Regularization and Variable Selection via the Elastic Net

Model, Algorithm and Application

Ruijuan Zhong Yue Zhou Wenxin JIANG

Department of Biostatistics
City University of Hong Kong

9 April 2024

Table of Contents

Table of Contents

香港城市大學
City University of Hong Kong

Linear regression model

Consider the usual linear regression model: given p predictors $x_1, ..., x_p$, the response y is predicted by

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \ldots + x_p\hat{\beta}_p \tag{1}$$

A model fitting procedure produces the vector of coefficients

$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, ..., \hat{\beta}_p\right)$. For example, the ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squares.

Criteria

- **Accuracy of prediction on future data**—it is difficult to defend a model that predicts poorly.
- **Interpretation of the model**—scientists prefer a simpler model because it puts more light on the relationship between the response and covariates.

It is well known that OLS often does poorly in both prediction and interpretation.

Penalization techniques to improve OLS

- **Accuracy of prediction on future data**—it is difficult to defend a model that predicts poorly.
- **Interpretation of the model**—scientists prefer a simpler model because it puts more light on the relationship between the response and covariates.

Ridge regression

As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias–variance trade-off. However, ridge regression cannot produce a parsimonious model, for it always keeps all the predictors in the model.

Best subset slection

Best subset selection in contrast produces a sparse model, but it is extremely variable because of its inherent discreteness.

Penalization techniques to improve OLS

Lasso

The lasso does both continuous shrinkage and automatic variable selection simultaneously.

- In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem.
- If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.
- For usual n>p situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression.

Table of Contents

Definition

Suppose that the data set has n observations with p predictors. Assume:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j = 1, 2, \ldots, p. \tag{2}$$

For any fixed non-negative λ_1 and λ_2, we define the naive elastic net criterion

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = ||y - X\boldsymbol{\beta}||^2 + \lambda_2||\boldsymbol{\beta}||^2 + \lambda_1||\boldsymbol{\beta}||_1, \tag{3}$$

where

$$||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2,$$

$$||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|.$$

Definition

The naive elastic net estimator $\hat{\beta}$ is the minimizer of equation (3):

$$\hat{\beta} = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta). \tag{4}$$

This procedure can be viewed as a penalized least squares method. Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, then solving $\hat{\beta}$ in equation (3) is equivalent to the optimization problem

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \|y - X\beta\|^2, \\ &\text{subject to } (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|^2 \leq t \text{ for some } t. \end{aligned} \tag{5}$$

Definition

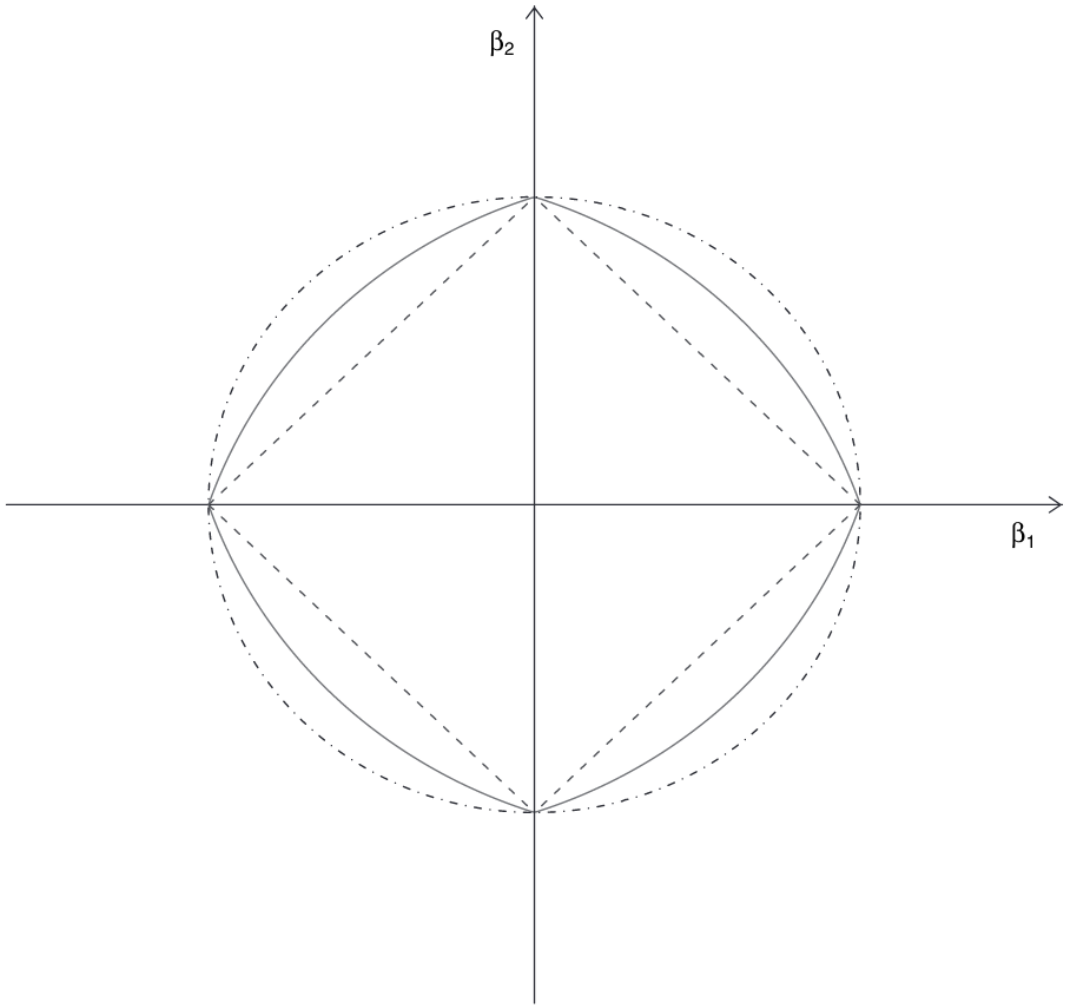


Figure: Two-dimensional contour plots

Solution

LEMMA 1

Given data set $(\mathbf{y}, \mathbf{X})$ and (λ_1, λ_2), define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\beta^* = \sqrt{(1+\lambda_2)}\beta$. Then the naive elastic net criterion can be written as

$$L(\gamma, \beta) = L(\gamma, \beta^*) = \|y^* - X^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1.$$

Let

$$\hat{\beta}^* = \arg\min_{\beta^*} L(\gamma, \beta^*); \quad \text{then} \quad \hat{\beta} = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\beta}^*.$$

Solution - Orthogonal case

In the case of an orthogonal design, it is straightforward to show that with parameters (λ_1, λ_2) the naive elastic net solution is

$$\hat{\beta}_i(\text{naive elastic net}) = \frac{(|\hat{\beta}_i^{(OLS)}| - \lambda_1/2)_+}{1+\lambda_2}\text{sgn}(\hat{\beta}_i^{(OLS)}) \tag{6}$$

where $\hat{\beta}^{(OLS)} = X^T y$ and z_+ denotes the positive part, which is z if $z \geq 0$ and 0 otherwise. The solution of ridge regression with parameter λ_2 is given by $\hat{\beta}^{(ridge)} = \hat{\beta}^{(OLS)}/(1+\lambda_2)$, and the lasso solution with parameter λ_1 is

$$\hat{\beta}_i(\text{lasso}) = |(\hat{\beta}_i^{(OLS)}| - \lambda_1/2)_+\text{sgn}(\hat{\beta}_i^{(OLS)}).$$

Solution - Orthogonal case

Proof: In orthogonal case, $X^T X = 1$, so $\hat{\beta}^{OLS} = X^T y$. Given these conditions, the loss function $L(\beta_i)$ is given by:

$$L(\beta_i) = (y - \beta_i)^2 + \lambda_2 \beta_i^2 + \lambda_1 |\beta_i|,$$

where y and β_i are elements of y and β, respectively. If we represent z as β_i, then the loss function $L(z)$ can be expanded as:

$$L(z) = (z - \hat{\beta}_i^{(OLS)})^2 + \lambda_2 z^2 + \lambda_1 |z|,$$

When $z \geq 0$,

$$\frac{dL(z)}{dz} = 2(z - \hat{\beta}_i^{(OLS)}) + 2\lambda_2 z + \lambda_1 = 0$$

Solution - Orthogonal case

(Continued) Given the above conditions, we deduce:

$$z^* = \frac{\hat{\beta}_i^{(OLS)} - \lambda_1/2}{1 + \lambda_2}$$

When $z \leq 0$,

$$L(z) = (z - \hat{\beta}_i^{(OLS)})^2 + \lambda_2 z^2 - \lambda_1 z$$

Subsequently, to find the minimum of the loss function, we take the derivative with respect to z and set it to zero:

$$\frac{dL(z)}{dz} = 2(z - \hat{\beta}_i^{(OLS)}) + 2\lambda_2 z - \lambda_1 = 0$$

Solution - Orthogonal case

(Continued) Subsequently, to find the minimum of the loss function, we take the derivative with respect to z and set it to zero:

$$\frac{dL(z)}{dz} = 2(z - \hat{\beta}_i^{(OLS)}) + 2\lambda_2 z - \lambda_1 = 0$$

