3.4 Shrinkage Methods

exclude β_0 from the penalty term

3.4.1 Ridge

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad \lambda > 0$$

normally standardize the inputs before solving the above formula

In matrix form: $RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

$-2X^T(y - X\beta) + 2\lambda\beta = 0$

$-X^Ty + X^TX\beta + \lambda\beta = 0$

$(X^TX + \lambda I)\beta = X^Ty$

Then $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$

For orthonormal inputs, $\hat{\beta}^{ridge} = \frac{\hat{\beta}}{1+\lambda}$

$$df(\lambda) = tr[X(X^TX + \lambda I)^{-1}X^T] = tr(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \qquad (3.50)$$

Exercise 3.6 prior dist. $\beta_j \overset{iid}{\sim} N(0, \tau^2)$

$$f(\beta|y) = \frac{f(\beta, y)}{f(y)} = \frac{f(y|\beta)f(\beta)}{\int_\beta f(y|\beta)f(\beta)d\beta} \propto f(y|\beta)f(\beta)$$

$\int_\beta f(y|\beta)f(\beta)d\beta$ ↓ Constant

$= N(\beta_0 + x_i^T\beta, \sigma^2) \cdot N(0, \tau^2)$

$f(\beta|y)$ is a function of $f(y|\beta)f(\beta)$ or like follow the same (scale different) distribution

$$\int_{\beta|y} C f(\beta|y)\, d\beta = 1$$

Thin SVD of X matrix.

$X = UDV^T$

$N \times p \quad p \times p \quad p \times p$

only calculated the p column vectors which correspond to p row vectors of V^T

Then $X\hat{\beta} = X(X^TX)^{-1}X^Ty = UU^Ty$

$UDV^T(VD^TU^TUDV^T)^{-1}VD^TU^Ty$

$UDV^T(V^T)^{-1}D^{-1}(D^T)^{-1}V^{-1}VD^TU^Ty$

V^T · 1 · $(V^T)^{-1}$, V · 1 · V^{-1}, D^{-1} · 1 · $(D^T)^{-1}$, D^T · 1 · D, U · 1 · U^T

Then ridge $X\hat{\beta}^{ridge} = X(X^TX+\lambda I)^{-1}X^Ty$

$= X(VD^2V^T+\lambda VV^T)^{-1}VDU^Ty$

$= X(V(D^2+\lambda I)V^T)^{-1}VDU^Ty$

$= X(V^T)^{-1}(D^2+\lambda I)^{-1}\underbrace{V^{-1}V}_{I}DU^Ty$

$= X(V^T)^{-1}(D^2+\lambda I)^{-1}DU^Ty$

$= UDV^T\underbrace{(V^T)^{-1}}_{I}(D^2+\lambda I)^{-1}DU^Ty$

$= UD(D^2+\lambda I)^{-1}DU^Ty$

$= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^Ty$

↓ Column of U

Sample covariance matrix: $S = X^TX/N$

Where $X^TX = \underline{VD^2V^T}$ → eigen decomposition

first principle component direction v_1: $z_1 = Xv_1$ has the largest (sample) variance.

$Var(z_1) = Var(Xv_1)$

$= v_1^T Var(X) v_1$

$= v_1^T \frac{X^TX}{N} v_1$

$= v_1^T \frac{VD^2V^T}{N} v_1 = \frac{d_1^2}{N}$

where $v_1^Tv = [1\ 0\ 0\cdots]$

$$Var(z_1) = \frac{[1\ 0\ 0\cdots]\begin{bmatrix} d_1^2 & & & \\ & d_2^2 & & 0 \\ & & \ddots & \\ 0 & & & d_p^2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}}{N} = \frac{d_1^2}{N}$$

