**Econometrics** **March 4, 2023**

# Topic 11: Lasso And Beyond: Convex Learning

*by Sai Zhang*

**Key points**:

**Disclaimer**:

## 11.1 Lasso

Lasso (Least absolute Shrinkage and Selection Operator), proposed by Tibshirani (1996), aims to minimize the **SSR *(sum of residual squares)*** subject to the $L1-$**norm *(sum of the absolute value)*** of the coefficients being less than a constant.

### 11.1.1 Set up

For data $(\mathbf{x}_i, y_i)_{i=1}^n$, where

- $y_i$ is the outcome for individual $i$
- $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$ is the $p \times 1$ vector of predictors

Then the Lasso estimator $\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right)$ is defined as

$$\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right) = \arg\min_{\alpha, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \qquad \text{s.t. } \sum_{j=1}^{p} |\beta_j| \leq t$$

for the $n \times 1$ response vector $\mathbf{y} = (y_1, \cdots, y_n)'$, the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$ where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$ is a $p \times 1$ vector. Here $\hat{\alpha} = \overline{y}$, w.l.o.g., let $\overline{y} = 0$ and omit $\alpha$ for simplicity.

In matrix form, we have

- constrained form:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \qquad \text{s.t. } \|\boldsymbol{\beta}\|_1 \leq t$$

- unconstrained form:

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

where the regularization parameter $\lambda \geq 0$:

- $\lambda \to \infty$: $\hat{\boldsymbol{\beta}}_{lasso} = \mathbf{0}$
- $\lambda = 0$: $\hat{\boldsymbol{\beta}}_{lasso} \to \hat{\boldsymbol{\beta}}_{OLS}$

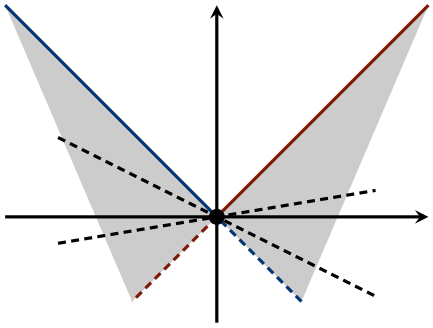
### 11.1.2 Solving Lasso

Lasso is essentially a quadratic optimization problem. Hence, the solution is given by taking the derivative (of the unconstrainted question) and set it equal to 0

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}\left(\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\right)=0$$

$$\xrightarrow{\text{KKT condition}} \quad \frac{1}{n}\underbrace{\mathbf{X}'}_{p\times n}\underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}_{=\boldsymbol{\epsilon},n\times 1}=\lambda\begin{cases}\text{sign}\left(\beta_j\right), & \beta_j\neq 0\\ [-1,1], & \beta_j=0\end{cases}$$

this result follows the fact the L-1 norm $\|\boldsymbol{\beta}\|$ is piecewise linear (convex)[1]:



L1-norm (1-dimension)

For each component of the vector of the L-1 norm $f(\beta_j)=|\beta_j|$, we have:
- $\beta_j>0$: $f'(\beta_j)=1$
- $\beta_j<0$: $f'(\beta_j)=-1$
- $\beta_j=0$: $\mathrm{d}f\in[-1,1]$ (shaded area)

which gives the results stated above.

Take another look at this result

**Proposition 11.1.1: Lasso Parameter Selection Rule**

$$\frac{1}{n}\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)=\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}=\lambda\begin{cases}\text{sign}\left(\beta_j\right), & \beta_j\neq 0\\ [-1,1], & \beta_j=0\end{cases}$$

which gives a parameter selection criterion: for $\beta_j\neq 0$, $\text{sign}(\beta_j)$ **must agree** with $\text{Corr}(\mathbf{x}_j,\boldsymbol{\epsilon})$, the correlation between the $j$-th variable $\mathbf{x}_j$ and (full-model) residuals $\boldsymbol{\epsilon}=\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$.

