Chin. Ann. Math. Ser. B
39(4), 2018, 607–620
DOI: 10.1007/s11401-018-0084-6

Chinese Annals of Mathematics, Series B

Model Selection Consistency of Lasso for Empirical Data*

Yuehan YANG[1] Hu YANG[2]

Abstract Large-scale empirical data, the sample size and the dimension are high, often exhibit various characteristics. For example, the noise term follows unknown distributions or the model is very sparse that the number of critical variables is fixed while dimensionality grows with n. The authors consider the model selection problem of lasso for this kind of data. The authors investigate both theoretical guarantees and simulations, and show that the lasso is robust for various kinds of data.

Keywords Lasso, Model selection, Empirical data
2000 MR Subject Classification 62J05, 62J07

1 Introduction

Tibshirani [14] proposed the lasso (least absolute shrinkage and selection operator) method for simultaneous model selection and estimation of the regression parameters. It is very popular for high-dimensional estimation due to its statistical accuracy for prediction and model selection coupled with its computational feasibility. On the other hand, under some sufficient condition the lasso solution is unique, and the number of non-zero elements of lasso solution is always smaller than n (see [15–16]). In recent years, this kind of data has become more and more common in most fields. Similar properties can also be seen in other penalized least squares since they have a similar framework of solution.

Consider the problem of model selection in the sparse linear regression model

$$y_n = X_n\beta_n + \epsilon_n,$$

where the detail setting of the data can be found in the next section. Then the lasso estimator is defined as

$$\widehat{\beta}_n(\lambda_n) \in \operatorname*{argmin}_{\beta\in\mathbb{R}^{p_n}} \Big\{\frac{1}{2}\|y_n - X_n\beta\|_2^2 + \lambda_n\|\beta\|\Big\},$$

where λ_n is the tuning parameter which controls the amount of regularization. Set $\widehat{S}_n \equiv \{j \in \{1, 2, \cdots, p_n\} : \widehat{\beta}_{j,n} \neq 0\}$ to select predictors by lasso estimator $\widehat{\beta}_n$. Consequently, $\widehat{S}_n$ and $\widehat{\beta}_n$

Manuscript received November 8, 2016. Revised February 24, 2017.

[1] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China. E-mail: yyh@cufe.edu.cn
[2] College of Mathematics and Statistics, Chongqing University, Chongqing 400044, China. E-mail: yh@cqu.edu.cn

*This work was supported by the National Natural Science Foundation of China (No. 11671059).

both depend on λ_n, and the model selection criteria results in the correct recovery of the set $S_n \equiv \{j \in \{1, 2, \cdots, p_n\} : \beta_{j,n} \neq 0\}$:

$$P(\widehat{S}_n = S_n) \to 1, \quad \text{as } n \to \infty.$$

On the model selection front of the lasso estimator, Zhao and Yu [22] established the irrepresentable condition on the generating covariance matrices for the lasso's model selection consistency. This condition was also discovered in [11, 20, 23]. Using the language of [22], irrepresentable condition is defined as $|C_{21}C_{11}^{-1}\mathrm{sign}(\beta_{(1)})| \leqslant \mathbf{1} - \eta$, where sign$(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero. The definitions of C_{21} and C_{11} can be seen in Section 2. When signs of the true coefficients are unknown, they need l_1 norms of the regression coefficients to be smaller than 1. Beyond lasso, regularization methods also have been widely used for high-dimensional model selection, e.g., [2, 4, 7–8, 10, 12, 17–19, 21, 24–25]. There has been a considerable amount of recent work dedicated to the lasso problem and regularization methods problem.

Yet, the study of model selection problem for empirical data is still needed. Stock data for instance, the Gaussian assumption of the noise term is always unsatisfied for these data. And the critical variables are extremely few contrast to the collected dimensionality. In this paper, we consider this kind of data: The sample size and the dimension are high, but the information of critical variable data is missing (the signs of the true β_n and the distribution of the noise terms are unknown) and the model is extremely sparse that the number of nonzero parameter is fixed. This kind of data is common in the empirical analysis hence we called it empirical data.

We consider the model selection consistency of lasso and investigate regular conditions to fit this data setting. Under conditions, the probability for lasso to select the true model is covered by the probability of

$$\{\|W_n\|_\infty \leqslant G_n\},$$

where $W_n = \frac{X_n' \epsilon_n}{\sqrt{n}}$ and G_n is a function of λ_n, n, q. Above inequality is simple and also easy to calculate its probability. Based on the train of thought of the proof, we analyze the model selection consistency of lasso under easier conditions than the irrepresentable condition for empirical data. In the simulation part, we discuss the effectiveness of lasso. Four samples are given, in which the irrepresentable condition fails for all the settings, but lasso still can select variables correctly in two of them when our conditions hold.

We discuss the different assumptions of noise terms $\epsilon_{i,n}$ for model selection consistency. Gaussian errors or the subgaussian errors[1] would be standard, but possess a strong tail. One basic assumption in this paper is that, errors are assumed to be identically and independently distributed with zero mean and finite variance.

The rest of the paper is organized as follows. In Section 2, we investigate the data setting, notations, and conditions. We introduce a lower bound to cover the case in which the lasso chooses wrong models when suitable conditions hold. Then, to demonstrate the advantages of this bound, we show the different settings and different assumptions of noise terms in Section 3.

[1] e.g. $P(|\epsilon_{i,n}| \geqslant t) \leqslant C\mathrm{e}^{-ct^2}$, $\forall t \geqslant 0$.

We show that the lasso has model selection consistency for empirical data with mild conditions. Section 4 presents the results of the simulation studies. Finally, in Section 5, we present the proof of the main theorem.

