Make the size of the coefficients small $\Rightarrow$ "a continuous version of sparsity".

Lecture 8

21 October 2019 08:59

With many thanks to Dr Giordano Scarciotti

② SHRINKAGE METHODS:

LINEAR REGRESSION (LS): GIVEN $X, \vec{y}$ ONE FIND $\vec{\beta}^*$

$$\min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 = \min_{\vec{\beta}} L_{LS}(\vec{\beta})$$

THE SOLUTION IS:

$$\vec{\beta}^* = \operatorname{ARGMIN} \|\vec{y} - X\vec{\beta}\|^2 = X^+ \vec{y}, \qquad X^+ = (X^T X)^{-1} X^T$$

$$\hat{y} = \hat{f}(\vec{x}_{IN}) = \vec{x}_{IN} \vec{\beta}^*$$

2.1 RIDGE REGRESSION

$$L_{RIDGE}(\vec{\beta}) \ \|\vec{y} - X\vec{\beta}\|^2 + \lambda \|\vec{\beta}\|^2, \qquad \lambda > 0$$

(THE IDEA IS TO "ELIMINATE" SOME DESCRIPTORS. IN PRACTICE WE WEIGHT THEM LOW)

$$\boxed{\min_{\vec{\beta}} L_{RIDGE}(\vec{\beta}) = \min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 + \boxed{\lambda \|\beta\|^2}}$$

Penalty term

||| EQUIVALENT

$$\boxed{\min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 \quad \text{SUBJECT TO} \quad \|\vec{\beta}\|^2 \leq t}$$

λ and t are inversely related

THIS PROBLEM CAN BE SOLVED EXPLICITLY:

$$L_{RIDGE}(\vec{\beta}) = \vec{y}^T\vec{y} - \vec{\beta}^T X^T \vec{y} - \vec{y}^T X \vec{\beta} + \vec{\beta}^T (X^T X + \lambda I)\vec{\beta}$$

$$\nabla_{\vec{\beta}} L_{RIDGE} = -2X^T\vec{y} + 2(X^T X + \lambda I)\beta$$

$$X^T \vec{y} = (X^T X + \lambda I)\vec{\beta}^*$$

$$\vec{\beta}^*_{RIDGE} = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

WE CAN CHECK THAT THE HESSIAN IS POSITIVE DEFINITE

BIAS: $\vec{y} = X\vec{\beta} + \vec{\varepsilon} \qquad \mathbb{E}[\vec{\varepsilon}] = 0_v$

$$\mathbb{E}\left[\vec{\beta}^*_{RIDGE}\right] = \mathbb{E}\left[(X^T X + \lambda I)^{-1}(X^T X)\vec{\beta} + (X^T X + \lambda I)^{-1} X^T \varepsilon\right] =$$

$$= (X^T X + \lambda I)^{-1}(X^T X)\vec{\beta}$$

$$(X^T X) \underset{\text{EIGENDECOMPOSITION}}{=} V D V^T \qquad V V^T = V^T V = I \qquad (X^T X)^{-1} = V D^{-1} V^T$$

D DIAGONAL

$\sigma(D) \equiv \sigma(X^T X)$

- V contains the eigenvectors as columns
- $D = \operatorname{diag}(d_i)$ has the eigenvalues on the diagonal.

$$\text{BIAS} = \mathbb{E}\left[\vec{\beta}^*_{\text{RIDGE}}\right] - \vec{\beta} = (X^TX+\lambda I)^{-1}X^TX\vec{\beta} - \vec{\beta} = V\left[(D+\lambda I)^{-1}D - I\right]V^T\vec{\beta}$$

$$\mathcal{D} = (D+\lambda I)^{-1}D - I \quad \longleftarrow \text{EVERYTHING IS DIAGONAL}$$

$$\mathcal{D}_{ii} = \left[\frac{d_i}{d_i+\lambda} - 1\right] = -\frac{\lambda}{d_i+\lambda}$$

$$\Rightarrow \text{BIAS} = -\lambda V\left[(D+\lambda I)^{-1}\right]V^T\vec{\beta} = -\lambda(X^TX+\lambda I)^{-1}\vec{\beta}$$

As expected as $\lambda \to 0$ BIAS $\to 0$ because we recover least squares

As $\lambda \to \infty$ BIAS $\to -\vec{\beta}$ ($\approx$ 100% error)

Variance:

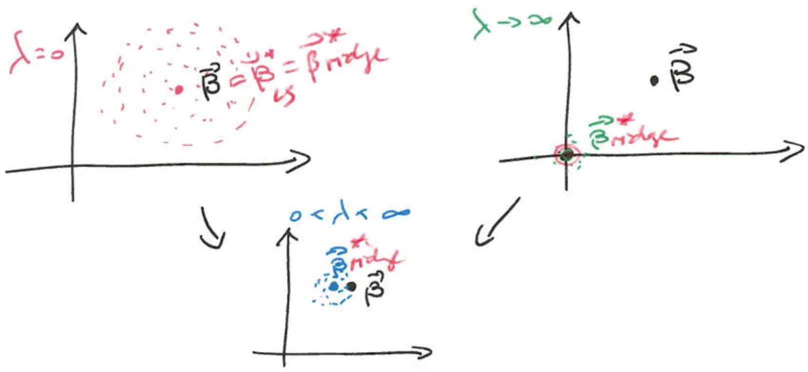
$$\text{VAR}(\vec{\beta}_{\text{RIDGE}}) = \mathbb{E}\left[\left(\vec{\beta}_{\text{RIDGE}} - \mathbb{E}\left[\vec{\beta}^*_{\text{RIDGE}}\right]\right)^2\right]$$

$$= \mathbb{E}\left[(X^TX+\lambda I)^{-1}X^T\underbrace{\vec{\varepsilon}\vec{\varepsilon}^T}_{\sigma^2 I}X(XX^T+\lambda I)^{-1}\right] =$$