### 11.1.3 Algorithm: LARS

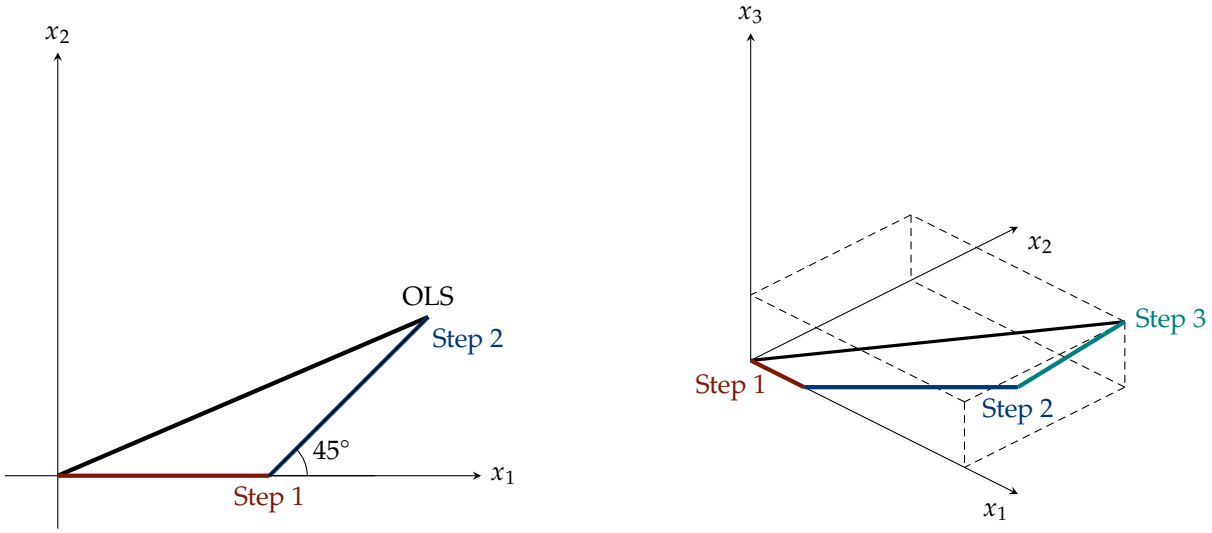
Mathematically, Lasso is quite intuitive, but computationally, it can be quite consuming. Efron et al. (2004) propose an algorithm that takes steps from a all-0 model to the biggest model (OLS), that is, **Least Angle Regression (LARS)**.

#### Intuition

The basic intuition of LARS is quite straight-forward: covariates are considered from the ***highest*** corerlation with $\mathbf{y}$ (*smallest* angle from $\mathbf{y}$) to the ***least*** correlated one (*largest* angle from $\mathbf{y}$) (illustrated below).

And the steps of the LARS algorithm are

[1] KKT condition gives the analytical optimization rule for **convex** function.




1 start with the null model $\hat{\boldsymbol{\beta}} = \mathbf{0}$: $\hat{\boldsymbol{\mu}} = \mathbf{X}'\mathbf{0} = \mathbf{0}$
2 calculate residual vector $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}$
3 determine the correlation vector between $\mathbf{r}$ and each parameter $\mathbf{x}_j, \forall j = 1, \cdots, p$: $\mathbf{X}'\mathbf{r}$
4 pick the largest correlation $\mathbf{x}^*_{\text{step1,1}}$, increase its $\hat{\beta}$ to the point where its correlation with $\mathbf{r}$ will be **equal** with that of another parameter $\mathbf{x}^*_{\text{step1,2}}$
5 next, increase the $\hat{\beta}$ for both $\mathbf{x}^*_{\text{step1,1}}, \mathbf{x}^*_{\text{step1,2}}$ in an **equiangular** direction between these two, until a third parameter becomes equally important

And keep looping this way, until all the predictors enter the model and eventually $\mathbf{X}'\mathbf{r} = \mathbf{0}$

**Properties of LARS**

LARS has several properties:

- geometrically travels in the direction of **equal** angle to all active covariates
- assume all covariates are independent
- computationally quick: only take $m$ steps, where $m$ is the number of parameters being considered

And it is in between 2 classic model-selection methods: **Forward Selection** and **Stagewise Selection**:

- **Forward Selection**
  - for $\mathbf{y}$, select the most correlated $\mathbf{x}_{j_1}$
  - regress $\mathbf{x}_{j_1}$ on $\mathbf{y}$, get the residuals
  - select the most correlated $\mathbf{x}_{j_2}$ with the redisual of $\mathbf{y}$ net of $\mathbf{x}_{j_1}$

  looping this, for a $k-$parameter linear model, it takes $k$ steps. Forward Selection is an aggressive fitting technique, can be overly greedy (some important predictors may be eliminated due to correlation with already selected variables).
- **Forward Stagewise**
  - also begin with $\hat{\boldsymbol{\mu}} = 0$
  - for a current Stagewise estimate $\hat{\boldsymbol{\mu}}$, the current residual vector is then $\mathbf{y} - \hat{\boldsymbol{\mu}}$, its correlation with $\mathbf{X}$ is then $\mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) \equiv \hat{\mathbf{c}}$

– next, heavily computational, go in the direction of the greatest current correlation, but by only a **small** step

$$\hat{j} = \arg\max|\hat{c}_j|,\ \hat{\boldsymbol{\mu}} \to \hat{\boldsymbol{\mu}} + \epsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}}$$

here, $\epsilon$ is a **small** constant, hence avoiding the greediness of Forward Selection, at a cost of computational efficiency[2].

