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

Jinzhu Jia[1] and Bin Yu[2]

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

Abstract: 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. They showed that it consistently selects the true model if and only if there exist suitable sequences $\lambda_1(n)$ and $\lambda_2(n)$ that satisfy EIC (which is defined later in the 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 are given 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 $q \log(p-q)$. We compare the variable selection performance of the Elastic Net with that of the Lasso. Through theoretical results and simulation studies, we provide insights into when the Elastic Net can consistently select the true model even when the Lasso cannot. We also point out through examples that when the Lasso cannot select the true model, it is very likely that the Elastic Net cannot select the true model either.

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 generate sparse models which are more interpretable. The Lasso provides a computationally feasible way for model selection (Osborne, Presnell, and Turlach (2000); Efron, Hastie, and Tibshirani (2004); Rosset (2004); Zhao and Yu (2007)). But it 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 show 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. This condition can be weakened to some moments condition (Zhao and Yu (2006)) or to some tail probability function condition (Ravikumar, Wainwright, Raskutti, and Yu (2008)). To simplify the proofs, we only consider Gaussian noise in the paper. Throughout, 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 the block-wise form:

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

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

The naïve Elastic Net estimate $\hat{\beta}$ is

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

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

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

Recent work (Zhao and Yu (2006); Zou (2006); Yuan and Lin (2007); Meinshausen and Yu (2008)) has been 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 follows.

Irrepresentable Condition (IC). There exists a positive constant $\eta > 0$ (which does not change with n), with

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

More precise results for the $p \gg n$ case are in Wainwright (2007), 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 the non-negative garotte, but give 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 follows.

Elastic Irrepresentable Condition (EIC). There exists a positive constant $\eta > 0$ (which does not change with n), with

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

Whether or not EIC holds depends on the data and the choice of parameters λ_1 and λ_2. EIC is exactly IC when $\lambda_2 = 0$ and C_{11} is invertible. EIC does not need C_{11} to be invertible. If C_{11} is invertible, and λ_2 is preselected and fixed, when λ_1

goes to ∞, as n goes to ∞, the Elastic Irrepresentable Condition reverts to the Irrepresentable Condition. Generally speaking, if the Irrepresentable Condition holds, then there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that the corresponding Elastic Irrepresentable Condition holds. The relationship between EIC and IC is further studied in Section 3.

Here we analyze the model selection consistency of the Elastic Net for general scalings of p, q, and n. The fixed p and q case is a special case. 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. We compare the model selection performance of the Elastic Net with that of the Lasso. We show that the Elastic Net can select the true model even when the Lasso cannot.

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 EIC is sufficient for the Elastic Net to select the true model. In Section 3, we compare the Elastic Net's model selection performance with the Lasso. Simulation studies are presented in Section 4 and we conclude in Section 5. The longer proofs can be found in the Appendix.

2. Model Selection Consistency

We follow the notation and definitions of sign consistency as found in Zhao and Yu (2006) and Wainwright (2007). Take $\hat{\beta} =_s \beta$, 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* $\hat{\beta}(\lambda_1, \lambda_2)$*, depending on the given parameters* λ_1 *and* λ_2*, for* (1.2) *with the property* $\hat{\beta} =_s \beta$*.*

Definition 2. *The Elastic Net estimate is* ***Sign Consistent*** *if there exists* $\hat{\lambda}_1, \hat{\lambda}_2$*, both of which are functions of* n *and depend on the data, such that*

$$\lim_{n\to\infty} P(\hat{\beta}(\hat{\lambda}_1, \hat{\lambda}_2) =_s \beta) = 1.$$

Note that the Elastic Net estimate $\hat{\beta}(\hat{\lambda}_1, \hat{\lambda}_2)$ is sign consistent if and only if $P[\mathcal{R}(X, \beta, \epsilon, \hat{\lambda}_1, \hat{\lambda}_2)] \to 1$ as $n \to \infty$.

