Least Angle Regression and Its Variants

Course Project of STAT 5361 Statistical Computing

*Biju Wang**

12/15/2018

Abstract

This project will show the details of least angle regression algorithm. Based on its idea, the variants of this algorithm is also presented.

1 Indroduction

Least Angle Regression (LARS) is an algorithm for fitting linear regression models to high dimensional data (Efron et al. 2004). Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients.

This algorithm is similar to forward stepwise regression (Hastie, Tibshirani, R., and Friedman, J. 2008). While forward stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set and then updates the least square fit to include all the active variables. Least angle regression only enters "as much" of a predictor as it deserves. At the first step, it identifies the variable most correlated with the response. Rather than fit this variable completely, LARS moves the coefficient of this variable continuously toward its least squares value. As soon as another variable "catches up" in terms of correlation with the residual, the process is paused. The second variable then joins the active set, and their coefficients are moved together in a way that keeps their correlations tied and decreasing. This process is continued until all the variables are in the model and ends at the full least-squares fit. The whole procedure will terminate after $\min\{N-1, p\}$ steps where N is the number of observations and p is the size of covariates.

In this project, I will write LARS algorithm for simulated data and compare the outcomes from my codes with the ones from package **lars**. I will also present two variants of LARS.

2 Least Angle Regression

The details of least angle regression is in algorithm 1 (Hastie, Tibshirani, R., and Friedman, J. 2008).

*bijuwang@uconn.edu

Algorithm 1 Least Angle Regression

1: Centralize the predictors to have mean zero. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}, \beta_1 = \cdots = \beta_p = 0$
2: Find the predictors $\mathbf{x}_j$ most correlated with $\mathbf{r}$
3: Move β_j from 0 towards its least-squares coefficient, 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

You may notice that in introduction, the idea of LARS is to make the correlation coefficients the same between residual and the predictor vectors at all times, while in the algorithm we say we should move toward the least square fitted vector. They are colored by red. One question naturally comes out, are these two directions the same? The following two lemmas answer this question.

Lemma 2.1. *Suppose we have three n-dimensional vecotrs* $\mathbf{y}, \mathbf{x}_1$ *and* $\mathbf{x}_2$ *with each of them has mean* 0. $\hat{\mathbf{y}}$ *is the projection of* $\mathbf{y}$ *on the space span*$\{\mathbf{x}_1, \mathbf{x}_2\}$ *which is also the lease square fitted vector.* $\lambda \in [0, 1]$. *We have the condition*

$$corr(\mathbf{y}, \mathbf{x}_1) = corr(\mathbf{y}, \mathbf{x}_2)$$

Then

$$corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1) = corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_2)$$

Proof. We use $\bar{\mathbf{y}}, \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$ to represent the mean vector for $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2$, P to represent the projection matrix corresponding to span$\{\mathbf{x}_1, \mathbf{x}_2\}$. And $\mathbf{y} \cdot \mathbf{x}_1$ denotes the inner product of two vectors.
In order to prove $corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1) = corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_2)$, we only need to show

$$\frac{[(\mathbf{y} - \lambda\hat{\mathbf{y}}) - (\bar{\mathbf{y}} - \lambda\bar{\hat{\mathbf{y}}})] \cdot (\mathbf{x}_1 - \bar{\mathbf{x}}_1)}{\sqrt{(\mathbf{x}_1 - \bar{\mathbf{x}}_1) \cdot (\mathbf{x}_1 - \bar{\mathbf{x}}_1)}} = \frac{[(\mathbf{y} - \lambda\hat{\mathbf{y}}) - (\bar{\mathbf{y}} - \lambda\bar{\hat{\mathbf{y}}})] \cdot (\mathbf{x}_2 - \bar{\mathbf{x}}_2)}{\sqrt{(\mathbf{x}_2 - \bar{\mathbf{x}}_2) \cdot (\mathbf{x}_2 - \bar{\mathbf{x}}_2)}}$$

Since we have $\bar{\mathbf{y}} = \bar{\mathbf{x}}_1 = \bar{\mathbf{x}}_2 = \mathbf{0}$, thus span$\{\mathbf{x}_1, \mathbf{x}_2\} \perp$ span$\{\mathbf{1}\}$. But $\hat{\mathbf{y}} = P\mathbf{y} \in$ span$\{\mathbf{x}_1, \mathbf{x}_2\}$, then $\bar{\hat{\mathbf{y}}} = \mathbf{0}$. The conclusion can be simplified as

$$\frac{(\mathbf{y} - \lambda\hat{\mathbf{y}}) \cdot \mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} = \frac{(\mathbf{y} - \lambda\hat{\mathbf{y}}) \cdot \mathbf{x}_2}{\sqrt{\mathbf{x}_2 \cdot \mathbf{x}_2}} \tag{1}$$

