1. Consider a linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad i=1,\cdots,n$$

Y and X are centered and standardized, ridge regression is

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

No assumption of X is of full rank. $\beta := [\beta_1, \cdots, \beta_p]^T$

(a) prove that β^{ridge} is biased estimator for β for given λ

Denote $P = \|Y - X\beta\|^2 + \lambda\|\beta\|^2$, which is convex, then

$$\frac{\partial P}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta$$

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY \qquad (X^TX+\lambda I) \text{ is always invertible}$$

then $\mathbb{E}\hat{\beta}^{ridge} = \mathbb{E}(X^TX+\lambda I)^{-1}X^TY$

$$= (X^TX+\lambda I)^{-1}X^TX\beta \neq \beta$$

Thus $\hat{\beta}^{ridge}$ is a biased estimator of β.

(b) find the bias and the variance of $\hat{\beta}^{ridge}$ for given tuning parameter λ;

$$\text{Bias}(\hat{\beta}^{ridge}) = \mathbb{E}\hat{\beta}^{ridge} - \beta = [(X^TX+\lambda I)^{-1}X^TX - I]\beta$$

$$\text{Var}(\hat{\beta}^{ridge}) = \text{Var}[(X^TX+\lambda I)^{-1}X^TY]$$

$$= (X^TX+\lambda I)^{-1}X^T[\text{Var}\, Y]X(X^TX+\lambda I)^{-1}$$

$$= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}$$

(c) Show that $\|\hat{\beta}^{ridge}\|$ increases as the tuning parameter $\lambda \to 0$.

Denote the SVD of X as $X = U\Lambda V^T$ where $\Lambda = \text{diag}(d_1, \cdots, d_p)$, $U \in \mathbb{R}^{n\times p}$ $V \in \mathbb{R}^{p\times p}$

$U := [u_1\ u_2 \cdots u_p]$ $\quad u_i \in \mathbb{R}^n$ $\quad u_i^Tu_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$

then $X^TX + \lambda I = V\Lambda^2 V^T + V\, diag(\lambda, \cdots, \lambda) V^T$

$$= V\, diag(d_1^2+\lambda, \cdots, d_p^2+\lambda) V^T$$

then $(X^TX + \lambda I)^{-1} = V\, diag\left(\frac{1}{d_1^2+\lambda}, \cdots, \frac{1}{d_p^2+\lambda}\right) V^T$

$$\hat{\beta}^{ridge} = (X^TX+\lambda I)^{-1} X^T Y = V diag\left(\frac{1}{d_1^2+\lambda}, \cdots, \frac{1}{d_p^2+\lambda}\right) V^T V \Lambda U^T Y$$

$$= V\, diag\left(\frac{d_1}{d_1^2+\lambda}, \cdots \frac{d_p}{d_p^2+\lambda}\right) U^T Y$$

$$\|\hat{\beta}^{ridge}\|^2 = [\hat{\beta}^{ridge}]^T [\hat{\beta}^{ridge}] = Y^T U\, diag\left(\frac{d_1}{d_1^2+\lambda}, \cdots, \frac{d_p}{d_p^2+\lambda}\right) V^T V\, diag\left(\frac{d_1}{d_1^2+\lambda}, \cdots, \frac{d_p}{d_p^2+\lambda}\right) U^T Y$$

$$= Y^T U\, diag\left(\frac{d_1^2}{(d_1^2+\lambda)^2}, \cdots \frac{d_p^2}{(d_p^2+\lambda)^2}\right) U^T Y.$$

$$= tr\left(Y^T U\, diag\left(\frac{d_1^2}{(d_1^2+\lambda)^2}, \cdots, \frac{d_p^2}{(d_p^2+\lambda)^2}\right) U^T Y\right)$$

$$= tr\left(diag\left(\frac{d_1^2}{(d_1^2+\lambda)^2}, \cdots, \frac{d_p^2}{(d_p^2+\lambda)^2}\right) \underbrace{U^T Y Y^T U}_{=: A (\perp\!\!\!\perp \lambda)}\right)$$

$$= \sum_{i=1}^{p} \frac{d_i^2}{(d_i^2+\lambda)^2} A_{ii}$$

which is a monotonely decreasing function wrt. λ.

