Generalized Least Angle Regression

George R. Terrell

Statistics Department

Virginia Polytechnic Institute and State University

Abstract:

Least-angle regression (LARS) is an algorithm that regularizes models by simultaneously selecting variables and shrinking predictions. It characterizes possible models as those for which the residual information about the dependent variable is equal for all active independent variables. We show that this algorithm may be generalized to regression problems for which the criterion of good fit (such as the log-likelihood) is twice differentiable and convex.

Key words: linear regression, model selection, LASSO

I. Introduction. The Least Angle Regression (LARS, Efron, et. al. 2004) algorithm is a modern regularized version of classical forward selection in the problem of parsimonious variable selection in least squares linear regression. By formulating this as requiring us to gradually shrink the maximum of a *t*-statistic for introducing each of the variables, we generalize LARS to a much larger class of linear models.

II. Forward Selection. We shall consider the usual linear model for predicting an *n*-vector of observations **y** by $\hat{\mathbf{y}} = \mathrm{X}\boldsymbol{\beta}$, where X is an *n* by *k* matrix each of whose *k* columns consists of the values of an independent variable possibly useful for prediction of **y**. It is easy to check that finding the usual least-squares solution is equivalent to solving the regression problem

$$\min_{\hat{\mathbf{y}}} \hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} \text{ subject to } \mathrm{X}^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} .$$

The *k* linear constraints are the *normal* equations.

When the number of independent variables *k* is large, there are familiar problems with this solution. The predictions tend to have high variance, and the prediction equation tends to be uninterpretable. There are two common approaches to *regularizing* the solution: we may *shrink* the solutions to smaller $\hat{\mathbf{y}}$ to reduce the variability of prediction, as in ridge regression; and we may seek a *parsimonious* solution using only a subset of the independent variables, as in step-wise variable selection. Least-angle regression (LARS) is an elegant way to pursue the two goals simultaneously.

So that we shall not have to worry about the overall mean of **y**, we shall center the independent variables by assuming $\mathrm{X}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$. We shall consider the potential further contribution of individual independent variables for a given proposed solution $\boldsymbol{\beta}$ by evaluating the statistics $t_j = \left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)^{-1/2}\mathbf{x}_j^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}})/\sigma$; where $\mathbf{x}_j$ is the column of the design matrix corresponding to independent variable *j*, and where σ is usually estimated from the data. Note that we have the least-squares solution when all $t_j = 0$. When, under the

usual error assumptions, t_j is surprisingly different from 0, we may conclude that the value of the coefficient β_j should be changed.

For purposes of variable selection, then, our index of how far $\boldsymbol{\beta}$ is from least squares will be $t = \max_j |t_j|$; if it is small enough to be readily explained by chance, then there is no statistical case for further adjustments to $\boldsymbol{\beta}$. For example, the classical method of *forward selection* starts with $\boldsymbol{\beta} = \mathbf{0}$, and at each stage selects a variable *j* so that $|t_j| = t$ and solves the normal equation for β_j so that then $t_j = 0$. We then recalculate $t = \max_j |t_j|$ at this new solution. The next stage selects another variable *l* such that $|t_l| = t$; and recalculates β_j and β_l using the two normal equations. We successively add variables in this way until *t* is small enough to be explained by chance. The result is therefore parsimonious; because unselected variables still have 0 coefficients. Our experience is that forward selection may be unstable: wildly different solutions may have similar values of *t*.

II. Least Angle Regression. LARS instead provides solutions to a modified regression problem. Choose a value of *t* and solve

$$\min_{\hat{\mathbf{y}}} \hat{\mathbf{y}}^{\mathbf{T}}\hat{\mathbf{y}} \text{ such that for all } j,\ |t_j| \le t.$$

Notice that for $t > 0$, this problem has a shrunken solution; since it relaxes the normal equation conditions $t_j = 0$, and therefore permits the squared length of $\hat{\mathbf{y}}$ to be smaller. But in addition

Proposition 1: If for any variable we have $|t_l| < t$, the solution will have $\beta_l = 0$.

Thus, we may have a parsimonious solution. As a consequence, for all $\beta_j \neq 0$, we have $|t_j| = t$.

Note that each t_j is proportion to the cosine of the angle between the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ and the vector of values of that independent variable $\mathbf{x}_j$. Thus a solution of the LARS problem consists of a choice of variables whose angles with the residual vector are smallest in size, justifying the name.

A major appeal of the LARS problem is that its solutions meet the LASSO (Tibshirani 1995) criteria. Another is that there is an algorithm, essentially a forward variable selection process, that evaluates all problems for different *t* simultaneously. It constructs a continuous path through $\boldsymbol{\beta}$ space for decreasing values of *t*. It will collect **active** variables for which $|t_j| = t$, one at a time, as *t* decreases, until we reach the least-squares solution at $t = 0$.

