Regularization and Variable Selection via the Elastic Net

A Comparative Study of Penalized Regression Methods

Ruijuan Zhong Yue Zhou Wenxin JIANG

Department of Biostatistics
City University of Hong Kong

9 April 2024

Table of Contents

1. rj
2. zy
3. Least Angle Regression and Coordinate Descent

Table of Contents

1. rj
2. zy
3. Least Angle Regression and Coordinate Descent

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
- Text visible on slides 3

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
- Text visible on slide 4

Table of Contents

1 rj

2 zy

3 Least Angle Regression and Coordinate Descent

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
- Text visible on slides 3

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2

- Text visible on slide 4

Table of Contents

1 rj

2 zy

3 Least Angle Regression and Coordinate Descent

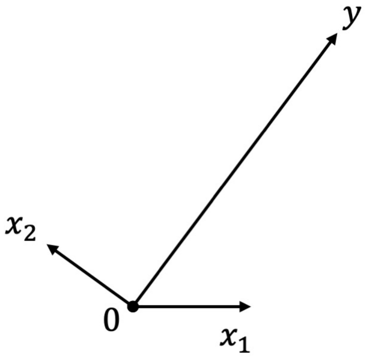
Least Angle Regression

1. Forward Stepwise Selection
2. Forward Stagewise Selection
3. Least Angle Regression

Forward Stepwise Selection

A simple example in the case of $p = 2$ predictors.

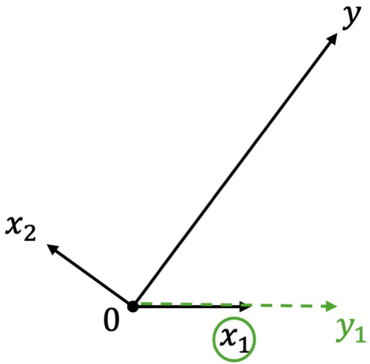
1. Start with a null model.



Forward Stepwise Selection

A simple example in the case of $p = 2$ predictors.

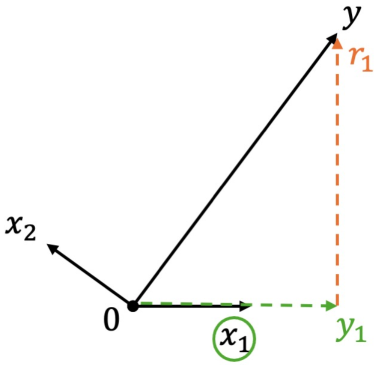
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.



Forward Stepwise Selection

A simple example in the case of $p = 2$ predictors.

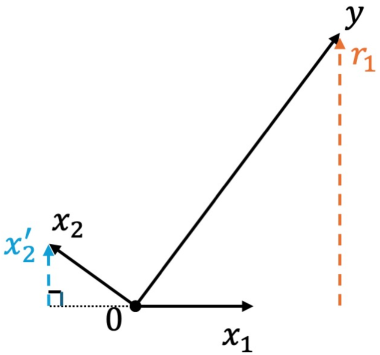
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.



Forward Stepwise Selection

A simple example in the case of $p = 2$ predictors.

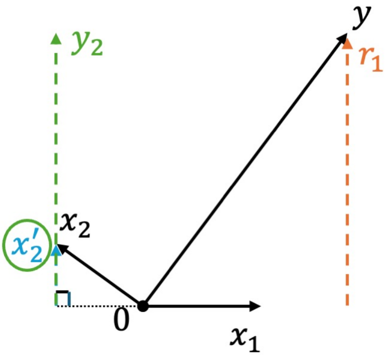
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.
4. Project other predictors orthogonal to the predictor selected in previous step.



Forward Stepwise Selection

A simple example in the case of $p = 2$ predictors.

1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.
4. Project other predictors orthogonal to the predictor selected in previous step.
5. Repeat steps $2 - 4$ until the stopping criterion is met.

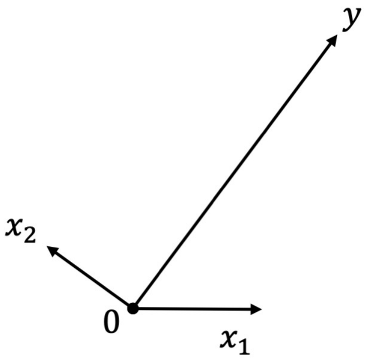


Forward Stagewise Selection

In contrast to forward stepwise selection, forward stagewise selection builds the model in successive small steps ε.

Let $\hat{\mu}$ be the current Stagewise estimate and $\hat{\mathbf{c}} = \mathbf{c}(\hat{\mu}) = X^T(y - \hat{\mu})$ be the vector of current correlations. Therefore, $\hat{c}_j$ is proportional to the correlation between the covariate x_j and the current residual vector.

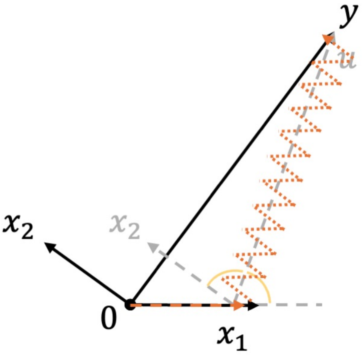
1. Start with $\hat{\mu} = 0$.



Forward Stagewise Selection

Let $\hat{\mu}$ be the current Stagewise estimate and $\hat{\mathbf{c}} = \mathbf{c}(\hat{\mu}) = X^T(y - \hat{\mu})$ be the vector of current correlations.