The solution is:

$$z^* = \frac{\hat{\beta}_i^{(OLS)} + \lambda_1/2}{1 + \lambda_2}$$

Solution - Orthogonal case

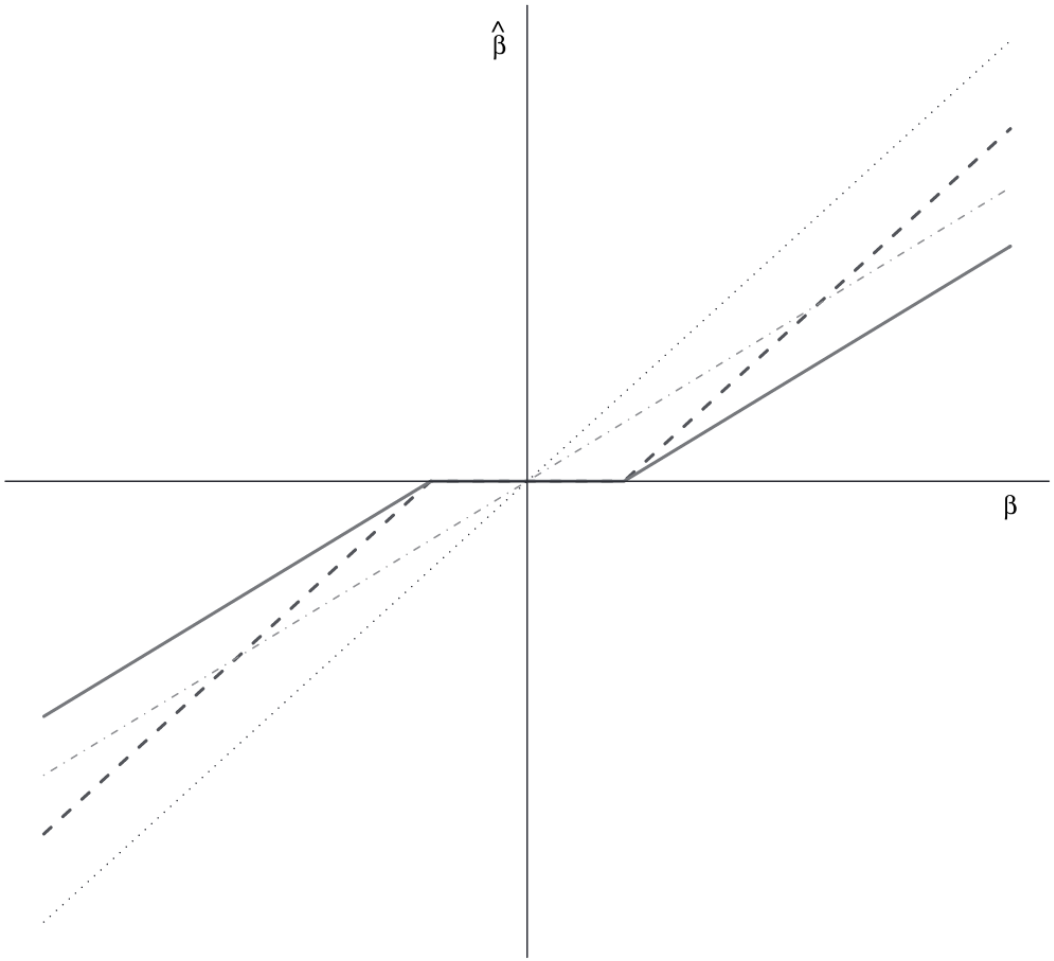


Figure: Exact solutions for the lasso, ridge regression and the naive elastic net in an orthogonal design

Grouping effect

We consider the generic penalization method

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda J(\beta) \tag{7}$$

where $J(\cdot)$ is positive valued for $\beta \neq 0$.

LEMMA 2

Assume that $\mathbf{x}_i = \mathbf{x}_j, i, j \in \{1, \ldots, p\}$.

1. If $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$.
2. If $J(\beta) = \|\beta\|_1$, then $\hat{\beta}_i \hat{\beta}_j \geq 0$ and $\hat{\beta}^*$ is another minimizer of equation (7), where

$$\hat{\beta}^*_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j, \end{cases}$$

for any $s \in [0, 1]$.

Grouping effect

A.1.1. Part (1)
Fix $\lambda > 0$. If $\hat{\beta}_i \neq \hat{\beta}_j$, let us consider $\hat{\boldsymbol{\beta}}^*$ as follows:

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ \frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = i \text{ or } k = j. \end{cases}$$

Because $\mathbf{x}_i = \mathbf{x}_j$, it is obvious that $\mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{X}\hat{\boldsymbol{\beta}}$; thus $|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*|^2 = |\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}|^2$. However, $J(\cdot)$ is strictly convex, so we have $J(\hat{\boldsymbol{\beta}}^*) < J(\hat{\boldsymbol{\beta}})$. Therefore $\hat{\boldsymbol{\beta}}$ cannot be the minimizer of equation (7), which is a contradiction. So we must have $\hat{\beta}_i = \hat{\beta}_j$.

A.1.2. Part (2)
If $\hat{\beta}_i \hat{\beta}_j < 0$, consider the same $\hat{\boldsymbol{\beta}}^*$ again. We see that $|\hat{\boldsymbol{\beta}}^*| < |\hat{\boldsymbol{\beta}}|$, so $\hat{\boldsymbol{\beta}}$ cannot be a lasso solution. The rest can be directly verified by the definition of the lasso, which is thus omitted.

Grouping effect

Theorem (1)

Given data (y, X) and parameters (λ_1, λ_2), the response y is centred and the predictors X are standardized. Let $\hat{\beta}(\lambda_1, \lambda_2)$ be the naive elastic net estimate. Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$. Define

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{\|y\|_1}\left|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)\right|;$$

then

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)},$$

where $\rho = x_i^T x_j$, the sample correlation.

Table of Contents

Deficiency of the naive elastic net

In the regression prediction setting, an accurate penalization method achieves good prediction performance through the bias–variance trade-off.

The naive elastic net estimator is a two-stage procedure: for each fixed λ_2 we first find the ridge regression coefficients, and then we do the lasso-type shrinkage along the lasso coefficient solution paths. It appears to incur a double amount of shrinkage.

Double shrinkage introduces unnecessary extra bias, compared with pure lasso or ridge shrinkage.

The elastic net estimate

Given data (y, X), penalty parameter (λ_1, λ_2) and augmented data (y^*, X^*), the naive elastic net solves a lasso-type problem

$$\hat{\beta}^* = \arg\min_{\beta^*} \|y^* - X^*\beta^*\|_2^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}} \|\beta^*\|_1 . \tag{8}$$

The elastic net (corrected) estimates $\hat{\beta}$ are defined by

$$\hat{\beta}(\text{elastic net}) = \sqrt{(1+\lambda_2)}\hat{\beta}^*. \tag{9}$$

Recall that $\hat{\beta}(\text{naive elastic net}) = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\beta}^*$; thus

$$\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net}). \tag{10}$$

Hence the elastic net coefficient is a **rescaled** naive elastic net coefficient.

The elastic net estimate

A strong motivation for the $(1+\lambda_2)$-rescaling comes from a decomposition of the ridge operator. Since the predictors X are standardized, we have

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ & 1 & \cdots & \cdot \\ & & \ddots & \cdot \\ & & & 1 \end{pmatrix}_{p \times p},$$

where $\rho_{i,j}$ is sample correlation. Ridge estimates with parameter λ_2 are given by $\hat{\beta}(\text{ridge}) = R\boldsymbol{y}$,

$$R = (X^T X + \lambda_2 I)^{-1} X^T.$$

The elastic net estimate

We can rewrite R as

$$R = \frac{1}{1+\lambda_2} R^* = \frac{1}{1+\lambda_2} \begin{pmatrix} \frac{1}{1+\lambda_2} & \frac{\rho_{12}}{1+\lambda_2} & \cdots & \frac{\rho_{1p}}{1+\lambda_2} \\ & \frac{1}{1+\lambda_2} & \cdots & \cdot \\ & & \ddots & \cdot \\ & & & \frac{1}{1+\lambda_2} \end{pmatrix}^{-1} X^T. \qquad (11)$$

R^* is like the usual OLS operator except that the correlations are shrunk by the factor $1/(1+\lambda_2)$, which we call *decorrelation*.

Hence from equation (11) we can interpret the ridge operator as decorrelation followed by direct scaling shrinkage.

The elastic net estimate

This decomposition suggests that the grouping effect of ridge regression is caused by the decorrelation step. When we combine the grouping effect of ridge regression with the lasso, the direct $1/(1+\lambda_2)$ shrinkage step is not needed and is removed by rescaling.

Although ridge regression requires $1/(1+\lambda_2)$ shrinkage to control the estimation variance effectively, in our new method, we can rely on the lasso shrinkage to control the variance and to obtain sparsity.