3.4.2 Lasso

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le t$$

equals

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

↓ nonlinear in the y_i — no closed form expression as in ridge

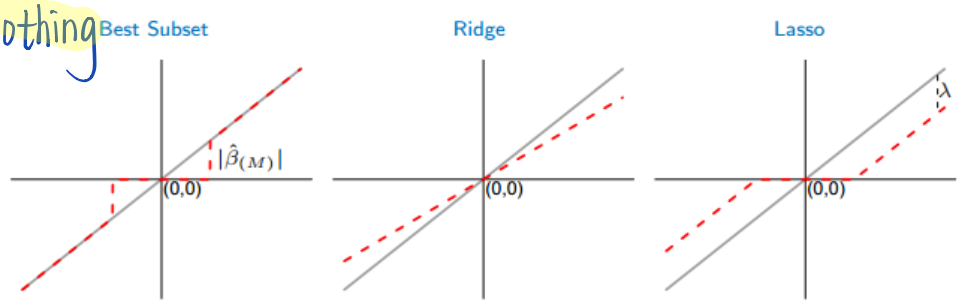
eg: $t_0 = \sum_1^p |\beta_j|$ if $t = \frac{t_0}{2}$, LSE coefficients are shrunk by about 50% on average.

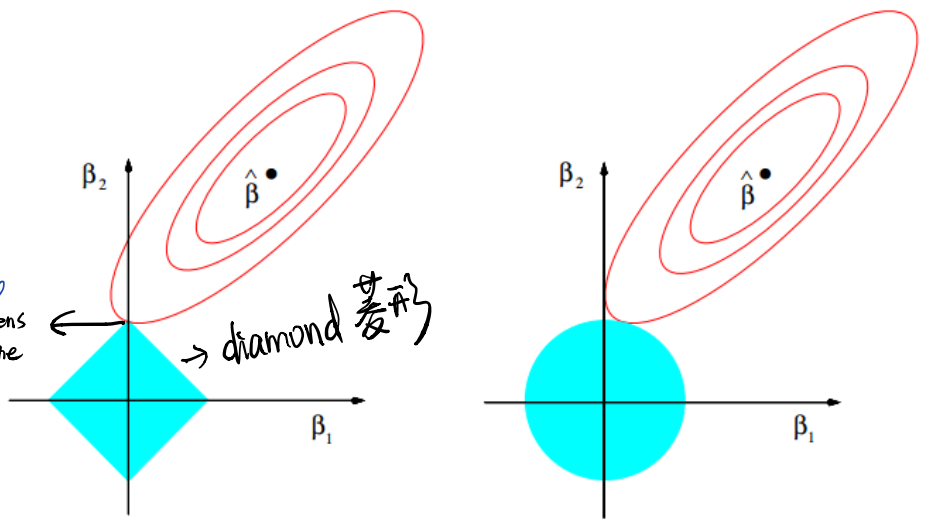
3.4.3 Generalization of Regularization

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques; sign denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I[\text{rank}(\lvert\hat{\beta}_j\rvert \le M)$ ← hard thresholding
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$ ← soft thresholding

example: section 5.9 ← wavelet-based smoothing ↓ 基于小波光滑





→ diamond 菱形

If the first point where the elliptical contours hit the constraint region happens to be at the corner, then it has one parameter β_j equals to zero. (Here we can find $\beta_1 = 0$)

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Generalize them and view them as Bayes estimates.

The criterion:

$$\tilde{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \quad \text{for } q \geq 0$$