LARS avoids the over-greediness of Forward Selection and computational heaviness of Forward Stagewise.

### 11.1.4 From LARS to Lasso

The Lasso algorithm is built upon LARS, with the constraint from the mathematical condition of Proposition 11.1.1: $\text{sign}(\beta_j)$ **must agree** with $\text{Corr}(\mathbf{x}_j, \boldsymbol{\epsilon})$.

**Theorem 11.1.2: Lasso Modification Condition**

If $\tilde{\gamma} < \hat{\gamma}$, stop the onging LARS step at $\gamma = \tilde{\gamma}$ and remove $j$ from the calculation of the next equiangular direction, where

- the path at any LARS step is

$$\boldsymbol{\beta}(\gamma),\ \beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j$$

  $\hat{d}_j$ specifies the **direction** to take the $j$-th component, $\gamma$ is **how far** to travel in the direction of $\hat{d}_j$ before adding in a new covariate
- $\hat{\gamma}$ represents the smallest **positive** value of $\gamma$ s.t. some new covariate joins the active set (the set of covariates used on path)
- $\tilde{\gamma}$ represents the first time $\beta_j(\gamma)$ **changes signs**.

The key point of 11.1.2 is that Lasso does **NOT** allow the $\hat{\beta}_j$ to change signs, if it changes sign, it will be substracted from the active set. Now, from this point of view, we can compare the 3 algorithms:

| | |
|---|---|
| LARS | no sign restrictions |
| Lasso | $\hat{\beta}_j$ agrees in sign with $\hat{c}_j$ |
| Stagewise | successive differences of $\hat{\beta}_j$ agree in sign with the current correlation $\hat{c} = \mathbf{x}_j'(\mathbf{y} - \hat{\boldsymbol{\mu}})$ |

Again, LARS requires the least steps but is most greedy, Stagewise is computationally consuming but robust. Lasso is in between.

## 11.2 Consistency of Lasso

Next, we want to establish the consistency of Lasso, by showing that Lasso selects exactly the relevant covariates asymptotically. We do this in 2 steps:

- show that Lasso at least captures all the relevant covariates
- asymptotically, under some conditions, Lasso selects exactly all the relevant covariates, not more

[2]Forward Selection is essentially choosing $\epsilon = |\hat{c}_{\hat{j}}|$

### 11.2.1 Overestimation

First, Lasso tends to select a superset of the relevant covariates.

Define the true relevant set Lasso selection estimation $\hat{S}_0$ aim to select as

$$S_0 = \left\{ j : \beta_j^0 \neq 0, j \neq 1, \cdots, p \right\}$$

and for some $C > 0$, define the relevant set w.r.t. $C$ as

$$S_0^{\text{relevant}(C)} = \left\{ j : |\beta_j^0| \geq C, j = 1, \cdots, p \right\}$$

then we have

**Theorem 11.2.1: Lasso Overestimation Condition**

$\forall 0 < C < \infty$

$$\mathbb{P}\left[ \hat{S}_0(\lambda) \supset S_0^{\text{relevant}(C)} \right] \xrightarrow{n \to \infty} 1$$

## Consistency

The consistency of Lasso is established by Meinshausen and Bühlmann (2006) as

**Theorem 11.2.2: Consistency of Lasso**

For a suitable $\lambda = \lambda_n \gg \sqrt{s_0 \log(p)/n}$, Lasso is consistent, i.e.

$$\mathbb{P}\left[ \hat{S}(\lambda) = S_0 \right] \xrightarrow{n \to \infty} 1$$

if and only if it satisfies the 2 properties:

- $\beta-$min condition (unselected coefficients non-trivial): $\inf_{j \in S_0^c} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$
- **irrepresentable** condition: $\mathbf{X}$ should **NOT** exhibit too strong a degree of linear dependence w.r.t. the selected covariates, that is, for some $0 < \theta < 1$

$$\left\| \hat{\Sigma}_{2,1} \hat{\Sigma}_{1,1}^{-1} \text{sign}\left( \beta_1^0, \cdots, \beta_{s_0}^0 \right) \right\|_\infty \leq \theta$$

**discussion on the irrepresentable condition** denote $\hat{\Sigma} = n^{-1}\mathbf{X}\mathbf{X}'$, and let the active set $S_0 = \left\{ j : \beta_j^0 \neq 0 \right\} = \{1, \cdots, s_0\}$ consists of the first $s_0$ variables, let

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1} & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{2,1} & \hat{\Sigma}_{2,2} \end{pmatrix}$$

where $\hat{\Sigma}_{1,1}$ is a $s_0 \times s_0$ var-cov matrix of the active variables $\mathbf{X}_1$, $\hat{\Sigma}_{2,2}$ is a $(p - s_0) \times (p - s_0)$ cov-var matrix of the other variables $\mathbf{X}_2$[3], then for a Lasso estimation $\hat{\boldsymbol{\beta}}$ that assign non-zero coefficients **only** to $\mathbf{X}_1$, we can,

[3]Here, $\mathbf{X}_1$ is $n \times s_0$, $\mathbf{X}_2$ is $n \times p - s_0$

following the Lasso result in Proposition 11.1.1, have

$$\begin{aligned} \frac{1}{n}\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) &= \lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1) \qquad \Rightarrow \frac{1}{n}\mathbf{X}_1'\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1) \\ \|\frac{1}{n}\mathbf{X}_2'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)\|_\infty &\le \lambda \qquad \Rightarrow \|\frac{1}{n}\mathbf{X}_2'\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right)\|_\infty \le \lambda \end{aligned} \tag{11.1}$$

Now, let's assume $\text{supp}(\hat{\boldsymbol{\beta}}) = \text{supp}(\boldsymbol{\beta}) = S_0$, then the true model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{bmatrix}+\boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1+\boldsymbol{\epsilon}$$

Put this back to the results in Equation 11.1, we have

- for the selected covariates $\mathbf{X}_1$

$$\begin{aligned} &\frac{1}{n}\mathbf{X}_1'\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1) \\ \Rightarrow &\frac{1}{n}\mathbf{X}_1'\left(\mathbf{X}_1\boldsymbol{\beta}_1+\boldsymbol{\epsilon}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1) \\ \Rightarrow &\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1\right) = \frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}-\lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1) \\ \Rightarrow &\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1 = \left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\left[\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}-\lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1)\right] \\ \Rightarrow &\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1+\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}-\underbrace{\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1)}_{\text{L1-norm regularization}} \end{aligned}$$

- for the non-selected covariates $\mathbf{X}_2$[4]

$$\begin{aligned} &\left\|\frac{1}{n}\mathbf{X}_2'\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right)\right\|_\infty \le \lambda \\ \Rightarrow &\left\|\frac{1}{n}\mathbf{X}_2'\left[\mathbf{X}_1\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1\right)-\boldsymbol{\epsilon}\right]\right\|_\infty \le \lambda \\ \Rightarrow &\left\|\frac{1}{n}\mathbf{X}_2'\left\{\mathbf{X}_1\left[\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}-\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\lambda\cdot\text{sign}(\hat{\boldsymbol{\beta}}_1)\right]-\boldsymbol{\epsilon}\right\}\right\|_\infty \le \lambda \\ \xRightarrow{\text{assume sign consistency}} &\left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left[\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}-\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\lambda\cdot\text{sign}(\boldsymbol{\beta}_1)\right]-\frac{1}{n}\mathbf{X}_2'\boldsymbol{\epsilon}\right\|_\infty \le \lambda \end{aligned}$$

  if we assume $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_n)$, i.e., $\epsilon_i$ i.i.d. $\sim\mathcal{N}(0,\sigma^2)$, and for each variable in the design matrix, we also assume standard normal $\mathbf{x}_j\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)$ (think this as normalizing each variable), then we have

$$\frac{1}{n}\mathbf{x}_j'\boldsymbol{\epsilon}\sim\mathcal{N}\left(0,\frac{\sigma^2}{n}\right),\forall j=1,\cdots,p$$

