1.5

Lasso と Ridge を比較して，λ が大きくなるにつれて各係数の絶対値が減少し，0 に近づく点では同じである．しかし，Lasso では λ の値がある一定以上になると各係数の値がちょうど 0 になり，その 0 になるタイミングが係数ごとに異なる．そのため Lasso は変数選択に用いることができる．

数学的な解析も行ってきたが，直感的な意味を幾何学的に把握してみよう．

$p = 2$ とし，$X \in \mathbb{R}^{N\times p}$ が $x_{i,1}, x_{i,2}(i = 1, \cdots, N)$ の 2 列からなっているとする．もちろん中心化を仮定する．最小 2 乗法では，

$$S := \sum_{i=1}^{N}(y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2$$

を最小にする β_1, β_2 を求めていた．それらを $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ とおく．ここで，各 i について $\hat{y}_i = x_{i,1}\hat{\beta}_1 + x_{i,2}\hat{\beta}_2$ とおくと，

$$\frac{\partial S}{\partial \beta_j}(\hat{\beta}) = -2\sum_{i=1}^{N} x_{i,j}(y_i - \hat{y}_i) = 0$$

が $j = 1, 2$ に対して成立していたことから

$$\sum_{i=1}^{N} x_{i,1}(y_i - \hat{y}_i) = \sum_{i=1}^{N} x_{i,2}(y_i - \hat{y}_i) = 0 \tag{1}$$

が成立することがわかる．また，任意の実数 β_1, β_2 に対して

$$y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2} = y_i - \hat{y}_i - (\beta_1 - \hat{\beta}_1)x_{i,1} - (\beta_2 - \hat{\beta}_2)x_{i,2} \tag{2}$$

が成立していることがわかるので，(1),(2) 式より最小にすべき $S = \sum_{i=1}^{N}(y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2$ が

$$\begin{aligned}
S &= \sum_{i=1}^{N}(y_i - \hat{y}_i - (\beta_1 - \hat{\beta}_1)x_{i,1} - (\beta_2 - \hat{\beta}_2)x_{i,2})^2 \\
&= \sum_{i=1}^{N}\Big\{(y_i - \hat{y}_i)^2 + (\beta_1 - \hat{\beta}_1)^2 x_{i,1}^2 + (\beta_2 - \hat{\beta}_2)^2 x_{i,2}^2 \\
&\qquad\qquad - 2(y_i - \hat{y}_i)(\beta_1 - \hat{\beta}_1)x_{i,1} + 2(\beta_1 - \hat{\beta}_1)x_{i,1}(\beta_2 - \hat{\beta}_2)x_{i,2} - 2(\beta_2 - \hat{\beta}_2)x_{i,2}y_i - \hat{y}_i\Big\} \\
&= (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{N} x_{i,1}^2 + 2(\beta_1 - \hat{\beta}_1)(\beta_2 - \hat{\beta}_2)\sum_{i=1}^{N} x_{i,1}x_{i,2} + (\beta_2 - \hat{\beta}_2)^2\sum_{i=1}^{N} x_{i,2}^2 + \sum_{i=1}^{N}(y_i - \hat{y}_i)^2
\end{aligned} \tag{3}$$

とかくことができる．上式において $(\beta_1, \beta_2) = (\hat{\beta}_1, \hat{\beta}_2)$ とすればもちろん最小値 $(= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2)$ が得られる．また，(3) 式で得られた式において，座標の変換と平行移動を用いることで S は楕円の式の標準系 (の左辺) になっていることがわかる．

ここで，Lasso，Ridge のそれぞれがある実数 $C, C' > 0$ を用いた制約条件 $\beta_1^2 + \beta_2^2 \leq C, |\beta_1| + |\beta_2| \leq C'$ の下での $S = S(\beta)$ の最小化問題となっている．このことを以下で確認する；

Ridge の場合について考える．$C \geq 0$ を任意にとって固定し，$g(\beta_1, \beta_2) = \beta_1^2 + \beta_2^2 - C$ とおく．このとき $g(\beta) \leq 0$ の下での $S(\beta)$ の最小化問題の解 β^* について考察にする．

