SHRINKAGE METHODS

29/03/21

They reduce prediction error by modifying the LS criterion by constraining / shrinking the coefficients

Two main approaches: ridge regression and lasso

RIDGE REGRESSION (1970, Tikhonov regularization)

$$\hat{\beta}_{LS} = \underset{\beta}{\operatorname{argmin}} \, (\underline{y} - X\underline{\beta})^t (\underline{y} - X\underline{\beta}) \rightsquigarrow \boxed{\hat{\beta}_{LS} = (X^tX)^{-1}X^t\underline{y}}$$

$$\hat{\beta}_{RIDGE} = \underset{\beta}{\operatorname{argmin}} \, (\underline{y} - X\underline{\beta})^t (\underline{y} - X\underline{\beta}) \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq c \quad \text{for } c \geq 0$$

CONSTRAINED OPTIMIZATION

RIDGE ESTIMATOR

$$\min_{\beta} (y - X\beta)^t (y - X\beta) \quad \text{st} \quad \sum_{j=1}^{p} \beta_j^2 \leq c \qquad (*)$$

$$\min_{\substack{\beta \\ \sum \beta_j^2 \leq c}} (y - X\beta)^t (y - X\beta) = \min_{\beta} \max_{\lambda \geq 0} \underbrace{(y - X\beta)^t (y - X\beta) + \overbrace{\lambda \left(\sum_{j=1}^{p} \beta_j^2 - c \right)}^{\lambda \sum_{j=1}^{p} \beta_j^2 - \lambda c}}_{\text{Lagrangian}}$$

λ: Lagrange multiplier

$$\hat{\beta}_{RIDGE} = \arg\min_{\beta} (y - X\beta)^t (y - X\beta) + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{PENALTY}} \quad \text{for } \lambda \geq 0 \qquad (**)$$

PENALIZED INFERENCE

$$\frac{d}{d\beta} \left((y - X\beta)^t (y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right) =$$

$$= -2X^t y + 2X^t X \beta + 2\lambda \beta = 0$$

$$\Rightarrow \quad X^t y - X^t X \beta - \lambda \beta = 0$$

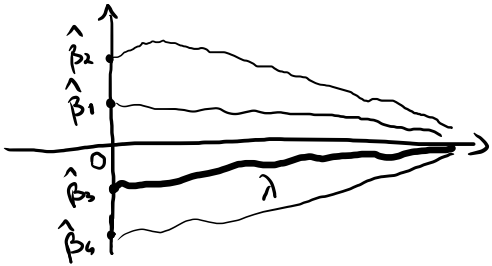
$$\Rightarrow \quad X^t y = (X^t X + \lambda I) \beta \quad \Rightarrow \quad \boxed{\hat{\beta}^{(\lambda)}_{RIDGE} = (X^t X + \lambda I)^{-1} X^t y}$$

THE ROLE OF λ (≥ 0)

1. $\lambda = 0$ ($c = +\infty$) $\rightarrow$ LS (unconstrained) solution
2. As λ increases (c decreases), the impact of the shrinkage penalty increases and the ridge coefficients will approach zero
3. λ stabilizes estimation for big data since it adds λ to the diagonal of X^tX. So $X^tX + \lambda I$ is now invertible also for $n < p$ (big data)

Remarks

① We have an estimate of $\hat{\beta}$ for each λ, so results are shown for the full path of solutions



$\sum_{j=1}^{4} \beta_j^2$ decreases with increasing λ

② λ controls the bias/variance trade off

$\lambda = 0$ OLS (no bias, high variance)

As $\lambda \to \infty$, the variance of ridge regression decreases and bias increases

BIAS / VARIANCE OF RIDGE ESTIMATOR

BIAS

$$\hat{\beta}_{RIDGE} = (X^tX + \lambda I)^{-1} X^t Y$$

$$= (R + \lambda I)^{-1} X^t Y$$

let $R = X^tX$

$$= (R + \lambda I)^{-1} R R^{-1} X^t Y$$

$$= (R + \lambda I)^{-1} R \underbrace{\left((X^tX)^{-1} X^t Y\right)}_{\hat{\beta}_{LS}}$$

$$= \left(R\left(I + \lambda R^{-1}\right)\right)^{-1} R \hat{\beta}_{LS}$$

$$= \left(I + \lambda R^{-1}\right)^{-1} \hat{\beta}_{LS}$$

$(AB)^{-1} = B^{-1}A^{-1}$

$$E[\hat{\beta}_{RIDGE}] = (I + \lambda R^{-1})^{-1} E[\hat{\beta}_{LS}] \overset{\text{LS unbiased}}{=} \underbrace{(I + \lambda R^{-1})^{-1}}_{\neq I \text{ unless } \lambda = 0} \beta \neq \beta$$

