STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 6 (April 12): This lecture continues the discussion of optimization and algorithmic aspects of Lasso and related methods, as discussed in the two papers by Osborne, Presnell, and Turlach (2000a,b) and Efron, Hastie, Johnstone, and Tibshirani (2003).

First we review two classical variable selection strategies for regression: *forward stepwise regression* (FStepR) and *forward stagewise regression* (FStageR).

Notation: Suppose that

$$\begin{aligned} \underline{Y} &= \mathbf{X}\underline{\beta} + \underline{\epsilon} \qquad \text{in } \mathbb{R}^n \\ &= \mu(\beta) + \underline{\epsilon} \end{aligned}$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ matrix, and each $\mathbf{x}_j$ is $n \times 1$. We assume that $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = 1$, $j = 1, \ldots, p$ and $\sum_{i=1}^n Y_i = 0$. For any estimator $\widehat{\beta}$ of β, let

$$\widehat{\mu} \equiv \widehat{\mu}(\beta) = X\widehat{\beta} = \sum_{j=1}^p \mathbf{x}_j \widehat{\beta}_j$$

be the predictor vector. Recall that the Lasso method seeks to

$$\text{minimize } \; S(\beta) = \frac{1}{2}\|Y - \mu(\beta)\|^2 \qquad \text{subject to } \; \|\beta\|_1 \le t$$

for $0 < t < t_0 \equiv \|\widehat{\beta}^{LS}\|_1$.

Forward stepwise regression: (Weisberg (2005), section 10.3; Miller (2002), section 3.2, p. 39 ff). *Step 1.* Let

$$\begin{aligned} j_1 &\equiv \text{argmax}_{1 \le j \le p} |\langle Y, \mathbf{x}_j \rangle| / \sqrt{\sum_1^n Y_i^2 \sum_1^n x_{ij}^2} \\ &= \text{argmax}_{1 \le j \le p} |\langle \widehat{Y}^{LS}, \mathbf{x}_j \rangle| / \sqrt{\sum_1^n Y_i^2 \sum_1^n x_{ij}^2} \end{aligned}$$

since $\langle Y - \widehat{Y}^{LS}, \mathbf{x}_j \rangle = 0$.
Step 2. Let

$$r_1 \equiv Y - \mathbf{x}_j \widehat{\beta}_{j_1}, \qquad X_{-j_1}^{(1)} = \Pi(X_{-j_1} | [\mathbf{x}_{j_1}]^\perp).$$

Step 3. Let

$$\begin{aligned} j_2 &\equiv \text{argmax}_{1\le j\le p, j\neq j_1} |\langle r_1, \mathbf{x}^{(1)}_{-j_1,j}\rangle| / \sqrt{\sum_1^n Y_i^2 \sum_1^n (x^{(1)}_{ij,-j_1})^2} \\ r_2 &= r_1 - x^{(1)}_{j_2} \widehat{\beta}_{j_2}. \end{aligned}$$

Then iterate this procedure.

Forward stagewise regression: Start with $\hat{\mu} \equiv \hat{\mu}^{(0)} = 0$. If $\hat{\mu} \equiv \hat{\mu}^{(m)}$ is the current estimate, then compute

$$\begin{aligned} &\underline{\hat{c}} \equiv c(\hat{\mu}) = X'(Y - \hat{\mu}) = \text{current correlations}, \\ &\hat{c}_j \propto \text{correlation between } \mathbf{x}_j \text{ and current residual vector.} \end{aligned}$$

Let $\hat{j} = \text{argmin}_j |\hat{c}_j|$ and set

$$\hat{\mu}^{(m+1)} = \hat{\mu}^{(m)} + \epsilon \text{sign}(\hat{c}_{\hat{j}}) \mathbf{x}_{\hat{j}}$$

where $\epsilon > 0$ is "small". Iterate this procedure. If we take $\epsilon = |\hat{c}_j|$ at each step, then we are back to forward stepwise regression.

LARS: Start with $\hat{\mu} \equiv \hat{\mu}^{(0)} = 0$.
Step 1. Find $\hat{j}_1 = \text{argmax}\langle Y, \mathbf{x}_j\rangle$.
Step 2. Take the largest step possible in the direction of this predictor, until some other predictor, say x_{j_2}, has the same amount of correlation with the current residual. Call this step $\mu^{(1)} = \hat{\mu}^{(0)} + \hat{\gamma}_1 \mathbf{x}_1$.
Step 3. Choose that direction equiangular between x_{j_1} and x_{j_2}; call this direction u_2.
Step 4. Follow this direction until a 3rd variable j_3 comes into the most correlated set, and continue ...

With two predictors, as in Figure xx,

$$\hat{\mu}^{(2)} = \hat{\mu}^{(1)} + \hat{\gamma}_2 u_2.$$

Claim: It is easy to calculate the step sizes $\hat{\gamma}_j$.

