J. R. Statist. Soc. B
67, Part 2, pp.301-320

Regularization and variable selection via the elastic net

2018 . 11 . 29 임소현

Hui Zou
(Stanford University, USA)
Trevor Hastile
(Stanford University, USA)

1. Introduction and motivation

- **OLS** often does poorly in both prediction and interpretation.

  -> need penalization techniques

- **Ridge** achieves better prediction through a bias-variance trade-off

  But, it always keeps all the predictors in the model.

- **Lasso** does both continuous shrinkage and automatic variable selection.

  But, it also has problems.

1. Introduction and motivation

Consider the following three scenarios.

(a) In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L_1-norm of the coefficients is smaller than a certain value.

(b) If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected. See Section 2.3.

(c) For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).

-> propose a new regularization technique called the **elastic net**

2.1 Definition

- *n* observations with *p* predictors

$$\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \qquad \mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}},\ j = 1, \ldots, p$$

- After a location and scale transformation, we can assume that the response is centered and the predictors are standardized
- For any fixed non-negative λ1,λ2, we define the naïve elastic net criterion

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1$$

where

$$|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2,$$

$$|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$$

2. Naïve elastic net

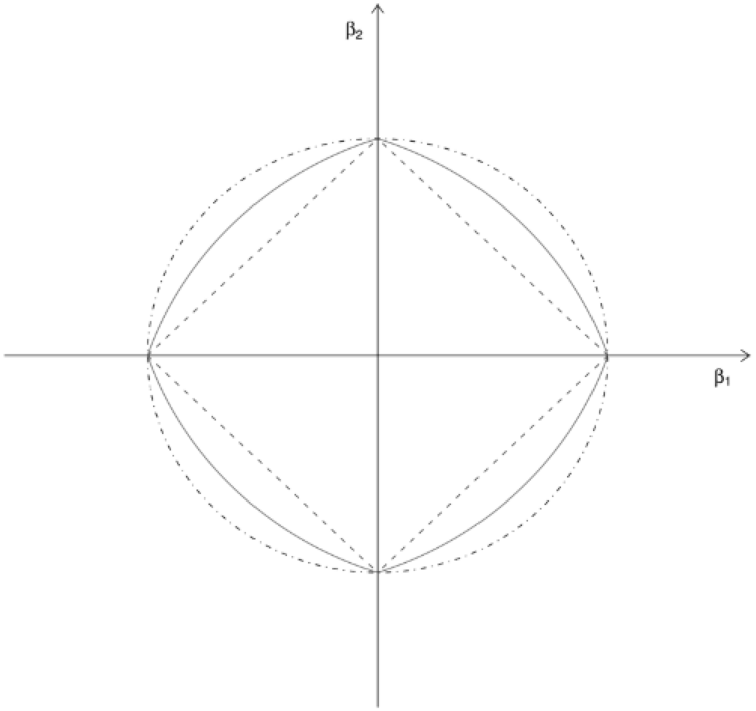
- The naïve elastic net estimator beta_hat

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \qquad \text{subject to } (1-\alpha)\,|\beta|_1 + \alpha\,|\beta|^2 \leqslant t \text{ for some } t.$$

where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$

- The function $(1-\alpha)\,|\beta|_1 + \alpha\,|\beta|^2$ is called the elastic net penalty, which is a convex combination of the lasso and ridge penalty
- alpha=0 -> lasso penalty
- alpha=1 -> ridge penalty

2.2 Solution



- Solving the naïve elastic net problem is equivalent to a lasso-type optimization problem

Lemma 1. Given data set $(\mathbf{y}, \mathbf{X})$ and (λ_1, λ_2), define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}, \qquad \mathbf{y}^*_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}.$$