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

$$= \sigma^2 V\left[\underbrace{(D+\lambda I)^{-1}D(D+\lambda I)^{-1}}_{\mathcal{P}}\right]V^T \qquad P_{ii} = \frac{d_i}{(d_i+\lambda)^2}$$

As $\lambda \to \infty$ $P_{ii} \to 0$ (quadratically)



2.2 LASSO (Tibshirani)

$$L_{\text{LASSO}}(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|^2 + \lambda\|\vec{\beta}\|_1 \qquad \boxed{\|\vec{\beta}\|_1 = \sum_{i=1}^{p}|\beta_i|}$$

$$\boxed{\min_{\vec{\beta}} L_{\text{LASSO}}(\vec{\beta})} \iff \boxed{\min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 \quad \text{subject to} \quad \|\vec{\beta}\|_1 \le t}$$

~~We can't find an analytical solution but this is a convex~~

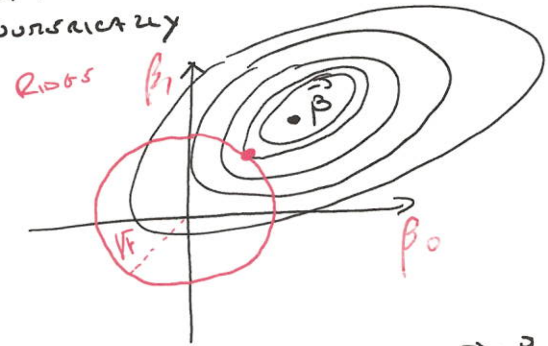
No analytical solution for LASSO (contrary to Ridge)

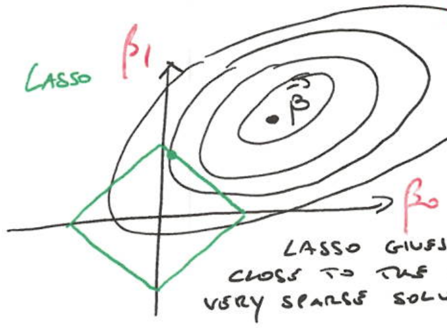
but it is a convex problem [optimize a convex function over a convex set]

So we can use convex optimisation techniques (quadratic programming) to optimise globally.

optimisation viewpoint:

~~WE USE~~ ... OPTIMISATION PROBLEM THAT HAS A SOLUTION THAT ... NUMERICALLY

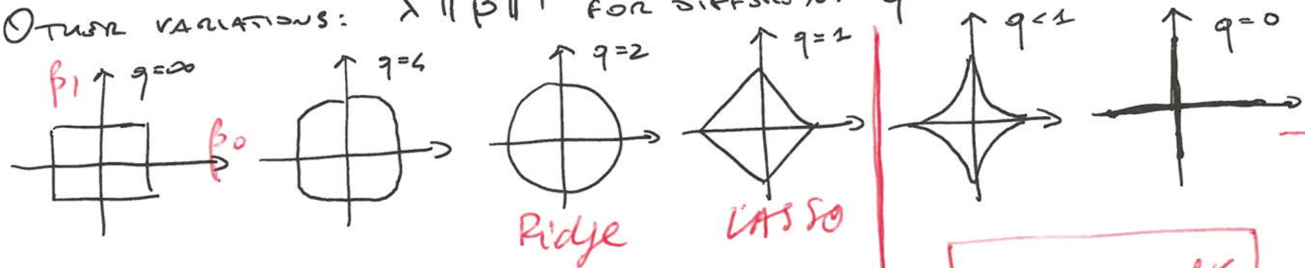




level of LS curves $\|\vec{y}-X\vec{\beta}\|^2 = K$

LASSO GIVES SOLUTIONS CLOSE TO THE AXES, i.e. VERY SPARSE SOLUTIONS.

OTHER VARIATIONS: $\lambda \|\vec{\beta}\|^q$ FOR DIFFERENT q



$\boxed{\text{convex sets}}$ ⟵ optimisation is doable

$\boxed{\text{Non-convex sets}}$ ⟸ optimisation is difficult

$q=0$ is the "ℓ_0"-pseudonorm case

↓

where we only have solutions where some of the parameters β_i are <u>zero</u> $\equiv$ $\boxed{\text{Subset selection}}$

⇓

i.e., "Sparse" models

↳ Difficult to optimise for sparsity

this whole area is called <u>regularisation</u>

or

controlling for model complexity

Another direction is to mix the penalty terms in objective function:

Example:

Elastic net:

$$L_{EN}(\vec{\beta}) = \| \vec{y} - X\vec{\beta} \|_2 + \lambda \left[\alpha \|\vec{\beta}\|_1 + (1-\alpha) \|\vec{\beta}\|_2 \right]$$