MATH 829: Introduction to Data Mining and Analysis
Least angle regression

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 29, 2016

Least angle regression (LARS)

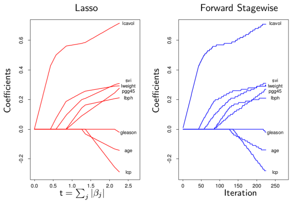
Recall the *forward stagewise approach* to linear regression:

1. Start with intercept $\overline{y}$, and centered predictors with coefficients initially all 0.
2. At each step the algorithm: identify the variable most correlated with the current residual.
3. Compute the simple linear regression coefficient of the residual on this chosen variable, and add it to the current coefficient for that variable.
4. Continued till none of the variables have correlation with the residuals.

- Greedy approach.
- However, the solution often looks similar to the lasso solution.
- Connection between the two methods?

Forward stagewise vs lasso

Example: Prostate cancer data (see ESL).



Efron et al., 2003

LARS

Least angle regression (LARS) is similar to forward stagewise, *but only enters "as much" of a predictor as it deserves.*

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.

ESL, Algorithm 3.2.

LARS (cont.)

- Let $\mathcal{A}_k$ be the current active set.
- $\beta_{\mathcal{A}_k}$ be the coefficients vectors at step k.
- Let $\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ denote the residual at step k.

Then, at step k, we move the coefficients in the direction

$$\delta_k = (\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k,$$

i.e., $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$.

Analysis of LARS

- How does the correlation between the predictors and the residuals evolve?

  $$\delta_k = (\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k \qquad \beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k.$$
- It remains the same for all predictors, and decreases monotonically.
- Indeed, suppose each predictor in a linear regression problem has equal correlation (in absolute value) with the response.

  $$\frac{1}{n}|\langle x_j, y\rangle| = \lambda \qquad j = 1, \dots, p.$$

  (Recall, we assume the predictors have been standardized.)
- Let $\hat{\beta}$ be the least-squares coefficients of y on X and let $u(\alpha) = \alpha X\hat{\beta}$ for $\alpha \in [0, 1]$.

Analysis of LARS (cont.)

We have $\frac{1}{n}|\langle x_j, y\rangle| = \lambda$ and $u(\alpha) = \alpha X\hat{\beta}$. Now,

$$\begin{aligned}\left(\frac{1}{n}|\langle x_j, y - u(\alpha)\rangle|\right)_{j=1}^{p} &= \frac{1}{n}|X^T(y - u(\alpha))| \\ &= \frac{1}{n}|X^T(y - \alpha X(X^TX)^{-1}X^Ty)| \\ &= \frac{1}{n}|(1-\alpha)X^Ty| \\ &= (1-\alpha)\lambda \cdot \mathbf{1}_{p\times 1}.\end{aligned}$$

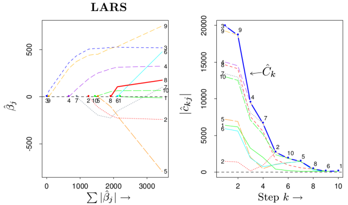
Therefore, the correlation between x_j and the residuals $y - u(\alpha)$ decreases linearly to 0.

In LARS, the parameter α is increased until a new variable becomes equally correlated with the residuals $y - u(\alpha)$.

The new variable is then added to the model, and a new direction is computed.

Analysis of LARS (cont.)

Example: $\hat{C}_k$ = current maximal correlation.



Efron et al., 2003.

Equiangular vector

Why " least angle" regression?

- Recal: $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$.
- Thus, $\hat{y}(\alpha) = X_{\mathcal{A}_k}\beta_{\mathcal{A}_k}(\alpha) = X_{\mathcal{A}_k}\beta_{\mathcal{A}_k} + \alpha \cdot X_{\mathcal{A}_k}\delta_k$.
- It is not hard to check that $u_k := X_{\mathcal{A}_k}\delta_k$ makes equal angles with the predictors in $\mathcal{A}_k$.

