Forward selection procedure

- Let $\mu = E(Y|X)$.
- Begin with a constant model $\hat{\mu} = 0$ assuming that y is centered and x_j's are standardized.
- Given a set of covariates, select x_j with the largest absolute correlation with y.
- Fit a linear model with x_j and update $\hat{\mu} \leftarrow \hat{\mu} + \hat{\beta}_j X_j$.
- Get a residual vector $r = Y - \hat{\mu}$ and project other X_j's orthogonally to X_j.
- Repeat the selection process.
- It can be very greedy, eliminating other correlated covariates.

Forward stagewise regression

- A cautious version of the forward selection
- An iterative procedure to build a regression function in successive small steps
- Begin with $\hat{\mu} = 0$.
- Define a vector of 'current correlations':

$$\hat{c} = c(\hat{\mu}) = X^\top (Y - \hat{\mu})$$

- Find the variable index $j = argmax|\hat{c}_j|$ and update

$$\hat{\mu} \leftarrow \hat{\mu} + \epsilon \cdot sign(\hat{c}_j)\, X_j,$$

  where ϵ is a small constant.
- Least squares boosting

Boosting for regression

- A technique for additive model building
- $y = f(x) + \epsilon$
- A family of basis functions ('base learners'):
  $\{b(x;\gamma), \gamma \in \Gamma\}$
  e.g. $\{b(x;j) = x_j, j = 1, \ldots, k\}$ in linear regression
- Consider an additive model of the form

$$\hat{f}(x) = \sum_{j=1}^{M} \beta_j b(x;\gamma_j).$$

- Boosting builds such a model in a stagewise fashion.

Least Squares Boosting

- Begin with $\hat{f}_0(x) = 0$.
- Let $\hat{f}_m(x)$ be the model at the mth step.
- At the $(m+1)$ step, find β and γ minimizing

$$\sum_{i=1}^{n} \{y_i - \hat{f}_m(x_i) - \beta b(x_i;\gamma)\}^2.$$

- Equivalently, with $r_i^{(m)} = y_i - \hat{f}_m(x_i)$

$$\min_{\beta,\gamma} \sum_{i=1}^{n} (r_i^{(m)} - \beta b(x_i;\gamma))^2.$$

- $\hat{f}_{m+1}(x) = \hat{f}_m(x) + \beta_{m+1} b(x;\gamma_{m+1})$

Introduction to Least Angle Regression

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004)
Least angle regression

- Motivated by the forward stagewise regression
- A computational shortcut to stagewise regression
- Explain a striking similarity between LASSO and stagewise regression
- Their connection as variants of LAR
- Geometrical interpretation
- An efficient computational algorithm for generating coefficient paths

Preliminaries for Least Angle Regression

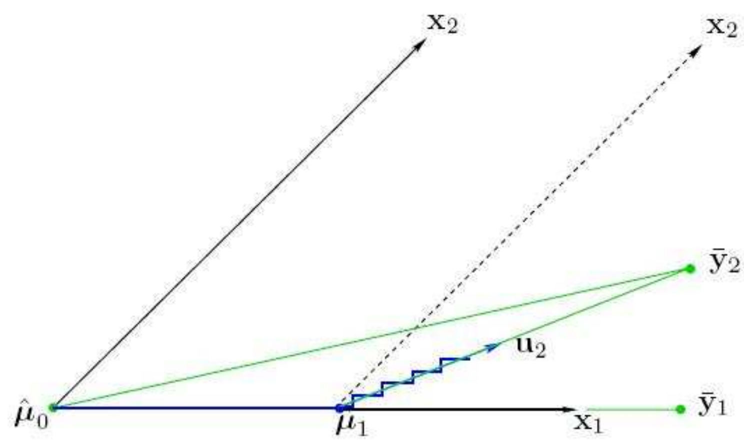
- $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1$ for $j = 1, \dots, k$
- Current residuals: $Y - \hat{\mu}$
- Current correlations: $c(\hat{\mu}) = X^\top (Y - \hat{\mu})$
- Let $\bar{Y}$ be the projection of Y onto the linear space $\mathcal{L}(X)$ spanned by X_j's.
- Then $c(\hat{\mu}) = X^\top (\bar{Y} - \hat{\mu})$
- The absolute correlations are related to the angles of the current residuals with X_j's.

Illustration of LAR when $k = 2$

Suppose that $X = [X_1, X_2]$.

- Begins at $\hat{\mu}_0 = 0$.
- $c(\hat{\mu}_0) = X^\top (\bar{Y}_2 - \hat{\mu}_0) = (X_1^\top \bar{Y}_2, X_2^\top \bar{Y}_2)^\top$
- The largest absolute correlation criterion is equivalent to choosing X_j with the least angle with Y.
- Suppose that X_1 has a smaller angle with Y.
  Then LAR updates
  $$\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 X_1,$$
  where $\hat{\gamma}_1$ makes $\bar{Y}_2 - \hat{\mu}_1$ equally correlated with X_1 and X_2.
- Let U_2 be the unit vector bisecting the angle between X_1 and X_2.
- $\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 U_2$ with $\hat{\gamma}_2$ chosen to make $\hat{\mu}_2 = \bar{Y}_2$

Geometry of LAR



Choice of $\hat{\gamma}_1$

- Forward selection:
  $\hat{\gamma}_1$ large enough to make $\hat{\mu}_1 = \bar{Y}_1$ where $\bar{Y}_1$: the projection of Y onto $\mathcal{L}(X_1)$
- Forward stagewise regression: some small value ϵ
- Least Angle Regression:
  an intermediate value so that the updated residual vector is equally correlated with X_1 and X_2

Least Angle Regression in general

- Let $\mathcal{A}$ be the set of indices corresponding to covariates in the current model.
- In what direction do we move to update the model?
  LAR steps are taken along 'equiangular vectors'.
- How far do we move along the direction?
  Until some new variable has the same current correlation as those in the active set

Equiangular vectors

- Let $X_{\mathcal{A}} = [\cdots s_j X_j \cdots]_{j \in \mathcal{A}}$, where $s_j = \pm 1$.
- $U_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}$:
  the unit vector making equal angles ($< 90^o$) with the columns of $X_{\mathcal{A}}$
- $X_{\mathcal{A}}^\top U_{\mathcal{A}} = A_{\mathcal{A}} 1_{\mathcal{A}}$ and $\|U_{\mathcal{A}}\| = 1$ where $A_{\mathcal{A}}$ is a constant.
- Then $w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} 1_{\mathcal{A}}$ and $A_{\mathcal{A}} = (1_{\mathcal{A}}^\top G_{\mathcal{A}}^{-1} 1_{\mathcal{A}})^{-1/2}$ where $G_{\mathcal{A}} = X_{\mathcal{A}}^\top X_{\mathcal{A}}$.

Current correlations

- $\hat{c} = X^\top (Y - \hat{\mu}_{\mathcal{A}})$
- Let $\hat{C} = \max_j \{|\hat{c}_j|\}$.
- $\mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}$
- Let $a = X^\top U_{\mathcal{A}}$.
- Consider $\mu(\gamma) = \hat{\mu}_{\mathcal{A}} + \gamma U_{\mathcal{A}}$. Then $c_j(\gamma) = X_j^\top (Y - \mu(\gamma)) = \hat{c}_j - \gamma a_j$.
- For $j \in \mathcal{A}$, $|c_j(\gamma)| = \hat{C} - \gamma A_{\mathcal{A}}$.