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

Table of Contents

Sample frame title

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

- Text visible on slide 1

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

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

香港城市大學
City University of Hong Kong

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

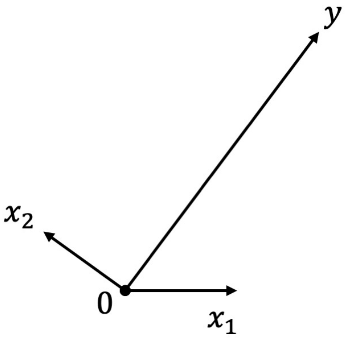
Table of Contents

Least Angle Regression

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

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

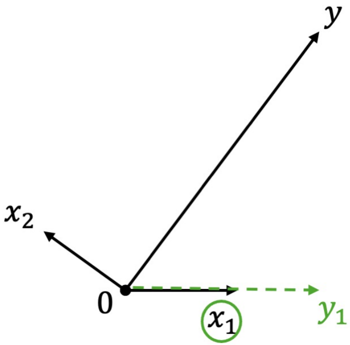
1. Start with a null model.



Forward Stepwise Selection

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

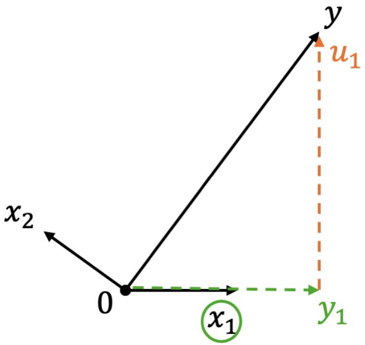
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.



Forward Stepwise Selection

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

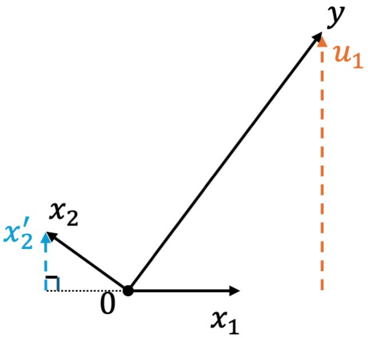
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.



Forward Stepwise Selection

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

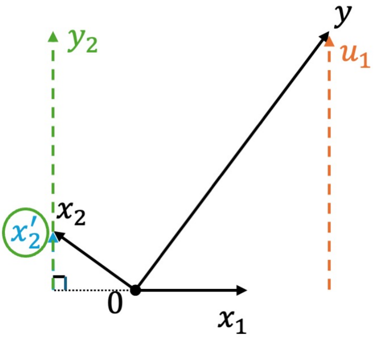
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.
4. Project other predictors orthogonal to the predictor selected in previous step.



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.

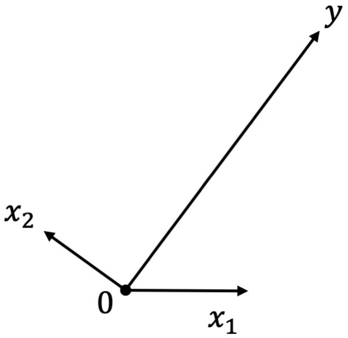


Forward Stagewise Selection

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.

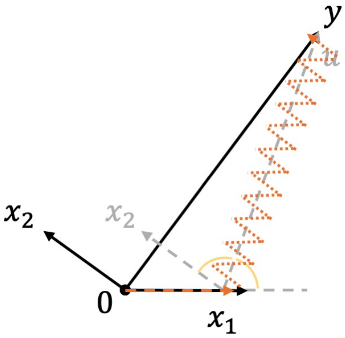
1. Start with $\hat{\mu} = 0$.



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.

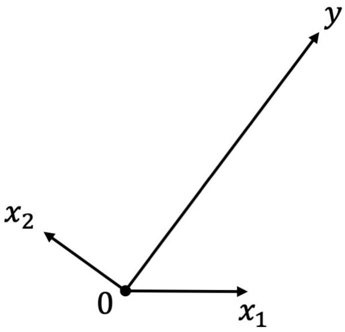
1. Start with $\hat{\mu} = 0$.
2. Find the predictor j that has the highest correlation that $\hat{j} = \arg\max_j |\hat{c}_j|$.
3. Update $\hat{\mu} \leftarrow \hat{\mu} + \varepsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}}$ and $\widehat{\mathbf{c}}$.
4. Repeat steps $2 - 3$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



Least Angle Regression: Example

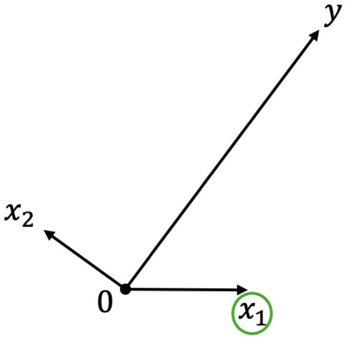
Least Angle Regression (LAR) is a stylized version of forward stagewise procedure that uses a simple mathematical formula to accelerate the computations.