Indeed,

$$X_{\mathcal{A}_k}^T u_k = X_{\mathcal{A}_k}^T X_{\mathcal{A}_k}\delta_k = X_{\mathcal{A}_k}^T X_{\mathcal{A}_k}(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}X_{\mathcal{A}_k}^T r_k = X_{\mathcal{A}_k}^T r_k.$$

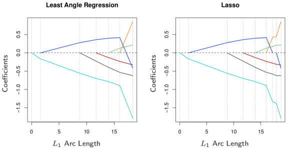
The entries of the vector $X_{\mathcal{A}_k}^T r_k$ are all the same since the predictors in $\mathcal{A}_k$ all have the same correlation with the residuals r_k (by construction).

Conclusion: u_k makes equal angles with the predictors in $\mathcal{A}_k$.

Problem: In general, given $v_1, \dots, v_k \in \mathbb{R}^n$, how do we find a vector that makes equal angles with $v_1, \dots, v_k$. When is this possible?

LARS and Lasso

- LARS is closely related to stepwise regression.
- There is also a connection to the Lasso.



ESL, Figure 3.15.

On the above figure, the lasso coefficient profiles are almost identical to those of LARS in the left panel, and differ for the first time when the blue coefficient passes back through zero.

LARS and Lasso (cont.)

The previous observation suggests the following LARS modification.

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.

ESL, Algorithm 3.2a.

Theorem: The modified LARS (lasso) algorithm (Algorithm 3.2a) yields the solution of the Lasso problem if variables appear/disappear "one at a time".

See Efron et al., *Least angle regression*, The Annals of Statistics, 2004.

Note: the theorem explains the *piecewise linear* nature of the lasso.

Consistency of the Lasso

- Recall: We proved before that the least-squares estimator is **consistent**, i.e.,

$$\hat{\beta}_n \to \beta$$

as the sample size n goes to infinity (under some assumptions).

- We now study analogous results for the lasso.
- **Assumptions:**
  - $X_1, \dots, X_p$ are (possibly dependent) random variables.
  - $|X_j| \leq M$ almost surely for some $M > 0$, $(j = 1, \dots, p)$.
  - $Y = \sum_{j=1}^p \beta_j^* X_j + \epsilon$ for some (unknown) constants β_j^*.
  - $\epsilon \sim N(0, \sigma^2)$ is independent of the X_j (σ^2 unknown).
  - Sparsity assumption (specified later).

Consistency of the Lasso (cont.)

We are given n iid observations

$$Z_i = (Y_i, X_{i,1}, \dots, X_{i,p})$$

of $(Y, X_1, \dots, X_p)$.

- Our goal is to recover $\beta_1^*, \dots, \beta_p^*$ as accurately as possible.
- Let
$$\hat{Y} = \sum_{j=1}^{p} \beta_j^* X_j,$$
the best predictor of Y if the true coefficients were known.
- Given $\tilde{\beta}_1, \dots, \tilde{\beta}_p$, let
$$\tilde{Y} = \sum_{j=1}^{p} \tilde{\beta}_j X_j.$$

Define the *mean square prediction error* by

$$\mathrm{MSPE}(\tilde{\beta}) = E(\hat{Y} - \tilde{Y})^2.$$

We will provide a bound on $\mathrm{MSPE}(\tilde{\beta})$ when $\tilde{\beta}$ is the lasso solution.

Consistency of the Lasso (cont.)

Given $K > 0$, let $\tilde{\beta}^K = (\tilde{\beta}_1^K, \dots, \tilde{\beta}_p^K)$ be the minimizer of

$$\sum_{i=1}^{n} (Y_i - \beta_1 X_{i,1} - \dots - \beta_p X_{i,p})^2$$

under the constraint

$$\sum_{i=1}^{p} |\beta_i| \leq K.$$

(The problem is equivalent to the lasso).

Theorem: Under the previous assumptions and assuming

$$\sum_{j=1}^{p} |\beta_j^*| \leq K \text{ for some } K > 0 \text{ (sparsity assumption)},$$

we have

$$\mathrm{MSPE}(\tilde{\beta}^K) \leq 2KM\sigma\sqrt{\frac{2\log(2p)}{n}} + 8K^2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

See Chatterjee, *Assumptionless consistency of the Lasso*, preprint, 2013.