上のステートメントを具体的にかくと

$$\min\ S(\beta_1, \beta_2) = \sum_{i=1}^{N}(y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2\ s.t.\ (\beta_1, \beta_2) \in A := \{(\beta_1, \beta_2) \mid g(x_1, x_2) \leq 0\}$$

となる．ここで，

$$A' := \{(\beta_1, \beta_2) \mid g(x_1, x_2) = 0\}$$
$$A'' := \{(\beta_1, \beta_2) \mid g(x_1, x_2) < 0\}$$

とおき，$\beta^* \in A', A''$ のそれぞれで場合分けして考える．

$\beta^* \in A''$ のとき

$g(\beta^*) < 0$ であったとする．このとき，g の連続性より β^* を中心とするある開距離球体 $B(\beta^*, \varepsilon)$ が存在して

$$\beta \in B(\beta^*, \varepsilon) \Rightarrow g(\beta) < 0 \tag{4}$$

が成立することがわかる．さらに β^* は制約付き最小化問題の解であることから

$$\beta \in B(\beta^*, \varepsilon) \Rightarrow S(\beta) \geq S(\beta^*) \tag{5}$$

が成立することがわかる．したがってこのとき (4),(5) 式より，β^* は S の極値であることがわかり，S の凸性よりこの場合は β^* が制約なしの最小化問題の解であることがわかる．

$\beta^* \in A'$ のとき

$g(\beta^*) = 0$ であったとする．このとき β^* は特に $g(\beta) = 0$ の下での条件付き極値問題の解であったことより，

(1)　β^* は g の特異点である．

(2)　ある実数 λ が存在し，$L_\lambda(\beta) := S(\beta) + \lambda g(\beta)$ としたときに

$$\frac{\partial}{\partial \beta_1} L_\lambda(\beta^*) = \frac{\partial}{\partial \beta_2} L_\lambda(\beta^*) = 0$$

のいずれかが成立することがわかる．曲線 $g(\beta) = 0$ は特異点を持たないことよりこの場合は (2) が成立していることがわかる．

また λ の正負について考えると，$\delta \in \mathbb{R}^2$ を

$$\nabla g(\beta^*)^T \delta = \frac{\partial g}{\partial \beta_1}(\beta^*)\delta_1 + \frac{\partial g}{\partial \beta_2}(\beta^*)\delta_2 < 0$$

となるようにとると，平均値の定理より十分小さい $\varepsilon > 0$ に対して

$$g(\beta^* + \varepsilon\delta) = g(\beta^*) + \varepsilon \nabla g(\beta^*)^T \delta < 0$$

が成立するので $\beta^* \in A$ がわかる．これと β^* が制約付き最小化問題の解であること，およびテーラーの定理よりある $c \in (0, 1)$ が存在して

$$\begin{aligned} S(\beta^*) &\leq S(\beta^* + \varepsilon\delta) \\ &= S(\beta^*) + \nabla S(\beta^* + c\varepsilon\delta)^T (\varepsilon\delta) \\ \therefore 0 &\leq \varepsilon \nabla S(\beta^* + c\varepsilon\delta)^T \delta \end{aligned}$$

が成立することがわかる．したがって $\varepsilon \to +0$ とすることで $0 \leq \nabla S(\beta^*)^T \delta$ がわかるので，これと条件 (2) より

$$\begin{aligned} \langle \nabla S(\beta^*) + \lambda \nabla g(\beta^*), \delta \rangle = \nabla S(\beta^*)^T \delta + \lambda \nabla g(\beta^*) \delta = 0 \\ \therefore \nabla S(\beta^*)^T \delta = -\lambda \nabla g(\beta^*) \delta \\ \therefore \lambda \geq 0 \end{aligned}$$

がわかる．以上より，制約 $g(\beta) \leq 0$ の下での $S(\beta)$ の最小化問題の解 β^* は

$$\begin{aligned} \frac{\partial (S + \lambda g)}{\partial \beta}(\beta^*) = 0 \\ \lambda \geq 0 \end{aligned}$$

