Least angle regression (LARS)

Recall the forward stagewise approach to linear regression:

1. Start with intercept 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. Continue 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).

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 \( r = y - \hat{y} \), \( \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p = 0 \).
2. Find the predictor \( x_j \) most correlated with \( r \).
3. Move \( \hat{\beta}_j \) from 0 towards its least-squares coefficient \((x_j, r)\), until some other competitor \( x_k \) has as much correlation with the current residual as does \( x_j \).
4. Move \( \hat{\beta}_j \) and \( \hat{\beta}_k \) in the direction defined by their joint least squares coefficient of the current residual on \((x_j, x_k)\), until some other competitor \( x_i \) 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.
Let $A_k$ be the current active set.

Let $\beta_{A_k}$ be the coefficients vectors at step $k$.

Let $r_k = y - X_{A_k} \beta_{A_k}$ denote the residual at step $k$.

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

$$\delta_k = (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k,$$

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

### Analysis of LARS (cont.)

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

$$\left(\frac{1}{n}|\langle x_j, y - u(\alpha) \rangle|\right)^p = \frac{1}{n} |X^T(y - u(\alpha))|$$

$$= \frac{1}{n} |X^T(y - \alpha X(X^T X)^{-1} X^T y)|$$

$$= \frac{1}{n} |(1 - \alpha)X^T y|$$

$$= (1 - \alpha) \lambda \cdot 1_{p \times 1}.$$

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

In LARS, the parameter $\alpha$ 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.)

- How does the correlation between the predictors and the residuals evolve?
  $$\delta_k = (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k \quad \beta_{A_k}(\alpha) = \beta_{A_k} + \alpha \cdot \delta_k.$$  
  
- 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.

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

  (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 $\alpha \in [0, 1]$.

#### Example:

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

---

---

---

---
Equiangular vector

Why “least angle” regression?

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

Indeed,
\[
X^T_{Ak} u_k = X^T_{Ak} X_{Ak} \delta_k = X^T_{Ak} X_{Ak} (X^T_{Ak} X_{Ak})^{-1} X^T_{Ak} r_k = X^T_{Ak} r_k.
\]

The entries of the vector \( X^T_{Ak} r_k \) are all the same since the predictors in \( A_k \) all have the same correlation with the residuals \( r_k \) (by construction).

Conclusion: \( u_k \) makes equal angles with the predictors in \( A_k \).

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

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.

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


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:**

- \( X_1, \ldots, X_p \) are (possibly dependent) random variables.
- \( |X_j| \leq M \) almost surely for some \( M > 0 \) (\( j = 1, \ldots, p \)).
- \( Y = \sum_{j=1}^p \beta_j^* X_j + \epsilon \) for some (unknown) constants \( \beta_j^* \).
- \( \epsilon \sim N(0, \sigma^2) \) is independent of the \( X_j \) (\( \sigma^2 \) unknown).
- Sparsity assumption (specified later).
We are given \( n \) iid observations
\[ Z_i = (Y_i, X_{i,1}, \ldots, X_{i,p}) \]
of \((Y, X_1, \ldots, X_p)\).

- Our goal is to recover \( \beta_1^*, \ldots, \beta_p^* \) 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 \( \tilde{\beta}_1, \ldots, \tilde{\beta}_p \), let
\[ \tilde{Y} = \sum_{j=1}^{p} \tilde{\beta}_j X_j. \]

Define the mean square prediction error by
\[ \text{MSPE}(\hat{\beta}) = E(\hat{Y} - \tilde{Y})^2. \]

We will provide a bound on \( \text{MSPE}(\hat{\beta}) \) when \( \hat{\beta} \) is the lasso solution.

**Theorem:** Under the previous assumptions and assuming
\[ \sum_{j=1}^{p} |\beta_j^*| \leq K \]
for some \( K > 0 \) (sparsity assumption),
we have
\[ \text{MSPE}(\hat{\beta}^K) \leq 2K M \sigma \sqrt{\frac{2 \log(2p)}{n}} + 8K^2 M^2 \sqrt{\frac{2 \log(2p^2)}{n}}. \]