On Model Selection Consistency of the Elastic Net When $p >> n$

Jinzhu Jia[1] and Bin Yu[2]

[1]*Peking University and* [2]*University of California, Berkeley*

Abstract: In this paper, we study the model selection property of the Elastic net. In the classical settings when p (the number of predictors) and q (the number of predictors with non-zero coefficients in the true linear model) are fixed, Yuan and Lin (2007) give a necessary and sufficient condition for the Elastic net to consistently select the true model, which is called the Elastic Irrepresentable Condition (EIC) in this paper. Here we study the general case when p, q and n all go to infinity. For general scalings of p, q and n, when gaussian noise is assumed, sufficient conditions on p, q and n ar given in this paper such that EIC guarantees the Elastic net's model selection consistency. We show that to make these conditions hold, n should grow at a rate faster than $q\log(p-q)$. For the classical case, when p and q are fixed, we also study the relationship between EIC and the Irrepresentable Condition (IC) which is necessary and sufficient for the Lasso to select the true model. Through theoretical results and simulation studies, we provide insights into when and why EIC is weaker than IC and when the Elastic net can consistently select the true model even when the Lasso can not.

Key words and phrases: Lasso; Elastic net; Model selection consistency; Irrepresentable Condition; Elastic Irrepresentable Condition.

1. Introduction

Regularization has been a popular technique for model fitting in statistical learning when the number of predictors p is large compared with the number of observations n. Regularization methods have been shown to have a better accuracy of prediction on future data (Tikhonov, 1943; Hoerl and Kennard, 1970). The Lasso (Tibshirani, 1996) which regularizes with an L_1 penalty, can also generates sparse models, which are more interpretable. The Lasso provides a computationally feasible way for model selection (Osborne et al, 2000; Efron et al 2004; Rosset, 2004; Zhao and Yu, 2007). But the Lasso does not perform well when the predictors are highly correlated or the number of predictors is much greater than the number of observations. Zou and Hastie (2005) proposed the

Elastic net, which also has the property of sparsity, to solve the above problems. Zou and Hastie (2005) state that the Elastic net regularization "is like a stretchable fishing net that retains all the big fish" and that "Simulation studies and real data examples show that the Elastic net often outperforms the Lasso in terms of prediction accuracy".

In this paper, we intend to understand the model selection performance of the Elastic net, relative to the Lasso. We obtain theoretical results showing that the Elastic net can select the true model consistently when the sparsity measure, the total number of predictors, and the sample size all go to infinity. We use both theoretical results and simulation studies to shed light on when and why the Elastic net can outperform the Lasso for model selection.

Assume our data consists of a design matrix $X \in R^{n\times p}$ and the response vector $Y \in R^n$. They follow a linear regression model

$$Y = X\beta + \epsilon, \tag{1.1}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of i.i.d. additive Gaussian noise with mean 0 and variance σ^2. Throughout this paper, the design matrix X is treated as a deterministic (non-random) matrix. For the random case all the conclusions can be obtained by conditioning on X. β is the vector of model coefficients. The model is assumed to be "sparse", i.e. most of the regression coefficients β are exactly zero corresponding to predictors that are irrelevant to the response. Without loss of generality, assume the first q elements of vector β are non-zeroes. Let $\beta_{(1)} = (\beta_1, \ldots, \beta_q)$ and $\beta_{(2)} = (\beta_{q+1}, \ldots, \beta_p)$, then $\beta_{(1)} \neq 0$ element-wise and $\beta_{(2)} = 0$.

Write $X_{(1)}$ and $X_{(2)}$ as the first q and the last $p-q$ columns of design matrix X respectively and let $C(n) = \frac{1}{n}X^TX$. For simplicity, $C(n)$ is denoted by C, which is a function of n. C can be expressed in a block-wise form:

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

where $C_{11} = \frac{1}{n}X_{(1)}^TX_{(1)}$,$C_{12} = \frac{1}{n}X_{(1)}^TX_{(2)}$,$C_{21} = \frac{1}{n}X_{(2)}^TX_{(1)}$ and $C_{22} = \frac{1}{n}X_{(2)}^TX_{(2)}$.

The naïve Elastic net estimate $\hat{\beta}$ is defined as

$$\hat{\beta}(\text{naïve}) = \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1, \tag{1.2}$$

where parameters λ_1 and λ_2 control the amount of regularization applied to the estimate. $\lambda_2 = 0$ leads the naïve elastic estimate back to the Lasso estimate.

Since the Elastic net estimate $\hat{\beta}$(Elastic net) is defined as $(1 + \lambda_2)\hat{\beta}$(naïve), it selects the same model as the naïve Elastic net estimate. In this paper, we will call the naïve Elastic net estimate ($\hat{\beta}$) the Elastic net estimate.

Recent works (Zhao and Yu, 2006; Zou, 2006; Meinshausen and Yu, 2007; Yuan and Lin, 2007) have worked precisely on the model selection consistency of the Lasso. It has been shown that in the classical case when p and q are fixed, a simple condition called the Irrepresentable Condition on the generating covariance matrices is necessary and sufficient for the Lasso's model selection consistency. IC is defined in Zhao and Yu (2006) as:

Irrepresentable Condition (IC). There exists a positive constant $\eta > 0$,

$$\left\| C_{21}C_{11}^{-1}\left(sign(\beta_{(1)})\right) \right\|_\infty \leq 1 - \eta, \tag{1.3}$$

where the inequality holds element-wise.

More precise results for the $p >> n$ case are in Wainwright (2006), which was the first to give conditions for the Lasso's model selection consistency in the case of general scalings of p, q and n. Yuan and Lin (2007) concentrate mainly on non-negative garotte, but contain a necessary and sufficient condition for the Elastic net to select the true model in the classical settings when p and q are fixed. EIC is defined as:

Elastic Irrepresentable Condition (EIC). There exists λ_1, λ_2 and a positive constant $\eta > 0$,

$$\left\| C_{21}(C_{11} + \frac{\lambda_2}{n}I)^{-1}\left(sign(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1}\beta_{(1)}\right) \right\|_\infty \leq 1 - \eta, \tag{1.4}$$

where the inequality holds element-wise.

