The lasso

The following explanations are based mainly on ideas drawn from the papers of Efron et al. (2004) and Tibshirani (2013). The analysis simplifies the explanations in DPlasso2010.pdf, my first attempt at understanding the LARS algorithm.

1 Least squares and lasso

In the classical least squares problem one starts from an $n \times 1$ vector y and an $n \times p$ matrix $X = (x_1, \dots, x_p)$. The task is:

<1> $$\text{find } \widehat{b} \in \mathbb{R}^p \text{ to minimize } Q(b) := \|y - Xb\|_2^2.$$

If X has rank p there exists a unique minimizing $\widehat{b}$, which can also be characterized by the equalities

$$x_j^T(y - X\widehat{b}) = 0 \qquad \text{for } j = 1, \dots, p.$$

If $\text{rank}(X) < p$ there are many solutions $\widehat{b}$, but they all make $X\widehat{b}$ equal to $\widehat{y}$, the orthogonal projection of y onto $\text{span}(X)$.

Tibshirani (1996) proposed a modification, by constraining the minimizing b to lie in a convex set $K_r = \{b \in \mathbb{R}^p : \|b\|_1 \le r\}$. The task became:

<2> $$\text{find } \widehat{b} \in K_r \text{ to minimize } Q(b) \text{ over } K_r$$

(Osborne et al., 2000) recast the problem in dual form. The task became:

<3> $$\text{find } \widehat{b} \in \mathbb{R}^p \text{ to minimize } G_\lambda(b) := \tfrac{1}{2}Q(b) + \lambda\|b\|_1.$$

The tuning parameters r and λ play parallel roles.

The LARS algorithm provides a continuous, piecewise linear set of estimates $\{\widehat{b}(\lambda) : 0 < \lambda < \infty\}$ for which $\widehat{b}(\lambda)$ minimizes G_λ, for each λ. That is,

<4> $$G_\lambda(\widehat{b}(\lambda)) = m(\lambda) := \min_{b\in\mathbb{R}^p} G_\lambda(b).$$

In general the $\widehat{b}(\lambda)$'s are not uniquely determined. However, as the next lemma shows, the corresponding fitted vectors $\widehat{y}(\lambda) = X\widehat{b}(\lambda)$ are unique.

<5> **Lemma.** *For a given $\lambda > 0$ suppose both $\widehat{a}$ and $\widehat{c}$ minimize G_λ. Then $X\widehat{a} = X\widehat{c}$ and $\|\widehat{a}\|_1 = \|\widehat{c}\|_1$.*

Proof Define $\widehat{b} = (\widehat{a} + \widehat{c})/2$. Define $f(z) = \frac{1}{2}\|y - z\|_2^2$, a strictly convex function of z. By convexity of f and $\|\cdot\|_1$,

$$\begin{aligned} G_\lambda(\widehat{b}) &= f(X\widehat{b}) + \lambda\|\widehat{b}\|_1 = f\left(\tfrac{1}{2}X\widehat{a} + \tfrac{1}{2}X\widehat{c}\right) + \lambda\|\tfrac{1}{2}\widehat{a} + \tfrac{1}{2}\widehat{c}\| \\ &\le \tfrac{1}{2}f(X\widehat{a}) + \tfrac{1}{2}f(X\widehat{c}) + \lambda\left(\tfrac{1}{2}\|\widehat{a}\|_1 + \tfrac{1}{2}\|\widehat{c}\|_1\right) \\ &= \tfrac{1}{2}G_\lambda(\widehat{a}) + \tfrac{1}{2}G_\lambda(\widehat{c}) = m(\lambda). \end{aligned}$$

The inequality would be strict if $X\widehat{a}$ were not equal to $X\widehat{c}$, which would lead to the contradiction that $G_\lambda(\widehat{b})$ is strictly smaller than the minimum of G_λ.

The equalities $G_\lambda(\widehat{a}) = m(\lambda) = G_\lambda(\widehat{c})$ and $\|y - X\widehat{a}\|_2^2 = \|y - X\widehat{c}\|_2^2$ then imply $\lambda\|\widehat{a}\|_1 = \lambda\|\widehat{c}\|_1$.

□

The Lemma shows that we may define quantities

$$\begin{aligned} &\widehat{y}(\lambda) = X\widehat{b} \quad \text{AND} \quad R(\lambda) = y - \widehat{y}(\lambda) \\ &q(\lambda) = \tfrac{1}{2}\|R(\lambda)\|_2^2 \quad \text{AND} \quad \ell(\lambda) = \|\widehat{b}\|_1 \end{aligned}$$

by choosing $\widehat{b}$ as any minimizer of G_λ. From Homework 9, both $q(\lambda)$ and $m(\lambda) = q(\lambda) + \ell(\lambda)$ are increasing functions of λ and $\ell(\lambda)$ is a decreasing function.