When p and q are fixed, Yuan and Lin (2007) have shown that the Elastic Net consistently selects the true model if and only if there exist suitable sequences of $\lambda_1(n)$ and $\lambda_2(n)$ that satisfy EIC. 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; Lemma 1 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*

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

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

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

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

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

These random variables 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, if λ_2 is chosen such that

$$\text{sign}\left(\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} X_{(1)}^T X_{(1)} \beta_{(1)}\right) = \text{sign}(\beta_{(1)}). \tag{2.7}$$

Condition (2.7) holds, if λ_2 is very small. Throughout this paper, we constrain λ_2 such that (2.7) holds.

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

$$\left| X_{(2)}^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[\operatorname{sign}(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1} \beta_{(1)} \right] \right| \le 1, \tag{2.8}$$

$$\operatorname{sign}\left(\left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} \left[X_{(1)}^T X_{(1)} \beta_{(1)} - \frac{\lambda_1}{2} \operatorname{sign}(\beta_{(1)}) \right] \right) = \operatorname{sign}(\beta_{(1)}). \tag{2.9}$$

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 $\sqrt{n}$ and $\epsilon \sim N(0, \sigma^2 I)$. Assume EIC (1.4) holds. Consider $q > 1$ and $p - q > 1$. If $\rho := \min \left| \left(C_{11} + \frac{\lambda_2}{n} I \right)^{-1} \left[C_{11} \beta_{(1)} \right] \right|$, $C_{min} = \Lambda_{min}(C_{11}) + \frac{\lambda_2}{n}$, where $\Lambda_{min}(\cdot)$ denotes the minimal eigenvalue, and λ_1, λ_2 are 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 result for general scalings of p, q, and n. In the classical setting where p, q, and β are fixed, if C_{11} converges to a non-negative definite matrix C_0, suitable choice of λ_2 makes ρ converge to a non-negative number ρ_0. Suppose $\rho_0 > 0$, then (a) is equivalent to $\lambda_1 / \sqrt{n} \to \infty$ and (b) is equivalent to $\lambda_1 / n \to 0$, if $C_{min} \ge \alpha$ for some $\alpha > 0$.

Corollary 1. *When p, q, and β are fixed, suppose that C_{11} converges to C_0, $\rho_0 > 0$, and $C_{min} \ge \alpha$ 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) $\lambda_1 / \sqrt{n} \to \infty$, and (b) $\lambda_1 / n \to 0$.*

Note that $\lambda_1 = \sqrt{n} \log n$ is a suitable choice for the fixed p and q case. 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.

When all three parameters (n, p, q) go 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 the following.

Corollary 2. *EIC implies that the Elastic Net has sign consistency if* (a) $\frac{\lambda_1^2}{n\log(p-q)} \to \infty$, (b) $\frac{\log q}{n} \to 0$, *and* $(c)\frac{\lambda_1\sqrt{q}}{n} \to 0$.

Proof. Note that $\left\| \left(C_{11} + \frac{\lambda_2}{n}I\right)^{-1} \overrightarrow{b} \right\|_\infty \leq C_{min}^{-1} || \overrightarrow{b} ||_2 = C_{min}^{-1}\sqrt{q}$. 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 $q\log(p-q)$. A suitable choice for λ_1 is $(\frac{\lambda_1\sqrt{q}}{n})^2 = (\frac{n}{q\log(p-q)})^{-\alpha}$, for some $0 < \alpha < 1$, with which we have $\lambda_1 = \frac{n^{1-\alpha/2}}{q^{(1-\alpha)/2}(\log(p-q))^{-\alpha/2}}$.

3. Comparison with Lasso

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. We have shown that in theory the Elastic Net can consistently select the relevant predictors, under conditions stated in Theorem 1. In this section, we compare the model selection performance of the Elastic Net with that of the Lasso. Obviously when the Lasso selects the true model, the Elastic Net can also select the true model. We also provide insights into when EIC is weaker than IC and when the Elastic Net can consistently select the true model even when the Lasso cannot.

Proposition 1. *IC implies that for any* $\lambda_1 > 0$, *there exists* λ_2, *such that EIC holds, but EIC does not imply IC.*

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

Proposition 1 suggests that when IC holds, under the conditions of Theorem 1, the Elastic Net can select the true model. From previous work (Wainwright (2007)), under similar conditions, IC makes the Lasso select the true model. So, 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 will not? It is hard to answer this question in general. But, in some situations, we can provide some insight into when EIC holds 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, and we can give a simple necessary and sufficient condition such that EIC holds.