を満たすことがわかり，$g(\beta) = \|\beta\|_2^2 - C$ であったことから

$$L := \sum_{i=1}^{N} (y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2 + \lambda \|\beta\|_2^2$$

としたときに

$$\frac{\partial L}{\partial \beta} = \frac{\partial (S + \lambda g)}{\partial \beta}(\beta^*) = 0$$

となることがわかる．よって β^* は L の制約条件なしの最小解となることがわかる．

また，ラグランジュの未定乗数法における乗数 λ は

$$\lambda = -\frac{\partial S}{\partial \beta_2}(\beta^*) / \frac{\partial g}{\partial \beta_2}(\beta^*)$$

とおけることから，C が小さくなるにつれて λ は大きくなる． □

図 1.8 の緑の範囲に最小二乗法の解 $(\hat{\beta}_1, \hat{\beta}_2)$ がくれば $\beta_1 = 0$ または $\beta_2 = 0$ が Lasso の解となる．特に楕円が円である場合の証明を考えよう．

Proof. S を変形した (3) 式が $\sum_{i=1}^{N} x_{i,1}^2 = \sum_{i=1}^{N} x_{i,2}^2 = 1, \sum_{i=1}^{N} x_{i,1} x_{i,2} = 0$ を満たしているとする．このとき

$$S(\beta_1, \beta_2) = (\beta_1 - \hat{\beta}_1)^2 + (\beta_2 - \hat{\beta}_2)^2 + \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

とかくことができる．T を大きくしていきながら $S(\beta_1, \beta_2) = T$ を満たす (β_1, β_2) の領域を考え，$D := \{(\beta_1, \beta_2) \mid |\beta_1| + |\beta_2| < C\}$ と $S(\beta_1, \beta_2) = T$ を満たす領域とが初めて空でなくなったとき，接点 (β_1', β_2') は

$$\|(\beta_1', \beta_2') - (\hat{\beta}_1, \hat{\beta}_2)\|_2 = \inf_{\beta \in D} \|\beta - (\hat{\beta}_1, \hat{\beta}_2)\|_2 = \|D - (\hat{\beta}_1, \hat{\beta}_2)\|_2$$

を満たすことがわかる．そのような点 (β_1', β_2') は，白の領域上では制約条件の直線上になり，緑の領域上では制約条件の特異点となることがわかり，したがって $\beta_1 = 0$ または $\beta_2 = 0$ が Lasso の解となる．

□

そして，見逃すことができない Ridge のメリットとして共線性がある．つまり，説明変数に類似する列が存在した場合に首尾よく動作することである．このことについて考察しよう．

与えられたデータ $X \in \mathbb{R}^{N \times p}$ と，そこから定まる回帰方程式による推定値 $\hat{y} \in \mathbb{R}^N$，およ

び実際に観測される値 $y \in \mathbb{R}^N$ に対し，決定係数 R の 2 乗 R^2 は

$$R^2 := 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

で定義され，この値が大きいほど実現値と推定値の相対的な差が小さいことを表す．また，j 番目のデータに対して VIF_j を，$X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \cdots, \boldsymbol{x}_p), y = \boldsymbol{x}_j$ として

$$VIF_j = \frac{1}{1 - R^2} = \frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

で定義する．この値が大きいほど j 列目の成分は他の列の成分で表されることを意味する．通常の線形回帰では，VIF の値が大きくなるような j が存在すると，推定された $\hat{\beta}$ の値が不安定になり，特に 2 列が完全に一致するときは $\mathrm{rank}(X^TX)$ が正則でなくなるので推定値が求まらない．また Lasso の場合，2 列が類似しているときは一方の係数が 0，他方の係数が非ゼロとして推定されることが多い．しかし Ridge の場合，$\lambda > 0$ であれば X の列 j, k が一致するときでも推定値が求まり，両者が一致するという性質がある．証明は教科書を読めばすぐに理解できる．