Elastic nets for the feature selection in linear regression models

Denis Dimitrov

Lomonosov Moscow State University

den.dimitrov@gmail.com

December 8, 2020

Overview

The Linear Regression Model

Model

$$y = X\beta^0 + \varepsilon$$

where

$y = (y_1, y_2, \ldots, y_n)^T$ is a response variable;

$\beta^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_p^0)^T$ is a vector of unknown true parameters;

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is an error vector, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_n)\ \forall i = \overline{1, n}$;

$X = (X_1, \ldots, X_p)$ is a $n \times p$ matrix with features $X_1, \ldots, X_p$.

Ordinary Least Squares

- $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Here errors ε_i for $\forall i = \overline{1, n}$ are independent
- $p(\varepsilon_1, \ldots, \varepsilon_n | \beta_1, \ldots \beta_p) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} =$

  $(\frac{1}{\sigma\sqrt{2\pi}})^n \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2} \xrightarrow{\beta} max$

  This problem is equivalent to the following one
- $\|y - X\beta\|_2^2 \xrightarrow{\beta} min$

  Solution
- $\hat{\beta} = \hat{\beta}(OLS) = (X^T X)^{-1} X^T y$

The pros and cons of OLS

- $(X^T X)^{-1}$ sometimes does not exist (e.g., in the case $p > n$)
- $\mathbb{E}\hat{\beta} = \beta^0$. So, OLS estimate is unbiased.
- $Var\hat{\beta} = \sigma^2 (X^T X)^{-1}$. So, OLS estimate have a large variance.

Which properties of estimation we would like to expect

- **Accuracy**. We wish to improve our prediction.
- **Interpretation of the model**. The goal is to determine a smaller subset in large set of predictors.
- **Group effect**. The aim is to have the coefficients, which close to each other when features are strong correlated.

Ridge Regression

- Assumption: $\beta = (\beta_1, \dots, \beta_p)$ where $\beta_i \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda_2})$ for $i \in \overline{1,p}$ (and are independent).

  We will expect that the most part of coefficients are close to zero in a sense
- $p(\beta_1, \dots \beta_p | \varepsilon_1, \dots, \varepsilon_n) \longrightarrow max$
- $p(\beta_1, \dots \beta_p | \varepsilon_1, \dots, \varepsilon_n) \sim e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i - \sum\limits_{j=1}^{p}\beta_j x_{ij})^2 - \frac{\lambda_2}{2\sigma^2}\sum\limits_{k=1}^{p}\beta_k^2}$

  This implies that
- $\|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 \xrightarrow{\beta} min$

  Solution
- $\hat{\beta} = \hat{\beta}(ridge) = (X^T X + \lambda_2 I_p)^{-1} X^T y$

Pros and cons of Ridge Regression

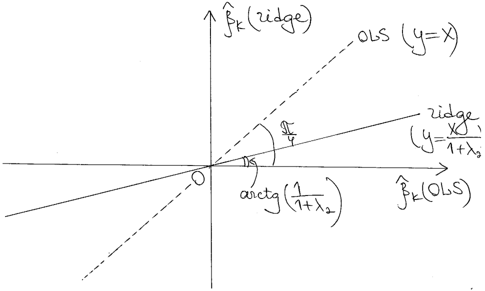
- $(X^TX + \lambda_2 I_p)^{-1}$ exists for some $\lambda_2 > 0$
- $\mathbb{E}\hat{\beta} = (X^TX + \lambda_2 I_p)^{-1}(X^TX)\beta^0$. So, RRE is biased
- $Var\hat{\beta} = \sigma^2(X^TX + \lambda_2 I_p)^{-1}(X^TX)(X^TX + \lambda_2 I_p)^{-1}$. So, one can modify the variance of RRE by changing the parameter λ_2

Simple case

Statement

If $x_1, x_2, \ldots, x_p$ is an orthonormal basis then

$$\hat{\beta}(ridge) = \frac{X^T y}{1+\lambda_2} = \frac{\hat{\beta}(OLS)}{1+\lambda_2}$$



Least Absolute Shrinkage and Selection Operator

- Assume that $\beta = (\beta_1, \dots, \beta_p)$, where $\beta_i \sim Lap(0, \frac{\lambda_1}{\sigma^2})$ for $i \in \overline{1, p}$ (and are independent).