非凸约束

log-prior density for β_j : $|\beta_j|^q$

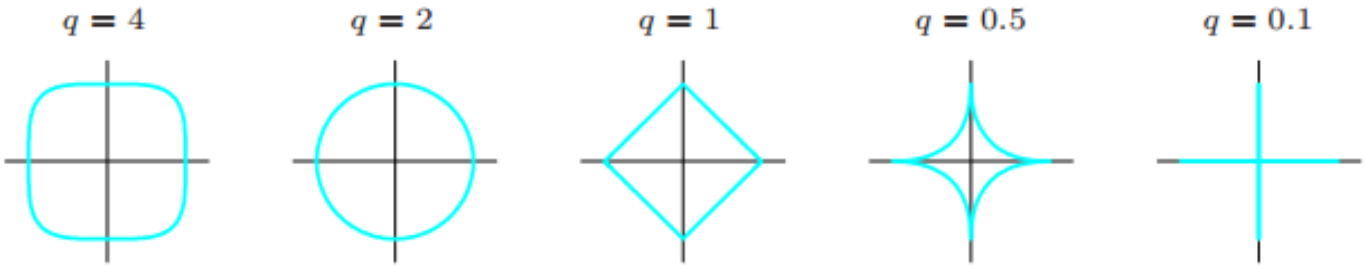


FIGURE 3.12. *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q.*

Notice that when $q \leq 1$, the prior is not uniform in direction, but concentrates more mass in the coordinate directions.

注意到 q ≤ 1，先验在各方向上不是均匀的，而是更多地集中在坐标方向上．对应 q = 1 情形的先验分布是关于每个输入变量是的独立的二重指数分布（或者 Laplace 分布），概率密度为 (1/2τ)exp(−|β|)/τ 并且 τ = 1/λ．q=1 的情形（lasso）是使得约束区域为凸的最小 q 值；非凸约束区域使得优化问题很困难．

Combination of Ridge & Lasso regression

Elastic Net $\lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ which is a different compromise between ridge & lasso.

↓ further in Section 18.4

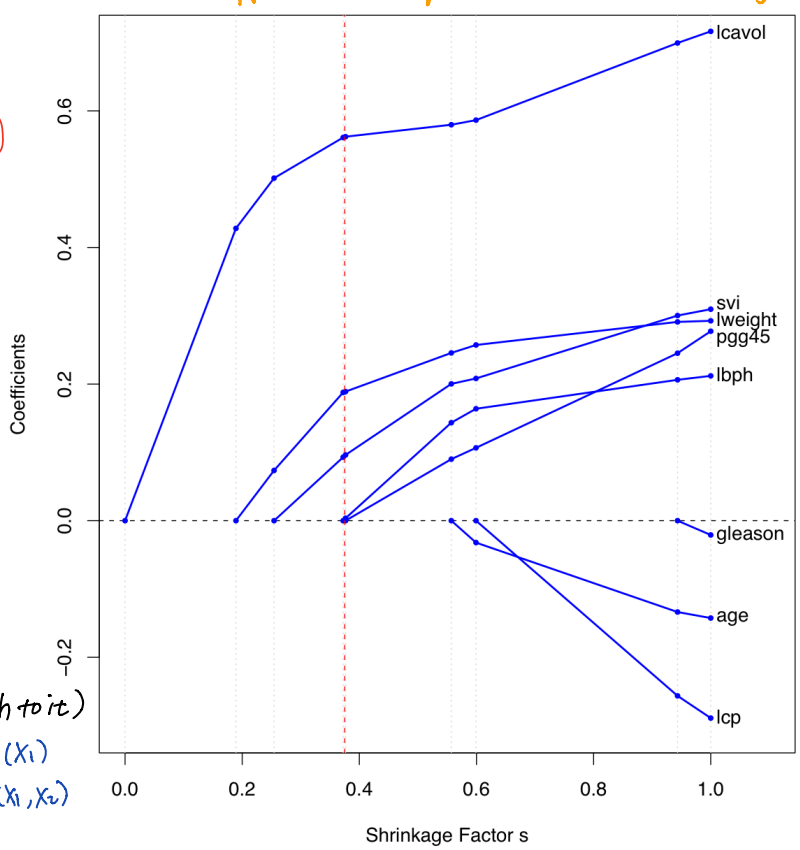
3.4.4 Least Angle Regression (LAR)

LAR can be viewed as a kind of "democratic" version of forward stepwise regression.

