

$$\begin{aligned} &P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} \neq \mathcal{M}_{\boldsymbol{\beta}}\right) \\ &\quad = O\Big[\exp\left(-C_{19}\left[n^{(1-q)/2-\alpha_1-\alpha_2-\eta}\right]^{2/(1-q)}\right) + \exp\left(-C_{20} n^{2(1/2-\alpha_1-\eta)}\right)\Big]. \end{aligned}$$

Theorem 4. *Assume* (C1′), (C2), (C3‴) $\|\mathbf{\Omega}\|_1 \le M$ *and* $p \asymp n^{\tau}$, *where* $\tau < \min\{l/2, l-1\}$, *and* (C4‴) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, *where* $\alpha_1 < 1/2 - \tau/l$ *and* $\alpha_2 < (1-q)/2 - \alpha_1$ *are positive constants. If* $\lambda_n = M_7 n^{-1/2}$ *for some* $M_7 > 0$, $t_n = M_8 n^{-\eta}$ *with constants* $M_8 > 0$ *and* $0 < \eta < \min\{1/2 - \tau/l - \alpha_1, (1-q)/2 - \alpha_1 - \alpha_2\}$, *and* $a_n - 1 \asymp (\log n)^{-1}$, *then there exists a positive constant* C_{25} *such that*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta},a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}\right) \\ &\quad = O\Big[\exp\left(-C_{25}\left[(\log n)^{-1} n^{(1-q)/2-\alpha_1-\alpha_2-\eta}\right]^{2/(1-q)}\right) \\ &\qquad + \left(\frac{1}{n}\right)^{2l(1/2-\tau/l-\alpha_1-\eta)} (\log n)^{2l} + \left(\frac{1}{n}\right)^{(l-1-\tau)/2}\Big]. \end{aligned}$$

If (3.2) *holds for* $h = 0$ *and* (C5′) $\liminf_{n\to\infty} n^{\kappa} \min_{j\in\mathcal{M}_{\boldsymbol{\beta}}} |\beta_j| > 0$ *for some constant* $0 < \kappa < \min\{1/2 - \tau/l - \alpha_1, (1-q)/2 - \alpha_1 - \alpha_2\}$, *and if* $t_n = M_8 n^{-\eta}$ *with* $M_8 > 0$ *and* $\kappa < \eta < \min\{1/2 - \tau/l - \alpha_1, (1-q)/2 - \alpha_2\}$, *then there exist a positive constant* C_{25} *such that*

$$\begin{aligned} P\left(\widehat{\mathcal{M}_{\boldsymbol{\beta}}} \neq \mathcal{M}_{\boldsymbol{\beta}}\right) = O\Big[&\exp\left(-C_{26}\left[n^{(1-q)/2-\alpha_1-\alpha_2-\eta}\right]^{2/(1-q)}\right) \\ &+ \left(\frac{1}{n}\right)^{2l(1/2-\tau/l-\alpha_1-\eta)} + \left(\frac{1}{n}\right)^{(l-1-\tau)/2}\Big]. \end{aligned}$$

4. Comparison of LASSO and RLASSO

We take $\boldsymbol{X}_1$ to be the s_0 columns of $\boldsymbol{X}$ corresponding to non-zero components of $\boldsymbol{\beta}$ and $\boldsymbol{X}_2$ to be those columns of $\boldsymbol{X}$ corresponding to zero components. Zhao and Yu (2006) showed the LASSO to be selection-consistent, if $\boldsymbol{S}$ satisfies the strong irrepresentable condition (SIC)

$$\eta_\infty = 1 - \|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\text{sign}(\boldsymbol{\beta}_1)\|_\infty \geq \gamma, \tag{4.1}$$

where $\boldsymbol{S}_{11} = \boldsymbol{X}_1'\boldsymbol{X}_1/n$, $\boldsymbol{S}_{21} = \boldsymbol{X}_2'\boldsymbol{X}_1/n$, γ is a positive constant, $\boldsymbol{\beta}_1 = \{\beta_j, j \in \mathcal{M}_{\boldsymbol{\beta}}\}$, and $\text{sign}(\boldsymbol{\beta}_1)$ is the vector whose components are the signs of components of $\boldsymbol{\beta}_1$. The SIC is also essentially necessary for LASSO to be selection-consistency.