EIC is exactly IC when when $\lambda_2 = 0$ and C_{11} is invertible. EIC does not need C_{11} to be invertible. If λ_2 is preselected and fixed, when λ_1 goes to ∞, the Elastic Irrepresentable Condition reverses back to the Irrepresentable Condition. Generally speaking, if the Irrepresentable Condition holds, then there exist some $\lambda_1 > 0$ and $\lambda_2 > 0$ such that the corresponding elastic Irrepresentable Condition holds. The relationship between EIC and IC will be further studied in Section 3.

In this paper, we analyze the model selection consistency of Elastic net for

general scalings of p, q and n. The fixed p and q case is a special case. For the classical settings, we do more analysis than that in Yuan and Lin (2007). Through special models and simulations, we study the relationship between EIC and IC; we show that EIC is weaker than IC and that the Elastic net can select the true model even when the Lasso can not. For the general case, we give sufficient conditions on the relationship of p, q and n such that EIC guarantees the Elastic net's model selection consistency.

The rest of the paper is organized as follows. In Section 2, we give our main results. For the general scalings of p, q and n, conditions on the relationship between p, q and n are given such that that EIC is sufficient for the Elastic net to select the true model. In Section 3, we compare the Elastic Irrepresentable Condition with the Irrepresentable Condition. Simulation studies are shown in Section 4. In Section 5, we conclude and propose the future directions for this research. The longer proofs can be found in the appendix.

2. Model Selection Consistency

We follow the notations and definitions of sign consistency defined in Zhao and Yu (2006) and Wainwright (2006). Define $\hat{\beta} =_s \beta$, if vector $\hat{\beta}$ and the true parameter β have the same sign element-wise.

Definition 1. ***Property*** $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$***:*** *There exists an optimal solution* $\hat{\beta}(\hat{\lambda}_1, \hat{\lambda}_2)$ *for model* (1.2) *with the property* $\hat{\beta} =_s \beta$.

Definition 2. *The Elastic net estimate is* ***Sign Consistent*** *if there exists* λ_1, λ_2 *such that*

$$\lim_{n\to\infty} P(\hat{\beta}(\lambda_1, \lambda_2) =_s \beta) = 1.$$

Note that the Elastic net estimate $\hat{\beta}(\lambda_1, \lambda_2)$ is sign consistent if and only if $P[\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)] \to 1$ as $n \to \infty$.

When p and q are fixed, Yuan and Lin (2007) have shown that EIC is a necessary and sufficient condition for the Elastic net to consistently select the true model. We show that when p, q and n all go to infinity, under some conditions on the relationship between p, q and n, EIC also guarantees that the Elastic net consistently selects the true model.

We first state necessary and sufficient conditions for property $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$

to hold in Lemma 1, which is a consequence of KKT (Karush-Kuhn-Tucker) conditions.

Lemma 1. *For any given $\lambda_1 > 0, \lambda_2 > 0$ and noise vector $\epsilon \in \mathbb{R}^n$, property $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$ holds if and only if*

$$\left| 2X_{(2)}^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T \epsilon - \frac{\lambda_1}{2} sign(\beta_{(1)}) - \lambda_2 \beta_{(1)} \right] - 2X_{(2)}^T \epsilon \right| \leq \lambda_1, \tag{2.1}$$

$$\left| \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T X_{(1)} \beta_{(1)} + X_{(1)}^T \epsilon - \frac{\lambda_1}{2} sign(\beta_{(1)}) \right] \right| > 0. \tag{2.2}$$

For shorthand, define $\overrightarrow{b} := \text{sign}(\beta_{(1)})$ and denote by e_i the vector with 1 in the $i'th$ position and zeroes elsewhere. For each index $i \in S = \{1, 2, \ldots, q\}$ and $j \in S^c = \{q+1, \ldots, p\}$, define the following random variables:

$$U_i := e_i^T \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T \epsilon - \frac{\lambda_1}{2} \overrightarrow{b} \right], \tag{2.3}$$

$$V_j := 2X_j^T \left\{ X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left(\frac{\lambda_1}{2} \overrightarrow{b} + \lambda_2 \beta_{(1)} \right) - \left[X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} X_{(1)}^T - I \right] \epsilon \right\}. \tag{2.4}$$

These random variables will play an important role in our analysis. In particular, condition (2.1) holds if and only if the event

$$\mathcal{M}(V) := \left\{ \max_{j \in S^c} |V_j| \leq \lambda_1 \right\} \tag{2.5}$$

holds. On the other hand, if we define $\rho := \min \left| \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T X_{(1)} \beta_{(1)} \right] \right|$, then the event

$$\mathcal{M}(U) := \left\{ \max_{i \in S} |U_i| < \rho \right\} \tag{2.6}$$

is sufficient to guarantee that condition (2.2) holds.

In the zero-noise setting ($\epsilon = 0$), the conditions in Lemma 1 will reduce to

$$\left| X_{(2)}^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[\text{sign}(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1} \beta_{(1)} \right] \right| \leq 1, \tag{2.7}$$

$$\left| \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T X_{(1)} \beta_{(1)} - \frac{\lambda_1}{2} \text{sign}(\beta_{(1)}) \right] \right| > 0. \tag{2.8}$$

When noises exist, under some conditions on the relationship between the scalings of p, q and n, the Elastic Irrepresentable Condition is still sufficient for the property of $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$ to hold with probability tending to 1 as $n \to \infty$:

Theorem 1. *Suppose that $Y = X\beta + \epsilon$, where each column of X is normalized to l_2-norm n and $\epsilon \sim N(0, \sigma^2 I)$. Assume EIC (1.4) holds. Define $\rho := \min \left| \left(C_{11} + \frac{\lambda_2}{n} I \right)^{-1} \left[C_{11} \beta_{(1)} \right] \right|$, and $C_{min} = \Lambda_{min}(C_{11}) + \frac{\lambda_2}{n}$, where $\Lambda_{min}(\cdot)$ denotes the minimal eigenvalue. If λ_1 is chosen such that*
(a) $\frac{\lambda_1^2}{n \log(p-q)} \to \infty$,

(b) $\frac{1}{\rho} \left\{ \sqrt{\frac{\log q}{n C_{min}}} + \frac{\lambda_1}{n} \left\| \left(C_{11} + \frac{\lambda_2}{n} I \right)^{-1} \overrightarrow{b} \right\|_\infty \right\} \to 0,$

then $P[\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)] \to 1$ as $n \to \infty$.

