## Lasso

1.1 節で述べたように

$$L := \frac{1}{2N}\|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{1.5}$$

を最小にする問題を Lasso という．まず簡単のために最初に

$$\frac{1}{N}\sum_{i=1}^{N} x_{i,j}x_{i,k} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases} \tag{1}$$

を仮定する．この仮定は $N \times p$ 行列の各列の 2 乗平均が 1(あるいは中心化されている $X$ に対し，各列が標準化) かつ異なる列が直交するということに他ならない．つまり，各 $i$ に対して $\boldsymbol{x}_i$ を縦ベクトルとして

$$X = (\boldsymbol{x}_i, \cdots, \boldsymbol{x}_p)$$

と書いたときに，$\|\boldsymbol{x}_i\|^2 = N$ かつ $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = 0 (i \neq j)$ が成立している状況を仮定している．
さらに $s_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}y_i$ とおく．これは $X$ の第 $j$ 列と $y$ との内積とみることができる．こうすることで計算が容易になる．

$L$ の定義より，各 $j$ に対して

$$L = \left(\beta_j^2\text{の係数が}\sum_{i=1}^{N} x_{i,j}^2\text{の 2 次式}\right) + \lambda|\beta_j| + (\beta_j^0\text{の項})$$

となり，凸関数の和で表されることから 4 ページ目の結果より $\beta_j$ に関する凸関数になる．したがって

$$L(\beta_1, \cdots, \beta_p)\ \text{が}\beta_j\text{において最小} \Rightarrow L\ \text{の}\beta_j\text{に関する劣微分が 0 を含む}$$

が成立するので，そのような $\beta_j$ を求める．
$L$ の $\beta_j$ に関する劣微分を求めると，微分可能な凸関数に対して劣微分が微分係数と一致す

ることから 1 ページ目の計算を用いて

$$
\begin{aligned}
(L\ \text{の}\beta_j\text{に関する劣微分}) &= -\frac{1}{N}\sum_{i=1}^{N} x_{i,j}\left(y_i - \sum_{k=1}^{p} x_{i,k}\beta_k\right) + \lambda \begin{cases} 1 & \beta_j > 0 \\ [-1,1] & \beta_j = 0 \\ -1 & \beta_j < 0 \end{cases} \\
&= -\frac{1}{N}\left(\sum_{i=1}^{N} x_{i,j}y_i + \sum_{k=1}^{p}\sum_{i=1}^{N} x_{i,j}x_{i,k}\beta_k\right) + \lambda \begin{cases} 1 & \beta_j > 0 \\ [-1,1] & \beta_j = 0 \\ -1 & \beta_j < 0 \end{cases} \\
&= \begin{cases} -s_j + \beta_j + \lambda & \beta_j > 0 \\ -s_j + \beta_j + \lambda[-1,1] & \beta_j = 0 \\ -s_j + \beta_j - \lambda & \beta_j < 0 \end{cases}
\end{aligned}
$$

となることがわかる．したがって最右辺を $\beta_j$ について解くと

$$
\beta_j = \begin{cases} s_j - \lambda & \beta_j > 0 \Leftrightarrow s_j - \lambda > 0 \\ 0 & \beta_j = 0 \Leftrightarrow -s_j \in [-\lambda, \lambda] \\ s_j + \lambda & \beta_j < 0 \Leftrightarrow s_j + \lambda < 0 \end{cases}
$$

が得られる．上式は $\mathbb{R}$ 上の関数

$$
\mathcal{S}_\lambda(x) := \begin{cases} x - \lambda & x > \lambda \\ 0 & |x| \leq \lambda \\ x + \lambda & x < -\lambda \end{cases} \tag{1.11}
$$

を用いることで $\beta_j = \mathcal{S}_\lambda(s_j)$ とかくことができる．つまり (1.9) 式の仮定の下では，(1.5) 式を最小にする $\hat{\beta} \in \mathbb{R}^p$ は定数 $s_1, \cdots, s_p$ を用いて

$$
\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\vdots} \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \mathcal{S}_\lambda(s_1) \\ \vdots \\ \mathcal{S}_\lambda(s_p) \end{pmatrix}
$$