An example shows a situation in which RLASSO works but LASSO fails. Suppose that, in (1.1), ϵ_i's are i.i.d. from $N(0,1)$, $\mu = 0$, $\sigma_i = 1$, $\boldsymbol{x}_i$'s are i.i.d. from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_{s_0} & \boldsymbol{B}' & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{I}_{k-s_0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{p-k} \end{pmatrix}, \tag{4.2}$$

and $\boldsymbol{B}$ is a $(k-s_0)\times s_0$ matrix satisfying $\|\boldsymbol{B}\|_\infty \geq 1+2\gamma$ for some $\gamma > 0$ and ensuring that $\boldsymbol{\Sigma}$ is positive definite. Take s_0 and k as fixed numbers, the first s_0 components of $\boldsymbol{\beta}$ equal to 1, and $n^{-1}\log p \to 0$. It is shown in the Web Appendix that

$$P(\|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\|_\infty \geq 1+\gamma) \to 1. \tag{4.3}$$

Thus the SIC fails with probability tending to 1, which implies that LASSO cannot be selection-consistent. On the other hand, it can be verified that (C1)−(C5) hold if we further assume $\log p = o(n^{1/3})$. Hence, RLASSO is selection-consistent. Since RLASSO uses a consistent estimator of $\boldsymbol{\Sigma}$, it avoids the requirement of a condition like the SIC.

We carried out another simulation to compare RLASSO and LASSO in an example from Zhao and Yu (2006). We took $n = 100$, $p = 32$, and $\boldsymbol{\beta} = (7,4,2,1,1,0,\ldots,0)$. We generated a covariance matrix $\boldsymbol{\Sigma}$ from the Wishart(p,p) distribution, then generated n i.i.d. $\boldsymbol{x}_i$'s from $N_p(\boldsymbol{0},\boldsymbol{\Sigma})$ and normalized $\boldsymbol{x}_i$'s to have mean 0 and variance 1. The matrix $\boldsymbol{X}$ containing $\boldsymbol{x}_i$ as its ith row was treated as fixed in 1,000 simulations. In each run, n i.i.d. ϵ_i's were generated from $N(0,1)$ and y_i was obtained from (1.1) with $\mu = 0$ and $\sigma_i^2 = 0.1$ for each i. In each simulation, we first ran LASSO to calculate its entire path to see if there was a model along the path that matched the true model. We then ran RLASSO with $\hat{\boldsymbol{\Sigma}}$ in (2.4) as an adaptive thresholding estimator of $\boldsymbol{\Sigma}$ and with tuning parameters chosen by (2.6). After 1,000 simulations, we calculated the percentages of times that LASSO and RLASSO selected the correct model. We

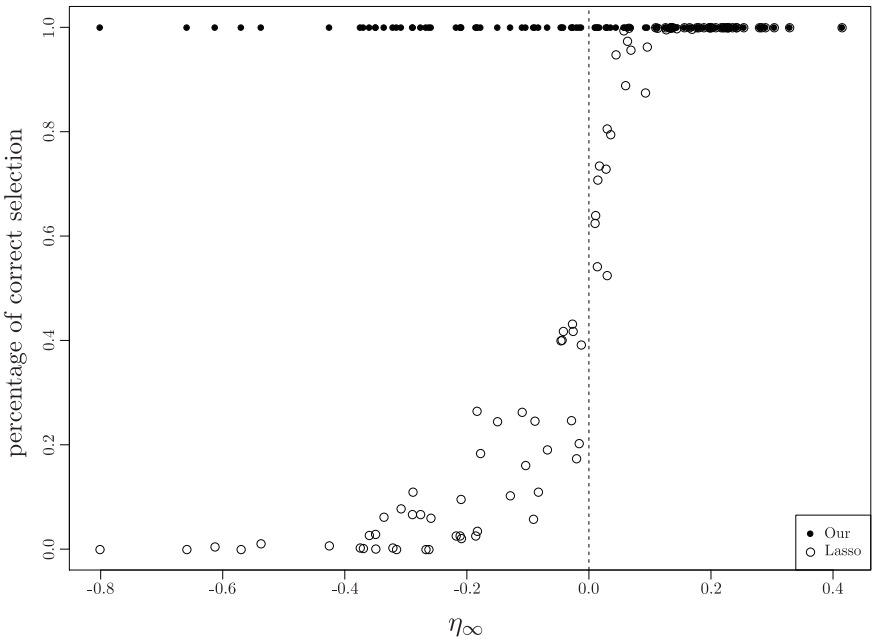

Figure 1. Comparison of LASSO and RLASSO.