2 Data Setting, Notations and Conditions

Consider the problem of model selection for specific data

$$y_n = X_n\beta_n + \epsilon_n,$$

where $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n}, \cdots, \epsilon_{n,n})'$ is a vector of i.i.d. random variables with mean 0 and variance σ^2, X_n is an $n \times p_n$ design matrix of predictor variables, $\beta_n \in \mathbb{R}^{p_n}$ is a vector of true regression coefficients and is commonly imposed to be sparse with only a small proportion of nonzeros. Without loss of generality, write $\beta_n = (\beta_{1,n}, \cdots, \beta_{q,n}, \beta_{q+1,n}, \cdots, \beta_{p,n})'$ where $\beta_{j,n} \neq 0$ for $j = 1, \cdots, q$ and $\beta_{j,n} = 0$ for $j = q+1, \cdots, p_n$. Then write $\beta_n^{(1)} = (\beta_{1,n}, \cdots, \beta_{q,n})'$ and $\beta_n^{(2)} = (\beta_{q+1,n}, \cdots, \beta_{p,n})$, that is, only the first q entries are nonvanishing. Besides, for any vector $\alpha = (\alpha_1, \cdots, \alpha_m)'$, we denote $\|\alpha\| = \sum\limits_{i=1}^{m} |\alpha_i|$, $\|\alpha\|_2^2 = \sum\limits_{i=1}^{m} \alpha_i^2$, and $\|\alpha\|_\infty = \max\limits_{i=1,\cdots,m} |\alpha_i|$.

For deriving the theoretical results, we write $X_n(1)$ and $X_n(2)$ as the first q and the last $p_n - q$ columns of X_n, respectively. Let $C_n = \frac{1}{n}X_n'X_n$. Partition C_n as

$$C_n = \begin{pmatrix} C_{11,n} & C_{12,n} \\ C_{21,n} & C_{22,n} \end{pmatrix},$$

where $C_{11,n}$ is $q \times q$ matrix and assumed to be invertible. Set $W_n = \frac{X_n'\epsilon_n}{\sqrt{n}}$. Similarly, $W_n^{(1)}$ and $W_n^{(2)}$ indicate the first q and the last $p_n - q$ elements of W_n. Suppose that $\Lambda_{\min}(C_{11,n}) > 0$ denotes the smallest eigenvalue of $C_{11,n}$ and consider that q does not grow with n. We introduce the following conditions:

(C1) For $j = q+1, \cdots, p_n$, let e_j be the unit vector in the direction of j-th coordinate. There exists a positive constant $0 < \eta < 1$ such that

$$\|e_j'C_{21}\|_2 \leqslant 1 - \eta.$$

(C2) There exists $\delta \in (0,1)$, such that for all $n > \delta^{-1}$ and $x \in \mathbb{R}^q$, $y \in \mathbb{R}^{p_n - q}$,

$$(x'C_{12,n}y)^2 \leqslant \delta^2(x'C_{11,n}x) \cdot (y'C_{22,n}y).$$

(C1) and (C2) play a central role in our theoretical analysis. Both conditions are easy to satisfy. (C1) for instance, it requires an upper bound on l_2-norm, which is much weaker than requires the upper bound on l_1-norm, i.e., irrepresentable condition and variants of this condition [6, 9, 11, 22, 25]. Another advantage of (C1) is that we do not need the signs of the true coefficients. (C2) requires that the multiple correlations between relevant variables and the irrelevant variables is strictly less than one. It is weaker than assuming orthogonality of the two sets of variables. This condition also has regular appeared many times in the literature, for example, [13].

Then we have the following theorem, which describes the relationship between the probability of lasso choosing the true model and the probability of $\{\|W_n\|_\infty \leqslant G_n\}$. Videlicet, it is a lower bound on the probability of lasso picking the true model.

Theorem 2.1 *Assume that* (C1)–(C2) *hold and* $\Lambda_{\min} \triangleq \Lambda_{\min}(C_{11,n})$. *Set* $\rho \in (0,1)$. *We have*

$$P(\widehat{S}_n = S_n) \geqslant P(\|W_n\|_\infty \leqslant G_n),$$

where $G_n = \min\left\{\sqrt{n}\Lambda_{\min}\left(\min\limits_{j=1,\cdots,q}|\beta_{j,n}| - \frac{\lambda_n\sqrt{q}}{\Lambda_{\min}\cdot n}\right),\ \frac{\lambda_n\rho}{\sqrt{n}}\right\}$.

Remark 2.1 Theorem 2.1 is a key technical tool in the theoretical results. It puts a lower bound on the probability of lasso selecting the true model, and this bound is intuitive to calculate. Besides that, considering about G_n, it is easy to find out that there exists a lower bound of non-zero coefficients $\min\limits_{j=1,\cdots,q}|\beta_{j,n}| > \frac{\lambda_n\sqrt{q}}{\Lambda_{\min}\cdot n}$. This bound can be controlled by the regularization parameter λ_n. It is also a regular assumption in the literature that the non-zero coefficients cannot be too small.

Remark 2.2 According to the proof of Theorem 2.1, we can find that it is also directly to obtain the sign consistency of the lasso (see the latter part of the proof). Besides, Theorem 2.1 can be applied in a wide range of dimensional setting. We will discuss the behavior of the lasso on model selection consistency under different settings in the next section.

3 Model Selection Consistency

Now we consider the decay rate of the probability of $\{\|W_n\|_\infty > G_n\}$. Different dimensions and different assumptions of noise terms are discussed in this section.

First, we consider general dimensional setting, i.e., $p_n = O(n^{c_1})$ where $0 < c_1 < 1$. Under this setting, we can obtain the model selection consistency of lasso by no constraint for the noise terms. Then, we consider ultra-high dimensional setting, i.e., $p_n = O(\mathrm{e}^{n^{c_2}})$ where $0 < c_2 < 1$. Under this setting, we need an assumption of ϵ_n to make the model selection of lasso. Gaussian assumption would be a simple and common one, but result in a strong tail. We prefer the more standard assumption that only i.i.d random variables of the noise terms.