While we already know that

$$\frac{\mathbf{y} \cdot \mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} = \frac{\mathbf{y} \cdot \mathbf{x}_2}{\sqrt{\mathbf{x}_2 \cdot \mathbf{x}_2}} \tag{2}$$

Then

$$\frac{\mathbf{y} \cdot P'\mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} = \frac{\mathbf{y} \cdot P'\mathbf{x}_2}{\sqrt{\mathbf{x}_2 \cdot \mathbf{x}_2}} \Rightarrow \frac{P\mathbf{y} \cdot \mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} = \frac{P\mathbf{y} \cdot \mathbf{x}_2}{\sqrt{\mathbf{x}_2 \cdot \mathbf{x}_2}} \Rightarrow \frac{\hat{\mathbf{y}} \cdot \mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} = \frac{\hat{\mathbf{y}} \cdot \mathbf{x}_2}{\sqrt{\mathbf{x}_2 \cdot \mathbf{x}_2}} \tag{3}$$

Now combine equation 2 and 3 together we can get equaiton 1. □

Lemma 2.1 has an intuitive meaning. Since we assume the mean of each vector is 0, in this case correlation coefficient of two vectors becomes the consine of angle of the two vectors. If $\mathbf{y}$ has the

same angle with $\mathbf{x}_1$ and $\mathbf{x}_2$, so does $\hat{\mathbf{y}} = P\mathbf{y}$ and any vector belongs to the plane determined by $\mathbf{y}$ and $\hat{\mathbf{y}}$.

Lemma 2.1 can be easily extended to multiple vectors. Here, we only state the lemma without proof.

Lemma 2.2. *Suppose we have $k+1$ n-dimensional vecotrs $\mathbf{y}, \mathbf{x}_1, \cdots, \mathbf{x}_k$ with each of them has mean 0. $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ on the space span$\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ which is also the lease square fitted vector. $\lambda \in [0, 1]$. We have the condition*

$$corr(\mathbf{y}, \mathbf{x}_1) = \cdots = corr(\mathbf{y}, \mathbf{x}_k)$$

Then

$$corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1) = \cdots = corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_k)$$

Since we now find a direction $\hat{\mathbf{y}}$ in the space span$\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ which has the same correlation coefficient with $\mathbf{x}_1, \cdots, \mathbf{x}_k$. We start from vector $\mathbf{0}$ and move along the direction $\hat{\mathbf{y}}$ until we reach $\hat{\mathbf{y}}$. The $|corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1)| = \cdots = |corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_k)|$ will monotonically decrease. In other words, $|corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1)| = \cdots = |corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_k)|$ is a monotone function as to $\lambda \in [0, 1]$. We accept this fact without proof.

Suppose we have another vector $\mathbf{x}_{k+1}$ with mean 0 and $|corr(\mathbf{y}, \mathbf{x}_{k+1})| < |corr(\mathbf{y}, \mathbf{x}_1)| = \cdots = |corr(\mathbf{y}, \mathbf{x}_k)|$. Our question is how to choose λ such that $|corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_{k+1})| = |corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_1)| = \cdots = |corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_k)|$? To answer this, we only need to solve the following equation

$$\left| \frac{(\mathbf{y} - \lambda\hat{\mathbf{y}}) \cdot \mathbf{x}_{k+1}}{\sqrt{\mathbf{x}_{k+1} \cdot \mathbf{x}_{k+1}}} \right| = \left| \frac{(\mathbf{y} - \lambda\hat{\mathbf{y}}) \cdot \mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} \right|$$

We can get

$$\lambda = min^+ \left\{ \frac{\mathbf{y} \cdot \left(\frac{\mathbf{x}_{k+1}}{\sqrt{\mathbf{x}_{k+1} \cdot \mathbf{x}_{k+1}}} - \frac{\mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} \right)}{\hat{\mathbf{y}} \cdot \left(\frac{\mathbf{x}_{k+1}}{\sqrt{\mathbf{x}_{k+1} \cdot \mathbf{x}_{k+1}}} - \frac{\mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} \right)}, \frac{\mathbf{y} \cdot \left(\frac{\mathbf{x}_{k+1}}{\sqrt{\mathbf{x}_{k+1} \cdot \mathbf{x}_{k+1}}} + \frac{\mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} \right)}{\hat{\mathbf{y}} \cdot \left(\frac{\mathbf{x}_{k+1}}{\sqrt{\mathbf{x}_{k+1} \cdot \mathbf{x}_{k+1}}} + \frac{\mathbf{x}_1}{\sqrt{\mathbf{x}_1 \cdot \mathbf{x}_1}} \right)} \right\}$$

where min^+ means we get the minimal positive value from the set.