repeated independently this process 100 times, and here plot in Figure 1 the 100 simulation percentages against $\eta_\infty = 1 - \|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\text{sign}(\boldsymbol{\beta}_1)\|_\infty$.

Figure 1 shows the performance of LASSO was good when $\eta_\infty > 0.05$; this occurred only 45 times in the 100 designs. In other 55 cases, LASSO did not perform well. While RLASSO performed well in all 100 cases, regardless of whether the SIC held or not.

Meinshausen and Yu (2009) introduced another sparsity condition on $\boldsymbol{S}$, called the incoherent design condition (IDC). If

$$\phi_{\min}(m) = \min_{\boldsymbol{z}:||\boldsymbol{z}||_{l_0}\le\lceil m\rceil} \frac{\boldsymbol{z}'\boldsymbol{S}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} \quad \text{and} \quad \phi_{\max}(m) = \max_{\boldsymbol{z}:||\boldsymbol{z}||_{l_0}\le\lceil m\rceil} \frac{\boldsymbol{z}'\boldsymbol{S}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}}$$

are the m-sparse minimal eigenvalue and m-sparse maximal eigenvalue of $\boldsymbol{S}$, respectively. $\boldsymbol{S}$ is said to satisfy the IDC if there exists a positive sequence e_n such that

$$\liminf_{n\to\infty} \frac{e_n\phi_{\min}(e_n^2 s_0)}{\phi_{\max}(s_0 + \min\{n, p\})} \ge 18, \tag{4.4}$$

where s_0 is the number of non-zero elements in $\boldsymbol{\beta}$. Meinshausen and Yu (2009) showed that if $\boldsymbol{S}$ satisfies the IDC, the true $\boldsymbol{\beta}$ is sparse, and $\boldsymbol{\beta}$'s minimal non-zero component does not converge to 0 too quickly, then LASSO followed by a thresholding could achieve selection-consistency.

If $p \gg n$, the IDC can fail for many $\boldsymbol{S}$. It essentially requires that $\lambda_{\min}(\boldsymbol{S}_0)$ cannot converge to zero too quickly, for any submatrix $\boldsymbol{S}_0$ of $\boldsymbol{S}$ with certain rank. When $p \gg n$, many submatrices of $\boldsymbol{S}$ can be singular or close to singular. Any $m \times m$ submatrix of $\boldsymbol{S}$ is singular if $m > n$, and for $m < n$, it is a hard to check condition.

When $\boldsymbol{\Sigma}$ is sparse, many regularized estimators $\hat{\boldsymbol{\Sigma}}$ in the literature, such as the adaptive thresholding estimator in Cai and Liu (2011), satisfy $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_2 \xrightarrow{P} 0$. If $\lambda_{\min}(\boldsymbol{\Sigma})$ is bounded away from 0, then $\lambda_{\min}(\hat{\boldsymbol{\Sigma}})$ is also asymptotically bounded away from 0. Hence, for any submatrix $\hat{\boldsymbol{\Sigma}}_0$ of $\hat{\boldsymbol{\Sigma}}$, $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_0) > 0$, and the IDC holds for $\hat{\boldsymbol{\Sigma}}$. The same conclusion can be made when $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse, $\lambda_{\max}(\boldsymbol{\Omega})$ is bounded away from ∞, and an L_2-consistent estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ is adopted.