Thus $\|\hat{\beta}^{ridge}\|$ increases as $\lambda \to 0$.

2. Consider the elastic-net optimization problem.

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda\left[\alpha \|\beta\|_2^2 + (1-\alpha)\|\beta\|_1\right]$$

(a) Show how the elastic-net optimization problem can turn this into a lasso problem.

Notice the $\|\beta\|_2^2$ can be rewritten as

$$\lambda\alpha \|\beta\|_2^2 = \|\sqrt{\lambda\alpha}\,\beta\|_2^2 = \|0 - \sqrt{\lambda\alpha}\, I_p \beta\|_2^2$$

We can put it into $\|y - X\beta\|_2^2$ by denoting

$$\tilde{y} = [y^T \; 0 1_p^T]^T, \quad \tilde{X} = [X^T \; \sqrt{\lambda\alpha} I_p]^T$$

then the elastic-net optimization problem can be rewritten as

$\min_{\beta} \|\tilde{y} - \tilde{X}\beta\|_2^2 + \lambda(1-\alpha)\|\beta\|_1$ with the same form of LASSO problem.

(b) Provide your own understanding about the effect of the elastic-net penalty on the param. estimate

From the above alternative problem of elastic-net, we know elastic-net penalty will shrink some of the parameters estimates to 0 just like LASSO penalty by setting $\alpha \neq 1$, and then the other parameters will still have a overall shrinkage encouraged by the augmented data with the same form of ridge penalty.

The turning parameter α is used to adjust the sparsity of parameter estimates.

3. Show the smallest λ such that the regression coefficients estimated by the LASSO are all equal to zero is given by $\lambda_{max} = \max_j |\frac{1}{N}\langle x_j, y\rangle|$

$\beta_\lambda = \min_\beta L(\beta) = \min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$ is the objective of LASSO problem.

then $\frac{\partial L}{\partial \beta}\Big|_{\beta=\beta_\lambda} = 0 \Rightarrow -X^T(y - X\beta_\lambda) + \lambda S(\beta_\lambda) = 0$

where $S(\beta_j) = \begin{cases} \mathrm{sgn}(\beta_j) & \text{if } \beta_j \neq 0 \\ S_j & \text{if } \beta_j = 0 \end{cases} \quad \exists\, S_j \in [-1, 1]$

Let λ_m is a λ such that $\beta_{\lambda_m} = 0 \in \mathbb{R}^p$

then we have $X^T y - \lambda_m S = 0$, $\exists\, S \in [-1,1]^p$

Then $\lambda_{max} = \min_{\lambda_m > 0} \lambda_m \quad$ s.t. $\begin{cases} \langle x_j, y\rangle - \lambda_m S_j = 0, & j = 1, \dots, p. \\ -1 \leq S_j \leq 1 \end{cases}$

Since for j, $\lambda_m = \frac{1}{S_j}\langle x_j, y\rangle \geq |\langle x_j, y\rangle|$

then $\lambda_{max} = \min_{\lambda_m > 0} \{\lambda_m : \lambda_m \geq |\langle x_j, y\rangle|, j = 1, \dots, p\}$

$= \min_{\lambda_m > 0} \{\lambda_m : \lambda_m \geq \max_j |\langle x_j, y\rangle|\}$

$= \max_j |\langle x_j, y\rangle|$

4. $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \gamma_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, $\gamma = (\gamma_1, \cdots, \gamma_N)$ are unknown constants

Consider minimization of $\min\limits_{\substack{\beta\in\mathbb{R}^p \\ \gamma\in\mathbb{R}^N}} \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \gamma_i\right)^2 + \lambda\sum_{i=1}^{N}|\gamma_i|$ ...... (2)

(a) Show this problem is jointly convex in β and γ

Denote $L(\beta,\gamma) := \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \gamma_i\right)^2 + \lambda\sum_{i=1}^{N}|\gamma_i|$

Denote $X := \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \in \mathbb{R}^{N\times p}$, $\beta = [\beta_1\ \beta_2 \cdots \beta_p]^T \in \mathbb{R}^{p\times 1}$