The elastic net estimate

Theorem (2)

Given data (y, X) *and parameters* (λ_1, λ_2)*, then the elastic net estimates* $\hat{\beta}$ *are given by*

$$\hat{\beta} = \arg\min_{\beta} \left[\beta^T \left(\frac{X^T X + \lambda_2 I}{1 + \lambda_2} \right) \beta - 2 y^T X \beta \right] + \lambda_1 \|\beta\|_1. \tag{12}$$

It is easy to see that

$$\hat{\beta}(\mathit{lasso}) = \arg\min_{\beta} \left[\beta^T (X^T X) \beta - 2 y^T X \beta \right] + \lambda_1 \|\beta\|_1. \tag{13}$$

The elastic net estimate

proof: Let $\hat{\beta}$ be the elastic net estimates. By definition and equation (10) we have

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{\sqrt{(1+\lambda_2)}} \|y^* - X^*\beta\|_2^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}} \|\beta\|_1 \right]$$

$$= \arg\min_{\beta} \beta^T \left(\frac{X^T X + \lambda_2 I}{1+\lambda_2} \right) \beta - 2y^T X\beta + \frac{\lambda_1}{1+\lambda_2} \|\beta\|_1 . \qquad (14)$$

Substituting the identities

$$X^{*T} X^* = \left(\frac{X^T X + \lambda_2 I}{1+\lambda_2} \right),$$

The elastic net estimate

(continued)

$$y^{*T}X^{*} = \frac{y^{T}X}{\sqrt{(1+\lambda_2)}},$$

$$y^{*T}y^{*} = y^{T}y$$

into equation (14), we have

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{1+\lambda_2}\left[\beta^{T}\left(\frac{X^{T}X+\lambda_2 I}{1+\lambda_2}\right)\beta - 2y^{T}X\beta + \lambda_1\|\beta\|_1\right] + y^{T}y$$

$$= \arg\min_{\beta} \beta^{T}\left(\frac{X^{T}X+\lambda_2 I}{1+\lambda_2}\right)\beta - 2y^{T}X\beta + \lambda_1\|\beta\|_1 .$$

Connections with univariate soft thresholding

The lasso is a special case of the elastic net with $\lambda_2 = 0$. The other interesting special case of the elastic net emerges when $\lambda_2 \to \infty$. By theorem 2, $\hat{\beta} \to \hat{\beta}(\infty)$ as $\lambda_2 \to \infty$, where

$$\hat{\beta}(\infty) = \arg\min_{\beta} \left[\beta^T \beta - 2y^T X\beta + \lambda_1 \|\beta\|_1 \right].$$

Connections with univariate soft thresholding

$\hat{\beta}(\infty)$ has a simple closed form

$$\hat{\beta}(\infty)_i = \left(|y^T X_i| - \frac{\lambda_1}{2} \right)_+ \text{sgn}(y^T X_i), \quad i = 1, 2, ..., p. \quad (15)$$

Observe that $y^T X_i$ is the univariate regression coefficient of the ith predictor and $\hat{\beta}(\infty)$ are the estimates by applying soft thresholding on univariate regression coefficients; thus equation (15) is called univariate soft thresholding (UST).

Computation: the algorithm LARS-EN

In a word, we do not explicitly use X^* to compute all the quantities in algorithm LARS. It is also economical to record only the non-zero coefficients and the active variables set at each LARS-EN step.

Choice of tuning parameters

For each fixed λ_2, the elastic net is solved by our algorithm LARS-EN; hence similarly we can use the number of the LARS-EN steps (k) as the second tuning parameter besides λ_2.

Table of Contents

Prostate Cancer Example

- 8 predictors: log(cancer volume), log(prostate weight), age, the logarithm of the amount of benign prostatic hyperplasia, seminal vesicle invasion, log(capsular penetration), Gleason score and percentage Gleason score 4 or 5.
- The response is the logarithm of prostate-specific antigen.

OLS, Ridge Regression, LASSO and Elastic Net

- OLS, ridge regression, the lasso, the naive elastic net and the elastic net were applied.
- Training set: 67 observations; Test set: 30 observations.
- Model fitting and tuning parameter selection by tenfold CV were carried out on the training data.

Comparison

Table 1. Prostate cancer data: comparing different methods

Method	*Parameter(s)*	*Test mean-squared error*	*Variables selected*
OLS		0.586 (0.184)	All
Ridge regression	$\lambda = 1$	0.566 (0.188)	All
Lasso	$s = 0.39$	0.499 (0.161)	(1,2,4,5,8)
Naïve elastic net	$\lambda = 1, s = 1$	0.566 (0.188)	All
Elastic net	$\lambda = 1000, s = 0.26$	0.381 (0.105)	(1,2,5,6,8)

- Elastic net is the winner in terms of both prediction accuracy and sparsity.
- OLS is the worst.
- Naive elastic net is identical to ridge regression.
- The prediction error: elastic net is about 24% lower than lasso.
- Elastic net is UST(Univariate Soft thresholding), because λ selected is very big.

Table of Contents

Simulation

- The simulated data comes from the true model: $y = x\beta + \sigma\epsilon, \epsilon \sim N(0, 1)$.
- Each simulated dataset is divided into training set/ validation set/ test set to serve. Models were fitted on the training set only, and the validation data were used to select the tuning parameters.
- The test error (the mean-squared error) was computed on the test set.

Simulation Example 1 and 2

- Simulation example 1: 50 data sets were simulated consisting of 20/20/200 observations and 8 predictors:

$$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0), \sigma = 3$$

$$\text{and } \text{cov}(x_i, x_j) = (0.5)^{|i-j|} \text{ for all } i, j = 1, \ldots, 8.$$

- Simulation example 2: Same as example 1, except $\beta_j = 0.85$ for all j.

Simulation Example 3

- Simulation example 3: 50 data sets were simulated consisting of 100/100/400 observations and 40 predictors:

$$\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$$

and $\sigma = 15, \text{cor}(x_i, x_j) = 0.5$ for all $i, j = 1, \ldots, 40$.

Simulation Example 4

- Simulation example 4: 50 data sets were simulated consisting of 50/50/400 observations and 40 predictors:

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25}), \text{and } \sigma = 15.$$

$$\mathbf{x}_i = Z_1 + \epsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1, \ldots, 5.$$

$$\mathbf{x}_i = Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, \ldots, 10.$$

$$\mathbf{x}_i = Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 11, \ldots, 15.$$

$$\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} N(0,1), \quad i = 16, \ldots, 40$$

$$\epsilon_i^x \overset{\text{i.i.d.}}{\sim} N(0, 0.01), \quad i = 1, \ldots, 15$$

Simulated Examples - Median MSE

Table 2. Median mean-squared errors for the simulated examples and four methods based on 50 replications†

Method	*Results for the following examples:*			
	Example 1	*Example 2*	*Example 3*	*Example 4*
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
Ridge regression	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Naïve elastic net	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

†The numbers in parentheses are the corresponding standard errors (of the medians) estimated by using the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors.

- Elastic Net is more accurate than the LASSO in all four examples, even when the LASSO is significantly more accurate than Ridge regression.
- The Naive Elastic Net performs very poorly with the highest mean-squared error in Example 1. In Example 2 and 3 it behaves very similar to Ridge regression, and in Example 4 it behaves similar to the LASSO.

Simulated Examples - Median MSE

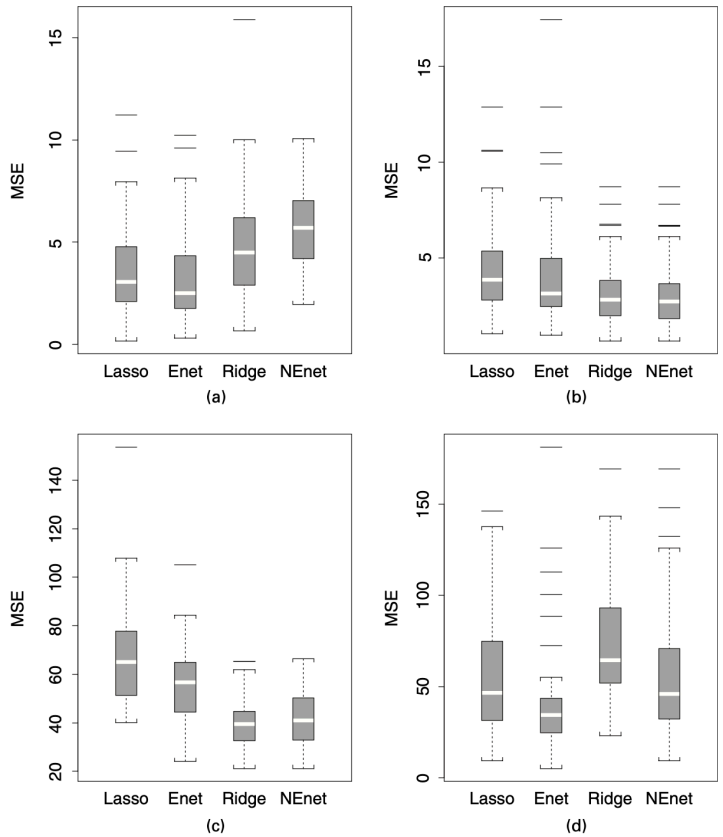


Fig. 4. Comparing the accuracy of prediction of the lasso, the elasic net (Enet), ridge regression and the naïve elastic net (NEnet) (the elastic net outperforms the lasso in all four examples): (a) example 1; (b) example 2; (c) example 3; (d) example 4

- Using the box-plot, the overall prediction performance of the LASSO, ridge, elastic net, and naive elastic net is compared for 4 example.

Simulated Examples - Variable Selection

Table 3. Median number of non-zero coefficients

Method	*Results for the following examples:*			
	Example 1	*Example 2*	*Example 3*	*Example 4*
Lasso	5	6	24	11
Elastic net	6	7	27	16

- Elastic Net selects more predictors than the LASSO due to the grouping effect.
- Elastic Net behaves like the ideal model in Example 4, where grouped selection is needed.
- Therefore, the Elastic Net has the additional ability to perform grouped variable selection, which makes it a better variable selection method than the LASSO.

