Lasso

Sunday, April 16, 2017 3:38 PM

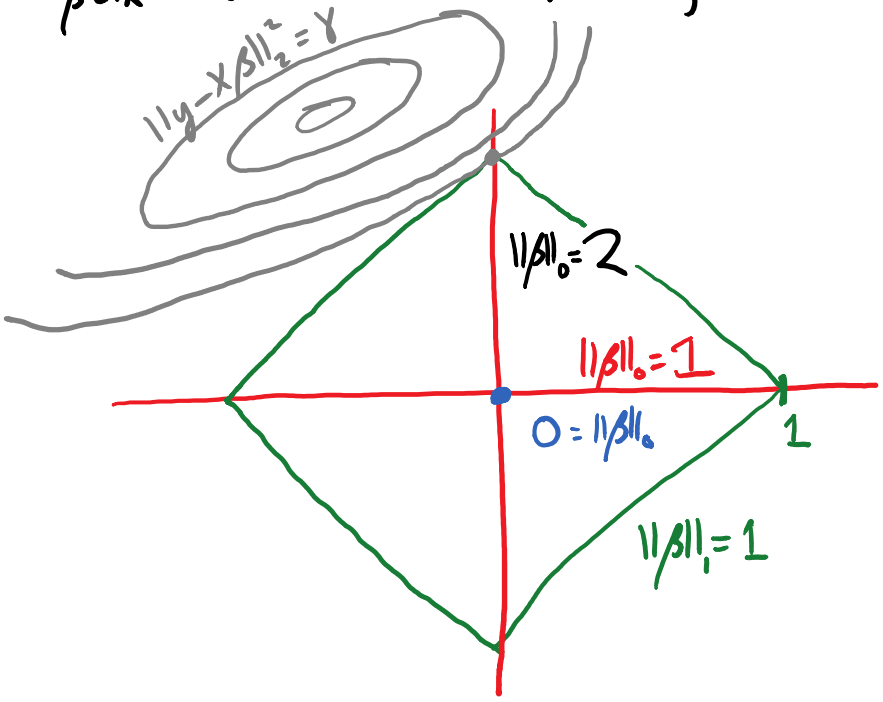
The **lasso estimator** for $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and tuning parameter $C > 0$ is:

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

(constrained form)

Compare this to best-subset selection

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



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

linear: $\|\beta\|_1 \neq A\beta$???

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

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

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ -1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & \cdots & 0 & -1 & 0 \cdots 0 \\ \vdots & & & & \\ 0 & \cdots & & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_{1+} \\ \vdots \\ \beta_{p+} \\ \beta_{1-} \\ \vdots \\ \beta_{p-} \end{bmatrix} \leq \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

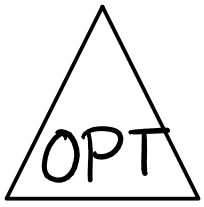
You can verify that lasso soln is argmin.

Claim: Let β_λ be the solution to

$$\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \quad \left(\begin{matrix}\text{regularized}\\ \text{form}\end{matrix}\right)$$

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

$$\left(\begin{matrix}\text{regularized}\\ \text{form}\end{matrix}\right) \quad \beta_\lambda = \beta \quad \left(\begin{matrix}\text{constrained}\\ \text{form}\end{matrix}\right).$$



More about Convex Opt.

(*) $\min_{\beta \in \mathcal{D}} R(\beta)$ s.t. $g(\beta) \leq 0$ (vector valued)
$h(\beta) = 0$

↳ Primal prob.

Define Lagrangian

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

for $\beta \in \mathcal{D}$, $\gamma \geq 0$.

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

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

(technically $\mathcal{L}$ is unbounded)

and if all constraints are satisfied

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

Hence,

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

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

(weak duality)

When is $(\dagger)$ an equality?

(called strong duality)

$(\star)$ is convex

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 lasso

Back to the Lasso

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

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

Let λ be the γ that max's

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

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

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

□

note: λ is dependent on C, X, y.

