LASSO REGRESSION

MA8701, 03.02.2023

• = added after class

$$\underset{N\times 1}{Y} = \underset{N\times p}{X}\,\underset{p\times 1}{\beta} + \underset{N\times 1}{\varepsilon} \qquad E(\varepsilon)=0, \quad \mathrm{Var}(\varepsilon)=\sigma^2 I$$

use centered Y and X to avoid β_0 in the model

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le t$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\arg\min}\, (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

For $t \ge t_0 = \sum_{j=1}^{p} |\hat{\beta}_{LS,j}| \Rightarrow \hat{\beta}_{lasso} = \hat{\beta}_{LS}$

$t \le t_{max} = \max_j X_j^T Y \Rightarrow \hat{\beta}_{lasso} = 0$

jth col of X

Group discussion: log PSA and 8 covariates

Discuss what do see

Ridge $X(X^TX+\lambda I)^{-1}X^T$

- X-axis: $df(\lambda) = \mathrm{tr}(H(\lambda))$ $[0,8]$
- $df \to 0 \Rightarrow \hat{\beta}_{ridge} = 0$
- $df = 8$ $\hat{\beta}_{ridge} = \hat{\beta}_{LS}$
- ranking of coeffs change not monotone

Lasso

- x-axis is $\frac{t}{\sum_{j=1}^{p} |\hat{\beta}_{LS}|}$ $[0,1]$
- $t=1$: $\hat{\beta}_{LS} = \hat{\beta}_{lasso}$
- when $\hat{\beta}$ shrink to 0, do not jump back to value >0.

- lines between knots ⇐ change in active set = non-zero coefs
- the nature of shrinkage is complex

PARAMETER ESTIMATION

In general no analytic solution to the $\hat{\beta}_{lasso}$ – except for 3 special cases: one covariate, two covariates, orthogonal design matrix.

ONE COVARIATE

$$\min_\beta \left(\overset{1;\ \frac{1}{N},\ \frac{1}{2N}}{\cdot} (\underset{N\times1}{Y} - \underset{N\times1}{X}\underset{1\times1}{\beta})^T (Y - X\beta) + \lambda |\beta| \right) \qquad \beta \text{ scalar}$$

$$\min_\beta \left(Y^TY - \underbrace{2\beta^T}_{\text{scalar}} \overbrace{X^TY}^{N\cdot\hat{\beta}_{LS}} + \beta^T \overbrace{X^TX}^{N} \beta + \lambda|\beta| \right)$$

$$\hat{\beta}_{LS} = (X^TX)^{-1}X^TY = \frac{1}{N} X^TY$$

$$\left(\sum x_i^2\right)^{-1} = N$$

standardized covs $x_i = \frac{x_i^{orig} - \bar{x}}{s}$

$\sum_{i=1}^N x_i = \frac{1}{s}\left(\overbrace{\sum x_i^{orig}}^{N\bar{x}} - N\bar{x} \right) = 0$

$\sum_{i=1}^N x_i^2 = \frac{1}{s^2}\left(\sum (x_i^{orig} - \bar{x})^2 \right)$ $= N$ or $N-1$ if $s^2 = \frac{1}{N}\sum_{i=1}^N (x_i^{orig} - \bar{x})^2$, $s^2 = \frac{1}{N-1}\sum_{i=1}^N (x_i^{orig} - \bar{x})^2$