Grouping Effect

- Important difference between the elastic net and LASSO: Grouping Effect
- Let Z_1 and Z_2 be two independent $U(0, 20)$ variables. The response y is generated as $N(Z_1 + 0.1Z_2, 1)$.
- $$\mathbf{x}_1 = Z_1 + \varepsilon_1, \quad \mathbf{x}_2 = -Z_1 + \varepsilon_2, \quad \mathbf{x}_3 = Z_1 + \varepsilon_3,$$

  $$\mathbf{x}_4 = Z_2 + \varepsilon_4, \quad \mathbf{x}_5 = -Z_2 + \varepsilon_5, \quad \mathbf{x}_6 = Z_2 + \varepsilon_6,$$
- where ε_i are independent identically distributed $N(0, 1/16)$.
- Within-group correlation is almost 1. Between-group correlations are almost 0.
- An oracle would identify the Z1-group as the important variates.

Grouping Effect

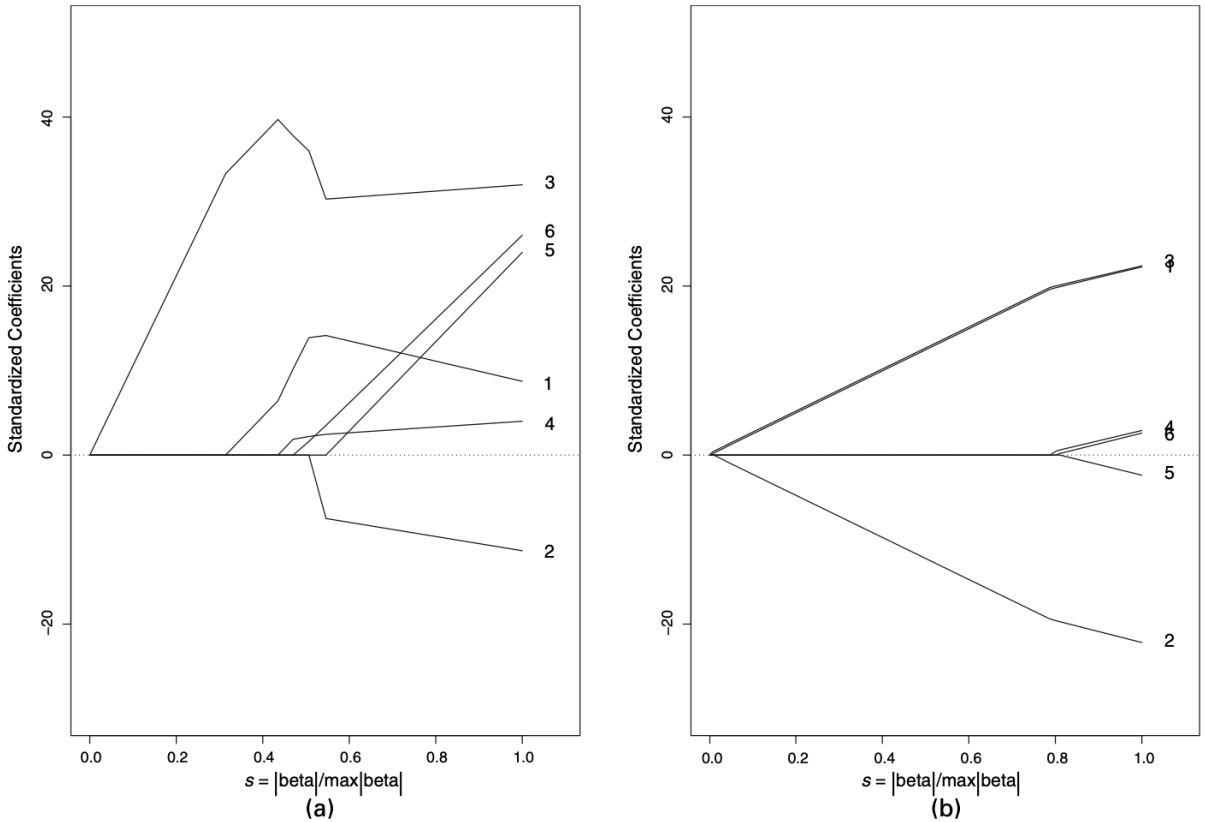


Fig. 5. (a) Lasso and (b) elastic net ($\lambda_2 = 0.5$) solution paths: the lasso paths are unstable and (a) does not reveal any correction information by itself; in contrast, the elastic net has much smoother solution paths, while clearly showing the 'grouped selection'—$\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ are in one 'significant' group and $\mathbf{x}_4$, $\mathbf{x}_5$ and $\mathbf{x}_6$ are in the other 'trivial' group; the decorrelation yields the grouping effect and stabilizes the lasso solution

- LASSO: Unstable; No correlation information
- Elastic Net: Smoother solution path; Grouped selection(x_1, x_2, x_3 in significant group, x_4, x_5, x_6 in trivial group)

Conclusion

- The LASSO can select at most n predictors in the $p > n$ case and cannot perform grouped selection. Furthermore, the ridge regression usually has a better prediction performance than the LASSO when there are high correlations between predictors in the $n > p$ case.
- The Elastic Net can produce a sparse model with good prediction accuracy, while selecting group(s) of strongly correlated predictors. It can also potentially select all p predictors in all situations. paths efficiently, similar to the LARS algorithm for LASSO.
- The Elastic Net has two tuning parameters as opposed to one tuning parameter like the LASSO, which can be selected using a training and validation set.
- Simulation results indicate that the Elastic Net dominates the LASSO, especially under collinearity.

Table of Contents

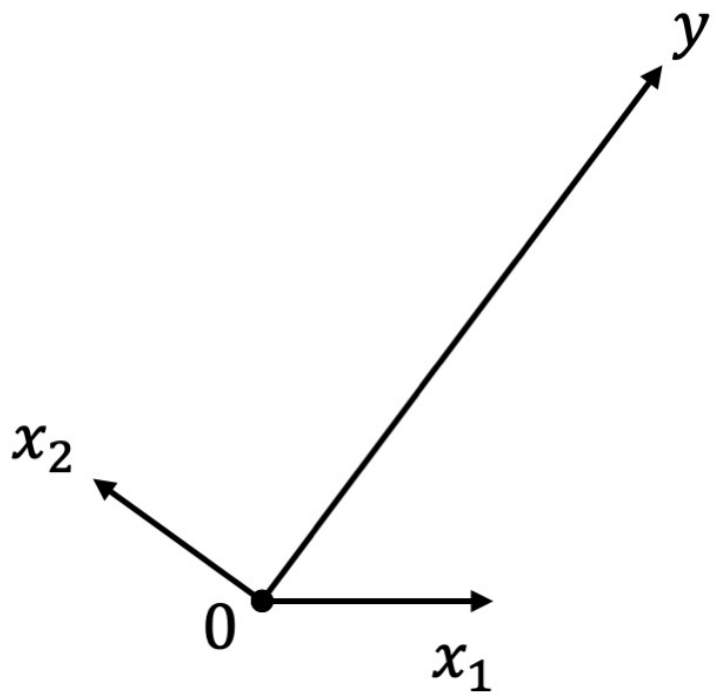
Least Angle Regression

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.

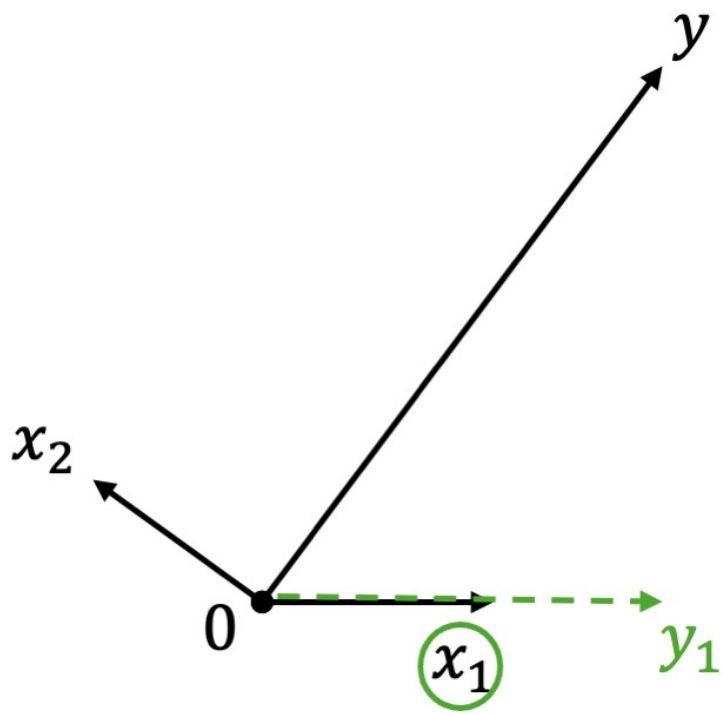
1. Start with a null model.



Forward Stepwise Selection

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

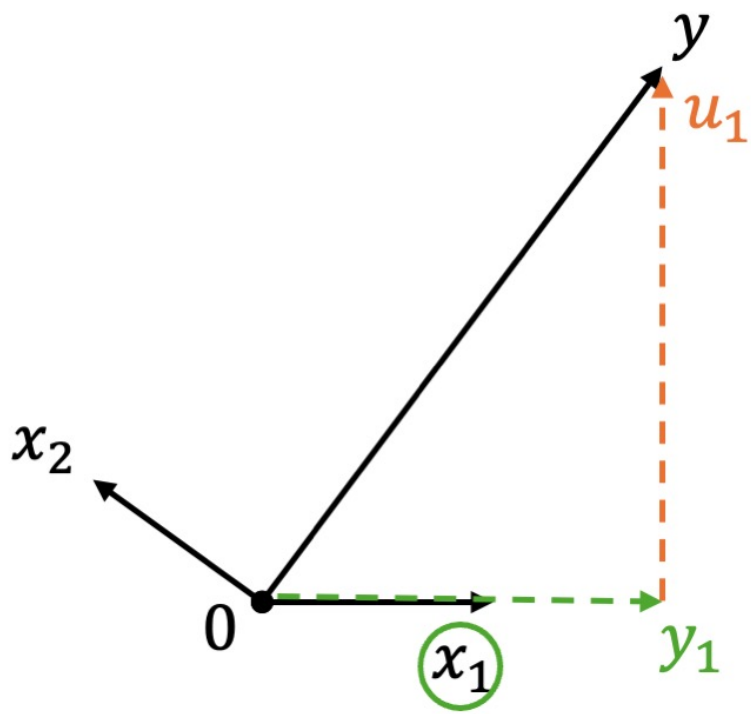
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.