There is another way to think about $\widehat{b}(\lambda)$ for $\lambda > 0$. By definition,

<6> $$\tfrac{1}{2}Q(\widehat{b}(\lambda)) + \lambda\|\widehat{b}(\lambda)\|_1 \le \tfrac{1}{2}Q(b) + \lambda\|b\|_1 \qquad \text{for all } b \text{ in } \mathbb{R}^p.$$

If b is a vector with $Xb = \widehat{y}(\lambda)$ then

<7> $$Q(\widehat{b}(\lambda)) = \|R(\lambda)\|_2^2 = Q(b).$$

Subtraction followed by cancellation of a λ factor give $\|\widehat{b}(\lambda)\|_1 \le \|b\|_1$. In other words,

$$\widehat{b}(\lambda) \text{ minimizes } \|b\|_1 \text{ over } \{b \in \mathbb{R}^p : Xb = \widehat{y}(\lambda)\}.$$

Even though this fact is not immediately useful it does give some insight into the behavior of $\widehat{b}(\lambda)$ as λ decreases to zero.

Piecewise linearity of $\lambda \mapsto \widehat{b}(\lambda)$ ensures that $b^* := \lim_{\lambda\to 0} \widehat{b}(\lambda)$ is a well defined vector in $\mathbb{R}^p$. By continuity and the fact that $\lambda \mapsto q(\lambda)$ is increasing we must have $Q(\widehat{b}(\lambda)) \downarrow Q(b^*)$. We also have

$$\begin{aligned}\tfrac{1}{2}Q(b^*) &= \lim_{\lambda\to 0}\left(\tfrac{1}{2}Q(\widehat{b}(\lambda)) + \lambda\|\widehat{b}(\lambda)\|_1\right)\\ &\le \lim_{\lambda\to 0}\left(\tfrac{1}{2}Q(b) + \lambda\|b\|_1\right) = \tfrac{1}{2}Q(b) \qquad \text{for every } b \in \mathbb{R}^p.\end{aligned}$$

The vector b^* minimizes $Q(b) = \|y - Xb\|_2^2$ over $\mathbb{R}^p$. It solves the least squares problem <1>. Moreover, for every other least squares solution $\widehat{a}$,

$$\tfrac{1}{2}Q(b^*) + \lambda\|\widehat{b}(\lambda)\|_1 \le \tfrac{1}{2}Q(\widehat{b}(\lambda)) + \lambda\|\widehat{b}(\lambda)\|_1 \le \tfrac{1}{2}Q(\widehat{a}) + \lambda\|\widehat{a}\|_1.$$

From the equality $Q(b^*) = Q(\widehat{a})$ we deduce that $\|\widehat{b}(\lambda)\|_1 \le \|\widehat{a}\|_1$. In the limit as λ tends to zero we must have $\|b^*\|_1 \le \|\widehat{a}\|_1$. That is, the limit of the LARS $\widehat{b}(\lambda)$'s as λ tends to zero gives the least squares solution with the smallest ℓ^1 norm.

2 Minimization of convex functions

Suppose G is a convex function (such as the G_λ from <8>) defined on $\mathbb{R}^p$, for which $G(p) \to \infty$ as $\|b\|_2 \to \infty$. Such a G must achieve its minimum, possibly at more than one point—the minimizer need not be unique.

> **Remark.** The convex function $g(t) = e^{-t}$ on the real line does not achieve its minimum. It gets arbitrarily close to zero but $g(t) \ne 0$ for all $t \in \mathbb{R}$. The assumption that $G(p) \to \infty$ as $\|b\|_2 \to \infty$ prevents analogous behavior in higher dimensions.