(0) Initially set $\boldsymbol{\beta} = \mathbf{0}$, and declare all independent variables **inactive.**

(1) Calculate $t = \max_j |t_j|$, and declare all variables *j* for which $|t_j| = t$ to be **active**. (Thus for all inactive variables *l*, $|t_l| < t$.)

(2) Move $\boldsymbol{\beta}$ in a direction **d** which is a linear combination of the active variables determined by the condition that it remains true that for all active variables *j*, $|t_j| = t$, and *t* decreases.

(3) At the first point in this process where it becomes true that for some *l* inactive, $|t_l| = t$, go to step (1).

The algorithm terminates at the point at which your criterion for a good model is achieved, or $t = 0$.

For LARS, the appropriate direction **d** is found as follows: let $\mathrm{X_J}$ be a matrix consisting of the columns $\mathbf{x}_j$ for the active variables. Then the vector of active *t*s is $\mathbf{t} = \mathrm{D}\left[\left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)^{-1/2}\right]\mathrm{X_J^T}(\mathbf{y} - \hat{\mathbf{y}})$ where D is a diagonal matrix indexed by *j*. If we shift the active regression coefficients by an amount **d**, then $\hat{\mathbf{y}}$ changes by $\mathrm{X_J}\mathbf{d}$; therefore **t** changes by $-\mathrm{D}\left[\left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)^{-1/2}\right]\mathrm{X_J^T X_J}\mathbf{d}$. Our requirement is that the direction keep all active |*t*|s equal to their maximum value; so we set the previous expression equal to the vector $\left(\mathrm{sgn}t_j\right)$, and solve to get $\mathbf{d} = -\left(\mathrm{X_J^T X_J}\right)\mathrm{D}\left[\left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)^{1/2}\right]\left(\mathrm{sgn}t_j\right)$.

The path is therefore continuous and piecewise linear, and terminates at the least squares solution.

Proposition 2: The algorithm provides the LARS solution for each value of *t* on the path.

IV. Generalizing LARS. We now wish to extend our selection principal to linear regression models evaluated by other criteria than least squares. We will usually require a log-likelihood of the form $l(\boldsymbol{\beta}) = l(\mathrm{X}\boldsymbol{\beta})$ such as generalized linear models. Assume *l* is convex and twice-differentiable. (Note that the algorithm we will develop will also work for *l* any appropriate measure of fit, such as a robust M-estimator.) From the point of view of a tentative solution vector $\hat{\boldsymbol{\beta}}$, we might measure the potential improvement in the fit from modifying a coefficient $\hat{\beta}_j$ as follows: Let $\hat{\boldsymbol{\beta}}_j$ be the solution to the maximum likelihood problem when β_j is allowed to vary but the other $\beta_{j\bullet}$ are held constant. Now let a (dimensionless) index of improvement be $2\left[l\left(\hat{\boldsymbol{\beta}}_j\right) - l\left(\hat{\boldsymbol{\beta}}\right)\right]$, twice the available increase in log-likelihood.

Our goal here will be to evolve a continuous improvement path, so the fact that this is not a statistic defined by local properties of the likelihood at $\hat{\boldsymbol{\beta}}$ will be inconvenient. Instead, if *l* is log-concave and twice differentiable, consider the empirical Rao (1948) score $t_j = \left(-\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \beta_j^2}\right)^{-1/2} \frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \beta_j}$. If our tentative solution is correct, this has asymptotically a standard normal distribution and is asymptotically equal to the signed square-root of the log-likelihood ratio suggested in the last paragraph. Note that for the classical linear regression model with i.i.d. normal errors, the general *t* is identical to that defined in Section II.

We will search for good parsimonious models using a Generalized Least Angle Regression (GLARe) algorithm that follows the steps in Section III, with the generalized definition of t_j. Since a GLARE path is continuous and piecewise smooth but not (except for the least-squares case) piecewise linear, constructing the path is more involved than with least-angle regression. We shall here propose an algorithm for updating the path in small increments, using partial derivatives. A given point on the path will be characterized by the vector $\boldsymbol{\beta}$ of current estimates of the coefficients of the *active* independent variables. We wish to update it to a further, nearby vector along the path $\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The condition for being on the path is that the vector $\mathbf{t}^* = \mathbf{t} - \boldsymbol{\delta}$ of active scores will still meet the condition all $\left|t_j^*\right| = t^*$ as t^* shrinks toward zero. Thus $\delta_j = \delta \mathrm{sgn}(t_j)$ where δ is a small decrement. Our procedure will be to select δ and approximate the corresponding ε for each step along the GLARE path.