Forward Stepwise Selection

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

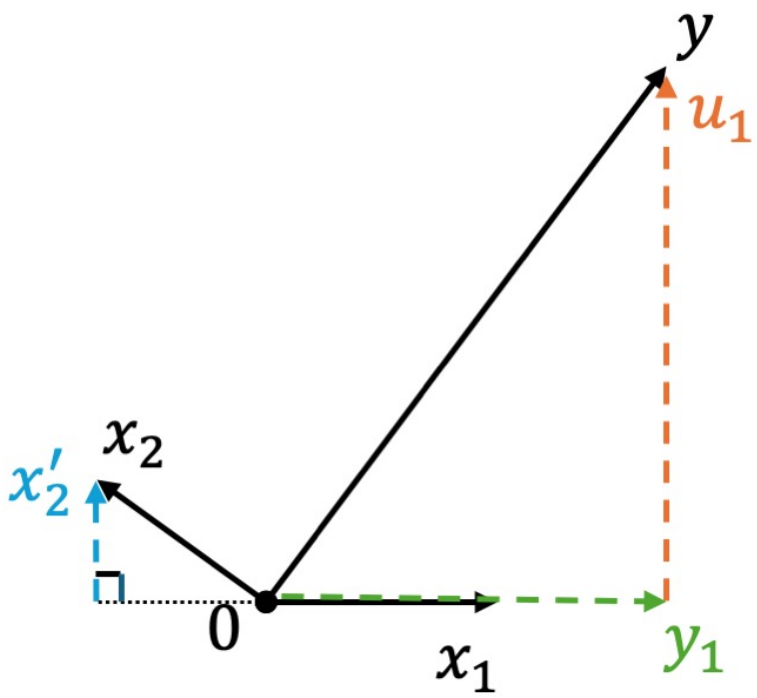
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.



Forward Stepwise Selection

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

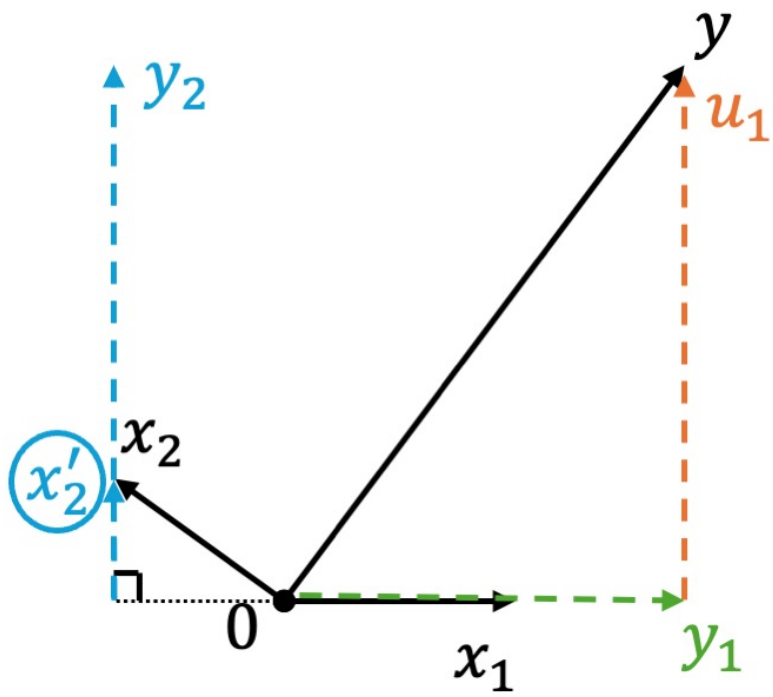
1. Start with a null model.
2. Find the predictor most correlated with the response and perform simple linear regression.
3. Set the residuals as the new response.
4. Project other predictors orthogonal to the predictor selected in previous step.



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.

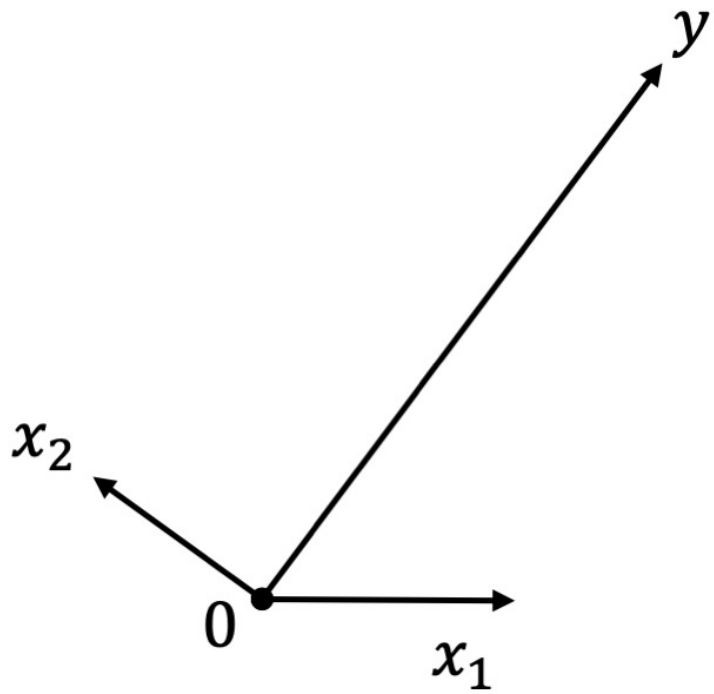


Forward Stagewise Selection

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 $\hat{\mu} = 0$ and a null model.

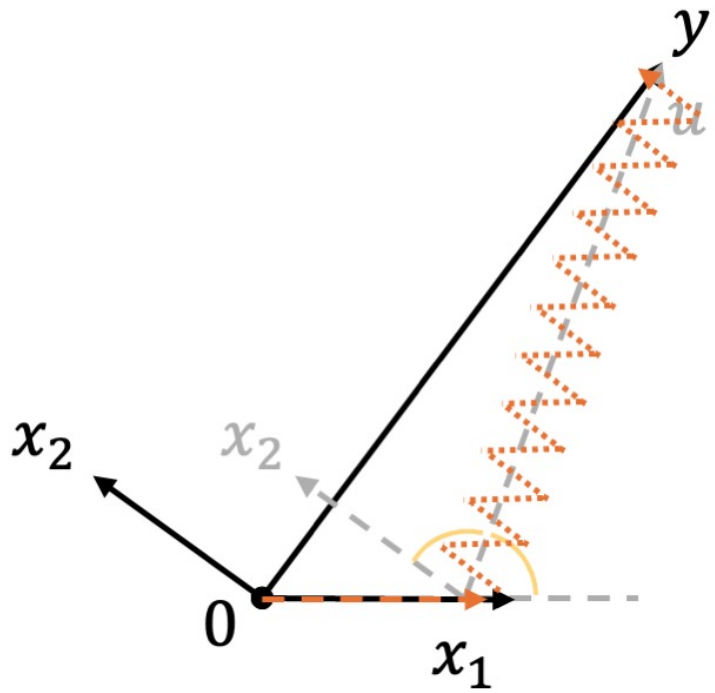


香港城市大學
City University of Hong Kong

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 $\hat{\mu} = 0$ and a null model.
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.