In step 4 and further, the LARS steps go along equiangular vectors: assume that $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are linearly independent. For $\mathcal{A} \subset \{1, \ldots, p\}$, let

$$\mathbf{X}_\mathcal{A} = (\cdots, s_j \mathbf{x}_j, \cdots)_{j\in\mathcal{A}}, \qquad \text{where } \; s_j = \pm 1, \tag{1}$$

$$M_\mathcal{A} = \mathbf{X}_\mathcal{A}^T \mathbf{X}_\mathcal{A}, \tag{2}$$

$$A_\mathcal{A} = (\underline{1}_\mathcal{A} M_\mathcal{A}^{-1} \underline{1}_\mathcal{A})^{-1/2}, \;\; \text{where} \tag{3}$$

$$\underline{1}_\mathcal{A} = (1, \ldots, 1)^T \in \mathbb{R}^{|\mathcal{A}|}. \tag{4}$$

Then

$$u_\mathcal{A} = \mathbf{X}_\mathcal{A} w_\mathcal{A} \equiv \mathbf{X}_\mathcal{A} A_\mathcal{A} M_\mathcal{A}^{-1} \underline{1}_\mathcal{A} \tag{5}$$

is the unit vector making equal angles ($< 90°$) with the columns of $\mathbf{X}_\mathcal{A}$: this results from

$$\mathbf{X}_\mathcal{A}^T u_\mathcal{A} = A_\mathcal{A} \underline{1}_\mathcal{A}, \qquad \|u_\mathcal{A}\|^2 = 1. \tag{6}$$

Here is a more detailed description of the LARS algorithm: start with $\hat{\mu}^{(0)} = 0$. Suppose that $\hat{\mu}_\mathcal{A}$ is the current LARS estimate, and set

$$\hat{c} = \mathbf{X}'(Y - \hat{\mu}_\mathcal{A}) = \text{vector of current correlations.} \tag{7}$$

Here $\mathcal{A}$ is the current "active set":

$$\mathcal{A} = \{j \in \{1, \ldots, p\} : \ |\hat{c}_j| = \hat{C}\}, \tag{8}$$

$$\hat{C} = \max\{|\hat{c}_j| : \ 1 \le j \le p\}. \tag{9}$$

Set $s_j = \text{sign}\{\hat{c}_j\}$, $j \in \mathcal{A}$, and compute $\mathbf{X}_\mathcal{A}$, $A_\mathcal{A}$, $u_\mathcal{A}$ as in (1) - (5), and let

$$a \equiv \mathbf{X}' u_\mathcal{A}.$$

Then the next step of LARS updates $\mu_\mathcal{A}^{(m)}$ to $\mu^{(m+1)}$ as

$$\mu_{\mathcal{A}_{m+1}}^{(m+1)} = \mu_\mathcal{A}^{(m)} + \hat{\gamma} u_\mathcal{A} \tag{10}$$

where

$$\hat{\gamma} \equiv \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j} \right\}; \tag{11}$$

here $\min^+$ means that the minimum is over only the positive components in each choice of j.

Here is an interpretation of (10) and (11): define

$$\mu(\gamma) = \hat{\mu}_\mathcal{A} + \gamma u_\mathcal{A}, \qquad \gamma > 0. \tag{12}$$

Then the current correlations are given by

$$c_j(\gamma) = \mathbf{x}_j'(Y - \mu(\gamma)) = \hat{c}_j - \gamma a_j. \tag{13}$$

For $j \in \mathcal{A}$, (6) - (9) imply that

$$|c_j(\gamma)| = \hat{C} - \gamma A_\mathcal{A}; \tag{14}$$

note that this shows that all of the maximal absolute current correlations decline equally as γ increases. For $j \in \mathcal{A}^c$, equating (13) with (14) shows that $c_j(\gamma)$ equals the maximal value at $\gamma = (\hat{C} - \hat{c}_j)/(A_\mathcal{A} - a_j)$. Similarly for $-c_j(\gamma)$, the current correlation for the reversed covariate vector $-\mathbf{x}_j$ achieves maximality at $(\hat{C} + \hat{c}_j)/(A_\mathcal{A} + a_j)$. Therefore $\hat{\gamma}$ in (11) is the smallest positive value of γ such that some new index $\hat{j}$ joins the active set; $\hat{j}$ is the minimizing index in (11), the new active set $\mathcal{A}^{(m+1)} = \mathcal{A}^{(m)} \cup \{\hat{j}\}$, and the new maximum absolute correlation is $\hat{C}^{(m+1)} = \hat{C}^{(m)} - \hat{\gamma} A_{\mathcal{A}^{(m)}}$.