$y = [y_1\ y_2 \cdots y_N]^T \in \mathbb{R}^{N\times 1}$, and we know $\gamma = [\gamma_1\ \gamma_2 \cdots \gamma_N]^{N\times 1}$

then $L(\beta,\gamma) = \frac{1}{2}(y - X\beta - \gamma)^T(y - X\beta - \gamma) + \lambda\|\gamma\|_1$

$$\frac{\partial L}{\partial \beta} = -X^T(y - X\beta - \gamma) \quad , \quad \frac{\partial L}{\partial \gamma} = -(y - X\beta - \gamma) + \lambda S(\gamma)$$

where $S(\gamma_i) = \begin{cases} \operatorname{sgn}(\gamma_i) & \gamma_i \neq 0 \\ S & \gamma_i = 0 \end{cases}$ $\quad \exists S \in [-1,1]$

$$\frac{\partial^2 L}{\partial \beta^2} = X^TX \quad \frac{\partial^2 L}{\partial\beta\partial\gamma} = X^T \quad \frac{\partial^2 L}{\partial\gamma\partial\beta} = X \quad \frac{\partial^2 L}{\partial\gamma^2} = I$$

so the Hessian matrix is

$$H = \begin{bmatrix} \partial^2 L/\partial\beta^2 & \partial^2 L/\partial\beta\partial\gamma \\ \partial^2 L/\partial\gamma\partial\beta & \partial^2 L/\partial\gamma^2 \end{bmatrix} = \begin{bmatrix} X^TX & X \\ X^T & I \end{bmatrix} \succcurlyeq 0$$

thus $L(\beta,\gamma)$ is jointly convex in β and γ.

(2) Huber's Loss function $\rho(t;\lambda) = \begin{cases} \lambda|t| - \lambda^2/2 & \text{if } |t| > \lambda \\ t^2/2 & \text{if } |t| \leq \lambda \end{cases}$

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{N} \rho\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j ; \lambda\right) \quad \cdots \quad (4)$$

Show that problems (2) & (4) have the same solutions $\hat{\beta}$.

(Continue on the next page).

Denote $E(\beta) = \sum_{i=1}^{N} \rho(y_i - \sum_{j=1}^{p} x_{ij}\beta_j ; \lambda)$

Since $$\frac{\partial \rho(y_i - \sum_{j=1}^{p} x_{ij}\beta_j ; \lambda)}{\partial \beta_k} = \begin{cases} -\lambda x_{ik} \operatorname{sgn}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j) & |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| > \lambda \\ -x_{ik}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j) & |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| \leq \lambda \end{cases}$$

$$= -x_{ik}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - e_i)$$

where $$e_i = \begin{cases} y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \lambda \operatorname{sgn}(y - \sum_{j=1}^{p} x_{ij}\beta_j) & \text{if } |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| > \lambda \\ 0 & \text{if } |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| \leq \lambda \end{cases}$$

then $$\frac{\partial E}{\partial \beta_k} = \sum_{i=1}^{N} \frac{\partial \rho(y_i - \sum_{j=1}^{p} x_{ij}\beta_j ; \lambda)}{\partial \beta_k} = -\sum_{i=1}^{N} x_{ik}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - e_i)$$

By (2), $$\frac{\partial L}{\partial \beta_k} = -\sum_{i=1}^{N} x_{ik}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - r_i)$$
$$\frac{\partial L}{\partial r_i} = -(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - r_i) + \lambda S(r_i)$$

$$\Rightarrow (y_i - \sum_{j=1}^{p} x_{ij}\beta_j - r_i) = \lambda S(r_i) \quad S(r_i) = \begin{cases} \operatorname{sgn}(r_i) & r_i \neq 0 \\ s & r_i = 0 \end{cases} \quad \exists s \in [-1,1].$$

if $r_i = 0 \Leftrightarrow -\lambda \leq y_i - \sum_{j=1}^{p} x_{ij}\beta_j \leq \lambda$ i.e. $|y_i - \sum_{j=1}^{p} x_{ij}\beta_j| \leq \lambda$

$$\left.\begin{array}{l} \text{if } r_i > 0 \Leftrightarrow y_i - \sum_{j=1}^{p} x_{ij}\beta_j > \lambda \\ \text{if } r_i < 0 \Leftrightarrow y_i - \sum_{j=1}^{p} x_{ij}\beta_j < -\lambda \end{array}\right\} \text{ i.e. } |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| > \lambda.$$

it has the same form with the solutions derived from optimization problem (2)