$r_k = y - X_{Ak}\beta_{Ak}$ is the current residual, then the direction for this step is

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

↓ just like $\hat{\beta} = (X^TX)^{-1}X^Ty$



Steps: 1st direction $\delta_1 = (X_{A_1}^T X_{A_1})^{-1} X_{A_1}^T y$

↳ (Just moving forward this direction, not reach to it)

2nd: get the direction for this step: $A_1 = (X_1)$

$\delta_2 = (X_{A_2}^T X_{A_2})^{-1} X_{A_2}^T (y - X_{A_1}\beta_{A_1})$ $A_2 = (X_1, X_2)$

coefficient profile evolves as

$\beta_{A_2}(\alpha) = \beta_{A_2} + \alpha \cdot \delta_2$

↳ this β_{A_2} just enters as 0 when X_2 is added

Ex. 3.23

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

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 $\mathbf{x}_j$.

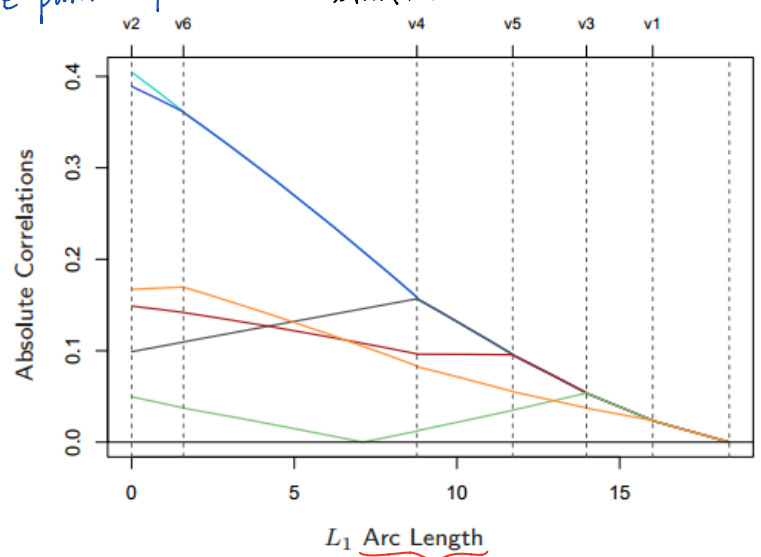
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

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.

Similar steps for fit vector $\hat{f}_k$, which would evolve as $\hat{f}_k(\alpha) = \hat{f}_k + \alpha \cdot u_k$, where $u_k = X_{Ak}\delta_k$

Ex 3.24

Whole path is piecewise-linear. 分段线性



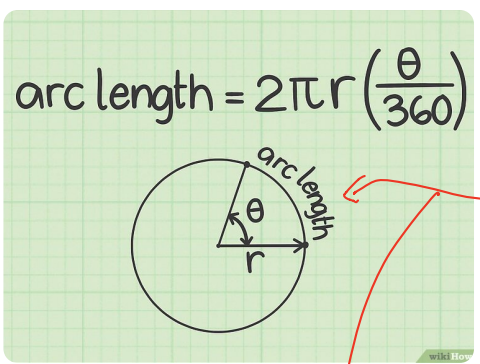


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

[2]The L_1 arc-length of a differentiable curve $\beta(s)$ for $s \in [0, S]$ is given by $\mathrm{TV}(\beta, S) = \int_0^S ||\dot{\beta}(s)||_1 ds$, where $\dot{\beta}(s) = \partial\beta(s)/\partial s$. For the piecewise-linear LAR coefficient profile, this amounts to summing the L_1 norms of the changes in coefficients from step to step.

when $p \gg N$. Osborne et al. (2000a) also discovered a piecewise-linear path for computing the lasso, which they called a *homotopy* algorithm.

