

Lasso

Sunday, April 16, 2017 3:38 PM

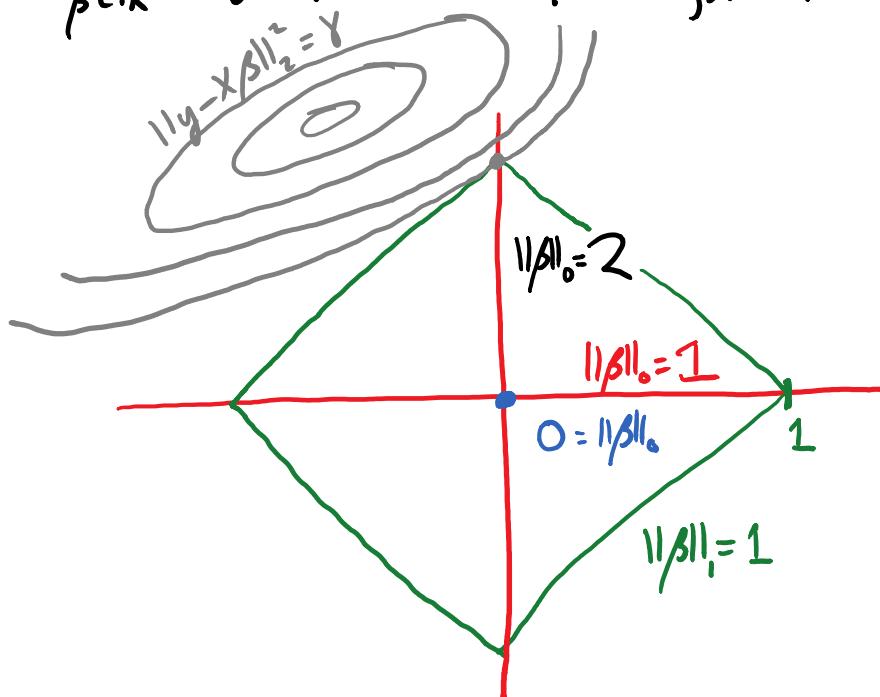
The **lasso estimator** for $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and tuning parameter $C > 0$ is

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \leq C$$

(constrained form)

Compare this to best-subset selection

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_0 = \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\} \leq C.$$



Claim The lasso is a Quadratic Program, i.e.

it can be written as

$$\min \underbrace{\beta^T Q \beta + \beta^T a}_{\text{quadratic}} \quad \text{s.t.} \quad \underbrace{A \beta \leq c}_{\text{linear}}$$

quadratic: $\|y - X\beta\|_2^2 = \beta^T X^T X \beta - 2y^T X \beta + y^T y$

linear: $\|\beta\|_1 \leq A\beta \quad ???$

solution: write $\beta_j = \beta_{j+} - \beta_{j-}$ w/ $\beta_{j+}, \beta_{j-} \geq 0$

then $\|\beta\|_1 = \sum_j \beta_{j+} + \beta_{j-}$

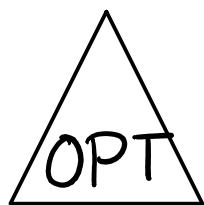
$$\begin{bmatrix} 1, 1, & \dots & & \dots, 1 \\ -1, 0, & \dots & & \dots, 0 \\ 0, 1, 0, & \dots & & \dots, 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & 0, -1, 0 & \dots & \dots, 0 \\ \vdots & & & \vdots & & \vdots \\ 0 & \dots & \dots & -1 & \dots & 0, -1 \end{bmatrix} \begin{bmatrix} \beta_{1+} \\ \vdots \\ \beta_{p+} \\ \beta_{1-} \\ \vdots \\ \beta_{p-} \end{bmatrix} \leq \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \square$$

Claim: Let β_λ be the solution to

$$\min_{\beta} \|\gamma - X\beta\|_2^2 + \lambda \|\beta\|, \quad (\text{regularized form})$$

then for every C there exists a λ s.t.

$$(\text{regularized form}) \quad \beta_\lambda = \beta \quad (\text{constrained form}).$$



More about Convex Opt.

$$(\star) \min_{\beta \in D} R(\beta) \quad \text{s.t.} \quad \begin{array}{l} g(\beta) \leq 0 \text{ vector valued} \\ h(\beta) = 0 \end{array}$$

Define \hookrightarrow Primal prob.
Lagrangean

$$L(\beta, \gamma, \mu) = R(\beta) + \gamma^T g(\beta) + \mu^T h(\beta)$$

for $\beta \in D, \gamma \geq 0$.