Orthogonal Design

Monday, April 17, 2017 3:04 PM

X is orthogonal if X^TX is diagonal

and if X is normalized then this means

$$X^TX = I.$$

Let's look at correlations: X normalized

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

$$= \text{Corr}(X_j, X_k)$$

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 $\sum_i x_{ij} x_{in} = 0 \Rightarrow$ Orthogonal.

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

$$\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\equiv \min_\beta \beta^T\beta - 2\tilde{y}^T\beta + \lambda\|\beta\|_1$$

for $\tilde{y} = X^T y$.

$$\equiv \min_\beta \|\tilde{y} - \beta\|_2^2 + \lambda\|\beta\|_1,$$

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 is subgradient of $|\beta_j|$]

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

$\frac{\partial}{\partial\beta}$ $-2(\tilde{y} - \beta) + \lambda s \overset{\text{set}}{=} 0$ then

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

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

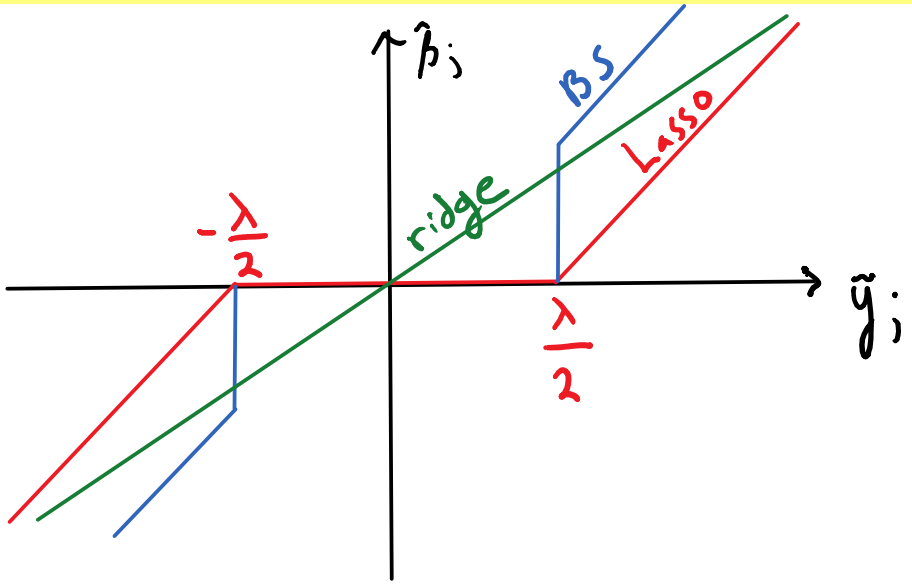
$\tilde{y}_j > \frac{\lambda}{2}$ then $\beta_j = \tilde{y}_j - \frac{\lambda}{2}$

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

So, lasso solution is soft-thresholding,

$\hat{\beta}$ = soft-$(\tilde{y}, \ldots, \tilde{y}_n)(\lambda, \ldots, \lambda)$

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



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

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

Best Subset: $(\tilde{y}_j - \beta_j)^2 + \frac{\lambda^2}{4} 1\{\beta_j \neq 0\}$

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

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 $t = 1, 2, \ldots$

Update β_t

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

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

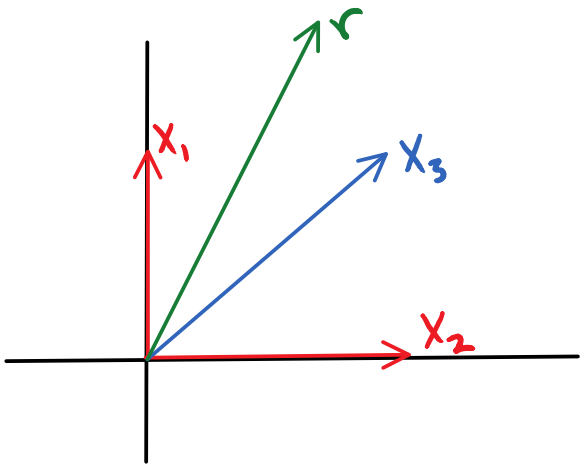
Recall stagewise update: change β_j for j with max $\text{Corr}(X_j, r)$, r residual