BIASED with bias increasing with λ

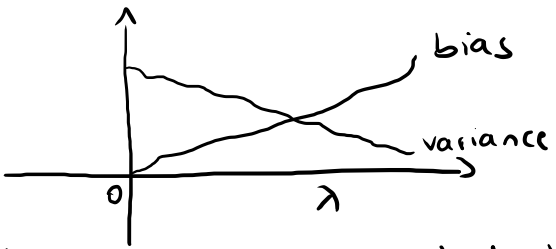
VARIANCE

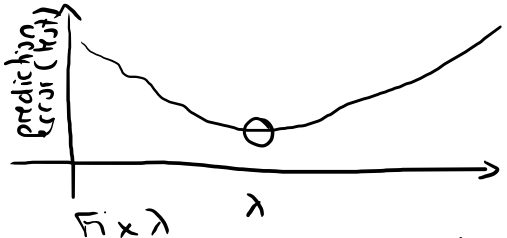
$$\hat{\beta}_{RIDGE} = (I + \lambda R^{-1})^{-1} \hat{\beta}_{LS}$$

$$Var(\hat{\beta}_{RIDGE}) = (I + \lambda R^{-1})^{-1} \underbrace{Var(\hat{\beta}_{LS})}_{\sigma^2 (X^t X)^{-1}} (I + \lambda R^{-1})^{-1}$$

$Var(A\underline{Y}) = A\, Var(\underline{Y})\, A^t$

$$= \underbrace{\sigma^2 (I + \lambda R^{-1})^{-1} (X^t X)^{-1} (I + \lambda R^{-1})^{-1}}_{\text{Decreases to 0 as } \lambda \to \infty}$$





λ can be chosen cross-validation
- Split data in K-folds
- $\hat{\beta}_{RIDGE}$ estimated from K-1 folds and make predictions on test set

SCALING OF DATA

The standard LS estimates are not affected by scaling: if we multiply X_j by a constant c, $\hat{\beta}_j$ will be scaled by a factor of $\frac{1}{c}$, so regardless of how X_j is scaled, $X_j \hat{\beta}_j$ remains the same

This does not happen for ridge regression. This is due to $\sum_{j=1}^{p} \beta_j^2$

For this reason, ridge regression is applied after standardizing the predictors (to mean 0 and variance 1).

As for the intercept, this is typically estimated beforehand ($\hat{\beta}_0 = \bar{y}$, unpenalized). The response variable is then centered to have mean zero and $\beta_1, \ldots, \beta_p$ are estimated by ridge regression (to standardized data)

LASSO

Least absolute shrinkage and selection operator

Reduces variance and performs automatic variable selection

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \; (y - X\beta)^t (y - X\beta) + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\|\beta\|_1} \quad \text{with } \lambda \geq 0$$

As before, this is equivalent to

$$\left.\begin{aligned} &\min \; (y - X\beta)^t (y - X\beta) \\ &\text{s.t.} \; \sum_{j=1}^{p} |\beta_j| \leq C \end{aligned}\right]$$

Compare this with subset selection:

$$\left.\begin{aligned} &\min \; (y - X\beta)^t (y - X\beta) \\ &\text{s.t.} \; \sum_{j=1}^{p} I(\beta_j \neq 0) \leq C \end{aligned}\right]$$

ROLE OF λ

1. $\lambda = 0 \iff$ LS
2. As λ increases (c decreases), more and more coefficients are shrunk to zero
3. One $\hat{\beta}_{Lasso}$ for each λ, so you can look at the path of solution, but no closed-form analytical solution
3. Choice of λ is crucial for bias/variance trade-off. So use CV to choose optimal value of λ.