Suppose $g_i(\beta) > 0$ for some i (violating constraint)

then $\max_{\gamma \geq 0} L(\beta, \gamma, \mu) = +\infty$

(technically L is unbounded)

and if all constraints are satisfied

$$\text{then } \max_{\gamma \geq 0, \mu} L(\beta, \gamma, \mu) = R(\beta)$$

Hence,

$$(\star) \equiv \min_{\beta \in D} \max_{\gamma \geq 0, \mu} L(\beta, \gamma, \mu)$$

$$(+) \geq \max_{\gamma \geq 0, \mu} \min_{\beta \in D} L(\beta, \gamma, \mu)$$

(weak duality)

When is (+) an equality?
(called strong duality)

clf there is $\beta \in D$ s.t. $g_i(\beta) < 0$
for non-affine g_i , and program
is feasible (called Slater's conditions),
then strong duality (\equiv) holds.

back to the lesson

Primal: $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2$ s.t. $\|\beta\|_1 - c \leq 0$

$$L(\beta, \gamma) = \|y - X\beta\|_2^2 + \gamma(\|\beta\|_1 - c)$$

Let λ be the γ that max'es

$$\max_{\gamma \geq 0} \min_{\beta} L(\beta, \gamma)$$

then

$$\min_{\beta} L(\beta, \lambda) = \min_{\|\beta\|_1 \leq c} \|y - X\beta\|_2^2$$

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda(\|\beta\|_1 - c)$$

□

note: λ is dependent on c, X, y .

Orthogonal Design

Monday, April 17, 2017 3:04 PM

X is orthogonal if $X^T X$ is diagonal

and if X is normalized then this means

$$X^T X = I.$$

Let's look at correlations: X normalized

$$(X^T X)_{jk} = X_j^T X_k = \frac{\sum_i (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_i (x_{ji} - \bar{x}_j)^2} \sqrt{\sum_i (x_{ki} - \bar{x}_k)^2}}$$
$$= \text{Corr}(X_j, X_k)$$

So, X is orthogonal if each variable is (empirically) uncorrelated.

Ex Y_i is stock log-returns and stock i belongs to an industry, j .

Define $X_{ij} = 1\{\text{stock } i \text{ in industry } j\}$,

then $\sum_i X_{ij} X_{ik} = 0 \Rightarrow \text{Orthogonal}$.

Lasso $\|y - X\beta\|_2^2 = y^T y - 2y^T X\beta + \beta^T \beta$

$$\begin{aligned} & \min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \\ & \equiv \min_{\beta} \beta^T \beta - 2 \tilde{y}^T \beta + \lambda \|\beta\|_1, \\ & \text{for } \tilde{y} = X^T y. \end{aligned}$$

Let $s_j \in \begin{cases} \{1\}, & \beta_j > 0 \\ [-1, 1], & \beta_j = 0 \\ \{-1\}, & \beta_j < 0 \end{cases}$ (1), so
 $[s_j \text{ is subgradient of } \|\beta\|_1]$

$$\equiv \min_{\beta, s} \|\tilde{y} - \beta\|_2^2 + \lambda s^T \beta \quad \text{s.t. (1)}$$

$\nabla \|\tilde{y} - \beta\|_2^2 + \lambda s \stackrel{\text{set}}{=} 0 \quad \text{then}$

$$\beta_j = \tilde{y}_j - \frac{\lambda}{2} s_j$$

$$|\tilde{y}_j| \leq \frac{\lambda}{2} \quad \text{then } \beta_j = 0 \text{ is possible}$$