Before discussing the detail rate of the probability of lower bound, we give the following regular condition:

(C3) $n^{-1}X'_{j,n}X_{j,n} \leqslant 1$ for $j = 1, \cdots, p_n$.

It is a typical assumption in sparse linear regression literature. It can be achieved by normalizing the covariates (see [9, 22]).

3.1 General dimensional setting $p_n = O(n^{c_1})$

In this part, we consider the general dimensional setting where p_n is allowed to grow with n and show the model selection consistency of lasso as follows.

Theorem 3.1 *Assume that* ϵ_i *are i.i.d random variables with mean* 0 *and variance* σ^2. *Suppose that* (C1)–(C3) *hold. For* $p_n = O(n^{c_1})$ *where* $0 < c_1 < 1$, *if* $\frac{\lambda_n}{\sqrt{n}} \propto n^{\frac{c_3}{2}}$ *where* $c_1 < c_3 < 1$ *and* $\min\limits_{j=1,\cdots,q}|\beta_{j,n}| > n^{\frac{c_3-1}{2}}$, *then we have*

$$P(\widehat{S}_n = S_n) \geqslant 1 - n^{c_1 - c_3} \to 1, \quad \textit{as } n \to \infty.$$

Proof Following the result in Theorem 2.1, we have

$$P(\widehat{S}_n = S_n) \geqslant P(\|W_n\|_\infty \leqslant G_n),$$

where

$$G_n = \min\Big\{\sqrt{n}\Lambda_{\min}\Big(\min_{j=1,\cdots,q}|\beta_{j,n}| - \frac{\lambda_n\sqrt{q}}{\Lambda_{\min}\cdot n}\Big),\ \frac{\lambda_n\rho}{\sqrt{n}}\Big\}.$$

Applying the setting of Theorem 3.1, hence for $n \to \infty$,

$$\Lambda_{\min}^{-1}\cdot\lambda_n\frac{\sqrt{q}}{n} = O(n^{\frac{c_3-1}{2}}) \to 0.$$

Then there exists a positive constant K_n that

$$G_n = \frac{\rho\lambda_n}{\sqrt{n}} = K_n n^{\frac{c_3}{2}}.$$

If (C3) holds, by Markov's inequality, we easily get

$$\begin{aligned} P(\|W_n\|_\infty > G_n) &\leqslant \sum_{j=1}^{p_n} P(|W_{j,n}| > G_n) \\ &= \sum_{j=1}^{p_n} P\Big(\Big|\frac{X'_{j,n}\epsilon}{\sqrt{n}}\Big| > K_n n^{\frac{c_3}{2}}\Big) \\ &\leqslant K_n^{-2} n^{-c_3}\cdot n^{c_1} \to 0, \quad \text{as } n \to \infty. \end{aligned}$$

The proof is completed.

The proof of Theorem 3.1 states that in this setting, lasso is robust and selects the true model with regular restrains. Similarly, if we consider the classical setting where p, q and β are fixed when $n \to \infty$, then we have the following result.

Corollary 3.1 *For fixed p, q and β, under regularity assumptions* (C1)–(C3)*, assume that ϵ_i are i.i.d random variables with mean* 0 *and variance σ^2. If λ_n satisfies that $\frac{\lambda_n}{\sqrt{n}} \to \infty$ and $\frac{\lambda_n}{n} \to 0$ when $n \to \infty$, then*

$$P(\widehat{S} = S) \to 1, \quad \textit{as } n \to \infty.$$

Similar with the argument of Theorem 3.1, Corollary 3.1 can be proved directly by Markov's inequality, hence the proof is omitted here.

Besides, if we assume that the noise term follows the Gaussian assumption, under the same setting of Theorem 3.1, then we have

$$\begin{aligned} P(\widehat{S}_n \neq S_n) &\leqslant P(\|W_n\|_\infty > G_n) \\ &\leqslant \sum_{j=1}^{p_n} P(|W_{j,n}| > K_n n^{\frac{c_3}{2}}) \\ &< n^{c_1-\frac{c_3}{2}}\mathrm{e}^{-\frac{1}{2}n^{c_3}} \to 0, \quad \text{as } n \to \infty, \end{aligned} \tag{3.1}$$

where the last inequality holds because of the Gaussian distribution's tail probability bound: $P(|\epsilon_i| \geqslant t) < t^{-1}\mathrm{e}^{-\frac{1}{2}t^2}$, $\forall t \geqslant 0$. It can be relaxed to subgaussian assumption, i.e., $P(|\epsilon_i| \geqslant t) \leqslant C\mathrm{e}^{-ct^2}$, $\forall t \geqslant 0$.

3.2 Ultra-high dimensional setting $p_n = O(\mathrm{e}^{n^{c_2}})$

In this part, we consider the ultra-high dimensional setting as $p_n = O(\mathrm{e}^{n^{c_2}})$ where $0 < c_2 < 1$ and discuss the different situation under different assumptions of noise terms (Gaussian assumption and non-Gaussian assumption). Theorem 3.2 shows the result under non-Gaussian assumption by applying Bernstein's inequality. Also, we show the model selection consistency of the lasso under the Gaussian assumption in Corollary 3.2.

We shall make use of the following condition:

(C4) Assume that $\epsilon_{1,n}, \cdots, \epsilon_{n,n}$ are independent random variables with mean 0 and the following inequality satisfies for $j = 1, \cdots, p_n$,

$$\frac{1}{\sqrt{n}} E|W_{j,n}|^m \leqslant \frac{m!}{2} L^{m-2}, \quad m = 2, 3, \cdots,$$

where $W_{j,n} = \frac{1}{\sqrt{n}} X'_{j,n}\epsilon_n$ and $L \in (0, \infty)$.