A proof of Theorem 1 can be found in the appendix.

Theorem 1 gives a general result for general scalings of p, q and n. In the classical setting where p and q are fixed, if C_{11} converges to a non-negative definite matrix C_0, ρ will converge to a non-negative number ρ_0. Suppose $\rho_0 > 0$, then condition (a) is equivalent to $\lambda_1 / \sqrt{n} \to \infty$ and condition (b) is equivalent to $\lambda_1 / n \to 0$, if $C_{min} \geq \alpha$ for some $\alpha > 0$.

Corollary 1. *When p and q are fixed, suppose that C_{11} converges to C_0, $\rho_0 > 0$ and $C_{min} \geq \alpha$ for some $\alpha > 0$, then EIC implies $P[\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)] \to 1$ as $n \to \infty$, if*
(a) $\lambda_1 / \sqrt{n} \to \infty$,

(b) $\lambda_1 / n \to 0$.

Note that $\lambda_1 = \sqrt{n} \log n$ is a suitable choice. A similar conclusion is also reached in Meinshausen and Buhlmann (2006), Zhao and Yu (2006), Zou (2006) and Wainwright (2007) for the Lasso to select the true model. Regarding constraints on λ_2: when C_{11} is invertible and $\Lambda_{min}(C_{11}) \geq \alpha$, for some $\alpha > 0$, any $\lambda_2 > 0$ can be chosen as long as it satisfies EIC; when C_{11} is not invertible, $\lambda_2 = \gamma n$ can be chosen, for any $\gamma > 0$ which satisfies EIC.

When all three parameters (n, p, q) grow into infinity, suppose that $C_{min} \geq \alpha$, for some $\alpha > 0$ and $\rho \geq \rho_0$, for some $\rho_0 > 0$. Then we have

Corollary 2. *EIC implies that the Elastic net has sign consistency if*
(a) $\frac{\lambda_1^2}{n\log(p-q)} \to \infty$,

(b) $\frac{\log q}{n} \to 0$,

(c) $\frac{\lambda_1\sqrt{q}}{n} \to 0$.

Proof. Note that $\left\| \left(C_{11} + \frac{\lambda_2}{n} I\right)^{-1} \overrightarrow{b} \right\| \leq C_{min}^{-1} || \overrightarrow{b} ||_2 = C_{min}^{-1}\sqrt{q}$. So, conditions (b) and (c) in Corollary 2 guarantee that condition (b) in Theorem 1 holds. □

The conditions $\frac{\lambda_1^2}{n\log(p-q)} (= (\frac{\lambda_1\sqrt{q}}{n})^2 \times \frac{n}{q\log(p-q)}) \to +\infty$ and $\frac{\lambda_1\sqrt{q}}{n} \to 0$ imply that the number of observations n must grow at a rate faster than $q\log(p-q)$.

3. Relationship between EIC and IC

As shown in Zou and Hastie (2005), the Elastic net can select the "important" variables for prediction and it often outperforms the Lasso in terms of prediction accuracy. Under some conditions, we have shown that in theory it consistently selects the relevant predictors. In this section, we will show theoretically that the Elastic net often outperforms the Lasso in terms of model selection consistency.

Proposition 1. *Irrepresentable Condition implies Elastic Irrepresentable Condition, but Elastic Irrepresentable Condition does not imply Irrepresentable Condition.*

This result is trivial, since $\lambda_2 = 0$ or small $\lambda_2 > 0$ leads EIC back to IC.

Proposition 1 shows that when the Lasso can select the true model, the Elastic net also can select the true model; the Elastic net often outperforms the Lasso in terms of model selection consistency. We have to point out that it may happen that in some situations neither the Lasso nor the Elastic net can select the true model, which can be seen by simulations in Section 4.

An interesting question is under what conditions, the Elastic net will do a much better job than the Lasso for model selection. In other words, what prior information about the model parameters would suggest that the Elastic net will select the true model while the Lasso does not? It is hard to answer this question in general. But, in some situations, we can provide some insight into when the EIC will hold while IC does not.

Consider the case $p-q=1$, that is, there exists only one irrelevant predictor. This is the simplest model selection problem. For this kind of problem, we can give a simple necessary and sufficient condition such that EIC holds.

Theorem 2. *In the case when $p-q=1$, EIC holds if and only if*

$$C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})\geq 1 \text{ and } C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}\beta_{(1)}<0, \tag{3.1}$$

or

$$C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})\leq -1 \text{ and } C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}\beta_{(1)}>0, \tag{3.2}$$

or

$$|C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})|\leq 1-\eta, \text{ for some } 0<\eta<1\ . \tag{3.3}$$

Proof. When $p-q=1$, $C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})$ and $C_{21}(C_{11}+\frac{\lambda_2}{n})^{-1}\beta_{(1)}$ are both scalers. Immediately, (1.4) is equivalent to conditions (3.1), (3.2) and (3.3) by choosing a suitable λ_1. □

Choosing the appropriate λ_2 for Theorem 2 requires difficult manipulations. Below, we give sufficient conditions to ensure that EIC holds and IC does not for *any* fixed value of λ_2.

