Erasmus+ machine Learning lecture, October 9, 2023

SVD analysis of OLS

Design matrix $X \in \mathbb{R}^{n \times p}$

$$X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0\,p-1} \\ x_{10} & & & \\ 1 & & & \\ 1 & & & \\ x_{n-1\,0} & \cdots & \cdots & x_{n-1\,p-1} \end{bmatrix}$$

$$X = \begin{bmatrix} X_0 & X_1 & X_2 & \cdots & X_{p-1} \end{bmatrix}$$

$$X = U \Sigma V^T$$

$$U \in \mathbb{R}^{n \times n}$$

$$V \in \mathbb{R}^{p \times p}$$

$$U U^T = U^T U = \underline{1}$$

$$V V^T = V^T V = \underline{1}$$

$$n \geq p$$

$$\Sigma = \begin{bmatrix} \sigma_0 & & & 0 \\ & \sigma_1 & & \\ & & \ddots & \\ 0 & & & \sigma_{p-1} \\ & & & 0 \end{bmatrix}$$

$$\Sigma \in \mathbb{R}^{n \times p}$$

$$\sigma_0 > \sigma_1 > \sigma_2 > \cdots > \sigma_{p-1} > 0$$

Example $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$

$\Sigma^T \Sigma \in \mathbb{R}^{p \times p}$ $\quad \Sigma \Sigma^T \in \mathbb{R}^{n \times n}$

$\Sigma^T \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = \frac{2}{n} X^T X$$

$$X^T X = (U \Sigma V^T)^T U \Sigma V^T$$

$$= V \Sigma^T \underbrace{U^T U}_{= I} \Sigma V^T$$

$$= V \Sigma^T \Sigma V^T$$

$$[X^T X] \cdot V = V \Sigma^T \Sigma V^T V = \underbrace{V}_{\mathbb{R}^{p \times p}} \underbrace{\Sigma^T \Sigma}_{\mathbb{R}^{p \times p}}$$

$$V = \begin{bmatrix} v_0 & v_1 & v_2 & \cdots & v_{p-1} \end{bmatrix}$$

$$v_i^T v_j = \delta_{ij}$$

$$[X^T X] V = V \begin{bmatrix} \sigma_0^2 & & & 0 \\ & \sigma_1^2 & & \\ & & \ddots & \\ 0 & & & \sigma_{p-1}^2 \end{bmatrix}$$

$$[X^T X] v_i = v_i \sigma_i^2$$

eigenvalues of $X^T X$ are the singular values of X. σ_i^2 with eigenvectors v_i. $\sigma_i^2 > 0 \Rightarrow X^T X$ is

a positive definite matrix $\Rightarrow$ convex optimization problem.

$\frac{1}{n} X^T X$ is the covariance matrix

For completeness

$$XX^T \in \mathbb{R}^{n \times n}$$

$$XX^T = U \Sigma \underbrace{V^T V} \cdot \Sigma^T U^T$$

$$= U \Sigma \Sigma^T U^T$$

$$[XX^T] U = U \Sigma \Sigma^T = U \begin{bmatrix} \sigma_0^2 & & & \\ & \sigma_1^2 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \cdot U$$

$$U = [u_0 u_1 u_2 \ldots u_{n-1}]$$

$$\Rightarrow [XX^T]\, u_i = \sigma_i^2 u_i$$

$$\sigma_0^2, \sigma_1^2 \ldots \sigma_{p-1}^2, 0, 0 \ldots \quad n-1$$

OLS

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

SVD

$$= U \Sigma V^T \left[V \Sigma^T \Sigma V^T \right]^{-1} V \Sigma^T U^T y$$

assume A and B, are square and both are invertible

V; $V^T = V^{-1}$

$\underline{\Sigma^T \Sigma} = \begin{bmatrix} \sigma_0^2 & & \\ & \sigma_1^2 & \\ & & \ddots \\ & & & \sigma_{p-1}^2 \end{bmatrix}$

is also invertible $\in \mathbb{R}^{p \times p}$

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

$$UV^T = V^T U = \underline{\mathbb{1}}$$

$$\left(V \Sigma^T \Sigma V^T\right)^{-1} = \left(V^T\right)^{-1} \left(\Sigma^T \Sigma\right)^{-1} V^{-1}$$

