Linear model for regression

Peida Wu

Institute of Math and Science

2025 年 5 月 6 日

Review

1. Least square
2. subset selection
3. Ridge regression
4. Lasso regression

Least angle regression

LAR was introduced in 2004, which is stable and highly interpreted model. It was introduced to solve the following problem:

1. In high dimension data, LAR select the most correlated variable to reduce the complexity of model to avoid overfitting.
2. More efficeint compute method, especially for a large scale of data.
3. Similar to Lasso regression, but it's more interpretable.

Least angle regression

Core idea: stepwisely select regression coefficients, introduce ther most correlated feature to the model.
Basic principle:

1. At each step, the variable direction most correlated with the current residual is selected and introduced into the model.
2. The algorithm moves forward in small steps until another variable's correlation with the residual becomes equal, at which point the new variable is introduced.
3. LARS always proceeds along the least angle direction, maximizing the explanatory power for the response variable.

Least Angle Regression

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_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 β_j and β_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.

Least angle regression

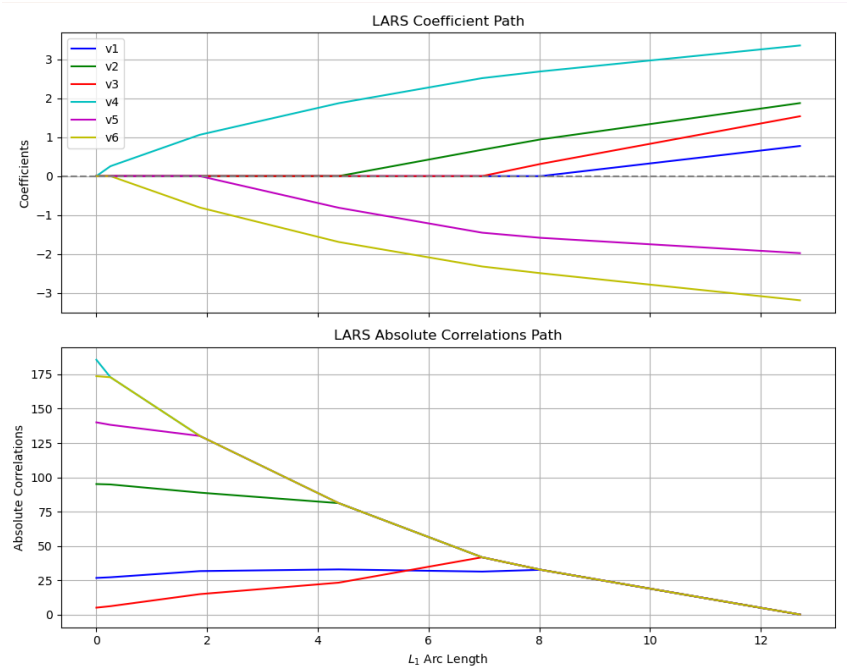
Suppose A_k is active set of variables in kth step.
β_{A_k} is coefficient vector, (k-1) nonzero vectors.
$r_k = y - X_{A_k}\beta_{A_k}$ is current residue.
The direction for the step is

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

The coefficient evlvoves as $\beta_{A_k}(\alpha) = \beta_{A_k} + \alpha\delta_{A_k}$
The fit vector at the beginning of the step is $\hat{f}_k$. It envolves as $\hat{f}_k(\alpha) = \hat{f}_k + \alpha u_k$, $u_k = X_{A_k}\delta_k$

Least angle regression

Denote $y = x_1 + 2x_2 + 1.5x_3 + 4x_4 - 2x_5 - 3x_6 + \epsilon$



Least square regression

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_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 β_j and β_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.

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.

Least square regression

Algorithm 3.2 with the lasso modification 3.2a is an efficient way of computing the solution to any lasso problem, especially when $p >> N$. $\mathcal{A}$: the active set of variables at some stage

$$\mathbf{x}_j^T(y - X\beta) = \gamma \cdot s_j, \quad \forall j \in \mathcal{A}, \tag{1}$$

where $s_j \in \{-1, 1\}$ indicates the sign of the inner-product, and γ is the common value. Also $|\mathbf{x}_k^T(y - X\beta)| \leq \gamma \; \forall k \notin \mathcal{A}$.
the lasso criterion

$$R(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{2}$$

Let $\mathcal{B}$ be the active set of variables in the solution for a given value of λ. the stationarity conditions give

$$\mathbf{x}_j^T(y - X\beta) = \lambda \cdot \text{sign}(\beta_j), \quad \forall j \in \mathcal{B}. \tag{3}$$

dimension reduction method

Original predictors: $X_1, \ldots, X_p$.
$Z_1, \ldots, Z_M$ represent $M < p$ linear combinations of original p predictors.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for some constants $\phi_{1m}, \ldots, \phi_{pm}$. Fitting the linear regression model:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im}, i = 1, \ldots, n$$

Notice that:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Principal component regression