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

$A_1 = \{3\}$

$A_2 = \{3, 1\}$

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



$A_{2.5} = \{1, 2\}$ (Add 2 and drop 3)

Lasso Path

Start at $\lambda = +\infty$: $\hat{\beta}_\lambda = 0$, $A_0 = \{\}$

Decrease λ until $\hat{\beta}_{j_1} \neq 0$, $A_1 = \{j_1\}$

Continue decreasing λ, updating the active set.

- "leaving" events (variables leaving A_t) 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 $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

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.

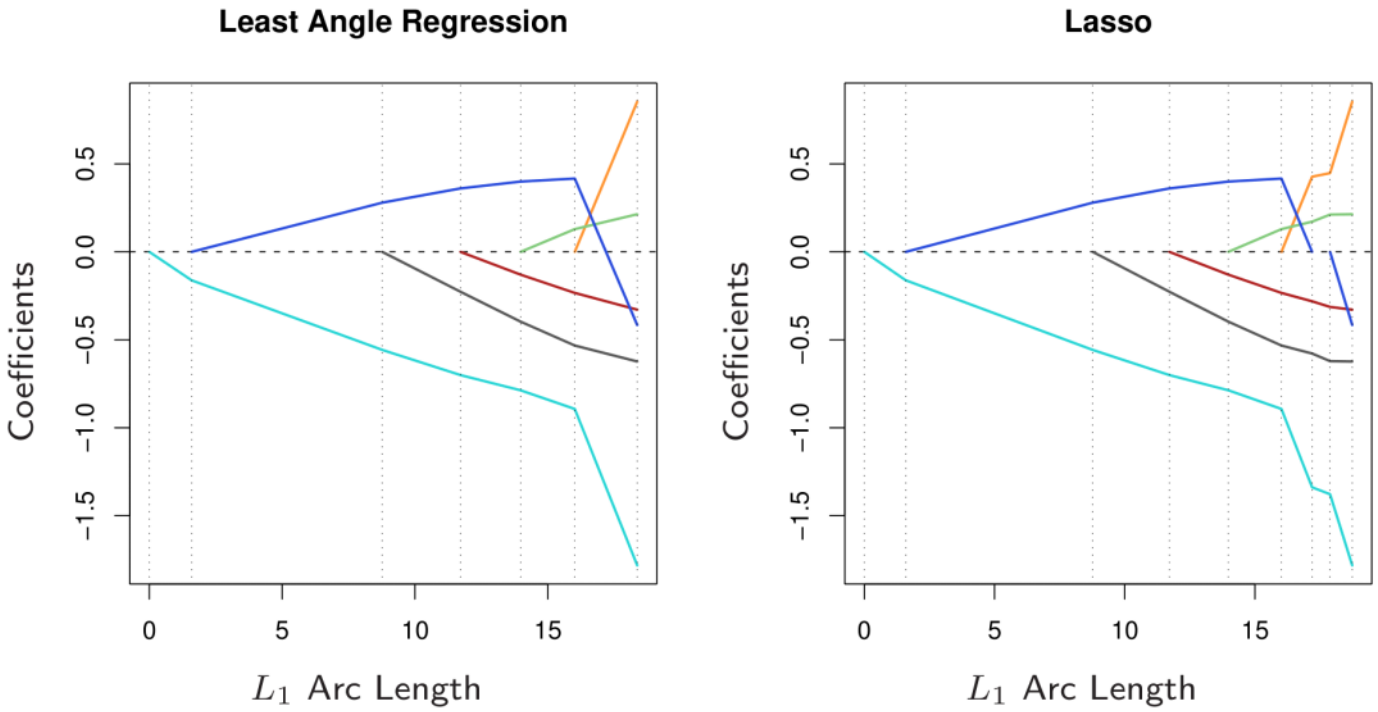


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about* 18.