For each b and u in $\mathbb{R}^p$, the directional derivative

$$D_G(b,u) = \lim_{t\downarrow 0}\frac{G(b+tu) - G(b)}{t}$$

is well defined. A point b minimizes G if and only if

$$D_G(b,u) \ge 0 \qquad \text{for all } u \in \mathbb{R}^p.$$

<8> **Example.** The penalized form of the lasso minimizes the convex function

$$G_\lambda(b) = \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1.$$

Abbreviate D_{G_λ} to D_λ. Then, as you showed on Homework 9,

$$D_\lambda(b,u) = \sum\nolimits_{j\in[p]} u_j \left(\lambda\mathcal{R}(\widehat{b}_j)\mathbb{1}\{u_j > 0\} + \lambda\mathcal{L}(\widehat{b}_j)\mathbb{1}\{u_j < 0\} - x_j^T(y - X\widehat{b})\right)$$

where

$$\begin{aligned}\mathcal{R}(t) &= \mathbb{1}\{t \ge 0\} - \mathbb{1}\{t < 0\} = \operatorname{sgn}(t) + \mathbb{1}\{t = 0\}\\ \mathcal{L}(t) &= \mathbb{1}\{t > 0\} - \mathbb{1}\{t \le 0\} = \operatorname{sgn}(t) - \mathbb{1}\{t = 0\}.\end{aligned}$$

It follows that $D_\lambda(\widehat{b}, u) \ge 0$ for all u if and only if

$$\begin{aligned} x_j^T(y - X\widehat{b}) &= \lambda \qquad \text{for all } j \text{ where } \widehat{b}_j > 0\\ x_j^T(y - X\widehat{b}) &= -\lambda \qquad \text{for all } j \text{ where } \widehat{b}_j < 0\\ |x_j^T(y - X\widehat{b})| &\le \lambda \qquad \text{for all } j \text{ where } \widehat{b}_j = 0.\end{aligned}$$

These inequalities are usually called the KKT conditions.

Temporarily write $C_j(\widehat{b}, \lambda)$ for $x_j^T(y - X\widehat{b})$ and $s_j(\widehat{b}, \lambda)$ for $C_j(\widehat{b}, \lambda)/\lambda$. The KKT conditions then become: $\widehat{b}$ minimizes G_λ if and only if

<9> $$|s_j(\widehat{b}, \lambda)| \le 1 \qquad \text{for all } j$$

<10> $$s_j(\widehat{b}, \lambda) = +1 \qquad \text{if } \widehat{b}_j > 0$$

<11> $$s_j(\widehat{b}, \lambda) = -1 \qquad \text{if } \widehat{b}_j < 0$$

It is often easy to check the KKT conditions for an explicitly specified $\widehat{b}$. It takes more effort to find the candidate $\widehat{b}$ in the first place. The LARS algorithm constructs $\widehat{b}(\lambda)$ by extending from a trivial case (very large λ, which forces $\widehat{b}$ to be zero) to smaller values of λ. As you will soon see, the KKT conditions themselves suggest the way in which $\widehat{b}(\lambda - t)$ should be related to $\widehat{b}(\lambda)$. □

The LARS algorithm starts with large λ. More precisely, we first consider

$$\lambda \ge L_1 := \max\nolimits_{j\in[p]} |x_j^T y|.$$

Try $\widehat{b} = 0$. The KKT conditions reduce to

$$|x_j^T y| \le \lambda \qquad \text{for all } j,$$

which clearly holds for $\lambda \ge L_1$.

> **Remark.** How would you prove that $\widehat{b} = 0$ is the only solution for this range of λ? Hint: Why must we have $\widehat{y}(\lambda) = 0$?

Just as clearly, the candidate $\widehat{b} = 0$ fails the KKT requirement if $\lambda < L_1$.

We need to move $\widehat{b}(L_1 - t)$ away from zero if $t > 0$.

For the diabetes data $L_1 \approx 949.4 = x_{bmi}^T y$. For all the other predictors $|x_j^T y| < L_1$. At least for small t we need only modify the bmi coefficient to maintain the KKT constraints. Consider

$$\widehat{b}_{bmi}(L_1 - t) = t \quad \text{AND} \quad \widehat{b}_j(L_1 - t) = 0 \qquad \text{for } j \neq \text{bmi}.$$

That gives

$$X\widehat{b}(L_1 - t) = x_{bmi}\widehat{b}_{bmi}(L_1 - t) = tx_{bmi}$$

and $R(L_1 - t) = y - X\widehat{b}(L_1 - t) = y - tx_{bmi}$. For this choice

$$x_{bmi}^T R(L_1 - t) = x_{bmi}^T y - t = L_1 - t.$$

The KKT constraint for $j = \text{bmi}$ is still satisfied. For $j \neq \text{bmi}$,

$$x_j^T R(L_1 - t) = x_j^T y - x_j^T x_{bmi} t.$$

For small enough positive t all the KKT constraints are still satisfied.

The $C_j(L_1 - t)$'s for the other predictors also change with t. Eventually $\max_{j \neq bmi} |s_j(\widehat{b}(L_1 - t)| = 1$, which forces a recalculation of $\widehat{b}$.