5. Numerical Results

5.1. Simulations

We report on several simulation studies to compare the following eight variable selection methods. RLASSO(AT): RLASSO with $\hat{\boldsymbol{\Sigma}}$ in (2.4) the adaptive thresholding estimator of $\boldsymbol{\Sigma}$. RLASSO(CLIME): RLASSO with $\hat{\boldsymbol{\Sigma}}$ in (2.4) the inverse of CLIME. RLASSO(GLASSO): RLASSO with $\hat{\boldsymbol{\Sigma}}$ in (2.4) the inverse of the Graphical LASSO estimator. LASSO: the ordinary LASSO method. LASSO+T: the ordinary LASSO followed by a thresholding step. Scout(1, 1): the Scout(1, 1) method in Witten and Tibshirani (2009). SLSE+T: the sparse least square estimator in (2.1) followed by a thresholding step; the estimator $\hat{\boldsymbol{\Sigma}}$ is the same as that in RLASSO(AT). SIS: the Sure Independence Screening method in Fan and Lv (2008).

Tuning parameters in all methods were chosen by the BIC described as (2.6). In particular, for SIS, we determined the number of selected variables by the BIC.

We examined the variable selection methods through four models. In each model, $\boldsymbol{y}$ was generated from (1.1) with $\mu = 0$ and $\sigma_i = 1$ for $i = 1, \ldots, n$, $\boldsymbol{x}_i$'s were i.i.d. $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $n = 100$, and $p = 5{,}000$. Parameters in each model were as follows.

Model 1: $\boldsymbol{\beta} = (2\boldsymbol{e}_8, \boldsymbol{0}_{p-8})$, where $\boldsymbol{e}_k$ is the k-dimensional vector with all components equal to 1 and $\boldsymbol{0}_k$ is the k-dimensional vector with all components equal to 0;

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{B}_{10\times 10} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p-10} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_{10\times 10} = \begin{pmatrix} \boldsymbol{I}_8 & 0.2\boldsymbol{e}_{8\times 2} \\ 0.2\boldsymbol{e}_{2\times 8} & \boldsymbol{I}_2 \end{pmatrix},$$

where $\boldsymbol{e}_{m\times n}$ is the $m \times n$ matrix with all elements equal to 1 and $\boldsymbol{I}_k$ is the k-dimensional identity matrix.

Model 2: $\boldsymbol{\beta} = (4, -1.2, 2.5, 1.5, 4.6, \boldsymbol{0}_{p-5})$, and $\boldsymbol{\Sigma} = \text{toeplitz}(1, 0.49, 0.44, 0.40, 0.36, 0.32, 0.29, 0.26, 0.23, \boldsymbol{0}_{p-9})$.

Table 1. Simulation results under Model 1.

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	93.00	99.98	91.00	31.00	8.62(1.72)
RLASSO(CLIME)	87.63	99.91	41.50	0.00	9.15(3.43)
RLASSO(GLASSO)	90.00	99.93	44.50	0.00	9.78(1.96)
Scout(1,1)	98.25	98.79	89.00	0.00	68.31(12.7)
LASSO	98.25	99.59	42.00	0.00	28.24(14.2)
LASSO+T	96.62	99.90	41.00	0.00	12.58(2.99)
SLSE+T	97.62	99.16	84.00	0.00	45.52(11.6)
SIS	81.50	99.09	26.50	0.00	51.92(26.4)

Model 3: $\boldsymbol{\beta} = (4, -1.2, 2.5, 1.5, 4.6, \mathbf{0}_{p-5})$, and $\boldsymbol{\Sigma}$ is the inverse of $\boldsymbol{\Omega} = \text{toeplitz}(1, 0.5, \mathbf{0_{p-2}})$.

Model 4: $\boldsymbol{\beta} = (\mathbf{1}_9, -7.2, \mathbf{0}_{p-10})$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.2\boldsymbol{I}_{10} + 0.8\boldsymbol{e}_{10}\boldsymbol{e}_{10}' & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-10} \end{pmatrix}.$$

In Models 1 and 4, both the covariance matrix $\boldsymbol{\Sigma}$ and its inverse $\boldsymbol{\Omega}$ are sparse. In Model 2, only $\boldsymbol{\Sigma}$ is sparse, while in Model 3, only $\boldsymbol{\Omega}$ is sparse. Model 1 is a setting in which the SIC in (4.1) can hardly hold, since (4.1) fails with $\boldsymbol{S}$ replaced by $\boldsymbol{\Sigma}$. Model 4 is motivated by a "false negative" example in Fan and Lv (2010), where (1.3) is violated; $\beta_{10} = -7.2$ in this model has a marginal effect of 0, but indeed has the largest effect. Condition (1.4) does not hold Models 1-3, but holds under Model 4.