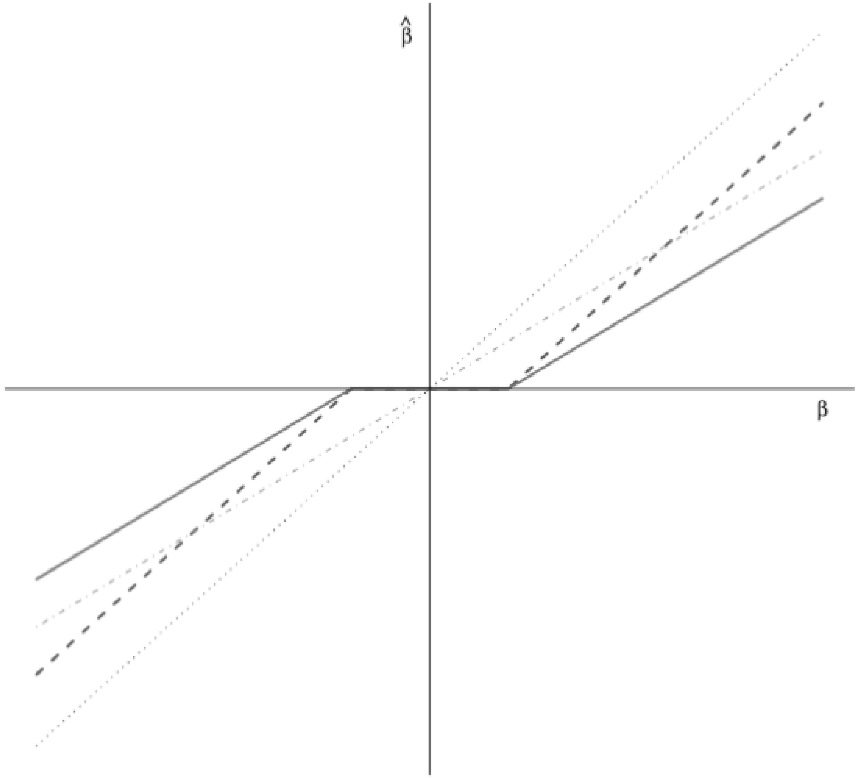
- X_star has a sample size of *(n+p)* and rank of *p*

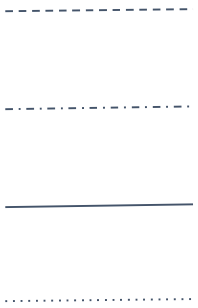
  -> the naïve elastic net can potentially select all *p* predictors

  -> overcomes the limitations of lasso in *p>n*

- can perform an automatic variable selection similar to lasso

2.2 Solution





: lasso

: ridge regression

: naïve elastic net

: OLS

- The naïve elastic net can be viewed as a two-stage procedure

  : a ridge-type direct shrinkage followed by a lasso-type thresholding

3.1 Deficiency of the naïve elastic net

- does not perform satisfactorily unless it is very close to either ridge regression or lasso
- two-stage procedure

  Stage 1. Find ridge regression coefficients

  Stage 2. Do the lasso-type shrinkage along the lasso coefficient solution path

  ⇨ incurs a double amount of shrinkage

3.2 The elastic net estimate

- the naive elastic net solves a lasso-type problem

$$\hat{\beta}^* = \arg\min_{\beta^*} |\mathbf{y}^* - \mathbf{X}^* \beta^*|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}} |\beta^*|_1$$

$$\hat{\beta}(\text{naïve elastic net}) = \{1/\sqrt{(1+\lambda_2)}\}\hat{\beta}^*$$

$$\hat{\beta}(\text{elastic net}) = \sqrt{(1+\lambda_2)}\hat{\beta}^*.$$

- the elastic net coefficient is a rescaled naive elastic net coefficient
- scaling transformation preserves the variable selection property of the elastic net and is the simplest way to undo shrinkage

4. Prostate cancer example

x	log(cancer volume)
	log(prostate weight)
	age
	amount of benign prostatic hyperplasia
	seminal vesicle invasion
	log(capsular penetration)
	Gleason score
	% Gleason score 4 or 5
y	prostate specific antigen

97 Observations of 9 variables

Train set : 67 obs -> model fitting
Test set : 30 obs -> compute their prediction mse

Method	*Parameter(s)*	*Test mean-squared error*	*Variables selected*
OLS		0.586 (0.184)	All
Ridge regression	$\lambda=1$	0.566 (0.188)	All
Lasso	$s=0.39$	0.499 (0.161)	(1,2,4,5,8)
Naïve elastic net	$\lambda=1, s=1$	0.566 (0.188)	All
Elastic net	$\lambda=1000, s=0.26$	0.381 (0.105)	(1,2,5,6,8)