3 The equicorrelation set

Suppose we have a $\widehat{b}(\lambda_0)$ that minimizes G_λ, for some specified λ_0. Tibshirani (2013, page 1464) called

$$\mathcal{E}(\lambda_0) := \{j : |x_j^T R(\lambda_0)| = \lambda_0\}$$

the equicorrelation set (at parameter λ_0).

> **Remark.** Notice that $\mathcal{E}(\lambda_0)$ depends on $\widehat{b}(\lambda_0)$ only through the vector $R(\lambda_0)$ of residuals.

For the moment let me abbreviate $\mathcal{E}(\lambda_0)$ to $\mathcal{E}$ when subscripts start getting too messy. The vector

$$s_{\mathcal{E}}(\lambda_0) := (s_j(\widehat{b}(\lambda_0), \lambda_0) : j \in \mathcal{E}(\lambda_0))$$

contains only values ±1. For j in $\mathcal{E}^c(\lambda_0)$ the KKT constraint forces $\widehat{b}_j(\lambda_0) = 0$, so that $R(\lambda_0) = y - X_{\mathcal{E}}\widehat{b}_{\mathcal{E}}$ and

<12> $$X_{\mathcal{E}}^T R(\lambda_0) = \lambda_0 s_{\mathcal{E}}.$$

For $j \in \mathcal{E}^c$ we must have $|x_j^T R(\lambda_0)| < \lambda_0$.

Now suppose we have another value of λ, say $\lambda_1 > \lambda_0$, with the same equicorrelation set and the same $s_{\mathcal{E}}$:

$$\mathcal{E}(\lambda_0) = \mathcal{E}(\lambda_1) = \mathcal{E} \quad \text{AND} \quad s_{\mathcal{E}}(\lambda_0) = s_{\mathcal{E}}(\lambda_1) = s_{\mathcal{E}}.$$

For $0 < \theta < 1$ define $\lambda_\theta := (1-\theta)\lambda_0 + \theta\lambda_1$ and $\widehat{a}(\theta) := (1-\theta)\widehat{b}(\lambda_0) + \theta\widehat{b}(\lambda_1)$. Notice that

$$\begin{aligned} X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\widehat{a}(\theta)) &= X_{\mathcal{E}}^T\Big((1-\theta)(y - X_{\mathcal{E}}\widehat{b}(\lambda_0)) + \theta(y - X_{\mathcal{E}}\widehat{b}(\lambda_1))\Big) \\ &= (1-\theta)X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\widehat{b}(\lambda_0)) + \theta X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\widehat{b}(\lambda_1)) \\ &= \lambda_\theta s_{\mathcal{E}} \end{aligned}$$

and, for $j \in \mathcal{E}^c$,

$$\begin{aligned} |x_j^T(y - X\widehat{a}(\theta))| &= |(1-\theta)x_j^T(y - X_{\mathcal{E}}\widehat{b}(\lambda_0)) + \theta x_j^T(y - X_{\mathcal{E}}\widehat{b}(\lambda_0)) \\ &< \lambda_\theta s_{\mathcal{E}}. \end{aligned}$$

Thus:

(i) The vector $\widehat{a}(\theta)$ satisfies the KKT conditions at λ_θ.

(ii) The vector $\widehat{a}(\theta)$ must minimize G_{λ_θ}.

(iii) The residual vector $R(\lambda_\theta)$ must equal $y - X_{\mathcal{E}}\widehat{a}(\theta)$.

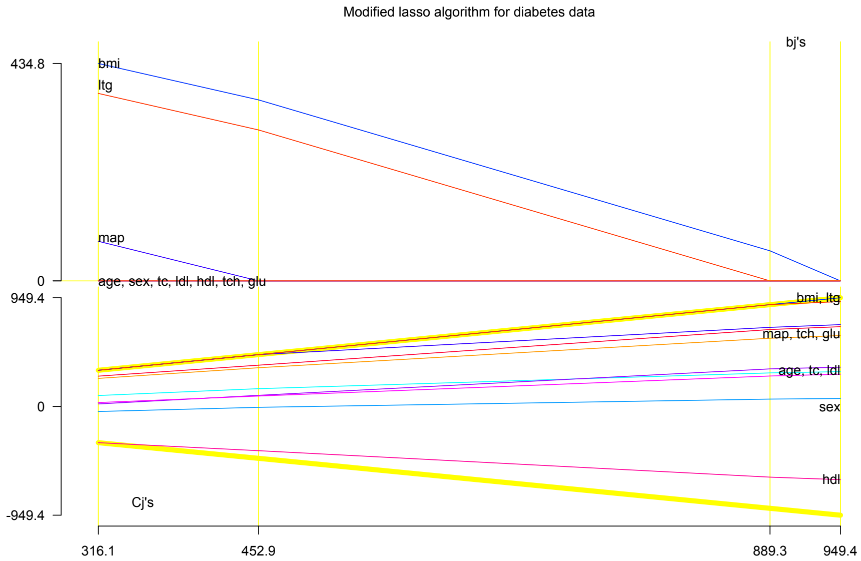
(iv) The equicorrelation set $\mathcal{E}(\lambda_\theta)$ must equal $\mathcal{E}$ and the vector $s_{\mathcal{E}}(\lambda_\theta)$ must equal $s_{\mathcal{E}}$.

