Marius Huy, Philipp Moehler, Jonas Kleine

1) Bias and variance of ridge regression:

$$\hat{\beta}_{RR} = (X^TX + \tau \mathbb{1})^{-1} X^T y$$

Use SVD to get a decomposition for X^TX:

Assume $\exists U, \Lambda, V$ s.t. $X = U \Lambda V^T$ and U, V are unitary.

$\Rightarrow$ • $X^TX = U\Lambda^2 V^T$

• $X^TX + \tau \mathbb{1} = U\Lambda^2V^T + \tau \underbrace{\mathbb{1}}_{= VV^T = V\mathbb{1}V^T} = U\Lambda V^T + \tau V \mathbb{1} V^T = V(\Lambda^2 + \tau \mathbb{1})V^T$

this isn't even used at any point?

Assume: $y = X\beta^* + \varepsilon$ and insert in $\hat{\beta}_{RR}$:

$$\Rightarrow \hat{\beta}_{RR} = (X^TX + \tau\mathbb{1})^{-1} X^T y \qquad (y = X\beta^* + \varepsilon)$$

$$= (X^TX + \tau\mathbb{1})^{-1} X^T(X\beta^* + \varepsilon)$$

$$= (X^TX + \tau\mathbb{1})^{-1} (X^TX\beta^* + X^T\varepsilon)$$

$$\Rightarrow \mathbb{E}(\hat{\beta}_{RR}) = \mathbb{E}\left((X^TX + \tau\mathbb{1})^{-1} X^TX\beta^*\right) + \underbrace{\mathbb{E}(X^T)\mathbb{E}(\varepsilon)}_{=0}$$

ε and Data are independent

$$= \mathbb{E}\left((X^TX + \tau\mathbb{1})^{-1} X^TX\beta^*\right)$$

$$= (X^TX + \tau\mathbb{1})^{-1} X^TX\beta^* = S_\tau^{-1} S \beta^*$$

training set is const.

$$\mathrm{Cov}(\hat\beta_\tau) = \mathbb{E}\left[(\hat\beta_{RR} - \mathbb{E}(\hat\beta_{RR}))(\hat\beta_{RR} - \mathbb{E}(\hat\beta_{RR}))^T\right]$$

$$= \mathbb{E}\left[(\hat\beta_{RR} - S_\tau^{-1} S \beta^*)(\hat\beta_{RR} - S_\tau^{-1} S \beta^*)^T\right]$$

$$= \mathbb{E}\left((S_\tau^{-1} x^T y - S_\tau^{-1} S \beta^*)(S_\tau^{-1} x^T y - S_\tau^{-1} S \beta^*)^T\right)$$

$$= \mathbb{E}\left(S_\tau^{-1}(x^T y - S\beta^*)\,(S_\tau^{-1}(x^T y - S\beta^*))^T\right)$$

$$= \mathbb{E}\left(S_\tau^{-1}(x^T y - S\beta^*)\,(x^T y - S\beta^*)^T S_\tau^{-T}\right)$$

$$= S_\tau^{-1}\, \mathbb{E}\left(\underbrace{(x^T y - S\beta^*)}(x^T y - S\beta^*)^T\right) S_\tau^{-T}$$

I like this order of calculation more than that of the sample solution :) It avoids excessively long terms like the one in line (19) of s.s.

$$\rightarrow \quad x^T y = x^T x \beta^* + x^T \varepsilon$$

$$\Rightarrow \quad x^T y - S\beta = x^T x \beta^* + x^T \varepsilon - x^T x \beta^*$$

$$= x^T \varepsilon$$

$$= S_\tau^{-1}\, \underbrace{\mathbb{E}\left((x^T\varepsilon)(x^T\varepsilon)^T\right)} S_\tau^{-T}$$

$$\Rightarrow \quad \mathbb{E}\left(x^T(\varepsilon - 0)(\varepsilon - 0)^T x\right)$$

$$= x^T\, \mathbb{E}\underbrace{\left[(\varepsilon - 0)(\varepsilon - 0)^T\right]}_{=\sigma^2} x = x^T x\, \sigma^2 = S\sigma^2$$

$$= S_\tau^{-1}\, S\sigma^2\, S_\tau^{-T} = S_\tau^{-1}\, S\, S_\tau^{-1} \sigma^2$$

$$\uparrow \quad S_\tau^{-T} = S_\tau^{-1}$$

$$S^T S_\tau^{-T} = x^T x \underbrace{(x^T x + \tau \mathbb{1})^{-T}}_{(x^T x + \tau \mathbb{1})^{-1}}$$

$$\Rightarrow \quad \mathrm{Cov}(\hat\beta_\tau) = S_\tau^{-1}\, S\, S_\tau^{-1} \sigma^2$$

$\square$

2) LDA-Derivation:

$$\widehat{\beta}_{\text{OLS}} = \text{argmin}_\beta \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 \quad \Longrightarrow \quad \widehat{\beta}_{\text{OLS}} = \tau \Sigma^{-1}(\mu_1 - \mu_{-1})^T$$