We measure the performance of each variable selection method by the following criteria. Sensitivity (SENS): the proportion of true non-zero β_j being estimated as non-zero. Specificity (SPEC): the proportion of true zero β_j being estimated as zero. Coverage probability (CP): $P(\mathcal{M}_{\boldsymbol{\beta}} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}})$. Hit probability (HP): $P(\mathcal{M}_{\boldsymbol{\beta}} = \widehat{\mathcal{M}}_{\boldsymbol{\beta}})$. Model size (SIZE): the size of selected model.

Note that $1-$SENS and $1-$SPEC are also called false negative error and false positive error, respectively. When HP is low, we should assess the performance of a selection method by jointly considering SENS, SPEC, and CP, not just by using one of them. Thus a method with 100% CP may not be good, since it is too conservative when SPEC is too low.

For each model, we carried out 200 simulation runs. The mean of the performance measures are reported in Tables $1-4$. The simulation standard deviation for SIZE is reported in parenthesis. Additional information regarding the computational time, and estimation accuracy of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ is available in the web supplementary material.

Although RLASSO(AT) performs well in terms of SENS and SPEC, its HP is low so that the asymptotic effect has not shown at $n = 100$ and $p = 5{,}000$.

Table 2. Simulation results under Model 2.

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	90.48	99.95	52.00	45.50	6.56(2.39)
RLASSO(CLIME)	95.24	99.95	76.00	7.50	6.94(1.47)
RLASSO(GLASSO)	89.76	99.97	48.50	7.50	5.85(1.29)
Scout(1,1)	100.00	99.58	100.00	0.00	26.37(3.50)
LASSO	83.33	99.89	16.00	0.00	8.68(4.34)
LASSO+T	82.38	100.00	11.00	11.00	4.18(0.39)
SLSE+T	100.00	99.10	100.00	0.00	40.88(11.1)
SIS	85.00	99.98	27.00	0.00	15.00(1.20)

Table 3. Simulation results under Model 3.

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	91.60	99.95	58.00	11.00	5.06(1.33)
RLASSO(CLIME)	93.40	99.98	68.50	17.00	5.71(0.81)
RLASSO(GLASSO)	99.80	99.90	99.00	15.00	6.00(0.35)
Scout(1,1)	100.00	99.59	100.00	0.00	9.06(1.41)
LASSO	81.80	99.65	9.50	0.00	7.61(1.91)
LASSO+T	81.20	99.92	6.50	3.00	4.90(0.90)
SLSE+T	100.00	99.22	100.00	0.00	12.74(4.26)
SIS	100.00	99.66	100.00	0.00	8.35(1.81)

Table 4. Simulation results under Model 4.

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	91.50	99.91	67.00	66.00	10.09(1.19)
RLASSO(CLIME)	96.20	93.30	62.00	0.00	18.99(2.54)
RLASSO(GLASSO)	93.60	90.94	46.50	0.00	18.08(7.55)
Scout(1,1)	94.30	90.01	53.00	0.00	27.29(7.93)
LASSO	82.80	97.24	42.00	0.00	35.63(10.1)
LASSO+T	56.90	100.00	2.00	2.00	5.72(2.32)
SLSE+T	96.60	90.34	66.50	0.00	19.26(9.83)
SIS	89.90	99.01	0.00	0.00	18.84(6.58)