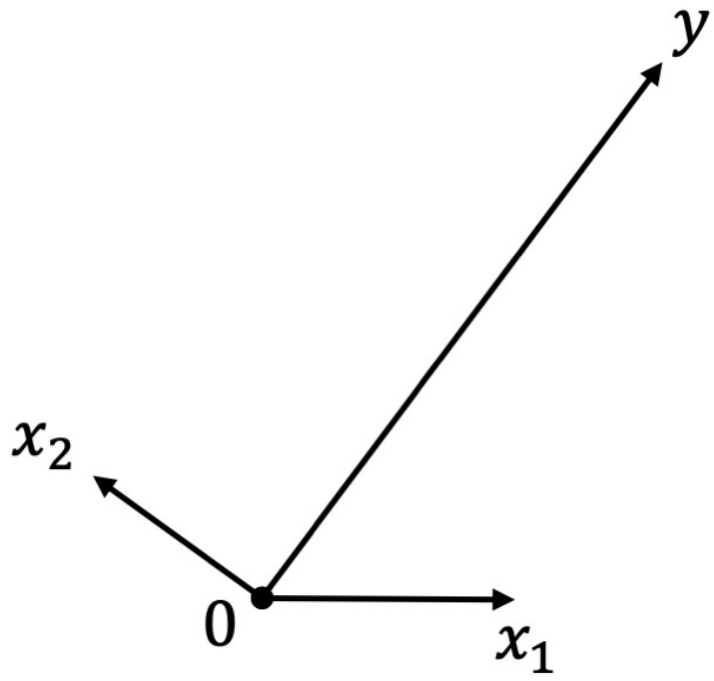


香港城市大學
City University of Hong Kong

Least Angle Regression: Example

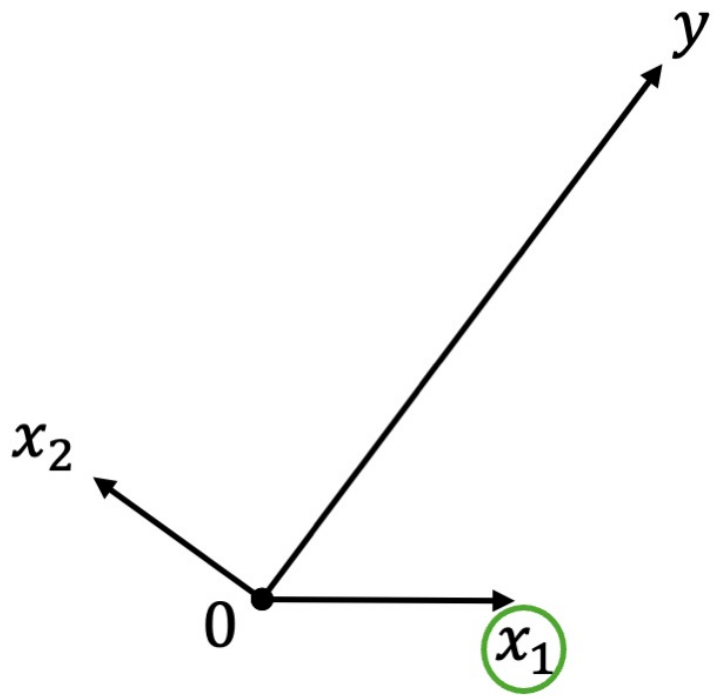
Least Angle Regression (LAR) is a stylized version of forward stagewise procedure that uses a simple mathematical formula to accelerate the computations.

1. Start with $\hat{\mu} = 0$ and a null model.



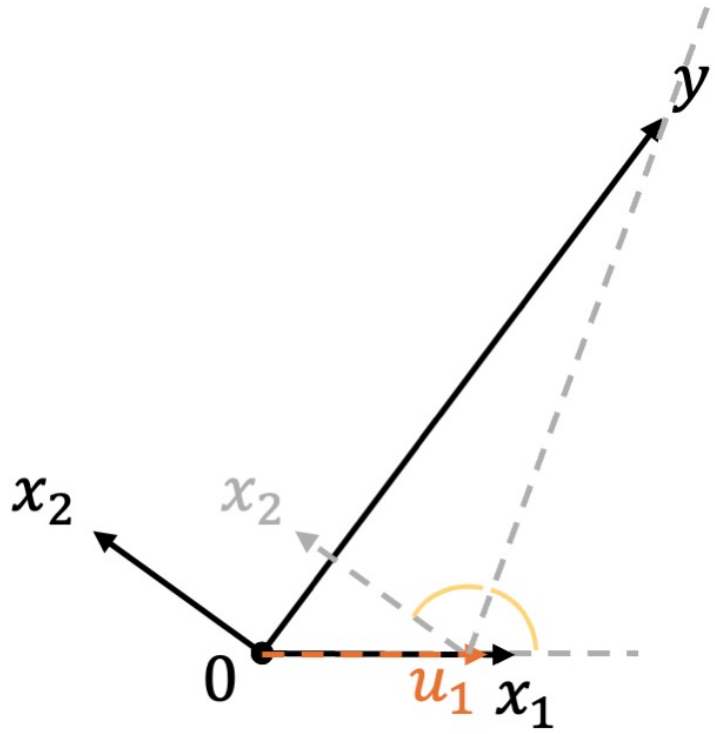
Least Angle Regression: Example

1. Start with $\hat{\mu} = 0$ and a null model.
2. Find the predictor most correlated with the response.



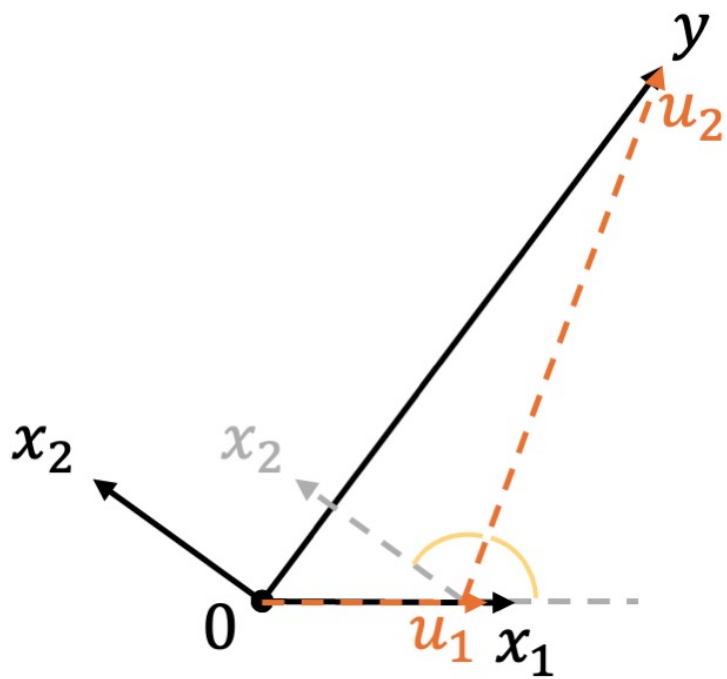
Least Angle Regression: Example

1. Start with $\hat{\mu} = 0$ and a null model.
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 $\hat{\mu} = 0$ and a null model.
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: L1 Arc Length

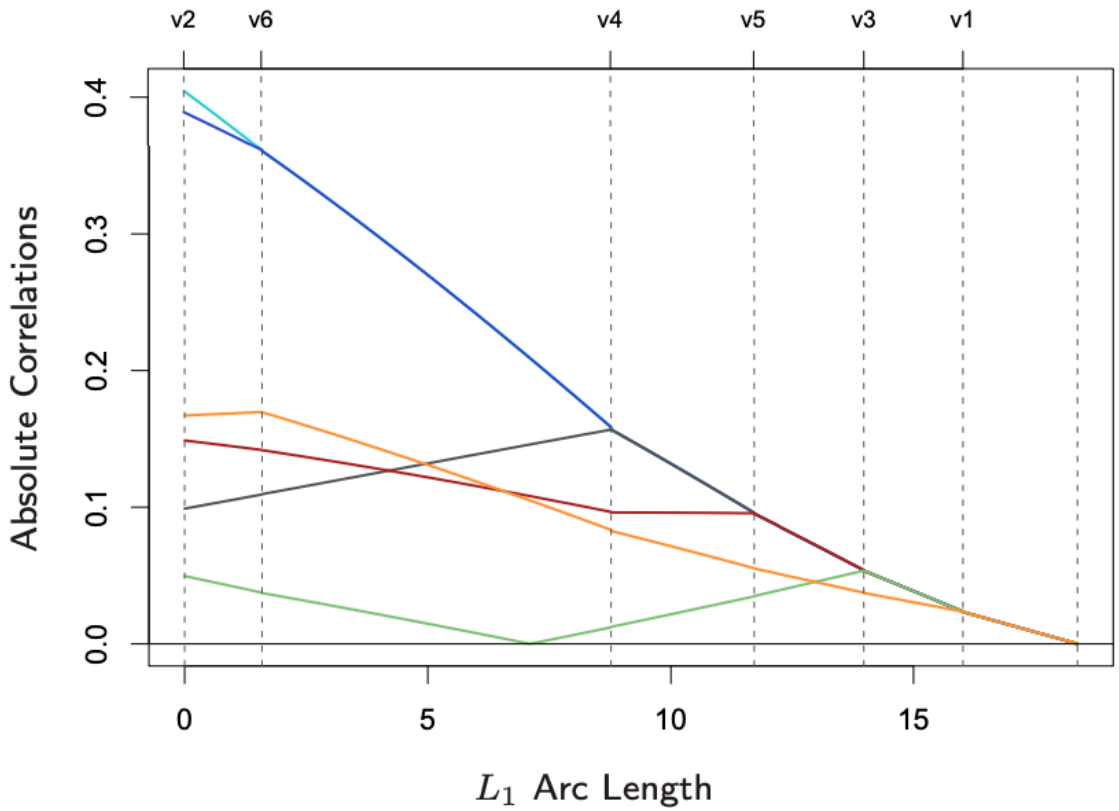


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

Least Angle Regression: The Equiangular Vector

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

$$g_\mathcal{A} = \mathbf{X}_\mathcal{A}^T \mathbf{X}_\mathcal{A} \quad \text{and} \quad A_\mathcal{A} = (\mathbf{1}_\mathcal{A}^T g_\mathcal{A}^{-1} \mathbf{1}_\mathcal{A})^{-1/2}, \tag{14}$$

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

$$\mathbf{u}_\mathcal{A} = \mathbf{X}_\mathcal{A} \omega_\mathcal{A}, \text{ where } \omega_\mathcal{A} = A_\mathcal{A} g_\mathcal{A}^{-1} \mathbf{1}_\mathcal{A}, \tag{15}$$

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: 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 (14) and (15). Also, compute the inner product $\mathbf{a} =: X^T\mathbf{u}_{\mathcal{A}}$. Updates $\hat{\mu}_{\mathcal{A}}$ as

$$\hat{\mu}_{\mathcal{A}} \leftarrow \hat{\mu}_{\mathcal{A}} + \hat{\gamma}\mathbf{u}_{\mathcal{A}}, \tag{16}$$

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.