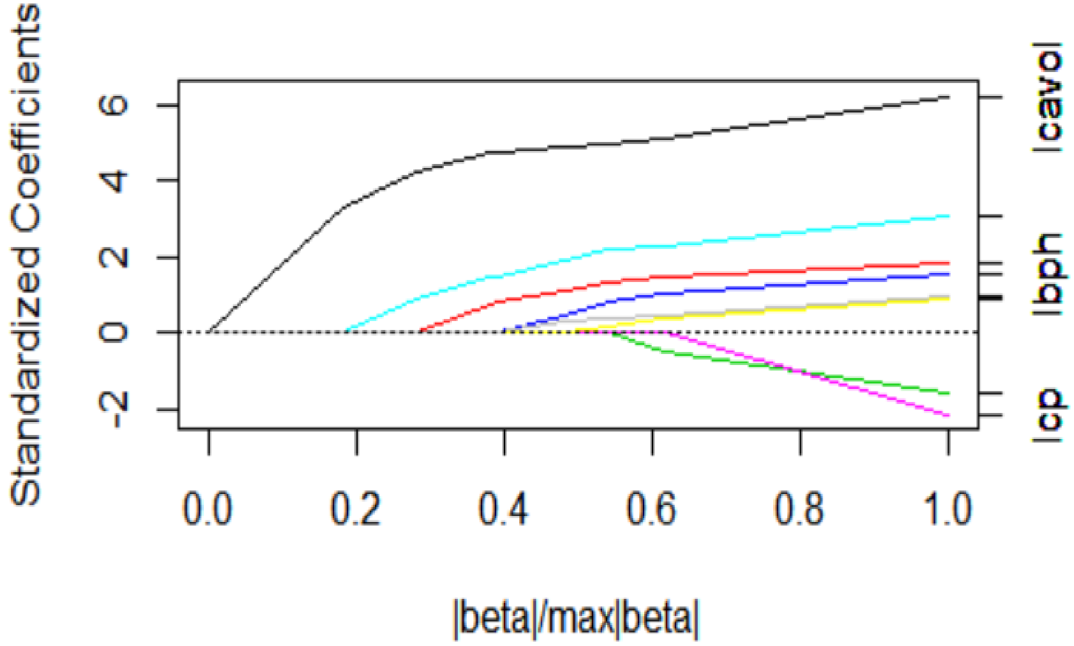
4. Prostate cancer example

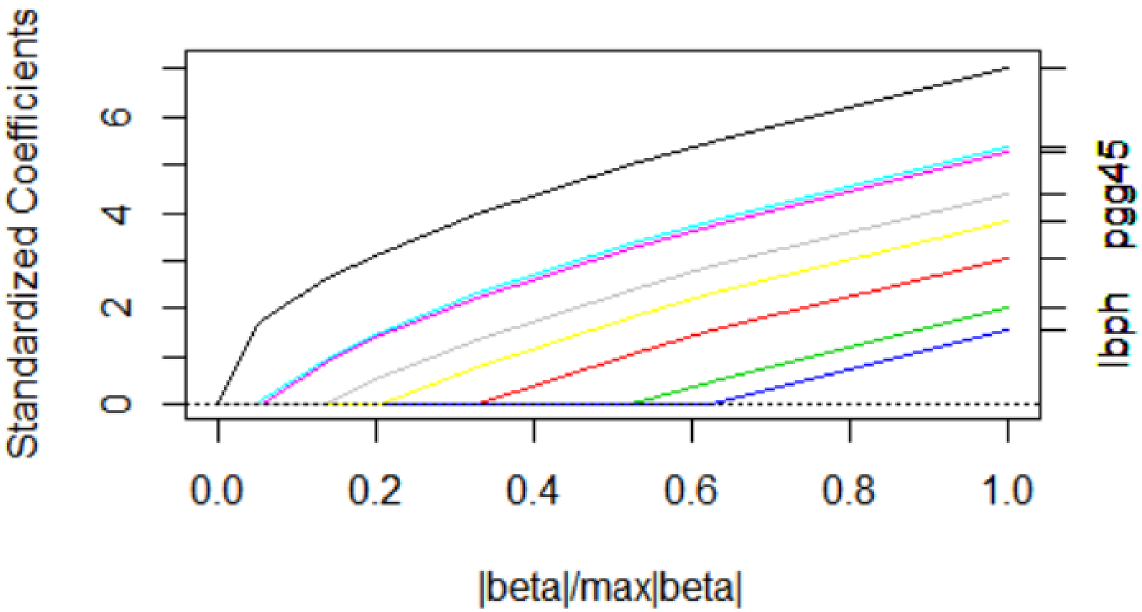
Prostate cancer data

The data in this example come from a study of prostate cancer.

```
> head(Prostate)
      lcavol  lweight age      lbph svi       lcp gleason pgg45       lpsa
1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0.4307829
2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0.1625189
3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0.1625189
4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0.1625189
5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0.3715636
6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0  0.7654678
```

4. Prostate cancer example





Method	Test mean-squared error
OLS	0.5518
Ridge regression	0.5108
Lasso	0.5353
Elastic net	0.4997

5. A simulation study

- **Purpose** the elastic net nominates the lasso in terms of prediction accuracy

  the elastic net is better variable selection procedure than the lasso
- Simulate data from the true model $\mathbf{y} = \mathbf{X}\beta + \sigma\varepsilon, \quad \varepsilon \sim N(0, 1)$

Table 3. Median number of non-zero coefficients

Method	*Results for the following examples:*			
	Example 1	*Example 2*	*Example 3*	*Example 4*
Lasso	5	6	24	11
Elastic net	6	7	27	16

5. A simulation study