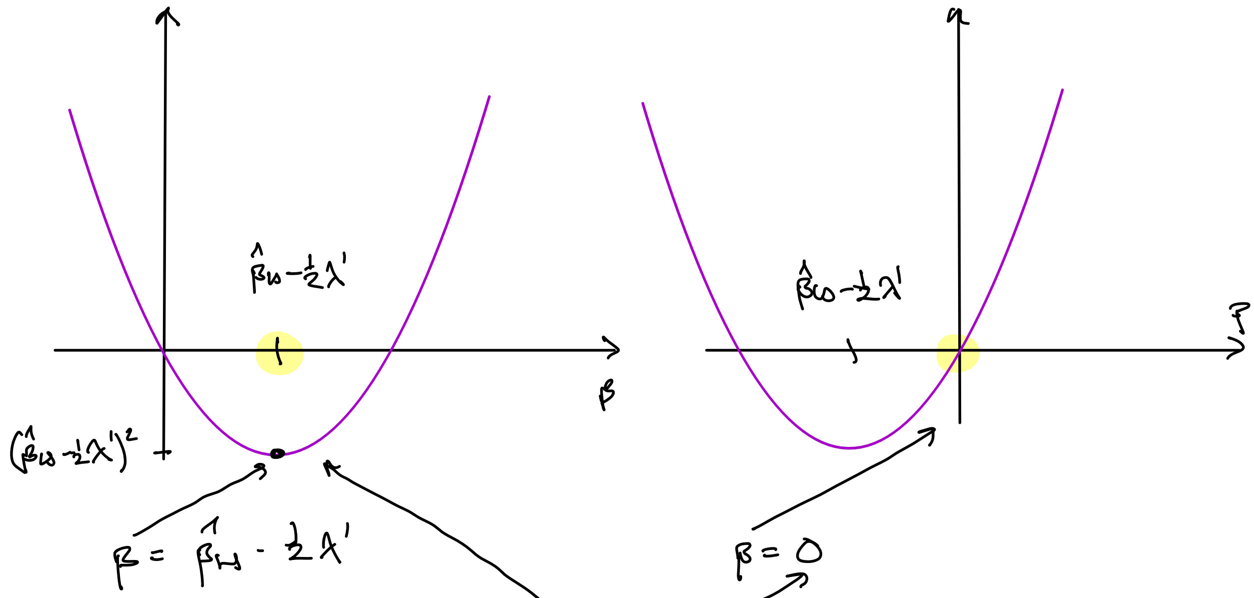
$$\min_\beta \left(-2\beta N\hat{\beta}_{LS} + N\beta^2 + \lambda|\beta| \right) = \min_\beta \left(-2\beta\hat{\beta}_{LS} + \beta^2 + \underbrace{\frac{\lambda}{N}}_{\lambda'}|\beta| \right)$$

Due to the $|\beta|$ term we look separately at $\beta \geq 0$, $\beta \leq 0$

$\underline{\beta \geq 0}$: loss $-2\beta\hat{\beta}_{LS} + \beta^2 + \lambda'\beta = \beta^2 - 2\underbrace{(\hat{\beta}_{LS} - \tfrac{1}{2}\lambda')}_{a}\beta$

$$= (\beta - (\hat{\beta}_{LS} - \tfrac{1}{2}\lambda'))^2 - (\hat{\beta}_{LS} - \tfrac{1}{2}\lambda')^2$$