Extend LAR to Lasso Regression

If a non-zero coefficient hits zero, drop its variable from the active set and recompute the current joint least squares direction. This is the modification to LAR for Lasso.
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 LAR to Stagewise Regression

If we modify the active set $\mathcal{A}$ (so that $\omega_{\mathcal{A}}$ would not have negative components), we can extend LAR to 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 the Solution Paths of LARS

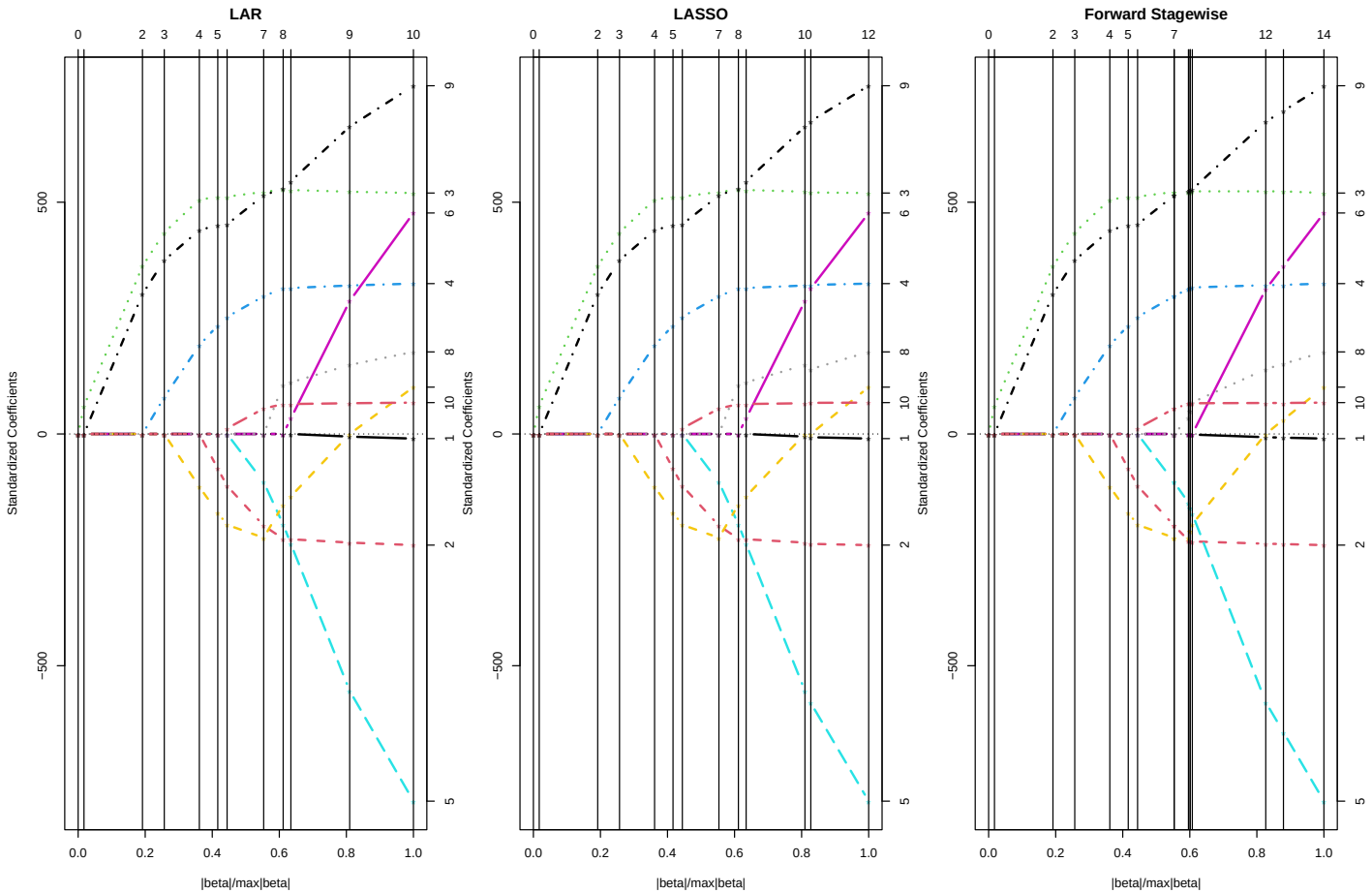


Figure: Solution paths of LAR, LAR-lasso and Forward Stagewise Selection for the diabetes data set.

Comparison of Computational Time

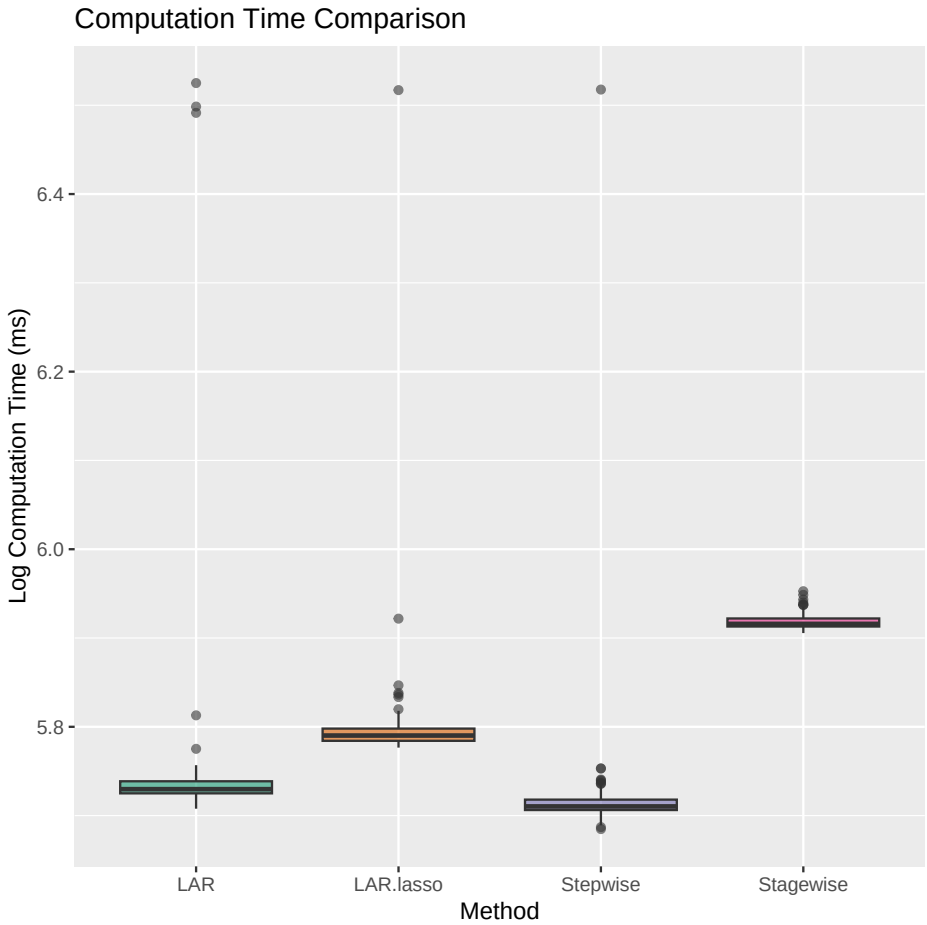


Figure: Comparison of computational time between LAR, LAR-Lasso, Forward Stagewise Selection, and Forward Stepwise Selection with the diabetes data set.

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 $\delta, i \Rightarrow f(x) = \min_z f(z)$ (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 $\delta, i \Rightarrow f(x) = \min_z f(z)$ (Here $e_i = (0, \ldots, 1, \ldots 0)$, the i-th standard basis vector) always hold?
Yes. **Proof**:

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

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?

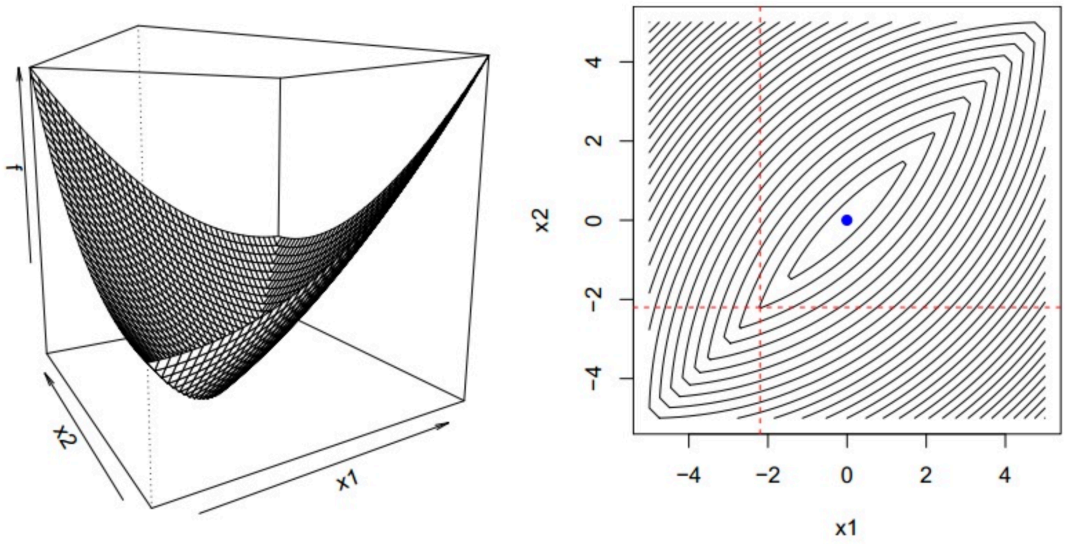


Figure: f is not differentiable along the diagonal, but is convex. The global minimizer is at the origin (centre).