(a) In example 1, we simulated 50 data sets consisting of 20/20/200 observations and eight predictors. We let $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ was set to be $\mathrm{corr}(i, j) = 0.5^{|i-j|}$.

(b) Example 2 is the same as example 1, except that $\beta_j = 0.85$ for all j.

(c) In example 3, we simulated 50 data sets consisting of 100/100/400 observations and 40 predictors. We set

$$\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$$

and $\sigma = 15$; $\mathrm{corr}(i, j) = 0.5$ for all i and j.

(d) In example 4 we simulated 50 data sets consisting of 50/50/400 observations and 40 predictors. We chose

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$$

and $\sigma = 15$. The predictors $\mathbf{X}$ were generated as follows:

$$\mathbf{x}_i = Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0, 1), \quad i = 1, \ldots, 5,$$
$$\mathbf{x}_i = Z_2 + \varepsilon_i^x, \quad Z_2 \sim N(0, 1), \quad i = 6, \ldots, 10,$$
$$\mathbf{x}_i = Z_3 + \varepsilon_i^x, \quad Z_3 \sim N(0, 1), \quad i = 11, \ldots, 15,$$
$$\mathbf{x}_i \sim N(0, 1), \quad \mathbf{x}_i \text{ independent identically distributed}, \quad i = 16, \ldots, 40,$$

5. A simulation study

- Example showing the important differences between the elastic net and the lasso.
- Let ***Z1*** and ***Z2*** be two independent $U(0,20)$ variables.
- ***y*** is generated as $N(Z1+0.1*Z2,1)$
- Suppose that