We differentiate the vector $\mathbf{t}$ with respect to the vector β to get the square matrix U where $U_{jm} = \left[-\frac{\partial^2 l}{\partial \beta_j^2}\right]^{-1/2} \frac{\partial^2 l}{\partial \beta_j \partial \beta_m} + \frac{1}{2}\left[-\frac{\partial^2 l}{\partial \beta_j^2}\right]^{-3/2} \frac{\partial l}{\partial \beta_j} \frac{\partial^3 l}{\partial \beta_j^2 \partial \beta_m}$. Then for δ small we have the approximation $\mathrm{U}\boldsymbol{\varepsilon} \cong -\boldsymbol{\delta}$. This gives the update rule for the estimated $\boldsymbol{\beta}$ coefficients $\boldsymbol{\varepsilon} \cong -\mathrm{U}^{-1}\boldsymbol{\delta}$.

For sufficiently small decrements to t (δ) this will give a good approximation to the GLARE path. It has the disadvantage that errors accumulate as we proceed, so that we require a δ so small that achieving an accurate final solution may be computationally burdensome. A slight modification will lead to an algorithm for which errors do not accumulate, and requiring less computation. At a given point on the path the vector $\mathbf{t}$ may not have precisely all $|t_j| = t$, but close enough for practical purposes. We will approximate a step along the GLARE path directed toward a vector $\mathbf{t}^*$ for which $\left|t_j^*\right| = t^*$, which is slightly decremented, so that $t - t^*$ is small. We will approximate that solution by $\boldsymbol{\beta}^* \cong \boldsymbol{\beta} + \mathrm{U}^{-1}\left(\mathbf{t} - \mathbf{t}^*\right)$. Since for any $\boldsymbol{\beta}$, $\mathbf{t}$ is calculated exactly, errors do not accumulate, and the overall computation of the path is stable.

The modified computation allows us, then, to use longer increments, since they now need only be sufficiently small that the endpoint of each step is acceptably close to the exact GLARE path. This substantially speeds computation. It raises another issue, though. Recall that whenever we reach a point at which for some inactive variable $|t_k| = t$, we add it to the active set and begin to increment β_k. In the first version of our procedure where increments are tiny, we can do exactly that. But in the improved version, increments may be large enough that at some step we will check that for the first time $|t_k| > t$ by a small but not negligible amount. We add k to the active set, and use the new formula to step to a new point $\mathbf{t}^*$ on the path for which to sufficient accuracy $\left|t_k^*\right| = t^*$ as well as for earlier enrollees $\left|t_j^*\right| = t^*$. We then proceed along the new segment.

V. Nuisance Parameters. In the LARS procedure, we would usually be fitting a model $\hat{y} = \beta_0 + \mathrm{X}^{\mathrm{C}}\boldsymbol{\beta}$ where X^{C} is a design matrix with centered columns, $\mathrm{X}^{\mathrm{CT}}\mathbf{1} = \mathbf{0}$. The mean parameter β_0 is not to be selected, but is present in every tentative fit. This is

no problem, because its value never affects the estimates of the slope coefficients, and is for least squares always $\hat{\beta}_0 = \bar{y}$. And though we estimate error variance σ^2 differently for each model, the same estimate is used in the *t*s at a given point on the path; so decisions about the sequence of variable selection are not affected by its value.

But in the non-least-squares case, estimates of β_0 interact with estimates of the β_j, and centering does not resolve the problem. The GLARE algorithm readily adapts to this circumstance: treat β_0 as always active, with the following differences: For each tentative model, starting with the null model $\boldsymbol{\beta} = \mathbf{0}$, estimate β_0 by requiring the maximum likelihood condition $\frac{\partial l}{\partial \beta_0} = 0$. Thus, it will always be the case that $t_0 = 0$, anywhere along the GLARE path. In our update algorithm for the next step along the path, we augment the vector $\boldsymbol{\delta}$ with an index-0 coordinate whose value is $\delta_0 = 0$ (since $t_0 = 0$ should not change). The U matrix now has one more row and column, and the estimate of β_0 is updated to approximately its new maximum likelihood solution. The stable modification of our algorithm is extended in the same way.

Centering of the design matrix is still useful here, (though unnecessary) because it allows β_0 to have a consistent interpretation, and to change slowly.

This modification also applies to control variables, which our design requires always to be in any tentative model. We fit them by maximum likelihood at $\boldsymbol{\beta} = \mathbf{0}$ for the variables available for selection, and update their coefficients at each step as we did for β_0.