$$= V \left(\Sigma^T \Sigma\right)^{-1} V^T \qquad \left(\in \mathbb{R}^{p \times p}\right)$$

$$\tilde{y} = U \Sigma \underbrace{V^T V}_{=I} \left(\Sigma^T \Sigma\right)^{-1} \underbrace{V^T V}_{=I} \Sigma^T U^T y$$

$$= U \Sigma \left(\frac{1}{\Sigma^T \Sigma}\right) \Sigma^T U^T y$$

$$\tilde{y} = \left[\sum_{j=0}^{p-1 \,(<n)} u_j u_j^T\right] y$$

$$\tilde{y} = u\Sigma\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T u^T y$$

$$\Sigma = \begin{bmatrix} \sigma_0^2 & & & & \\ & \sigma_1 & & & \\ & & \sigma_2 & & \\ & & & \ddots & \\ & & & & \sigma_{p-1} \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{m \times p}$$

$$u = [u_0\ u_1\ u_2 \cdots u_{m-1}]$$

$$\Sigma^T u^T \in \mathbb{R}^{p \times m}$$

$$\Sigma^T U^T = \begin{bmatrix} \sigma_0 u_0 & \sigma_1 u_1 & \cdots & \sigma_{p-1} u_{p-1} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\Sigma^T = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} u_{00} & u_{01} & u_{02} \\ u_{10} & u_{11} & u_{12} \\ u_{20} & u_{21} & u_{22} \end{bmatrix} \qquad n = 3$$

$$= \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix}$$

$$\tilde{y} = \left[\sum_{j=0}^{p-1} u_j u_j^T \right] y$$

$$p \leq m \left[\sum_{j=0}^{m-1} u_i u_j^T \right] y = U U^T y$$

$$= y$$

projection of y in terms of the orthogonal vectors U_i but including only the singular values different from zero.

X^TX when it is singular cannot invert, cheap trick is to simply a positive small constant λ to the diagonals.

$$X^TX \in \mathbb{R}^{p \times p}$$

$$X^TX + \lambda I_{p \times p} \Rightarrow$$

$$\tilde{y} = \underset{\in \mathbb{R}^{n \times p}}{X}\left(X^TX + \lambda I\right)^{-1} X^T y$$

Ridge Regression

Cost function (objective)

$$C(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=0}^{p-1} \beta_j^2$$

— Regularization or shrinkage

$$= \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=0}^{p-1} \beta_j^2$$

$$= \frac{1}{n}\|(y - X\beta)\|_2^2 + \lambda \|\beta\|_2^2$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\frac{\partial C}{\partial \beta} = -\frac{2}{n} x^T (y - x\beta) + 2\lambda\beta = 0$$

$$\lambda \geq 0$$

$$\lambda \cdot n = \lambda^2 \Rightarrow \lambda$$

$$0 = -x^T y + x^T x \beta + \lambda \beta \Rightarrow$$

$$\hat{\beta} = (x^T x + \lambda I)^{-1} x^T y$$

Ridge Regression

λ is a hyperparameter (shrinkage parameter)

can show that (exercises this week)

$$\tilde{y} = \left[\sum_{j=0}^{p-1} u_j u_j^T \frac{\nabla_j^2}{\nabla_j^2 + \lambda} \right] y$$