とかくことができるのである．

ここで，$\mathbb{R}$ 上の関数 $\mathcal{S}'_\lambda(x)$ を

$$
\mathcal{S}'_\lambda(x) := \mathrm{sgn}(x) \cdot \max\{|x| - \lambda, 0\} \quad (\text{ただし sgn は符号関数})
$$

によって定義すると，この関数は (1.11) 式で定義された関数 $\mathcal{S}_\lambda(x)$ と一致する．著書のプログラムはこの事実を用いて書かれている．

次に，仮定が無い場合にはどうなるか考える．まず $X, y$ は中心化されているとする．このとき，L の $\beta_j$ における劣微分を考えると

$$
\begin{aligned}
(L\ \text{の}\beta_j\text{に関する劣微分}) &= -\frac{1}{N}\sum_{i=1}^{N} x_{i,j}\left(y_i - \sum_{k=1}^{p} x_{i,k}\beta_k\right) + \lambda \begin{cases} 1 & \beta_j > 0 \\ [-1,1] & \beta_j = 0 \\ -1 & \beta_j < 0 \end{cases} \\
&= -\frac{1}{N}\sum_{i=1}^{N} x_{i,j}\left(y_i - \left(x_{i,j}\beta_j + \sum_{k\neq j} x_{i,k}\beta_k\right)\right) + \lambda \begin{cases} 1 & \beta_j > 0 \\ [-1,1] & \beta_j = 0 \\ -1 & \beta_j < 0 \end{cases} \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_{i,j})^2\beta_j - \frac{1}{N}\sum_{i=1}^{N} x_{i,j}\left(y_i - \sum_{k\neq j} x_{i,k}\beta_k\right) + \lambda \begin{cases} 1 & \beta_j > 0 \\ [-1,1] & \beta_j = 0 \\ -1 & \beta_j < 0 \end{cases}
\end{aligned}
\tag{2}
$$

となることがわかる．ここで，最右辺第 2 項は教科書の通り $r_{i,j} = y_i - \sum_{k\neq k} x_{i,k}\beta_k$ とすることで $s_j := \frac{1}{N}\sum_{i=1}^{N} x_{i,j}r_{i,j}$ と表される．さらに第 1 項について，$\frac{1}{N}\sum_{i=1}^{N}(x_{i,j})^2$ が 1 であれば $\beta_j$ について解くことは安易なのでこれを仮定する．つまり，今 $X$ は

$$
X = \begin{pmatrix} \frac{x_{1,1}-\bar{x}_1}{\sigma_1} & \cdots & \frac{x_{1,p}-\bar{x}_p}{\sigma_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{N,1}-\bar{x}_1}{\sigma^1} & \cdots & \frac{x_{N,p}-\bar{x}_p}{\sigma^p} \end{pmatrix}, \quad \sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{i,j}-\bar{x}_j)^2}
$$

と書かれていると仮定する．このことは $X$ の各列が標準化されていることに他ならない．

上の仮定のもとでは，(2) 式は $\beta_j$ について解くことができて

$$
\beta_j = \begin{cases} s_j - \lambda & (s_j > \lambda) \\ 0 & (|s_j| \leq \lambda) \\ s_j + \lambda & (s_j < -\lambda) \end{cases} = \mathcal{S}_\lambda(s_j) \tag{3}
$$

となることがわかる．

ここで注意すべきは，上式の右辺は $\beta_k(k \neq j)$ を用いて表していることである．すなわち，$\beta_k(k \neq j)$ を固定した下で $L$ を最小にする $\beta_j$ は求めることができるが，$L$ を最小にす

る $\beta \in \mathbb{R}^p$ を一度に求めるのは上の手法では不可能である．そこで，以下に述べる座標降下法という手法を用いて $\beta$ を求めることにする．

**座標降下法**

以下では，$L$ を $\beta_j$ に関する 1 変数関数 $L(\beta_j)$ とみている時には $L_j$ とかくことにする．

(i) $\beta \in \mathbb{R}^p$ を任意にとる．(解 $\hat{\beta}$ の近くであれば収束は速いが，一般には定まらないので教科書では初期値 $\mathbf{0}$ をとっている．) このときとった $\beta$ の各成分を $\beta_{1,1}, \cdots, \beta_{p,1}$ と書くことにする (つまり $\beta = (\beta_{1,1}, \cdots, \beta_{p,1})$).