[4]Here, additionally assume **sign consistency**:

$$\begin{aligned} \text{sign}(\hat{\beta}_j) = \beta_j \neq 0, \forall j\in S_0 \qquad &\text{selected coviates} \\ \text{sign}(\hat{\beta}_j) = \beta_j = 0, \forall j\in S_0^C \qquad &\text{non-selected coviates} \end{aligned}$$

and also $\max_{1\le j\le p}\mathbf{x}_j \sim \sqrt{2\log p}$. This gives

$$\left\|\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty \sim \sqrt{2\log p}\cdot\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{2\log p}{n}}\sigma$$

hence, it can be bounded by $\lambda = \sqrt{\frac{C\log p}{n}}\sigma$, where the constant $C \ge 2$, then with large probability

$$\left\|\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty \le \frac{1}{3}\lambda \Rightarrow \begin{matrix}\left\|\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}\right\|_\infty \le \frac{1}{3}\lambda \\ \left\|\frac{1}{n}\mathbf{X}_2'\boldsymbol{\epsilon}\right\|_\infty \le \frac{1}{3}\lambda\end{matrix}$$

now go back to the condition of non-selected covariates $\mathbf{X}_2$

$$\begin{aligned}
&\left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left[\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon} - \left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\lambda\cdot\operatorname{sign}(\boldsymbol{\beta}_1)\right] - \frac{1}{n}\mathbf{X}_2'\boldsymbol{\epsilon}\right\|_\infty \\
\le&\left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}\right\|_\infty + \left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\operatorname{sign}(\boldsymbol{\beta}_1)\right\|_\infty\cdot\lambda + \left\|\frac{1}{n}\mathbf{X}_2'\boldsymbol{\epsilon}\right\|_\infty \\
\le&\left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\cdot\underbrace{\left\|\frac{1}{n}\mathbf{X}_1'\boldsymbol{\epsilon}\right\|_\infty}_{\le\frac{1}{3}\lambda} + \left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\cdot\underbrace{\left\|\operatorname{sign}(\boldsymbol{\beta}_1)\right\|_\infty}_{=1}\cdot\lambda + \underbrace{\left\|\frac{1}{n}\mathbf{X}_2'\boldsymbol{\epsilon}\right\|_\infty}_{\le\frac{1}{3}\lambda} \\
\le&\left\|\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\cdot\frac{4}{3}\lambda + \frac{1}{3}\lambda \le \lambda
\end{aligned}$$

for the last part ($\le \lambda$) to stand, a necessary condition is $\left\|\mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty \le \frac{1}{2}$, or more generally, the **irrepresentable condition**:

$$\left\|\mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\cdot\operatorname{sign}(\boldsymbol{\beta}_1)\right\| \le \left\|\mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\cdot\left\|\operatorname{sign}(\boldsymbol{\beta}_1)\right\| = \left\|\hat{\Sigma}_{2,1}\hat{\Sigma}_{1,1}^{-1}\operatorname{sign}\left(\beta_1^0,\cdots,\beta_{s_0}^0\right)\right\|_\infty \le \theta$$

The idea of ***irrepresentable*** is that the **maximum** correlation between noise features and important variables needs to be **bounded**.

### 11.2.2 Oracle

Consistency is part of the requirements of an oracle procedure

**Definition 11.2.3: Oracle Property**

For a fitting procedure $\delta$, and the estimation $\hat{\beta}(\delta)$, then if $\delta$ is an oracle procedure if $\hat{\beta}(\delta)$ asymptotically has the following properties

- **consistency** (identifying right subset model): $\left\{j:\hat{\beta}_j\neq 0\right\} = S_0$
- **optimal estimation rate** (asymptotically normal): $\sqrt{n}\left(\beta(\delta)_{S_0} - \beta_{S_0}^0\right)\xrightarrow{d}\mathcal{N}(0,\Sigma_0)$, where $\Sigma_0$ is the *true* subset covariance matrix

The oracle property gives consistency, and asymptotic normality. Ideally, this is what you need.

## 11.3 Variants of Lasso

Lasso is great, but there is still space for improvement

- for high dimension $p > n$ cases, lasso selects **at most** $n$ variables
- for a group of variables with large **pairwise correlation**, then the lasso tends to select **only one**
- for $n > p$ cases, if there are high correlations between predictors, lasso is dominated by ridge regression

Here are some popular variants people came up with to tackle these issues.

