

# Least Angle Regression (LARS)

# LASSO

- LASSO is a constrained version of Ordinary Least Squares (OLS)
- Let  $x_1, x_2, \dots, x_p$  be the variables/predictors and  $y$  be the response
- $X = [x_1 \ x_2 \ \dots \ x_p]$ , the matrix with columns containing the predictors.
- If the regression coefficients,  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2 \dots \hat{\beta}_p)'$  give the estimated response  $\hat{y}$ , then

$$\hat{y} = \sum_{j=1}^p x_j \hat{\beta}_j$$

- LASSO chooses  $\hat{\beta}$  by minimizing total squared error subject to a constraint on the coefficients, i.e

$$\min ||y - \hat{y}||^2 \quad \text{subject to} \quad \sum_{j=1}^p |\hat{\beta}_j| \leq t$$

# Forward Stagewise Regression

- A subset selection method (select a subset of variables with linear regression)

- Idea:

Repeat

- Select the predictor having largest absolute correlation with residual vector
- Update the estimated response to move in the direction of the correlated variables

# Stagewise

- Iterative technique that begins with  $\hat{y} = 0$  and builds up the regression function in successive small steps.
- At any step, let  $c(\hat{y})$  be the vector of *current correlations*

$$\hat{c} = c(\hat{y}) = X'(y - \hat{y})$$

where  $\hat{c}_j$  is proportional to the correlation between predictor  $x_j$  and current residual vector  $(y - \hat{y})$

- The Stagewise algorithm moves the prediction in the direction of the greatest current correlation

$$\hat{j} = \operatorname{argmax} |\hat{c}_j| \text{ and } \hat{y} = \hat{y} + \epsilon * \operatorname{sign}(\hat{c}_j) * x_j$$

where  $\epsilon$  is a small constant

# Least Angle Regression

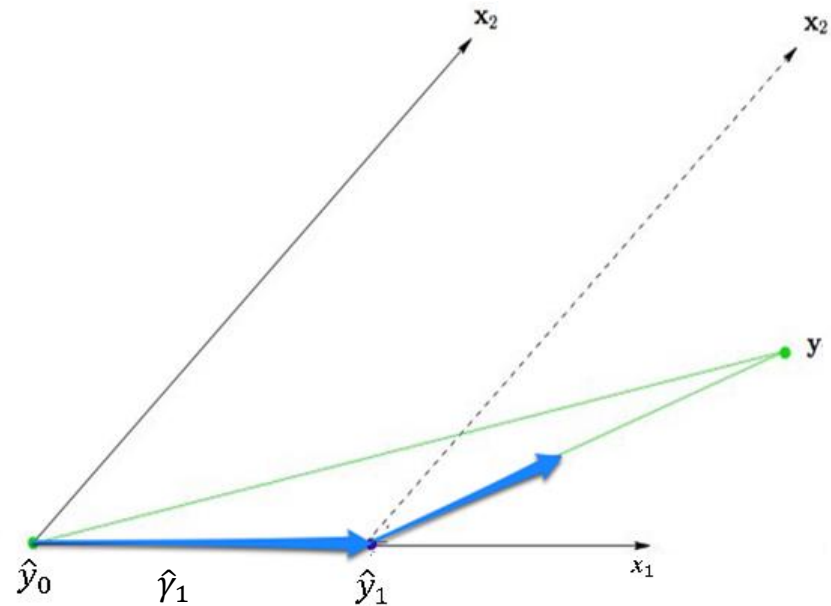
- Both LASSO and Stagewise are variants of a basic procedure called Least Angle Regression (LARS)
- LARS is a stylized version of the Stagewise procedure
- Only  $p$  steps are required for the full set of solutions ( $p$  is the number of variables) for LARS
- The advantages of LARS compared to Stagewise procedure
  - The prediction movement is in a direction which is equiangular to all the most equally correlated variables (**optimal directions**)
  - optimum stepsize (i.e  $\epsilon$ ) is taken such that the next variable is equally correlated with the previously taken variables (**optimal sized leaps**)

# Least Angle Regression

- Assume standardized predictors in the model (mean 0 and unit variance)
- Algorithm
  - Start with no predictors in the model
  - Find the predictor  $x_1$  most correlated to the residual (equivalently, the variable making **least angle** with the residual)
  - Keep moving in the direction of the most correlated predictor until another predictor  $x_2$  becomes equally correlated with the residual.
  - Move in a direction equiangular to both the predictors
  - Continue until all the predictors are in the model

# Example

- Example with 2 correlated variables  $x_1$  and  $x_2$
- $y$  is the response,  $\hat{y}_0.. \hat{y}_2$  are estimated responses at each step ,  $\hat{\gamma}_1... \hat{\gamma}_2$  are the optimal step sizes
- At  $\hat{y}_0 = 0$ , the residual vector  $y - \hat{y}_0$  is most correlated to  $x_1$  (least angle)  
$$\hat{y}_1 = \hat{y}_0 + \hat{\gamma}_1 x_1$$
- Select  $\hat{\gamma}_1$  such that the residual  $y - \hat{y}_1$  is equally correlated with  $x_1$  and  $x_2$
- Then,  $\hat{y}_2 = \hat{y}_1 + \hat{\gamma}_2 u_2$ , where  $u_2$  is the unit bisector (equiangular vector)
- Here,  $\hat{y}_2 = y$



# LARS (contd)

- At  $k^{th}$  step, let  $A_k$  be the set of active variables/predictors and  $\beta_{A_k}$  be the coefficient vectors and  $\hat{y}_k$  be the estimated response.
- $X_{A_k}$  be the active variables i.e those variables whose absolute correlation with the residual vector equals the maximal achievable absolute correlation.
- The current residual will be  $r_k = y - \hat{y}_k = y - X_{A_k}\beta_{A_k}$
- The coefficients are moved in the direction  $\delta_k = (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k$   
i.e  $\beta_{A_k}(\gamma) = \beta_{A_k} + \gamma\delta_k$



# Equiangular vector

- $\beta_{A_k}(\gamma) = \beta_{A_k} + \gamma \delta_k$
- Thus  $\hat{y}(\gamma) = X_{A_k} \beta_{A_k}(\gamma) = X_{A_k} \beta_{A_k} + \gamma * \underbrace{X_{A_k} \delta_k}_{u_k}$

$u_k$ :

makes equal angles with the predictors in  $A_k$

- Prove that  $u_k$  makes equal angles with the active predictors

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

- The entries of the vector  $X_{A_k}^T r_k$  are all the same since the predictors have same correlation with residuals (by construction)

# LAR algorithm:

Least Angle Regression 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  $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$ ,  $\beta_1, \beta_2, \dots, \beta_p = 0$ .
  2. Find the predictor  $\mathbf{x}_j$  most correlated with  $\mathbf{r}$ .
  3. Move  $\beta_j$  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  $\beta_j$  and  $\beta_k$  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.
-

# Modification of LAR to solve LASSO problem

---

**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.
-

# References

- Efron, Bradley, et al. "Least angle regression." *The Annals of statistics* 32.2 (2004): 407-499.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2008