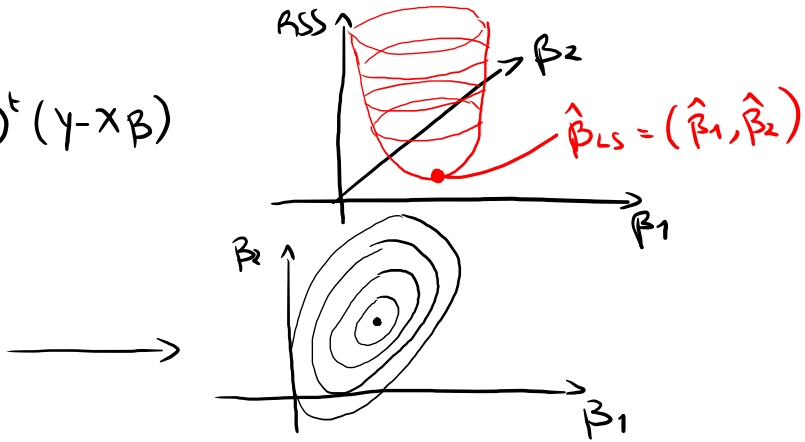
WHY LASSO ESTIMATES CAN BE EXACTLY ZERO

We say that lasso yields sparse models, that is the models have only a subset of predictors.

In order to see this, let us take two predictors, so we wish to estimate β_1 and β_2.

PLOT OF

$$RSS = (y - X\beta)^t (y - X\beta)$$



CONSTRAINED OPTIMAZATION

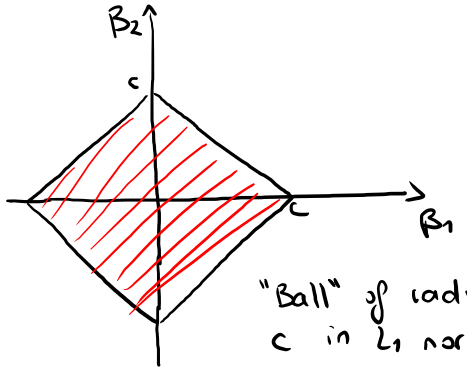
LASSO

$\sum_{i=1}^{p} |\beta_i| \leq C$

$|\beta_1| + |\beta_2| \leq C$

$\Leftrightarrow$

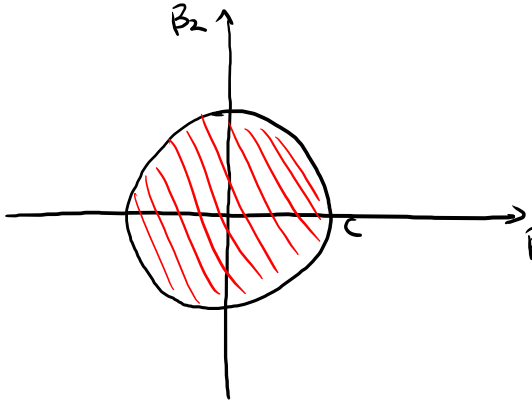
$$\begin{cases} \beta_1 + \beta_2 \leq C & \beta_1, \beta_2 \geq 0 \\ \beta_1 - \beta_2 \leq C & \beta_1 \geq 0, \beta_2 \leq 0 \\ -\beta_1 + \beta_2 \leq C & \beta_1 \leq 0, \beta_2 \geq 0 \\ -\beta_1 - \beta_2 \leq C & \beta_1 \leq 0, \beta_2 \leq 0 \end{cases}$$



"Ball" of radius c in L_1 norm

RIDGE

$\beta_1^2 + \beta_2^2 \leq C$



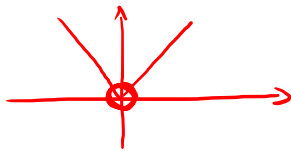
"Ball" of radius c in L_2 norm

EXAMPLE

Let's assume that we have one predictor ($p = 1$)

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \underbrace{(y - X\beta)^t (y - X\beta)}_{\sum_{i=1}^{n} (y_i - x_i \beta)^2} + \lambda |\beta|$$

$$|\beta| = \begin{cases} \beta & \beta \geq 0 \\ -\beta & \beta < 0 \end{cases}$$



DERIVATIVE w.r.t β

$$-2 X^t (y - X\beta) + \lambda \, \text{sign}(\beta) = 0$$

where $\text{sign}(\beta) = \begin{cases} 1 & \beta > 0 \\ -1 & \beta < 0 \\ [-1,1] & \beta = 0 \text{ (sub-differential)} \end{cases}$

$$X^t(y - X\beta) - \frac{\lambda}{2}\,\text{sign}(\beta) = 0$$

$\hat{\beta}_L > 0 \Rightarrow X^t(y - X\hat{\beta}_L) = \frac{\lambda}{2}$

$\Rightarrow X^t y - X^t X \hat{\beta}_L = \frac{\lambda}{2} \Rightarrow (X^t y - \frac{\lambda}{2}) = X^t X \hat{\beta}_L$