Calculate:

$$\nabla_\beta \sum_{i=1}^{N} (y_i^* - x_i \beta)^2$$

$$= 2 \sum_{i=1}^{N} (y_i^* - x_i \beta) \, \nabla_\beta (x_i \beta)$$

$$(\nabla_\beta (x_i \beta))_j = \frac{\partial}{\partial \beta_j} \sum_{r=1}^{D} x_{ir} \beta_r$$

$$= \sum_{j=1}^{D} x_{ir} \delta_{rj}$$

$$= x_{ij} \implies \nabla_\beta (x_i \beta) = \underline{\underline{x_i^T}}$$

$$= -2 \sum_{i=1}^{N} (y_i^* - x_i \beta) x_i^T$$

$$\stackrel{!}{=} 0$$

$$\iff \sum_{i=1}^{N} (y_i^* x_i^T - x_i \beta_{ols} x_i^T) \stackrel{!}{=} 0$$

$$\iff \sum_{i=1}^{N} y_i x_i^T = \sum_{i=1}^{N} x_i \beta_{ols} x_i^T$$

not necessary with correct derivative

$$\underbrace{x_i \beta_{OLS}}_{\in \mathbb{R}} x_i^T.$$

$$\iff x_i^T x_i \beta_{OLS} = x_i \beta_{OLS} x_i^T$$

$$= \left(\sum_{i=1}^{N} x_i^T x_i \right) \beta_{ols}$$

Use $\sum_{i=1}^{N} y_i^* x_i^T = \sum_{i:y_i^*=1} x_i^T - \sum_{i:y_i^*=-1} x_i^T$

$$= (N_1 \mu_1 - N_{-1}\mu_{-1})^T$$

$$\underset{N_1 = N_{-1} = \frac{N}{2}}{=} \frac{N}{2}(\mu_1 - \mu_{-1})^T$$

use (2)

Look at

$$\sum_{i=1}^{N} x_i^T x_i = \sum_{i:y_i^*=1} x_i^T x_i + \sum_{i:y_i^*=(-1)} x_i^T x_i$$

$$\overset{!}{=} N\left(\Sigma + \frac{1}{4}(\mu_1 - \mu_{-1})^T(\mu_1 - \mu_{-1})\right)$$

$$\overset{?}{=} N\left(\Sigma - \mu_1^T \mu_{-1}\right)$$

$$= \sum_{i:y_i^*=1} \underbrace{(x_i - \mu_{+1})^T (x_i - \mu_{+1})}_{x_i^T - \mu_{+1}^T (x_i - \mu_{+1})} + \sum_{i:y_i^*=(-1)} (x_i - \mu_{+1})^T (x_i - \mu_{-1}) - N\mu_1^T \mu_{-1}$$

$$= x_i^T x_i - x_i^T \mu_{+1} - \mu_{+1}^T x_i + \mu_{+1}^2$$

This is incredibly tedious to read but appears sort of maybe right. Some explanation would go a long way.

$$\Rightarrow \sum_{i:y_i^*=1} x_i^T x_i - \underbrace{\sum_{i:y_i^*=1} x_i^T \mu_1 - \sum_{i:y_i^*=1} \mu_1^T x_i}_{} + \sum_{i:y_i^*=1} \mu_1^T \mu_1$$

$$\underbrace{- \frac{N}{2}\mu_1^T \mu_1 - \frac{N}{2}\mu_1^T \mu_1 + \frac{N}{2}\mu_1^T \mu_1}_{}$$

$$- N\mu_1^T \mu_1 + \frac{N}{2}\mu_1^T \mu_1 = -\frac{N}{2}\mu_1^T$$

$$= \sum_{i: y_i^* = 1} x_i^T x_i - \frac{N}{2} \mu_1^T + \sum_{i: y_i^* = (-1)} x_i^T x_i - \frac{N}{2} \mu_{-1}^T$$

$$- N \mu_1^T \mu_{-1}$$

$$= \sum_{i=1}^{n} x_i^T x_i - N \underbrace{\left(\frac{1}{2} \mu_1^T \mu_1 + \frac{1}{2} \mu_{-1}^T \mu_{-1} + \mu_1^T \mu_{-1} \right)}$$

$$= \frac{1}{2} \underbrace{\left(\mu_1^T \mu_1 + \mu_{-1}^T \mu_{-1} + 2 \mu_1^T \mu_{-1} \right)}$$

$$\underset{\text{da vllen}}{=} (\mu_1 + \mu_{-1})^T (\mu_1 + \mu_{-1})$$

$$= \sum_{i=1}^{n} x_i^T x_i - \frac{N}{2} (\underbrace{\mu_1 + \mu_{-1}}_{= 0})^T (\mu_1 + \mu_{-1})$$

Therefore equation is true:

$$\Rightarrow \quad \Sigma \cdot \beta + \overbrace{\frac{1}{4}(\mu_1 - \mu_{-1})^T \cdot (\mu_1 - \mu_{-1})}^{:= \tau'} \cdot \beta = \frac{1}{2}(\mu_1 - \mu_{-1})^T$$

$$\Leftrightarrow \quad \Sigma \cdot \beta = \underbrace{\left(\frac{1}{2} - \frac{\tau'}{4} \right)}_{:= \tau} (\mu_1 - \mu_{-1})^T$$

$$\Leftrightarrow \quad \beta_{OLS} = \Sigma^{-1} \tau (\mu_1 - \mu_{-1})^T$$

✓