These four conclusions tell us a lot about the LARS solution path. The set $(0, \infty)$ must be partitioned into finitely many intervals on each of which the equicorrelation set $\mathcal{E}(\lambda)$ and the vector $s_{\mathcal{E}}(\lambda)$ stay constant. Moreover, once the solution path changes the $\mathcal{E}(\lambda)$ or the $s_{\mathcal{E}}(\lambda)$ it can never return to the same set of values. Even better, we can take $\widehat{b}(\lambda)$ to be linear on each of the intervals.

It now just a matter of finding those intervals and determining how $\widehat{b}(\lambda)$ behaves on each of them.

4 The LARS solution path for the diabetes data

At the end of Section 2 we left LARS as it started to explore values of λ smaller than $L_1 \approx 949.4$. The coefficient $\widehat{b}_{bmi}$ was happy to be increasing linearly; the other coefficients were staying zero. The $C_{bmi}(\widehat{b}(\lambda), \lambda)$ was cruising down the upper boundary of the constraint set. The first disruption occures at $\lambda = L_2 \approx 889.3$, where $C_{ltg}(\widehat{b}(\lambda), \lambda)$ hits the upper boundary, adding a second predictor to the equicorrelation set.



By an as-yet unexplained method, the coefficients $\widehat{b}_{bmi}$ and $\widehat{b}_{ltg}$ then switch to different linear regime, while the other coefficients stay at zero. At $\lambda = L_3 \approx 452.9$ the map predictor joins the equicorrelation set.

Modification of the linear regime is also needed when one of the coefficients for the equicorrelation set hits zero. I'll skip over this aspect of the LARS algorithm and instead focus on what happens when a new variable joins $\mathcal{E}(\lambda)$, say at $\lambda = \lambda_1$. Abbreviate $\mathcal{E}(\lambda_1)$ to $\mathcal{E}$ and $s_{\mathcal{E}}(\lambda_1)$ to $s_{\mathcal{E}}$. We know that

$$X_{\mathcal{E}}^T R(\lambda_1)/\lambda_1 = X_{\mathcal{E}}^T (y - X_{\mathcal{E}} \widehat{b}_{\mathcal{E}})/\lambda_1 = s_{\mathcal{E}}.$$

From HW9.1, there exists a vector W for which

$$X_{\mathcal{E}}^T X_{\mathcal{E}} W = s_{\mathcal{E}}.$$

In fact every $X_{\mathcal{E}}$ has full rank for the diabetes example, so the W can be written as

$$W = (X_{\mathcal{E}}^T X_{\mathcal{E}})^{-1} \mathfrak{X}_{\mathcal{E}}^T R(\lambda_1)/\lambda_1.$$

Consider the effect of taking

$$\widehat{b}_{\mathcal{E}}(\lambda_1 - t) = \widehat{b}_{\mathcal{E}}(\lambda_1) + tW \qquad \text{for } t > 0,$$

with the other $\widehat{b}_j$'s left at zero.

$$\begin{aligned} X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\widehat{b}_{\mathcal{E}}(\lambda_1 - t)) &= X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\widehat{b}_{\mathcal{E}}(\lambda_1)) - X_{\mathcal{E}}^T X_{\mathcal{E}}(tW) \\ &= \lambda_1 s_{\mathcal{E}} - ts_{\mathcal{E}} = (\lambda_1 - t)s_{\mathcal{E}}. \end{aligned}$$

Voila!

The new C_j's for $j \in \mathcal{E}$ are now heading along the appropriate $\pm\lambda$ boundaries. The KKT conditions are again satisfied until a new predictor wants to enter the equicorrelation set and we have extended $\widehat{b}(\lambda)$ to some interval $[\lambda_0, \lambda_1]$.

And so on.

References

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), pp. 407–451. With discussion, 452–499.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*(2), 319–337.

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Statist. 7*, 1456–1490.