### 11.3.1 Adaptive Lasso

Adaptive Lasso replaces the $L_1-$penalty for a **re-weighted** version:

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left( \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|^{\gamma}} \right)$$

this is a 2-step procedure:

- Step 1: run an initial estimation (could be OLS or Lasso) as $\hat{\boldsymbol{\beta}}_{init}$, if $\hat{\beta}_{init,j} = 0$, then $\hat{\beta}_{adapt,j} = 0$
- Step 2: run the reweighted estimation, which is still an $L1-$penalization, still a **convex** optimization problem (allowing the same Lasso algorithm to solve)

**Theorem 11.3.1: Properties of Adaptive Lasso**

The adaptive Lasso has the following properties

- If $|\hat{\beta}_{init,j}|$ is large (*important* features), then the adaptive Lasso uses a small penalty
- The adaptive lasso enjoys the oracle properties
  - **consistency**: $\lim_n \mathbb{P}\left(\hat{S}(\lambda) = S_0\right) = 1$
  - **asymptotic normality**: $\sqrt{n}\left(\beta_{S_0} - \beta^*_{S_0}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \times C^{-1}_{1,1})$

  if $\lambda_n$ is chosen properly, i.e., $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Notice: the adaptive lasso's consistency **DOES NOT** require the irrepresentable condition.

### 11.3.2 Naive Elastic Net

The idea is to do a convex combination of Lasso and Ridge: $\forall \lambda_1, \lambda_2 \geq 0$, the naive elastic net criterion is

$$\mathcal{L}(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \underbrace{\|\boldsymbol{\beta}\|_1}_{=\sum_{j=1}^{p}|\beta_j|} + \lambda_2 \underbrace{\|\boldsymbol{\beta}\|_2^2}_{=\sum_{j=1}^{p}\beta_j^2}$$

Now, notice that both $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{\beta}\|_2^2$ are both **quadratic**, hence, we can reframe the qeustion as a lasso-type optimization, with an augmented data $(\mathbf{X}^*, \mathbf{y}^*)$ where

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

then we have

**Definition 11.3.2: Naive elastic net solution**

Let $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, then the naive elastic net criterion can be rewritten as

$$\mathcal{L}(\gamma, \boldsymbol{\beta}) = \mathcal{L}(\gamma, \boldsymbol{\beta}^*) = \|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \gamma\|\boldsymbol{\beta}^*\|_1$$

which gives use the estimation for the augmented data

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \mathcal{L}(\gamma, \boldsymbol{\beta}^*)$$

and the estimation for the true model of the original data $(\mathbf{X}, \mathbf{y})$ is

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$$

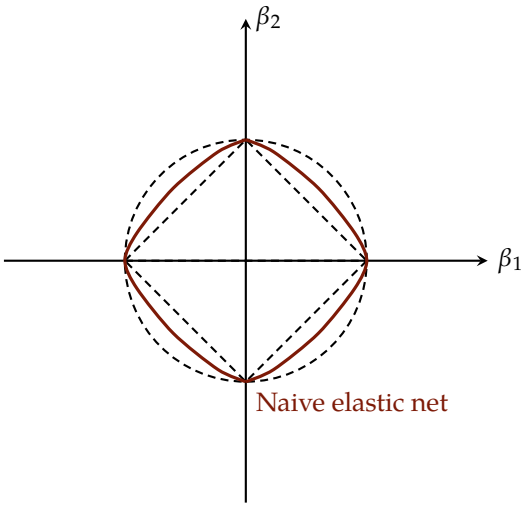
Naive elastic net enjoys 2 advantages:

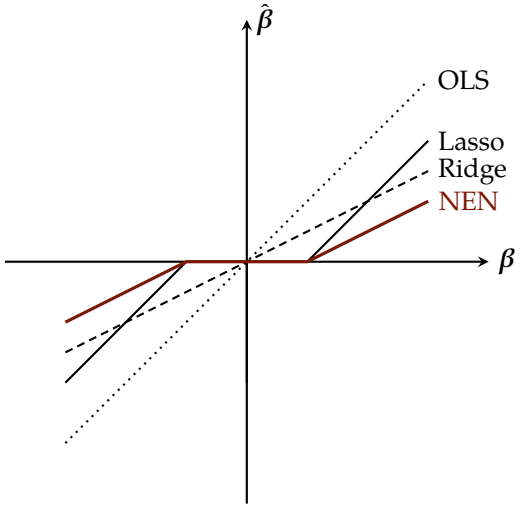
- The estimation is done on the augemented data $(\mathbf{X}^*, \mathbf{y}^*)$, and now $\mathbf{X}^*$ is $(n+p)\times p$, therefore, $\text{rank}(\mathbf{X}^*) = p$, hence all $p$ predictors **can** be selected, instead of just $n$, like in Lasso.
- Computationally, it is similarly efficient as Lasso