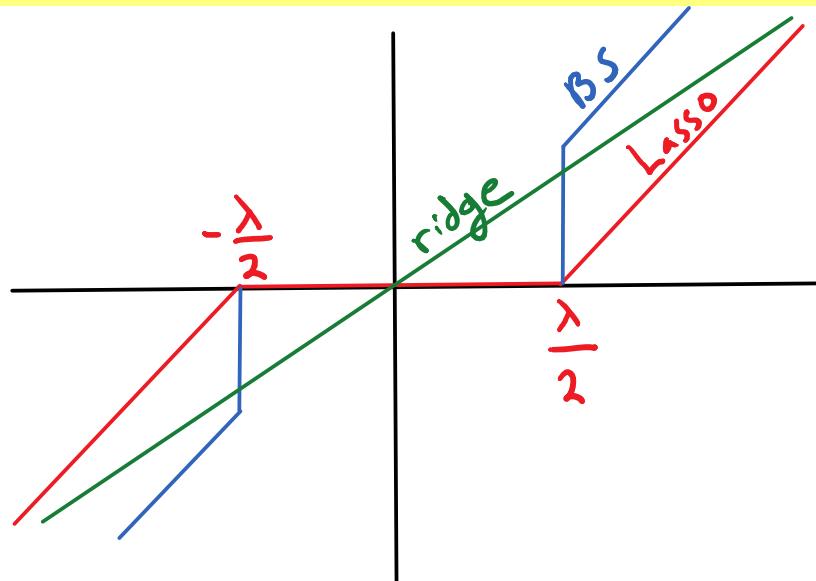
$$\tilde{y}_j > \frac{\lambda}{2} \quad \text{then } \beta_j = \tilde{y}_j - \frac{\lambda}{2}$$

$$\tilde{y}_j < \frac{\lambda}{2} \quad \text{then } \beta_j = \tilde{y}_j + \frac{\lambda}{2}$$

So, lasso solution is soft-thresholding,

$$\hat{\beta} = \text{soft } (\tilde{y} - X^T y) / \lambda$$

$$\hat{\beta}_j = \text{soft}_{\frac{\lambda}{2}}(\tilde{y}_j) = \text{sign}(\tilde{y}_j) \left(|\tilde{y}_j| - \frac{\lambda}{2} \right)$$



Ridge: $\|\tilde{y} - \beta\|_2^2 + \lambda \|\beta\|_2^2 \xrightarrow{\partial \beta} 2(\tilde{y} - \beta) + 2\lambda \beta \stackrel{\text{set}}{=} 0$

$$\hookrightarrow \hat{\beta}_j^{\text{ridge}} = \frac{1}{1+\lambda} \tilde{y}_j$$

Best Subsets: $(\tilde{y}_j - \beta_j)^2 + \lambda \mathbb{1}\{\beta_j \neq 0\}$

min'd at $\hat{\beta}_j^{\text{BS}} = \text{hard}_{\frac{\lambda}{2}}(\tilde{y}_j) = \tilde{y}_j \cdot \mathbb{1}\{|\tilde{y}_j| > \frac{\lambda}{2}\}$

Lasso Path

Monday, April 17, 2017 2:44 PM

For a single C (or λ) can solve the Lasso with QP, proximal gradient, alternating direction method of multipliers, etc.
... but λ is a tuning parameter.

Consider an iterative method, of the form

For $t = 1, 2, \dots$

Update β_t

Define $A_t = \text{supp}(\beta_t)$, the active set

then forward methods (stepwise, stagewise)
have growing active sets $A_1 \subseteq A_2 \subseteq \dots$

Recall stagewise update : change β_j for j with $\max_{\tau \text{ residual}} \text{corr}(X_j, r)$

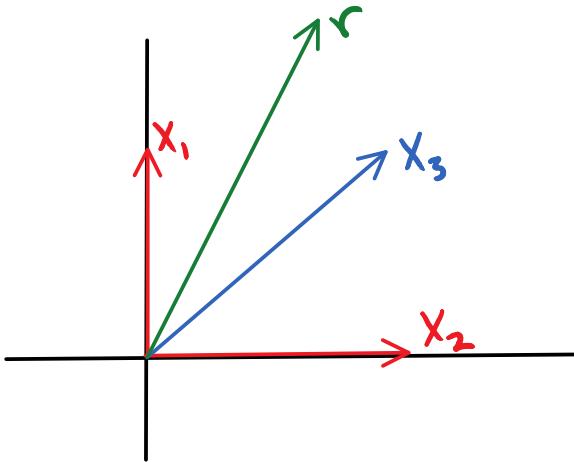
Suppose $X_3 = X_1 + X_2$ is the stagewise update direction, but X_3 is redundant!