1. Start with all coefficients equal to zero.



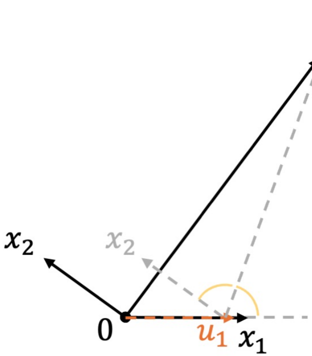
Least Angle Regression: Example

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



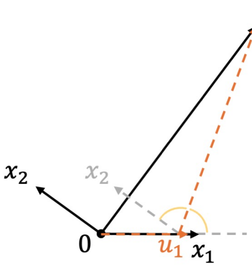
Least Angle Regression: Example

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: Example

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: Notation

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

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

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

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

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 $\|\mathbf{u}_{\mathcal{A}}\| = 1$.

香港城市大學 City University of Hong Kong

Least Angle Regression: Algorithm

1. Initialize all the coefficients $\hat{\mu}_0$ as 0, and let the residual $\mathbf{u} = \mathbf{y}$.
2. Suppose that $\hat{\mu}_{\mathcal{A}}$ is the current estimate of coefficients 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

$$\hat{\mu}_{\mathcal{A}} \leftarrow \hat{\mu}_{\mathcal{A}} + \hat{\gamma}\mathbf{u}_{\mathcal{A}}, \tag{3}$$

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 step 2 until the stopping criterion is met.

Comparison the Solution Paths of LARS

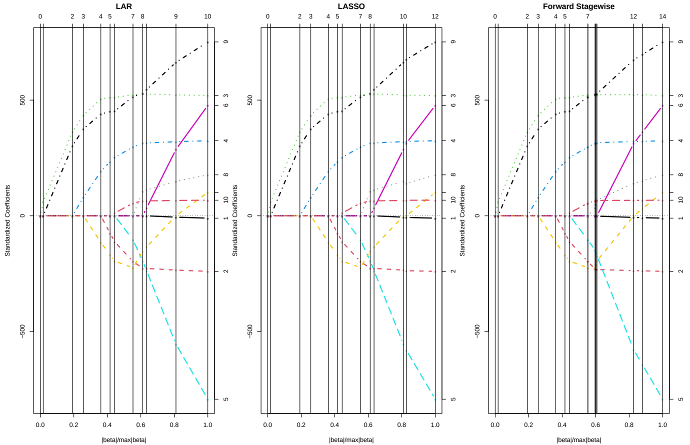


Figure: Solution paths of LAR, LAR-lasso and Forward Stagewise Selection for the diabetes data set.

Extend LARS to Lasso and Stagewise Regression

Define $\hat{\mathbf{d}}$ to be the m-vector equaling $s_j\{A_\mathcal{A} g_\mathcal{A}^{-1} \mathbf{1}_\mathcal{A}\}_j$ for $j \in \mathcal{A}$ and zero elsewhere. Let

$$\tilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\},$$

where $\gamma_j = -\hat{\beta}_j / \hat{d}_j$, we have the following modification to LAR for Lasso:

LASSO MODIFICATION

If $\tilde{\gamma} < \hat{\gamma}$, stop the ongoing LARS at $\gamma = \tilde{\gamma}$ and remove $\tilde{j}$ from the active set. Then continue the LARS path from the current point.