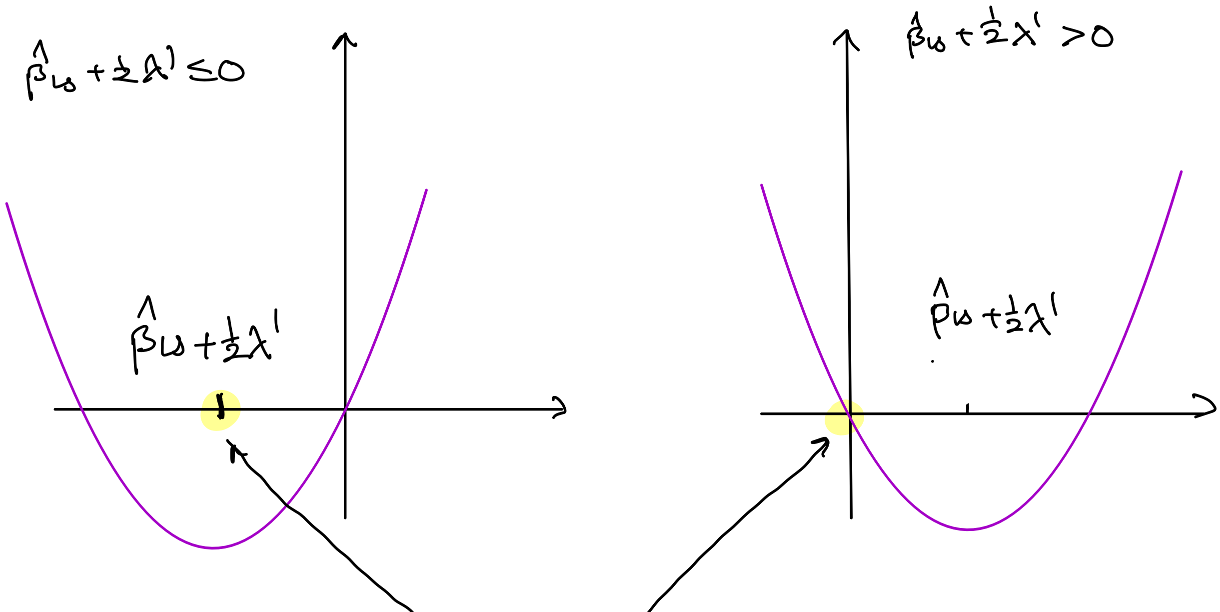
$\beta^2 - 2a\beta = (\beta - a)^2 - a^2$



$\beta = \max(\hat\beta_{LS} - \frac{1}{2}\lambda', 0)$ is minimum for $\beta \geq 0$

$\beta \leq 0$: loss $-2\beta\hat\beta_{LS} + \beta^2 - \lambda'\beta = \beta^2 - 2(\hat\beta_{LS} + \frac{1}{2}\lambda')\beta$

$= (\beta - (\hat\beta_{LS} + \frac{1}{2}\lambda'))^2 - (\hat\beta_{LS} + \frac{1}{2}\lambda')^2$

