Econometrics **February 28, 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

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

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 α for simplicity.

In matrix form, we have

- constrained form:

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

- unconstrained form:

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

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

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

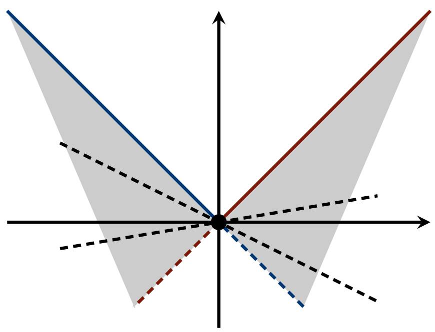
11.1.2 Solving Lasso

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

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

$$\Rightarrow \qquad \frac{1}{n}\underbrace{\mathbf{X}'}_{p\times n}\underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}_{=\boldsymbol{\epsilon},n\times 1}=\lambda\begin{cases}\operatorname{sign}(\beta_j), & \beta_j\neq 0\\ [-1,1], & \beta_j=0\end{cases}$$

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



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

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

which gives a parameter selection criterion: for $\beta_j\neq 0$, $\operatorname{sign}(\beta_j)$ **must agree** with $\operatorname{Corr}(X_j,\boldsymbol{\epsilon})$, 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: from LARS to Lasso

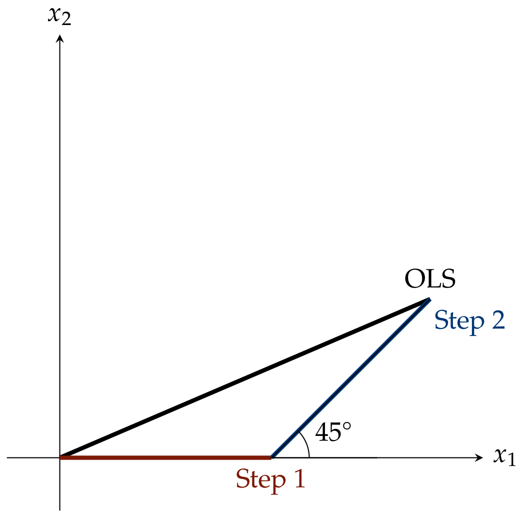
Mathematically, Lasso is quite intuitive, but computionally, it can be quite consuming. Efron et al. (2004) propose an algorithm that takes steps from a all-0 model to the biggest model (OLS), that is, **Least Angle Regression (LARS)**.

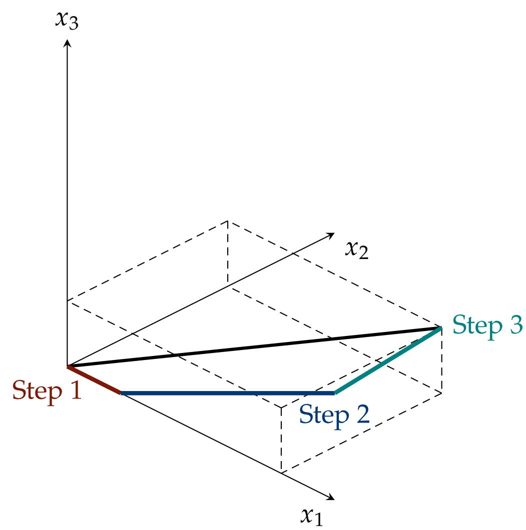
Intuition

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 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

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

11.2 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

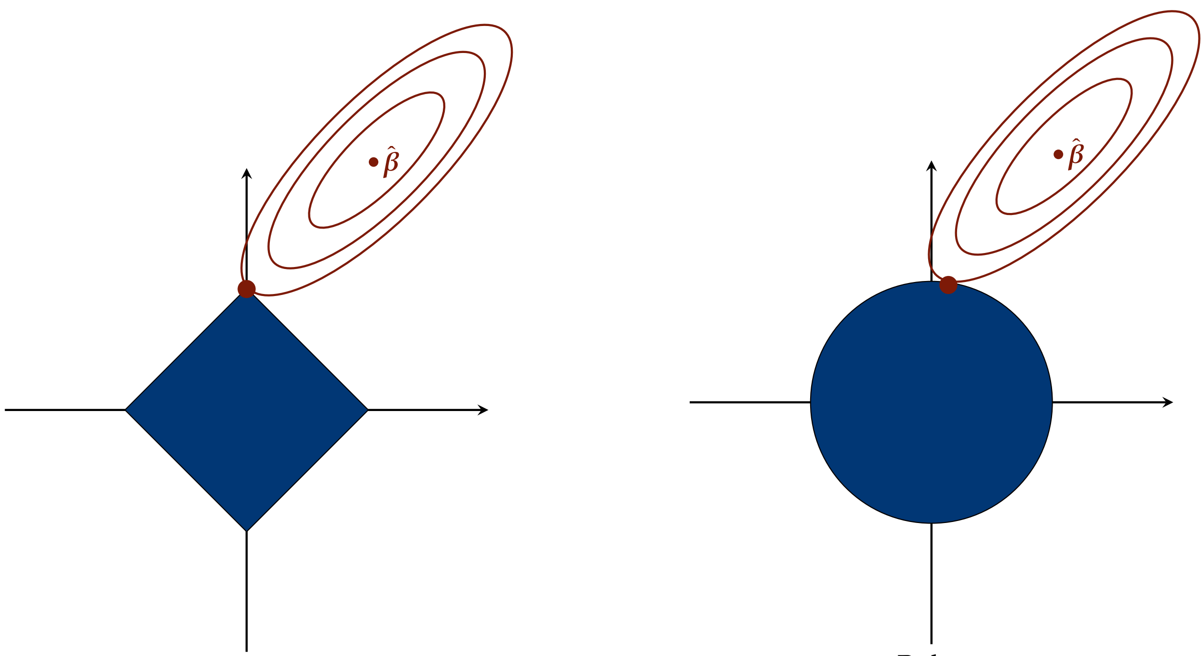
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right) \right\}$$

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



Lasso Ridge

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.

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