Corollary 3. *Suppose C_{11} invertible, in the case when $p-q=1$, for any fixed value λ_2, when n is very large, EIC holds while IC does not if*

$$C_{21}C_{11}^{-1}sign(\beta_{(1)})\geq 1 \text{ and } C_{21}C_{11}^{-1}\beta_{(1)}<0 \tag{3.4}$$

or

$$C_{21}C_{11}^{-1}sign(\beta_{(1)})\leq -1 \text{ and } C_{21}C_{11}^{-1}\beta_{(1)}>0 \tag{3.5}$$

Proof. When $\lambda_2=0$, (3.1) is exactly condition (3.4) and (3.2) is exactly condition (3.5). λ_2 is not allowed to be 0, a small λ_2 or a small $\frac{\lambda_2}{n}$ can be chosen, such that conditions (3.1) and (3.2) hold, each of which is sufficient for EIC to hold. □

Denote Ψ by the estimated regression of the linear model $X_{(2)}=X_{(1)}\psi+noise$. It can be the OLS estimate $C_{11}^{-1}C_{12}$ as in Corollary 3 or the ridge regression estimate $(C_{11}+\frac{\lambda_2}{n})^{-1}C_{12}$ as in Theorem 2. Theorem 2 and its corollary

(Corollary 3) suggest that if the Lasso does not select the true model, it is because $|\Psi^T sign(\beta_{(1)})|$ is too large. But the Elastic net might be able to conquer this problem by introducing another penalty term $\Psi^T\beta_{(1)}$ on $\Psi^T sign(\beta_{(1)})$ such that the absolute value of the new term $\Psi^T sign(\beta_{(1)}) + \alpha\Psi^T\beta_{(1)}$ is not very large for some $\alpha > 0$. The small absolute value of the new term implies that the EIC holds, and therefore the Elastic net can consistently select the true model.

In the situations when $p - q \geq 2$, explanations about EIC are complicated. But conditions (3.1) and (3.2) are necessary conditions such that EIC holds. We state it as a corollary of Theorem 2.

Corollary 4. *In the case when $p - q > 1$, EIC holds only if*

$$\left[C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\beta_{(1)}\right]_i < 0 \text{ when } \left[C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})\right]_i \geq 1, \tag{3.6}$$

and

$$\left[C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\beta_{(1)}\right]_i > 0 \text{ when } \left[C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}sign(\beta_{(1)})\right]_i \leq -1, \tag{3.7}$$

where, $[\cdot]_i$ denote the $i-th$ element of a vector.

Proof. When condition (3.6) or (3.7) does not hold, then

$$\begin{aligned}\left|C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}[\frac{2\lambda_2}{\lambda_1}\beta_{(1)} + sign(\beta_{(1)})]_i\right| &\geq \left|[C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\beta_{(1)}]_i\right| \\ &\geq 1\end{aligned}$$

which violates EIC. □

4. Simulations

Zou and Hastie (2005) contain many experiments to show that the Elastic net performs much better than the Lasso, OLS and ridge regression in terms of prediction accuracy, but they did not compare the model selection performances between the Lasso and the Elastic net. Yuan and Lin (2007) also have no example to show the differences of the performance on the model selection consistency

between the Lasso and the Elastic net. In this section, some simulations are provided to show that when the Lasso can not select the true model, the Elastic net may still select the true model. When $p >> n$, especially when $q > n$, the Lasso can select at most n variables before the model saturates. So, when $q > n$, the Lasso theoretically can not select all of the true predictors . We will give an example to show that the Elastic net might be able to solve this kind of problems.

In the first 3 examples, p and q are small compared to $n = 1000$. These examples can be treated as fixed p and q cases. Because of the large number of observations, the results are consistent and the plots appear the same for multiple simulations. From Corollary 1 and Corollary 3, it can be seen that the choice of λ_2 is not very important. In these examples, we take $\lambda_2 = 100$. We did many simulations with different λ_2's, and did not see much effect of λ_2 on the performance of model selection consistency.

Example 1. The first example has the same settings as Zhao and Yu (2006). They gave an example with $p = 3$ to show that when the Irrepresentable Condition holds there is a consistent Lasso solution and when the Irrepresentable Condition does not hold, there is no consistent Lasso solution.

X_1, X_2, e and ϵ are first generated from the standard normal distribution with mean 0 and variance 1. X_3 is generated to be correlated with X_1 and X_2 by

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

which also has a standard normal distribution. The true linear model is:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon.$$

Now, consider two settings: (a) $\beta_1 = 2, \beta_2 = 3$ and (b) $\beta_1 = -2, \beta_2 = 3$. In both settings, $X_{(1)} = (X_1, X_2), X_{(2)} = X_3$ and it is easy to check that $C_{22}C_{11}^{-1} = (\frac{2}{3}, \frac{2}{3})$. So, setting (b) makes Irrepresentable Condition hold, while setting (a) does not. The Lasso and the Elastic net are applied to both settings (a) and (b) respectively and the solution pathes are shown in Figure 4.1 and Figure 4.2. Figures 4.1 and 4.2 show that in setting (a), neither the Lasso nor the Elastic net can select the true model and in setting (b), both the Lasso and the Elastic net can select the true model.

Example 2. This example is used to illustrate that when the Irrepresentable

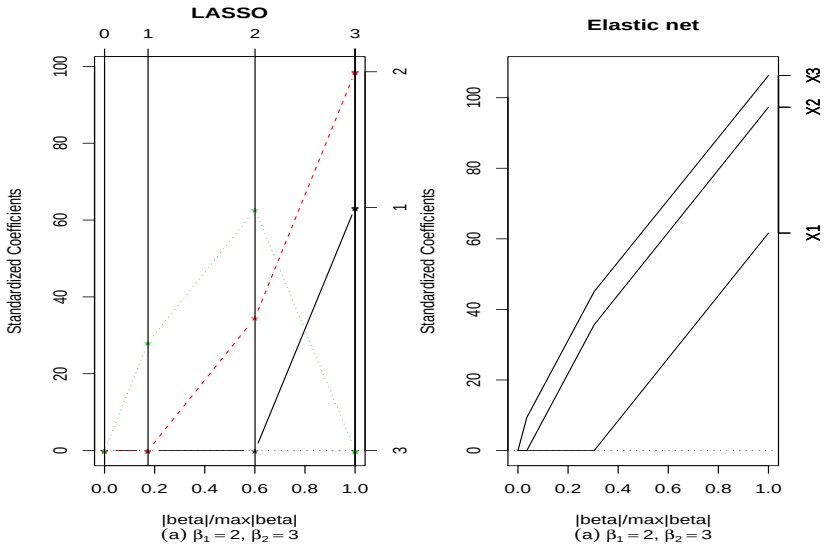

Figure 4.1: the Lasso solution paths for setting (a)