No. We can see that the cross-point is minimized for each axis, but only the origin is the global minimizer.

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)?

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

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

We now want to proof

$$\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i) \geq 0. \tag{17}$$

Coordinate Descent: Motivation Question 3

We now want to proof

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

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

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

then by definition of subgradient:

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

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:

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

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

$$\ldots$$

$$x_n^{(k)} \in \arg\min_{x_n} f\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n\right)$$

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

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:

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

Note that the nonsmooth part is separable: $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$.
We can perform coordinate descent by repeatedly minimize over β_i for solving:

$$0 = X_i^T (X_i\beta_i + X_{-i}\beta_{-i} - y) + \lambda s_i, \tag{18}$$

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

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

Coordinate Descent: Lasso Regression

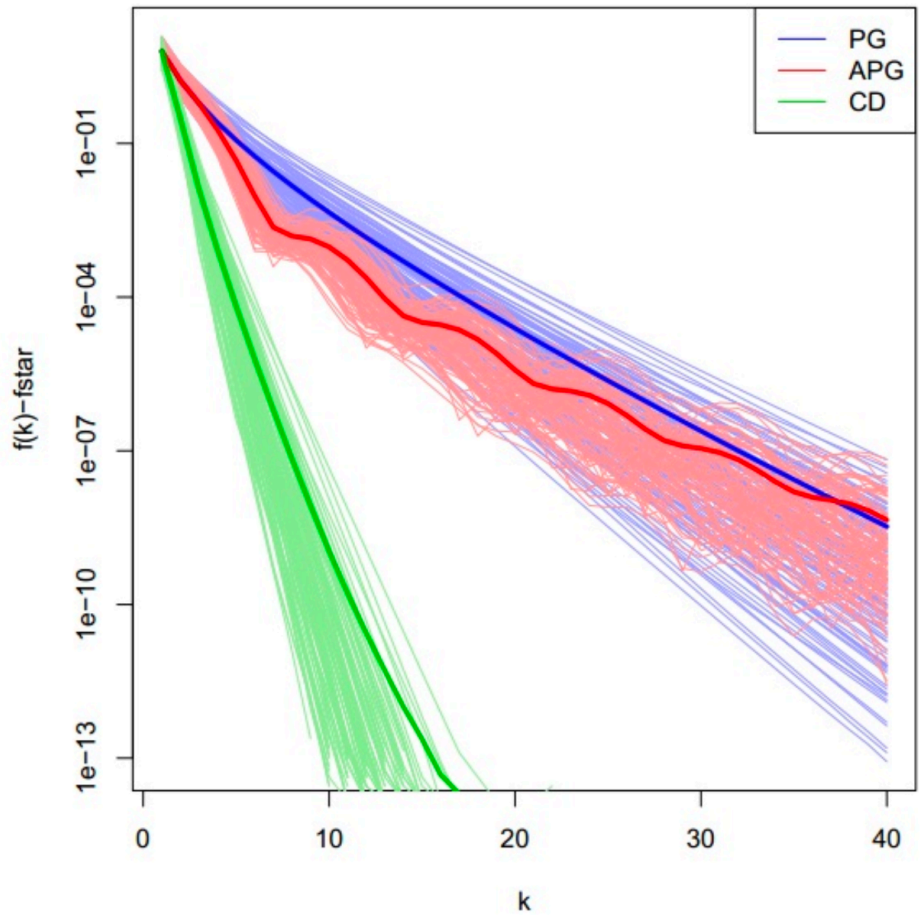
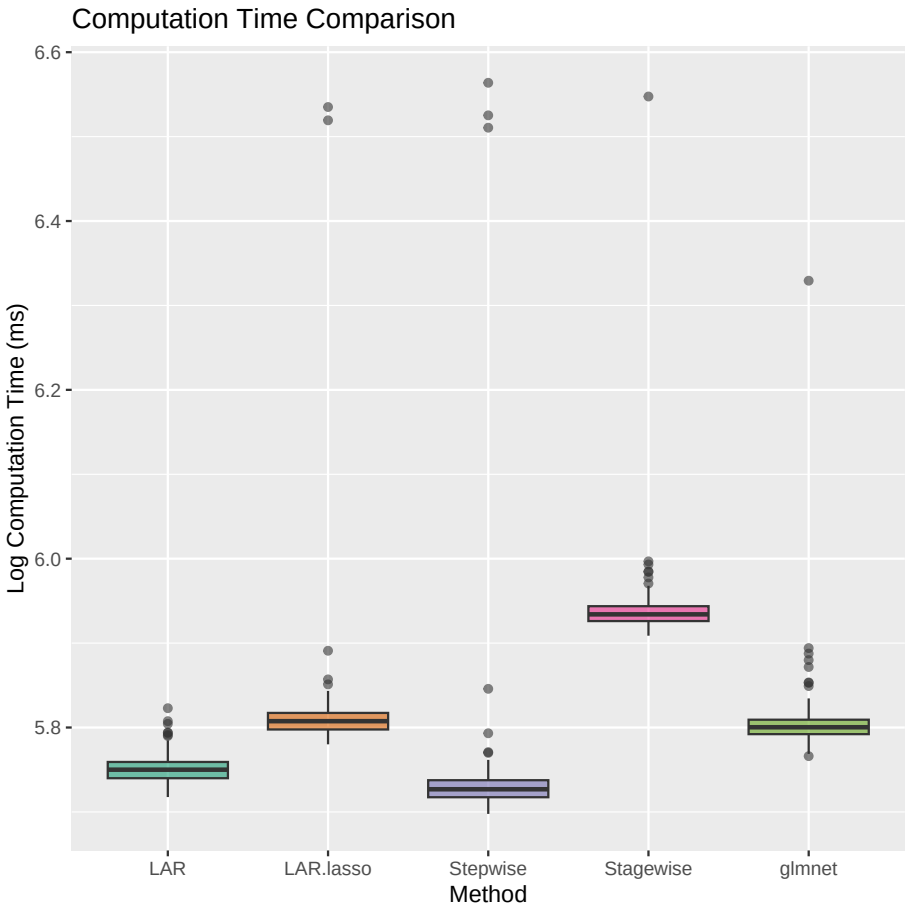


Figure: Coordinate descent and (accelerated) proximal gradient descent for lasso regression with $n = 100, p = 20$. Note that both GD and CD cost $O(np)$ operators in one cycle.

LARS VS Coordinate Descent: Computational Time



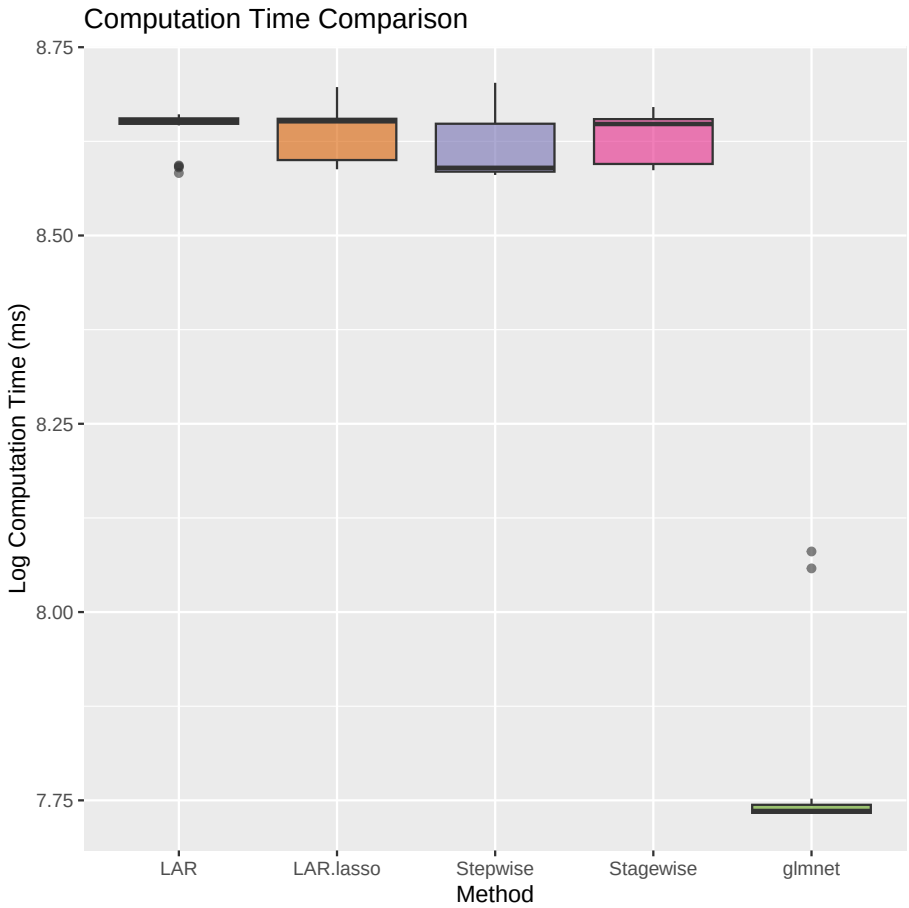


Figure: Comparison of computational time between LAR, LAR-Lasso, Forward Stagewise Selection, Forward Stepwise Selection and glmnet.lasso with the diabetes data set ($n = 442, p = 10$) and a simulated data set ($n = 1e4, p = 200$).

References

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

[5] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. New York: springer, 2009.