First we give some regularity conditions on the model. These conditions are easily fulfilled and they make our proof easier.

$$0 < L_{\min} \le \Lambda(C_{11}) \le L_{\max}, \tag{3.1}$$

$$\|\beta\|_2 \ge c_1, \text{ for some constant } c_1 > 0, \tag{3.2}$$

$$\|[C_{21}]_i\|_2 \ge c_2, \text{ for some constant } c_2 > 0, \tag{3.3}$$

where $\Lambda(\cdot)$ denotes the eigenvalues of a matrix, and $[\cdot]_i$ denotes the ith row of a matrix, i= $1, \ldots, p-q$. To simplify the proof, we consider β and C (and correspondingly C_{11}, C_{21}) as fixed, and they don't change with sample size n.

Theorem 2. *Let* (3.1)*,* (3.2)*, and* (3.3) *hold, and suppose that $p-q=1$. When IC does not hold, for the sequence of λ_1 with $\lambda_1\sqrt{q}/n \to 0$, there exists λ_2 such that EIC holds when n is very large, if and only if one of*

$$C_{21}C_{11}^{-1} sign(\beta_{(1)}) \ge 1 \textit{ and } C_{21}C_{11}^{-1}\beta_{(1)} < 0, \tag{3.4}$$

$$C_{21}C_{11}^{-1} sign(\beta_{(1)}) \le -1 \textit{ and } C_{21}C_{11}^{-1}\beta_{(1)} > 0. \tag{3.5}$$

A proof of Theorem 2 can be found in the Appendix.

When $p-q \ge 2$, it is difficult to give a necessary and sufficient condition such that EIC holds, but (3.4) and (3.5) are necessary conditions such that EIC holds. We state it as a corollary of Theorem 2.

Corollary 3. *Under* (3.1)*,* (3.2) *and* (3.3)*, suppose $p-q>1$. When IC does not hold, for a sequence of λ_1 with $\lambda_1\sqrt{q}/n \to 0$, there exists λ_2 such that EIC holds when n is very large, only if, for all $i = 1, \ldots, p-q$,*

$$[C_{21}]_i C_{11}^{-1}\beta_{(1)} < 0 \textit{ when } [C_{21}]_i C_{11}^{-1} sign(\beta_{(1)}) \geq 1, \tag{3.6}$$

$$[C_{21}]_i C_{11}^{-1}\beta_{(1)} > 0 \textit{ when } [C_{21}]_i C_{11}^{-1} sign(\beta_{(1)}) \leq -1, \tag{3.7}$$

Proof. This is a straightforward result of Theorem 2. We get this necessary result from Theorem 2 by considering only one irrelevant variable (X_{q+i}) at a time. □

Corollary 3 is useful to detect the case when neither the EIC nor the IC holds, which suggests neither the Elastic Net nor the Lasso selects the true model.

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, simulations are provided to do this comparison. When $p \gg n$, especially when $q > n$, the Lasso can select at most n variables before the model saturates (Zou and Hastie (2005)). So, if $q > n$ the lasso never selects the true model . We give an example to show that the Elastic Net might be able to solve this kind of problem. Here the R packages "lars" (Efron, Hastie and Tibshirani (2004); http://cran.r- project.org/web/packages/lars/index.html) and "elasticnet" (Zou and Hastie (2005); http://cran.r-project.org/web/packages/elasticnet/index.html) are used to compute the Lasso and the Elastic Net solution paths.