VI. Example. Harry Khamis reports Alcohol, Cigarette, and Marijuana usage of Dayton high school seniors (reported in Agresti (2013), p. 381). The table has been collapsed over Race, because so few non-Whites appeared, and their reports were consistent with Whites:

		Female		Male	
	Cig/Mar	Yes	No	Yes	No
Yes	Yes	428	291	483	247
Alcohol	No	15	237	29	219
No	Yes	1	18	2	25
	No	1	129	1	150

We attempt to build a Poisson loglinear model using parametrization 1/–1 for the two levels of each variable. The variables available for selection were then all possible main effects and associations, for a total of $p = 15$.

At the null point, $t = 136$, the main effect A(lcohol) became active.

At the point $t = 24.8$, the association AC(igarettes) became active.

At the point $t = 21.4$, the association CM(arijuana) became active.

At the point $t = 9.4$, the main effect M became active.

At the point $t = 3.3$, the main effect C and the association MAS(ex) became almost simultaneously active.

At the point $t = 2.7$, the association MA became active.

At the point $t = 2.1$, the association MAC became active.

Stopping at the naïve stopping point $t = 1.0$, we had found the log-linear model

Log(count) = 4.13 + 1.00A + .60AC + .57MC – .57M + .11C – .05MAS + .20MA + .05MAC.

Appendix.

Proof of Proposition 1: Let l be an inactive variable so that $|t_l| < t$. Then we can decompose $\hat{\mathbf{y}} = \mathbf{y}^* + \beta_l \mathbf{x}_l^*$ where y* is in the linear span of the active variables and $\mathbf{x}_l^*$ is the projection of $\mathbf{x}_l$ into the linear space orthogonal to the active variables. We may assume that $\mathbf{x}_l^* \neq \mathbf{0}$ (because otherwise l is redundant for constructing a parsimonious model). Now consider the vector $\hat{\mathbf{y}} + \varepsilon \mathbf{x}_l^*$. This corresponds to a possible model for which $|t_j| = t$ still hold for all active variables, by the orthogonality of the $\mathbf{x}_j$ to $\mathbf{x}_l^*$. Furthermore, we may choose $\varepsilon \neq 0$ small enough in absolute value that it remains true for all inactive variables that $|t_l| < t$. Then we let the sign of ε be opposite that of β_l; by the Pythagorean theorem, $\| \hat{\mathbf{y}} + \varepsilon \mathbf{x}_l^* \| < \| \hat{\mathbf{y}} \|$. Then $\hat{\mathbf{y}}$ cannot be a solution to the LARS problem. To avoid a contradiction, it is necessary that $\beta_l = 0$. QED

Proof of Proposition 2: First note that, because we centered the **y**s, the starting point of the LARS path $\boldsymbol{\beta} = \mathbf{0}$ has $\hat{\mathbf{y}}^{\mathrm{T}} \hat{\mathbf{y}} = 0$; therefore it is a solution to our optimization problem. Furthermore, at any point on the LARS path for which there is only a single active variable (of index j), we find ourselves on the linear segment in $\boldsymbol{\beta}$ space from $\hat{\mathbf{y}}^{\mathrm{T}} \hat{\mathbf{y}} = 0$ to the least squares solution for that variable. All points for which $|t_j| = t$ are orthogonal to the LARS path, so by the Pythagorean theorem, all have larger $\hat{\mathbf{y}}^{\mathrm{T}} \hat{\mathbf{y}}$. Thus any such point is a solution to our optimization problem.

Now consider any point on the LARS path with $t > 0$, and with more than one active variable. Assume it is a solution to our optimization problem. We will reason that further progress along the path can only construct further solutions. The reasoning for Proposition 1 shows that slightly modifying the βs for inactive variables can only

increase the length of $\hat{\mathbf{y}}$. Therefore, we will consider nearby points that do not change *t* that involve adjustment to some active coefficient β_j. Then necessarily since we must have $|t_j| < t$, since the point at which we started was a solution, $\hat{\mathbf{y}}^T\hat{\mathbf{y}}$ has increased. We now know the sign of the corresponding adjustment to β_j.

Now make a small step forward along the LARS path. It will therefore decrease all the active $|t_j| = t$ by an amount small enough not to create any new active variables. But then the earlier argument says that further small adjustments to the coefficients that leave the new *t* unchanged must increase $\hat{\mathbf{y}}^T\hat{\mathbf{y}}$. Therefore, small steps along the path create new solutions to our optimization problem. Therefore the entire path consists of solutions.

References:

Agresti, A. (2013) **Categorical Data Analysis** *3rd ed.* Wiley

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** v. 2 pp. 407-451.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.* **58** pp. 267–288.