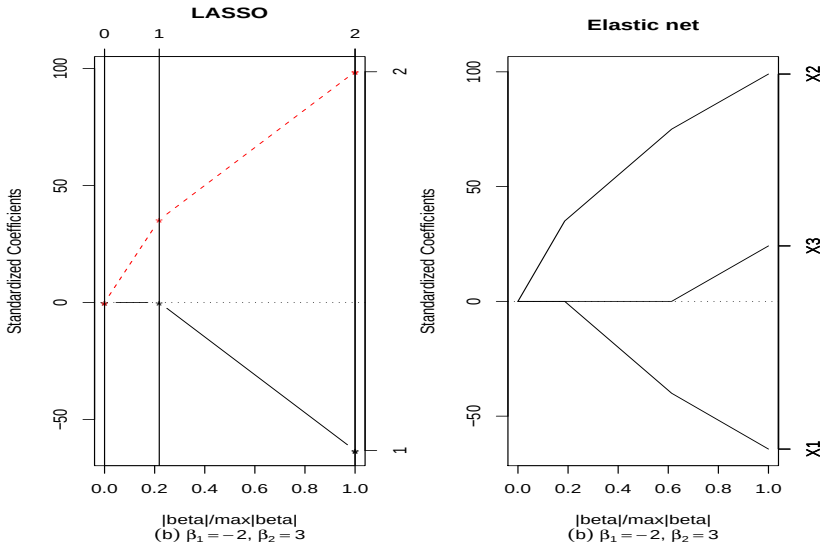

Figure 4.2: Elastic net solution paths for setting (b)

Condition does not hold, the Elastic Irrepresentable condition may hold and the Elastic net will select the true model, while the Lasso does not. In this example, $p = 6$. $X_1, X_2, X_3, X_4, X_5, e$ and ϵ are first generated from the standard normal distribution. X_6, also from the standard normal distribution, is generated to be correlated with X_1, X_2, X_3, X_4 and X_5 by

$$X_6 = \frac{1}{8}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4 + \frac{1}{2}X_5 + \eta e,$$

where the constant $\eta = \frac{\sqrt{11}}{8}$ is used to make X_6 have variance 1. The regression model is

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon.$$

Now $X_{(1)} = (X_1, X_2, X_3, X_4, X_5)$, $X_{(2)} = X_6$ and it is easy to check that $C_{21}C_{11}^{-1} = (\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Suppose that $\beta_1 < 0, \beta_2 < 0, \beta_3 > 0, \beta_4 > 0$ and $\beta_5 > 0$. It is easy to check that

$$|C_{21}C_{11}^{-1}(sign(\beta_{(1)})| = \frac{9}{8} > 1,$$

so, the Irrepresentable Condition does not hold.

In the settings above, a sufficient condition can be given such that the Elastic net select a consistent model. This condition is a direct consequence of Corollary 3.

In the settings of Example 2, the Elastic Irrepresentable Condition holds if

$$-(\beta_1 + 2\beta_2) > 4\beta_3 + 4\beta_4 + 4\beta_5 \tag{4.1}$$

Now let $\beta_1 = -4, \beta_2 = -2, \beta_3 = 0.5, \beta_4 = 0.6$ and $\beta_5 = 0.7$. It is easy to check that inequality (4.1) holds. The Lasso is first used to get the solution path shown in Figure 4.3 (a) and then the Elastic net is used to get the solution path shown in Figure 4.3 (b). The figure shows that the Lasso does not select the true model while the Elastic net does.

Example 3. As reported in Zou and Hastie (2005), when predictors are highly correlated, the Lasso tends to select only one of these highly correlated predictors. Especially, when there are two predictors which are the same, theoretically, the Lasso can not select both of them. In this example, we will show that the Elastic net can select both of them and can select the true model. By this example, we also show that when C_{11} is not invertible, we can still consider

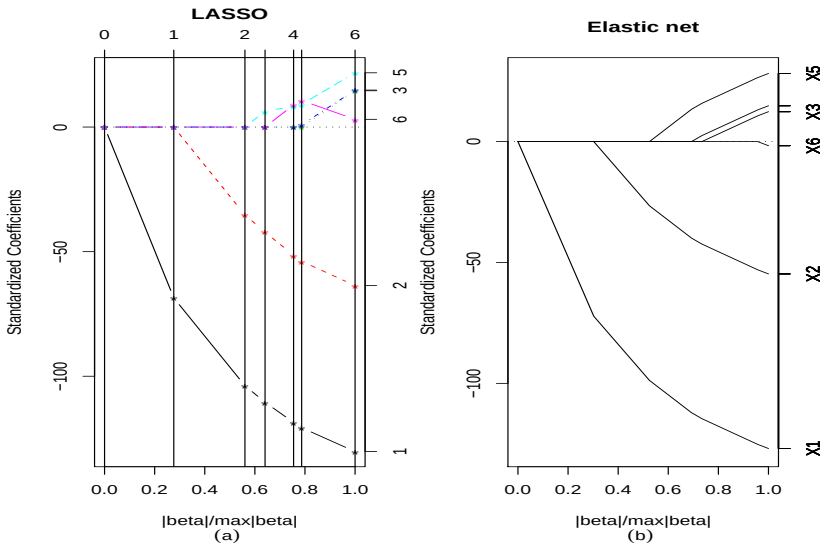

Figure 4.3: Lasso and Elastic net solution paths

the consistency of the Elastic net. While the consideration of consistency of the Lasso needs the assumption that C_{11} is invertible.