Here, we need to mention that the requirement that all the vectors have mean 0 ($\bar{\mathbf{y}} = \bar{\mathbf{x}}_1 = \cdots = \bar{\mathbf{x}}_k = \mathbf{0}$) or to say $\mathbf{y} \perp \mathbf{1}, \mathbf{x}_1 \perp \mathbf{1}, \cdots, \mathbf{x}_k \perp \mathbf{1}$ plays an important role. We should note that in least angle regression algorithm, the key is to find a vector $\mathbf{x}$ in the space span$\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ such that $corr(\mathbf{y} - \mathbf{x}, \mathbf{x}_1) = \cdots = corr(\mathbf{y} - \mathbf{x}, \mathbf{x}_k)$ which is equavalent to finding a vector $\mathbf{x}$ in the sapce span$\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ such that $corr(\mathbf{x}, \mathbf{x}_1) = \cdots = corr(\mathbf{x}, \mathbf{x}_k)$. In general, $\mathbf{x}$ is not necessary equal to $\hat{\mathbf{y}}$. But if we have the requirement above, then $\mathbf{x} = \hat{\mathbf{y}}$. This makes us consider two variants of LARS which can be found in section 3.

Since if the mean of each vector is 0, then the correlation coefficient which measures the association between two vectors is the consine value of the angle of the two vectors. We want to find the most associated preditor vector with response vector which means absolute value of consine value of their angle is the smallest. The is the source of the name "least angle".

Another thing needs to be noticed in algorithm 1 is that predictor vectors are centralized. Strictly speaking, after we did some transformations, the linear model should change. But in the situation here, what is the relationship between the original linear model and the new linear model? Now considering the following two linear models

$$\mathbf{y} = \mu\mathbf{1} + \beta_1\mathbf{x}_1 + \cdots + \beta_p\mathbf{x}_p + \boldsymbol{\varepsilon} \tag{4}$$

$$\mathbf{y} = \mu'\mathbf{1} + \beta_1'(\mathbf{x}_1 - \bar{\mathbf{x}}_1) + \cdots + \beta_p'(\mathbf{x}_p - \bar{\mathbf{x}}_p) + \boldsymbol{\varepsilon} \tag{5}$$

Obviously, $\text{span}\{\mathbf{1}, \mathbf{x}_1, \cdots, \mathbf{x}_p\} = \text{span}\{\mathbf{1}, \mathbf{x}_1 - \bar{\mathbf{x}}_1, \cdots, \mathbf{x}_p - \bar{\mathbf{x}}_p\}$. The projection of $\mathbf{y}$ on these two spaces must be the same. Therefore

$$\hat{\mu}\mathbf{1} + \hat{\beta}_1\mathbf{x}_1 + \cdots + \hat{\beta}_p\mathbf{x}_p = \hat{\mu}'\mathbf{1} + \hat{\beta}_1'(\mathbf{x}_1 - \bar{\mathbf{x}}_1) + \cdots + \hat{\beta}_p'(\mathbf{x}_p - \bar{\mathbf{x}}_p)$$

The relationship between two sets of least square estimates are

$$\begin{cases} \hat{\beta}_1 = \hat{\beta}_1' \\ \vdots \\ \hat{\beta}_p = \hat{\beta}_p' \\ \hat{\mu} = \hat{\mu}' - \hat{\beta}_1'\bar{x}_1 - \cdots - \hat{\beta}_p'\bar{x}_p \end{cases}$$

Also notice that, in model (5), $\text{span}\{\mathbf{1}\} \perp \text{span}\{\mathbf{x}_1 - \bar{\mathbf{x}}_1, \cdots, \mathbf{x}_p - \bar{\mathbf{x}}_p\}$. Every vector in the orthogonal space of span{1} has mean 0, if we take the residual of $\mathbf{y}$ after it projects on $\text{span}\{\mathbf{1}\}$, then everything is all set and we can apply least angle regression focusing on the space $\text{span}\{\mathbf{y} - \bar{\mathbf{y}}, \mathbf{x}_1 - \bar{\mathbf{x}}_1, \cdots, \mathbf{x}_p - \bar{\mathbf{x}}_p\}$. The paths of $\hat{\beta}_1', \cdots, \hat{\beta}_p'$ can represent the paths of $\hat{\beta}_1, \cdots, \hat{\beta}_p$.