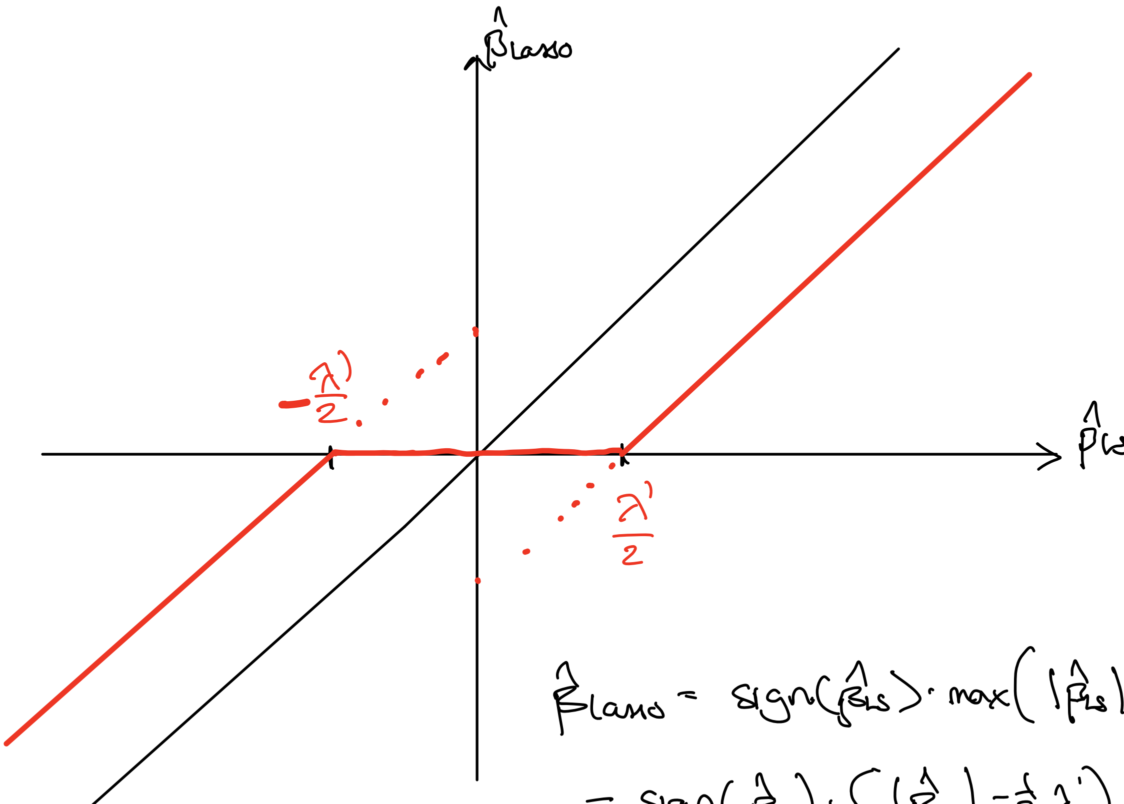


$\beta = \min(\hat\beta_{LS} + \frac{1}{2}\lambda', 0)$ for $\beta \leq 0$

---

as a function
Next: combine and do this conditional on $\hat{\beta}_{LS}$ not β

$$\left.\begin{array}{ll} \beta = \max\left(\hat{\beta}_{LS} - \tfrac{1}{2}\lambda', 0\right) & \beta \geq 0 \\ \beta = \min\left(\hat{\beta}_{LS} + \tfrac{1}{2}\lambda', 0\right) & \beta \leq 0 \end{array}\right\}$$



$$\hat{\beta}_{Lasso} = \text{sign}(\hat{\beta}_{LS}) \cdot \max\left(|\hat{\beta}_{LS}| - \tfrac{1}{2}\lambda', 0\right)$$

$$= \text{sign}(\hat{\beta}_{LS}) \cdot \left(|\hat{\beta}_{LS}| - \tfrac{1}{2}\lambda'\right)_+$$

Soft threshold operator: $S_\lambda(x) = \text{sign}(x)\,(|x| - \lambda)_+$

Ortogonal design matrix

$X^TX = I = (X^TX)^{-1}$

$\hat{\beta}_{LS} = (X^TX)^{-1}X^TY = X^TY$

$$\min_\beta \left((Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

$$= \min_\beta \left(Y^TY - 2\beta^T \underbrace{X^TY}_{\hat{\beta}_{LS}} + \beta^T \underbrace{X^TX}_{I} \beta + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

$$= \min_\beta \left(-2\beta^T \hat{\beta}_{LS} + \beta^T\beta + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

$\beta_1 \hat{\beta}_{LS,1} + \beta_2 \hat{\beta}_{LS,2} + \dots$ $\qquad$ $\beta_1^2 + \beta_2^2 + \dots$

$$= \min_\beta \left(\sum_{j=1}^{p} \left(-2\beta_j \hat{\beta}_{LS,j} + \beta_j^2 + \lambda |\beta_j| \right) \right)$$

$$= \sum_{j=1}^{p} \min_{\beta_j} \left(-2\beta_j \hat{\beta}_{LS,j} + \beta_j^2 + \lambda |\beta_j| \right)$$

$\Rightarrow$ can handle each β_j separately

$\hat{\beta}_{lasso}$: we find $\operatorname{sign}(\hat{\beta}_{LS,j}) \left(|\hat{\beta}_{LS,j}| - \frac{1}{2}\lambda \right)$

Pseudocode for cyclic coordinate descent

λ given

$t=0$: initialize $\beta_1^t, \dots, \beta_p^t$

define the ordering of the $\mathcal{J} = \{k_1, k_2, \dots, k_p\}$

for(t in 1:convergence)   [or while]

{

    for (j in $\mathcal{J}$)

$$\tilde{Y} = Y - X_{-j}\,\beta^t_{[-j]} \qquad \text{partial residuals}$$

$$\underset{\beta_j}{\text{minimize}} \quad (\tilde{Y} - X_j\beta_j)^T(\tilde{Y} - X_j\beta_j) + \lambda\,|\beta_j|$$

    } $\rightarrow$ soft thresholding $\beta_j^{(t+1)}$

}