(C4) is the precondition for the non-Gaussian assumption (The model selection consistency of the lasso under the Gaussian assumption does not need this condition). It is applied here for the Bernstein's inequality. According to (C4), we have

$$E \exp\Big[\frac{X'_{j,n} \cdot \epsilon_n}{L_0}\Big] \leqslant \exp\Big[\frac{n}{2(L_0^2 - L \cdot L_0)}\Big],$$

where $L_0 > L$. This bound leads to Bernstein's inequality as given in [1]. Then we have the following result.

Theorem 3.2 *Assume that ϵ_i are i.i.d random variables with mean 0 and variance σ^2. Suppose that (C1)–(C2) and (C4) hold. If $\frac{\lambda_n}{\sqrt{n}} \propto n^{\frac{c_3}{2}}$ where $c_2 < c_3 < 1$ and $\min\limits_{j=1,\cdots,q} |\beta_{j,n}| > n^{\frac{c_3-1}{2}}$. We have*

$$P(\widehat{S}_n = S_n) \geqslant 1 - \mathrm{e}^{-n^{c_2}} \to 1, \quad \textit{as } n \to \infty.$$

Proof By Bernstein's inequality, let $t > 0$ be arbitrary, we have

$$P(W_{j,n} \geqslant n^{\frac{c_3}{2}}(Lt + \sqrt{2t})) \leqslant \mathrm{e}^{-tn^{c_3}} \leqslant \mathrm{e}^{-n^{c_2}}.$$

Applying the result of Lemma 14.13 from [3], when (C3) holds, we have

$$P\Big(\max_{1 \leqslant j \leqslant p_n} |W_{j,n}| \geqslant n^{\frac{c_3}{2}}(Lt + \sqrt{2t} + \alpha(L, n, p_n))\Big) \leqslant \mathrm{e}^{-n^{c_2}}.$$

Following the setting of p,

$$\alpha(L, n, p_n) = \sqrt{\frac{2\log 2p_n}{n}} + \frac{L\log(2p_n)}{n} \to 0.$$

Let $J \in (0, \infty)$ to make the following inequalities hold for all $t > 0$,

$$\alpha(L, n, p_n) < J, \quad Lt + \sqrt{2t} \leqslant Jt.$$

Then we have

$$P(\|W_n\|_\infty > J(1+t)n^{\frac{c_3}{2}}) \leqslant \mathrm{e}^{-n^{c_2}},$$

which completes the proof.

Similarly as in general high-dimensional setting, we have the following result under Gaussian assumption. Since the proof of Corollary 3.2 is direct, we just state the result here without proof.

Corollary 3.2 *Assume that* ϵ_i *are i.i.d Gaussian random variables. Let* $p_n = O(\mathrm{e}^{n^{c_2}})$ *where* $0 < c_2 < 1$. *Suppose that* (C1)–(C3) *hold. If* $\frac{\lambda_n}{\sqrt{n}} = O(n^{\frac{c_3}{2}})$ *where* $c_2 < c_3 < 1$ *and* $\min\limits_{j=1,\cdots,q} |\beta_{j,n}| > n^{\frac{c_3-1}{2}}$. *The Gaussian assumption of noise terms is considered in the following*

$$\begin{aligned} P(\widehat{S}_n \neq S_n) &\leqslant P(\|W_n\|_\infty > G_n) \\ &\leqslant \sum_{j=1}^{p_n} P(|W_{j,n}| > G_n) \\ &= O(n^{-\frac{c_3}{2}} \mathrm{e}^{n^{c_2} - \frac{1}{2} n^{c_3}}) = o(\mathrm{e}^{-n^{c_2}}) \to 0, \quad \textit{as } n \to \infty. \end{aligned}$$

4 Simulation Part

In this section, we evaluate the finite sample property of lasso estimator with synthetic data. We start with the behavior of lasso under different settings, then consider the relationship between n, p, q and then consider the different noise terms.

4.1 Model selection

This first part illustrates two simple cases (low dimension vs high dimension) to show the efficiency of lasso. Following cases describe two different settings to lead the lasso's model selection consistency and inconsistency when (C1) and (C2) hold and fail. As a contrast, we introduce the irrepresentable condition in this part, and it fails in all the settings.

Example 4.1 In the low dimensional case, assume that there are $n = 100$ observations and the values of parameters are chosen as $p = 3$, $q = 2$, that is,

$$\beta = \{2, 3, 0\}.$$

We generate the response y by

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \epsilon,$$

where X_1, X_2 and ϵ are i.i.d random variables from Gaussian distribution with mean 0 and variance 1. The third predictor X_3 is generated to be correlated with other parameters as the following two cases:

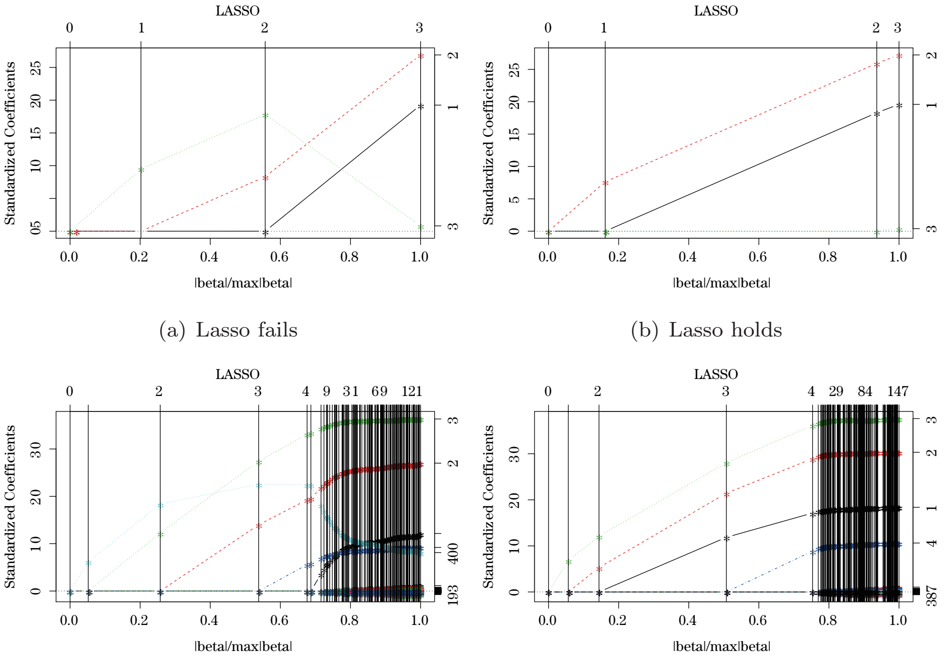
$$X_3 = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}e$$

and

$$X_3 = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{\sqrt{2}}e,$$

where e is i.i.d random variable with the same setting as ϵ.

We can find that the lasso fails for the first case when (C1) and (C2) fail, and selects the right model for the second case when (C1) and (C2) hold. The different solutions are illustrated by Figure 1. Since the irrepresentable condition fails in both cases, it shows that the lasso suits more kinds of data even if the irrepresentable condition is relaxed.



(a) Lasso fails

(b) Lasso holds

(c) Lasso fails

(d) Lasso holds

Figure 1 An example to illustrate the efficiency of lasso's (in)consistency in model selection. The above two graphs are constructed in a low dimensional setting. The below graphs are constructed in a high dimensional setting. The left graphs are set where (C1) and (C2) fail, and the right graphs are set where (C1) and (C2) hold.

Example 4.2 We construct a high dimensional case with $p = 400$, $q = 4$ and $n = 100$. The true parameters are set as

$$\beta = \{2, 3, 1, 4, 0, 0, \cdots, 0\}$$

and the response y is generated by

$$y = X\beta + \epsilon,$$

where

$$X = (X_1, \cdots, X_p)$$

is 100×400 matrix, and the elements of X are i.i.d random variables from Gaussian distribution with mean 0 and variance 1 except X_{400}. The last predictor X_{400} is generated in the following two settings respectively,

$$X_{400} = \frac{7}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{8}X_3 + \frac{1}{8}X_4 + \frac{1}{8}X_5 + \frac{1}{8}X_6 + \frac{1}{8}X_7 + \frac{1}{8}e$$

and

$$X_{400} = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 + \frac{1}{4}X_4 + \frac{1}{4}X_5 + \frac{1}{4}X_6 + \frac{1}{4}X_7 + \frac{3}{4}e,$$

where e follows the same setting as Example 4.1. Hence X_{400} is also constructed from Gaussian distribution with mean 0 and variance 1. We find that our conditions also fail for the first high dimensional case but hold for the second. Besides that, irrepresentable condition fails for both two situations.

We get different lasso solutions for above four cases in Figure 1 (the lasso path is got by lars algorithm in [5]). As shown in Figure 1, both graphs on the left satisfy neither irrepresentable condition nor (C1)–(C2), and lasso cannot select variables correctly (both graphs select other irrelevant variables, e.g., X_4 in the first graph and X_{400} in the second). In contrast, both graphs on the right select the right model in the settings that (C1)–(C2) hold and irrepresentable condition fails.

Besides that, the above examples are all constructed based on the synthetic data, in which the unknown parameter is actually known. In the empirical analysis, the true model cannot be known in advance. We should recognize a situation in which lasso can be used without precondition.

4.2 Relationship between p, q and n

In this part, we give a direct view to show the relationship between n, p and q, or to say, how the sparsity and the sample size affect the model selection of lasso.

The nonzero elements $\beta^{(1)}$ are set as

$$\{9, 6, 8, 12, 19, 8, 19, 9, 6, 8, 12, 19, 8, 19\}.$$

If the number of nonzero elements is less than 14, we select the number in sequence. The rest of the other elements in this gather are shrunk to zero. The number of observations and the parameters are chosen as Table 1. The predictors are made from Gaussian random generation. Among this table, lasso selects the right variables in the first six items in the list and selects the wrong variables in the remaining items in the list.

Table 1 Example settings

Example	n	p	q	Example	n	p	q
1	100	400	4	7	100	400	5
2	100	500	5	8	100	500	6
3	200	500	7	9	100	500	7
4	200	1000	7	10	100	1000	7
5	500	500	14	11	100	2000	7
6	500	2000	14	12	300	2000	14

The high dimensional settings are considered. The results indicate that q is always required to be small enough for the efficiency of the lasso. When the number of critical factors increases, the sample size needs to be increased too to make sure the lasso chooses the right model. In contrast, the number of zero elements has less influence on the lasso's (in)consistency in model selection.