Example 1. In this example, we want to illustrate that if $p \gg n$, and EIC holds, then conditions in Corollary 2 of Theorem 1 guarantee that the Elastic Net can 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 $q\log(p-q)$, here $50 \times \log 2 = 35$. So we chose $n = 46$. The design matrix X was generated as $N(0, I_{p\times p})$. We set $\lambda_2 = 0.01$ and simulated X, that satisfied $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1} \times \mathbf{1} < 1$, $\mathbf{1}$ a column vector with all entries 1. With $\beta = [\beta_{(1)}, \beta_{(2)}]$, $\beta_{(1)}$ a $q-$vector with all entries 1 and $\beta_{(2)}$ a $(p-q)-$vector with all entries 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 is some λ_1 such that EIC holds. The true

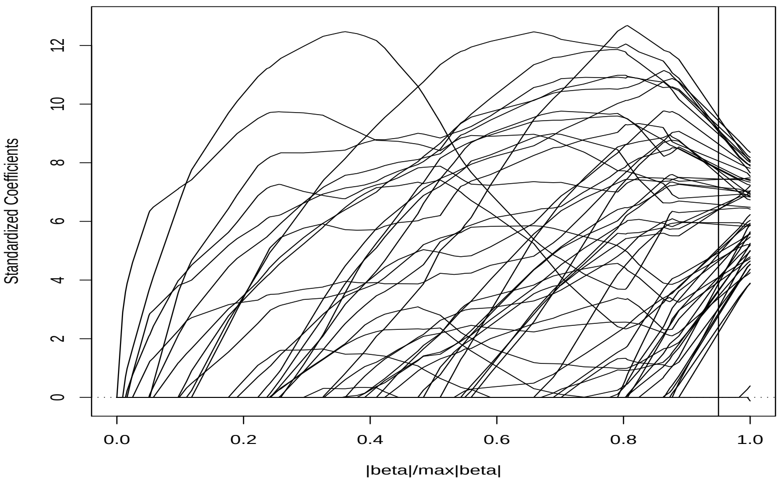

Figure 4.1: Elastic Net solution paths for $p = 52, q = 50, n = 46$. The solution corresponding to the vertical line recovers exactly the first 50 predictors.

model is: $Y = X\beta + 0.04 \times \epsilon$. The Elastic Net was applied. Solution paths are shown in Figure 4.1.

After examining the solutions, we see that the solution corresponding to the vertical line in Figure 4.1 recovered exactly the first q non-zero predictors. Theoretically, the Lasso can select at most $n = 46$ variables (Zou and Hastie (2005)) and so does not perform well on these data. Applying the Lasso, we found that it could select 45 variables at most.

Example 2. In this example, we compare the model selection performance of the Lasso and the Elastic Net by simulation. From Theorem 2, we see that when the Lasso does not select the true model consistently, the Elastic Net might. In this example, we set $p = 200, q = 10$, and took the sample size $n = 100$. $X_i' s$ were independently simulated from the standard normal. To make IC not hold, we set $X_p = \frac{1}{8}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4 + \frac{1}{2}X_5 + \frac{\sqrt{11}}{8}e$, where e was also from the standard normal distribution and independent of X. Thus, X_p was standard normal, but correlated with some of the relevant predictors. We took $\beta_1 = \beta_2 = -4, \beta_3 = \beta_4 = \beta_5 = 0.5, \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 1, \beta_i = 0,$ for all $i > 10$, and the response $Y = X\beta + 0.04\epsilon$, with ϵ standard normal and independent of all the predictors. We ran the simulation 500 times. Each time the Lasso and the Elastic Net were applied to check if they selected the true model or not. When the Elastic Net was applied, we set $\lambda_2 = 0.01$. The Elastic Net selected the true model 22.6% of the time, the Lasso 19%, the Elastic Net slightly better than the Lasso in term of model selection consistency.