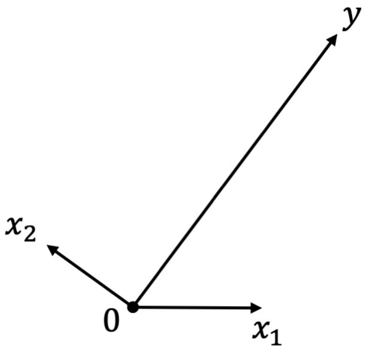
1. Start with $\hat{\mu} = 0$.
2. Find the predictor j that has the highest correlation that $\hat{j} = \arg\max_j |\hat{c}_j|$.
3. Update $\hat{\mu} \leftarrow \hat{\mu} + \varepsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}}$ and $\widehat{\mathbf{c}}$.
4. Repeat steps $2 - 3$ until the stopping criterion is met.



Least Angle Regression

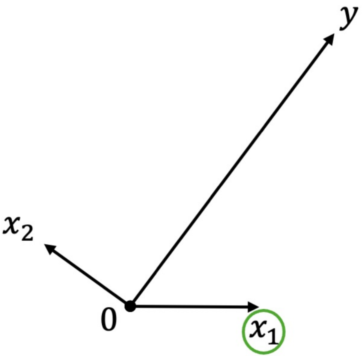
Least Angle Regression (LARS) is a stylized version of forward stagewise procedure that uses a simple mathematical formula to accelerate the computations. Here shows the idea of LARS.

1. Start with all coefficients equal to zero.



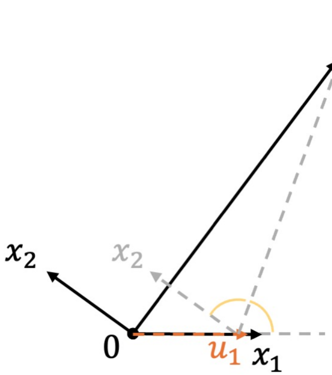
Least Angle Regression

1. Start with all coefficients equal to zero.
2. Find the predictor most correlated with the response.



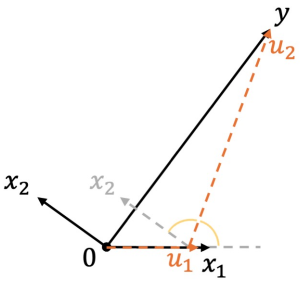
Least Angle Regression

1. Start with all coefficients equal to zero.
2. Find the predictor most correlated with the response.
3. Take the largest step possible in the direction of this predictor until some other predictor has as much correlation with the current residual.



Least Angle Regression

1. Start with all coefficients equal to zero.
2. Find the predictor most correlated with the response.
3. Take the largest step possible in the direction of this predictor until some other predictor has as much correlation with the current residual.
4. The new direction is the equiangular vector of the two predictors. Move in until a third predictor earns its way into the "most correlated" set.
5. Repeat steps 3 – 4 until met the stopping criterion.



Least Angle Regression

Assume that $\mathbf{x}_1, \dots, \mathbf{x}_p$ are linearly independent and for $\mathcal{A}$ a subset of indices $\{1, \dots, p\}$, define the matrix $\mathbf{X}_{\mathcal{A}} = (\dots, s_j\mathbf{x}_j, \dots)_{j \in \mathcal{A}}$ where signs s_j equal ± 1. Let

$$\}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \quad \text{and} \quad A_{\mathcal{A}} = \left(\mathbf{1}_{\mathcal{A}}^T \}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}\right)^{-1/2}, \tag{1}$$

where $\mathbf{1}_{\mathcal{A}}$ is a vector of ones of length $|\mathcal{A}|$. The equiangular vector $\mathbf{u}_{\mathcal{A}}$ is defined as

$$\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} A_{\mathcal{A}} \}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}, \tag{2}$$

is the unit vector making equal angles, less than 90°, with the columns of $\mathbf{X}_{\mathcal{A}}$ satisfying $\mathbf{X}_{\mathcal{A}}^T \mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}}$ and $\|\mathbf{u}_{\mathcal{A}}\| = 1$.

Least Angle Regression

Then the algorithm of LARS comes as follows:

1. Initialize all the coefficients as 0, the residual $\mathbf{u} = \mathbf{y}$ and the active set $\mathcal{A} = \emptyset$.
2. Suppose that $\hat{\mu}_{\mathcal{A}}$ is the current estimate of the response and $\hat{\mathbf{c}} = \mathbf{c}(\hat{\mu}_{\mathcal{A}}) = X^T(y - \hat{\mu}_{\mathcal{A}})$ are the current correlations. The active set $\mathcal{A}$ is the set of indices corresponding to covariates with the greatest absolute correlations, i.e., $\mathcal{A} = \{j : |\hat{c}_j| = \hat{\mathbf{C}}\}$ and $\hat{\mathbf{C}} = \max_j |\hat{c}_j|$. Let $s_j = \text{sign}(\hat{c}_j)$ for $j \in \mathcal{A}$, and compute $A_{\mathcal{A}}$, and $\mathbf{u}_{\mathcal{A}}$ as in 1 and 2. Also, compute the inner product $\mathbf{a} =: X^T\mathbf{u}_{\mathcal{A}}$. Updates $\hat{\mu}_{\mathcal{A}}$ as

$$\hat{\mu}_{\mathcal{A}} \leftarrow \hat{\mu}_{\mathcal{A}} + \hat{\gamma}\mathbf{u}_{\mathcal{A}},$$

   where $\hat{\gamma} = \min^{+}_{j \in \mathcal{A}^c} \left(\frac{\hat{\mathbf{C}} - \hat{c}_j}{A_{\mathcal{A}} - \mathbf{a}_j}, \frac{\hat{\mathbf{C}} + \hat{c}_j}{A_{\mathcal{A}} + \mathbf{a}_j} \right)$; "min+" denotes the minimum taken over only positive quantities.
3. Repeat steps 2 until the stopping criterion is met.