4.3 Different noise terms

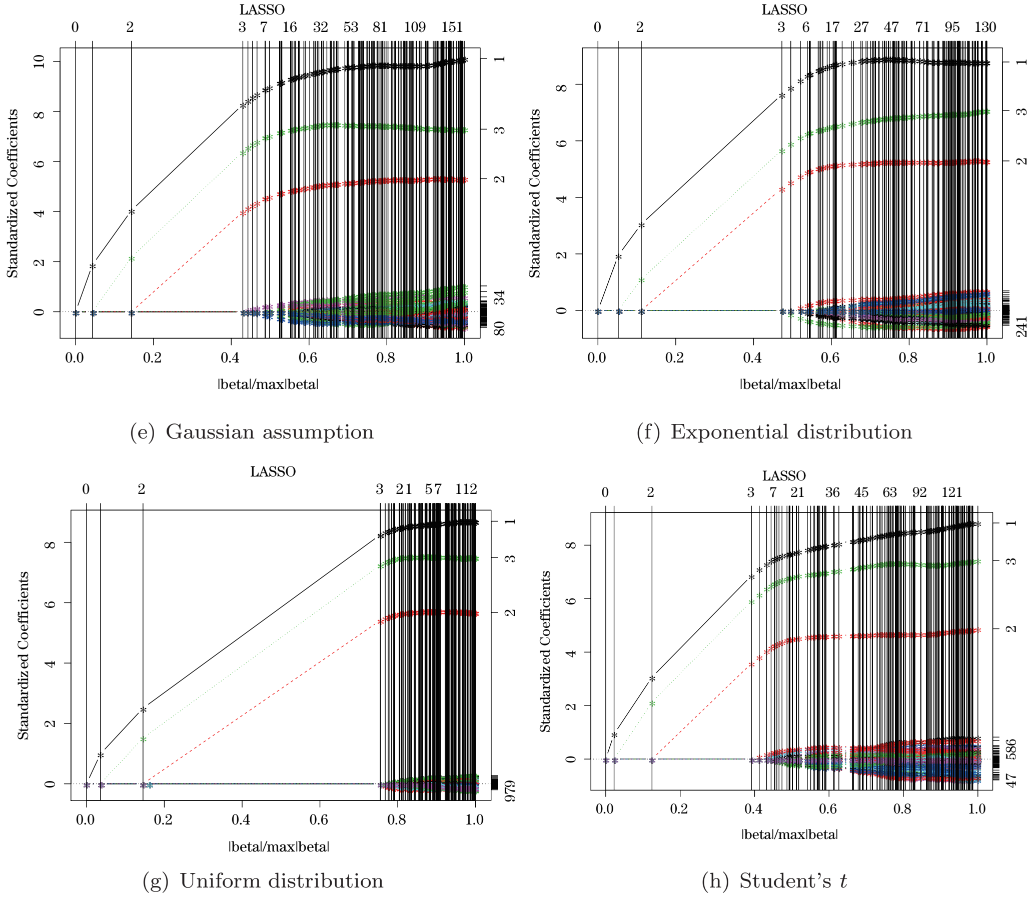
In this part, we consider a high dimensional example with different noise terms. Data from the high-dimensional linear regression model is set as

$$y_i = X_i'\beta + \epsilon_i, \quad i = 1, \cdots, n,$$

where the data have $n = 100$ observations and the value of parameter is chosen as $p = 1000$. The true regression coefficient vector is fixed as

$$\beta = \{9, 6, 8, 0, \cdots, 0\}.$$

For the distribution of the noise ϵ, we consider four distributions: Gaussian assumption with mean 0 and variance 1; exponential distribution with rate 1; uniform distribution with minimum 0 and maximum 1; student's t with degrees of freedom 100.



(e) Gaussian assumption

(f) Exponential distribution

(g) Uniform distribution

(h) Student's t

Figure 2 An example to illustrate the lasso's behavior in the high dimensional setting with different assumptions of noise terms. It reflects that in a situation with standard data and strong sparsity, lasso always chooses the right model no matter the distribution of noise terms.

The results are depicted in Figure 2. It reflects that in a situation with standard data and strong sparsity, lasso always chooses the right model no matter the distribution of noise terms.

5 Proof of Theorem 2.1

Review the lasso estimator

$$\widehat{\beta}_n(\lambda_n) \in \underset{\beta\in\mathbb{R}^{p_n}}{\mathrm{argmin}}\Big\{\frac{1}{2}\|y_n - X_n\beta\|_2^2 + \lambda_n\|\beta\|\Big\}.$$

Let $\widehat{u}_n = \sqrt{n}(\widehat{\beta}_n - \beta_n)$ and

$$F_n(\beta_n) = \frac{1}{2}\|y_n - X_n\beta_n\|_2^2 + \lambda_n\|\beta_n\|.$$

Defining $V_n(\widehat{u}_n) = F_n(\widehat{\beta}_n) - F_n(\beta_n)$, $C_n = \dfrac{1}{n}X_n'X_n$ and $W_n = \frac{1}{\sqrt{n}}X_n'\epsilon_n$, $V_n(\widehat{u}_n)$ can be written as

$$V_n(\widehat{u}_n) = \frac{1}{2}\widehat{u}_n'C_n\widehat{u}_n - \widehat{u}_n'W_n + \lambda_n\Big(\|\beta_n + \frac{\widehat{u}_n}{\sqrt{n}}\| - \|\beta_n\|\Big).$$

Let $\widehat{\beta}_n^{(1)}$, $\widehat{\beta}_n^{(2)}$ and $W_n^{(1)}$, $W_n^{(2)}$ be the first q and last $p_n - q$ elements of $\widehat{\beta}_n$ and W_n, respectively. Similarly, $\widehat{u}_n^{(1)}$ and $\widehat{u}_n^{(2)}$ denote the first q and last $p_n - q$ elements of $\widehat{u}_n$.