Now suppose that LARS has completed step $k-1$ with $\hat{\mu}_{k-1}$ the current predictor: the new active set $\mathcal{A}_k$ will have k elements, yielding $\mathbf{X}_k \equiv \mathbf{X}_{\mathcal{A}_k}$, $M_k \equiv M_{\mathcal{A}_k}$, $u_k = u_{\mathcal{A}_k}$. Let

$$\overline{Y}_k = \Pi(\underline{Y}|[\mathbf{X}_k]).$$

Since $\hat{\mu}_{k-1} \in [\mathbf{X}_k]$,

$$\overline{Y}_k = \hat{\mu}_{k-1} + X_k M_k^{-1} X_k(\underline{Y} - \hat{\mu}_{k-1}) = \hat{\mu}_{k-1} + \frac{\hat{C}_k}{A_k} u_k; \tag{15}$$

the last equality follows from (6) and the fact that the signed current correlations in $\mathcal{A}_k$ all equal $\hat{C}_k$,

$$X_k'(\underline{Y} - \hat{\mu}_{k-1}) = \hat{C}_k \underline{1}_{\mathcal{A}}. \tag{16}$$

But since u_k has length 1, (15) says that $\overline{Y}_k - \hat{\mu}_{k-1}$ has length

$$\overline{\gamma}_k \equiv \frac{\hat{C}_k}{A_k}.$$

Comparing these with (10) shows that $\hat{\mu}_k$ lies on the line from $\hat{\mu}_{k-1}$ to $\overline{Y}_k$:

$$\hat{\mu}_k - \hat{\mu}_{k-1} = \frac{\hat{\gamma}_k}{\overline{\gamma}_k}(\overline{Y}_k - \hat{\mu}_{k-1}). \tag{17}$$

Since $\hat{\gamma}_k$ is always less than $\overline{\gamma}_k$, $\hat{\mu}_k$ is closer to $\hat{\mu}_{k-1}$ than $\overline{Y}_k$. Thus it seems that the successive LARS estimate $\hat{\mu}_k$ always approach the ordinary least squares estimator $\overline{Y}_p$. But the exception is the very last step: $\mathcal{A}_p$ contains all p covariates, and hence (11) is not defined. By convention the algorithm takes $\hat{\gamma}_p = \overline{\gamma}_p = \hat{C}_p/A_p$, and this yields $\hat{\mu}_p = \overline{Y}_p$, with $\hat{\beta}_p =$ the ordinary LS estimate for the full set of p covariates.

LARS - lasso relationship: Let $\hat{\beta}$ be a lasso solution and let $\hat{\mu} = \mathbf{X}\hat{\beta}$. Then if $\beta_i \neq 0$,

$$\text{sign}(\hat{\beta}_j) = s_j = \text{sign}(\hat{c}_j). \tag{18}$$

this is proved in EHJT, lemma 8, page 434, and it also follows from the Karush-Kuhn Tucker conditions characterizing the solution given in Osborne, Presnell, and Turlach (2000a). This is what OPT (2000a), page 393, call *sign feasibility* of the solution. Although the LARS algorithm does not enforce the sign-feasibility restriction (18) required of lasso solutions, it can be modified to achieve this, as follows:

Suppose that we have completed a LARS step, with new active set $\mathcal{A}$ and the that the corresponding LARS estimate $\hat{\mu}_{\mathcal{A}}$ corresponds to a lasso solution $\hat{\mu} = \mathbf{X}\hat{\beta}$. Let

$$w_{\mathcal{A}} = A_{\mathcal{A}} M_{\mathcal{A}}^{-1} \underline{1}_{\mathcal{A}},$$

and define $\hat{d}$ to be the $p-$ vector which equals $s_j w_{\mathcal{A},j}$ for $j \in \mathcal{A}$ and zero elsewhere. Moving in the positive γ direction along the LARS line (12) we find that

$$\mu(\gamma) = \mathbf{X}\beta(\gamma), \qquad \text{where} \qquad \beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j,$$

for $j \in \mathcal{A}$. Thus $\beta_j(\gamma)$ will change sign at

$$\gamma_j = -\hat{\beta}_j / \hat{d}_j,$$

the first such change occurring at

$$\tilde{\gamma} \equiv \min_{\gamma_j > 0} \{\gamma_j\},$$

say for covariate $x_{\tilde{j}}$; take $\tilde{\gamma} = \infty$ if there is no $\gamma_j > 0$. If $\tilde{\gamma} < \hat{\gamma}$, then $\beta_j(\gamma)$ cannot be a lasso solution for $\gamma > \tilde{\gamma}$ since then sign-feasibility is violated: $\beta_{\tilde{j}}(\gamma)$ has changed sign while $c_{\tilde{j}}(\gamma)$ has not. (Note that the function $c_{\tilde{j}}(\gamma)$ cannot change sign within a single LARS step since $|c_{\tilde{j}}(\gamma)| = \hat{C} - \gamma A_{\mathcal{A}}$.) These considerations lead to the following Lasso modification of LARS:
Lasso modification: If $\tilde{\gamma} < \hat{\gamma}$, stop the ongoing LARS step at $\gamma = \tilde{\gamma}$ and remove $\tilde{j}$ from the calculation of the next equiangular direction. That is,

$$\hat{\mu}_{\mathcal{A}_+} = \hat{\mu}_{\mathcal{A}} + \tilde{\gamma} u_{\mathcal{A}}, \qquad \text{and} \qquad \mathcal{A}_+ = \mathcal{A} \setminus \{\tilde{j}\}.$$

Theorem 1. Under the Lasso modification, and assuming the "one at a time condition" discussed below, the LARS algorithm yields all Lasso solutions.