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 $p \gg n$.

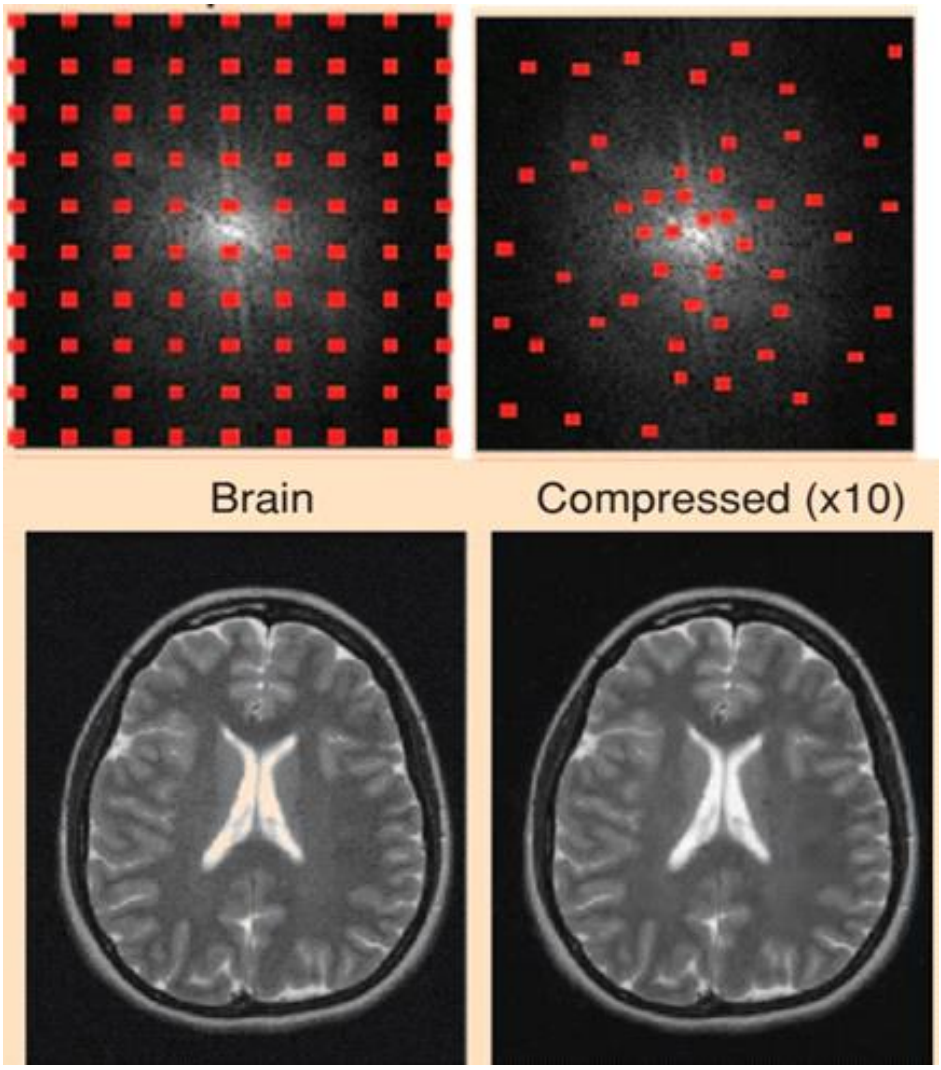
Restricted Isometry Condition (in words):

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

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

- Given the restricted isometry cond. where $|\text{supp}(\beta)| = s$ for $y = X\beta + \varepsilon$ 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