Due to $\big\{\|W_n\|_\infty \leqslant \rho\frac{\lambda_n}{\sqrt{n}}\big\}$, by (C2), we have

$$\begin{aligned} V_n(\widehat{u}_n) \geqslant{}& \frac{1-\delta}{2}[\widehat{u}_n'^{(1)}C_{11,n}\widehat{u}_n^{(1)} + \widehat{u}_n'^{(2)}C_{22,n}\widehat{u}_n^{(2)}] - \widehat{u}_n'W_n \\ & - \frac{\lambda_n}{\sqrt{n}}\|\widehat{u}_n^{(1)}\| + \sum_{j=q+1}^{p_n}|\widehat{u}_{j,n}|\Big(\frac{\lambda_n}{\sqrt{n}} - |W_{j,n}|\Big). \end{aligned}$$

Since $\widehat{u}_n'^{(2)}C_{22,n}\widehat{u}_n^{(2)} \geqslant 0$ and $\Lambda_{\min}(C_{11,n})$ denotes the smallest eigenvalue of $C_{11,n}$, we have

$$\begin{aligned} V_n(\widehat{u}_n) \geqslant{}& \|\widehat{u}_n'^{(1)}\|_2\Big\{\frac{1-\delta}{2}\Lambda_{\min}(C_{11,n})\|\widehat{u}_n^{(1)}\|_2 - \|W_n^{(1)}\|_2 - \frac{\lambda_n}{\sqrt{n}}\sqrt{q}\Big\} \\ & + \sum_{j=q+1}^{p_n}|\widehat{u}_{j,n}|\Big(\frac{\lambda_n}{\sqrt{n}} - |W_{j,n}|\Big) \\ \geqslant{}& \|\widehat{u}_n'^{(1)}\|_2\Big\{\frac{1-\delta}{2}\Lambda_{\min}(C_{11,n})\|\widehat{u}_n^{(1)}\|_2 - \frac{\lambda_n}{\sqrt{n}}\sqrt{q}(1+\rho)\Big\} \\ & + \frac{\lambda_n}{\sqrt{n}}(1-\rho)\cdot\sum_{j=q+1}^{p_n}|\widehat{u}_{j,n}|. \end{aligned}$$

Define

$$M_n \equiv \frac{2}{1-\delta}\cdot\frac{\lambda_n\sqrt{q}}{\sqrt{n}}(1+\rho)\cdot\Lambda_{\min}^{-1}(C_{11,n}).$$

Then $V_n(\widehat{u}_n) > 0$ depends on

$$\{\|\widehat{u}_n^{(1)}\|_2 > M_n\}.$$

Since $V_n(0) = 0$, the minimum of $V_n(\widehat{u}_n)$ cannot be attained at $\|\widehat{u}_n^{(1)}\|_2 > M_n$. Then assume that $\{\widehat{u}_n \in \mathbb{R}^{p_n} :\ \|\widehat{u}_n^{(1)}\|_2 \leqslant M_n,\ \widehat{u}_n^{(2)} \neq 0\}$ and (C1) holds. Set e_j to be the unit vector in the direction of j-th coordinate. Then the following inequality holds uniformly:

$$\begin{aligned}V_n(\widehat{u}_n) - V_n(\widehat{u}_n^{(1)}, 0) &= (\widehat{u}_n^{(1)})' C_{12,n} \widehat{u}_n^{(2)} + \frac{1}{2}(\widehat{u}_n^{(2)})' C_{22,n} \widehat{u}_n^{(2)} \\ &\quad + \frac{\lambda_n}{\sqrt{n}} \|\widehat{u}_n^{(2)}\| - (\widehat{u}_n^{(2)})' W_n^{(2)} \\ &> \sum_{j=q+1}^{p_n} |\widehat{u}_{j,n}| \Big[\frac{\lambda_n}{\sqrt{n}} - |W_{j,n}| - |((\widehat{u}_n^{(1)})' C_{12,n})_j| \Big] \\ &\geqslant \sum_{j=q+1}^{p_n} |\widehat{u}_{j,n}| \Big[\frac{\lambda_n}{\sqrt{n}}(1-\rho) - M_n \|C_{12,n} e_j\|_2 \Big] \\ &> 0. \end{aligned} \tag{5.1}$$

Set $\eta > 0$ such that $1 - \eta = \frac{1-\rho}{1+\rho} \cdot \frac{1-\delta}{2} \Lambda_{\min} q^{-\frac{1}{2}}$. By (C1), the last inequality of (5.1) holds. Then the minimum of $V_n(u_n)$ cannot be attained at $u_n^{(2)} \neq 0$ too, hence we have

$$\operatorname*{argmin}_{\widehat{u}_n \in \mathbb{R}^{p_n}} V_n(\widehat{u}_n) \in \{u_n \in \mathbb{R}^{p_n} : \|\widehat{u}_n^{(1)}\|_2 \leqslant M_n, \widehat{u}_n^{(2)} = 0\}.$$

After discussing the model selection consistency of $\widehat{\beta}_n^{(2)}$, we now consider about the model selection consistency of $\widehat{\beta}_n^{(1)}$. According to the definition of $\widehat{u}_n$ and the solution of the lasso, if we want $\widehat{\beta}_n^{(1)} \neq 0$ or $\operatorname{sign}(\widehat{\beta}_n^{(1)}) = \operatorname{sign}(\widehat{\beta}_n^{(1)})$, the following hold,

$$C_{11,n} \cdot \widehat{u}_n^{(1)} - W_n^{(1)} = -\frac{\lambda_n}{\sqrt{n}} \operatorname{sign}(\beta_n^{(1)}),$$

$$\Big| \frac{\widehat{u}_n^{(1)}}{\sqrt{n}} \Big| < |\beta_n^{(1)}|.$$

Combining above two restraints of $\widehat{u}_n^{(1)}$, the existence of such $\widehat{u}_n^{(1)}$ is implied by

$$|C_{11,n}^{-1} W_n^{(1)}| < \sqrt{n}\Big(|\beta_n^{(1)}| - \frac{\lambda_n}{n} |C_{11,n}^{-1} \cdot \operatorname{sign}(\beta_n^{(1)})| \Big).$$