Extend LARS to Lasso and Stagewise Regression

Define

$$P \equiv (N_1, \ldots, N_p)/N \quad \mathcal{C}_{\mathcal{A}} = \left\{ \mathbf{v} = \sum_{j \in \mathcal{A}} s_j \mathbf{x}_j P_j, P_j \geq 0 \right\}$$

where $N_j \equiv \#\{\text{steps with selected index j}\}$. Then we have the following modification to LAR for Stagewise Regression:

STAGEWISE MODIFICATION

Replace the $\mathbf{u}_{\mathcal{A}}$ in LAR with $\mathbf{u}_{\hat{\mathcal{B}}}$, the unit vector lying alone the projection of $\mathbf{u}_{\mathcal{A}}$ into $\mathcal{C}_{\mathcal{A}}$.

Comparison of Computational Time

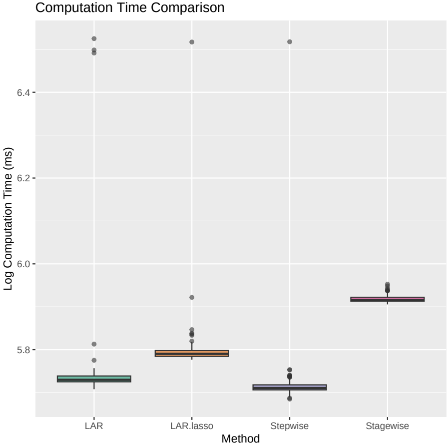


Figure: Comparison of computational time between LAR, LAR-Lasso, Forward Stagewise Selection, and Forward Stepwise Selection with the diabetes data set.

Table of Contents

香港城市大學
City University of Hong Kong

Coordinate Descent: Motivation Question 1

To motivate the objective function we would like to deal with using coordinate descent, let's consider these questions first:

Q1: Does $f(x+\delta e_i) \geq f(x)$ for all $\delta, i \Longrightarrow f(x) = \min_z f(z)$ (Here $e_i = (0, \ldots, 1, \ldots 0)$, the i-th standard basis vector) always hold?

In other words, given convex, differentiable $f : R^n \to R$, if we are at a point x such that $f(x)$ is minimized along each coordinate axis, then have we found a global minimizer?

Coordinate Descent: Motivation Question 1

Q1: Does $f(x+\delta e_i) \geq f(x)$ for all $\delta, i \Longrightarrow f(x) = \min_z f(z)$ (Here $e_i = (0, \ldots, 1, \ldots 0)$, the i-th standard basis vector) always hold?

Yes. **Proof**:

$$f(x+\delta e_i) \geq f(x) \Longrightarrow \frac{\partial f}{\partial x_i}(x) = 0,$$

which means

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) = 0$$

Then we get $f(x) = \min_z f(z)$.

Coordinate Descent: Motivation Question 2

Q2: Same question, but f is convex, not differentiable?

Coordinate Descent: Motivation Question 2

Q2: Same question, but f is convex, not differentiable?

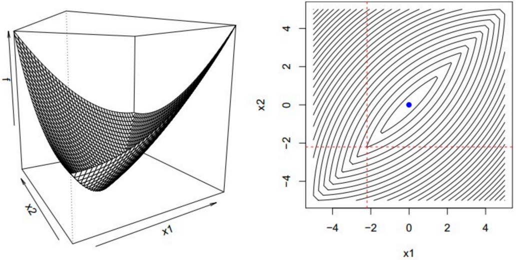


Figure: Illustration of the question.

No. We can see that the whatever the cross-point goes any direction along the axis, the criterion value will increase.

Coordinate Descent: Motivation Question 3

Q3: Same question again, but now $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$, where $g(x)$ is convex, differentiable and each h_i is just convex (Here the non-smooth part is called separable)?

Q3: Same question again, but now $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$, where $g(x)$ is convex, differentiable and each h_i is just convex?

Yes. **Proof**: Since $g(x)$ is convex, differentiable, for any y, we have

$$
\begin{aligned}
f(y) - f(x) &= g(y) + \sum_{i=1}^{n} h_i(y_i) - \left[g(x) + \sum_{i=1}^{n} h_i(x_i) \right] \\
&\geq \nabla g(x)^T (y - x) + \sum_{i=1}^{n} [h_i(y_i) - h_i(x_i)] \\
&= \sum_{i=1}^{n} \left(\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i) \right)
\end{aligned}
$$

We now want to proof

$$
\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i) \geq 0.
$$

We now want to proof

$$\nabla_i g(x)\left(y_i - x_i\right) + h_i\left(y_i\right) - h_i\left(x_i\right) \geq 0.$$

Consider $f_i\left(x_i\right) = g\left(x_i; x_{-i}\right) + h_i\left(x_i\right)$, we have

$$f\left(x + \delta e_i\right) \geq f(x) \Rightarrow 0 \in \partial f_i\left(x_i\right) = \nabla_i g(x) + \partial h_i\left(x_i\right) \Rightarrow \nabla_i g(x) \in -\partial h_i\left(x_i\right),$$

then by definition of subgradient:

$$h_i\left(y_i\right) \geq h_i\left(x_i\right) - \nabla_i g(x)\left(y_i - x_i\right).$$

Thus, we can conclude that for any y, $f(y) - f(x) \geq 0$.