1 - - - | $\nearrow r$

$$A_1 = \{3\}$$

$$A_2 = \{3, 1\}$$

$$A_3 = \{3, 1, 2\}$$



$$A_{2,0} = \{1, 2\} \quad (\text{Add 2 and drop 3})$$

Lasso Path

Start at $\lambda = +\infty : \hat{\beta}_\lambda = 0, A_\infty = \{ \}$

Decrease λ until $\hat{\beta}_\lambda \neq 0, A_\lambda = \{j, \}$

Continue decreasing λ , updating
the active set.

- ▷ "leaving" events (variables leaving A_λ) can happen
- ▷ $\hat{\beta}_{\lambda_j}$ is piecewise linear in λ with knots at "hitting" and "leaving" events

Algorithm 3.2 Least Angle Regression.

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \dots, \beta_p = 0$.
 2. Find the predictor \mathbf{x}_j most correlated with \mathbf{r} .
 3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor \mathbf{x}_k has as much correlation with the current residual as does \mathbf{x}_j .
 4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor \mathbf{x}_l has as much correlation with the current residual.
 5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.
-

Algorithm 3.2a Least Angle Regression: Lasso Modification.

- 4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.
-

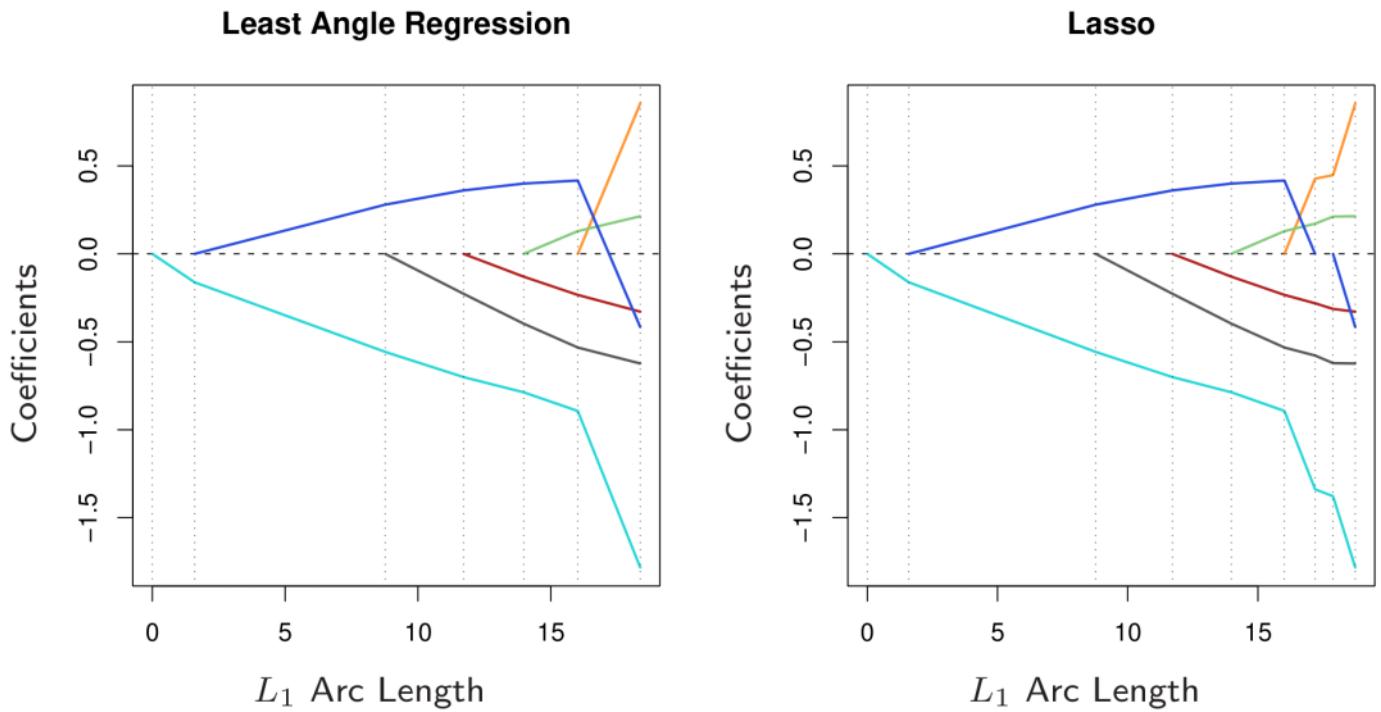


FIGURE 3.15. Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.