If we further assume an orthogonal design ($\mathbf{X}'\mathbf{X} = \mathbf{I}_p$), we can easily compare the 3 estimations

$$\begin{aligned}
\hat{\beta}_j \text{ (naive elastic net)} &= \frac{\left(\left|\hat{\beta}_j^{OLS}\right| - \frac{\lambda_1}{2}\right)_+}{1+\lambda_2} \cdot \text{sign}\left[\hat{\beta}_j^{OLS}\right] \\
\hat{\beta}_j \text{ (ridge)} &= \frac{\hat{\beta}_j^{OLS}}{1+\lambda_2} \\
\hat{\beta}_j \text{ (lasso)} &= \left(\left|\hat{\beta}_j^{OLS}\right| - \frac{\lambda_1}{2}\right)_+ \cdot \text{sign}\left[\hat{\beta}_j^{OLS}\right]
\end{aligned}$$

This comparison is illustrated graphically below.



Now consider another problem: ***grouping effect***, that is, for a generic penalization method

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda J(\boldsymbol{\beta})$$

where $J(\cdot)$ is positive valued for $\boldsymbol{\beta} \neq 0$, then for a group of highly correlated variables, the regression coeffeicients tend to be **equal**.

**Lemma 11.3.3: Grouping effect**

Assume $\mathbf{x}_i = \mathbf{x}_j, i, j \in \{1, \cdots, p\}$, then

- If $J(\cdot)$ is **strictly convex**, then $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$
- If $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$, then $\hat{\beta}_1, \hat{\beta}_2 \geq 0$ and $\hat{\boldsymbol{\beta}}^*$ is another minimizer of the generic penalization function, where

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

for any $s \in [0, 1]$

As shown in Lemma 11.3.3, there is a clear distinction between ***strictly convex*** penalty functions and the Lasso penalty (not strictly convex). For the naive elastic net penalty, $\lambda_2 > 0$ gives strict convexity.

And now, to take this grouping effect into consideration, we want: for two variables $\mathbf{x}_i, \mathbf{x}_j$ that are closely correlated, we should expect the coefficient paths of them to converge, that is

**Theorem 11.3.4: Grouping effect requirement**

For the naive elastic net estimate $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$, WLOG, suppose that $\hat{\beta}_j(\lambda_1, \lambda_2), \hat{\beta}(\lambda_1, \lambda_2) > 0$, then define the ***difference*** between the coefficient paths of predictors $i$ and $j$ as

$$D_{\lambda_1,\lambda_2}(i, j) = \frac{1}{\|\mathbf{y}\|_1} |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|$$

then the grouping effect requires

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

where $\rho = \mathbf{x}_i' \mathbf{x}_j$ is the sample correlation.

### 11.3.3 Elastic Net

One downfall of naive elastic net estimation is that it incurs **double** shrinkage (2-stage procedure, with each stage having a regularizer). To address this issue, we can rescale the naive elastic net estimation to get the **elastic net estimate**:

**Definition 11.3.5: Elastic net solution**

Again for $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, we have

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \mathcal{L}(\gamma, \boldsymbol{\beta}^*) = \arg\min_{\boldsymbol{\beta}^*} \|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \gamma\|\boldsymbol{\beta}^*\|_1$$

and the elastic estimate of $\beta$ is

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

The rescaling from $\hat{\boldsymbol{\beta}}_{\text{naive elastic net}}$ to $\hat{\boldsymbol{\beta}}_{\text{elastic net}}$ preseves the variable selection property (able to select all $p$ variables), while solve the double-regularization shrinkage problem.

But how do we make sense of the scaler $1+\lambda_2$? Remember the ridge estimator is

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

which leads to a Lasso-looking rewritting of the elastic net estimation

**Theorem 11.3.6: Elastic net solution: A modified version of Lasso**

Given data $(\mathbf{y}, \mathbf{X})$ and regularization parameters $(\lambda_1, \lambda_2)$, then the elastic net estimation is

$$\hat{\boldsymbol{\beta}}_{\text{elastic net}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}' \left(\frac{\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}}{1+\lambda_2}\right)\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1$$

comparing to Lasso estimation

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1$$

hence, elastic net is a **stabilized** version of Lasso.