when λ is large, we "shrink" the contribution from a given component

- 5 -

Example

$X = \mathbb{1}$ $\qquad \tilde{y} = \beta$; $n = p$

OLS : $C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \beta_i)^2$

$\tilde{y} = X\beta = \beta$

minimize wrt to β $\Rightarrow$

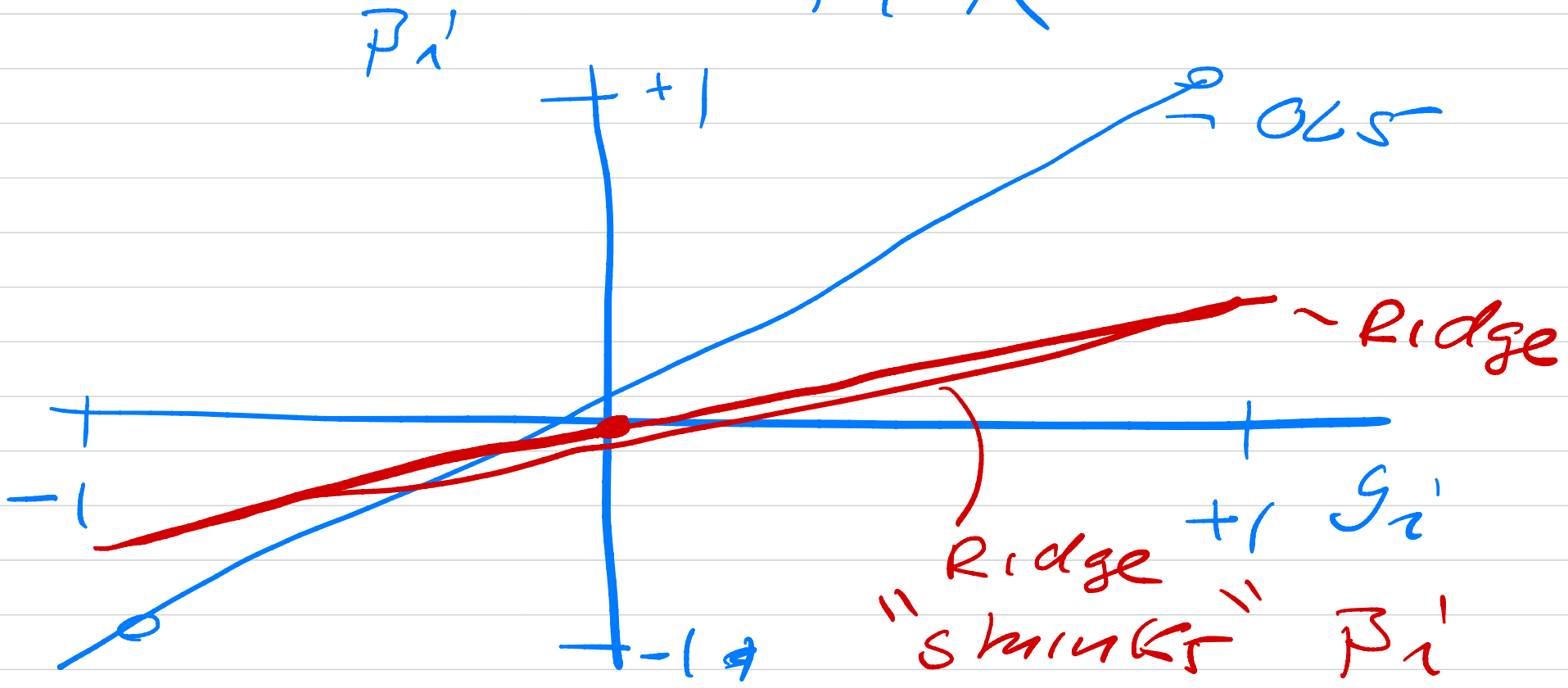
$$\boxed{y_i = \beta_i}$$

Ridge : $C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \beta_i)^2$

$$+ \lambda \sum_{i=0}^{n-1} \beta_i^2$$

$$\frac{\partial C}{\partial \beta_i'} = 0 = -\frac{2}{n}(y_i' - \beta_i') + 2\lambda\beta_i'$$

$$\hat{\beta}_i^{Ridge} = \frac{y_i'}{1+\lambda}$$



Lasso Regression

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \sum_{j=0}^{p-1} x_{ij} \beta_j \right)^2$$

$$+ \lambda \sum_{j=0}^{p-1} |\beta_j|$$

$$= \frac{1}{n} \| (y - X\beta) \|_2^2 + \lambda \| \beta \|_1$$

we cannot find an analyti-cal expression for $\hat{\beta}^{Lasso}$ due to

$$\frac{d|x|}{dx} = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Example $\tilde{y} = \beta$

$$C(\beta) = \frac{1}{n} \sum_i (y_i - \beta_i)^2$$

$$+ \lambda \sum_i |\beta_i|$$

$$\frac{\partial C}{\partial \beta_i} = -\frac{2}{n}(y_i - \beta_i) + \lambda \frac{\beta_i}{|\beta_i|}$$

$$\beta_i^{lasso} = \begin{cases} y_i - \lambda/2 & \text{if } y_i > \lambda/2 \\ y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & \text{if } |y_i| \le \lambda/2 \end{cases}$$