3 Two Variants of LARS

As we discussed in section 2, they may exist two variants of LARS.

One variant is we insist on finding $\mathbf{x}$ has the equal correlation coefficient with $\mathbf{x}_1, \cdots, \mathbf{x}_k$, but in this case the final SSE may not be the least since $\mathbf{x} \neq \hat{\mathbf{y}}$. The other variant is we use the projection of y on $\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ and use $corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \hat{\mathbf{y}})$ to represent the association between $\mathbf{y} - \lambda\hat{\mathbf{y}}$ and $\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ and find λ such that $corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \hat{\mathbf{y}}) = corr(\mathbf{y} - \lambda\hat{\mathbf{y}}, \mathbf{x}_{k+1})$. The advantage of this variant is SSE is the least square SSE.

4 Simulation

We first use the following code to generate the data.

```
data.simu <- function(n = 100, p = 4, coefs, sd.x = 1, snr = 5){

  X <- matrix(rnorm(n * p, 0, sd.x) ,nrow = n, ncol = p)
  coefs <- runif(p + 1, 1, 4) * sample(c(-1, 1), size = p + 1, replace = TRUE)

  y <- cbind(rep(1, n), X) %*% coefs
  y.mean <- y

  sd.e <- sd(y.mean) / snr
  e <- rnorm(n, 0, sd.e)
  y <- y.mean + e

  list(y=y, X=X, y.mean=y.mean, coefs=coefs)
}
```

We use the following code to implement the least angle algorithm.

```
lambda.cal <- function(y, y.hat, x1, xk){

  candidate1 <- sum(y * (xk / sqrt(sum(xk^2)) - x1 / sqrt(sum(x1^2)))) /
    sum(y.hat * (xk / sqrt(sum(xk^2)) - x1 / sqrt(sum(x1^2))))
  candidate2 <- sum(y * (xk / sqrt(sum(xk^2)) + x1 / sqrt(sum(x1^2)))) /
    sum(y.hat * (xk / sqrt(sum(xk^2)) + x1 / sqrt(sum(x1^2))))

  lambda <- c(candidate1, candidate2)
  lambda <- lambda[lambda >= 0]
  lambda <- min(lambda)

  lambda
}

mylars <- function(y, X){

  X <- scale(X, scale = F)
  y <- y - mean(y)
  n <- nrow(X)
  p <- ncol(X)

  #use step.path to store which variabe enters in each step
  step.path <- c()
  lambda.path <- c()
  index <- 1:p

  select <- which.max(abs(cor(y, X)))
  step.path <- c(step.path, index[select])
  index <- index[-select]

  for (i in 1:(min(n - 1, p) - 1)) {

    X.new <- X[,step.path]
    y.hat <- lm(y ~ X.new - 1)$fitted.values

    lambda <- c()
    for (j in 1:(p - length(step.path))) {

       if(length(step.path) == 1){
         x1 <- X[,step.path]
       }
      else{x1 <- X[,step.path][,1]}

      if(p - length(step.path) == 1){
        xk <- X[,-step.path]
      }
```

```
      else{xk <- X[,-step.path][,j]}

       lambda[j] <- lambda.cal(y, y.hat, x1, xk)
  }

    select <- which.min(lambda)[1]
    lambda.path <- c(lambda.path, min(lambda))
    step.path <- c(step.path, index[select])
    index <- index[-select]
    y <- y - min(lambda) * y.hat

  }

  list(step.path=step.path, lambda.path=lambda.path)
}
```

There is a package called **lars** can implement the algorithm. Now let's compare if the function we coded gives the same results with the ones obtained from function lars().

```
set.seed(1)
data <- data.simu()
mylars(data$y, data$X)
```

```
## $step.path
## [1] 2 3 4 1
##
## $lambda.path
## [1] 0.09120289 0.40597861 0.65805845
```

```
library(lars)
```

```
## Loaded lars 1.2
```

```
lars(data$X, data$y, type = "lar")
```

```
##
## Call:
## lars(x = data$X, y = data$y, type = "lar")
## R-squared: 0.952
## Sequence of LAR moves:
##
## Var  2 3 4 1
## Step 1 2 3 4
```

We can see the entering sequence of variables are the same.

References

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2).

Hastie, T., Tibshirani, R., and Friedman, J. 2008. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.