### 11.3.4 Other Variants

There are also some other useful variants of Lasso

- **Positive Lasso**: Constrains the $\hat{\beta}_j$ to enter the prediction equation in their **defined** directions, non-negative here

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right\} \qquad \text{s.t. } \|\boldsymbol{\beta}\|_1 \le t \text{ and } \beta_j > 0, \forall j$$

- **LARS-OLS hybrid**: Use the covariates selected by LARS, but use $\hat{\boldsymbol{\beta}}$ from the OLS model
- **Main effects first**:
  - Step 1: run LARS for a model, considering **only** main effects
  - Step 2: run LARS again, with the chosen main effects, and **all possible interactions** between them
- **Backward Lasso**: start from the **full** OLS model, and eliminate covariates **backwards** (by the order of correlation going 0 the earliest)

## 11.4 Penalized Least Square Estimation

Lasso is one special class of Penalized Least Square (PLS) Estimation. For the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, if $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, we have PLS as

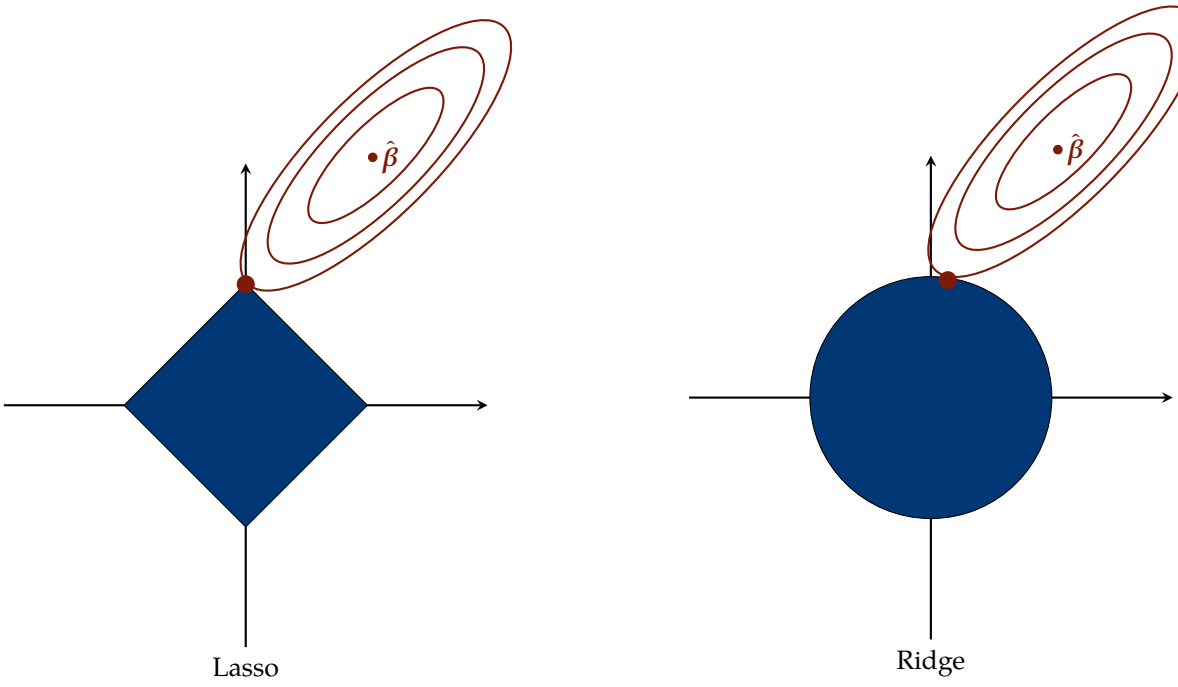
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right) \right\}$$

where $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. Antoniadis and Fan (2001) showed that the PLS estimator $\hat{\boldsymbol{\beta}}$ has the following properties:

- **sparsity**: if $\min_{t\geq 0}\left\{t + p'_\lambda(t)\right\} > 0$
- **approximate unbiasedness**: if $p'_\lambda(t) = 0$ for $t$ large enough
- **continuity**: iff $\arg\min_{t\geq 0}\left\{t + p'_\lambda(t)\right\} = 0$

In general

- the **sigularity** of penalty function at the origin, $p'_\lambda(0_+) > 0$ is needed for generating **sparsity** in variable selection
- the **concavity** is needed to reduce the bias



## References

Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

Nicolai Meinshausen and Peter Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Ann Stat*, 34:1436–1462, 2006.

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