Coordinate Descent: Update Rule

Q3 suggests that for $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$, where $g(x)$ is convex, differentiable and each h_i is just convex, we can use coordinate descent to find a minimizer: start with some initial guess $x^{(0)}$, and repeat:

$$
\begin{gathered}
x_1^{(k)} \in \arg\min_{x_1} f\left(x_1, x_2^{(k-1)}, \ldots, x_n^{(k-1)}\right) \\
x_2^{(k)} \in \arg\min_{x_2} f\left(x_1^{(k)}, x_2, \ldots, x_n^{(k-1)}\right) \\
\ldots \\
x_n^{(k)} \in \arg\min_{x_n} f\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n\right)
\end{gathered}
$$

for $k = 1, 2, 3 \ldots$

香港城市大學
City University of Hong Kong

Coordinate Descent: Notes

Here is several things worth to notice:

- The **order of cycle** through coordinates is arbitrary, we can use any permutation of $1, 2, \ldots, n$. If only we visit linear number of updates x_i before going to update x_j (eg. update $2n$ times, but cannot be n^2), the algorithm can converge.
- We can replace individual coordinates with **blocks of coordinates** in everywhere.
- "**One-at-a-time**" update scheme is critical, and "all-at-once" scheme does not necessarily converge. In other words, after solving for $x_i^{(k)}$, we use its new value from then on.

Coordinate Descent: Lasso Regression

Given $y \in R^n$, and $X \in R^{n \times p}$ with columns $X_1, \ldots, X_n$, consider lasso regression:

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

We can perform coordinate descent by repeatedly minimize over β_i for $i = 1, 2, \ldots, p, 1, 2, \ldots$ Here β_i can be gotten by solving:

$$0 = X_i^T \left(X_i \beta_i + X_{-i} \beta_{-i} - y \right) + \lambda s_i,$$

where $s_i \in \partial \left| \beta_i \right|$. Then by using soft-thresholding we get

$$\beta_i = S_{\lambda / ||X_i\|_2^2} \frac{X_i^T \left(y - X_{-i} \beta_{-i} \right)}{X_i^T X_i}$$

香港城市大學
City University of Hong Kong

Coordinate Descent: Lasso Regression

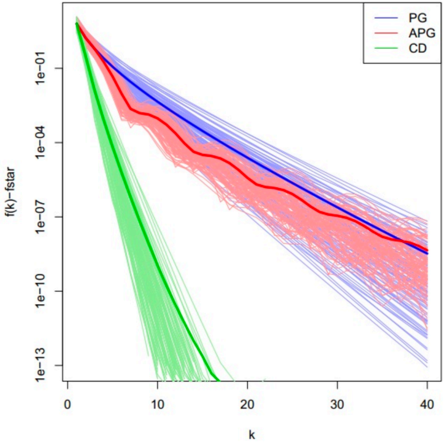


Figure: Coordinate descent and (accelerated) proximal gradient descent for lasso regression with $n = 100, p = 20$.

References I

[1] H. Zou and T. Hastie, *Regularization and Variable Selection Via the Elastic Net*, Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[2] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., vol. 32, no. 2, Apr. 2004, doi: 10.1214/009053604000000067.

[3] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent*, J. Stat. Soft., vol. 33, no. 1, 2010, doi: 10.18637/jss.v033.i01.

[4] T. Hastie, R. Tibshirani, and R. Tibshirani, *Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons*, Statist. Sci., vol. 35, no. 4, Nov. 2020, doi: 10.1214/19-STS733.

[5] J. Pena, R. Tibshirani, *Coordinate Descent. Convex Optimization: Fall 2016* Retrieved April 6, 2024, from `https://www.stat.cmu.edu/~ryantibs/convexopt-F16/`