(ii) $L = L(\beta_1, \cdots, \beta_p)$ に $\beta_2 = \beta_{2,1}, \beta_3 = \beta_{3,1}, \cdots, \beta_p = \beta_{p,1}$ を代入することで，$L$ を $\beta_1$ に関する 1 変数関数 $L(\beta_1)$ とする．

(iii) (ii) で生成した $L_1$ を最小にする $\beta_1$ を (3) 式を解いて求める．そこで得られた解を $\beta_1'$ とおく．このとき，

$$L(\beta_1', \beta_{2,1}, \cdots, \beta_{p,1}) \leq L(\beta) \tag{4}$$

となる．

(iv) 次に $L$ に $\beta_1 = \beta_1', \beta_3 = \beta_{3,1}, \cdots, \beta_p = \beta_{p,1}$ を代入して 1 変数 $L_2$ を生成する．

(v) (iv) で生成した $L_2$ を最小にする $\beta_2$ を (3) 式を解いて求める．そこで得られた解を $\beta_2'$ とおく．このとき

$$L(\beta_1', \beta_2', \cdots, \beta_{p,1}) \leq L(\beta_1', \beta_{2,1}, \cdots, \beta_{p,1}) \tag{5}$$

となる．

(vi) 上で行った操作を $p$ までについて行う．すなわち，$\beta_{j-1}'$ $(j \leq p)$ まで得られているとき，$L$ に $\beta_1 = \beta_1', \cdots, \beta_{j-1} = \beta_{j-1}', \beta_{j+1} = \beta_{j+1,1}, \cdots, \beta_p = \beta_{p,1}$ を代入して $\beta_j$ に関する 1 変数関数 $L_j$ を生成し，$L_j$ を最小にする $\beta_j$ を (3) 式を解いて求め，得られた解を $\beta_j'$ として $\beta_{j+1}'$ を求める．このとき

$$L(\beta_1', \cdots, \beta_{j-1}', \beta_j', \beta_{j+1,1}, \cdots, \beta_{p,1}) \leq L(\beta_1', \cdots, \beta_{j-1}', \beta_{j,1}, \beta_{j+1,1}, \cdots, \beta_{p,1}) \tag{6}$$

となることに注意する．

(vii) (vi) により新たに $\beta' := (\beta_1', \beta_2', \cdots, \beta_p')$ が得られる．このとき，(4),(5),(6) 式より

$$L(\beta) \geq L(\beta')$$

が成立することがわかる．したがって $\beta'$ を新たに初期値として (ii) からの操作を施すというループを無限回行うことで

$$L(\beta) \geq L(\beta') \geq L(\beta'') \geq \cdots \geq L(\beta^{(n)}) \geq \cdots$$

を満たす $\mathbb{R}$ 上の数列 $\{L(\beta^{(n)}\}_{n=1}^{\infty}$ が得られる．任意の自然数 $n$ について

$$L(\beta^{(n)}) \geq \inf_{\beta \in \mathbb{R}^p} L(\beta) \geq 0$$

が成立することから数列 $\{L(\beta^{(n)})\}_{n=1}^{\infty}$ は下に有界な単調減少列であるので極限 $\lim_{n\to\infty} L(\beta^{(n)})$ が存在し，それは $\inf_{\beta \in \mathbb{R}^p} L(\beta)$ に一致する．この極限を $\hat{\beta}$ とすれば，$\forall \beta \in \mathbb{R}$ に対して

$$L(\beta) \geq L(\hat{\beta})$$

となるので，求める $\hat{\beta}$ を生成することができる． □

上記の手順で $L$ を最小にする $\hat{\beta}$ を生成できるが，それには無限試行という手順が必要なので現実には有限回の操作で打ち切って近似値を求めることになる．