X_1, X_2, e and ϵ are first generated from a normal distribution with mean 0 and variance 1. Let $X_3 = X_2$. X_4 is generated to be correlated with X_1, X_2 and X_3 by

$$X_4 = \frac{2}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 + \frac{1}{3}e,$$

which also has a standard normal distribution. The true linear model is:

$$Y = -2X_1 + X_2 + X_3 + \epsilon.$$

The Lasso and the Elastic net are applied separately and the solution paths are shown in Figure 4.4. This figure shows that the Elastic net selects the true model while the Lasso does not.

Example 4. In this example, we want to illustrate that if $p >> n$, and EIC holds, then conditions in Corollary 2 of Theorem 1 guarantee that the Elastic net can select the true model. In the $p > n$ case, the Lasso selects at most n variables before it saturates. So if $q > n$, the Lasso cannot select the true model.

Set $q = 50$ and $p = 52$. From the comments after Corollary 2, n is supposed to grow at a rate faster than $q \log(p - q)$, which is equal to $50 \times \log 2 = 35$. So

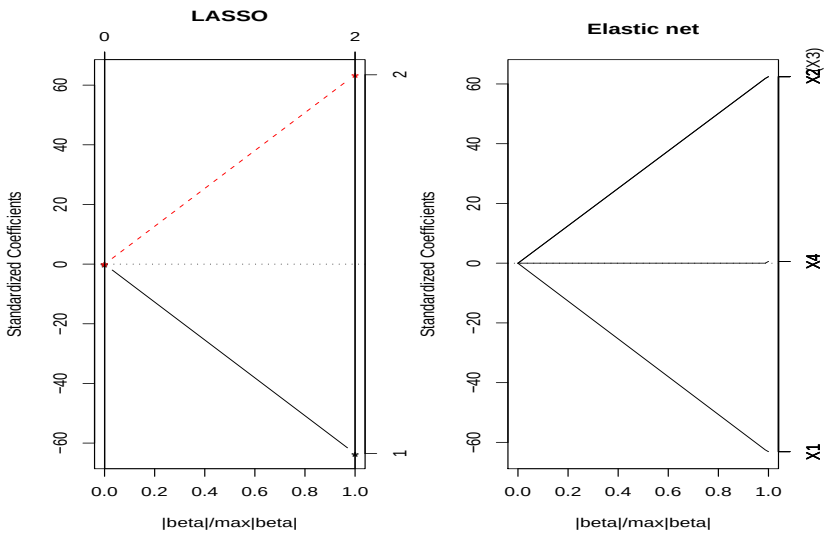

Figure 4.4: Lasso and Elastic net solution paths; $X_2 = X_3$

here we choose $n = 46$ which is less than q. The design matrix X is generated from joint standard normal distribution $N(0, I_{p\times p})$. Set $\lambda_2 = 0.01$ and simulate X, such that X satisfies $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1} \times \mathbf{1} < 1$, where $\mathbf{1}$ is a column vector with all entries being 1. Let $\beta = [\beta_{(1)}, \beta_{(2)}]$, where $\beta_{(1)}$ is a $q-$vector with all entries being 1 and $\beta_{(2)}$ is a $(p-q)-$vector with all entries being 0. Since $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\left(sign(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1}\beta_{(1)}\right) = (1 + \frac{2\lambda_2}{\lambda_1})C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1} \times \mathbf{1}$, there exists some λ_1 such that EIC holds. The true model is: $Y = X\beta + 0.04 \times \epsilon$. Then the Elastic net is applied. The solution path is shown in Figure 4.5.

After examining the solution on the path, we find that the solution corresponding to the vertical line in Figure 4.5 recovers exactly the first q non-zero predictors. Theoretically, the Lasso can select at most $n = 46$ variables and so the Lasso does not perform well on this data. After applying the Lasso on this simulated data, we find that it can only select 45 variables at most.

5. Conclusion

In this paper, we have discussed the ability of the Elastic net to recover the sparsity pattern of regression coefficients β. EIC is crucial for the Elastic net's model selection consistency. In the classical case when p and q are fixed, EIC is

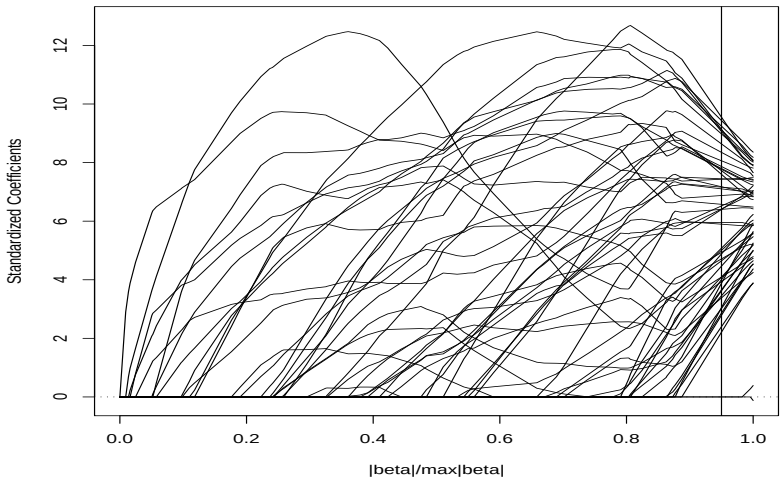

Figure 4.5: Elastic net solution paths for large p, q, small n

necessary and sufficient for the Elastic net to consistently select the true model (Yuan and Lin, 2007). When p and q both grow as n grows, EIC is not sufficient any more. Some conditions between the relationship of p, q and n are required. In this paper, for our consistency results, it is required that n grows at a rate faster than $q \log(p - q)$. When $p > n$, as in Example 4, the Elastic net performs better than the Lasso.