Example 3. In Example 2, we saw that the Elastic Net did better than Lasso, but not by much. To understand this, we did some simulation to see how strong the necessary conditions for EIC are. We used 4 designs to study the necessary conditions. The first design has $p = 6, q = 5, n = 1000$. The first 5 relevant predictors were i.i.d. normal and the X_6 was $\frac{1}{8}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4 + \frac{1}{2}X_5 + \frac{\sqrt{11}}{8}e$,where e was also normal and independent of X. We took $\beta_1 = -4, \beta_2 = -2, \beta_3 = 0.5, \beta_4 = 0.6, \beta_5 = 0.7, \beta_6 = 0$. Under this setting, through simple calculations, we have $\Sigma_{21}\Sigma_{11}^{-1} sign(\beta) = 1.125$ and $\Sigma_{21}\Sigma_{11}^{-1}\beta = -0.1$, where $\Sigma = E(X'X/n)$ is the population covariance matrix. This means that the necessary condition for EIC applied in the population value is satisfied. Since $p - q = 1$, it is also a sufficient condition. But due to noise, $C \neq \Sigma$, and EIC might not hold. The second design is similar to the first design with the only difference that, for the second design, we sampled $\beta_i, i = 1, \ldots, q$, i.i.d. from a uniform distribution $U[-10, 10]$. In the third and the fourth designs, $p = 200, q = 10, n = 100$. $X_i's$ were normal with mean 0 and variance 1, and with correlation $\rho_{ij} = \rho$ for all pairs of X_i and X_j. $\beta_i, i = 1, \ldots, q$, were i.i.d. from a uniform distribution $U[-10, 10]$. In the third design, $\rho = 0.3$ and in the fourth design, $\rho = 0.8$. For each design, we ran 100 simulations. We use "IC YES" to denote the count of times when IC held and 'IC NO" denotes the count of times when IC did not hold. "NC YES" is used to denote the count of times when the necessary conditions held and "NC NO" the count of times when the necessary conditions did not hold. We call the 100 simulations an experiment and we did 10 experiments. After the 10 experiments, the mean of the counts and their corresponding standard errors are shown in Table 4.1.

From Table 4.1 we see that roughly speaking, when IC did not hold, it was more likely that EIC did not hold either. So when the Lasso did not select the true model, it was more likely that the Elastic Net did not select the true model either.

Design		I	II	III	IV
IC YES		0(0)	77(2.8)	31(3.4)	25(5.2)
IC NO	NC YES	94(1.8)	0(0)	0(0)	0(0)
	NC NO	6(1.8)	23(2.8)	69(3.4)	75(5.2)

Table 4.1: The mean of the counts and their corresponding standard errors for different designs. "IC YES" is used to denote the count of times when IC held, 100 simulations. "IC NO" means that IC did not hold. "NC YES" is used to denote the count of times when the necessary conditions held and "NC NO" denotes the count of times when the necessary conditions did not hold. For Design I, $p = 6, q = 5, n = 1000$; X and β were designed such that EIC held theoretically. For Design II, the design matrix X was the same as that in design I, but $\beta_i, i = 1, \ldots, q$, were i.i.d. generated randomly from a uniform distribution $U(-10, 10)$. In Design III and Design IV, $p = 200, q = 10, n = 100$ and $\beta_i, i = 1, \ldots, q$, were i.i.d. generated from a uniform distribution $U(-10, 10)$. In the last two designs, $X_i's$ were standard normal, but correlated with each other. In Design III, the correlation between each pair was 0.3 and in Design IV, it was 0.8.

5. Conclusion

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,the condition that there exist suitable sequences $\lambda_1(n)$ and $\lambda_2(n)$ such that EIC holds is 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 no longer sufficient. Some conditions on the relationships of p, q, and n are required. In this paper, for our consistency results, n should grow at a rate faster than $q \log(p-q)$. When $p > n$, in our examples, the Elastic Net performed better than the Lasso.

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 was 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 $2X^TX\hat\beta - 2X^TY + 2\lambda_2\hat\beta + \lambda_1\hat z = 0$, where

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

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

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

Condition $\mathcal{R}(X,\beta,\epsilon,\lambda_1,\lambda_2)$ holds if and only if we have $\hat\beta_{(2)} = 0$, $\text{sign}(\hat\beta_{(1)}) = \text{sign}(\beta_{(1)})$, and $|\hat z_{(2)}| \le 1$. From these conditions and using (1), we conclude that the condition $\mathcal{R}(X,\beta,\epsilon,\lambda_1,\lambda_2)$ holds if and only if

$$\begin{aligned} 2X_{(2)}^TX_{(1)}(\hat\beta_{(1)} - \beta_{(1)}) - 2X_{(2)}^T\epsilon &= -\lambda_1\hat z_{(2)}, \\ 2X_{(1)}^TX_{(1)}(\hat\beta_{(1)} - \beta_{(1)}) - 2X_{(1)}^T\epsilon + 2\lambda_2\hat\beta_{(1)} &= -\lambda_1\text{sign}(\beta_{(1)}). \end{aligned}$$