Effective degrees of freedom of the fitted vector $\hat{y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N)$ as

Section 7.4-7.6 $df(\hat{y}) = \frac{1}{\sigma^2}\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$ (3.60)

predicted value $\hat{y}_i$ → corresponding outcome value y_i

3.5 Methods Using Derived Input Directions

3.5.1 Principal Components Regression

Components $Z_1, Z_2, \cdots Z_M$ are orthogonal — Standardized X

$\hat{y}^{\mathrm{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$ $\quad \hat{\beta}^{\mathrm{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$ Where $Z_m = X v_m$

$\hat{y} = X\hat{\beta} = Z\hat{\theta}^{\mathrm{pcr}} = XV\hat{\theta}$ if $\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_m \\ \vdots \\ \hat{\theta}_n \end{bmatrix}$

the mth principal component direction v_m solves:

$$\max_{\alpha} \mathrm{Var}(\mathbf{X}\alpha) \quad (3.63)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S} v_\ell = 0,\ \ell = 1, \ldots, m-1,$$

where $\mathbf{S}$ is the sample covariance matrix of the $\mathbf{x}_j$. The conditions $\alpha^T \mathbf{S} v_\ell = 0$ ensures that $\mathbf{z}_m = \mathbf{X}\alpha$ is uncorrelated with all the previous linear combinations $\mathbf{z}_\ell = \mathbf{X}v_\ell$.

3.5.2 Partial Least Squares — Standardized X

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 $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\mathrm{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

PLS's solution path is a nonlinear fcn of y.

The mth PLS direction $\hat{\varphi}_m$ solves:

$$\max_{\alpha} \mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha) \quad (3.64)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S} \hat{\varphi}_\ell = 0,\ \ell = 1, \ldots, m-1.$$

Ex. 3.14
Ex. 3.15
Ex. 3.18

3.6 Discussion: A Comparison of the Selection and Shrinkage Methods

From Frank and Friedman (1993):

Method	Properties
Best subset	
Ridge	shrinks all direction, but the lower-variance direction more
Lasso	middle between ridge & Best subset, enjoy some of the properties of each
PCA	leaves M high-variance directions only and discard the rest
PLS	shrinks low-variance directions, & also inflate some high-var directions

a little unstable, a little higher prediction error compared to ridge

3.7 Multiple Outcome Shrinkage and Selection

I. Canonical correlation analysis (CCA)

Ex. 3.20

Algorithm 3.4 *Incremental Forward Stagewise Regression—FS_ϵ.*

1. Start with the residual $\mathbf{r}$ equal to $\mathbf{y}$ and $\beta_1, \beta_2, \ldots, \beta_p = 0$. All the predictors are standardized to have mean zero and unit norm.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$
3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}[\langle \mathbf{x}_j, \mathbf{r} \rangle]$ and $\epsilon > 0$ is a small step size, and set $\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j$.
4. Repeat steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.

increase a little bit everytime, similar to LAR

(β的移动方向始终与最小二乘法方向一致)

$\text{sign}[\langle x_j, r \rangle]$

Ⅱ. FS_0 ☆.

Algorithm 3.2b *Least Angle Regression: FS_0 Modification.*

4. Find the new direction by solving the constrained least squares problem

$$\min_b ||\mathbf{r} - \mathbf{X}_{\mathcal{A}} b||_2^2 \text{ subject to } b_j s_j \geq 0,\ j \in \mathcal{A},$$

where s_j is the sign of $\langle \mathbf{x}_j, \mathbf{r} \rangle$.

Ⅲ. Group Lasso

The grouped-lasso minimizes the convex criterion

$$\min_{\beta \in \mathbb{R}^p} \left(||\mathbf{y} - \beta_0 \mathbf{1} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \beta_\ell||_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\beta_\ell||_2 \right), \tag{3.80}$$

where the $\sqrt{p_\ell}$ terms accounts for the varying group sizes, and $||\cdot||_2$ is the Euclidean norm (not squared). Since the Euclidean norm of a vector β_ℓ is zero only if all of its components are zero, this procedure encourages sparsity at both the group and individual levels. That is, for some values of λ, an entire group of predictors may drop out of the model.

Generalizations include more general L_2 norms $||\eta||_K = (\eta^T K \eta)^{1/2}$, as well as allowing overlapping groups of predictors (Zhao et al., 2008).

3.8.5 To be Continued.