We compared the ability of the Elastic net to select the true model with that of the Lasso. EIC is weaker than IC. So, the Elastic net always performs better than the Lasso in terms of model selection consistency. From Example 2, it can be seen that when the Lasso can not select the true model, the Elastic net may select the true model. But we also see that in some situations, neither the Lasso nor the Elastic net selects the true model (see Example 1). Example 3 is used to show that when the true predictors are highly correlated, the Lasso does not select all the highly correlated variables. Yet, the Elastic net can select all of them.

At last, we propose future directions for this research. From Theorem 1 and its corollaries, the choice of λ_2 does not affect the Elastic net's ability to

select the true model when EIC hold. This suggests that, in practice, we can find a suitable λ_2 such that the EIC holds before the Elastic net is applied to do model selection and prediction. In some situations (see Corollary 3), any fixed λ_2 satisfies EIC. But in general, how to choose a suitable λ_2 such that EIC holds should be studied further.

Acknowledgments

Jinzhu Jia is supported by a fellowship from China Scholarship Council and he thanks the Statistics Department at UC Berkeley for hosting his visit during which this work is done. Bin Yu is partially supported by NSF grant DMS-0605165, ARO grant W911NF-05-1-0104, NSFC grant 60628102, and a grant from MSRA. We thank Jing Lei, Pradeep Ravikumar, Guilherme Rocha, Karl Rohe and Ying Xu for their helpful comments and suggestions on this paper.

Appendix: Proofs

Proof of Lemma 1. By standard (KKT) conditions for optimality in convex program, the point $\hat{\beta}$ is optimal if and only if

$$2X^TX\hat{\beta} - 2X^TY + 2\lambda_2\hat{\beta} + \lambda_1\hat{z} = 0. \tag{1}$$

Here

$$\hat{z} = \begin{cases} \text{sign}(\hat{\beta}_i) & \hat{\beta}_i \neq 0 \\ \text{any real number which} \in [-1,1] & \hat{\beta}_i = 0. \end{cases}$$

Substituting Y by $X\beta + \epsilon$ yields:

$$2X^TX(\hat{\beta} - \beta) - 2X^T\epsilon + 2\lambda_2\hat{\beta} + \lambda_1\hat{z} = 0. \tag{2}$$

Since condition $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$ holds if and only if we have

$$\hat{\beta}_{(2)} = 0, \hat{\beta}_{(1)} \neq 0, \text{ and } \hat{z}_{(1)} = \text{sign}(\beta_{(1)}), |\hat{z}_{(2)}| \leq 1.$$

From these conditions and using equation (2), we conclude that the condition $\mathcal{R}(X, \beta, \epsilon, \lambda_1, \lambda_2)$ holds if and only if

$$2X_{(2)}^TX_{(1)}(\hat{\beta}_{(1)} - \beta_{(1)}) - 2X_{(2)}^T\epsilon = -\lambda_1\hat{z}_{(2)}, \tag{3}$$

$$2X_{(1)}^TX_{(1)}(\hat{\beta}_{(1)} - \beta_{(1)}) - 2X_{(1)}^T\epsilon + 2\lambda_2\hat{\beta}_{(1)} = -\lambda_1\text{sign}(\beta_{(1)}). \tag{4}$$

By these two equations, we may solve for $\hat{\beta}_{(1)}$ and $\hat{z}_{(2)}$ to conclude that

$$-\lambda_1 \hat{z}_{(2)} = 2X_{(2)}^T X_{(1)}(X_{(1)}^T X_{(1)} + \lambda_2 I)^{-1}(X_{(1)}^T \epsilon - \frac{\lambda_1}{2}\text{sign}(\beta_{(1)}) - \lambda_2 \beta_{(1)}) - 2X_{(2)}^T \epsilon,$$

$$\hat{\beta}_{(1)} = (X_{(1)}^T X_{(1)} + \lambda_2 I)^{-1}(X_{(1)}^T X_{(1)}\beta_{(1)} + X_{(1)}^T \epsilon - \frac{\lambda_1}{2}\text{sign}(\beta_{(1)})).$$

The conditions $\hat{\beta}_{(1)} \neq 0$ and $|\hat{z}_{(2)}| \leq 1$ yield conditions (2.1) and (2.2) respectively. □

Before proving Theorem 1, we state without proof one well-known comparison result on Gaussian maxima (see Ledoux and Talagrand, 1991).

Lemma 2. *For any Gaussian random vector $(X_1, \ldots, X_n)$, we have*

$$E \max_{1 \leq i \leq n} |X_i| \leq 3\sqrt{\log n} \max_{1 \leq i \leq n} \sqrt{EX_i^2} \tag{5}$$

Proof of Theorem 1.

1. Analysis of $\mathcal{M}(V)$

Note that V_j is Gaussian with mean

$$\mu_j = E(V_j) = X_j^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} (\lambda_1 \overrightarrow{b} + 2\lambda_2 \beta_{(1)}).$$

Recall that the Elastic Irrepresentable Condition is:

$$\left| X_{(2)}^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \left[\text{sign}(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1}\beta_{(1)}\right] \right| \leq 1 - \epsilon. \tag{6}$$

By condition (6), $|\mu_j| \leq (1 - \eta)\lambda_1$.

Define $\widetilde{V}_j := 2X_j^T \left[I - X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} X_{(1)}^T\right] \epsilon$, then $V_j = \mu_j + \widetilde{V}_j$. Note $\mathcal{M}(V)$ holds if and only if $\frac{\max_{j \in S^c} V_j}{\lambda_1} \leq 1$ and $\frac{\min_{j \in S^c} V_j}{\lambda_1} \geq -1$. Since

$$\frac{\max_{j \in S^c} V_j}{\lambda_1} = \frac{\max_{j \in S^c} \mu_j + \widetilde{V}_j}{\lambda_1} \leq (1 - \eta) + \frac{1}{\lambda_1} \max_j \widetilde{V}_j, \text{ and} \tag{7}$$

$$\frac{\min_{j \in S^c} V_j}{\lambda_1} = \frac{\min_{j \in S^c} \mu_j + \widetilde{V}_j}{\lambda_1} \geq -(1 - \eta) + \frac{1}{\lambda_1} \min_j \widetilde{V}_j, \tag{8}$$

now we need to show that

$$P\left[\frac{1}{\lambda_1} \max_{j \in S^c} \widetilde{V}_j > \eta, \text{ or } \frac{1}{\lambda_1} \min_{j \in S^c} \widetilde{V}_j < -\eta\right] \to 0. \tag{9}$$