ESL pg. 75

Compressed Sensing

Monday, April 17, 2017 6:11 PM

Fact: to compute Lasso solution only need $X_{A_t}^T X_{A_t}$ invertible for A_t in Lasso path. (Can happen if $|A_t| < n$)

Conclusion: Lasso can work in $p \gg n$.

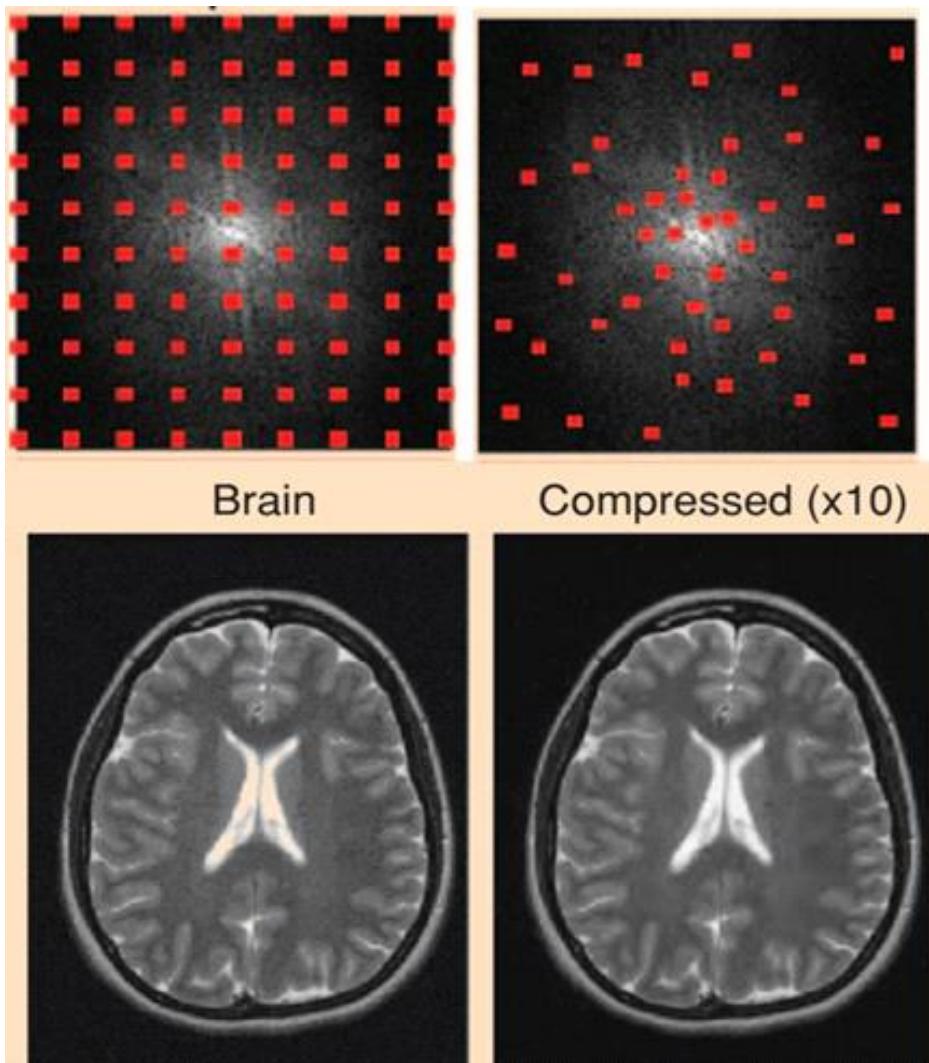
Restricted Isometry Condition (in words):

max eigenvalue of $(X_A^T X_A)^{-1}$ and $X_A^T X_A$ is close to 1 for all $|A| \leq s$.

Summary of results [eg. Candes Et Tao '05]

- ▷ Given the restricted isometry cond. where $|\text{supp}(\beta)| = s$ for $y = X\beta + \epsilon$ then Lasso has $\text{supp}(\beta) = \text{supp}(\hat{\beta})$ for some λ .
- ▷ If X has iid random Gaussian entries then it has the restricted iso. cond.

[Lustig et al. '08]



Compressive sensing typically requires the ability to design X with minimal restrictions. Can happen in

- ▷ medical, astronomical images
- ▷ signal compression
- ▷ some genomics applications