$\mathbb{R}$ 上の収束列 $\{L(\beta^{(n)})\}_{n=1}^{\infty}$ は特に $\mathbb{R}$ 上の Cauchy 列であることから，任意の $\varepsilon > 0$ に対して十分大きな自然数 $m, n$ が存在して

$$|L(\beta^{(m)}) - L(\beta^{(n)})| < \varepsilon$$

が成立する．したがって十分大きな $m$ のもとで

$$\begin{aligned} |L(\beta^{(m)}) - L(\hat{\beta})| &= \lim_{n\to\infty} |L(\beta^{(m)}) - L(\beta^{(n)})| \\ &\leq |L(\beta^{(m)}) - L(\beta^{(m+1)})| + \lim_{n\to\infty} |L(\beta^{(m+1)} - L(\beta^{(n)})| \\ &< |L(\beta^{(m)}) - L(\beta^{(m+1)})| + \varepsilon \end{aligned} \tag{7}$$

となることから $|L(\beta^{(m)}) - L(\beta^{(m+1)})|$ の差が小さくなるまでループを繰り返すことで任意の精度の近似値が得られる．あるいは $L$ が ($\mathbb{R}^p$ と $\mathbb{R}$ に通常の位相が入った上で) 連続関数となることから，任意の $\varepsilon > 0$ に対して

$$\|\beta^{(m)} - \beta^{(m+1)}\| < \delta \Rightarrow |L(\beta^{(m)}) - L(\beta^{(m+1)})| < \varepsilon$$

を満たすような $\delta > 0$ が存在する．これと (7) 式を考えれば，$\|\beta^{(m)} - \beta^{(m+1)}\|$ の差が小さくなるまでループを繰り返すことで精度の良い近似値を得られることができる．著書では後者を採用している．

以上より各列が標準化された $X$ と，中心化された $y$ について $L$ を最小にする $\hat{\beta}$ を求めることができた．これを用いることで一般の $X$ について $L$ を最小にする $\hat{\beta}, \hat{\beta}_0$ を求める；

$X$ 中心化したものを $\overline{X}$，標準化したものを $X'$ とかく．さらに $p$ 次正方行列 $C$ を

$$C = \begin{pmatrix} \frac{1}{\sigma_1} & \cdots & \frac{1}{\sigma_p} \\ \vdots & \cdots & \vdots \\ \frac{1}{\sigma_1} & \cdots & \frac{1}{\sigma_p} \end{pmatrix}$$

で定める．このとき $X' = \overline{X}C$ であることに注意すれば

$$\frac{1}{2N}\|y - X'\beta\|^2 + \lambda\|\beta\|_1 \text{が} \hat{\beta} \text{において最小} \Leftrightarrow \frac{1}{2N}\|y - \overline{X}\beta\|^2 + \lambda\|\beta\|_1 \text{が} \beta = C\hat{\beta} \text{において最小}$$

となることがわかるので，中心化した $X$,$y$ について $L$ を最小にする $\beta$ が

$$\hat{\beta}' = C\hat{\beta} = \begin{pmatrix} \hat{\beta}_1/\sigma_1 \\ \vdots \\ \hat{\beta}_p/\sigma_p \end{pmatrix}$$

とかけることが示せた．あとは (1.4) 式を用いることで，一般の $X, y$ について $L = \frac{1}{2N}\|y - X\beta - \beta_0\|^2 + \lambda\|\beta|_1$ を最小にする $\hat{\beta}, \hat{\beta}_0$ は

$$\hat{\beta} = \hat{\beta}', \quad \hat{\beta}_0 = \bar{y} - \sum_{i=1}^{p} \bar{x}_j \hat{\beta}'_j$$

となることがわかる． □