In fact, it is sufficient to show that $P\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_1} > \eta\right] \to 0$. By applying Markov's inequality and Gaussian comparison results (5), we have

$$P\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_1} > \eta\right] \le \frac{E[\max_{j\in S^c}|\widetilde{V}_j|]}{\lambda_1\eta} \le \frac{3\sqrt{\log(p-q)}}{\lambda_1\eta}\max_j\sqrt{E[\widetilde{V}_j^2]}. \quad (10)$$

Straightforward computation yields that

$$\begin{aligned}
\frac{1}{4}E[\widetilde{V}_j^2] &= \sigma^2 X_j^T\left[I - X_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^T\right]^2 X_j \\
&= \sigma^2 X_j^T\left[I - 2X_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^T\right]X_j \\
&+ \sigma^2 X_j^TX_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}(X_{(1)}^TX_{(1)})\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^TX_j \\
&\le \sigma^2 X_j^T\left[I - 2X_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^T\right]X_j \\
&+ \sigma^2 X_j^TX_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}(X_{(1)}^TX_{(1)})\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^TX_j \\
&+ \sigma^2 X_j^TX_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}\lambda_2 I\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^TX_j \\
&= \sigma^2 X_j^T\left[I - X_{(1)}\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^T\right]X_j \\
&\le \sigma^2 X_j^TX_j = n\sigma^2.
\end{aligned}$$

Put it into inequality (10), we have

$$P\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_1} > \eta\right] \le \frac{6\sigma\sqrt{n\log(p-q)}}{\lambda_1\eta},$$

so, condition (1) in Theorem 5 guarantees that $P\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_1} > \eta\right] \to 0$, and hence $P(\mathcal{M}) \to 1$.

2. Analysis of $\mathcal{M}(U)$

Define $Z_i = e_i^T\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}X_{(1)}^T\epsilon$, then

$$\begin{aligned}
\max_i|U_i| &= \max_i|Z_i - \frac{1}{2}e_i^T\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}\lambda_1\overrightarrow{b}| \\
&\le \max_i|Z_i| + \frac{1}{2}\lambda_1\left\|\left(X_{(1)}^TX_{(1)} + \lambda_2 I\right)^{-1}\overrightarrow{b}\right\|_\infty.
\end{aligned}$$

Note Z_i is Gaussian with mean 0 and variance

$$\begin{aligned} var(Z_i) &= \sigma^2 e_i^T \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} (X_{(1)}^T X_{(1)}) \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} e_i \\ &\leq \sigma^2 e_i^T \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} e_i \\ &\leq \frac{\sigma^2}{nC_{min}} \end{aligned}$$

So by standard Gaussian comparison theorem (5), we have

$$E[\max_i |Z_i|] \leq 3\sqrt{\frac{\sigma^2 \log q}{nC_{min}}}.$$

$$\begin{aligned} & 1 - P\left[\left|\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \left[X_{(1)}^T X_{(1)} \beta_{(1)} + X_{(1)}^T \epsilon - \frac{\lambda_1}{2} \text{sign}(\beta_{(1)})\right]\right| > 0\right] \\ & \leq P\left[\max_i |U_i| \geq \rho\right] \\ & \leq P\left[\frac{1}{\rho}\left\{\max_i |Z_i| + \frac{1}{2}\lambda_1 \left\| \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \overrightarrow{b} \right\|_\infty \right\} \geq 1\right] \\ & \leq \frac{1}{\rho}\left\{E\left[\max_i |Z_i|\right] + \frac{1}{2}\lambda_1 \left\| \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \overrightarrow{b} \right\|_\infty \right\} \\ & \leq \frac{1}{\rho}\left\{3\sqrt{\frac{\sigma^2 \log q}{nC_{min}}} + \frac{1}{2}\lambda_1 \left\| \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \overrightarrow{b} \right\|_\infty \right\}. \end{aligned}$$

So, condition (2) in Theorem 1 guarantees that $P(\mathcal{M}) \to 1$. □

References

Donoho, D., Elad M., and Temlyakov, V. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory.* **52**, NO. 1, JANUARY, 6-18.

Efron, B., Hastie, T., and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics.* **32**, 407-499.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics.* **12** 55-67.

Ledoux, M. and Talagrand, M. (1991) Probability in Banach Spaces: Isoperimetry and Processes. Springer-Verlag, New York, NY.

Meinshausen, N. and Buhlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics.* **34(3)**, 1436-1462.

Meinshausen, N. and Yu, B. (2007) Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics.* To appear.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000) On the lasso and its dual. *Journal of Computational and Graphical Statistics.* **9(2)**, 319-37.

Rosset, S. (2004) Tracking curved regularized optimization solution paths. *NIPS.*

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.* **58(1)**, 267-288.

Tikhonov, A. N. (1943) On the stability of inverse problems. *Dokl. Akad. Nauk SSSR.* **39**, No. 5, 176-179.

Wainwright, M. (2006) Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Technical Report, statictics Department, UC Berkeley.*

Yuan, M and Lin, Y. (2007) On the Nonnegative Garrote Estimator. *J. R. Statist. Soc. B.* **69**, Part 2, 143-161.

Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research.* **7**, 2541-2563.

Zhao, P. and Yu, B. (2007) Stagewise Lasso. *The Journal of Machine Learning Research.* **8**, 2701-2726.

Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association.* Volume 101, Number 476, 1418-1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* **67**, Part 2, 301-320.

School of Mathematical Sciences, Peking University, Beijing, P. R. China.
E-mail: jzjia@math.pku.edu.cn

Department of Statistics, University of California, Berkeley, CA, USA.
E-mail: binyu@stat.berkeley.edu