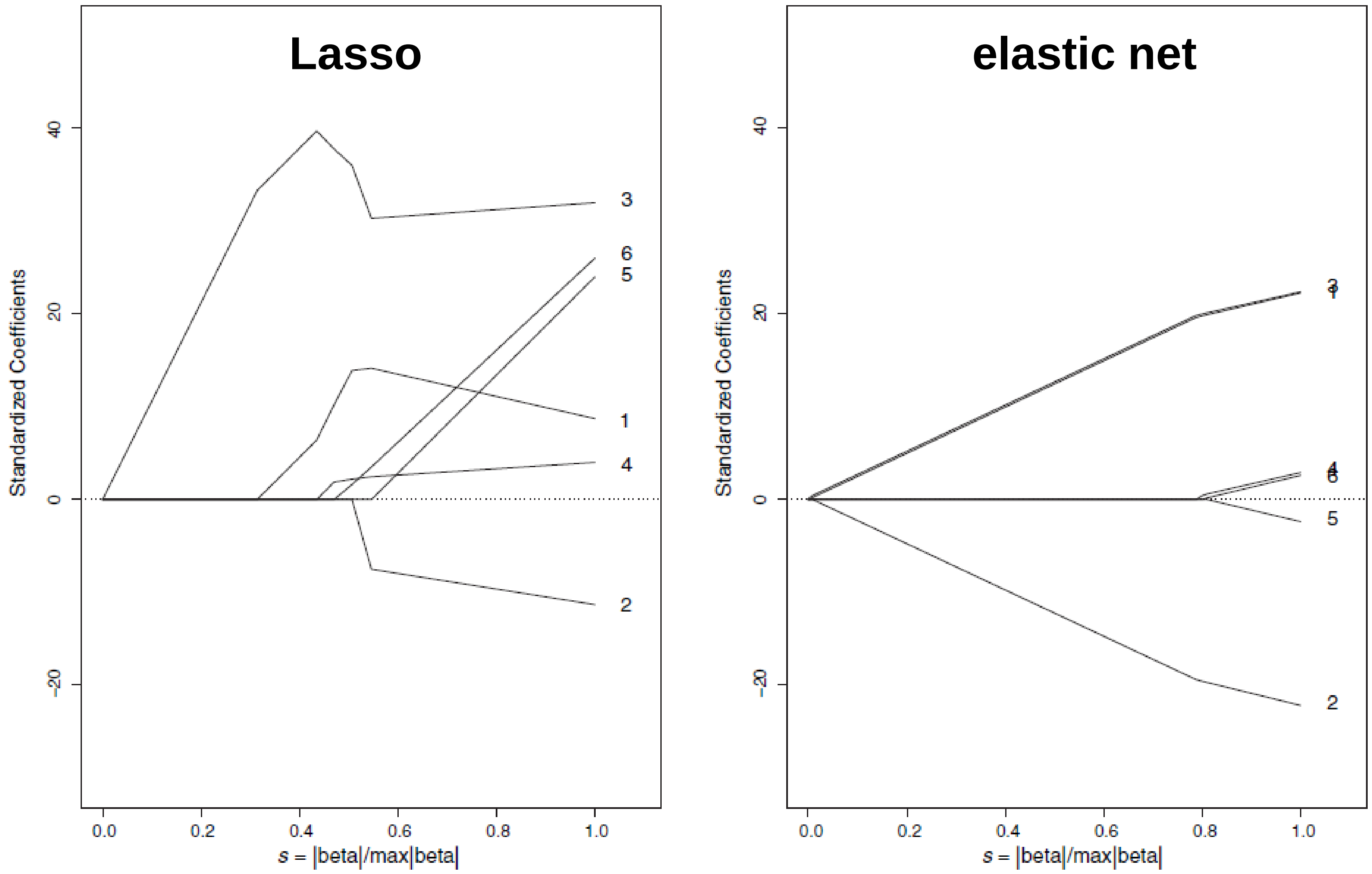
$$\mathbf{x}_1 = Z_1 + \varepsilon_1, \quad \mathbf{x}_2 = -Z_1 + \varepsilon_2, \quad \mathbf{x}_3 = Z_1 + \varepsilon_3,$$
$$\mathbf{x}_4 = Z_2 + \varepsilon_4, \quad \mathbf{x}_5 = -Z_2 + \varepsilon_5, \quad \mathbf{x}_6 = Z_2 + \varepsilon_6,$$

$$\varepsilon_i \sim N(0,1/16)$$

- within-group correlations are almost 1

  between-group correlations are almost 0

5. A simulation study



6. Conclusion

- Propose the elastic net, a new regularization and variable selection method.
- Real data show that the elastic net often outperforms the lasso.
- The elastic net is particularly useful,

  when the number of predictors is much bigger than the number of observations.
- This results offer other insights into the lasso, and ways to improve it