$X \in \mathbf{R}^{N \times p}$, by SVD,

$$X = UDV^T$$

The sample covariance matrix is given by

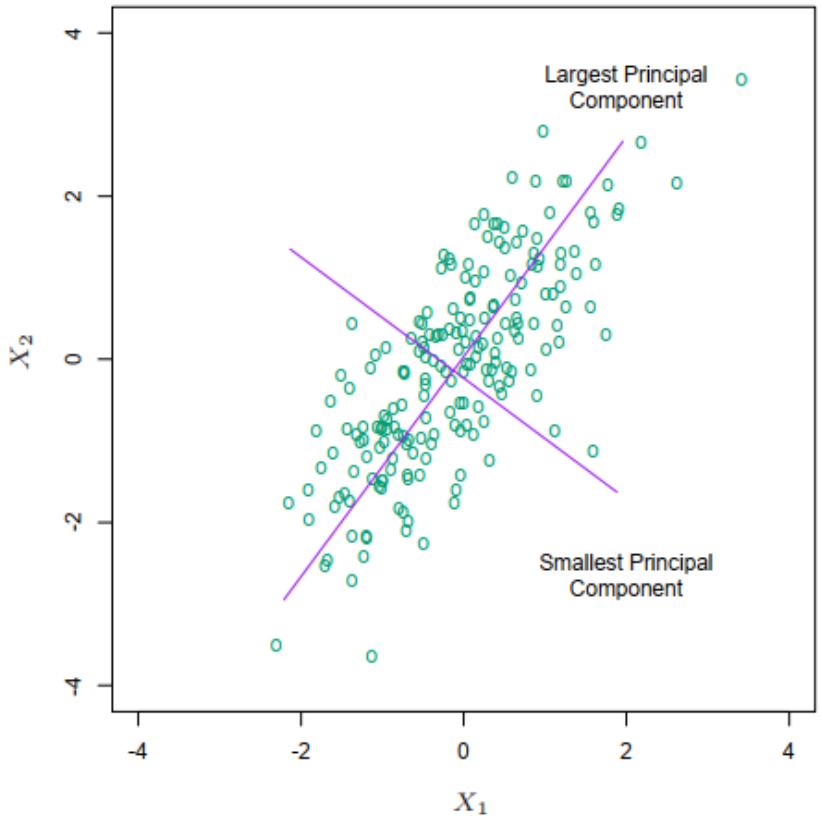
$$S = X^T X/N = VD^2 V^T/N$$

The first principal component v_1 has the property $z_1 = Xv_1$ has the largest sample variance.

$$Var(z_1) = Var(Xv_1) = \frac{d_1^2}{N}$$

$z_1 = Xv_1 = u_1 d_1$, where z_1 is the first principal component, u_1 is normalized first component.
Subsequent principal components z_j have maximum variance d_j^2/N, subject to being orthogonal to the earlier ones.

Principal component regression regression



Partial least regression

The coefficients of PCA only consider the distribution of the observed measurement X. Its purpose is to find the components with the widest variation in X.
In regression systems, there are two requirements for the data:

1. The range of variation for both observed data and predicted data should be large.
2. There should be a correlation between observed data and predicted data.

To address these two points, Partial Least Squares Regression (PLS) is introduced.

Partial least regression

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \ldots, p$.
2. For $m = 1, 2, \ldots, p$
   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.
   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.
   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.
   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle]\mathbf{z}_m,\ j = 1, 2, \ldots, p$.
3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Partial least regression

The mth principal component direction v_m solves:

$$\max_{\alpha} \mathrm{Var}(X\alpha)$$
$$\text{subject to} \|\alpha\| = 1, \alpha^T S v_\ell = 0, \ell = 1, \ldots, m-1,$$

The mth PLS direction φ_m solves:

$$\max_{\alpha} \mathrm{Corr}^2(y, X\alpha)\,\mathrm{Var}(X\alpha)$$
$$\text{subject to} \|\alpha\| = 1, \alpha^T S \hat{\varphi}_\ell = 0, \ell = 1, \ldots, m-1.$$

Pathwise Coordinate Optimization

An alternative algoritm to the LARS for computing the lasso problem.
Main idea: The idea is to fix the penalty parameter in the Lagrangian form and optimize successively over each parameter, holding the other parameters fixed at their current values.

Pathwise Coordinate Optimization

Suppose the predictors are standardized to mean zero and unit norm. The objective function is:

$$R(\vec{\beta}(\lambda), \beta_j) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k(\lambda) - x_{ij}\beta_j\right)^2 + \lambda\sum_{k\neq j}|\hat{\beta}_k(\lambda)| + \lambda|\beta_j|$$

which is a univariate lasso problem. The partial residue is $y_i - \bar{y}_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k(\lambda)$ The update become:

$$\tilde{\beta}_j(\lambda) \leftarrow S\left(\sum_{i=1}^{N} x_{ij}(y_i - \bar{y}_i^{(j)}), \lambda\right).$$

where $S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$

Thanks!