Since its average SIZE is quite close to the true size (8, 5, 5, 10 for Models 1−4, respectively), the low HP is caused by selecting of 1 or 2 unnecessary variables or missing 1 or 2 important variables. In Tables 1 and 4 where both $\mathbf{\Sigma}$ and its inverse $\mathbf{\Omega}$ are sparse, RLASSO(AT) is, in general, better than RLASSO(CLIME) and RLASSO(GLASSO). The same is true in Table 2, where $\mathbf{\Sigma}$ is sparse but $\mathbf{\Omega}$ is not, a situation that is in favor of RLASSO(AT). In Table 3, $\mathbf{\Omega}$ is sparse but $\mathbf{\Sigma}$ is not, both RLASSO(CLIME) and RLASSO(GLASSO) are better than RLASSO(AT). RLASSO(CLIME) and RLASSO(GLASSO) perform similarly, as they only differ in the estimation of $\mathbf{\Omega}$. Overall, RLASSO(AT) performs better than SLSE+T, indicating that the L_1 penalty in (2.4) is worthwhile. Note that SLSE+T is a special case of RLASSO with $\lambda_n = 0$.

In Tables 2−3, any of the three RLASSO's is better than LASSO. In Table 1, LASSO has a higher SENS but is too conservative in having an average model size 28.24, much larger than the true model size 8. In Table 4, RLASSO(AT) is always better than LASSO; LASSO has a SPEC higher than those of RLASSO(CLIME) and RLASSO(GLASSO), but it has much worse SENS and CP. Thus, the overall performance of LASSO is worse than any RLASSO and sometimes is much worse. LASSO+T in general improves LASSO, since LASSO is too conservative in all tables. However, LASSO+T is still worse than any of RLASSO in general, indicating the importance of regularizing the estimation of $\mathbf{\Sigma}$.

Since the marginal effect $\boldsymbol{\beta}_M$ differs very much from $\boldsymbol{\beta}$ in Tables 1−3, condition (1.4) does not hold, SIS is too conservative. On the other hand, in Table 4 where condition (1.4) holds but condition (1.3) does not hold, SIS has zero CP.

5.2. Data analysis

Sinnaeve et al. (2009) studied the relationship between Coronary Artery Disease (CAD) and gene expression patterns. In their study, each subject's coronary artery disease index (CADi) was measured; this is a validated angiographical measure of the extent of coronary atherosclerosis. Gene expression profiles were obtained by using Affymetrix U133A chips. The raw dataset is available under the name "GSE12288" in Gene Expression Omnibus.

We regressed CADi on the expression of genes that are listed in Kyoto Encyclopedia of Genes and Genomes (KEGG). There were $n = 110$ subjects and $p = 4{,}260$ genes involved in the analysis. To select important genes, we applied RLASSO(AT), RLASSO(CLIME), RLASSO(GLASSO), Scout(1,1), LASSO, LASSO+T, and SIS. Tuning parameters of each method were chosen by the BIC method described in Section 2. The selection of genes by various methods is listed in Table 5.

Gene PECAM1 is selected by all methods except SIS. Stevens et al. (2008) showed that PECAM1 plays a role as a critical mediator of atherosclerosis, which accounts for the vast majority of fatal and non-fatal CAD events. Besides PECAM1, at least one of EPHA4, TPM2, and TSHR is selected by all methods except SIS, and some methods select two or all three of them. Without a second thresholding step, Scout(1,1) and LASSO select too many genes. LASSO+T is much more reasonable. The selection by SIS is very different from the others, which may be caused by the fact that it ignores the correlations among genes.

Overall, the analysis shows the importance of PECAM1 for this data set, followed by a few more genes such as EPHA4, TPM2 and TSHR.

Table 5. Genes Selections for GSE12288.