Solve for $\hat\beta_{(1)}$ and $\hat z_{(2)}$ to conclude that

$$\begin{aligned} -\lambda_1\hat z_{(2)} &= 2X_{(2)}^TX_{(1)}(X_{(1)}^TX_{(1)} + \lambda_2 I)^{-1}(X_{(1)}^T\epsilon - \frac{\lambda_1}{2}\text{sign}(\beta_{(1)}) - \lambda_2\beta_{(1)}) - 2X_{(2)}^T\epsilon, \\ \hat\beta_{(1)} &= (X_{(1)}^TX_{(1)} + \lambda_2 I)^{-1}(X_{(1)}^TX_{(1)}\beta_{(1)} + X_{(1)}^T\epsilon - \frac{\lambda_1}{2}\text{sign}(\beta_{(1)})). \end{aligned}$$

Conditions $\text{sign}(\hat\beta_{(1)}) = \text{sign}(\beta_{(1)})$ and $|\hat z_{(2)}| \le 1$ are exactly (2.1) and (2.2). □

Before proving Theorem 1, we state without proof a 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*

$$E\max_{1\le i\le n} X_i \le 3\sqrt{\log n}\max_{1\le i\le n}\sqrt{EX_i^2}. \tag{2}$$

With this lemma, we have when $n > 1$,

$$\begin{aligned} E\max_{1\le i\le n}|X_i| &\le E|X_1| + 2E\max_{1\le i\le n} X_i \\ &\le \sqrt{EX_1^2} + 6\sqrt{\log n}\max_{1\le i\le n}\sqrt{EX_i^2} \\ &\le 8\sqrt{\log n}\max_{1\le i\le n}\sqrt{EX_i^2}, \end{aligned} \tag{3}$$

where the first inequality comes from Ledoux and Talagrand (1991), the second from (2), and the third from the fact that $2\log(n) > 1$ when $n > 1$.

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:

$$\left| X_{(2)}^T X_{(1)} \left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1} \left[\mathrm{sign}(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1}\beta_{(1)}\right]\right| \le 1 - \eta.$$

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

$$\begin{aligned}
\frac{\max_{j\in S^c} V_j}{\lambda_1} &= \frac{\max_{j\in S^c} \mu_j + \widetilde{V}_j}{\lambda_1} \le (1-\eta) + \frac{1}{\lambda_1}\max_j \widetilde{V}_j, \text{ and} \\
\frac{\min_{j\in S^c} V_j}{\lambda_1} &= \frac{\min_{j\in S^c} \mu_j + \widetilde{V}_j}{\lambda_1} \ge -(1-\eta) + \frac{1}{\lambda_1}\min_j \widetilde{V}_j,
\end{aligned}$$

we need to show that

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

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 (3), we have

$$P\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_1} > \eta\right] \le \frac{E[\max_{j\in S^c}|\widetilde{V}_j|]}{\lambda_1 \eta} \le \frac{8\sqrt{\log(p-q)}}{\lambda_1\eta}\max_j \sqrt{E[\widetilde{V}_j^2]}. \quad (4)$$

Straightforward computation yields

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

Then from (4) we have

$$P\left[\frac{\max_{j \in S^c} |\widetilde{V}_j|}{\lambda_1} > \eta \right] \leq \frac{16\sigma\sqrt{n \log(p-q)}}{\lambda_1 \eta}.$$

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

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

Let $Z_i = e_i^T \left(X_{(1)}^T X_{(1)} + \lambda_2 I \right)^{-1} X_{(1)}^T \epsilon$, so that