Since $C_{11,n}^{-1} W_n^{(1)} = C_{11,n}^{-1} \cdot \frac{1}{\sqrt{n}} (X_n^{(1)})' \epsilon$, considering that ϵ_i are i.i.d random variables with mean 0 and variance σ^2 and

$$\Big\| C_{11,n}^{-1} \cdot \frac{1}{\sqrt{n}} (X_n^{(1)})' \Big\|_2^2 = C_{11,n}^{-1},$$

we have

$$P([C_{11,n}^{-1} W_n^{(1)}]_j > t) \leqslant P([W_n^{(1)}]_j > t \cdot \Lambda_{\min}),$$

where $\Lambda_{\min} = \Lambda_{\min}(C_{11,n})$. Besides, we also have

$$|C_{11,n}^{-1} \operatorname{sign}(\beta_n^{(1)})| \leqslant \|C_{11,n}^{-1}\|_2 \cdot \|\operatorname{sign}(\beta_n^{(1)})\|_2 \leqslant \sqrt{q} \cdot \Lambda_{\min}^{-1}.$$

By Bonferroni's inequality, we know that if we want to prove

$$P(\forall j \in S_n, \widehat{\beta}_{j,n} = 0) \to \mathrm{e}^{-nt}, \quad \text{as } n \to \infty,$$

it suffices to show that for every $j \in S_n$,

$$P(\widehat{\beta}_{j,n} = 0) \to \mathrm{e}^{-nt}, \quad \text{as } n \to \infty.$$

Hence, we have

$$P(\widehat{\beta}_{j,n} = 0) \leqslant P\Big([|W_n^{(1)}|]_j \geqslant \sqrt{n} \cdot \Lambda_{\min}\Big(|\beta_{j,n}| - \frac{\lambda_n}{n}\Lambda_{\min}^{-1}\sqrt{q}\Big)\Big).$$

Let $G_n = \min\big\{\sqrt{n} \cdot \Lambda_{\min}\big(|\beta_{j,n}| - \frac{\lambda_n}{n}\Lambda_{\min}^{-1}\sqrt{q}\big),\ \frac{\rho\lambda_n}{\sqrt{n}}\big\}$. Then we have

$$P(\widehat{S}_n = S_n) \geqslant P(\|W_n\|_\infty \leqslant G_n),$$

which completes the proof.

References

[1] Bennet, G., Probability inequalities for sums of independent random variables, *Journal of the American Statistical Association*, **57**, 1962, 33–45.

[2] Bickel, P. J., Ritov, Y. and Tsybakov, A. B., Simultaneous analysis of lasso and Dantzig selector, *Annals of Statistics*, **37**(4), 2009, 1705–1732.

[3] Buhlmann, P. and Van de Geer, S., Statistics for Hhigh-dimensional Data, Methods, Theory and Applications, Springer-Verlag, Heidelberg, 2011.

[4] Candes, E. and Tao, T., The Dantzig selector: Statistical estimation when p is much larger than n, *Annals of Statistics*, **35**(6), 2007, 2313–2351.

[5] Efron, B., Hastie, T., Johnstone, L., et al., Least angle regression, *Annals of Statistics*, **32**(2), 2004, 407–451.

[6] Fan, J. Q., Fan, Y. Y. and Barut, E., Adaptive robust variable selection, *Annals of Statistics*, **42**(1), 2014, 324–351.

[7] Fan, J. Q. and Li, R. Z., Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**(456), 2001, 1348–1360.

[8] Fan, J. Q. and Peng, H., Nonconcave penalized likelihood with a diverging number of parameters, *Annals of Statistics*, **32**(3), 2004, 928–961.

[9] Huang, J., Horowitz, J. L. and Ma, S., Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Annals of Statistics*, **36**(2), 2008, 587–613.

[10] Lv, J. C. and Fan, Y. Y., A unified approach to model selection and sparse recovery using regularized least squares, *Annals of Statistics*, **37**(6), 2009, 3498–3528.

[11] Meinshausen, N. and Bühlmann, P., High-dimensional graphs and variable selection with the lasso, *Annals of Statistics*, **34**(3), 2006, 1436–1462.

[12] Meinshausen, N. and Bühlmann, P., Stability selection, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), **72**(4), 2010, 417–473.

[13] Meinshausen, N. and Yu, B., Lasso-type recovery of sparse representations for high-dimensional data, *Annals of Statistics*, **37**(1), 2009, 246–270.

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

[15] Tibshirani, R. J., The lasso problem and uniqueness, *Electronic Journal of Statistics*, **7**, 2013, 1456–1490.

[16] Wainwright, M. J., Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming (lasso), *IEEE Transactions on Information Theory*, **55**(5), 2009, 2183–2202.

[17] Wu, L. and Yang, Y. H., Nonnegative elastic net and application in index tracking, *Applied Mathematics and Computation*, **227**, 2014, 541–552.

[18] Wu, L., Yang, Y. H. and Liu, H. Z., Nonnegative-lasso and application in index tracking, *Computational Statistics & Data Analysis*, **70**, 2014, 116–126.

[19] Yang, Y. H. and Wu, L., Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling, *Journal of Statistical Planning and Inference*, **174**, 2016, 52–67.

[20] Yuan, M. and Lin, Y., Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B*, **68**, 2006, 49–67.

[21] Zhang, C. H., Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, **38**(2), 2010, 894–942.

[22] Zhao, P. and Yu, B., On model selection consistency of lasso, *Journal of Machine Learning Research*, **7**, 2006, 2541–2563.

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

[24] Zou, H. and Hastie, T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B*, **67**, 2005, 301–320.

[25] Zou, H. and Zhang, H. L., On the adaptive elastic-net with a diverging number of parameters, *Annals of statistics*, **37**(4), 2009, 1733–1751.