  We will expect the most coefficients are close to zero in a sense
- $p(\beta_1, \dots \beta_p | \varepsilon_1, \dots, \varepsilon_n) \longrightarrow max$
- $p(\beta_1, \dots \beta_p | \varepsilon_1, \dots, \varepsilon_n) \sim e^{-\frac{1}{2\sigma^2} \sum\limits_{i=1}^{n} (y_i - \sum\limits_{j=1}^{p} \beta_j x_{ij})^2 - \frac{\lambda_1}{\sigma^2} \sum\limits_{k=1}^{p} |\beta_k|}$

  This implies that
- $g(\beta) = \frac{1}{2} \left\| y - X\beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_1 \xrightarrow{\beta} min$

Solution of the problem in simple case

- If $f : \mathbb{R}^p \to \mathbb{R}$ is convex then $x \in absmin \Leftrightarrow 0 \in \partial f(x)$
- $\partial g(\beta) = X^T(X\beta - y) + \lambda_1 S = \beta - X^T y + \lambda_1 S$ where $S = S_1 \times S_2 \times \times \ldots \times S_p$: $S_k = \begin{cases} [-1, 1], \text{ if } \beta_k = 0, \\ sign(\beta_k), \text{ if } \beta_k \neq 0. \end{cases}$
- When $0 \in \partial g(\beta)$?
- $\hat{\beta}_k = (\mathbb{S}_{\lambda_1}((X^T y)))_k = \begin{cases} (X^T y)_k - \lambda_1, \text{ if } (X^T y)_k > \lambda_1, \\ 0, \text{ if } |X^T y|_k \leqslant \lambda_1, \\ (X^T y)_k + \lambda_1, \text{ if } (X^T y)_k < -\lambda_1. \end{cases}$

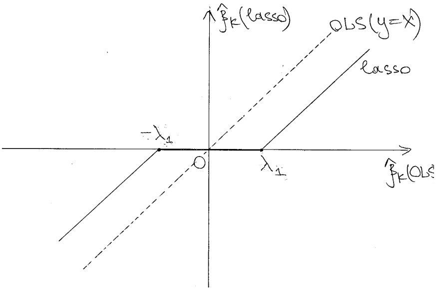
  Operator $\mathbb{S}_{\lambda_1}$ is called a **soft-thresholding operator**.

Comparison with OLS

Statement

If $x_1, x_2, \ldots, x_p$ is an orthonormal basis then

$$\hat{\beta}_k(lasso) = sign(\hat{\beta}_k(OLS)) \cdot (|\hat{\beta}_k(OLS)| - \lambda_1)_+ \;\; \forall k \in \overline{1,p}.$$

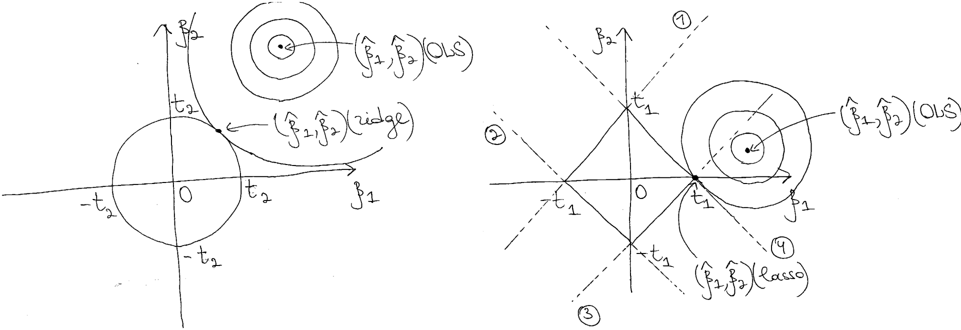


Geometrical explanation of selection properties of LASSO in case $p = 2$

In view of **Karush–Kuhn–Tucker theorem**

1. Ridge Regression: $\begin{cases} \|y - X\beta\|_2^2 \xrightarrow{\beta} min \\ \|\beta\|_2^2 \leqslant t_2^2 \end{cases}$
2. LASSO: $\begin{cases} \|y - X\beta\|_2^2 \xrightarrow{\beta} min \\ \|\beta\|_1 \leqslant t_1 \end{cases}$

Geometrical explanation of selection properties of LASSO in case $p = 2$

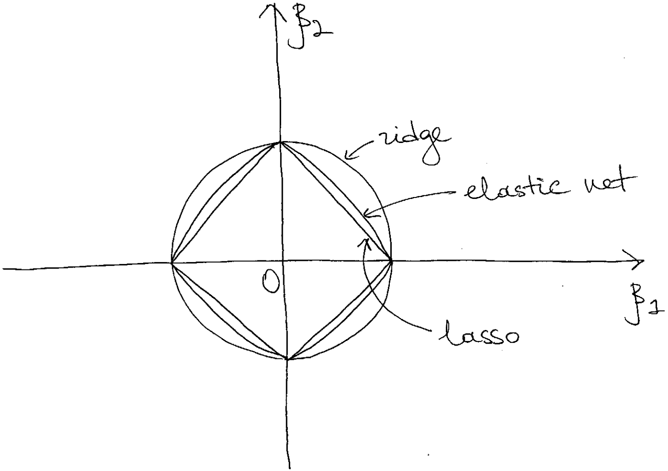


Elastic net

- *Hui Zou* and *Trevor Hastie* had introduced **Elastic net** regularization and corresponding optimization problem
$\mathbb{L}(\beta, \lambda_1, \lambda_2) = \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \xrightarrow{\beta} min$
- One can consider the problem from a different view-point
$$\begin{cases} \|y - X\beta\|_2^2 \xrightarrow{\beta} min \\ \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \leqslant t_3 \end{cases}$$