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

Note Z_i is Gaussian with mean 0 and variance

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

From (3) we have

$$E[\max_i |Z_i|] \leq 8\sqrt{\frac{\sigma^2 \log q}{nC_{min}}}.$$

$$\begin{aligned}
&1 - P\left[\left|\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1}\left[X_{(1)}^T X_{(1)}\beta_{(1)} + X_{(1)}^T \epsilon - \frac{\lambda_1}{2}\text{sign}(\beta_{(1)})\right]\right| > 0\right] \\
&\leq P\left[\max_i |U_i| \geq \rho\right] \\
&\leq P\left[\frac{1}{\rho}\left\{\max_i |Z_i| + \frac{1}{2}\lambda_1\left\|\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1}\overrightarrow{b}\right\|_\infty\right\} \geq 1\right] \\
&\leq \frac{1}{\rho}\left\{E\left[\max_i |Z_i|\right] + \frac{1}{2}\lambda_1\left\|\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1}\overrightarrow{b}\right\|_\infty\right\} \\
&\leq \frac{1}{\rho}\left\{8\sqrt{\frac{\sigma^2 \log q}{nC_{min}}} + \frac{1}{2}\lambda_1\left\|\left(X_{(1)}^T X_{(1)} + \lambda_2 I\right)^{-1}\overrightarrow{b}\right\|_\infty\right\}.
\end{aligned}$$

Now, (b) of Theorem 1 guarantees that $P(\mathcal{M}(U)) \to 1$. □

Proof of Theorem 2.

Proof. **"If" part.** Suppose (3.4) holds, and let $\lambda_2 = -\frac{\lambda_1 C_{21}C_{11}^{-1}sign(\beta)}{2C_{21}C_{11}^{-1}\beta}$. Since $\lambda_1/n \to 0$, $\lambda_2/n \to 0$. Then we have $C_{21}(C_{11} + \frac{\lambda_2}{n}I)^{-1}\left(sign(\beta_{(1)}) + \frac{2\lambda_2}{\lambda_1}\beta_{(1)}\right) \to 0$, as $n \to \infty$. Thus EIC holds when n is very large. In the same way, when (3.5) holds, EIC holds when n is very large for some λ_2.

"Only if" part. First, we show that $\lambda_2/n \leq c$, for some constant c. Or else, we may assume $\lambda_2/n \to \infty$, as $n \to \infty$. Note that both

$$\begin{aligned}
|C_{21}(C_{11} + \lambda_2/n)^{-1}sign(\beta)| &\leq \frac{1}{L_{\min} + \frac{\lambda_2}{n}}\|C_{21}\|_2\sqrt{q} \\
&\leq \frac{n}{\lambda_2}\|C_{21}\|_2\sqrt{q},
\end{aligned}$$

$$\begin{aligned}
|C_{21}(C_{11} + \lambda_2/n)^{-1}\beta| &\geq \frac{1}{L_{\max} + \frac{\lambda_2}{n}}\|C_{21}\|_2\|\beta\|_2 \\
&\geq \frac{n}{2\lambda_2}\|C_{21}\|_2\|\beta\|_2 \\
&\geq \frac{nc_1}{2\lambda_2}\|C_{21}\|_2,
\end{aligned}$$

when n is large enough. We used the fact that $\lambda_2/n \to \infty$, and then $\lambda_2/n > L_{\max}$ when n is large enough.

From the above two inequalities, we have

$$\begin{aligned}
& |C_{21}(C_{11}+\lambda_2/n)^{-1}(sign(\beta)+\frac{2\lambda_2}{\lambda_1}\beta)| \\
\geq\ & \frac{2\lambda_2}{\lambda_1}|C_{21}(C_{11}+\lambda_2/n)^{-1}\beta| - |C_{21}(C_{11}+\lambda_2/n)^{-1}sign(\beta)| \\
\geq\ & \frac{2\lambda_2}{\lambda_1}\frac{nc_1}{2\lambda_2}\|C_{21}\|_2 - \frac{n}{\lambda_2}\|C_{21}\|_2\sqrt{q} \\
=\ & \frac{nc_1}{\lambda_1}\|C_{21}\|_2 - \frac{n}{\lambda_2}\|C_{21}\|_2\sqrt{q}.
\end{aligned}$$

Since $\lambda_1\sqrt{q}/n \to 0$, $\frac{n}{\lambda_1} \geq M\sqrt{q}$, for any $M > 0$, when n is very large. Since $\lambda_2/n \to \infty$, we have when n is large enough, $n/\lambda_2 \leq 1$. So,

$$|C_{21}(C_{11}+\lambda_2/n)^{-1}(sign(\beta)+\frac{2\lambda_2}{\lambda_1}\beta)| \geq \|C_{21}\|_2\sqrt{q}(Mc_1-1),$$

from which we see that when n is very large there is no λ_2 such that EIC holds. So, to make EIC hold, λ_2/n must be less than a constant c.