Method	Selection of Genes
RLASSO (AT)	GABBR1 PECAM1 PRKAB2 PTCRA TPM2 TSHR UBE3A
RLASSO (CLIME)	EPHA4 FGF14 PECAM1
RLASSO (GLASSO)	EPHA4 PECAM1 TPM2 TSHR
Scout(1,1)	ABCC6 ACSL5 ADRB1 AKT3 APOA4 ATP6V0A4 ATR CA6 CBLB CBR3 CCNA1 CD14 CLDN14 COX4I1 COX6A1 CSF2 CYBA DAD1 DSE EPHA4 FGF14 GABBR1 GLS2 GNG12 GNG13 H2AFY2 HSPBP1 LRP2 MC2R MLNR MYO10 NFAT5 NOS1 OAS1 OR2J2 PAPSS1 PDSS2 PECAM1 PIGB PNLIPRP2 PPP2R5A PRKAB2 PTCRA RARS2 RRAS SERPINB4 SPRY2 STX2 SUCLA2 THY1 TPM2 TRAF3 TSHR UBE3A WNT10B
LASSO	ABCC6 ACSL5 ADRB1 AKT3 APOA4 ATP6V0A4 ATR CA6 CBLB CBR3 CCNA1 CD14 COX4I1 COX6A1 CSF2 CYBA DAD1 DSE EPHA4 FGF14 GABBR1 GLS2 GNG13 H2AFY2 HSPBP1 LRP2 MC2R MLNR MYO10 NFAT5 NOS1 OAS1 OR2J2 PDSS2 PECAM1 PIGB PNLIPRP2 PPP2R5A PRKAB2 PTCRA RARS2 SERPINB4 SPRY2 STX2 SUCLA2 THY1 TPM2 TRAF3 TSHR UBE3A WNT10B
LASSO + T	CSF2 EPHA4 MC2R PECAM1 TPM2 TSHR
SIS	A2M ABP1 C1QB C3 C5 C6 CD55 CFB CFH CPB2 CYP2A6 CYP3A7 DPYS F13A1 F2R F7 F8 F9 FGB FGG IMPDH2 KLKB1 MASP2 PLAU SERPINA5 TFPI THBD UCKL1 UGT2B15 UGT2B28 UPB1 XDH

6. Discussion

We propose a regularized LASSO that replaces ill-behaved $\boldsymbol{X}'\boldsymbol{X}/n$ with a sparse estimator of $\boldsymbol{\Sigma}$ or its inverse and adds a thresholding step to handle variable selection with p much larger than n. Theoretical and empirical properties of the proposed method, RLASSO, are discussed. In applications, we have to decide which of $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Omega}$ is sparse and choose one version of RLASSO accordingly. In a situation where no information regarding the sparsity of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ is available, we recommend RLASSO(AT) based on its empirical property. We may also try three different RLASSO methods as we did in the data example. The computation of RLASSO with CLIME $\hat{\boldsymbol{\Omega}}$ is lengthy when p is as large as 5,000. Some improvement may be developed in our future research.

The idea of replacing ill-behaved $\boldsymbol{X}'\boldsymbol{X}/n$ by a sparse estimator of $\boldsymbol{\Sigma}$, or its inverse, can be applied to variable selection in more complicated models. For example, we are investigating how to use this approach to variable selection in linear mixed-effect models and some nonlinear or nonparametric models.

As with almost all asymptotic methods, the proposed RLASSO requires the signal from non-zero components of $\boldsymbol{\beta}$ to be large enough to have the good properties predicted by the asymptotic theory. One strategy is to carry out simulation studies to check finite sample performance in a setting close to the actual situation. There are some procedures that work well in the low signal cases; see, for example, Ji and Jin (2012).

Acknowledgement

The authors thank two referees and an associate editor for their comments and suggestions. The research was partially supported by the NSF Grant DMS-1007454.

References

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672-684.

Cai, T., Liu, W. and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106**, 594–607.

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, 2118–2144.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40**, 73-103.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246-270.

Sinnaeve, P. R., Donahue, M. P., Grass, P., Seo, D., Vonderscher, J., Chibout, S. D., Kraus, W. E., Sketch, M., Nelson, C., Ginsburg, G. S., Goldschmidt-Clermont, P. J. and Granger, C. B. (2009). Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS ONE* **4**, e7037.

Stevens, H. Y., Melchior, B., Bell, K. S., Yun, S., Yeh, J.-C. and Frangos, J. A. (2008). Pecam-1 is a critical mediator of atherosclerosis. *Disease Models & Mechanisms* **1**, 175-181.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

Witten, D. M. and Tibshirani, R. (2009). Covariance–regularized regression and classification for high dimensional problems. *J. Roy. Statist. Soc. Ser. B* **71**, 615–636.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.
E-mail: quefeng@stat.wisc.edu

School of Finance and Statistics, East China Normal University, Shanghai, 200241, China.
Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.
E-mail: shao@stat.wisc.edu

(Received January 2013; accepted May 2014)