An equivalent formulation: $\begin{cases} \|y - X\beta\|_2^2 \xrightarrow{\beta} min \\ \alpha \|\beta\|_2^2 + (1-\alpha) \|\beta\|_1 \leqslant t \end{cases}$

Geometrical comparison of Ridge Regression, LASSO and Elastic Net

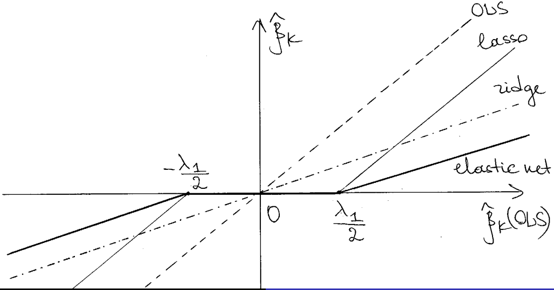


Comparison with OLS, RR and LASSO

Statement

If $x_1, x_2, \ldots, x_p$ is an orthonormal basis then

$$\hat{\beta}_k(elastic\,net) = sign(\hat{\beta}_k(OLS)) \cdot \frac{(|\hat{\beta}_k(OLS)| - \frac{\lambda_1}{2})_+}{1+\lambda_2} \quad \forall k \in \overline{1,p}.$$



'Compromise' density

Introduce more general density

$$p_{\lambda,\alpha}(\beta_i) = C(\lambda,\alpha) \cdot e^{-\lambda(\alpha|\beta_i|^2 + (1-\alpha)|\beta_i|)}$$

Convex regularization

Let us consider the case of general **convex regularization** $\hat{\beta} = \underset{\beta}{argmin}\left(\|y - X\beta\|_2^2 + \lambda\, J(\beta) \right)$ where $J(\beta)$ is convex and symmetric function and $\lambda > 0$.

Convex regularization

Theorem (Group effect)

Suppose that $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ *is a solution of the convex problem and* $x^i = x^j$, *for some* $i, j \in \overline{1, p}$.
1) If a function $J(\beta)$ *is strictly convex then* $\hat{\beta}_i = \hat{\beta}_j \ \forall \lambda > 0$.
2) If $J(\beta) = |\beta|_1$ *(it means that we are dealing with LASSO regularization) then* $\hat{\beta}_i \cdot \hat{\beta}_j \geqslant 0$ *and one could find another minimizer* $\hat{\beta}^* = (\hat{\beta}^*_1, \ldots, \hat{\beta}^*_p)^T$ *of* $\|y - X\beta\|_2^2 + \lambda\, J(\beta)$*:*

$$\hat{\beta}^*_k = \begin{cases} \hat{\beta}_k, & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot s, & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s), & \text{if } k = j, \end{cases}$$

for each $s \in [0, 1]$.

Elastic Net regularization

Theorem (Group effect of Elastic Net)

Assume that we have a standard sample (y, X). Let $\hat{\beta}(\lambda_1, \lambda_2)$ be a minimizer in linear regression problem with Elastic Net regularization. Also we have assumed that $\hat{\beta}_i \cdot \hat{\beta}_j > 0$ (otherwise one can consider $-x^i$ instead of x^i).
Let us define

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{|\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)|}{\|y\|_2}.$$

Then

$$D_{\lambda_1,\lambda_2}(i,j) \leqslant \frac{1}{\lambda_2}\sqrt{2(1-\rho)}$$

where $\rho = (x^i, x^j)$.

Auxiliary fact

Lemma

1) Let $\hat{\beta}^{enet} = \underset{\beta}{argmin}\left(\mathbb{L}^{enet}(\beta, \lambda_1, \lambda_2)\right)$ *where* $\mathbb{L}^{enet}(\lambda_1, \lambda_2) = \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$.
2) All eigenvalues of matrix $\frac{1}{n}X^T X$ *are bounded, so* $0 \leqslant b \leqslant \lambda_{min} \leqslant \lambda_{max} \leqslant B$.
Then the following inequality is valid

$$E(\|\hat{\beta}^{enet} - \beta^0\|_2^2) \leqslant 4\frac{\lambda_2^2 \left\|\beta^0\right\|_2^2 + Bpn\sigma^2 + \lambda_1^2 p}{(bn + \lambda_2)^2}.$$

Consistency of Elastic Net estimation

Theorem (Consistency)

1) Let the conditions of the previous lemma be satisfied.
2) Suppose that $\lim_{n \to +\infty} \frac{p}{n} = 0$ *then* $\hat{\beta}^{enet}$ *is a consistent estimate of* β^0.

Conclusions

In the work:

- we considered Ridge Regression, LASSO and discussed their advantages and disadvantages. To eliminate the disadvantages the Elastic net regularization was introduced.
- the application of the Elastic Net regularization to feature selection problems is provided.
- the consistency of the Elastic net estimates was established.

Thank you for attention!