$\Rightarrow \underline{\hat{\beta}_L = (X^t X)^{-1}(X^t y - \frac{\lambda}{2}) = \underbrace{(X^t X)^{-1} X^t y}_{\hat{\beta}_{LS}} - (X^t X)^{-1}\frac{\lambda}{2} =}$

$\underline{= \hat{\beta}_{LS} - (X^t X)^{-1}\frac{\lambda}{2}}$

$\hat{\beta}_L < 0 \Rightarrow X^t(y - X\hat{\beta}_L) = -\frac{\lambda}{2} \Rightarrow \hat{\beta}_L = \hat{\beta}_{LS} + (X^t X)^{-1}\frac{\lambda}{2}$

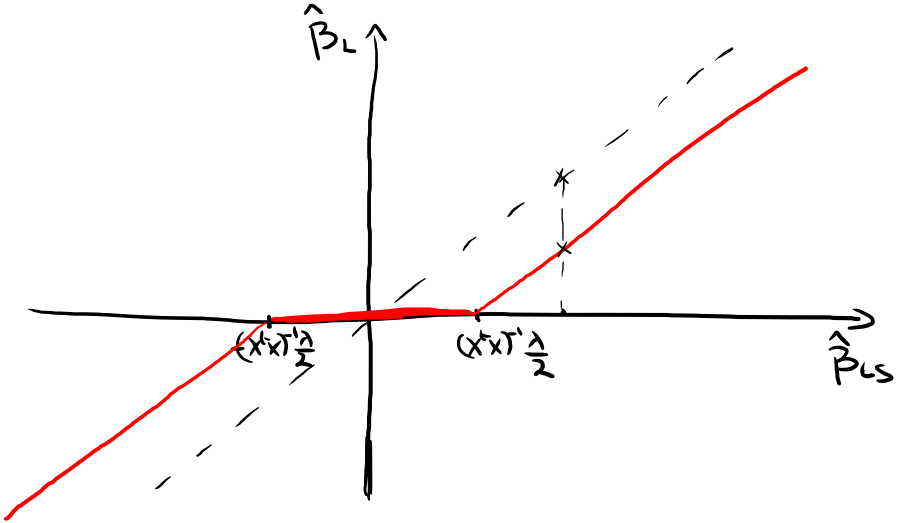
$\underline{\hat{\beta}_L = 0}$

$\Rightarrow X^t(y - X\hat{\beta}_L) \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]$

$\Rightarrow X^t y - X^t X \hat{\beta}_L \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]$

$\Rightarrow \underbrace{(X^t X)^{-1} X^t y}_{\hat{\beta}_{LS}} - \cancel{(X^t X)^{-1}(X^t X)}\hat{\beta}_L \in [-(X^t X)^{-1}\frac{\lambda}{2}, (X^t X)^{-1}\frac{\lambda}{2}]$

$\Rightarrow \hat{\beta}_{LS} \in [-(X^t X)^{-1}\frac{\lambda}{2}, (X^t X)^{-1}\frac{\lambda}{2}]$



$p = 1$

$$(X^TX)^{-1} = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

SOFT-THRESHOLDING Lasso penalty translates the value towards zero, making them exactly zero when small enough

LASSO ALGORITHM (COORDINATE DESCENT)

Obj function $f(\beta) = (y - X\beta)^t (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$

$= f(\beta_1, \beta_2, \ldots, \beta_p)$

Optimise wrt to β_j while keeping all the other values fixed to the value in the previous iteration.

The algorithms are extremely efficient!

FINAL COMMENTS

1. Both ridge and lasso admit solutions also when X^TX is not invertible so they can be used on high-dimensional data ($p \gg n$)

2. Regularization reduces prediction error, in terms of reducing variance but at the expense of bias.
   λ choice is crucial and needs CV

3. Lasso produces sparse solutions compared to ridge, though with higher variance.

4. Correlated predictors $\rightsquigarrow$ ridge likely to do better
   Few associated predictors $\rightsquigarrow$ lasso might do better and be more useful than ridge.

5. They have been extended in different directions $<$ penalties / glm