We now prove that λ_2/n must go to 0, as $n \to \infty$. Or else we may assume with no loss that $\lambda_2/n \to c_3$, for some $c_3 > 0$. Then we would have

$$|C_{21}(C_{11}+\lambda_2/n)^{-1}sign(\beta)| \ \leq\ \frac{1}{L_{\min}+c_3/2}\|C_{21}\|_2\sqrt{q},$$

$$|C_{21}(C_{11}+\lambda_2/n)^{-1}\beta| \ \geq\ \frac{1}{L_{\max}+2c_3}\|C_{21}\|_2\|\beta\|_2$$

and, from the two inequalities,

$$\begin{aligned}
& |C_{21}(C_{11}+\lambda_2/n)^{-1}(sign(\beta)+\frac{2\lambda_2}{\lambda_1}\beta)| \\
\geq\ & \frac{2\lambda_2}{\lambda_1}|C_{21}(C_{11}+\lambda_2/n)^{-1}\beta| - |C_{21}(C_{11}+\lambda_2/n)^{-1}sign(\beta)| \\
\geq\ & \frac{2\lambda_2}{\lambda_1}\frac{1}{L_{\max}+2c_3}\|C_{21}\|_2\|\beta\|_2 - \frac{1}{L_{\min}+c_3/2}\|C_{21}\|_2\sqrt{q}.
\end{aligned}$$

Note that $\frac{\lambda_2}{\lambda_1} = \frac{\lambda_2}{n}\frac{n}{\lambda_1} \geq \frac{c_3}{2}M\sqrt{q}$, for any $M > 0$, when n is large enough. We used the fact that $\frac{\lambda_2}{n} \to c_3$ and $\frac{\lambda_1\sqrt{q}}{n} \to 0$.

Now we have, for any $M > 0$,

$$\begin{aligned} & |C_{21}(C_{11} + \lambda_2/n)^{-1}(sign(\beta) + \frac{2\lambda_2}{\lambda_1}\beta)| \\ \geq \quad & \frac{c_1 c_2 c_3}{2} M \sqrt{q} \frac{1}{L_{\max} + 2c_3} - \frac{c_2}{L_{\min} + c_3/2}\sqrt{q}, \end{aligned}$$

from which we see that when n is very large, EIC does not hold.

At last, we show that if condition (3.4) or (3.5) does not hold, EIC will not hold for any λ_2. Suppose condition (3.4) does not hold. Since $\lambda_2/n \to 0$, we have $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1} sign(\beta) \to C_{21}(C_{11})^{-1} sign(\beta) > 1$ and $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\beta \to C_{21}(C_{11})^{-1}\beta > 0$ when n is very large. Therefore, $C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1} sign(\beta) + \frac{2\lambda_2}{\lambda_1} C_{21}(C_{11} + \frac{\lambda_2}{n})^{-1}\beta > 1$ when n is very large, and EIC does not hold. The proof for condition (3.5) is the same, and the proof is completed. □

References

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.

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

Meinshausen, N. and Yu, B. (2008) Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**, 246 - 270.

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

Ravikumar, P., Raskutti, G., Wainwright, M., and Yu, B. (2008). Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of l_1-regularized MLE. *NIPS.*

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, B* **58**, 267-288.

Tikhonov, A. N. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **39**, 176-179.

Wainwright, M. (2007). Sharp thresholds for high-dimensional and noisy recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE transactions on information theory* **55**, 2183 - 2202.

Yuan, M and Lin, Y. (2007). On the Nonnegative Garrote Estimator. *J. R. Statist. Soc. B* **69**, 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* **101**, 1418-1429.

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